Work 41 (2012) 188-197
DOI: 10.3233/WOR-2012-0155-188
IOS Press

Application of diffusion maps to identify human factors of self-reported anomalies in aviation

Chris Andrzejczak[a], Waldemar Karwowski[a], and Piotr Mikusinski[b]

[a]*Department of Industrial Engineering and Management Systems, 4000 Central Florida Blvd, P.O. Box 162993, Orlando, FL 32816-2993, United States of America*

[b]*Department of Mathematics, 4000 Central Florida Blvd, P.O. Box 161364 Orlando, FL 32816-1364, United States of America*

Abstract. A study investigating what factors are present leading to pilots submitting voluntary anomaly reports regarding their flight performance was conducted. Diffusion Maps (DM) were selected as the method of choice for performing dimensionality reduction on text records for this study. Diffusion Maps have seen successful use in other domains such as image classification and pattern recognition. High-dimensionality data in the form of narrative text reports from the NASA Aviation Safety Reporting System (ASRS) were clustered and categorized by way of dimensionality reduction. Supervised analyses were performed to create a baseline document clustering system. Dimensionality reduction techniques identified concepts or keywords within records, and allowed the creation of a framework for an unsupervised document classification system. Results from the unsupervised clustering algorithm performed similarly to the supervised methods outlined in the study. The dimensionality reduction was performed on 100 of the most commonly occurring words within 126,000 text records describing commercial aviation incidents. This study demonstrates that unsupervised machine clustering and organization of incident reports is possible based on unbiased inputs. Findings from this study reinforced traditional views on what factors contribute to civil aviation anomalies, however, new associations between previously unrelated factors and conditions were also found.

Keywords: Data Mining; Dimensionality Reduction; Text Records; Clustering; Incident Reports

1. Introduction

The demand for worldwide air travel continues to increase. In addition, the business model of commercial aviation is continually evolving, with more direct flights between city-pairs on smaller aircraft replacing more traditional “spoke and hub” operations where large aircraft connect major cities, thus requiring de-boarding and connecting flights on smaller aircraft. The changing operational environment comprises longer flight times between city-pairs, increased aircraft capability and complexity, and more airplane traffic, all of which, taken together, create many opportunities for anomalous events.

Accidents in day-to-day aviation operations are rare. However, aircraft anomalies occur much more frequently. These anomalies mimic conditions that lead to accidents. Understanding what factors are present and how they are associated with anomalies can lead to methods aimed at reducing or otherwise managing factors before they lead to serious incidents.

Nagel [1] reports that 90 percent of aviation mishaps are labeled as and attributed to human error. Studies conducted by Lautman and Gallimore [2] report that about 70 percent of accidents in commercial jet transport can be attributed to crew error. This percentage is consistent over any time period under review, and has not changed in recent times despite implementations of new technologies and findings from human factors and related safety disciplines. An understanding of what factors are present when anomalous events occur will strongly aid in managing or preventing future anomalies.

The Federal Aviation Administration (FAA) and National Transportation Safety Board (NSTB) keep detailed accident reports of commercial aircraft incidents, in an effort to use knowledge gleaned from such incidents to prevent future problems. Analysis of this data using methods from the fields of traditional statistical analysis, human factors studies, clustering, and dimensionality reduction may yield new information on causes and provide insights into what conditions are present for similar error types. This information may be used to influence design, training, or operations that have the potential to further reduce error.

Today's high performance computers and the vast storage capabilities of these computers constitute unprecedented opportunities for data creation and archiving. Many accident and incident databases exist, yet pertinent information may be overlooked in all of these data. Methods are required to reduce the dimensionality and "noise" in all of these data while leaving relevant structural information intact.

Maintaining vast stores of information is only useful if this information is organized and retrievable when needed. Classifying the data using human operators is tedious and possibly inaccurate, as two individuals may classify the same record differently. Semi-supervised and unsupervised accurate, reliable classification algorithms and applications would greatly increase the value of maintaining vast data stores of incident or anomaly data.

Dimensionality reduction is a topic that has received recent attention. There is a staggering amount of data being created every day. Accessing, categorizing, and using this information requires organization. To handle "real-world," often unstructured, high-dimensional data accurately and with minimal computational load is a challenge for mathematicians and computer scientists alike. In a review by van der Maaten et al [3] efficient data representations should have a dimensionality level that approaches the intrinsic dimensionality of the original data. Fukunaga (1990) defines intrinsic dimensionality as the minimum number of parameters necessary to represent the functional properties of the data. The below displays a taxonomy of dimensionality reduction techniques reproduced from van der Maaten [3].

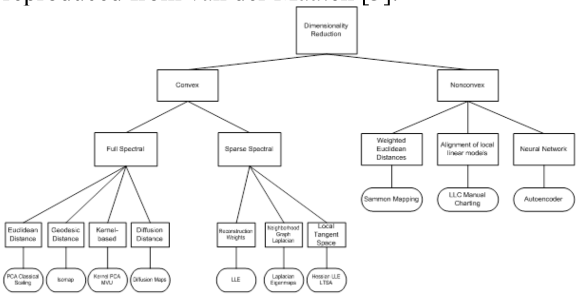


Figure 1: Taxonomy of dimensionality reduction techniques (van der Maaten, 2009)

Diffusion maps are one of many techniques for dimensionality reduction. When dealing with multiple categories of data or many factors, visualizing the data to deduce meaning can be challenging or impossible. Dimensionality reduction assumes that the data observed has some sort of structure or logical order to it, and therefore could be reduced to a dimensional level low enough to be represented or visualized. For this to occur it is assumed that the interesting data can be represented on a non-linear manifold, or mathematical space. A manifold is an abstract mathematical concept where points exist in a domain that resembles Euclidean space. Manifolds of a sufficiently low dimensional level can be plotted or otherwise represented visually. Dimensionality reduction techniques have been successfully employed in machine learning, mapping, and clustering activities. When reducing dimensionality the goal is to maintain any underlying structures or patterns in the high-dimensional data.

Diffusion maps work by embedding high dimensional data onto low dimension Euclidean space. This is done through the eigenvectors of defined random walks performed on the data. The data is assumed to be randomly sampled from an underlying general probability distribution:

$$p(\mathrm{x}) = e^{-U(\mathrm{x})} \tag{1}$$

As the number of samples approaches infinity, eigenvectors of each diffusion map converge to the eigenfunctions of a corresponding differential operator defined on the support of the probability distribution [5]. Diffusion maps have the added benefit that, when properly employed, they quickly converge on a meaningful scheme or result. In most applications diffusion maps are unsupervised when employed. Coifman et al [6] present a general framework for diffusion maps, and demonstrate diffusion maps' effectiveness in exploring geometry, statistics, and functions of data. The authors also demonstrate how diffusion maps afford a low-dimensional embedding of high-dimensional data. This process is naturally suited for visualization, clustering, and regression.

The diffusion map algorithm process as described by Bah [7] employs four major steps, summarized below. It assumes the data has already been modeled by a weighted graph.

1. An adjacency matrix of the graph is created
2. A diagonal k × k normalization matrix and Laplacian matrix of the graph are calculated
3. Eigenvalues and eigenvectors are computed of these two matrices
4. The lowest value initial eigenvector is dropped, and the next m eigenvectors are used to represent the n-dimensional space.

2. Method

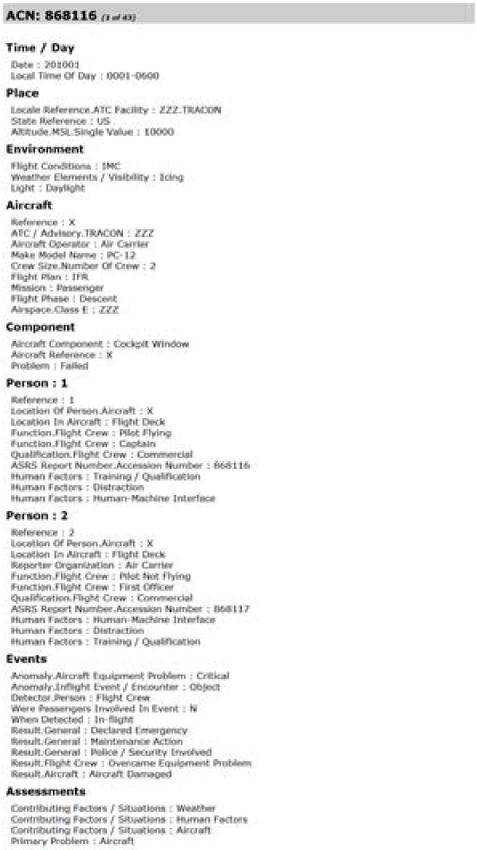
Data was provided via an ASCII text dump of the Oracle database employed by ASRS. These data were extracted, merged, and categorized by type. All available data as of March 15, 2010 was used.

The below figure shows a sample ASRS record. The record contains identifiers, information on weather, pilot experience, contributing factors to the incident, professional subjective assessments, and a narrative description written by the report submitter (usually the pilot experiencing the incident). The submission process is completely anonymous, and after entry into the electronic database the original submitted paper record is destroyed.

One fascinating aspect pointed out by Bah (2008) is that diffusion maps may mirror biological functions. For example, the human brain is constantly bombarded by unstructured, highly dimensional stimuli. Diffusion maps may model the biological analogs that perform natural clustering and categorizing applications to make sense of surroundings.

Lafon and Lee (2006) describe a unified framework for employing diffusion maps to reduce dimensionality and cluster documents according to the words contained within them. Use of k-means clustering in diffusion space allows this categorization, and the authors also propose a measure of clustering accuracy used to assess the results given by the algorithm.

Dimensionality reduction has found recent use in extracting information from a corpus of text documents. Underhill (2007) states that manipulating large amounts of text data can be extremely computationally intensive. A reduced dataset with relevant meanings intact would be extremely useful in information extraction efforts. In addition, such information extraction could be unsupervised and automated, providing a way to manage the incredible amounts of information being generated.



ACN: 868116 (1 of 43)

Time / Day

Date : 201001
Local Time Of Day : 0001-0600

Place

Locale Reference.ATC Facility : ZZZ.TRACON
State Reference : US
Altitude.MSL.Single Value : 10000

Environment

Flight Conditions : IMC
Weather Elements / Visibility : Icing
Light : Daylight

Aircraft

Reference : X
ATC / Advisory.TRACON : ZZZ
Aircraft Operator : Air Carrier
Make Model Name : PC-12
Crew Size.Number Of Crew : 2
Flight Plan : IFR
Mission : Passenger
Flight Phase : Descent
Airspace.Class E : ZZZ

Component

Aircraft Component : Cockpit Window
Aircraft Reference : X
Problem : Failed

Person : 1

Reference : 1
Location Of Person.Aircraft : X
Location In Aircraft : Flight Deck
Function.Flight Crew : Pilot Flying
Function.Flight Crew : Captain
Qualification.Flight Crew : Commercial
ASRS Report Number.Accession Number : 868116
Human Factors : Training / Qualification
Human Factors : Distraction
Human Factors : Human-Machine Interface

Person : 2

Reference : 2
Location Of Person.Aircraft : X
Location In Aircraft : Flight Deck
Reporter Organization : Air Carrier
Function.Flight Crew : Pilot Not Flying
Function.Flight Crew : First Officer
Qualification.Flight Crew : Commercial
ASRS Report Number.Accession Number : 868117
Human Factors : Human-Machine Interface
Human Factors : Distraction
Human Factors : Training / Qualification

Events

Anomaly.Aircraft Equipment Problem : Critical
Anomaly.Inflight Event / Encounter : Object
Detector.Person : Flight Crew
Were Passengers Involved In Event : N
When Detected : In-flight
Result.General : Declared Emergency
Result.General : Maintenance Action
Result.General : Police / Security Involved
Result.Flight Crew : Overcame Equipment Problem
Result.Aircraft : Aircraft Damaged

Assessments

Contributing Factors / Situations : Weather
Contributing Factors / Situations : Human Factors
Contributing Factors / Situations : Aircraft
Primary Problem : Aircraft

Figure 2: Typical ASRS Record

Text data is inherently unstructured and contains data that, if plotted, is of high dimension. Each word in a text document, for example, could be considered a dimension. Thus a 100-word record has 100 dimensions with which to contend. To manage this, methods were modified from those used by Underhill [8] and Martinez [9]. An unsupervised approach to dimensionality reduction was chosen, as the literature indicated that a need exists for unsupervised dimensionality reduction in text mining. Dimensionality reduction techniques have seen varied uses in clustering and categorizing data. For example, Higgs et al [10] successfully employed dimensionality reduction through diffusion maps to classify brain images according to species of animal from which brain scans were taken.

The method employed is that described by Lafon and Lee [11] where a diffusion distance between terms can be approximated to a level of precision given by δ by observing the first few $q(t)$ nontrivial eigenvalues in the following relation:

$$D_t^2(x,z) \cong \sum_{j=1}^{q(t)} \lambda_j^{2t}\left(\psi_j(x)-\psi_j(z)\right)^2$$

Lafon and Lee [11] explain that the above relation can be interpreted as a Euclidean distance in the linear map $\mathbf{R}^{q(t)}$ if the right eigenvectors are selected with λ_j^t coordinates on the data. The following diffusion map:

$$\Psi_t : x \rightarrow \begin{pmatrix} \lambda_1^t \psi_1(x) \\ \lambda_2^t \psi_2(x) \\ \vdots \\ \lambda_{q(t)}^t \psi_{q(t)}(x) \end{pmatrix}$$

describes the relation below

$$D_t^2(x,z) \cong \sum_{j=1}^{q(t)} \lambda_j^{2t}\left(\psi_j(x)-\psi_j(z)\right)^2 = \|\Psi_t(x)-\Psi_t(z)\|^2$$

The dimensionality reduction and selection of the relevant eigenvectors is dictated by the fall-off of the eigenvalues and other factors described in more detail in Lafon and Lee [11]. The main idea of this process is that the distance measures between Anomaly records are preserved in the dimensionality reduction, these measures then afford classification by *k*-means clustering.

The Term-Frequency Inverse-Document Frequency equation specified by Underhill [8] was employed to create the input data, and dimensionality reduction using diffusion maps was carried out on the resulting matrix. The computed document difference matrix was chosen as a measure of document dissimilarity. The underlying theory was that the measures of differences between the documents, when reduced, would suggest what level of dimensionality is required to categorize the documents. The below figure, modified from its original version found in Underhill [8] describes the process.

The dimensionality reduction and selection of the relevant eigenvectors is dictated by the fall-off of the eigenvalues and other factors described in more detail in Lafon and Lee [11]. The main idea of this process is that the distance measures between Anomaly records are preserved in the dimensionality reduction, these measures then afford classification by *k*-means clustering.

The Term-Frequency Inverse-Document Frequency equation specified by Underhill [8] was employed to create the input data, and dimensionality reduction using diffusion maps was carried out on the resulting matrix. The computed document difference matrix was chosen as a measure of document dissimilarity. The underlying theory was that the measures of differences between the documents, when reduced, would suggest what level of dimensionality is required to categorize the documents. The below figure, modified from its original version found in Underhill [8], describes the process.

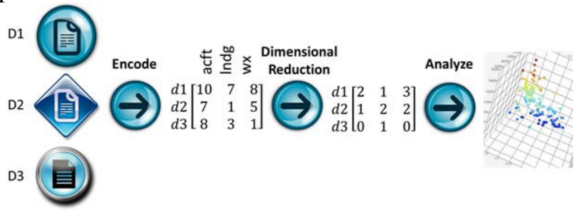


Figure 3: Dimensionality reduction approach (adopted from Underhill, 2007)

Diffusion maps aim to transform distance matrices that highlight local relationships between points [6]. These preserved relationships are based on the number of paths that exist between two data points; they describe how anomaly types are connected [8].

PASW Modeler 13 was used to extract the most common words associated across all records with pilot anomalies. This action returned 127,766 records. Over 5,000 commonly-occurring words were extracted. The words were sorted in descending order by the number of documents that contained them. For example, the most commonly occurring word was “acft” (aircraft), and it was found in 41 percent, or 52,502, records. Due to memory and software constraints which will be described later, only 100 of these words could be used for dimensionality reduction. The 100 words chosen were the most frequently occurring words; these 100 words served as the input dataset for the clustering activity.

A Microsoft Excel document was created that listed these 100 words across the top row arranged by columns, each column containing a word. PASW Modeler 13’s category extraction feature was employed to create a sparsely populated term-document frequency matrix that indicated presence or absence as well as frequency of encounter of a given word within the anomaly record.

This matrix was then used to calculate a document feature. The method chosen was described by Underhill [8], and is called the weighted term-document matrix. To create this matrix, a slightly modified version of the Term-Frequency Inverse-Document Frequency formula was employed:

$$T_{i,j} = (t/T) * ln(D/d)$$

In this equation, t is the frequency value of a word j appears in document i. The sum total of all words of interest (row sums) that appeared in a given record was term T. The term D is the total number of documents (127,766), and d is the number of documents that contain the term j. This equation led to the creation of a term-document matrix, which then could be visualized and its dimensionality reduced in MATLAB.

MATLAB r2007b was used to carry out the dimensionality reduction, with the actual computation carried out using freely distributable example code developed by Ann B. Lee, Associate Professor within the Department of Statistics at Carnegie Mellon University at the time of writing of this work. The code was accessed from Professor Lee’s personal webpage [12]. The code was modified in MATLAB to accept 1,000 anomaly report records containing 100-item term-document matrix. A random sample of 1000 anomaly reports was selected. The below table displays a truncated, representative sample of the input data.

3. Results

The dimensionality reduction classification activity was performed on 100 of the most commonly occurring words. This was done to make the effort as “unsupervised” as possible, meaning that no investigator input was required to choose the inputs. The 100 words were transformed into a document difference matrix, and this matrix was reduced by the Diffusion Maps method described by Lafon and Lee (2006). The results of the classification are shown below. The diffusion maps algorithm creates large matrices, namely a pairwise distance matrix using MATLAB’s pdist() command. This command has memory and computation limitations that become quickly apparent when large datasets are employed. To accommodate this limitation a reduced dataset of only 1,000 records was used.

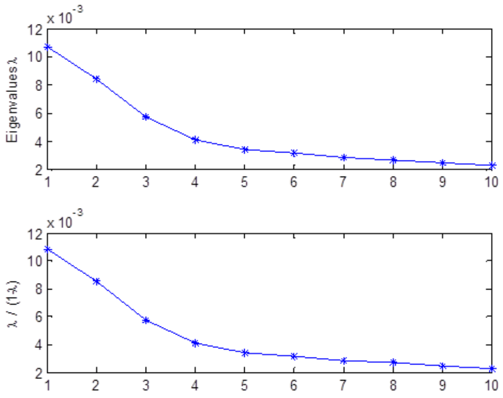


Figure 4: Falling eigenvalues plot from dimensionality reduction

The dimension reduction activity produced the above figure. The first nontrivial eigenvalues describe paths that the underlying structure of the data implies. The various vectors that describe the data in the forms of word presence or absence may follow common paths, and these eigenvalues describe those vectors that account for much of the underlying structure [11]. The fall-off of these eigenvalues dictates the dimensionality reduction and weighting of relevant eigenvectors.

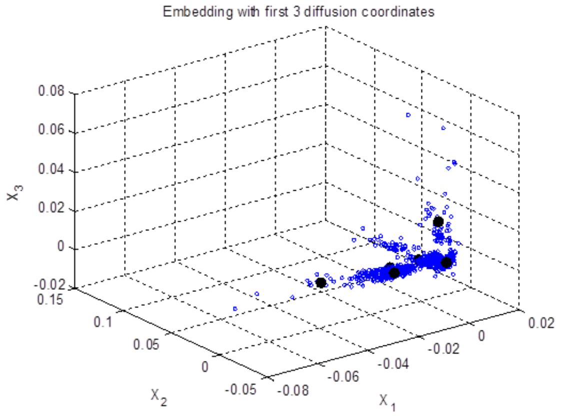


Figure 5: Diffusion map of the first three embedding coordinates

The above figure displays the embedding of the first three diffusion coordinates. The data are clustered around these coordinates. These arrangements describe the underlying structure of the data. These data were employed in a *k*-means clustering algorithm to classify the anomaly records by words contained within them according to anomaly types. The above graph is a realization of a cloud of points where the rescaled eigenvectors are the coordinates. The above graph is a lower dimensionality representation of the data that reveals underlying structure.

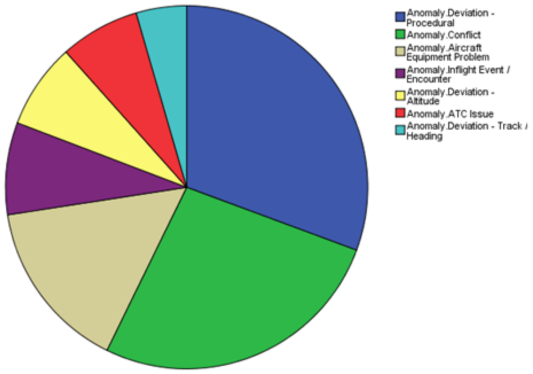


Figure 61: Composition of the dimensionality reduction dataset – actual anomalies

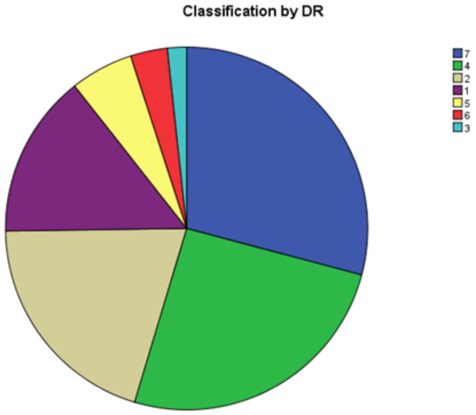


Figure 72: Composition of the dimensionality reduction dataset – classification results

The reduced 1,000 record dataset contained a representative sample of the data. The above figures and tables describe the data in terms of frequencies and visual distributions. After running the algorithm, a *k*-means labeling of the outputs was run to classify the records.

The algorithm created a distribution of records very similar to the records classification distribution given by the ASRS database administrators. When comparing the output of the algorithm to the predetermined categories assigned by the ASRS database administrators the accuracy is 23.8 percent. These findings suggest that dimensionality reduction can be used to classify anomaly reports in an unsupervised fashion. By using more salient, relevant keywords in the process, this method could yield far better results as the algorithm focuses on using these features to classify the data. The fact that the frequencies of classification are so similar to classification activities performed by the ASRS database administrators suggests that the algorithm identifies features in the dataset. Most records had multiple classifications and contributing factors (however only the first one was used) and the dimensionality reduction activity may focus on a feature that was not used by the human ASRS administrator.

4. Discussion

Dimensionality reduction was performed using the Diffusion Maps algorithm within code freely distributed by Ann Lee [12]. The algorithm demonstrated a limited ability to classify records simply based by word frequencies. This operation can be greatly improved with appropriate selection of words, as their meanings were not considered for the extraction process. This method was completely unsupervised. The findings demonstrate a proof-of-concept that suggests completely unsupervised classification of data is possible using just dimensionality reduction algorithms. Furthermore, the text data fed to the algorithm can be raw. This method assumes that the text data does exhibit meaningful underlying structure.

The study limited the number data sources to a single large repository, the NASA Aviation Safety Reporting System. This repository contained only voluntary reports submitted by pilots flying both private and commercial operations. No military data was included in this dataset. The database was chosen because the literature suggested that, aside from NASA's internal metrics and monthly publication summarizing relevant, current anomalies with commentary, no large-scale analysis had been conducted on the data.

The encoded nature of the ASRS database, with its many abbreviations, excluded the possibility of a context-link analysis using a traditional dictionary. The data would have had to be decoded; this was deemed unfeasible due to the number of encoded terms. In addition, much of aviation terminology is rife with acronyms, abbreviations, and non-standard technical terms. A context analysis, though very powerful, is usually limited to full-text sources such as web pages and interview or survey data.

The diffusion maps aspect of the study did not have an official, well-developed software application. Diffusion maps for document clustering are a relatively novel idea at the time of this study; most other studies that classified documents had fewer numbers of larger documents containing full text. Of those studies, the classifications were usually binary in nature. Other studies also reported limited success in classifying documents with diffusion maps; e.g., Underhill [8] reported a 27 percent accuracy rate. The rate reported in this study was 23.8 percent using a completely unsupervised input dataset. Records from the ASRS dataset usually had multiple classifications, and it is entirely possible that the unsupervised classification algorithm focused on a different feature or identified a feature that the human classifier did not.

The techniques used to create the document difference matrices were not exceedingly sophisticated. Dimensionality reduction performs best with large samples of data, while the method used to handle the vectors of word presences was insufficient. Microsoft Excel was used for the task, and the software's limits were quickly reached when attempting to create a matrix sufficient to encompass the 100 columns of words and 127,776 rows of text records. A reduced dataset of 1000 records was used in its stead.

Future studies in this area might investigate partitioned data to identify additional opportunities for insight that could be gleaned from focusing on specific areas or eras of aviation. Other databases, ones not limited to aviation, could also be used. Many organizations employ feedback systems and surveys, and these domain-specific text records could be analyzed in a similar fashion.

Other dimensionality reduction techniques could also be employed. There are many techniques, both linear and nonlinear, that can be applied once data is properly encoded. One technique often cited in the literature was Principle Component Analysis PCA; it was not employed in this study. Additionally, the method of encoding anomaly reports could be altered to generate better results, or else a more sophisticated or better-equipped tool for dimensionality reduction could be employed. A software package that could encode raw text, apply the appropriate dimensionality reduction, and suggest or perform appropriate analyses on the output would be highly beneficial to this and other domains employing Literature Based Discovery (LBD).

Finally, this study employed unsupervised, semi-supervised, and supervised approaches. The approaches that provided the most meaningful results were those that were supervised by a human analyst. Future studies could develop wholly unsupervised methods and techniques. The method of performing keyword extraction was highly rudimentary, especially for the dimensionality reduction element of the study. Future studies may perform more in-depth and comprehensive keyword extraction, in order to create the most accurate possible relationships and associations, as well as providing a richer feature set for the dimensionality reduction activity to act upon.

Although the possibility of achieving a perfect safety record and zero percent accident rates for the aviation industry is highly unlikely, it is necessary to reduce the current accident rate to accommodate the up and coming drastic increases in worldwide air traffic. The aviation market demands steady increases in performance and safety, and only through diverse, multidisciplinary, and systematic employment of findings from studies such as these will this demand be realized. Sheridan [13] describes that changes to engineering itself are necessary; claiming that engineers are often focused on designing *things*, when really their focus should be designing *relationships to people*. It is this kind of thinking that often results in the "wild, unpredictable, *organic*" human to be blamed for incidents rather than the orderly and digital technology. Understanding that the human can also be the savior of the system and designing with the intention of exploiting these "saving features" should be the focus of design, training, and research. Only through collaboration across disciplines and integration of meaningful findings can powerful, self-correcting, and sustainable practices emerge that will guide the aviation field into its exciting future.

Table 1

Truncated input data for MATLAB

apch	rwy	flt	turn	twr	acft	clred
0	1	0	0	0	0	0
1	0	1	1	0	0	0
0	1	0	0	0	1	0
0	1	0	1	0	0	0
0	1	0	1	0	1	0
1	0	0	0	0	1	0
0	0	0	0	0	1	0
0	1	0	0	0	1	0
0	1	0	0	0	1	0
0	1	1	0	0	0	1
0	0	0	0	0	0	0
0	1	0	0	1	1	1
0	1	0	0	0	1	0
0	0	0	1	0	1	0
0	0	1	0	0	1	0

Table 2: Actual anomaly frequencies from reduced dataset

Anomaly Type	Frequency	Percent
Anomaly.Deviation - Procedural	307	30.6
Anomaly.Conflict	266	26.6
Anomaly.Aircraft Equipment Problem	153	15.3
Anomaly.Inflight Event / Encounter	82	8.2
Anomaly.Deviation - Altitude	76	7.6
Anomaly.ATC Issue	71	7.1
Anomaly.Deviation - Track / Heading	45	4.5
Total	1000	100.0

Table 3: Category frequencies as created by *k*-means labeling

k-means label	Frequency	Percent
7	292	29.1
4	255	25.5
2	201	20.1
1	146	14.6
5	56	5.6
6	33	3.3
3	17	1.7
Total	1000	100.0

References

[1] Nagel, D. (1988). Human error in aviation operations. In E. Weiner and D. Nagel (Eds.), Human factors in aviation. Academic Press: San Diego, CA.

[2] Lautman, L. G., & Gallimore, P. L. (1987). Control Of The Crew-Caused Accident : Survey Of 12 Airline Operators Reveals Techniques That Contribute To Safe Operations. Air Line Pilot, 56(10).

[3] Van der Maaten, L., Postma, E., & Van den Herik, H. (2007). Dimensionality reduction: A comparative review. Published online, 10(February), 1–35.

[4] Fukunaga, K. (1990). Introduction to statistical pattern recognition. Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[5] Nadler, B., Lafon, S., Coifman, R.R., Kevrekidis, I. (2006). Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. Applied Computational Harmonics Analysis. Vol 21, 2006, pp 113-127.

[6] Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., & Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proceedings of the National Academy of Sciences of the United States of America, 102(21), 7426 -7431. doi:10.1073/pnas.0500334102

[7] Bah, B. (2008). Diffusion Maps: Analysis and Applications. Master’s Thesis submitted to University of Oxford.

[8] Underhill, D.G. (2007). Exploring dimensionality reduction for text mining. Trident Scholar project report no 362. United States Naval Academy, Annapolis, Maryland

[9] Martinez, A. R., and Wegman, E. J., (2002). Text stream transformation for semantic-based clustering. Proceedings of Computing Science and Statistics, 34, 2002.

[10] Higgs, B., Weller, J., & Solka, J. (2006). Spectral embedding finds meaningful (relevant) structure in image and microarray data. BMC Bioinformatics, 7(1), 74. doi:10.1186/1471-2105-7-74

[11] Lafon, S., & Lee, A. B. (2006). Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 28(9), 1393–1403.

[12] Lee, Ann. (2010). Personal Webpage. http://www.stat.cmu.edu/~annlee/software.htm

[13] Sheridan, T.B. (2010). The system perspective on human factors in aviation. In E. Sala and D. Maurino (Eds.), Human factors in aviation. Academic Press. San Diego, CA.