Data Mining Applications in
Aeronautics & Space Exploration Workshop

Clustering & Recurring Anomaly Identification: Recurring Anomaly Detection System (ReADS)

Dawn McIntosh

June 21, 2006

Discovery and Systems Health Technical Area
NASA Ames Research Center - Computational Sciences Division

Contact: Dawn McIntosh, 650-604-0157
Dawn.M.McIntosh@nasa.gov

Credit given to those involved

- NASA ARC, Intelligent Systems Division, Discovery & Systems Health Area, Intelligent Data Understanding (IDU) Group
  - Dr. Ashok Srivastava
  - Eugene Turkov
  - Brett Zane-Ulman
- NASA ARC, Intelligent Systems Division, Advanced Engineering Network (AEN) Group
  - Dr. David Bell
  - Mohana Gurram
  - Peter Tran
  - Jenessa Lin
  - Chris Knight

Agenda

- What we are trying to accomplish
- What we HAVE accomplished
- Demo ReADS

Problem Introduction

NASA programs have large numbers (and types) of problem reports.

- ISS PRACA: 3000+ records, 1-4 pages each;
- ISS SCR: 28,000+ records, 1-4 pages each;
- Shuttle CARS: 7000+ records, 1-4 pages each;
- ASRS: 27000+ records, 1 paragraph each

These free text reports are written by a number of different people, thus the emphasis and wording vary considerably

With so much data to sift through, analysts (subject experts) need help identifying any possible safety issues or concerns and to help them confirm that they haven't missed important problems.

- Unsupervised clustering is the initial step to accomplish this;
- We think we can go much farther, specifically, identify possible recurring anomalies.
  - Recurring anomalies may be indicators of larger systemic problems.

Text Mining - ReADS

Recurring Anomaly Detection System (ReADS):

The Recurring Anomaly Detection System (ReADS) is a tool to analyze text reports, such as aviation reports and maintenance records.

- – Text clustering algorithms group large quantities of reports and documents.
  - Reduces human error & fatigue
- – Identifies interconnected reports;
  - Automates the discovery of possible recurring anomalies;
- – Provides a visualization of the clusters and recurring anomalies

We have illustrated our techniques on data from Shuttle and ISS discrepancy reports, as well as ASRS data.

ReADS has been integrated with a secure online search tool: NX

Contact: Dawn McIntosh, 650-604-0157
Dawn.M.McIntosh@nasa.gov

ReADS Text Mining Algorithms

Unsupervised Clustering:

Spherical k-means → modified von Mises Fisher.

Recurring Anomaly Identification:

1. Identify reports which mention other reports as a recurring anomaly
   a. Using regular expressions to search documents and identify mention of other reports by name.

2. Detect recurring anomalies,
   a. find the similarity between documents to detect recurring anomalies using cosine distance similarity measure,
   b. then according to the similarity measure, run a hierarchical clustering algorithm to cluster the recurring anomalies.

Contact: Dawn McIntosh, 650-604-0157
Dawn.M.McIntosh@nasa.gov

Recurring Anomaly Algorithm

1. Cosine similarity measure;
2. Hierarchical Clustering
   - After calculating the distance between each document, the algorithm applies single linkage, i.e., nearest neighbor, to create a hierarchical tree representing connections between documents.
     - Also generates an ‘inconsistency coefficient’ which is a measure of the relative consistency of each link in the tree.
   - The hierarchical tree is partitioned into clusters by setting a threshold on the inconsistency coefficient.
     - A high inconsistency coefficient implies that the reports could be very different and still be sorted into the same cluster.
   - Currently the inconsistency coefficient threshold is set very low, which returns many smaller clusters of very similar reports.
     - Clusters consisting of only one document are excluded from the recurring anomaly results.

Shuttle CARS dataset →Toy Dataset

Shuttle Corrective Action Reporting System (CARS)	**Real Dataset (analyzed by experts)**	**Toy Dataset**	**Algorithm Results using Toy Dataset (similarity measure clustering threshold = 0.2)**	**Algorithm Results using Toy Dataset (similarity measure clustering threshold = 0.4)**
# of Documents	7440	344	344	344
# of RA Clusters	366	20	RegEx: 28 SimMeasure: 18	RegEx: 28 SimMeasure: 33
# of Total Documents in RA Clusters	1570	70	RE+SM = 92+56 = 118	RE+SM = 92+116 = 208

Min & Max size of RA Clusters	Min = 2
	Max = 48
		Min = 2
		Max = 10
			Min = 2
			Max = 8
				Min = 2
				Max = 9

Evaluation of Clustering Results

- #1 Goal: Don't miss documents identified by the experts as a Recurring Anomaly
- #2 Goal: Get the same results as the experts
- #3 Goal: Find Recurring anomalies missed by the experts.

- Criteria:
  - To meet #1, the ReADS RAs only have to overlap with the experts. The same documents don't have to fall into the same RA clusters. Therefore, if an expert RA cluster contains Docs A, B, & Z, and those documents fall into two ReADS clusters, this is still a success:

    Expert Cluster: A, B, Z
    ReADS Cluster: A, Z
    ReADS Cluster: B, P, M
  - To meet #2, an Expert RA cluster should be identical to a ReADS RA cluster.

    Expert Cluster: C, L, R, T
    ReADS Cluster: C, L, R, T
  - To meet #3, ReADS correctly identifies a set of documents which the Experts did not.

    Experts Unused Document Cluster: D, E, F, G, H, I, J, K, M, N, O, P, Q, S, U, V, W, X, Y

    ReADS Cluster: F, I, N, D

Contact: Dawn McIntosh, 650-604-0157
Dawn.M.McIntosh@nasa.gov

Shuttle CARS dataset →Toy Dataset

Shuttle Corrective Action Reporting System (CARS)	Real Dataset (analyzed by experts)	Toy Dataset (selected from Real CARS dataset)	Comments
# of Total Documents	7440	333	344-70=274, selected randomly from 7440-1570 non-RA reports.
# of RA Clusters	366	20	Toy clusters selected to match, as much as possible, a variety of the types of RAs identified by NESC.
# of Total Documents in RA Clusters	1570	70	

Min & Max size of RA Clusters	Min = 2
	Max = 48
		Min = 2
		Max = 10
			Toy Dataset RA clusters didn't cover the breadth of the cluster sizes, but the large clusters were rare.

Contact: Dawn McIntosh, 650-604-0157
Dawn.M.McIntosh@nasa.gov

ReADS stats on Toy Dataset

Shuttle Corrective Action Reporting System (CARS)	Experts Results using Toy Dataset	ReADS Results using Toy Dataset	ReADS Results using Toy Dataset
Similarity Measure Clustering Threshold	NA	0.2 (documents must be very similar to qualify)	0.4 (a less conservative threshold)
# of Total Documents	333	333	333
# of RA Clusters	20	RegEx: 28 SimMeasure: 18	RegEx: 28 SimMeasure: 33
# of Total Documents in RA Clusters	70	RE+SM = 92+56 = 118 (note: There's overlap!)	RE+SM = 92+116 = 208 (note: There's overlap!)

Min & Max size of RA Clusters	Min = 2
	Max = 10
		Min = 2
		Max = 8
			Min = 2
			Max = 9

Toy Dataset Confusion Matrix: ReADS similarity measure vs. Experts

	Shuttle CARS Toy Dataset	ReADS Recurring Anomaly Clusters																																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	RAs missed by ReADS	subtotals =	
Subject Experts Recurring Anomaly Clusters	1											5																								5	
	2												2																							2	
	3																															2				2	
	4			4		1																														5	
	5														2																					2	
	6																2																			2	
	7																																		2	2	
	8					2																														2	
	9																																		2	2	
	10																										7								2	9	
	11	4					2																													6	
	12																	3																		3	
	13						3																													3	
	14																																		2	2	
	15																					1		5											1	7	
	16		1											1																						2	
	17		1											1																						2	
	18																																		2	2	
	19																																		2	2	
	20																						7												1	8	
	RAs missed by Experts		3		2	1		5	2	2	2					3			2	2	3	8			2	3	2	2	3	5	2		3	3	203	263	
subtotals =		4	5	4	2	4	5	5	2	2	2	5	2	2	2	3	2	3	2	2	3	9	7	5	2	3	9	2	3	5	2	2	3	3	217	TOTAL =	333
																																				333	

Exact Matches between Experts & ReADS (Goal #2)

ReADS clusters completely missed by Experts (Goal #3)

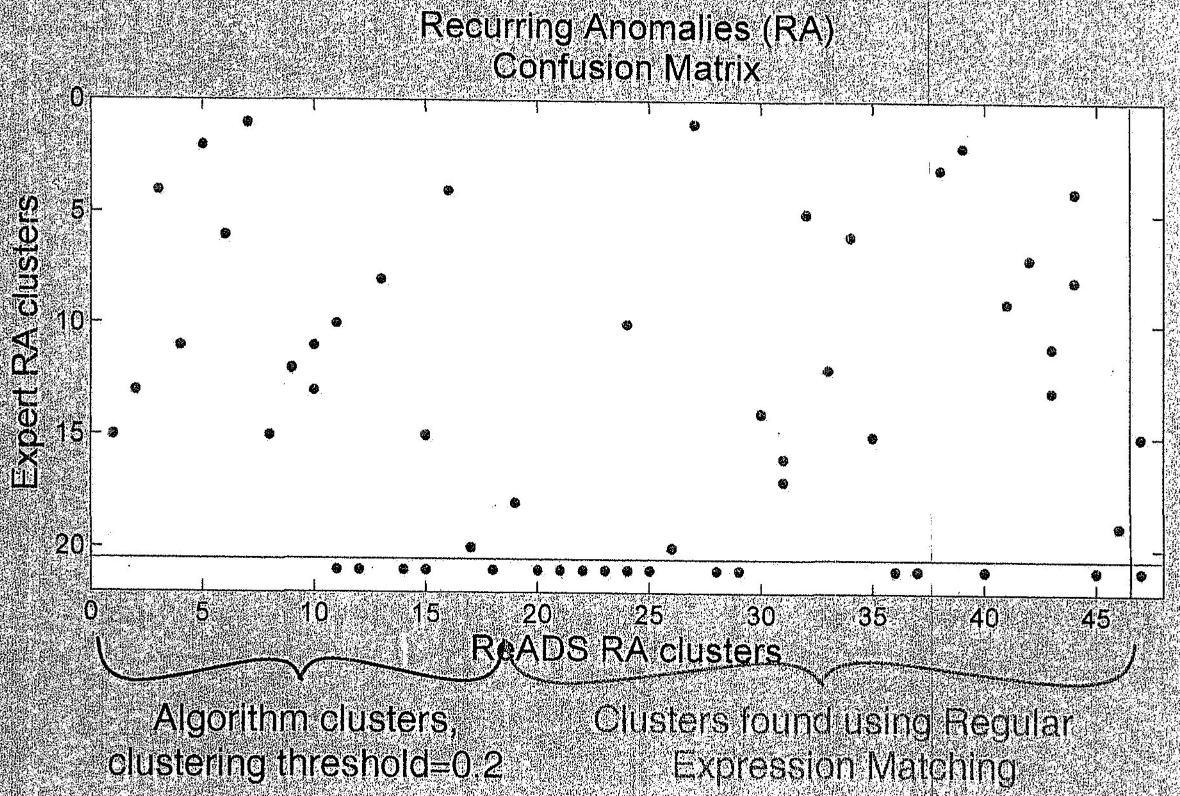
Expert Clusters missed by ReADS similarity measure algorithm, but caught by the Regular Expression matching (partial failure of Goal #1)

Only document in the toy dataset completely missed by ReADS (failure of Goal #1)

Discovery and Systems Health Technical Area
NASA Ames Research Center - Computational Sciences Division

Contact: Dawn McIntosh, 650-604-0157
Dawn.M.McIntosh@nasa.gov

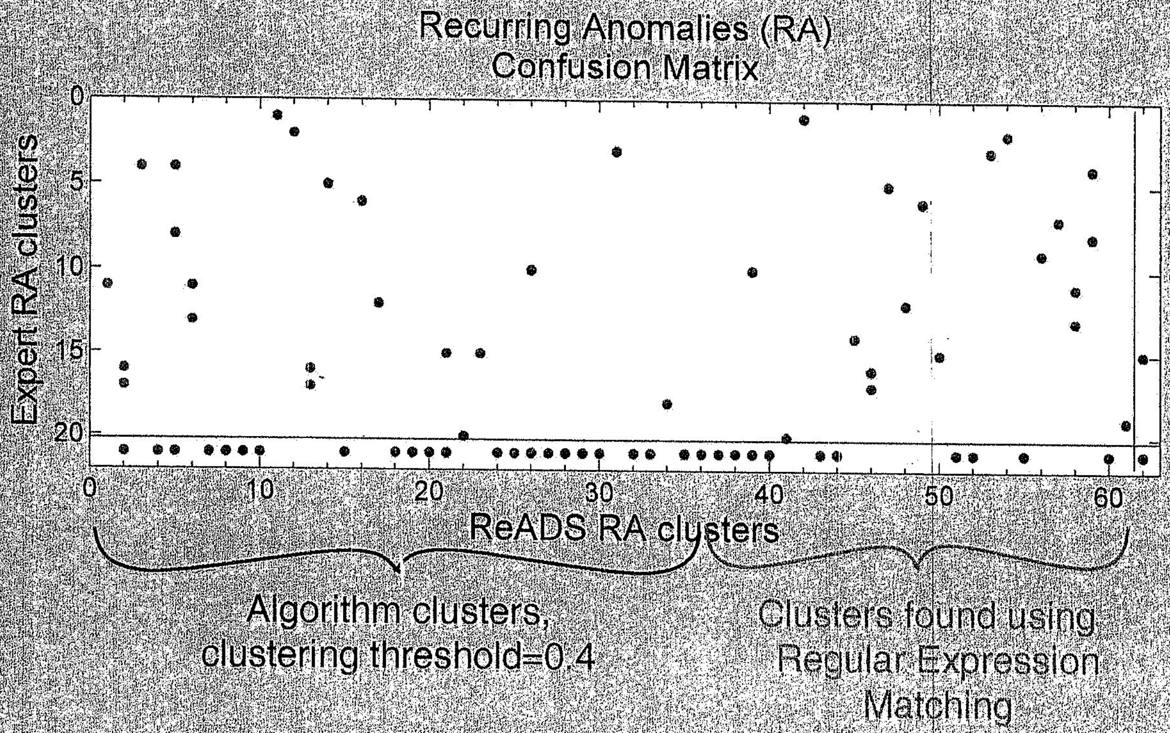
Preliminary Toy Dataset Results: Using a conservative clustering threshold



Contact: Dawn McIntosh, 650-604-0157
Dawn.M.McIntosh@nasa.gov

Preliminary Toy Dataset Results: Less conservative clustering threshold

Recurring Anomalies (RA) Confusion Matrix



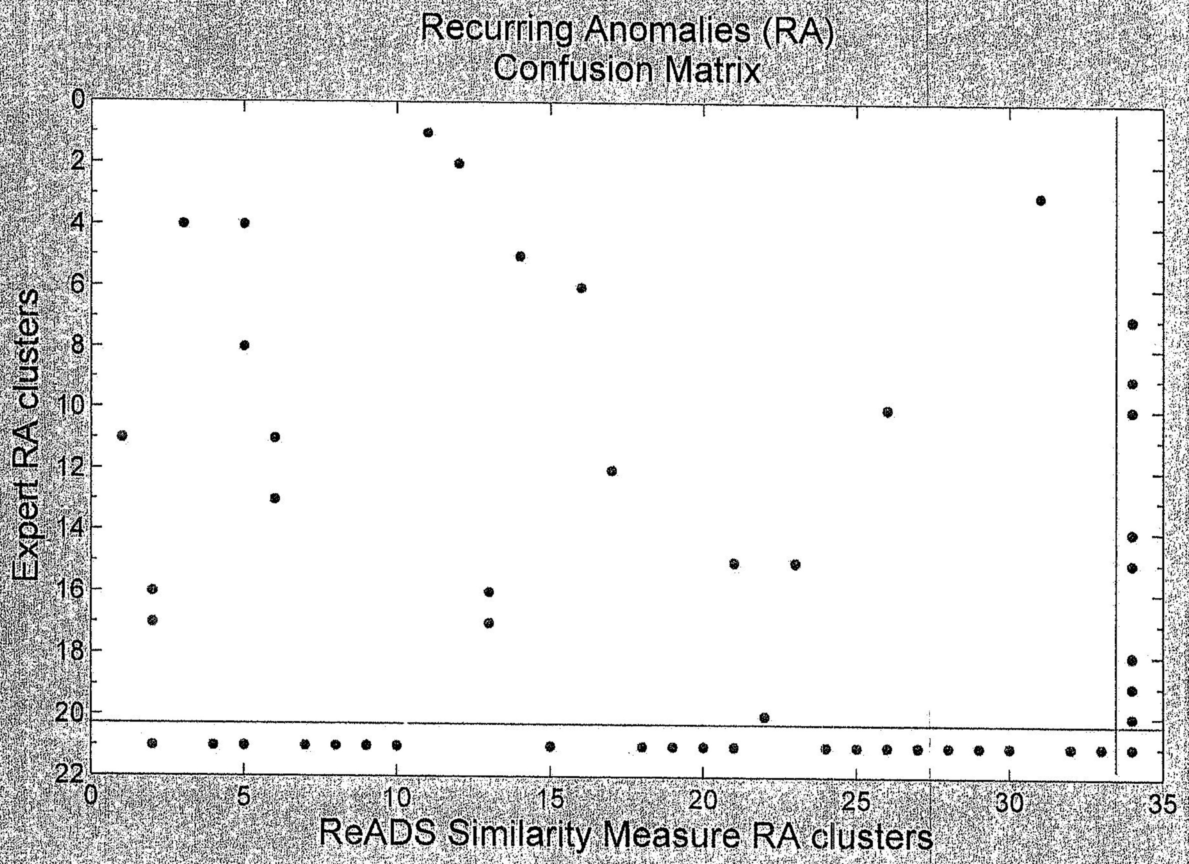
Algorithm clusters, clustering threshold=0.4

Clusters found using Regular Expression Matching

Discovery and Systems Health Technical Area
NASA Ames Research Center - Computational Sciences Division

Contact: Dawn McIntosh, 650-604-0157
Dawn.M.McIntosh@nasa.gov

Preliminary Toy Dataset Results: SimMeasure Algorithm Only (No RegEx Matching)

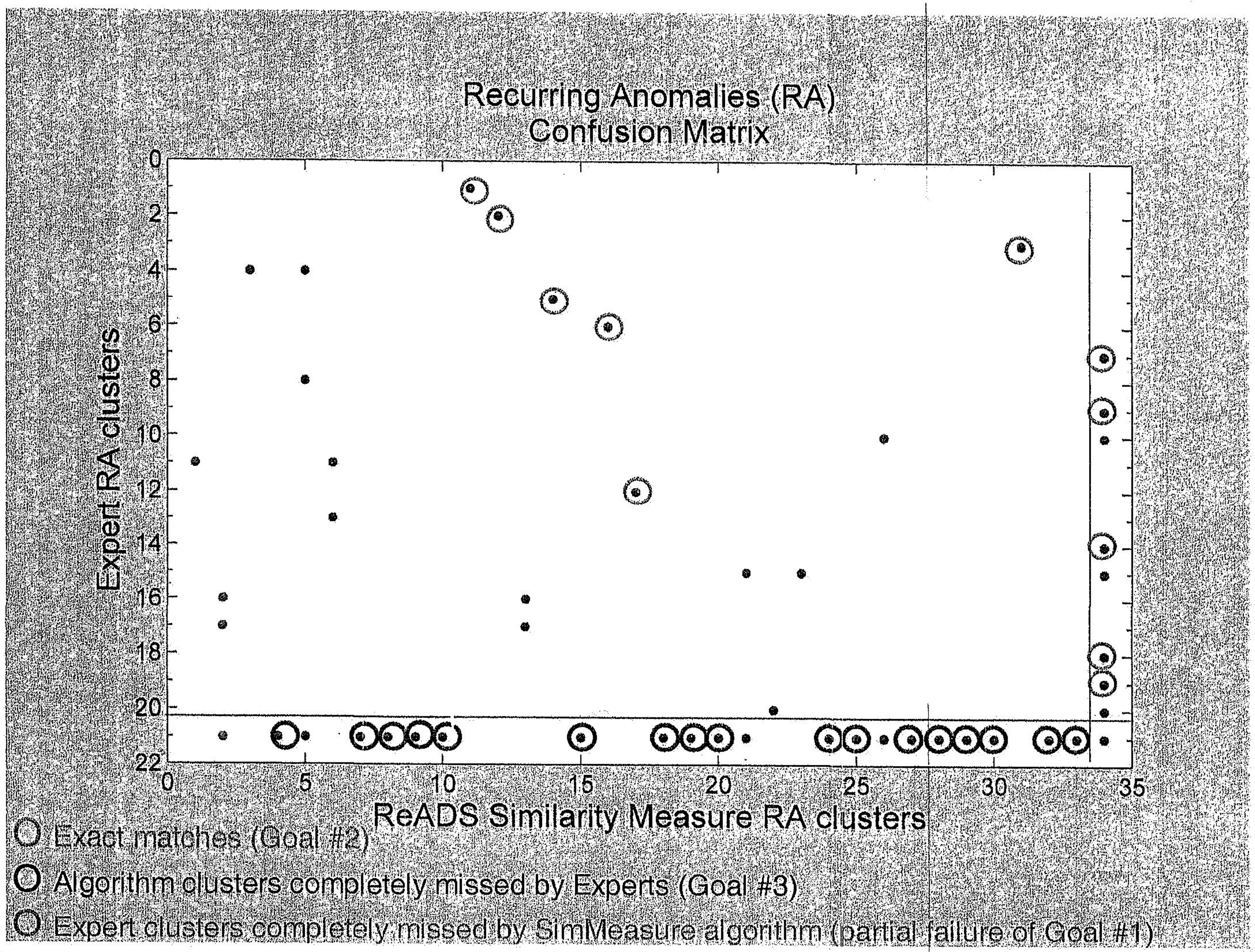


Clustering threshold = 0.4

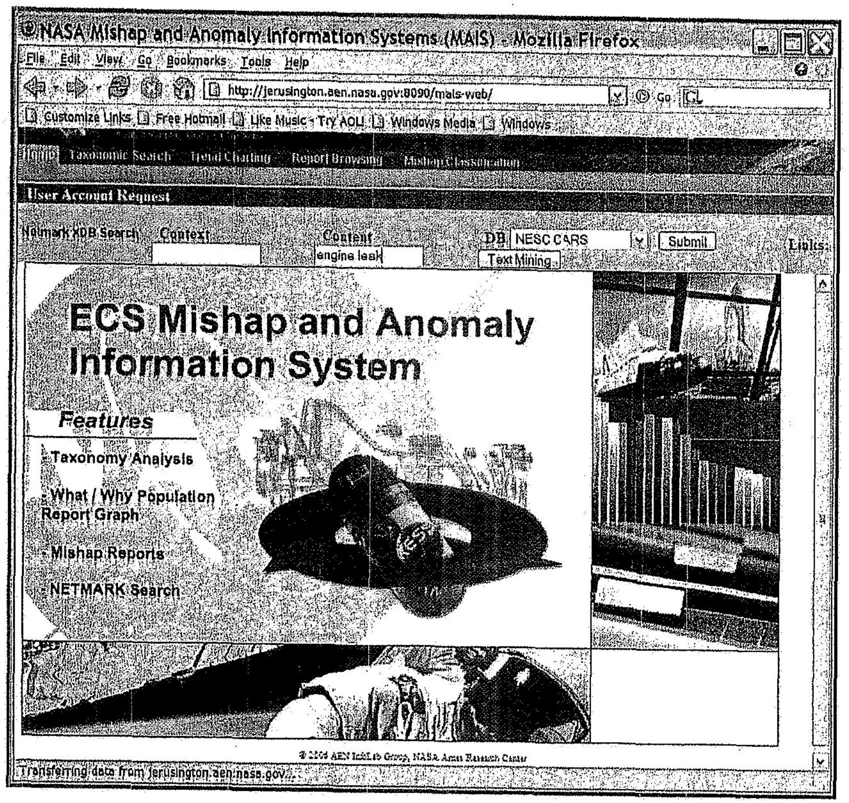
Discovery and Systems Health Technical Area
NASA Ames Research Center - Computational Sciences Division

Contact: Dawn McIntosh, 650-604-0157
Dawn.M.McIntosh@nasa.gov

Preliminary Toy Dataset Results: SimMeasure Algorithm Only (No RegEx Matching)



ReADS System & Interactive Visualization


NASA Mishap and Anomaly Information Systems (MAIS) - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://jerusington.aen.nasa.gov:8090/mais-web/
Go
Customize Links Free Hotmail Like Music - Try AOL! Windows Media Windows
Home Taxonomic Search Trend Charting Report Browsing Mishap Classification
User Account Request
Netmark XDB Search Context Content DB Submit Links
Text Mining
ECS Mishap and Anomaly Information System
Features
Taxonomy Analysis
What / Why Population Report Graph
Mishap Reports
NETMARK Search
© 2004 AEN InfoLab Group, NASA Ames Research Center
Transferring data from jerusington.aen.nasa.gov...

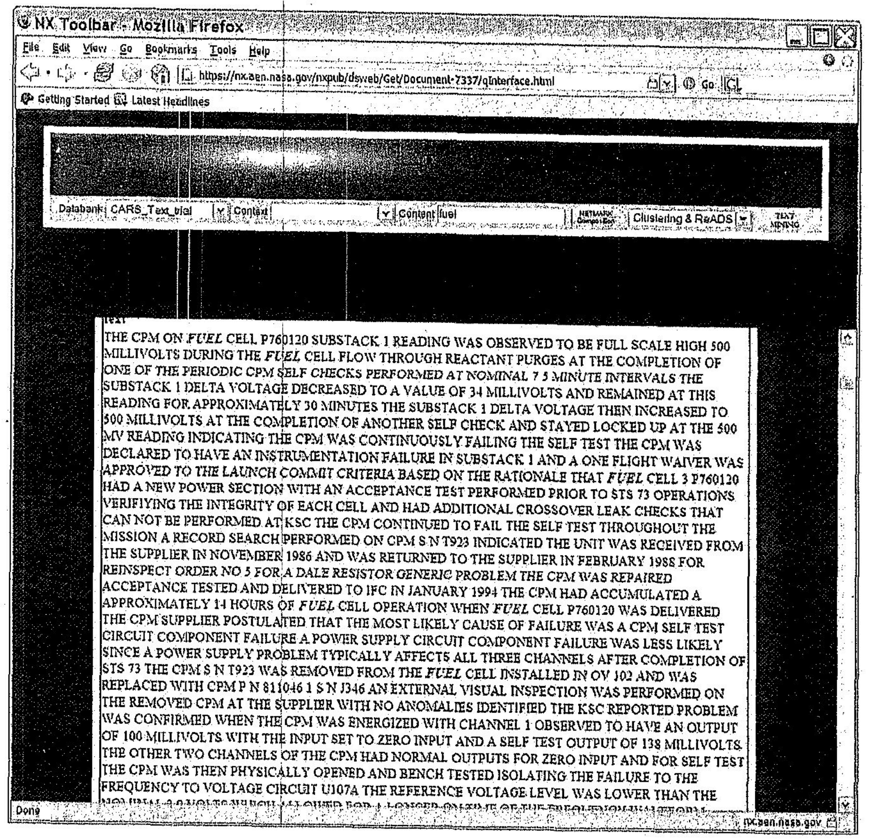
Content: engine leak

DB: NESC CARS

Web URL:
http://jerusington.aen.nasa.gov

Currently integrating w/ NX
https://nx.aen.nasa.gov/nxpub

Online secure search & text mining system
Multiple DBs available for search & text mining


NX Toolbar - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
https://nx.aen.nasa.gov/nxpub/dsweb/Get/Document-7337/qInterface.html
Go
Getting Started Latest Headlines
Databank Context Content
NETMARK
TEXT MINING
Text
Done
nx.aen.nasa.gov

Databank: CARS_Text_bial

Content: fuel

Clustering & ReADS

THE CPM ON *FUEL* CELL P760120 SUBSTACK 1 READING WAS OBSERVED TO BE FULL SCALE HIGH 500 MILLIVOLTS DURING THE *FUEL* CELL FLOW THROUGH REACTANT PURGES AT THE COMPLETION OF ONE OF THE PERIODIC CPM SELF CHECKS PERFORMED AT NOMINAL 7 5 MINUTE INTERVALS THE SUBSTACK 1 DELTA VOLTAGE DECREASED TO A VALUE OF 34 MILLIVOLTS AND REMAINED AT THIS READING FOR APPROXIMATELY 30 MINUTES THE SUBSTACK 1 DELTA VOLTAGE THEN INCREASED TO 500 MILLIVOLTS AT THE COMPLETION OF ANOTHER SELF CHECK AND STAYED LOCKED UP AT THE 500 MV READING INDICATING THE CPM WAS CONTINUOUSLY FAILING THE SELF TEST THE CPM WAS DECLARED TO HAVE AN INSTRUMENTATION FAILURE IN SUBSTACK 1 AND A ONE FLIGHT WAIVER WAS APPROVED TO THE LAUNCH COMMIT CRITERIA BASED ON THE RATIONALE THAT *FUEL* CELL 3 P760120 HAD A NEW POWER SECTION WITH AN ACCEPTANCE TEST PERFORMED PRIOR TO STS 73 OPERATIONS VERIFIYING THE INTEGRITY OF EACH CELL AND HAD ADDITIONAL CROSSOVER LEAK CHECKS THAT CAN NOT BE PERFORMED AT KSC THE CPM CONTINUED TO FAIL THE SELF TEST THROUGHOUT THE MISSION A RECORD SEARCH PERFORMED ON CPM S N T923 INDICATED THE UNIT WAS RECEIVED FROM THE SUPPLIER IN NOVEMBER 1986 AND WAS RETURNED TO THE SUPPLIER IN FEBRUARY 1988 FOR REINSPECT ORDER NO 5 FOR A DALE RESISTOR GENERIC PROBLEM THE CPM WAS REPAIRED ACCEPTANCE TESTED AND DELIVERED TO IFC IN JANUARY 1994 THE CPM HAD ACCUMULATED A APPROXIMATELY 14 HOURS OF *FUEL* CELL OPERATION WHEN *FUEL* CELL P760120 WAS DELIVERED THE CPM SUPPLIER POSTULATED THAT THE MOST LIKELY CAUSE OF FAILURE WAS A CPM SELF TEST CIRCUIT COMPONENT FAILURE A POWER SUPPLY CIRCUIT COMPONENT FAILURE WAS LESS LIKELY SINCE A POWER SUPPLY PROBLEM TYPICALLY AFFECTS ALL THREE CHANNELS AFTER COMPLETION OF STS 73 THE CPM S N T923 WAS REMOVED FROM THE *FUEL* CELL INSTALLED IN OV 102 AND WAS REPLACED WITH CPM P N 811046 1 S N J346 AN EXTERNAL VISUAL INSPECTION WAS PERFORMED ON THE REMOVED CPM AT THE SUPPLIER WITH NO ANOMALIES IDENTIFIED THE KSC REPORTED PROBLEM WAS CONFIRMED WHEN THE CPM WAS ENERGIZED WITH CHANNEL 1 OBSERVED TO HAVE AN OUTPUT OF 100 MILLIVOLTS WITH THE INPUT SET TO ZERO INPUT AND A SELF TEST OUTPUT OF 138 MILLIVOLTS THE OTHER TWO CHANNELS OF THE CPM HAD NORMAL OUTPUTS FOR ZERO INPUT AND FOR SELF TEST THE CPM WAS THEN PHYSICALLY OPENED AND BENCH TESTED ISOLATING THE FAILURE TO THE FREQUENCY TO VOLTAGE CIRCUIT U107A THE REFERENCE VOLTAGE LEVEL WAS LOWER THAN THE

Discovery and Systems Health Technical Area
NASA Ames Research Center - Computational Sciences Division

Contact: Dawn McIntosh, 650-604-0157
Dawn.M.McIntosh@nasa.gov

Summary

- The ReADS text mining work
- Using the ReADS text mining system on the toy dataset:
  - Only one document was identified by the experts and missed by ReADS.
  - On the other hand, ReADS found many interesting clusters which are possible Recurring Anomalies that the experts may wish to reevaluate.
  - Moreover, by identifying possible recurring anomalies the analysts can quickly focus in on the subset of documents worthy of their time and energy.
    - For the toy dataset of 344 documents, our worst case scenario meant the experts had to read ~208 of those documents (still saves the experts from having to read ~136 documents).
    - Our better scenario has the experts only having to read less than 118 documents – less than 1/3 of the size of the original dataset – a much more manageable set of reports to review!

References

Srivastava, A.N., et al., "Enabling the Discovery of Recurring Anomalies in Aerospace Problem Reports using High-Dimensional Clustering Techniques," IEEE Aerospace Conference, Big Sky, MT, March 2006.

Srivastava, A.N. and B. Zane-Ulman, "Discovering Recurring Anomalies in Text Reports Regarding Complex Space Systems," IEEE Aerospace Conference, Big Sky, MT, March 2005.

Banerjee, A. et al., "Generative Model-based Clustering of Directional Data," SIGKDD '03, Washington, D.C., August 2003.

The End

Thank you.