Your Knowledge Partner™

Application of PolyAnalyst to Flight Safety Data at Southwest Airlines

Proof-of-Concept Demonstration of Data and Text Mining

Prepared by:
Megaputer Intelligence
120 W. 7th Street, Suite 310
Bloomington, IN 47404

Project Manager:
Sergei Ananyan s.ananyan@megaputer.com

Primary Analysts:
Richie Kasprzycki rkasprzy@megaputer.com
Vijay Kollepara vkollepara@megaputer.com

GAIN®

SOUTHWEST AIRLINES®

Megaputer
Your Knowledge Partner™

January 2004

DISCLAIMER

This project report and the results are based on a time-constrained proof-of-concept demonstration carried out by Megaputer Intelligence for Southwest Airlines and involve analysis of only a subset of data. Therefore, the results should not be used to form any conclusions or business decisions.

Table of Contents

Acknowledgements

Megaputer Intelligence Inc. would like to thank all the individuals and members of Working Group B (Analytical Methods and Tools) of the Global Aviation Information Network (GAIN) and Southwest Airlines, who made this report possible.

A special mention needs to be made of the following people for their active support and timely advice that were received at different stages of the project and in the preparation of this report.

Southwest Airlines Project Participants

Captain Mike Hinnenkamp, ASAP Manager, Southwest Airlines
Tim Logan, Director, Flight Operational Safety, Southwest Airlines
Wy Teter, ASAP Administrator, Southwest Airlines
Tem Youngblood, Senior Database Administrator, Southwest Airlines
Alex Hwang, Senior Database Administrator, Southwest Airlines

GAIN Working Group B Members

Carolyn Edwards, Federal Aviation Administration
Christina Hunt, Phaneuf Associates Incorporated
Andy Muir, Federal Aviation Administration
Lee Nguyen, Federal Aviation Administration
Grant Schneemann, Abacus Technology Corporation

Report Authors

Dr. Sergei Ananyan, Megaputer Intelligence
Richie Kasprzycki, Megaputer Intelligence
Vijay Kollepara, Megaputer Intelligence

Project sponsored by the Federal Aviation Administration, Office of System Safety.

Executive Summary

Background

Southwest Airlines and Megaputer Intelligence conducted a joint proof-of-concept project in support of the Global Aviation Information Network (GAIN) Working Group B efforts to facilitate and promote the use of automated data and text mining tools in the aviation safety community.

Megaputer Intelligence provided the analytical software (PolyAnalyst™) and its usage expertise, and Southwest Airlines provided de-identified reports from their Aviation Safety Action Program (ASAP) database along with guidance and insight on the relevancy of the results. The ASAP database contains voluntary feedback from pilots on abnormal occurrences in different phases of flight.

The project spanned a six-week period wherein different analytical methodologies were demonstrated to the Flight Safety Officers (FSOs) to identify potential safety issues from the pilot narratives. The emphasis of the project was to demonstrate new knowledge by conducting a comprehensive analysis of all available data to discover trends, associations, and correlations.

Results

PolyAnalyst provided Southwest Airlines with a framework to conduct analysis across the entire ASAP report database. Both the structured and the unstructured (narrative) portions of the database were analyzed. The Southwest Airlines participants found the following specific capabilities of PolyAnalyst, which were demonstrated and used in the analysis, to be of particular value:

- <u>Automatic extraction of important patterns of terms from pilot narratives</u>. PolyAnalyst enabled rapid extraction of stable patterns of terms occurring together in separate pilot narratives and presented results of the analysis in a convenient visual form.
- <u>Focusing the analysis on specific categories of interest</u>. PolyAnalyst allowed the extraction and summarization of all terms related to a particular category – for example, "equipment" – without the user specifying in advance all possible equipment pieces that might be encountered in the text.
- <u>Simultaneous analysis of structured and unstructured data.</u> The capability of PolyAnalyst to analyze all available data at once, both structured and unstructured, helped generate additional insights in the analysis of safety related events.
- <u>Ability to process large amounts of data</u>. By providing automated tools for the analysis of narrative data, PolyAnalyst allowed FSOs to explore more ASAP reports during an analysis session than was previously possible.
- <u>Quick generation of a variety of useful graphs capturing key findings</u>. PolyAnalyst provided several efficient graphing capabilities for the visualization of the relationships discovered in the ASAP data, providing easy means of detecting potential safety concerns.

- Intuitive drill-down from the graphs to original reports. The drill-down capabilities of PolyAnalyst enabled visual dynamic querying of the database, which had a significant timesaving effect over form-based static querying.

The results helped the Southwest Airlines safety managers locate some potential issues across different aircraft types and airports. The safety managers concluded that a number of direct tangible benefits of deploying the system could be achieved:

- A wide range of techniques and visualizations that can enhance the quality and range of the analysis.
- Ability to identify hidden knowledge and patterns in the ASAP narratives.
- Ability to provide explicit and actionable results that can be used by the FSO to quickly identify anomalies and rectify them.
- Higher speed of analysis and drill down, making the process efficient.

Overall the project demonstrated that a synergetic combination of automated text analysis and visual presentation of discovered clusters and correlations can significantly reduce potential biases in the analysis, automate the most time-intensive operations and increase the thoroughness and quality of the results. Identifying hidden issues and root causes of problems can be done proactively rather than waiting for these issues to exacerbate into accidents. The project demonstrated an avenue to analyze all available data (both structured and textual) in order to derive maximum value from information technology and data collection investments.

1.0 Introduction

1.1. Purpose of the Proof-of-Concept Demonstration

This proof-of-concept demonstration is part of the Global Aviation Information Network (GAIN) Working Group B's (Analytical Methods and Tools) efforts to facilitate and promote the use of automated data and text mining tools in the aviation community for improving overall flight safety performance. The project proposes new techniques and methodologies to conduct timely analysis of flight safety data to reveal associations and trends that may otherwise be difficult and time consuming to identify. It is also GAIN's desire to share the knowledge of this demonstration with others in the aviation community.

Aviation safety experts surmise that accidents are usually a culmination of a series of unsafe events that had gone unnoticed. For every accident and major event that is thoroughly investigated, there can be as many as 300 minor events or hazard reports (Heinrich's pyramid[1]) that could have contained some information about the impending event. The industry has placed significant investments into collecting and collating this aviation safety information from multiple sources.

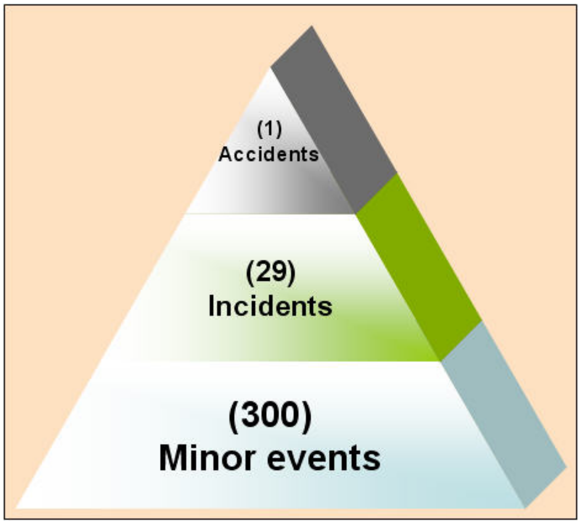


Figure 1: Heinrich's Pyramid

Though these databases contain significant amounts of critical data, there have been enormous challenges in analyzing the information. Analysis has primarily been focused on only the structured portion of the database. Yet experts estimate that over 80% of all information in a report is in the textual (unstructured) format and could contain nuggets of valuable knowledge.

This project demonstrates an avenue to analyze all available data (both structured and textual) in order to derive the maximum value from these safety data collection investments. Therefore, the primary objectives were to develop a broad methodology of analysis, demonstrate how knowledge dispersed in a large collection of Aviation Safety Action Program reports can be easily revealed, and outline the value of this knowledge.

[1] The accident pyramid, also referred to as the safety triangle was derived from a 1931 study by H. W. Heinrich and detailed in his book, *Industrial Accident Prevention: A Scientific Approach*. Widely accepted in the industry, the pyramid serves to illustrate Heinrich's theory of accident causation: unsafe acts lead to minor injuries, and over time to major injury.

1.2 Overview of Participating Organizations

The proof-of-concept demonstration is the culmination of a joint effort by GAIN Working Group B, Megaputer Intelligence, and Southwest Airlines. Megaputer Intelligence provided the analytical software system PolyAnalyst™ and its usage expertise. Southwest Airlines provided de-identified safety data from the Aviation Safety Action Program (ASAP) database, as well as guidance and insight on the relevancy of results. A brief overview of the participating organizations follows.

Global Aviation Information Network (GAIN)

GAIN is an industry and government initiative to promote and facilitate the voluntary collection and sharing of safety information by and among users in the international aviation community to improve safety. GAIN was first proposed by the Federal Aviation Administration (FAA) in 1996, but has now evolved into an international industry-wide endeavor that involves the participation of professionals from airlines, air traffic service providers, employee groups, manufacturers, major airframe and equipment suppliers and vendors, and other aviation organizations. GAIN Working Group (WG) B, Analytical Methods and Tools, facilitates and promotes the use of analytical methods and tools in the aviation community.

Southwest Airlines

Southwest Airlines is one of the largest airlines in America with $5.5 billion in revenues, and transports over 60 million customers annually. It operates over 2,800 flights serving 59 airports around the U.S, and has a fleet of 375 Boeing 737 aircraft.

Megaputer Intelligence

Megaputer Intelligence provides solutions for analyzing both structured and textual data. Megaputer offers tools for data mining, text mining, and web data analysis, and serves customers worldwide, including organizations in the insurance, aerospace, financial, educational and government sectors.

1.3 Overview of PolyAnalyst Text and Data Mining System

PolyAnalyst is a text and data mining system that provides capabilities ranging from data importing, cleansing and manipulation, to visualization, modeling, scoring and reporting. PolyAnalyst can access data stored in major commercial databases and some proprietary data formats (Excel, SAS), as well as popular document formats. It offers a selection of semantic text analysis, clustering, prediction, and classification algorithms, link analysis, transaction analysis, and visualization capabilities. PolyAnalyst can directly access data from any major commercial database through standard OLE DB (Object Linking and Embedding for Database) or ODBC (Open Database Connectivity) protocols.

Results obtained with PolyAnalyst can provide key insights into different aviation processes, helping safety officers to:

a) Reveal hidden issues (irrespective of data type – structured or unstructured)
b) Generate strategic overview charts for management across different parameters
c) Identify bottlenecks in processes and highlight aircraft part quality or part supplier related issues.

PolyAnalyst provides a set of tools that can address many analytical tasks that safety officers are facing and can be tailored to a specific application domain. A major portion of the user's involvement is in providing direction to the analysis process and defining their areas of interest. User-defined parameters for running analysis engines are entered in the corresponding dialog boxes.

In more advanced implementations of PolyAnalyst with the WebAnalyst™ integration platform, users of the system can record reusable analytical scripts for typical data exploration scenarios. Business users can then execute these scripts with a push of a button and view resulting reports in a preset template format.

1.4 Input Data: Aviation Safety Action Program (ASAP) Data

Southwest Airlines, as part of its Aviation Safety Action Program (ASAP), collects voluntary feedback from pilots on abnormal occurrences in different flight phases. These reports are in a standard format and comprise details about the flight, pilot, weather, and aircraft. They also contain a pilot's description of the event in more detail. Each report is then categorized by a Flight Safety Officer based on a predefined set of categories and entered into the database.

The objective of ASAPs[2] is to enhance aviation safety through the timely identification and correction of potential safety problems before they lead to accidents. Under an ASAP, safety issues are resolved through the implementation of corrective actions rather than through regulatory or disciplinary procedures. ASAP safety data, much of which would otherwise be unobtainable, is used to develop corrective actions for identified safety concerns, and to educate the appropriate parties to prevent a reoccurrence of the same type of safety event.

1.5 Current Processes for Analyzing ASAP Data

Typically, analytical processes currently existing in the industry are primarily appropriate for handling structured data. The analysis revolves around looking for and quantifying known hypotheses and problems. Quantities such as number of 'altitude deviations' or 'fatigue issues' for a particular time period can be easily calculated when the knowledge base (database) is properly categorized into known broad concepts and dynamically updated over time. Currently, analysis is conducted on the categorical data only by manual query, keyword search, and reporting software.

Southwest Airline's ASAP data resides in an Oracle database. Data are organized in different tables and comprise the ASAP input report; the allocated categories as recommended by the safety officer from a set of predefined Voluntary Aviation Safety Information (VASI) categories; and the associated feedback to pilots.

The ASAP reports comprise 64 fields (e.g., Pilot Age Group, Flight Hours, Duty Period, Weather Conditions, Visibility Conditions, Navigation System, Location, Aircraft Type, Mission, Flight Plan, and Pilot Narrative). The narrative part of the ASAP reports has a 4000-word limit.

[2] Please refer to FAA Advisory Circular No 120-66B for more details on ASAP

The ASAP reports are manually keyed into the system using a custom developed software system. Each report can have multiple categories e.g., altitude deviation, fatigue, equipment issues, navigational issues, etc). These categories and subcategories are then analyzed at a later time to highlight specific issues and trends. For example, an analyst may group all incidents categorized as "altitude deviations" and "crossing restrictions," and then determine what keywords are associated with these categories. The current system also provides basic analysis features such as histograms and pie charts that could be generated from predefined VASI categories.

Figure 2 shows the current analytical process that mainly consists of search and querying for known problems and issues from the ASAP database. It involves reading and manually categorizing the report. The accuracy of such analysis depends on the strength of the original categorization and whether it incorporated all issues and problems mentioned in the narrative part of the report. Over time, this manual coding and categorization may have to be redone to maintain relevancy.

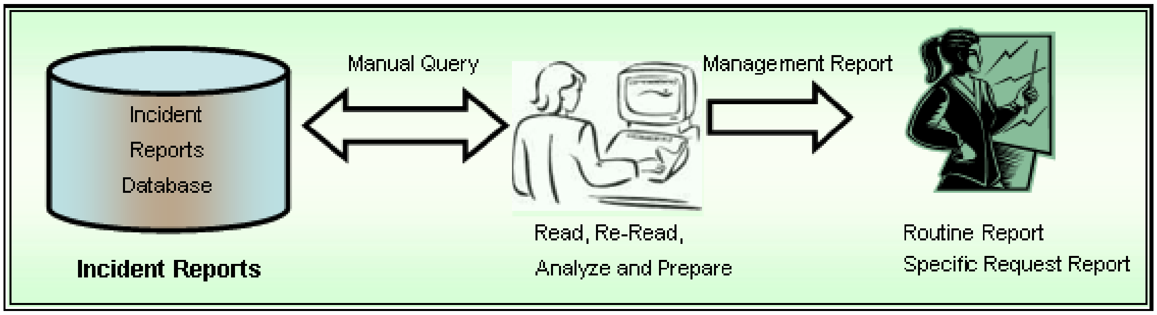


Figure 2: Traditional Analytical Processes

This is not only a laborious process, but it relies heavily on the analyst's expertise and memory to identify new anomalies and trends. Additionally, when a manual process is utilized for narrative analysis, the following problems arise:

- Potential for human error and bias.
- A major portion of the analyst's time is spent on the routine processing of raw data and not on carrying out root cause analysis for various problems.
- The user has to anticipate the right answer before he can formulate an intelligent question.
- Existing systems cannot produce aggregate visual views of results across all data.
- Analysis is only available to the user, restricting access to other decision makers.

These problems limit the timely and efficient analysis of safety event data.

1.6 New Process Using PolyAnalyst Text and Data Mining System

The project objective for Southwest Airlines was to gain insight into pilot's concerns and develop new, more efficient methodologies for the analysis of safety event data with the ultimate goal of improving long-term flight safety performance.

Thus, the project aims at demonstrating another viable methodology for users to search for unknown issues and potential problems. The emphasis was on using both the structured and the narrative parts of the pilot reports and applying advanced text and data mining technologies to identify potential problem areas. The new methodology is a more efficient process for analysts to obtain results.

PolyAnalyst also aims at freeing up valuable analyst time to allow a user to focus on identifying new problems, trends, and issues occurring in the enterprise rather than repeating manual routine analysis. Time is of paramount importance when the analyst's manual interpretation of a particular situation is the key factor in identifying safety concerns.

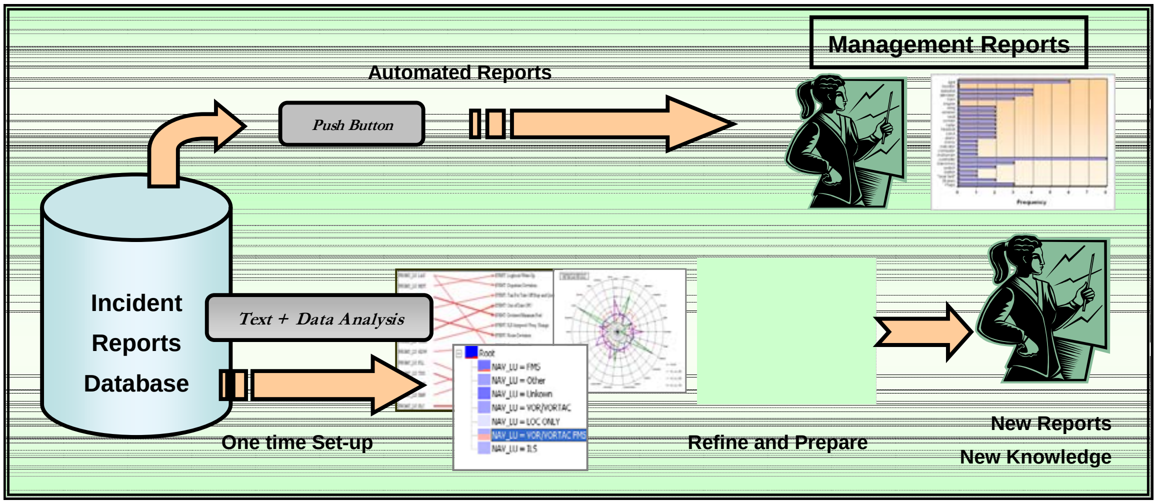


Figure 3: New Analytical Processes

Figure 3 shows how the new range of analytical systems can simplify an analyst's pursuit for new knowledge. The top portion shows that the existing manual process described in Figure 2 can be automated. This would free up valuable analyst time to focus on searching for inherent patterns. Note, that the analyst's time and effort is not on the repetitive task of creating patterns, but on refining, fine tuning, and ascertaining the accuracy of new automatically generated reports, trends, and patterns.

The key advantage to the 'new knowledge discovery' process is realized when tools have the ability to automatically 'tickle' the interest of the user by showing them patterns and charts first, and then allowing the user to refine the analysis based on his expertise.

1.7 Project Methodology

The proof-of-concept project was carried out during six weeks and comprised the following broad phases:

a) Understanding Southwest Airline's ASAP Data and Current Analysis Methodology (described in sections 1.4 and 1.5),
b) Data Cleansing and Transformation, and
c) Data Analysis.

Pilot safety event reports are collated into the ASAP database that resides in the Southwest Airline's Flight Operations Department in Dallas, Texas. The confidential nature of these reports necessitated that the entire data analysis be performed on-site.

<u>Data Cleansing and Transformation</u>

The de-identified ASAP data was directly imported from the Oracle database into PolyAnalyst using a built-in Data Import Wizard and an appropriate SQL statement. The corresponding Look-Up tables facilitating mappings of values of categorical attributes were inserted using documentation provided by Southwest Airlines. Data consisted of 4600 ASAP reports. However, the analysis was performed on only 2000 reports due to time constraints.

In the process of text analysis, PolyAnalyst is recognizing and counting encountered concepts: individual words and stable word combinations (collocations). PolyAnalyst recognizes individual terms encountered in the text based on a comprehensive system of universal semantic dictionaries of English built into the system. To further enhance the quality of text analysis, PolyAnalyst provides means for the user to define and incorporate in the project three additional special user dictionaries for data cleansing: dictionary of abbreviations and synonyms, dictionary of phrases, and dictionary of ignored terms.

Once created, these user dictionaries become an integral part of the overall decision support system, holding valuable domain knowledge. PolyAnalyst facilitates developing, saving and reutilizing such user dictionaries in future projects involving the analysis of text from the same domain. The utilization of user dictionaries in the Southwest Airlines project is discussed below.

1) Dictionary of abbreviations and synonyms

To enhance the quality of the analysis of text containing non-standard terms, PolyAnalyst offers a capability of importing user-created dictionaries reflecting domain-specific jargon, which are then used by the system in concert with the underlying comprehensive dictionary of English.

To ensure correct interpretation of non-standard terms encountered in natural language narratives of ASAP reports, specific dictionaries containing domain-specific synonyms and abbreviation expansions were developed and imported in the system. This was necessary because the narrative text was interspersed with acronyms and industry jargon, which could be a source of errors made by both the end user and the text analysis engine.

To accomplish the development of domain-specific dictionaries, one first had to generate a list of words not recognized by the system, which were either misspelled words or acronyms, and then develop a collection of rules for correcting misspelled words and expanding acronyms. An initial list of unrecognized non-standard terms was obtained by running the PolyAnalyst text analysis engine on ASAP narratives in an undirected mode. The resulting list of unrecognized terms was turned into a dictionary containing domain-specific abbreviations based on a manually created collection of rules mapping encountered misspelled words and acronyms into their correct full form. The process of creating a collection of mapping rules for the aerospace domain was a collaborative effort of analysts of Southwest Airlines, FAA and Megaputer Intelligence. A few examples of the developed mapping rules are provided below:

-	w/o	→	without
-	t/o	→	take-off
-	a/c	→	aircraft
-	alt	→	altitude
-	freq	→	frequency
-	rwy	→	runway
-	nm	→	nautical miles
-	VFR	→	Visual Flight Rules
-	MSL	→	Mean Sea Level

It is important to note that in addition to misspelled words and acronyms, one had to sift out abbreviations for airport codes and navigation fixes frequently encountered in pilot narratives.

2) Dictionary of phrases

Based on their background knowledge of the field, Flight Safety Officers of Southwest Airlines had a collection of phrases that they wanted to be recognized by the automated system as inseparable word combinations. Correspondingly, a list of such phrases (word sequences) that were of particular interest to analysts was created and imported in PolyAnalyst as a dictionary of phrases in the beginning of the project. A few examples of phrases included in this list are provided below:

- Altitude Conflict
- Departure Deviation
- ATC Clearance
- Logbook Error
- Aircraft Separation
- Bird Strike

Now whenever PolyAnalyst encounters one of these word combinations in the text during the analysis, it identifies it as a single concept and does not attempt to recognize and count individual words comprising this concept.

3) Dictionary of ignored terms

Finally, one more piece of useful information summarizing either the background knowledge of the investigated domain or the results of iterative analysis of data, is a list of ignored terms – terms that are unconditionally excluded from further analysis. The content of this list depends on the project being undertaken. For example, when investigating all safety events that happened at take-off, one might want to add the term ‘*runway*’ to the dictionary of ignored terms because this term occurs in nearly every description of take-off events and does not help in further categorization of these events. Yet, the term ‘*runway*’ by itself might be very informative when studying other sets of safety reports.

For the Southwest Airlines project, the list of ignored terms was created through manual evaluation of the results of iterative analysis. The resulting terms that were found to provide little information were excluded from further analysis with the help of the PolyAnalyst “ignored terms” dictionary.

Data Analysis

The overall analysis methodology for the proof-of-concept was to conduct focused text analysis to identify specific sets of concepts of interest to the user. This was to be followed by other analysis and visualization engines to help reveal patterns and trends in safety events. The obtained results would then enable the analyst to efficiently drill down to the potential issues.

The following analytical tasks were to be addressed in the project:

1) Revealing all mechanisms and devices mentioned in pilot narratives and using the results of the analysis for tagging the corresponding records.
2) Comparing the distributions of various device-related problems across different types of aircraft.
3) Checking for possible correlations between device-related problems and specific phases of the flight.
4) Discovering stable patterns and clusters of terms occurring more frequently than expected in pilot descriptions of safety events and drilling down to the corresponding records.
5) Developing classification rules for processing text records based on the analysis of pre-categorized data.

PolyAnalyst provides sixteen different analytical engines and several visualization techniques that can be used either independently or sequentially to derive new knowledge from data. This broad range of analytical engines allows the user to conduct the analysis irrespective of the type of data (Numeric, Boolean, Date, Categorical or Textual).

The following seven of the sixteen analysis and visualization techniques available within PolyAnalyst were used in the proof-of-concept project, as described in more detail in section 2.0:

a) Text Analysis: Analyze and extract important concepts from pilot narratives. The technique was used in the supervised and unsupervised mode.
b) Summary Statistics: Gain an understanding of initial data and intermediate text analysis results. The tabulated results and pie chart visualizations are also used to formulate the 'next steps' in analysis process
c) Snake/ Radar Chart: Provide a comparative overview of chosen concepts across different attributes.
d) Link Charts: Carry out one-dimensional correlation analysis between different attributes.
e) Link Analysis: Conduct an *n*-dimensional correlation analysis to find correlation between multiple attributes.
f) Link Terms: Quickly identify clusters of words and phrases that were predominant in the narrative. The system visually displays the clusters and nodes.
g) Decision Tree: Perform 'root cause' analysis based on the concepts extracted from narratives.

The following sections illustrate these techniques as they were applied to the analysis of Southwest's ASAP reports. The analysis methodology generated insightful correlation and radar graphs that tickled the Flight Safety Officer's curiosity. The results were investigated and refined for accuracy and relevance.

2.0 Application of PolyAnalyst Techniques

Southwest Airlines FSOs were primarily interested in utilizing PolyAnalyst capabilities to automate the manual process described in Section 1.5 and find trends/relationships throughout both structured and textual fields. Narrative analysis would enable investigators to further direct their attention to potential safety issues and pilot concerns. In addition, there was an interest in using the software to find trends/relationships in the data that may not be related to preset categories, letting the intelligence of the software find the relationships. The FSO could then continue to refine and drill down into the data for further analysis.

Textual data often contains over 80% of useful information, but analysts often lack efficient tools for textual analysis. A broad methodology to carry out a typical text analysis project is shown in Figure 4. Once data is imported into the system, different analysis techniques can be applied to the database depending on the type of data and desired end results. Text analysis is an iterative process, therefore, an important feature for a user is the capability to drill down and refine output reports.

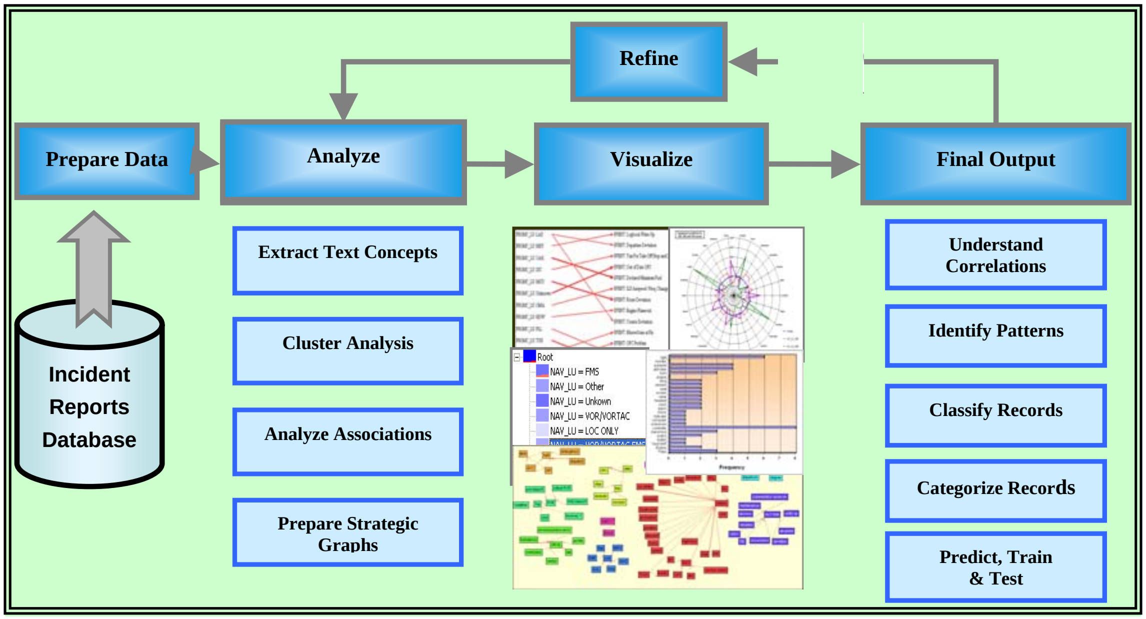


Figure 4: Text Mining Methodology

2.1 Text Analysis (TA) Engine

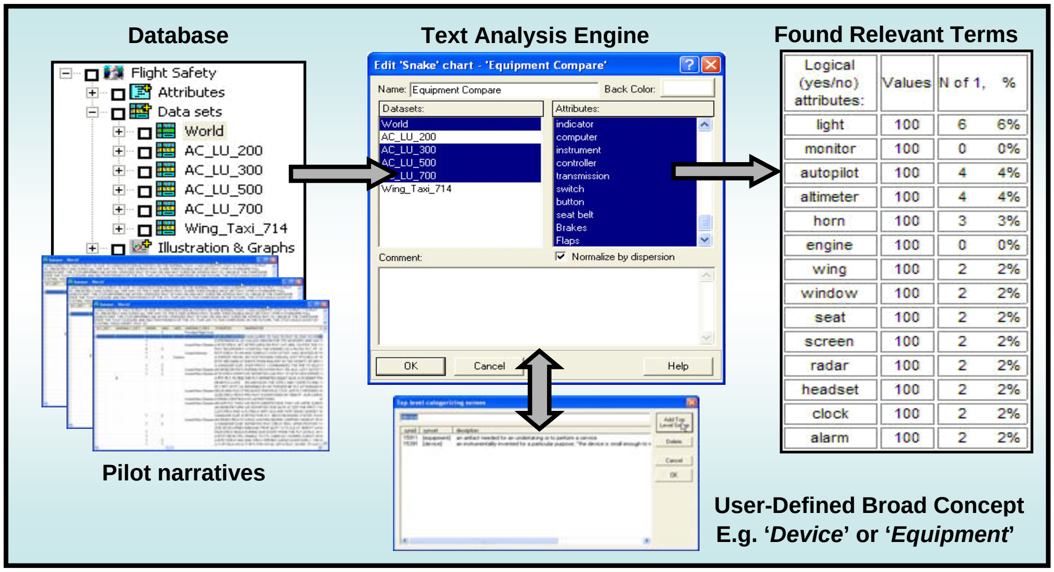
The process of gaining knowledge from narratives involves two main steps: discovery and interpretation of results. PolyAnalyst Text Analysis (TA) engine aids in this process by identifying important concepts discussed in the narratives. The analysis can be carried out in two modes:

i) Unsupervised TA Mode: In this mode, the TA engine extracts all important concepts occurring in the text based on the analysis of their relative frequencies, without any additional input from the user. These concepts may be either:
  - single words like *'light'*, *'monitor'* and *'autopilot'* (as shown in Figure 5) or
  - word sequences like *'course deviation'*, *'missed altitude restriction'* and *'RNAV departure'* (as shown in Figure 12).

ii) Supervised TA Mode: The user can guide the TA engine to only search and extract all particular instances of selected concepts of interest. Let's consider this technique in more detail using the Southwest example below.

The Southwest Airlines FSO wanted to gain insight into concerns discussed in the ASAP reports. An intuitive way of doing this would be to extract all text concepts occurring in the narratives and then manually glean through them. However, a far more powerful mechanism would be to search for specific concepts and let the text analysis engine do the work of identifying and reporting only the relevant chosen concepts.

An example of this type of supervised analysis is to identify only 'equipment'-related issues mentioned in ASAP report narratives.



Edit 'Snake' chart - 'Equipment Compare'

Name: Equipment Compare

Datasets: World, AC_LU_200, AC_LU_300, AC_LU_500, AC_LU_700, Wing_Taxi_714

Attributes: indicator, computer, instrument, controller, transmission, switch, button, seat belt, Brakes, Flaps

Comment:

☑ Normalize by dispersion

Logical (yes/no) attributes:	Values	N of 1,	%
light	100	6	6%
monitor	100	0	0%
autopilot	100	4	4%
altimeter	100	4	4%
horn	100	3	3%
engine	100	0	0%
wing	100	2	2%
window	100	2	2%
seat	100	2	2%
screen	100	2	2%
radar	100	2	2%
headset	100	2	2%
clock	100	2	2%
alarm	100	2	2%

Figure 5: Identify Relevant Concepts of Interest

For this example, PolyAnalyst was instructed to focus on any specific instances of '*equipment*' and '*device*'. The Text Analysis (TA) engine then sifted through the narratives of all 2000 ASAP reports and automatically returned a list of concepts like *'autopilot'*, *'altimeter'*, *'horn'*, and *'engine'*.

The user could now create a tabular report with the help of the Summary Statistics engine of PolyAnalyst and view simple pie charts, or employ other graphical visualization techniques like Snake Charts (section 2.2) or Correlation Charts (section 2.3). Alternatively, the frequency table of equipment concerns could be exported into Microsoft Excel for further visualization.

2.2 Visualization Techniques – Snake/Radar Charts

PolyAnalyst incorporates different visualization techniques to enable the user to generate explicit and actionable results. The user can employ graphs and charts to better understand patterns of terms and relations between the terms during text analysis.

One analytical objective of the Southwest FSO was to gain insight on how pilot concerns vary across different attributes or categories. Snake/Radar charts can be combined with supervised text analysis processes to produce insightful views of the investigated issues.

For this analysis a Snake/Radar Chart was generated based on equipment concerns identified from ASAP reports against different aircraft types. The methodology used to generate the 'Snake/Radar Chart' is shown in Figure 6. The top part represents the different equipment issues like *'engine'*, *'brake'*, and *'wing'*. The bottom part shows that the dataset easily divided into different aircraft types like *AC300*, *AC500* and *AC700*.

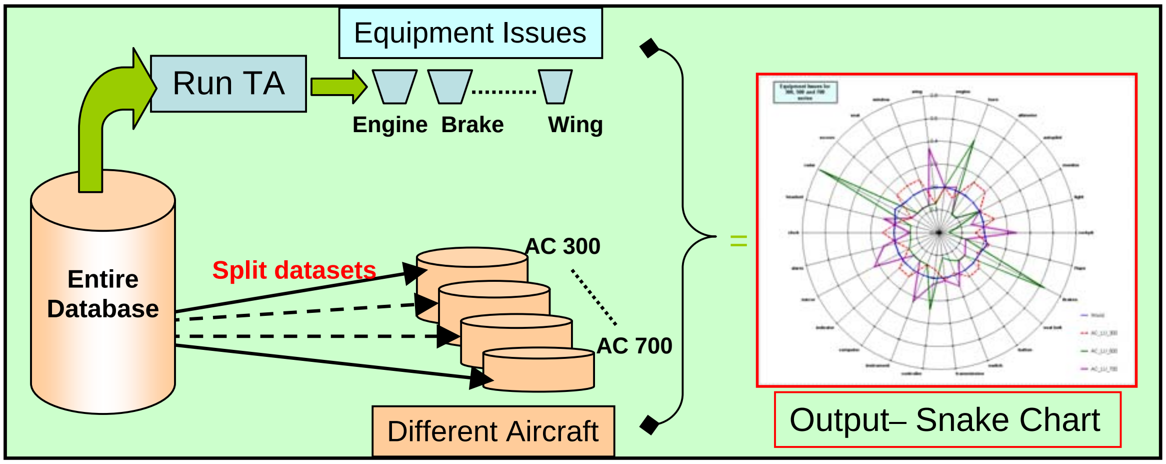


Figure 6: Broad Process for Generating a Snake/Radar Chart

Figure 7 demonstrates different concepts from pilot narratives that could be potential issues pertaining to aircraft types AC 300, AC 500 and AC 700. Pilots flying AC 500s predominately mention *'radar', 'brakes', 'horn', and 'controller'*, whereas pilots flying AC 700s mention *cockpit', 'instrument', 'mirror' and 'wing'* frequently. The World line represents the overall average of the 2000 ASAP reports in the analysis.

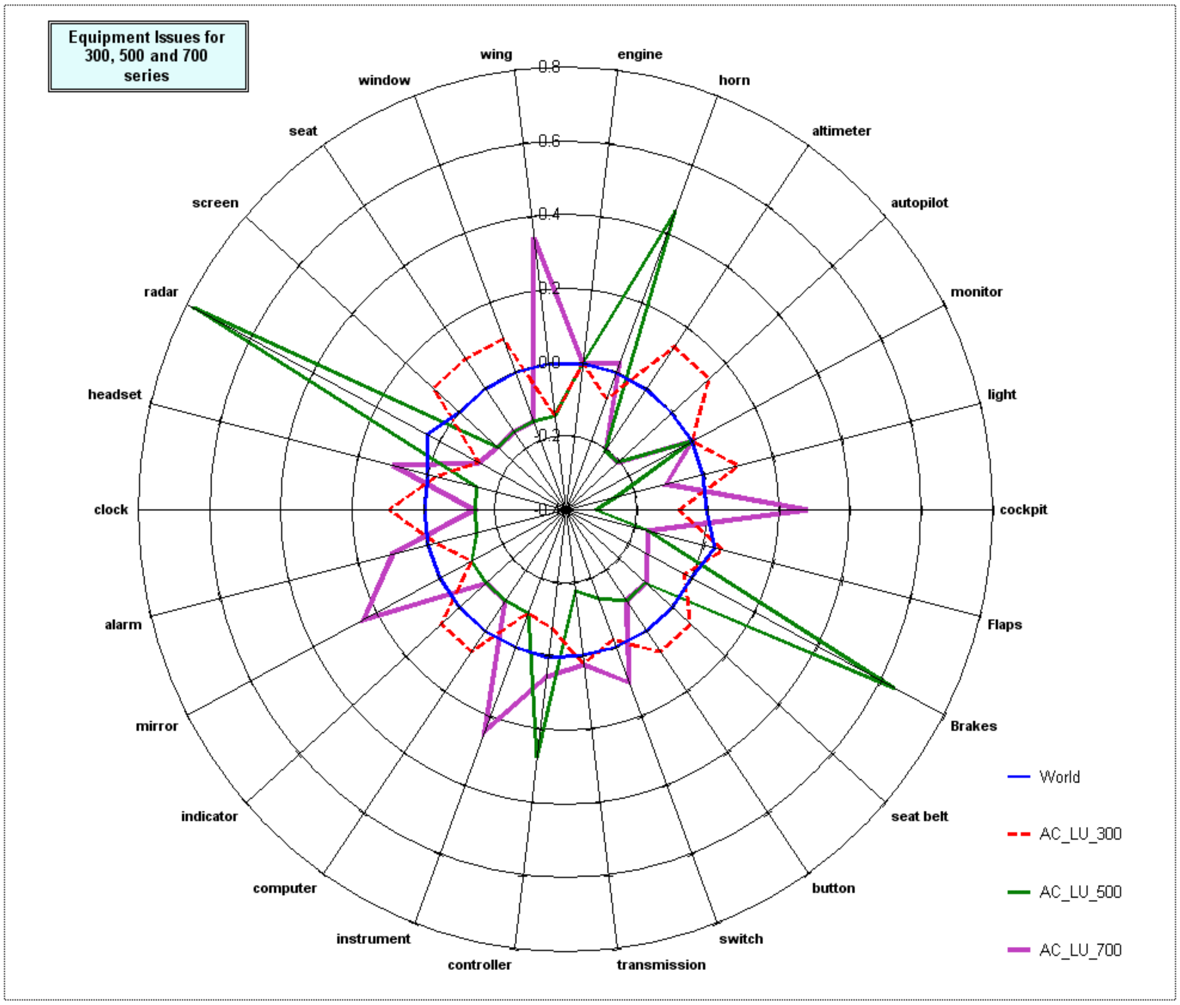
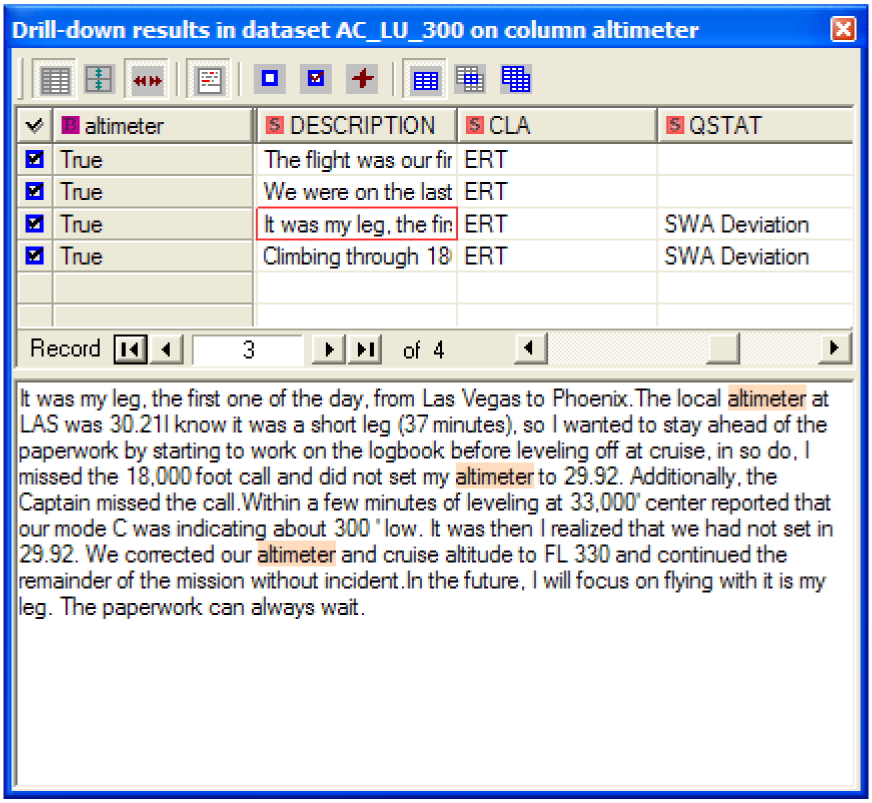


Figure 7: Equipment Concepts by Aircraft Type

The user can drill down on a term of interest for a certain data set to view the contents of the corresponding records with the terms of interest highlighted in them. For example, clicking on the intersection of the red dashed line representing AC 300 aircraft with the radial line corresponding to the term '*altimeter*', one invokes the drill-down window (Figure 8) displaying all records related to AC 300 where pilot narratives contain the term '*altimeter*':

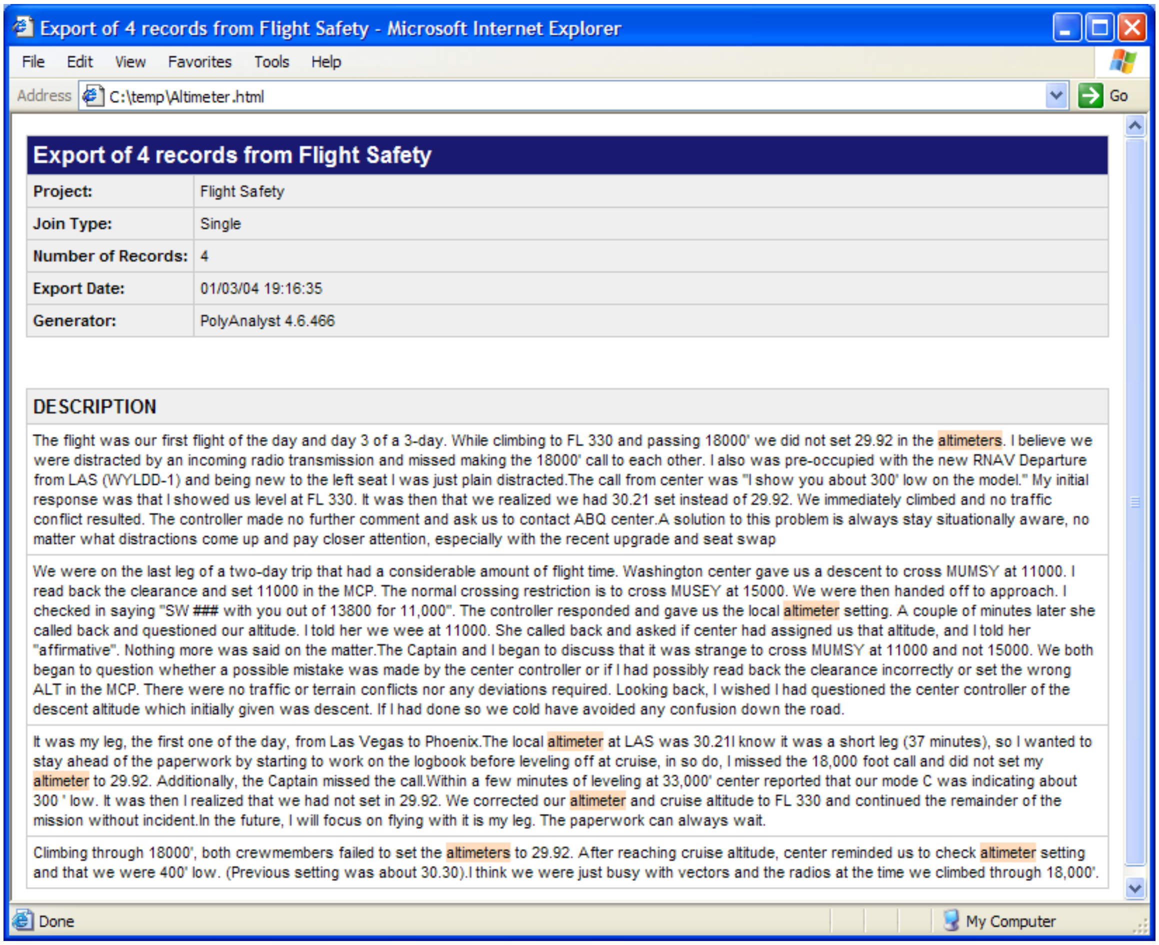


✓	altimeter	DESCRIPTION	CLA	QSTAT
☑	True	The flight was our fir	ERT	
☑	True	We were on the last	ERT	
☑	True	It was my leg, the fir	ERT	SWA Deviation
☑	True	Climbing through 18	ERT	SWA Deviation

It was my leg, the first one of the day, from Las Vegas to Phoenix.The local altimeter at LAS was 30.21I know it was a short leg (37 minutes), so I wanted to stay ahead of the paperwork by starting to work on the logbook before leveling off at cruise, in so do, I missed the 18,000 foot call and did not set my altimeter to 29.92. Additionally, the Captain missed the call.Within a few minutes of leveling at 33,000' center reported that our mode C was indicating about 300 ' low. It was then I realized that we had not set in 29.92. We corrected our altimeter and cruise altitude to FL 330 and continued the remainder of the mission without incident.In the future, I will focus on flying with it is my leg. The paperwork can always wait.

Figure 8: Drill Down Window for 'Altimeter' and AC 300 Aircraft

In this case, there are four records mentioning altimeter for AC 300. The drill-down window allows the user to browse through individual records, select or deselect records of interest and export selected records to separate data sets for further analysis or to an HTML report. Right clicking on records listed in the drill-down window and selecting appropriate export options, one obtains the HTML report as shown in Figure 9.



Export of 4 records from Flight Safety

Project:	Flight Safety
Join Type:	Single
Number of Records:	4
Export Date:	01/03/04 19:16:35
Generator:	PolyAnalyst 4.6.466

DESCRIPTION
The flight was our first flight of the day and day 3 of a 3-day. While climbing to FL 330 and passing 18000' we did not set 29.92 in the altimeters. I believe we were distracted by an incoming radio transmission and missed making the 18000' call to each other. I also was pre-occupied with the new RNAV Departure from LAS (WYLDD-1) and being new to the left seat I was just plain distracted.The call from center was "I show you about 300' low on the model." My initial response was that I showed us level at FL 330. It was then that we realized we had 30.21 set instead of 29.92. We immediately climbed and no traffic conflict resulted. The controller made no further comment and ask us to contact ABQ center.A solution to this problem is always stay situationally aware, no matter what distractions come up and pay closer attention, especially with the recent upgrade and seat swap
We were on the last leg of a two-day trip that had a considerable amount of flight time. Washington center gave us a descent to cross MUMSY at 11000. I read back the clearance and set 11000 in the MCP. The normal crossing restriction is to cross MUSEY at 15000. We were then handed off to approach. I checked in saying "SW ### with you out of 13800 for 11,000". The controller responded and gave us the local altimeter setting. A couple of minutes later she called back and questioned our altitude. I told her we wee at 11000. She called back and asked if center had assigned us that altitude, and I told her "affirmative". Nothing more was said on the matter.The Captain and I began to discuss that it was strange to cross MUMSY at 11000 and not 15000. We both began to question whether a possible mistake was made by the center controller or if I had possibly read back the clearance incorrectly or set the wrong ALT in the MCP. There were no traffic or terrain conflicts nor any deviations required. Looking back, I wished I had questioned the center controller of the descent altitude which initially given was descent. If I had done so we cold have avoided any confusion down the road.
It was my leg, the first one of the day, from Las Vegas to Phoenix.The local altimeter at LAS was 30.21I know it was a short leg (37 minutes), so I wanted to stay ahead of the paperwork by starting to work on the logbook before leveling off at cruise, in so do, I missed the 18,000 foot call and did not set my altimeter to 29.92. Additionally, the Captain missed the call.Within a few minutes of leveling at 33,000' center reported that our mode C was indicating about 300 ' low. It was then I realized that we had not set in 29.92. We corrected our altimeter and cruise altitude to FL 330 and continued the remainder of the mission without incident.In the future, I will focus on flying with it is my leg. The paperwork can always wait.
Climbing through 18000', both crewmembers failed to set the altimeters to 29.92. After reaching cruise altitude, center reminded us to check altimeter setting and that we were 400' low. (Previous setting was about 30.30).I think we were just busy with vectors and the radios at the time we climbed through 18,000'.

Figure 9: Output Report in HTML Format

Such exported HTML reports summarizing the results obtained by different PolyAnalyst exploration engines help FSOs substantiate their conclusions with concrete examples, as well as facilitate sharing the results of the analysis with other decision makers across the organization.

A similar analysis was performed on 'Crossing restriction' cases, a type of "Altitude Deviation" event that occurs when the aircraft did not cross a specific point in space at a predetermined altitude. Figure 10 shows potential equipment concerns only for records classified as 'crossing restriction". Note that AC 200 type aircraft may have potential concerns related to *‘approach control’*, *‘center controller’ ‘plate’*, *‘press’*, *‘selector’*, *‘wheel’*, *‘previous controller’*, *’speed brake’ and ‘radial’* compared to other issues. The spikes indicate that these terms are identified more frequently. Also note that the terms *‘cockpit’, ‘needle’ and ‘safety’* are cited more frequently in AC 700 than AC 200 aircraft.

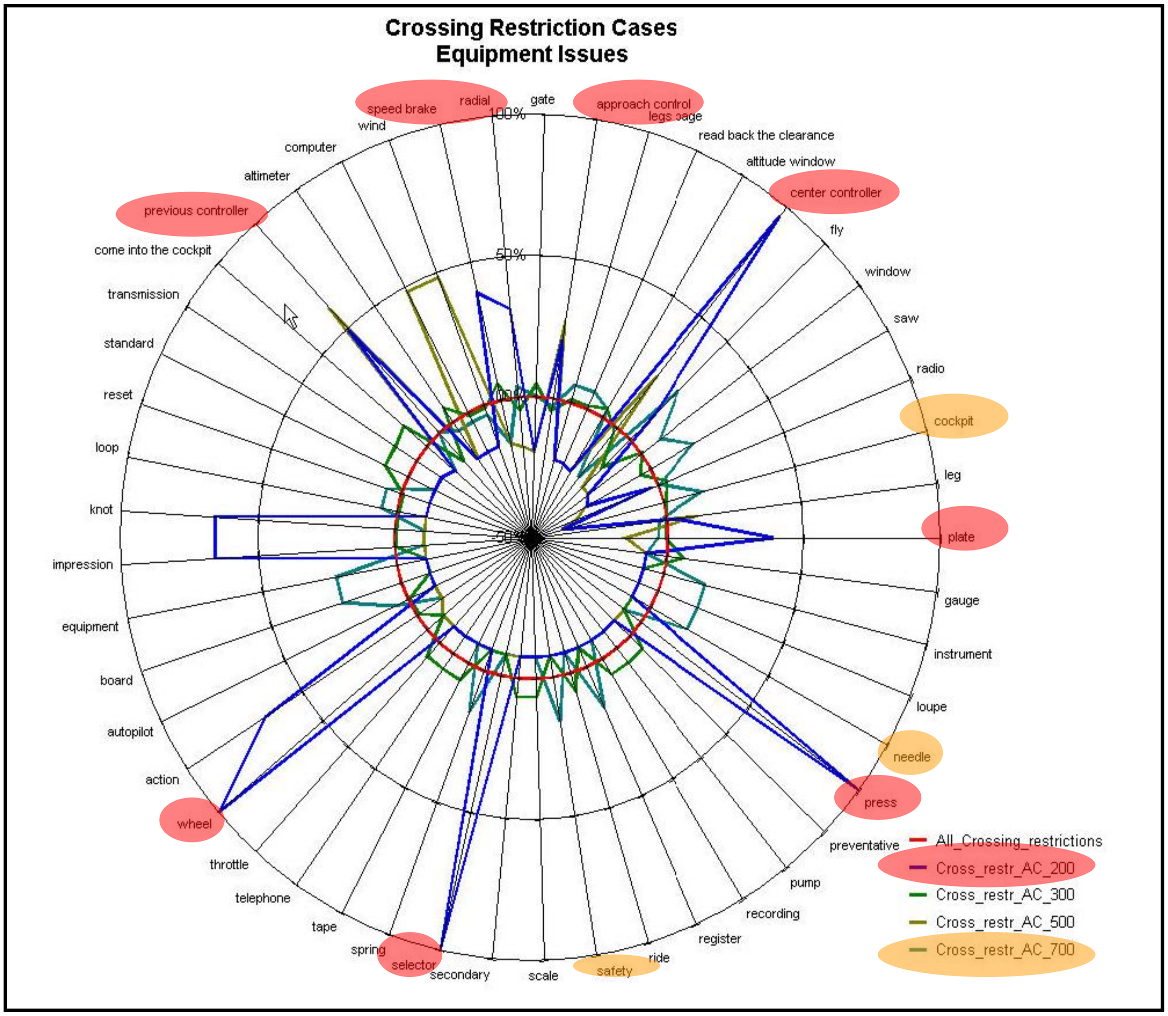


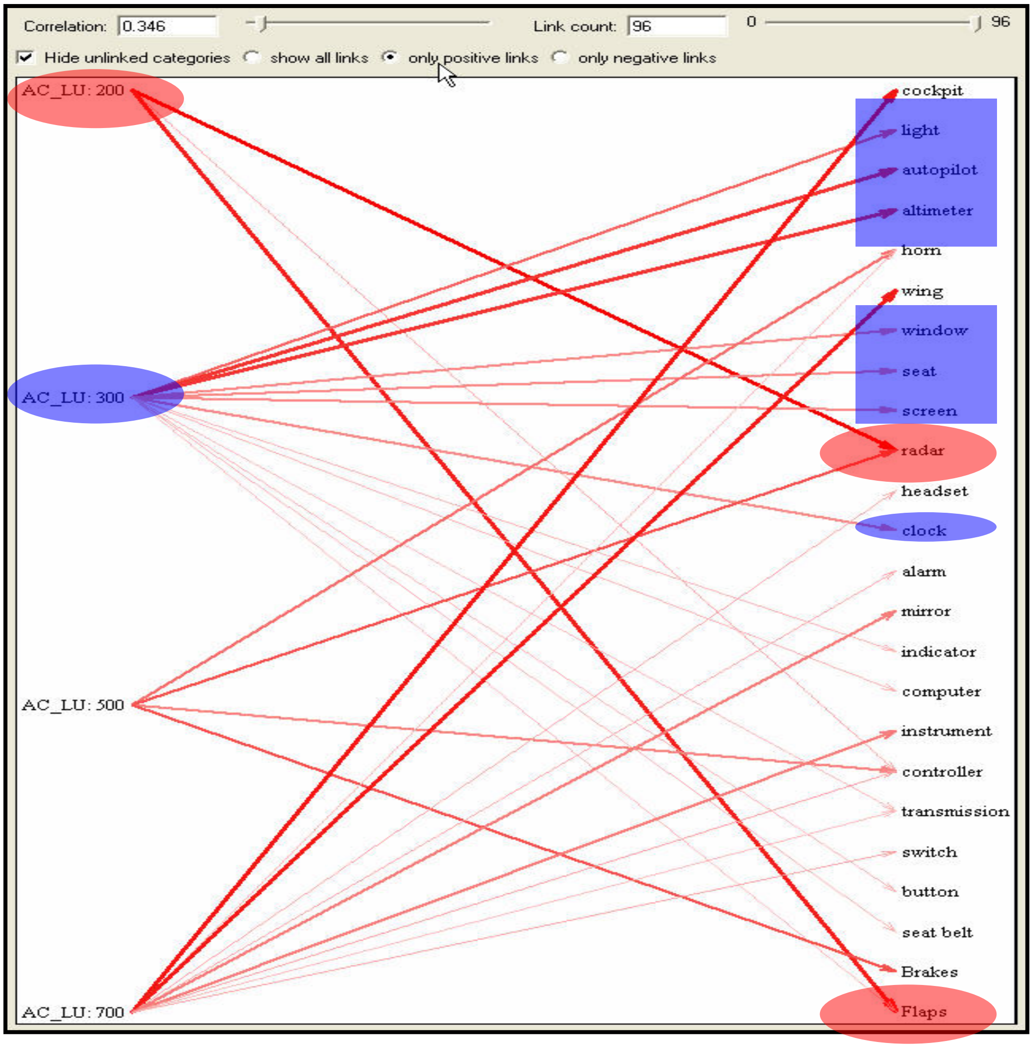
Figure 10: Chart for Crossing Restriction Issues Only

2.3 Correlation Analysis (Link Charts, Link Analysis, Link Terms)

Calculating and visualizing correlations of attribute values gives the user knowledge of stable patterns of co-occurrences of individual attributes.

2.3.1 Link Charts

Link Charts display both positive and negative correlations. In PolyAnalyst, a red line indicates positive correlation and a blue line indicates negative correlation of the link. The intensity of the line is indicative of the strength of the correlation. The darker and thicker the line, the higher is the strength of the correlation. Link charts in Figures 11 and 12 show a quick way to view the most important correlations between items of interest.



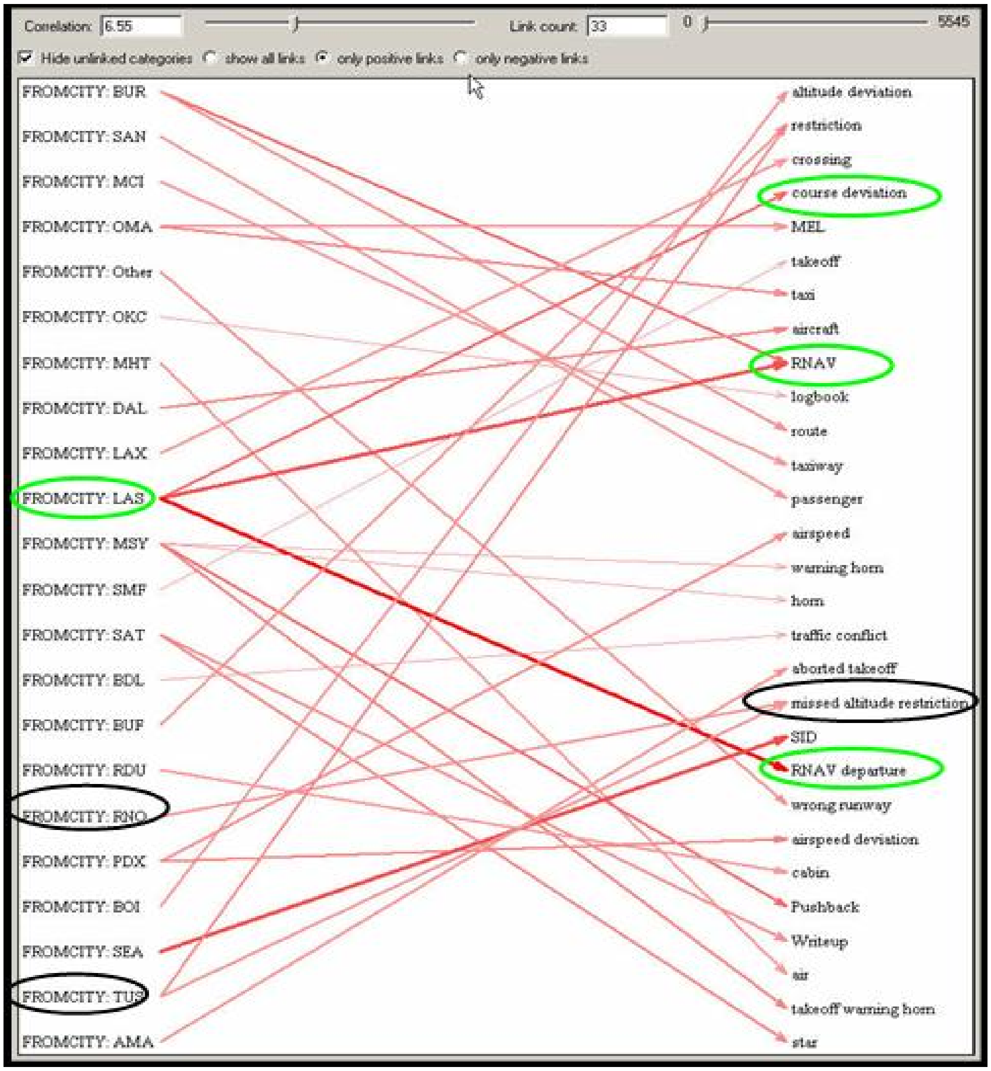
Correlation: 0.346 Link count: 96 0 96

☑ Hide unlinked categories ○ show all links ● only positive links ○ only negative links

Figure 11: Correlation – Aircraft Type and Terms in Pilot Narratives

Figure 11 shows the discovered correlations between the aircraft type and the terms extracted from ASAP narratives.

Strong correlations occurred, for example, between aircraft type *AC 200* and the terms *'radar'* and *'flaps'*. Similarly, pilots flying Aircraft type *AC 300* mention the words '*autopilot' and 'altimeter'* more often than other terms. Additionally, these pilots are mentioning the words *'window'*, *'seat'*, *'screen'*, *'clock'* and *'light'*. These terms are represented by dim lines, signifying a weaker correlation.



Correlation: 6.55 Link count: 33 0 5545

☑ Hide unlinked categories ○ show all links ● only positive links ○ only negative links

Figure 12: Correlation Between Airport and Event Type

Figure 12 shows how different airports and event types are related. To illustrate the interpretation of this graph, let us consider all links connected to a single object on the left hand side, and then do the same for an object on the right hand side.

- LAS (Las Vegas, Nevada) Airport is strongly correlated with *'RNAV','RNAV departure'* and *'course deviation'* events.

- Airports RNO (Reno, Nevada) and TUS (Tucson, Arizona) are correlated with *‘missed altitude restriction’* events.

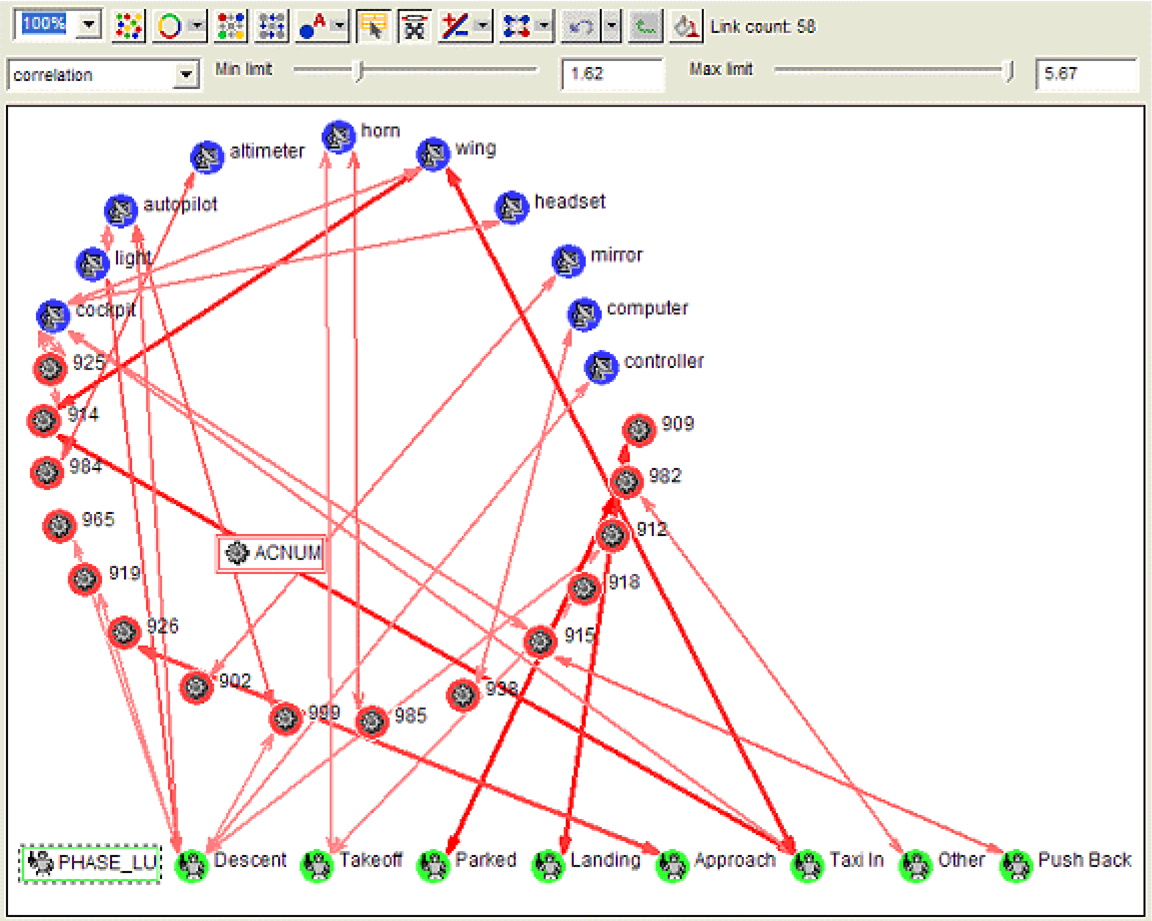
The user can easily visualize correlations between important items from both structured and unstructured parts of the database.

2.3.2 Link Analysis

Link Analysis was used to find correlations across multiple attributes. This helps an analyst check for existence of any inherent patterns of co-occurrences.

Figure 13 shows a Link Analysis chart with correlation between the following three attributes:

~ *Specific aircraft* (AC NUM)
~ *Particular instances of the concept ‘device’, extracted from pilot narratives*
~ *Different flight phases* (PHASE_LU)



100%

Link count: 58

correlation

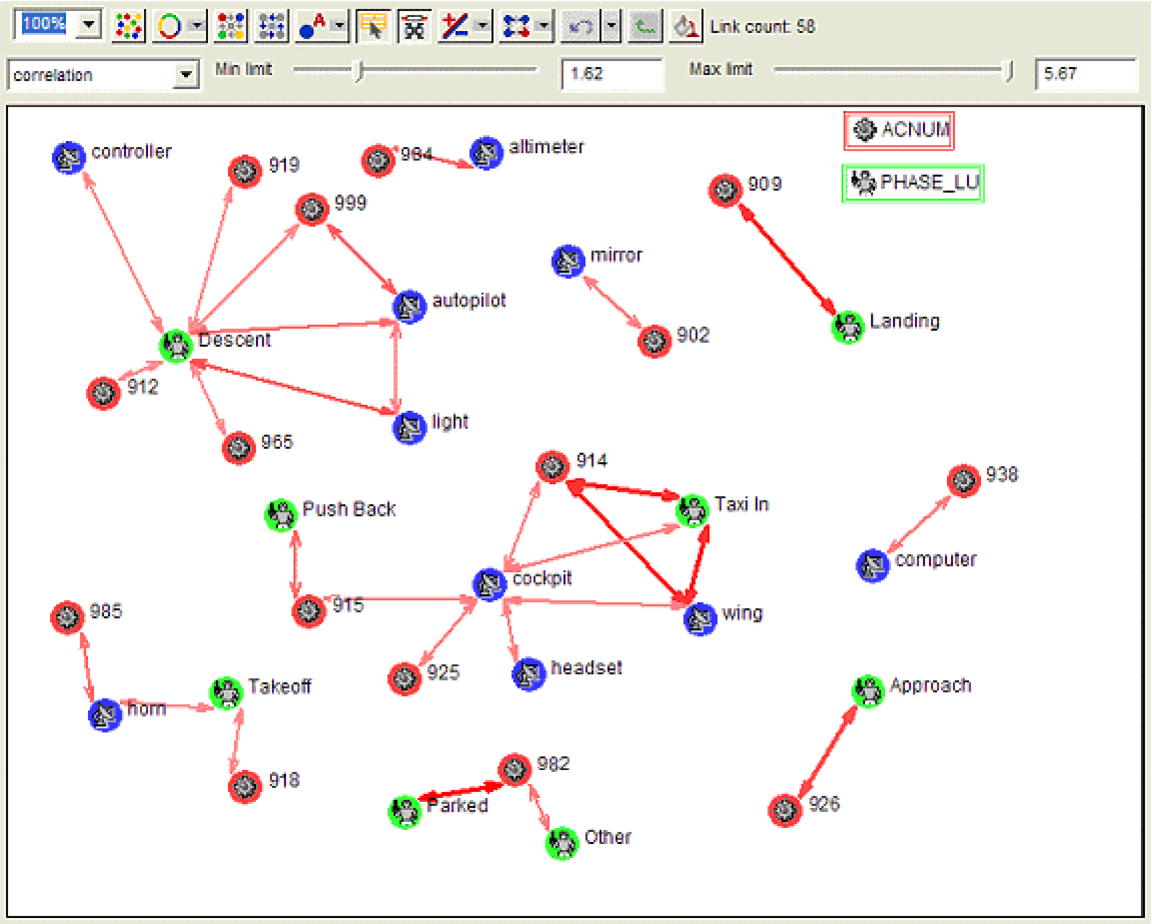
Min limit 1.62

Max limit 5.67

Figure 13: Multi-Dimensional Link Analysis Chart

From this chart, one can identify strong correlation patterns. The marked pattern shows that Aircraft number ‘*914*’ was in the ‘*Taxi-in*’ flight phase and had ’*wing*’ identified in the narrative. This pattern has a stronger correlation than other correlations and hence is highlighted. One can also isolate all instances of this pattern for further drill down and analysis.

PolyAnalyst can automatically visually isolate clusters representing different stable patterns discovered by the system. Figure 14 below presents the same information as Figure 13 – but with the automated layout performed by the system. Automated layout helps the user conveniently visualize and interpret clusters of related objects.



100%

Link count: 58

correlation

Min limit 1.62

Max limit 5.67

Figure 14: Multi-Dimensional Link Analysis Chart to view Clusters of Terms

2.3.3 Link Terms

PolyAnalyst Link Terms engine can be used to reveal clusters of important terms from the narrative portion of the ASAP reports. Link Terms engine conducts ‘n-dimensional’ correlation analysis and provides a visual layout of the results to help reveal close associations and patterns of terms in the data. An ability to capture patterns of pilot concerns without reading through all of the records can provide valuable insights into past occurrences, therefore, saving time for more advanced analysis. Figure 15 illustrates the clusters generated by Link Terms.

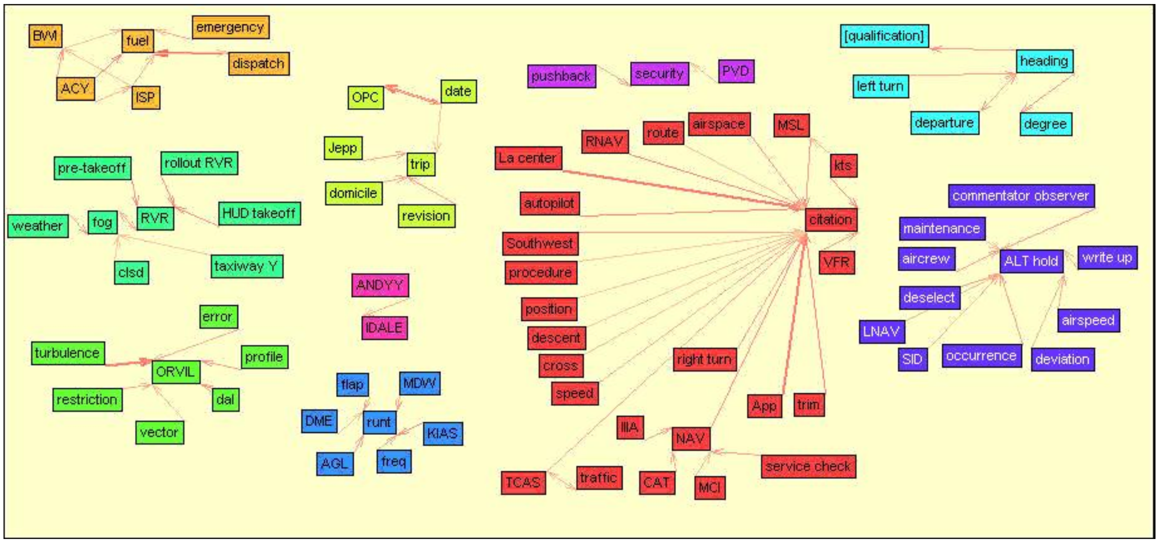


Figure 15: Identifying Clusters Using Link Terms

Link Terms produced ten clusters, each denoted by different colors. These clusters now prompt the user to further investigate the relationships and ask questions such as why is:

- *‘Pushback’* mentioned together with *‘security’* and *‘PVD (Providence, Rhode Island)’*
- *‘ORVIL’*[3] (an altitude restriction point) is associated with ‘DAL’,*‘error’, ‘turbulence’, restriction’, ‘profile’ and ‘vector’*

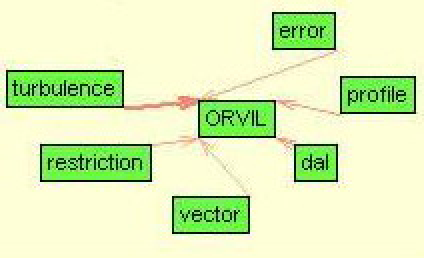
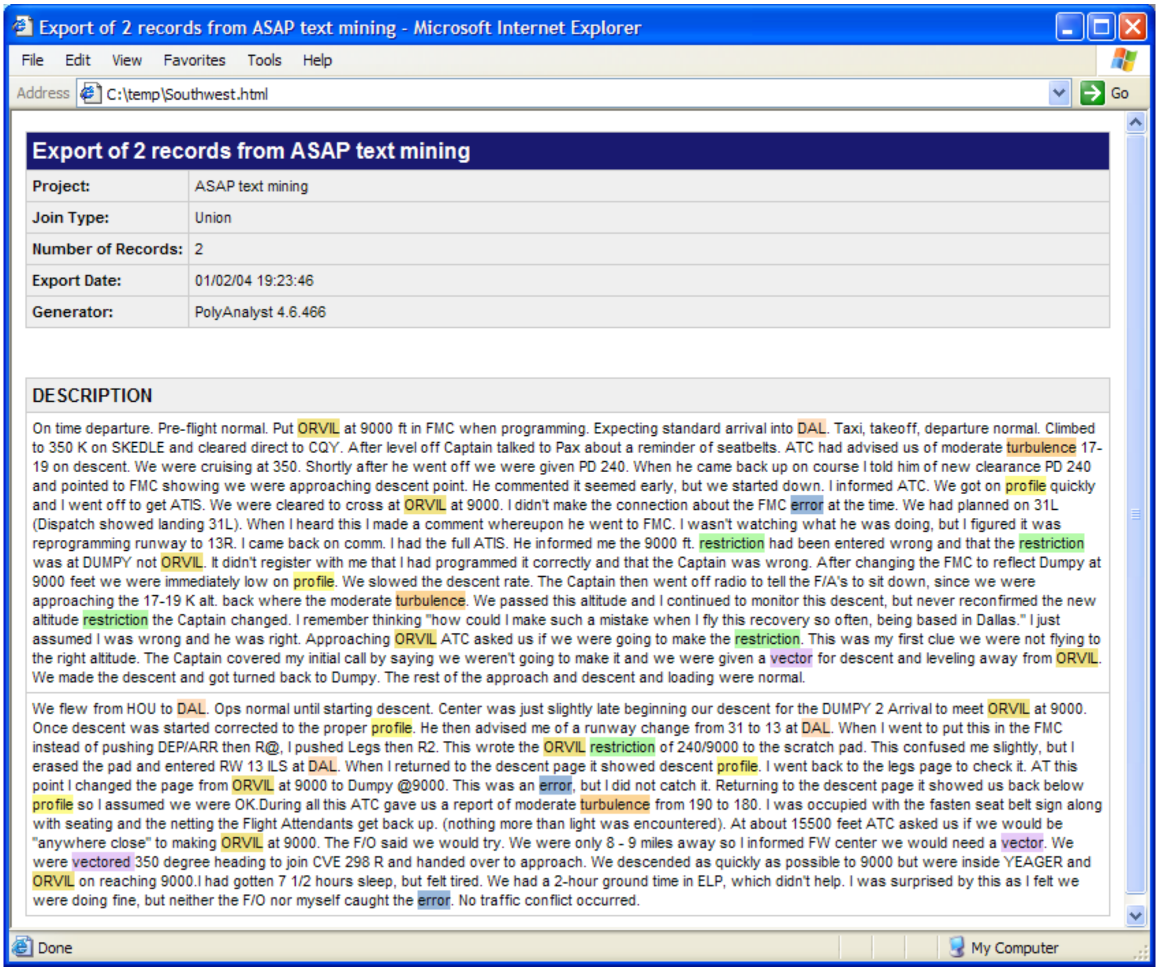


Figure 16: Link Terms Cluster of ‘ORVIL’

This can be accomplished by clicking on the ‘ORVIL’ cluster and drilling down into the corresponding ASAP records. Figure 17 presents two records corresponding to the cluster

[3] ORVIL is an altitude restriction point and must be dialed in by the pilot. It is a point in space that requires a particular profile to be flown.

‘ORVIL’ with the clustered terms highlighted (the colors indicate the different terms of the cluster).


Export of 2 records from ASAP text mining - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address C:\temp\Southwest.html
Go
Done
My Computer

Export of 2 records from ASAP text mining

Project:	ASAP text mining
Join Type:	Union
Number of Records:	2
Export Date:	01/02/04 19:23:46
Generator:	PolyAnalyst 4.6.466

DESCRIPTION
On time departure. Pre-flight normal. Put ORVIL at 9000 ft in FMC when programming. Expecting standard arrival into DAL. Taxi, takeoff, departure normal. Climbed to 350 K on SKEDLE and cleared direct to CQY. After level off Captain talked to Pax about a reminder of seatbelts. ATC had advised us of moderate turbulence 17-19 on descent. We were cruising at 350. Shortly after he went off we were given PD 240. When he came back up on course I told him of new clearance PD 240 and pointed to FMC showing we were approaching descent point. He commented it seemed early, but we started down. I informed ATC. We got on profile quickly and I went off to get ATIS. We were cleared to cross at ORVIL at 9000. I didn't make the connection about the FMC error at the time. We had planned on 31L (Dispatch showed landing 31L). When I heard this I made a comment whereupon he went to FMC. I wasn't watching what he was doing, but I figured it was reprogramming runway to 13R. I came back on comm. I had the full ATIS. He informed me the 9000 ft. restriction had been entered wrong and that the restriction was at DUMPY not ORVIL. It didn't register with me that I had programmed it correctly and that the Captain was wrong. After changing the FMC to reflect Dumpy at 9000 feet we were immediately low on profile. We slowed the descent rate. The Captain then went off radio to tell the F/A's to sit down, since we were approaching the 17-19 K alt. back where the moderate turbulence. We passed this altitude and I continued to monitor this descent, but never reconfirmed the new altitude restriction the Captain changed. I remember thinking "how could I make such a mistake when I fly this recovery so often, being based in Dallas." I just assumed I was wrong and he was right. Approaching ORVIL ATC asked us if we were going to make the restriction. This was my first clue we were not flying to the right altitude. The Captain covered my initial call by saying we weren't going to make it and we were given a vector for descent and leveling away from ORVIL. We made the descent and got turned back to Dumpy. The rest of the approach and descent and loading were normal.
We flew from HOU to DAL. Ops normal until starting descent. Center was just slightly late beginning our descent for the DUMPY 2 Arrival to meet ORVIL at 9000. Once descent was started corrected to the proper profile. He then advised me of a runway change from 31 to 13 at DAL. When I went to put this in the FMC instead of pushing DEP/ARR then R@, I pushed Legs then R2. This wrote the ORVIL restriction of 240/9000 to the scratch pad. This confused me slightly, but I erased the pad and entered RW 13 ILS at DAL. When I returned to the descent page it showed descent profile. I went back to the legs page to check it. AT this point I changed the page from ORVIL at 9000 to Dumpy @9000. This was an error, but I did not catch it. Returning to the descent page it showed us back below profile so I assumed we were OK.During all this ATC gave us a report of moderate turbulence from 190 to 180. I was occupied with the fasten seat belt sign along with seating and the netting the Flight Attendants get back up. (nothing more than light was encountered). At about 15500 feet ATC asked us if we would be "anywhere close" to making ORVIL at 9000. The F/O said we would try. We were only 8 - 9 miles away so I informed FW center we would need a vector. We were vectored 350 degree heading to join CVE 298 R and handed over to approach. We descended as quickly as possible to 9000 but were inside YEAGER and ORVIL on reaching 9000.I had gotten 7 1/2 hours sleep, but felt tired. We had a 2-hour ground time in ELP, which didn't help. I was surprised by this as I felt we were doing fine, but neither the F/O nor myself caught the error. No traffic conflict occurred.

Figure 17: ASAP Reports Involving ORVIL.

The number of reports submitted by pilots grows over time causing the relevance of concepts to change too. Thus, links between terms may assume more or less significance as time progresses and can serve as a valuable tool for knowing whether there are changing patterns in the data.

2.4 Decision Tree

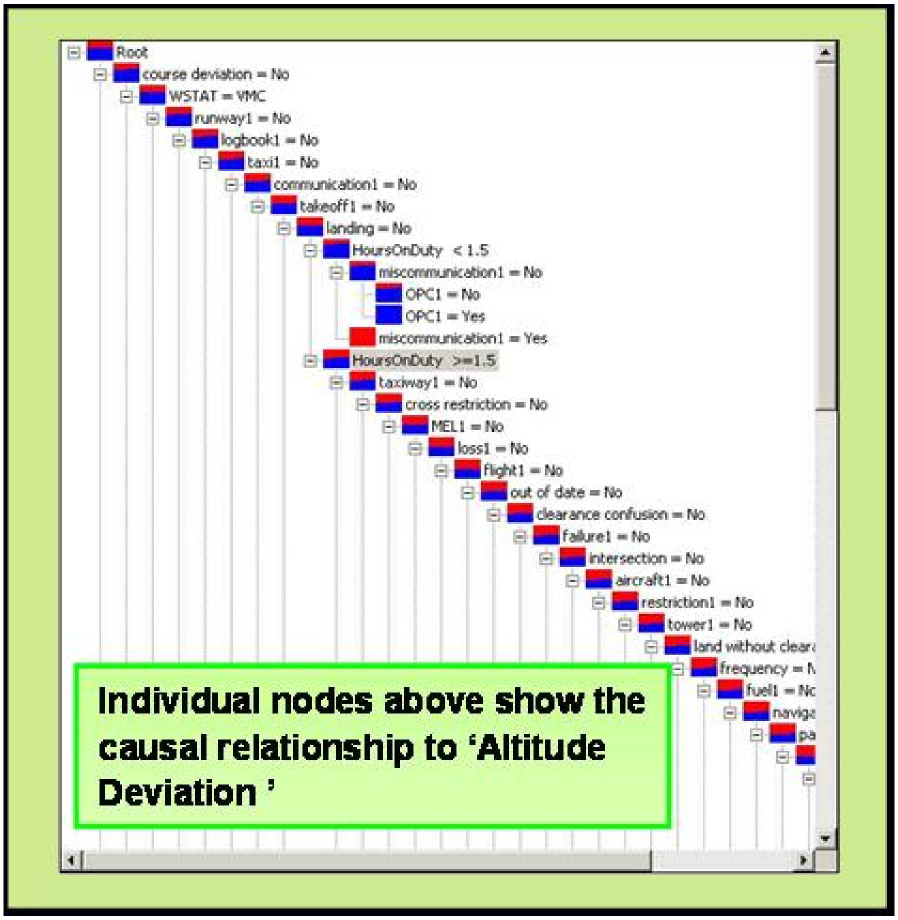
A typical task involved in the processing of a new report is the categorization of the report to one of predetermined classes based on the description of a situation provided in the pilot narrative and other associated characteristics of the situation. The Decision Tree[4] engine of PolyAnalyst elicits a viable classification rule based on automated learning from a large collection of manually categorized historical reports. In the process of self-learning, the system selects sequentially the best parameters and thresholds on these parameters allowing the Decision Tree engine to separate training records to classes of interest with a minimal number of classification errors. The developed hierarchical classification rule is conveniently represented in the form of a tree, justifying the name of the algorithm. This classification rule can be used for automated categorization of newly filed reports.

The Decision Tree report helps the analyst interpret the developed classification rule. It shows the FSO an ordered list of the most influential factors affecting the categorization decision. These factors can be:

- different flight parameters like altitude, weather, navigation system, pilot experience, duration of flight that constitute the structured portion of the ASAP database and
- specific patterns of words and word combinations occurring in pilot narratives - the unstructured part of the ASAP database.

Figure 18 shows a small section of a sample decision tree generated to classify ‘altitude deviation’ events from other records.

[4] A decision tree is an exploratory method used to classify individual records into separate target classes based on automatically identified splitting criteria. This classification model can show how different attributes (text already extracted from narratives as well as existing structured attributes) predict a dependent attribute.



Individual nodes above show the causal relationship to 'Altitude Deviation '

Figure 18: Identify Causal Relationships Using Decision Tree

The data used contained both the terms extracted from narratives as well as the original structured data. Each node (box) on the tree represents a set of records formed by splitting the parent node by a Decision Tree-identified criterion to get a more homogeneous set of records of 'altitude deviation'. The selected splitting criterion is mentioned to the right of each node. The percentage of 'altitude deviation' records in a node is pictorially represented by red (the share of 'altitude deviation' cases) and blue (the share of non- 'altitude deviation' cases) colors. Thus, for example, a totally red node denotes a set of records all related to 'altitude deviation', while a totally blue node represents no 'altitude deviation' records.

In the "leaves" of the tree – terminal nodes with no children – the decision of the record belonging to one of the classes is made. For example, let us consider the first terminal nodes found on the fragment of a classification tree in Figure 18. The first branch of the decision tree rule illustrates that if one concentrates on records where 'Weather' was described as *VMC* and pilot narratives did not contain any of the terms:
'*course deviation*',
'*runway*',
'*logbook*',
'*taxi*',
'*communication*',
'*takeoff*',

‘*landing*’,

and ‘Hours on Duty’ were less than 1.5, then one can efficiently categorize these records to those associated with ‘altitude deviation’ and not by looking at the following factors:

a) If the pilot narrative contains a term ‘*miscommunication*’ then one can be certain that the record is related to ‘altitude deviation’.
b) When the narrative does not contain ‘*miscommunication*’, then with a high probability this record is not related to ‘altitude deviation’. And if the term ‘Operational Process Computer (OPC)’ is additionally present in the narrative, one can be absolutely certain that the record is not related to ‘altitude deviation’.

And records that did not fit some of the criteria listed above will be categorized with the help of other branches of the developed decision tree.

Notice that both structured attributes and pilot narratives were utilized for developing the best classification rule.

The same methodology can be used to highlight hidden co-occurrences in structured and narrative information (that could either confirm or challenge current thinking) and provide an avenue to initiate new incident monitoring practices.

3.0 Assessment of Results by Southwest Airlines

The following assessment of the results of this proof-of-concept was made by the following individuals from Southwest Airlines: Captain Mike Hinnenkamp, ASAP Manager; Mr. Tim Logan, Director, Flight Operational Safety; and Ms. Wy Teter, ASAP Administrator.

The analysis conducted during the proof-of-concept delivered useful and intriguing results. The following features of PolyAnalyst were found to be of particular value:

- **Automatic extraction of important patterns of terms from pilot narratives**. The existing analysis process required initial manual reading and coding of pilot narratives, so that incidents related to a certain issue could be grouped together through a simple database query. Subsequent analysis was performed through a precise keyword search and thus relied on the analyst to anticipate the results of the analysis prior to asking a question. PolyAnalyst enabled the user to quickly perform unsupervised extraction of stable patterns of terms occurring together in separate pilot narratives and presented results of the analysis in a convenient visual form.

- **Focusing the analysis on categories of interest**. PolyAnalyst allowed the user to extract and summarize all terms related to a particular category – for example, “equipment” – without specifying all possible equipment pieces that might be encountered in the text. The extraction mechanism of PolyAnalyst performed generalizations based on the underlying structure of semantic dictionaries of English. Southwest amended these dictionaries with project-specific extensions, listing certain domain-specific terms. The category-based extraction was an obvious improvement over the standard exact keyword based search, which saved the Flight Safety Officer from performing the analysis numerous times each analysis session.

- **Analysis of structured and unstructured data simultaneously**. Traditional analysis provided means for separate analysis of either structured data or manual coding and keyword search of textual narratives. The capability of PolyAnalyst to analyze all available data at once, both structured and unstructured, helped generate additional insights in the analysis of safety events. The Text Analysis algorithm extracted key terms from pilot narratives, and Link Charts and Link Analysis algorithms permitted analysts to easily see correlations between the occurrence of these terms and structured attributes such as airport code, aircraft type, and type of navigation in use.

- **Ability to process large amounts of data**. Traditional reading and manual analysis of pilot narratives is time consuming, therefore limiting the amount of data that can be included in any particular investigation. Data are extracted primarily based on the codes assigned by a FSO during the initial manual analysis of pilot narratives. By providing automated tools for narrative analysis, PolyAnalyst allowed FSOs to explore more ASAP reports during an analysis session than was previously possible.

- **Quick generation of a variety of useful graphs capturing key findings**. The saying that a picture is worth a thousand words holds true in the analysis of safety event data. PolyAnalyst provided several efficient graphing capabilities for the visualization of the relationships discovered in the ASAP data. In particular, Link Charts, Link Diagrams, and Snake Charts facilitated quick navigation and better comprehension of the results. A quick assessment of visually represented results of automated analysis directed the flight safety officer's attention to potential issues. These visualization capabilities optimized the usage of the FSOs time, and helped create easy-to-understand executive reports, thus making the results available to more decision makers in a timely manner.

- **Intuitive drill-down from the graphs to original data records**. The FSO could visually interact with the results of the automated analysis by drilling down to the corresponding ASAP reports. Key terms identified during narrative analyses were highlighted in the reports, which simplified the analysis process. The drill-down capabilities of PolyAnalyst enabled visual dynamic querying of the database, which had a significant timesaving effect over form-based static querying.

The wide range of analysis techniques built into the system opens up an avenue to explore the data and expands the scope of analysis to carry out automated categorization of ASAP reports. However, the data analysis project conducted at Southwest, with the help of PolyAnalyst, currently represents a multi-step interactive process requiring inputs from the analyst working on the project. Southwest Airlines' desire would be to automate and streamline this process to enable analysts/investigators to easily direct the investigations.

An area where some customization and development efforts would be useful is in the ability to refine charts based on the results discovered during drill down. This will help FSOs generate more accurate and publishable reports.

The chosen analysis methodology for discovering knowledge from data is intuitive, especially for narrative analysis. The ability to efficiently interact with visual reports to drill down and verify importance is a critical step in the search for hidden trends and patterns.

Based on the proof-of-concept, some direct tangible benefits of deploying the system could be achieved in:

a) Efficient utilization of FSO analysis time.
b) Identifying hidden safety issues, trends, and relationships.

A customized interface with value-add functionalities will enable users to routinely use the system and conduct other analyses. Overall, the systems ability to bypass the need for pre-categorizing concepts for the purpose of visualization enables it to be used as a very good reporting tool, even for non-technical users.

4.0 Summary

The purpose of this proof-of-concept project was to evaluate practical usefulness of data and text mining tools applied to the analysis of aviation safety incident reports, and to develop methodologies for the analysis of de-identified data in the ASAP database of Southwest Airlines.

Textual data from ASAP reports (pilot narratives) were analyzed with the help of semantic text analysis algorithms of the PolyAnalyst data and text mining system. Extracted patterns of terms were utilized in further knowledge discovery processes together with the structured data in the database. A variety of machine learning and visualization algorithms were then utilized during this process.

The results proved to be useful to the Flight Safety Officers (FSOs) for identifying potential safety issues across different multiple attributes, such as aircraft type, airport, and type of navigation used. The ability to automatically view textual terms extracted from the narratives, across multiple attributes gave the FSOs insight into the data. Multi-correlation analyses augmented the process by visually showing strong associations between attributes of interest.

The project demonstrated that significant value could be generated through:

- Capturing previously unanticipated knowledge from raw data
- Efficient use of analyst's time
- Automation of repetitive processes
- Quick, intelligent analysis of textual data, and
- Consistent and comprehensive utilization of *all* data (structured and unstructured).

Additionally, it is likely, based on the proof-of-concept, that similar data and text mining analysis techniques can be applied to information derived from other data sources in an aviation organization, such as airport safety and aircraft maintenance, to deliver a more comprehensive picture of overall safety issues.