# CMSC601 Research Proposal Outline

Patrick Trinkle
Dept. of Computer Science and Electrical Engineering,
University of Maryland Baltimore County,
Baltimore, MD, 21250
tri1@umbc.edu

March 28th, 2010

1. Problem Statement - you've already written most of this
2. Survey of Related work - from the 10+ papers you've read
3. What could still be done? What would be new? What's your secret weapon?
4. Why is your idea plausible? What evidence do you have, or could hope to get?
   (a) You can describe evidence that you don't yet have or even know you can get
5. What would need to be done to get from here to there?

## 1 Introduction

A primary problem with large data sets is accessing the information you specifically need to find. More complex data sets have relational structuring linking certain data together. To derive useful information from sometimes terabytes of data can require viewing and reviewing the whole picture. An effective method for information retrieval against large data sets is the iterative process of querying and refining.

Documents have many attributes of features. These features can be contained as metadata for the document including its categories, its term space, its nearest neighbors, its tags, its cluster views.

## 2 Previous Work

Todo.

- Because of refocusing I need to dive back into the literature.

### 2.1 Clustering

- Recent cluster work (SVMs)
- Relational cluster work
- Temporal cluster work

- Refining clusters (semi-supervised)
- Need to Search for Papers on Metadata Clustering
- Need to Search for cluster on demand
- Need to Search for query clustering

To build initial metadata for the documents as well as making indexes for querying and clustering, I'll be using a modernized approach using Latent Semantic Analysis [1]. Approaching certain parts of clustering with linear algebra is fairly straightforward. However, because there is relational structure and items that need to be "linked," this may be insufficient. Part of clustering will likely be done with k-Nearest Neighbor specifically, because it was shown to be a very effective somewhat quick method [2]. However, the proofs were done with considerably smaller data sets. Also, I have to determine if we're working with fully unsupervised categorizations or semi-supervised text categorization. The document category can easily be part of the metadata and is an attribute of the document.

Yin et al [3] approach the issue of relational clustering with help from the users. Data sets with relational structures may not have obvious correlations. These correlations may be easily noticed via human users. It falls onto the issue of language processing, whereby a computer may not have the semantic information to fully link information.

### 2.2 User Interface

An appropriate user interface needs to be developed to support interactive querying and clustering.

- Need to look into other adaptations of Scatter/Gather.

Alonso et al [4] extended Scatter/Gather to support user interactions and adjustments to the clusters. The prototype they built is effective, but has shortcomings if the data set is absurdly large. The view is all in a tree, listing clusters. It was important because we may be able to extend Scatter/Gather in a similar way, but more simiarly to the following paper by desJardins et al.

DesJardins et al [5] built a prototype user interface wherein information was portrayed as clusters. This system was defined as Interactive Visual Clustering and makes an attempt to cluster the data to address a user's goals. The interface makes an update to operate on user interactions with the clusters in two dimensions. The data is clustered based on its attributes, which is similar to what I'm approaching with my user interface.

There may be other important papers on this matter.

## 3 Improvement

It's been shown that there is latent semantic information in text data. Given a non-trivial large data set that has relational structure, we will build cluster information from metadata. The metadata will be built automatically by parsing the data and storing this data. Users of the system will also be able to provide metadata, via tagging. Each query against the system will query both the data and then the corresponding metadata will be clustered.

This clustering information will be protrayed to the user via an interactive user interface. The user interface needs to allow the user to interact with the clusters, but less in an edit way. The user needs to be able to rotate the view of the clusters, in 3 dimensions. The points on the sphere need to be spaced and grouped as they are clustered. The attributes associated with documents need to be editable, so that a user can add to or edit or remove features. The basic initial features for a document stored as metadata will be basic and not be solved with an NLP-hard problem. The view is superior to a list because you can browse and interact with a much larger data set without just scrolling through lists.

It may also be useful to indicate to users previously investigated and saved cluster views. Effectively if a user of the system tracks down and groups documents, and decides it's important it can be saved and easily indicated to other users of the system.

- Need to consider making images/graphics.

## 4 Requirements

Lots of equipment. Someone to work with me on the user interface. And a couple years' time.

## References

[1] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407, 1990.

[2] Yiming Yang. An evaluation of statistical approaches to text categorization. *Inf. Retr.*, 1:69–90, May 1999.

[3] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Cross-relational clustering with user's guidance. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 344–353, New York, NY, USA, 2005. ACM.

[4] Omar Alonso and Justin Talbot. Structuring collections with scatter/gather extensions. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 697–698, New York, NY, USA, 2008. ACM.

[5] Marie desJardins, James MacGlashan, and Julia Ferraioli. Interactive visual clustering. In *Proceedings of the 12th international conference on Intelligent user interfaces*, IUI '07, pages 361–364, New York, NY, USA, 2007. ACM.