Safety Science 89 (2016) 72–76

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/ssci

Visual analytics for text-based railway incident reports

Miguel Figueres-Esteban *, Peter Hughes, Coen van Gulijk

University of Huddersfield, Institute of Railway Research, Queensgate, Huddersfield, UK

ARTICLE INFO

Article history:
Received 25 November 2015
Received in revised form 23 May 2016
Accepted 23 May 2016
Available online 11 June 2016

Keywords:
Close call
Visual analytics
Railway safety
Risk analysis
Network text analysis

ABSTRACT

The GB railways collect about 150,000 text-based records each year on potentially dangerous events and the numbers are on the increase in the Close Call System. The huge volume of text requires considerable human effort to its interpretation. This work focuses on visual text analysis techniques of Close Call records to extract safety lessons more quickly and efficiently. This paper treats basic steps for visual text analysis based on an evaluation test using a pre-constructed test set of 150 Close Call records for "Trespass", "Slip/Trip hazards on site" and "Level crossing". The results demonstrate that visual text analysis can be used to identify the risks in a small-scale test set but differences in language use by different cohorts of people interferes with straightforward risk identification in larger sets. This work paves the way to machine-assisted interpretation of text-based safety records which can speed up risk identification in a large corpus of text. It also demonstrates how new possibilities open up to develop interactive visualisations tools that allow data analysts to use text analysis techniques for risk analysis.

1. Introduction

The benefits of analysing Close Call/near misses reports have been proved in many industries (Bliss et al., 2014; Gnoni and Lettera, 2012; Macrae, 2014). In the GB Railways, two systems are in operation today to exploit these benefits: in the Close Call System (CCS) and the Confidential Incident Reporting and Analysis System (CIRAS).

The Close Call System collects about 150,000 text-based records each year on potentially dangerous events and the numbers are on the increase. The huge volume of text from Close Call records requires considerable human effort and time to its interpretation. Computer-assisted Text Analysis (TA) provides alternative techniques that can facilitate the extraction of safety knowledge and reduce the human effort. Three fundamentally different approaches can be found for TA: thematic, semantic and networks (Popping, 2000). Network analysis is in the emerging field of Visual Analytics (VA). VA combine automated data analysis techniques from massive, inconsistent and conflicting data with human knowledge by means of interactive visualisations for an effective understanding, reasoning and decision making (Keim et al., 2010, 2008; Thomas and Cook, 2005). This paper describes the initial steps for using VA techniques and demonstrates a way forward to develop interactive visualisations tools but also demonstrates some of the difficulties on the way ahead.

It is beyond the scope of this paper to analyse the benefits of different methods that can be used for text analysis (e.g. see Popping (2000) for overview). This paper uses the analysis method proposed by Paranyushkin (2011). It demonstrates the benefits of a method for representing normalised text as a graph and using network analysis for detecting contextual clusters and key concepts that are junctions for meaning within a text. This VA approach suits the aim of this work. It is used to support interpretation of large amounts of text by graphical representation techniques that reduce the analyst's workload (Crow et al., 1994). The method is based on visual text analysis by means of graphs: terms (words and multi words) are nodes, and their relationships are links in word based graph networks (Drieger, 2013; Paranyushkin, 2011; Popping, 2003). This way of working allows analysis of the type and strength of relationships between the main concepts from a text, and thus, allows information extraction from the graph. To date, no references about using this technique in safety science were found.

2. Methodology

Although it is desirable to analyse all Close Calls in one go, we believe that there are many obstacles that have to be addressed before this is possible. This paper explores the basic principles by analysing a sample of Close Call records that describes three risk scenarios in order to identify them. A pre-constructed dataset of 150 records was constructed by selecting the first 50 records from the Close Call database classified as "Trespass", "Slip/Trip hazards

* Corresponding author.
E-mail address: m.figueres@hud.ac.uk (M. Figueres-Esteban).

http://dx.doi.org/10.1016/j.ssci.2016.05.009
0925-7535/

Table 1
Example of cleaned text, tagged text and tokenised text. The latter is used for the analysis.

Cleaned record
Emailed report from LOM Date: 08/09/13 Time: 1900 ELR: LEN3 59m 14ch Issue – Trespasser on the line in the Hartburn Junction area. Trains cautioned, reported all clear by MOM @ 1930 Action – Fencing to be checked 09/09/13 DU: Newcastle
Cleaned and tagged record in lowercase
Emailed report from local operation manager date _date_ time _time_ elr_code distance_tag issue trespasser on the railway line in the geo_place junction area trains cautioned reported all clear by mobile operations manager _time_ action fencing to be checked _date_ geo_place
Cleaned, tagged and tokenised record without stopwords and in lowercase
Email report from local operate_ manager_ date _date_ time _time_ elr_code distance_tag issue trespasser on railway_ line_ in geo_place junction_ area_ train_ warning_ reported all clear_ by mobile_operations_manager_ _time_ action fence_ check_ _date_ geo_place

on site" and "Level crossing". These records were cleaned of non-desired characters using the NLTK toolkit in Python (Bird et al., 2009) in order to generate the text source to process (cleaned record in Table 1). The "*tagging process*" and "*tokenisation process*" described in Hughes et al. (2015) was used to create the two sets of text for visualising. The visual analysis of the tagged-text (cleaned and tagged record in lowercase in Table 1) provided information to tailor the tokenisation process (removing main stopwords and stemming plurals or verbs), avoiding obscuring main concepts in the tokenised-text network (cleaned, tagged and tokenised record without stopwords in Table 1).

The final tokenised text is composed of terms that are (1) tags related to places, codes or measured entities (i.e. *geo_place*, *elr_code* and *distance_tag*, respectively), (2) tokens that link relevant adjacent words or represent stem verbs and nouns (e.g. *mobile_operations_manager_*, *check_* or *junction_*) and (3) words from the original text (e.g. *trespasser*).

The final text can be transformed into a network building its adjacency matrix of words (aka word by word co-occurrence matrix). An adjacency matrix shows how the nodes of a graph are connected into pair of nodes and it is the input of visualisation tools. In the evaluation test, the adjacency matrix to visualise is the addition of two matrices: one for a context window of size two and one for a context window of size five. The two-gap context window identifies relevant adjacent words such as *access* and *gate* (Fig. 1.2). The five-gap context window takes into account the proximity of the words that are slightly further apart such as *press* and *button* (the sequence would be *press stop_ button*, Fig. 1.2) but it also amplifies the adjacent words by double counting. Gephi software was the visualisation tool selected for the visual representation of the adjacency matrix. The visualisation was made using the Force Atlas layout with the parameters *Inertia* = 0.1, *Repulsion* = 10,000, *Attraction strength* = 10, *Maximum displacement* = 10, *Autoslab Strength* = 80, *Autoslab sensibility* = 0.2.

In order to gather knowledge from the networks two key centrality measures were analysed, the *degree of a node* and the *betweenness of nodes*. The degree of a node is the number of links connecting a node (Lewis, 2011; Newman, 2010). It is represented by the size of the node in Fig. 1 and is an indicator of the importance of the node (for instance *cross_* in Fig. 1.1 or *barrier_* in Fig. 1.2). The betweenness of nodes is defined by Freeman (1978) as the frequency with which a node falls between pairs of other nodes on the shortest paths connecting them (like the *stop_* in the *press stop_ button* sequence in Fig. 1.2). In the text analysis context, the betweenness gives information about the nodes that connect clusters (Paranyushkin, 2011; Popping, 2000). Thus it provides information about the overlap of clusters as shown in Fig. 2. Although the betweennes cannot be expressed in the Fig. 1, the strongest betweenness is considered in the cluster interpretation.

The Louvain method for community detection was applied to detect clusters in the text network. A resolution of 1.5 was given in order to discover large clusters (Blondel et al., 2008).

3. Results

The resulting text network is an undirected graph of 775 nodes and 16,563 edges. The Louvain method identified four clusters with a modularity of 0.611 (Fig. 1).

The first, second and third clusters have the highest degree nodes with a high betweeness (*cross_, geo_place, distance_tag, location, barrier_, access_*, gate_ and *road_vehicle_*) and contain a great quantity of high and medium degree nodes related to level crossings (*elr_code*, *level_crossing, road, driver_, red_, light_, flash_, warning_, miss_, padlock_, unsecure, point, track, trackside_, lock, open_, enter, safe_* or *authorised*). These three clusters present differences regarding the nodes that represent people and the topics that the higher degree nodes describe. The first cluster encloses nodes related to technical staff (for example *network_rail_, operative or signaller)* and operational railway terms such as *box_, signal_, cctv_, elr_code, cess, main_, delay_, safe_, line_, dn_, up_, platform_, bridge_, station_*or *downside*. The second cluster contains two high degree nodes with high weight that describe the general public (*member_, public_* or *pedestrian_*) and diverse road safety terms such as *road_vehicle, barrier_, light_, red_, descend_, stop_, button*, *press, pass_* or *stopped_*. As with the first cluster, the third cluster shows many nodes related to technical staff (for example *mobile_operations_manager, operational, telecommunications, manager* or *engineer*) and operational work terms such as *close_, call_, access_, gate_, miss_, padlock_, unsecure, point, track, trackside_, lock, open_, enter, control_* or *authorised.*

The fourth cluster displays high degree nodes for example *hazard_, potential_, trespass_ or sliptripfall_*, nodes related to people like *worker_* or *member_of_staff* and terms related to the workforce environment such as *tool_, gap_, wall_, sticking, cable_, fence_, boundary_, overgrown_* or *vegetation_*.

4. Discussion

Four clusters were found from the five-word gap tokenised network using a resolution of 1.5. The choice of the resolution influences how many clusters are determined by the Louvain method for community detection. As a guideline, it is accepted that small values (less than 1) generates too many small clusters to extract sensible learning from the data. A value greater than 1 means fewer clusters are created but they tend to be larger in the sense that there are more nodes in a cluster. As we are interested in identifying three risk clusters, a value of 1.5 was used.

The resulting clusters have a modularity of 0.611. According to Paranyushkin (2011) this is higher than the threshold value of 0.4 to indicate stable clusters. Stable means that this is an allowed use of the modularity algorithm. De facto, the connectivity of the terms within a cluster is higher than with other clusters in the network.

The graphs still show some words that could be considered stopwords (e.g. *that, could* or *which*). This is a shortcoming of the method used in this paper. The words identified in the graphs could be re-evaluated and made part of the text cleaning rules in

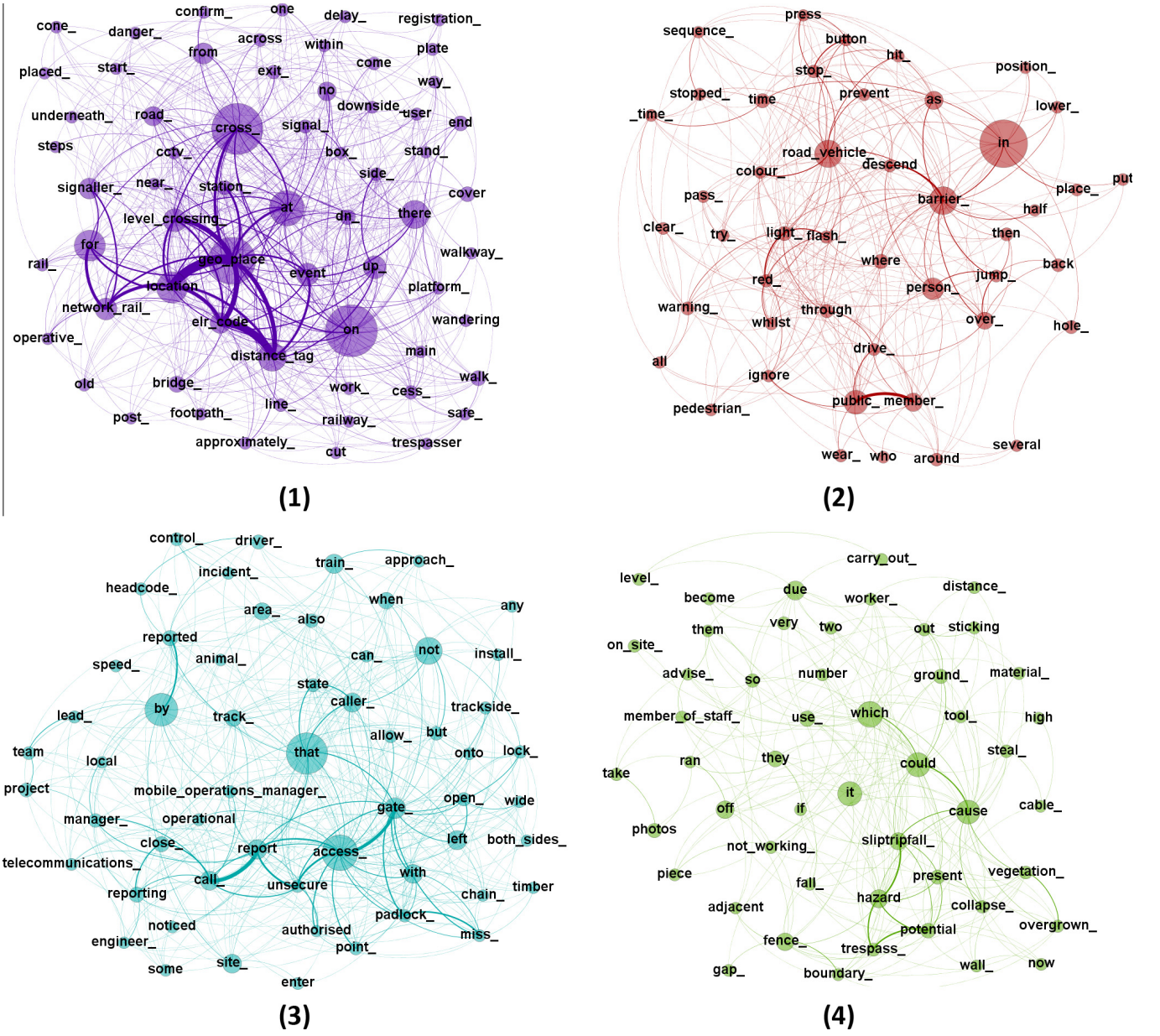

Fig. 1. Sub-networks that represent the clusters of degree nodes (1 = 29.49%, 2 = 21.2%, 3 = 26.27% and 4 = 23.04% of nodes from left to right) from the five-word gap – tokenised network (resolution = 1.5; modularity = 0.611). Filtered by 20 degree node.

a second iteration for the creation of the adjacency matrix. How many iterations would be optimal for cleaning the text is context dependent and is beyond the scope of this paper.

4.1. Word clusters vs risk scenarios

The clustering method looks for relationships amongst written terms that describe *level crossing, trespasses* and *slips, trips and fall* safety scenarios. It is used to identify groups of nodes (called clusters or word-clusters) that are interconnected. The word-clusters do not return the three risk scenarios as originally expected but different uses of language by different cohorts of staff is mixed into the equation. The clearest language distinction is between technical railway staff and non-technical staff or lay-people. Technical staff tends to use operational railway terms, which yields high-degree nodes in technical vernacular. Non-technical staff describe risks in their own words but tend to be less precise about the hazard they are describing. This means that VA machine-based risk identification from text requires additional interpretation by a human analyst. As humans we can get a lot of meaning beyond the literal content of text and we have no problem in inferring context that text describes. However, as easy and obvious as this task is for humans, it is difficult for machines. This finding suggests that it is likely that risk identification by VA is sped up but it cannot be fully automated without human guidance.

The findings in the four clusters are interpreted as follows. In the first, second and third clusters the high degree nodes (key concepts) are mainly related to level crossings terms (*level_crossing_*, *barrier_*, *road_vehicle_*, *access_* and *gate_*). Moreover, these nodes have a high betweeness value, indicating that the three clusters are closely connected. The difference probably arises from different cohorts of people who are entering records. Thus, what we are visualising is the way that different people describe a level crossing scenario rather than identifying the level crossing scenario itself. The first and third cluster show more technical and operational railway terms that technical staff (e.g. *signaller_*, *manager_* or *engineer*)

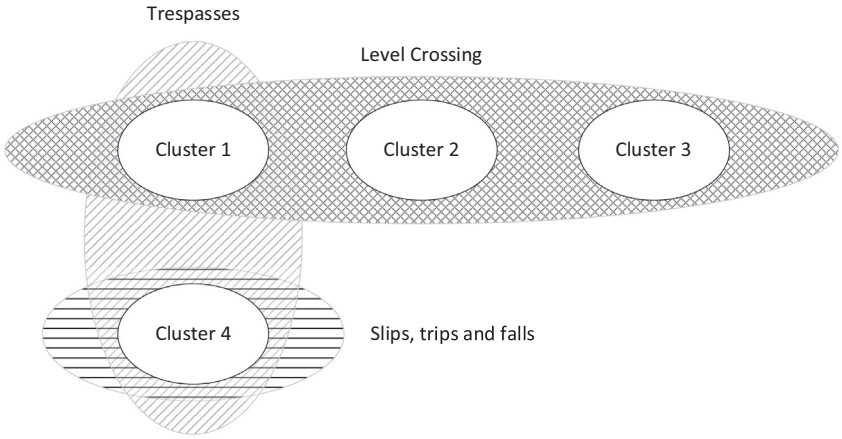

Fig. 2. Mapping of safety scenarios with network clusters. Three clusters are related to level crossings, the fourth cluster is related to STF events and the trespasses events are overlapped with a level crossing cluster and the STF cluster.

use for reporting (e.g. *elr_code, box_, signal_, main_, line_, up_, dn_, cctv_, track_side, authorised, unsecure, lock, miss_* or *padlock_*) whilst the second suggest that the records where the general public are involved are related to road safety issues (*road_vehicle_, pedestrian, ignore, warning_, red_, flash_, light_* or *press, stop_ button*). Moreover, the first cluster includes a medium degree node associated with trespasses (*trespasser*). This finding might mean that the terms used to describe a level crossing may also be commonly used to describe trespasses (*cross_, geo_place, location, station_ or platform_*).

The fourth cluster shows very high degree nodes related to trespass and STF events (*trespass_ and sliptripfall_*). It is theorised that these records are made by track workforce (*worker_* or *member_of_staff*) who mainly report about STFs. Moreover, the terms reported may be associated with work activities (e.g. *tool_, gap_, fence_, wall_, sticking, collapse_, boundary_, fall_, overgrown_* or *vegetation_*). The high degree nodes related to trespasses might mean that workers are the ones who most commonly report trespasses. This finding may also explain why the first cluster, also related to technical staff, contains nodes about trespasses: they report about both level crossing and trespass risks scenarios.

Thus, although words which may be seem to be inherent to a one type of risk scenario these words are also used to describe other scenarios. People essentially use the same words to describe different types of scenarios, the risk categories have been not completely lost in the analysis and it is possible to map the clusters and the risk scenarios. However, it takes additional interpretation from a risk analyst to extract risk scenarios from the data (Fig. 2).

4.2. Close Call vs risk scenarios in Railways

The GB railways are currently representing risk scenarios by means of bow-ties diagrams. A bow-tie diagram represents causes and consequences of potential accidents (Hudson, 2010), and can be created from the Safety Risk Model (SRM) that consist of a set of fault and event trees for hazardous events (Marsh and Bearfield, 2008).

This work suggests that people creating Close Call records are not familiar with the nomenclature and formal classification of risks: people do not use the definitions or terms used in the SRM model. This means that it may be challenging for automated data analysis (e.g. machine learning or clustering techniques) to map different risks automatically from the free text in Close Call System to SRM models. Some form of human interpretation may always be needed. Despite that, it is expected that visual analytics will speed up the analysis process greatly.

As an alternative to pre-determined risk scenarios and bow-ties, visual analytics techniques can guide the building of alternative scenario descriptions: scenarios that are purely based on the perception of the cohort of reporters. This would represent perceived risks and allows for a completely new perspective on risks on the rails.

4.3. Visual analytics for network text analysis

Thomas and Cook (2005) state that text analysis should be able *"to infuse visual analytics data representations with semantic richness"*. But this work shows that even with open-access tools such NLTK and Gephi is not straightforward and requires significant contextual safety knowledge by the analyst using the tools. However, if such an analyst is trained in the use of such tools the Gephi tool provides essential interactions for the visual analysis of risk-text. More interactivity with the NLP techniques should be developed to support visual text analysis, such as improving the cleaning of the text and updating the list of stopwords removing selected nodes during the visual text analysis, or creating new tokens linking nodes directly from the network.

4.4. Consolidation and future steps

It was expected that the input of three types of events would produce three clusters, instead four were found. This unexpected finding was achieved even though a small set of records that were preprocessed to avoid obscuring main concepts. The difference in number of clusters has arisen from the different reporting styles of authors. This finding demonstrates the power of the technique to not only obtain information on the event, but on who was reporting it, which in turn gives some context to the report beyond what is actually written. This result is consistent with the research based on security and social structure networks where the interaction between people and their language is analysed (Diesner and Carley, 2004; Diesner, 2013). However, it needs to be addressed and resolved before we can go to larger numbers of text records because this research focuses on identifying risks without such interference.

Though the machine helps to speed up text interpretation, the human remains vital in the process. Firstly, during the tagging and tokenisation process analysts intend to condensate nodes that could obscure the analysis, that is, to detect anomalies in the text that could complicate the finding of risk scenarios (Diesner and Carley, 2004). For example, in the pre-constructed data set, information such as places, codes and measures was considered

superfluous to the analysis. Once this process is performed, it requires little human intervention and can be fully automated.

Secondly, human interpretation is required to guide the automated cluster generation process. The visual analysis of the clusters enable a human to interpret whether risk scenario's appear twice or whether risk scenarios are mixed in a single cluster. Also, Close Call records can contain details that are not relevant for the identification of risk scenarios, such as job titles, that could create clusters that do not reflect risks but jobs. Inclusion of this information could be obscuring the meaning of the records for detecting the risk scenarios. That means that much of the classification can be done automatically but a human is required to check whether the results make sense. Note that the anomalies like job titles can be coded and washed out in the cleaning process to improve the speed and the accuracy of the risk identification process. The fact that the human plays an important role in machine-assisted learning was clearly identified by Keel (2006) and this work reinforces that view.

The acquired knowledge from the application of VA for text analysis opens noticeable new research line for safety science. The VA technique described in this work has demonstrated pathways quicker analysis of text bodies but also demonstrates the need for human interaction to obtain safety knowledge. It is thought that the human interaction can be introduced into computer systems for improving future automated analysis of big data sets. A common technique to describe this knowledge in machine-readable representations is to build ontologies of that knowledge domain. The network text analysis provides a suitable environment to support the building of ontologies from alternative sources such as Close Call data. In particular, VA shows very good benefits in order to extract terms and relationships that would be "atoms of knowledge" of the ontologies. These ontologies would be the way to support interactive machine learning techniques that could improve the detection and classifications of Close Call records.

5. Conclusion

This paper discusses a new research framework where VA analytics techniques are applied in the safety and risk domain to obtain insight from unstructured text data. The results demonstrate that mature network analysis techniques for text analysis can be applied to safety-relevant text and compete with purely human based analysis. VA methods cannot be used without additional interpretation by a risk specialist but their work load could be significantly reduced (the alternative being to manually read all reports). In this work, the graph analysis based on text networks from close call records demonstrates that visual text analysis mixes the way in which people use language with pre-defined risk scenarios. This is mostly due to the fact that most workers do not know the formal risk models and its associated jargon. It is expected that this language use could be codified in ontologies to bypass this shortcoming.

This work paves the way to machine-assisted interpretation of text-based safety records which can speed up risk identification in a large corpus of text. It also demonstrates how new possibilities open up to develop interactive visualisations tools that allow data analysts to use text analysis techniques for risk analysis.

Acknowledgements

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

References

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python. O'Reilly Media Inc.

Bliss, J.P., Rice, S., Hunt, G., Geels, K., 2014. What are close calls? A proposed taxonomy to inform risk communication research. Saf. Sci. 61, 21–28. http://dx.doi.org/10.1016/j.ssci.2013.07.010.

Blondel, V.D., Guillaume, J., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of community hierarchies in large networks. Networks 1–6. http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

Crow, V., Pottier, M., Thomas, J., 1994. Multidimensional Visualization and Browsing for Intelligence Analysis. Pacific Northwest Lab, Richland, WA (United States).

Diesner, J., 2013. From texts to networks: detecting and managing the impact of methodological choices for extracting network data from text data. KI – Künstliche Intelligenz 27, 75–78. http://dx.doi.org/10.1007/s13218-012-0225-0.

Diesner, J., Carley, K.M., 2004. 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.

Drieger, P., 2013. Semantic network analysis as a method for visual text analytics. Procedia – Soc. Behav. Sci. 79, 4–17. http://dx.doi.org/10.1016/j.sbspro.2013.05.053.

Freeman, L.C., 1978. Centrality in social networks conceptual clarification. Soc. Networks. http://dx.doi.org/10.1016/0378-8733(78)90021-7.

Gnoni, M.G., Lettera, G., 2012. Near-miss management systems: a methodological comparison. J. Loss Prev. Process Ind. 25, 609–616. http://dx.doi.org/10.1016/j.jlp.2012.01.005.

Hudson, P.T., 2010. Integrating organisational culture into incident analyses: extending the bow tie model. In: SPE International Conference on Health, Safety and Environment in Oil and Gas Exploration and Production. Society of Petroleum Engineers.

Hughes, P., Van Gulijk, C., Figueres-Esteban, M., 2015. Learning from text-based close call data. Proceedings of the 25th European Safety and Reliability Conference. ESREL, p. 8.

Keel, P.E., 2006. Collaborative visual analytics: inferring from the spatial organization and collaborative use of information. In: IEEE Symp. Vis. Anal. Sci. Technol. 2006, VAST 2006 – Proc., pp. 137–144. http://dx.doi.org/10.1109/VAST.2006.261415.

Keim, D., Andrienko, G., Fekete, J.D., Görg, C., Kohlhammer, J., Melançon, G., 2008. Visual analytics: definition, process, and challenges. In: Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 154–175. http://dx.doi.org/10.1007/978-3-540-70956-5_7.

Keim, D., Kohlhammer, J., Ellis, G., Mansmann, F., 2010. In: Mastering the Information Age Solving Problems with Visual Analytics, 57–86.

Lewis, T.G., 2011. Network Science: Theory and Applications. John Wiley & Sons.

Macrae, C., 2014. Close calls: Managing Risk and Resilience in Airline Flight Safety. Palgrave Macmillan.

Marsh, D.W.R., Bearfield, G., 2008. Generalizing event trees using Bayesian networks. Proc. Inst. Mech. Eng. Part O J. Risk Reliab. 222, 105–114. http://dx.doi.org/10.1243/1748006XJRR131.

Newman, M., 2010. Networks an Introduction. Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199206650.001.0001.

Paranyushkin, D., 2011. Identifying the pathways for meaning circulation using text network analysis. Nodus Labs, 1–26.

Popping, R., 2003. Knowledge graphs and network text analysis. Soc. Sci. Inf. http://dx.doi.org/10.1177/0539018403042001798.

Popping, R., 2000. Computer-Assisted Text Analysis. New Technologies for Social Research Series. SAGE Publi. http://dx.doi.org/10.4135/9781849208741.

Thomas, J.J., Cook, K.A., 2005. Illuminating the path: the research and development agenda for visual analytics. IEEE Computer Society.