Mining Causal Relations and Concepts in Maritime Accidents Investigation Reports.

Santosh Tirunagari, Maria Hänninen, Kaarle Ståhlberg, and **Pentti Kujala.**
Department of Applied Mechanics, Aalto University, Espoo, Finland.
santosh.tirunagari@aalto.fi

ABSTRACT

Text mining is a process of extracting information of interest, from the text. In here, we applied text mining methods to extract causal patterns from the maritime accident reports collected from the Marine Accident Investigation Branch (MAIB). These causal patterns from the accident reports provide information on various mechanisms behind accidents. These include human and organisational concepts. A careful and manual investigation of causal patterns extracted from the reports provided opportunity to collect a list of concepts present in an accident according to the investigation. In this paper we discuss the statistics of the accidents that are caused by the list of concepts that were collected in this research work and also apply Self Organising Maps for visualization.

NOMENCLATURE

SOM	Self Organising Maps
NLTK	Natural Language Tool kit
MAIB	Marine Accident Investigation Branch
VSM	Vector Space Model

1. INTRODUCTION

Text mining is an emerging technology that can be used to augment existing data in any textual databases by making unstructured text data available for analysis [4]. Text mining is a process of extracting information of interest from the text. Such a method includes techniques from various areas, e.g. Information Retrieval (IR) [12], Natural Language Processing (NLP), and Information Extraction (IE) [8]. In this paper, text mining methods are applied to extract causal patterns from the maritime accident investigation reports collected from the Marine Accident Investigation Branch (MAIB[1]) web pages. MAIB examines and investigates all types of maritime accidents to or on board United Kingdom (UK) ships worldwide, and other ships in UK territorial waters. These reports consist of a narrative and analysis of the circumstances relating to the accident, written in natural language.

According to Grech and Horberry [1], causal patterns are one of the fundamental reasoning that is necessary for decision making. These causal patterns from the accident reports provide information on various mechanisms behind accidents. Unfortunately, in the maritime field no standard reporting formats exist and data collection from the textual reports is a laborious task. Text mining provides a means for efficient and informative scanning of accident cases of interest without reading the actual report. Therefore, text mining in this context is seen as a useful tool in building understanding on accidents and their influencing concepts. Analysis in this paper is limited to groundings, collisions, machinery failures and fire with the objective to enable analyzing maritime accident investigation reports by means of extracting concepts and patterns, describing the human and organizational elements in accidents. A careful investigation of accident reports provides opportunity to improve and manage safety in the future [3].

Louise Francis et al. [6] applied text mining methods on two text databases, a road accident description and survey databases. They extracted new variables from the unstructured text which later used for predicting the likelihood of attorney involvement and the severity of claims. Text mining also identified interesting themes in the responses of the survey data. Thus, useful information that would not otherwise be available was derived from both the databases using text mining methods. Yiu, [9] investigated and validated a novel text mining methodology for occupational accident analysis and prevention. He also suggested that adoption of text mining analysis is probably most feasible for large organisations that can more easily absorb the labour-intensive steps required to conduct the most meaningful text mining analysis of occupational injury data. Another paper by Zheng et al. [10] used a text data mining technique called attribute reduction from accident reports to extract most frequent factors which were considered as the reasons leading to human errors in ship accidents. A paper by Artana et al [4] developed and evaluated software using text mining algorithms for encountering marine hazards. This essential risk management system, covered both, organizational and human resources. A paper by Tirunagari et al. [8] used NLP methods to cluster the maritime accident reports. The results show that a new report could be classified as a particular type of accident based on the causal relations extracted from the text.

The literature suggests that, text mining could be applied on accident investigation reports. However, application of text mining is a complex task as it involves dealing with the text data which are very unstructured and fuzzy [8]. Moreover, there are quite a few challenges when dealing with accident reports. The reports are written in the natural language with no standard template.

[1] http://www.maib.gov.uk/

Misspellings and abbreviations are often found. Detecting of multi words such as "safety culture", "spirit status", etc are difficult because, which multi word is of greater importance is not known. The words "safety" and "culture" have a different meaning when considered as different words, but has a completely different meaning when considered as a single word. Therefore, context and semantics also play an important role in text mining. Limitations with regard to the experiments in this paper are explained in the section 6.

In this paper, we extract few concepts that are present in the accident investigation reports. Later, these causal patterns may be utilized, e.g. in constructing causal Bayesian networks for risk management [2]. The objectives of this paper are as follows:

- To extract causal relations from the maritime accident investigation reports.
- To extract concepts involved in the accident investigation reports.
- To link the concepts to the type of accident.

2 CAUSAL RELATIONS

A causal relation is usually a sentence consisting of a reason and a result phrase. Reason is the cause and result is the effect or impact of the cause. These relations are usually connected by a transition or conjunction [19, 20]. Table-I shows the terms which serve as a transition from one sentence to the next. Transition introduces an effect of situation in a causal sentence. The examples here are taken from the grammar-quizzes[2] website.

Table I. Cause (reason) and effect (result) with transition.

CAUSE (REASON)	TRANSITION	EFFECT (RESULT)
She had no other options.	***Consequently,***	she married at thirteen.
She was not protected.	***As a result,***	she had a baby at thirteen.
She had no access to health education or medical clinics.	***Therefore,***	she was more likely to get HIV.
There was poor sanitation in the village.	***As a consequence,***	she had health problems.
The water was impure in her village.	***For this reason,***	she suffered from parasites.
She had no shoes, warm clothes or blankets.	***For all these reasons,***	she was often cold.
She had no resources to grow food.(land, seeds, tools)	***Thus,***	she was hungry.
She had not been given a chance,	***so***	she was fighting for survival.

In the Table-II, “because” and other conjunctions, join one clause with another clause. Conjunction introduces a cause (reason) for the situation stated in the other clause.

Table II. Effect (result) and cause (reason) with conjunction.

EFFECT (RESULT)	CONJUNCTION	CAUSE (REASON)
She married at thirteen	***because***	she had no other options.
She had a baby at thirteen	***as***	she was not protected.
She was more likely to get HIV	***since***	she had no access to health education.
She had health problems	***because of***	poor sanitation in the village.
She suffered from parasites	***on account of***	the impure water in her village.
She was often cold	***due to***	not having shoes, warm clothes or blankets.
She was hungry	***for the reason that***	she had no resources to grow food.
She was fighting for survival	***since***	she had not been given a chance.

Girju et al. [21] extracted causal relations which included verb phrases. They classified the verb phrases present in causal relations in to four categories: 1) Low ambiguity and high frequency (LAHF). 2) Low ambiguity and low frequency (LALF). 3) High ambiguity and low frequency (HALF). 4) High ambiguity and high frequency (HAHF). The verb phrases which have LAHF are as follows: “cause”, “affect”, “induce”, “produce”, “generate”, “effect”, “arouse”, “elicit”, “lead to”, “trigger”, “derive”, “associate”, “relate to”, “link”, “originate”, “bring on”, and “result”. But in this paper, we concentrated on verb phrases such as “cause” and “result”, since they have no ambiguity. Table-III shows causal relations with verb phrases. Here the verb phrase introduces the effect in the cause and result expressions. Both verbs “cause” and “result” are used in the active form.

[2] http://www.grammar-quizzes.com/19-2.html

Table III. Cause (reason) and effect (result) with verb phrases.

CAUSE (REASON)	VERB PHRASE	EFFECT (RESULT)
Poor childhood education	***causes***	illiteracy.
Poor childhood education	***results***	in illiteracy.

In Table-IV, both verbs "cause" and "result" are used to introduce a cause. The verb cause may be used in the passive form with a "by phrase". The verb result does not take the passive form. Instead, it is followed by a prepositional phrase "from".

Table IV. Effect (result) and cause (reason) with verb phrases.

EFFECT (RESULT)	VERB PHRASE	CAUSE (REASON)
Illiteracy	***is caused***	by poor childhood education.
Illiteracy	***results / is resulted by***	from poor childhood education.

3 SELF ORGANISING MAPS

Self Organizing Map (SOM) introduced by Kohonen [5] is an unsupervised [23] neural network method which has both clustering [8, 11, 12, 13] and visualization properties. In the field of "machine-learning", the algorithms are, either 'supervised' [25] or 'unsupervised' [24] or 'reinforced' [26]. The distinction is drawn from how the learner classifies data. In supervised algorithms, the classes are predetermined. These classes can be conceived of as a finite set, previously arrived at by a human. The machine learner's task is to search for patterns and construct mathematical models. These models then are evaluated on the basis of their predictive capacity in relation to measures of variance in the data itself [25]. Decision tree induction, naive Bayes, etc are examples of supervised learning techniques [25]. Unsupervised learners are not provided with classifications. In fact, the basic task of unsupervised learning is to develop classification labels automatically. Unsupervised algorithms seek out similarity between pieces of data in order to determine whether they can be characterized as forming a group [25]. SOM can be considered as an algorithm that maps a high dimensional data space, to a lower dimension, generally 2 to 3 dimensions, called a map. This projection enables the input data to be partitioned into "similar" clusters while preserving their topology.

4 DOCUMENT REPRESENTATION

Vector Space Model (VSM) is a classical approach applied on text documents to obtain a matrix of numbers [18]. The vector space model is based on linear algebra and treats documents as vectors of numbers, containing values corresponding to occurrence of words (also called terms) in respective documents [18]. Let t be size of the terms set, and n be the size of the documents set. Then, all documents D_i, $i=1,\cdots,n$ may be represented as t-dimensional vectors: where coefficients a_{ik} represent the values of term k occurrences with in document D_i [18]. Thus both documents and terms form a term-document matrix$A_{(n\times t)}$. Rows of this matrix represent documents, and columns represent term vectors. Let us assume that position a_{ik} is set equal to 1, when term k appears in document i, and to 0 when it doesn't appear in it. Then for example documents collection corresponding to a query "king" we could create a corresponding term-document matrix.

Documents set:

D_1: The King University College

D_2: King College Site Contents

D_3: University of King College

D_4: King County Bar Association

D_5: King County Government Seattle Washington

D_6: Martin Luther King

Terms set: The, King, University, College, Site, Contents, of, County, Bar, Association, Government, Seattle, Washington, Martin, Luther

Table V. Term-document matrix.

Doc/ Term	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}	T_{12}	T_{13}	T_{14}	T_{15}
D_1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
D_2	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0
D_3	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0
D_4	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0
D_5	0	1	0	0	0	0	0	1	0	0	1	1	1	0	0
D_6	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1

From the Table V, we see that term "King" (T_2) appears in documents D{1,2,3,4,5,6} i.e the term "King" is present in all the documents. But, the term "Luther" is only present in D_6.

The Linear algebra as the basis of the model is a merit [18]. After transforming documents to vectors linear algebraic mathematical operations can be easily applied. Simple, efficient data structures may be used to store data. Representation of documents in the vector space

model is very simple. However, often these vectors are sparse, i.e. most of contained values are equal to 0. Hence, sparse vectors could be used to save memory and time.

In basic vector space model, only occurrence of terms in documents is of importance and their order is not considered. It is the main reason why this approach is often criticized [14, 15], as the information about the proximity between words (their context in sentence) is not utilized. Consider for example two documents: one containing a phrase "White House", which has a very specific meaning, and another containing a sentence "A white car was parked near the house" Treating documents simply as sets of terms we only know that words "white" and "house" occur in both documents, although their context there is completely different. However, this problem can be easily by supplementing this model, using phrases in addition to terms in document vectors, as described in [16, 17].

5. DATA AND EXPERIMENTS

5.1 DATA

The document collection used in the experiment is 'MAIB accident investigation reports'. There are 11 categories of accidents in the collection. We concentrated on only 4 types of accidents with 135 documents as shown in the Table VI.

Table VI. Document Collection.

Accident Type	Documents
Collisions	55
Groundings	44
Machinery failures	21
Fire	15
Total	**135**

5.2 EXPERIMENT

As a preprocessing step, the document collection is parsed from pdf to text files. Puntuation symbols are then removed and the text is converted to a lower case. Using linux command "grep"[3] the sentences which contains the transitions, conjunctions and verb phrases listed in Table I, Table II, Table III and Table IV are collected. A accident report typically consists of 60 pages. The causal relations extracted from a report is on average 10 sentences. Hence a 60 page report is transformed to a half page text document. Some example causal relations extracted from a single 60 pages report, are as follows:

Cause 1: "In assessing that Boxford was overtaking the fishing vessel, it is clear that the master misinterpreted the lights he saw. ***Consequently***, his alteration to starboard to keep clear of Admiral Blake only served to reduce an already small CPA, thereby exacerbating the close-quarters situation."

Cause 2: "The master did not activate Saffier's general alarm or alert the crew in any other way. ***Consequently*** they had limited warning to prepare for, or react to, the subsequent damage."

Cause 3: "No fire detection or fire suppression systems were fitted. ***As a result***, the fire was able to develop undetected for about minutes."

Cause 4: "The distortion and subsequent cracking of the furnace tube in the auxiliary boiler was ***due to*** sustained overheating."

Cause 5: "The scenario that the fire was ***caused*** when hot debris from the hotwork on the hopper came into contact with the conveyor belt."

Cause 6: "Actions to reduce, or stop, the sheer, were insufficient to counteract the forces acting on the hull. ***Therefore***, control of Arold was lost and a collision with the approaching Anjola ensued."

A total of 1775 causal relations are collected out of 135 accident investigation reports. A total of 19 concepts are generated manually after carefully reading the extracted causal relations. The manually generated concepts are as follows: "tiredness", "confusion", "background-noise", "labour-relationship", "bad-light", "health", "chemical", "alcohol", "traffic", "awareness", "visibility", "electrical", "fuel", "information", "fatigue", "machine-fault", "speed", "bad-weather" and "equipment". The causal relations are then given as an input to a NLTK-Python program. Natural Language Tool Kit (NLTK[4]) is a leading platform for building Python[5] programs to work with human language data. This program extracts nouns and adjectives, a total of 6547 words. These words are compared with the 19 concepts based on the NLTK word-net[6] similarity. The words which have 60% or more similarity with the concept are replaced by the concept itself in the extracted causal relations. Few example words replaced with the concepts are listed in the Table-VII. Using these concepts, the document-term matrix is constructed, which is then given as an input to the SOM program[7]. The SOM program outputs a text visualization map of concepts and its involvement with the accidents. In this experiment we have chosen 5X5 SOM grid for visualization. The entire experimental procedure is shown as a flow chart in Figure 1.

[3] http://linux.die.net/man/1/grep

[4] http://nltk.org/
[5] http://www.python.org/
[6] http://nltk.googlecode.com/svn/trunk/doc/howto/wordnet.html
[7] http://www.cis.hut.fi/somtoolbox/

Table VII. Words that are replaced by concepts.

Word	Concept
abuse	labour-relationship
acetone	chemical
acceleration	speed
arc-flash	electrical
asphyxia	fatigue
blackout	visibility
brake	equipment
buckling	machine-fault
deterioration	health
diesel	fuel
smoking	awareness

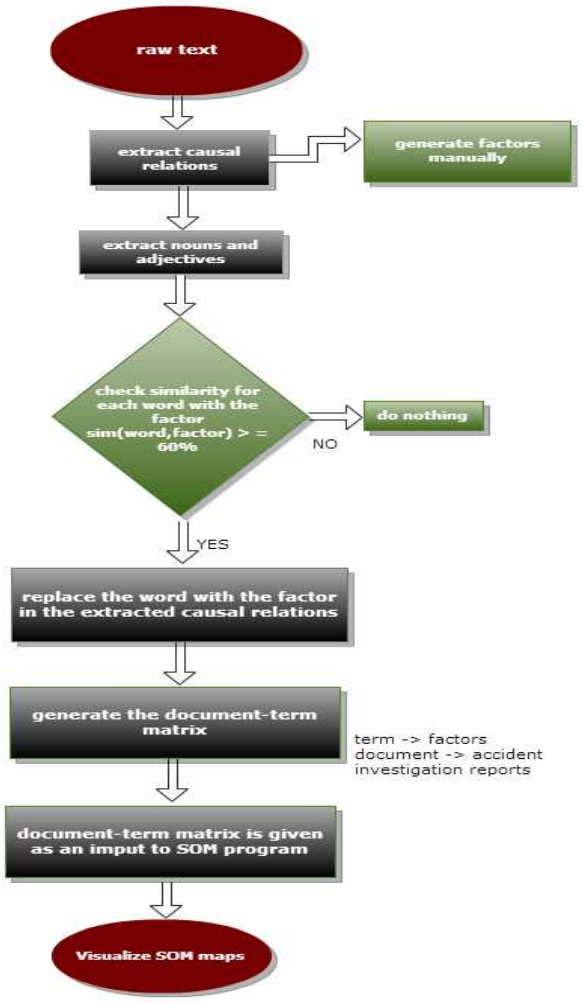

Figure 1: Flow chart showing the experiment procedure.

6. RESULTS AND DISCUSSIONS

The resulting SOM maps show how the concepts such as, "visibility", "fatigue", "bad-weather" etc. are involved in the accident investigation reports. C, G, M and F represent collisions, groundings, machinery failures and fire accidents respectively. The colour bar in the map shows the correlation between the concept and the accident type. From the Figure 2, it can be seen that "traffic density" as a concept, is involved in many collision and grounding accidents. The darker the gray colour in the SOM map the more frequent is concept within the accident investigation reports.

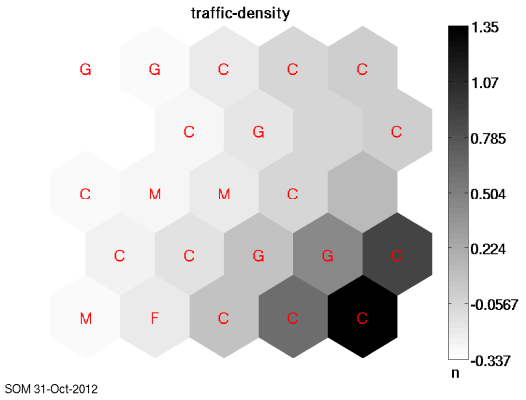

Figure 2: SOM map showing how "traffic density" as a concept correlates with the accident types C,G and how frequent the concept is involved within the accidents (C: collision, G: grounding, F: fire, M: machinery failure).

From the Figure 3, it can be seen that "speed" as a concept involved in many collision accidents. It can also be seen that "speed" as a concept is associated with few grounding related accidents. "Chemical" as a concept is highly correlated with machinery failures and also been associated with fire accidents in the investigated reports. "Confusion" as a concept is strongly correlated with collisions accidents and obviously "machine-fault" as a concept involved in machinery failures. "Equipment" as a concept involved in collisions and groundings in the analyzed accident investigation reports.

Figure 4, shows the concepts and their involvement in the four major accidents type. The larger font in the figure conveys the importance of the concept to the particular type of accident. "Health" as a concept involved in all the four accidents viz. Collision, grounding, fire and machinery failure, though it is strongly correlated to machinery failure and fire accidents. "Bad-weather" as a concept involved in collision and grounding accidents. The importance of the concept is determined by the frequency of its occurrence in the accident investigation reports.

Figure 5 shows the contribution of concepts in the four major accidents type investigated in this paper. 44% of the concepts are involved in collisions, 38% in groundings, 10% in machinery failures and 8% involved in fire accidents for the analyzed accident investigation reports.

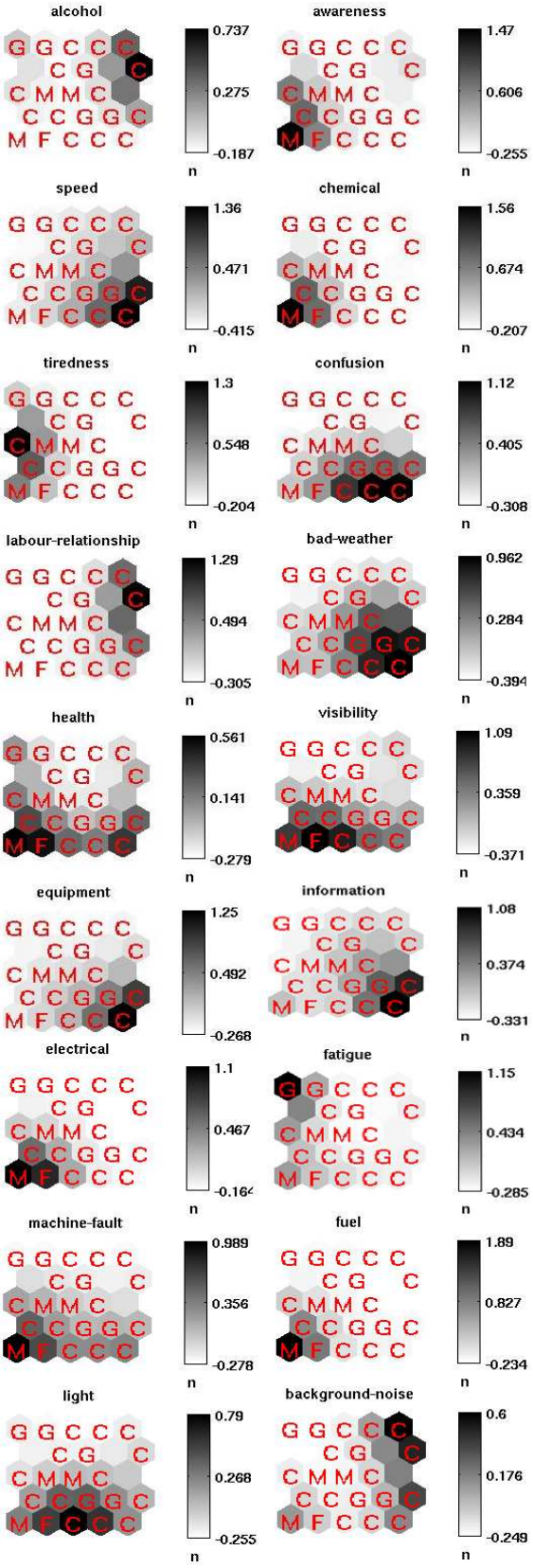

Figure3: SOM maps showing how different concepts correlate with the accident type and how frequent the concepts are within the accidents (C: collision, G: grounding, F: fire, M: machinery).

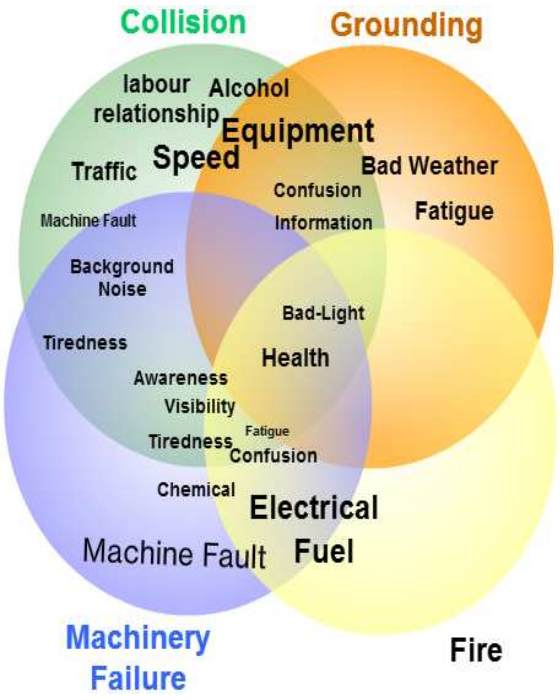

Figure 4: Venn diagram showing the concepts and their involvement in the four major accidents type.

Figure 6, reveals the frequency of concepts present in the accident investigation reports. Frequency here means number of occurrences of a concept with in all the accident investigation reports.

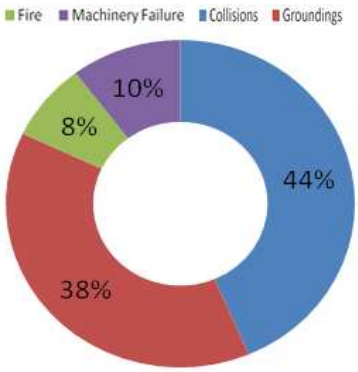

Figure 5: Doughnut chart showing the contribution of concepts in the four major accidents.

The concepts such as, “equipment”, “fatigue”, “information”, “machine-fault”, “speed”, “fuel”, “electricity” and “visibility” have greater frequency when compared to the others with more than 100 occurrences.

The result in this section contradicts the western researchers’ research, that up to 80% of the marine accidents are related to the human error [30]. A report issued by USCG stated that between 75–96% of marine casualties are caused at least in part by some form of human error [27]. Esbensen et al. [29] in their report

stated that 43% of the accidents reported to the U.S. Coast Guard cite human error as the primary cause. The authors also stated that actual figure of incidents involving human error may be as high as 80%. Wagenaar et al. [28] analyzed 100 accidents from the Dutch Shipping Council between 1982 and 1985 and cited 2,250 causes for the accidents. The authors have also stated that in 96 of 100 cases, the people involved could have prevented the accident; however they were rarely caused by just one human error [30].

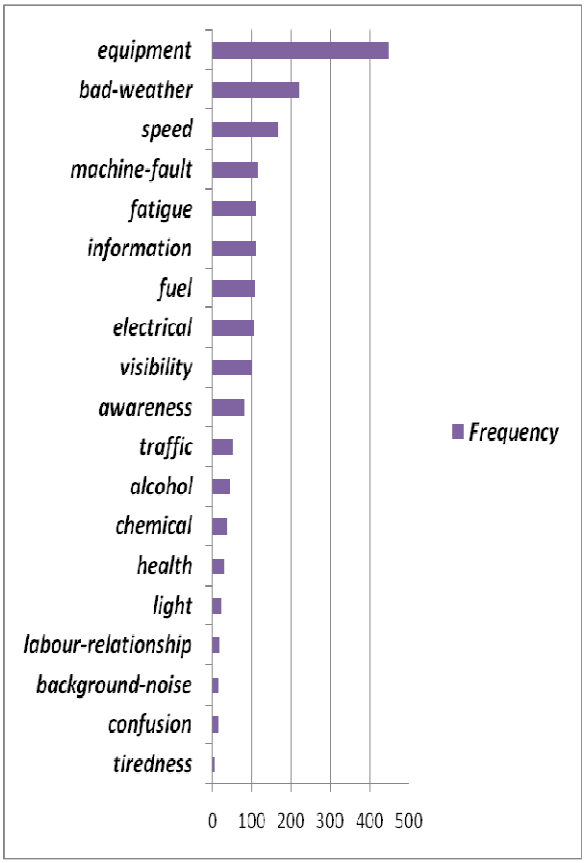

Figure 6: Bar chart showing the concepts in accidents and their frequencies in the analysed MAIB investigation reports.

Schröder-Hinrichs et al. [22] investigated 41 accident investigation reports of fire in machinery spaces using HFACS as analysis tool. Their investigation revealed that technical failures were more dominating in the marine accidents than human factors. Their paper also pointed out that accident investigation reports only present a limited number of all facts gathered during an accident investigation. The results in this paper also show that "equipment", "bad-weather", "machine-fault" and "electrical" as concepts occur more frequently in the analyzed marine accident reports which are not related to any human errors. Hence within these results the technical failures and bad-weather are by far the most dominating concepts.

Experiments using text mining in this regard has some limitations as well. The word "blackout" (Table VII) generally means a period during a massive power failure with the lack of electricity and it also means alcohol-related amnesia, where a person loses his memory. This rises an ambiguity to what concept it should be linked. Hence, the use of contextual information is required here to link the word "blackout" to a "visibility" concept or an "alcohol" concept. Sometimes the accident investigation report mention sentences such as: "fatigue was not a cause in this accident", but in our experiments, the concept "fatigue" is also considered as an occurrence. Hence, sentiment analysis should be applied to analyse the causal relations. A concept should be mined if it has a positive sentiment with regard to the context.

7. CONCLUSIONS

This study shows the significance of text mining methods in order to infer useful information from marine accident investigation reports. The results obtained using SOM analysis provide a good understanding of the causal concepts by showing the underlying correlations within different accident types investigated in this paper. The results in this paper reveals, concepts such as, "equipment", "fatigue", "information", "machine-fault", "speed", "fuel", "electricity" and "visibility" have greater frequency when compared to the rest of the concepts with more than 100 occurrences. Technical concepts such as "equipment", "machine-fault", "electrical", "fuel" and "information" are by far the most dominating concepts in the analyzed accident investigation reports. We hope this research will provide a starting point for developing effective tools and methodologies for identifying human and organizational concepts present in the accident investigation reports.

8. ACKNOWLEDGEMENTS

The study was conducted as a part of CAFE project, financed by the European Union - European Regional Development Fund - Regional Council of Päijät-Häme, the City of Kotka, Kotka-Hamina regional development company Cursor Ltd., Kotka Maritime Research Association Merikotka and the following members of the Kotka Maritime Research Centre Corporate Group: Port of Hamina Kotka, Port of Helsinki, Aker Arctic Technology Inc. and Arctia Shipping Ltd.

9. REFERENCES

1. Grech, M.R. and Horberry, T. and Smith, A., Human error in maritime operations: Analyses of accident reports using the leximancer tool. Proceedings of the Human Concepts and Ergonomics Society Annual Meeting 46:19, 1718–1721 (2002).

2. Kristiansen.S., A BBN approach for analysis of maritime accident scenarios. Reliability, Risk and Safety ISBN 978-0-415-60427-7 (2010).

3. Schröder-Hinrichs, J.U. and Baldauf, M. and Ghirxi, K.T., Accident investigation reporting deficiencies related to organizational concepts in machinery space fires and explosions. Accident Analysis & Prevention 43:3, 1187–1196 (2011).

4. K.B. Artana, DD Putranta, I.K. Nurkhalis, and YD Kuntjoro. Development of simulation and data mining concept for marine hazard and risk management. In Proceedings of the 7th International Symposium on Marine Engineering (24-28 October), 2005.

5. Kohonen, T.: self-organizing maps, Springer-Verlag, 1997.

6. L. Francis and M. Flynn. Text mining handbook. In Casualty Actuarial Society E-Forum, Spring 2010, page 1, 2010.

7. A.H. Tan et al. Text mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases, pages 65–70, 1999.

8. S. Tirunagari, M. Hanninen, A. Guggilla, K. Stahlberg, and P. Kujala. Impact of similarity measures on causal relation based feature selection method for clustering maritime accident reports. Journal of Global Research in Computer Science, 3(8):46–50, 2012.

9. W.L. Yiu et al. Investigation and validation of a novel text mining methodology for occupational accident analysis and prevention— nova. the university of newcastle's digital repository. 2011.

10 B. Zheng and Y. Jin. Analysis on concepts leading to human fault in marine accidents based on attribute reduction [j]. Journal of Shanghai Maritime University, 1:026, 2010.

11. MacQueen, J.: Some methods for classification and analysis of multivariate observations. The fifth Proc. Berkeley symposium on Math., stat. and Prob., Vol. 1, 281-297, 1965.

12. Paukkeri, Mari-Sanna, Ilkka Kivimäki, Santosh Tirunagari, Erkki Oja, and Timo Honkela. "Effect of Dimensionality Reduction on Different Distance Measures in Document Clustering." In *Neural Information Processing*, pp. 167-176. Springer Berlin/Heidelberg, 2011.

13. Diday, E. and Schroeder, A. and Ok, Y.: The Dynamic Clusters Method in Pattern Recognition. IFIP Congress 1974, 691-697.

14. O. Zamir and O. Etzioni. Web document clustering: A feasibility demonstration. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 46-54. ACM, 1998.

15. D. Weiss, D. Weiss, and S.I. Oprogramowania. A clustering interface for web search results in polish and english. 2001.

16. Y.S. Maarek, D.M. Berry, and G.E. Kaiser. An information retrieval approach for automatically constructing software libraries. Software Engineering, IEEE Transactions on, 17(8):800{813, 1991.

17. Y.S. Maarek, R. Fagin, I.Z. Ben-Shaul, and D. Pelleg. Ephemeral document clustering for web applications. 2000.

18. G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11):613-620, 1975.

19. Huddleston, Rodney and Geoffrey K. Pullum, et al. The Cambridge Grammar of the English Language (CaGEL) . Cambridge: Cambridge University Press, 2002. Print.

20. Quirk, Randolph and Sidney Greenbaum. A Comprehensive Grammar of the English Language (CoGEL) . 7th ed. New York: Longman Group, 1989. Print

21. Text Mining for Causal Relations; Roxana Girju and Dan Moldovan; FLAIRS-02 Proceedings, American Association for Artificial Intelligence, 2002.

22. Schröder-Hinrichs, Jens U., Michael Baldauf, and Kevin T. Ghirxi. "Accident investigation reporting deficiencies related to organizational concepts in machinery space fires and explosions." Accident Analysis & Prevention 43, no. 3 (2011): 1187-1196.

23. Barlow, Horace B. "Unsupervised learning." Neural Computation 1, no. 3 (1989): 295-311.

24. Gentleman, R., and V. J. Carey. "Unsupervised machine learning." Bioconductor Case Studies (2008): 137-157.

25. Kotsiantis, S. B., I. D. Zaharakis, and P. E. Pintelas. "Supervised machine learning: A review of classification techniques." Frontiers in Artificial Intelligence and Applications 160 (2007): 3.

26. Sutton, Richard S., and Andrew G. Barto. "Reinforcement learning: An introduction (Adaptive computation and machine learning)." (1998).

27. Rothblum, A.R. (2000, October 13–20). Human error and marine safety. Paper presented at the National Safety Council Congress and Expo, Orlando, FL.

28. Wagenaar, W. A., & Groeneweg, J. (1987). Accidents at sea: Multiple causes and impossible consequences. International Journal of Man-Machine Studies, 27, 587−598.

29. Esbensen, P., Johnson, R. E., & Kayten, P. (1985). The importance of crew training and standard operating procedures in commercial vessel accident prevention. Paper presented at the Tenth ship technology and research (STAR) symposium, Norfolk.

30. Hetherington, Catherine, Rhona Flin, and Kathryn Mearns. "Safety in shipping: The human element." Journal of Safety Research 37, no. 4 (2006): 401-411.

31. Pang, Bo, and Lillian Lee. "Opinion Mining and Sentiment Analysis." Information Retrieval 2, no. 1-2 (2008): 1-135.

9. AUTHORS BIOGRAPHY

Santosh Tirunagari holds the current position of research assistant at Aalto University, Department of Applied Mechanics. He is responsible for analysing the human and organizational concepts causing the accidents in maritime domain using text mining methods. His previous experience includes informational retrieval and image analysis using machine learning methods.

M.Sc. (Tech.) Maria Hänninen is a researcher and a doctoral student at Aalto University, Department of Applied Mechanics. She received her Master's degree from the Department of Automation and Systems Technology in Helsinki University of Technology in 2004. Since 2007, she has worked in EU-funded research projects addressing the safety of maritime traffic. Her research focuses on developing a Bayesian-network based tool for maritime safety management and accident prevention.

Kaarle Ståhlberg is a researcher and a doctoral student at Aalto University, Department of Applied Mechanics. He graduated as Master of Science in Naval Architecture in the year 2010 from Aalto University. He is responsible for modelling of ship behaviour before accidents.

Pentti Kujala is a professor of marine technology (safety) at the Aalto University, School of Engineering in Finland. He has about 35 years of research experience related to ice-going vessels and structures. He has been working before e.g. at Lloyd´s Register of Shipping in London, VTT in Finland and Aker Yards in Finland. He got the degree of Doctor of Technology in Naval Architecture at Helsinki University of Technology in the year 1994. His main research interests have been devoted to the analysis of ice-induced loads and their statistical nature on ships and development of innovative structural solutions for various types of ships.