# 概念索引を用いたテキストマイニング
## ー(2)対話的マイニングー

伊藤 山彦　　相川 勇之　　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

企業における情報共有の必要性が強調され、大量の電子化文書が蓄積される中、蓄積した文書を有効に活用したいという要求が高まっている。近年、CRM (顧客関連管理)においては、Web アンケート等によって顧客の声を大量に収集することが可能になった。収集した顧客の声を製品やサービスに反映させることは、企業にとって重要な課題となっている。こうした要求に対し、大量の電子化文書を分析して、製品の開発やサービスの改善に活用可能な情報を抽出するテキストマイニング S/W が注目されている。

これまでに、種々のテキストマイニング S/W が提案されているが[1]、顧客の類似した意見を抽出するための類義語辞書の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、アンケート中に潜在する顧客のニーズを十分に引き出せていなかった。我々は、これらの課題に対し、(1)概念索引を用いることにより類義語辞書を用いずに文書の類似性を判定できる、(2)対話的にマイニングにより分析結果を蓄積し再利用できる、という特徴をもつテキストマイニング方式を提案する。本稿では対話的マイニングを中心に記す。

## 2. アンケート自由記述分析の課題

アンケートの回答には、年齢や性別、設問に対する回答項目のように選択肢で指定されるもの(本稿では属性と呼ぶ)と、自由記述で記載されるものがある。自由記述には、アンケートの作成者が想定していなかった顧客の意見が記載されるため、非常に重要な情報である。従来、テキストの分析には、テキストからキーワードや句、または文末表現を抽出し、他の属性との相関関係を求める等の処理が行われていたが、分析結果を知識として蓄積し、以降の分析に活用する機能は十分でなかった。

本稿で提案する対話的マイニングは、分析の過程で得られた知識を基に、ユーザがテキスト中の情報を類別して新たな属性を定義する作業を支援する。新たに定義された属性は、予めアンケートの設問として備わっていた属性と同様に以降の分析に利用できる。本手法を用いることにより、アンケートの作成者が予め調査項目としていなかった属性をテキストから抽出して蓄積することが可能になる。

## 3. 概念索引を用いた対話的マイニング

### 3.1 概念索引の利用

図 1 に、本テキストマイニング方式の全体構成を示す。本方式では、文書から抽出した単語の共起頻度表に対して線形代数計算により次元を圧縮した概念ベクトルを格納した概念索引を用いる[2]。文書中で出現の仕方が類似した単語同士は類似した概念ベクトルを持つ。単語に対する概念ベクトルを合成することにより、文書や、共通の属性を持つ文書集合に対しても概念索引を作成することができる。本方式では、概念索引を利用することにより、類義語辞書を用いずに、キーワード間、キーワードー属性間、属性間の種々の関連度を測定できる。

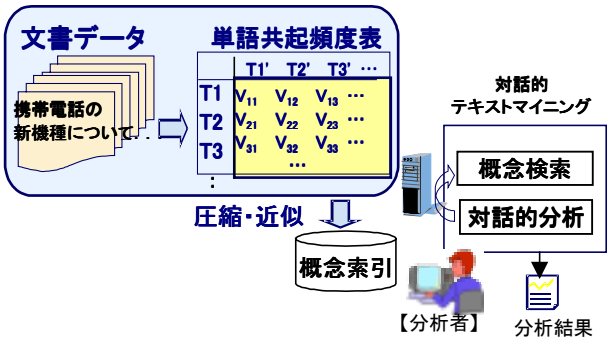



図 1 概念索引を用いた対話的マイニング

Text Mining Using Concept Index. -(2) Interactive Mining -
Takahiro ITO, Takeyuki AIKAWA, Yasuhiro TAKAYAMA, Katsushi SUZUKI
Information Technology R&D Center, Mitsubishi Electric Corporation
5-1-1 Ofuna, Kamakura, Kanagawa, 247-8501, JAPAN.

### 3.2 対話的マイニング

対話的マイニングとは、利用者が検索や分析を行った結果を保存し、その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機能である。以下、対話的マイニングの処理の概要を記す。

#### (1)分析結果の文書集合

利用者が検索やマイニングによって得た文書集合は、名前を付与して保存できる。図 2 は、キーワードを組み合わせて分析を行った結果の文書集合に対して名前を付与し、保存した例である。例えば「上司、取引先、報告・・」等のキーワードを含む記述に対し「ビジネス」という文書集合名を付与したことを示す。

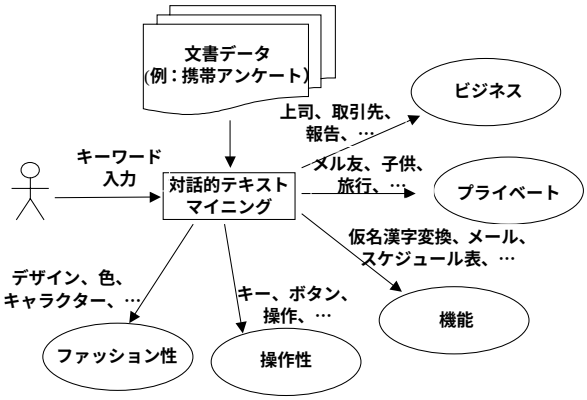
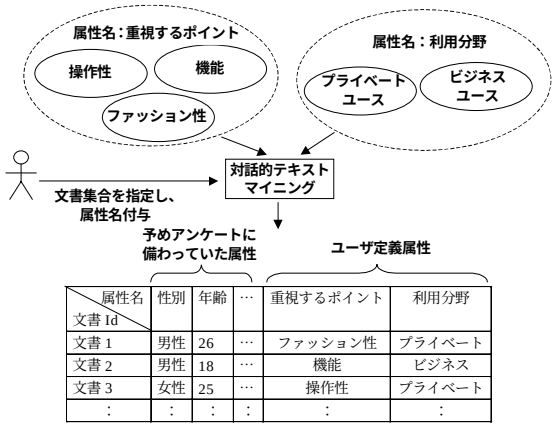



**図 2 分析結果の文書集合**

上記によって保存された文書集合には、各文書と分析に使用したキーワードとの関連度の値がスコアとして付与される。スコアは各文書とキーワードの概念ベクトルの類似度によって計算される。

#### (2)ユーザ定義属性

文書集合に共通の属性名を付与して、新たな属性を定義することができる。例えば、図 2 の文書集合のうち、「ファッション性」「操作性」「機能」の 3 つに対して「重視するポイント」という属性名を定義する。また、「ビジネス」と「プライベート」の 2 つに対して「利用分野」という属性名を定義する。それぞれの文書がどの文書集合に最も近いかを、関連度のスコアから判定する。この処理によって、「重視するポイント」および「利用分野」という属性名に対して、各文書がどの文書集合に属するかを、属性値として付与する(図 3)。属性は、他アプリとの連携を考慮し、汎用性の高い XML 形式で記述される。




| 属性名 / 文書 Id | 性別 | 年齢 | … | 重視するポイント | 利用分野 |
|---|---|---|---|---|---|
| 文書 1 | 男性 | 26 | … | ファッション性 | プライベート |
| 文書 2 | 男性 | 18 | … | 機能 | ビジネス |
| 文書 3 | 女性 | 25 | … | 操作性 | プライベート |
| ： | ： | ： | ： | ： | ： |

**図 3 ユーザ定義属性**

#### (3)ユーザ定義属性を用いたマイニング

上記(2)で定義したユーザ定義属性は、以降のマイニングで利用することができる。図 4 は「重視するポイント」と「利用分野」の間の相関を分析した例である。図の「関連度」は、それぞれの属性値を持つ文書集合間の近さを表す。図 4 の例では、分析の結果「ファッション性」と「プライベート」の相関が高く、「機能」と「ビジネス」の相関が高いという結果が得られたことを示している。

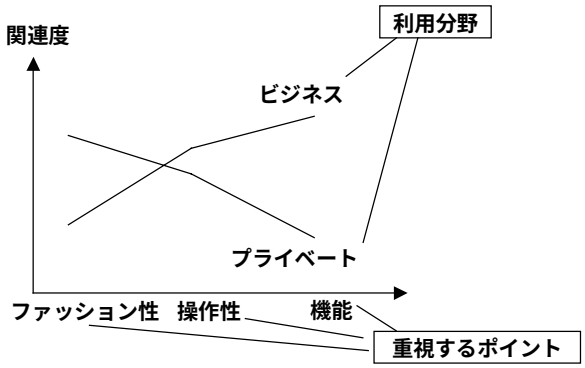



**図 4 ユーザ定義属性を用いたマイニング**

## 4. まとめ

本稿では、概念索引を用いた対話的テキストマイニング方式について提案した。テキストの分析結果を関連付けて新たな属性を定義し、以降の分析に利用可能とすることが可能になる。今後、本方式を実業務に適用することにより、有効性を検証する予定である。

参考文献
(1)市村ほか:テキストマイニングー事例紹介, 人工知能学会誌, Vol.16,No2(2002).
(2)相川ほか: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).