

テキストマイニングの研究

森 ゼミ

中 濱 崇 史 (I02V059)

1 はじめに

従来のデータマイニング手法では、構造化されていないテキストデータを扱うことは困難であった。しかし、近年、技術の進歩により、日本語の分かち書き処理の技術も進み、これらがテキストマイニングに応用できるようになってきた。定性的なデータを解析するため、テキストデータの解析に関心が集まり、技術の進歩とあいまって、テキストマイニングが注目されると同時に、ビジネスの分野でも実用されるようになってきた。

本研究では、テキストデータを解析し、そこから新たな知見を見出すことを目的とし、テキストマイニングツールに **WordMiner** (日本電子計算, 2003) を用い、実際の自由記述文の解析を行うことにする。

2 研究方法

以下の方法によって研究を行う。

- (1) 自由記述文の解析について整理する。
- (2) テキストマイニングツールについて整理する。
- (3) 実データの解析をする (就活アンケートの後輩へのアドバイス)。

3 テキストマイニングについて

テキストマイニングが生まれる以前のマイニング技術は、数値・属性データのみを扱うものであった。やがて、テキストデータを分析する必要に迫られ、テキストデータを解析する技術であるテキストマイニングが生まれた。

テキストマイニングには、言語処理・データ分析・可視化の 3 つの手法が用いられる。

また、テキストマイニングは、自由記述文を解析するため、ビジネスや Web 調査などの分野に用いられている。

代表的なテキストマイニングツールの特徴を示すと、表 1 のようになる。

表 1: 代表的なテキストマイニングツールの特徴

製品名	機能				
	単体での動作	他のツールとの連携	係り受け解析	CSV ファイルへの出力	類義語の自動抽出
WordMiner	○			○	
TRUE TELLER	○		○	○	
Text Mining Studio	○	○	○	○	○
Text Mining for Clementine		○	○	○	
DeskTopClusterring	○*			○	

*ただし、インターネットサービスと共に提供される。

4 実データの解析

ここでは、いくつか分析を行った中から、2002年度岡山理科大学就職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

分ち書き処理とキーワード抽出処理を行い、テキストを分析可能な「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対する用語をもれなく分析するように、置換と削除という方法を用いて、データのクリーニングを行った。

(2) 対応分析

事前処理によって分解された、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2の布置図を得た。また、軸の解釈により、所属学科をグループ分けすると表2のようになった。



図1:構成要素と質的変数の同時布置（一部を拡大）

表2:軸の解釈による所属学科のグループ分け

自己分析・具体的な試験対策重視 (第1象限)	応用化学科・精密化学専攻, 数理情報学科, 応用数学科, 生物化学科・生物化学専攻, 電子工学科, シミュレーション物理学
自己分析・外面重視 (第2象限)	生物地球システム学科 応用物理学
メンタル面・外面重視 (第3象限)	生物化学科・臨床生物化学専攻 化学科
メンタル面・具体的な試験対策重視 (第4象限)	情報工学科, 応用化学科・応用化学専攻, 社会情報学科, 機械システム工学科, 基礎理学科

(3) 考察

多くの学科が、第1象限と第4象限に分布しており、多くの学科の学生が、具体的な試験対策に重きを置いていることが分かる。

また、同じ学科であっても、専攻によって傾向が違ってくる。

5 おわりに

テキスト型データ解析ソフトウェア WordMiner により、就職活動における後輩へのアドバイスについて書かれた自由記述アンケートを解析した。それにより、特に自由記述のような、一見ただけでは規則性を発見することが困難なデータから、知見を見出すことができた。その結果、テキストマイニングの有効性をみることができた。

なお、テキストマイニングはクリーニングの際における削除・置換辞書の作成法により、大幅に解析結果が変わることがあり、ある一定の基準の下、クリーニングを行うことが求められる。そういった方法論の研究が、今後の課題として挙げられる。