テキストマイニング

−最新技術動向と応用事例−

日本行動計量学会第33回大会
特別セッション

石岡 恒憲 先生（オーガナイザ）
大学入試センター

2005年8月26日～29日
長岡技術科学大学

大隅 昇
統計数理研究所
ohsumi@ss.iij4u.or.jp

All rights reserved. Copyright by Noboru Ohsumi, ISM Professor Emeritus.

各トークに対して＜意見・感想＞を…

- 「ユーザ，データアナリストの視点」から考察，コメントしたい（「データ科学」の観点から，「データ」とは？）
- 個々の話題，4件に共通したコメント（TMとして共通項の整理）
- 一般的なテキスト・マイニング（TM）についての私見
- 事前にいただい予稿と文献の範囲から，スライドは記述
- 討論者として，立ち入った数理的，論理的な部分は十分に理解・消化できていない（知識と学習が不十分）
- 時間が限られているので一方通行，消化不良，となるおそれ

発表演題は4件

- テキスト･マイニング技術の動向

  - Key semanticsマイニング，動的トピック分析によるKnowledge Organization -

  森永 聡（日本電気株式会社　インターネットシステム研究所）

- 関連性理論に基づく文章内容の自動図解化

  石塚 隆男（亜細亜大学経営学部）

- アカデミックライティングへのJess導入の試み

  井上 達紀･佐渡島 沙織（早稲田大学大学院アジア太平洋研究科）

- 文科系学生が作成した投稿文の統計的な分析とその結果を活用した学習事例

  生田 和重（徳島文理大学文学部），　石岡 恒憲（大学入試センター研究開発部）

- **いずれも，大変に興味ある話題（大いに参考になった）．**

4件のトークの内容

- 始めの2件（話題，キーワードが多すぎる）
  - テキスト・マイニング（TM）の「ある部分」のレビュー・紹介
  - 主に自然言語処理の技術要素・技法の統合化，実装化
  - 文脈評価（文章内容理解，文脈分析），図解化の提案と応用例
- 残り2件はJess（Japanese Essay Scoring System）を用いた事例
  - 小論文，投稿文の評価分析へのJess応用例
- 全体に，比較的限定された範囲の話題に見える
- TMの守備範囲はさらに広い，茫漠としている（私見）
  - 方法論：データマイニングの亜種（？），様々な関連技法の適用
  - 電子的なデータ取得が容易となった（TMかどうか意識することなく）
  - 多様な分野に，多様な形で展開　⇒　例：質的心理学研究など
  - 古典的な（？）「質的研究」の見直し，あるいは変容（電子化が関係）

例えば，TMに関連あると思われる話題・範囲として，…

- 定性情報の分析・質的研究（Qualitative Research）の見直し
  - 市場調査，…
  - エスノグラフィー（集団観察他）
  - 福祉・看護，…
  - 質的心理学研究，…
- 内容分析（Content Analysis）
  - 研究の長い歴史がある分野
  - CACA（ Computer-assisted Content Analysis）
- コーディング処理
- データ取得環境の観点からは，…（電子的取得，多様化）
  - 調査の自由回答，FG・OFG，…
  - ディスコース分析（発語・発話分析）
  - 日記形式（看護，福祉，…），ブログ，聞き取り調査，…
  - コールセンター，コンタクトセンター，…

「データの様相」が多様化，これをどう考える

- 電子的取得　⇒　何でも集まる，集められる（そうみえる）
- 入力データの多様性から，TMを一元的に議論できない
- 「意図的に集める」vs「既に集められたもの／集まってくる」
- 「構造的」vs「非構造的」
  - 構造的　⇒　RDB，DWH，…（なかなか得られない）
  - 非構造的　⇒　E-mail・Web，コール／コンタクト・センター，…
- 「等質」vs「非等質」
- 本日のトーク内容からも，このことを実感
- TMでは，データの様相と分析方法のミスマッチ
- どう取得したか（収集方式），何のためか（目的）（データ履歴）

各発表に共通した技術要素を見ると，…

- 自然言語処理の諸要素
  - 形態素解析，構文解析，係り受け処理，共起処理，…
  - 部分順序木処理（組み合わせ論的？）…
- 統計的・確率的方法論，その関連指標化
  - マルコフ・モデル／N-gram, bigram, trigram,…
  - 情報量を測る諸指標（確率的コンプレキシティ，石塚氏：情報量，…）
  - 特徴度，…
- 多次元データ解析手法
  - LSI: Latent Semantic Indexing ⇒ データ表（単語の関連行列）のSVD他
  - 混合分布モデル（mixture model），他
- 各種の記述統計的指標
  - データ加工処理の過程で登場する諸指標
  - Jessで言うメトリック指標各種，他
  - 語彙数，文書数，語彙生起頻度…（df，tf-idf，…）

現状のテキスト･マイニングのある部分が見える

「計量化」による探査

※種々の分析手法の流行？

数値型データ
質的データ（名義，順序）
量的データ（区間，比例）

方法論の諸要素の組合せ

- 自然言語処理技法の諸要素
- 統計的･確率的方法論，その関連指標化
- 多次元データ解析手法
- 各種の記述統計的指標，…
- 概念モデル，付帯する諸条件

特徴，傾向，規則性
単純なﾊﾟﾀｰﾝの発見
計量化の限界？

利用者の要求とのギャップ？

- ﾃﾞｰﾀの量，質，型
- 構造化か非構造化か
- 加工･変換による情報の変容

テキスト型データ
非数値型データ

ユーザの期待･理想（将来）
画期的な発想が必要？
（こうできそうとの過剰期待）

真の目標
意味ある複雑な情報の発見
真の内容分析
意味ある知見･知識発見

データ科学（林･大隅）
①実験の計画
②独自のデータ取得法
③分析方法論，モデル化

※①，②が欠けていないか？

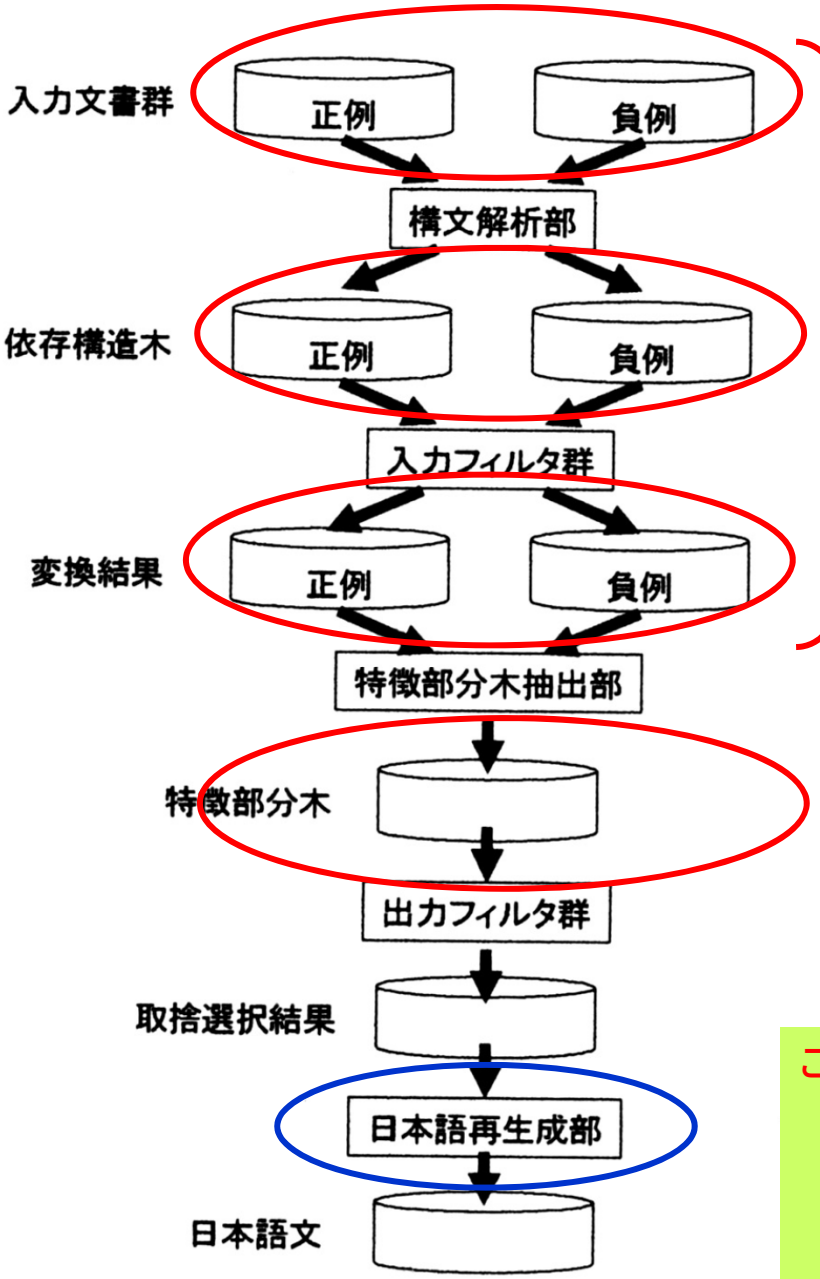
少しでも接近する可能性があるのか？

暫定策がデータ科学からのアプローチか？

1. テキスト・マイニング技術の動向

森永 聡（日本電気株式会社　インターネットシステム研究所）

- 副題：Key semanticsマイニング，動的トピック分析による Knowledge Organization
- この分野の先端研究成果をご紹介
- 主に自然言語処理系の技術要素・技法の統合化
- 面白い着想が多々ある　⇒　知りたいことも多々
- システムとして実装化（商品化）に意義（もの作りは大変）
  - Key semanticsマイニング　⇒　SurveyAnalyzer
  - 動的トピック分析　⇒　TopicScope



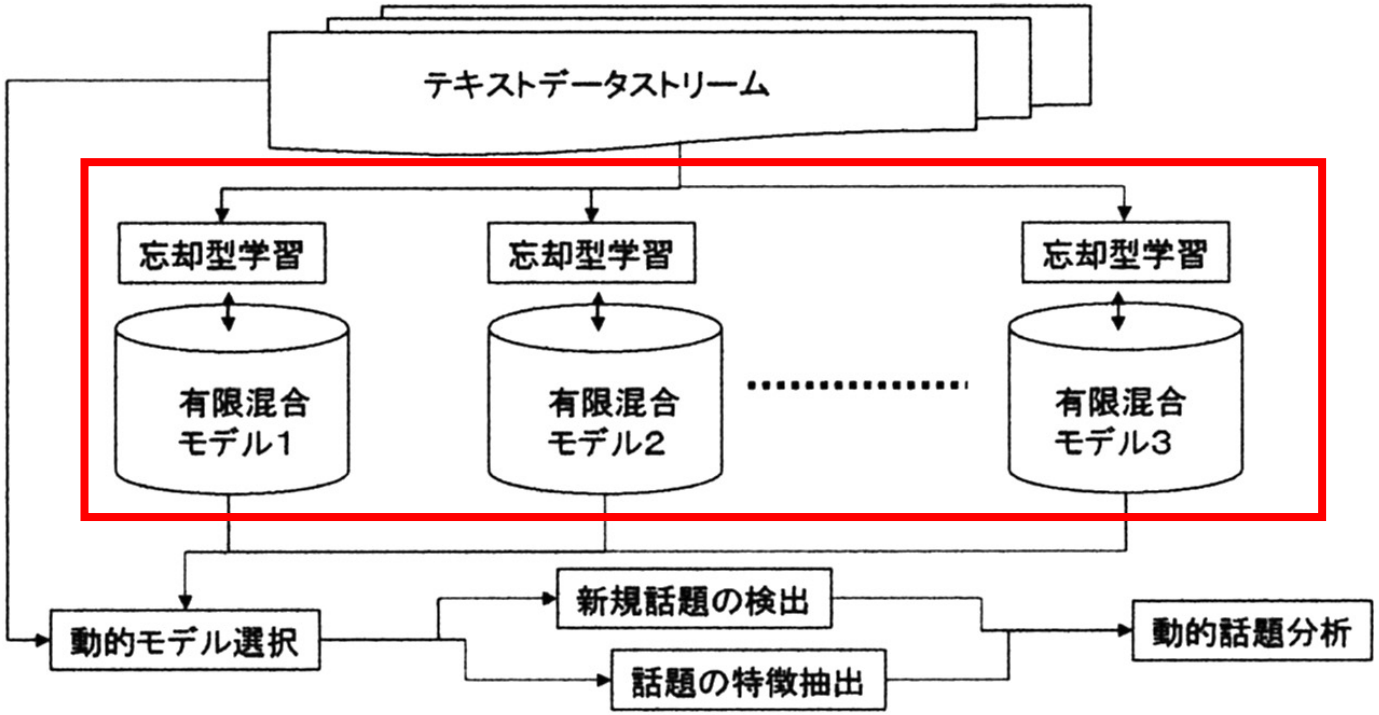
- Key semanticsマイニング
  - 形態素解析，構文解析
  - 入力フィルタ処理
  - 特徴部分木抽出部
  - 出力フィルタ処理
  - 日本語再生部

データ入力・選択の方法は？
データの作り方，選び方（正例，負例）
形態素解析，構文解析，辞書ツール等の適用法

組合せ数の増大，解除？
大量データの処理可能性は？

この発想は面白い．
候補文の選択は？
それをどう測るか（bigramなどの機能）？
（解釈の）客観性の担保
入力データへの依存性？

SurveyAnalyzer ※IBIS2004から引用



TopicScope

※IBIS2004から引用

- 動的トピック分析
  - トピック構成の同定
    - 忘却型／忘却学習によるトピック構成
  - 新規トピックの検出
  - 動的モデル選択（mixture modelの動的利用）
  - トピックの特徴抽出

•この着想も面白い
•時間軸にそったオンライン分析
•課題として，…
　　パラメータが多い
　　どのようにチューニングするのか
　　最適モデルの時間的遷移の制御

感想：テキスト・マイニングのある部分に焦点

- 種々の方法論，要素の統合体としての機能評価
  - 全体が統合化（integrate）されたときに，どう機能するのだろうか
  - システムのパーフォーマンス，スケーラビリティ，…
  - 各過程・段階での処理内容・調整が結果にどう影響するかをどう知るか
    （各モデル内パラメータ調整，各種指標の調整，…）
  - 例：分布混合モデル（mixture model，1960年代後半登場）の諸パラメータ
    - コンポーネント数（＝トピック数） ⇒ 時間軸に沿って変化？
    - 混合比パラメータ推定
    - 所属確率の評価 ⇒ クラス判定に関係
    - データ構造の仮定（正規分布，他），分布の仮定がシビア？
      ⇒テキスト・マイニングで扱うデータの多くは歪みのある分布
- 結果解釈の客観的な評価から知識へどうつなげる（⇒どう使う？）
  （⇒討論者の理解・知識不足）
- 大規模・大量データの処理への適用可能性

2. 関連性理論に基づく文章内容の自動図解化

石塚 隆男（亜細亜大学経営学部）

- 概念的，仮説的な論理展開と簡潔すぎる事例（とみえる？）
- 「文章内容理解」「文脈を分析」という意図（かなり壮大な目標？）
- テキスト（長文，平文形式）の文章内容の自動図解化・可視化
- 文章構造の同定化と構成要素の確定（要素の関連性）
  - パラグラフのつながり・流れの追跡，文の構造を知る⇒パラグラフ間の関連？
  - エンティティ（実体） ⇒ 重要単語・語句・キーフレーズの抽出
  - 関連性理論 ⇒ 認知行動，語用論⇒文章処理労力最小／文脈効果最大
- 「単語×段落」マトリクス（パラグラフの分析）を使った例示（ここは具体的）
  - 対象テキストの形態素解析とパラグラフ別の単語出現表
  - 情報量指標で測る⇒文章構造の同定と可視化が可能とあるが，そう簡単か？
- 自動図解をどう客観的に評価・解釈するのか ⇒ 知識へ？
- 一部，「内容分析」の諸研究との類似性を感じる
- 第1のトーク（森永氏）のトピック抽出にも関連する要素あるか？

後半2件はJess（日本語小論文採点システム）の応用例

- Jess（Japanese Essay Scoring System）の構成の確認
  - 引用：石岡恒憲・亀田雅之（2003）：コンピュータによる小論文の自動採点システム　Jessの試作（計算機統計学，第16巻，第1号）
- Jessの内部的な構造は正確に理解できていない
- ２つのトークに共通したこと
  - 小論文作成の習得方法と結果評価
  - 単文作成法の習得，比較評価
- 分析対象の範囲，目的が具体的に絞られている
- その意味で「分析・処理の見通しがよい」（成功のチャンスが大）
- TM全体から見るとやや特別な話題となる？

3．アカデミックライティングへのJess導入の試み

井上達紀・佐渡島沙織（早稲田大学大学院アジア太平洋研究科）

- 自動採点システムをJessのWindows版で実装化する例
- 個々の限定された「課題」の中の分析（TMとして効果的）
  - 第1回：思考の単位を書く（一文一義）
  - 第2回：明確な語句で書く（概念・意味範囲と限定），…，第15回
- 「Jess活用指針」の各回評価，相互関係（情報）をどう連結させるのか⇒ある種の総合指標化が必要なのか？
- 非等質な各回の目標（課題目標・内容）の違いから生じる「回答特徴の相対的比較評価」はどう行うのか
- 別の解析方法も用いた分析を行う必要性⇒回帰分析？
  - 総合評価と個人別の特徴評価の方法が必要か？
  - Jessをエンジンとした解析の仕組みの再構築の可能性
- 評定の客観性の確保，学習効果⇒Jessの寄与

4. 文科系学生が作成した投稿文の統計的な分析とその結果を活用した学習事例　　生田　和重（徳島文理大学），石岡　恒憲（大学入試センター）

- 質問文（の入力）は，どう作るのか，同じ課題提供を行う？
- 「学生の投稿文」と「一般人投稿文」の比較分析の意味は？
- 素材・題材（テーマ）が異なる内容の比較とならないのか
- 同じ学生の「投稿文」と「採択された掲載文」の比較検証が必要か
- 朝日新聞「声」欄，選者の志向・指向に依存するか
  - 例：「私の視点」欄も類似の傾向にある（経験から）
  - 朝日独自の編集方針の保持（書き手の意向から逸れるおそれ）
  - 書き手の迎合現象は？（類似：パネル調査などの疲労に類似）
- これは「調査の質問文の標準化」操作に類似する
  - 同質の質問（同じ課題内容）で比較評価する必要があるか，かつ反復的
  - 掲載率を基準は目安となるか　⇒　“記者風”の書き方がよい，となる懸念
- データ取得の設計は？

（表にはでない）共通の分析指針として，…

- いずれも，テキスト型データ（≒質的データ，定性情報）の計量化の方向
- 事前処理（データ加工から入力データ確定）
  - スクリーニング，ランドリ，…
  - 分かち書き，形態素解析，…
  - 例外処理，はずれ値対応，ゴミの除去，…
- 記述的アプローチ・初動探査
  - これも一種の計量化（指標化），測れる内容の限界
  - 各種統計値の工夫と利用の範囲，客観性の担保，…
- モデル化とそれによる解析，評価・検証，仮説・制約
  - 精緻化と適応・分析力のトレードオフをどう考えるか
  - モデルの精緻化≠分析（データ解析）の精緻さ（⇒感度，頑健性）
- 情報の表現形式の検討
  - 統計値，グラフィカル表現，可視化，…
  - SOM，Spring Graph，Category Connecting Map，…
- 総合的な考察，結果の客観的解釈から知識化へ（どんな方法？）

重要な操作
（意識すべきこと）

モデル化は
重要だが…

個々の固有技術からみた課題もある

- 自然言語処理
  - 形態素解析，構文解析，係り受け，…
  - 辞書，コーパス（コーポラ），シソーラス，…
  - 例外処理，はずれ値対策，…
  - 個別の技術要素が抱える多数の問題があるのでは？
- 多次元データ解析
  - はずれ値の手当，パラメータ設定，…
  - データの種類（型）に対応した対応
  - 量的データ対応の手法が多い（質的から量的への変換操作）
- テキスト型データ以外の情報活用（この説明が少なかった）
  - 例：個人情報，デモグラフィック要因，定量的，…
  - 情報量の補填，増大となるはず

形態素解析を「例」にとると，…

- ４件とも「形態素解析」ありきで議論している
  （認めた上で議論しているようだが問題はないのか）
- ツールによって結果が異なる（５種を比較検証した経験）
  - バージョン・更新，設計思想などの違い
  - 分かち書きの結果，品詞分類区分，不確定語の扱い，…
  - 初動探査・分析で現れる不具合への対処策（表現曖昧性，同義語，分類不能や例外処理，…）
- 換言すると，出発時点で異なれば分析結果は異なる
- これらの多様性・曖昧さと使い方の均衡（TMの特徴の一つ）

何が必要か or 指摘があって欲しかったこと？

- 現実は「**レシピ・ライク**」に対応できない場面が多々ある
- 対象・現象別に**テイラード方式**（tailored design）**指向**をとること
- 要は「**うまくできた**」だけでなく，「**失敗例**」に注目すべき
- どこに**分析の困難性**があるか（あったか）の情報の開示
- **適用可能性の範囲**（出来ること，出来ないこと）を明示
- 「**データ**」**とは何か**，それをどう考えるか？
- 現象解明の分析対象とする**データの吟味・議論**の重要性
- **現象解明の客観性**の担保は？⇒知識へ？

TMに期待する現象解明をどう考えるか？

- 現象に合った適応的なアプローチ
  - 「データ科学」の提案（林・大隅，1996～）
  - 実験の計画（design of experiment）
  - 独自のデータ取得法（data collection mode）
  - 分析方法論，モデル化（model-like approach）

この2点の議論が重要

この部分の議論に偏る

- 探索的・記述的方法の重要性
  - データ主導型であるべき
  - 常に発見的である
  - 実験⇒データ取得⇒仮説発見⇒仮説検証⇒実験⇒…
  - 実験を看過する傾向（実施の困難性もある）

このきめ細かい操作

- どう分析するか，何を使うか　⇒　何を分析しているのか

現状のテキスト･マイニングのある部分が見える

「計量化」による探査

※種々の分析手法の流行？

数値型データ
質的データ（名義，順序）
量的データ（区間，比例）

方法論の諸要素の組合せ

- 自然言語処理技法の諸要素
- 統計的･確率的方法論，その関連指標化
- 多次元データ解析手法
- 各種の記述統計的指標，…
- 概念モデル，付帯する諸条件

特徴，傾向，規則性
単純なﾊﾟﾀｰﾝの発見
計量化の限界？

利用者の要求とのギャップ？

- ﾃﾞｰﾀの量，質，型
- 構造化か非構造化か
- 加工･変換による情報の変容

テキスト型データ
非数値型データ

ユーザの期待･理想（将来）
画期的な発想が必要？
（こうできそうとの過剰期待）

真の目標
意味ある複雑な情報の発見
真の内容分析
意味ある知見･知識発見

データ科学（林･大隅）
①実験の計画
②独自のデータ取得法
③分析方法論，モデル化

※①，②が欠けていないか？

少しでも接近する可能性があるのか？

暫定策がデータ科学からのアプローチか？

ありがとうございました.