テキストマイニング

野　村　義　明

Text mining

Yoshiaki Nomura

はじめに

定量調査においても、「ご意見を自由にお書き下さい」といったように自由回答が組み込まれてることが多い。このデータは従来分析方法が確立しておらず、データとして分析されず眠っていることが多い。しかし、これらのデータには消費者または患者の生の声として貴重なデータであり、この中に多くのヒントが隠されている。

このような自由回答のデータをテキストデータとして分析する方法としてテキストマイニングが注目を浴びている。

テキストマイニングとは

言語学においてはテキストデータを分析する方法として自然言語処理が存在していた。自然言語とは日本語や英語など人が通常コミュニケーションに用いている言語でC言語、S言語等のコンピュータ言語を人工言語または計算機言語としてそれに対応した用語である。自然言語処理にはどこまでテキストを区切って分析するかによって、またどこまで意味を分析するかによって、形態素解析、構文解析、文脈解析、内容分析など様々な技法が存在する。特に助詞や助動詞の種類や頻度などから書き手のクセを判定し源氏物語の著者を推定する方法などは真贋の科学や計量文献学として知られている[1)]。

自然言語処理において日本語は英語等の言語と比較して膠着語であること、べた書きであることなどの特徴がありその処理は非常に困難であった。日本語は仮名漢字変換の開発も非常に困難であったようにコンピュータでは扱いにくい言語である。余談であるが、かな漢字変換は隠れマルコフモデルの採用によって大きく改善され実用化にいたったという経緯がある。しかし、奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座が日本語を処理するソフトを開発した。現在バージョンアップも進み無料で公開されているソフトで「茶筅」という命名がなされている[2)]。現在市販されている多くのテキストマイニングツールはこの「茶筅」をベースにしているものが多い。実際に「このような自由回答のデータをテキストデータとして分析する方法としてテキストマイニングが注目を浴びている。」という文章を「茶筅」で処理を行うと表のような結果が得られる。このようにして困難であった日本語の分かち書き処理も大きく改善され実用化される段階に至った。

分かち書き処理後の分析方法としては名詞、動詞、形容詞を中心としたキーワード抽出を行い、単語の発現頻度や単語同士に共出現の分析をクロス集計やバスケット分析により行い記述統計としてのデータを得る。その後は単語の出現を1,0で

【著者連絡先】
〒230-8501 神奈川県横浜市鶴見区鶴見2-1-3
鶴見大学歯学部　予防歯科学講座　野村義明
TEL：045-581-1001　FAX：045-573-9599

表し、行列を作成する。その後は種々の次元短縮法を用いるが、行列自体が巨大でしかも疎であることからクラスター分析やコレスポンデイング分析が用いられることが一般的である。

大学院生を対象としたワークショップ感想文のコレスポンデイング分析の結果を図に示した。四角で示した単語はユニットに与えられたテーマである。大学院の制度というテーマが与えられたユニットで有意義と感想文に述べた者がいることなど各ユニットごとに様々な傾向を読みとることができる。

まとめ

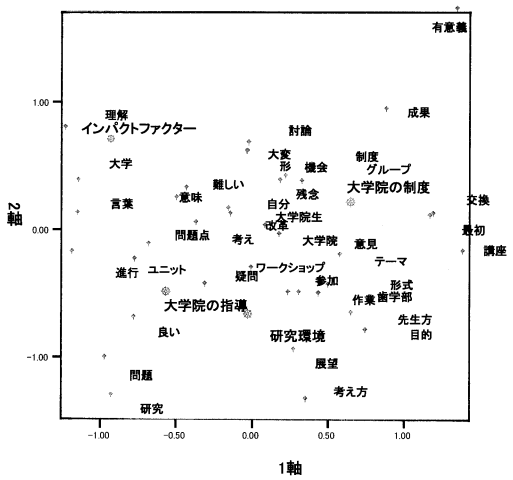
このように実用段階に至ったテキストマイニングにおいては質的研究との融合など様々な可能性を秘めた分析方法と言える。しかしその位置づけはあくまでも仮説探索型の分析であり、定量調査によって仮説を検証してゆくための前段階で用いてこそ、その威力が発揮できるものであり、定量調査に取って変わるものではないことに留意したい。

文　献

1）村上征勝行動計量学シリーズ〈6〉真贋の科学－計量文献学入門　朝倉書店（1994-09-25出版）
2）http://chasen.aist-nara.ac.jp/

茶筅による分かち書き処理の結果

この	連体詞
よう	名詞-非自立-助動詞語幹
な	助動詞
自由	名詞-形容動詞語幹
回答	名詞-サ変接続
の	助詞-連体化
データ	名詞-一般
を	助詞-格助詞-一般
テキスト	名詞-一般
データ	名詞-一般
として	助詞-格助詞-連語
分析	名詞-サ変接続
する	動詞-自立
方法	名詞-一般
として	助詞-格助詞-連語
テキスト	名詞-一般
マイニング	名詞-サ変接続
が	助詞-格助詞-一般
注目	名詞-サ変接続
を	助詞-格助詞-一般
浴び	動詞-自立
て	助詞-接続助詞
いる	動詞-非自立



ワークショップ感想文のコレスポンデイング分析による単語の布置図

テキストマイニング

Text mining

Yoshiaki Nomura
（Department of Preventive Dentistry and Public Health, Tsurumi University）

Japanese language is difficult to treat by the computer because of its nature of the agglutinative word or writing without spaces. However, with the development of free software Chasen, these problems were overcome. In this paper, the concept of the text mining and some examples of the text minig were presented. Text minig is a attractive method and powerful tool to analyse the textual data, however, its should be used for exploring the hypothesis.