MP 03W0000238

MITRE PRODUCT

Application of Aviation Safety Data Mining Workbench at American Airlines

Proof-of-Concept Demonstration of Data and Text Mining

November 2003

Zohreh Nazeri

MITRE
Center for Advanced Aviation System Development
McLean, Virginia

CAASD

MP 03W0000238

MITRE PRODUCT

Application of Aviation Safety Data Mining Workbench at American Airlines

Proof-of-Concept Demonstration of Data and Text Mining

November 2003

Zohreh Nazeri

Sponsor: Federal Aviation Administration **Contract No.:** DTFA01-01-C-00001
Dept. No.: W903 **Project No.:** 02044314-AA

Approved for public release; distribution unlimited.

MITRE
Center for Advanced Aviation System Development
McLean, Virginia

Abstract

This paper describes the application of The MITRE Corporation's Aviation Safety Data Mining Workbench to American Airline's Aviation Safety Action Program (ASAP) data, to demonstrate the usefulness of data and text mining tools in the analysis of aviation safety data and assess the ability of these tools to enhance internal airline safety analysis.

This work was done in support of the efforts of the Analytical Methods and Tools Working Group B of the Global Aviation Information Network (GAIN). In compliance with the Working Group B's objectives, the objective of this project was to apply MITRE's Aviation Safety Data Mining Workbench to ASAP data and demonstrate the ease of installation and use of the tool, the usefulness and effectiveness of the tool and its capabilities, the quality of the results generated by the tool, and the applicability of the results to address areas of interest to the flight safety community.

Prior to beginning the project, MITRE and American Airlines signed a License Agreement and a Non-Disclosure Agreement (NDA) to not disclose proprietary and company-confidential information.

KEYWORDS: Analytical Tools, Aviation Safety, Data Mining, Flight Safety Reports, Knowledge Discovery

Acknowledgments

The MITRE Corporation would like to thank American Airlines, the Analytical Methods and Tools Working Group B of the Global Aviation Information Network (GAIN), and the Federal Aviation Administration (FAA), Office of System Safety, who made this project possible.

The author would like to especially thank the following people for their support and helpful comments throughout the project including preparation of this report:

Captain Larry Kelly, American Airlines

Carl Halford, American Airlines

Hugh Schoelzel, TWA Limited Liability Corporation, and Member GAIN Steering Committee

Peggy Sterling, Corporate Safety, American Airlines

Carolyn Edwards, Federal Aviation Administration, and GAIN Working Group B

Geoff Gosling, Aviation System Planning Consultant, and Co-Chair, GAIN Working Group B

Christina Hunt, Phaneuf Associates Incorporated, and GAIN Working Group B

Andy Muir, Federal Aviation Administration, and GAIN Working Group B

Grant Schneemann, Abacus Technology Corporation, and GAIN Working Group B

Eric Bloedorn, Washington C3 Center, The MITRE Corporation

Gerard Eldering, Technology Transfer Office, The MITRE Corporation

Wallace Feerrar, Center for Advanced Aviation System Development, The MITRE Corporation

Earl Harris, Washington C3 Center, The MITRE Corporation

John Pyburn, Center for Advanced Aviation System Development, The MITRE Corporation

Max Rosen, Contracts, The MITRE Corporation

This project was sponsored by the FAA, Office of System Safety.

Table of Contents

List of Figures

Section 1

Introduction

1.1 Objective of the Proof-of-Concept Demonstration

This paper describes the application of The MITRE Corporation's Aviation Safety Data Mining Workbench to American Airline's Aviation Safety Action Program (ASAP) data. The purpose of this proof-of-concept project was to demonstrate the usefulness of data and text mining tools in the analysis of aviation safety data and assess the ability of these tools to enhance internal airline safety analysis. This work was done in support of the efforts of the Analytical Methods and Tools Working Group B of the Global Aviation Information Network (GAIN).

The objective of GAIN Working Group B is to identify and increase awareness of existing analytical methods and tools, as well as promote the development and validation of these methods and tools [1]. To assess the usefulness and usability of existing tools, Working Group B seeks partnerships with airlines and tool vendors/developers to demonstrate the use of analytical tools, including data mining and text mining tools as described in this report, for the analysis of safety data. It is also GAIN's desire to share the knowledge of these demonstrations with others in the aviation community.

In compliance with the GAIN Working Group B's objectives, the objective of this project was to apply MITRE's Aviation Safety Data Mining Workbench to ASAP data and demonstrate the ease of installation and use of the tool, the usefulness and effectiveness of the tool and its capabilities, the quality of the results generated by the tool, and the applicability of the results to address areas of interest to the flight safety community.

Prior to beginning the project, MITRE and American Airlines signed a License Agreement and a Non-Disclosure Agreement (NDA) to not disclose proprietary and company-confidential information. Due to the sensitivity of the data, the majority of the work that involved data loading and data analysis was done on-site at American Airlines' facility.

The proof-of-concept project took about six weeks, and included customization of the Workbench for American Airlines' ASAP data, preparation of the data, analysis of the data, and writing of reports and other documents.

MITRE also provided American Airlines with a complimentary copy of the Workbench, a Users' Guide document to explain features of the Workbench, and two short training sessions on how to use the tool.

1.2 Aviation Safety Data Analysis and the Use of Data/Text Mining Tools

The process of aviation safety data analysis typically starts with the submission of incident reports by the cockpit crewmembers, cabin crewmembers, maintenance personnel or other sources. The safety office at the airline receives the reports and enters them into a database. The reports are then compiled and analyzed at the safety office; they might also be routed to other departments in the airline for further analysis and investigation. The safety office may then send summary reports and recommendations (for prevention of similar incidents in the future) to management and other departments.

Typically the process of analyzing the safety data is not very automated. Often there are query tools available to the safety officer to search safety report databases for specific issues but at most airlines there are few or no tools to help them undertake a more comprehensive analysis that might reveal underlying patterns in the data. Thus the analysis process often relies heavily on the safety officer's memory and past experiences. While data mining tools do not replace the human analyst, they help the analyst with the analytical task. The tools have been demonstrated to save time by presenting results that otherwise could only be obtained through repeated queries. Furthermore, data mining tools can make discoveries that cannot be accomplished by queries alone. For example, the analysis does not have to start with a hypothesis or previous knowledge of an incident. While safety officers have a good knowledge of their domain and know how to examine the data, data mining can still discover patterns or anomalies in the data that have not been thought of before.

1.3 Overview of Aviation Safety Data Mining Workbench

The Aviation Safety Data Mining Workbench is the result of an internal research project at MITRE. Since its completion, the Workbench has been applied to a few airlines' safety data and has proved to be very useful [2]. The use of Workbench has been useful to airlines of different sizes. Larger airlines have bigger databases and can benefit from the automated discoveries in their data made by the Workbench. Smaller airlines do not have a large amount of data but they do not have a large number of analysts to process the data either and therefore could benefit from using the Workbench.

The Workbench runs on the PC and takes only a few minutes to install. Working with the Workbench is very easy and its training session takes only a few hours. Currently the Workbench does not directly read the data from a database. The input data needs to be saved in a flat text file. The Workbench will then read the text file and load the data. The Workbench includes three tool modules: FindSimilar, FindAssociations, and FindDistributions as described below.

FindSimilar. This tool searches both the structured fields and free-text narratives in the data and finds reports that are similar to a report selected by the user (as the target). For example, consider a case where the user is focusing on a report that involves an altitude deviation due to distraction of the cockpit crew. The user could enter the report's ID as the

target and run the FindSimilar tool to see what similar cases exist in the data and what has been the cause of distraction in each case.

FindAssociations. This tool searches the structured fields in the reports. FindAssociations would be used when users want to discover outstanding associations in the data. The tool could be run on the entire data or on a selected subset. In either case, the user does not need to specify which associations in the data to look for. The tool examines all possible associations and returns the ones that are above the specified thresholds.

FindDistributions. This tool searches the fields in the reports, and identifies unusual distributions of incidents. To run this tool, users need to select the field they want to focus on (Focused Attribute). For example, to search for anomalies in distribution of MONTH, select MONTH as the Focused Attribute. The tool then calculates distribution of subsets of incidents over the selected field (MONTH in this case). Those subsets that differ most from the overall distribution are identified as unexpected.

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

The input data for this project were American Airlines air safety reports filed by pilots under the Aviation Safety Action Program (ASAP). In recent years, the Federal Aviation Administration (FAA) in the U.S. and the air transportation industry have sought improved ways to address safety problems and identify potential safety hazards. In an effort to increase the flow of safety information, a number of airlines, in cooperation with the FAA, have established ASAP programs, following the guidance of FAA Advisory Circular-120-66B [3]. American Airlines was the first airline to establish an Aviation Safety Action Program. The airline is now planning to enhance its reporting system by collecting a broader set of data fields and adding a new web-based reporting system

ASAP data records are voluntary reports of safety issues and events that come to the attention of employees (cabin crew, cockpit crew, etc.). The objective of ASAP 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]

---

[1] The MITRE access to and analysis of safety reports was in full compliance with the objectives, spirit, intent, and confidentiality of the American Airlines' existing ASAP Memorandum of Understanding (MOU) and its Data and Information Policy. All proprietary information shall remain so, and is for the sole use of compliance with the purpose, intent, and requirements of the ASAP MOU to assist in the proactive prevention of incidents.

Section 2

Data Preparation and Workbench Customization

2.1 Data Preparation

For this project, about three years' worth of data or 12,000 reports were selected from the American Airlines ASAP database of 40,000+ reports and analyzed using the three data and text mining tools in the Workbench. The most recent data (past three years) were selected as an adequate sample size, that represented three different periods of operations, that had been reviewed and analyzed (without the Workbench) by American Airlines ASAP staff. All the selected reports were submitted by cockpit crew. Each report consisted of two text fields and eight structured fields selected from ASAP database by American Airlines. The remaining data fields were not selected for the analysis to limit the test findings to the most significant items and preserve the anonymity of the reporters.

The two selected text fields were the report synopsis and the full event description, as written by the pilot who submitted the report. The eight selected structured fields were report ID, date of the incident, date of the report, aircraft ID, aircraft type, departure, destination, and the category of incident. Three additional fields, YEAR, MONTH, and DAY were extracted from the date field and added to the data. The reports were divided into three time periods for the analysis: pre September 11, 2001, post September 11, 2001, and the most recent months since December 2002. These dates reflect periods of different flight conditions, policies and procedures and therefore it was more meaningful to analyze incidents of each period separately. For example, many of the reports in the post-September-11 period were related to security-related regulations that did not exist before then.

To prepare the ASAP data for loading into the Workbench, the data was first saved into comma-delimited text files. The data resided in two separate tables (one containing the structured fields and the other containing the narratives). Unique IDs were used to relate corresponding records in the two files. A parser program was written (in C language) to read both text files, combine the two sections into a single record, and write them to a single text file.

Another program was written to clean the data; for example, any END_OF_LINE or CARRIAGE_RETURN characters were removed from the free text field since these would cause the tools to stop reading the rest of the text in that field. For the structured fields, the program checked their values, and entered a "NULL" if the value was missing. This input file was then reformatted to meet the Workbench input requirements; for example, the DATE field was parsed into separate fields of YEAR, MONTH, and DAY.

2.2 Workbench Customization

To use the Workbench on a new database, the Workbench modules need to be customized for that database. This is a one-time effort that needs to be done before loading the new data (with new schema) into the Workbench. After the customization, new sets of data from the same database (with the same data schema) can be loaded by running the load option of the Workbench (clicking on the LOAD button).

Customization of the Workbench involves creating a set of database tables matching the new data schema, writing code for cleaning the data and preparing it for loading into the Workbench, extracting additional features (e.g., YEAR, MONTH from the DATE field), changing the interface screens that display database-specific fields (such as browse screens, and field selection screen), and adding queries for automated reports and charts.

The current version of Workbench does not directly read the data from a database. The selected data for analysis needs to be saved in a delimited (by commas or other separators) text file. After the Workbench was customized for the specific ASAP data schema for this project, the selected ASAP data, prepared and saved in a comma-separated text file, was loaded into the Workbench and the three tool modules were applied to it as described below.

Section 3

Application of Data Mining Workbench to ASAP Data

The purpose of this application of the Workbench at American Airlines was to analyze the ASAP data in order to show the usefulness of the Workbench in discovering trends and patterns that might be useful for improving aviation safety and preventing future incidents. American Airlines was also interested to run the Workbench on the data they had analyzed before (without using the Workbench) and compare the results of the Workbench with their findings. Testing the Workbench's ease of use and speed of execution were also of interest.

3.1 Data Selection

After loading the data into the Workbench, the user could select all or different subsets of the loaded data for analysis; Figure 3-1 indicates the data-selection screen. The Workbench provides a query screen to select subsets of data. The selection criteria for the query could be values of structured fields, words in text fields, a date range, or a combination of the three. For example, the user could select the reports that are pilot-deviations from January through March of 2001.

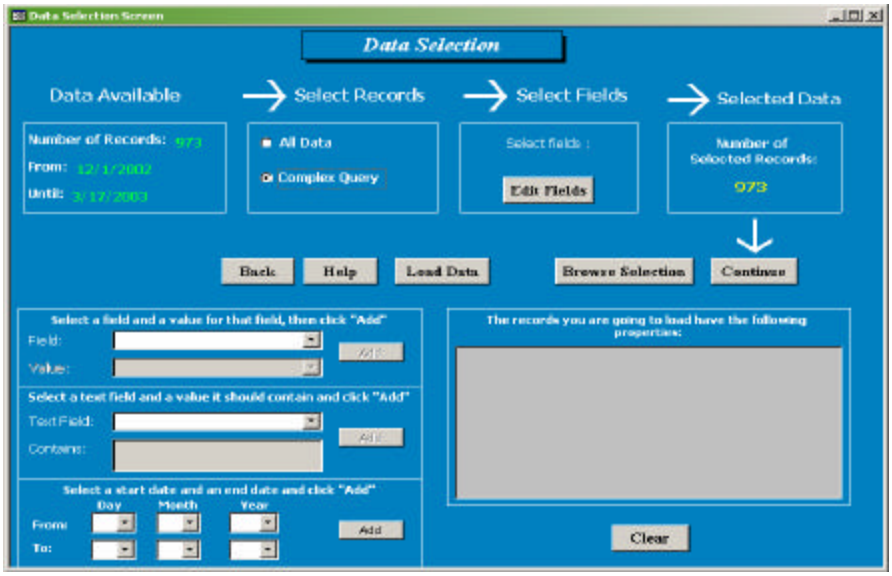

Figure 3-1. Data Selection Screen of the Workbench

Furthermore, the fields to be included for analysis by each tool could be selected by the user. For example, the user could select some structured and text fields to be used by the FindSimilar tool, and some of those fields to be used by the FindDistributions and FindAssociations tools, as shown in Figure 3-2.

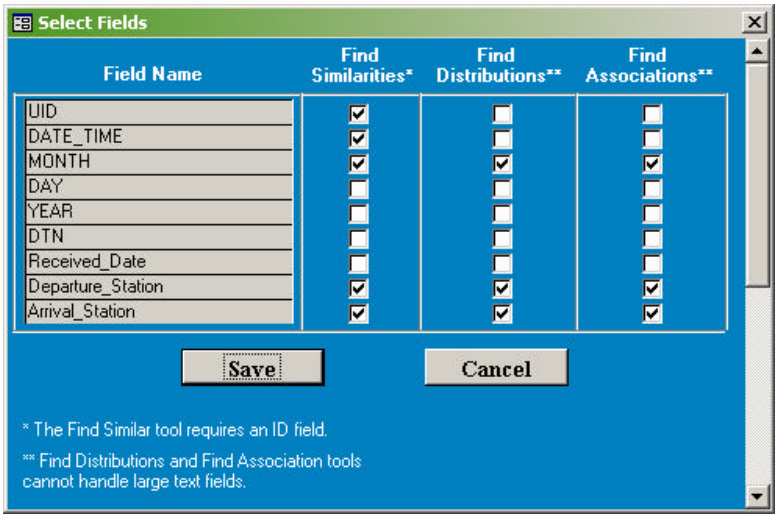

Figure 3-2. Field Selection Screen of the Workbench

3.2 Application of FindSimilar Tool Module

The FindSimilar tool searches both the structured fields and free-text narratives in the data and finds reports that are similar to a report selected by the user (as the target). For example, consider a case where the user is focusing on a report that involves an altitude deviation due to distraction of the cockpit crew. The user could enter the report's ID as the target and run the FindSimilar tool to see what similar cases exist in the data and what has been the cause of distraction in each case.

When searching the reports, the tool uses "stemming" to determine whether the words should be considered similar. Stemming is the process of removing affixes from words and leaving the stem. A thesaurus containing lists of inflectional endings such as "s," "ed," and "ing" is used for stemming. For example, the words OPEN, OPENS, OPENED, and OPENING are all stemmed to "OPEN-" and therefore are considered similar by the tool.

The tool also uses a list of "stop-words," the words that are ignored when searching for similar words in the reports. Stop-words are a set of English words such as AND, IS, A, THAT, THIS which do not add any specific importance to the text and therefore similarity of the reports should not be based on the usage of the stop-words. In addition to the standard English stop-words, other words that should be ignored in the search for similarities could be

added to the list. For example, if the data set to be analyzed consists of all runway incursion reports, then the word RUNWAY is used in all of them and is not specific to the individual ones. Therefore, the word RUNWAY could be added to the stop-words. Figure 3-3 shows the beginning of the stop-words list.

```
"A","About","Above","Accordingly","Across","After","Afterwards",
"Again","Against","All","Allows","Almost","Alone","Along","Already",
"Also","Although","Always","Am","Among","Amongst","An","And","Another",
"Any","Anybody","Anyhow","Anyone","Anything","Anyway","Anyways",
"Anywhere","Apart","Appear","Appropriate","Are","Around","As","Aside", …
```

Figure 3-3. The Beginning of the Stop-Words List

The two files in the Workbench, stemming and stop-words, are populated with initial lists; additional entries to the files could be made as needed.

3.2.1 User Inputs for FindSimilar Tool Module

To run the FindSimilar tool, the user needs to specify the following options.

Set Weights. The user could indicate which fields (structured or free text) are more important in determining the similarity by assigning weights to the fields. The value entered for the weight should be a positive number. Assigning a weight of zero for a field indicates the field should not be considered for comparison and determining similarity. On the other hand, a weight of 1 or higher indicates the field should be considered. To consider all the fields equally for determination of similarity, assign a weight of "1" to all fields. If a field is assigned a higher weight than others, then a similarity in that field is considered more important. For example, if a user were interested in events that occurred at a particular destination, then the user would assign a weight of "2" to the field "Destination" and "1" to all other fields. This would instruct FindSimilar to return reports with a higher similarity in the field "Destination" versus the other fields.

The Workbench provides an interface for assigning and saving the weights, which can be accessed by pressing Select Weights Set (as seen on the left side of the screen shown in Figure 3-4).

Minimal Match Threshold. Users also need to specify the THRESHOLD, a percentage between 0 and 1, to show the minimum degree of similarity they want to see. A higher threshold will limit the discovered reports to the highly similar ones only.

The FindSimilar tool searches the fields with a weight of 1 or higher in all reports and compares them to the specified target. Depending on the frequency of the common words and the weight of each field, a similarity score is computed for each report. The reports with a similarity score equal to or higher than the THRESHOLD will be returned as discovered matches [4].

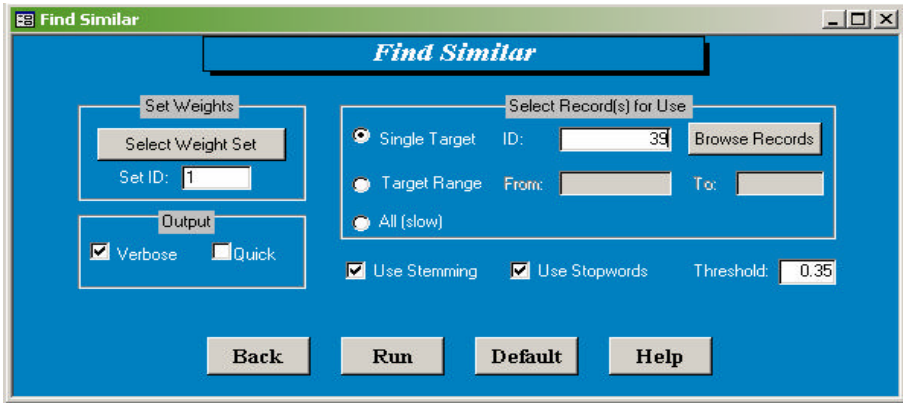

Figure 3-4. The FindSimilar User Interface Screen

3.2.2 Example Application of FindSimilar to American Airlines ASAP Data

TEXT DESCRIPTIONS, SHOWN IN FIGURES 3-5 THROUGH 3-8, ARE NOT ACTUAL REPORTS FROM AMERICAN AIRLINE'S ASAP DATABASE SINCE ASAP REPORTS ARE PROPRIETARY, BUT ARE REPRESENTATIVE EXAMPLES SIMILAR TO THE REAL DATA. To represent the reports, actual data is altered in various ways to de-identify the information and protect the anonymity of the reporter.

In these examples, altitude deviation reports were selected as the data subset to analyze. The weights in the FindSimilar Weight Set were selected such that only the narratives in the reports were compared for similarity and the other fields were not considered. This is because we were interested in similar causes (described in the narratives) and not, for example, similar flight dates or airports. The following report was selected as the target.

text description: … AIRPORT CLEARANCE TO 7000 FEET AFTER T.O. WE WERE DISTRACTED BY DEPT CONTROL CLEARANCE DIRECT TO A POINT AT LEVEL OFF TIME. FO WAS BUSY ENTERING IN THE FMS AND CA WAS BUSY CHECKING TO SEE WHAT THE FO WAS DOING AND LOOKING OUT THE WINDOW FOR TRAFFIC. A/C WAS FLOWN UP TO 7400 FEET UNTIL WE DESCENDED BACK DOWN TO 7000 FEET.

Figure 3-5. Selected Target Record for the FindSimilar Tool

The target report describes an altitude deviation due to the cockpit crew being busy with other tasks and therefore distracted from monitoring the altitude. The first similar report returned by the tool is shown in Figure 3-6. This report also describes a similar cause, the cockpit crew being busy and distracted.

text description: … ASSIGNED 11000- AIRCRAFT WENT TO 11-300 DUE TO PILOT FLYING DISTRACTED BY PROBLEMS WITH RADIO. PF WAS HANDFLYING HAD A NEW FO. SHOULD HAVE USED AUTOPILOT TO REDUCE WORKLOAD- BETTER HELP NEW FO.

Figure 3-6. Similar Report Returned by the FindSimilar Tool

Figure 3-7 shows an altitude deviation report selected as the target for a second run of the FindSimilar tool. Figure 3-8 shows a similar report returned by the tool. Both reports indicate the cause of distraction is described as cockpit crew being busy with other routine tasks they have to perform.

text description: …CENTER CLEARED US TO CROSS MEDOW INTER-SECTION AT FL230 – WE WERE AT FL330 APPROXIMATELY 40 MILES FROM THE FIX WITH A TAILWIND OF 170KTS. I RECOMMENDED THE FO START DOWN WHICH HE DID. I WAS OFF MAKING A PA AND FAILED TO NOTICE THAT THE DESCENT RATE THE FO WAS USING WOULD BE INADEQUATE TO MAKE THE CROSSING. RETURNING TO THE LOOP AFTER MAKING THE PA I NOTICED THAT WE WERE DESCENDING THROUGH FL240 AND WERE AT HANDY INT. WE LEVELED AT FL230 APPROXIMATELY 5 MILES LATE. CENTER QUESTIONED OUR ALTITUDE AT THIS POINT. NO OTHER COMMUNICATION WITH CENTER WAS MADE REGARDING THE MISSED CROSSING AND WE SHORTLY THEREAFTER SWITCHED TO APPROACH CONTROL IN RETROSPECT ESPECIALLY WITH AN INEXPERIENCED FO I SHOULD HAVE WAITED TO MAKE MY PA UNTIL I WAS SURE THAT THE CROSSING RESTRICTION WAS GOING TO BE MADE.

Figure 3-7. A Target Report for the FindSimilar Tool

text description: …IT WAS THE FIRST OFFICER*S LEG- WHEN ATC GAVE US A PILOT*S DISCRETION/PD/ FOR DESCENT TO FL 240. AS WE WERE LEVELING AT FL 240- WE WERE GIVEN A CLEARANCE TO CROSS CCT AT FL 200. AS WE LEVELED FL200-ATC GAVE US A CLEARANCE TO CROSS MELOW AT 11-000. AS WE DESCENDED THROUGH FL200- THE CAPTAIN LEFT THE RADIOS TO MAKE A PA TO THE PASSENGERS. AFTER THE PA WAS COMPLETE AND NEARING MELOW WE RESET ALTI-METERS FROM 29.92 TO 30.57 IN ADDITION ATC AMENDED OUR CLEARANCE TO MAINTAIN 11-000 AS WE CROSSED MELOW AT APPROXIMATELY 11-400.

Figure 3-8. Similar Report Returned by the FindSimilar Tool

3.2.3 Findings

The FindSimilar tool was applied to the ASAP data from second and third periods. The first period (prior to September 11, 2001) did not have enough text descriptions for the analysis. The second period (after September 11, 2001 through November 2002) had thousands of ASAP reports and it took a couple of minutes for the tool to execute them. The third period (after November of 2002) had hundreds of ASAP reports and running the tool on them took a few seconds. The execution time on this tool increases with the increase in number of reports. Among three tool modules of the Workbench, this tool is the one that will take longer to execute since it searches the text fields in addition to the structured fields.

The tool correctly grouped similar reports in a short amount of time. The results were displayed with the target reports on the left side of the screen and the discovered similar reports on the right side. The similar reports are sorted with the most similar one on the top, and the user could scroll down to the least similar report found.

The speed of the process, using the Workbench, was much faster than the traditional way of grouping similar reports. The number of similar reports returned by the tool depended on the level of threshold. When the threshold was set to a high percentage, the number of returned similar reports was limited to less than ten reports. When lowering the threshold, the tool returned more number of similar reports with the less similar ones at the end of the list. The "high" and "low" values for the threshold are relative and depend on the data being analyzed.

3.3 Application of FindAssociations Tool Module

The FindAssociations tool searches the structured fields in the reports. FindAssociations would be employed when users want to discover strong relationships in the data. The tool could be run on the entire data set or on a selected subset. In either case, the user does not need to specify which associations to look for in the data. The tool examines all possible associations and returns the ones that are above the user-specified thresholds. The user will

need to review the discovered associations and determine which ones are expected and which ones are unexpected and require further investigation. Depending on the level of thresholds and the number of data fields being analyzed, a large number of associations, not all of which candidates for further attention, could be returned to the user.

3.3.1 User Inputs for FindAssociations Tool Module

To run the FindAssociations tool module the user must specify two items: SUPPORT and CONFIDENCE, as shown in Figure 3-9. No other values need to be specified by the user. The tool will search all combinations of attribute values and returns outstanding associations. Outstanding associations are the ones that meet the specified SUPPORT and CONFIDENCE values.

SUPPORT is the minimum number of times a field value (or combination of values) should exist in the data in order to consider its associations. For example, a support of 0.5 (or 50%) indicates the user is interested in associations of field values that appear in at least 50% of the data records.

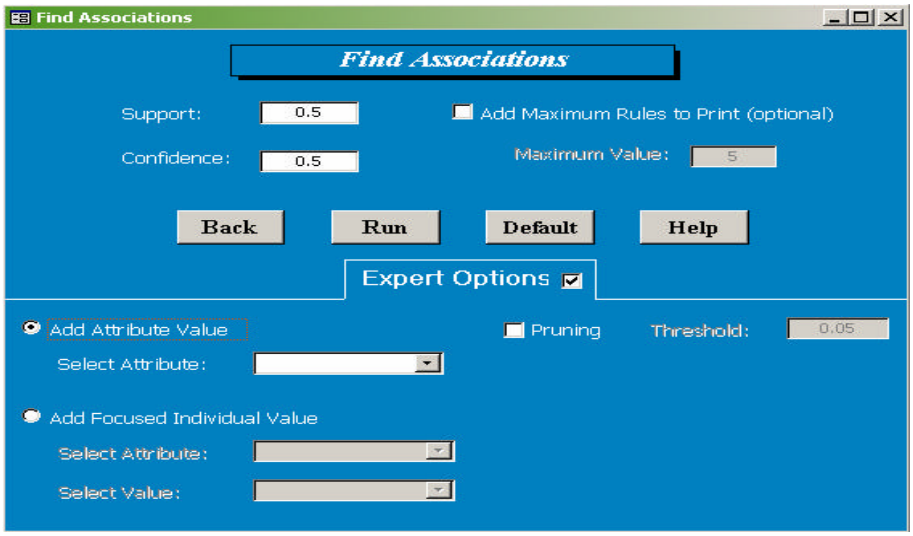

Figure 3-9. The FindAssociations User Interface Screen

CONFIDENCE indicates the strength of the association. For example, a confidence of 0.5 means the user wants to see associations for field values that appear together at least 50% of the time.

Users could also check the '**Expert Options**' box and access the additional options described below.

Add Attribute Value. Allows users to select a field to focus on the associations of values of that particular attribute. For example, selecting the individual field *phase* will search for associations of different *phases* with other fields.

Add Focused Individual Value. Allows users to select a value of a field to focus on. For example, selecting *phase* with the value *takeoff* will search for associations of (*phase = takeoff)* with other fields.

3.3.2 Example Application of FindAssociations to American Airlines ASAP Data

Following are examples of types of findings that could be obtained using the FindAssociations tool. PLEASE NOTE THAT THE FIELDS AND VALUES IN THESE EXAMPLES ARE NOT ACTUAL DATA FROM AMERICAN AIRLINE'S ASAP DATABASE SINCE ASAP REPORTS ARE PROPRIETARY, BUT ARE REPRESENTATIVE EXAMPLES SIMILAR TO THE REAL DATA. Actual data is altered in various ways to de-identify the information and protect the anonymity of the reporter.

Examples of what the FindAssociations tool could discover, using a SUPPORT of 0.25 and a CONFIDENCE of .05, are shown in Figure 3-10.

55% of {event = ALT_DEVIATION, Aircraft Series = 300} coincide with {phase=APPROACH}
(55% of altitude deviations with 300-series aircraft occurred during the APPROACH phase of flight)

78% of {departure = FLORIDA} coincide with {event = ALT_DEVIATION}
(78% of reported incidents with flights departing Florida have involved altitude deviations)

Figure 3-10. The FindAssocations Findings Examples

The findings might be explained by other facts about the data (such as total number of 300-series aircraft in the airline's fleet, and total number of flights departing Florida in the time period under analysis) and therefore high associations might be expected. It is also possible that the findings do not have an obvious explanation and further investigation might be necessary to determine the cause of high associations. For example, further investigations, focusing on flights departing from Florida in the time period under analysis, might reveal that a certain problem in communication with the tower, certain equipment malfunction in the aircraft, or certain pilot behavior could be causing the deviations.

Note that the values in the above findings were NOT specified by the user ahead of time. For example, the user did not ask for associations between altitude deviations and Florida departures. Only the thresholds are specified by the user. The tool identifies outstanding associations among all values and brings them to the user's attention.

3.3.3 Findings

The FindAssociations tool was applied to all three periods of ASAP reports. The tool was easy to run and re-run on different sets of data. The results matched findings of the traditional methods previously used at American Airlines and confirmed the effectiveness and accuracy of the tool. Also, smaller subsets of data were selected for analysis based on findings of the other two tools in the Workbench and the results confirmed findings of the other tools and helped narrowing down the analysis on the data set of interest. In addition to the accuracy, applying the tool to thousands of reports took only a few seconds, which was a big improvement in the speed of the analysis.

3.4 Application of FindDistributions Tool Module

The FindDistributions tool searches the structured fields in the reports, and identifies unusual distributions of incidents. It provides a method for monitoring over time the expected versus actual rate of occurrence of events. For example, it might assist in identifying seasonal trends, or in identifying an emerging problem caused by changes in policy.

3.4.1 User Inputs for FindDistributions Tool Module

To run the FindAssociations tool, the user needs to specify the focused attribute and set the rest of the options or use the provided default values.

FOCUSED ATTRIBUTE. Users need to select the field they want to focus on (Focused Attribute, as shown in Figure 3-11). For example, to search for anomalies in the distribution of incidents over different months of the year, select MONTH as the Focused Attribute. The tool then calculates the distribution of subsets of incidents over the selected focused field (MONTH in this case). Those subsets that differ most from the overall distribution are identified as unexpected [5].

The tool uses three other parameters as explained below. Users could use the default values provided for these parameters or change them as desired.

Count. A positive number determining the minimum number of values to be included in any subset of data returned for the focused attribute. To understand this parameter better, consider the following formula that calculates the "count" for each subset of data whose distribution is being compared to the total (or expected) distribution. Assuming the focused attribute is MONTH, the formula calculates the "count" for the subsets for each MONTH.

The following example shows the calculation for the month of January.

Count = (number of reports in the subset * number of reports in January) /
(Total number of reports for all months)

If this calculated "count" for the subset is less than the specified "count" on the FindDistributions screen, the subset will be ignored. Setting a higher "count" results in fewer returns from this module, allowing the user to focus on the larger data subsets.

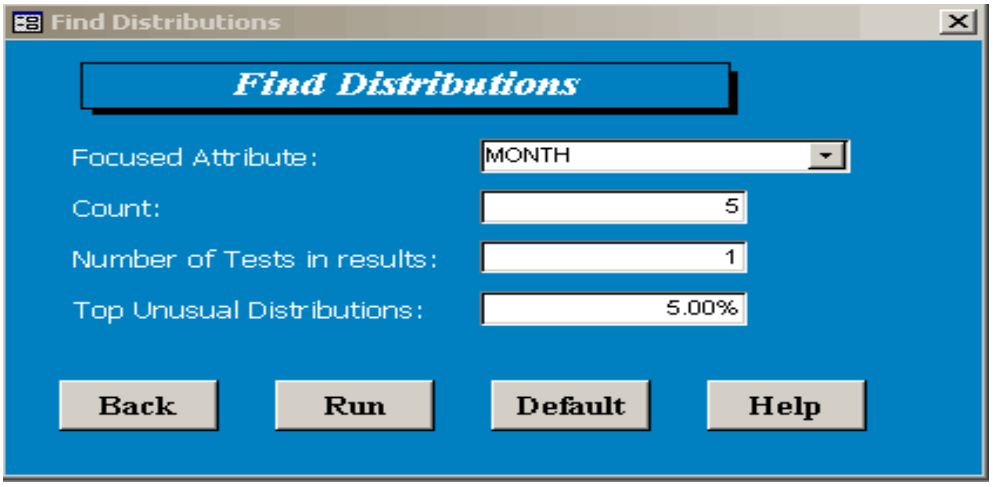

Figure 3-11. The FindDistributions User Interface Screen

Number of tests in results. A positive number indicating the maximum number of fields combined in a subset. For example, a value of 1 in this parameter, will instruct the tool to look at subsets of data that have common values for single fields (e.g., incidents of type *Runway Incursion, incidents that occurred in MIAMI, etc.*); a value of 2 in this field, allows looking at data subsets that have common values in two fields in addition to subsets with one common field (e.g., incidents of *Runway Incursion* type in *MIAMI airport*).

Top unusual distributions. A number determining whether the tool shows only the very unusual distributions or shows slightly unusual distributions as well. Enter a number between 0 and 1 (exclusive) or a percentage between 0% and 100% (exclusive) in this field. A small number in this field returns only the distributions that are very unusual. A larger number allows less unusual distributions to be displayed.

3.4.2 Example Application of FindDistributions to American Airlines ASAP Data

Figure 3-12 shows an example of the FindDistributions findings. PLEASE NOTE THAT THE FIELDS AND VALUES IN THESE EXAMPLES ARE NOT ACTUAL DATA FROM AMERICAN AIRLINE'S ASAP DATABASE SINCE ASAP REPORTS ARE PROPRIETARY, BUT ARE REPRESENTATIVE EXAMPLES SIMILAR TO THE REAL DATA. Actual data is altered in various ways to de-identify the information and protect the anonymity of the reporter.

In this example, the field MONTH was selected as the Focused Attribute. The findings returned by the tool are displayed in a Microsoft Excel spreadsheet where the user can use the Chart option and generate graphs from the returned information. The graph presented in Figure 3-12 was generated from the information returned by the tool for the selected attribute, MONTH. The black and white bars on the chart indicate the expected distribution and the solid black bars indicate the actual distribution for taxi-in/taxi-out incidents. The expected distribution is calculated by counting the number of all incidents (not just taxi-in/taxi-out) for each month. These counts are then scaled down for comparison with the taxi-in/taxi-out subset of incidents. Therefore, the black and white bars do not show the number of the incidents but are relative percentages indicating the pattern of distribution of all incidents over different months. Since actual and expected distributions for the taxiing incidents do not follow the same pattern, the tool has identified it as an anomaly and brought it to the user's attention.

In viewing the chart in Figure 3-12, questions that may occur to the user include: Why is the number of incidents during taxi-in/taxi-out much higher than expected in the month of May and down to zero in August? Is the total number of flights in these months a factor? Have the same airports been used during the four months under analysis? Have different taxiways been used in the months of July and August? Or has there been a change in the taxiways, policies, or pilot training immediately before the month of August? These questions could be investigated further to find the explanation for the unusual increase/decrease of taxiing incidents in these four months. Further investigation might identify causal factors of the incidents, which could then be used for prevention of these incidents.

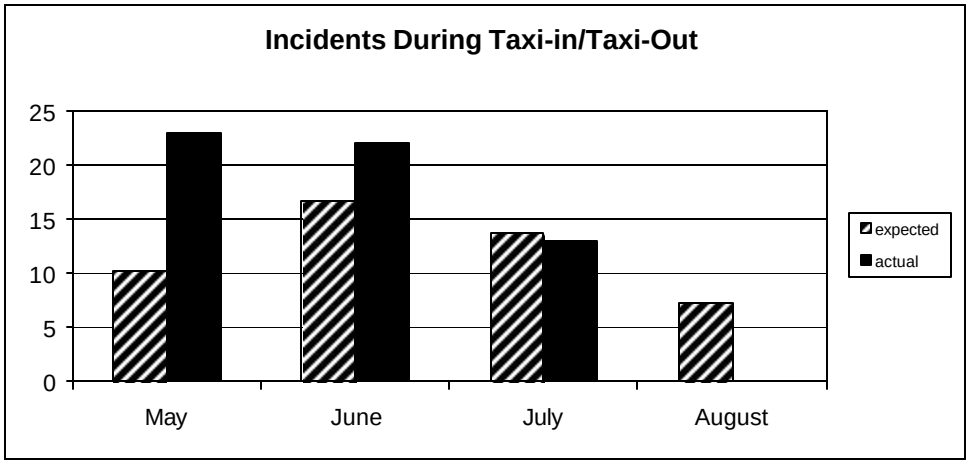

Figure 3-12. Example Output of the FindDistributions Tool

3.4.3 Findings

The FindDistributions tool was applied to all three periods of ASAP reports. The tool was easy to run and the results were easy to understand. Among the findings, some were the same as those previously obtained by the ASAP staff using traditional analysis methods. They found this correlation confirmed the accuracy of the tool, and were greatly pleased with the significant reduction in time needed to derive the findings. Running the tool on thousands of reports took only a few seconds to execute; the results returned by the tool were in a format ready to generate graphs in seconds.

Section 4

Conclusion

This proof-of-concept project applied the Aviation Safety Data Mining Workbench to American Airlines ASAP reports in order to assess the usefulness of the Workbench. The project demonstrated that installation of the Workbench is very easy and it is easy to learn how to work with the Workbench. The three tool modules in the Workbench were found effective in making accurate and useful discoveries in the data. The generated results were easy to understand and interpret by the analyst. Other key conclusions from this proof-of-concept demonstration with American Airlines include:

1. The three tools of the Data Mining Workbench (FindSimilar, Find Associations, and FindDistributions) offer benefits when used independently or in conjunction with each other (e.g., FindAssociations was used to focus on associations of a field whose distribution was found unusual by FindDistributions.)
2. The FindSimilar tool provided immediate grouping of reports having similar factors.
3. The FindAssociations and FindDistributions tools can provide unexpected correlations that call for further investigation.
4. This project identified specific problem areas previously recognized by American Airlines analysts, validating the accuracy and usefulness of the tool.
5. The Data Mining Workbench was much quicker than other analytical systems in use at the airline in identifying specific issues of interest. American considers this increase in efficiency/productivity to be critical to their safety analysis work.

It is important to note that data mining tools help the human analysts but do not eliminate the need for them. Once the discoveries are made by the tools, a human analyst is needed to go over the findings and interpret them in the context of the data being analyzed.
American Airlines provided the following comments regarding application of the Data Mining Workbench to ASAP data.

The three tools offer benefits when used independently or in conjunction with each other. The Distributions and Associations tools can provide unexpected correlations that call for further investigation. The distributions in this review showed various apparently sporadic deviations from expected distributions. While such results do highlight areas warranting review, they can be seen as affirming a lack of consistent undetected or unattended to weaknesses. The distribution peaks revealed in this look-back study did relate to specific problem areas that had been recognized and addressed with corrective actions by the Event Review Team. This validated the accuracy and usefulness of the tool. It is worth noting that the tool produced indications and graphic products in seconds that had taken days of work

with the staff and tools currently available to the ASAP manager. Distribution comparisons using the MITRE tool over extended time periods in the data may reveal seasonal factors not previously recognized.

The results of the FindAssociations show great promise for application to the broader set of data fields coming with the company's new web-based reporting system. This should lead to statistical support for relationships between particular errors and related crew and situational factors. Presenting documentation of these associations should assist our operational department managers in identifying root causes that lead to modifying training, procedures, or the operating environment to improve performance. The FindSimilar tool did provide immediate grouping of reports having similar factors. In seconds it achieved what was previously hours of work to collect reports for study or presentation in support of a risk warning or recommended change to policy.

In short, we (the partnering airline) are thrilled to have the speed, flexibility, and accuracy of these tools - especially for application on our coming field rich data base of self-reported crew errors and safety concerns. They will greatly enhance our responsiveness to analyze and report significant concerns, deviations, and correlations.

List of References

1. Global Aviation Information Network (GAIN) website: www.gainweb.org
2. Nazeri, Z., Bloedorn, E., and Ostwald, P., "Experiences in Mining Aviation Safety Data", Association for Computing Machinery (ACM) Special Interest Group on Management of Data (SIGMOD) Santa Barbara, California, May 21-24, 2001.
3. Federal Aviation Administration, March 2000, *Aviation Safety Action Program*, Advisory Circular 120-66B, Washington, D.C.
4. Bloedorn, E., "Mining Aviation Safety Data: A Hybrid Approach," Armed Forces Communications and Electronics Association (AFCEA) First Federal Data Mining Symposium, Washington D.C., March 28-29, 2000.
5. Harris, E., Bloedorn, E., Rothleder, N., "Recent Experiences with Data Mining in Aviation Safety," Special Interest Group on Management of Data, Data Mining and Knowledge Discovery (SIGMOD-DMKD) Workshop, Seattle, Washington, June 1998.