概念索引を用いたテキストマイニング
－(1)アンケート分析への適用－

相川 勇之　伊藤 山彦　高山 泰博　鈴木 克志
三菱電機株式会社　情報技術総合研究所

1. はじめに

インターネットの普及に伴い企業内の文書の電子化および共有化が進んでいる。また、消費者による電子メールや掲示板の利用が増えており、顧客の生の声がテキストデータとして蓄積されるようになってきた。従来の全文検索では、これらの大量テキストから定性的なデータを抽出して商品開発やサービスの改善に活用することはできなかったため、全体の傾向を把握したり特徴的な部分を深く分析するためのテキストマイニングが注目を集めている[1]。

しかし従来のテキストマイニングには、顧客の類似意見を抽出するための類義語辞書開発コストが大きい、あるいは、分析結果を組み合わせて分析する機能がない等の課題があり、潜在する顧客ニーズを充分には引き出せていない。本稿では、同義性や類義性など単語間の潜在的な関係を、類義語辞書を用いずに自動的に抽出した概念索引を使い、分析結果を組み合わせて再利用できる対話的テキストマイニング方式を提案する。

2. アンケート分析の課題

顧客からの生の声を調べる方法の一つに、アンケート調査がある。図 1 にアンケート調査業務の流れを示す。テキストマイニングには図１に示した分析過程の支援が期待されている。

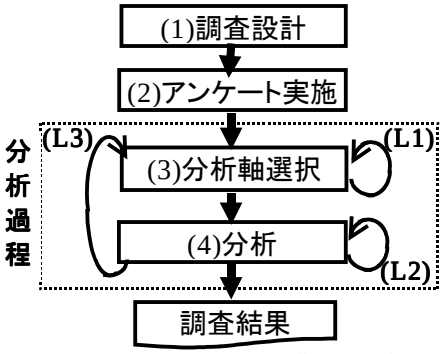

図１ アンケート分析業務の流れ

Text mining method using concept index.
Takeyuki AIKAWA, Takahiro ITO, Yasuhiro TAKAYAMA, Katsushi SUZUKI
Information Technology R&D Center, Mitsubishi Electric Corporation
5-1-1 Ofuna, Kamakura, Kanagawa, 247-8501, JAPAN.

分析過程の支援には、分析対象（アンケート回答中の自由記述テキスト）から類似意見を抽出するための類似性判定処理が必須である。従来のテキストマイニングには、単語の出現頻度に重み付けをした文書ベクトルを用いる方式や、文構造を照合することにより類似性を判定する方式がある。しかし、これらの方式では出現単語そのものに基づいて類似性を判定するため、表現の異なる類似意見を抽出するには類義語辞書が必要であり、開発コストが大きいという課題があった。

また、分析過程では、分析軸の選択(L1)や分析処理の個々の作業(L2)が繰り返されるのに加え、分析軸の選択と分析全体(L3)も繰り返される。分析過程の支援にあたり、以下の要求に応える必要があるが、既存のシステムでは、充分に応えられていない。

(1) 調査設計（設問設定）に関連付けた分析をしたい。
(2) 分析軸の抽出を支援して欲しい。
(3) 試行錯誤的な分析作業を効率化したい。
(4) 分析軸抽出と分析の反復を支援して欲しい。

3. 概念索引によるマイニング方式

3.1. システム構成

2 章で述べた課題を解決するため、我々は、線形代数演算によって単語共起頻度表から特徴的な次元を自動抽出して概念索引を生成する手法[2]を採用した。概念索引を用いた対話的テキストマイニング方式の全体構成を図 2 に示す。

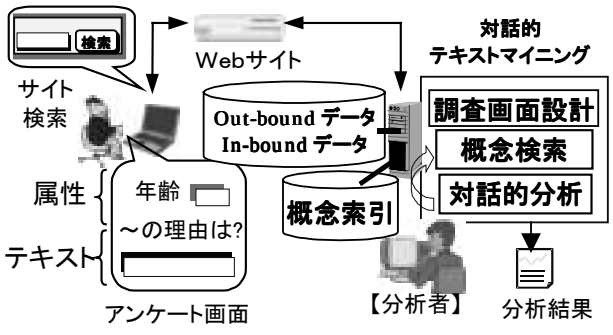

図 2 概念索引に基く対話的テキストマイニング

Web サイトにおいて顧客の声を収集し、データベースに蓄積する。分析者は蓄積されたテキストデータをテキストマイニング支援のもとで分析し、分析結果を商品開発やサービスの改善に活用する。

本方式では、分析対象のテキストから生成する概念索引を用いて分析を行なう。2 章で述べた各要求に対応して、下記の機能を有する。

(1) 調査設計に対応したアンケート画面設計
(2) 分析軸抽出のための概念検索
(3) 試行錯誤的な分析を支援するための分析結果を保存・再利用する機能
(4) 分析軸の異なる分析結果を組み合わせて別の観点で分析する機能

以下では、概念索引の生成方法、および対話的マイニング方式の概要について述べる。

3.2. 概念索引の生成

概念索引は、単語と概念との対応を格納する概念辞書、及びアンケートの各回答に対して概念を対応付ける文書索引からなる。

(1) 概念辞書の学習（図 3 の①および②）

本方式では概念辞書として、単語の共起頻度行列を線形代数演算により圧縮した概念空間を採用している[2]。各概念は、概念空間における座標を示す概念ベクトルとして表現する。

この方式では、単語の概念的な性質をその単語と共起する単語の統計として定義する。例えば、「画面」と「液晶」は「きれい」「明るい」などの単語との共起傾向が類似するので、類似の概念をもつとする。

(2) 文書索引の生成（図 3 の③）

登録対象テキストを形態素解析により単語に分割し、各単語の概念ベクトルより文書ベクトルを合成する。合成の際には、各単語に対する tf・idf 重みを適用する。

分析軸抽出を支援するための概念検索機能では、この文書ベクトルと検索文から同じく概念辞書を参照して得られる検索ベクトルとの余弦値を類似度として検索を行なう。

3.3. 対話的マイニング

アンケート分析では、選択式の回答データを集計し、クロス集計やコレスポンデンス分析などの統計処理により分析する手法が一般的である。本稿では、これらの分析で用いる選択式の回答データを属性と呼ぶ。

本方式では、文書索引に登録された文書ベクトルを合成することにより、単一の回答データだけではなく、共通の属性を持つ回答の集合に対しても概念ベクトルを定義することができる。

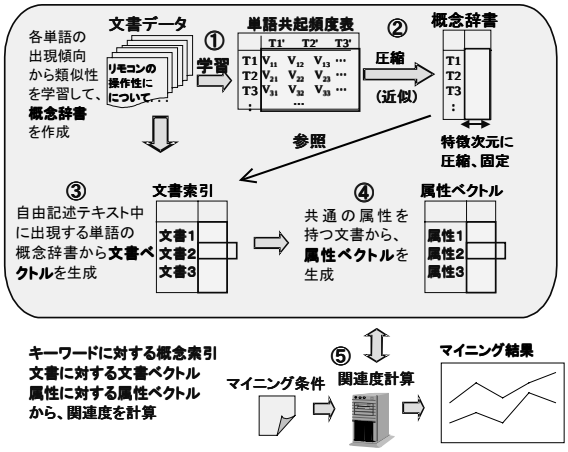

図 3 概念索引の生成と対話的マイニング

したがって、各回答データ同士の類似性だけではなく、キーワードと属性の間の相関や、属性同士の相関など、種々の関係を同一の尺度である概念ベクトルによって分析することができる（図 3 の④および⑤）。以下に分析機能の例を示す。

(1) 属性－キーワード相関
属性とキーワードを入力して相関を分析する。地域ごとの嗜好の違いなどを分析できる。

(2) 属性－属性相関
２つの属性およびキーワードを入力し、属性間のキーワードの偏りを分析する。

(3) 時系列分析
キーワードおよび期間を入力し、指定期間における注目キーワードの推移を分析する。

対話的マイニングでは、試行錯誤により得られる分析結果を保存し、分析結果を組み合わせて新しい分析の基準となる属性を作成し、以降の分析に活用することができる。

4. まとめ

テキスト形式で寄せられた顧客の声を分析する過程を支援するテキストマイニング方式を提案した。今後は、実際のアンケート調査業務での評価を進めていく予定である。

参考文献

[1] 市村由美,他："テキストマイニング－事例紹介", 人工知能学会誌, Vol.16, No.2(2002)

[2] 相川 勇之,他:"大規模検索システムにおける概念辞書自動更新",FIT2002, D-37.(2002)