CALIFORNIA STATE SCIENCE FAIR
2010 PROJECT SUMMARY

Name(s)

Dylan Freedman

Project Number

S1603

Project Title

A Novel Approach to Text Compression Using N-Grams

Abstract

Objectives/Goals

This project investigated a novel approach to text compression using principles of natural language. After observing traditional algorithms compress small text files and yield inefficient results due to a lack of useful redundant content, I wondered whether compression exploiting redundancies inherent in human language could offer any advantages.

Methods/Materials

To implement this concept, I first obtained a copy of Google's English N-Gram database, a comprehensive linguistic model for examining how often commonly observed sequences of words occur. To extract useful information, I optimized this database and sorted it alphabetically and by frequency so that information could be retrieved efficiently through a binary search. I then wrote an undemanding program able to quickly deduce the relative probability of a word occurring given a few preceding words as context. Compression was achieved by first converting each word of an input file into a ranking sorted by the word's respective probability of occurrence compared to other words that could have occurred. Then, preexisting compression algorithms were applied to the resultant rankings of the encoded file.

Results

This algorithm significantly outperformed multiple existing compression algorithms, working particularly well in comparison with other methods when compressing small English text files. Even files less than a few hundred bytes compressed to an average of 25% of their original size, an unprecedented ratio.

Conclusions/Discussion

With increased global dissemination of small text files such as cell phone text messages, emails, and chats, this method implemented properly could significantly reduce the environmental strains of data storage worldwide.

Summary Statement

My project investigates an effective, novel natural language approach to small text compression using n-grams from a comprehensive linguistic database

Help Received

Professors Craig Martell and George Dinolt from Naval Postgraduate School reviewed my project and algorithms to validate the effectiveness and soundness of my approach