New Capabilities of PolyAnalyst Text and Data Mining Applied to STEADES Data at the International Air Transport Association (IATA)

A Technology Demonstration

In Partnership with the Federal Aviation Administration and the Global Aviation Information Network (GAIN)

Report Prepared by:

Dr. Sergei Ananyan
President
Megaputer Intelligence

Mr. Michael Goodfellow
Assistant Manager, STEADES
International Air Transport Association (IATA)

October 2004

Disclaimers; Non-Endorsement

All data and information in this document are provided "as is," without any expressed or implied warranty of any kind, including as to the accuracy, completeness, currentness, noninfringement, merchantability, or fitness for any purpose.

The views and opinions expressed in this document do not necessarily reflect those of the Global Aviation Information Network or any of its participants, except as expressly indicated.

Reference in this document to any commercial product, process, or service by trade name, trademark, servicemark, manufacturer, or otherwise, does not constitute or imply any endorsement or recommendation by the Global Aviation Information Network or any of its participants (e.g., FAA) of the product, process, or service.

This project report and the results are based on a time-constrained proof-of-concept demonstration carried out by Megaputer Intelligence for the International Air Transport Association (IATA) STEADES group and involved analysis of only a subset of data. Therefore, the results should not be used to form any conclusions or business decision.

Notice of Right to Copy

For further information on this Technology Demonstration

Dr. Sergei Ananyan
President
Megaputer Intelligence
120 W. 7th Street, Suite 310
Bloomington, IN 47404 USA
s.ananyan@megaputer.com
+1 812-330-0110
www.megaputer.com

Mr. Andy Muir
GAIN Program Office
Federal Aviation Administration / FSAIC
800 Independence Avenue, SW
Washington, DC 20591 USA
andy.muir@faa.gov
+1-202-267-9180
www.gainweb.org

Table of Contents

Acknowledgements

This project was funded by the US Federal Aviation Administration, Office of System Safety (known as "ASY" within the FAA) to facilitate the application of advanced methods and tools in the analysis of aviation safety data with the goal of improving aviation safety industry-wide. The project also involved the support and guidance of the Global Aviation Information Network (GAIN), an industry-led international coalition of airlines, manufacturers, employee groups, governments (including FAA) and other aviation organizations formed to promote and facilitate the voluntary collection and sharing of safety information by and among users in the international aviation community to improve aviation safety. Specifically, the project was guided by principles developed by GAIN's Working Group B, "Analytical Methods and Tools" (WG B), which has been tasked with fostering the use of existing analytical methods and tools and the development of new tools that elicit safety information out of aviation data.

Megaputer Intelligence Inc. would like to thank all the individuals who supported this project at the FAA, in GAIN, and at International Air Transport Association. 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.

International Air Transport Association Project Participants

Mike Goodfellow, Assistant Manager, STEADES, IATA
Jill Sladen-Pilon, Manager, Safety Data Management & Analysis, IATA
John Denman, Manager, Airside Safety, IATA
David Mawdsley, Director, Safety, IATA
Martin Maurino, Safety Analyst, IATA
Henri Guay, Safety Intern, IATA

Federal Aviation Administration

Carolyn Edwards, Office of System Safety
Andy Muir, Office of System Safety
Chris Hart, Office of System Safety
Lee Nguyen, Aircraft Certification Service

Report authors

Sergei Ananyan, Megaputer Intelligence
Richie Kasprzycki, Megaputer Intelligence

Report contributors

Mike Goodfellow, Assistant Manager, STEADES, IATA
Jill Sladen-Pilon, Manager, Safety, IATA
Martin Maurino, Safety Analyst, IATA
Henri Guay, Safety Intern, IATA

Executive Summary

Background

IATA STEADES and Megaputer Intelligence conducted a joint proof-of-concept project in conjunction with FAA and GAIN Working Group (WG) B's efforts to facilitate and promote the use of automated data and text mining tools in the aviation community.

Megaputer Intelligence analytical software PolyAnalyst™ was applied to a de-identified sample of reports describing TCAS events from the IATA Safety Trend Evaluation Analysis and Data Exchange System (STEADES) database. STEADES is a global safety event database providing analysis of incidents, with the goal of reducing accident potential. It is based on open, non-punitive reporting and consists of about 300,000 reports from approximately 40 airlines. IATA also provided guidance and insight on the relevancy and type of results.

The total project spanned a sixteen week duration wherein different analytical methodologies were demonstrated to IATA safety analysts to identify hidden issues from pilot narratives.

Project outline

Megaputer Intelligence carried out the analysis of a mixture of structured attributes and text narratives in reports related to TCAS events. The main goals were to automate various steps of the current manual data analysis process at IATA and study the feasibility of employing text mining technology to:

1) Perform automated monitoring of safety reports for known issues and patterns of interest
2) Demonstrate techniques for visual interactive analysis of text narratives
3) Discover possible unexpected patterns and trends suggested by data.

The project concentrated on the possibility of solving four specific tasks:

1) Carrying out automated categorization of report narratives with respect to WinBASIS hierarchical taxonomy of categories (see Sections 1.6 and 5.2).
2) Exploring pilot narratives for TCAS events using text-based multi-dimensional analysis techniques (see Section 5.4.4).
3) Revealing unanticipated patterns and trends present in data (see Sections 5.3.3 and 5.4.1-5.4.3).
4) Finding reports describing historical events similar to the event under consideration (see Section 5.4.6).

PolyAnalyst exploration engines based on a combination of linguistic, statistical, machine learning and visual data analysis techniques were employed in the analysis of TCAS-related reports covering about 10,000 events. The project was aimed at assessing the applicability and value in aviation safety data analysis of the following PolyAnalyst

engines: Taxonomy Categorization, Text OLAP, Link Analysis, Text Clustering, and Similarities Finder.

Obtained results

Aviation dictionary. To enhance the results of the analysis, Megaputer expanded its existing dictionary of terms and abbreviations specific to aviation domain through automated analysis of multi-airline data and clarifying the meaning of unknown terms in cooperation with IATA specialists. The resulting dictionary included over 1,100 standard abbreviations, airport codes, standard misspells and stop words. A more detailed description of the created dictionary is provided in Section 5.1.1. The project illustrated that this domain-specific dictionary, while being quite far from comprehensive and requiring more work, can significantly enhance the quality of the obtained results.

Taxonomy Categorization. Historically, categorization of safety events to WinBASIS system of categories utilized by IATA was carried out by representatives of participating airlines based on manual evaluation of report narratives. Taxonomy-based Categorization was designed for automating the process of the analysis of narratives to improve the speed and objectiveness of report categorization. PolyAnalyst Taxonomy Categorization was performed on TCAS report narratives. The developed taxonomy branch with defined categorization node patterns was able to accurately categorize 77% of considered reports. Megaputer analysts defined node patterns for categorizing TCAS-related reports in terms of special Pattern Definition Language (PDL) operators. Taxonomy Categorization is discussed in Sections 5.2 and 5.3.

Text OLAP (On-Line Analytical Processing). Tracking safety events across a number of predefined dimensions can help analysts quickly find answers to a large number of possible questions. In this project, PolyAnalyst Text OLAP engine was utilized for generating interactive multi-dimensional reports on a mixture of textual and structured data found in TCAS related reports. A multi-dimensional reporting matrix developed for this project accounted for categorizing 79% of all TCAS related safety reports. The performed automated categorization proved to be correct in 72% of categorized narratives. An interactive visual reporting feature of PolyAnalyst Text OLAP was found to have high potential for aviation safety data analysis applications. Text OLAP is discussed in Section 5.4.4.

Link Analysis. PolyAnalyst Link Analysis engine was utilized for revealing and visualizing patterns of correlations involving values of structured attributes and terms in report narratives. Drill-down feature allowed visual aggregation of all reports supporting the discovered patterns. The Time Filter feature was employed for monitoring time evolution of the discovered patterns. Link Analysis is discussed in Sections 5.4.3 and 5.4.5.

Text Clustering. PolyAnalyst Link Terms engine was employed for finding and visualizing stable combinations of terms in report narratives. IATA analysts found this tool to be useful not for identifying new trends and correlations, but for easily acquiring grouped records that were related. An example of this would be the ability to single out

all TCAS RA reports that mention *closure rate* for further analysis. Link Terms is discussed in Sections 5.4.1 and 5.4.2.

Similarities Finder. Often safety analysts require a capability to quickly identify historical reports that are similar to the considered report. This feature might help identify things that assisted in resolving a similar situation in the past. PolyAnalyst Similarities Finder indexed all TCAS related reports to provide the user with a list of similar past events, each with the corresponding pattern of similar terms highlighted. Similarities Finder is discussed in Section 5.4.6.

IATA feedback

IATA undertook to evaluate carefully and objectively the merits and drawbacks of using the PolyAnalyst system for its STEADES program. The goal of STEADES is to provide members of the service with regular, unbiased industry analysis of global safety trends and concerns. IATA was seeking tools for automating the process of text analysis to save analyst time and increase quality of obtained results.

IATA found that PolyAnalyst Text OLAP feature, described in the Appendix, shows the greatest potential for STEADES analysis. Its ability to combine both structured and unstructured data into a single analysis module offers flexibility to analysts looking to analyze multiple categories of mutually exclusive concepts. The automated assignment of descriptors, as referred in Section 2.3 of this report, also shows future promise in assisting analysts validate that the submitted reports have been properly classified.

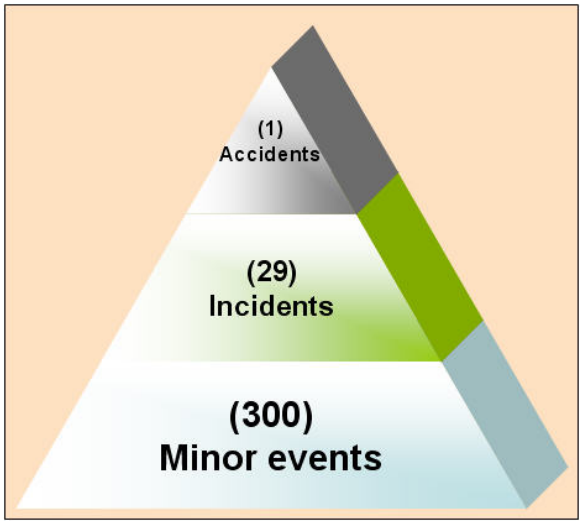
IATA also discovered an inherent limitation in the PolyAnalyst software that will require addressing: the dictionary included in the current product lacks many domain-specific terms and definitions of words pertinent to the global aviation safety arena. Enhancing aviation domain specific dictionaries to the point where complex automated analysis can be carried out while minimizing human intervention will require a considerable investment in time and resources from the industry. Once this single global reference dictionary exists, the capabilities of text-mining software systems such as PolyAnalyst should be greatly enhanced.

This Page Intentionally Left Blank

1. Introduction

1.1. Purpose of the Proof-of-Concept Demonstration

This proof-of-concept demonstration is part of the FAA Office of System Safety and the Global Aviation Information Network (GAIN) Working Group B's (Analytical Methods and Tools) effort 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 in order to reveal associations and trends that may otherwise be difficult and time consuming to identify. The FAA in cooperation with GAIN desires to share the knowledge of this demonstration with others in the aviation community.



Heinrich's Pyramid

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

Though these databases contain significant amounts of critical data, there have been substantial challenges in analyzing the information. Analysis has primarily been focused on only the structured portion of the database (aircraft type, flight phase, etc.), yet experts estimate that over 80% of all useful information in a report resides in the textual (unstructured) format and could contain valuable knowledge.

This project demonstrates an avenue for analyzing all available data, both structured and textual, to derive maximum value from safety data collection investments. Therefore, the primary objectives were to develop a broad methodology of analysis and to demonstrate how knowledge dispersed in a large collection of safety reports can be easily revealed.

[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 the FAA, GAIN Working Group B, Megaputer Intelligence, and the International Air Transport Association (IATA). Megaputer Intelligence provided the analytical software system PolyAnalyst™ and its usage expertise. This software was applied to de-identified safety data from the Safety Trend Evaluation, Analysis and Data Exchange System (STEADES) provided by IATA. 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.

International Air Transport Association (IATA)
Originally founded in 1919, IATA brings together approximately 280 airlines, including the world's largest. Flights by these airlines comprise more than 95 percent of all international scheduled air traffic. IATA allows airlines to operate more efficiently. It offers joint means – beyond the resources of any single company – of exploiting opportunities, reducing costs and solving problems.

Megaputer Intelligence
Megaputer Intelligence provides business intelligence solutions for analyzing both structured and textual data and discovering valuable knowledge in large volumes of data. As a leader in delivering data mining and text mining tools, Megaputer serves hundreds of customers worldwide, including organizations in the insurance, aerospace, financial, healthcare, educational and government sectors.

1.3. Overview of the 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, 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 and analysts to:

a) Reveal hidden issues (irrespective of data type – structured or unstructured)
b) Generate strategic overview charts for management
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 analysts face and can be tailored to a specific application domain. A major portion of the users' involvement is in providing direction to the analysis process and defining their areas of interest. User-defined parameters for analysis are entered into the system throughout the process.

In more advanced implementations of PolyAnalyst, 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: Safety Trend Evaluation, Analysis and Data Exchange System (STEADES)

PolyAnalyst was applied to the analysis of safety event reports in IATA's Safety Trend Evaluation, Analysis and Data Exchange System (STEADES). The STEADES database is a pool of incident reports submitted to IATA by a large number of international airlines. The project aimed at discovering useful patterns and relations contained in the mixture of structured and unstructured data.

STEADES is a global safety event database providing analysis of incidents, with the goal of reducing accident potential and, therefore, costs. It is based on open, non-punitive reporting. The database currently consists of data from approximately 40 airlines and is growing at a rate of about 50,000 records per year. It is IATA's goal that STEADES will eventually cover approximately 95% of all international commercial air traffic as well as a very substantial amount of domestic traffic data.

STEADES members forward their Air Safety Reports (ASRs) to IATA on a quarterly basis, using specifically developed software. This information is collated with data from all other participating airlines and analyzed for trends and issues of concern. IATA analysts then generate and distribute to STEADES members Trend Reports capturing results of the analysis.

1.5. Analytical Objectives

IATA performs aggregation and joint analysis of incident reports received from all participating airlines to reveal and report back to members the discovered trends and current industry-wide patterns and risks in aviation safety. This represents a unique and

valuable service for participating airlines, which becomes possible only upon the pooling of de-identified safety data from multiple airlines.

The main objectives of the analysis are as follows:

1. Periodically determine an industry-wide list of important current problems and trends.
2. Monitor trends and patterns related to known issues of high importance or high risk, such as TCAS related events.
3. Investigate causes, consequences, risk factors and other patterns related to the discovered group of most important events.
4. Track performance and industry acceptance of selected policies and technologies.
5. Compare new events and patterns to previous analysis periods, identify key trends, and predict future developments.

Then IATA analysts summarize their conclusions in the form of easy to comprehend reports delivered to safety officers at participating organizations.

1.6. Data Contents

De-identified summaries of incident reports in a pre-defined format are received in electronic form from over 40 participating airlines and stored in the WinBASIS system from British Airways. There are about 50,000 reports filed each year, with predicted growth in the future. The system has accumulated almost 300,000 reports dating back to 1998. The current project involves the analysis of about 9,800 TCAS event reports covering the time period from 1998 to June of 2003. Records involving the term TCAS either in event title or event summary were included in this data set, which is referred to as TCAS data set in Figure 1.

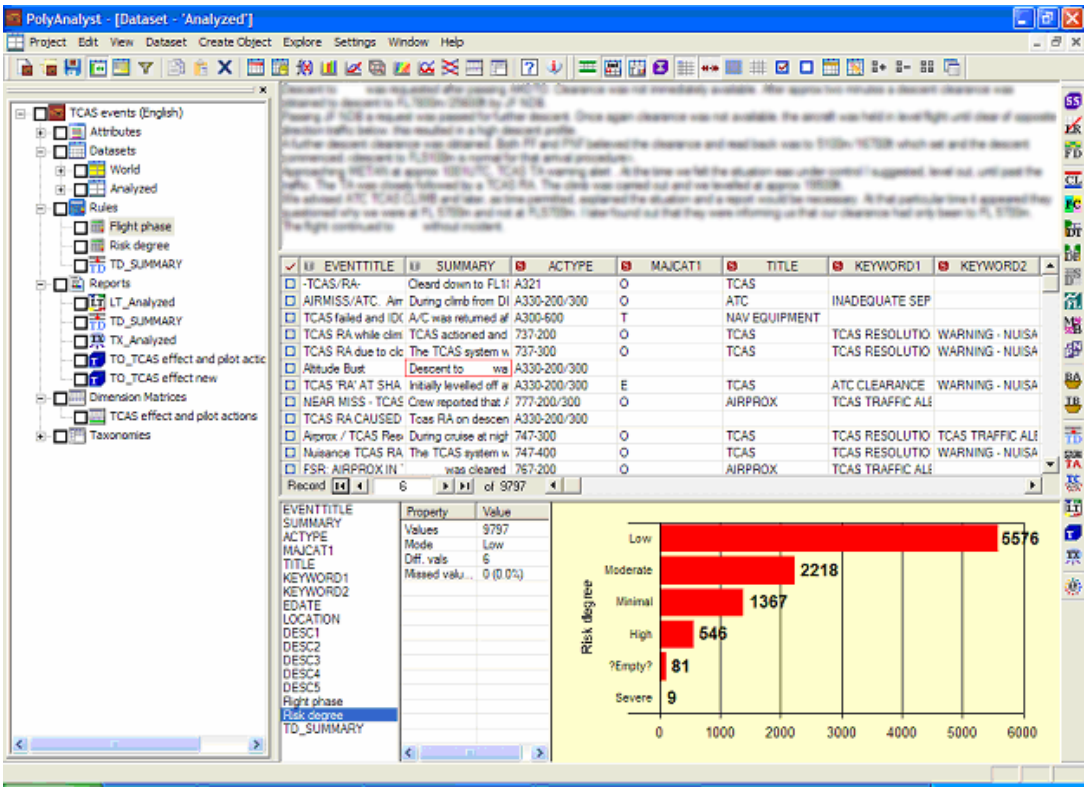


Figure 1. A view of STEADES data imported in PolyAnalyst. A histogram in the bottom right pane is an interactive representation of the distribution of the risk degree of events.[2]

Through the analysis of TCAS incidents, IATA analysts wish to monitor strong and weak points of the TCAS system, its effectiveness, ATC versus system instructions and system limitations. Megaputer analysts developed a taxonomy to recognize and track issues of interest to IATA specialists, which arise in TCAS related situations.

Structured data fields and summaries

The STEADES data is a combination of both structured data and textual descriptions of safety events. Data fields include the title and summary of the event provided by the contributing airline representative, as well as the event date, location, type of the aircraft involved, flight phase, and the risk level assigned by the airline to the event. In addition to these attributes, a system of *descriptors* exists for categorizing each event according to a preset taxonomy in order to facilitate future retrieval and analysis.

Descriptors and risk factors

Flight Safety Officers at participating airlines can specify the risk degree for each reported safety event, as well as assign to each event one or more relevant categories from an event categorization list included in BASIS.

[2] All PolyAnalyst screenshots in this report were edited to disguise any sensitive information and preserve privacy of data.

To ensure the highest quality of analysis, IATA analysts continuously work with participating airlines toward improving the analysis process, including data submission. IATA and British Airways developed a standard classification system with individual categories called descriptors. The descriptor system is organized into four levels: Operational Effect, Immediate Effect, Event Type, and Event Descriptor. Pairs of Event Type and Event Descriptor form a two-level taxonomy for categorizing safety events, where Event Descriptor nodes have Event Type parents. Typically, a safety record is assigned from one to six descriptors.

In the absence of text mining tools, manual categorization of safety reports by airlines was the only mechanism for preparing them for further trend analysis. Upon the introduction of text mining capabilities, it may become possible for IATA to develop an automated process for centralized categorization of safety reports to the descriptors. This potentially standardizes and facilitates quick and convenient monitoring of trends and patterns in data. An initial evaluation of this capability can be found in the IATA Feedback section below.

2. Analytical Process

2.1. Current Analytical Process at IATA

Current data analysis processes at IATA involve extensive manual data analysis. The main source of information about every incident report is the natural language summary of these incidents. The analysis of incident summaries requires in-depth manual analysis in the absence of efficient text mining tools. In fact, most of the work performed by IATA analysts involves the exploration of text fields containing titles and summaries of safety events: reading through event descriptions and manually determining trends and patterns of interest.

IATA analysts focus their research on groups of safety events based on industry concerns and member inputs. Dedicated analysis for a particular issue can be carried out on a request from a participating airline.

The process of safety data analysis contains three main steps:

1) **Problem areas identification**. First, incident reports are aggregated and compared across all descriptors to determine problem areas. Second, certain categories of key importance, for example TCAS events, are monitored on an ongoing basis. Third, safety events that have high risk level assigned to them by an airline are focused on (the risk level is assessed based on a combination of the severity of the event and the probability of its reoccurrence). An example of the high risk incident would be a high-speed rejected take-off.

2) **Analysis for causes, consequences, trends and patterns**. With the help of a built-in descriptor search system, IATA analysts extract reports related to the selected groups

determined to be most important. The corresponding reports including textual summaries are printed out and then manually investigated by IATA safety analysis experts for patterns, trends, causes and consequences. This involves reading event descriptions, discovering relationships in data, and identifying prevention strategies. This is the most manual labor intensive step of data analysis carried out at IATA.

3) **Results summarization and reporting**. The results of the analysis and analysts' conclusions are captured in comprehensive reports including graphs, text and tables, delivered to Flight Safety Officers at participating airlines.

2.2. Challenges of Current Analytical Process

The main challenge of the current analytical process derives from its heavy dependence on manual data processing. This challenge is common across the aerospace industry: the volume of data requiring analysis is huge and growing quickly. The fact that the most useful information is located in natural language text narratives further complicates the task.

Individual challenges

1) **Proliferation of text data**. Airlines are accumulating data about minor safety incidents at the rate of up to 5,000 reports per year. A large portion of useful information about a safety event is typically stored in the text narrative describing the event. This is inevitable in the description of any complex system or situation because no preset structured framework can account for all possible future event characteristics.

   The implementation of event descriptors is the first step in removing the necessity for IATA analysts to read ALL text reports, achieved by delegating this responsibility along with the task of structuring (classifying) all reports to representatives of member airlines. While being the most convenient arrangement at the time, this methodology introduces a set of new challenges.

2) **The descriptor system**. In order to promote standard organization of safety reports across all participating airlines, IATA and British Airways developed an extensive system of descriptors, as discussed above, for comprehensive classification of safety events. The price to pay for this comprehensiveness is that it is sometimes difficult even for a professional analyst to select the best set of descriptors for an accurate characterization of all important aspects of the considered safety event out of several hundred of available descriptors. This can cause unintended omissions and errors in event coding.

3) **Third party classification using descriptors**. While the system of descriptors is devised by IATA and British Airways, the process of actually tagging reports with descriptors has to be carried out by representatives of participating airlines. This can create additional confusion and cause unintended errors in event coding.

4) **Recent change in the system of descriptors**. The current IATA standard classification system of event descriptors is quite new. Until 2002 there was an older event classification system based on keywords, which was far too generic for the purposes of global safety analysis. IATA is encouraging STEADES program members to move to the new system. This has three implications.

    a. *Old system confusion*. The inflexibility of the old keyword system caused many airlines to be quite creative in how they classified events. There is a wealth of information that IATA analysts have difficulty getting to in the older system because many of the events were poorly classified due to the limitations of the old system.

    b. *Challenges with the new system*. Due to some initial confusion over how the new system works and people's varying ideas at the time of classification, many events have either not been classified at all, or worse, poorly classified even in the new system of descriptors. An example might be an incident that gets classified as an altitude deviation during descent, where the real cause of the event is a TCAS alert. If an event is improperly classified this way, it would possibly remain unnoticed by IATA analysts. IATA analysts suspect that there are quite a few errors of this kind in the data. However, such errors are extremely difficult to find.

    c. *Mapping between old and new systems*. Even under the assumption that all events in both the old and new systems are classified 100% correctly, the transition to the new event categorization system causes other temporary difficulties. There is no set of unequivocal rules providing the mappings between the old and new sets of categories. This further complicates joint analysis of old and new data, making trend analysis very difficult during the transitional period.

5) **Unexpected patterns and trends**. Descriptor taxonomies are very useful for monitoring data for known issues of interest, but they provide no help at all in discovering unexpected patterns. A newly developing (and thus highly important) trend can easily reveal itself in a collection of events scattered across different categories in a preset taxonomy. In this situation, relying on purely taxonomy-based categorization might further mask the new trend. The task of finding clusters of highly correlated issues throughout a mix of structured data and text narratives can be more efficiently solved with a combination of modern text mining and link analysis techniques. IATA offers further comments on this issue in Section 4.

6) **Creating and testing hypotheses**. Upon the determination of top problem categories, the real work for a safety analyst is only starting. Next, one has to pinpoint key patterns, trends, causes and consequences in data. For IATA analysts today, a typical batch of reports from a single category requiring further heavy manual analysis can contain up to 400 incident summaries. It takes thorough

industry expertise and lots of diligent manual work to discover reliable and valuable knowledge in this data.

7) **Summarizing analysts' findings in reports**. Upon completing the analysis of safety report summaries, IATA analysts generate a STEADES report outlining the latest trends and patterns and highlighting risk factors across the aviation industry for participating airlines. These regular reports are distributed to airline representatives by e-mail at the moment. IATA is also publishing these reports at a password-protected web site to facilitate on-demand "any time – anywhere" delivery of analytical reports to its members. The current process of generating a report relies heavily on manual efforts and is time consuming. IATA analysts have to summarize and record their findings and use other tools such as Excel to generate the corresponding graphs. Deeper integration of the data analysis and report generation processes is highly desirable.

2.3. Capabilities of PolyAnalyst

Megaputer Intelligence develops text mining tools and methodologies for automating the most labor-intensive steps of the safety report analysis process. The application of the Megaputer data and text mining system PolyAnalyst to the analysis of safety data at IATA is expected to provide the following main effects:

1. Automating preliminary stages of data analysis and saving analysts' time for interpreting the results (e.g. removing duplication and/or short summaries).
2. Current validation of previously classified events. There may be the potential of shifting the responsibility of initial data categorization from airline representatives to computer algorithms monitored periodically by IATA analysts in the future. IATA's assessment of this capability can be found in Section 3.1.
3. Providing more elaborate and quick reporting capabilities.

3. Assessment of Results by IATA

The assessment of the results of this proof-of-concept was provided by Mike Goodfellow, Assistant Manager, STEADES at IATA. Jill Sladen-Pilon, Martin Maurino and Henri Guay, all of IATA, also contributed to this section.

The analysis conducted with the help of PolyAnalyst during the project offered some useful and interesting insights into both the state of the industry in text-mining capabilities and what is required to successfully implement this type of solution at an organization such as IATA. IATA also used data from beyond the TCAS scope of this proof of concept to assess the software's abilities in a wider aviation safety analysis field. IATA recognizes that its database is truly unique in the world, and not necessarily similar to what would be found at an airline's flight safety office. It does however prove to be a strong challenge to the capabilities of PolyAnalyst or any other data mining system.

Although there are many different analysis options in the software, IATA focused on a few areas that were identified as likely to have the most potential for industry safety analysis:

3.1. Automated assignment of descriptors

For the STEADES program, IATA currently uses a classification system that was co-developed by IATA and British Airways to classify Air Safety Reports (ASRs) as they are entered into an airline's safety database. PolyAnalyst was able to classify a sample of TCAS data with reasonable accuracy and some false-positives. The taxonomy engine successfully classified 77% of the sample data set. The pre-built taxonomy supplied by Megaputer was modified to account for TCAS jargon likely to be used in ASRs. This module shows eventual promise for being able to help classify earlier events not coded in the descriptor classification system and to assist in accommodating other electronic safety reporting systems not using the STEADES descriptor system and wishing to participate in STEADES. The other sections of the descriptor classification system were not sufficiently defined in PolyAnalyst and therefore tended to produce erratic results. Considerable time in the development of the complete taxonomy and dictionary used in PolyAnalyst would be required to extend this categorization functionality to the entire STEADES descriptor classification system.

3.2. Text OLAP (On-line Analytical Processing)

IATA has been using OLAP-like analysis tools for the STEADES program in the production of the STEADES Safety Trend Analysis Reports since the program started. However, these capabilities have been focused only on the structured data in the database and not the text narratives or titles. The PolyAnalyst Text OLAP feature was carefully evaluated to see how it can be applied to routine STEADES analysis. The OLAP model developed by Megaputer was evaluated to determine: a) what percentage of ASRs was ignored because PolyAnalyst was unable to determine a suitable category and b) what percentage of the records not ignored was correctly assigned into a category by PolyAnalyst. IATA's tests showed: a) 21% of records were ignored by PolyAnalyst because they did not match any pattern defining individual categories in the proof-of-concept project and b) 72% of the remaining records were properly assigned correct categories. Overall, this feature shows potential in separating clear-cut, mutually exclusive categories from the text narratives such as the class of airspace a TCAS RA occurred in. With respect to the STEADES database, further development of the dictionaries would be required for the user to take better advantage of this tool's capabilities. Refer to Appendix A for more details on the Text OLAP feature.

3.3. Link Terms

This tool proved to be useful not for identifying new trends and correlations, but for being able to easily acquire grouped records that were related. An example of this would be the ability to single out all TCAS RA reports that mention *closure rate* for further

analysis. Expected correlations such as altitude deviation did not immediately identify themselves due to differing terminologies used in the reports.

3.4. General Comments

The system was generally easy to use and quite intuitive in certain aspects. Once models were set up, minimal training was required for analysts to begin using the software. Importing data into the system was simple for someone with database experience. Exporting data via HTML is very simple, although exporting into non-HTML formats was not intuitive. The current non-HTML data export process requires the user to create a new data set from the drill-down results and then export it, which adds extra steps to the analysis process.

The software's ability to combine both structured and unstructured data into the same analysis model shows potential for an analysis of TCAS events at particular airfields. Also, the included pattern definition language used to build taxonomies, rules, and OLAP cubes was intuitive and simple for non-technical users to grasp. The system discovered several patterns already known to IATA analysts that confirmed its ability to identify trends.

3.5. Limitations

IATA strongly feels that the dictionary needs to be further enhanced to offer a truly automated system for global analysis. The STEADES database currently houses records from about 40 contributing airlines over five continents with varying reporting styles and cultures, as well as differing terminology for similar events. Although the previous work from the Southwest Airlines project was applied in the base dictionary, further work is still required to create a global reference of airline terminology and contexts. This work will undoubtedly take some time to perform and perfect. That said, once a quality aviation-specific global dictionary exists, automated text mining should provide valuable insights into the airline industry's safety concerns.

4. Summary

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

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 used in later processes together with the structured data in the database. A variety of machine learning and visualization algorithms were utilized during this process.

The project demonstrated that value is generated through:

- Using the software to extend the analytical capabilities beyond the existing classification system.
- Efficient use of analyst's time for many tasks
- Automation of repetitive processes
- Quick, intelligent analysis of textual data
- Consistent and comprehensive use of both structured and unstructured data.

IATA undertook careful and objective evaluation of the merits and drawbacks of using the PolyAnalyst system for its STEADES program. IATA found that PolyAnalyst's Text OLAP feature, described in the Appendix, shows the most potential for STEADES analysis. Its ability to combine both structured and unstructured data into a single analysis module offers great flexibility to analysts looking to analyze multiple categories of mutually exclusive concepts. The automated assignment of descriptors, as referred in Section 2.3 of this report, also shows future promise in assisting analysts in the validation of the fact that the submitted reports have been successfully classified.

IATA also discovered some inherent limitations in the PolyAnalyst software that will require addressing. The dictionaries, as they pertain to global aviation terminology, need development to make them truly representative of the global aviation safety arena. This will require a considerable investment of time and resources from the industry to upgrade these dictionaries to the point where complex automated analysis can be carried out with a minimum of human intervention. Once this single global reference dictionary exists, the capabilities of text-mining software systems such as PolyAnalyst should be greatly enhanced.

5. Appendix: Results of analysis with PolyAnalyst

The entire process of safety data analysis at IATA can be split in five major steps:

1) Data preprocessing
2) Safety report categorization
3) Problem areas ranking
4) Discovery of trends and patterns
5) Report generation

PolyAnalyst provides tools for assisting investigators through different steps of the data analysis process.

5.1. Data Preprocessing

Typically, safety data is stored in a form that requires some preprocessing before in-depth machine analysis can be performed. The first steps of intelligent analysis involve understanding the data, evaluating data quality and performing data cleansing and integration. The success of further analysis critically depends on the effectiveness of these first steps.

5.1.1. Domain-specific dictionary

The application of text mining techniques in a particular domain can be further enhanced by providing the system with additional background knowledge about the considered domain. The PolyAnalyst Dictionary Editor allows the user to specify and reuse relevant terms and relationships broadly utilized in the explored domain.

Megaputer carried out the development of a dictionary specific to the context of this proof-of-concept project for TCAS analysis in cooperation with IATA domain experts. The dictionary will require extensive development to be truly useful in global safety analysis. It incorporates background knowledge of domain experts limited to TCAS and can be re-used in future data analyses performed with the help of PolyAnalyst. It enhances the quality of analysis and the ease of comprehension of the results.

The following dictionaries were developed:

1) *List of abbreviations and other unknown terms*. First, a list of unrecognized terms and abbreviations was created by PolyAnalyst during the first pass through the data. Industry expertise of flight safety analysts at IATA, along with experience gained from previous studies with Southwest Airlines and FAA, was utilized for providing expansions for standard airspace abbreviations and jargon terms.
2) *List of frequently encountered terms that are synonyms within the aviation field*. A list of the most frequent terms was shown to aviation specialists to provide possible synonyms to these terms.
3) *List of stable phrases in aviation field*. A list of stable word collocations discovered by PolyAnalyst in raw data was checked by aviation specialists to leave only significant word combinations.
4) *List of airport codes and navigational fixes*. Megaputer analysts collected from public sources a comprehensive list of world airport codes and their expansions. Navigational fixes were interpreted by airspace specialists.

The dictionaries were developed through a series of iterative steps. Using PolyAnalyst to interpret known terms with the help of underlying semantic dictionaries, Megaputer analysts developed a list of terms that were not recognized by the system. Megaputer provided a list of unknown terms and abbreviations, and a list of terms suggested to be added to the "ignore" list (terms from this list encountered in the text are not included in the analysis) based on the results of automated analysis of text. IATA analysts helped decipher the meaning of the corresponding terms, for example, they helped interpret common airspace abbreviations for standard terms and navigational fixes. Airport codes and navigational fixes were expanded to their full form in those cases when there was no other stable expansion of the corresponding abbreviation frequently used in safety reports. The resulting dictionary included over 1,100 standard abbreviations, airport codes, standard misspells and stop words.

Transformations. Transformations allow the user to define sets of synonyms, provide expansions for abbreviations in the text, and supply corrections to typical misspells to be corrected by the system automatically. Transformations ensure that different representations of the same object are counted as the same term, as well as help the user

better interpret the results of the analysis. For example, pilots from different airlines might use different abbreviations for the word *aircraft*: *A/C*, *ac*, *acft*, *acfts*, etc. During the analysis all these abbreviations should be mapped to the word *aircraft* in order to obtain accurate results.

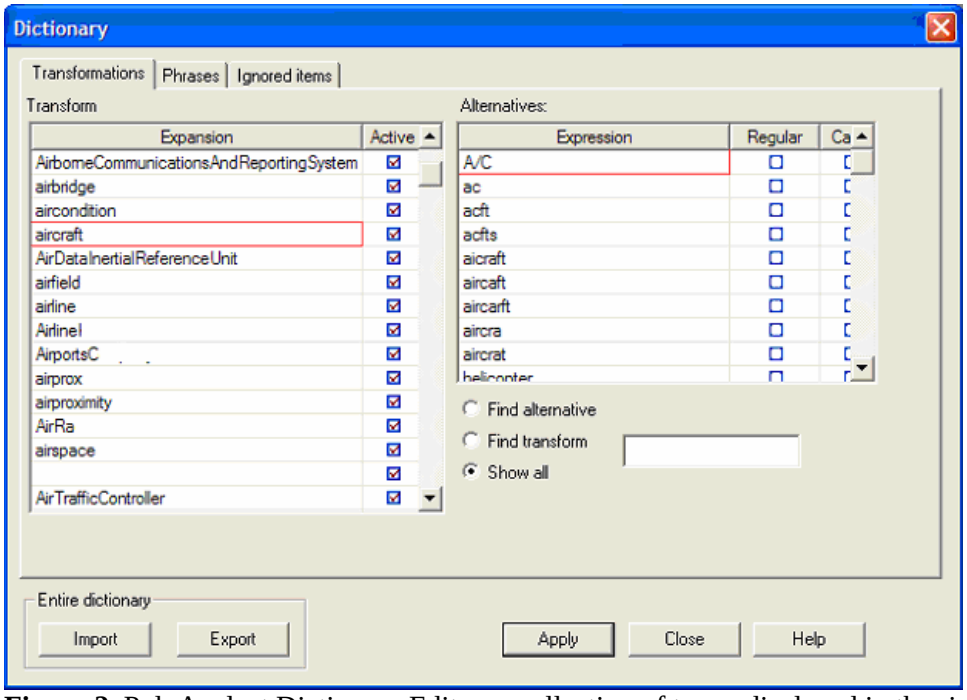


Figure 2. PolyAnalyst Dictionary Editor: a collection of terms displayed in the right pane, such as *A/C*, *ac*, *acft*, *acfts*, etc. are mapped to the word *aircraft*.

CeilingAndVisibilityOK	CAVOK
checked	CHKD
checked	CK'D
checked	CKED
configuration	CONF
configuration	CONFIG
continue	CONT

Figure 3. A fragment of dictionary mappings utilized in the project.

Phrases. This mechanism helps the user define a list of stable phrases specific to the considered application domain, which are not split into individual words during the analysis.

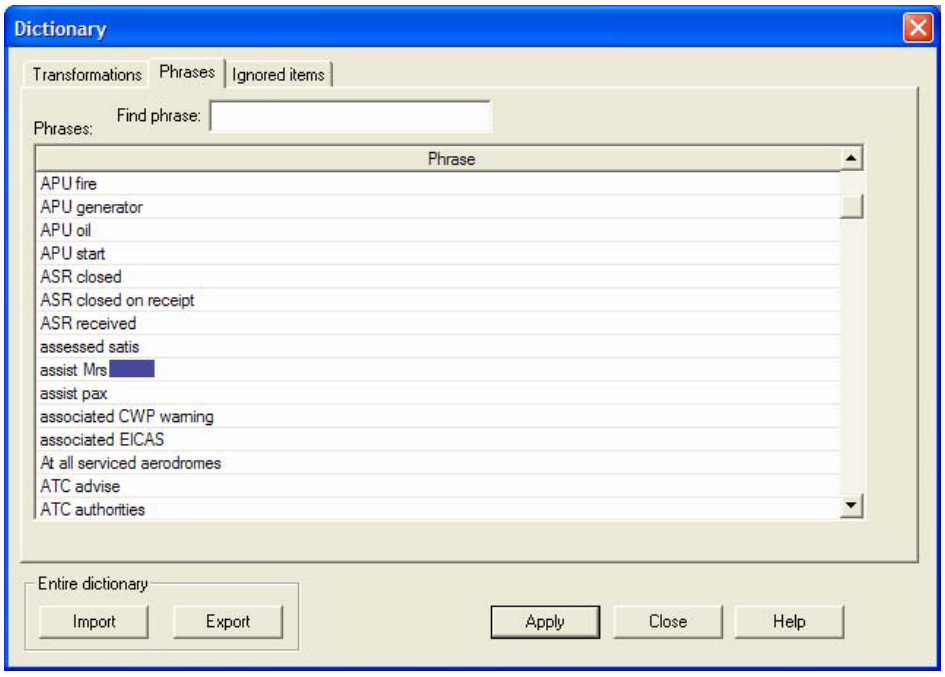


Figure 4. Collection of stable phrases PolyAnalyst captured during the analysis of safety reports.

Ignored items. This mechanism allows the user to specify terms and phrases that should be ignored in the analysis.

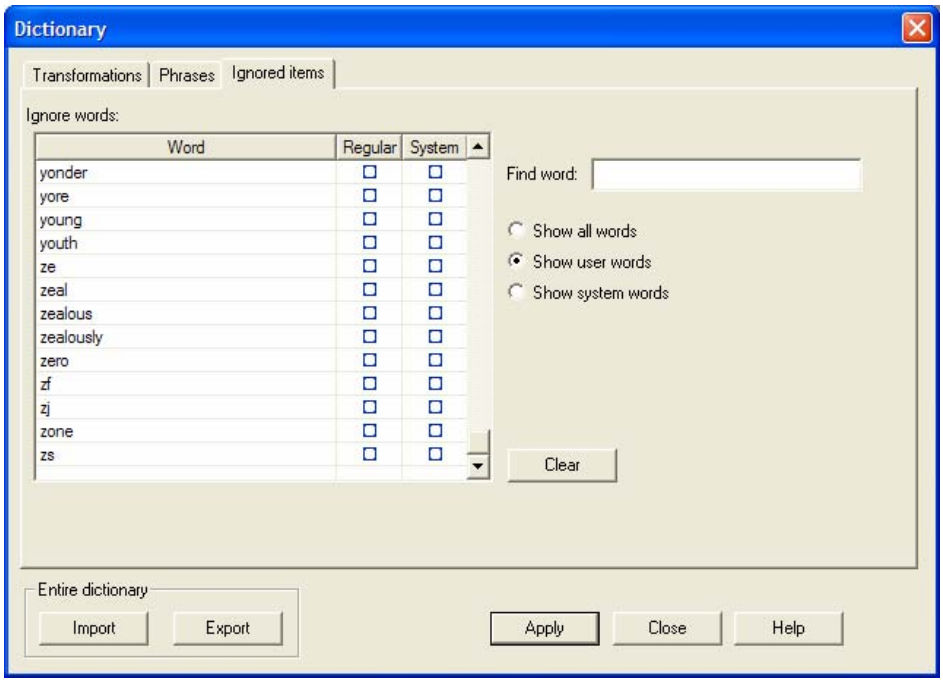


Figure 5. A fragment of the list of ignored terms built into the system.

PolyAnalyst's Regular Expressions Machine allows the user to extract entities matching certain lexicographic patterns. For example, currency amounts, dates, time stamps, and telephone, social security and driver license numbers can all be easily extracted by this mechanism. As another example, the user can use the pattern **B7?7** to extract all typical references to commercial aircraft manufactured by Boeing.

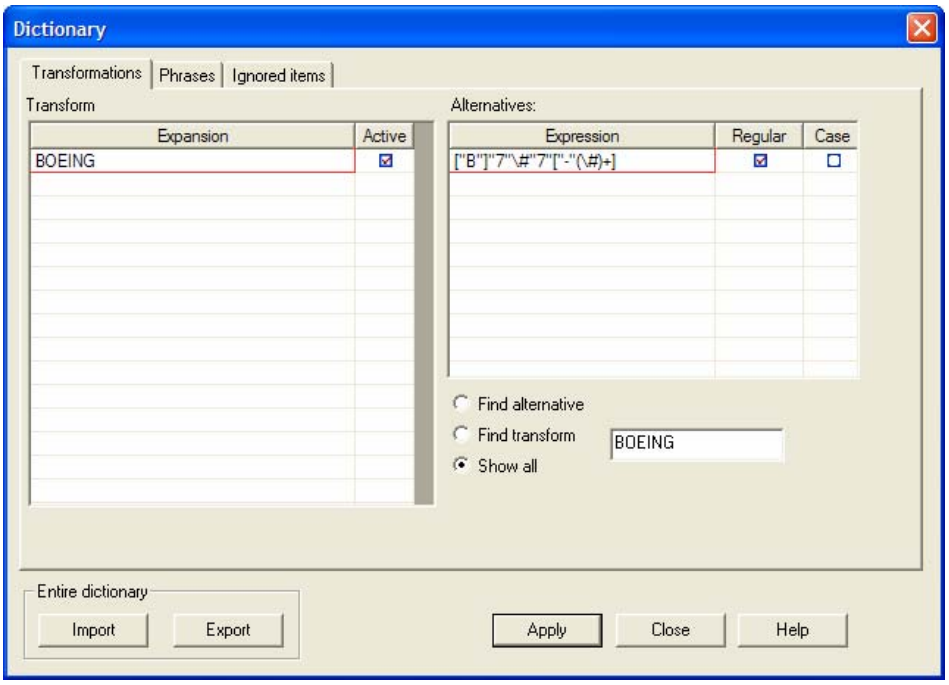


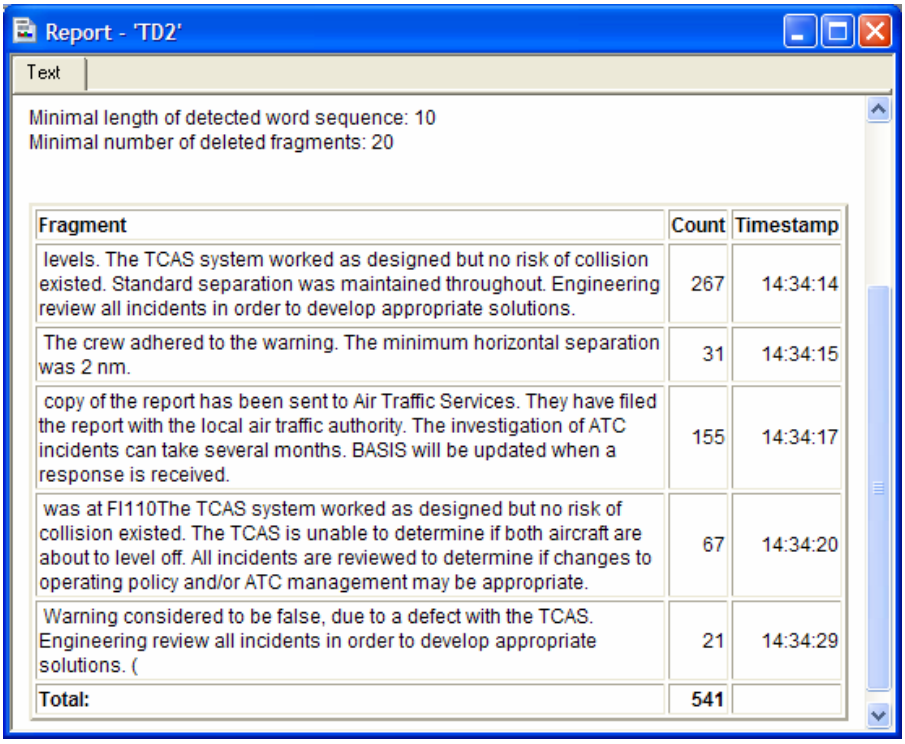
Figure 6. PolyAnalyst mechanism of regular expressions allows the user to define any lexicographic patterns to be matched.

The definition of an expression named BOEING, displayed in Figure 6, in the language of Regular Expressions states that it should be matching any fragments containing 7?7 with a possible B in front of this combination or a number of digits after a dash following this combination. A question mark "?" in this construction stands for any digit coming between two sevens. Thus defined expression was recognizing symbol sequences matching the defined pattern, such as 747, 737-400, B767, etc., as instances of BOEING aircraft.

5.1.2. Duplicate texts and almost identical fragments

To clean data from duplicate text records, a new text analysis engine was added to PolyAnalyst: *Eliminate Duplicate Texts*. This engine utilizes a robust algorithm for comparing and eliminating duplicate records. The de-duplicating engine allows compared records to be different from each other by no more than certain percent of terms, as defined by the user. This PolyAnalyst engine discovered that about 20% of provided records were duplicates of other records. This result is consistent with the WinBASIS "Standard Event" feature.

Another feature of the STEADES data was that text narratives sometimes contained standard repeating fragments, which in many cases were not informative for the purpose of the analysis. To discover, count and eliminate, if necessary, such fragments, one more text analysis engine was added to PolyAnalyst: *Text De-repeater*.



Fragment	Count	Timestamp
levels. The TCAS system worked as designed but no risk of collision existed. Standard separation was maintained throughout. Engineering review all incidents in order to develop appropriate solutions.	267	14:34:14
The crew adhered to the warning. The minimum horizontal separation was 2 nm.	31	14:34:15
copy of the report has been sent to Air Traffic Services. They have filed the report with the local air traffic authority. The investigation of ATC incidents can take several months. BASIS will be updated when a response is received.	155	14:34:17
was at FI110The TCAS system worked as designed but no risk of collision existed. The TCAS is unable to determine if both aircraft are about to level off. All incidents are reviewed to determine if changes to operating policy and/or ATC management may be appropriate.	67	14:34:20
Warning considered to be false, due to a defect with the TCAS. Engineering review all incidents in order to develop appropriate solutions. (	21	14:34:29
Total:	**541**	

Figure 7. Text De-repeater finds frequently occurring almost identical fragments of text. The count value shows how many times the corresponding fragment was encountered in data.

To verify the results obtained by this engine, the user can export all records containing almost identical text fragments to a CSV or HTML file with similar fragments highlighted.

5.1.3. Attribute values mapping

To save storage space, databases frequently have categorical values represented by numbers, with mapping rules stored in associated look-up tables. When performing data analysis, these numbers should be mapped back to categories in order to facilitate simple interpretation of the results. PolyAnalyst allows the user to define such inverse mapping rules and use them consistently throughout the project.

The same mechanism can be employed when the user wants to lump together several different values of an attribute representing similar concepts. For example, one might want to consider Aircraft Type values of B747, Boeing-737, B-757 and B767 collectively as Boeing aircraft.

5.2. Air Safety Report Categorization

5.2.1. Automated assignment of descriptors

To facilitate easier retrieval of safety reports at the time of the analysis, safety reports are currently manually categorized to a predefined taxonomy by airline analysts reading text narratives provided in the safety reports. At the moment, IATA is in the process of migrating from an old keyword-based taxonomy to a new taxonomy based on "descriptors".

One goal of the current proof-of-concept project was to demonstrate the capability of automated categorization based on utilizing patterns of terms defining individual taxonomy nodes. This should help automating the classification of safety reports to different taxonomy nodes in the descriptor classification system.

An analyst starts by defining for each node patterns of terms determining which records will be categorized to this node. This is similar to defining and storing a search query for each node in the taxonomy. In some cases, patterns defining individual nodes can be very simple, containing just a single word. In other cases, the user has to define more advanced patterns expressing various logical, lexical, affinity and sequential relations between terms.

Patterns of terms are defined by means of a simple but powerful Pattern Definition Language (PDL). This language allows the user to utilize either exact terms or their morphological variations (through Stemming mechanism) or particular instances (through built-in or user-added Semantic Dictionaries). PDL has a number of logical (AND, OR, XOR and NOT), lexical (STEM and THESAURUS) and "geometric" (FOLLOW, SENTENCE, etc.) pattern definition operators helping the user to express any complex pattern.

Upon completing the initial taxonomy node definition, the analyst applies the taxonomy to data and evaluates the results of categorizing representative subsets of data with the initial taxonomy, iteratively perfects the initial node definitions several times, and then relies on the developed model to automatically categorize the bulk of safety reports. Once developed, a defined taxonomy typically needs little additional efforts to maintain it. In the case of the descriptor system, it can be new industry developments that necessitate a revision of taxonomy node definitions.

For the proof-of-concept project, Megaputer adopted the existing STEADES descriptor classification system and generated a taxonomy for automated narrative categorization for a fragment of the descriptor system. Megaputer analysts defined nodes in a two-level Event Type-Descriptor taxonomy with tentative patterns of terms based on the provided definitions of categories and common sense. A fragment of the resulting defined taxonomy accounting for various TCAS-related events is shown in Figure 8.

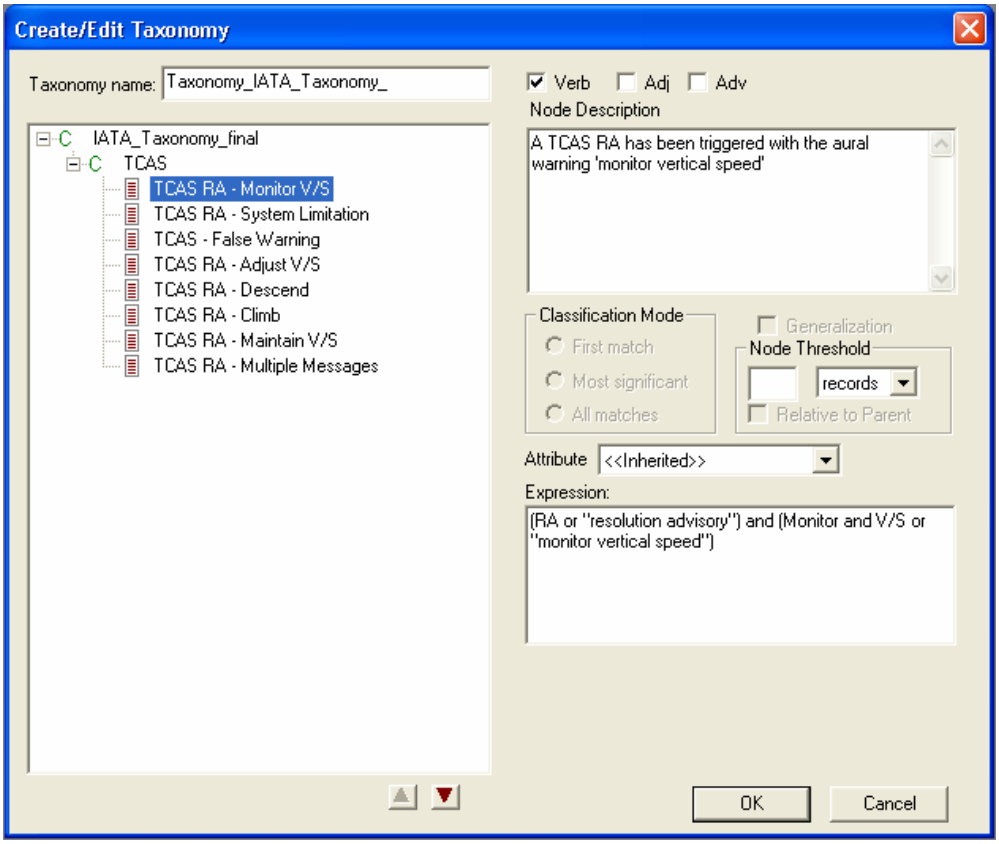


Figure 8. PolyAnalyst Taxonomy Editor with imported IATA-British Airways taxonomy of event types and descriptors for TCAS: each taxonomy node in the left pane is defined through the corresponding pattern in PDL displayed in the "Expression" text box on the right.

For example, the "TCAS RA – Monitor V/S" node is defined by the following pattern: a report is categorized to this node if it contains a combination of one of the terms *RA* or *resolution advisory* and one of the terms *Monitor* and *V/S* or *monitor vertical speed* (see Figure 8). PolyAnalyst can import/export taxonomies in either CSV or XML format.

Applying the defined taxonomy to the selected subset of records from IATA safety reports results in categorizing the majority of these records to one or more categories in the defined taxonomy. For example, the "TCAS RA – Monitor V/S" node harvested 60 records, as can be seen in Figure 9.

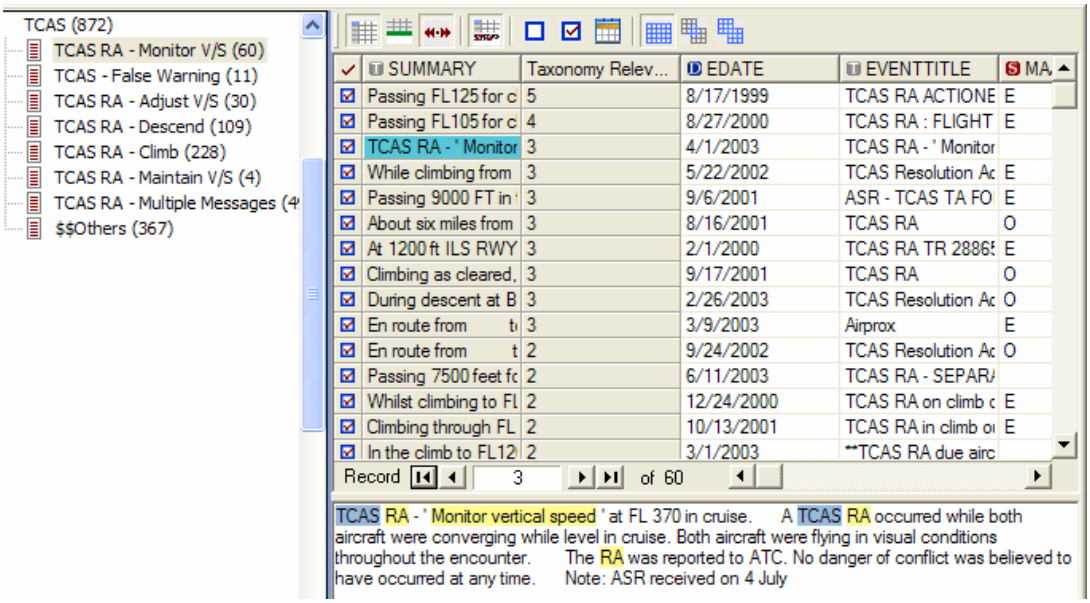


Figure 9. The results of taxonomy-based categorization performed by PolyAnalyst: 60 records matched the definition of *TCAS RA => monitor vertical speed* node.

PolyAnalyst allows the user to browse through all relevant records with the patterns that triggered categorization highlighted and to export the results in CSV or HTML formats, as shown below.

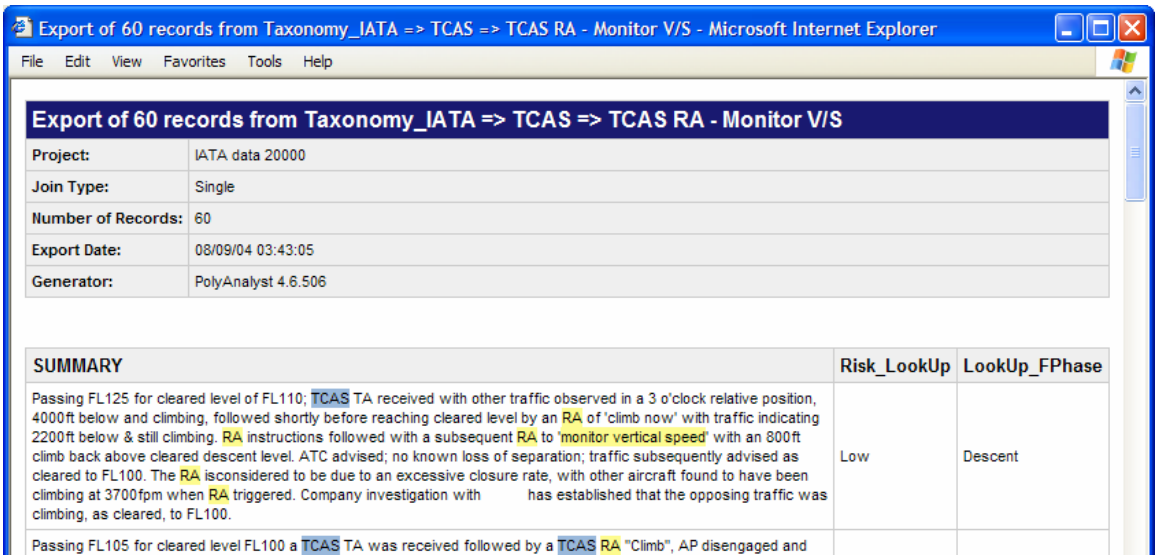


Figure 10. Exported to HTML results of the drill-down to original records matching the pattern defining the *TCAS RA => monitor vertical speed* taxonomy node.

5.2.1. Detection of possible errors in human coding

Upon developing a taxonomy, defining and polishing patterns of terms representing individual nodes, one can obtain a system for accurate automated categorization of safety reports. Comparing the results of automated categorization to the results of manual categorization, the system helps highlighting possible errors in the latter.

5.3. Problem areas ranking

5.3.1. Determination of key problem areas

PolyAnalyst Taxonomy-based categorization allows the user to apply separate taxonomy nodes to different data attributes and to monitor relative importance of events and execute certain pre-defined actions when a user-defined threshold for the number of records categorized a selected node is surpassed.

For example, parameters supplied in the taxonomy creation window shown in Figure 11 instruct PolyAnalyst to highlight in bold the node counting incident summaries where *RA* is mentioned prior to *ATC* when this node contains 50 or more reports related to *Aircraft 2*.

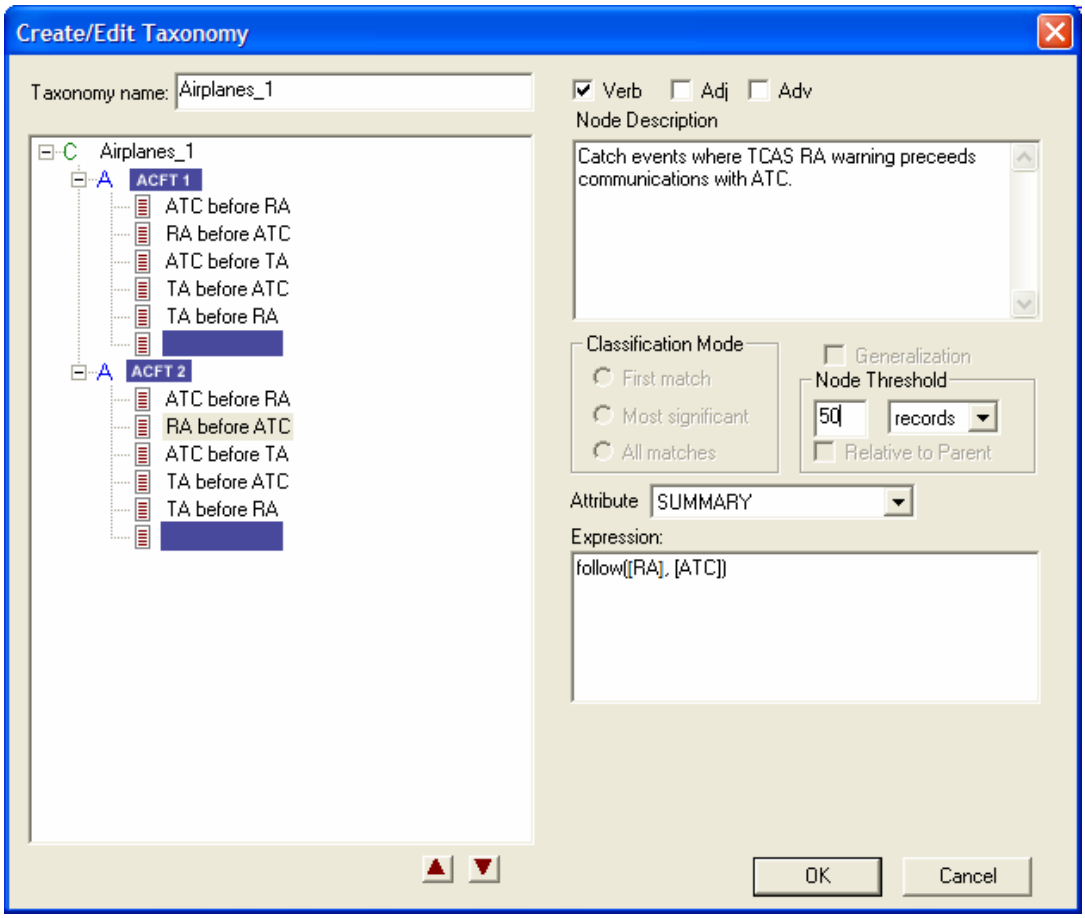


Figure 11. Setting a taxonomy for comparing the TCAS system performance across different aircraft models. The "RA before TA" node will be highlighted in bold in the results of categorization if it captures 50 or more reports related to *Aircraft 2*.

5.3.2. Learning rules for risk degree assignment

During the proof-of-concept project, we investigated whether PolyAnalyst can automatically determine a reliable set of text categorization rules based only on the analysis of a large collection of manually categorized records. The goal was to predict the risk degree of events based on text narratives. This task was solved by expanding the self-learning categorization capabilities of PolyAnalyst Decision Trees to processing text notes. At each branching point, the classification rule splits the collection of reports into smaller subsets for each of which a better classification decision can be made and the number of incorrect classifications is minimized.

Automated learning of categorization rules based on pre-categorized examples may help save manual labor and further increase the usefulness of available historical safety data.

5.3.3. Taxonomy building

Taxonomy-based Categorization serves the purpose in those cases when one has a detailed list of issues of concern. However, taxonomies can be "outgrown" over time and need refinement. In such cases, analysts can be helped by the PolyAnalyst Taxonomy Builder. It suggests tentative schemes for categorizing investigated text records. A sample taxonomy built from the analysis of TCAS related data is presented in Figure 12.

A data-driven machine-generated taxonomy located in the left pane can be saved as a PolyAnalyst taxonomy and further edited by an analyst. The results viewer in the right pane assists in navigating records supporting individual nodes of the created taxonomy. Terms highlighted in different colors in the results viewer correspond to patterns defining related nodes at different levels in the taxonomy: for example, *TCAS => antenna => history*.

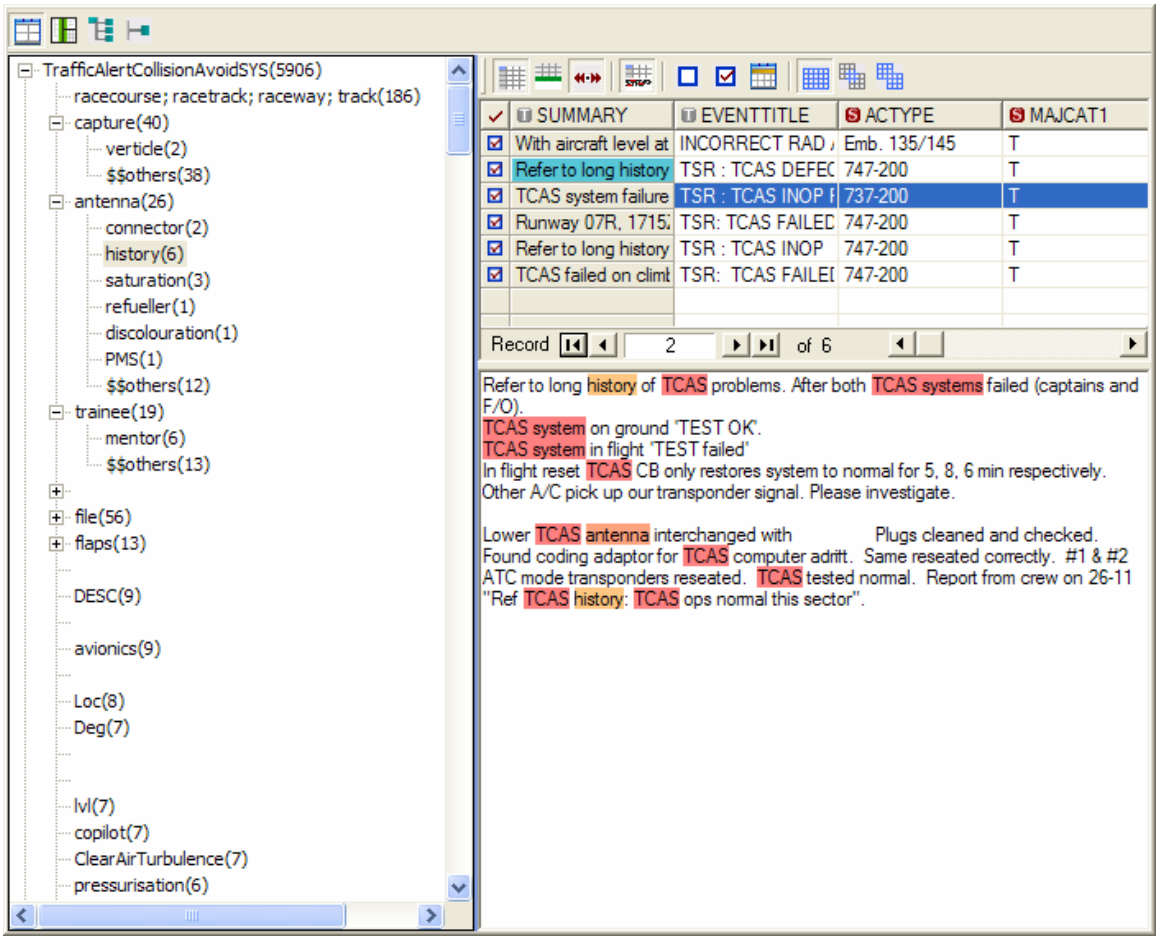


Figure 12. Tentative taxonomy created by PolyAnalyst based on unsupervised linguistic and statistical analysis of report narratives.

5.4. Discovery of trends and patterns

5.4.1. Strongest correlations between risk degrees and narrative terms

Link Chart engine offers calculating and visualizing pair-wise correlations between individual values of different attributes. Figure 13 below represents the strongest correlations between different risk degree values and terms extracted through text analysis from event summaries. The heavier the line on the graph the stronger the correlation between the corresponding objects on the graph.

For example, *High* risk degree is strongly correlated with terms (in decreasing strength order) *intruder*, *safety comment*, *same level*, *opposite traffic*, *safety* and *air traffic controller*. On the other hand, *Minimal* risk degree is strongly correlated with terms *flight*, *flight crew* and *notify*.

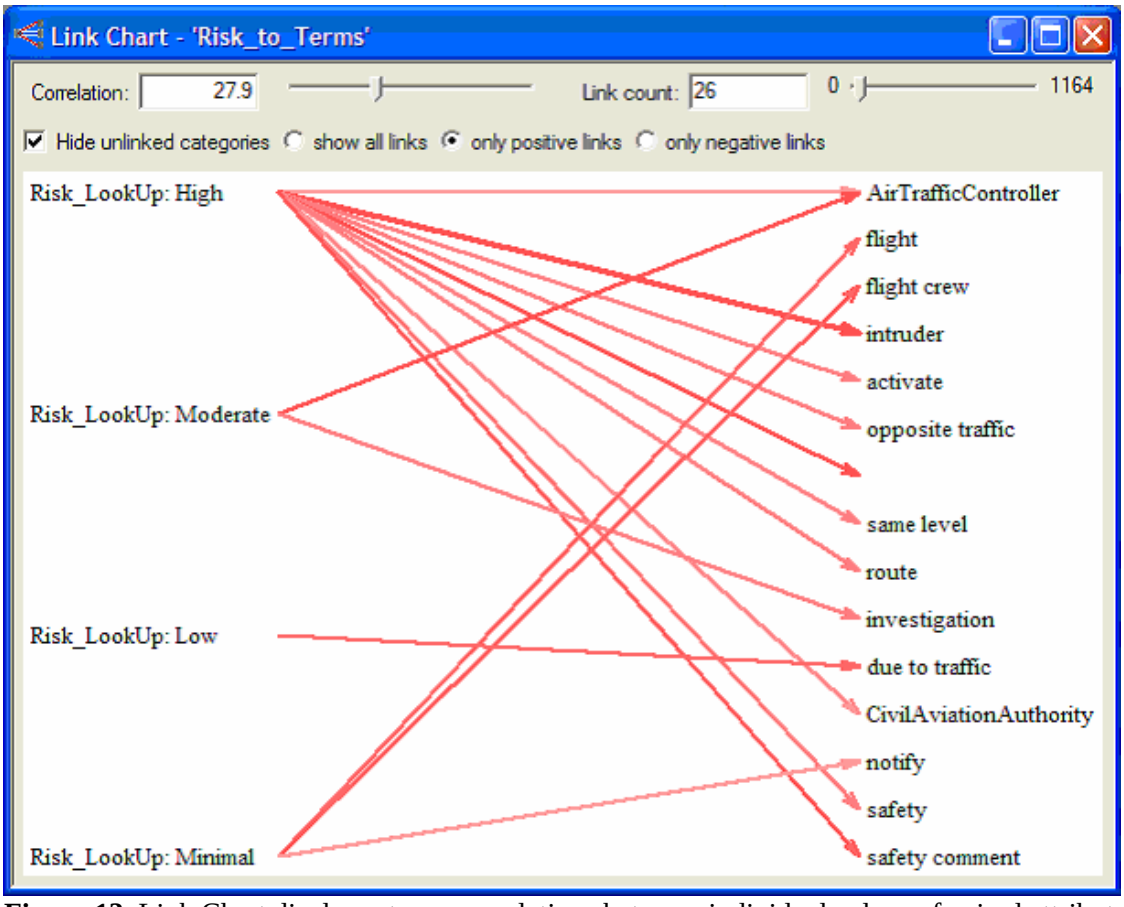


Figure 13. Link Chart displays strong correlations between individual values of paired attributes.

5.4.2. Stable patterns of terms in text descriptions

Stable multi-dimensional patterns and correlations of terms can be discovered by the PolyAnalyst Link Terms engine. Patterns of terms occurring in a combination with a particular event can represent a characteristic signature of this event and can help trace nontrivial cause and consequence relationships for some events.

The Link Terms diagram in Figure 14 displays several clusters of strongly correlated terms. Individual clusters are shown in different colors to facilitate simple visual identification of these clusters. For example, the cluster shown in yellow indicates that *minimum separation distance* to *intruder* is measured in *nautical miles*. The red cluster containing the *Ground Proximity Warning System* (GPWS) node suggests that, as expected, GPWS is strongly correlated with *radio altimeter*, *terrain*, and *warning triggered by traffic*, but this *warning* is often considered to be a *nuisance* when associated with a TCAS alert.

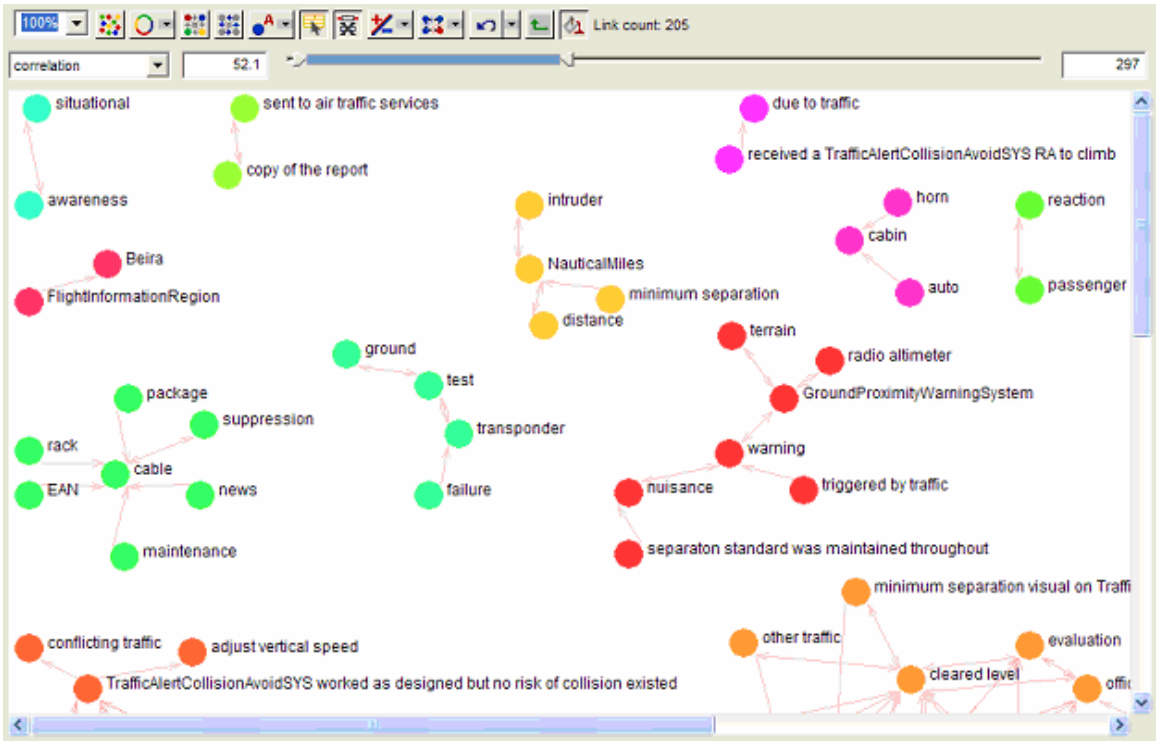


Figure 14. Link Terms Diagram visually displays patterns of terms encountered in safety event textual descriptions.

5.4.3. Correlations between structured attributes and narratives

PolyAnalyst's Link Analysis engine allows the user to visualize multidimensional correlations between values of structured attributes included in the analysis and patterns of terms in the corresponding event description summaries.

Figure 15 displays the strongest correlations between *Risk Degree*, *Flight Phase* and summary terms extracted from TCAS related events. One can observe that the flight phases *Climb* and *Descent* have strong correlations with the phrase *monitor vertical speed*, while *Climb* is correlated in addition with *warning* and *rate of climb*. At the same time, *Moderate* risk degree is strongly correlated with the terms *TA*, *heading*, *air traffic controller*, *opposite traffic*, *flight* and *action*.

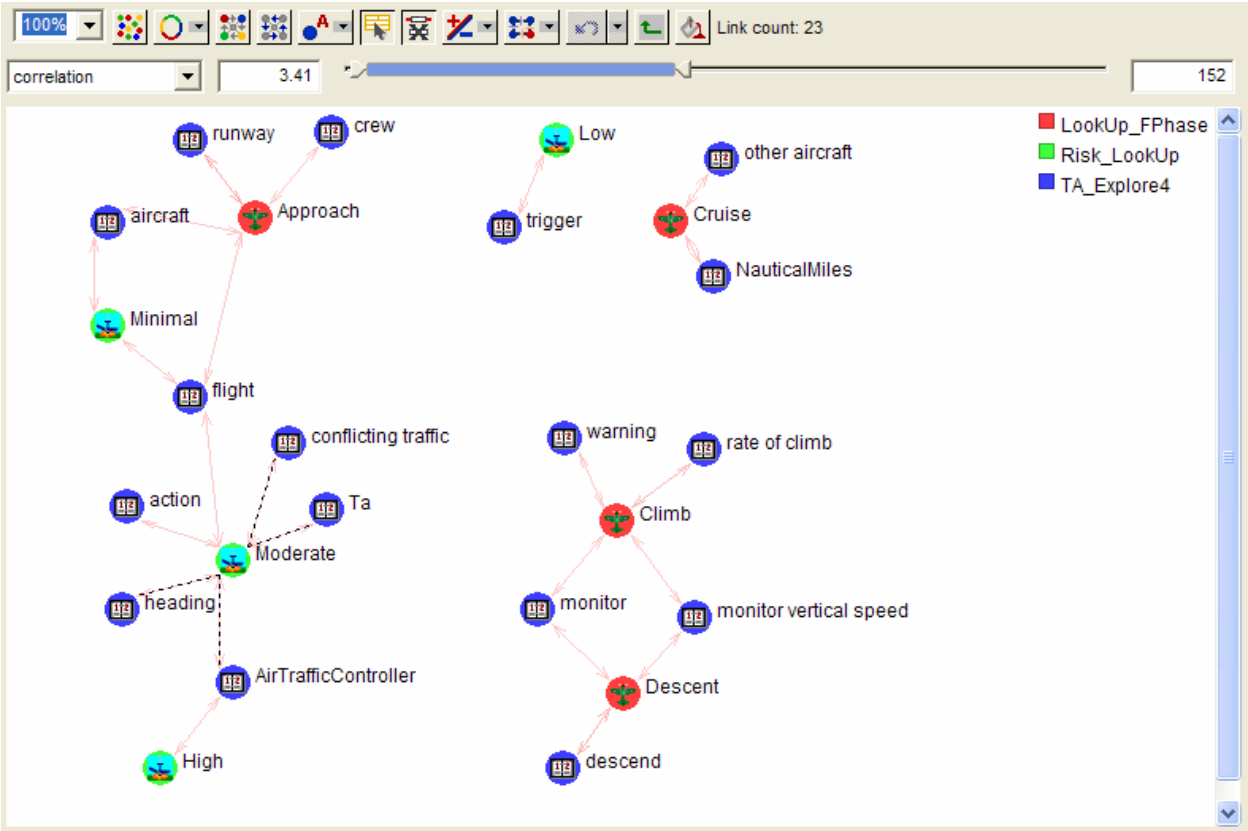


Figure 15. Link Analysis Diagram visualizes multi-dimensional correlations between values of structured attributes and patterns of terms extracted from textual narratives.

All PolyAnalyst Link Analysis engines, including Link Terms and Link Chart, support drill-down capabilities. By clicking on the link of interest, the user displays all reports supporting the chosen link with the corresponding terms highlighted in the narratives. The user can select how to perform the drill-down: selecting more than one link on a graph, the user can then choose to see an intersection or union of all records supporting the considered collection of links.

For example, a user may select a number of links connecting *Moderate* risk degree value with strongly correlated narrative terms: *conflicting traffic*, *TA*, *heading* and *air traffic controller* and select the intersection drill-down mode. Figure 16 represents one of three records that support all selected links at once.

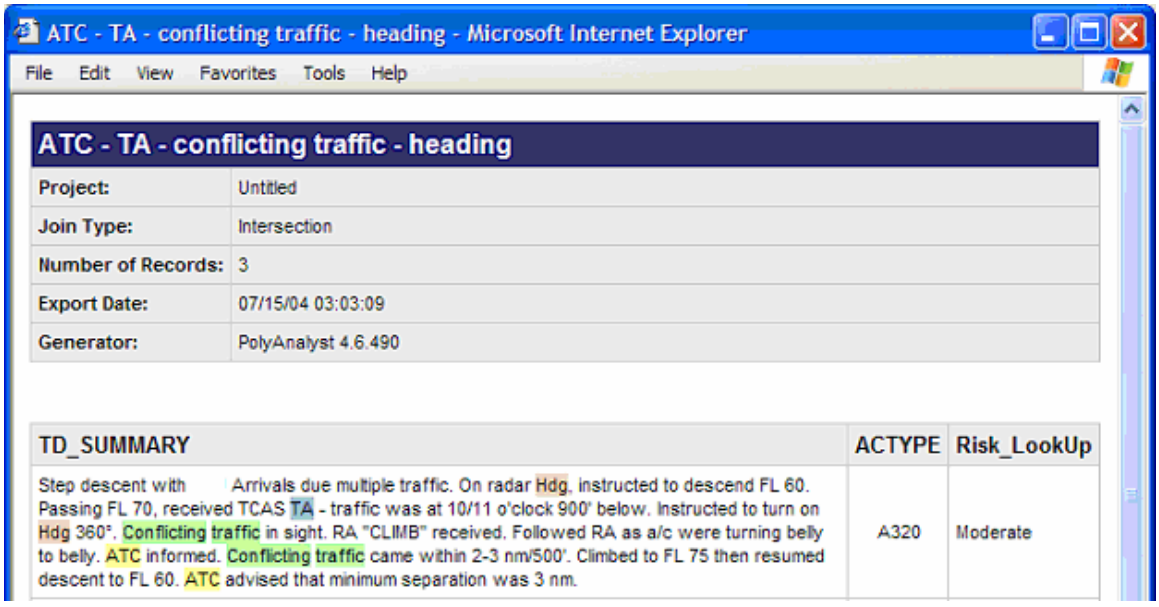


Figure 16. Visual drill-down capability allows the user to select collections of original records supporting the selected link patterns.

5.4.4. Interactive multi-dimensional narrative investigation – Text OLAP

Similarly to the analysis of structured data, defining dimensions of interest and developing a multi-dimensional cube holding data presents new opportunities for the analysis of free form text. Manipulating the defined cube, the user can quickly slice and dice data across different dimensions, rotate data, and perform drill-down to original records with terms of interest highlighted.

To build an interactive Text OLAP report, a user of PolyAnalyst defines a matrix of dimensions utilized for reporting on either textual or structured data fields. A dimension matrix used in this project is displayed in Figure 17.

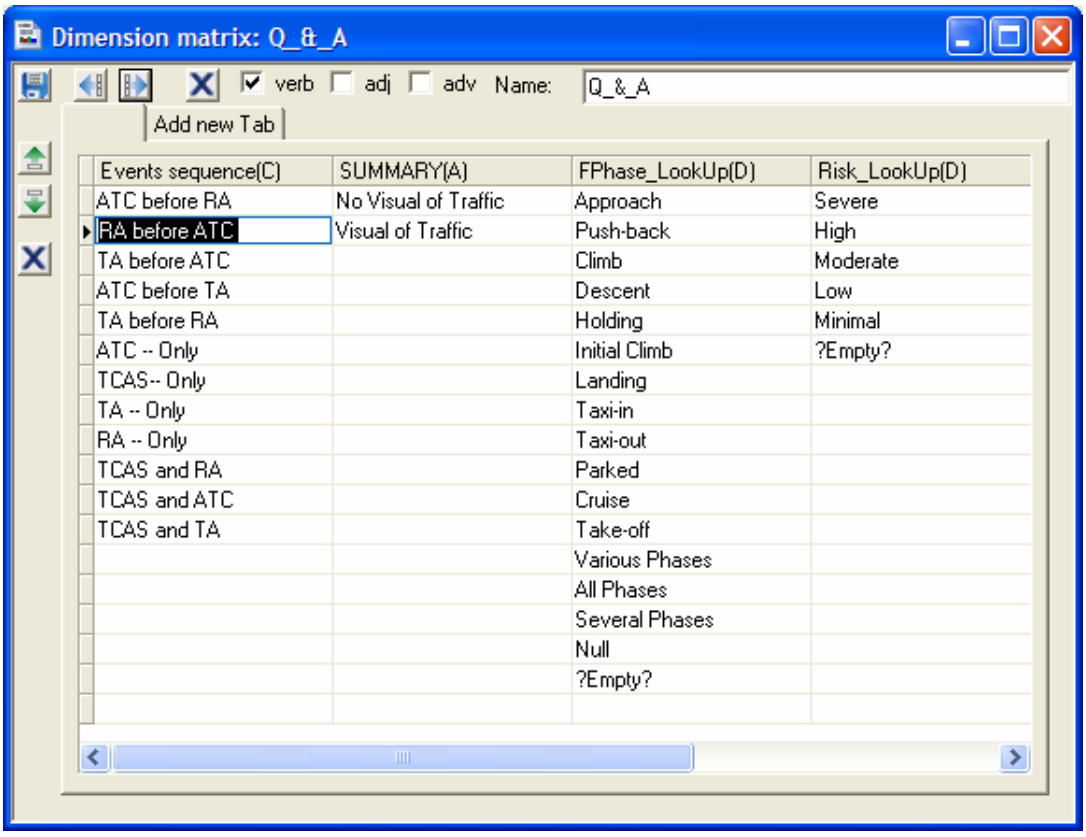


Figure 17. The dimension matrix defines a set of dimensions used for interactive generation of analytical reports within PolyAnalyst Text OLAP.

Values of individual cells in dimension matrices can be defined with the help of the same Pattern Definition Language (PDL) that was described when defining patterns for separate taxonomy nodes above. For example, the cell capturing RA events occurring prior to ATC can be defined through the following simple PDL expression: *follow*([RA], [ATC]).

Upon applying the developed dimension matrix to a collection of TCAS reports, one obtains the Text OLAP report depicted in Figure 18. The first column reports on the time

sequence of events: for example, whether ATC warning was received prior to Traffic Alert (TA) or Resolution Advisory (RA) warnings coming from the TCAS system, or ATC was informed about the situation afterwards. Figure 18 illustrates that in the majority of cases (301 cases), RA warning appears prior to ATC, while in 264 cases ATC is mentioned prior to RA. Similarly, in 92 cases, TA precedes RA.

Drilling down on the latter cases, one sees the distribution of records *Visual/No Visual* contact with the approaching traffic, as well as the distributions across other dimensions. In 31 of the cases where TA preceded RA, there was a visual contact with the approaching traffic, while in 13 cases there was no visual contact (Figure 19). Continuing the drill-down to events with visual contact, one sees the corresponding distributions of flight phases and risk degrees. Drilling down further to *Descent* flight phase, and *Moderate* Risk Degree, one can view a list of records supporting the entire selected pattern (one report in our case).

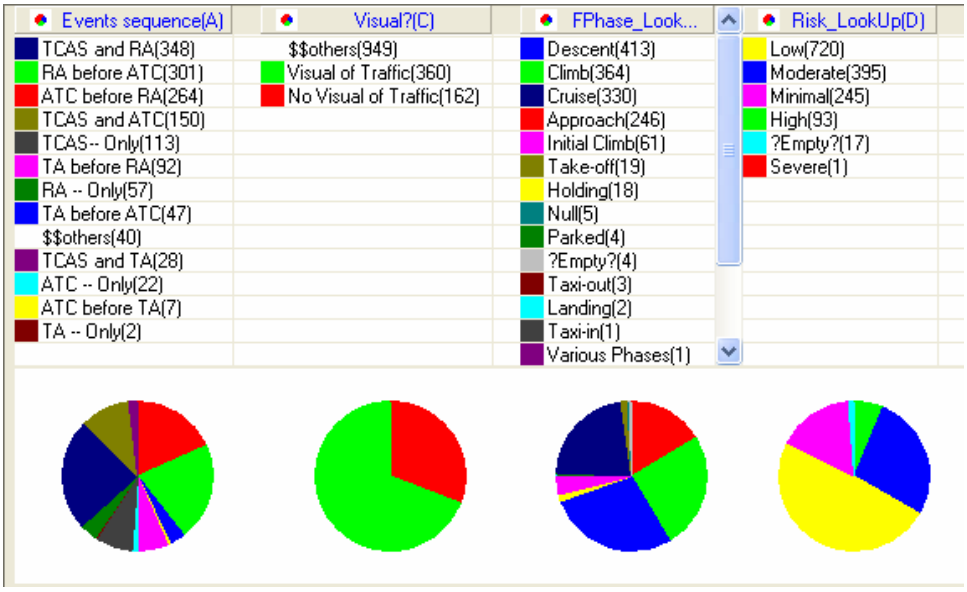


Figure 18. PolyAnalyst Text OLAP report reveals the distributions of safety reports across different dimensions defined on structured data as well as natural language descriptions of events.

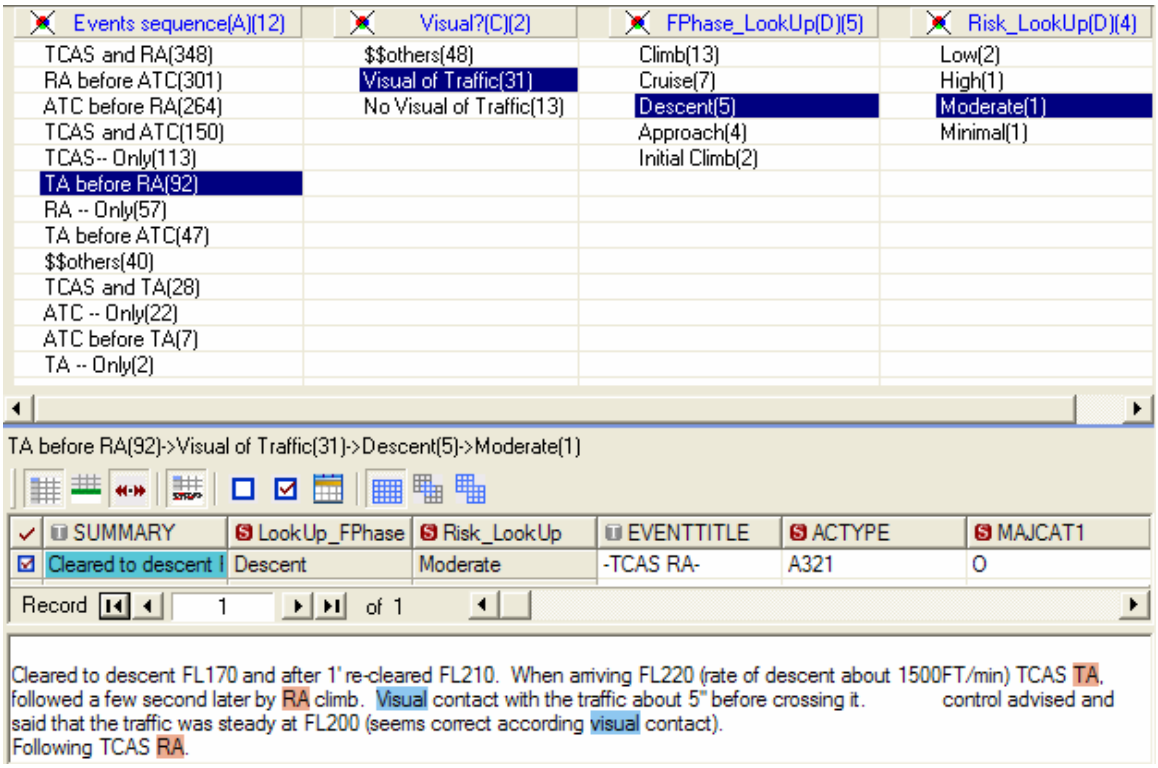


Figure 19. Drilling down in the report, one focuses on an event where a TA was received prior to a RA, conflicting traffic was seen visually and the problem occurred during the descent and was assigned moderate risk degree.

The drill-down path bar shows the analyst the current path of navigation with the corresponding populations of intermediate nodes. The text viewer on the bottom of the window allows the user to browse through individual text records with the corresponding patterns of terms highlighted. An interactive grid in the middle of the window allows the user to export a CSV or HTML report capturing his/her findings.

Note that only a single path for diving into the results was followed out of a large number of possible paths providing answers to other important questions analysts might have. By empowering analysts with an ability to readily answer a variety of questions through simple selection of a drill-down path, PolyAnalyst can significantly increase the productivity of safety analysts.

5.4.5. Evolution of discovered patterns

As traffic conditions, equipment sophistication and other parameters evolve over time, it is important to track time evolution of particular situations. For example, one could track how the distribution of false alerts caused by TCAS evolves.

PolyAnalyst provides interactive time filter capabilities. The user can specify time intervals of interest and observe the evolution of the discovered patterns. The user utilizes

a slider to define and shift the time window and observe the corresponding evolving patterns.

Two snapshots representing subsets of the main event correlation pattern seen on Figure 15 are presented on Figures 20 and 21. The first figure displays patterns representing safety events that occurred in April-May of 1998. The second figure displays patterns representing events from February-March of 1999.

Figure 20. PolyAnalyst time filter helps monitor evolution of patterns. Displayed are the main safety event patterns recorded during April and May of 1998.

Figure 21. The main safety event patterns recorded during February and March of 1999. Compare to the patterns displayed in Figure 20.

5.4.6. Finding similar reports

In order to better classify a new safety event, Flight Safety Officers frequently need to locate similar events in past reports. Due to the volume of safety data, analysts cannot always locate similar reports in a timely manner. PolyAnalyst offers an efficient mechanism for carrying out this task automatically: the Find Similar engine can retrieve safety reports similar to a specified report. The system highlights those patterns of terms that show why the retrieved reports are similar to the investigated report and calculates

the measure of similarity of retrieved reports. The similarity of reports is assessed with the help of a Case Based Reasoning algorithm.

Figure 23 presents a historical safety event description summary that the system identified as the most similar to a selected safety report displayed in Figure 22. One can observe that PolyAnalyst considered these summaries to be similar because they contain the same terms *wake turbulence*, *experienced*, and *glideslope*.

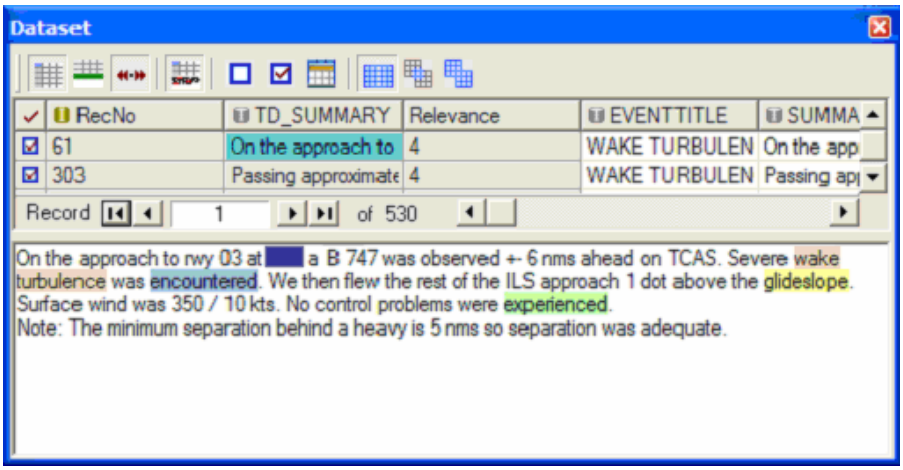


Figure 22. An arbitrary safety report selected by the user. PolyAnalyst Find Similar engine identified a historical report most similar to the selected record, as illustrated below.

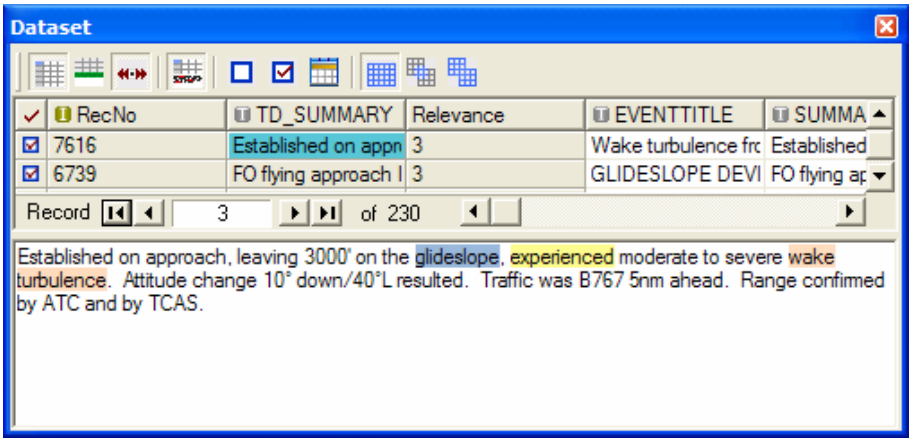


Figure 23. Report identified by the system as most similar to the selected report.

5.5. Generating safety analysis reports

Upon discovering patterns and trends of interest in the considered data, analysts may need to share their findings with colleagues. For IATA STEADES, the results of analysis need to be delivered to safety officers in all member organizations.

PolyAnalyst offers a number of ways to generate business reports capturing important findings throughout the project:

1) Graphical objects can be saved in one of several graphical formats (BMP, JPG, PNG or WMF).
2) Results of the analysis support drill-down capability.
3) The results of drill-downs can be saved in HTML format or in new Excel data sets.

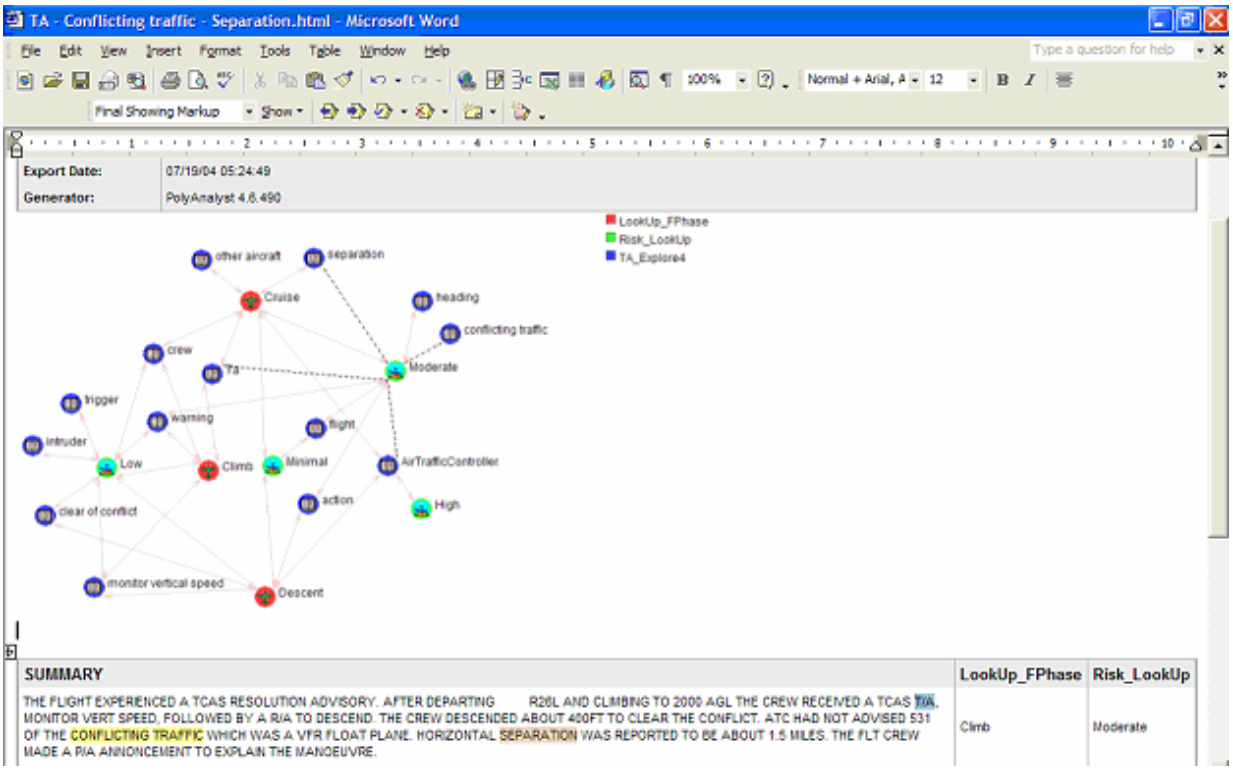


Figure 24. PolyAnalyst allows the user to create HTML reports capturing the results of the analysis.

A combination of these mechanisms allows the user to build business reports with graphical elements, as illustrated in Figure 24. A more comprehensive business reporting system allowing the user to edit and annotate all objects sent to pre-set business report templates will be available in the next release of PolyAnalyst in the third quarter of 2005.