データマイニング講座

講師：清水裕士

Word Miner を用いたテキストマイニング

近年、テキストデータの分析手法として「テキストマイニング」と呼ばれる方法論が注目を集めています。本講座では、主に WordMiner と呼ばれるテキストマイニング専用ソフトウェアを用いた分析手法について解説し、その意義と限界について触れたいと思います。

◆テキストマイニングとは

○これまでのデータ分析と何が違うのか

・データには大きく分けて、質的データと量的データの２種類がある。
　量的データ：間隔が等しい数値が用いられる。多変量解析に向いている
　質的データ：カテゴリーやテキストが用いられる。意味の分析に向いている

・質的データの分析手法･･･エスノメソドロジー、KJ 法、グランデッドセオリー
→研究者の経験や主観によって分類・解釈を行う
　つまり、研究者の力量や関心によって得られるものが違ってしまう可能性がある

・テキストマイニング･･･テキストデータを「誰がやっても極力おなじ結果が見出される」手法で分析することができる
→研究者の力量を問わず、得られるものが「大体」同じになるのがメリット

※テキストデータを、量的に分析・解釈することができる方法

○どんなときに効力を発揮するのか

・とても数値ではとりだせないような繊細なデータをとりたい！
・すでにテキストとしてあるデータ（歴史的資料など）を分析したい！
・テキストデータと数値データの関連を見たい！
・これから尺度を作りたいが、その前に予備分析がしたい！

などなど

つまり、これまでお手上げだった（面倒である、分析手法がわからない、客観的な分析ができない、など）テキストデータを積極的に分析し、多変量解析に応用できる。

データマイニング講座

講師：清水裕士

〇どうやって分析するのか

・テキストマイニングは、まずテキストを単語に分解するところから始まる
→テキストを形態素に分解する（茶筅などの機能を活用）

形態素分解の例
私 / は / 清水裕士 / である /。

・分解した単語の頻出を 0・1 でコード化する
→つまり、単語頻出の有無をカテゴリーデータに変換する

・0・１コードに基づき、クロス表を作成する
→変数×回答者のクロス表･･･テキストマイニングでは回答者もカテゴリーとみなす

・クロス表を基に数量化分析を行い、構造を明らかにする
→林の数量化理論や、対応分析、等質性分析を用いる

・数量化分析によって、回答者のテキストに数値をあてはめる
→他の量的データとの分析ができるようになる

〇どこまでわかるのか・主張できるのか

・個人に特有であるテキストデータを、多くの人との共通点を見出すことができる
→事例的な分析にならず、ある程度一般化可能な主張ができる

・テキストを単語に分解し、その共頻関係を見出し、構造化する
→実はテキストの意味そのものを扱っていない。また、情報が単語に還元される

しばしば質的研究で指摘される、「事例研究だ」という批判は退けることができるが、テキストデータを「そのまま」分析するこれまでの手法に比べ、情報がかなり落ちてしまうことを意識する必要がある。

※KJ 法のような意味分析が客観的にできるのでなく、頻度を分析している点に注意
※丁寧なデータ処理によって、分析結果はテキストの意味を反映したものに近づく

講師：清水裕士

◆WordMiner でテキストマイニングを行う

○WordMiner(ワードマイナー)とは

・日本電子計算株式会社が作成した、テキストマイニング専用ソフトウェア
　→大隈先生が中心となって開発された

・比較的安価（15 万）で、テキストマイニングを実行するのに十分な機能を備えている
　→分析結果のエクスポートも簡単にできることも評価できる

○どんな機能があるのか

・分かち書き（形態素分解）機能（Happiness）
　→茶筅のパワーダウン版
・質的変数の読み込み・作成
　→テキストデータだけでなく、カテゴリーデータも分析可能
・構成要素変数のリファイン
　→分かち書きしたデータを分析しやすいようにきれいにする作業
・クロス表作成・有意性検定（正規近似した等頻度検定）
　→0・1 にコード化されたデータをクロス表にする機能　検定も可
・対応分析（等質性分析）
　→いわゆる、数量化Ⅲ類を実行する機能　回答者のデータを連続量に変換できる
・クラスター分析
　→数量化した単語をクラスター化し、さらにクラスター変数を作成する

○分析の方法

・データ入力
　→エクセルにデータを入力
　　ID、性別などのカテゴリーデータ、そして自由記述データを各セルに入力
　　欠損値は空白ではなく、ピリオドか何か、わかりやすいものを入れる
　　1 行目には変数名を入れておくと便利

・データの読み込み
　→csv 形式で保存して、読み込む

・分かち書き
→Happiness という形態素分解ソフトが起動（正直なところ、性能はよくない）
分かち書き処理とキーワード検出処理の二つがある

・データのリファイン
→分かち書きデータを整えて、データをきれいにする作業
意味不明な単語がでてきたとき　→　「コンコーダンスと検索」で原文を確認
わかれてほしくない単語　→　分かち書き回避機能
意味が近いのでまとめたい　→　置換辞書の作成と保存
いらない単語・記号がある　→　削除辞書の作成と保存

※置換・削除辞書の保存は、手続きの明確化のために重要である

・クロス表の作成
→リファインしたデータの頻度を回答者・カテゴリーごとに表示
その後の数量化分析の結果を解釈するためにも、一度見ておくことが重要

・多次元データ解析
→クロス表に基づいた対応分析を実行し、クラスター化も行う
分解した単語同士の関係を明らかにすることができる
クラスター化して、データを要約することもできる
数量化したデータをもとに、他の質的データとの関連を見ることもできる

〇WM におけるテキストマイニングの応用的活用法

・データを自由記述そのまま入力するのではなく、KJ で分類してから WM で分析する
メリット
情報を単語に還元せずに分析ができる
意味的な同一性をある程度確保して分析できる
分かち書きやデータのリファイン作業労力が大幅に軽減できる
デメリット
せっかくテキストマイニングを行うのに、研究者の主観が混在してしまう
データ量が多いと、KJ 自体が大変になる

※分類というよりは、同一の意味を事前にまとめておく、というぐらいがよい？

・テキストではなく、単語でデータを収集する
メリット
リファイン作業労力を大幅に軽減できる
単語分解による情報の消失をふせげる
デメリット
収集できる情報量がそもそも少ない

・テキストデータとカテゴリーデータを併合して、同時に分析する
メリット
カテゴリーデータは影響力が大きいので、軸を固定化することができる
結果の解釈がしやすい
デメリット
カテゴリーデータに依存した結果である

◆テキストマイニングの実際

〇テキストマイニングを用いた研究

・集団における社会的表象の研究
岡本・藤原(2006)の SIMSOC における自由記述データの分析
・死生観の研究
藤井(2005)の、死へのイメージの分析
・「社会」のイメージについての研究
小川・斉藤(2005)における、中学生の自由記述データの分析
・ジェンダーの研究
湯川・清水・廣岡(2006)における、ジェンダー語からの連想の変容の分析

〇テキストマイニングは有効か！？

・テキストデータの分析における、手続きの明確化・客観化
→質的分析の熟練者というより、これからはじめる人にオススメ
・かなりの情報量を捨てていることを意識する必要
→単語への還元、対応分析・クラスター分析による要約･･･

しっかりと限界を意識して使えば、説得力のある分析を行うことができる！