Abstract of Paper to be published in the Proceedings of
Complexity, Cybernetics, and Informing Science and Engineering PORTUGAL- -
an IIIS Conference collocated with InSITE

Statistical Properties of Alpha-Beta Coding

Vilius Normantas
Institute of Mathematics AS Republic of Tajikistan,
Dushanbe, Tajikistan

vilius@norma.lt

Abstract

In this paper a type of text coding is presented. The essence of alpha-beta coding is that letters of every word-token of a given text are sorted in a specific way to create a code of that word. List of codes obtained by scanning text corpora is stored in a database together with words that could be transformed into each code and word frequencies. Decoding is performed by transforming scrambled words according to the algorithm of the coding and finding the most frequent word corresponding to the resulting code. As more than one word may result in the same code, decoding is inherently ambiguous. Three different types of alpha-beta coding have been studied: A - all letters are being sorted; A1 - all letters, except the first; A1N - all letters, except the first and the last. I used percentage of successfully decoded word-tokens for the given corpus as a measure of quality of decoded text. According to this measure coding A1N performed best (99.8%), followed by A1 (99.4%) and A (97.4%). This type of coding could be used to introduce resilience against certain distortions of text, for example typing errors, in applications where indexing of textual data is needed.

Keywords: alpha-beta coding, text coding

Biography

Vilius Normantas is a software engineer who specializes in solutions relating to active trading in the financial markets. He holds a Bachelor's degree in Computer Science and at present moment is a PhD student at the Institute of Mathematics AS Republic of Tajikistan. Vilius hopes to merge his computer programming skills with general interest in languages to make a meaningful contribution in the field of Information Theory.