

単語抽出法による次世代データ圧縮法の開発

東京大学 情報科学科 岡野原 大輔

abracadabra → abra cad abra

VZV05226@nifty.com

<http://member.nifty.ne.jp/DO/>

目的

- ・データをより圧縮したい。
- ・復元時に、高速・低メモリ使用・軽い実装

背景

- LZ法 × 圧縮率
- BWT法、PPM法 × 低速 高メモリ使用
重い実装

最新の自然言語処理技術・パターン認識技術をデータ圧縮に

- ・Suffix Arrayを用いてデータを最も圧縮率が高くなるように高速に分解
このアルゴリズムは、前提を用いていないので、あらゆるデータに適用可能
- ・分解されたデータをClass Modelを用いて分類、Trigger Model によって共起関係を調べ、データ構造を抽出。

特徴

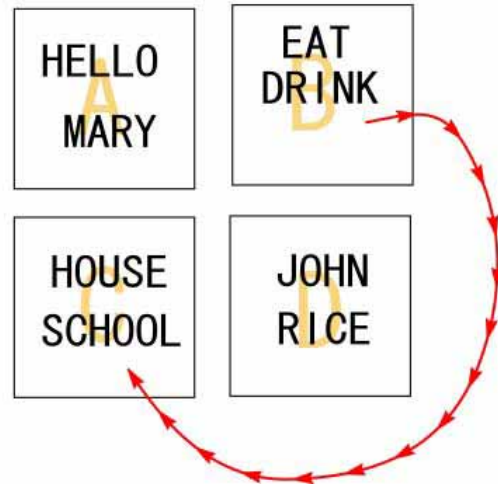
圧縮時に、複雑な計算、準備を全て行い、復元時には簡単な計算、少ないメモリで行うことができる。

単語抽出法による次世代データ圧縮法の開発

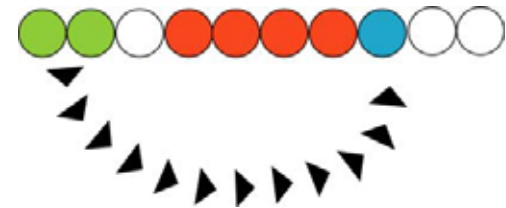
abracadabra

↓
a
ab
abra
abrac
abracad
abracadabr
acadabra
acadabraa
acadabraab
adabra
adabraa
adabraab
bra
braa
braac
braacad
braacadab
braacadabra
cadabra
cadabraa
cadabraab
dabra
dabraa
dabraab
ra
raa
raab
raacad
raacadab
racadabra
racadabraa
racadabraab

(1) Suffix Array
から単語を抽出

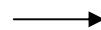


(2) 抽出した単語をClass Model
により自動分類



全ての部分において、いかなる前提知識、ヒューリスティックな情報を用いていないので、あらゆるデータ(ゲノム情報、XML、バイナリデータ)にそのまま対応できる。

データ圧縮のみではなく、自然言語処理、ゲノム情報解析などに用いることができる。



平成15年度未踏プロジェクト
「汎用的データにおける、確率的言語モデルの抽出、その利用」開発中