Using visual analytics to make sense of railway Close Calls

Miguel Figueres-Esteban, Peter Hughes and Coen Van Gulijk

Institute of Railway Research, University of Huddersfield

Abstract

In the big data era large and complex data sets will exceed scientists' capacity to make sense of them in the traditional way. New approaches in data analysis, supported by computer science, will be necessary to address the problems that emerge with the rise of big data. The analysis of the Close Call database, which is a text-based database for near-miss reporting on the GB railways provides a test case. The traditional analysis of Close Calls is time-consuming and prone to differences in interpretation. This paper investigates the use of visual analytics techniques, based on network text analysis, to conduct data analysis and extract safety knowledge from 500 randomly selected Close Call records relating to worker slips, trips and falls. The results demonstrate a straightforward, yet effective, way to identify hazardous conditions without having to read each report individually. This opens up new ways to perform data analysis in safety science.

Keywords

Close call; visual analytics; railway safety; risk analysis; network text analysis

Introduction

Great Britain is accelerating the digital modernisation of technological railway systems that produce massive amounts of data [1–3]. Ambitious programs as ORBIS (to address complex data assets) or SMIS+ (to develop an on-line system for safety management) bring a change in railway organisations to improve the acquisition, storage and use of asset information for a truly digitised railway.

This data contains information that may be relevant for safety and risk in the GB railways. The Institute of Railway Research at University of Huddersfield attempts to use this data to improve safety and risk management in the Big Data Risk Analysis (BDRA) program [4]. BDRA investigates big data analytics techniques for efficient methods of risk management in the future. Moving to BDRA is not simply a matter of scaling-up existing analysis techniques. BDRA has to coordinate and combine a wide range of sources with different types of data and accuracy, and that is not straight-forward.

This paper treats a specific challenge of the BDRA program: extracting safety knowledge from the GB railway's Close Call database, that is, obtaining information from unstructured text-based data. The approach is to work with a sample of Close Call records to explore the potential benefits and shortcomings of visual analytics of word-nets.

Theoretical background

The benefits of analysing near-misses have been proved in different industries for safety management [5–7]. The GB railways wish to exploit the potential of near-miss reporting with the Close Call system. Workers in the GB railway industry can report concerns about hazardous situations by entering free-text descriptions of hazards into the Close Call database. This reporting method provides the freedom for workers to describe any safety concern. With more than 150,000 Close Calls being reported each year it is a success from the reporting culture point of view but analysing them all is no easy feat. The task is complex in the sense that close calls are related to a vast number of different parts of a railway system. It is also time-consuming to extract the critical information from this large body of data. Currently, the analysis of the database relies on expert knowledge from safety analysts [8], but this approach has its limits: a single analyst can unintentionally introduce biases into the results [9, 10]; different analysts may have very different skills and knowledge about the domain and, depending on their risk perception and understanding of goals, the interaction with the system and the way to represent and communicate the results might be different for each user. A consolidated work process that is assisted by visual analytics could circumvent such shortcomings [11] and speed up the process.

Visual Analytics

The term '*visual analytics'* (VA) arose around 2005, being defined as a combination of "*…automated analysis techniques with interactive visualisations for effective understanding, reasoning and decision making on the basis of very large and complex data sets"* [12]. By definition, VA shares many similarities with big data analytics, and it might be said that VA is a variant of 'big data analytics' supported by interactive visualisation techniques [13]. VA encompasses five mature disciplines: data management; data analysis;

human-computer interaction; dissemination and communication; and information/scientific visualisation to enhance the analysis and discovery of information from data (e.g. recognition of distributions, patterns or trends). VA engages creative interpretations in humans beyond those that a computer can detect automatically. It is also a valuable tool for reducing the subjectivity that can occur within groups of users, and can ultimately lead to better decision-making [11, 14, 15]. In this paper, we are interested in the use of VA to perform data analysis and enable interactive learning.

Network text analysis

Three different computer-based text analysis approaches are possible for retrieve information from text: thematic, semantic and networks [16]. Thematic analysis has been the main approach for a long time and it is based on the frequency of concepts (e.g. words or "bag of words") that allows classification of the topics of texts. Semantic analysis also takes into account the relationships among the concepts encoding the semantic grammar (e.g. subject, verb and object). Network analysis is based on network text analysis to obtain semantically linked concepts.

Network text analysis is a method that represents text as a graph: the words or concepts are the nodes, and their relationships are the edges [17–19]. This analysis provides a richer analysis than word frequency analysis, since it is possible to analyse the strength of relationships among main concepts from a text. Paranyunshkin demonstrated the benefits of this technique in text analysis. His work shows that the best results were found when the normalised text is presented as a graph using a context window of two and five words. That is, considering the relationships between the words within a window of two words in a first stage and a window of five words in a second stage, and using network analysis for detecting contextual clusters and key concepts that are junctions for meaning within a text [19]. The network analysis was done using the measures of *degree* and *betweenness of a node.* The degree is the number of edges connecting a node [20]. The betweenness is a centrality measure that is the frequency with which a node falls between pairs of other nodes on the shortest paths connecting them [21]. If these concepts were translated into a map, the degree would be the number of entrances or exits to a city, and the betweenness would be the frequency with which a city falls on the shortest route between two other cities. The number of entrances to a city would be a reference of the importance of that city (a concept in text analysis), and the betweenness would be the value of the city in connecting other places (words that connect and therefore belong to different contexts). These attributes form the backbone of the analysis in this paper.

Methodology

A data set of 500 records was constructed selecting a random sample of 12,171 Slip, Trip & Fall Close Call records. These records were pre-processed using the NLTK toolkit in Python [22] in order to eliminate anomalies that could obscure the text analysis [23]. The records were *cleansed* of stopwords (e.g. *a, an, and* or *the), tokenised* and *tagged* as described in Hughes et al. to create the text for visualising [8]. The *tokenisation process* is based on a non-standard lightweight ontology in the form of list of terms that creates unique tokens from multi-words related to railway safety. The ontology is non-standard in the sense that it was custom-made by railway safety experience of the authors since standard linguistic ontologies were unable to deal with jargon. For example, essential multi-words *TRIPPING_HAZARD_* and *NETWORK_RAIL* were hard to find with standard linguistic ontologies. Table 1 demonstrates a *NETWORK_RAIL_* instance that normal parsers could not capture. The *tagging process* condenses information into tags such name of places, codes, numbers or measured entities. For example, name of places and codes are condensed into the tags *GEO_PLACE* and *_CODE_*, see Table 1.

The cleansed text was transformed into a network by creating a word-per-word co-occurrence matrix following the method of Paranyushkin [19] for a context window of two words. The nodes of the matrix are words of the text, tokens or tags. This matrix shows how the nodes of a network are linked into pairs of nodes and it is a common input for visualisation tools. Because we are not considering the direction of the link between words the network is undirected. That means that the co-occurrence matrix counts the frequency of relationships between adjacent words, that is, between the prior and posterior word.

Source record
"A N/R Supervisor called to report while working on VS148 Signal someone has run a power cable going across the middle of the ladder causing a tripping hazard. Location, Penge East, ELR VIR - Aprox 7m 15chain.
Cleansed record
A NETWORK_RAIL Supervisor called to report while working on _CODE_ Signal someone has run a power cable going across the middle of the ladder causing a TRIPPING_HAZARD_. Location, GEO_PLACE, ELR _CODE_ Aprox DISTANCE_TAG.

Table 1. Example of the pre-process of a Close Call record. Non-desired characters are deleted and the *tokenization* and *tagging* process are applied.

Text visualisation

Following Paranyushkin [19] the resulting co-occurrence matrix was the input for the visual representation of the network. Gephi software was selected for the visualization [24]. It is an open graph visualization platform that allows exploratory data analysis by network manipulation in real-time. *Force-directed graph drawing* algorithms from Gephi were used to draw the network. For simplicity, the mature Force Atlas technique was chosen. Through experimentation, we found that the following parameters suited our purposes best: *Inertia*=0.1, *Repulsion*=10000, *Attraction strength*=10, *Maximum displacement*=10, *Autoslab Strength*=80, *Autoslab sensibility*=0.2. Setting these parameters allow us to see the nodes and the links how the Figures 1, 2 and 3 look.

Clustering

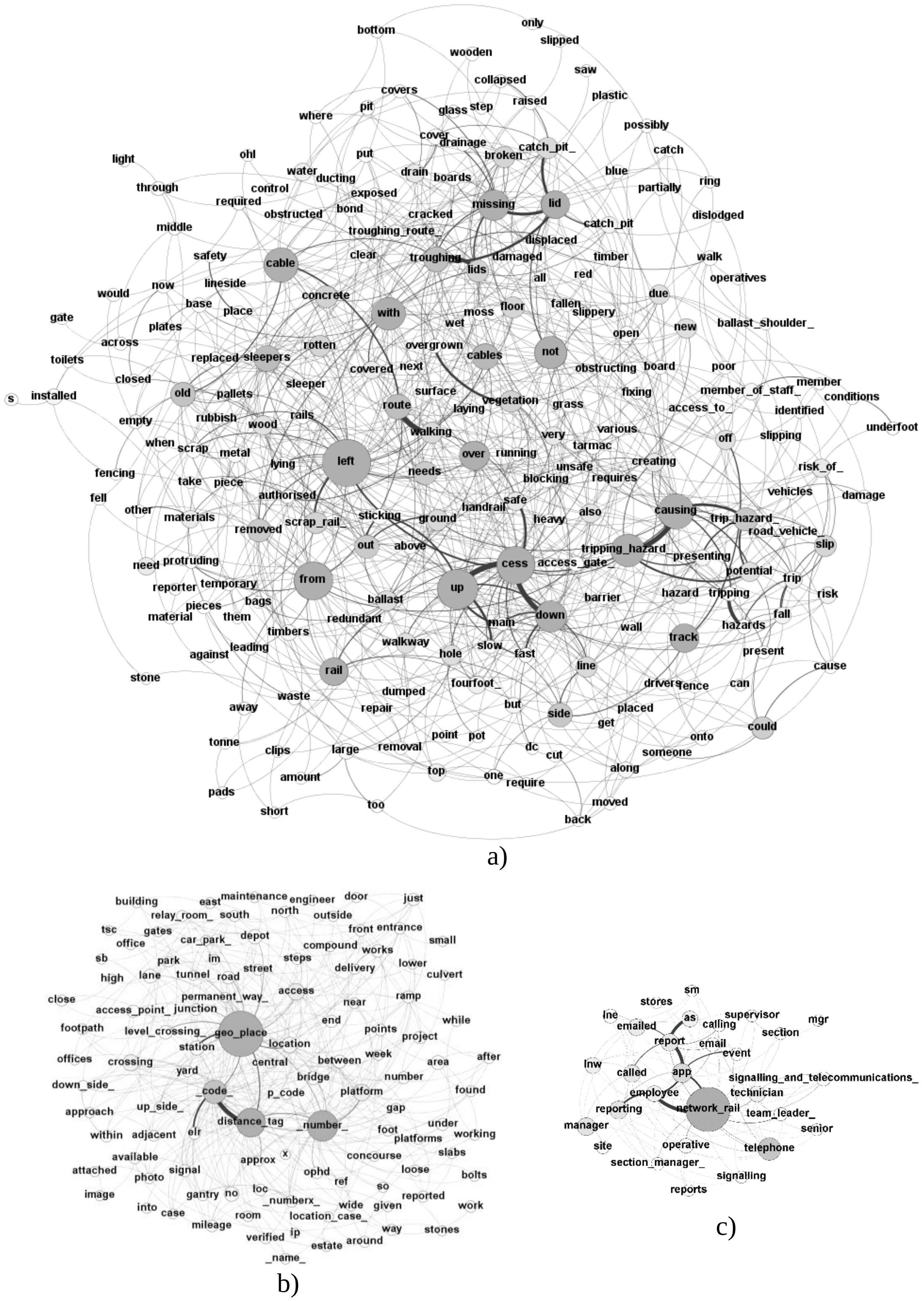
The Louvain Method (LM) for community detection was applied by Paranyushkin to detect contextual clusters in the text network with high accuracy. This clustering method is one of the most popular and it has been used with success in different social science studies in order to discover clusters and zoom within these clusters to discover sub-clusters [25]. It also present good characteristics of scalability, speed and performance [26].

LM extracts clusters of a network based on modularity optimisation. Modularity is a measure of the structure of networks that provide information about the division of a network into modules. High modularity means high density between the nodes of a cluster but sparse connections between different clusters. In plain words, it groups nodes in density areas to create clusters.

LM uses resolution as a sliding ruler to identify clusters but there is no exact answer to what the resolution has to be other than a rule of thumb that it should be higher than 1.0 to identify larger clusters. Above the threshold, higher values of resolution create a few large clusters. Lower values (close to 1.0) create many small clusters that we found to be irrelevant for identifying safety concerns. In this paper, resolutions of 2.0 and 2.5 were found to discover large clusters that capture safety-relevant information [27]. The largest cluster, that contained safety information, was extracted as an independent network in order to detect secondary clusters.

Results

The resulting text network is an undirected graph of 1844 nodes and 5002 edges. The primary clustering with the Louvain method identified three large clusters with a modularity of 0.430 (Figure 2) that represents 98.43% of the network. The secondary clustering yielded four large clusters with a modularity of 0.428 (Figures 3 and 4) that represents 99.15% of the primary safety cluster.



a)

b)

c)

Figure 1. Sub-networks that represent the primary clusters of degree nodes from the cleansed network: a) Safety cluster (56.89%); b) Location cluster (33.35%); c) Staff and method of reporting cluster (8.19%). Resolution=2. Modularity=0.430. Filtered by 5 degree node.

Figure 1 shows the primary clusters. Figure 1.a has high degree nodes such as *missing*, *left*, *cess*, *up*, *down*, *track*, *cable*, *cables*, *sleepers*, *lid*, *troughing*, *not*, *over*, *causing* and *tripping_hazard*. Moreover, the nodes *missing*, *left*, *cable*, *not*, *cess*, *up*, *down*, *over*, *track* and *tripping_hazard* are among the top 25 betweenness nodes. This cluster shows safety hazards. The highest degree and betweenness nodes in Figure 1.b include *geo_place*, *_code_*, *distance_tag* and *number*. In addition, there are numerous medium- and low-degree nodes such as *access*, *platform*, *area*, *depot*, *yard*, *bridge*, *station*, *level_crossing*, *junction*, *permanent_way*, *tunnel*, *car_park* and *relay_room*. This cluster describes locations. Figure 1.c has high-degree nodes such as *network_rail* and *telephone*. Furthermore, it has medium- and low-degree nodes such as *technician*, *employee*, *team_leader*, *signalling_and_telecommunications_*, *operative*, *section_manager*, *manager*, *email*, *calling*, *app*, *reporting* and *report*. This cluster describes people and the method of reporting.

Figures 2 and 3 show the secondary clusters that were derived from the primary safety cluster (Figure 1.a). Figure 2.a contains high-degree nodes such as *left*, *cable*, *rail*, *cess*, *up*, *down*, *route*, *walking* and *sleepers*. In addition, it has medium- and low-degree nodes such as *overgrown*, *vegetation*, *pallets*, *ballast*, *hole*, *tarmac*, *timbers*, *materials*, *rubbish* and *surface*. This cluster shows abandoned objects. Figure 2.b shows a cluster has high- and medium-degree nodes such as *missing*, *lid*, *lids*, *catch_pit*, *troughing*, *wood*, *rotten*, *timber* and *boards*. This cluster shows hazards that arise from missing covers. Figure 2.c has high-degree nodes such as *trip_hazard_*, *tripping_hazard_*, *tripping*, *trip*, *slip*, *slippery*, *fall*, *damage*, *potential*, *hazard* and *risk*. This cluster shows the risk scenario under consideration: slip-trips and falls. Finally, Figure 3 has just three high- medium-degree nodes, *not*, *track* and *member_of_staff*, and many low degree nodes such as *secured*, *protected*, *supported*, *filled*, *happened*, *sited*, *banded*, *member_of_public_*, *inspection*, *inspections*, *obscuring*, *track*, *workers*, *third* and *party*. This cluster refers to procedural errors.

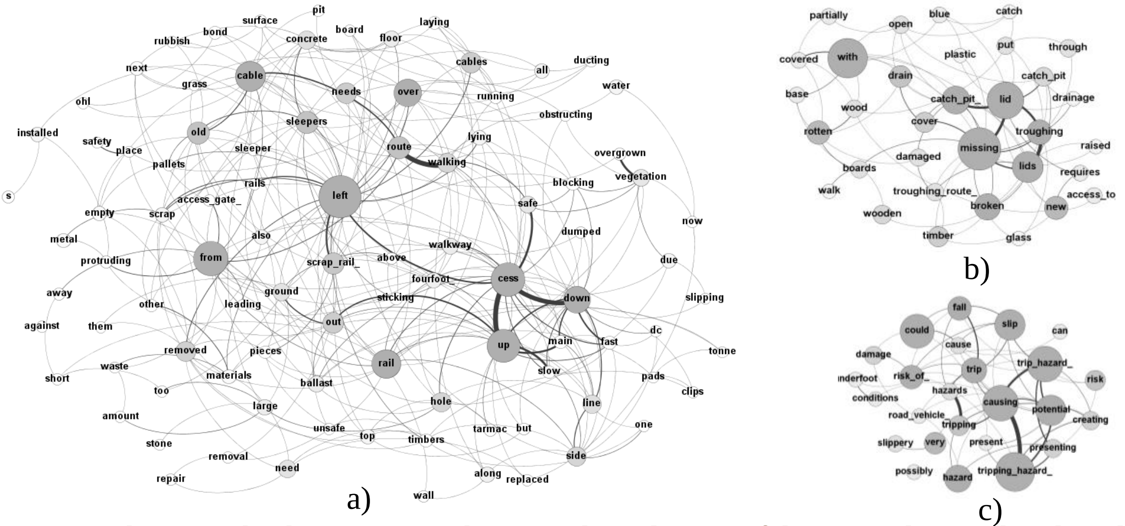


Figure 2. Sub-networks that represent the secondary clusters of degree nodes: a) abandoned objects cluster (50.91%); b) missing covers cluster (18.88%); c) risk scenario cluster (19.16%). Resolution=2.5 Modularity=0.428. Filtered by 5 degree node.

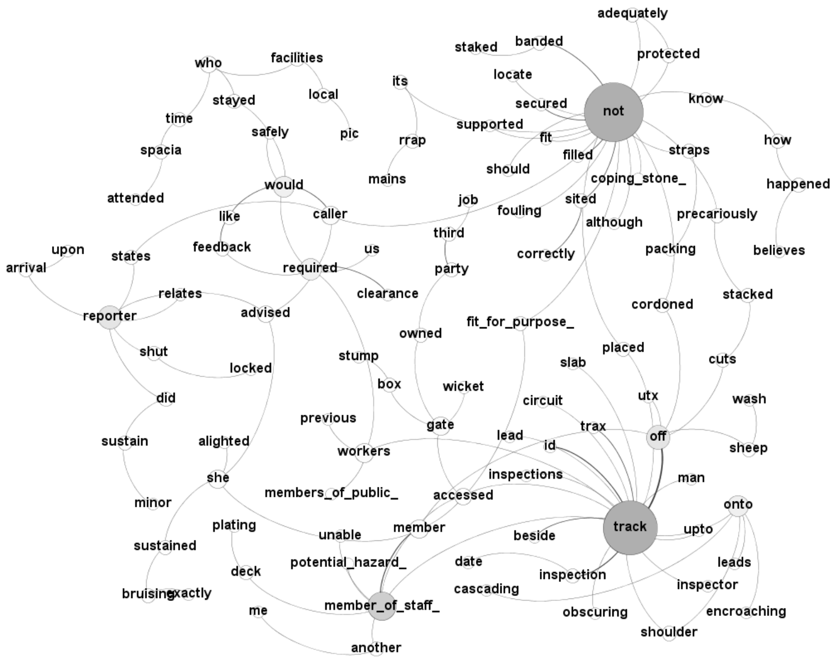


Figure 3. Sub-network that represent a secondary cluster of degree nodes. Procedural errors cluster (10.2%). Resolution=2.5. Modularity=0.428. Not filtered.

Discussion

Text analysis results

The word-net investigation of 500 Close Calls on slip-trip fall incidents identifies a number of issues on the railway: abandoned objects, missing covers and procedural errors. The reported abandoned objects are overwhelmingly leftovers from work on railway systems (scrap rail, sleepers and old cables) along railway tracks (up cess and down cess) but also include: rubbish, overgrown vegetation and scrap. Procedural errors are mostly related to track work procedures. The method identifies concerns that the reporters consider important enough to report about in their working environment. The reports treat hazardous situations during work and since these situations also appear in accident reports the findings may cautiously be considered to be causal factors. In that sense the analysis method can identify safety risks on the railway. However, we observe that the concerns of staff are limited. One notable observation is that natural factors of the environment are absent in these records: weather conditions, night-time and the shape of the cess (e.g. a narrow cess) do not feature in the records. This indicates that Close Call reporters overwhelmingly report man-made factors, which leads to the speculation that that the reporters are mainly focussed on the responsibilities of railway staff for good housekeeping. This may indicate that housekeeping rules are poor but it could also mean that it is an easy way to blame other railway staff and/or colleagues. It also suggests that the Close Call system is not being used to report the full spectrum of causal factors of slip-trip-fall hazards.

VA in network text analysis

Although network analysis is a consolidated area of computing text processing, the visual analytics method that is introduced in this paper is new for railway risk analysis. Since the scope of this exercise was focused on risk identification a certain amount of information loss was accepted during the pre-processing of the text (like geo_place_). This focussed the attention to identifying slip-trip-fall hazards and not on high-hazard hotspots. However, depending on the purpose of the analysis, different types of pre-processing rules could be used to detect different trends and/or anomalies [23]. Thus, the network text analysis has the interesting property of allowing to show data in different perspectives depending on the level of aggregation or disaggregation of the information. With more elaborate tokenisation and tagging rules it might be possible to process specific problems completely automatically by statistical means, thus relinquishing the necessity for visual analytics. Elaborate tokenisation schemes can also elevate text analysis to the field of knowledge representation and sematic networks [28, 29].

After the pre-processing the graphs still show some words that could be considered stopwords (e.g. from and over) or the same concept that represent another node (e.g. lid / lids, cable / cables, trip and tripping_hazard / trip_hazard_). These nodes could be removed in refined cleansing rules but since they did not interfere with the identification of reported hazardous situations they were left in to enable a more rapid identification process.

The analysis was based on the network measures of degree and betweeness. The degree provided information of the importance of the words within the cluster. The betweeness provided information of the relationship of the nodes between the different clusters. That means that nodes with high betweenness (e.g. *geo_place, _code_, distance_tag* and *number)* could be removed in order to integrate large clusters such as the safety cluster and the location cluster, however their inclusion points the way to further analysis that could be used to uncover new findings. These measures have proved to be enough for risk identification, but additional centrality measures such as closeness or eigenvectors could support particular analysis for safety management in railways.

Making sense of Close Calls

For the Railways, the network text analysis of close calls records has highlighted which hazardous conditions railway staff is concerned about when it comes to slip-trip and fall risks on tracks. The rapid analysis helps the railway duty holder to identify major concerns of staff quickly and could even be used to develop automated processes for trend analysis (which was beyond the scope of this paper). For the analysis of the Close Call database it offers a process to speed up the analysis of over 10,000 in for every monthly report. The methods in this paper can be refined to identify low-frequency hazard detection, trend analysis (for instance to monitor the effect of safety programmes), and mixing in alternative text-documents such as accident reports. It seems unlikely that this method can help to detect emergency signals that require immediate attention such as a fridge dumped on the track or a suicide attempt.

For safety science, this method offers an investigation tool to link hazardous situations to accidents. Safety science has struggled with the link between precursors and accidents. Heinrich was the first one to assign a causal relationship between non-injury events and accidents. He considered these non-injury events to be causally linked with accidents since he considered causal factors to be the same [30]. Fault trees can make the causal link between causes and accidents clearer but they do not always work when the precursors are not defined well. Hollnagel is more cautious in assigning causality in the sense that the relation between an observable cause and its observable effect is a non-observable "metaphysical" process [31]. Finally, near misses and Close Calls can be considered in the sense of "weak signals" for safety management [6]. In this school of thought, Close Calls identify areas of trouble in an organization that might indicate that the probability of (particular) accidents is on the rise due to the rise of troublesome preconditions for those accidents. If that perspective is chosen, Close Calls can help to understand the *risk space* to identify areas of trouble but not to match causes to a defined sequence of an accident.

For cognitive sense-making, this paper sheds some light in the relation between cognitive processes and the way data-analysis tools can be supported humans. It is beyond the scope of this paper to provide an overview so we work from Grolemund & Wickham's paper to provide an overview of this research area. They argue that traditional sense-making, where a mental model is constructed in the mind of a single analyst, cannot work with the amount of data that has to be considered in contemporary scientific problems [11]. For safety analysis, this is relevant since it is a multi-disciplinary domain in which many different sources of data have to be considered. Grolemund & Wickham propose that analysts should work from explicit models throughout an investigation. A number of analysts can add their viewpoints and interpretation to develop the model into a substantiated theory based on their interactions. VA is one of these tools that can assist such group efforts by allowing them to apply their perceptual abilities to deal with large quantities of data [14, 32]. This is exactly the way that the authors worked in this investigation.

Conclusions

The paper considers the use of VA to identify potential causal factors in 500 randomly chosen Close Call records about slip-trip-fall hazards. The paper shows VA to be a powerful technique that makes it relatively straightforward to identify hazardous conditions on the railways without having to read each of the 500 Close Calls. This is a very useful property as the number of close calls increases to 150,000 and beyond, although further work would be required to refine the method for such numbers of records. Nonetheless, this approach speeds safety analysis up considerably for large data sets. At the same time, the method is a tool that allows for analysis by groups of investigators rather than a single one which should reduce interpretation bias. The method can also be developed further for low-frequency risk detection automated trend analysis, and mixing different text-based data-sources.

This work paves the way to model-assisted sense-making that enables the analysis of huge amounts of data that cannot realistically be handled by manual analysis. Close Call analysis for the GB Railway benefits from a quicker analysis. The technique is not limited to Close Call reports, the method can be tuned for any text-based source such as accident reports, safety procedures and standards and accident reports in any industry.

Acknowledgements

The work reported in this paper was undertaken as part of the Strategic Partnership agreement between the University of Huddersfield and RSSB.

References

[1] CCS. http://www.closecallsystem.co.uk/maximo/webclient/login/login.jsp?welcome=true (2015, accessed 5 June 2015).

[2] ATOC. http://www.atoc.org/about-atoc/national-rail-enquiries/access-to-information (2015, accessed 5 June 2015).

[3] NR. http://www.networkrail.co.uk/data-feeds (2015, accessed 5 June 2015).

[4] Van Gulijk C, Hughes P, Figueres-Esteban M, et al. Big Data Risk Analysis for Rail Safety? In: Podofillini L, Sudret B, Stojadinovic B, et al. (eds) *Safety and Reliability of Complex Engineered Systems*. London: Taylor & Francis Group, 2015, pp. 643–650.

[5] Bliss JP, Rice S, Hunt G, et al. What are close calls? A proposed taxonomy to inform risk communication research. *Saf Sci* 2014; 61: 21–28.

[6] Gnoni MG, Lettera G. Near-miss management systems: A methodological comparison. *J Loss Prev Process Ind* 2012; 25: 609–616.

[7] Macrae C. *Close calls: managing risk and resilience in airline flight safety*. Palgrave Macmillan UK, 2014.

[8] Hughes P, Van Gulijk C, Figueres-Esteban M. Learning from text-based close call data. In: Podofillini L, Sudret B, Stojadinovic B, et al. (eds) *Safety and Reliability of Complex Engineered Systems*. London: Taylor & Francis Group, 2015, pp. 31–38.

[9] Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. *Science* 1974; 185: 1124–1131.

[10] Tversky A, Kahneman D. The framing of decisions and the psychology of choice. *Science (80-)* 1981; 211: 453–458.

[11] Grolemund G, Wickham H. A Cognitive Interpretation of Data Analysis. *Int Stat Rev* 2014; 82: 184–204.

[12] Keim D, Andrienko G, Fekete J-D, et al. Visual Analytics: Definition, Process, and Challenges. In: Kerren A, Stasko JT, Fekete J-D, et al. (eds) *Information Visualization: Human-Centered Issues and Perspectives*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 154–175.

[13] Figueres-Esteban M, Van Gulijk C, Hughes P. *Visualisation and Risk Communication in Railway Big Data Risk Analysis (BDRA): Literature Review. Report 110-113*. 2015.

[14] Card SK, Mackinlay JD, Shneiderman B. *Readings in Information Visualization: Using Vision to Think*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1999.

[15] Tory M, Möller T. Human factors in visualization research. *IEEE Trans Vis Comput Graph* 2004; 10: 72–84.

[16] Popping R. *Computer-Assisted Text Analysis*. London: SAGE Publications Ltd, 2000. Epub ahead of print 2000. DOI: 10.4135/9781849208741.

[17] Drieger P. Semantic network analysis as a method for visual text analytics. *Procedia - Soc Behav Sci* 2013; 79: 4–17.

[18] Popping R. Knowledge Graphs and Network Text Analysis. *Social Science Information* 2003; 42: 91–106.

[19] Paranyushkin D. Identifying the Pathways for Meaning Circulation using Text Network Analysis. *Nodus Labs* 2011; 26.

[20] Newman M. *Networks an Introduction*. Epub ahead of print 2010. DOI: 10.1093/acprof:oso/9780199206650.001.0001.

[21] Freeman LC. Centrality in social networks conceptual clarification. *Social Networks* 1978; 1: 215–239.

[22] Bird S, Klein E, Loper E. *Natural Language Processing with Python*. O'Reilly Media, Inc., 2009. Epub ahead of print 2009. DOI: 10.1097/00004770-200204000-00018.

[23] Diesner J, Carley KM. Using Network Text Analysis to Detect the Organizational Structure of Covert Networks. In: *Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference*. 2004.

[24] Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third Int AAAI Conf Weblogs Soc Media* 2009; 361–362.

[25] Labatut V, Dugué N, Perez A. Identifying the community roles of social capitalists in the twitter network. In: *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, 2014, pp. 371–374.

[26] Papadopoulos S, Kompatsiaris Y, Vakali A, et al. Community detection in Social Media. *Data Min Knowl Discov* 2012; 24: 515–554.

[27] Blondel VD, Guillaume J, Lambiotte R, et al. Fast unfolding of community hierarchies in large networks. *Networks* 2008; 1–6.

[28] Walker GH, Stanton NA, Salmon PM. Cognitive compatibility of motorcyclists and car drivers. *Accid Anal Prev* 2011; 43: 878–888.

[29] Houghton RJ, Baber C, Cowton M, et al. WESTT (workload, error, situational awareness, time and teamwork): An analytical prototyping system for command and control. *Cogn Technol Work* 2008; 10: 199–207.

[30] Heinrich H. W. *Industrial accident prevention*. McGraw Hill New York, 1931.

[31] Hollnagel E. *Barriers and Accident Prevention*. Ashgate UK, 2004.

[32] Cleveland WS, McGill R. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *J Am Stat Assoc* 1984; 79: 531–554.