



Chen, Lin and Nayak, Richi (2007) A Case Study of Failure Mode Analysis with Text Mining Methods. In Ong, K.-L. and Li, W. and Gao, J., Eds. *Proceedings 2nd International Workshop on Integrating Artificial Intelligence and Data Mining (AIDM 2007)* CRPIT, 84, pages pp. 49-60, Gold Coast, Qld..

© Copyright 2007 Australian Computer Society
Copyright © 2007, Australian Computer Society, Inc. This paper appeared at the Second Workshop on Integrating AI and Data Mining (AIDM 2007), Gold Coast, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 84, Kok-Leong Ong, Junbin Gao and Wenyuan Li, Ed.
Reproduction for academic, not-for profit purposes permitted provided this text is included.

A case study of Failure Mode Analysis with Text Mining Methods

Lin Chen

Faculty of Information Technology
CRC for Integrated Engineering Asset Management
Queensland University of Technology, Brisbane

l33.chen@student.qut.edu.au

Dr Richi Nayak

Faculty of Information Technology
CRC for Integrated Engineering Asset Management
Queensland University of Technology, Brisbane

r.nayak@qut.edu.au

Abstract

The maintenance dataset provided by SunWater contains information about failed assets also known as components and their corresponding failure modes. Currently, extraction of this information from the dataset been conducted in a manual manner, which is very tedious, time consuming and cumbersome work. It is necessary to discover an automatic method to decide/extract the failure mode. This paper presents three methods that were attempted in an effort to solve this problem. The performance of each method is analysed in detail and suggestions for how the outcomes can be improved are also proposed.

Keywords: Text Mining, Clustering, Semantic Network, Association Rule Mining, pump failure mode.

1 Introduction

With the civilization of human society, text has been the most common way to store all types of information, communication, ideas, stories and knowledge. However, because of the limitation of a person's reading speed, it is impossible to grasp the key information in a short time when there is a large amount of text involved. Text mining is one approach to deal with large amounts of unstructured data. It facilitates extraction of useful and important concepts from the text data and categorizes the information contained within.

This paper presents a case study detailing the use of text mining methods for extracting useful information from the maintenance dataset that is collected by SunWater. SunWater is a leading company providing water infrastructure and supply solutions and services to customers throughout Queensland, Australia and the world. http://www.sunwater.com.au/about_overview.htm

SunWater has a data set that records the maintenance activities. This dataset implicitly contains information about failed assets also known as components and their corresponding failure modes. Information about the various failure modes, which include pump failure, valve failure, motor failure and cooling system failure, can be derived from this SunWater maintenance data set.

Identification of the failure mode and even predicting the type of failure in future can give extensive economic benefit to SunWater. However, all the failure modes are currently decided manually from the dataset. To change the manual decision process, automatic decision processes are proposed through the use of text mining technology. The optimal results for the analysis of failure mode should be as follows. The failure mode could be drilled down into sub failure modes. If it is pump failure, which type of pump failure should the record belong to? Table 1 shows the details of the pump failure mode.

Failure Mode	Sub Failure Mode
Pump	Pump
	Pump bearing
	Pump Seal
Valve	Suction
	Discharge
Motor	Motor
	Motor Bearing
	Motor Slipping
	Motor winding
Cooling water system	Cooling Water System
	Cooling Water Pump
	Cooling water pump bearing
	Cooling water motor
	Cooling water motor bearing
	Cooling water heat

Table 1: A list of Failure Modes

Learning on the given data set falls into the category of unsupervised learning, as instances in the given data set have no labelling of failure modes. Due to the lack of labelled data with failure modes, this project uses unsupervised learning approaches such as clustering, association rule mining and semantic networks to solve the problem faced.

The task here is to identify the key terms that reflect the failure modes as listed in Table 1. If a record contains a group of keywords reflecting a failure mode, the record can automatically be identified as a maintenance record

that can further be utilised in analysis or seeks immediate attention. In this paper, we utilise various clustering and semantic network based methods to come up with a group of terms that indicate a particular failure mode. We also utilise association rule mining to list rules that would indicate that the records, having terms of these association rules, are maintenance record as a result of failure, not as a result of routine inspection.

The data set can be divided into three groups according to the maintenance records belonging to each of the power stations namely Awoonga, Bocooolima and Wooderson. Table 2 summarises the number of unique terms without and with stemming and with and without stop word removal in the dataset. Naturally, the use of stopword removal and stemming techniques reduced the number of terms to be processed, but even when utilising such methods, the number of terms still exceeded 1600 for a typical pump station. This shows that these datasets are highly dimensional and there exists a need of utilising text mining methods to identify terms indicative of failure mode. With such dimensionality it can be difficult to find similarities or associations between data entries, as there are so many different terms.

Station	Stopword Removal	Porter Stemming	Total No. of terms
Awoonga	True	True	2180
	True	False	2494
	False	True	2613
	False	False	2927
Bocooolima	True	True	1814
	True	False	2052
	False	True	2128
	False	False	2363
Wooderson	True	True	1607
	True	False	1820
	False	True	1892
	False	False	2102

Table 2: Number of unique terms in the datasets

The structure of paper is as follows. The second section will introduce the text mining methods that are used in solving the problem. In the third section, results of various methods will be presented and discussed further. Finally, the potential improvements for the research will be suggested in conclusion.

2 Text Mining Methods

Text Mining is discovery of new, previously unknown information, by automatically extracting it from different written (text) resources. The key element is linking of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation (Hearst, 2003). Text mining includes the steps of processing the input text, deriving patterns within the newly processed data and finally the evaluation and interpretation of the output. Major study areas includes text categorization, text

clustering, and concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization and entity relation modelling.

The SunWater case study is a typical text mining problem. Some of the fields are structured, such as Functional Location, Priority, System Status, Main Work Centre, Reported By and Notification Date. But unfortunately, this data is mostly an identifier and is not useful in deciding the pump failure mode. At the same time, there exist unstructured fields such as Short Description, Equipment Description, Damage Text, Cause Text and Long Text. The unstructured data can be more helpful in determining the pump failure mode. "Short Description" and "Long Text" are the descriptions for the pump failure events. "Damage Text" and "Cause Text" may also be helpful for detecting the type of failure. But identification of the failure mode is not the traditional text categorization problem. No training data (instances labelled as failure modes) is provided or available. Therefore, it is impossible to decide which category the record should belong to. Clustering techniques, on the other hand, do not need a training data set. Semantic Network as well as Association Rule Mining are unsupervised learning techniques. These three methods are potentially suitable for this text mining task. Below are details of these methods and the software that we used in analysing the SunWater maintenance data.

2.1 Clustering

Two types of clustering methods and tools are used namely, Ward's agglomerative method with SAS Text Miner and the Similarity Histogram Clustering method. The major difference between these two methods is that Ward's method is a hierarchical approach that produces a dendrogram of the dataset and thus the output can change at different levels of similarity. An agglomerative approach starts with each document being in its own cluster (the leaves of the dendrogram) and the repeatedly merges them together until a stop condition is met. It can be argued that a hierarchical approach produces nested clusters (due to the structure of the dendrogram but at the same time does not actually produce any clusters, but rather a tree-based representation of the data. The Similarity Histogram Clustering method (SHC) is a partitional approach. Partitional approaches do not produce a dendrogram, but rather cut the dataset into partitions at a single level. This makes them more desirable than hierarchical approaches when large datasets are involved. The biggest downside to partitional approaches is that they usually require the user to specify the number of clusters desired, which can be difficult to know in advanced. The SHC approach however does not require the user to specify the number of clusters desired and thus avoids this problem.

2.1.1 SAS Text Miner

The SAS Text Miner software is used for text analysis (SAS Institute, 2003). The process of SAS Text Miner includes converting unstructured text into structured data,

clustering the documents/data entries and viewing the concept links. Conversion from unstructured text to structured data allows to convert unstructured text into structured numeric statistical summaries of the elements of a document and to convert unstructured text into frequency distributions of words. Stop words such as “and”, “the”, “is” are removed. Some other words, which appear frequently but don’t contain much meaning or are not related to the results in the context, are also removed. Some examples of these words are “Bocoolima”, “Awoonga” (the pump station name), “ARMS”, “AWC”, “ACP” (short name for the pump).

The singular value decomposition (SVD) is used in reducing the dimensions of the input data set. SVD is a powerful technique for automatically relating similar terms and documents, eliminating the exhaustive need to manually generate industry specific ontologies or synonym lists (SAS Institute, 2007). The whole dataset includes a total of 843 records. Each record contains information about the stations name, long text and short text which describes the action taken after failure. Each record has the terms ranging from 1 to 50. If SVD is used, the complexity can be reduced. The function rollout term can reduce the dimensionality by taking the n highest weighted terms.

Clustering is performed according to the content of the text data. During the clustering process, several clustering algorithms are used. Spatial clustering techniques are served to maximize expectation. In SAS Text Miner, the documents are grouped into hierarchical clusters. There exist parent and child relationships among documents. Ward’s agglomerative method (Ward, 1963) makes the hierarchical clusters possible.

2.1.2 Similarity Histogram Clustering

The idea to use histogram clustering is to provide to the expert an automated way to trace/determine the failure mode. If the ratio of the similarity of records is high enough, record are deemed to belong to the same cluster. The similarity is based on a pair-wise similarity score between documents/data entries (Hammouda & Karnel, 2003). The concept for the similarity histogram clustering is as following. The histogram is composed of a number of predefined bins. Each bin corresponds to a similarity value interval. Each bin has the count of the number of pair-wise similarities for that cluster that fall within the associated interval. To make the similarity histogram work, high cohesiveness needs to be achieved through building and maintaining a concise statistical representation of the cluster using the pair-wise similarities of the documents within the cluster. When a new record is being added into the dataset, it is also possible for records already assigned to be reassigned, in which a record already in one cluster could be moved to another if both associated clusters would benefit from such reassignment (Hammouda & Karnel, 2003; Shaw, 2006). This paper uses the two versions of similarity histogram clustering: one is original (Hammouda & Karnel, 2003) and another one is enhanced similarity

histogram clustering using Intra centroid vector similarity (Shaw, 2006).

2.1.2.1 Original Histogram Clustering

The contents of the bins should be skewed towards the high similarity values in order to achieve highly a coherent cluster. When a record is to be added to an existing cluster, its pair-wise similarities with the documents already in the cluster are added to the histogram. The change in the histogram will then determine if the document is to be added to that cluster or not. If the record causes the histogram degrade too much or drop below a pre-specified threshold then the document cannot be added.

The algorithm for the implementation of the original SHC algorithm is as follows:

```

1: L ← Empty List {Cluster List}
2: for each document D do
3:   for each cluster C in L do
4:     HRold = HRC
5:     Simulate adding D to C
6:     HRnew = HRC
7:     if (HRnew ≥ HRold) OR ((HRnew
      > HRmin) AND (HRold - HRnew
      < ε)) then
8:       Add D to C
9:     end if
10:  end for
11:  if D was not added to any cluster then
12:    Create a new cluster C
13:    ADD D to C
14:    ADD C to L
15:  end if
16: end for

```

Figure 1: Original Histogram Clustering

The algorithm calculates the histogram ratio by using the records that exists in a cluster before the addition of a new record and after the addition of the new record. The after histogram ratio is calculated by simulating the addition of the document to the cluster. Then the new ratio is compared with the old ratio. If the new ratio is less than the old ratio, but the degradation is less than predefined value and the new ratio is above a predefined minimum value/threshold, then the document can be added. If the new ratio doesn’t meet the minimum value/threshold or the degradation is too much then document cannot be added to that cluster. If there is no cluster for which a document can be added to, then a new cluster is generated and the document is added to the new cluster.

2.1.2.2 Enhanced Similarity Histogram Clustering Using Intra Centroid Vector Similarity (ESHC-IntraCVS)

The difference between the original version of histogram and ESHC-IntraCVS lies in taking into account the centroid vector of a cluster in the enhanced implementation. The centroid vector is a vector space

model (VSM) representation of the centre or average of the cluster and the documents it contains. Unlike the original Histogram algorithm, which places the document in the cluster that would receive the best histogram ratio change, this algorithm puts the new document into the cluster that has the most similar centroid vector to the document. The performance of this algorithm should be better than that of the original algorithm by the theory that the cluster with the highest similarity to the document will have the greatest number of similar documents in it and would be the best cluster to place the document in. The cohesiveness of the clusters can be maintained and any tendency for a cluster to spread out over vector space can be limited. As a result, it should give more tightly packed clusters, which are more distinct from each other with minimal overlap between them. In the test conducted by Shaw (2006), the overall average F-Measure was improved by 6.3% by using ESHC-IntraCVS. The detailed algorithm of ESHC-IntraCVS is as follows.

```

1: L ← Empty List {Cluster List}
2: for each document D do
3:   for each cluster C in L do
4:     HRold = HRC
5:     Simulate adding D to C
6:     HRnew = HRC
7:     if (HRnew ≥ HRold) OR ((HRnew > HRmin)
        AND (HRold - HRnew < ε)) then
8: SimDC = Cosine similarity between D and CV
        of C (without updating to include D)
9: store details in List (P)
10:   end if
11: end for
12:   order List P in decreasing SimDC
13:   take the first entry in List P and
        determine C
14:   Add D to C (which was the first entry in
        List P)
15:   if D was not added to any cluster then
16:     Create a new cluster C
17:     ADD D to C
18:     ADD C to L
19:   end if
20: end for

```

Figure 2: ESHC-IntraCVS

The new and old ratios are decided in the same manner as the original SHC approach. If the new histogram ratio is all right, then the similarity between the new document and the cluster centroid (or average) is calculated. A list stores the similarities in decreasing order for each cluster. By going through all the clusters and checking/comparing, the document will go to the cluster, which is most similar to it. The document then is added to the first cluster in candidate list. If no cluster is similar to the new document, the document is assigned to a new cluster.

2.2 Semantic Network

Semantic features can be used in check the text coherence through the notion of “isotopy”. The notion behind this is the recurrence within a given text section of the same semantic feature through different words (Dutoit & Poibeau, 2002). The recurrence of these features throughout a text allows extraction of the topic(s) of interest and some other points that are marginally tackled in the text. It provides an interesting way to glance at the text without a full reading of it and it also helps in the interpretation.

Leximancer is a software program used in the experiment for extracting the main concepts, which are generated from seed words, contained within documents/data entries. It has been used in several error tackling experiments. In order to analyse the problem of a lack of situation awareness by mariners, Leximancer is introduced for use. It is reported that the results of manual coding and Leximancer on 26 reports revealed a very close proximity between hit rates (Smith, et. al., 2002). In another article (Smith, 2006), Leximancer provides decent performance which was validated by doing stability, reproducibility, correlative, functional test/s. Leximancer is reported in an earlier article (Smith, 2003) that it can even tell the optimistic and negative tones.

No training data is needed in Leximancer since it is an unsupervised ontology discovery. The results provide analysis of direct or indirect associations of the concepts. It might help in the decision making process for the domain expert. If the pump failure mode is related to words or concepts such as electric, flood; then the existence of electric and/or flood means pump failure to certain extent (Smith, 2003).

In this experiment, Leximancer is used to grasp the global context and the significance of the concepts. Avoiding fixation on particular anecdotal evidence is another use. The important terms are selected based on word frequency and co-occurrence usage in the unified body of text. These terms later are used for classifiers. Then the classifiers extend the seed word definitions and finally form the concepts. Between the concepts, links represent the relationships and relatedness of any two concepts. In the experiment, “Functional location Description”, “Short Description”, “Damage Code, Damage Text”, and “Long Text” are the attributes for extracting important concepts. Functional location Description provides the information about the location of the failure mode. Short Description and Long text usually contain the most critical information about the failure event.

2.3 SAS Association Rule Mining

The purpose of association rule mining is to discover elements that co-occur frequently within a data set. It is largely used for marketing purposes. Buying product A, what is the percentage of the customers who likely to buy product B? Similarly, using term A, how likely is the term B used in the same record? For example, the terms “Recirculation Pump fails” are related to the term

“Bearing fails”. Pump bearing failure is one of the failure modes. By discovering the term associations, it helps the experts to determine the failure mode much easier. In this experiment, the researcher hopes to find out the relations within and get the percentage rate of these relations.

SAS Association Rule Mining utilizes the Apriori algorithm (Agrawal, et. al., 2003) to discover the important rules. This means that each record (row) should be deemed as one transaction. The text in each record should be separated into separate terms wherein each term represents as an item. To implement the rule mining, all transaction ID’s and the terms from different records will appear in the binary styled table. If the term exists in the particular record, then “1” will appear in the binary table. Therefore, it is necessary to separate the sentence to individual words before importing the data to SAS for association rule mining. The stop word removal and stemming approaches are used to pre-process the text. The repeated occurrences of the same word are noted as the existence of term in the record.

3 Analysis and Results

3.1 SAS Text Miner Results

In SAS Text Miner, “Long Text” from three stations is set to “Target”. All other attributes; “Equipment Description”, “Damage Code” and “Damage Text” are set as “Input”. The clustering in Text Miner uses SVD (singular value decomposition) for dimension reduction and an agglomerative clustering method that facilitates automatic grouping of documents into taxonomies in hierarchical style. As a result, the whole dataset is clustered into 13 categories or clusters. Table 3 lists the details of the clusters.

The ID is the cluster identification in Table 3. Descriptive Term is the representative terms of the cluster which has been chosen by SAS. Freq is the number of documents in the cluster. Coverage is the number of documents in a cluster divided by the total number of documents in the collection. For example, the record which “Long Text” contains “*Additional Work----RWR 150 SAP 2021946 Instructions: Recommission security system (unauthorised removal of key) est Materials \$5.00est Labour \$125.63est Plant \$11.16est. Overheads \$74.530ther ARMS Details----Centr*”, belongs to cluster 1 since it has the descriptive terms “sap, work, overhead, plant, labour”. Any term that is preceded with “+” means that it is a stemmed term and can be expanded to include all of its original terms. For example “+provide” in cluster 3 means that “provides”, “providing”, “provided” all belong to the cluster 3’s descriptive terms. The two largest clusters are cluster 7 and cluster 12, which have 13% of records in each cluster. The descriptive terms for cluster 7 are “new”, “replace”, “with”, “faulty” and “remove”. Cluster 12 includes the terms “line”, “find”, “flag”, “fault” and “reset”.

Some of the results are quite impressive. For example, when cluster 8 is viewed (as listed in Table 4) from the contents in the “Long Text” field (only “Long Text” is

displayed, “short Text”, “ID” and “Station” etc are ignored), the main idea for the records in this cluster can be derived and is about the replacement of network cable.

However, some of the results do not seem to be so accurate. One record has long text “*Slipping burnt. Spare slipping fitted. Sent for repairs and slipping repaired*”. In another record, long text includes “*isolate motors. Remove the covers on the heat exchanger tube. Cleaned the tubes with wire brush and flushed out. Replace covers and turned motors to service*”. Both records belong to cluster 7. But the first record depicts “slipping burnt” as a problem. The second record describes a “tube” problem. The two problems appear to be different, however, have been put together in a cluster due to the commonality of the terms being used for description of the problem. A close observation of clustering terms also indicates that there are no clusters about the cooling water system in the text miner results.

ID	Descriptive Term	Freq	Coverage
1	sap, work, overhead, plant, labour	30	4%
2	+valve, +work, as, +order, no	59	7%
3	station, + provide, between, project, see	18	2%
4	details, arms, centre, work, completed	73	9%
5	slip rings, carbon, +ring, + brush	23	3%
6	centre, maintenance, type, code: location, other	80	10%
7	new, +replace, with, faulty, +remove	106	13%
8	input, +transmitter, low, temperature, type	13	2%
9	+motor, dowdoing, +test, mills, hv	38	5%
10	purpose, fire, +pump, project, +station	62	7%
11	vibration analysis, analysis, +carry, vibration, manager	56	7%
12	+line, +find, +flag, +fault, +reset	108	13%
13	+order, +work, compltd, +complete, +instruction	64	8%

Table 3: Text Miner Clustering Results

The SAS Text Miner also lists the terms, their frequency, and their importance in clustering process. Table 5 shows an example of the statistics of the terms after stopword removal and stemming have been applied. “Keep” denotes whether the term will be kept or not for the SVD and clustering. If the term is not kept, then the term is not important. In this experiment, a total of 65 terms are deemed as not important (Keep = N). These terms do not frequently appear in different records. Therefore the weights of these terms are low and are deemed not worthy of being kept. The SAS Text Miner uses tf-idf (term frequency-inverse document frequency) to evaluate the importance of the terms. Term frequency denotes the total number of occurrences a term has in a record. Inverse document frequency denotes the fractional number of records that includes the given term

in comparison to all records in the data set. Role means which part of speech such as noun, verb, etc the term belongs to.

See also notification 10044145.This project comprises the replacement of cables that provide a communication network between each of the PLCs in the three pumpstations.Awoonga Station has,
See also notification 10044145.This project comprises the replacement of cables that provide acommunication network between each of the PLCs in the three pumpstations.Awoonga Station has,
SEE ALSO NOTIFICATION 10050992See also attached notifications. This project comprises the replacement of cables that provide acommunication network between each of the PLCs in the three pum
See also notification 10044145.This project comprises the replacement of cables that provide acommunication network between each of the PLCs in the three pumpstations.Awoonga Station has,
See also notification 10044145.This project comprises the replacement of cables that provide acommunication network between each of the PLCs in the three pumpstations.Awoonga Station has,
See also notification 10044145.This project comprises the replacement of cables that provide acommunication network between each of the PLCs in the three pumpstations.Awoonga Station has,
See also notification 10044145.This project comprises the replacement of cables that provide acommunication network between each of the PLCs in the three pumpstations.Awoonga Station has,
See also notification 10044145.This project comprises the replacement of cables that provide acommunication network between each of the PLCs in the three pumpstations.Awoonga Station has,
See also notification 10044145.This project comprises the replacement of cables that provide acommunication network between each of the PLCs in the three pumpstations.Awoonga Station has,
See also notification 10044145.This project comprises the replacement of cables that provide acommunication network between each of the PLCs in the three pumpstations.Awoonga Station has,
See also notification 10044145.This project comprises the replacement of cables that provide acommunication network between each of the PLCs in the three pumpstations.Awoonga Station has,
See also notification 10044145.This project comprises the replacement of cables that provide acommunication network between each of the PLCs in the three pumpstations.Awoonga Station has,
See also notification 10044145.This project comprises the replacement of cables that provide acommunication network between each of the PLCs in the three pumpstations.Awoonga Station has,

Table 4: Records in cluster 8

Term	Freq	# Doc	Keep	Weight	Role
+actuator	2	2	Y	0.897	Noun
+address	6	6	Y	0.734	Verb
+affect	2	2	Y	0.897	Verb
+alarm	20	19	Y	0.566	Noun
+allow	4	4	Y	0.794	Verb

Table 5: Terms

As to the testing of the clustering results from text miner, it is recommended to let the domain experts decide the accuracy of the clustering. The domain experts can readjust the setting of the SAS Miner until they get the desired results. When they find the terms in the results are not important, they can reset “Y” to “N” for the “Keep” attribute. Text Miner also allows the user to reassign the weight of the terms. The domain experts need to decide which terms are more important than other terms and then reassign the weight to improve the accuracy. The evaluation of the results of text miner is only available when there is testing data. More functions

in text miner could be used for future work. By using a “Memory-Based Reasoning” node, “Data Set Attributes” and/or “SAS Code” to get the Precision and Recall ROC chart.

3.2 Similarity Histogram Clustering Results

For this clustering experiment, several parameters need to be mentioned, as they are user provided. One is the Minimum Similarity Threshold and it dictates whether the pair-wise similarities between two entries (such as an existing entry in the cluster and a new entry being added) contribute positively towards the histogram ratio. If a pair-wise similarity is above the minimum similarity threshold, then that score will increase the ratio when it is recalculated. If it is below the threshold then it will cause the ratio to decrease. It is related to and has a direct effect on the cohesiveness of the cluster. The similarity threshold should be given before clustering. Another parameter is Minimum Histogram Ratio which is the minimum value every cluster must maintain its histogram ratio above. Its function is to stop clusters completely degrading and lacking cohesiveness, thus resulting in poor quality clusters. The last parameter is Maximum Degrade per Data Entry. This specifies the maximum degradation ratio allowed when a new data entry is added to a cluster. Thus stopping entries degrading the quality of a cluster by too much and too fast is its main function.

Due to the pair-wise computation between each pair of records, this method is quite expensive. These cannot take a combined dataset including the maintenance records of the three pump stations, namely Awoonga, Bocoollima, and Wooderson as the inputs. That is why, in these experiments, maintenance data of each station is handled separately.

Experiments conducted here used 0.1, 0.2, 0.3, 0.4 and 0.5 for the Min Similarity and Min Histogram Ratio. There were 503 separate tests and files generated containing the results with a combination of different options (different Min Similarity / Min. Hist. Ratio value pair, with or without stop word removal, with or without porter stemming) across the three datasets, each of which is a separate pump station (Awoonga, Bocoollima and Wooderson).

The evaluation of the results of these clustering algorithms is made difficult by the fact that there is no information provided to indicate which category a data entry belongs to (thus no training data). Without this knowledge it is impossible to evaluate the quality of the clustering using the traditional recall, precision and F-measure techniques as all of them rely on external knowledge. This external knowledge is lacking in this situation. Thus internal measures must be used. The first is to simply use the number of clusters built and compare that result against the number of categories as suggested by the SAS Text Miner (which is 13). The second is to measure the cluster self-similarity scores for each cluster. This is done by simply averaging the pair-wise scores that are contained within a given cluster. The self-similarity for each cluster can then be averaged to determine the average cluster self-similarity score for the

clustered dataset. The third technique is to determine the cluster cohesiveness for each cluster and average this across all the clusters. The method chosen to determine the cohesiveness was the weighted similarity of the internal cluster similarity and the following formula (Steinbach, et. al., 2000) outlines how the cohesiveness for a cluster was determined.

$$\frac{1}{|S|^2} \sum_{\substack{d \in S \\ d' \in S}} \text{cosine}(d', d)$$

Min Sim. / Min. Hist. Ratio	#Clusters with Stopword Removal and No Stemming	#Clusters with Stopword Removal and Stemming	#Clusters with Stopword Removal But No Stemming	#Clusters with Stopword Removal But Stemming
0.5/0.5	44	139	40	151
0.4/0.5	33	100	28	96
0.3/0.5	26	54	22	54
0.2/0.5	14	24	13	24
0.1/0.5	6	11	8	6
0.5/0.4	42	133	38	146
0.4/0.4	31	90	28	89
0.3/0.4	25	43	22	44
0.2/0.4	11	9	10	12
0.1/0.4	5	8	12	4
0.5/0.3	40	127	38	134
0.4/0.3	31	81	28	78
0.3/0.3	23	37	22	41
0.2/0.3	8	7	7	11
0.1/0.3	3	3	8	3
0.5/0.2	40	111	38	116
0.4/0.2	31	72	28	62
0.3/0.2	19	28	20	31
0.2/0.2	4	9	5	4
0.1/0.2	3	2	2	3
0.5/0.1	40	78	38	94
0.4/0.1	31	34	28	45
0.3/0.1	11	18	20	9
0.2/0.1	4	4	4	4
0.1/0.1	3	2	2	3

Table 6: Clusters for Bocoilima Station using SHC

Tables 6 and 7 show some of the results achieved using the SHC and ESHC-IntraCVS algorithms with term frequency used during the building of the vector space model. Overall the enhanced algorithm (ESHC-IntraCVS) seems to perform slightly better than the original SHC from the view of building a smaller number of clusters for a given dataset.

Results in Tables 6, and 7 also show the effect of employing the text pre-processing in clustering. Results

show the number of clusters produced when either or both of the stopword removal and stemming techniques was included. The higher the Min. Similarity / Min. Hist. Ratio, the higher quality clusters generated in terms of cohesiveness. However, the clusters are stricter and therefore tend to try to be perfect. This can result in a large number of small (contain only a handful of entries) tight clusters being generated for a dataset. An example is first line of Table 6, where the number of clusters is very high due to production of highly cohesive clusters. The Low Min Similarity/Min Hist. Ratio makes the clustering process easier as the clusters are less strict and entries can more readily be added to a cluster. However, this has the side effect of reducing the cohesiveness of the clusters thus making them looser or low purity.

Min Sim. / Min. Hist. Ratio	#Clusters with No Stopword Removal and No Stemming	#Clusters with Stopword Removal and Stemming	#Clusters with Stopword Removal But No Stemming	#Clusters with No Stopword Removal But Stemming
0.5/0.5	44	39	41	41
0.4/0.5	33	28	27	31
0.3/0.5	25	21	25	21
0.2/0.5	14	11	14	12
0.1/0.5	8	8	8	8
0.5/0.4	41	37	38	39
0.4/0.4	31	28	25	31
0.3/0.4	24	18	21	19
0.2/0.4	11	9	10	11
0.1/0.4	5	4	5	5
0.5/0.3	39	37	38	39
0.4/0.3	31	28	24	31
0.3/0.3	20	15	19	21
0.2/0.3	8	6	7	8
0.1/0.3	3	2	3	3
0.5/0.2	39	37	38	39
0.4/0.2	31	28	26	31
0.3/0.2	14	8	12	19
0.2/0.2	4	4	4	4
0.1/0.2	3	2	3	3
0.5/0.1	39	37	38	39
0.4/0.1	31	28	22	31
0.3/0.1	6	4	4	5
0.2/0.1	4	4	4	4
0.1/0.1	3	2	3	3

Table 7: Clusters for Bocoilima Station using ESHC-IntraCVS

In Table 6 and 7, the number of clusters is smaller when stopword removal is implemented than when stopword removal is not used. This is due to the removal of common words that are common and thus reduces the dimensionality of each entry. Also, the number of

clusters generated when stemming is used is also smaller than when stemming is not used. Stemming reduces words/terms to their common stem and thus reduces the number of unique words for each entry and can reduce the pair-wise similarity scores due to their being fewer unique terms being matched between entries.

From the results obtained, the Minimum Similarity value has the biggest effect on not only the number of clusters produced but also the self similarity and cohesiveness of those clusters. However, with a high Minimum Similarity value many small clusters are produced as opposed to having a low Minimum Similarity value producing few large clusters. The challenge is therefore to find the most suitable value for the Minimum Similarity to achieve the best outcome possible and this can vary from one dataset to another.

Min Sim. / Min. Hist. Ratio	No Stopword Removal and No Stemming		Stopword Removal and Stemming		Stopword Removal and No Stemming		No Stopword Removal but Stemming	
	CSS	CC	CSS	CC	CSS	CC	CSS	CC
0.5/0.5	0.58	0.17	0.59	0.20	0.58	0.20	0.57	0.17
0.4/0.5	0.49	0.18	0.53	0.13	0.50	0.13	0.51	0.19
0.3/0.5	0.40	0.11	0.39	0.08	0.42	0.11	0.37	0.08
0.2/0.5	0.30	0.06	0.30	0.03	0.30	0.09	0.28	0.07
0.1/0.5	0.29	0.03	0.30	0.08	0.28	0.02	0.32	0.04
0.5/0.4	0.56	0.17	0.59	0.20	0.56	0.20	0.56	0.17
0.4/0.4	0.49	0.18	0.53	0.13	0.50	0.13	0.51	0.19
0.3/0.4	0.38	0.11	0.34	0.11	0.34	0.08	0.34	0.08
0.2/0.4	0.28	0.04	0.27	0.02	0.23	0.04	0.29	0.06
0.1/0.4	0.20	0.00	0.28	0.04	0.24	0.00	0.29	0.00
0.5/0.3	0.55	0.17	0.59	0.20	0.56	0.20	0.56	0.17
0.4/0.3	0.49	0.18	0.43	0.13	0.49	0.13	0.51	0.19
0.3/0.3	0.28	0.04	0.32	0.09	0.26	0.00	0.35	0.07
0.2/0.3	0.25	0.05	0.19	0.00	0.18	0.03	0.26	0.09
0.1/0.3	0.16	0.00	0.13	0.00	0.16	0.00	0.18	0.00
0.5/0.2	0.55	0.17	0.59	0.20	0.56	0.20	0.56	0.17
0.4/0.2	0.49	0.18	0.53	0.13	0.50	0.13	0.51	0.19
0.3/0.2	0.22	0.06	0.22	0.05	0.17	0.00	0.37	0.12
0.2/0.2	0.20	0.02	0.18	0.00	0.17	0.03	0.21	0.05
0.1/0.2	0.16	0.00	0.13	0.00	0.16	0.00	0.18	0.00
0.5/0.1	0.55	0.17	0.59	0.20	0.56	0.20	0.56	0.17
0.4/0.1	0.49	0.18	0.53	0.13	0.51	0.13	0.51	0.19
0.3/0.1	0.15	0.00	0.21	0.05	0.15	0.00	0.21	0.05
0.2/0.1	0.21	0.04	0.18	0.00	0.18	0.03	0.23	0.05
0.1/0.1	0.16	0.00	0.13	0.00	0.16	0.00	0.18	0.00

Table 8: Average cluster self similarity and cluster cohesiveness for Bocoilima Station using ESHC-IntraCVS

3.3 Semantic Network Results

Using the default settings in Leximancer is particularly useful for those who are not familiar with the text data on hand and is similar to the Grounded Theory approach, in which the important themes in the text are meant to appear from the text alone, independent of the observer (Smith, et. al., 2002).

The first experiment feeds the dataset from the three stations as free text for the semantic analysis. In the pre-processing step, the common stop words are removed and the terms that contain the same stems are merged into one word. The list of important concepts is showed in Table 9. The results generated by Leximancer include important concepts containing: pump, suction, valve, discharge, motor, bearing, and spurling. Concept “pump” is more important than “bearing” which is identical to the theory that bearing class belongs to pump class. Similarly, the concept “valve” gets a higher percentage than the concept “discharge” and the concept “suction” in that discharge and suction are subclasses of valve. It is interesting to see that the concept “cooling” is deemed as moderate related (37.4%) to this whole dataset. The concept “water” also appears in the results. Usually the term “code” would not be deemed important due to its common use. The concept “code” is selected as the most important concept based on its frequency counts in the data set and due to its significance...The “code” here reflects the kind of work that has been done in that particular maintenance record. The code number is implicitly related to a failure mode; however, a non-expert cannot infer the relationship. For example, J00458BIL is the one of the technical codes.

Each concept in table 9 is related to other concepts, which appear in the concept lists. Table 10 is a list of the concepts related to the concept “discharge”. From Table 10, relativeness of concept “valve” to concept “discharge” and “actuator” are quite high. From the raw dataset, we find out the terms “discharge” and “valve” and “actuator” are often used together. In the raw dataset, there is a sentence in the “Long Text” field to illustrate the situation: “*Investigated alarm on SCADA pump locks out at Awwonnga pump station. Travelled to Awoonga to reset flags and check discharge valve. Returned to collide*”. In the “Equipment Description” field, it normally shows the text that *discharge valve and actuator* when the term “discharge” appears in the field.

Figure 5 shows the associated concept map. The bigger the circle on the map the more important that theme is compared to other themes. Theme is the representative concepts amongst the collection of concepts. From the concept map, we can tell the theme “pump”, “code”, and “replace” are important themes. Themes can be overlapped. One concept can belong to multiple themes because of the relativeness between concepts. Click concept node on the concept map and it will show the related concepts by linking them together. Figure 5 also shows the concept links when the concept “valve” is chosen. The shorter of the concept link means the more relevant the other concept is to the selected concept. The distance from “valve” to “charge” is shorter than the distance from “valve” to “cooling”. So the relationship

between “valve” and “charge” is closer than the relationship between “valve” and “cooling”.

Concept	Absolute Count	Relevant Count
code	219	100%
pump	182	83.1%
purpose	128	58.4%
motor	117	53.4%
text	115	52.5%
replace	112	51.4%
work	105	47.9%
cooling	82	37.4%
Flowmeter	79	36.10%
provide	65	29.6%
pstn	56	25.5%
inspection	53	24.2%
project	45	20.5%
valve	43	19.6%
damage	43	19.6%
discharge	32	14.6%
closed	32	14.6%
actuator	31	14.1%
condition	30	13.6%
contractor	29	13.2%
water	25	11.4%
upgrade	24	10.9%
cables	20	9.1%
switch	17	7.7%
bearing	15	6.8%
thermocouple	14	6.3%
temperature	14	6.3%
conditioner	13	5.9%
batteries	11	5%
voltage	11	5%
supplied	10	4.5%
satisfactory	9	4.1%
suction	9	4.1%
exchanger	8	3.6%
heat	8	3.6%
Isolated	7	3.1%
Pumpstat	6	2.7%
Maintenance	6	2.7%
sparling	6	2.7%
ventilation	6	2.7%

Table 9: Ranked concepts

Since small themes can be contained within bigger themes, such as the “valve” and “discharge” themes should be under “replaced” themes (Figure 5). For example, one record contains “Flowmeter, ults bestobell sparing Electeq investigated fault. Reset 425 u/v flags due to induced surge. Test complete. Control supply reinstated program logic contr pump No1 replaced plc batteries discharge valve and actuator program logic” in the “Long text” tuple. Since the terms “replaced”, “discharge” and “valve” appear together, this record can

be regarded as belonging to the “replaced” theme. By using this characteristic, small themes can be classified into big themes. All the records should belong to 11 themes at the end, if each big circle counts for 1 theme as figure 5 shows. This would help reducing the number of themes (clusters).

Concept	Absolute Count	Relevant Count
discharge	32	100%
valve	30	93.7%
actuator	29	90.6%
code	20	62.5%
work	17	53.1%
pump	16	50%

Table 10: Related Concept with “Discharge”

Semantic Network identifies the important concepts in experiment. It discovers the unnoticed relationship between the concepts. The distances between the concepts can be found out from the links on the concept map. Small themes can be included in big themes.

3.4 Association Rule Mining Results

For the SAS association rule mining, the minimum support threshold is set as 4% for minimum transaction frequency to support associations. The reason to set a low threshold is that some important words that reflect the failure mode are not frequently occurring in the whole dataset. For the given problem, it is valuable to discover the association rule of the infrequent terms if any of the failure modes (Table 1) appear in the rule. In general, Support and Confidence are the most commonly used indicators to evaluate association rules. Low support or low confidence rules should not be used. A low support and/or low confidence means the rule is not seen or not occurring often and could be a random event or coincidence. However, in the proposed problem, even a rule does not have enough supporting records, but if it contains the terms reflecting the failure mode, the rule can be deemed as important.

A total of, 65536 association rules are generated due to the lower support and confidence threshold. An example of generated association rule is pump→maintenance. The rule states that if the record contains the term “pump” then it is high likely that the record will also contain the term “maintenance”. These rules become interesting if any of the rules include the terms that indicate the pump failure modes as listed in Table1. For example, rules such as pump→maintenance and maintenance→pump indicate that records supporting these rules may be the records that indicate the “pump failure mode”.

For example a subset of rules that include the term “cooling” are extracted to find out the pump failure mode of “cooling water system”. Some of rules that were discovered include the following:

Water→cooling, pump→cooling, cooling→clean, pipework→cooling, pig→cooling, lines→cooling, remove→cooling, debris→cooling, several→cooling, refilled→cooling, drained→cooling, clear→cooling, blew→cooling, waterpump→cooling, vaccum→cooling, transformer→cooling, thermographic→cooling, report→cooling, minor temperature→cooling, imagining→cooling, flow→cooling, current→cooling, circuit→cooling, cables→cooling, breaker→cooling, addressed→cooling, cooling→abnormalities, service→cooling.

The rules that contain the key words that appear as the pump failure mode are direct indicative of the rules importance. Examples are water→cooling and pump→cooling. However these rules do not assist in finding other associative terms appearing with the failure mode terms. These associative terms, in turn, can further be utilised in identifying the maintenance records as failure of a component rather than a routine maintenance. The rules such as debris→cooling, cables→cooling, drained→cooling, clear→cooling, temperature→cooling are more interesting. Terms like debris, cables, drained and clear do not directly reflect that they indicate a failure mode. However the association of these terms with a term (that is a failure mode) indicate that the presence of these terms in a record might give a trace about the failure mode. For example, if there is a coexistence of cable and cooling in records, then they are possibly an indication of a cooling water motor problem. Or if temperature→cooling, then the records which contain temperature and cooling are clustered as cooling water heat failure mode.

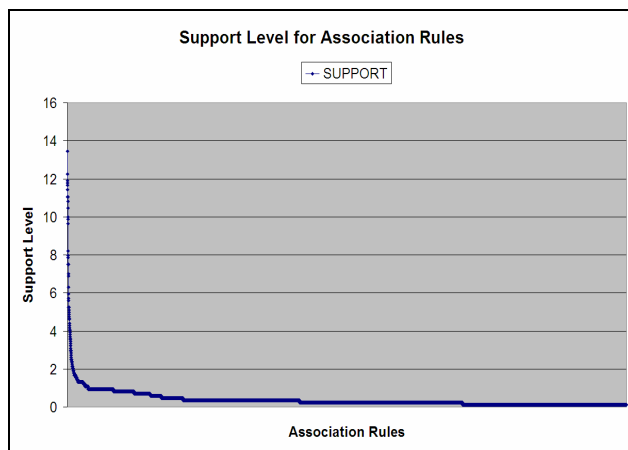


Figure 3: Support

The figures above and following (3 and 4) are the support and confidence for the association rules discovered in this experiment. These figures show that most of the records have a low support and low confidence. From figure 3, only the first ten to twenty rules have a support above 10%. The rest of rules are near to 1%. In figure 4, only 20% or fewer points reach the confidence over 50%. The confidence of most points ranges from 0 to 30%.

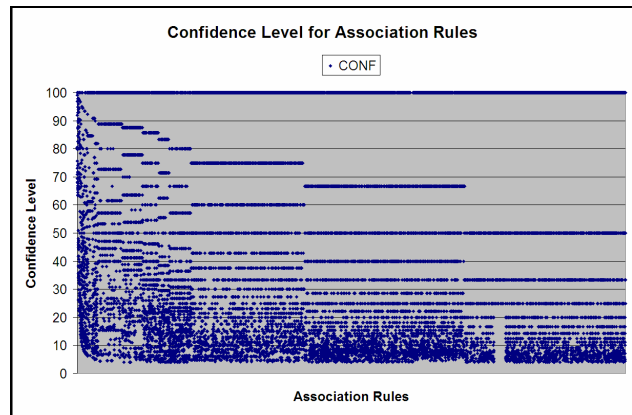


Figure 4: Confidence

Since in this section, we are interested in the term associations of failure modes, rules with low support and low confidence still reflect the terms that may indicate the cause of a failure mode.

4 Discussion and Conclusion

Three methods have been employed to analyse the SunWater dataset in order to classify the failure mode for each record. SAS Text Miner makes use of the traditional text mining technology to analyse the data. Stop word removal, stemming, and same word as different part of speech are used in the parsing process. SAS Text Miner distils key concepts and analyses relationships between isolated terms or phrases. Therefore, the results are not just the collection of keywords. It identifies the cognitive terms or concepts. Experiments with SAS reveal that 13 clusters exist in the data set including maintenance records of all three pump stations. There are some impressive results that identify the network cable failure successfully. At the same time, it can't identify the cooling system failure. As for Histogram clustering methods, two types of clustering methods were implemented and there were 503 separate results from the various conditions tested. Results significantly varied according to the parameter settings. The number of clusters can be varied from 2 to 151. Due to the scalability problem, datasets from three stations can't be combined into one dataset. Leximancer seems to present better solutions. In the results, it contains the anticipated outcome. Terms such as "pump", "bearing", "motor", "cooling" appear as the concepts in the results. The concept map presents the relativeness between concepts in the straightforward way. If the small themes can be included into big theme, then it is much easier to find out the right cluster. But the condition is that the big themes are identified correctly and as the requirement. Association Rule mining was also used in an attempt to find out the relations between important terms. However, the experiment results in poor performance. Most confidence and support levels appearing in results are low. Further concerns should be given to finding the relevant rules giving the tolerance to the low support and low confidence.

Several lessons can be learnt from the 3 experiments.

- The application of tf-idf to short documents that only contain a few sentences doesn't work as well as expected. SAS, which uses tf-idf, fails to identify some important concepts or terms such as cooling water system. The reason is that the important terms, while appearing infrequently across the dataset often do not appear frequently within those few entries that contain them. Thus the tf-idf score is reduced due to the low tf component.
- A Semantic Network approach might be quite useful in terms of identifying important concepts. It utilizes the co-occurrence of terms/concepts to find the important terms/concepts. Using the weights generated from semantic network software might be another way to improve the clustering results.
- It is hard to evaluate unsupervised learning. The precision and recall, which are only for classification, cannot be used for clustering, as there is no prior knowledge or training data available. Cluster self-similarity and cluster cohesiveness are a good option to evaluate the results of the pair-wise clustering when this external information is lacking.
- Stop word removal and stemming is an important pre-processing step. Learnt from the figures of CSS and CC, the performance without stop word removal and stemming is worse than when stop word removal and stemming are used.

The performance would be better, if more data is provided. "Long Text" and "Short Description" are the tuples to describe the failure event. However, the failure events about the pump station are not just written up by one person. Therefore, the terms to describe the failure events are quite different. If there were some standard to describe the failure events, the performance of the experiments would be better as well. If some terms and cluster relationships are known/available as domain knowledge and set up before the experiment, the performance/outcome would also be improved and should be closer to the desired results. For example, when there is short term "recirculation pump fails/failure" in "Short description" or "Long Text", then the failure event is always related to the failure mode "pump bearing problems".

As for future work, it would include obtaining a dataset of failure modes, which have previously been categorized by human experts making it possible to compare automated approaches against the manual approach, as well as each other. Such comparisons could be done using standard evaluation measures such as recall, precision and F-measure. Secondly, identifying the big themes correctly in semantic network should be carried out in order to accurately and correctly include the smaller themes into them.

Acknowledgements

Thanks to SunWater for the funding of the research. Thanks to the support from Professor Colin Fidge and Associate Professor Lin Ma. Thank you to Gavin Shaw for providing the source code for the SHC and ESHC algorithms.

5 References

- Agrawal, R., Imielinski, T. and Swami, A. (1993): Mining association rules between sets of items in large databases. *Proc. ACM SIGMOD International Conference on Management of Data*, Washington DC, USA, **22**:207-216, ACM Press.
- Dutoit, D. & Poibeau, T. 2002. Inferring Knowledge from a large semantic network. <http://arxiv.org/ftp/cs/papers/0501/0501072.pdf> Accessed 27 May 2007.
- Hammounda, K.M. & Karnel, M.S. 2003. Incremental Document Clustering Using Cluster Similarity Histograms. In *Proceedings of the IEEE/WIC Web Intelligence 2003*, 597-601.
- Hearst, M. 2003. *What is Text Mining?* <http://www.ischool.berkeley.edu/~hearst/text-mining.html>. Accessed 20 April 2007.
- SAS Help and Documentation. 2003. USA. SAS Institute Inc.
- SAS Institute: SAS Text Miner 3.1 Capitalize on the value hidden in textual information. <http://www.sas.com/technologies/analytics/datamining/textminer/factsheet.pdf>. Accessed 4 May 2007.
- Shaw, G. 2006 Enhancing an Incremental Clustering Algorithm for Web Page Collections. (Honours diss., Queensland University of Technology).
- Smith, A. E., Grech, M., and Horberry, T. 2002. Application of the Leximancer Text Analysis System to Human Factors Research. <http://www.leximancer.com/documents/hfes2002.pdf>. Accessed 20 April 2007.
- Smith A. E. 2003. Automatic Extraction of Semantic Networks from Text Using Leximancer. <http://www.leximancer.com/documents/hlt2003.pdf> Accessed 20 May 2007.
- Smith A. E. Humphreys, M. S. 2006. Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. <http://www.leximancer.com/documents/B144.pdf> Accessed 28 May 2007,
- Steinbach, M, Karypis, G., & Kumar, V. 2000. A Comparison of Document Clustering Techniques. In *Proceedings of Knowledge Discovery in Databases Workshop on Text Mining 2000*. Boston, MA, USA.
- Ward, J. H. 1963. Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association*. 58: 234-244.

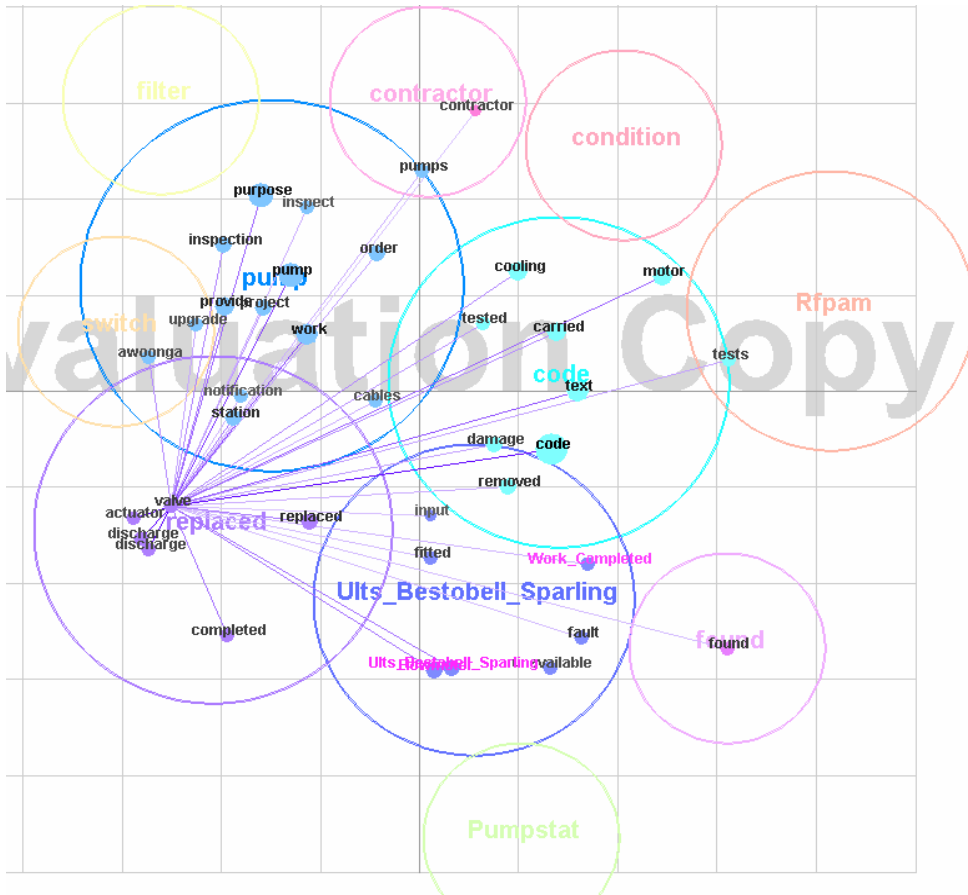


Figure 5: Concept Map