平成15年度未踏ソフトウェア創造事業

汎用的データにおける確率的言語モデルの抽出及びその利用

－ データマイニングの新たな可能性 －

岡野原　大輔　(東京大学　理学部　情報科学科)
e-mail:　VZV05226@nifty.com
website: http://member.nifty.ne.jp/DO/index.htm

あらゆるデータに対し、用意された辞書や文法など構造情報などを使わずに、そのデータ自体が持つ統計的情報を確率的言語モデルによって解析し、データを最小構成要素 (Unitと呼ぶ) に分解(WX法)、分類 (Class Model) 、そして共起関係 (Trigger Model) を抽出する一連のアルゴリズム、及びシステムを開発した。

本システムは、辞書には見られない単語を専門用語ごとに分類することや、関係のある単語対を抽出することが可能。また、この技術を応用して、復元時の計算量、領域量が少ない上、任意の位置から復元可能な非歪み圧縮アルゴリズムを開発した。