Safety Science 88 (2016) 123–128

Contents lists available at ScienceDirect

Safety Science

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

Visual representation of safety narratives

S.D. Robinson [*]

Parks College of Engineering, Aviation and Technology, Saint Louis University, Saint Louis, MO 63103, USA

ARTICLE INFO

Article history:
Received 4 January 2016
Accepted 11 May 2016

Keywords:
LSA
Adaptive taxonomy
Safety
Isometric mapping
Machine learning

ABSTRACT

A computational method for the visualization of text-based safety narratives on a two dimensional plane is shown. This multi-step approach utilizes latent semantic analysis to first infer higher-order structures and then isometric mapping to reduce the projection to two dimensions. Metadata may then be overlaid on the projection. Demonstrated is the application of this process to the human coded primary-problems identified and the phase of flight for a sample of the Aviation Safety Reporting System database. It is evident that this approach provides additional insight for the analysis of large inter-related corpora commonly found in safety programs.

1. Introduction

The role of the safety analyst is to look for significance in any given set of data. Those instances, which in the analyst's experience are outliers, draw the focus of attention. The response by the analyst to anomalous reports are dependent on perceived risk. What is learned from the data facilitates an intervention whose goal is the perceived reduction in harm.

The seeming simplicity of this description belies the complexity of understanding what conditions are indicative of risk. Many paradigms for understanding and interpreting safety have been developed, but primarily form two approaches. In general, these models provide either a taxonomy for subsequent numerical analysis or a framework for interpretation of events.

The need for improved tools within safety research to identify similarity, estimate occurrence frequency, and visually interpret reports has been expressed most recently by Tanguy et al. (2015). In their discussion, highlighted is the unreleased potential of topic modeling applied to safety corpora. Through the use of topic modeling, they intimate that, greater value may be extracted from databases that do not have extensive post-reported classification or metadata. The authors suggest that through natural language processing (NLP), a user should be able to quantify the characteristics of a report, conduct temporal analysis, and view the links between factors. Thus identify trends or connections that may be indicative of safety issues.

The Aviation Safety Reporting System (ASRS) database provides a corpus of narratives lending itself to NLP techniques, as it is focused in both the reporter and event type. Robinson et al. (2015) demonstrated through the use of NLP that narratives may be compared numerically delivering significantly similar results to that of a thematic qualitative process. The procedure shown, applied latent semantic analysis (LSA) to facilitate document clustering. This approach provides an additional tool to reduce the efforts required of narrative analysis and the biases introduced by human interpretation. However, this approach does not provide a direct method for the comparison of multiple narratives simultaneously, nor does it directly indicate narrative uniqueness relative to all those narratives found in the complete database.

The visual representation of connections in language was first conducted by Louwerse et al. (2007) in order to demonstrate the symbolic encoding of relationships within language. At first limited to the discussion of symbolic and embodied language within the cognitive science domain, Louwerse and Zwaan (2009) went on to demonstrate the relationship between city locations and spatial information present in language. Not presented in their work was to use of LSA and a dimensional reduction technique to visualize narratives within a corpus.

The combination of LSA with the dimension reduction method of isometric mapping allows for the representation of safety narratives in a two-dimensional plot. This approach permits the assessment of context for individual reports within a corpus, review of the contextual topography of a corpus, a graphical method for intuitive navigation of categorical data found in the reports, and perhaps insight into mesoscale lexicological features found within a focused data set. Thus relationships only visible at the corpus level may present themselves.

[*] Address: Parks College of Engineering, Aviation and Technology, 3450 Lindell Blvd., Saint Louis, MO 63103, USA.

E-mail address: robinssd@slu.edu

http://dx.doi.org/10.1016/j.ssci.2016.05.005
0925-7535/© 2016 Elsevier Ltd. All rights reserved.

2. Theory

The amalgamation of LSA with isometric mapping for safety reports provides a two-dimensional plane for direct visual comparison of narratives. Where narratives are proximate as a bag-of-words, they appear closer together in the projection. Where a narrative shares little lexical content with other reports, greater separation from other points is evident. Thus, outliers within the corpus are visually represented. Metadata may then be overlaid providing the opportunity for further comparison.

2.1. ASRS taxonomy

The ASRS database, established in 1976, is one of the earliest examples of a non-punitive reporting system. This repository was developed without the benefit of knowledge of prior reporting programs. As a direct result of its pioneering status, there are a number of characteristics of the framework that, in hindsight, are seen as limitations. As with any taxonomy, shortcomings discovered through usage are increasingly difficult to overcome as more data is added. Any changes to the information retained or its taxonomy, render all prior work on the data necessarily obsolete.

The taxonomy of the ASRS database consists of a narrative and a series of categories and sub-categories each with a series of codes available. The assessment category provides an assessment by a minimum of two subject matter experts of the event narrative. The expert analysts then allocate codes to the sub-categories, *contributing factors* and *primary problem*. The taxonomy published by ASRS does not provide explicit guidance for code usage for analysts. The consistency of results and agreement between analysts is a matter of speculation in the literature as such details remain unpublished by ASRS (Beaubien and Baker, 2002).

2.2. Latent semantic analysis

The methodology of latent semantic analysis, developed by Deerwester et al. (1990), has seen many applications since original publication. This process has subsequently been applied to diverse fields of inquiry, from educational theory to automated document classification (Landauer and Dumais, 1997). LSA may be applied to incident analysis narratives which are commonly used to manually discriminate factors developed through traditional categorical statistical methods.

LSA is a mathematical technique for inferring relations between words within bodies of text. The LSA process first extracts the occurrence of words in a text and creates a term frequency document matrix. A singular value decomposition (SVD) is applied to the resulting matrix. The central matrix from the SVD is then truncated by the substitution of the lowest values with zero. This truncated, or reduced space, form of the matrix then provides the inferred relationships between terms used in similar contexts. In this reduced space, term associations are made that are not present in any single document. Thus, latent contextual relationships are revealed by this method.

2.3. Isometric mapping

Isometric mapping is a mathematical method for reducing the dimensionality of a set of data in a non-linear best-fit manner. Developed by Tenenbaum et al. (2000), this process allows analysts to find lower dimensional patterns within higher-dimensional data sets. If the vector produced by LSA is considered as a point in a high dimensional space, it may be compared to its neighbors locally and mapped onto a non-linear manifold for comparison in a reduced dimensional space.

The multi-step process required to produce the lower dimensional representation may be demonstrated by a simple analogous example transforming a three dimensional pyramid into a two-dimensional fit. The coordinates for the four vertices of a triangle based pyramid may each be represented by a vector with three dimensions. The euclidean distance between each of the vertices may then be calculated.

Provided a notecard for each vertex, the index of the vertex and the distance between that vertex and its neighbors may be written down. Four notecards, each with their index and the distance to their indexed neighbors, will result. The four notecards may then be placed on a table, a two-dimensional plane, and arranged for their best fit with the distances on the cards. Since the pyramid cannot be replicated on a two-dimensional plane, no positioning of the cards will result in an arrangement where all distances on the cards are correct. Given any arrangement of note cards on the table, a measure of fit may be calculated. Through an iterative process, the arrangement may be revised until the best fit is found. The final calculated fit is termed the reconstruction error.

3. Method

Narratives are taken from the given database and parsed to create a dictionary. The dictionary words may then be stemmed (e.g. ‘engines’ becomes ‘engine’), stop words removed (e.g. I, am, the), and subject specific translations made (e.g. ‘twr’ becomes ‘tower’). Although given a sufficiently sized and focused training set these techniques are not strictly necessary, they often result in significant improvements in overall accuracy for machine learning tasks. However, each method requires tuning to the specific learning task.

For each document within the corpus an array is generated describing the number of occurrences of the words found within the dictionary. The dictionary may be truncated to remove the least common words and reduce the sparsity of the array for each document. To prioritize less common words found in the array and to generate a normalized length array form of the corpus the term frequency inverse document frequency (TFIDF) of Salton and Buckley (1988) may be used. This approach is a customary frequency-driven method for preprocessing a corpus to avoid the need for the other more time-intensive methods of preparing the corpus (Aggarwal and Zhai, 2012, pg. 50, 80).

The array form of the corpus is then subject to a singular value decomposition in which the central matrix is reduced. The degree of term reduction in the central matrix is a question of optimization for the given corpus. Where the task is not one with a desired outcome, as with many machine learning tasks, the choice of terms to retain may be made based on the information complexity maximization approach of Robinson et al. (2015).

The resulting document arrays in the semantic space are used by the isometric mapping process. The isomap requires dimensionality of the manifold along with the number of neighbors, calculated by euclidean distance, to develop the manifold geometry. In this approach the manifold is restricted to two-dimensions and the number of nearest neighbors must be determined from the dataset.

3.1. Corpus selection

Where a principal goal of the visualization process is to express relationships between otherwise latent factors, the selection of the corpus is of primary importance. Narratives within the corpus are necessarily encoded into written word by the individual who experienced the event. Thus, language and experience form layers of interpretation in the narrative. Where relations are inferred