ASSERT – Internal Progress Report

Brian Rea – 13.02.07

Overview of Progress

Due to delays in the appointment of a new researcher to carry out this project did not start until 1st December 2006. After resent discussions with Alison Turner it was agreed that the official start date of the project, and all subsequent dates, could be changed accordingly. The official start and end dates of the project are now respectively 01/12/06 and 30/11/08. All project documentation has been updated to take this into account.

The project activities are fully on schedule with some significant developments being delivered early. As the test data was supplied early by EPPI it has been possible to develop an early prototype demonstrator for the document clustering work packages. This shows some of the functionality that will be available, though this will be extended over the coming months. By having such a tool at this early stage we gain several benefits in terms of assisting in requirements gathering and evaluation framework development through demonstrated use and feedback from stakeholders. It also allows us to ensure high levels of quality and usability by allowing changes to made to the system based on user experience, which is often more difficult and costly at later stages of development.

A project web site has been developed for ASSERT and is now available at http://www.nactem.ac.uk/assert.php . As part of the dissemination work package this will be updated regularly showing progress and results and all project documentation will be made available, although certain personal and financial details will be omitted from the public versions.

After lengthy discussion it has been decided that the information extraction and document clustering work packages will be swapped around to allow for earlier development of tools and to allow the results of the clustering tools to assist in the development of the information extraction tools. This will have no significant bearing on the project plan and will serve to gather more interest at the earlier stages of the project and to more actively engage the end users as the project develops rather than at final deliverable dates. The updated work packages are available on the website at http://www.nactem.ac.uk/assert/reports/workPackages120207.pdf .

Detailed Progress to Date

Requirements Analysis

The following requirements came out of the procedures employed in manual systematic reviews, based upon our discussions with our EPPI partners. The requirements are further analyzed from the point of view of possible automatic text mining techniques that could assist human reviewers through the three phases of systematic reviews: search, screen and synthesize.

Search

After a review problem has been defined human reviewers need to come up with some very complex search strategies in order to locate those relevant studies. The number of search results need to be limited to any available labour resources, whilst giving a complete account of the area under investigation. It is non-trivial to devise such search strategies. Here text mining can help with query definition through term and acronym expansion in order to find relevant documents, without worrying too much about the number of hits, since there will be help from the other text mining techniques to reduce the load for humans at each of the later stages.

As part of the visualisation of the search results the users have requested query highlighting to assist the user in detection of appropriate documents. Finally, where possible, duplicate documents should be removed from the collection before reaching the user. This can be resolved by checking of the metadata and looking at the document similarities to identify all documents within a given threshold of 100%.

Screening

During the screening phase the reviewer has to go through all of the documents one by one and decide which ones to include or exclude against a set of predefined inclusion / exclusion criteria. Text mining can assist here by discovery of the information referenced in the criteria through named entity recognition techniques, followed by a combination of information extraction techniques to identify the sets of documents that are limited by the criteria.

Synthesis

During the last phase, a review is created, summarising all the findings from the documents screened out of the search result. TM can help with this using extractive summarisation to list the most significant sections of the documents most relevant to the enquiry.

Cluto Tool

This powerful, open-license, clustering tool has a wide range of features beyond standard clustering tools. Not only can this cluster the documents but it can also clusters the terms within the documents by identifying the co-occurring features that most clearly characterize a given cluster. It is capable of outputting simple sets of features for each cluster or generating hierarchical tree diagrams for the documents or terms (or both simultaneously as shown below). This is both fast and scalable but relies upon the user identifying the number of clusters within a set; this is generally unknown but can be found through experimentation.

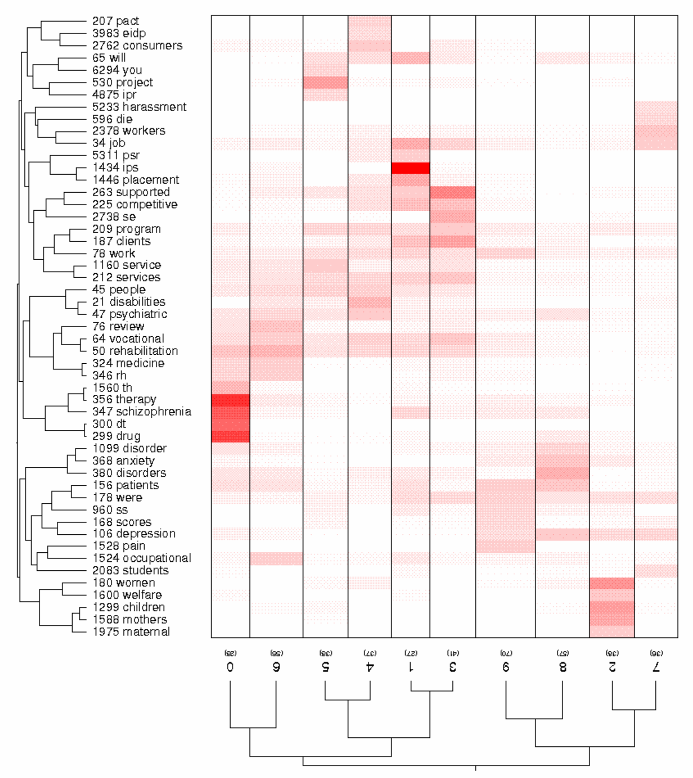


Figure 1: Sample pictorial output from Cluto with Mental Health test data

Cluto supports many different types of output ranging from tabular details of the clusters with similarity details or key features to pictorial formats showing the hierarchies of clusters and strengths of features (as shown in figure 1 above). Each of these types of results can offer the user different views of the collection of documents and are useful in different ways. Any service built upon this should be able to provide the flexibility to access each different method to allow the best possible overview of the collection and to assist in the experimental process of discovery of the best number of clusters.

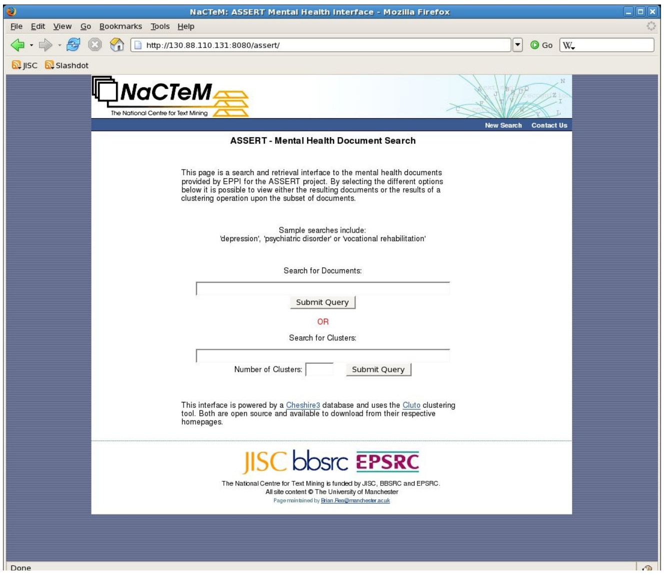
Cheshire Tool

This acts as the main store of the documents after ingest and reformatting as well as linking between the web interface and the component tools. To assist in a more focused investigation within the domain of study it allows for searching within the document and filtering before sending the sub-set of documents to other tools. This not only benefits in faster execution of the analysis but also improves the accuracy, as the tools will be working with less 'noise' from other unrelated sub-topics.

Since the use of Cheshire in the construction of the prototype demonstrator a number of questions have been raised concerning the suitability of it for this project. The strict requirement of a single operating system and lack of supporting documentation for porting to other platforms, mean that we would be limiting the number of potential users beyond the web interface. As such we feel it necessary to investigate alternative information retrieval systems at this stage rather than have to replace the entire engine at a later date. The current main competitor is Lucene which is described in a later section. More information on the issues involved and the justification for this decision will be made available in a separate document once this investigation has been carried out.

Cheshire/Cluto Demonstrator

This prototype demonstrator has been developed over the last month to act as an early point of reference to the stakeholders and to elicit discussion for assisting in the requirements gathering stage and development of evaluation metrics. As it is still in development stages the demo is currently only available locally at Manchester, but we will open the site for those with a specific need upon request.



NaCTeM: ASSERT Mental Health Interface - Mozilla Firefox
File
Edit
View
Go
Bookmarks
Tools
Help
http://130.88.110.131:8080/assert/
Go
JISC
Slashdot
NaCTeM
The National Centre for Text Mining
New Search
Contact Us
Submit Query
OR
Submit Query
JISC
bbsrc
EPSRC
Done

ASSERT - Mental Health Document Search

This page is a search and retrieval interface to the mental health documents provided by EPPI for the ASSERT project. By selecting the different options below it is possible to view either the resulting documents or the results of a clustering operation upon the subset of documents.

Sample searches include:
'depression', 'psychiatric disorder' or 'vocational rehabilitation'

Search for Documents:

Search for Clusters:

Number of Clusters:

This interface is powered by a Cheshire3 database and uses the Cluto clustering tool. Both are open source and available to download from their respective homepages.

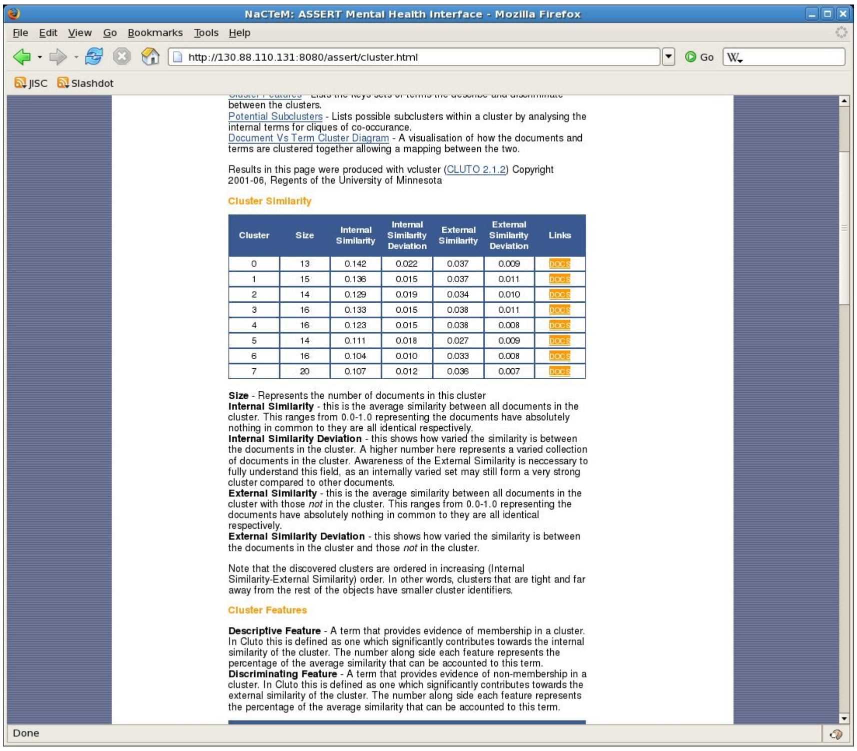
The National Centre for Text Mining is funded by JISC, BBSRC and EPSRC.
All site content © The University of Manchester
Page maintained by Brian.Rea@manchester.ac.uk

Figure 2: Prototype Demonstrator Interface

The demonstrator brings together the power of information manipulation and retrieval from Cheshire and the clustering capabilities of Cluto to create a web interface for examining a collection of documents. It is currently operating with the mental health data provided by EPPI, with the intention of extending this out to other sets as more users become involved. The interface, pictured below, allows the user to choose between examining the documents based upon a query, like familiar search engines, or to view the clusters predicted by Cluto. The IR side acts as a filtering mechanism only analysing an appropriate subset of the collection reducing processing requirement and irrelevant ‘noise’ from the analysis.

The results are split into four sections:

- *Cluster Similarity* - Indicates the strength of the clusters, showing how internally consistent and mutually exclusive they are.
- *Cluster Features* - Lists the keys sets of terms that describe and discriminate between the clusters.
- *Potential Sub-clusters* - Lists possible sub-clusters within a cluster by analysing the internal terms for cliques of co-occurrence.
- *Document and Term Cluster Diagram* - A visualisation of how the documents and terms are clustered together allowing a mapping between the two.


NaCTeM: ASSERT Mental Health Interface - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://130.88.110.131:8080/assert/cluster.html
Go
JISC
Slashdot
Done

Cluster Features - Lists the keys sets of terms the describe and discriminate between the clusters.
Potential Subclusters - Lists possible subclusters within a cluster by analysing the internal terms for cliques of co-occurance.
Document Vs Term Cluster Diagram - A visualisation of how the documents and terms are clustered together allowing a mapping between the two.

Results in this page were produced with vcluster (CLUTO 2.1.2) Copyright 2001-06, Regents of the University of Minnesota

Cluster Similarity

Cluster	Size	Internal Similarity	Internal Similarity Deviation	External Similarity	External Similarity Deviation	Links
0	13	0.142	0.022	0.037	0.009	DOCS
1	15	0.136	0.015	0.037	0.011	DOCS
2	14	0.129	0.019	0.034	0.010	DOCS
3	16	0.133	0.015	0.038	0.011	DOCS
4	16	0.123	0.015	0.038	0.008	DOCS
5	14	0.111	0.018	0.027	0.009	DOCS
6	16	0.104	0.010	0.033	0.008	DOCS
7	20	0.107	0.012	0.036	0.007	DOCS

Size - Represents the number of documents in this cluster
Internal Similarity - this is the average similarity between all documents in the cluster. This ranges from 0.0-1.0 representing the documents have absolutely nothing in common to they are all identical respectively.
Internal Similarity Deviation - this shows how varied the similarity is between the documents in the cluster. A higher number here represents a varied collection of documents in the cluster. Awareness of the External Similarity is neccessary to fully understand this field, as an internally varied set may still form a very strong cluster compared to other documents.
External Similarity - this is the average similarity between all documents in the cluster with those *not* in the cluster. This ranges from 0.0-1.0 representing the documents have absolutely nothing in common to they are all identical respectively.
External Similarity Deviation - this shows how varied the similarity is between the documents in the cluster and those *not* in the cluster.

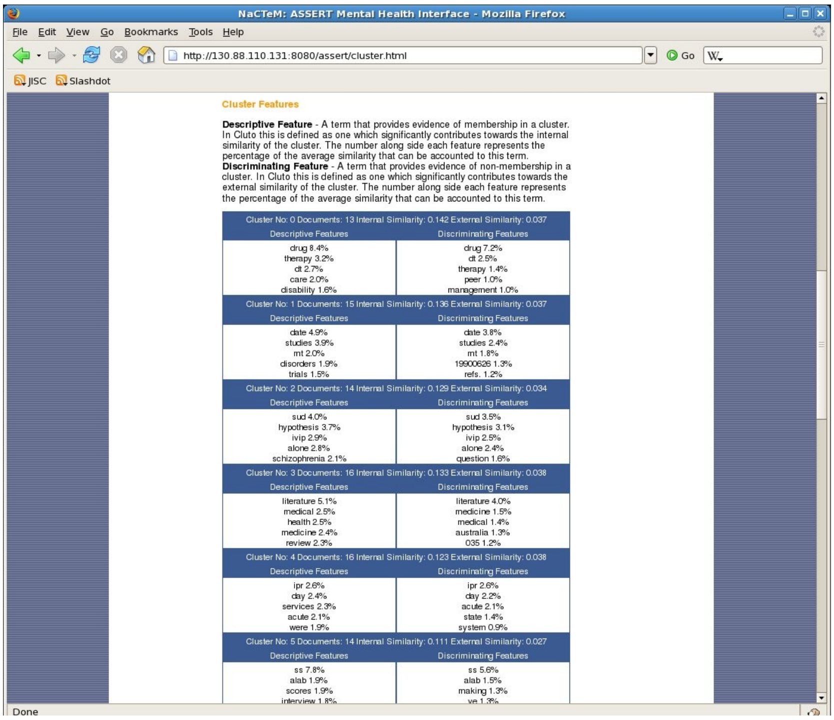
Note that the discovered clusters are ordered in increasing (Internal Similarity-External Similarity) order. In other words, clusters that are tight and far away from the rest of the objects have smaller cluster identifiers.

Cluster Features

Descriptive Feature - A term that provides evidence of membership in a cluster. In Cluto this is defined as one which significantly contributes towards the internal similarity of the cluster. The number along side each feature represents the percentage of the average similarity that can be accounted to this term.
Discriminating Feature - A term that provides evidence of non-membership in a cluster. In Cluto this is defined as one which significantly contributes towards the external similarity of the cluster. The number along side each feature represents the percentage of the average similarity that can be accounted to this term.

Figure 3: Cluster Similarities Interface

The ‘Cluster Similarities’ section describes the general details of each cluster including information about the number of documents it contains, metrics of internal consistency, external distance and a link to display the set of documents. This is useful to gain an overview of the strength and suitability of the set of clusters before looking at the more detailed analysis provided by the other sections.


NaCTeM: ASSERT Mental Health Interface - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://130.88.110.131:8080/assert/cluster.html
Go
W
JISC
Slashdot
Done

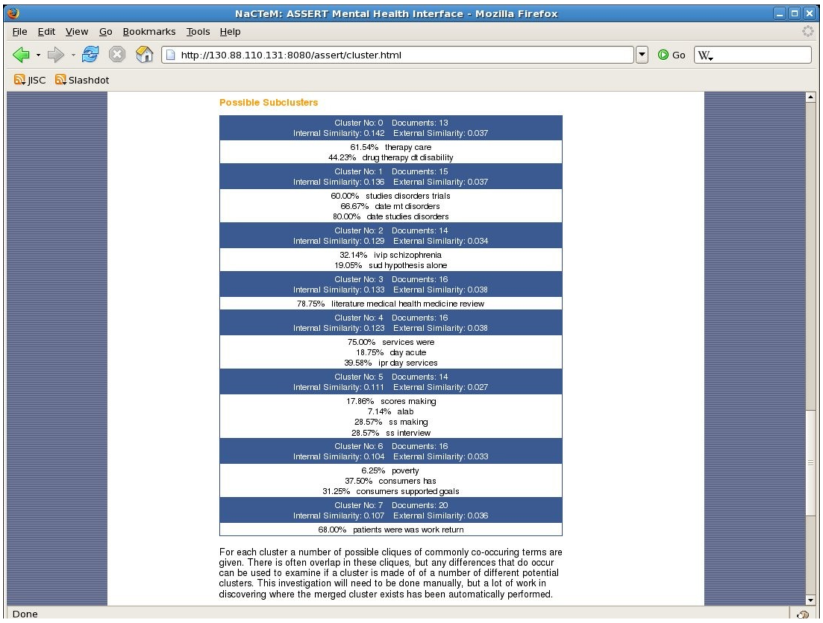
Cluster Features

Descriptive Feature - A term that provides evidence of membership in a cluster. In Cluto this is defined as one which significantly contributes towards the internal similarity of the cluster. The number along side each feature represents the percentage of the average similarity that can be accounted to this term.
Discriminating Feature - A term that provides evidence of non-membership in a cluster. In Cluto this is defined as one which significantly contributes towards the external similarity of the cluster. The number along side each feature represents the percentage of the average similarity that can be accounted to this term.

Cluster No: 0 Documents: 13 Internal Similarity: 0.142 External Similarity: 0.037	
Descriptive Features	Discriminating Features
drug 8.4% therapy 3.2% dt 2.7% care 2.0% disability 1.6%	drug 7.2% dt 2.5% therapy 1.4% peer 1.0% management 1.0%
Cluster No: 1 Documents: 15 Internal Similarity: 0.136 External Similarity: 0.037	
Descriptive Features	Discriminating Features
date 4.9% studies 3.9% mt 2.0% disorders 1.9% trials 1.5%	date 3.8% studies 2.4% mt 1.8% 19900626 1.3% refs. 1.2%
Cluster No: 2 Documents: 14 Internal Similarity: 0.129 External Similarity: 0.034	
Descriptive Features	Discriminating Features
sud 4.0% hypothesis 3.7% ivip 2.9% alone 2.8% schizophrenia 2.1%	sud 3.5% hypothesis 3.1% ivip 2.5% alone 2.4% question 1.6%
Cluster No: 3 Documents: 16 Internal Similarity: 0.133 External Similarity: 0.038	
Descriptive Features	Discriminating Features
literature 5.1% medical 2.5% health 2.5% medicine 2.4% review 2.3%	literature 4.0% medicine 1.5% medical 1.4% australia 1.3% 035 1.2%
Cluster No: 4 Documents: 16 Internal Similarity: 0.123 External Similarity: 0.038	
Descriptive Features	Discriminating Features
ipr 2.6% day 2.4% services 2.3% acute 2.1% were 1.9%	ipr 2.6% day 2.2% acute 2.1% state 1.4% system 0.9%
Cluster No: 5 Documents: 14 Internal Similarity: 0.111 External Similarity: 0.027	
Descriptive Features	Discriminating Features
ss 7.8% alab 1.9% scores 1.9% interview 1.8%	ss 5.6% alab 1.5% making 1.3% ve 1.3%

Figure 4: Cluster Features Interface

The 'Cluster Features' section lists the distinguishing features of each of the clusters, these are categorised as descriptive and discriminating. The descriptive features are those that contribute most to the internal similarity of the cluster and the discriminating features are those that contribute to the external differences. This can hint towards how appropriate the given number of clusters is, and can offer an insight into why the documents are clustered in this fashion.



NaCTeM: ASSERT Mental Health Interface - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://130.88.110.131:8080/assert/cluster.html
Go
JISC
Slashdot
Done

Possible Subclusters

Cluster No: 0 Documents: 13
Internal Similarity: 0.142 External Similarity: 0.037
61.54% therapy care
44.23% drug therapy dt disability

Cluster No: 1 Documents: 15
Internal Similarity: 0.136 External Similarity: 0.037
60.00% studies disorders trials
66.67% date mt disorders
80.00% date studies disorders

Cluster No: 2 Documents: 14
Internal Similarity: 0.129 External Similarity: 0.034
32.14% ivip schizophrenia
19.05% sud hypothesis alone

Cluster No: 3 Documents: 16
Internal Similarity: 0.133 External Similarity: 0.038
78.75% literature medical health medicine review

Cluster No: 4 Documents: 16
Internal Similarity: 0.123 External Similarity: 0.038
75.00% services were
18.75% day acute
39.58% ipr day services

Cluster No: 5 Documents: 14
Internal Similarity: 0.111 External Similarity: 0.027
17.86% scores making
7.14% alab
28.57% ss making
28.57% ss interview

Cluster No: 6 Documents: 16
Internal Similarity: 0.104 External Similarity: 0.033
6.25% poverty
37.50% consumers has
31.25% consumers supported goals

Cluster No: 7 Documents: 20
Internal Similarity: 0.107 External Similarity: 0.036
68.00% patients were was work return

For each cluster a number of possible cliques of commonly co-occuring terms are given. There is often overlap in these cliques, but any differences that do occur can be used to examine if a cluster is made of of a number of different potential clusters. This investigation will need to be done manually, but a lot of work in discovering where the merged cluster exists has been automatically performed.

Figure 5: Sub-Cluster Detection Interface

The ‘Sub-cluster Detection’ section of the interface analyses the documents in each of the clusters to find cliques of terms that can suggest the overall content of the cluster. This assists the user to evaluate each cluster for possible undetected sub-clusters. If the terms presented in each of the cliques are not semantically related or suggest multiple topics the user may wish to perform the analysis again with a larger number of clusters. This interface has been designed with speed as a key issue, to ensure that such multiple sets of analysis may be carried out in a reasonable time frame. Compared with the manual alternatives this is therefore a significant improvement.

Finally the visualisation section creates a diagram clustering the documents and terms hierarchically and representing the clusters in a visual format. An example of this is shown in figure 1 above, whilst a worked example of how this can be used is shown below.

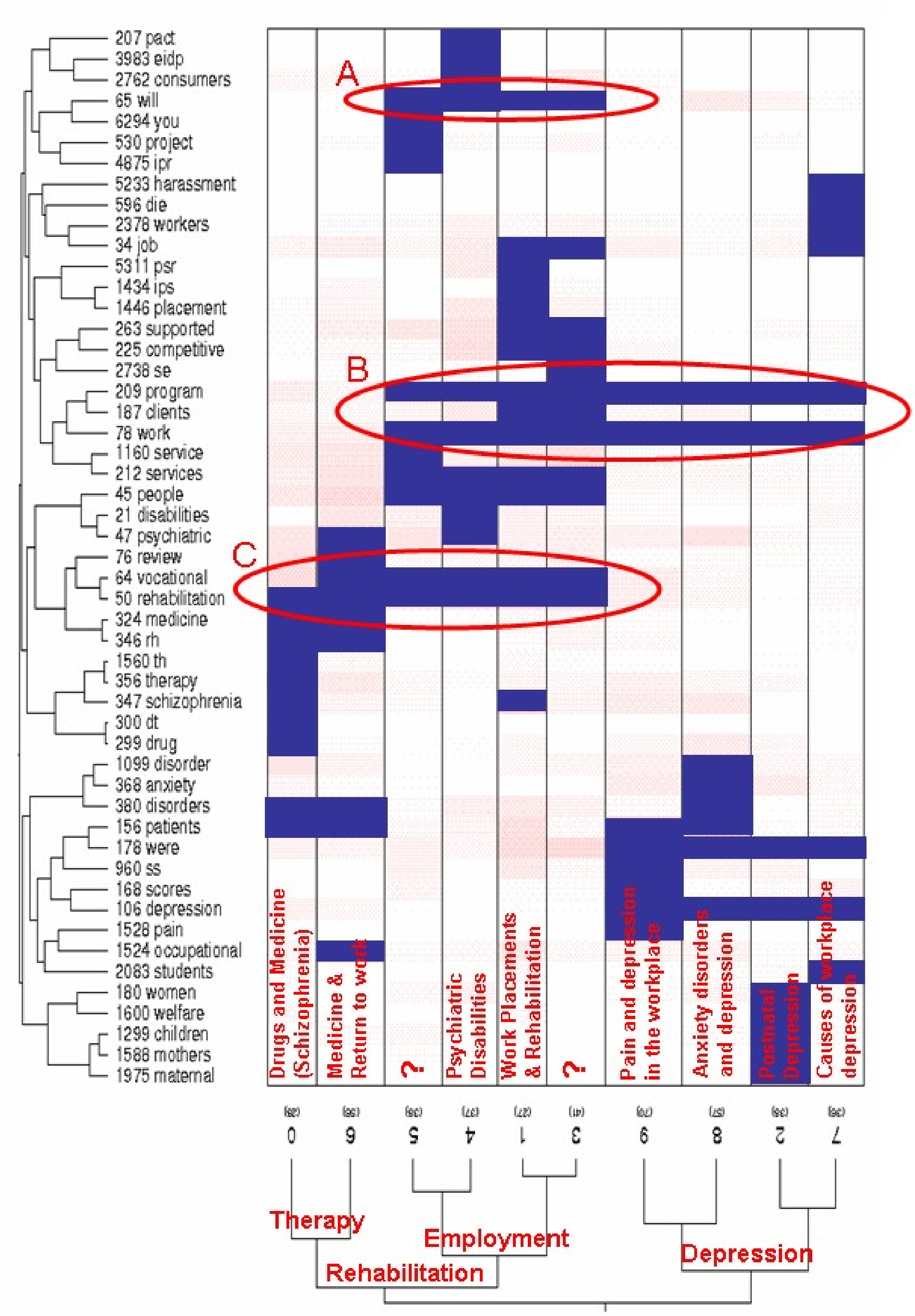


Figure 6: Manually annotated pictorial output from demonstrator

This version of the cluster diagram is the same picture as in figure X but with additional manual annotations to highlight some of the features. These additions are not part of the demonstrator but represent a few minutes work looking between the diagram and other cluster information for a given query. The labels on each cluster represent the theme of the cluster based upon the features with the highest strengths. From this information it is possible to then label the cluster trees to show the main topics of each of the branches. As can be observed two of the clusters do not contain enough information to say with any certainty what the main topic is. When we bring this together with the similarity information it seems that these clusters are more internally consistent than some of the others, so it is not a matter that they are made up of documents on many different subjects. On the other hand it can be seen that these clusters are based upon the final separations, from the height of the division. This suggests that clusters 1 & 3 and 4 & 5 are the closest matches and that perhaps the number of clusters given by the user may be incorrect. Perhaps nine or eight clusters would give better results.

Overall this shows that the tool can support this type of investigations and can show an overview of the collection as a whole, in a matter of seconds. It also shows that the identified problem of an unknown number of clusters can be at least partially resolved with a small amount of manual assessment, clearly less than would be required for someone to carry out the whole process by hand. Features that occur across many of the clusters can be classified two ways, the first being a significant summary of the collection (annotations B and C) and the second being a word with little information that appears in many documents (annotation A). The next stage of development of this tool will look at removing these valueless words (or stop words as they are commonly called) as well as assigning scores to the other words to weight them on individual contribution to the collection.

Lucene

This tool is a scalable information retrieval library in a similar vein to Cheshire. What it lacks in some of the flexibility in internal objects it more than makes up for in terms of maturity, stability and documentation. Written in Java (but also available in other languages), Lucene is platform independent and its open source nature combined with the reputation of its Apache branding make it the most popular free Java based IR library. The large user base and strong community support for Lucene ensure that the major APIs are well known and used widely. For us this means that anything we build with Lucene will be much easier for others to understand or implement themselves given our documentation. This means that if we accept this as the main IR component we can expect a better chance of the software being sustainable and used by the UK and international communities.

So far the development and testing of this software for this project has proved it to be most useful, but this is based upon simple experiments and observations rather than complex services. Over the coming months we intend to test this much further and consider the implications of its use, such as losing the distributed computing components or gaining access to community driven extensions, before any final decision is made. This may take some time at this stage of the project, but would take much longer at later stages if we had to redevelop all of the IR code.

MIMA

This clustering tool takes input from the document collection and clusters the documents based upon similarity of the words (and terms). It allows for the documents to be weighted by the user based upon source, a useful feature for the evidence gathering and analysis stage. For small sets of documents this tool works very well allowing user oriented topic maps to be automatically generated from the clusters. However the layout algorithm struggles with the small viewing space and very large document sets and the results can become hard to read. Despite this it is still useful within the system when combined with the filtering techniques of an information retrieval tool.

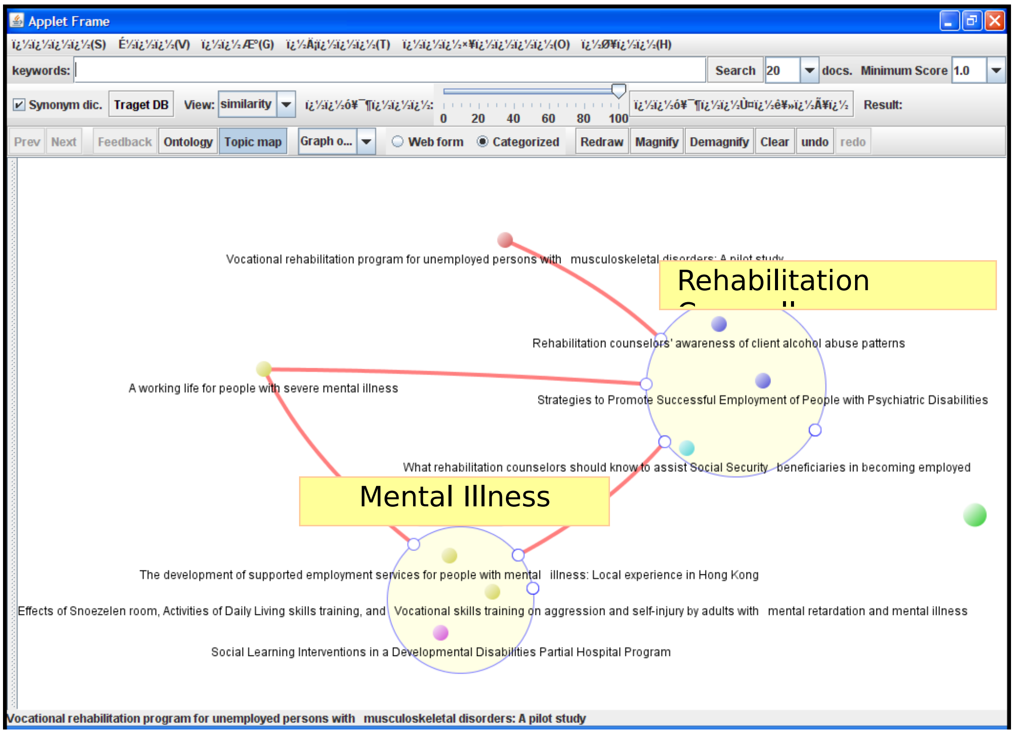

Applet Frame
keywords:
Search
20
docs.
Minimum Score
1.0
☑ Synonym dic.
Traget DB
View: similarity
0 20 40 60 80 100
Result:
Prev
Next
Feedback
Ontology
Topic map
Graph o...
○ Web form
● Categorized
Redraw
Magnify
Demagnify
Clear
undo
redo
Rehabilitation
Mental Illness
Vocational rehabilitation program for unemployed persons with musculoskeletal disorders: A pilot study
Rehabilitation counselors' awareness of client alcohol abuse patterns
A working life for people with severe mental illness
Strategies to Promote Successful Employment of People with Psychiatric Disabilities
What rehabilitation counselors should know to assist Social Security beneficiaries in becoming employed
The development of supported employment services for people with mental illness: Local experience in Hong Kong
Effects of Snoezelen room, Activities of Daily Living skills training, and Vocational skills training on aggression and self-injury by adults with mental retardation and mental illness
Social Learning Interventions in a Developmental Disabilities Partial Hospital Program
Vocational rehabilitation program for unemployed persons with musculoskeletal disorders: A pilot study

Figure 7: Example output from MIMA with mental health documents

Further investigation of this tool is required before a decision can be made of it use in this project. It is clear that the interface itself offers many benefits, such as not needing to know the number of clusters, but there is no current way of accessing the resulting documents for further analysis.