

# **Anupam Gupta**

**Assistant Professor**

**Computer Science Department**

**SCS**

**August 29<sup>th</sup>, 2007 3:00pm**

## **Research Topic**

Dimensional Reduction

## **Research Problem**

How do we do dimensional reduction on very large sets (100,000 elements or more) with an acceptable level of loss?

## **Problem Statement**

Given very large sets of points or values of some kind, construct an algorithm for performing dimensional reduction that will result in a set whose values will not exceed a predefined maximum level of data loss.

*Very large sets:* collections of many elements of some type, usually containing (in this context) 100,000 elements or more.

*Dimensional reduction:* A process by which the elements within a set are consolidated into a smaller set, where similar elements are matched with, or merged into, a single element which represents all of the similar elements. This effectively reduces the size of the set but retains a reasonable representation of the original set.

*Data loss:* removing elements from a set results in the loss of the information associated with each element which is removed.

## **Problem Description**

Given a set of points and distances between those points, how do we best process questions dealing with those distances? For example, how do you do the near-neighbors problem given a set of millions (or billions) of points? The resulting set considers the set of connectors, with a dimension of 100,000 or more. The research looks into dimensional reduction to map the dimensions to much smaller and manageable number, such as 10. The goal of the research is find algorithms that find the smallest number of dimensions such a set could be mapped to with a level of acceptable loss.

## **Computer Science Perspective**

This research deals with relationships between objects, such as all the documents on the web, and how to effectively manage data that size. If this research is successful, new ways to sort through and relate large amounts of data will be discovered, allowing for far more accurate storage, retrieval, and searching.

## **Disciplines actively involved**

Complexity theory; Mathematics; Algorithms; Graph Theory; Network Theory.

Actively Involved Discipline: a discipline in which further research can result as a result of the successful completion of this research.

**Description of Disciplines Involved**

This research deals with algorithms, complexity theory, graph theory, and network theory many of which have a strong mathematical basis.

**References**

Presenter web page:

<http://www.cs.cmu.edu/~anupamg/>

Ultra-Low-Dimensional Embeddings for Doubling Metrics  
dim\_red.pdf (found in the online DB)

**By** mnovakou

**Updated** mnovakou