Data Mining and Analysis

Kevin Samms
Kennedy Space Center
Major: Computer Science

Data Mining and Knowledge Discover
Spring Session
Date: 04 24 2015

Data Mining Development

Kevin O. Samms[1]
University of Central Florida, Orlando, FL, 32816

The Data Mining project seeks to bring the capability of data visualization to NASA anomaly and problem reporting systems for the purpose of improving data trending, evaluations, and analyses. Currently NASA systems are tailored to meet the specific needs of its organizations. This tailoring has led to a variety of nomenclatures and levels of annotation for procedures, parts, and anomalies making difficult the realization of the common causes for anomalies. Making significant observations and realizing the connection between these causes without a common way to view large data sets is difficult to impossible. In the first phase of the Data Mining project a portal was created to present a common visualization of normalized sensitive data to customers with the appropriate security access. The tool of the visualization itself was also developed and fine-tuned. In the second phase of the project we took on the difficult task of searching and analyzing the target data set for common causes between anomalies. In the final part of the second phase we have learned more about how much of the analysis work will be the job of the Data Mining team, how to perform that work, and how that work may be used by different customers in different ways. In this paper I detail how our perspective has changed after gaining more insight into how the customers wish to interact with the output and how that has changed the product.

Nomenclature

SAS	=	Enterprise Miner (statistical analysis software)
SME	=	Subject Matter Expert
Dashboard	=	A view of data in Tableau in which the view is composed of a number of smaller specialized views.
VBScript	=	A light version of Microsoft's programming language Visual Basic. The default scripting language in ASP (Active Server Pages)
P-value	=	A measure of significance of a term or factor to the overall statistical model

I. Introduction

The Data Mining project at NASA Kennedy Space Center (KSC) has moved through the stages of portal creation, visualization development, data analysis development, and customer feedback. It is in these last two stages that we currently find the project. The remaining work involves refinement of the data analysis method and continuing to collect feedback from the customers, the SMEs (subject matter experts), regarding the types of anomalies and issues they are experiencing and how those experiences correlate to the larger picture across the target domain.

Our project has primarily used four tools up to this point: SAS (statistical analysis software), Tableau Desktop (data visualization software), Tableau Reader (data visualization reader), and Microsoft Excel. Anomaly reports of various types have been entered into their respective databases and SAS is used to normalize and combine those data then analyze the result and generate output. The “data developer” then creates a view of that output using the Tableau Desktop data visualization software and uses Microsoft Excel for analysis if SAS is used in an additional stage of processing to output a spreadsheet of the data organized by topic. When SAS analyzes a report it considers the part of speech that a word fits into and considers the proximity of that word to other words. Whether a word becomes a keyword and what keywords are then used to generate a topic depends on how often that word is used, how close that word is to other analyzed words, and whether that word is a noun, proper-noun, adjective, noun-group, etc. A topic generated by SAS is not the single topic word that the reader of this report might expect. For software to sum up an entire document into a single descriptive word or sentence would probably require a human

[1] IT-G Intern, IT-G, KSC, and University of Central Florida.

level of reading comprehension, not to mention experience within the problem domain, and that is a technology that unfortunately does not yet exist. What SAS does is it scores words based on how often they are used and their part of speech then gathers the most important keywords, i.e. the highest scoring keywords, and strings them together into a comma separated chain that becomes the topic of that anomaly report. Many similar reports would be gathered under one topic generated based on their common contents. For example, "data, report, SAS, data mining, internship", could be a topic generated by SAS under which you might find this paper in a database of intern papers.

SAS uses a statistical approach to document analysis and this approach is important for its independence from human bias. It looks for what is there, not for what you want to find. We should take a minute to understand that part. When reading a document with a particular word (anomaly or concept) in mind, for example "reboot", that word may stand out to the analyst and there may be other computer terms in there as well. If the analyst is focused on this idea of computer reboots then just about anything they read that happens to contain those terms may cause the analyst to place that document under a "computer reboot" topic since it seems like that document has something to do with computer reboots. But perhaps in a more thorough and less biased reading it may turn out that the document has more to do with hard drive failures than the kinds of software anomalies that typically lead to a reboot of the computer as a solution. If the document mentions "hard drive" and other related problems to a greater degree than reboots then it is likely that the analyst has placed the document in the wrong category. In essence, the document gets categorized based on a bias of what the analyst wants to find rather than what is really there. It is assumed that an anomaly report will mention many terms but in the overall discussion, or description of the anomaly, the terms most frequently mentioned will highlight the overall intent and topic of the anomaly and therefore highlight to what topic the anomaly actually pertains. There is also the complication that an anomaly may belong under the heading of many different topics. The danger involved in a human categorization of documents is that bias may be unavoidable and we get subjective topics under which we get records that are similar but have no real underlying correlation. A statistical analysis, like that used by SAS, is better suited to look across many documents to find correlations between word usages in many documents and group those documents accordingly. Afterwards, the analyst can search the related documents for noteworthy trends to be further investigated for common parts, assemblies or conditions that identify a common problem or a common cause.

The documents as data become records in an Excel spreadsheet. As we got deeper into the analysis we ran into issues such as to how to categorize records, what keywords might belong to certain issues that we were searching the data for, how to efficiently search topics of interest when there were many records that fell under multiple topics thus leading to re-reading the same records multiple times, and how to determine what records were most relevant to a topic that is or might be of interest to an SME. The project evolved to solve these problems and change the output to the SME from a customized tool with which they would perform research to the research already performed and the output produced for their review.

II. A Change of Product

In the original process, Tableau Desktop was intended to create views of the data called Dashboards that would be customized to suit the needs and focus of a particular SME then output those views to them so they could use Tableau Reader with those views to conduct their own research of the database. Tableau Reader is free while the Tableau Desktop is prohibitively expensive to just casually install on the computer of each customer. Tableau Desktop is data visualization software meant to help solve the "Big Data" problem. It allows the data developer to manipulate how the underlying data is related, like creating multiple concurrent database queries, and then visually represent that relationship with graphics and colors that can expose commonalities and meaning in the underlying data. Reading hundreds or thousands of records to accomplish the same is practically impossible for a human. And using simple database queries is a one at a time endeavor that does not improve the situation much. The consequence as it pertains to NASA is that common causes of anomalies may then be missed because no one can or will read the many records that must be read in order to connect the dots.

Tableau Desktop for the data developer and Tableau Reader for the customer were meant to solve that problem and were doing quite well in the exploratory first phase of the project. However, there is a learning curve involved in using the reader and although the developers became very familiar with the tool, it was discovered that the end users of our output, the SMEs, would likely have neither the time nor enough collective interest in making the necessary investment in time and effort to learn how to use Tableau Reader. Their schedules were quite busy enough already. The project workflow had to shift from producing visualization dashboards for the SMEs to conduct their own research to using the tool ourselves and giving the end resulting output to the SMEs for their investigation by reading. This replaced our step of producing visualization dashboards with three new steps: find interesting trends

on our own, investigate those trends for common issues, and then produce output for the SMEs to read but break that output up into a short list of highly important records grouped by some sort of topic and a long list for further investigation if the SME desired a deeper look. Basically the meat of the investigation, the part where interesting common anomalies are found and output, would have to be done by the data developers, or the Data Mining team itself. The interesting anomalies are the ones trending upward, or happening frequently, versus trending downward, or happening less frequently. Upward trends are interesting because they are indicative of a problem that has not been solved or a problem that perhaps NASA does not have the work force to keep up with. Most of the second half of this project concerned finding the proper way of meeting the shift in workflow and refining both our methods and our output to the SMEs.

III. Phase II Data Analysis

At the beginning of this internship I began familiarizing myself again with Tableau, the data visualization tool used by our Data Mining team to search for and investigate anomalies. I watched online training videos to increase my proficiency in producing visualizations for the customer – the SMEs. This data analysis phase of the project was intended to create and customize the data visualization tool by which the SMEs would explore available data, recognize anomalies within their domain, and see connections to other anomalies that would help to expose root causes and paths to solutions. We planned training sessions to familiarize the SMEs with this new tool and the support options we offered towards helping them to get over the learning curve associated with its use. Quick reference sheets and guides were planned. However, opinions from the SMEs lead to an understanding that while new additions to the workforce may have an interest in learning the new tool, the veterans may not be convinced that adding this new tool to their workflow would be worth the investment in time and effort.

As a result we quickly chose a new route. Use Tableau ourselves to search for anomalies that were trending upward then dive into those trends and pull together related records to be output to the SMEs for their review. We divided up the work relative to the technical backgrounds of our Data Mining team members in order to better spot commonalities between anomalies. Whether electrical, engineering, or computer related we could spot connections between anomaly descriptions on technical or logical grounds. However, as the SMEs were primarily the ones to recognize connections, any descriptions that were not obvious connections but could not be ruled out as having no connection at all would be included in the output for SME review.

We began by aggregating anomaly reports under topics. We created those topics according to what issues we decided to investigate and what suggestions we received from the SMEs. The output from Tableau was sent to Microsoft Excel spreadsheets with the anomaly report data on one sheet and the visual analytics pasted onto another. These reports contained descriptions of the anomaly and its associated data. A typical anomaly report is in PDF form but from a database perspective each report, as it was in the Excel spreadsheet, consumed many fields of one row of the spreadsheet table and were thus thought of as records instead of reports. Our Tableau Dashboard, the configuration of visual tools used to search and analyze data, was developed to allow us to perform “OR” and “AND” type logical searches of the data based on keywords. The keywords themselves came from a combination of our own creativity, experience, and a SAS output that showed us all the different ways that a term is used across different databases and disciplines. The latter was referred to as the “terms tool” and it comes from SAS performing a word analysis and creating word clusters; words that are often used within close proximity to each other within a document. This SAS output was visualized in Tableau and used as a starting tool to help find key terms and ideas upon which to conduct a search. Using the terms tool and the Dashboard along with some creativity, we found issues worth investigating. We established a naming convention that would describe the terms that we used for a search and the topic to which that search pertained. A typical file under which records would be aggregated was “orSearch_Term1_Term2_Term3_andSearch_Term_TopicName.xlsx.” We could do keyword searches using 3 ORs and one AND per search.

IV. Search Work and Utilization

My background is in computers and software so my share of the work began with performing the terms search based on “software”. My mentor and the Data Mining Team Lead worked other parts of software and the engineering related searches respectively. Using the terms tool I researched the information brought up under the key terms generated by SAS, one-by-one. I also searched based on terms that I would create on my own based on what I saw in the data. Each key term that I clicked in the terms tool sorted its Dashboard view by that term and brought up related documents. I then read the anomaly descriptions that appeared in order to discern the overall idea

of what the anomaly was about. I noted the idea, specific terms that were repeated, and noted the words that came to me as I interpreted what appeared. I then used my notes and the search Dashboard of our data visualization of the database in order to find anomaly reports, or records. The OR search is one that takes 3 terms and searches all of the text in all of the fields of each record and pulls in a record if it has any of those terms. This is coupled with the AND search which takes only 1 term and pulls in only records that have that one term existing in all. This AND term is what I call a unifying term since an OR search on "air", "quality", "fail" will bring in any record that has "air" in it somewhere and similarly will bring in any record with "quality" or "fail" in it somewhere. However, entering an AND term of "monitor" means the "air" document must have both "air" and "monitor" in it; the "quality" document must have both "quality" and "monitor"; the "fail" record must have "fail" and "monitor". Thus, the AND term is the unifying term that will find records with "air", "quality", and "fail" that also include "monitor" in the discussion. It is now likely that we will records in which the "air quality monitor" device is being discussed along with the issue of having some sort of failure. This is how the search has been used to go from general search results, pulling in a lot of data, to a more focused search that pairs that data down to a more specific search. That search is then output to an Excel file so that the Description column for each anomaly record can be read row by row. We quickly discovered that even a paired down search can return a lot of data and reading through that data can be very time consuming. The team realized that we were going to have to use more technology in order to sort through and make sense of what we found. We knew that these first baby steps into sorting the data and producing useful results for the SMEs would be a learning experience.

We had teleconferences with a few SMEs for the purpose of getting ideas about how to use our capabilities for their benefit. One wanted a very low level look at the data, from an engineering perspective, that we could not provide since that level of detail does not exist in the data but he provided us with some good higher level use cases that would make this tool useful to him. He also provided us with some specific topics upon which to conduct a search. He seemed very enthusiastic about how this tool could help him and his team to be more productive and possibly make the connection between what he believed were an increase in anomalies and a series of design changes. He also brought up the topic of safety criticalities vs. mission criticalities. What was a critical item for providing safety to personnel may not be considered critical to completing a mission. Not to say that safety is not important but to highlight the fact that a broken safety item may prevent the use of that system but not bring an entire mission to a halt. Another SME was concerned about proving the case for needing more support in critical areas. He hinted that sometimes an increase in the number of anomalies, an upward trend, can indicate a lack of personnel necessary to keep up with known problems or expected anomalies such as regular patching and updates to software.

Overall we learned that the product of our work would be used differently and in more ways than we had originally thought. In any case we had to come up with an efficient process for investigating an issue brought to us by an SME.

V. Search Methods

A typical search using our early method would be to look at all of the different ways, for example, that the word "pump" is used by using the terms tool to look at the terms related to pump and noting the ones that turned up records when used in a search. Then note those terms on paper along with the most often used terms that showed up in the problem descriptions. Finally, use the search Dashboard by inputting terms that proved useful, and noting the associated number of records which contain all terms; the AND term as well as any useful OR search terms. This information along with the p-value, trend line, and years covered are output to an Excel spreadsheet. This term group could then be considered a topic and the file holding the associated records would be named after the main AND search term, the OR search terms, and the main subject of the search – "pump" for this example.

After this exploratory search the Dashboard is used to analyze trending for those topics that trend upward. Essentially searching each separate topic (e.g. the various usages of "pump") for common issues between the records under that topic for any common links. A topic has all of its records under one file, or "bin". All of the records in that bin had to be read and highlighted according to whether those anomaly reports shared a common link, were possible links, or could not possibly be a link perhaps because they had nothing to do with that topic. Each topic had some number of the latter. Some records would be captured within a topic simply because they contained keywords relative to that topic but upon reading, it would turn out that the record really had little to no applicability to the current topic. I eventually changed the method to reading through the records binned under a topic and highlighting them to pulling individual records from a topic and placing them in what I thought were the appropriate bins. If a bin did not exist then I created one that I thought was appropriate. This whittling down of a set of records

into their respective bins also proved time consuming since moving records from one Excel file to another was not simple and mostly required that the two files, source and destination, had the same column and field structure.

Each topic search can capture any number of records, from the very few to hundreds of records. It can take an entire day or more to read through that much information. Very painstaking work and very slow to find actual connections between anomalies with this method. We knew that we would need more technology to get to a more efficient workflow. Tableau is good for showing data and basic numerical correlations but finding correlations based on text; common problems, keywords, etc., requires more intelligent algorithms to correlate words and ideas. We found that using SAS to search and sort the data would be a better idea than doing that ourselves.

VI. SAS Adjustment

My mentor ran SAS on the database to produce another list of terms and their associated terms (i.e. word clusters). To improve SAS output the team made adjustments to the list that SAS uses to build synonyms, noun-groups, etc. The power of the statistical analysis performed by SAS is that it finds terms and associates them based on what is actually in the documents and how they are used. It could search for what we want or expect it to find by performing what is called a "supervised" search but that is for when we are looking for something specific. If we want to generate topics and capture their associated records then we assume no knowledge of the target data set and allow SAS to show us what is actually there, with no subjective prejudgments or assumed foreknowledge of what we think is there. The idea is to let SAS create the topics and "bin" the records, then after we know more about what is actually there we can perform more directed searches. This way subjective bias is mostly removed from the search itself and the output is an accurate reflection of the input. Some training is necessary however since SAS does not know jargon, or the nonstandard words that are particular to a discipline. SAS has to know *something* about the jargon and other words that are typically used in various disciplines but may have meanings not typically found in a dictionary. SAS may also piece together words on its own that we do not see as being related. So our modification to the file that SAS uses to define word types and connections was not a subjective interference upon a pure statistical process but a necessary enhancement to it.

While the SAS process was being prepared to begin taking more of the load for organizing topics, I continued work using the first method. Search topics, then deep dive into the ones that were trending upwards. While searching the topics created through the current process I realized that many records were showing up in multiple topics. At the time it was a surprise and manifested as reading and recognizing some of the same records over and over. Many of these repeat records rightfully belonged to multiple topics due to the fact that their contents could apply to multiple systems and concepts. So these repeat records would often have to be placed in multiple different bins. This compounded the time consuming nature of this first method of research and made it even more obvious that we had to develop something different.

VII. Topics and Relevance Scores

We started using SAS to search the database and generate topics under which it would place anomaly records and give them each a score that judged, statistically, how relevant that record is to that topic. This means that every record has a relevance score in every topic generated by SAS but those scores are very likely different for each topic and means that a record will have its highest score in some specific topic. So the problem of a record belonging to many topics is solved by simply considering the fact that its highest score guarantees that it does in fact belong to some specific topic. Searching for records and moving them from one bin to another is no longer necessary. The problem of having to place a record in multiple bins is also solved. Now we can just sort all the records in a topic, which is now a column in an Excel sheet, from highest to lowest relevance score and we can follow the scores downward to the point where the relevance score is higher in some other topic than it is in the current one. Thus we have a boundary at which one topic ends and another begins. This greatly cuts the time to produce an output of records relative to some topic of interest. A record still belongs to many topics, hence the difficulty in subjectively assigning a record to a topic, but statistically it is most appropriate to some specific topic. In this method the topics are also generated from the statistical analysis of the anomaly records, further removing the problems associated with subjective judgment. The analyst no longer has to make up topic names. But of course there is still more to it than that and we would continue to learn more and continue to improve this method as well.

There was the question of how many topics to generate under a particular search such as "pump." We can use SAS to search the database for everything to do with pumps (water pumps, fuel pumps, etc.) thus creating a data subset. Then SAS can generate 25 topics to 100 topics or more under that subset of data. A lower number of topics

and we get more records under a topic, however, many of those records may not have a lot to do with that topic. Too many topics and we get few records per topic, which means those records are definitely relevant to that topic but then we may lose correlation between records that fall out of that topic, due to a low relevance score, but are still related to records within the topic. In the latter case this means that we miss the correlation between anomalies that we were trying to find in the first place. The topics and records output by SAS were sent to an Excel spreadsheet where the records under each topic could be sorted from highest to lowest relevance score.

We tested various numbers of topics against the thoroughly human reviewed topics from the first research method in order to find the best number of topics to have generated by SAS. We came to the conclusion that 10 to 25 topics was best based on how many records there were in each topic, or each subset of some major search. It was important to note that this high level search had to be done first in order to see what is going on in the data before targeted searches could be done. A high level search results in topics, or bins of interest that can then be quickly analyzed with the visualization tool, Tableau, to find upward trending anomalies for investigation. A Targeted search (a search based on a specific device or anomaly) can then be performed based on an understanding of what we find that we *should* search for instead of what we *think* we should search for. Targeted searches can still be performed separately of the high level bin and investigate method but that is best done based on terms coming from an SME; an individual with experience in the domain under search. An SME targeted search means that a search is still done with an understanding of what is happening in the problem domain rather than being done under the subjective speculation of someone conducting an unfocused search of the data for related problems. It could be said that an SME directed search is a supervised search on par with a statistical search by SAS in that both are a function of the realities of the different problem domains rather than being a function of speculation.

VIII. New Method of Investigation

Just as with the earlier method, this new method of using SAS to do most of the heavy lifting in terms of creating topics and binning records would also evolve as it was used. In my first try at the new method I read records near what we thought would be a good relevance score to ensure that records belonged to the current topic and were good for a “first look” at what was occurring within that topic. I created a separate file where I recorded notes about each topic and what I found. This output to my mentor would be the basis of the decision to cap an output to the SME at a particular relevance score and thus a particular number of records. The idea was to generate a first look output that would not be an overwhelming amount of reading for the SME. If they wanted more after that first look then additional output could be given or a deeper investigation performed. In my output to my mentor I described what was going on above and below a certain relevance score and then highlighted the notes that contained my recommended baseline. We eventually changed that method to creating a 25 record boundary and evaluating the content of a record to determine whether it truly belonged to that topic and whether the boundary should be extended to include it, thereby extending the relevance score boundary as well.

For the 25 record boundary modification to the method, my notes included recording the highest relevance score in a topic (the score of the first record in that topic) and the relevance score at the 25 record boundary. After the records were studied I included the recommended baseline in terms of the number of records and the relevance score at that number of records. I also included important snippets from the record, an area to describe how the number of primary terms are changing throughout the topic, and a notes area briefly describing my overall decision to set a boundary at some point or to point out any significant observations about the anomaly occurrences within the topic. After the first few major topics I also began color coding snippets based on whether they were linked by concept or linked by a reference to another record.

IX. Twenty Five Record Boundary

Our second evolution of the new method was to begin with a target of the 25 most relevant documents within a title. This gave me the task of checking the records just outside of that limit to see if some of those records needed to be included within the limit, thus extending the limit past 25 to whatever I found that it needed to be. Most topics tend to have at least that many records and if it had less records then they would simply be output since, in that case, few records may be related to each other but are still a good first look output for an SME to see the top 25 or less issues most relevant to a topic of interest. At first this process seemed pretty simple; read the records just outside of the 25 record boundary for any links to records within the boundary and extend the boundary as necessary. Records that seemed to represent the same overall idea I called concept-links and records that were elevations of other records I called record-links. An example of an elevation is an anomaly documented at a low level type of anomaly

report then documented at a higher level where more personnel are involved in finding a solution. The higher level report refers to the lower one by number and may contain the same information along with more details. In that case both records should be included in the output for the SME to see that an elevation of the anomaly occurred. The concept-links are important as they could be a second or third occurrence of an anomaly which is obviously important to see where anomalies are reoccurring.

As mentioned earlier, under this new method the records under a topic are sorted from high to low relevance which makes it easier to read through them and find correlations. Not to mention all of the other problems from our first method that are now solved in this new method. These records are output from SAS into an Excel sheet and the topic terms (remember topic terms are word clusters) are color coded wherever they appear in the record. So the topic words in the column header are color coded and wherever any of those words appear in the document they are also color coded. The color coding was done with a tool that was a user form created by my mentor in VBScript. This color coding serves to both show how terms would drop-off as the relevance score decreases and to make it easier to read the records for correlations. Using the color coding tool, other terms found could be colored as needed to help find where they appeared in other records.

X. The Running Boundary Problem

A couple of new problems were also discovered during this new method of investigation. What I called the "running boundary" problem and the "diffuse boundary" problem. The running boundary came about as I read through records at the 25 record boundary and upwards. I found that the concept-links tended to be close together, usually less than 10 records apart, while the record-links could be many tens of records apart; sometimes hundreds. Due to the nature of the conversation going on in a record it could statistically belong to a different topic while having a record-link with something in the current topic being investigated. This lead to the following problem: concept-links outside the 25 record boundary are close together causing some necessary extension of the boundary to include those records, however, the newly included records may be record-links jumping far away from the current position, thus including more concept-links that may not be appropriate for a first look, such as a 4th occurrence or a first occurrence outside of what we intended to be a first look. Any additional record-links found within those newly included records could again move the boundary far from its current position. I detailed this problem to my mentor and offered a solution.

Running Boundary

Context
A 25 record boundary is established for a "first look" into the 25 most relevant anomalies under a particular topic. This first look will be output to the SME for their examination.
This **boundary** is **not** extended, *instead* a second boundary is created for any records that need to be included with the 25 record first look output, based on the sanity check described below.

Sanity check
Some records outside the 25 record boundary (i.e. the 26th to the 30th record) are read to see if the anomaly presented in any of those records has any conceptual link to - or is an elevation of - any record that exists inside of the 25 record boundary (i.e. the 25th record and lower).

Conceptual link defined:

1. A record considered to be a second or third occurrence of some anomaly mentioned or discussed in some other record that is the first occurrence of the anomaly.
2. A record mentions or discusses a [possible] remedy to some anomaly (this record must be included in the output regardless of how many occurrences of the anomaly are already being output for the first look).

Elevation defined:

1. Any record that has a corresponding MOD AR, PART IFI or PART PRACA within that topic. Call this a record-link.

- For example: if record #20 is an MOD AR and record #36 is the corresponding PART IFI then this is a record-link, and both records need to be included in the output to the SME (with limitations*).
- * Since record #20 is within the 25 record boundary, and thus is part of the first look output, record #36 *becomes* part of the first look and thus establishes the second boundary.

Boundary *Limitations

Conceptual link
Records must be included in the output only up to the 3rd occurrence of an anomaly. Additional occurrences are not necessary for a first look.

Elevation
Records linked to each other by elevation are only included if one of them is conceptually linked to a record within the 25 record boundary or if one of them already exists within the 25 record boundary.

Boundary extension problem, and a possible fix
If a second boundary is established due to a record-link then any records in-between the second and first boundary are of course included in the output, but if any of those records in turn have a record-link to another record that is *outside* of this new second boundary then they get included, thus extending the boundary further. If each boundary extension contains more links then we get a running boundary.

Running boundary problem
RA_1 = the 25th record, and all records below 25.
Some record, RA_2 is just outside the first boundary (of 25 records) and is conceptually linked to some record *inside* the first boundary or it has a record-link to some record *inside* the first boundary.

Therefore we create a second boundary so that RA_2 can be included in the output.
Thereby our output = 2nd boundary + 1st boundary

Some record, RA_{21} is between this new second boundary and the first boundary (of 25 records)

Regardless of whether it is conceptually linked to some record under the first boundary or not, it is included in the output by default since it is within the second boundary.
However, it also has a record-link with some other record that exists *outside* of the second boundary.
This creates a third boundary.
Thereby our output = 3rd boundary + 2nd boundary + 1st boundary

Some record, RA_{32} is between this new third boundary and the second boundary

Regardless of whether it is conceptually linked to some record under the second boundary or not, it is included in the output by default since it is within the third boundary.
However, it also has a record-link with some other record that exists *outside* of the third boundary.
This creates a fourth boundary.
Thereby our output = 4th boundary + 3rd boundary + 2nd boundary + 1st boundary

This could continue on until we exhaust all records that belong to the topic (records with their highest relevance score in this topic and not in some other topic).

This is the problem that I think I am running into and I believe the solution is to consider only a first boundary (of 25 records) and a second boundary. The second boundary is created by records *outside* of the first boundary that are conceptually linked to records *inside* the 25 record boundary (up to a maximum of the 3rd occurrence) or the second boundary is created by a record that mentions a solution to some anomaly inside the 25 record boundary. Any records found in-between the second and first boundary that have a record-link to a record outside of the second boundary are **not** considered as a reason to create any new boundaries.

Basically if the output needs to be increased beyond 25 records then we do so based on conceptual links to records within the 25 record boundary but pick up any record-links only if they are linked to or conceptually linked to the 25th record or a record **below** the 25th record.

To this I later added that record-links *should* be considered as a reason to create new boundaries but with the provision that record-links did not move the current boundary too far away from its current position. I believed this would prevent a running boundary and prevent the inclusion of too much data that was off topic or not appropriate to a first look output. The method was thus updated again to consider this running boundary problem.

Another issue I encountered was a diffuse boundary. I had thought that the boundary at which one topic ends and another begins was definite. Upon further investigation I discovered that a topic would have a boundary that was definite only up to a point after which the topic would have records belonging to other topics interspersed with its own records. For an example of this, consider a topic of "electrical, relay, amp" having all records from 1 to 35 that have their highest relevance scores in that topic column. It seems the boundary ends at 35, however, we have records 36, 39 and 40 that have their highest scores in another topic, "current, pump, switch" whereas more records belonging to the initial topic, "electrical, relay, amp", are found at 37, 38, and 41. Thus the boundary for the initial topic is not definite. It actually becomes diffuse at some point. To remedy this issue and to speed up the process of determining a proper boundary I used the highlighting tool to find all of the record-links first, then find where the boundary became diffuse. I then used the records in that diffused area as a basis to pick a point outside of the 25 record boundary and begin investigating for concept-links. I read from a distance downward towards the 25 record boundary instead of reading from the boundary upwards. Establishing a high level view of the record-links and boundary was another modification to the method that helped to speed up the process of finding an appropriate cutoff point for a first look output to the SMEs.

XI. Conclusion

We knew at the outset that this would be a learning experience and that we would have to develop procedures along the way. There were no prewritten procedures or a model for us to follow. This endeavor was new to our customers as well, hence the expected enthusiastic response to the prospect of what we could do for them but the unexpected response at the prospect of having to adjust to using a new tool. With so many unknowns there were bound to be surprises that would necessitate flexibility on our part as well as an ability to step back from the details and keep an eye on the big picture. I believe that as a team we learned the necessity of having a backup plan if the customer wants a significant change in the product. We also learned to use all of the available technology to do most of the work where possible and to keep looking for greater efficiency.

Regarding efficiency, we learned that it is important to avoid getting buried in the details. One has to step back from the details and perform a sanity check every so often. Always ask, is there a better way to do this, do the results match the reality, does the procedure produce results that are close to an ideal model? In our case the ideal model did not exist, so we worked with what we thought would be that ideal model from a manual sorting and binning of records. This is the procedure that we started with and thought would be correct albeit time consuming. We thought it best to use the technology to replicate the results of the manual procedure but do it at a greater speed. This served to actually expose the subjectivity in the manual procedure and changed the goal to improving the manual procedure instead of just replicating it. The benefit of the manual check now was to verify that the output of the technology produced reasonable results and that those results were as close to an ideal as possible. We learned that the ideal model was something that had to be developed along the way.

In the end we produced the output desired; the most relevant anomalies under topics of interest to an SME as well as the ability to perform targeted searches based on the reality of what is in the data. This project closes temporarily where the current procedure needs more refinement at the point of establishing the proper baseline for the output under a topic. The replacement of this manual work seems to require better artificial intelligence algorithms or an augmented procedure for the current SAS output. The only remaining goal for a future phase of the project seems to be the refinement or creation of algorithms that can establish that baseline at least nearly as well as a human.