

テキスト型データの解析

森 ゼミ

岡 本 佳 子 (I00V012)

1 はじめに

現在，コンピュータやインターネットの普及によって，あらゆるデータがデジタル化・データベース化され，容易に入手・保存できるようになった。ところが，大部分のデータは構造化されていないテキストデータであり，従来のデータマイニング手法で扱うことができなかった。

一方，近年，テキストを分析するために必要なツールの進展によって，テキストデータの解析も実用段階になってきており，テキストマイニングが注目されている。

そこで，本研究では，テキストデータを収集・解析し，隠れた情報や特徴，傾向，相関関係を明らかにすることを試みる。解析には，テキスト型データ解析ソフトウェアの 1 つである WordMiner を用いて，頻度に対する対応分析を行う。

2 研究方法

- (1) テキストマイニングについて整理する。
- (2) ソフトウェアについてまとめる。
- (3) 対応分析について整理する。
- (4) 実データの解析

3 テキストマイニングについて

テキストマイニングとは，大量のテキストデータをさまざまな観点から分析し，「隠れた」情報や特徴，傾向，相関関係を探し出す技術である。同じ目的をもつ研究にデータマイニングがあるが，データマイニングで扱うデータがデータベース・スキーマによってきれいに整理されているという前提があるのに対し，テキストマイニングでは，形式化されていないテキストという生のデータからのマイニング（知識・情報を見つけ出すこと）を目的としている。

4 ソフトウェアについて

WordMiner は，社会調査や市場調査等の実施者や研究者の抱える自由記述型データ，あるいは，マーケティング・リサーチなどで収集されるテキスト型データなど，様々な場面における「現場のデータ解析」を支援するツールである。テキスト型データの解析時に発生する様々な事象を想定した記述的多次元データ解析を設計指針として，形態素解析と多次元データ解析の要素技術を有機的に結びつけた新たなデータ手法を提案している。特に，属性項目・選択肢型設問等を併用する自由回答を含めたデータ解析に適したソフトウェアである。

5 対応分析について

対応分析は、質的データを分析する多変量解析法のひとつである。集計済みのクロス集計結果を用いて、行の要素と列の要素の相関関係が最大になるように数量化し、その行の要素と列の要素を多次元空間（散布図）に表現する。

関連の強いカテゴリーは近くに、弱いカテゴリーは遠くにプロット（布置）されるため、集計表や通常のグラフ表現だけでは簡単に読み取れないようなデータの傾向を直観的に把握することができる。縦軸と横軸がクロスする中央近くに布置される場合は、そのカテゴリーが含まれる表頭（行の要素）や表側（列の要素）の他カテゴリーと比較して突出した特徴がないと読む。

6 実データの解析

ここでは、いくつかの分析を行ったうち、2003年3月卒業生を対象に実施された就職に関するアンケートの自由記述項目「後輩へのアドバイス」について解析を行った結果を報告する。

(1) データの事前処理

まず、分かち書き処理とキーワード抽出処理を行うことで、テキスト型データを分析が可能な「構成要素」に分解する。

次に、構成要素について、就職活動におけるアドバイスに関する用語をまねなく分析するよう、置換と削除によりデータのクリーニングを行った。

(2) 対応分析

構成要素（就職アドバイス）×質的変数（所属学科）の頻度に対して対応分析を行った結果、図1の布置図を得た。軸の解釈により、次のように所属学科をグループ分けすることができる。

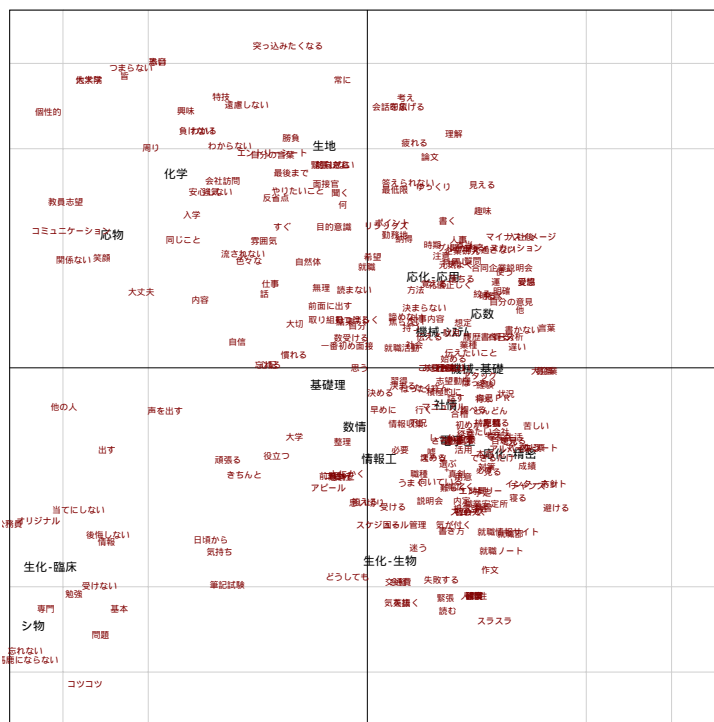


図1 構成要素×質的変数の同時布置

自分をよく見せる・試験対策（第1象限）	応用数学科，応用化学科・応用化学専攻 機械工学科・機械システム工学専攻
精神的・自分をよく見せる（第2象限）	化学科，応用物理学科，生物地球システム学科
情報を集める・精神的（第3象限）	基礎理学科，生物化学科・臨床生物化学専攻 数理情報学科，シミュレーション物理学科
試験対策・情報を集める（第4象限）	生物化学科・生物化学専攻，電子工学科 応用化学科・精密化学専攻，情報工学科 機械工学科・機械基礎工学専攻，社会情報学科

7 おわりに

テキスト型データ解析ソフトウェア WordMiner を用い、実データの解析を行った。これにより、テキストデータの隠れた特徴や関係を明らかにすることを試みた。その結果、1つ1つデータを見ていっても分からなかったであろう傾向や関係を明らかにすることができた。

WordMiner により、分析者の主観による影響度を軽減し、客観的な分析が可能となったが、データクリーニングの段階での影響は依然として残っている。データクリーニングの基準や処理方法の研究をすることが今後の課題として挙げられる。