Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 130 (2018) 98–102

Procedia Computer Science

www.elsevier.com/locate/procedia

The 9th International Conference on Ambient Systems, Networks and Technologies
(ANT 2018)

A Comparison of LSA and LDA for the Analysis of Railroad Accident Text

Trefor Williams[a]*, John Betak[b, c]

[a]Rutgers University, 96 Frelinghuysen Road,Piscataway,NJ,USA
[b]Collaborative Solutions LLC, 726-33 Tramway Vista Dr., NE, Albuquereque, NM,USA
[c]Center for Risk Management and Insurance, The University of Texas at Austin, 2110 Speedway Stop B6500, Austin, TX, USA

Abstract

Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation(LDA) were used to identify themes in a database of text about railroad equipment accidents maintained by the Federal Railroad Administration in the United States. These text mining techniques use different mechanisms to identify topics. LDA and LSA identified switching accidents, hump yard accidents and grade crossing accidents as major accident type topics. LSA identified accidents with track maintenance equipment as a topic. Both text mining models identified accidents with tractor-trailer highway trucks as a particular problem at grade crossings. It was found that the use of the two techniques was complementary, with more accident topics identified than with the use of a single method.

Keywords: railroad;text mining;accidents

1. Introduction

There are various causes of railroad accidents. In the United States, railroads are required to report accidents with damage costing greater than US $9,200 in 2010 to $10,500 in 2015 to the Federal Railroad Administration. The accident form includes a field for a textual description of the accident. Text mining is defined in the context of discovering previously unknown information that is implicit in the text but not immediately obvious[1]. There are various methods of text mining with distinct methods of identifying underlying topics in the text. This paper compares the results of applying Latent Dirichlet Allocation (LDA), a type of probabilistic topic modeling, and Latent Semantic Analysis (LSA), a natural language processing technique, to the text field in a database of Federal Railroad Administration Equipment Accident Reports.

There are few existing examples of the application of text mining to the railroad industry. Williams and Betak[123] have used LDA analysis, to study grade crossing accidents, equipment accidents and accident investigation reports. Brown[4] has used LDA models to mine text from railroad accident reports. The output from these LDA models are used to enhance data analytic models that predict the cost of extreme accidents. Previous railroad studies have only used LDA. The goal of this paper is to compare and contrast the output of LSA models with LDA to determine if use of the two models leads to additional insights when applied to railroad data.

*Corresponding Author. Tel: +1-848-445-2880 ; fax +1-732-445-0577
Email address: tpw@soe.rutgers.edu

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2018.04.017

The data used for this study are from the Federal Railroad Administration's railroad equipment accident database available on line at http://safetydata.fra.dot.gov/officeofSafety/publicsite/Query/AccidentByStateRailroad.aspx. The database includes text fields that describe each accident. The length of the text field varied from a few words to paragraphs with several sentences. These text data were mined to reveal additional knowledge about railroad accidents. Data for the railroad equipment accidents were collected from January 2010-February 2015. There were 12,447 accidents reported during this period.

2. Text Mining Algorithms

The details of the LSA and LDA algorithms are presented in this section. This section illustrates how the two algorithms use different mechanisms to automatically generate the topics in the text corpus. A topic is a grouping of related words.

2.1. Latent Semantic Analysis

Text can be characterized by the semantic content it carries. Over the past two decades computational models have been developed to create semantic representations for words encountered in text. One such model is Latent Semantic Analysis (LSA)[5,6].

LSA is a computational model that works on the notion that words with similar meanings tend to appear in similar contexts. It creates semantic representations for words by analyzing the pattern with which words occur together in documents across thousands of text samples provided to it in a training corpus. Then, from an analysis of the words that do and do not co-occur in the corpus, the model estimates what words should occur in similar documents (i.e., contexts) and are, therefore, close to each other in the semantic space [7].

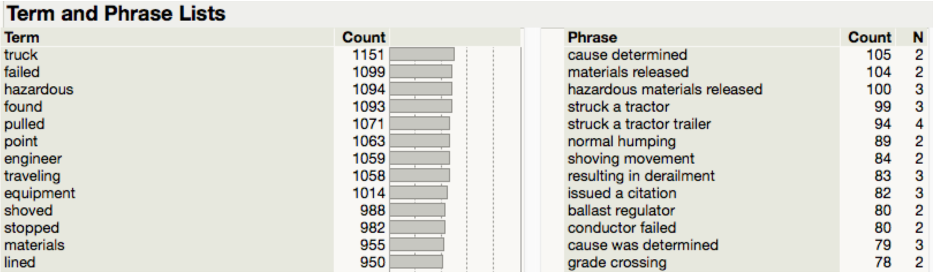
The LSA algorithm generates a semantic space from a statistical analysis of the frequencies with which words co-occur in a large collection of documents (i.e., contexts). The process by which LSA builds a semantic space from the document collection is called 'training'. After training, the semantic space comprises a set of vectors containing the semantic features for each word encountered in the document collection. We refer to the vectors in the semantic space as, semantic vectors. Generally speaking, the more documents contained in a training corpus, the more contextual information the system has to semantically differentiate or align words[7].

2.2. Latent Dirichlet Allocation

Topic modeling algorithms are statistical methods that analyze the words of unstructured original texts to automatically discover the themes that run through them. Topic models automatically organize a text collection into its major themes. A frequently used topic-modeling algorithm is Latent Dirichlet Allocation (LDA). Details of the LDA Algorithm are given by Blei[8]. LDA is a generative probabilistic model for collections of discrete data such as text corpora.

The underlying assumption of LDA is that a text document will consist of multiple themes. LDA is a three-level hierarchical Bayesian model where each item of a collection of text is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. For text modeling, the topic probabilities provide an explicit representation of a document[9]. Additionally, a topic model generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers per-document discrete distributions over topics. In other words, the LDA algorithm automatically identifies words that occur in the accident reports and forms then into ranked topics. In this case, we are using the LDA algorithm's ability to find themes in the FRA accident reports.

The groups of words automatically generated (the topics) may represent the most frequent types of railroad accidents that occur. In addition, the LDA ranks the topics by their probability of occurrence and also ranks the words in each topic by their probability of occurrence within the topic.



Term and Phrase Lists

Term	Count	Phrase	Count	N
truck	1151	cause determined	105	2
failed	1099	materials released	104	2
hazardous	1094	hazardous materials released	100	3
found	1093	struck a tractor	99	3
pulled	1071	struck a tractor trailer	94	4
point	1063	normal humping	89	2
engineer	1059	shoving movement	84	2
traveling	1058	resulting in derailment	83	3
equipment	1014	issued a citation	82	3
shoved	988	ballast regulator	80	2
stopped	982	conductor failed	80	2
materials	955	cause was determined	79	3
lined	950	grade crossing	78	2

Fig. 1. Frequent Words and Phrases

3. Basic Text Mining

The Initial phase of analyzing the railroad accident text was to remove stop words and to tokenize the text into individual words and phrases up to four words in length. Using the JMP text mining software, an initial analysis of
the text provides a count of the most frequently occurring words and phrases. The JMP software also provided the facility to add stop words. In processing the text, frequently occurring words like "the" and "and" are automatically removed from the word list. The software uses a standard list of these stop words. It was found that adding the most frequently occurring words in the FRA accident text to the stop word list improved the clarity of the topics. Therefore, words like rail and track were deleted from the analysis.

Figure 1 shows the most frequently occurring words, and phrases after stop words were removed. There are several different frequently occurring phrases that illustrate the nature of many accidents. For example, one of the most frequently occurring phrases is "struck a tractor trailer." This indicates that accidents at highway grade crossing between a train and a tractor-trailer highway truck. Another interesting term that occurs frequently is "ballast regulator." This is a type of railroad maintenance equipment that is used to shape and distribute the stone ballast that acts as a foundation for railroad ties. Recently there have been several serious accidents involving maintenance-of-way equipment being struck by trains. The phrase "hazardous materials released" occurs frequently and indicates that in many accidents there are releases of liquid or gaseous chemicals.

4. LSA and LDA Model Outputs

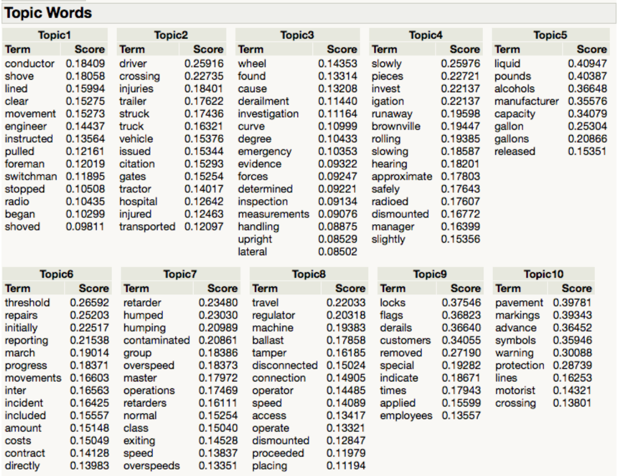
The topics generated by the LSA and LDA models are shown and their meaning in a railroad accident context are discussed in this section.

4.1. Latent Semantic Analysis Topics

LSA can be used to automatically generate topics from the words contained in the corpus of text of the railroad equipment accidents. Each topic is a listing of words associated with a particular accident theme. The number of topics to generate is selected by the user. Through experimentation it was found that 10 topics yielded the most useable results. Figure 2 shows the topics generated from the analysis. Figure 2 shows a table of terms in each topic that have the largest scores in absolute value. Each topic is sorted in descending order by the absolute value of the score

The LSA analysis yielded several interesting topics. They include:

- Grade crossing accidents. Two topics clearly addressed grade crossing accidents. Topic 2 shows that there is a grouping of truck and tractor-trailer accidents at highway grade crossings. Topic 10 mentions is a text grouping that includes accidents related to highway pavement markings and signs at grade crossings.
- Accidents related to maintenance equipment particularly including tampers and ballast regulators (Topic 8).
- Accidents relating to switching when cars are shoved (Topic 1).
- Accidents in hump yards. Hump yards are yards where freight cars pushed up a small hill and then roll down to be automatically switched to the correct track (Topic 7). Many major classification yards in the United States are hump yards.
- Accidents related to liquid spills (Topic 5).
- Accidents caused by wheel problems (Topic 3).



Topic Words

Topic1		Topic2		Topic3		Topic4		Topic5	
Term	Score	Term	Score	Term	Score	Term	Score	Term	Score
conductor	0.18409	driver	0.25916	wheel	0.14353	slowly	0.25976	liquid	0.40947
shove	0.18058	crossing	0.22735	found	0.13314	pieces	0.22721	pounds	0.40387
lined	0.15994	injuries	0.18401	cause	0.13208	invest	0.22137	alcohols	0.36648
clear	0.15275	trailer	0.17622	derailment	0.11440	igation	0.22137	manufacturer	0.35576
movement	0.15273	struck	0.17436	investigation	0.11164	runaway	0.19598	capacity	0.34079
engineer	0.14437	truck	0.16321	curve	0.10999	brownville	0.19447	gallon	0.25304
instructed	0.13564	vehicle	0.15376	degree	0.10433	rolling	0.19385	gallons	0.20866
pulled	0.12161	issued	0.15344	emergency	0.10353	slowing	0.18587	released	0.15351
foreman	0.12019	citation	0.15293	evidence	0.09322	hearing	0.18201		
switchman	0.11895	gates	0.15254	forces	0.09247	approximate	0.17803		
stopped	0.10508	tractor	0.14017	determined	0.09221	safely	0.17643		
radio	0.10435	hospital	0.12642	inspection	0.09134	radioed	0.17607		
began	0.10299	injured	0.12463	measurements	0.09076	dismounted	0.16772		
shoved	0.09811	transported	0.12097	handling	0.08875	manager	0.16399		
				upright	0.08529	slightly	0.15356		
				lateral	0.08502				

Topic6		Topic7		Topic8		Topic9		Topic10	
Term	Score	Term	Score	Term	Score	Term	Score	Term	Score
threshold	0.26592	retarder	0.23480	travel	0.22033	locks	0.37546	pavement	0.39781
repairs	0.25203	humped	0.23030	regulator	0.20318	flags	0.36823	markings	0.39343
initially	0.22517	humping	0.20989	machine	0.19383	derails	0.36640	advance	0.36452
reporting	0.21538	contaminated	0.20861	ballast	0.17858	customers	0.34055	symbols	0.35946
march	0.19014	group	0.18386	tamper	0.16185	removed	0.27190	warning	0.30088
progress	0.18371	overspeed	0.18373	disconnected	0.15024	special	0.19282	protection	0.28739
movements	0.16603	master	0.17972	connection	0.14905	indicate	0.18671	lines	0.16253
inter	0.16563	operations	0.17469	operator	0.14485	times	0.17943	motorist	0.14321
incident	0.16425	retarders	0.16111	speed	0.14089	applied	0.15599	crossing	0.13801
included	0.15557	normal	0.15254	access	0.13417	employees	0.13557		
amount	0.15148	class	0.15040	operate	0.13321				
costs	0.15049	exiting	0.14528	dismounted	0.12847				
contract	0.14128	speed	0.13837	proceeded	0.11979				
directly	0.13983	overspeeds	0.13351	placing	0.11194				

Fig. 2. LSA Topics

4.2. Latent Dirichlet Analysis Topics

The LDA software requires the number of topics to be input by the user. Best results were found using 20 topics. Table 1 shows the most interesting topics generated from the LDA modelling. Topic 1 is a topic related to hump yard accidents. Topics 2, and 7 include switching accidents. Topic 3 indicates that there is a grouping of accidents involving cars being shoved. Topic 4 appears to include accidents involving signals. Topic 5 indicates that accidents involving brakes are a major railroad equipment accident theme. Topic 8 is very interesting because it suggests that highway-rail grade crossing accidents involving trucks with trailers is a common accident type. Topic 14 includes the words "hazardous" and "materials" and suggests that accidents involving hazardous materials occur in yards when cars are shoved or kicked. Words like "derail" and "derailed" are major words in Topics 9, 10, and 19.

5. Comparison and Analysis of the Topic Modeling Techniques

Both LDA and LSA yielded useful information about the nature of the railroad equipment accidents. There is significant overlap in the topics generated by the two methods. Table 2 shows the major accident themes identified by each method.

Table 1. LDA Topics

Topic	Top Words
1	bowl utlx hump humped car gatx humping tilx tank operations class retarder fuel cut gallons group foul stalled speed system
2	switch lined point movement ran crossover move run line failed previously reverse switches improperly zone split route running points pulled
3	track cars shoving shoved cut shove pulled job made movement move foreman joint making make coupling clear double couple switchman
4	signal stop damaged fire operator causing pantograph stopped reported bridge control hit failed wire tra machine time ballast contact ck
5	train found emergency mph mp inspection brake speed investigation revealed air traveling curve brakes excessive slack handling grade upright experienced
8	struck crossing truck trailer injuries unit driver vehicle impact tractor road stop lead injured semi fouling gates northbound rig front
9	derailment cars derailed loaded wheels caused empty determined investigation coal curve load inside high grain flat resulted csx hopper center
10	derailed cars head loads ns pulling empties derailing tons units engines st westward shoving locomotives dttx wc eastward sou ft
14	cars released due track hazardous materials yard shoving derailed failure impacted articulated trk rco control contained pulling kicking mt irregular

Both techniques clearly show accidents related to hump yards, grade crossings, wheels and switching/shoving. The LSA technique identified accidents related to track maintenance equipment particularly ballast regulators and tampers, while the LDA did not clearly identify this topic. The LDA analysis found braking accidents. The LDA also found topics where the highest ranked word was derailment or derailed.

Table 2. Comparison of Identified Topics

Topic Meaning	LSA	LDA
Shoving/Switching	X	X
Grade Crossing Accidents	X	X
Wheels	X	X
Liquids Released	X	
Hazardous Materials		X
Track Maintenance Equipment	X	
Braking Accidents		X
Hump Yards	X	X
Derailment/Derailed		X

6. Conclusions

The analysis shows that that both techniques find the most frequently occurring accident types. However, the text mining techniques generated several topics that are not as widely known in the railroad industry including the identification of accidents involving ballast maintenance equipment, and the prominence of tractor-trailer highway trucks in grade crossing accidents. This illustrates how the text mining tools can be used to identify problems that require further investigation. It also illustrates that text mining yields information not observable in the numeric data used in most railroad accident statistical analyses. This further suggests that a very rich field of analysis lies in using these tools on additional railroad databases, such as track inspection reports, railroad incident reports filed for each employee involved in an incident and so on.

The use of the two text mining techniques complements each other. LSA and LDA are in agreement for many of the major accident topics, yet they each generated some topics that the other method didn't identify. This indicates that using more than one text mining technique that use different mechanisms to identify topics can result in a more meaningful analysis and better identification of accident causes from the text.

References

1. William T, Betak J, Nelson, C. Applying Topic Modeling to Railroad Grade Crossing Accident Reports. In: Joint Rail Conference 2015, Mar 23-26, San Jose, CA; New York: American Society of Mechanical Engineers, 2015.
2. Williams T, Betak, J, Findley, B. Text Mining Analysis of Railroad Accident Investigation Reports. In: Joint Rail Conference 2016, April 12-15, Columbia, SC; New York: American Society of Mechanical Engineers, 2016.
3. Williams, T, Betak J. Identifying Themes in Railroad Equipment Accidents Using Text Mining and Text Visualization, 2016 In: International Conference on Transportation and Development, Jun 26-29, 2016; Washington: American Society of Civil Engineers, 2016.
4.Brown, D. Text Mining the Contributors to Railroad Accidents. *IEEE Transactions on Intelligent Transportation Systems* 2016;**17**(2):346-355.
5. Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol Rev* 1997;**104**(2):211-240.
6. Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. *Discourse Processes* 1998 01/01;**25**(2-3):259-284.
7. Kwantes PJ, Derbentseva N, Lam Q, Vartanian O, Marmurek HHC. Assessing the Big Five personality traits with latent semantic analysis. *Personality and Individual Differences* 2016 November 2016;**102**(Supplement C):229-233.
8. Blei DM, Ng AY, Jordan MI, Lafferty J. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 2003 05/15;3(4):993-1022.
9. Blei DM. Probabilistic Topic Models. *Communications of the ACM* 2012;**55**(4):77-7.