

Research Statement

April Kontostathis

September, 2007

My current research is in Information Retrieval. As a PhD student, I focused on answering the question “Why does Latent Semantic Indexing (LSI) work?” Lately, I am focusing on a more important and interesting question: “When does LSI work?” The primary focus of my research is determining how the semantic information captured by LSI is defined and identifying how it is captured. LSI has been shown to work for a variety of Information Retrieval (IR) applications across a wide variety of languages, including English, French, Chinese, German, and Greek, but prior research has also shown that LSI does not improve retrieval performance (as measured by precision and recall) for all collections. A strong theoretical foundation is necessary to apply LSI in a way that consistently and correctly identifies the relationships between terms.

My research group is pursuing two main directions:

1. **Development of a theoretical basis for identifying and using term relationship data.** We are focusing on a careful analysis of dimensionality reduction applications, particularly LSI, to understand term relationships.

In prior work we have proven that the Singular Value Decomposition (SVD) algorithm which underlies LSI encapsulates term co-occurrence information. We have also shown that LSI performance correlates with second- and third-order term co-occurrences. More recently we have developed a model for understanding which values in the reduced dimensional space contain the term relationship information. We have also begun testing the model.

We hypothesize that a full understanding of the term relationships in a dimensionality reduction context will allow us to update and enhance our model for relationships between terms. Once our model is finalized, we can develop an algorithm that quickly identifies closely related terms. This algorithm will be integrated into a variety of information retrieval applications.

2. **Use of undergraduate students to develop and evaluate information retrieval applications.** In prior work, the PI developed a prototype system that correctly identifies trends in several small document collections. Indro De, Ursinus 2005, De tested this system on the Topic Detection and Tracking 3 (TDT3) datasets and was able to validate the process. Harry Schwartz, Ursinus 2007, has worked on enhancements to this application.

Daniel Waegel, Ursinus 2006, developed the Text Mining Operations Library and Environment (TextMOLE) application. TextMOLE is designed to quickly analyze a corpus of documents and determine which parameters will provide maximal retrieval performance. Enhancements to TextMOLE will depend on student interests.

Scott Kulp, Ursinus 2008, is currently exploring methods for improving search and retrieval in the Legal domain by participating in the summer 2007 in the Text Retrieval Conference (TREC) Legal competition. Scott has developed innovative ways to modify queries to improve retrieval performance, and has also explored some promising new methods for document normalization. Scott also has looked into the efforts of OCR detection methods on retrieval performance.

A critical piece of my research is the suitability of these research activities for undergraduate research experiences at a liberal arts college, like Ursinus College. Ursinus has a long history of undergraduate research, and, in fact, does not offer courses in the summer so that professors can focus on undergraduate research during those months. Furthermore, all students at Ursinus are required to complete an “independent learning experience,” and many choose to complete a research project for credit.

Many textual collections are very large and thus cumbersome to process. To complete evaluation of the term clusters developed by my algorithm using large data sets, I will need to develop a parallel version of the program. In the future, I would like to expand my research focus to include the development of parallel algorithms for other text mining and data mining problems. Finally, my 13 years of industry experience provided exposure to formal methods for project management and processing engineering. Thus, the field of software engineering holds a particular interest for me, and I plan to expand my research to include this field also.