Basic Operations in CiteSpace

Version: CiteSpace 2.0 R10
Date: Sept 15, 2006
Author: Chaomei Chen

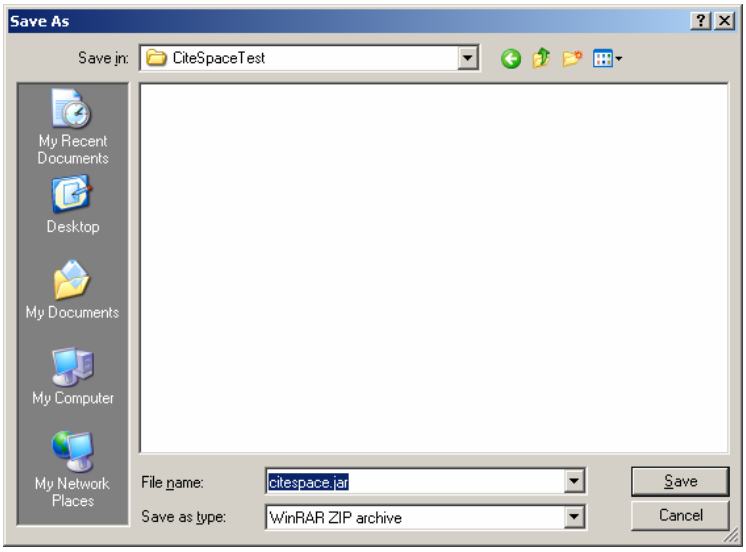
Setup:	Steps 1-4
Run:	Steps 5-7
Run with Burst Terms:	Steps 5-9
Run with Noun Phrases:	Steps 5-7, 10-11
Customize Configurations:	Step 12*
Run EM:	Step 13

*Reported bugs. To be fixed.

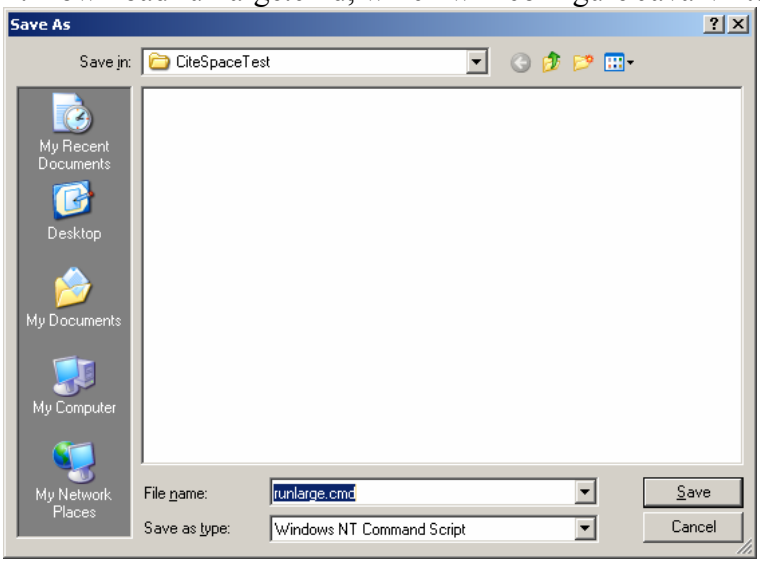
1. Download citespace.jar
2. Download runlarge.cmd
3. Create a 'resources' subdirectory
4. Download resources files for part of speech analysis
5. Launch CiteSpace
6. Create a Project in CiteSpace: 6a) Create the data directory; 6b) Create a Project
7. Start data processing (Restart CiteSpace if this is the first time ever you created a project)
8. Select the burst terms option and generate burst terms if necessary
9. Generate visualizations with burst terms
10. Extract noun phrases with part-of-speech tagging
11. Generate networks with noun phrases
12. Configuration files: ExportSpace (it has a bug)
13. EM

1. Download citespace.jar to a directory.

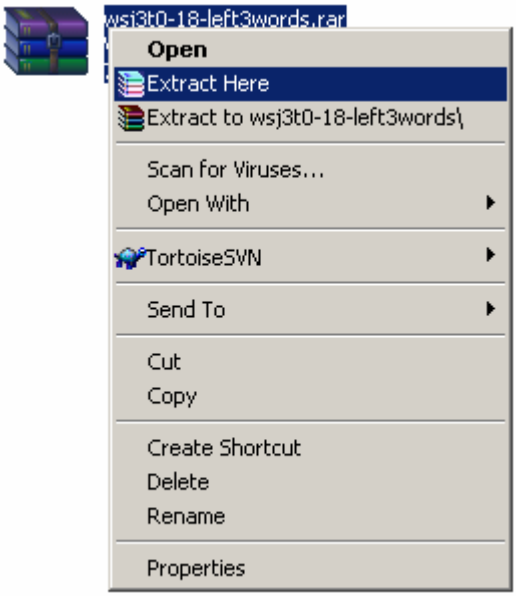
The directory could have any name. In this example, it is called CiteSpaceTest.



2. Download runlarge.cmd, which will configure Java Virtual Machine's memory.

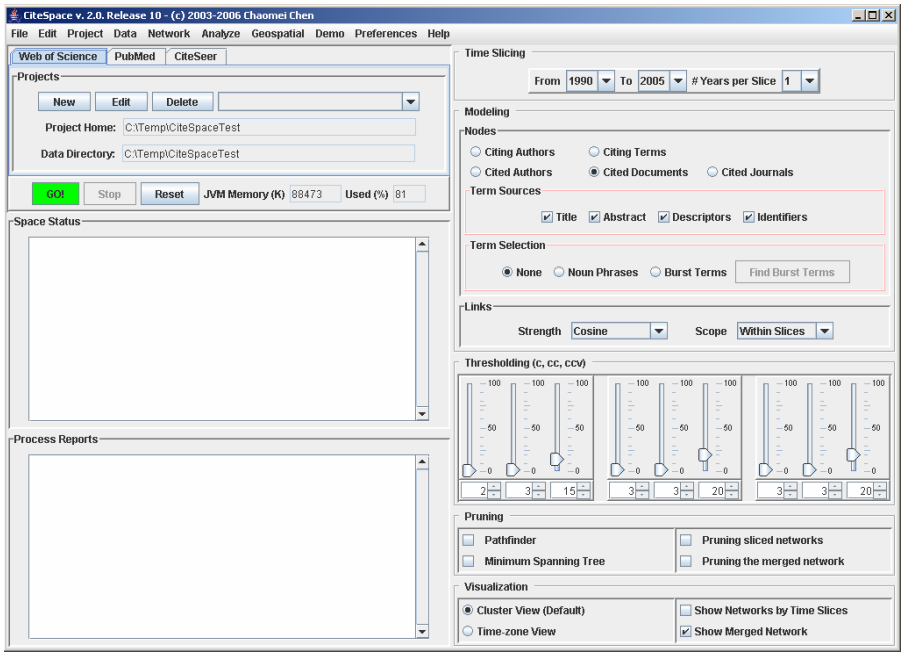


3. Create a subdirectory called 'resources'.
4. Download http://cluster.cis.drexel.edu/~cchen/citespace/resources/wsj3t0-18-left3words.rar.
5. Extract the rar file. Use 'Extract Here'. You can delete the rar file after the extraction.



6. Launch CiteSpace by double clicking on runlarge.cmd.

You should see the CiteSpace GUI as shown below.

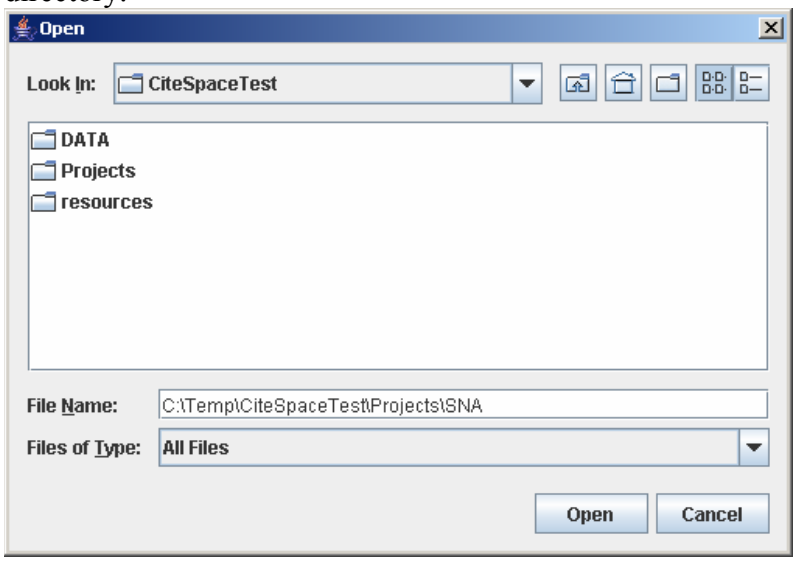


6. Create a Project in CiteSpace.

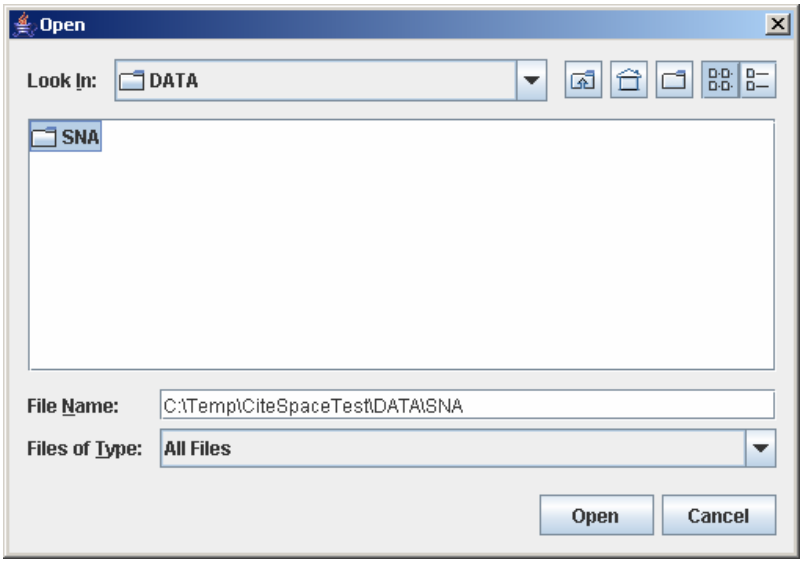
You need to specify two directories for a new project: one as the project home directory and the other as the data directory.
You need to get the data directory and data files ready before you proceed.
Suppose that you have a set of data files stored in a directory CiteSpaceTest/DATA/SNA and you want to make a directory CiteSpace/Projects/SN as the project home directory.

Go straight to step 6b if you already have a data directory with data files.
6a. Create the data directory CiteSpaceTest/DATA/SNA. Note that the data directory does not have to be a subdirectory under CiteSpaceTest.
SNA = Social Network Analysis, which is the illustrative example about.
This data set is available at:
http://cluster.cis.drexel.edu/~cchen/citespace/data/SocialNetworks1996-2002.zip
Use 'Extract Here' to unpack the files to the SNA directory.

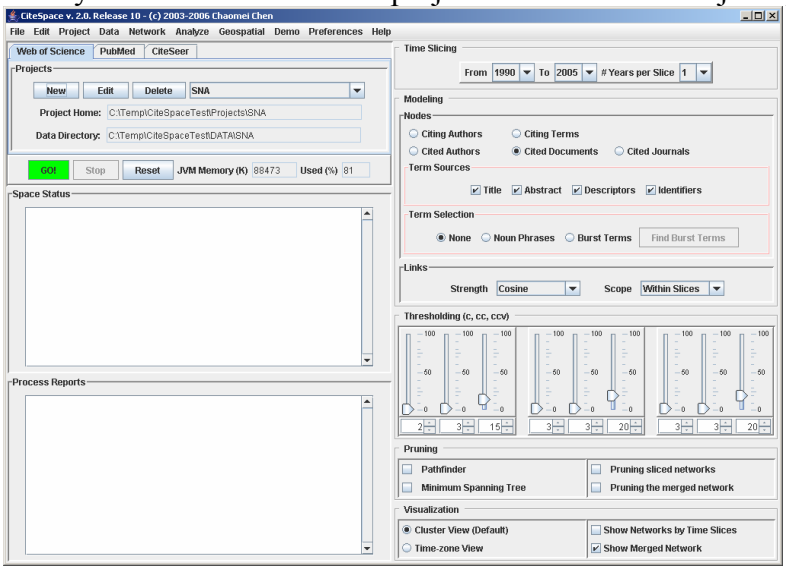
7b. Create CiteSpace projects directory if you do not have one: CiteSpaceTest/Projects. Note that the Projects directory does not have to be a subdirectory of CiteSpaceTest either.
In the Open dialog box, type in the new project SNA as a subdirectory of the Projects directory.



Next, select the DATA/SNA as the data directory for the new project.

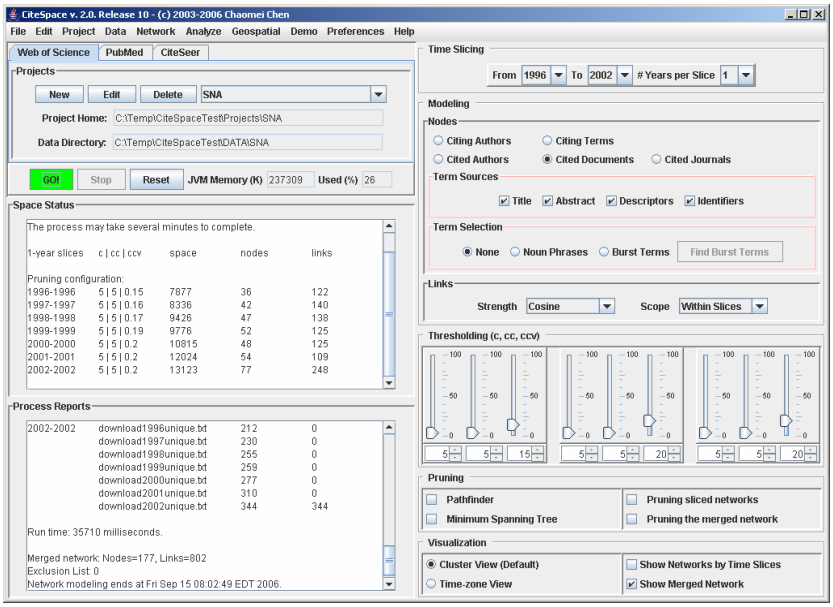


Now you should see the SNA project is listed in the Projects panel in CiteSpace.

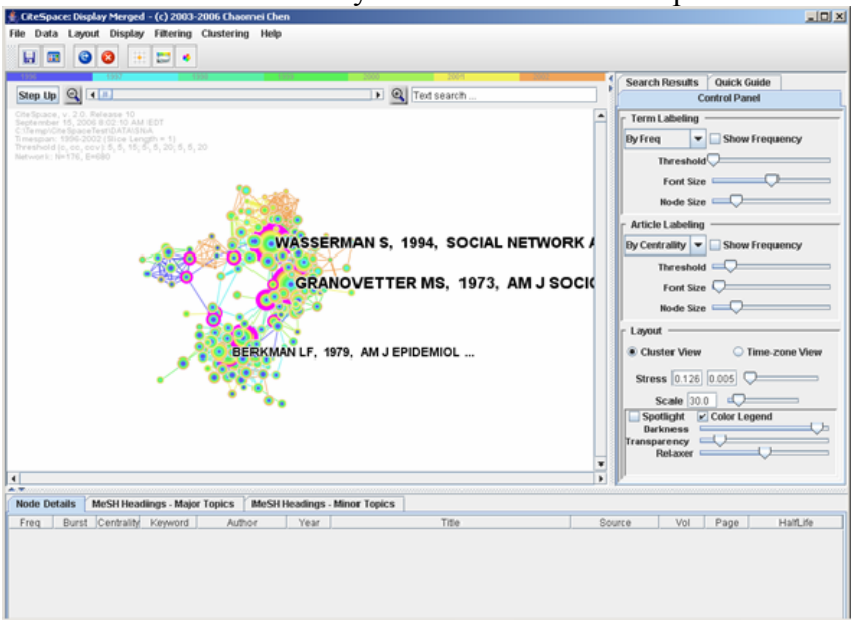


7. Processing the project in CiteSpace. For the SNA data set, change the following parameters:
Time Slicing: from 1996 to 2002
Thresholding: 5, 5, 15; 5, 5, 20; 5, 5, 20
Then press the GO button.

Possible bug: If you are creating the very first ever new project in CiteSpace, there seems to be a bug – the visualization window may not show up. Restarting CiteSpace should resolve this problem.



You should see a visualization window as shown below, without adjusting any controls. Each run of CiteSpace will open up a new visualization window. It is a good idea to close visualization windows that you do not need to free up more memory.

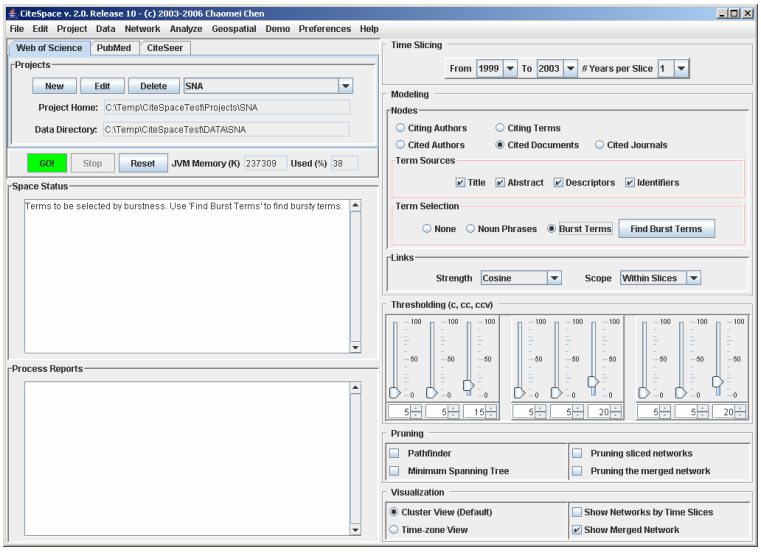


8. Select Burst Terms in the Term Selection panel.

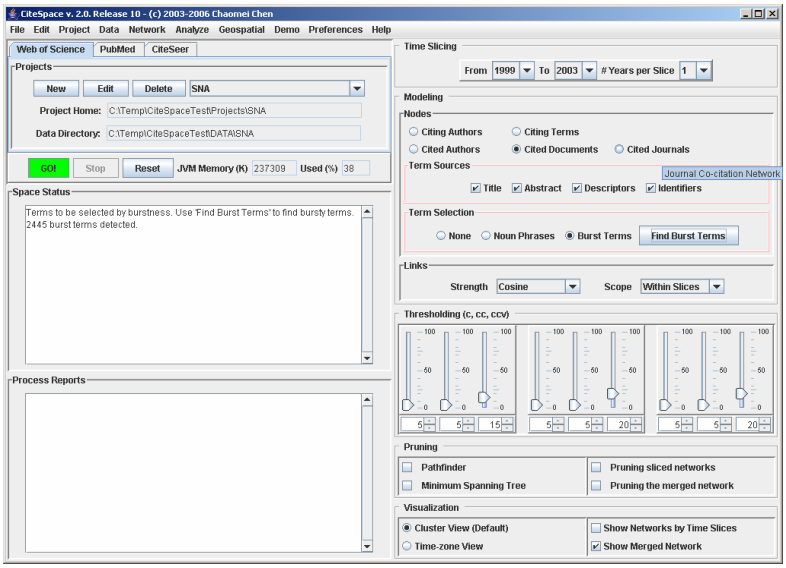
The Space Status area will display one of the following two messages:

a) Terms to be selected by burstness. Use 'Find Burst Terms' to find bursty terms.
b) CiteSpace found existing burst terms. You may use them, or re-run the detection process, which may take several seconds to complete.

If you see a), it means you need to press the “Find Burst Terms” button first to extract burst terms from the data files.

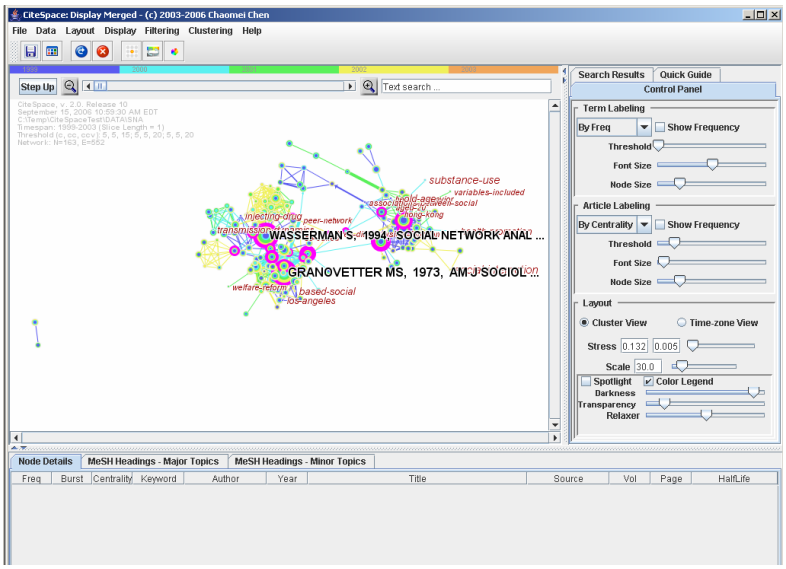


The number of burst terms will be shown in the Space Status area as soon as the process is completed, which may take a couple of seconds or much longer, depending on the volume of your data. The following figures shows that CiteSpace has found 2,445 burst terms.



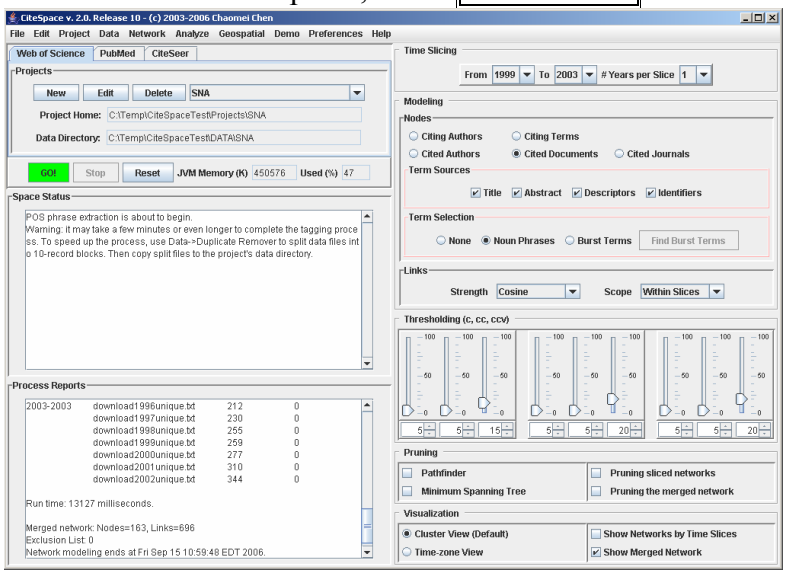
If you see b), then there are burst terms available for use. If you like to generate the burst terms again, just press the Find Burst Terms button.

9. Generate visualizations with burst terms. Make sure you select the Burst Terms radio button and the Space Status area shows that burst terms are available.
Then press the GO button



10. Extract noun phrases

In the Term Selection panel, select Noun Phrases radio button.



The Space Status area will display one of the following two messages, depending on the status:

a) No extracted noun phrases available.

POS phrase extraction is about to begin. Warning: it may take a few minutes or even longer to complete the tagging process. To speed up the process, use Data->Duplicate Remover to split data files into 10-record blocks. Then copy split files to the project's data directory.

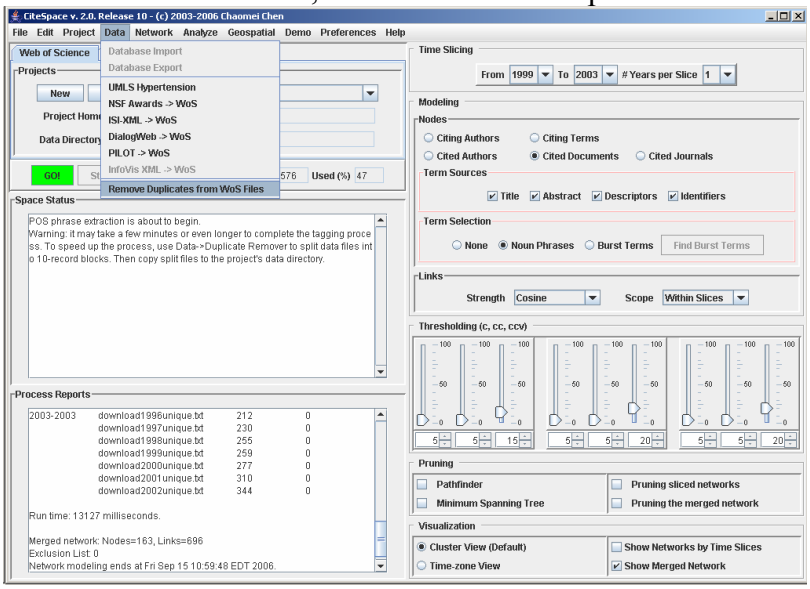
b) Previously extracted noun phrases available.

Previously extracted POS phrases are found in the project home directory. They will be used as default. If you do not wish to use them, go to the home directory and remove files with names ending with '_pos.txt' before 'GO'.

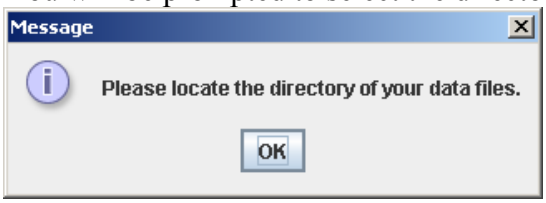
Warning: it may take a few minutes or even longer to complete the tagging process. To speed up the process, use Data->Duplicate Remover to split data files into 10-record blocks. Then copy split files to the project data directory.

The noun phrase extraction works faster with shorter files than longer files.
10a. Split data files to smaller 10-record files.
10b. Remove the original files from the data directory, or move new files to a new project.

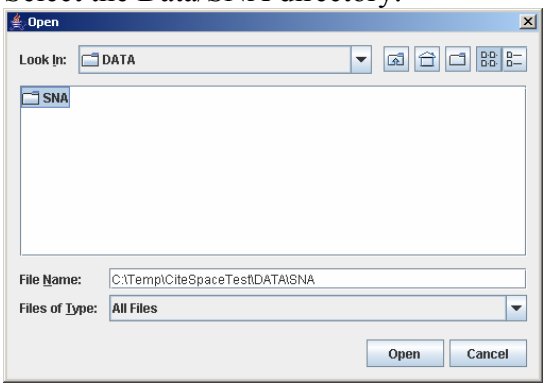
10a. From the menu bar, Data → Remove Duplicates from WoS files.



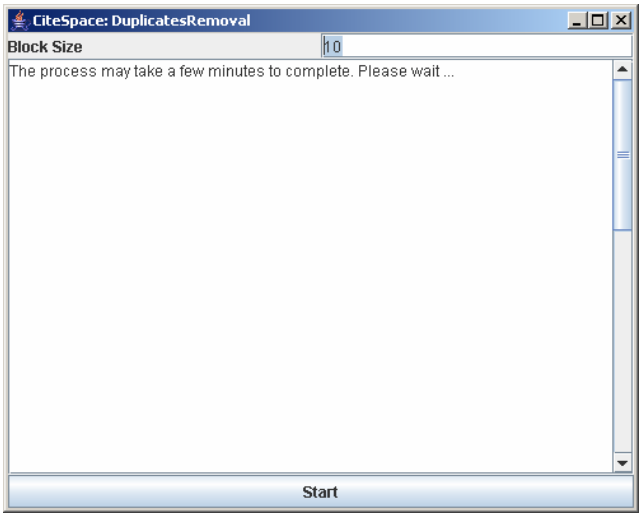
You will be prompted to select the directory of data files for the conversion.



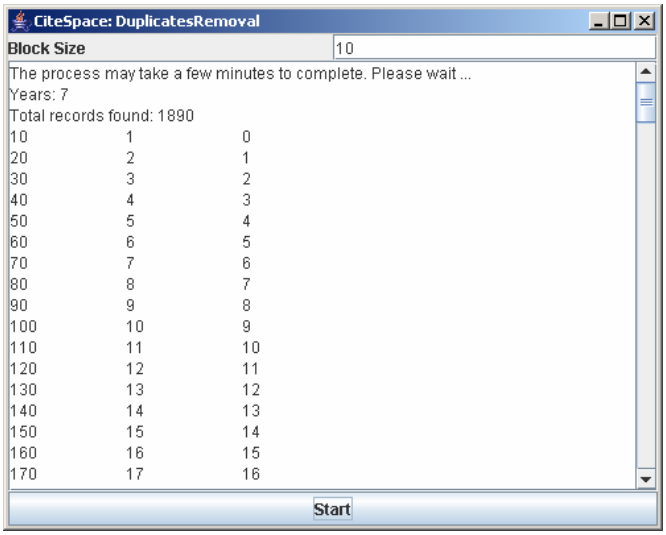
Select the Data/SNA directory.



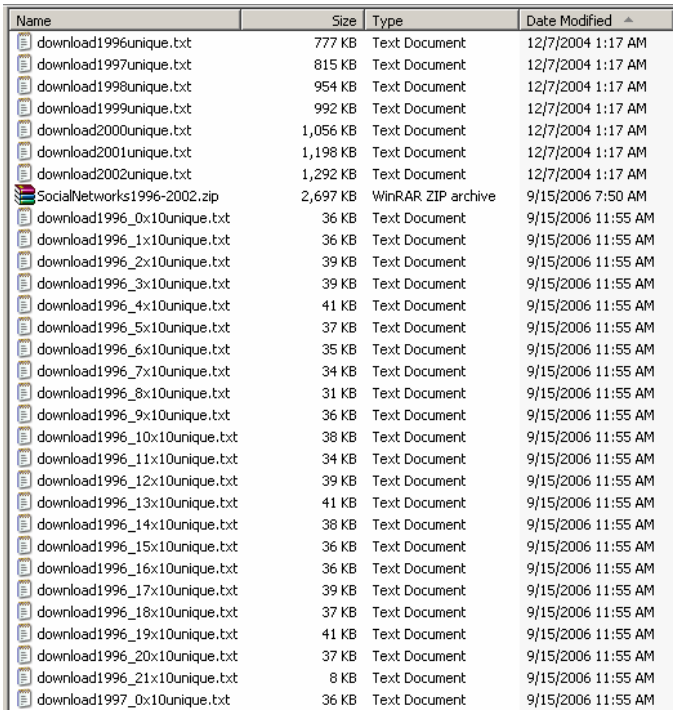
Type 10 in the Block Size field. This will generate data files containing 10 records per file.



Press Start once. The result report will show the total number of years covered by the data set as a whole, the breakdown of how each year's records are reorganized into 10-record files.



Now go to the Data/SNA directory. You should see the original data files and the newly generated small-sized files with names such as **`download1996_0x10unique.txt`**.



Name	Size	Type	Date Modified
download1996unique.txt	777 KB	Text Document	12/7/2004 1:17 AM
download1997unique.txt	815 KB	Text Document	12/7/2004 1:17 AM
download1998unique.txt	954 KB	Text Document	12/7/2004 1:17 AM
download1999unique.txt	992 KB	Text Document	12/7/2004 1:17 AM
download2000unique.txt	1,056 KB	Text Document	12/7/2004 1:17 AM
download2001unique.txt	1,198 KB	Text Document	12/7/2004 1:17 AM
download2002unique.txt	1,292 KB	Text Document	12/7/2004 1:17 AM
SocialNetworks1996-2002.zip	2,697 KB	WinRAR ZIP archive	9/15/2006 7:50 AM
download1996_0x10unique.txt	36 KB	Text Document	9/15/2006 11:55 AM
download1996_1x10unique.txt	36 KB	Text Document	9/15/2006 11:55 AM
download1996_2x10unique.txt	39 KB	Text Document	9/15/2006 11:55 AM
download1996_3x10unique.txt	39 KB	Text Document	9/15/2006 11:55 AM
download1996_4x10unique.txt	41 KB	Text Document	9/15/2006 11:55 AM
download1996_5x10unique.txt	37 KB	Text Document	9/15/2006 11:55 AM
download1996_6x10unique.txt	35 KB	Text Document	9/15/2006 11:55 AM
download1996_7x10unique.txt	34 KB	Text Document	9/15/2006 11:55 AM
download1996_8x10unique.txt	31 KB	Text Document	9/15/2006 11:55 AM
download1996_9x10unique.txt	36 KB	Text Document	9/15/2006 11:55 AM
download1996_10x10unique.txt	38 KB	Text Document	9/15/2006 11:55 AM
download1996_11x10unique.txt	34 KB	Text Document	9/15/2006 11:55 AM
download1996_12x10unique.txt	39 KB	Text Document	9/15/2006 11:55 AM
download1996_13x10unique.txt	41 KB	Text Document	9/15/2006 11:55 AM
download1996_14x10unique.txt	38 KB	Text Document	9/15/2006 11:55 AM
download1996_15x10unique.txt	36 KB	Text Document	9/15/2006 11:55 AM
download1996_16x10unique.txt	36 KB	Text Document	9/15/2006 11:55 AM
download1996_17x10unique.txt	39 KB	Text Document	9/15/2006 11:55 AM
download1996_18x10unique.txt	37 KB	Text Document	9/15/2006 11:55 AM
download1996_19x10unique.txt	41 KB	Text Document	9/15/2006 11:55 AM
download1996_20x10unique.txt	37 KB	Text Document	9/15/2006 11:55 AM
download1996_21x10unique.txt	8 KB	Text Document	9/15/2006 11:55 AM
download1997_0x10unique.txt	36 KB	Text Document	9/15/2006 11:55 AM

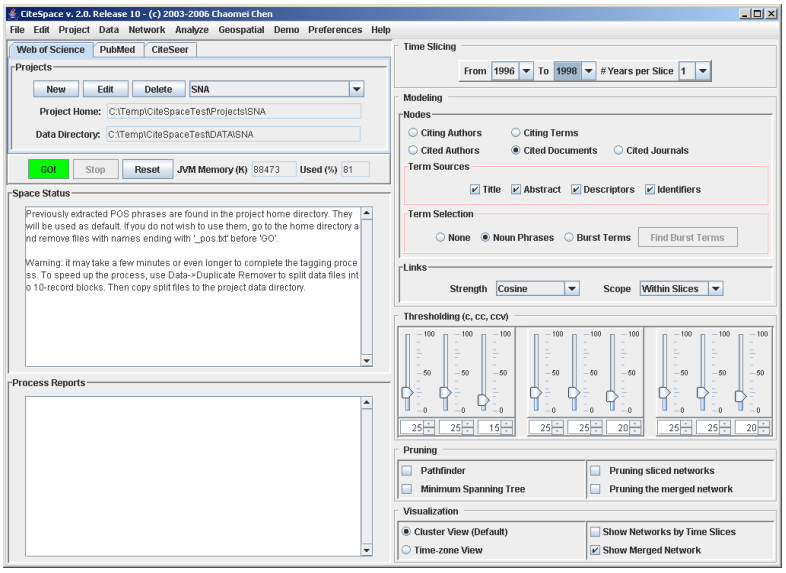
Now you can either remove the original data files or move these small-sized files to a different directory and create a new project for them.
We will remove the original data files in this example.

11. Generate networks with noun phrases
Go to the CiteSpace main interface. Set Time Slicing to the interval of 1997 and 1998. Press GO. Resultant files are saved in the project home directory.

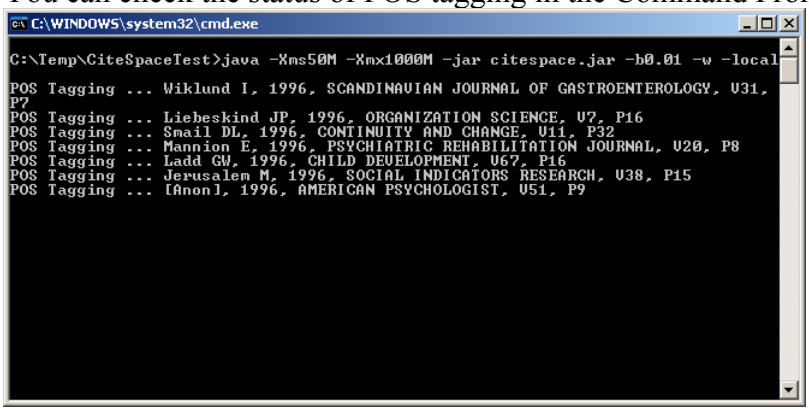
Warning: This is a time consuming process. Time taken for the SNA dataset is as follows. Expect longer waiting for longer datasets.

1996: 12min
1997: 13min
1998: 14min

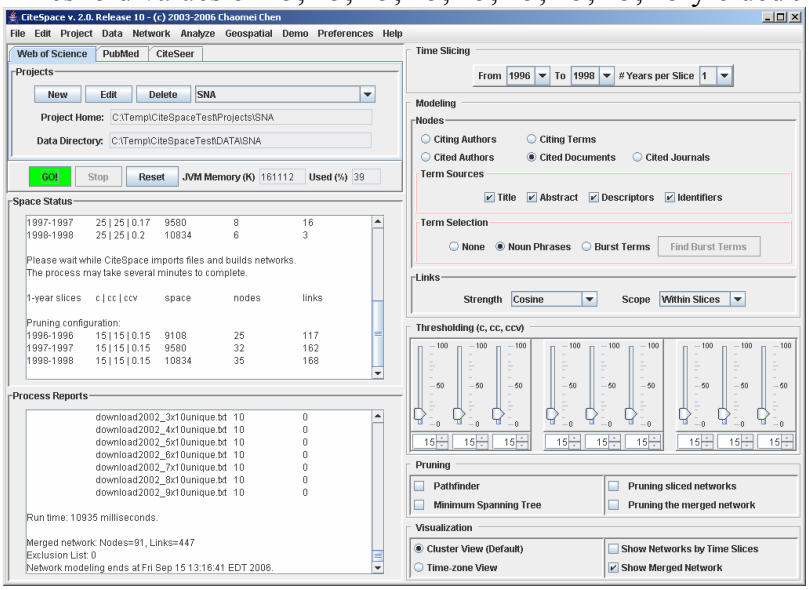
The number of noun phrases tends to be much larger than the usual number of qualified articles or burst terms. You'd better walk down from higher threshold values to lower ones, for example, starting with thresholds of 25, 25, 15; 25, 25, 20; 25, 25, 20.



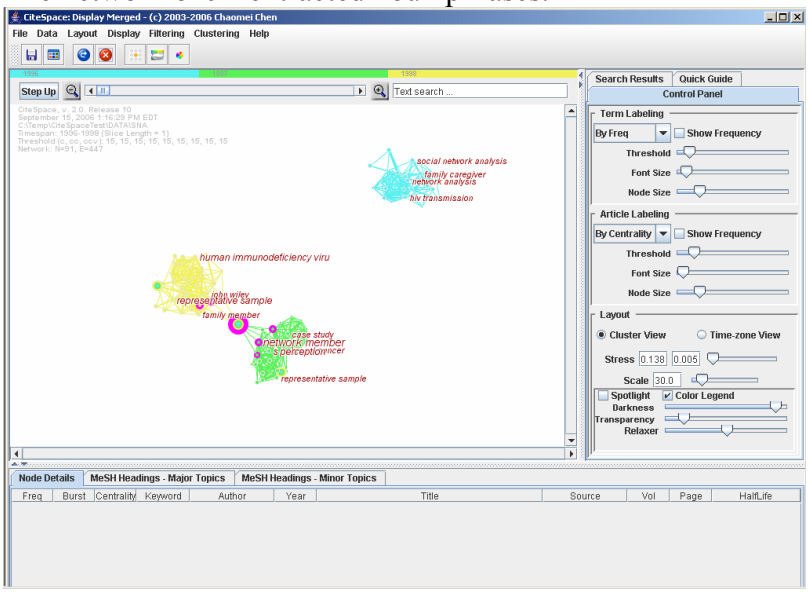
You can check the status of POS tagging in the Command Prompt window.



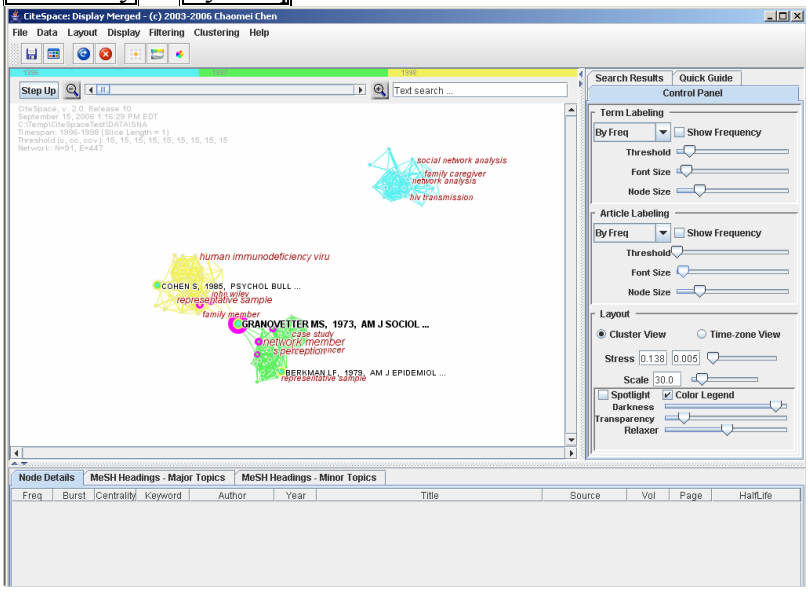
Threshold values of 15, 15, 15; 15, 15, 15; 15, 15, 15 yielded 91 unique phrases.



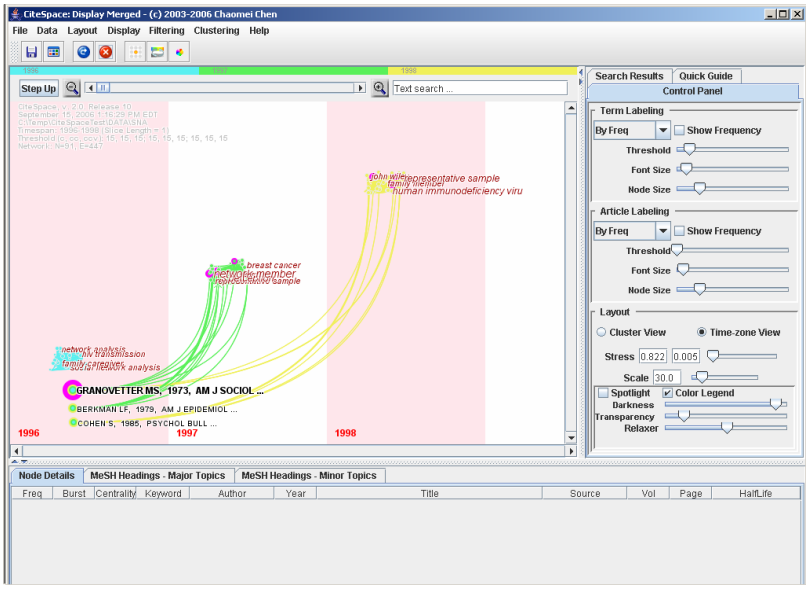
The network of 91 extracted noun phrases.



The Article Labeling in the Control Panel controls the labels. Switch from By Centrality to By Freq. You will see 3 articles in this network. All others are phrases.



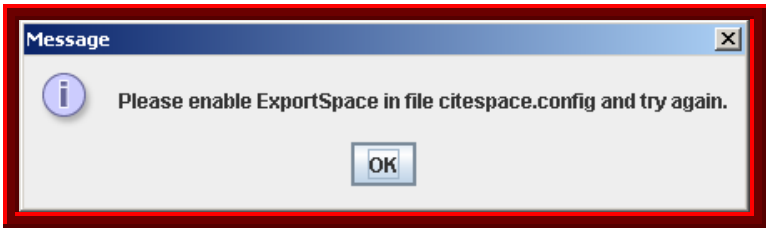
You may switch to Time-zone view and get the visualization as shown below.



12. Turn ExportSpace on
A configuration file CiteSpaceTest\Projects\SNA\citespace.config is automatically generated by CiteSpace.

```
AliasList=on
ExclusionList=on
ExportSpace=off
JDIC=on
```

To switch ExportSpace on, make the following change in the file:
`ExportSpace=on`

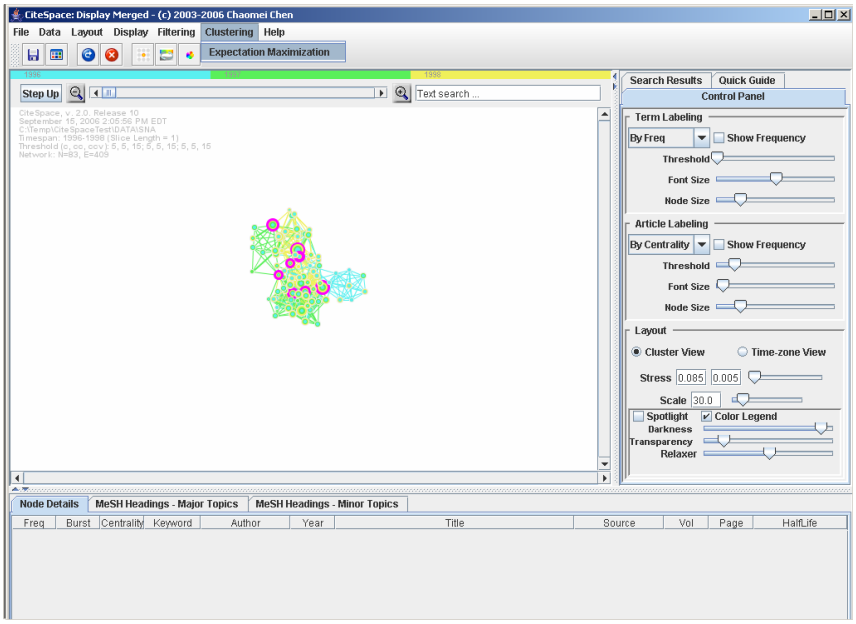


Bug: Currently, there is a bug here.

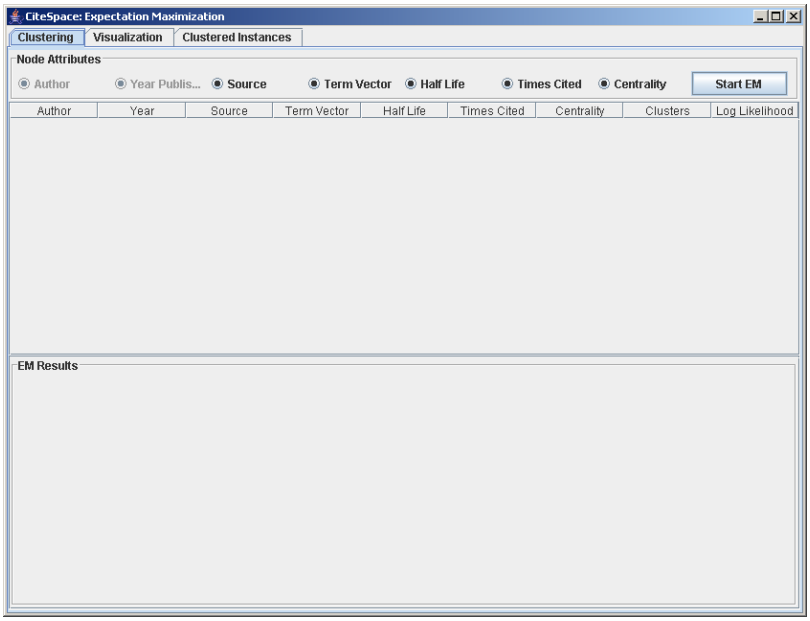
13. EM
Choose Clustering → Expectation Maximization (EM) in the Visualization Window.

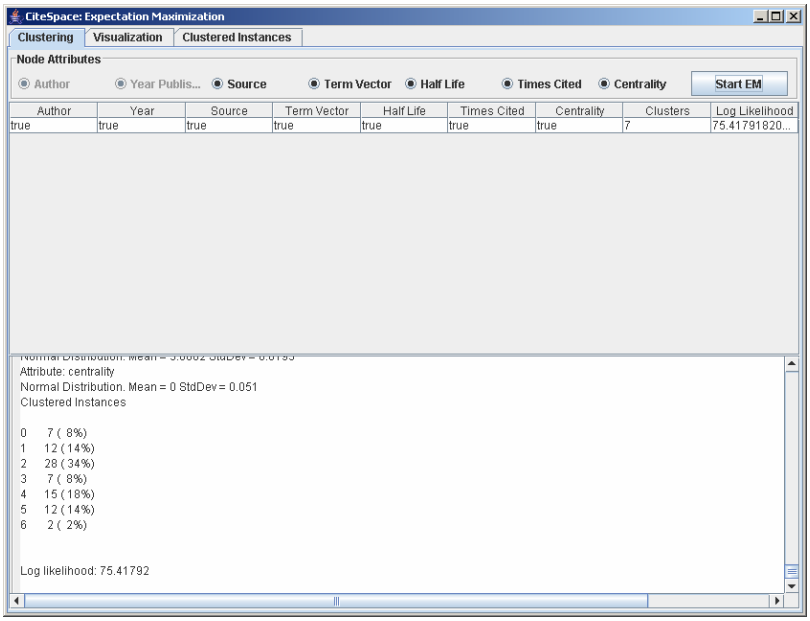
Warning: This could be a time consuming process for large datasets.

Try this example with thresholds of 5, 5, 15; 5, 5, 15; 5, 5, 15 between 1996 and 1998.

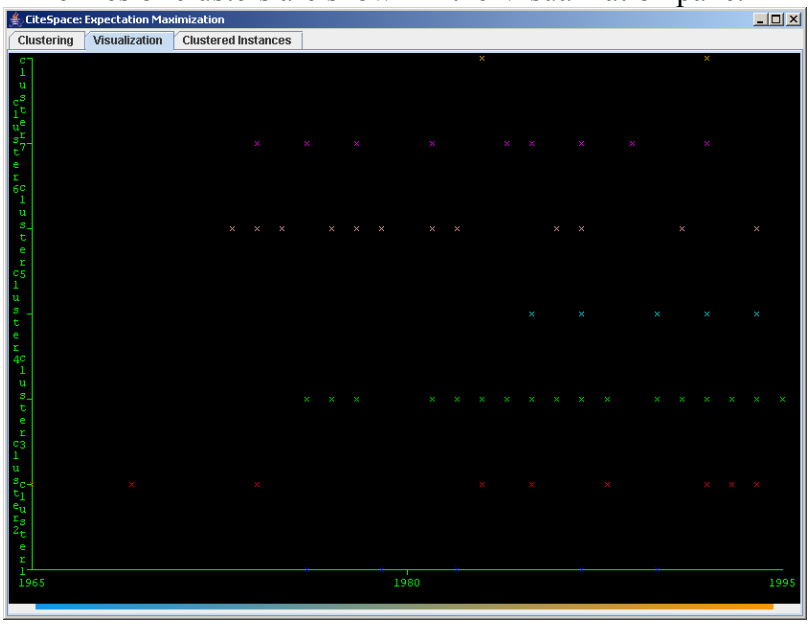


Select or de-select node attributes. Then press Start EM. (It takes 77 seconds for this example).





Timelines of clusters are shown in the Visualization pane.



The tabular data is available in the Cluster Instances pane.