1A2-4

テキストデータの構造化を支援する対話的マイニングシステム

Interactive text-mining system for structuring high dimensional text data

根本啓一[*1] Keiichi Nemoto
大西健司[*1] Takeshi Onishi
増市博[*1] Hiroshi Masuichi

[*1] 富士ゼロックス株式会社 研究技術開発本部 コミュニケーション技術研究所
Communication Technology Laboratory, Research & Development Group, Fuji Xerox Co., Ltd.

The advancement of Information Technology and social media increases the massive amount of unstructured data in enterprises. Many systems have been developed so that enterprises can take advantage of such big data. In this paper, we focus on text as unstructured data and propose a text mining system, which supports making unstructured data into structured data in order to handle it with other quantitative data. We employ an interactive user interface in the system so that analysts can explore optimal structuring level. In addition, the proposed system enables multiple analysts to explore the data simultaneously in order to analyze the data from multiple perspectives.

1. はじめに

近年, ICT の発達やソーシャルメディアの台頭により, 大量のテキストデータから顧客や市場の要望, 課題などを抽出するテキストマイニングの研究が注目されている.

テキストマイニングの研究では, 予め定められた分類体系に従ってテキストを分類する手法[Sebastiani 2002]や, テキスト集合をクラスタリングにより集約したり[Iwayama 1995], テキスト集合からトピックを特定する手法[Blei 2003]など, テキストを分類することによって構造化する様々な手法が開発されている. 一般に, テキストを分類, 構造化した後に, 要望, 課題などの特徴的な話題をマイニング結果として抽出する処理を実施する.

大量のテキストデータの一例である顧客の声（Voice of Customer: VOC）は, 自由記述のテキストが, 5 段階評価の結果や顧客の年齢等のアンケートデータ(数値データ)とともに蓄積される. このようなテキストデータとアンケートデータの両者を利用した分析では, アンケートの数値データから統計的に得られる結果を補足するために人手でテキストが読み込まれることや, テキストマイニングから得られた要望や課題のプロファイリングとして, アンケートデータを利用するといった方法が行われている. 例えば, ある要望は 30 代の男性に多く見られるといったように, テキスト以外のコンテキスト情報と対応付けることで意味を推定しやすくなる. また, 特定のコンテキストでしか現れない話題は, テキストデータだけから発見することは難しい. このように, テキストデータから有用な話題を抽出するために, テキストのもつコンテキスト情報を参照しながらマイニングを行うことは有用である. しかしながら, テキストとコンテキストを相互に行き来しながらテキストを分類, 構造化しマイニングする方法の研究は十分に行われていない.

そこで本稿では, VOC のようなテキストデータとアンケートデータの両者を含むデータソースに対して, コンテントとコンテキスト, さらに時間軸の 3 軸により整理するシステムを提案する. それぞれの軸で, ユーザが対話的なインタフェースにより分類体系の変更を可能とすることで, 最適な分類体系の発見を促すことを目的とする. さらに, そのようなプロセスを多人数で実施可能とすることで, より広い観点から分類, 構造化を実施できるシステムの実現を目指す.

連絡先：根本啓一, 富士ゼロックス(株) 研究技術開発本部, keiichi.nemoto@fujixerox.co.jp

2. 関連研究

本稿で提案するマイニングシステムに関連する研究として, 2.1 節では対話的なインタフェースを取り入れた研究について, 2.2 節では多人数で協調的に行われるマイニング手法について述べる.

2.1 対話的データマイニング

従来の多くのテキストマイニング手法では, 事前に決められた分類体系に従ってテキストを分類, 構造化し, マイニングを行う. しかし, 分類体系を事前に決めることが難しいケースや, 分類体系を特定するプロセス自体が試行錯誤であり, 研究者ではなく, データを扱う現場の知識が必要な場合が多い. そこで, 近年ではマイニングのアルゴリズムと可視化などのプレゼンテーションによりユーザが逐次的に分析を進め, 知識発見を支援する対話的な仕組みを取り入れたシステムが考案されている. そのような手法では, 分析者が分類のためのルールを記述したタグ付けを行う方法[楠村 2008]や, システム側がクラスタリングや多次元尺度構成法などのボトムアップな手法でテキストを分類し, それに対して分析者がタグを付与するといった方法が検討されている[田淵 2009].

このように, 対話的プロセスへ積極的にデータの分析者を取り入れることで, 機械のみに頼るのではなく人と機械が協調したマイニングが可能になる.

2.2 多人数データマイニング

知識発見のプロセスを多人数で行う取り組みも行われている. ManyEyes[Viegas 2007]や sense.us[Heer 2007]では, ウェブ上でのデータの可視化を通して, 多人数がウェブ上の掲示板で議論し, 知識発見を行う仕組みが提供されている. これにより, 他者の発見をトリガーとして新たな分析視点を得ることが可能となる. また, WeFold[Khoury 2014]では, 三次元のタンパク質折り畳み構造という複雑な問題を多人数で解決できる仕組みを取り入れることで, 並行して課題解決を促進する仕組みを提供している.

テキストマイニングにおいても, 有用な分類を特定する作業は上記の例と同様に複雑なタスクであり, また, 結果を定量化することが難しいため, 計算機による自動化が難しいタスクである. WeFold のように多人数によって異なる問題空間を探索すること

で，大規模なテキストに対して有用な分類の発見を促進することができると考える．

さらに，VOC のテキストマイニングで有用な分類を発見するためには，マイニング対象の背景知識が必要である．個々の課題に対して背景知識を持った分析者が，個々の観点でデータを掘り下げ，分類する仕組みが必要であると考えられる．

3. テキストマイニングプロセス

3.1 コンテント・コンテキスト・時間軸

従来のテキストマイニングでは，コンテント(テキストデータ)のみに着目して，マイニング結果である話題の抽出が行われていた．しかし，局所的に発生する話題，例えば特定のコンテキストや，特定の時間で発生する話題を抽出することは困難であった．そこで本稿では，コンテント軸，コンテキスト軸，および時間軸の 3 軸に基づいて，テキストデータを分類，構造化しマイニングするシステムを提案する(図 1)．テキストに存在する話題は，この 3 つの軸で定義された空間のある局所点を探索することになる．以下，各軸の詳細を記述する．

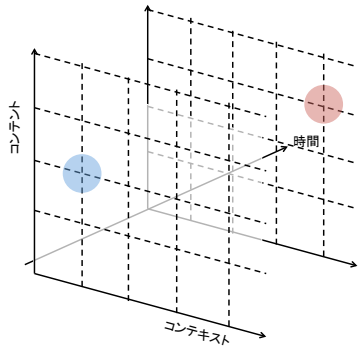

図 1 コンテント軸・コンテキスト軸・時間軸でのマイニング

(1) コンテント軸

コンテントとは，テキストデータそのものの内容を表す軸である．大量のテキスト集合がある場合，それらをある類似度に基づき，分類し構造化することで，テキストデータを集計可能とし，他の定量データとともに扱うことができる．分類の方法には様々な手法が存在し，テキストの分類も，分類体系の取り方に応じて無限に存在する．どのような分類体系で分類するかは分析者の経験や視点に依存する．たとえば，「太郎は犬を飼っている．」と「次郎は猫が好きだ．」といった文があるとき，前者は犬という分類，後者は猫という分類とすることもできるが，両者ともペットという分類にすることもできる．このように，どのような分類体系とするかによって得られる結果は大きく異なる．分類体系は分析者の分析意図や，コンテキスト軸・時間軸での分布を参照して決める必要がある．

(2) コンテキスト軸

コンテキストとは，各テキストが持つ背景情報である．コンテント自身から推定される文章コンテキスト，コンテントを発した主体から推定される主体属性コンテキスト，そして，コンテントを発した状況から推定される状況コンテキストに大別できる．文章コンテキストの例として，テキスト自体がポジティブな内容かネガティブな内容かを示す極性情報をあげることができる．主体属性コンテキストは，40 代男性のテキストといった年齢，性別などの情報があげられる．状況コンテキストは，例えば一連のサービスの利用者へのアンケート結果であれば，どのサービスの利用時のテキストであるかといった情報があげられる．

(3) 時間軸

多くのテキストデータは時間情報を保持しており，テキストデータから抽出される話題は時間によって変化している[Cui 2011]．そこで，時間軸を設けることで，ある特定の時間においてのみ存在する一過性の話題や，時間軸を通じて常に存在する話題などを個別に抽出することが可能となる．

3.2 マイニングプロセス

テキストのコンテントはコンテキストとの関係性と合わせて解釈される．分析者にとって，コンテントとコンテキストの価値ある組み合わせを得る作業は，多次元から構成される問題空間を探索することになる．そのようなプロセスは対話的なインタフェースによって提供され，結果を適時確認しながら最適な分類を発見することを目指す(図 1 の青丸や赤丸)．

また，そのような発見的プロセスを多人数で実施することで，より広範囲の空間を探索し，マイニングすることができる．

4. 提案システム

4.1 システム構成

提案システムでは，コンテント軸，コンテキスト軸，時間軸それぞれについて，分類を実施するモジュールを適時実行できる構成とした(図 2)．テキストの分類，構造化を再帰的に実行可能な仕組みを取り入れ，任意の粒度の話題を利用者が抽出可能とする．時間軸は，時系列変化を算出する仕組みを取り入れることで，特異点の抽出などを可能とする．

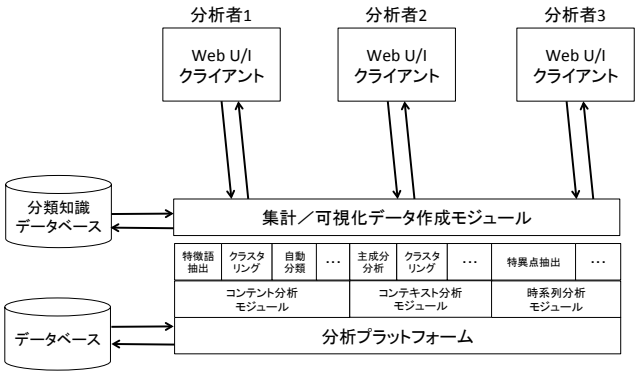

図 2 システム構成図

4.2 ユーザインタフェース

図 3 にユーザインタフェースを示す．分析者は，はじめに左ペイン部で，分析対象とするデータを絞り込む．次に，どのような分類を利用して，コンテントを分類し，構造化するかを決定する．さらに，どのようなコンテキストに着目するかを選択することで，中央ペイン部にコンテントとコンテキストの 2 軸で集計されたマトリックスが表示される．マトリックスの各セルをクリックすることで，右ペイン部に分類されたテキストが表示される．ユーザはテキストに対して更なる分類の操作を行うことが可能であり，インタラクティブに操作を繰り返すことができる．例えば，分類されたテキストから特定のトピックのものをチェックし，新たにタグ名を設定することで，下位の分類を作成できる．中央ペイン下部では，選択した分類への該当テキスト数の時系列変化を表示し，時間軸での変化を可視化する．

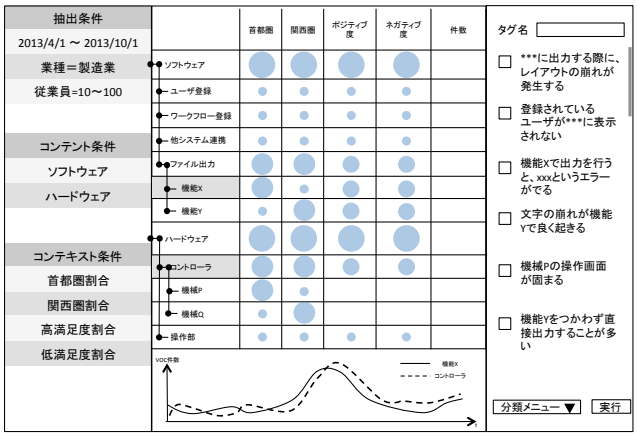

図 3　ユーザインタフェース

4.3 テキスト構造化手法

本システムでは，テキストデータを構造化する手法として，既存の様々な分類手法を取り入れており，新たな構造化手法も容易に追加することができるように設計されている．以下に，システムに取り入れた代表的な分類手法を記述する．

- 単語リストによる分類：分析者が指定した単語を含むか否かの条件でテキストを分類とする
- 係り受け関係に基づく分類：単語ペアの係り受け関係を考慮し，指定された単語ペアと係り受け関係を含むか否かの条件でテキストを分類とする
- クラスタリングに基づく分類：各種クラスタリング手法や LDA 等のトピックモデリング手法により，ボトムアップに類似した部分集合を抽出し，それらを分類のベースとする
- 教師あり機械学習による分類：分析者が文集合から選択した，同一分類とするテキストリストをもとに，類似するテキスト集合を抽出し，分類とする

4.4 ユースケース

本節では，あるソフトウェアとハードウェアからなるサービスに関する VOC データを分析するユースケースを記述する．VOC データには，サービス利用に関するテキストデータに加えて，表 1 に示す属性データが保存されている．

表 1　VOC データ構成例

軸	属性	例
時間軸	入力日	2014-02-28 14:30:10
コンテント軸	対象商品名	***ソフトウェア
	テキスト	***に出力する際に，レイアウトの崩れが発生する
コンテキスト軸	業種	製造業
	規模	300 人
	利用期間	120 日
	住所	大阪府***

分析者はあらかじめ対象とする期間やコンテキストを絞り込んだ状態で分析することができる．例えば，分析者が製造業での中手市場に関心をもっている場合，従業員規模が 10 名から 100 名の製造業のデータのみを対象とするといった操作である．次に，コンテント軸での分類を試みる．ここで，複数の分析者が異なる視点で個別に分析を行うことが可能である．例えば分析者 A はソフトウェアに関する分析を行い，分析者 B はハードウェアに関する分析を行う場合，個々の分析者が，まず特定の商品名 のリストからなる分類を作成する．同時に，着目したい幾つかのコンテキスト分類を作成する．例えば，地域毎の比較を行うため，首都圏，関西圏といった分類を作成したり，テキスト自体の極性を判別し，ポジティブとネガティブのそれぞれの割合を表す属性を作成するといった形で分類を作成する．

次に，分析者はそれぞれにコンテント軸での新たな構造化試行する．分析者 A が，クラスタリングの手法により分類された 5 つのテキストクラスタ内のテキストを参照することによって，ソフトウェアの利用フェーズ毎のクラスタに分類されていることが確認できた．例えば，ユーザを作成し登録するフェーズ，ワークフローを登録するフェーズ，ファイルを登録するするフェーズ，他システムへファイル共有するフェーズ，ファイルを出力するフェーズの 5 つの分類である．ここで，該当する件数を見ると，ファイルを出力するフェーズでの件数が多いことを知る．そこで，次にどのような機能を用いて出力しているかを知るため，出力フェーズのテキストを機能毎に分類する．機能毎に人手で分類のタグ付けを行い，教師あり機械学習によりタグ付けされたテキストと類似したテキストの分類を行う．そこから，ある機能 X を用いた出力に関するVOC件数は首都圏で，機能 Y を用いた出力に関する VOC 件数は関西圏で，それぞれ多いことが分かった．

同様な構造化をハードウェアの VOC を対象に行っていた分析者 B の結果から，機械とソフトウェアとを連携させるためのコントローラに関する VOC 件数が多く，関東圏では機械 P，関西圏では機械 Q のように，件数が機械の種類によって異なることが分かった．

そこで分析者 A と B が，ソフトウェアの出力機能の視点で作成した分類と，ハードウェアのコントローラに関する視点で作成した分類を対象に，時間軸での変化を見てみると，両者の件数増加傾向が一致した．そこで，これらの分類を統合した新たな分類を作成したところ，機能 X と機械 P のコントローラと，機能 Yと機械 Q のコントローラの組み合わせ時に，テキストの極性がネガティブに大きく偏ることが分かり，顕在化していなかった課題が明らかになった．

今回のケースでは，ネガティブ度の大きいテキストのみから話題の抽出を試みても，様々な要因が含まれているため，そこから本ユースケースで発見に至ったような特定の機能に関するテキストを抽出することは困難である．分析者が，分類仮説を立て，対話的に分類を行い，実際に分類されたテキストが持つコンテキスト，今回の例では首都圏と関西圏での偏りなどを頼りに構造化していくことや，時間軸での変化を見ることで，局所的であるが有用な情報のマイニングが行える．

さらに，このようなユースケースでは，テキストを分類，構造化するためには分析対象サービスの知識が必須である．ソフトウェアとハードウェアの分析をそれぞれ異なる担当者が行うことで，個々の仮説を生成しながら適切な構造化が可能となる．多人数での分析を支援する提案システムのような仕組みは，試行錯誤から課題を発見するという目的に対して有用であると考える．

5. まとめ

本稿では，VOC データのように，コンテントとコンテキストの両者を含むデータの集合から，有用な話題を抽出するテキストマイニングシステムを提案した．従来，コンテントのみを対象に話題を抽出する仕組みは存在していたが，そのコンテキストや時間の軸を加えた局所的な話題の抽出を行うことは困難であった．また，多次元にわたる問題空間の探索となるため，抽出を定量的にスコア化することが難しく，したがって，抽出処理の自動化も難しかった．そこで，背景知識を持つ現場の分析者が対話的

に分析を行うことを可能とし，複数の分析者が並行して探索することができる仕組みを提案した．

今後，具体的なデータに適用した実証実験を通じて，システムの有用性や課題を明らかにしていきたい．

参考文献

[楠村 2008] 楠村幸貴, 神谷俊之: 対話的テキストマイニングのためのタグ付け用検索基盤，情報処理学会研究報告，2008．

[砂山 2011] 砂山渡, 高間康史, BOLLEGALA: テキストデータマイニングのための統合環境―TETDM プロジェクト―，電子情報通信学会技術研究報告，2011．

[大塚 2004] 大塚 裕子，内山 将夫，井佐原 均: 自由回答アンケートにおける要求意図判定基準，言語処理学会，11(2), 21-66, 2004．

[田淵 2009] 田淵史郎, 鍛治伸裕, 吉永直樹. 大規模コーパスからの語義のマイニング．日本データベース学会論文誌，Vol. 8, No. 1, pp. 77–82, 2009.

[Blei 2003] Blei, D., Ng, A, and Jordan, M.: Latent dirichlet allocation, The Journal of Machine Learning Research, 3, p.993-1022, 2003.

[Boley 2013] Boley, M., Mampaey, M., Tokmakov, P., and Wrobel, S.: One Click Mining—Interactive Local Pattern Discovery through Implicit Preference and Performance Learning, IDEA'13, August 11th, 2013.

[Cui 2011] Cui, W., et al: TextFlow: Towards Better Understanding of Evolving Topics in Text, IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 17, NO. 12, DECEMBER 2011.

[Heer 2007] Heer, J., Viegas, F., and Wattenberg, M.: Voyagers and Voyeurs: Supporting Asynchronous Collaborative Information Visualization, In Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07).

[Iwayama 1995] Iwayama, M., Tokunaga, T.: Cluster-based text categorization: a comparison of category search strategies, Proc. of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, p.273-280, July 09-13, 1995.

[Khoury 2014] Khoury, G., Liwo, A., et al: WeFold: A Coopetition for Protein Structure Prediction, Proteins: Structure, Function, and Bioinformatics, 2014.

[Sebastiani 2002] Sebastiani, F.: Machine learning in automated text categorization, ACM Computing Survey, 34(1), 1-47, 2002.

[Viegas 2007] Viegas, F., Wattenberg, M., et al: ManyEyes: a Site for Visualization at Internet Scale. IEEE Transactions on Visualization and Computer Graphics 13, 6 2007.