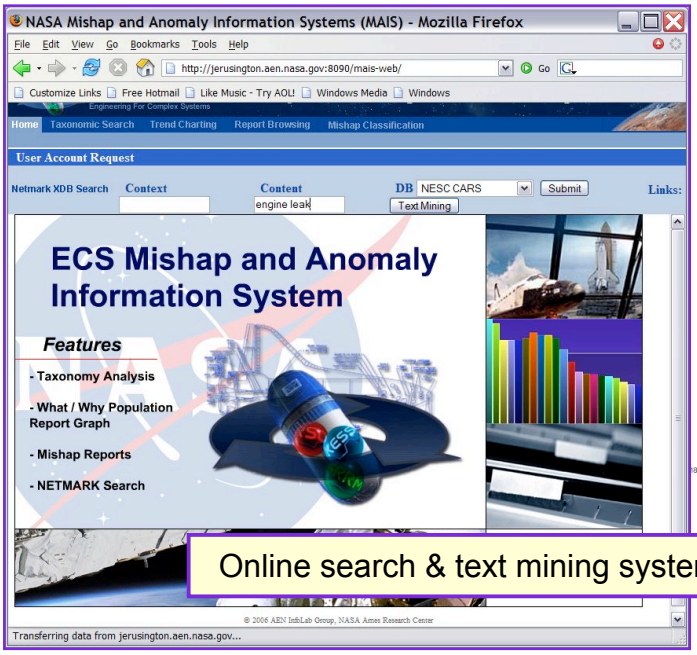
Clustering & Recurring Anomaly Identification: Recurring Anomaly Detection System (ReADS)

A tool to analyze text reports, such as aviation reports and maintenance records.

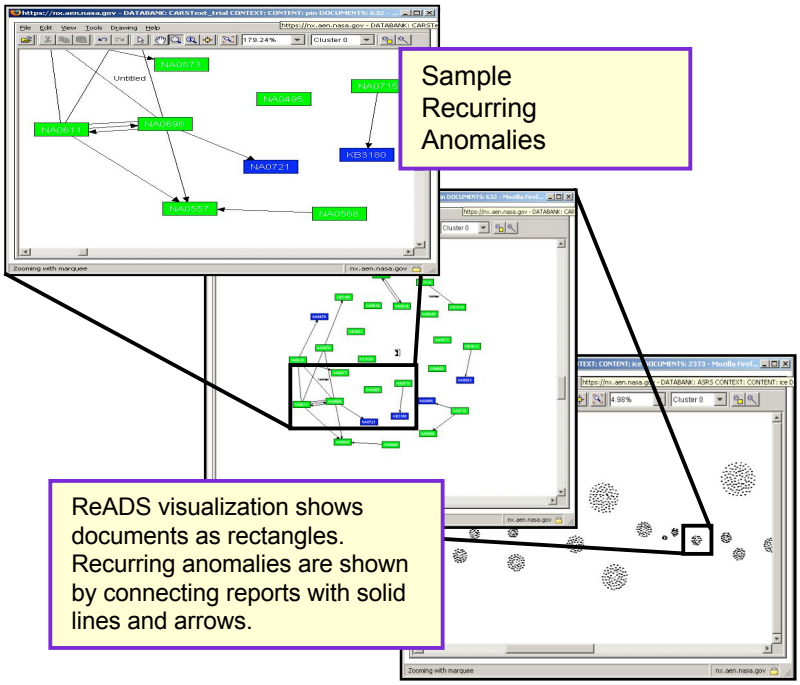
Problem Introduction

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



Research Overview:

The NASA Engineering & Safety Center (NESC) has commissioned our group to develop a tool to help them cluster NASA documents and identify recurring anomalies. ReADS can analyze text reports, such as aviation reports and problem or maintenance records. ReADS uses text clustering algorithms to group loosely related reports and documents, this reduces human error and fatigue. Plus, ReADS identifies interconnected reports; automating the discovery of possible recurring anomalies. ReADS provides a visualization of the clusters and recurring anomalies. ReADS has been integrated into a secure web-based search tool to allow uses to perform their own text mining.



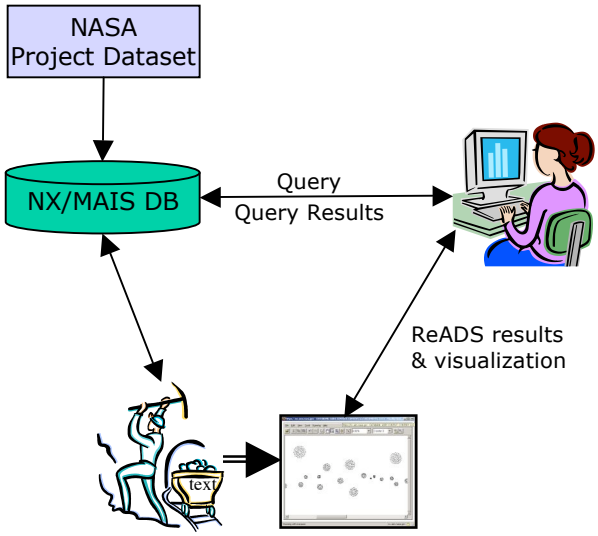
Recurring Anomaly Identification

ReADS identifies reports which mention other reports as a recurring anomaly using regular expressions to search documents and identify references of other reports by name. ReADS also detects recurring anomalies by determining the similarity between documents using a cosine distance similarity measure. Then according to the similarity measure, ReADS will run a hierarchical clustering algorithm to detect the recurring anomalies. The hierarchical tree is partitioned into clusters by setting a threshold. A low threshold implies that the reports must be very similar to be sorted into the same cluster.

Enabling discovery of anomalous trends in complex aerospace systems.

Goals
Our goal is to provide the NESC with a text mining tool that can be applied to any NASA dataset containing text. The tool will cluster the documents and provide a list of possible recurring anomalies, greatly reducing the amount of tedious effort required by the NESC subject experts.

Example of Application
Application Use Flowchart



ReADS - text mining

Evaluation of Recurring Anomaly Results
In the evaluation of the recurring anomaly identification results, we have three goals. The first goal is to catch all of the documents identified by experts as recurring anomalies. The second goal is to identify exactly the same results as the experts. The third goal is to find recurring anomalies missed by the experts.
When we used the ReADS text mining system on the toy dataset of 333 reports. We eliminated the need for experts to read approximately 60% of the dataset by identifying those reports as non-recurring anomalies. We found many interesting clusters not identified by the experts. These are possible recurring anomalies and false positives. Because review by experts is still necessary, false positives are not a large problem. Only one document was identified by the experts and missed by ReADS.

Application
The algorithms have been applied to two NASA-specific datasets, as well as multiple public datasets. Initial analysis of the results for one of the NASA datasets returned several recurring anomalies not found by the subject experts. A more thorough statistical result will be calculated this year.

Milestones
Unsupervised clustering using a modified version of the von Mises Fisher algorithm;
Recurring anomaly identification using agglomerative clustering applying cosine similarity.

Using ReADS:
ReADS has been integrated with a secure online search tool.

Mishap Anomaly Investigation System:
https://nx.aen.nasa.gov/textmining
- and login to access ReADS

Points of Contact:
Dawn McIntosh
NASA Ames Research Center
Moffett Field, CA
Telephone: 650-604-0157
E-Mail: Dawn.M.McIntosh@nasa.gov

Ashok N. Srivastava Ph.D.
Principal Scientist and Group Leader, Intelligent Data Understanding Group.
Telephone: 650-604-2409
E-Mail: ashok@email.arc.nasa.gov
Web: http://ti.arc.nasa.gov/people/ashok

Group Web Page:
datamining.arc.nasa.gov