CS 194 Project Proposoal

Determining Relatedness of Sets of Data Using the Fast Johson-Lindenstrauss Transform

Advisor: Professor Rafail Ostrovsky

Justin Lu

The Problem

- How, given large sets of arbitrary data (e.g. text from webpages), to determine their relevance to one another
- Has applications in things like better searching and datamining
- How to deliver a good level of quality while still remaining computationally feasible

The Focus

- Will focus on implementing a solution to relating sets of textual data (e.g webpages, articles etc...) to each other
- But, hopefully, will yield some general results applicable to other scenarios as well

The Idea

- Use a bit vector to model each piece of data
- Each field may represent the presence or lack of a particular element, in this case, it can be a word in English
- Each vector then represents a point in n-space, where n is the length of the vector
- Vectors can be averaged to represent a set of data, which can then be compared
- Unfortunately, for practical purposes, n need be very large

The Idea

- The Johnson-Lindenstrauss lemma says we can reduce the dimensionality of n points to $k = O(\varepsilon^{-2} \log n)$ dimensions while incurring a distortion of at most $1+\varepsilon$
- The running time of this with previous methods is $O(d\,\varepsilon^{-2} \log n)$
- However the Fast Johnson-Lindenstrauss Transform has a running time of $O(d \log d + \varepsilon^{-3} \log^2 n)$ for embedding into l_1^k

The Goals

- Short term
  - Implement the FJLT and test it on some simple data
  - Utilize the FJLT to implement a comparison system for textual data as described above
  - Get some preliminary results on simplified data sets; verify that it works in a general sense
  - Time frame: End of fall quarter

The Goals

- Long Term
  - Improve performance in terms of speed
    - May need to change things to make it feasible, hard to know until the first part is complete
  - Improve quality of results
    - Simply using word counts isn't likely to produce good results, may need to tailor the process to the date
    - Explore other options, perhaps giving elements different weights or including other pieces of data beside word usage
  - Time frame
    - Hard to ascertain, tentatively end of winter quarter, but, as Professor Ostrovsky commented, this may evolve to become a longer project