テキストマイニング技術とその応用

Text Mining Technology and its Applications

林 田 英 雄，脇 森 浩 志

要 約 コンピュータの処理能力向上により，テキストデータを用いた分析に対するニーズが高まり，テキストデータのマイニングが注目されている．数値化，コード化されたデータに対してのマイニングは従来から行ってきたが，1997 年以来テキストデータもマイニングの対象に加えるために，プロダクト開発とその適用を通じて，テキストマイニング技術を培ってきた．

本論文では，テキストマイニングの適用分野，技術，その製品開発，適用事例について述べることで，テキストマイニング全般について解説する．まず，2 章でテキストマイニングの適用分野について述べ，技術の利用範囲を規定する．3 章では，想定された利用分野で有効なテキストデータからの情報抽出方法と分析方法について述べる．4 章では，その技術を利用した二つの製品「MiningPro 21 文書マイニング・システム」と「TopicExplorer」について触れ，その適用事例を示すことで技術と製品の有効性を示す．更に，5 章ではテキストマイニングの適用分野であるコールセンターにおける「顧客の声」分析システムをモデル化して開発した「CVPro 活用系」に触れ，製品開発・適用を通じて培った知見について述べることとする．

Abstract The enhanced throughput of the computer system increases the need for the analysis using text data, and consequently the mining of text data attracts the attention. Since 1997, we have developed analytical techniques for not only numeric or coded data, but also text data, in development and its application of text-mining system product.

This paper explains the text mining as a whole by describing the field of application, technology, product development, and the application practice of the text mining. First of all, chapter 2 discusses the field of application of text mining, and defines the range of use of the technology. Chapter 3 describes the method of information extraction and analysis from text data useful to the assumed application field. Chapter 4 introduces two products using the technology“ MiningPro 21 Document Mining System ”and“ TopicExplorer ”, and presents the application experiences to prove the effectiveness of the technology and the product. In addition, chapter 5 discusses“ CVPro Utilization System ”that has been implemented though modeling“ Customer ’ s Voice analysis system ” in the call center that is one of the application field of text mining, as well as the knowledge obtained by the product development and its application.

1. は じ め に

コンピュータの処理能力の向上と安価なメモリにより，従来では不可能であった大量データを対象にした分析業務が行えるようになってきた．データマイニング，特にテキストマイニングはその恩恵を最も受けているものの一つである．従来は蓄積されたデータの分析といえば数値化，コード化されているデータを対象に，各種の集計やデータマイニングなどを駆使してそ

の傾向を発見していくことであった.ところが近年のコンピュータの処理能力の向上を受けて,蓄積されたテキストデータを分析する業務が行えるようになってきた．数値化，コード化されたデータは，値の範囲，コードのバリエーション，それらの意味を予め人が規定した構造的なデータである．一方テキストデータは，そこに何が書かれるか予め人が規定することができない非構造的データである．このテキスト中から効率的に意味のある傾向を抽出することで，従来では気づかなかった情報を得ることができる.

本稿ではテキストデータ中の単語や同義語などを用いてデータを自動仕分けやグルーピングをする MiningPro 21 文書マイニング・システム（文書マイニング・システム），テキストデータ中の係り受けを抽出，集計する TopicExplorer の開発を通じて培ったテキストマイニングの技術と適用事例，そしてその技術をコールセンターへの適用を通じて作成した「顧客の声」活用モデルとそのアプリケーション CVPro について述べたものである.

2. テキストマイニングの概要と適用分野

本稿で扱うテキストマイニングは，最近注目を浴びるようになってきた分野である．この章では，テキストマイニング出現の背景と，その適用分野について述べる.

2.1 テキストマイニング製品出現の背景

従来行われてきた分析は，売上高や年齢層などの数値化，コード化されたデータを対象に分析し，その法則性を発見しようとしていた．しかし，企業内には数値化，コード化されたデータよりも多くのテキストデータが存在する．たとえば，商品の購買データがあった場合，感想欄や備考欄に書かれているクレームなどの情報はテキストデータとして保存されている．このテキストデータは，そこに何が書かれるかを予め規定されていないため，従来見過ごしていた規則性を発見できる可能性があり，近年その活用ニーズが高まっている．一方で，コンピュータの処理能力の向上によりテキストデータから得られる情報の処理が可能になってきた．テキストマイニングはこのような市場のニーズとコンピュータ処理能力の向上の双方から注目されるようになった分野である.

2.2 テキストマイニングの適用分野

テキストマイニングは，CRM （Customer Relationship Management）分野における分析業務で利用されることが多い．レジス・マッケンナによれば，CRM は「顧客および業界内の関係者集団との関係を築き，維持すること．企業が行うデザイン，開発，製造，販売のプロセスの中に顧客を取り込むこと」[7]といわれている．CRM のシステムでは顧客のニーズやウォンツを探り，商品開発などに結びつけるため，顧客との接点で様々なデータが採取される．企業にとって消費者への能動的な接点で取られるデータには消費者アンケートのデータがある．また,消費者からの受動的な接点で取られるデータには顧客コンタクトにおける対話歴がある．採取されるデータには数値化，コード化されたデータだけでなく，テキストデータも存在する．非構造的なデータであるテキストデータを分析することで，今まで気づかなかった傾向を発見しようという試みが様々な企業で行われている．このためテキストマイニングの適用分野で最も多いのは，このアンケートの自由記述欄とコールセンターの対話歴の分析である.

2.3 テキストマイニング製品の種類

それぞれの適用分野に合わせてテキストマイニング製品も二つに大別される．一つはアンケートデータを手早く分析するための製品と，もう一つはコールセンターの対話歴を分析するための製品である．どちらもテキストデータ中の文章を単文に分解し，そこから抽出される単語，同義語，後述の係り受けなどを拠所にして分析をする．

アンケートデータの分析では，個々のアンケート対象が異なるため，そのテキストデータの内容も異なり，継続性が無い．このため，アンケートデータを分析する製品では，手間を掛けずにその内容を把握し，分析できる必要がある．多くの製品では，単語，同義語，係り受けを抽出し，どれが同じテキストデータに出現するのかを分析する共起関係分析や，そのテキストデータに付随するその他の項目と併せたときの規則性の分析を行う．

コールセンターデータの分析では，製品やサービスは継続的に提供されており，そのテキストデータの内容も，毎月同じ対象について述べられ継続性がある．このため，コールセンターデータを分析する製品では，抽出した単語，同義語，係り受けを拠所にして，テキストデータを文意の似通ったグループに自動的に分類した後に，そのグループとテキストデータに付随するその他の項目との関係を分析する．

3. テキストマイニングの要素技術

2 章ではテキストマイニングの適用分野を中心に述べたが，本章ではそれらの分野でテキストマイニングを行うためテキストデータから情報を抽出する技術とその分析技術について述べる．

3.1 テキストからの情報抽出

テキストデータはコンピュータから見ればただの文字コードの集まりに過ぎない．この文字コードの集まりから意味のある情報を抽出する手段として，単語の抽出を行う形態素解析，表現のゆれを吸収して形態素の集約を行う同義語，テキストデータ中の単語の並びを抽出するワードパターン，抽出した形態素を文節や係り受けにまとめる係り受け解析の技術がある．

3.1.1 単語を抽出するための形態素解析

形態素とはこれ以上細かくすると意味がなくなってしまう最小の文字列をいう．また文を最小の文字列に分解し，その品詞を特定することを形態素解析という[4]．日本文は，英文のように空白で形態素が区切られていないため，形態素に分解する際に，文の切り方などで複数の形態素の抽出結果が得られる．日本語の形態素解析は，これらをルールベースもしくは，統計モデルを用いて，効率よく，納得性が高い結果を得る技術である．表 1 に「日本の車は品質が高い」という文を形態素解析した例を示す．

表 1　形態素解析の例

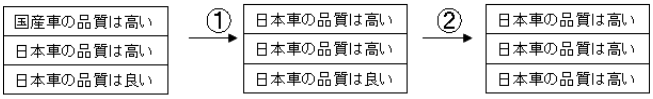
見出し	読み	品詞
日本車	ニホンシャ	名詞
の	ノ	助詞
品質	ヒンシツ	名詞
は	ハ	助詞
高い	タカイ	形容詞

形態素解析は，形態素の見出し，読み，品詞，活用形などを記録した形態素の辞書を用いながらテキストデータから形態素を抽出する．「日本車は品質が良い」という文を形態素解析した結果，「日本車」が「日本」と「車」と解析された場合，「日本車」という単語を形態素の辞書に加える．この「日本車」のように複数の単語が合わさってできたものを「複合語」という．形態素辞書にはこのように必要に応じて複合語やデータに固有の単語などを登録する．

3.1.2 表現のゆれを吸収するための同義語

統制された言葉で記述されていない文は，同じ内容が様々な言葉で表現される．例えば同じ日本の車を指す言葉には，「国産車」という表現もあれば，「日本車」という表現もある．このような表現のゆれを吸収するため，同義語辞書を作成し，表現のゆれを吸収した一つの代表語に集約する．

図１は「日本車」と「国産車」，「良い」と「高い」を同義語登録した例である．



①「日本車」と「国産車」を「日本車」として同義語登録
②「高い」と「良い」を「高い」として同義語登録

図１　同義語の例

この例では，どの文も本質的に言いたいことは同じであり，同義語登録することで表現のゆれを解消できる．ただし，何を同義語とすべきであるかは，取り扱うデータによって決まることになるので，テキストデータを見ながら決定する必要が有る．図１の例では，「良い」と「高い」は，全ての文で品質に関する表現であるため同義語として良いが，他のデータに「金額が高い」などがあれば，同義語とするべきではない．

このようにして同義語を指定することで，表現のゆれを吸収して分析の精度を上げることができる．

3.1.3 近傍に出現する形態素を抽出するためのワードパターン

ワードパターンは，「あるか？」や「思えない」などの疑問形や否定形を抽出する場合や「あるか」だけでなく「ありますか？」など表現がゆれた場合の同義語を抽出する場合に利用する．

ワードパターンは，正規表現[8]のマッチングにより複数の単語の並びを抽出する仕組みである．正規表現とは文字列のパターンを表現する表記法である．正規表現は通常の文字と，メタキャラクタと呼ばれる特別な意味を持った記号で表記される．メタキャラクタには行頭を表す「＾」，任意の１文字を表す「.」，直前の要素の１回以上の繰り返しを表す「＋」などがある．正規表現を使えば，文字列を直接指定せず，特徴（パターン）を指定することができるため，表記の揺れを吸収した抽出が行える．

しかし，日本語の場合，動詞や形容詞には活用がある．例えば「ありますか？」と「あるか？」では「ある」という単語が「あり」と「ある」という異なる活用をしている．このためこの両者を抽出する場合，文字列のままで正規表現を用いたマッチングを行うと「あ」という文字の

後ろにある「か？」という文字の出現を探すことになり，意図しない「あえますか？」という表現も抽出してしまう．そこでワードパターンにより近傍に出現する形態素を抽出するときは，文字列を形態素に分解して，分解した形態素に対して活用をそろえた後に正規表現でマッチングを行う．先の例では，まず抽出対象のテキストである「あるか？」と「ありますか？」を形態素解析して，形態素ごとに“＜”と“＞”で囲い，「＜ある，動詞＞＜か，助詞＞＜？，記号＞」，「＜ある，動詞＞＜ます，助動詞＞＜か，助詞＞＜？，記号＞」を記録する．このテキストに対して，「ある」の後に１単語以内に連続して「か」「？」が出現する正規表現「＜ある，動詞＞（？：＜［＾＞］＋＞）{0,1}＜か，助詞＞＜？，記号＞」を指定して，マッチングをかける．このようにすることで，漏れが少なく不要なものも少ない情報抽出が実現できる．

3.1.4　係り受け解析による係り受けの抽出

係り受け解析は，テキストから形態素を抽出し，抽出した形態素を文の意味がわかる程度にできるだけ短い単位の文節にまとめ，更にまとめた文節間の主語と述語の関係，修飾語と被修飾語の関係などの係り受けを判定する技術である．

図２は「明日はきっと良い天気になるだろう．」という文から，単語を抽出し，文節にまとめて，文節間の係り受けを解析した例である．

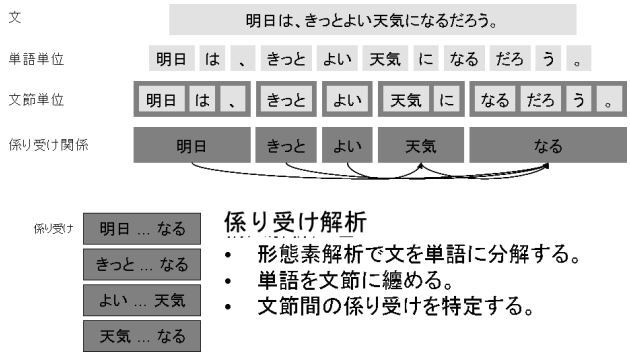


図２　係り受け解析の例

係り受け解析では，まず形態素を抽出する．次にその品詞などの属性と形態素の出現位置を基にして，各形態素の直後で文節が切れるかどうかを判定する．更に，各文節の属性を基にして，文節間の係り受け関係を判定する．

文節の判定，文節間の係り受け判定にはSVM （Support Vector Machine）を用いて行う方法[5]がある．SVMには過去に手作業で作成した文節区切りや係り受けの正答値が学習してあり，その基準を入力とし，結果を推定することができる．

3.2　分析

これらの方法でテキストデータから単語，同義語，ワードパターン，係り受けなどを抽出したが，次に抽出した情報を用いて様々な分析を行う．分析には，抽出した情報である単語や同義語，ワードパターン間の関連を分析する方法，抽出した情報を基にテキストデータのグループを作りその傾向を分析する方法，係り受けに注目してその係り元や係り先を集計し，テキストデータの傾向を分析する方法などがある．また更に踏み込んでテキストマイニングの結果と

データに付随するコードや数値情報を併せて行う 2 次分析もある．

3.2.1　単語間の関連分析

単語 A と単語 B が同時に出現することを単語 A と単語 B が共起しているという．単語間の関連分析では，この単語間の共起関係に注目し，お互いに結びつきの強い単語を見ることで，テキストデータ全体の傾向を分析する．例えば，特定の製品名と「破損」などの単語の関連が強ければ，その製品の破損について記述されているデータが多いことがわかる．単語の関連度にはいろいろな規定の仕方がある．単純に同じ文書に単語 X と単語 Y が同時に出現する確率を用いているものもあれば，抽出された単語を軸とするベクトル空間に対して潜在的意味インデクシングを行うことで，文意を元にした関連度を規定しているものもある．テキスト集合に対する検索などでは主要語を軸としたベクトル空間法が有名であるが，軸の数が多いため高次元となる．このベクトル空間を特異値分解などによりデータの次元を縮小したものが潜在的意味インデクシングである．

図 3 には潜在的意味インデクシングでできた空間上に単語を配置した「単語マップ」の例を示す．

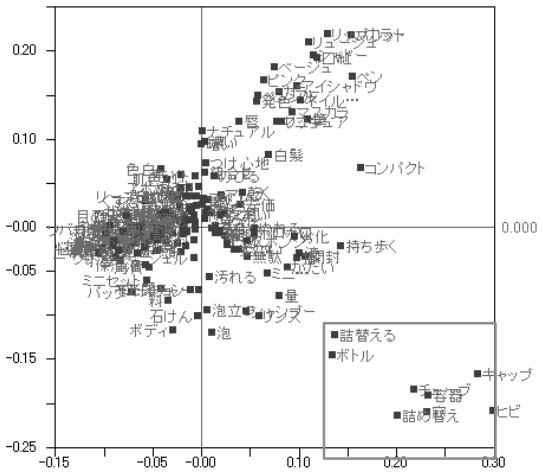


図 3　単語マップの例

このマップの右下には「詰め替える」「ボトル」「詰め替え」「容器」「キャップ」「ヒビ」などの単語がある．これらの単語から「詰め替え用容器」と「容器の破損」について述べられたテキストデータのグループがあることが類推できる．また，この例でも分かるように「詰替える」,「詰め替え」というほとんど同じ意味の単語が近くに配置されている．この二つの単語が同時に使われることはほとんど無いが,「容器」などの同時に出現する単語の傾向が似通っているため近くに配置されている．これが潜在的意味インデクシングの特徴である．

3.2.2　テキストデータの文意による分類

テキストデータの文意による分類では，テキストデータを意味の似通ったグループにまとめて，出来上がったグループの件数などを見ることで，テキストデータ全体の傾向を把握する．

テキストデータの分類には，次の 2 種類の方法がある．一つ目はテキストデータの分類基準

が既にあり，一部データについて既に分類済みであるときに，未分類データを同様に仕分けする「クラシファイ」，二つ目はテキストデータの分類基準が無いときに意味の似通ったデータのグループに分類する「クラスタリング」である．

1）　クラシファイによるテキストデータの自動仕分け

テキストデータを自動的に仕分けして分類コードを付与できれば，継続的に発生するデータをいつでも集計や分析をすることができるようになる．分類コードの付与は，予め仕分けした結果を基に自動仕分けルールを作成し，データ発生時にそのルールを適用することで実現できる．

最も単純な自動仕分けの仕組みには，メールソフトの自動仕分けがある．これは特定の文字列が出現した場合に，どのフォルダーに仕分けるかというルールを設定しておき，データ発生時にこれを検査して自動で仕分ける仕組みである．例えば，件名に文字列「見積」が含まれていて文字列「不要」が含まれていない場合に，見積フォルダーにメールを移動する場合には，次のような指定を行う．

```
if　(文字列'見積'が件名に出現)　then
    if　　not(文字列'不要'が件名に出現)　then
          '見積'フォルダーに移動
    endif
endif
```

しかしこのような単純なルールでは誤って仕分けてしまうことも多く，それに対応するためには「不要」のような複雑な例外ルールを定義しなくてはならず，限界がある．

テキストマイニングを用いた自動仕分けでは，テキストデータから抽出した単語・同義語・ワードパターンなどの情報を基に特定のグループに属するかどうかを判別する．判別するモデルも，判別関数[10]を用いるものもあれば，SVM を用いるものなどもある．以下の例では，判別関数 P（見積）を定義し，その判別関数値により，仕分け先を判断している．

```
if　(P(見積) ≥ 0.8)　then
   '見積'フォルダーに移動
endif
P(見積) = 0.6×単語'見積'出現有無(0,1)
          −0.4×単語'不要'出現有無(0,1)
          +0.2
```

この様に複数の単語の出現を見て判別するときには，各単語に重み付けを行い，判別関数を作成する．この重み付けは事前判定結果の付いたテキストデータさえあれば，統計的に算出することができ，ユーザは使用する単語さえ決定すれば，判別関数を作成することができる．

2）　クラスタリングによる文書グループの作成

予めテキストデータをどのようなグループに仕分けるかが決まっていない場合，クラスタリングを行い，テキストデータのグループを作成することができる．文意が似通ったものでグループを作成するには，前出の潜在的意味インデクシングを利用してできた空間上

にテキストデータを配置してクラスタリングを行う．図4は図3の単語マップ上にテキストデータをプロットして，そのテキストデータに対してクラスター分析[10]を用いてマップ上の近いものをまとめてグループを作成した例である．

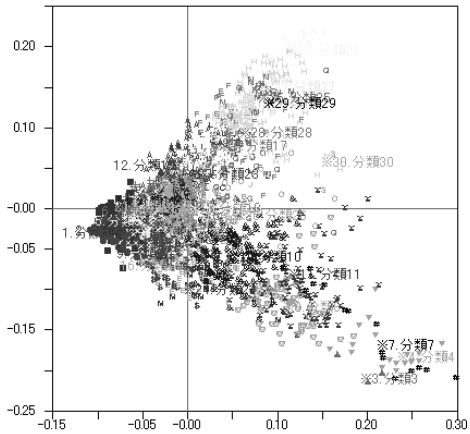


図4　テキストマップと分類結果の例

このように潜在的意味インデクシング後の空間でクラスタリングを行うことで，文意の似通ったテキストデータをグループにまとめることができる．

3.2.3　テキストデータの係り受けの分析

係り受け解析をすると「肌」→「合う」などの文節間の関係を係り受けとして抽出する．係り受けには係り元と係り先があり，「肌」→「合う」では「肌」が係り元であり，「合う」が係り先になる．係り受けの分析では，特定の係り元に注目して係り先ごとの件数を見ることで，その対象がどのように扱われているかがわかり，係り先に注目すると何に対して感じているのかがわかる．図5は，係り元，係り先それぞれの視点で見る係り受け分析の例である．

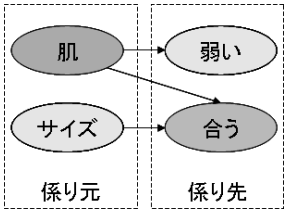


図5　係り受けの分析の例

この例では，係り元である「肌」に着目し，係り先を見ると，「合う」「弱い」などがあり，「肌」についてどのように感じているのかがわかる．また係り先の「合う」に着目し，その係り元をみると，「肌」「サイズ」などがあり，何に対して「合う」と感じているのかがわかる．

3.2.4　テキストマイニング結果を使った2次分析

テキストマイニングの結果と，そのデータに付随する他の項目を併せて集計，分析すること

でより有益な分析が可能になる．例えば，現場レベルのコールセンターでは，内容別に分類された問合せとその日時を基に，時系列に内容別問合せ件数の推移を監視している．その結果，ある製品に関する問い合わせが増加していれば，それを FAQ に掲載してコール数を減らすことで運営費を低減させ，更に社内の関連部門にその情報を提供している．また，マーケティング部では，問合せの内容に顧客の R（Recency：最新購入日），F（Frequency：累積購入回数），M（Monetary：累積購入金額）や年齢層，購買開始からの期間など顧客情報を加えることで，離脱原因分析などの分析を行っている．本稿の 4.3 節に，その事例を記述したので，具体例はそちらで述べる．

4.　テキストマイニング製品の開発

テキストマイニングの適用分野に合わせて，テキストマイニングの技術を使ったシステム「MiningPro 21 文書マイニング・システム」と「TopicExplorer」を開発した．

4.1　MiningPro 21 文書マイニング・システム

MiningPro 21 文書マイニング・システム（以下文書マイニング・システム）には，クラシファイを行う判別機能と，クラスタリングを行う分類機能がある．

4.1.1　判　別　機　能

判別機能では単語，同義語，ワードパターンの有無を用いた判別関数を用いてクラシファイを行う．判別関数の結果は 0 点から 100 点に調整され，その結果により特定のグループに仕分けるべきかどうかを判断する．

更に文書マイニング・システムでは，判定結果の信頼性を評価し，誤り無く判断できる部分だけ判断するようにしている．テキストデータから求めた判別関数の結果を 0 点〜10 点，10 点〜20 点，…，90 点〜100 点という 10 点毎に区分化して，各区分でデータの事前判定結果がどちらかに偏っていれば，その偏った値に自動的判断し，偏っていない場合には，判定不能とする．こうすることで無理やり全てを自動的に判断せず，信頼性の高い区分だけを自動化し，精度と効率を両立させている．図 6 は，苦情の判別結果を区分化し，区分毎にユーザの事前判定をグラフ化したものである．

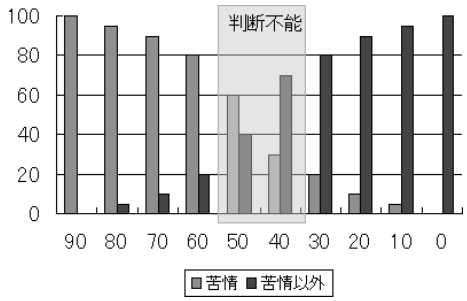


図 6　判別結果の区分別帰属率

この例では，判別確率が 60 以上は苦情の割合が 80% 以上であるため苦情とし，40 未満は苦情でないものの割合が 80% 以上であるため苦情ではないとする．そして残りの 40〜60 では，

苦情とそうでないものの割合が偏らないため判断不能として扱う．

あるコールセンターに対して，問合せデータに判別機能による自動仕分けを適用したところ，その正答率は 90% となり，その結果として約 60% のコストカットができた．このコールセンターも，従来はメールソフトの自動仕分けのような仕組みと最終的な人による補正を行っていたが，判断の個人差などもあり 70% までしか精度が上がっていなかったが，判別機能により精度とコストを両立させている．

4.1.2 分 類 機 能

分類機能では単語，同義語，ワードパターンの有無を用いてテキストデータのクラスタリングを行う．文書マイニング・システムでは，単純なベクトル空間を用いずに，潜在的意味インデクシングを行う方法を採用しているため，文意を反映したクラスタリングが行える．

また，分類を応用することで類似検索も行える．事前に検索対象データをテキストマップ上に配置しておき，検索時には検索文をテキストマップ上に配置して，その近傍の検索対象データから出力することで，類似検索を実現する．表 2 は，PC サポートセンターの問合せを類似検索した例である．

表 2　潜在的意味インデクシングを用いた類似検索例

番号	問い合わせ内容	類似度
1	FA-0050 PCを新しいPCに交換したので、Exchangeの個人用アドレス帳をコピーしたいがどのファイルか?	0.0
2	EL-0134 旧PCの個人用フォルダの内容を新PCへ移行したい。	0.9
3	EK-0284 古いPCから新しいPCへ個人用フォルダの内容を引き継ぎたい。	1.0
4	EJ-0016 本体交換の時、受信トレイ(エクスチェンジ)の場合、今まで使用していたアドレス帳のインポートは可能なのか?	1.3
5	FG-0134 PCを交換し、サーバーにバックアップしておいたアドレス帳 個人フォルダをコピーしたところ、受信トレイに自動的に	1.4
6	FH-0188 旧PCから新PCにメールを移行したが、受信トレイに古い個人用フォルダーと新しい個人用フォルダーが2個できてし	1.4
7	EK-0454 Outlookで個人用アドレス帳の内容を旧PCから新PCへ移行しようとしたが、エクスプローラで検索しても、「mailbpxpa	1.4
8	FJ-0279 新規PCで旧PCからコピーしたpst, pabファイルを個人用フォルダと個人用アドレス帳を使用したい。	1.7
9	FD-0205 本社勤務の人専用のPCでプロファイル設定をしている。起動すると個人用アドレス帳がないというMsgが出る。	1.7
10	HK-0020 連絡帳を以前使用していたPabファイルにもどしたい。	1.7
11	FJ-0165 Exchangeで使用していた個人用アドレス帳を、Outlookのインストールの際「連絡先」に入れてしまった。元に戻す方法	1.7
12	FD-0185 4/10に5F南から15F北に移動する際に、固定アドレスを使用していたサーバにDHCPでアドレスを割り当ててしまった	1.7
13	EL-0005 PCの移動をしたら、個人用フォルダが消えてしまった。	1.8
14	EK-0377 古いPCで使用していた個人用フォルダの内容を新しいPCでも参照したい。	1.8
15	FB-0280 以前使用していた、受信トレイアシスタントの内容を現在使用しているPCに移行したい。	1.8
16	HK-0239 アドレス帳に連絡先を表示させたいが、連絡先のフォルダのプロパティで電子メールのアドレス帳にこのフォルダを	1.9
17	FB-0200 情報システム部のホームページの[現在のエイリヤス]には、申請されているようになっているが、「NISEシステム」でも	1.9
18	HK-0127 アドレス帳の連絡先が三つ表示されており、二つは空白となっている。	2.0
19	EH-0128 <EH-0112>の続き。「配信不能」で戻れないメールの差出人をダブルクリックすると、プロパティの[電子メールアドレス	2.1
20	HK-0252 連絡先の内容を個人用フォルダにコピーしたい。	2.1

この例の番号 10「連絡帳を以前使用していた Pab ファイルに戻したい」という文は検索文である番号 1「PC を新しい PC に交換したので Exchange の個人用アドレス帳をコピーしたいがどのファイルか」とは，そこに含まれる単語の重なりは無いものの潜在的意味インデクシングを用いているので関連問合せとして抽出されている．

4.2 TopicExplorer

アンケートのように毎回，データの内容が異なり継続性が無いテキストデータは，手早く全体を概観することが望まれる．そこで文書の内容を文意というレベルまではまとめてはいないもののテキスト中の係り受けを使って傾向を把握するシステムとして TopicExplorer を開発した．TopicExplorer では，係り元係り先の関係分析と係り受けとその他の属性との分析を行

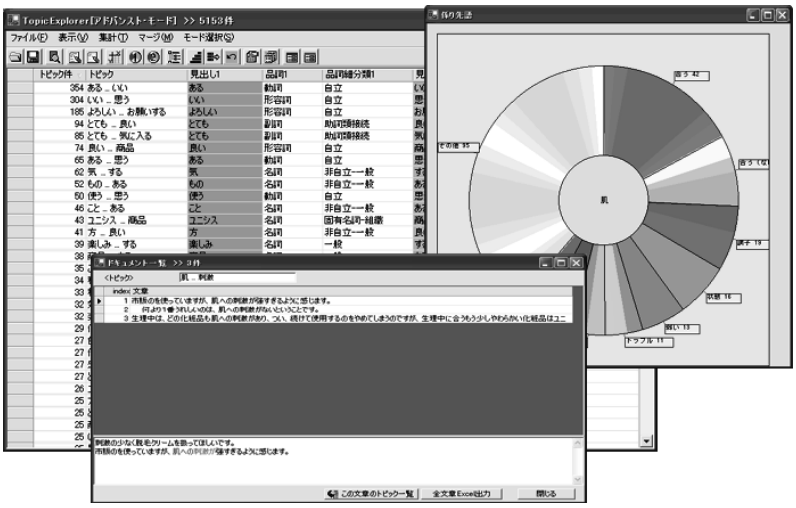


図 7　TopicExplorer の画面例

い，手早く文書全体の傾向を把握することができる．図 7 は，TopicExplorer で係り受けを抽出し，係り元係り先の関係分析を行った画面である．

TopicExplorer ではこのように係り受けを一覧表示して，気になる係り受けを掘り下げ，その係り元や係り先を探索していくことで，全体のテキストの内容を概観する．係り受けを用いた分析で注意しなくてはならないのは，係り受けの件数が多いからといって，そのテキストデータで一番述べたい内容かどうか分からないということである．表現としてはよく出てくるが，そのテキストデータで本当に述べたいことは他にあることがあり，テキストデータそのものを確認しながら評価する必要が有る．このため TopicExplorer では，単純に係り受けを集計するだけでなく，元のテキストデータとその中での係り受けを強調表示する機能を用意している．

4.3　文書マイニング・システムの分類結果と他の属性を使った 2 次分析

文書マイニング・システムはテキストデータの自動分類を行うが，内容毎の件数だけでは限定的な分析しか行えない．例えばコールセンターで自動分類をした件数を見て FAQ などの掲載を行うこともできるが，これはコールセンターの運営費を減らすことにはなっても，顧客を全て同じとみなしており，CRM の観点からすると初歩的な分析を行ったに過ぎない．

米国では，サービスや商品の不備による不満が原因で企業から離れてしまいそうな顧客を CAR （Customer At Risk）として管理している．顧客には苦情などの問合せでコールセンターにコンタクトしてきても，それが原因で企業から離脱する顧客と離脱しない顧客がある．

そこで，A 通販会社のコールセンターに対して，文書マイニング・システムを使って問合せを内容で分類し，そこに年齢層などの顧客属性や購買履歴を加えて離脱分析を行った．また単純な内容別の分析ではなく，そこに顧客の属性や企業との付き合いの長さなどを併せて細かく分析を行うことにより，顧客の離脱を防ぐためにはどのような問合せ内容から改善して行くべきかを分析した．図 8 は，コールセンターの問い合わせ内容別の維持率である．

この結果から，問合せをする顧客のほうが問合せをしない顧客よりも維持率が高い（離脱率が低い）ことがわかる．更に問合せのあった顧客を問合せの種類でみると，一般には不満の度合いが高いと思われる苦情→要望→質問→感謝の順に維持率が高く，また問合せの内容でみて

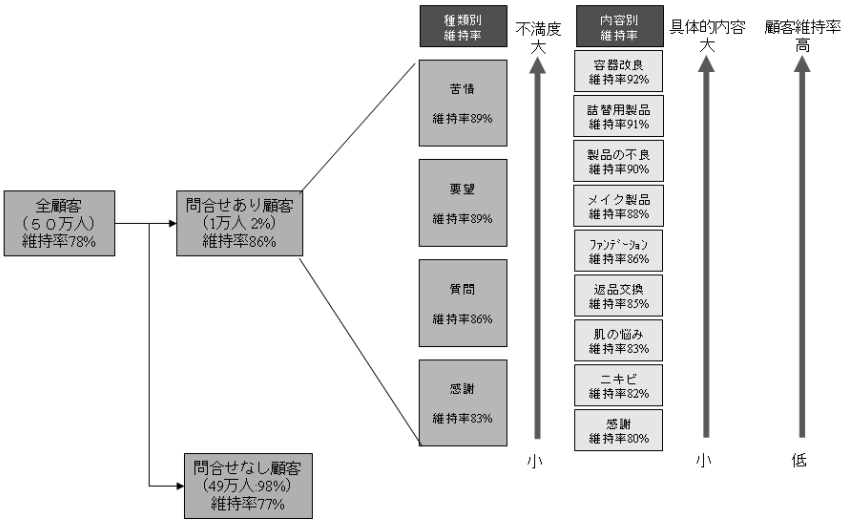


図 8　問合せ内容別維持率

も，より具体的な内容の方がより維持率が高いことがわかる．

次にこれを購買が始まってからの年数という顧客の属性で分析した．図 9 は，図 8 で問い合わせのあった顧客を，企業と顧客との付き合いの長さで離脱率を分析した例である．

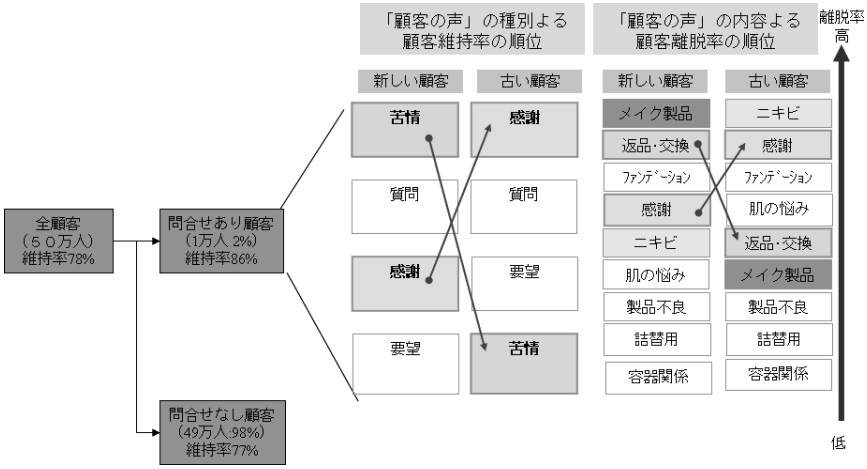


図 9　購買期間別離脱率

この結果からは購買経験の浅い顧客の離脱率は，購買経験を無視した分析結果での離脱率とは異なることがわかる．購買経験の浅い顧客（新しい顧客）は，苦情が原因でその企業との取引を止めてしまう場合が多い．企業にとっては，せっかくプロモーション費用を掛けて得た顧客が離脱すると非常にインパクトが大きい．また，更に内容別に見て行くと，購買経験の浅い顧客は返品交換で離脱するが，長期購買者（古い顧客）の場合には相対的に離脱率が下がってくる．長期購買者は購買経験が浅い顧客よりも肌の悩みで離脱する場合が多いことも分かる．

このように問合せをしてきた顧客がどのような属性であるかに合わせて対応を行わないと顧客の維持，拡大はできないことがわかる．購買経験が浅い顧客が返品・交換の苦情を言ってく

れば，丁寧な対応が求められるし，購買経験が長い顧客が肌の悩みを述べているときには，ちゃんと相談に乗らなくては離脱してしまう．この企業では分析結果を基にインターネットサイトの入り口を購買経験の浅いユーザとそうでないユーザで分け，購買経験の浅いユーザには企業や製品を理解してもらうためより詳しい情報提供と丁寧なサービスを行っている．

5.　CRM における「顧客の声」活用方法とシステム化

テキストマイニングの適用先として，最も多いのは，CRM 分野におけるコールセンターである．コールセンターに蓄積された問合せは「顧客の声」として蓄積されており，企業が今まで気づかなかった内容が含まれているため，「宝の山」と呼んでいる企業もある．コールセンターに対して適用して行くうちに顧客の声を活用するために必要な要件も把握できてきた．そこで本章では顧客の声を活用するための要件を定義し，その要件を満たすシステムである「顧客の声」システム CVPro 活用系について述べる．

5.1　企業の「顧客の声」の活用方法

企業ではそれぞれの部門で顧客の声の活用方法が異なっているが，顧客の声が変化したときに重要な情報が含まれていることが多い．顧客の声の変化は顧客の声の件数の変化でとらえることができる．コールセンターでは，顧客の声の変化を全体の件数増減として捉えており，件数が増加していればその件数を増やしている原因を見つけ，回答マニュアルの作成，FAQ の Web への掲載を行い，問合せ件数を減少させて，運営コストを削減している．商品を開発や担当している部門では，顧客の声の変化を商品毎の件数増減として捉え，各自の担当商品の件数増減に注視している．特に新製品はクローズアップして注視されている．マーケティング部門では，顧客を見込み客，新規顧客，通常顧客，優良顧客，離脱顧客に区分化しているが，顧客の声の変化をこの区分間の移り変わりの件数で捉えている．そのため優良顧客から通常顧客，離脱顧客に，または，通常顧客から離脱顧客になった顧客の問い合わせ内容を分析し，改善を行うことで離脱などを防いでいる．

このように企業各部門では変化を様々な視点で捉えており，そのために顧客の声に分類コードを付与して記録している．しかし，分類コード付与のルールは定めているものの，その判断には個人差があり，多くの企業では正確な分類コードが付与できていない．またこの変化を捉えるタイミングも，月次レポートなどでは対応が遅れてしまうため，毎日確認している企業もあるが，その一方で変化を見るべき各現場は忙しく，なかなか定着していない．

5.2　「顧客の声」活用システムの要件

企業における顧客の声活用の現状から，顧客の声活用のためには次のような要件が必要であることがわかる．

最初の要件は，顧客の声の変化にすぐに気づくことができることである．顧客の声の変化はその件数の変化から分かるので，件数の変化がすぐに，しかも忙しい現場で活用できるように手間を掛けずに見ることができる必要がある．

2 番目の要件は，細かく件数の変化を調べて変化の要因を探ることができることである．件数の著しい変化が見つかった場合，その変化の要因を探るため，様々な分類で更に細かく件数の変化を掘り下げて，どの商品・サービスのどの部分にその変化があるのかを詳しく調べられ

る必要がある．

3 番目の要件は，顧客の声を見るために切り口を提供できることである．変化の要因となっている商品・サービスが見えてくると，問合せそのものを参照してその原因を正確に把握する必要がある．参照する件数が多いと，全ての問合せを最初から最後まで読んでいくことは不可能である．そのため問合せの内容を表した単語や係り受けなどの切り口を提供できる必要がある．

4 番目の要件は，離脱の要因を探り，そこから離脱防止のヒントを得られることである．商品担当は 1 番目から 3 番目の要件を通じて商品視点での顧客の声の活用を行えるが，マーケティング担当には顧客という視点が必要となる．このため問合せデータに顧客のデータを加えて，離脱している顧客の原因を集計結果から探り，最終的にその切り口からその問合せを読んでいく．

5 番目の要件は，顧客の声を様々な視点で捉えるための自動仕分けができることである．コールセンター，商品担当，マーケティング部など企業内の様々な部門で変化を見る視点が異なっており，顧客の声に自動的に分類コードを付与する必要が有る．人手による分類も可能であるが，判断の個人差による仕分けのぶれを防ぐためにも，自動化された仕組みが必要である．

5.3 「顧客の声」システム CVPro 活用系

コールセンターに溜まった顧客の声を全社レベルで活用するために，これらの要件を満たす「顧客の声」システム CVPro 活用系（以降 CVPro 活用系）を開発した．CVPro 活用系は，前日までに寄せられた顧客の声をその内容や種別毎に集計して，その変化を素早く検知する速報システム，様々な切り口で顧客の声を集計して，その変化の要因分析や月次の定型分析を行う報告システム，顧客の声に購買情報を加えた顧客分析を行う分析システムの三つのモジュールがある．また，CVPro 活用系の基盤システムとして顧客の声を自動的に内容や種別で仕分けする MiningPro 21 文書マイニング・システムが，速報システム，報告システム，分析システムの集計結果から，その問合せを読む切り口として係り受けを提示する TopicExplorer が利用されている．

図 10 は CVPro 活用系を使った顧客の声活用システムである．

CVPro 活用系システムを利用すると図 10 の様に顧客の声を全社レベルで活用できる．まず，コールセンターに寄せられた顧客の声は MiningPro 21 文書マイニング・システムでその内容，種別が付与されて活用系データベースに蓄積される．次に企業の各部門では速報システム上でそれぞれの視点で問合せの件数を集計し，その変化に気づく．更に変化に気づくと報告システムを使い，様々な視点で集計して変化の要因を掘り下げる．最後にその要因を正確に把握するため問合せそのものを読むことを行う．全件は読まずに TopicExplorer が抽出した係り受けを切り口として手早く読み，適切な対処を行い，商品開発のヒントを得ることもできる．またマーケティング部では分析システムで顧客の声にその顧客の購買データを加えて，見込み客，新規顧客，通常顧客，優良顧客，離脱顧客の変化を集計して，離脱などの要因を探り，適切な対応を行うヒントを得る．

このように，CVPro 活用系は顧客の声の活用手順が企業内の広い組織で利用されることを想定しており，コールセンターのデータを全社レベルで活用することができるようになっている．

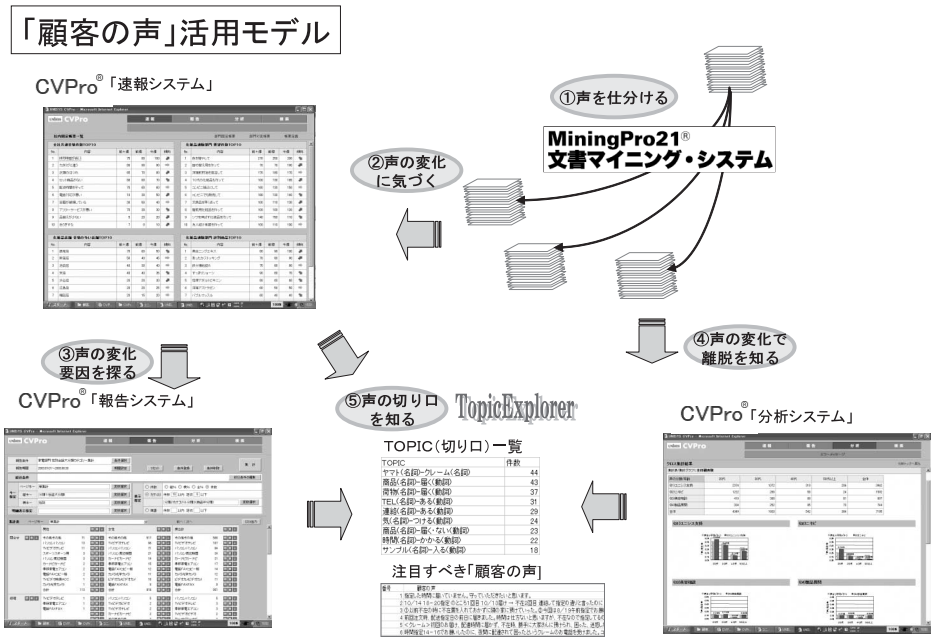


図 10 CRM における「顧客の声」システム CVPro 活用系

6. ま と め

ある業態では，初めて購買した顧客が次の購買を行う確率は約 20% であり，5 回購買を行った顧客が次の購買を行う確率は約 90% といわれている．このように顧客は購買回数が増えれば増えるほど離脱しにくくなる．このためどれほど多くの初回購買顧客に継続的に購買をさせられるかが，企業収益に影響する非常に重要な課題となっている．

CVPro 活用系は，単なる問い合わせ集計システムではなく，CRM という観点で顧客の声からマーケティング活動のためにナレッジを抽出することを目的としている．CVPro 活用系を利用することで，1 回で購買を止めてしまった顧客の離脱原因を知り，顧客維持を図ることができる．このように CVPro 活用系は顧客の声と購買情報を結びつけることで，企業収益を生み出すために活用されている．

CVPro 活用系は，CRM という観点で顧客の声を分析している業種に広く適用でき，コールセンターなどで顧客を管理している様々な企業への今後の適用が期待される．

参考文献
［1］ 北　研二著　言語と計算-4　確率的言語モデル　東京大学出版会
［2］ 徳永健伸著　言語と計算-5　情報検索と言語処理　東京大学出版会
［3］ 金明哲，村上　征勝，永田　昌明，大津　起夫，山西　健司著，統計科学のフロンティア 10　言語と心理の統計　岩波書店
［4］ 松本裕治，影山太郎，永田昌明，斉藤洋典，徳永健伸著，岩波講座言語の科学　単語と辞書　岩波書店
［5］ 工藤　拓，松本裕治著，チャンキングの段階適用による係り受け解析，情報処理学会研究報告 2000-NL-142, pp.97-104, March 2001.
［6］ 株式会社グロービス［編著］，MBA マネジメントハンドブックダイヤモンド社
［7］ レジス・マッケンナ著『ザ・マーケティング「顧客の時代」の成功戦略』ダイヤモンド社刊
［8］ Jeffrey E.F. Friedl（原著），歌代和正，鈴木武生，春遍雀来（翻訳），詳説正規表現，オライリー・ジャパン

［9］ 山口和範，高橋淳一，竹内　光悦著，よくわかる多変量解析の基本と仕組み
［10］ 田中　豊，脇本和昌著　多変量統計解析法　現代数学社

執筆者紹介 林　田　英　雄（Hideo Hayashida）
1964 年生．1987 年 3 月神戸商科大学管理科学科卒業．同年日本ユニシス(株)入社．意思決定支援システム，情報検索システムの開発を行い，現在日本ユニシス・ソリューション(株)データサイエンスビジネス部にて MiningPro 21 文書マイニング・システムの開発，適用を担当．OR 学会員．

脇　森　浩　志（Hiroshi Wakimori）
1980 年生．2003 年 3 月慶應義塾大学理工学部管理工学科卒業．同年日本ユニシス(株)入社．現在日本ユニシス・ソリューション(株)データサイエンスビジネス部にて MiningPro 21 文書マイニング・システムの開発，適用，コールセンター呼量予測システムの開発，適用を担当．