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

Dawn McIntosh

June 21, 2006
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

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

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

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

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

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

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

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

Contact: Dawn McIntosh, 650-604-0157
Dawn.M.McIntosh@nasa.gov
## Shuttle CARS dataset → Toy Dataset

<table>
<thead>
<tr>
<th>Shuttle Corrective Action Reporting System (CARS)</th>
<th>Real Dataset (analyzed by experts)</th>
<th>Toy Dataset</th>
<th>Algorithm Results using Toy Dataset (similarity measure clustering threshold = 0.2)</th>
<th>Algorithm Results using Toy Dataset (similarity measure clustering threshold = 0.4)</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Documents</td>
<td>7440</td>
<td>344</td>
<td>344</td>
<td>344</td>
</tr>
<tr>
<td># of RA Clusters</td>
<td>366</td>
<td>20</td>
<td>RegEx: 28 &lt;br&gt;SimMeasure: 18</td>
<td>RegEx: 28 &lt;br&gt;SimMeasure: 33</td>
</tr>
<tr>
<td># of Total Documents in RA Clusters</td>
<td>1570</td>
<td>70</td>
<td>RE+SM = 92+56 = 118</td>
<td>RE+SM = 92+116 = 208</td>
</tr>
<tr>
<td>Min &amp; Max size of RA Clusters</td>
<td>Min = 2 &lt;br&gt;Max = 48</td>
<td>Min = 2 &lt;br&gt;Max = 10</td>
<td>Min = 2 &lt;br&gt;Max = 8</td>
<td>Min = 2 &lt;br&gt;Max = 9</td>
</tr>
</tbody>
</table>

**Discovery and Systems Health Technical Area**<br>NASA Ames Research Center - Computational Sciences Division

**Contact:** Dawn McIntosh, 650-604-0157<br>Dawn.M.McIntosh@nasa.gov
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.
  - ReADS Cluster: F, I, N, D
### Shuttle CARS dataset → Toy Dataset

<table>
<thead>
<tr>
<th>Shuttle Corrective Action Reporting System (CARS)</th>
<th>Real Dataset (analyzed by experts)</th>
<th>Toy Dataset (selected from Real CARS dataset)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Total Documents</td>
<td>7440</td>
<td>333</td>
<td>344-70=274, selected randomly from 7440-1570 non-RA reports.</td>
</tr>
<tr>
<td># of RA Clusters</td>
<td>366</td>
<td>20</td>
<td>Toy clusters selected to match, as much as possible, a variety of the types of RAs identified by NESC.</td>
</tr>
<tr>
<td># of Total Documents in RA Clusters</td>
<td>1570</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>Min &amp; Max size of RA Clusters</td>
<td>Min = 2, Max = 48</td>
<td>Min = 2, Max = 10</td>
<td>Toy Dataset RA clusters didn't cover the breadth of the cluster sizes, but the large clusters were rare.</td>
</tr>
</tbody>
</table>

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

Discovery and Systems Health Technical Area
NASA Ames Research Center - Computational Sciences Division
# ReADS stats on Toy Dataset

<table>
<thead>
<tr>
<th>Shuttle Corrective Action Reporting System (CARS)</th>
<th>Experts Results using Toy Dataset</th>
<th>ReADS Results using Toy Dataset</th>
<th>ReADS Results using Toy Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity Measure Clustering Threshold</td>
<td>NA</td>
<td>0.2 (documents must be very similar to qualify)</td>
<td>0.4 (a less conservative threshold)</td>
</tr>
<tr>
<td># of Total Documents</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td># of RA Clusters</td>
<td>20</td>
<td>RegEx: 28 SimMeasure: 18</td>
<td>RegEx: 28 SimMeasure: 33</td>
</tr>
<tr>
<td># of Total Documents in RA Clusters</td>
<td>70</td>
<td>RE+SM = 92+56 = 118 (note: There's overlap!)</td>
<td>RE+SM = 92+116 = 208 (note: There's overlap!)</td>
</tr>
<tr>
<td>Min &amp; Max size of RA Clusters</td>
<td>Min = 2</td>
<td>Min = 2</td>
<td>Min = 2</td>
</tr>
<tr>
<td></td>
<td>Max = 10</td>
<td>Max = 8</td>
<td>Max = 9</td>
</tr>
</tbody>
</table>

---

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

Contact: Dawn McIntosh, 650-604-0157  
Dawn.M.McIntosh@nasa.gov
Toy Dataset Confusion Matrix: 
ReADS similarity measure vs. Experts

<table>
<thead>
<tr>
<th>Subject Experts Recurring Anomaly Clusters</th>
<th>ReADS Recurring Anomaly Clusters</th>
<th>RAs missed by ReADS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1, 2, 3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2, 3, 4</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3, 4, 5</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4, 5, 6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5, 6, 7</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>6, 7, 8</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>7, 8, 9</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>8, 9, 10</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>9, 10, 11</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>10, 11, 12</td>
<td>11</td>
</tr>
<tr>
<td>11</td>
<td>11, 12, 13</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>12, 13, 14</td>
<td>13</td>
</tr>
<tr>
<td>13</td>
<td>13, 14, 15</td>
<td>14</td>
</tr>
<tr>
<td>14</td>
<td>14, 15, 16</td>
<td>15</td>
</tr>
<tr>
<td>15</td>
<td>15, 16, 17</td>
<td>16</td>
</tr>
<tr>
<td>16</td>
<td>16, 17, 18</td>
<td>17</td>
</tr>
<tr>
<td>17</td>
<td>17, 18, 19</td>
<td>18</td>
</tr>
<tr>
<td>18</td>
<td>18, 19, 20</td>
<td>19</td>
</tr>
<tr>
<td>19</td>
<td>19, 20, 21</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>20, 21, 22</td>
<td>21</td>
</tr>
<tr>
<td>RAs missed by Experts</td>
<td></td>
<td>22</td>
</tr>
</tbody>
</table>

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)

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

Recurring Anomalies (RA) Confusion Matrix

Algorithm clusters, clustering threshold = 0.2
ReADS RA clusters
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: Less conservative clustering threshold

Recurring Anomalies (RA) Confusion Matrix

Algorithm clusters, clustering threshold = 0.4
Clusters found using Regular Expression Matching

Contact: Dawn McIntosh, 650-604-0157
Dawn.M.McIntosh@nasa.gov
Preliminary Toy Dataset Results:
SimMeasure Algorithm Only (No RegEx Matching)

Recurring Anomalies (RA)
Confusion Matrix

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)

Recurring Anomalies (RA)
Confusion Matrix

○ Exact matches (Goal #2)
○ Algorithm clusters completely missed by Experts (Goal #3)
○ Expert clusters completely missed by SimMeasure algorithm (partial 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
ReADS System & Interactive Visualization

Web URL: http://jerusalem.aen.nasa.gov
Currently integrating w/ NX https://nx.aen.nasa.gov/nxpub

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

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

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


The End

Thank you.