Jenn Bernstein
jenn_bernstein@yahoo.com
Visualizing Information
Winter 2006
Final project Presentation

For my final project presentation, I though I would leave you with a list of techniques I used for my project in the hopes that they might be useful to you should you embark on a similar adventure.

LATENT SEMANTIC INDEXING

What: A type of information retrieval that improves upon the typical keyword search by taking word context and association into account.

Why: You wanted to search documents for concepts rather than just terms. LSA can take care of things like the same prefix having different suffixes, one word having multiple meanings, multiple words having one meaning, etc.

How: First, a term-frequency matrix is generated (how many times a single term is used in each "document"). Then the sparse matrix is made less sparse through SVD (singular value decomposition). Then a document-to-document similarity matrix is created.

What it isn't so good for: Discerning meaning in otherwise very similar documents

Resources: The "Text to Matrix Generator" is a free Matlab add-in available from some nice Greek guys at http://scgroup.hpclab.ceid.upatras.gr/scgroup/Projects/TMG/. Mac users beware, however. This add-in can create your term-document matrix and perform LSI, as well as a variety of clustering techniques (which, while they might have the same name as the ones that come with Matlab, are in fact different). It also comes with a common word dictionary, used to remove some of the "noise" from the documents. Singular value decomposition comes with Matlab.

Another option for basic use is the LSA @ CU Boulder site (http://lsa.colorado.edu/). It has a variety of LSA-associated functions, various dictionaries, and a handful or articles. Once again, Mac users beware. You have less of an ability to set your parameters than with Matlab, hence extremely user friendly.

CLUSTERING

What: Groups your similarity data into clusters, which makes it easier to visualize.

I don't know enough to advise anyone on clustering.

MULTIDIMENSIONAL SCALING

What: A spatial approach to finding underlying patterns in multivariate data.

Why: Good if you have a data set with a number of attributes, and want to find the primary factors that differentiate the data points from one another.

How: After entering your data, the first decision you make is how many dimensions to use for visualization. The more you use, the more information you capture, but the less generalized it becomes and the harder it is not represent in 2 dimensions. Typically this is done through a creating a scree plot, where can visually see the cutoff between the dimensions that explains a lot of information and the ones that explain much less.

What it isn't so good for: If you do use more than 2 dimensions, MDS squishes the information into 2, which can misrepresent proximity. Once again, the technique is good at representing significant differences, but not so good as representing subtle ones.

Resources: The Kruskal; and Wish text, "Multidimensional Scaling", is the classic, if a bit wordy. The statistical package, SPSS, has a very usable MDS function.