情　報　学　基　礎　70－1
デジタル・ドキュメント　38－1
（２００３．３．２８）

概念抽出型テキストマイニングによるアンケート分析手法の提案

相川勇之　伊藤山彦　高山泰博　鈴木克志　今村誠

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み，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．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ためのテキストマイニングが注目を集めている．本稿では，同義性や類義性など単語や複合語間の潜在的な関係を自動的に抽出した概念索引を用いることを特徴とする概念抽出型テキストマイニングによるアンケート分析手法を提案する．本手法の特長は，概念索引により類義語辞書を用いずに類似意見を抽出できる点，および，対話的テキストマイニングにより分析結果を組み合わせて再利用できるという点である．

A proposal of a method of Analysis of Questionnaires using Text mining based on Concept Extraction

Takeyuki Aikawa , Takahiro Itoh , Yasuhiro Takayama , Katsushi Suzuki , Makoto Imamura

We propose a method of text mining to support analysis of questionnaires for marketing research. This method is called text mining based on concept extraction.　The word and the compound word associations are represented by concept vectors that automatically extracted from word co-occurrence in open answers of questionnaires.　In this method, the synonym dictionary registration is not required for text analysis and the analyzed results registered as the concept vectors are interactively re-usable for the next analysis step.

三菱電機株式会社 情報技術総合研究所
Information Technology R&D Center, Mitsubishi Electric Corporation

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や複合語間の潜在的な関係を自動的に抽出した概念索引を用いることを特徴とする概念抽出型テキストマイニングによるアンケート分析手法を提案する．本手法の特長は，概念索引により類義語辞書を用いずに類似意見を抽出できる点，および，対話的テキストマイニングにより分析結果を組み合わせて再利用できるという点である．

2. アンケート分析とマイニング

顧客からの生の声を調べる方法の一つとしてアンケート調査がある．最近では，大量の回答を迅速に得られる Web アンケートなどのネット調査が盛んになりつつある．

アンケートデータには，選択式で回答する定型項目と，自由文で回答する自由記述項目とがある．定型項目についてはデータベースに登録することで，データマイニングツールなどにより機械的な集計および分析が可能である．一方，自由記述項目については自動処理には限界があり，人手で分析するとコストが非常に大きくなるため，必ずしも有効活用されてはいない．

しかし，選択式の回答からでは，いま売れている商品に対する情報は得られても，今後どのような商品が必要とされているかをつかむことは難しい．選択式の場合は設問をあらかじめ用意する必要があるので，顕在化していない顧客ニーズを先行してとらえることは困難である．

このような潜在的な顧客ニーズを掘り起こすために，自由記述式の回答を分析した結果が求められている．そこで，人手では分析しきれない大量の自由記述の分析を支援するための，テキストマイニング技術の活用が期待されている．

3. 従来のテキストマイニング方式

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．

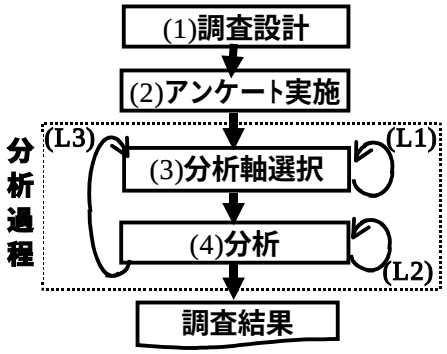


図１　アンケート分析業務の流れ

分析過程の支援で利用するテキストマイニングには，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，文書の自動分類を基本とする方法，カテゴリを記述した自立語の辞書を用いる方法，あらかじめ設定したパタンによりパタン照合を行う方法などがある [1][2][3] ．

しかし，これらの方式では分析対象の分野ごとに辞書を構成する各語にカテゴリを付与する必要がある．また，出現単語そのものに基づいて類似性を判定するので，表現の異なる類似意見を抽出するため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) 分析軸の抽出を支援して欲しい.
(2) 試行錯誤的な分析作業を効率化したい.
(3) 分析軸抽出と分析の反復を支援して欲しい.

4. 概念抽出型テキストマイニング

3 章で述べた課題を解決するため, 同義性や類義性など単語間の潜在的な関係を自動的に抽出した概念索引を用いる概念抽出型テキストマイニングによるアンケート分析手法を提案する。

4.1. 概念索引に基づくテキストマイニング

本手法には, 自動抽出した概念索引により, 類義語辞書を用いずに類似意見を抽出できるという特徴がある. 図 2 に概念抽出型テキストマイニングの概要を示す。

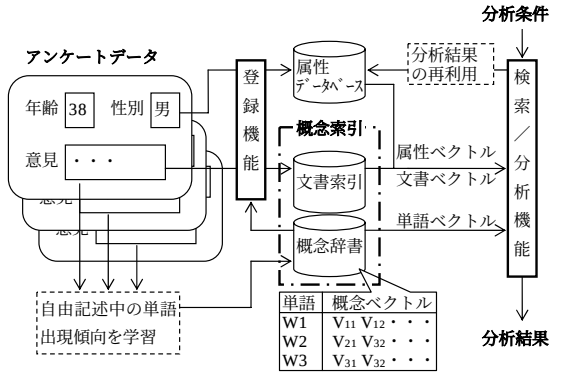


図 2 概念抽出型テキストマイニング

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. 概念索引は概念辞書および文書索引からなり, それぞれ単語と概念ベクトル, および文書と概念ベクトルとを対応づけるための索引である. 以下では, 単語に対する概念ベクトルを単語ベクトルと呼び, 文書に対する概念ベクトルを文書ベクトルと呼ぶ.

概念ベクトルとは, 単語の出現傾向を示す統計データ（行列）を特異値分解により次元圧縮することにより得られるベクトルである. この概念ベクトルの近さ(余弦値)により, 各単語, 各文書, および各文書集合の概念の類似性を計算する.

分析には, 単語ベクトル, 文書ベクトル, および属性ベクトルを用いる. 属性ベクトルとは, その属性値をもつ文書の文書ベクトルを合成することにより得られる概念ベクトルである.

分析条件では, 上記のどのベクトルを組み合わせて分析するかを指定し, 概念ベクトル間の近さを関連度として視覚化することで分析結果を得る.

概念辞書は, 分析対象データの単語の出現傾向を学習することにより自動生成するので, 類義語辞書開発のような手作業は不要であり, 開発コストが小さい.

4.2. 対話的テキストマイニング

前節で説明した概念索引を用いることにより, 3 章で述べた各要求に対応して, 以下の対話的テキストマイニング機能を実現できる.

(1) 分析軸抽出のための概念検索
概念ベクトルの類似性により, 入力した検索文と類似する内容の文書を検索できる概念検索機能を提供する.
本手法で得られる概念ベクトルは単語間の潜在的な関係もとらえることができるので, その文書のテキスト中に検索要求中の語自体が含まれていなかったとしても, 概念的に検索要求と関連していると判定した文書を検索することができる.

(2) 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
分析条件を記録して再利用するためのマクロ機能を提供する. また, 検索結果を保存して分析の母集団としたり, 再度呼び出して別の絞込み条件を追加する機能も提供する.

(3) 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
検索結果や分析結果から得られる情報を, 新たなユーザ定義属性として属性データベースに登録する機能と, これらのユーザ定義属性から得られる属性ベクトルを用いた分析実行の機能を提供する.
分析機能において使用するベクトルは, すべて基底が同一の概念ベクトルである. そのため, 各種の分析結果を自在に組み合わせて, 独自の観点での分析を行うことが可能である.

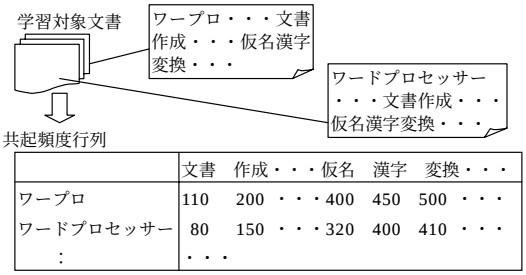
以下では, 概念索引の生成方法, および対話的テキストマイニングの詳細について述べる.

5. 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.

5.1. 概念辞書の学習

本方式では概念辞書として, 単語の共起頻度行列を特異値分解により圧縮した概念空間を採用している[4][5][6]. 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（図３）.



	文書	作成・・・仮名	漢字	変換・・・
ワープロ	110	200 ・・・400	450	500 ・・・
ワードプロセッサー	80	150 ・・・320	400	410 ・・・
：	・・・			

図３　共起頻度行列

例えば,「ワープロ」と「ワードプロセッサー」は「文書」「作成」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. 概念辞書の学習処理は以下の 2 段階で行なう.

A)　共起頻度収集

まず, 学習対象文書を形態素解析し, 単語に分割する. このうち、統計的に特徴の強い単語の並びを複合語として検出する. さらに同一段落中に共起する単語および複合語の頻度を集計し, 共起頻度行列を作成する.

B)　特異値分解

上記で得た共起頻度部分行列に対して特異値分解を実行する（図４）. 特異値分解の結果得られる k 次元の３つ組（U_k, Σ_k, V_k）のうち, 左特異ベクトルU_k を概念辞書として使用する. 各単語に対する概念ベクトルは長さが１となるよう正規化する.

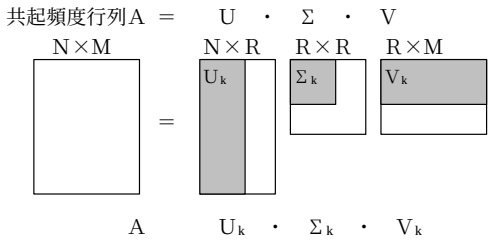


図４　特異値分解の適用

5.2. 文書索引の生成

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して文書索引とする. 合成の際には, 各単語に対する tf・idf 重みを適用する. アンケートに複数の自由記述欄がある場合には, 各記述欄ごとに文書索引を作成する.

自由記述欄以外の定型項目（属性情報）についてはデータベースに格納しておき、後述の属性ベクトル生成処理において参照する。

5.3. 概念索引による分析

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. したがって, 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. 以下に分析機能の例を示す.

(1) 属性－キーワード相関

属性値とキーワードを入力して相関を分析する機能である. 地域ごとの嗜好の違いなどを分析できる.

入力した各属性値に対応する属性ベクトルを生成し, 各キーワードの単語ベクトルとの類似性を計算して棒グラフなどで視覚化する（図５）.

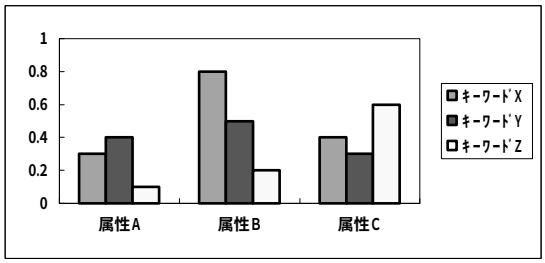


図５　属性－キーワード相関の結果例

(2) 属性－属性相関

２つの属性値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．性別による好みの違いなどを分析する際に使用する．

まず，入力した２つの属性値に対応する属性ベクトルを生成する．つぎに各キーワードに対して，その単語ベクトルと２つの属性ベクトルとの余弦値を，それぞれＸ座標，Ｙ座標として２次元上にマッピングして視覚化する（図６）．

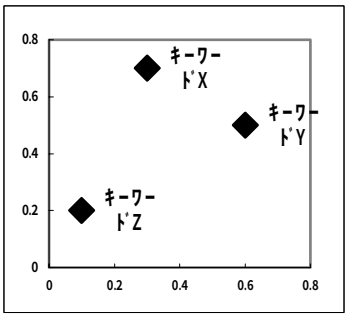


図６　属性－属性相関の結果例

(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．

回答データを日付データなどにしたがって時分割し，各区間ごとに属性ベクトルを生成する．各属性ベクトルと各キーワードベクトルとの類似性を，折れ線グラフなどで視覚化する．

(4) コレスポンデンス分析

テキストおよび属性値を入力とし、各属性値ごとの検索結果件数をクロス集計表としてコレスポンデンス分析を実行する．

分析結果として，入力したテキストおよび属性値を２次元上にマッピングしたデータが得られる．

上記の各機能において，キーワードのかわりに文，文章などのテキストを入力することもできる．この場合は単語ベクトルのかわりに，入力したテキストを形態素解析して単語に分割し，各単語の概念ベクトルを合成して上記と同様の処理を行なう．

6. 対話的マイニング

概念索引を用いることにより，以下の対話的テキストマイニング機能を実現できる．

6.1. 概念検索

アンケート分析においても，試行錯誤的な要素が多いため検索機能は必須である．本方式では，分析軸の選択の支援のために，概念検索機能を用いる．

ベクトル空間モデルを用いた情報検索自体を概念検索と呼ぶことがあるが，多くは文書中から抽出した単語そのものを基底とするベクトルを用いたものである．ここでは，5.1 節で説明した概念辞書を用いて検索要求に対する概念ベクトルを求め，文書索引中の近接する文書ベクトルを求める手法を概念検索と称する．

文書の検索時には，検索要求を形態素解析して，各単語の概念ベクトルを合成して検索ベクトルとする．この検索ベクトルと文書索引中の各文書ベクトルとの余弦値の順にソートすることにより，検索結果を求める．

この手法の利点は，その文書のテキスト中に検索要求中の語自体が含まれていなかったとしても，概念的に検索要求と関連していると判定した文書を検索することができる点にある．

6.2. 分析結果の保存・再利用

アンケート分析では，ある程度の仮説をもって分析する場合が多い．しかし，自由記述の分析においては，予期せぬ回答を見出すことも多く，試行錯誤の連続となる場合がある．

分析作業の効率をあげるためには，ある分析結果を別の分析作業において利用するための機能が必須である．

ここでは，分析条件を記録して再利用するためのマクロ機能を提供する．また，検索結果を保存して分析の母集団としたり，再度呼び出して別の絞込み条件を追加する機能も提供する．

過去の分析履歴を参照し, 分析の母集団を指定することもできるので, ドリルダウン分析を自然な形で実現できる.

6.3. 分析結果の組合せ

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, これらの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.

何種類かの検索結果に対して属性値を与え, これらをひとつの属性として定義するのが, もっとも単純な例である. こうして定義した属性値に対しても属性ベクトルを生成できる.

分析機能において使用するベクトルは, すべて基底が同一の概念ベクトルである. そのため, 各種の分析結果を自在に組み合わせて, 別の観点での分析を行うことが可能である.

7. アンケート分析業務への適用

概念抽出型テキストマイニングによるアンケート分析システムの構成例を図７に示す.

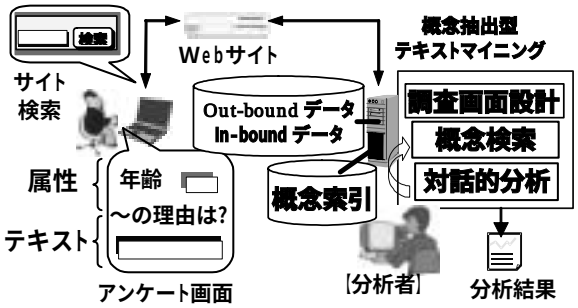


図７ アンケート分析システムへの適用例

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概念抽出型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, 類義語辞書の開発が不要なので, 多種多様な分野のアンケート分析に対して迅速なサービスが提供可能となる.

8. おわりに

本稿では, 同義性や類義性など単語や複合語間の潜在的な関係を自動的に抽出した概念索引を用いることを特徴とする概念抽出型テキストマイニングによるアンケート分析手法を提案した. 現在, 本方式によるテキストマイニングソフトウェアの試作を進めている. 今後は, 同ソフトウェアを実際のアンケート分析業務に適用し, 本方式の有効性を検証するための評価を行なう予定である.

[参考文献]

[1] 市村, 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] 諸橋, 他, “テキストマイニング：膨大な文書データからの知識獲得－意図の認識－“, 情報処理学会第 57 回全国大会, 5K-3, Vol3,pp.75-76, 1998.

[3] 市村, 他, “日報分析システムの開発, 信学技報“, NLC2000-26, pp.31-35, Oct. 2000.

[4] Y. Takayama, R. S. Flournoy and S. Kaufmann, “Information Mapping – Concept-based Information Retrieval based on Word Associations”, Report No. CSLI-98-203, CSLI Publications, Stanford, August 1998.

[5] H. Schütze, and J.O. Pederse, “A coocurrence-based thesaurus and two applications to information retrieval. Information Processing & Mangement”, Vol.33, No.3, pp.307-318, 1997.

[6] 相川, 他 “大規模検索システムにおける概念辞書自動更新”, 情報科学技術フォーラム,FIT2002, D-37, 2002.