局所化指向テキストマイニングの実践と評価

2012年度

竹内 広宜

学位論文　博士(工学)

局所化指向テキストマイニングの実践と評価

2012年度

慶應義塾大学大学院理工学研究科

竹内 広宜

要旨

テキストマイニングでは，特定の文書集合においてキーワードの出現頻度を求め，傾向や規則を発見することが行われている．この時，分析に用いる観点や辞書を準備する必要があり，それらの初期設定や選択が実践上の課題となっている．本論文では，分析観点や辞書の設定や利用に局所化手法を適用した．そして，市場分析および会話分析に対して，局所化を利用した分析手法を提案し，実践例を通してその有用性の評価を行った．

以下，本論文の構成について述べる．

はじめに，第１章において，本論文の背景，課題，目的について述べる．

第２章では，本研究の関連技術としてテキストマイニングで用いられる自然言語処理技術およびテキストマイニングの活用手法について述べるとともに，これらの関連研究について述べる．

第３章では，会話分析を対象とし，タスクを持った会話からタスクの成功につながる発言パターンの抽出を行う．その際，局所化手法として，冗長な発言を含む会話データからタスクの成功に寄与する重要発言区間を同定する手法と，同定した重要発言区間からタスクの成功に関連するキーワードを抽出する手法を提案した．実践例としてコンタクトセンター受託企業で収集されたレンタカーの予約会話データを対象とした．そして，顧客が予約した車を取りに来る/来ないと結果が異なる予約会話間の差異分析を行った．長い会話の中から結果に影響を与える発言区間として顧客の最初の発言および提案時の発言を同定し，その中から結果に関連する発言パターンを抽出した．そして抽出した発言パターンを元にオペレーターへの教育に行い，予約された車の利用率を約３％向上することができた．

第４章では，市場分析を対象とし，自由回答および選択回答形式のアンケートデータから次期購買層の発見につながるルールの抽出を行う．その際，分析するキーワードを限定する局所化手法として，順序関係を持った顧客属性に対して頻度が増加・減少する傾向を持つキーワードをランキングする手法を提案した．そして，データマイニングによるルール発見の結果から，テキストマイニングで関連があると分析したキーワードと顧客属性の組を含むルールをフィルタリングする手法を提案した．実践例として，生ごみ処理機の市場分析を目的とした購買者・非購買者へのアンケートデータから次期購買層の発見につながるルールの抽出を行った．実践例では，提案手法により，マーケティング専門家が解釈・評価を行うルール数を，精度を保ちながら約1/3に削減することができた．

最後に第５章で，本論文のまとめと今後の課題および展望について述べる．

Title: Practice and Evaluation on Localization-Oriented Text Mining

Abstract:

In the text mining analysis, we usually try to get the frequencies of keywords in a selected document set. For such an analysis, it is needed to define view points and prepare dictionaries in advance. In this thesis, we considered to apply localization methods for preparing dictionaries and selecting analysis view points. We proposed localization-oriented text mining for marketing analysis and conversation analysis.

This thesis is organized as follows:

Chapter1 describes backgrounds, issues and purpose of this study.

Chapter2 introduces text mining technologies and researches on text mining applications as related technologies and studies around this thesis.

Chapter 3 describes the conversation analysis where we try to find utterances leading to insights that improve business from the conversation data. As a localization method, we proposed a method to identify important segments from the conversations which are often long and redundant, and extract effective expressions from the important segments to define the viewpoints. We applied the method to the conversation data from a car rental reservation center. We identified customers' first utterance and utterances in the proposal as important segments and extracted expressions relating to the reservation conversation where customers picked up the reserved cars. Through the education for the operators based on the extracted utterance patterns, we could improve the picked-up ratio of the reserved cars by about 3%.

Chapter 4 describes the market analysis to find rules for potential customers from a questionnaire survey data for a product. As a localization method, we proposed a method for ranking keywords correlating to segments that have ordering. We also proposed a method for filtering data mining results by using the trend analysis results by text mining. We applied the methods to the market analysis of a garbage disposal. In this analysis, the number of extracted rules that the marketing expert had to assess was reduced into about 1/3.

Finally, in chapter5, we conclude this thesis and point out future directions.

目 次

図 目 次

表 目 次

第1章　序論

1.1　背景

ビジネスの現場では，大量に蓄積されたデータを活用する試みが昨今盛んに行われている．蓄積されたデータの中には，販売履歴データのようにデータ型やスキーマを定義してデータベースに格納した構造化データだけでなく，テキスト文書のように，内容に関する情報を別途付加しなければ分析が困難な非構造データもあり，両者を活用することが求められている [17]．そして，構造データを活用する技術としてデータマイニング，非構造データであるテキストデータを活用する技術としてテキストマイニングが広く研究開発されている．これらの技術をビジネスの現場に適用する活動は Business Intelligence（BI）と呼ばれ，様々な研究がなされてきた [27]. BI で行われてきたデータマイニングやテキストマイニングの適用では，大量のデータからの情報抽出とその可視化が中心であり，様々なシステムが研究開発されている [14][20]．その一方で，抽出され可視化された情報から知見を見出すのは分析者にゆだねられていた．

そのような状況に対し，近年 Business Analytics(BA) と呼ばれる活動が企業内で試みられている [30]．BA は，情報を整理するだけでなく，将来を予測しビジネスを最適化させる活動である．ビジネスを最適化させることが BA の目標であり，多くの場合，対象分野の専門家が分析者となる．したがって情報抽出や可視化だけでなく，有効なビジネスアクションにつながる知見を得るための分析手法が必要となってくる．

有効なビジネスアクションを立案するために，テキストマイニングを用いて何らかの知見を得るためには，単に情報抽出をするだけでは不十分である．通常，テキストマイニング分析においては，まず何らかの分析観点を定義する．分析観点は抽出された単語（キーワード）や係り受け表現に割り当てる意味ラベルであり，例えば，製品名，ソフトウェア名，金額表現，要望表現などがある．分析観点を定義した後，各観点に関係するキーワードや表現を辞書として登録する．これら分析観点および辞書は対象データおよび分析目的に応じて準備する必要がある．分析では，選択した観点に関係する文書がどのくらいあるのかを把握することができ，観点間の相関情報を得ることができる．そして，特定の文書集合において強い相関関係を持つ観点の組み合わせを見つけることが，テキストマイニン

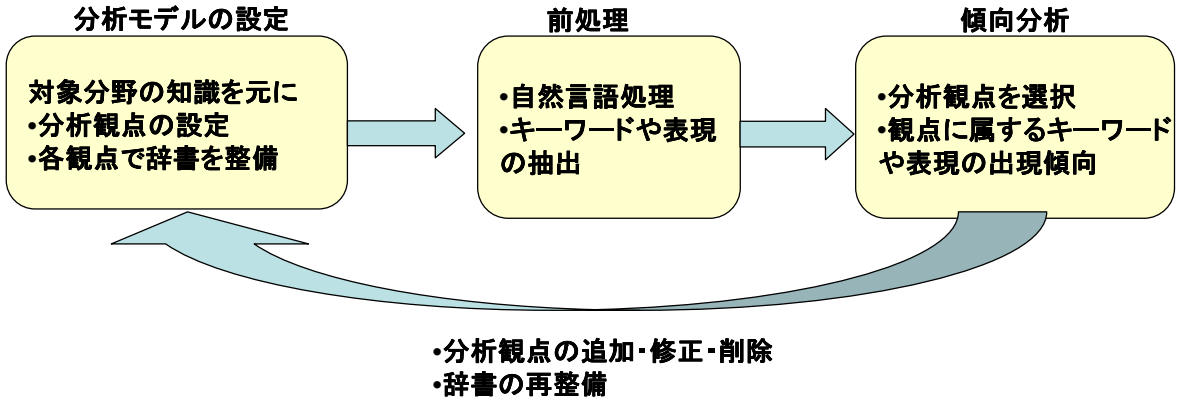

図 1.1: テキストマイニングの分析ループ

グ分析で試みられている [87]．例えば，PC ヘルプセンターの問い合わせデータから,「不具合」という分析観点の辞書に含まれる表現（「動かない」,「フリーズする」など）について出現頻度を製品ごとに集計し，製品固有の問題の発見につながる傾向を得ることができる．

この分析観点の設定や辞書の整備は，分析前に行うだけでなく，分析結果を元に追加・修正を行う必要がある．そして，その後，再度分析を行うという分析ループが回る (図 1.1)．ここで，分析観点や辞書は分析モデルと考えられるが，その初期設定は分析者の対象分野に対する知識に依存する．適切な分析モデルを初期設定できないと満足な結果が得られない場合がある．分析時においては，どの分析観点を選択すれば有効な傾向情報が得られるか分からないため試行錯誤しながら分析を行うことになる．また，テキストマイニングで得られる結果は，単語や表現の出現頻度であり，具体的なアクションにつなげるためには詳細な分析が必要な場合もある．

データマイニングの実践では，マイニングのプロセスにおける前処理，マイニング，後処理の比は，およそ 7:1:2 であるとされている [76]．テキストマイニングの実践においても，分析観点や辞書の初期設定（前処理）や分析時における分析観点の利用や分析結果の解釈（後処理）が大きな割合を占める．そのため，有効な分析を行うためには，効果的に前処理および後処理を行うことが必要となっている．

1.2 目的

本研究では，テキストマイニングの実践における，前処理や後処理において，局所化手法の利用を考える．局所化手法とは，分析目的に応じて対象範囲を限定する手法である．

テキストマイニングの前処理において，分析者が対象データのサンプルを目視して，分析観点を決めることが多い．この時，各文書のサイズが大きい場合，分析者は文書を読み込み必要があり，分析観点の設定にコストがかかる．そこで，局所化手法を適用し対象データを限定し，限定した範囲から分析観点を定義し，辞書を作成することで，前処理を効果的に行うことを考える．後処理でも，分析結果を精査し，知見に結びつけることができるか評価する必要があり，分析結果が多くなるにつれてコストがかかる．そこで，情報抽出で得られた様々な分析結果を局所化することで絞り込み，専門家が精査すべき分析結果を削減することで，後処理を効果的に行う分析手法を考える．

本研究では，会話分析および市場分析において，これら 2 つの局所化手法を利用した分析手法を実践し，実践例を通してその有用性を検証する．

1.3 本論文の構成

本論文の構成は以下の通りである．

- 第 1 章　序論
  本研究の背景及び目的を明らかにし，本研究が目指すところについて述べる．また，本論文の構成についても述べる．

- 第 2 章　背景知識・関連研究
  第 2 章で，本研究の関連技術として，テキストマイニングシステムを構築するにあたって必要となる技術について述べる．また，テキストマイニング技術の活用について，関連研究を述べる．

- 第 3 章　有効な分析観点の設定と表現の自動抽出と会話分析への適用
  第 3 章では，会話分析を対象とし，タスクを持ったビジネス会話から成功につながる発言パターンの発見を試みる．その際，局所化手法として，冗長な発言を含む会話から分析に寄与する重要発言箇所の同定する手法を提案する．そして，同定された重要発言箇所から，有用な表現を抽出し，分析観点や辞書の構築を支援する手法を提案する．レンタカーの予約会話への適用を通して，提案手法により，分析観点

や辞書の初期設定を分析者の知識に依存せずに行え，有用な分析結果が得られることを示す.

- 第 4 章　市場分析におけるテキストマイニングを活用したデータマイニングの実践
  第 4 章では，市場分析を対象とし，アンケートの定型回答データのデータマイニング結果と合わせて次期購買層の発見を試みる．ここで，局所化手法として，データマイニングで得られる特定の顧客層にテキストマイニングの傾向分析結果を対応付けるとともに，データマイニングで得られる様々な顧客層をフィルタリングする手法を提案するまた，テキストマイニングの傾向分析の際に，顧客属性と順序相関性の高い分析観点を選択する手法を提案する．生ごみ処理機という普及が進んでいない製品に対する，購買者・非購買者のアンケートデータを用いた市場分析への適用を行う．そして，提案手法により，専門家によるデータマイニング結果の解釈・評価が効果的に行えることを示す.

- 第 5 章　結論
  最後に第 5 章では，本論文のまとめと今後の課題および展望について述べる.

第2章　背景知識・関連研究

2.1　テキストマイニングシステム

2.1.1　テキストマイニングシステムの概要

テキストマイニングシステムには，テキストデータを文字列の集合ととらえ，大量の文字列の中から頻出する文字列パターンを抽出するというアプローチがある [82]．しかしながら，通常，テキストマイニングシステムでは，対象となる入力テキストデータからあらかじめキーワードや表現などの情報を抽出し（情報抽出），その結果をデータベースなどに格納し，分析に用いることが多い [15]．分析者は抽出済みのデータにアクセスし，集計処理を行う（ランタイム分析）．本研究で使用するテキストマイニングシステムも同様な構成である．その全体図を図 2.1 に示す．

2.1.2　テキストデータからの情報抽出

本節では，テキストマイニングで必要となるテキストデータからのキーワードや表現を抽出する技術について述べる．多くのテキストマイニングシステムでは，大量のテキストデータを分析対象として扱うため，分析前に一括して入力テキストからキーワードや表現

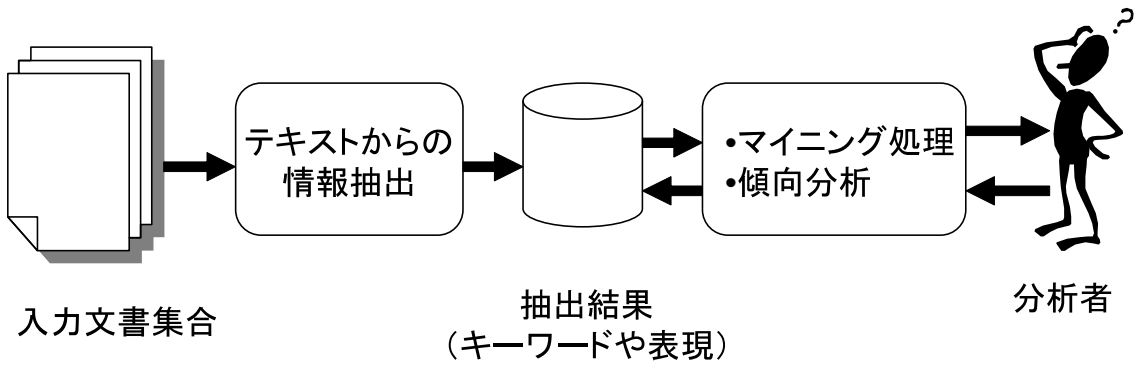

図 2.1: 本研究で使用するテキストマイニングシステムの全体図

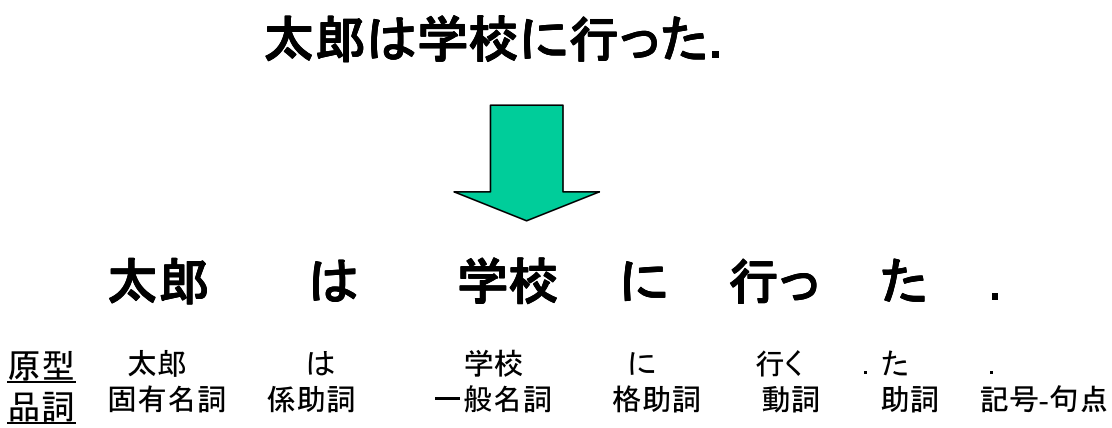

図 2.2: 形態素解析結果の例

などの情報を抽出することが多い.

この入力テキストデータから，一括して情報抽出を行う部分では，テキストデータに対して，形態素解析，構文解析といった自然言語処理が適用される.

形態素解析

形態素解析 (Morphological Analysis) は，自然言語で書かれた文を形態素と呼ばれる最小単位の列に分割し，分割した形態素に原形や品詞を付与する処理である [89]. 図 2.2 に形態素解析の結果例を示す. 現在，形態素解析の精度は 90%を越えるようになり，様々な応用に利用されるようになっている. 日本語を対象にした形態素解析器として，JUMAN[28] や ChaSen[6] といったツールが公開されている.

構文解析

構文解析 (Syntactic Analysis) は，ある文章の文法的な関係を抽出することであり，自然言語処理だけでなく，プログラミング言語などの形式言語の解析にも使用される. 自然言語処理では，形態素または文節間の関係を抽出することである. 図 2.3 に構文解析で得られた係り受け情報の例を示す.

係り受け関係の抽出手法には，語彙機能文法 (LFG) や主辞駆動句構造文法 (HPSG) といった文法を使用した手法 [89] や，統計学的な手法 [36] など，様々な手法がある. 構文解析技術の精度も近年向上しているが，自然言語で書かれた記述には，曖昧性があるものが存在する. 例えば,「美しい水車小屋の乙女」という文において,「美しい」の係り先として

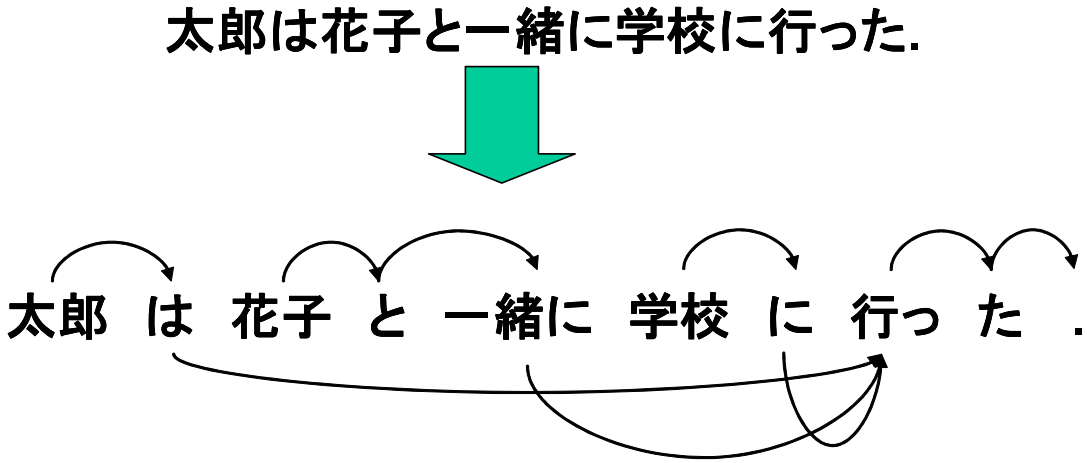

図 2.3: 係り受け情報の例

「水車小屋」と「乙女」の両方が考えられる．単語の意味から曖昧性が解消される場合もあるが，周りの文脈や書き手の意図を考慮しなければ曖昧性が解消されないものがある．構文解析器の結果では，係り先は 1 つに定まるため，書き手が意図した結果が得られないことがある．

形態素解析器と同様，日本語を対象にした構文解析器についても KNP[29] や CaboCha[3] といったツールが公開されている．

固有表現抽出

現実世界に存在するテキストのほとんどにおいて，人名，地名などの固有表現 (Named Entity) が記述されている．辞書に登録されていない固有表現は形態素解析などの結果では，未知語となる．固有表現抽出 (Named Entity Extraction) は，テキストから自動的に固有表現を抽出する技術である．

例えば，IREX(Information Retrieval and Extraction Exercise)[50] では，固有表現として，組織名 (ORGANIZATION)，人名 (PERSON)，地名 (LOCATION)，日付表現 (DATE)，時間表現 (TIME)，金額表現 (MONEY)，割合表現 (PERCENT) ，固有物名 (ARTIFACT) の全 8 種類を定義している．図 2.4 に固有表現の抽出例を示す．固有表現抽出では，定義された固有表現を明示したコーパスデータを準備し，機械学習アルゴリズムを用いて抽出器を作成することが行われている．日本語の固有表現抽出では，Support Vector Machine(SVM) を用いることで 80%の精度で抽出が可能となっている [81].

ユーザー辞書適用による意味ラベル付与

分析対象データに情報抽出処理を適用する前に，辞書を作成する．ここでの辞書は対象分野に固有な語をまとめたものである．様々な観点（カテゴリ）ごとに語をそれらの同義語と共に登録する．例えば，PC のコールセンターでのテキストマイニングであれば，ソフトウェアという観点に対し,「Windows XP」というキーワードを正規形とし「WinXP」といった同義語とともに登録する．このユーザーが定義した辞書を適用し，複数の同義語をある一つの正規形にまとめ，単語に観点 (カテゴリ) を付与する処理が行われる．図 2.5 に形態素解析結果にユーザー辞書を適用した例を示す.

本処理を行うには，辞書の作成では，まず観点を定義し，分析対象データのサンプルに形態素解析を適用して得られた名詞や未知語から辞書エントリーを作成していく．サンプル文書から得られた名詞のうち，頻出する語は一般の文書でも出てくる語であり対象分野の専門用語でない場合が多い．文書から専門用語を抽出する手法として，新聞記事などのコーパスデータにおける出現頻度との比較を元に，専門用語を抽出する手法がある [48]. このような手法は仕様書などからの技術用語の抽出などで用いられている [19][54].

一方，コンタクトセンターなどで人が書いたテキストデータなどには様々な異表記が存在し，辞書では同義語として正規形に紐付けをする必要がある．そのため，同義語候補を自動抽出する技術が開発されている [40][84]．また，テキストマイニングにおける辞書作成作業を支援するツール [92] も開発されている．

パターンマッチングによる意味ラベル付与

辞書によって，単語に意味ラベルを付与することができるが，単語への意味ラベルだけでは十分な情報が得られないことがある．例えば，PC コールセンターのテキストマイニングにおいて辞書を用いて「CD ドライブ」という単語に対して「部品・機能」といった観点 (意味ラベル) を付与することができる．しかしながら,「CD ドライブ」に関して，どのような不満・希望・疑問を持っているかどうかはわからず，単語間の関係も考慮した情報抽出が必要となる.

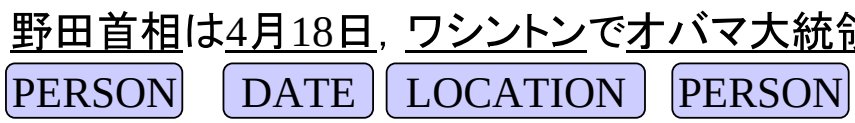

図 2.4: 固有表現抽出の例

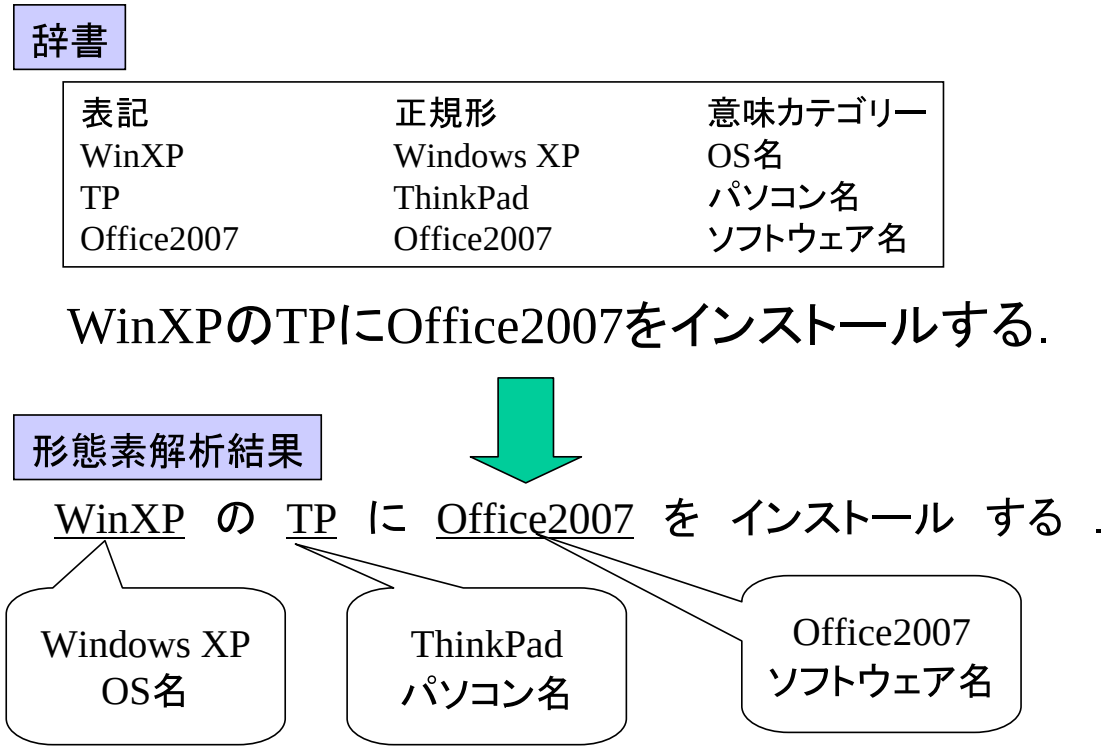

図 2.5: ユーザー辞書適用結果の例

このような目的に対して，特定の単語列や係り受け構造に対して付加情報を付与するパターンマッチング処理も行われる．例えば,「数字＋円」という単語列パターンに対して「金額表現」と情報を付与することで，異なる金額表現を同一視して集計することが可能になる．また,「名詞 → が → 動詞」という係り受けパターンを適用することで,「何がどうする」という表現を収集することができる．テキストから表現を抽出する手法として，形態素解析結果に対する正規言語を元にしたパターン記述言語と抽出器が開発されている [35][9].

述語項構造分析による意味情報の取得

対象データの特徴を利用し，特定の意味表現を自動獲得する手法が考えられ始められており．その一つとして構文解析技術を適用して得られた述語項構造から，名詞や動詞を意味情報とともに抽出することが試みられている．ここでは，その一例として，要求仕様書のひとつであるユースケース記述から名詞や動詞を意味情報とともに抽出し，ユースケース記述のモデル化とその活用し，記述をモデル化する試みについて述べる．

ユースケースは，あるシステムの機能について，システムとその利用者との間のインタ

ラクションを記述したものである [10]．構築するシステムを利用する顧客視点でシステムの振る舞いが記述されているため，開発者とエンドユーザーの意思疎通に役立ち，要求を引き出しやすいという利点がある．要求の獲得後，ユースケースを元に，外部設計，詳細設計が行われる．そのため，ユースケースの記述が不明確であると，間違った設計が行われる可能性がある．ユースケースは利用者に見える振る舞いを示しているため，本来ユースケースで意図していた内容と違う理解で設計がなされた場合，最終的にできあがるシステムはユーザーが求めるものと異なる可能性が高い．そのため，ユースケースに書かれている内容はどんな単純な記述であっても，読み手によって一意に決まる内容かどうかを吟味する必要がある．そのため，レビューなどを通してその品質を高める必要があるが，大規模システムになると大量のユースケースが作成され，頻繁に更新される．したがって，人手を中心としたレビューを限られた期間内に行うことが難しくなっている．また，仕様書を元にテストケースを作成することが可能であるが，前述の通り，仕様書は頻繁に更新されるため，ユースケースとテストケースとの間のトレーサビリティが確保されなくなる可能性がある．

そこで，ユースケースをモデル化し，ユースケース記述の品質チェックやテストケースの生成などに利用することを考えられている．ユースケースに対して自動的にモデル化することができれば，品質チェックの一部が機械化でき，レビューを効果的に行うことができる．ユースケース記述のモデルとして Text-To-Test[58] で定義されたモデル (Use Case Description Model) を用いて記述をモデル化する．モデル化の例を図 2.6 に示す．モデル化では，ユースケース記述から述語項構造を抽出し，そこからユーザーまたはシステムの振る舞いである Action を同定し，意味ラベルを付与する．また，Action の動作主 (initiator) になる Actor に USER・SYSTEM といった意味ラベルを付与する．意味ラベルの付与に際して，[62] では，ユースケース記述はユーザーとシステムの振る舞いが記載されているという特徴に基づいて意味制約を用いた手法を提案している．そこでは 13 の意味制約を定義することで，Action，Actor に意味ラベルが示されている．

抽出された意味ラベルを用いて，ユースケースの再利用化に向けた品質分析やユースケースからテストケースの自動生成などが試みられている [38] [79]．

その他の表現抽出技術

辞書および構文的な構造パターンを抽出するだけでなく，対象データに特有の表現を抽出することが行われている．Web 上にある個人が作成した文書には，自分の考え，感想といったものを記述したものが多い．そのため，評価や意図を表す表現を抽出することで，

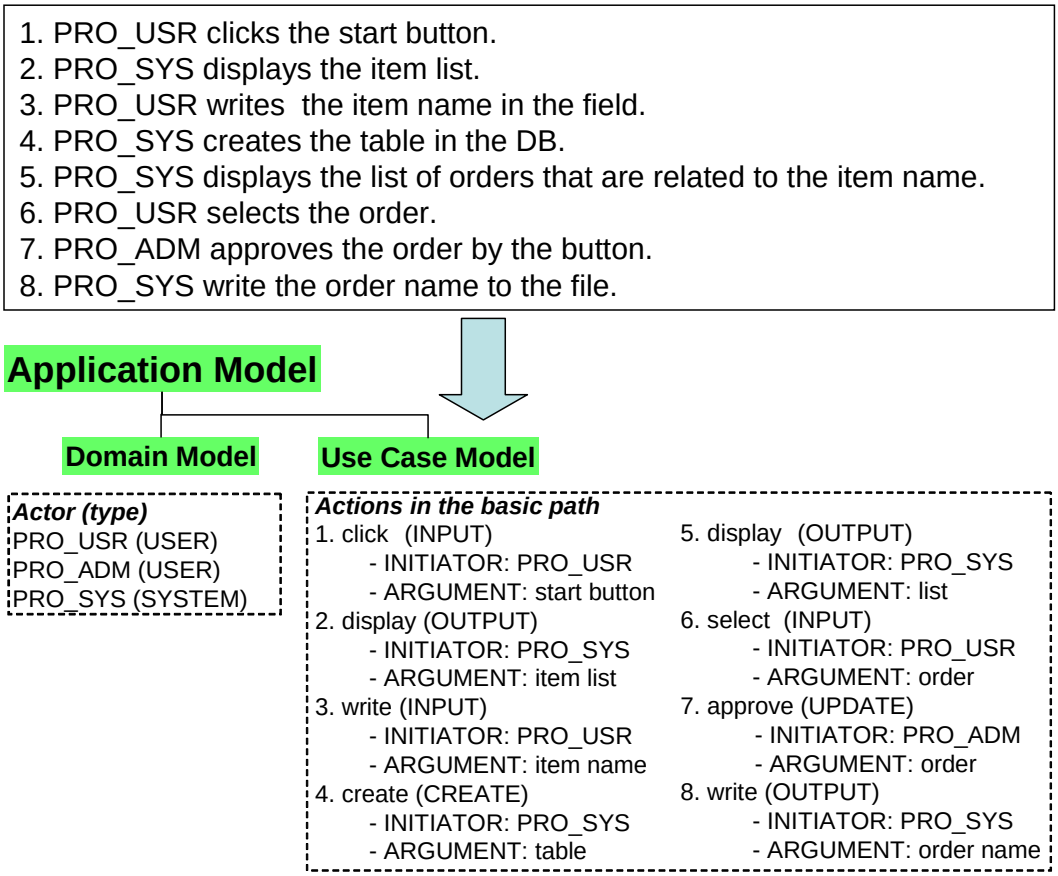

図 2.6: ユースケース記述のモデル化の例

何が好意的な評価を得ているか，といった分析が可能になる．そこで,「名詞+が → 形容詞」といった係り受け表現に肯定・否定といった極性情報を加えた評価表現を抽出することが広く行われている [77].

一方，特許や製品紹介などの技術文書には記載されている技術によって，実現される効果などが書かれている．このような情報を抽出し，整理することで，技術動向調査などに活用できることが考えられる．[93] では，技術文書から，技術の特長を示す表現を抽出する試みがされている．

テキストマイニングシステムにおける情報抽出処理

以上で述べた情報抽出を順次行うことで，図 2.7 に示すように，入力テキストから様々なレベルの情報を抽出することが可能になる．

このようなテキストマイニングシステムでは，テキストデータから情報抽出を行うた

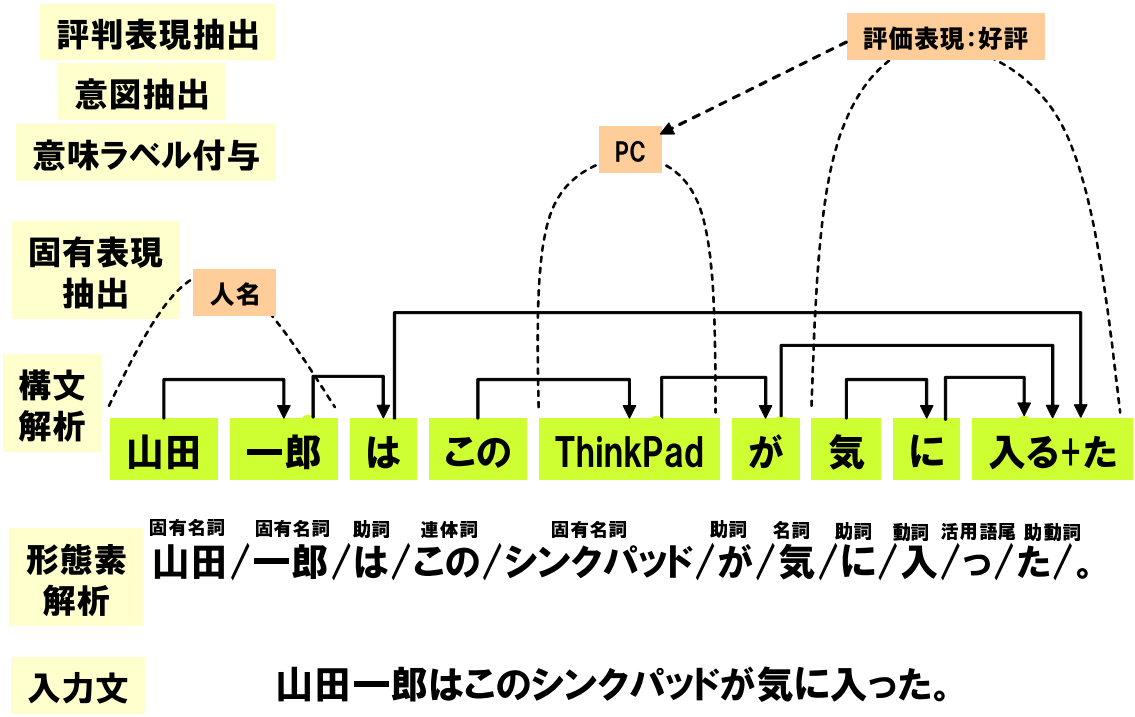

図 2.7: 情報抽出技術の適用結果例

めに，様々な自然言語処理を適用する必要がある．自然言語処理には，共通のフレームワークがなかったため，過去の研究成果を組み合わせて，より高度な処理システムを作るということが容易ではないという課題があった．Unstructured Information Management Architecture (UIMA) は，このような多様なツールを連結したいというニーズのもと，テキストを中心とした非構造情報を処理するモジュールの共通フレームワークとして開発された [16]．UIMA ではそれまでデータ構造の互換性がなかったために相互運用できなかった様々な自然言語処理ツール（形態素解析，構文解析，パターン抽出）を CAS (Common Analysis Structure) という共通のデータ構造を用いることによって統合している．UIMA は現在オープンソース化されて Apache Software Foundation のもとで開発が続けられ Java で実装された非構造情報分析アプリケーション開発用の SDK (Software Development Kit) がフレームワークとして配布され，利用されている [67][80]．これにより，形態素解析や構文解析といった独立した処理コンポーネントを組み合わせた情報抽出システムを構築することが容易になっている．

UIMA ではテキストデータをはじめとした非構造データを CAS(Common Analysis Structure) という共通のデータ構造で扱い，Text Analysis Engine (TAE) と呼ばれる共通のインターフェースを持つ処理モジュールで処理を行う．共通のデータ構造を用いるため，UIMA

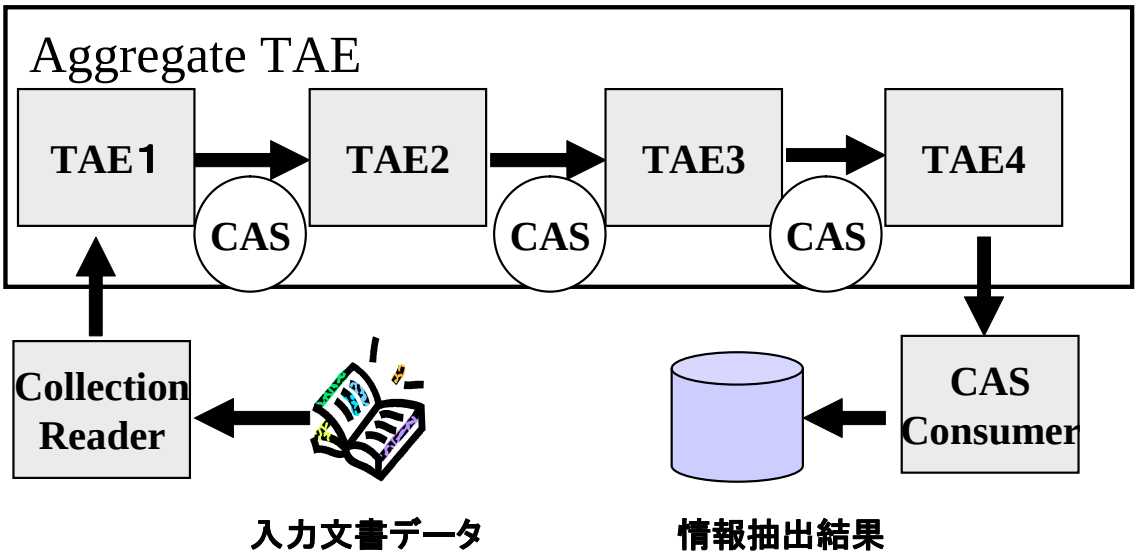

図 2.8: UIMA を用いた情報抽出システムのアーキテクチャ

上で処理した結果をデータベースに格納し，構造化データと同様に処理する，といったことが可能となる．このことから，UIMA は非構造データと構造化データの架け橋の役割を果たしていると考えることができる．一つの TAE は特定の文字列を抽出する処理や形態素解析、構文解析といった自然言語処理を行うモジュールとなる．TAE は単独だけでなく複数適用することもでき，TAE をある目的に合わせて決まった順番に適用されるようにひとまとまりに構成したものを Aggregate TAE と呼ぶ．各 TAE は処理結果をアノテーションという形式で CAS に追加する．例えば形態素解析を行う TAE の場合，文を単語に切り分けた後，各単語に対して正規形や品詞といった情報をアノテーションとして付与する．TAE を複数適用する場合，処理ごとに CAS に対してアノテーションが追加される．

UIMA の重要な特徴の一つは TAE の相互運用性である．これは TAE がどのような情報をアノテーションとして付与するのか，その定義情報を公開することで，1 つの TAE が付与した CAS 上のアノテーションを利用し，さらに別の TAE が新たな分析を行えるようにしているためである．TAE の適用順序あるいは TAE 間の依存性に注意し，TAE を柔軟に組み合わせることで複雑な情報抽出が実現できる．例えば，カーネギーメロン大学 (CMU) では，40 個の TAE が UIMA Component Repository として公開されている [49].

テキストマイニングシステムでは，既存の自然言語処理コンポーネントをそれぞれ呼び出す TAE を，Aggregate TAE 上に配置することで情報抽出システムを実現することができる．UIMA を用いた情報抽出システムのアーキテクチャを図 2.8 に示す．情報抽出の結果，各文書データから，単語やパターンに合致した特定の表現が，意味ラベルのような観

点（カテゴリ）情報および記述中の出現箇所情報とともに抽出される．

2.1.3 ランタイム分析システム

情報抽出の結果，抽出された表現・キーワードを特定の形式で保存することで，頻出語や文書内に含まれる語の集合を得るといった分析が可能となる．この部分は通常ランタイム分析システムと呼ばれる．

ランタイム分析システムではまず情報抽出結果に対してインデックスを作成する．抽出された各表現に対して，出現する文書の ID と出現箇所を対応付けたインデックスファイルと，各文書 ID に対して，抽出された表現を対応付けたインデックスファイルを作成する．また，観点（機能名，金額表現といったカテゴリ）ごとに抽出された表現もインデックスファイルとして保持する．これにより，以下の検索・集計処理を行うことが可能となる [5][56].

- 特定の表現を含む文書数
- ある文書集合において特定の観点に属する表現の出現頻度の分布
- ある文書集合における特定の表現の出現頻度

これらの検索・集計処理を利用することで，キーワードや表現の出現傾向を求めることができる．例えば，2 番目の集計処理を行うことで，対象文書全体 10000 件において，要望表現に属する表現の出現分布を調べることができる．また，1 番目の検索処理を行った後で，2 番目の集計処理を行うことで,「AAA」という製品名を含む文書の 1000 件における要望表現に属する表現の出現分布を調べることができる．これら 2 つの出現分布を比較することで，特定の製品に多く出てくる表現を見つけることが分析の 1 つとして行われている [86]．例えば,「音量…調整したい」という要望を表す係り受け表現が全体で 500 件,「AAA」という製品名を含む文書で 150 件出現している場合を考える．[86] では，それぞれの文書集合での言及頻度の比として，式 (2.1) で示す相対頻度を定義している．

$$\text{相対頻度 } Rf = \frac{\text{特定の文書集合での出現確率}}{\text{全体での出現確率}} \tag{2.1}$$

上述の例では，$Rf = \frac{500/10000}{150/1000} = 3$ となる (図 2.9)．そして，相対頻度が高い表現が，注目している文書集合に関係性が高い表現として抽出している．

例えば,「欲しい」を含む文書集合に対し，機能名という観点の属する表現の出現頻度分布を調べることができる．また，文書集合を絞り込む検索質問を変更・追加すること

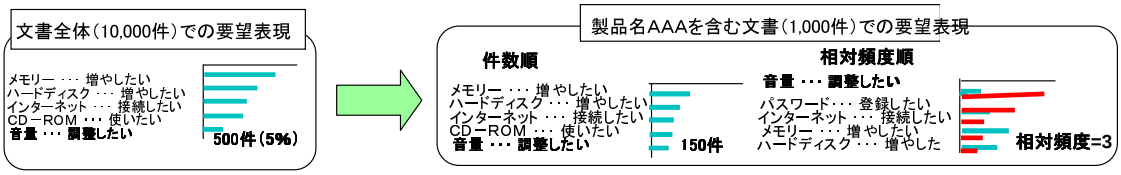

図 2.9: 相対頻度を用いた分析例

表 2.1: 2 次元表による分析例

不具合を表す表現	製品名 (文書数)		
(言及されている文書数)	A(100)	B(100)	C(100)
異音 (31)	15	11	5
発熱 (23)	4	6	13
エラー (17)	6	7	5

や，出現頻度の分布を見るための観点を変更することで，分析者は様々な観点から繰り返し分析を行うことができる．同様に，文書集合間の差を求める分析の一つとして，ある観点 (カテゴリ) に属する表現の出現頻度を比較することが行われている．例えば製造業のコールセンターでは部品名，苦情，要望，質問といった観点とそれぞれに該当する表現を事前に集め辞書に登録し分析に用いられている．こうすることで製品ごとに分けられた文書集合を分析する際，設定した観点に属する表現を縦軸，製品名を横軸にとり，両属性を満たす文書数を表す 2 次元表を作成することができる．表 2.1 は製品ごとに不具合を表す表現が言及されている文書数を集計した例である．この分析例では,「異音」は製品 A と B で,「発熱」は製品 C で多く言及されている．この結果から各製品がどのような不具合を持っている可能性があるかを推察することができる．このように観点を選択し，2 次元表を作成することで，分析者は文書集合間の関係を概観し，知見につながる傾向を得ることができる．

通常，テキストマイニングで扱うデータには，テキスト以外に日時，作成者名（書き手），問い合わせタイプ，顧客の性別や年代といった定型情報が付与されている．このような定型情報と組み合わせることで，単語や共起する単語対の出現頻度といった傾向だけでなく，以下のような分析が可能となる．

- 日時情報と組み合わせ，単語の時系列的な出現頻度情報
- 特定の文書作成者に関連性の強い単語の抽出

- 各顧客属性に関連性の強い単語の抽出

2.2 テキストマイニング技術の利用

現在，様々なテキストマイニングシステムが開発され [14]，様々な対象データに適用されている．ここでは，いくつかの対象データごとに既存の研究を概観する．

2.2.1 Web 上のテキストデータの分析

インターネット上には，口コミサイト，掲示板，ブログ (Weblog) など様々な意見が書かれた Web ページが存在する．[7] [8] では，これらの Web ページデータを分類することが行われている．これらの Web ページ上のテキストから，評価表現を抽出し，どのような商品が好意的に受け止められているのか？商品の何が不評なのか？といった意見や評判の分析が行われている [46]．また，テキストとして書かれた情報について，それらの間に非明示的に存在する，同意，対立，根拠といった意味的関係を抽出・可視化する研究が行われている [41]．Web 上の個人の日記であるブログ (Weblog) に対しては，データを定期的に収集し，時系列的な分析をし，非常に盛り上がっている話題に出てくるキーワードなどを可視化するシステムが研究開発されている [42]．

また，tweet と称される短文を投稿し閲覧できる Twitter と呼ばれるサービスが近年展開され，多数の利用者が身の回りの出来事を中心に様々なデバイスから短文を投稿している．Twitter 上のテキストデータは文字数制限があるため，内容が的確に記載されている tweet も多数あり，分析がしやすいという利点がある．この Twitter 上の投稿データを分析し活用する研究が始まっている [34][53]．

2.2.2 医療論文データからの知識発見

MEDLINE は米国国立医学図書館が医学を中心とする生命科学関係の論文情報を収集したオンラインデータベースである．データベースにアクセスする PubMed と呼ばれる検索エンジンが提供されているだけでなく，データ自身も公開され入手可能となっている．データには論文の要約の他に，書誌情報や該当分野などを表すカテゴリ情報が付与されている．MEDLINE には大量の論文情報が蓄積されており，広く医薬系研究者に利用されておりテキストマイニングの重要な対象となっている [11][60]．

MEDLINE データからのパターン発見に関する研究として，複数の文献内におけるキー

ワードの間接的な共起出現を抽出するものがある [61]．この研究の適用によって,「マグネシウム」と「片頭痛」との間の間接共起が見つかり，従来の文献には書かれていなかったマグネシウムと片頭痛の間の因果関係が発見できたことが報告されている．また，遺伝子やタンパク質の間の相互作用を文献情報から抽出し，可視化する試みがなされている [12][65].

MEDLINE のテキストデータは論文の要約であり，遺伝子・タンパク質といった専門用語が記載されている．これらの専門用語は複雑な複合語であることも多いため，通常の形態素解析や構文解析が失敗することが多い．そのため，専門用語辞書を作成し，それを活用した言語処理を行う必要がある．医療生命科学のエリアでは Gene Ontology など様々な知識体系情報が構築されている．このような情報を言語リソースとして言語処理に活かし，テキストマイニングシステムが研究開発されている [68][90].

2.2.3 コンタクトセンターにおける顧客の声の分析

企業において，顧客への対応業務を専門に行う部門がコンタクトセンターである．外部からの電話対応業務が中心であったため，コールセンターとも呼ばれるが，E メールや Web を利用した問い合わせもあるため，コンタクトセンターと呼ばれるようになってきている．複数のチャネルで顧客からの問い合わせが来るが，電話での問い合わせが多く，企業によっては問い合わせ数は毎月数万件になる．

コンタクトセンターにおける電話対応業務では，オペレータが対応内容のメモをコールログという形式で残し，膨大なコールログが電子的に蓄積されている．このようなコールログを対象としたテキストマイニング分析が行われている．情報抽出の結果，抽出された情報をデータベースに格納することで，構造化データに対して行われてきた頻出パターンマイニング [25] が適用できる．例えば，抽出された係り受け表現の中から頻出パターンを発見し，FAQ 作成支援に用いるという試みが行われている [37]．頻出パターンを自動的に発見するのではなく，文書集合ごとに頻出するキーワードや表現を比較可能な表形式で可視化する試みもある [43]．このようなアプローチによって，特定の製品に出現している表現を同定し，問題の早期発見につなげられた実践例が報告されている [86].

一方で，コンタクトセンターにおいて，オペレータが残すコールログだけでなく，顧客との会話を直接録音し，そのデータを分析活用する試みもはじまっている．基本的な活用例として，分類技術を利用した，コール種別の判定 [66][75] や問い合わせ先の自動判別 [22][32] がある．コールログ（要約）の作成支援 [13]，オペレータの対話支援 [39]，領域知識の構築支援 [51] といったシステムも研究開発されている．また，近年，企業における問

い合わせ対応には様々な法規制や，ビジネス損失を回避するためのガイドラインがある．そのような規制・ガイドラインに沿ってオペレータが会話できるているかをモニタリングする研究も行われている [23] [69]．会話データからの知見の導出を目的とした分析研究としては，会話データから頻出する対話パターンを抽出する研究が行われている [45]．

第3章　有効な分析観点の設定と対象概念の自動抽出と会話分析の適用

3.1　会話データの分析

近年，蓄積されたテキストデータを活用するテキストマイニング技術が研究・開発されている．特にCRM(Customer Relationship Management)の分野では蓄積された顧客の声をテキストマイニング技術を通して分析し，ビジネスに活用することが行われている[87]．従来，分析の対象となっていた顧客の声はコールセンターの電話応対者（オペレーター）が顧客との会話の後に会話内容を記したコールメモが中心であった．一方で，音声認識技術の向上により自動的にテキストに書き起こされた生の会話をテキストマイニングの対象にすることが可能となってきており，会話データの分類などの研究がなされている[51][75]．本章ではコールセンターでの会話を分析するシステムを構築し，有用な知見を得ることを試みる．

コールセンターで行われているテキストマイニング技術を用いた分析では，あらかじめ，製品名や問題表現といった意味ラベルを分析観点として定義し，コールメモに含まれる各観点に関連するキーワードや表現を辞書として登録する．そして，分析時には，製品名と問題表現といった観点に属するキーワードや表現の間の関係を２次元の表形式で表示するといったことがあげられている[86]．このような分析によって，ある製品のみに特徴的に出現している問題表現を概観することができ，知見につながる傾向を容易に得ることが可能となっている．書き起こされた会話のデータに対しても自然言語処理技術を適用し有用な表現を抽出することで同様の分析を行うことができる．

様々な角度から蓄積されたテキストデータを分析するためには観点(カテゴリ)を設定し，各観点ごとに該当する概念(キーワード，表現)を集めて辞書を構築する必要がある．従来のコールメモの分析では各文書が対話中の重要な内容のみで比較的短く記述されているため，観点の設定や各観点ごとの辞書構築はサンプルデータ中に頻出する語を手がかりにすることで比較的容易に人手で行うことができた[86]．しかしながら会話データの場合，各会話のデータサイズが大きくなり冗長な表現も多く含まれる．そのため，高頻度語

の情報だけから，どういった観点が分析に有効であるのかを事前に設定したり適切な辞書を準備することは難しい．営業活動や問題解決のように目的と結果を伴う会話において何が成功に寄与しているかといった要因分析は，生産性の向上が期待できることから，テキストマイニングの魅力的なアプリケーションである．この要因分析においては，冗長性の高い会話の一体どこに着目すれば有益な知見の獲得につながるかの判断が重要であるが，分析者の勘に依存しながら試行錯誤していては効率が悪いうえ，たとえ要因が存在しても，そこに気づけるとは限らない．そのため分析者の知識や経験に依存しない分析システム・手法が必要となっている．

本章では，会話データを処理するためのマイニングシステムとそれを用いてコールセンターにおける目的を持ったビジネス会話から有用な知見を得るための分析手法について述べる．本章の構成は以下のようになっている．3.2 節でコールセンターにおける目的を伴う会話とその特徴について述べ，3.3 節で従来のテキストマイニングシステムを拡張した会話分析システムと会話データの分析における問題点について述べる．この問題点を解決するために，3.4 節で目的を持ったビジネス会話を顧客満足度の観点からモデル化し，3.5 節で定義したモデルを用いた会話データ間の差の要因を見つけ出す要因分析の手法を提案する．3.6 節および 3.7 節で提案した要因分析手法を実際のデータに適用し，提案した会話モデルおよび分析手法を検証する．適用結果などについて 3.8 節で考察し，最後に 3.9 節で本章のまとめを行う．

3.2 コールセンターにおける目的をもったビジネス会話

本研究では企業のコールセンターなどに寄せられる目的をもった会話を対象とする．目的をもった会話の例として，電話による商品の予約・購入といったものが挙げられる．図 3.1 にレンタカー予約センターにおける会話データの例を示す．このような会話には以下の特徴がある．

1. 会話は顧客とオペレーター（コールセンターの電話応対者）との間のやりとりで構成される
2. 会話の流れはある程度事前に定義されている
3. ビジネス結果が各会話データに割り当てられている

前に述べた，レンタカーの予約センターでは開始，予約詳細，提案，顧客情報の取得，予約再確認と必須事項の確認，終了，といった会話の流れが定義されオペレーターの研修な

```
OPERATOR: Welcome to CarCompanyA. My name is Albert.  How may I help you?

CUSTOMER: Aah ok I need it from New York.

OPERATOR: For what date and time.

.........

OPERATOR: Wonderful so let me see ok mam so we have a 12 or 15 passenger van avilable on this location on
those dates and for that your estimated total for those three dates just 300.58$ this is with taxes with
surcharges and with free unlimited free milleage.

OPERATOR: That is fine.

.........

OPERATOR : Tehe confirmation number for your booking is 221 384 699.

CUSTOMER: OK ok thanks you.

OPERATOR: Thank you for calling CarCompanyA and you have a great day good bye.
```

図 3.1: 会話データ (例)

どで用いられている．また，各会話データに予約成立・予約不成立といったビジネス結果が付与され，予約成立に対しては pick up（顧客がカウンターに来る），not pick up（顧客がカウンターに来ない）が最終的に付与される．なお，ヘルプセンターにおける会話のように，顧客が苦情や要望など様々な問い合わせをし，会話の進め方が一様でない会話データは本研究の対象としない．

3.3 会話データを分析するための会話分析システムと分析における課題

3.3.1 会話分析システム

図 3.2 は本研究で構築した会話分析の構成を示している．書き起こされた会話に対して，文分割，形態素解析といった自然言語処理が行われる．抽出された単語に対して正規形に置き換え，意味カテゴリを付与する処理や定義したパターンに適合する複数の語からなる表現を抽出する処理を前処理という形で行う．テキストから抽出されたデータは会話データにあらかじめ付与されている定型項目（日付，オペレーターの名前，ビジネスの結果など）と共に随時蓄積する．

蓄積されたデータに対して定期的にインデックスを張り，従来のテキストマイニング分

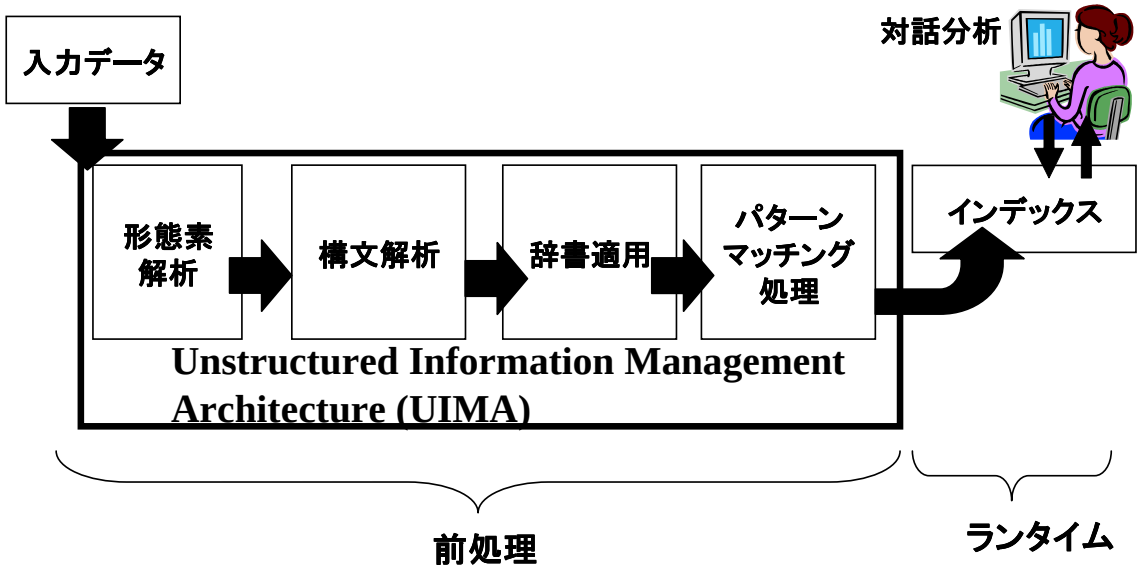

図 3.2: システムの全体図

析で行われている様々な観点からデータを対話的に俯瞰できるようなシステム [43] を会話データに対して構築した．本会話分析システムで行える対話分析の結果の例を図 3.3 に示す．これは 3.2 節で述べたレンタカー予約成立会話における 2 次元の分析結果である．この 2 次元表の各軸はオペレーターの発言の中に出てきた車名カテゴリに属するキーワードと顧客が予約した車を取りに来たかどうかという定型情報との関係を示しており，各セルはそれぞれの属性を満たす会話数を示している．この結果から例えば “Vehicle A” についての会話 17 件のうち 12 件が pick up されたことがわかり，他の車名に比べて pick up に偏っているということがわかる．コールセンターの解析者はこのような傾向分析を対話的に繰り返し行い，コールセンター業務の改善につなげようとしている．

3.3.2 会話分析における課題

従来のテキストマイニング分析の対象はコールメモであった．これはオペレーターが顧客との会話の後に書く会話内容の要約であり通常数文程度に内容が端的に記載されている．そのため，どのような観点で分析するかという点について解析者は容易に文書をサンプリングして中身を見ることで判断でき，設定した各観点 (カテゴリー) に属するキーワードや表現を対象データ中の高頻度語を元に辞書を作り分析に用いることができた．例えば，製品の苦情・質問が寄せられるコールセンターのコールメモ分析であれば，製品名や質問・苦情を分析観点とし，該当する表現を辞書として登録することで有用な分析が行

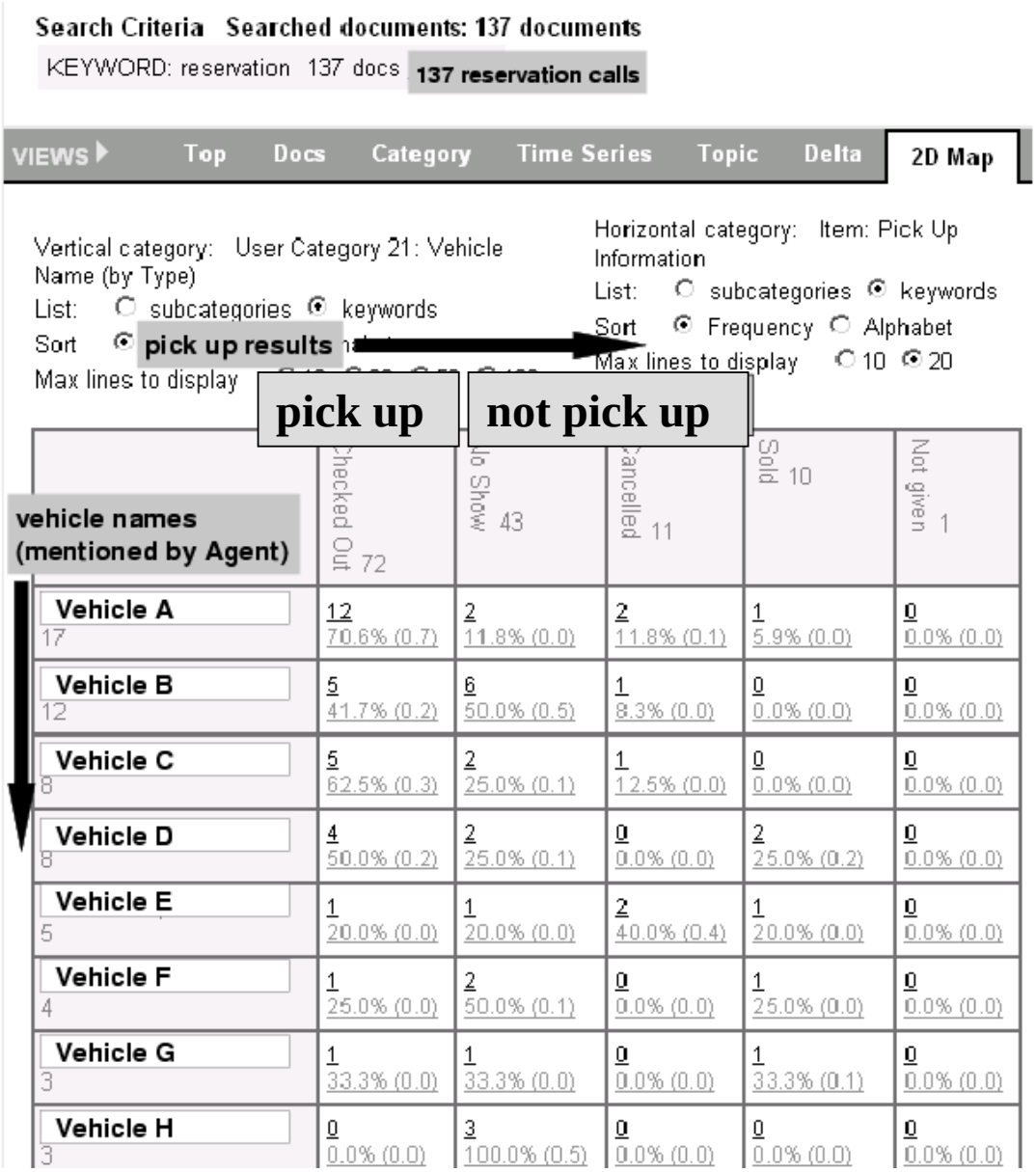

図 3.3: 分析の例 (車名と pick up 情報との相関分析)

える.

しかしながら対象データが会話全体となると，各データには挨拶などの本題から外れた冗長な部分も多く含む．従って必ずしもデータ中に出てくる語が分析に役立つ語であるわけではない．また，一つのデータが会話内容全体であるためサンプリングして注目すべき箇所や分析観点を人手で設定することは難しい．会話データの中から重要箇所を抽出する手法は [39] などで行われているが，この手法はある会話独自の主題を見つけるものであり，本研究で扱うような同じ流れに従う会話の分析に直接用いることは難しい．また従来行われている特定の文書集合における特徴語の抽出方法 [24] では，特徴語のみを抽出するため会話中のどの箇所の表現が有効であるかということは具体的にわからないため分析観点の設定が難しい．どういった観点に注目すれば有用な差異分析を行えるかどうかを解析者の勘に頼りながら試行錯誤することは効率が悪く，差異の要因を網羅的に見つけ出せるとは限らない.

そのため，分析者の経験や勘に頼らない統一的な分析手法が会話分析では必要不可欠になっている．以下では目的を持ったビジネス会話の分析手法について述べ，その有効性を実データを用いた分析で示す.

3.4 目的を持ったビジネス会話のモデリング

3.4.1 会話における参加者の意図と Expectation-Disconfirmation Model

一般に会話には複数の参加者が存在し，少なくとも一人以上の参加者は何らかの意図を持ち会話に参加している．そして参加者は自分の発言に対する他者の発言を考慮しながら自分の意図を達成するために会話を進めていくと考えられている [21].

一方，マーケティングの分野では，顧客満足は，顧客の購買体験を元にして形成される態度もしくは感情であり，製品やサービスを利用する際の状況的要因の全てが満足感の形成に関係していると考えられている．顧客の満足がどのように形成されるかについて，顧客満足に関する研究で広く採用されてきたものとして Expectation-Disconfirmation Model(期待不一致モデル) がある [44]．このモデルでは，顧客が事前に抱いていた期待 (E) の大きさと，経験された成果・製品パフォーマンス (P) とを比較し，$E > P$ であれば不満，$E \leq P$ であれば満足と決定されると仮定している．つまり，顧客が持つ期待値と実現値との差と顧客満足度の間に相関があるというモデルである.

本研究では目的を持ったビジネス会話においてもこれらの仮定が成立するとする.

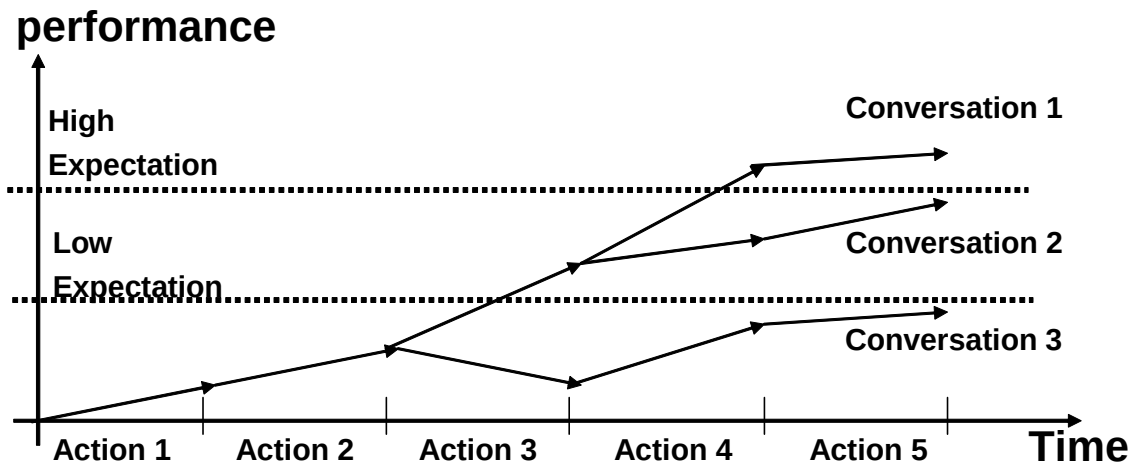

図 3.4: 目的を持ったビジネス会話のモデル

3.4.2 目的をもったビジネス会話のモデル

本研究で扱う目的を持ったビジネス会話において，オペレーター側の意図は目的の成功 (=ビジネスの成功) という共通のものであるのに対し，顧客側には様々な意図があると考えられる．顧客側はある目的のためにコールセンターに電話をかけるが，その際に何らかの期待をもっていると考えられる．例えば，商品購入という目的の電話では顧客は以下のような意図を期待として持っていると考えることができる．

意図 1 商品についてよく知っているので早急に購入したい

意図 2 商品についての情報 (値段、制限事項など) を確認してから購入したい

一方で，コールセンターにおける目的を持ったビジネス会話は話の流れおよび内容があらかじめ決まっていることが多く，それぞれの場面におけるオペレーターの発言内容 (アクション) が顧客が得る成果であると考えられる．そして，会話の流れが一通り終わった時点で得られた成果が最初に持っていた期待を超えていれば顧客は満足し，オペレーターにとって好ましい結果を出すという会話のモデルを立てることができる．このモデルを元にした会話推移の例を図 3.4 に示す．この例ではオペレーターは順番が固定された 5 つのアクションを必ず取る．それぞれのアクションにおいて顧客に伝えるべき事柄の大枠は事前に既定されているが，オペレーターの発言内容・発言の仕方は微妙に違うため，顧客に与える成果は異なったものになる．その結果，会話を終えた時点で顧客が得る成果は会話ごとに若干異なったものになる．図 3.4 で Conversation 2 は Low Expectation の顧客の場合

に顧客は満足し結果は成功となるが，High Expectation の顧客の場合は不満足となり失敗となる．

顧客側は会話を始めるにあたって様々な期待を持っていると考えられ，会話終了時点における成果が期待を上回っているかどうかで結果（ビジネス成功・不成功など）が付与される．アクションの数を n とし，会話 i における顧客の期待値 E_i，j 番目のアクションにおける成果を pfm_{ij} とすると，結果 r_i は以下のようになる．

$$r_i = \begin{cases} 1 \ : \text{success} & (\sum_{j=1}^{n} pfm_{ij} - E_i > 0) \\ 0 \ : \text{failure} & (\sum_{j=1}^{n} pfm_{ij} - E_i \leq 0) \end{cases} \tag{3.1}$$

例えば前述の商品購入にあたっては意図 1 を持った顧客は期待値が低く (Low Expectation)，高い成果を与えなくても会話終了時点では満足すると考えられる．逆に意図 2 を持った顧客は得られた情報によって購入を決めるため高い期待値 (High Expectation) を持っており，会話終了時点で得られる成果が高くないと満足しないと考えられる．

3.5 分析手法

3.5.1 分析目的と分析手順

本研究で扱う会話は話す内容および流れがあらかじめ決まっており，何らかの結果が付与されるビジネス会話である．このような会話は一見するとどれも同じ内容であり，何が結果に影響を与えているかを見つけるのは難しい．そこで「ビジネスに成功した会話集合と成功しなかった会話集合の差は何か?」といったような結果の要因を分析することは非常に重要となる．そこで，このような要因分析を分析目的とする．

3.4 節で述べたように顧客側は会話に臨むにあたって何らかの期待があり，その大きさによって結果は大きく変わると考えられる．よって，分析手順は次の 2 段階になる 1 番目については顧客が抱いている期待が明示的に会話で表されるかどうかはビジネス会話によって異なるが，まず初めに相手に会話の要件を伝える意図から顧客の期待が発言に現われる可能性があるのは，最初の発言であると考えられる．よって顧客の最初の発言が何らかの期待の大きさを含み，結果に影響を与えるかどうかを検証する．2 番目については会話が進んでいく中で，結果に影響を与える発言区間と有効な表現を見つけることになる．同定された結果に影響を与える発言内容を用いることで，例えば高い期待値を持っている顧客 (ビジネスを成功に導くのが難しい) に満足してもらうにはどんな発言が必要かといった分析が可能になる．

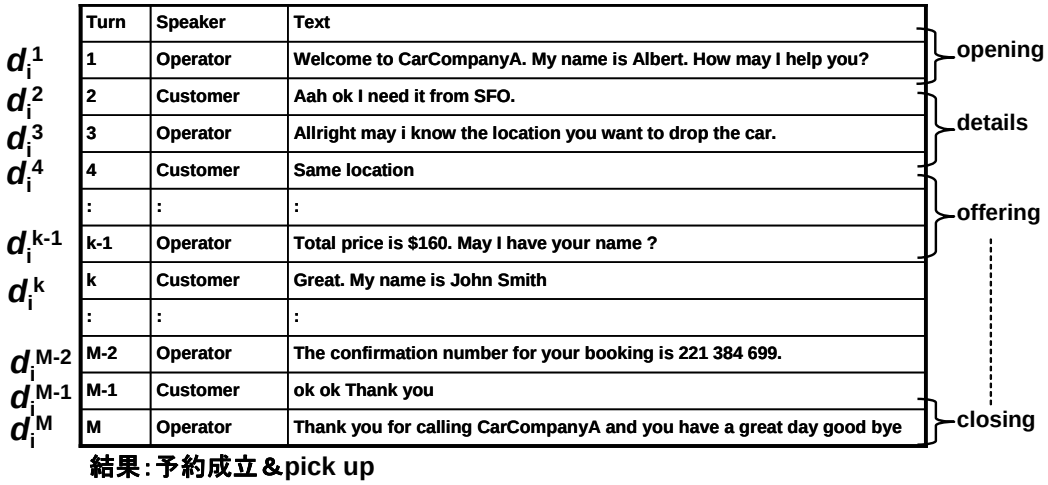

図 3.5: 1 つの会話データ $\vec{d_i}$

これらの分析手順は，ビジネス会話から重要発言箇所および有効な表現を抽出する [64] ことで実現できる．以下でその詳細を説明する．

3.5.2 データモデル

本研究で扱う会話データは，顧客とオペレーターとの交互のやり取りで構成され，3.2 節で述べた性質を持つ．会話 $\vec{d_i}$ におけるやり取りの数（turn 数）を M_i とすると各会話データは以下のように表される．

$$\vec{d_i} = \vec{d_i^1} + \vec{d_i^2} + \cdots + \vec{d_i^{M_i}} \tag{3.2}$$

図 3.5 に式 (3.2) で表される 1 つの会話データ $\vec{d_i}$ の例を示す．

ここで会話の最初から j 番目の発言までを考慮したデータを考え，$\vec{d_i^{\sim j}} = \vec{d_i^1} + \vec{d_i^2} + \cdots + \vec{d_i^j}$ で表す．そして m_k 番目の発言までを考慮した各会話データ $(\vec{d_i^{\sim m_k}})$ を集め，D_k とする．図 3.6 は $\vec{d_i}$，$\vec{d_i^{\sim m_k}}$，D_k を図示したものである．そして，m_k をいくつか設定し，時系列累積データ D_1, D_2, $\cdots$, D_k, $D_{all}(= D)$ を作成する．例えば，$m_1 = 1$，$m_2 = 2$，$m_3 = 5$，$m_4 = 10$，$m_5 = 15$ とした場合，時系列累積データは，

1. D_1: 最初の発言 (オペレーターの最初の発言) を全ての会話から集めたデータ
2. D_2: 最初から 2 番目の発言まで (オペレーターおよび顧客の最初の発言) を全ての会話から集めたデータ

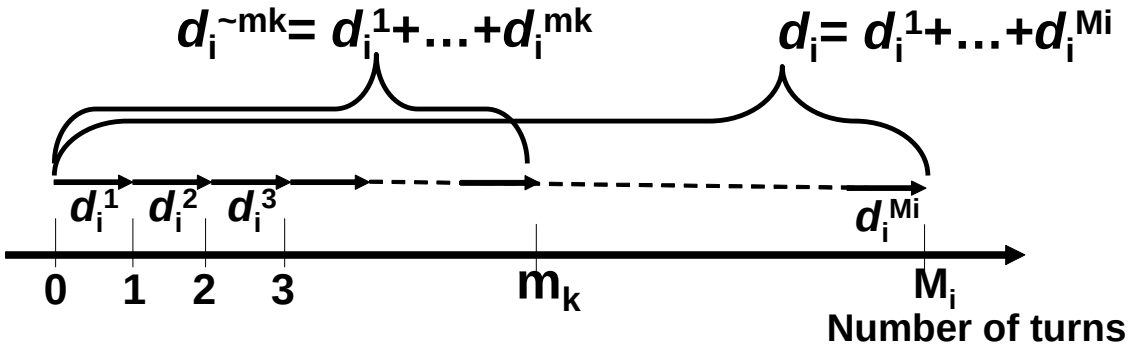

図 3.6: 会話データのモデル

3. D_3: 最初から 5 番目の発言までを全ての会話から集めたデータ

4. D_4: 最初から 10 番目の発言までを全ての会話から集めたデータ

5. D_5: 最初から 15 番目の発言までを全ての会話から集めたデータ

となる．図 3.7 に時系列累積データの例 D_1，D_2，D_3，D_4，D_5 を示す.

通常，コールセンターにおける会話はオペレーター側からの挨拶によって会話が始める．そこで，$m_1 = 1$ と $m_2 = 2$ とし，時系列累積データ D_1 および D_2 を構成することにより，D_1 はオペレーターの最初の発言を表し，D_2 は顧客の最初の発言までを集めたデータとなる．D_3 以降については会話の流れを考慮して m_i を定義する必要があるが，いくつかのサンプル例から各アクションに相当する発言数を決定して時系列累積データを作成していく．

3.5.3 特徴発言箇所同定

会話 $\vec{d_i}$ に付与される結果 r_i は 2 値 (例えば成功・失敗) であるとする．各時系列累積データ D_k に対してデータを学習データ (D_k^{train}) とテストデータ (D_k^{test}) に分ける．D_k^{train} を用いて分類器を構成しその性能を D_k^{test} を用いて求める．分類器の性能は精度 [71] を用いて求める．テストデータ D_k^{test} に含まれるデータ数を $|D_k^{test}|$，テストデータに対する分類で付与された結果 r を正しく予測できたデータ数を a とした場合，D_k における精度 $acc(categorizer(D_k))$ は,

$$acc(categorizer(D_k)) = \frac{a}{|D_k^{test}|} \tag{3.3}$$

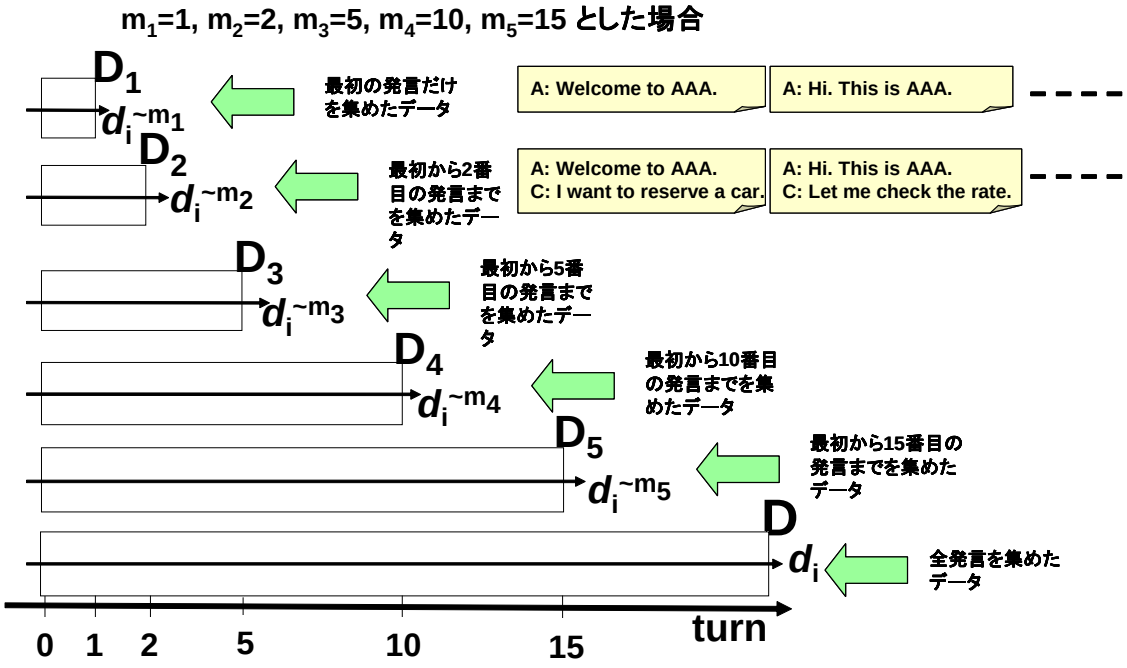

図 3.7: 時系列累積データの例

となる．

得られた分類器の精度 $acc(categorizer(D_k))$ を時系列 (turn 数 (m_i)) に沿ってプロットする．D_k は時系列に沿って累積的にデータを分割したものであるので，精度は，文書全体 D に対する分類器の精度に収束する．本研究で対象とする会話データは話しの流れや内容が事前に定義されているため，どの会話もほぼ同期して会話が進んでいると考えられる．会話中の発言の中には結果に影響を与える表現が出現する一方，結果に影響を与えない表現も多く出現する．会話の流れ全体を考えた場合，結果に影響を与える発言箇所と影響を与えない箇所がある．したがって，結果に影響を与える表現を含んでいる箇所が D_k に加わった場合には精度は増加する．一方，結果に影響を与えない表現はノイズとなるため，そのような表現を多く含む箇所が D_k に加わった場合には精度は減少する．これを模式的にあらわしたのが図 3.8 である．

ここで分類器精度が増加している区間を trigger 区間 seg(開始地点, 終了地点) として抽出とする．図 3.8 の例では $seg(m_1, m_2)$ および $seg(m_4, m_5)$ が trigger 区間として同定される．trigger 区間の同定は分類器の精度の増減を元に行うため，使用する分類器には大きく依存しないと考えられる．本研究では SVM を用いた文書分類器を用いた [26].

3.5.2 節をもとに $m_1 = 1$，$m_2 = 2$ とした．$seg(1, 2)$ が trigger 区間として抽出された場

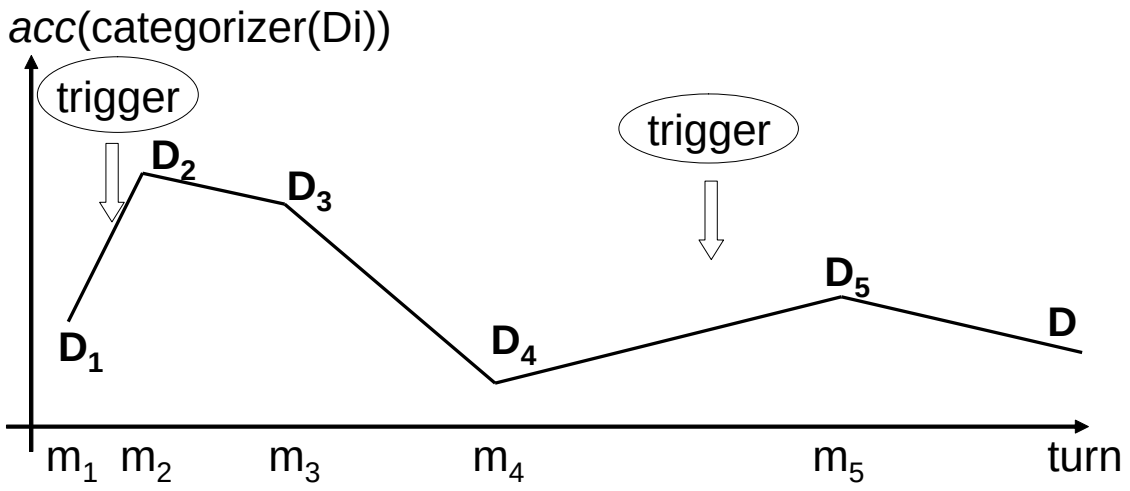

図 3.8: 特徴発言区間の同定

合，これは顧客の最初の発言内容から結果を予測できることを示しており，期待の大小が顧客によって存在し，それが結果に影響を与えるということになる．そのため，trigger 区間 $seg(1,2)$ から結果に影響を与える顧客の発言中の表現を同定し，顧客の期待に関する表現を見つける必要がある．

他の trigger 区間は，オペレーターの発言内容や発言の仕方により顧客に与える成果が大きく変わる発言箇所と考えられる．従って，結果に関連する何らかの表現が結果に影響を与えていると考えられ，そのような有用なキーワードや表現を同定する必要がある．

3.5.4 特徴表現抽出

本節では，trigger 区間における特徴表現の抽出を考える．trigger 区間 $seg(m_{k-1}, m_k)$ における特徴表現は D_{k-1} にデータが加わり D_k になってはじめて特定のラベル (A または not-A) を持つ文書に多く出現する表現と定義する．文書集合における特徴表現の抽出はさまざまな方法が提案されている [24][72]．本研究では特徴的に出現する表現を抽出する尺度として χ^2 統計量を用い，D_k 中の各キーワード・表現に対して統計量を求める．全文書数 N のデータにおいて，あるキーワード (kwd) が含まれる文書数の分布が表 3.1 であったとする．この時，χ^2 統計量は式 (3.4) で表される．

$$\chi^2 = \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{(n_{11}+n_{12})(n_{11}+n_{21})(n_{12}+n_{22})(n_{21}+n_{22})} \tag{3.4}$$

また，注目している区間に特徴的に偏って出現している表現を抽出するため以下の尺度

表 3.1: キーワード kwd の分布

文書ラベル	kwd を含む文書の数	kwd を含まない文書の数
A	n_{11}	n_{12}
not-A	n_{21}	n_{22}

を用いる．

$$new(kwd) \quad = \quad \frac{freq_{D_{k-1}}(kwd)}{freq_{D_k}(kwd)} / \frac{m_{k-1}}{m_k} \times sign(freq^{A}_{D_k} - freq^{notA}_{D_k}) \tag{3.5}$$

ここで $freq_{D_k}(kwd)$ は D_k における kwd の出現頻度，m_k は D_k の turn 数 (発言のやりとり数)， $freq^{A}_{D_k}(kwd)$ は D_k でラベル A を持つ文書集合における kwd の出現頻度，$sign(\cdot)$ は符号関数をあらわす．D_k において kwd がラベル A を持つ文書においてはじめて出現した場合，本尺度によるスコアは 1 以上になる．これら 2 つの尺度を組み合わせた

$$\chi^2(kwd) \times new(kwd) \tag{3.6}$$

を用いて特徴表現を抽出する．この尺度を用いることで，偏在性と新規性を持った表現を抽出することができる．

3.6 分析実験

3.6.1 分析目的とデータ

分析実験では，インドにおいて電話オペレーションセンターを運営する会社で収集した会話データを用いた．現在，多くの企業で管理部門などで行われいる特定のビジネスプロセスを専門企業に外部委託することが行われている．このような外部委託はビジネス・プロセス・アウトソーシング (BPO:Business Process Outsourcing) と呼ばれている．データを収集した BPO 受託企業は，様々な企業からコンタクトセンター運営の実作業を行っている．その中で，本実験では，レンタカー会社の電話予約業務で収集された会話を対象とした．会話データは会話の録音を人手で書き起こしたデータを用いた．図 3.1 のデータはその一例である．

レンタカーの電話予約で収集される会話データの概要を図 3.9 に示す．会話データは，まず予約成立会話と予約不成立会話に分けられる．予約成立会話は，提案内容に納得した

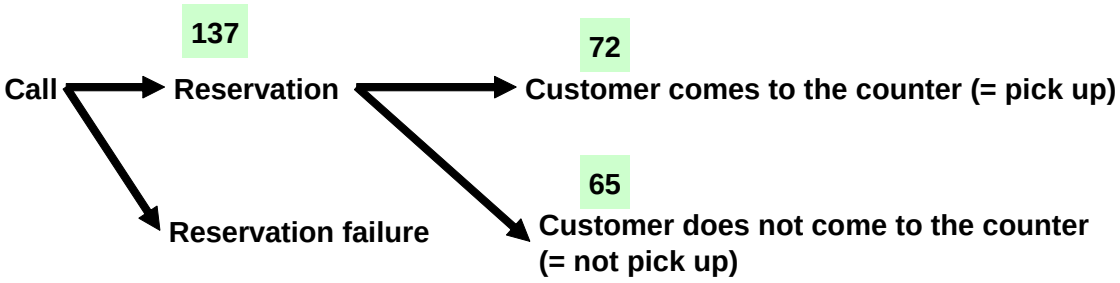

図 3.9: データの分類

顧客に対して，予約に必要な情報を顧客から収集し，予約番号を顧客に提示し，終了した会話を指す．顧客は，この予約番号をレンタカーを実際に借りる場所の受付カウンターで提示することで，車を借りることができる．これに対して，予約番号を顧客に知らせることができずに会話が終了する会話が予約不成立会話となる．予約不成立会話の中には，提案内容を聞いた段階で電話を切るものや，予約番号を発行する直前に，顧客が断るケースなどが含まれる．

本実験では会話データのうち予約が成立した 137 会話を分析の対象とした．予約成立会話は，さらに分類される．顧客が車を予約した日時に受け取りカウンターに車を取りに来た (pick up) および取りに来ない (not pick up) の 2 種類に分類される．BPO 受託企業が予約センターを請け負っているレンタカー会社では，顧客側の予約の取り消しに対して，キャンセル料は発生しない．そのため，予約したにも関わらず，顧客がカウンターに車を取りに来ないケース (not pick up) が多く発生する．実験データの予約成立会話における pick up と not pick up はそれぞれ 72 会話，65 会話となっている．

本実験を行った BPO 受託企業は，顧客企業（レンタカー会社）の予約業務を一部，つまり，特定の時間帯および特定のエリアからの予約電話のみを扱っている．そして，予約業務の請負範囲を拡大するためには，人件費が安く予約業務にかけるコストを削減できるというだけでなく，予約業務ビジネスが十分に行えることを示す必要があった．予約センターにおけるビジネス改善が必要となっている．

レンタカー予約業務の改善として，予約成立会話を増やすことがあげられる．これに対しては，オペレーターを，多く電話が来る時間に多く配置し，予約回線が埋まってしまい顧客からの電話がつながらなく事態を，避けることが行われている．また，会話内容を定型化し，オペレーターが会話をスムーズに進め，顧客との会話にかける時間を短縮し，数多くの案件を受け付ける試みが行われている．

一方で，別の改善策として，予約成立会話に含まれる pick up のケースを増やし，not

pick up のケースを減らすことが考えられる．このような改善を行うためには，まず，同じ予約成立であるにも関わらず，pick up と not pick up に分かれる要因を知る必要がある．その上で，見つかった要因に対応した改善策を実施する必要がある．

このような改善をしていく上で，予約業務を行うインドの BPO 受託企業には以下の課題があった．

1. 電話応対を行う担当者および管理者の離職率が高く，経験を積んだ担当者が少ない
2. 各電話応対者は予約成立件数を増やすために，数多くの電話応対することが求められており，個々の予約受付会話の質を上げる工夫をする時間が取れない

そこで，会話データを収集しテキストマイニングを適用し，データから客観的な傾向を抽出し，ビジネスの改善につなげる必要があった．

テキストマイニングによる分析目的として，まず，予約不成立会話において予約が成立しない原因を見つけることが考えられる．予約不成立会話の中には，顧客側が，提案内容に対して意見を述べた後，予約を断り電話を終了するケースもあるが，電話を突然切る場合が多く，会話データの比較が難しい．そこで，同じ予約成立会話にも関わらず結果が異なる理由は何かという分析 (要因分析) を行った．

同じ進め方に従っている予約成立会話から結果の違いの要因を見るけるために，提案した重要発言箇所の同定手法と特徴キーワード抽出を適用する．そして，得られた重要発言箇所とキーワードから分析観点と辞書を作成し分析モデルを構築する．得られた分析モデルを用いて pick up を改善するための知見を得るための要因分析を行う．

3.5.2 節より，m_1=1 および m_2=2 として D_1 および D_2 を作成した．D_3 以降については，会話の長さがほぼどれも同じであり，その流れは予約詳細，提案，顧客情報の取得，予約再確認と必須事項の確認という形で事前に定義されていることから，m_3=5, m_4=10, m_5=15, m_6=20 を設定し，$D_3, \ldots, D_6$ および D を作成した．各会話データは，名詞，複合名詞，形容詞+名詞といった特定の名詞句および動詞を属性としたベクトルで表されている．各時系列累積データ D_i において抽出されたキーワードおよび表現の数を表 3.2 に示す．

表 3.2: 各時系列累積データの属性数

	D_1	D_2	D_3	D_4	D_5	D_6	D
キーワード・表現の数	42	193	442	1046	1265	1473	2182

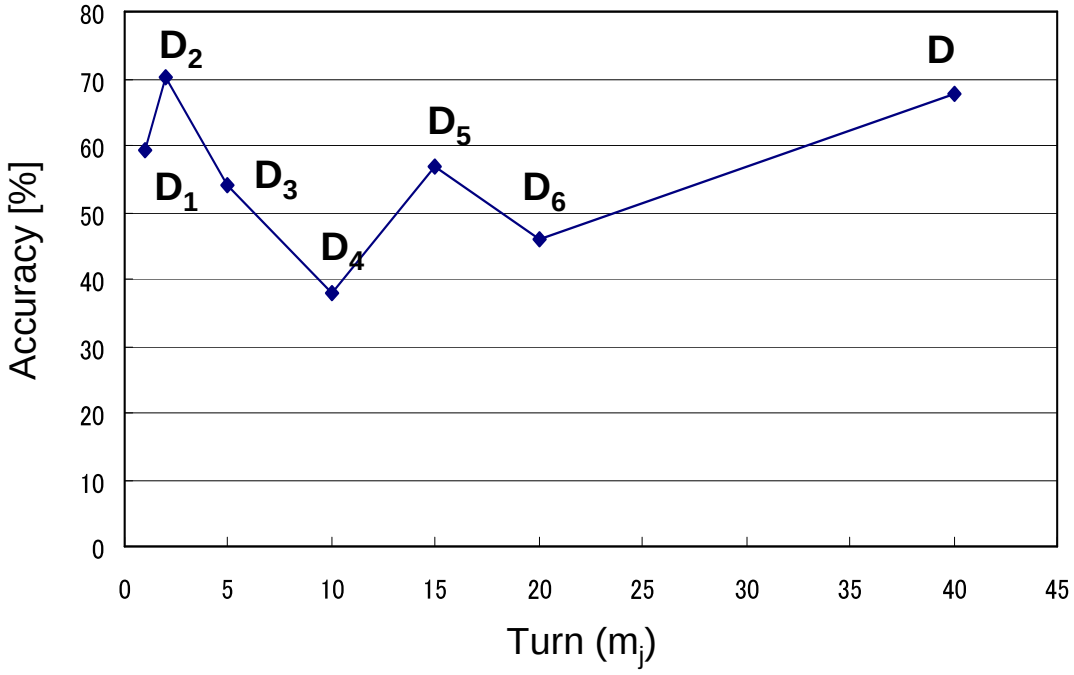

図 3.10: 各 D_k における $acc(categorizer(D_k))$

3.6.2 特徴発言箇所の同定と特徴表現の抽出を利用した分析モデルの構築

それぞれの D_k において会話データを pick up/not pick up に分類する分類器を作成し精度を求めた．図 3.10 は SVM を用いた分類器 [26] によって得られた精度の推移を示している．この結果から D_1 と D_2 の間である $seg(1,2)$ と，D_4 と D_5 の間である $seg(10,15)$ が trigger 区間として同定される．このことから，顧客は結果に影響を与える期待を事前に持っていて，それが最初の発言に出ていることが検証される．また，$seg(10,15)$ の区間の発言も結果に影響を及ぼしていることがわかる．

次に，各 trigger 区間ごとに 3.5.4 節で定義した尺度を用いて特徴表現を抽出する．表 3.3 は各 trigger 区間における高いスコアを持つ表現を示している．抽出された結果から，顧客の最初の発言 ($seg(1,2)$) および提案内容中の発言 ($seg(10,15)$) に含まれる表現と関連があることが予想される．顧客の最初の発言については，実データの該当箇所を見ることで，“would like to **make** a **reservation**” や “**check** the **rate**” といった表現が特徴表現であることがわかった．前者は予約する意思がある (low expectation) ことを示す表現と考えられ，一方，後者は値段を調べようとしていること (high expectation) を示している表現と考えられる．また，提案内容中の発言については，ディスカウント (discount) への

表 3.3: 各 trigger 区間ごとに抽出された特徴表現

Trigger	抽出された特徴表現	
	pick up	not pick up
$seg(1,2)$	make, return, tomorrow, day, airport, look, assist, reservation, tonight	rate, check, see want, week
$seg(10,15)$	number, corporate program, contract, card, have, tax surcharge, just NUMERIC dollars, discount, customer club, good rate, economy	go, impala

直接的な言及だけでなく，ディスカウントを可能にするプログラム (corporate program, customer club, contract number) に関する表現が特徴表現であることがわかる．そして，“good rate”, “just NUMERIC dollar”(NUMERIC は数字を置き換えたもの) のように，よい提案内容であることをアピールする表現も抽出されている．

比較のため，会話データ全体 D に対して従来の特徴語抽出手法を適用する．以下は χ^2 統計量を用いて抽出された特徴語のうち上位 20 語である．

corporate program, contract, counter, September, mile, rate, economy, last name, valid driving license, BRAND NAME, driving, telephone, midsize, tonight, use, credit, moment, airline, recap, afternoon

この結果からでは，ディスカウントに関連した表現 (corporate program) が結果に関連があるということがわかる程度である．従来手法に比べ，本提案手法は分析観点の設定に有効であることがわかる．

顧客の最初の発言が結果に影響を及ぼすことがわかったので，何らかの期待を持って会話を始めていることが検証された．そこで，Customer intention at start of call という分析観点を定義し，いくつかのサンプルを元に表現パターンを辞書に入れ，顧客の最初の発

言から以下の 2 種類の該当表現を抽出する.

1. strong start: *would like to make a booking*, *need to pick up a car*, ...
2. weak start: *would like to check the rates*, *want to know the rate for vans*, ...

strong start の顧客は実際に予約する意思を持っており，weak start の顧客は値段を調べているだけである可能性が高いという仮説を立て，それぞれの顧客タイプを **booking customer** および **rates customer** と定義する．なお，これらに分類できないものや意思を示していない発言もある．この場合，顧客の期待を推定することは難しいためそのような会話データからは顧客タイプを推定しなかった.

次にオペレーターのアクションについて以下を抽出する.

1. Discount-related phrases: *discount, corporate program, motor club, buying club* といったディスカウントやディスカウントを可能にするプログラムを表す表現をディスカウント表現として辞書に登録し，オペレーターがディスカウントについて言及しているかどうかという分析観点を定義する.
2. Value selling phrases: 以下のような提案している車種や値段が魅力的であることをアピールする表現をそれぞれ辞書に登録し，それぞれについて言及しているかどうかという分析観点を定義する.
   (a) good rates: *good rate*, *wonderful price*, *save money*, *just need to pay this low amount*, ...
   (b) good vehicles: *good car*, *fantastic car*, *latest model*, ...

これらの分析観点を用いた分析手順は次のようになる.

1. **booking customer** は pick up，**rates customer** は not pick up となる傾向があるという仮説は正しいか？
2. **rates customer** のうち，オペレーターがディスカウントの言及や良い提案内容だとアピールすることによって pick up になる確率はあがるのか?
3. **booking customer** が not pick up となってしまう場合はどのような時なのか？

これを図示すると図 3.11 となる.

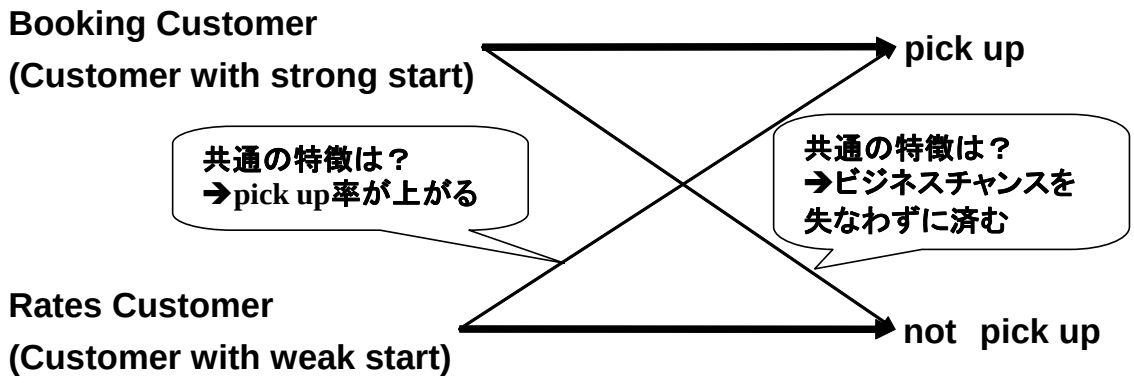

図 3.11: 分析モデル

表 3.4: 顧客タイプと pick up 情報との関係

顧客の最初の発言から抽出した顧客タイプ情報	Pick up 情報	
	pick up	not pick up
booking customer (w/ strong start) (70)	47	23
rates customer (w/ weak start) (37)	13	24

3.6.3 テキストマイニングシステムを使った分析結果

3.6.2 節で得られた分析モデルに基づき，分析観点と該当する概念（表現）を辞書として準備し，予約成立会話間の差異を分析するテキストマイニングシステムを構築し，オペレーターの生産性を改善する知見の取得を試みた．

表 3.4 は 137 の予約データにおける 2 次元の相関分析の結果であり，顧客の最初の発言内容から抽出された顧客タイプと pick up 情報 (顧客が予約した車を取りに来たかどうか) との関係を示している．この表から booking customer の 67% (47/70) が予約した車を取りに来ている (pick up) が，一方で rates customer は 35% (13/37) しか車を取りに来ない (not pick up) ということがわかる．この結果から，顧客の最初の発言内容から顧客が予約後に車を取りに来るかどうかを予測することが可能だということがわかる．

次に，rates customer の会話において，pick up 情報とディスカウントの分析観点に属しているキーワードの出現との関係を分析した．得られた分析結果を図 3.12 に示す．rates customer の場合，最終的に pick up された会話では，ディスカウントを可能にするプログラムに関連した表現が多く言及されていることがわかる．表 3.5 は，rates customer およ

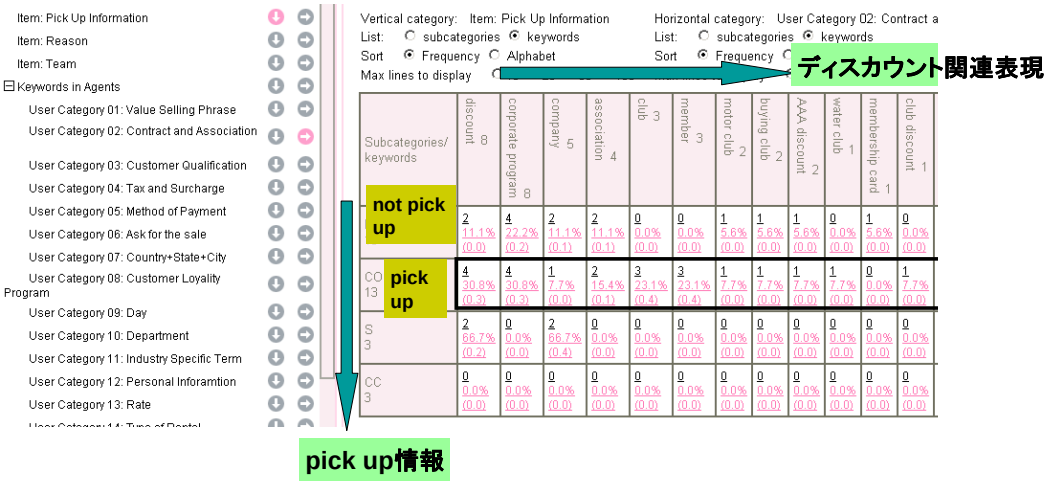

図 3.12: rates customer に対するディスカウントの言及と pick up 情報についての分析結果

び booking customer それぞれの場合におけるオペレーターによるディスカウント関連表現の言及と pick up 情報との関係を示している．この結果より，ディスカウント関連表現に言及する方が rates_customer が車を取りに来る確率が高い (P(pick up)=0.476) ことがわかる．図 3.13 に分析で得られた会話データを示す．このデータでは，顧客の最初の発言から，値段を調べに来た rates_customer であると予想される．この会話では，途中，オペレーター (Agent) がディスカウントを可能にするメンバーシップクラブの例をあげ，顧客にディスカウントを提案している．その結果，顧客に対して予約を成立させることができ，最終的に顧客はカウンターに予約した車を取りに来た (pick up) という結果が得られている．

良い提案であることを示す表現 (value selling phrase) に関しては表 3.6 で示す傾向が得られた．booking_customer に対して，よい提案内容であることをアピール (value selling) し安心させることが必要であることがわかる．

3.6.4 得られた知見とその評価

これらの分析結果から以下のような知見を得られる．

1. 最初の発言に注目することで顧客が事前に持っている期待を予測することができる．予約する顧客には 2 種類 (booking customer と rates customer) あり，それぞれ車を取りに来る傾向が異なる．

2. rates customer の場合，積極的にディスカウントを提案することで予約後に実際に

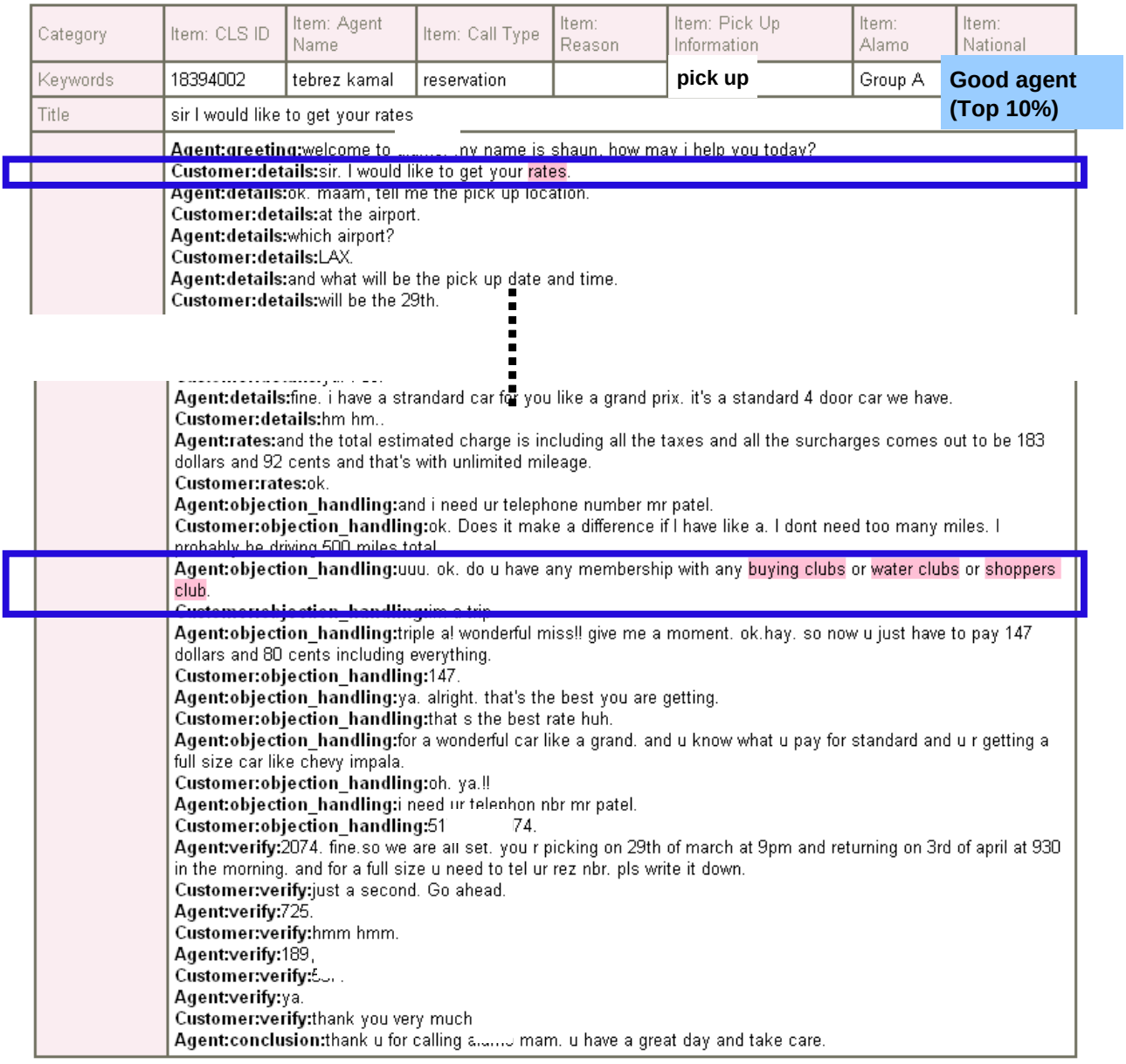

図 3.13: rates_customer に対してディスカウントに言及している会話例

表 3.5: オペレーターによるディスカウント関連表現の言及と pick up 情報との関係

Rates customer	Pick up 情報	
ディスカウント関連表現の言及	pick up	not pick up
あり (21)	10	11
なし (16)	3	13
Booking customer	Pick up 情報	
ディスカウント関連表現の言及	pick up	not pick up
あり (40)	30	10
なし (30)	17	13

車を取りに来る確率を改善することができる可能性がある.

3. booking customer の場合，予約後に確実に車を取りに来ていただけるよう提案内容が良いということをアピールし安心させる必要がある.

得られた知見を用いることで，レンタカー予約の業務を改善できるかどうかを検証した. 具体的には，得られた知見を適用することで，予約件数に対する pick up 件数の割合である pick up ratio が改善できるかどうかを評価した．レンタカーの電話予約センターに所属する 83 人のオペレーターを 2 つのグループに分け，一方のグループ A(オペレーター数=22 人) に対して，知見を元にした教育研修を行った．残りの一方のグループ B(オペレーター数=61 人) に対しては，研修などを行わず，また得られた知見を提供しなかった．研修を適用する前，これらの 2 つのグループ間で，オペレーターの成績に大きな差はなかった．1ヶ月間，これら 2 つのグループで業務を行った．もし，得られた知見を適用することで，オペレーターの成果が向上するとすれば，グループ A の pick up ratio が 1ヶ月間で上昇することが期待できる.

研修を受けた後，グループ A に属するオペレーターの pick up ratio の平均は 1ヶ月間で 4.75%上昇した．一方，グループ B に属するオペレーターの pick up ratio の平均は 1ヶ月間で 2.08%上昇した．レンタカー予約は，旅行シーズンなどの影響で，予約件数および pick up 件数は毎月上下に変動する．グループ B における pick up ratio の上昇は，この季節変動による影響だと考えられる．これらより，テキストマイニング分析によって得られた知見を適用することで，pick up ratio が約 2.67%上昇したと考えることができる.

表 3.6: オペレーターによる良い提案であることを示す表現 (value selling phrase) の言及と pick up 情報との関係

Rates customer	Pick up 情報	
良い提案であることを示す表現の言及	pick up	not pick up
あり (17)	9	8
なし (20)	8	12
Booking customer	Pick up 情報	
良い提案であることを示す表現の言及	pick up	not pick up
あり (28)	21	7
なし (42)	25	17

2 つのグループ間における pick up ratio の上昇の差が有意であるかどうか t 検定を行った結果，p 値として 0.0675 が得られた．5%有意とはならなかったが，10%有意となっており，有意差があると考えられる．以上の評価結果をもとに，対象のレンタカー予約センターではテキストマイニング分析で得られた結果を元にした研修を全オペレーターに展開し，業務改善を行った．

3.7 音声認識データを用いた実験

本節では，3.6 節で用いた会話データの一部のデータに対し，音声認識 (Automatic Speech Recognition:ASR) 技術を適用し自動書き起こしデータを取得した．そして，そのデータを用いて同様の分析結果を得られるかどうかを検証した．

3.7.1 データ

近年，認識学習用データの増加に伴い，音声認識の精度は向上し，様々なアプリケーションが利用できるようになっている．しかしながら，電話での会話を対象とした場合，異なる話者の発話を認識し，書き起す必要がある．そのような電話会話データに対する音声認識の精度は未だに十分ではない [45].

音声認識のエラーは以下の 3 種類に分けられる．

表 3.7: 認識エラーの例

エラータイプ	正解データ	音声認識結果
置換	ya on what date and time	ok on what date and time
削除	just a moment here we are open for twenty four hours	just a moment we are open twenty four hours
挿入	this is what I am saying	this is what i am saying nine booking

1. 置換：ある語が別の語として認識される
2. 削除：ある語が認識されない
3. 挿入：存在しない語が認識され挿入される

表 3.7 で，これら 3 種類のエラーについて例を示す．音声認識結果の精度を測る指標として単語誤り率 (Word Error Rate: WER) が一般的に用いられている．WER は以下のように計算される [73].

$$WER(\%) = \frac{S + D + I}{N} \times 100 \tag{3.7}$$

ここで，S は置換エラーの数，D は削除エラーの数，I は挿入エラーの数を示し，N は正解データ中の語の数を示す.

3.6 節で用いたレンタカー会社の電話予約センターで収集された会話の内，57 の予約成立会話について手作業および音声認識技術で書き起しを行い，比較可能なデータセットを作成した．分析対象となる 57 の予約成立会話の pick up 情報の内訳は，pick up が 34 件，not pick up が 23 件であった.

分析対象の会話データについて，音声認識技術および手作業で書き起されたデータの比較を行った．表 3.8 に書き起し手法で得られた対応する発話データの例を示す．手作業による書き起しデータにはミススペリングなどの誤りが含まれるが，この人手による書き起しデータを正解データとみなし，音声認識結果の精度を WER として計算することができる．評価の結果，$WER = 46.7\%$ であった．テレビのニュース番組のような読み上げデータの場合，WER が 4%から 9%である一方，電話の会話の WER は 20%から 30%であるされている [45]．近年，コールセンターは世界中の様々な場所に設置されており，英語で会話が行われているものの，実際は国際電話となっていることが多い．このため，会話は異

表 3.8: 手作業および音声認識技術による書き起こしの例

人手による書き起し	音声認識結果
i am calling to rent a car on tuesday	umm ok ok and umm i am calling to come to rent a car on that tuesday
what is the pick up date and time	what date and time would that be
i would like to pick the car up in dallas at the airport	i like to pick the car in dallas at the airport
kindly give me the confirmation number please	can you give me the confirmation number please

なるアクセントや訛りを含み，それが音声認識結果の精度が低下している要因となっていると考えられる．

3.7.2 分析結果の比較

3.6 節で得られた分析観点を用いて，人手による書き起こしデータを分析して得られた傾向分析結果と同様の結果を音声認識で得られた会話データから得られるかどうかを調査した．

表 3.9 は，顧客タイプと pick up 情報との相関について，人手による書き起しデータと音声認識による書き起しデータによる分析結果を比較したものである．両方の分析結果から，booking_customers は予約した車を pick up しやすく，rates_customers は予約した車を pick up しにくいという結果が得られている．表 3.10 は，オペレーターによるディスカウントへの言及と pick up 情報との相関，そして，表 3.11 は，オペレーターによる提案内容のよさのアピールと pick up 情報との相関について分析した結果を示している．これらの表での比較から，人手による書き起しデータと音声認識による書き起しデータの両方から同様の分析結果が得られることがわかる．

3.8 考察

分析にあたって別途，BPO 受託企業における電話予約担当者がレンタル場所，車種などの観点を設定し辞書を作成したが，これらの分析観点は本要因分析においては有効では

表 3.9: 顧客タイプと pick up 情報との関係 (手作業による書き起こしデータと音声認識による書き起こしデータの比較)

	人手による書き起こしデータ		音声認識データ	
顧客の最初の発言から抽出した顧客タイプ情報	Pick up 情報		Pick up 情報	
	pick up	not pick up	pick up	not pick up
booking customer (w/ strong start)(29)	18	11	19	10
rates customer (w/ weak start) (17)	7	10	4	6

表 3.10: オペレーターによるディスカウント関連表現の言及と pick up 情報との関係 (手作業による書き起こしデータと音声認識による書き起こしデータの比較)

Rates customer	人手による書き起こしデータ		音声認識データ	
ディスカウント関連表現の言及	Pick up 情報		Pick up 情報	
	pick up	not pick up	pick up	not pick up
あり	5	6	4	4
なし	1	5	0	2

Booking customer	人手による書き起こしデータ		音声認識データ	
ディスカウント関連表現の言及	Pick up 情報		Pick up 情報	
	pick up	not pick up	pick up	not pick up
あり	14	5	13	4
なし	4	6	3	9

表 3.11: オペレーターによる良い提案であることを示す表現 (value selling phrase) の言及と pick up 情報との関係 (手作業による書き起こしデータと音声認識による書き起こしデータの比較)

Rates customer	人手による書き起こしデータ		音声認識データ	
良い提案であることを示す表現の言及	Pick up 情報		Pick up 情報	
	pick up	not pick up	pick up	not pick up
あり	5	11	3	5
なし	0	1	1	1

Booking customer	人手による書き起こしデータ		音声認識データ	
良い提案であることを示す表現の言及	Pick up 情報		Pick up 情報	
	pick up	not pick up	pick up	not pick up
あり	15	9	11	9
なし	3	2	5	4

なかった．提案した会話モデルと分析手法によって，対象業務に深く精通していなくても分析に有効な観点の設定や表現をデータから半自動的に取得し，有効な分析を容易に行うことができるようになると考えられる．お客様が第一声でどう発言したかという，通常コールメモにも残さないような些細な情報が結果を左右する要因になっていたという知見を得ることができたことは，生の会話を分析対象とする会話分析の有用性を示している．また，お客様が第一声が会話の結果に大きな影響を与えるという分析結果は，取引などの会話にあたっては，会話の最初の部分が重要な役割を果たすという Simons の仮説 [57][59] がコールセンターにおけるビジネス会話で成立しているということを意味している．

3.7 節の結果から，音声認識で得られた書き起こしデータからも人手による書き起しデータの場合と同様の分析結果が得られることがわかった．これは，テキストマイニングによる語や表現の出現頻度を元にした分析では，会話の全てが正しく書き起される必要はなく，出現数の大小が得られればよいからである．実際に，人手で書き起された会話データに対し，人工的にノイズを付与したデータ用いた実験の結果，出現頻度を用いた分析はノイズに対して頑健であることが報告されている [1][88]．音声認識の精度は今後も向上することが期待できるが，分析に有用な情報を自動書き起こしデータといったノイズの多いテキストデータからロバストに抽出する技術が必要であると考えられる．

本章で行った会話データのテキストマイニングでは，ビジネス会話には決まった流れが

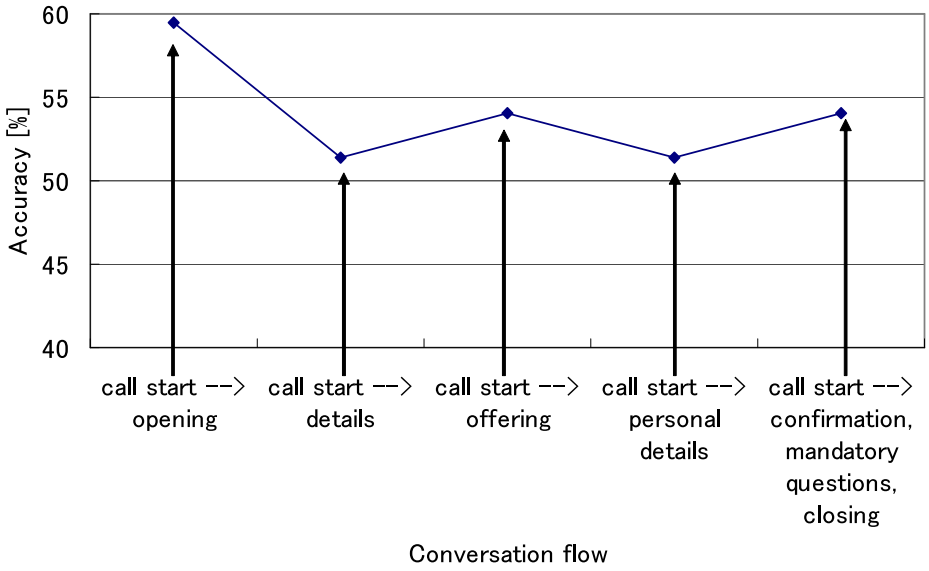

図 3.14: 場面情報を用いた時系列累積データにおける $acc(categorizer(D_k))$ の推移

あり，ほぼ同数の発言数で会話の場面転換が行われるという特徴を利用し，時系列累積データを作成した．会話における場面転換を抽出できればその情報を元に時系列累積データを作成することができる．3.6 節で用いたデータに対して，人手で「opening」,「offering」といった場面情報を付与した．場面情報を用いて，時系列累積データを作成し，トリガー検出を行った．図 3.14 は，$acc(categorizer(D_k))$ の推移を示している．ここで各会話データに付与した場面情報は 3.6 節で定義した D_k と完全に一致はしないが，以下のような対応関係がある．

1. call start $\longrightarrow$ opening : D_2
2. call start $\longrightarrow$ details : D_4
3. call start $\longrightarrow$ offering: : D_5
4. call start $\longrightarrow$ personal: details : D_6
5. call start $\longrightarrow$ confirmation, mandatory questions, closing : D

図 3.14 で得られる傾向は，図 3.10 で得られる傾向と似ている．この結果から,「opening」や「offering」といった場面がトリガー区間であることがわかる．一般に場面転換の抽出

のためには学習データが必要であるという課題がある [4] が，逆に本手法を会話データの場面転換部分の抽出に用い自動セクション分けに応用することも期待できる．

本研究では，目的を持ったビジネス会話には決まった流れがある性質を利用した分析手法を提案したが，本手法は会話データに限らない．何かしらの結果が付与され，決まった流れで内容が展開される文書データにも適用可能である．例えば，文書全体の形式がある程度決まっている報告書の分析 [63] で利用できると考えられる．会話データ以外への本手法の適用と拡張も今後の課題である．

3.9 本章のまとめ

本章では，コールセンターで得られる生の会話データを分析の対象としたテキストマイニングを検討した．生の会話データを分析対象とした場合，各データが冗長部分も含み大きいため，従来のテキストマイニング分析で広く行われいた分析観点の設定や辞書の構築を人手で行うことは困難になるという課題があった．この課題を解決するため，コールセンターなどにおけるビジネス会話はほぼ同じ流れに沿っているという特徴と顧客満足度の観点に基づき会話をモデル化し，結果に影響を与える会話データ中の特徴的な発言区間を同定し，特徴表現を抽出する方法を用いた分析モデルを提案した．実データに提案した分析手法を適用し，分析に有効と思われる分析観点（カテゴリ）および該当表現を容易に取得でき，効果的な分析ができることが検証した．会話の場面推定への活用や，本研究で扱った目的を持ったビジネス会話と同様の性質を持つ会話データ以外への適用が今後の課題となっている．

第4章　市場分析におけるテキストマイニングを活用したデータマイニングの実践

4.1　背景

近年 Business Analytics(BA) と呼ばれる活動が企業内で試みられている [30]．BA は，情報を整理するだけでなく，将来を予測しビジネスを最適化させる活動である．従来のデータマイニングやテキストマイニングの適用では，大量のデータからの情報抽出とその可視化が中心であり，様々なシステムが研究開発されている [14][20]．その一方で，抽出され可視化された情報から知見を見出すのは分析者にゆだねられていた．データマイニングの実践研究として，製品の不具合事象の抽出 [85] や顧客行動パターンの抽出 [83] などがある．また，テキストマイニングについても，企業に関するニュース記事や特許文書からの企業動向に関する傾向分析 [15] や医療文献から分子間の相互作用の抽出 [52] といった実践的な試みが行われている．しかしながら，分析結果から有効な知見を得るための手段について検討した実践的な研究はまだ少ない．

例えば，データマイニングでは，標準的な分析プロセスとして CRISP-DM[1] が定義されている．しかしながら，得られた結果の分析については，試行錯誤をしながら進める，と述べられているだけで具体的な解決策は提示されていない．また分析対象によっては利用可能な知識源が存在するが，それを有効に活用し，再利用する仕組みは存在しなかった．このような課題に対する議論は [18] で行われている．そこでは，工学だけにとどまらず，社会科学や経営の場におけるデータマイニングの適用において以下が必要になると述べられている．

- 計算機科学に精通していないユーザーが適宜データを更新し，モデルを変更しながら分析をすることができるシステム

[1] http://crisp-dm.org

- 単一の分析ツールではなく，データの収集，管理なども同時に扱う統一したシステム
- 分析を完全に代行するのではなく，分析者の作業を支援するシステム（=人間系を含んだ分析システム）

このような要求に対して，例えば，単一のデータマイニング結果から意思決定を行うのではなく，得られた結果を元に分析者が別の分析を再度行い，よりよい意思決定に結びつける分析システムが提案されている [33]．このようなアプローチにおいても，次の分析ステップに移るためには何らかの気づきを現時点の結果から得る必要がある．本研究で扱う，データマイニングを用いたルール発見では，PDCA サイクルのアクションにつながる内容で構成された分析結果（ルール）が多数得られる．結果を元に何らかの決定を行う場合，分析者は各結果に対して，対象分野において有効なものであるか検討する必要がある．しかしながら，対象分野の専門家が抽出された結果を評価した結果，活用できるとしたものはごく少数であったという経験もある．その結果，実践的なデータマイニングでは，抽出された結果のほとんどが知見に結びつかず，利用されないということが起きている．

そこで本研究では，市場分析におけるデータマイニング実践においてテキストマイニングを活用することを考える．単一のデータだけを分析するのではなく，テキストを含め様々なデータを分析するためのシステムが考えられている [55] が，具体的な活用方法などについては，ほとんど研究がされていない．テキストマイニングでは，注目している文書集合で，特定のキーワードや表現の出現頻度が他の文書集合に比べ多い・少ないといった傾向を分析することができる．また，コールセンターにおける分析では，顧客が書いたテキストを参照することができるため，分析結果の解釈が容易である可能性が高い．そこで，本研究では，テキストマイニングで得られる傾向分析の結果を元にデータマイニングの結果から意味解釈が容易なものをフィルタリングすることを考える．そのためには，テキストマイニングにおいて，分析結果を効果的に導く手法が必要になる．本研究ではテキストマイニングによる傾向分析における分析観点の選択手法について提案する．そして，生ごみ処理機という未普及の製品について，実際の市場分析のために取得した実データに適用し，その有効性を検討する．また，決定木分析をはじめとした，様々な条件の組み合わせで構成されるルールを多数抽出するデータマイニング分析において，テキストマイニングによる傾向分析結果を元に結果をフィルタリングする．これにより，対象分野に精通していない分析者でもある一定の分析結果を得られることを実践例を元に検討する．

本章の構成は以下の通りである．4.2 節において，市場分析におけるデータマイニングとテキストマイニングについて述べる．そして，4.3 節において，データマイニング実践におけるテキストマイニングの活用方法を述べ，課題となる分析観点の選択について説明

する．4.4 節では，分析観点の選択手法を提案する．そして 4.5 節で市場分析を目的として実際に収集したデータを用いて，提案手法を用いてテキストマイニング分析が効果的に行えることを示す．また，テキストマイニング結果を用いたデータマイニングの実践例を示す．最後に，4.6 節で実践結果を考察し，4.7 節で本章のまとめを行う．

4.2 市場分析におけるデータマイニングとテキストマイニング

市場分析では，データからビジネスを拡大するための知見を得ることが行われている．例えば普及初期の商品がある場合，効果的な販売促進をするための施策が求められる．近年，Web を使った意識調査など顧客の意見を取得することが容易になってきており，市場分析においても顧客アンケートを分析する試みが多くなされている．

具体的には，マーケティングの分野では，サービスや商品の品質を測定するための顧客側感覚尺度として以下の 5 つの観点 (5D) が SERVQUAL として提案されている [47].

- Tangible(有形性): 商品・物的施設・設備の内容，接客員の外見等
- Reliability(信頼性): 商品提供やサービスの遂行に関して信頼し，期待する結果が実際に提供されること
- Assurance(確実性): 商品提供やサービスの遂行に関して十分な技能を持つ，顧客への丁重さを持つ，リスクを抱かないこと
- Responsiveness(反応性): 商品・サービスを素早く提供する意欲を感じること
- Empathy(共感性): 十分な情報提供，積極的な顧客理解

この 5 次元の観点を分析対象の製品やサービスに対して拡張し，質問項目を作成し，対象製品・サービスに対して顧客が感じている品質感覚を整理する試みが行われている．

また，製品やサービスの構成要素についてマーケティング理論では，マーケティングミックスという概念が定義されている．マーケティングミックスとは，企業が対象とする市場での目的を達成するために用いるマーケティング要素の組み合わせである [31]．従来，製品提供のためのマーケティング要素として，

- Product: 製品，サービス自身
- Price: 価格

- Promotion: 販売促進活動，広告
- Place: 提供場所，立地，流通範囲

の 4 つが定義されていた (4P)．これを元に，サービスの場合，

- Physical Evidence: デザイン，機能性
- People: 販売員，サービスを提供する全ての要員
- Process: サービスを提供する手段

の 3 つを加えた 7 つのマーケティング要素が定義されている (7P)．

顧客の人口動態属性 (性別，年代，収入など) と共に上記の 5D を元にした顧客の感覚や 7P(4P) を元にした製品・サービスの構成要素に対する意識をアンケートをとって分析することが行われている．そして得られたアンケートから，例えば対象製品に対する市場のニーズを把握することが試みられている．

このようなアンケートを用いた市場分析において，データマイニングを用いることで，例えば，対象製品を購入している顧客を特徴付ける属性の組み合わせをルールとして抽出することができる．分析で得られる抽出されたルールの中から従来の市場分析ではわからなかった新規顧客層を見つけられる可能性がある．本研究では，新規顧客層を見つけるためにデータマイニングを適用し，ある望ましい顧客層 (例えば製品の購買者) の特徴を求めることを考える．具体的には，製品・サービスについて市場分析で得た購買者および非購買者のアンケート内の定型質問の回答から，新規顧客層を特徴付けるルールを抽出する．本研究では，数多くのデータマイニング手法に対する Java API を提供している WEKA[70] を用いて分析システムを構築した．分析システムでは決定木学習アルゴリズム C4.5 の実装である J48 を用い，購買者・非購買者を決定するルールを抽出する．ここで，データマイニングで対象とするデータは，顧客属性を含め選択式質問に対する回答である．そのため，各質問に対する回答には 1，2 といった名義尺度を対応付けることができる．分析システムでは，日本語で書かれた質問・回答と名義尺度の間の対応表を別途作ることで，言語によらない名義尺度データだけでデータマイニングを行うことができる．決定木学習において，これらの名義尺度の分割に 2 分分割を用いた．

得られる決定木からのルール抽出について以下に述べる．決定木の例を図 4.1 に示す．図 4.1 で，$R1$〜$R7$ で示したノードが条件，末端のリーフが判定例を示し，(a, c, e, g) が正例 (購買者)，(b, d, f, h) が負例（非購買者）と判定されたものする．ノード間およびノードとリーフの間のエッジが条件が取る値 (例えば $R1$ では r_{11} と r_{12}) となり，条件と条件

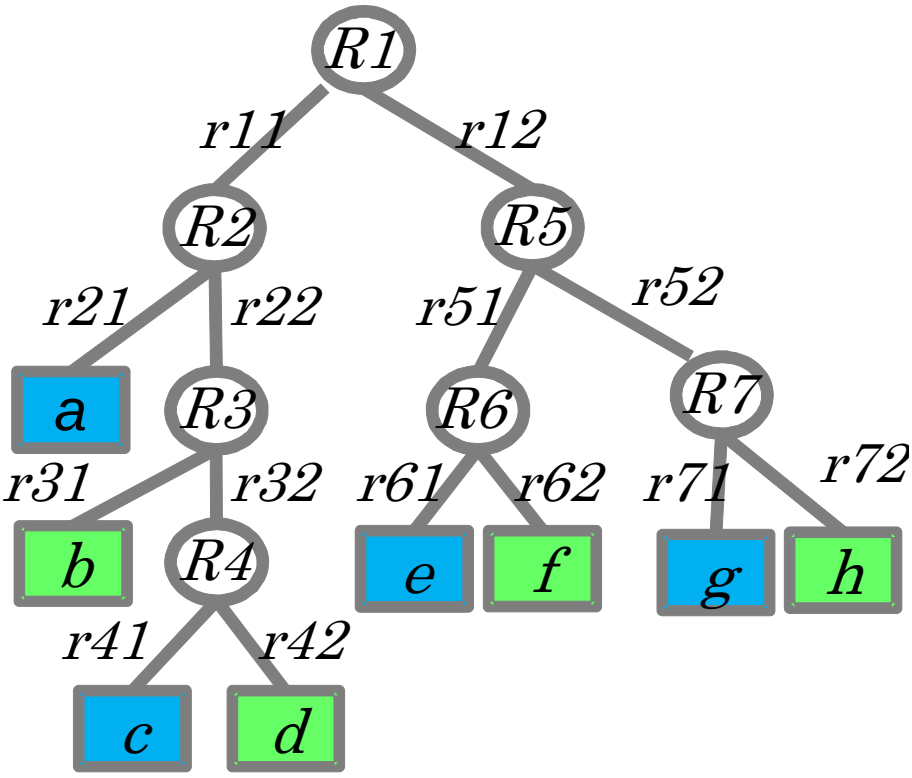

図 4.1: 決定木の例

が取る値の組がルールの構成要素となる．目的が，新規購買層に関する知見の発見であるため，購買者と非購買者の違いを決定付ける特徴を抽出することになる．そこで，子がリーフのみである末端ノードに注目し，ルートからそのノードまでをルールとして抽出する．図 4.1 の例からは $\{R4|R1 = r_{11}, R2 = r_{22}, R3 = r_{32}\}$，$\{R6|R1 = r_{12}, R5 = r_{51}\}$，$\{R7|R1 = r_{12}, R5 = r_{52}\}$ がルールとして抽出される．各抽出結果において，ルートから末端ノードまでの条件で表される顧客層を施策の適用先と考える．そして，顧客層を示す条件と末端ノードの条件から非購買者を購買者に変えるような施策につながる知見を得る．

一方，市場分析では，顧客の製品・サービスに対する顧客が自由に書いたコメントデータに対して，テキストマイニングを適用することができる [78]．製品・サービスに対するコメントデータとしては口コミサイトのデータやアンケートなどを用いて分析者が能動的に取得したテキストデータが考えられる．

市場分析におけるテキストマイニング分析では，分析観点として前述した 5D，7P および人口動態属性を利用できる．観点（カテゴリ）ごとに該当する概念 (キーワード) を辞書

に登録することで，それぞれの観点でテキスト中に出現する表現をまとめあげることできる．例えば,「Promotion に関するコメントが何件あるか？」といった集計が可能になる．各表現の出現頻度が少数である場合，傾向を見つけることは困難であるが，観点レベルで集約することで，出現頻度が多い・少ないといった傾向が見つかる可能性が出てくる．

また，複数の観点を組み合わせることで，観点に属する概念間の相関関係を調べることができる．通常，製品・サービスの提供側の視点であるマーケティング要素 (7P) の組み合わせに対して，顧客側が期待している品質 (5D) が形成され，顧客属性の組み合わせによっても顧客が抱く期待品質が変わると考えられている [74]．従って，顧客がサービスや製品について経験や知識を持っている場合，

- 7P$\Longleftrightarrow$5D
- 5D$\Longleftrightarrow$ 顧客属性

の 2 種類の関連を分析することで，顧客と製品・サービスの間の関係を品質という観点で可視化することができる．その結果，特定の顧客層が製品・サービスのある要素に対して，どのような品質 (5D) を持っているという傾向を得ることができ，製品・サービスの改善やマーケティングにつながる知見を導出することが期待できる．一方，新規の製品やサービスの場合，顧客は対象製品やサービスに対して十分な経験や知識を持っていない．そのため，顧客が期待する品質は憶測を含み，信頼性の高い情報を得ることが難しい．この場合はマーケティング要素 (7P) と顧客属性の間の関連より，市場分析につながる知見を得ることになる．本研究では，未普及の製品の市場分析を実践例とするため，マーケティング要素 (7P) と顧客属性の間の関連から知見を得ることを考える．

4.3　データマイニング実践におけるテキストマイニングの活用と課題

データマイニングによるルール発見の実践では多数のルールが抽出される．例えば 4.2 節で述べた決定木を用いた分析では，末端ノード数のルールが抽出される．抽出結果から，施策につなげる知見を得るためには専門家が抽出結果を評価する必要がある．

各ルールに含まれる条件は顧客属性やサービス・製品の構成要素（マーケティングミックス）に対する顧客意識調査に対する回答である．しかしながら，条件の組み合わせが，必ずしも意味ある顧客層を表すわけではない．また，顧客層を特徴づける理由が専門分野に精通した分析者でもわからない場合も多い．分析結果を元に何らかの販売施策を立てる

ためには分析結果を裏付ける根拠が必要である．専門分野の知識に基づいた仮説が立てられない分析結果を施策に採用することはできない．その結果，抽出されたルールを専門家が多大な時間をかけて精査した結果，知見として採用されるルールは少数であることが多い．そのため，抽出されたルールを精査するコストの削減や，ルールの解釈を支援する仕組みが必要である．

このような課題を解決する手段として，顧客の自由なコメントにテキストマイニングを適用して得られる，特定の顧客層に関する大まかな傾向を活用することを考える．4.2節で述べたようにテキストマイニングの結果として例えば2つの分析観点間の関係 (例えば，世代と Promotion に関連するキーワードの共起が高い) が得られる．観点間の相関を具体的な知見に結び付けられるかどうかは，該当するキーワードを含む文書を読む必要がある．前処理で抽出された表現ごとに，出現する文書 ID と出現箇所をインデックスファイルとして保持しているため，閲覧する際に，該当するキーワードの出現箇所をハイライトすることができる．ハイライトした箇所の前後を中心に読むことで具体的な知見が得られそうか否かの判断は容易になると考えられる．一方，データマイニングで得られる各ルールを構成する条件はアンケートの質問であり，顧客属性やマーケティングミックスに対応する．これらは，テキストマイニングで用いた分析観点（カテゴリ）に対応させているため，テキストマイニングで得られた分析結果とデータマイニングで得られたルールを関連付けることができる．

そこで，本研究では，テキストマイニングによる分析で得られた観点間の関連を用いて抽出されたルールをフィルタリングする．まず，各ルールを構成している条件に対して顧客属性，またテキストマイニングで作成した辞書を用いてマーケティングミックス (7P) を関連付ける．例えば「世帯年収=1000 万円以上」という条件に対して顧客属性の「世帯年収」が，そして「宣伝を見て購入した=yes」という条件に対してマーケティングミックスの「Promotion」が関連付けられる．その後，テキストマイニングで関連付けられた2観点ごとに，ルールにそれら2観点が含まれるかどうかを判定する．2観点が含まれる場合，ルールは特定のテキストマイニングの結果に関連するルールと判断する．もし，いずれのテキストマイニングの結果とも関係しない場合，そのルールは本フィルタリングで除かれる．

図 4.2 にフィルタリングの例を示す．本例では，ルール B はテキストマイニングで関連があるとされた観点を条件に含まない．そのためルール B は知見を見出すのが難しいと判断され，フィルタリング結果では除外される．

このように，テキストマイニングの結果と組み合わせることで，データマイニングで得られる多数の結果を，解釈可能性を元に振り分けることができる．また，テキストマイニ

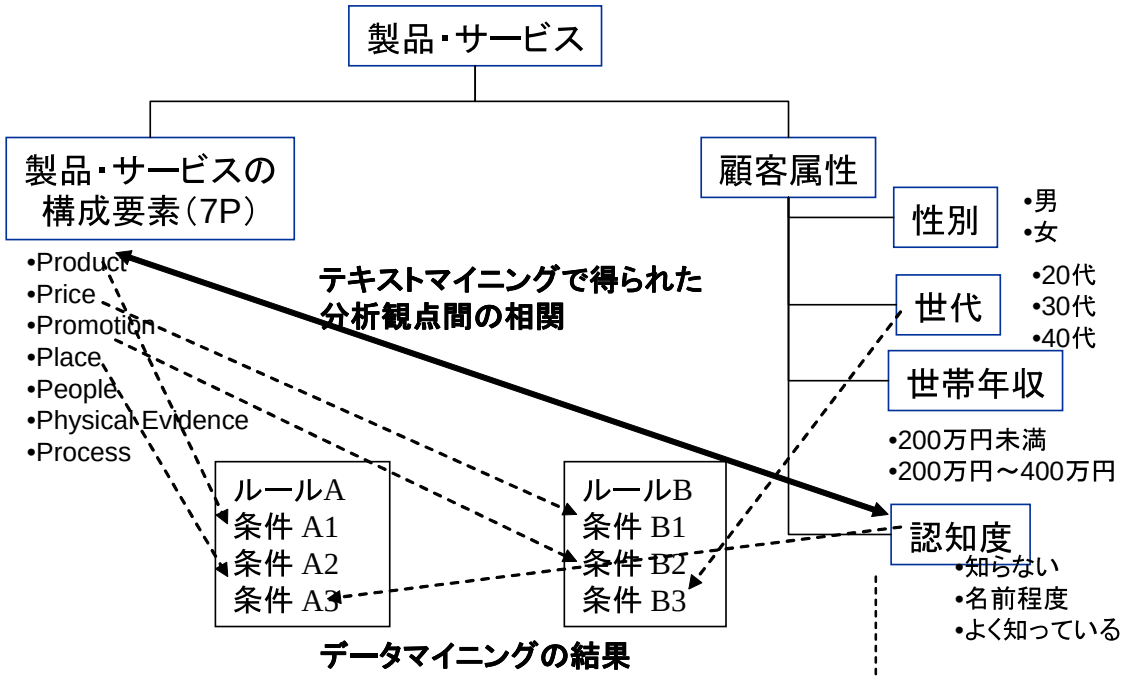

図 4.2: フィルタリングの例

ングでは，ある特定の顧客属性を持つ文書集合での，単語や表現の出現傾向しか分からなかったが，データマイニングの結果と関連付けることにより，アクションにつながるルールを得る可能性がでてくる．しかし，今度は，テキストマイニングを用いた分析において，有用な分析観点の組み合わせを効果的に見つけることが課題となる．顧客属性とマーケティング要素との間の組み合わせの中で，どの観点を選択すれば知見に結びつく分析結果が得られるかどうかは，未知である．そのため，分析者は経験を元に様々な観点の組み合わせを選択し，何らかの関連があるかどうかを逐一調べる必要があり，多大な時間とコストを要する．

この課題に対して，従来，特定の文書集合で頻出する表現を自動的に抽出する方法があり [24]，注目している文書集合の特徴を分析するために用いられている．また，分析だけでなく，文書分類においても特定の文書集合における出現に関する特徴量を利用し，文書分類に有効な素性を選択することが行われている [72]．文書クラスタリングにおいても各クラスタのラベル付けの指標として用いられている [2]．しかしながら，こういった手法で上位にランキングされる表現は，注目している文書集合で非常に多く出ると判定されるものである．そのため，市場分析では，ある顧客層の文書集合で特定の語が頻出する理由を専門家が検討する必要がある上，必ずしも有用な知見につながらないことが多かった．

観点の組み合わせが多数考えられる中，知見につながる傾向が得られる可能性が高い順に観点の組み合わせをランキングすることができれば，効果的に分析を進めることが期待できる．そこで 4.4 節において，市場分析におけるテキストマイニングにおける有効な分析観点の選択を行う指標について述べる．

4.4 テキストマイニングにおける分析観点の選択

4.4.1 市場分析に有効な分析観点の性質

市場分析では，特定の顧客層（セグメント）だけが持つ特徴や，顧客属性の違いに合わせて変化する特徴から施策を立案することができる．例えば，製品の販売拡大を考えた場合，ターゲットとなる顧客層が顕著に持つ特徴が分かれば，その顧客層に最適な販売促進計画を立てることができる．また，年齢や居住地域などの顧客属性の違いに合わせて変化する特徴を捉えることができれば，それに影響を与える施策を顧客属性に応じて変化させ効果的に結果を生み出す可能性がある．本研究では，顧客属性の違いに合わせて変化する特徴をテキストマイニング分析によって同定する手法を考える．具体的には，キーワードの出現頻度を特徴とし，多数あるキーワードの集合から，顧客属性の違いに合わせて，出現頻度が変化するものをランキングし，分析に用いる手法を考える．顧客属性のほとんどは，年齢，世帯年収をはじめとして何らかの順序関係を持つものが多い．この性質を用いて，有効となる分析観点を効果的に見つける手法を以下で述べる．

4.4.2 順序関係を持つ属性を考慮した分析観点のランキング

あるキーワード・表現 k_j の対象テキストデータにおける言及の度合い（言及頻度）を $f_c(k_j)$ とする．データ数 (文書数) が N，k_j を含む文書数が m の時，$f_c(k_j) = m/N$ となる．1 から n まで連続の値を取りセグメント数が n であるセグメント軸 (seg) での分析を考える．セグメント軸の値が seg_i であるデータ内での k_j の言及頻度を $f_c(k_j|seg_i)$ とする．セグメント軸の値が順序関係

$$seg_1 \prec seg_2 \prec seg_3 \prec \ldots \prec seg_n$$

を持っている場合に，

$$f_c(k_j|seg_1) > f_c(k_j|seg_2) > \ldots > f_c(k_j|seg_n)$$

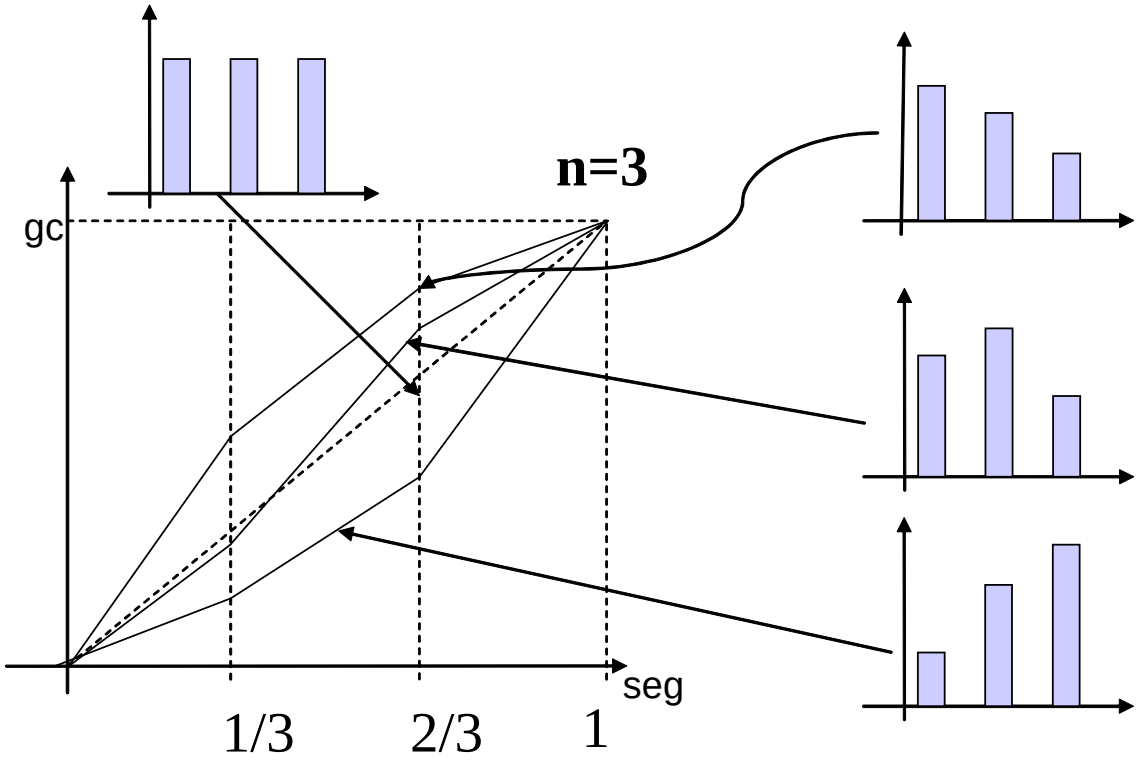

図 4.3: 正規化累積頻度

または

$$f_c(k_j|seg_1) < f_c(k_j|seg_2) < \ldots < f_c(k_j|seg_n)$$

なる傾向を持つ k_j を上位にランキングする指標を考える．このとき，$f_c(k_j|seg_i)$ から

$$g_c(k_j|seg_i) = \sum_{h=1}^{i} \frac{f_c(k_j|seg_h)}{n f_c(k_j)} \tag{4.1}$$

を求め (正規化累積頻度)，i/n ごとにプロットすると，図 4.3 で示すような結果が得られる．

$f_c(k_j|seg_i)$ がセグメント値によらない場合，$g_c(k_j|seg_i)$ は傾き 1 の直線となる．この直線に対して，$f_c(k_j|seg_i)$ がセグメント値の増加に対して単調に減少する傾向を持つ場合は上に凸，単調に増加する傾向を持つ場合は下に凸の折れ線となる．この折れ線と傾き 1 の直線で囲まれる面積を求め上に凸の場合は正，下に凸の場合は負の符号をつけ，選択したセグメント軸の特徴量 (S_{seg}) とする．$f_c(k_j|seg_i)$ が，増加と減少の傾向を両方持つ場合は，傾き 1 の直線に対して下に凸の領域と上に凸の領域ができる場合もある．その場合は，それぞれの領域の面積に符号を付け，合計したものが S_{seg} となる．ある k_j について S_{seg} は

表 4.1: 特徴量の比較

	SegA	SegB	SegC
χ^2統計量	5.7	4.5	4.2
S_{seg}	−0.00036	−0.085	−0.11

式 (4.2) のように表される．

$$S_{seg} = \sum_{i=1}^{n} \frac{1}{n} \{ (g_c(k_j|seg_i) + g_c(k_j|seg_{i-1}) - \frac{2i-1}{n} \} \tag{4.2}$$

これにより，$f_c(k_j|seg_i)$ がセグメント軸に対して単調ではないが，全体的には増加または減少している傾向を持つものについてもスコアが与えられる．S_{seg} の符号および大きさに注目することで，セグメント値の増加に対して注目している概念の言及頻度が，増加または増加傾向を持つ順にランキングすることができる．

分析する際には，製品・サービスの構成要素 (7P) に関する観点を縦軸，順序属性を持つ顧客属性の観点を横軸にとり，各観点の組み合わせに対する S_{seg} の表を作成する．この表を見ることで，どの観点の組み合わせが特徴的かどうかを見ることが概観できる．具体的にはある 7P の観点と顧客属性の組み合わせで S_{seg} が正の高い値を示していれば，強い単調減少を示していることを予測することができる．

S_{seg} を用いた分析観点の選択例を以下で示す．各文書に SegA, SegB, SegC と呼ばれるセグメント情報が定型項目として付与された文書集合を考える．そして各セグメント情報は 5 つの値をとり，それらの間には順序関係があるとする（例えば年代など）．ここで，5 つに分かれた文書集合でのキーワードの出現確率が図 4.4 である例を考える．キーワードの出現確率がある文書集合において特徴的であるかを測る指標として，セグメント値によらない平均確率からの乖離を元にした χ^2 統計量を用いることが行われている [24]．χ^2 統計量と S_{seg} との比較を表 4.1 に示す．S を比べることで，セグメント軸に沿って増加する傾向が強い順にランキングすることができることがわかる．あるキーワードの出現確率について，セグメント値が増加する方向に出現数が増加・減少するといった傾向は，市場分析において,「顧客属性○○が大きい（高い）ほど，マーケティング要素△△に言及している」といった傾向に相当する．このような傾向から施策につながる知見を得ることは容易であると考えられる．表 4.1 の結果から S_{seg} を用いて，あるセグメント軸に対して多数ある観点をランキングすることで，このような傾向を効果的に得られると考えられる．

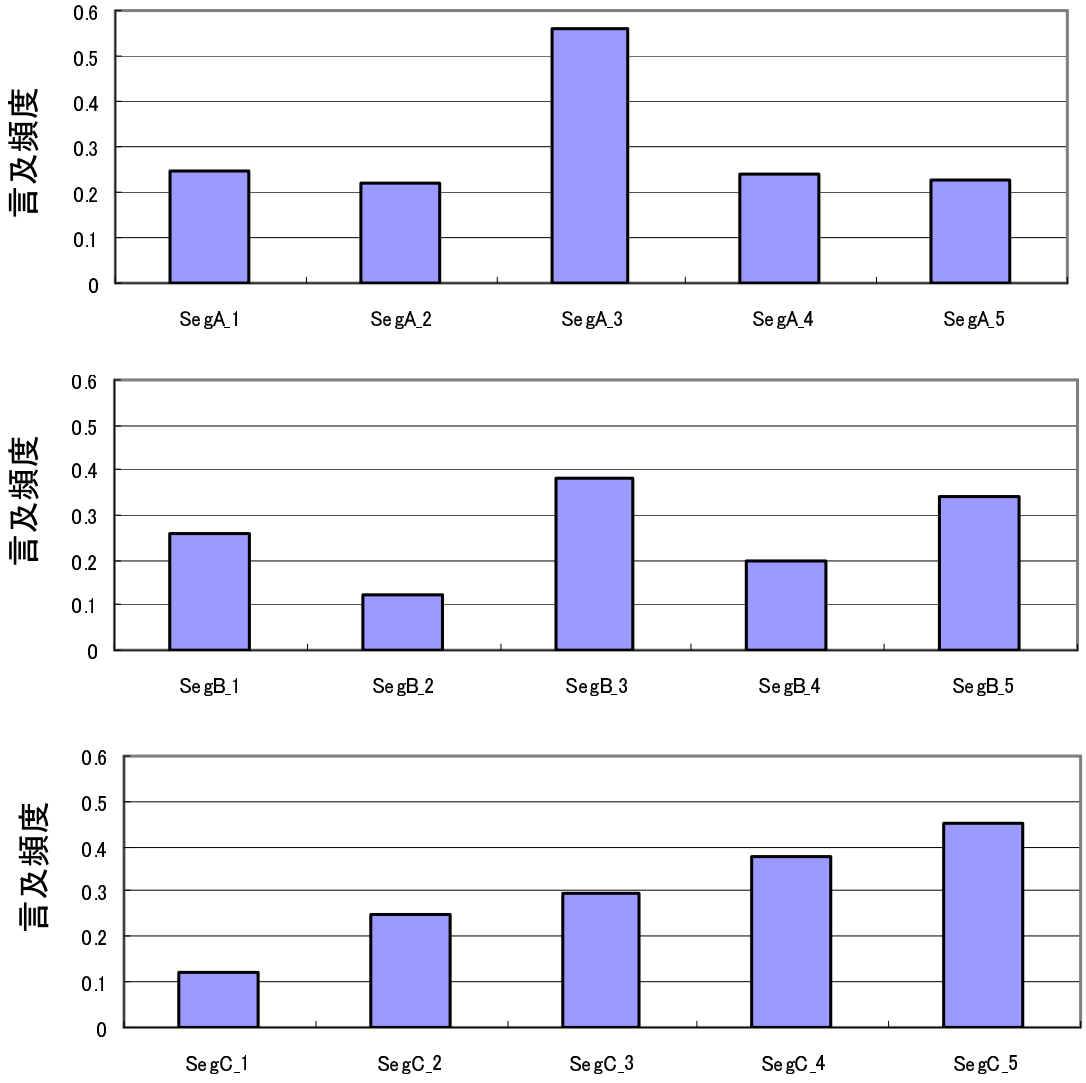

図 4.4: 擬似データによる比較

表 4.2: 相関係数との比較に用いた出現頻度データ

	Seg1	Seg2	Seg3	Seg4	Seg5
キーワード A	0.080	0.16	0.20	0.25	0.30
キーワード B	0.53	0.60	0.63	0.70	0.75

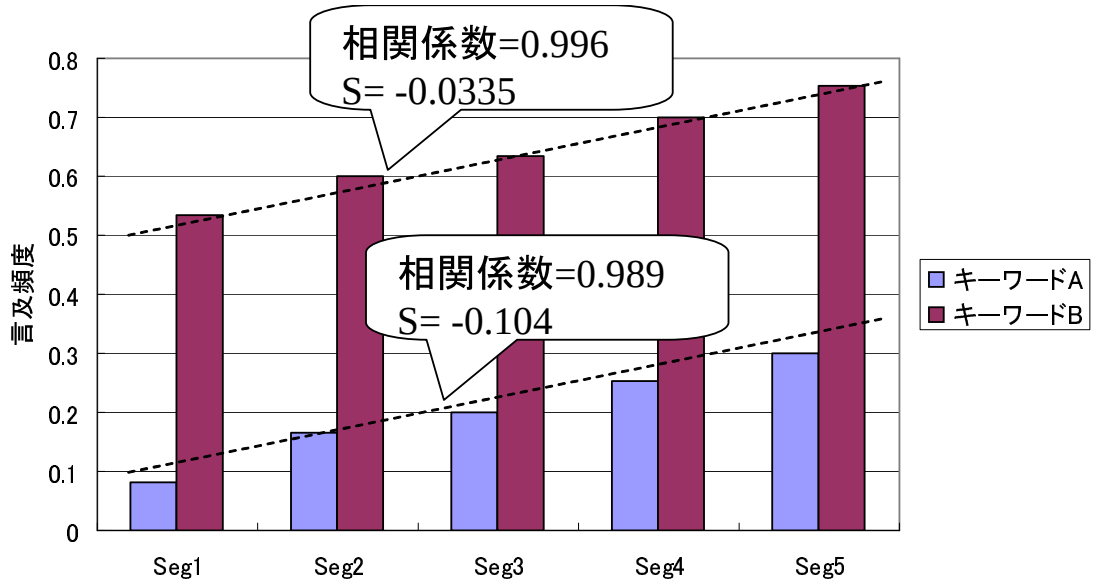

図 4.5: 相関係数と S_{seg} との比較

セグメントの方向に対して，増加または減少傾向を検出する指標として，相関係数の利用が考えられる．表 4.2 に示すデータを用いて相関係数との違いについて述べる．相関係数を求めたところ，キーワード A，キーワード B それぞれ 0.989 および 0.996 とほぼ同じ結果が得られた．一方，S_{seg} はキーワード A，キーワード B それぞれ-0.104 および-0.335 となり，大きな差が得られた．図 4.5 に示した通り，キーワード A，キーワード B の言及頻度は両方共ほぼ同じ傾きを持った増加傾向を示している．この結果から提案した指標 S_{seg} は，言及頻度が増加または減少の傾向を持ち，さらに増加または減少の量が言及頻度に対して大きな割合を占めるものを上位にランキングする特徴を持つことがわかる．

セグメント軸に対して，各キーワードの S_{seg} を計算することで，セグメントの方向に対して増加または減少傾向を示している順にキーワードをランキングすることができる．したがって，ある特定のセグメント軸に対して，どのキーワードや観点で分析をすればよ

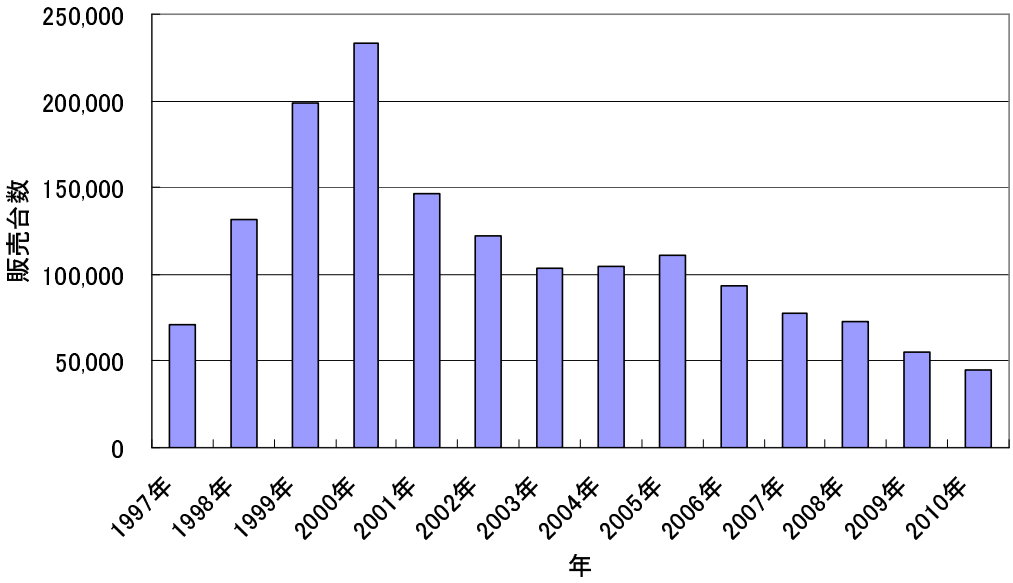

図 4.6: 生ごみ処理機の販売台数の推移 [2]

いか，分析者は優先付けすることができる．その結果，分析者の経験や勘により分析観点を選択する必要はなくなり，分析を効果的に進めることが期待できる．

次節において，実際のデータを用いた分析実践例を示す．

4.5　市場分析の実践例

4.5.1　分析目的とデータ

市場分析の例として，未普及製品を販売促進するための施策につながる知見を顧客アンケートデータから得ることを考える．分析対象の製品は生ごみ処理機とした．生ごみ処理機は，電気と特殊な触媒で，生ごみを分解し肥料にする，機械である．生ごみ処理機は 1990 年代に開発され，1997 年から販売されている．生ごみ処理機の販売台数の推移を図 4.6 に示す．2000 年をピークに減少し続けている理由として，それまでは，４社ほどがが，家庭用生ごみ処理機市場に参入していたが，事業の選択と集中の結果，現在製品を供給しているのは、2 社のみとなっているからである．生ごみ処理機の 2010 年までの累計台数は約 150 万台であり，世帯普及率は約 3%となっている．本製品と同時期に販売が開始されている家電製品に，食器洗い乾燥機，空気清浄機がある．これらの製品との累計販売台数

表 4.3: 販売実績の比較 [2]

	食器洗い乾燥機	空気清浄機	生ごみ処理機
販売台数（台）	7,813,372	4,744,380	1,565,526
売り上げ金額（百万円）	364,041	70,547	62,335

および売り上げ金額の比較を表 4.3 に示す． 食器洗い乾燥機、空気清浄機は現在，家庭に普及し始めつつあると考えられるが，販売台数を比較すると 3-5 倍の差がある．このように生ごみ処理機は，依然導入期であり，成長期に入っていない．そのため，成長期に入るためにはいかに普及率を上げていくかが，家庭用生ごみ処理機市場の課題となっている．

生ごみ処理機の販売を促進し，普及率をあげる施策につながる知見を顧客へのアンケートから得ることを試みる．この製品を購入している人（購買者），購入していない人（非購買者）を対象に，アンケートを作成し，Web アンケート会社に依頼し，データを収集した．データの収集では，まず，データが特定の顧客層に偏らないよう，Web アンケートシステムに登録されているユーザーからランダムに選択した 1 万人に対し，生ごみ処理機の購入の有無および年代・性別を事前に質問し回答を得た．その結果を元に，性別および年代が偏らないようにしながら，購買者，非購買者とも 300 人を同定した．これら合計 600 人を被験者として，本調査となるアンケートを実施した．アンケートでは，家族人数，世帯年収，住居スタイル，製品の認知度（全く知らない，名前だけを知っている程度，機能を把握している，購入済）などの顧客の属性情報（15 項目）を取得し，4.2 節で述べたマーケティングミックス (4P) に関係する質問 (25 項目) をもとに製品に関する意識調査を行った．例として，製品に対する意識に関する質問を図 4.7 に示す．得られた 600 件のデータにおける顧客属性の分布の内，性別・年代・家族人数・世帯年収の分布を図 4.8 に示す.

さらにこれらの選択式のアンケートデータの他に，テキストデータとして,「どういった課題が解決されれば生ごみ処理機を買うか？(購入者の場合は，何が購入する前に気になったか？)」という質問に対する回答を自由記述形式で取得した．この時，回答を必須とするとともに，別の製品を題材とした回答文例を複数提示し，できるだけ同じ長さの回答を得られるようにした．得られた 600 件のテキストデータの統計情報を表 4.4 に示す.

取得したデータを元に，対象製品（生ごみ処理機）を普及させるための施策につながる

[2]一般社団法人日本電機工業会による統計 http : //www.jema – net.or.jp/Japanese/data/ka01.html

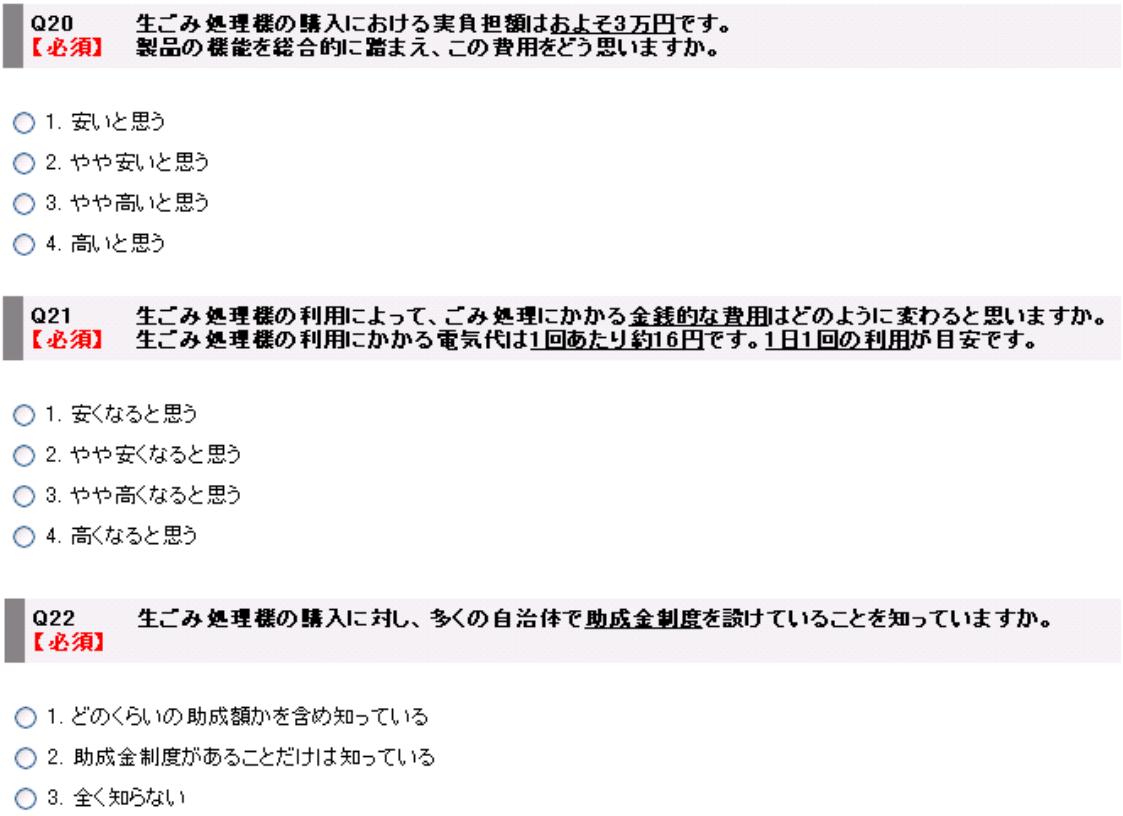

図 4.7: Web アンケートの質問例

知見を得ることを試みる．まず，テキストマイニングを用いて顧客層と製品意識との間の傾向を分析する．次に，データマイニングを用いて購買者と非購買者を決定づけるルールを抽出する．そして，抽出されたルールの中から，テキストマイニングで得られた結果に関連するもののみをフィルタリングし，出力する．

4.5.2　テキストマイニングの結果

テキストマイニング分析では，分析観点を定義し，観点ごとに属する単語や表現を辞書として整備することが行われている．これにより，テキスト中から抽出された個々の単語や表現の出現回数ではなく抽象化したレベルで分析することができ，知見につながりやすい分析結果を得ることが期待できる．本分析では，4.2 節で述べたマーケティングの視点を分析観点に利用する．具体的には製品提供におけるマーケティングミックス（4P）を分析観点として定義した．そして，対象データからランダムに抽出した 100 文書から抽出した名詞をマーケティングミックス (4P) を観点（カテゴリ）として分類し，辞書を作成

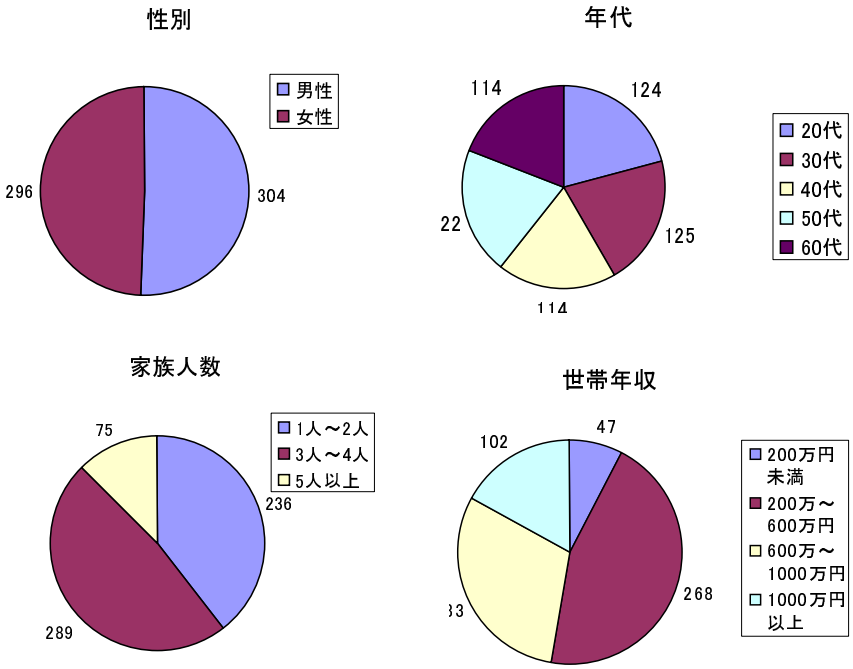

図 4.8: アンケート対象の分布

した．観点ごとの辞書エントリーを表 4.5 に示す．なお，Place に相当する名詞は得られなかった．

これにより，例えば，Product に関する言及をしている文書数を調べることができる．各表現の出現頻度が少数で傾向を把握することが困難である場合でも，観点レベルで出現頻度を集約し，顧客属性との関係を分析することが可能となる．

アンケート中の順序関係を持つ顧客属性に 4.4.2 節で提案した分析観点のランキング手法を適用し，得られた分析観点の組み合わせと特徴量 S_{seg} を表 4.6 に示す．

Price, Product, Promotion といったマーケティング要素への言及がそれぞれのセグメント軸でどういった傾向を示すのかを知るには，マーケティング要素とセグメント軸の 2 次元の表を作成し，言及頻度を求め傾向を観測する必要があった．通常，どのセグメント軸を選択すればよいかわからないため，従来は必要と思われる全てのセグメント軸に対して分析を行う必要があった．それに対して，本研究で提案する手法を用いることによって，表 4.6 から Promotion については世帯年収が高くなるにつれて言及が減少するという傾向を予測できる．一方，製品認知度というセグメント軸では，製品を知っているほど

表 4.4: テキストデータの統計情報

	平均	標準偏差
文数	2.28	0.21
単語数	26.7	6.28
文字数	51.2	10.9

表 4.5: 作成した辞書のエントリー（一部）

観点	辞書エントリー
Product	処理，電気代，におい，手間，無臭，大きさ 処理時間，音，ランニングコスト
Price	価格，値段，購入価格，金額，低価格 購入代金，安価
Promotion	無料，助成金，補助金，無料配布
Place	(該当する辞書エントリー無し)

Promotion の言及頻度が下がり，Product の言及頻度があがっていることが予想される．実際に，セグメント軸として世帯年収と製品認知度を選択した場合の言及頻度の傾向分析結果を図 4.9 に示す.

世帯年収が低い顧客や製品を知らない顧客ほど，Promotion(販売促進策) が気になるという傾向が実際に確認できる．この傾向は，販売促進策は主に価格に関係することから，世帯年収と関連があることは既知の傾向だと考えられる．また，製品を知っている消費者ほど，販売促進策に対する不満点や期待が少なくなるが，逆に Product(製品自体) に対する不満点や期待を持つという傾向を確認することができる．この傾向は既知のものではなく，詳細に調べることで，例えば製品のどういう点が重要視されているのか，どこに不満が存在するのか，といった有用な知見を得られる可能性がある．そこで Product に属するキーワードと製品の認知度とを調べ，詳細な傾向分析を行った．表 4.7 に各製品認知度の顧客コメントにおける Product に関する概念への言及頻度を示す.

この結果，製品の認知度に関わらず,「におい」を懸念事項にあげる記述があるが，製

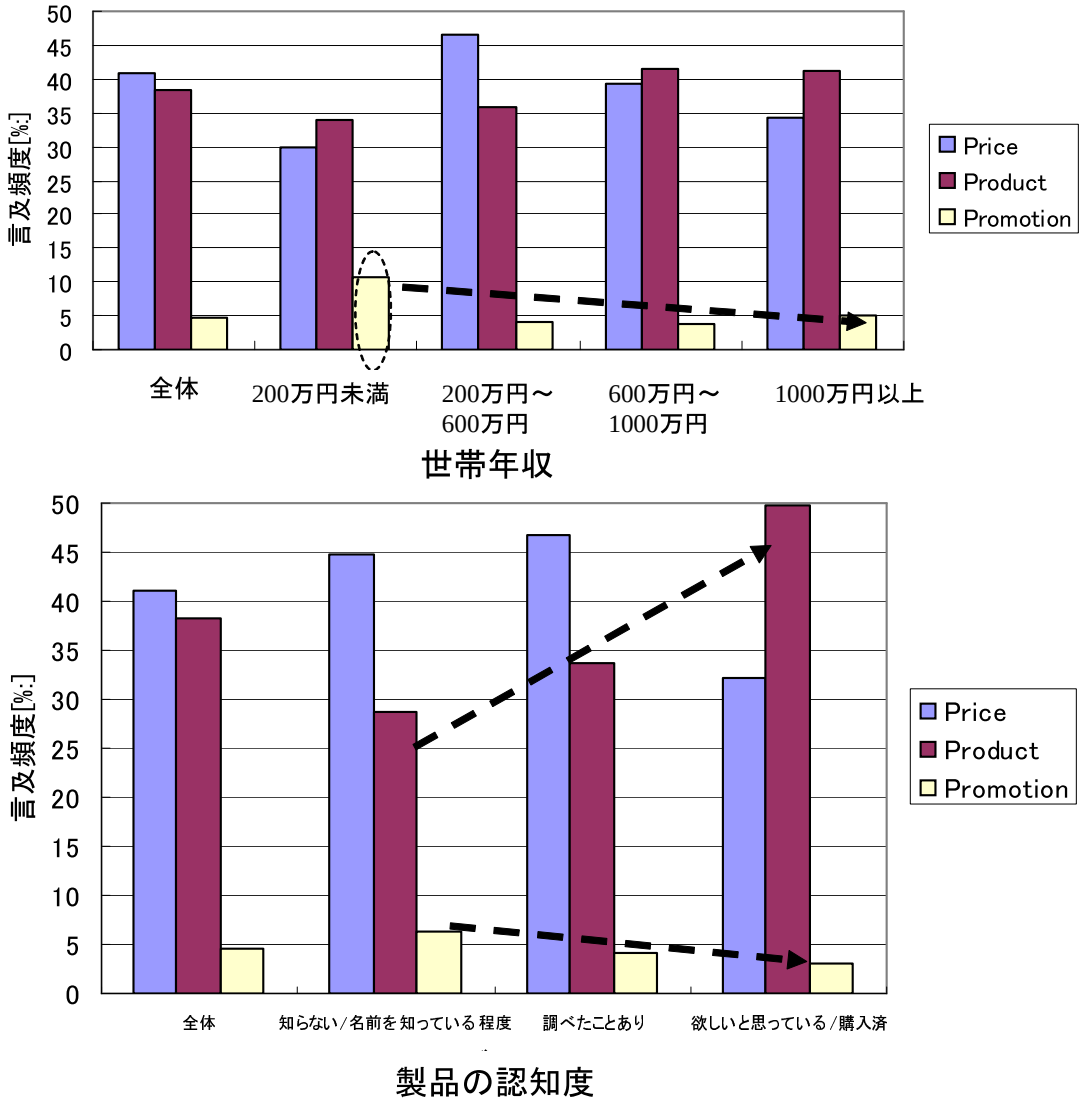

図 4.9: セグメント軸を世帯年収（上）と製品認知度（下）にした場合のマーケティング要素の言及頻度

表 4.6: 顧客属性に対して増加・減少の傾向を示す分析観点（マーケティングミックス）

増加傾向			減少傾向		
観点	顧客属性	S_{seg}	観点	顧客属性	S_{seg}
Product	製品認知度	−0.0693	Promotion	世帯収入	0.214
Price	世帯収入	−0.0366	Promotion	製品認知度	0.0642
Product	世帯収入	−0.0240	Price	製品認知度	0.0366
Product	世代	−0.00452	Promotion	家族人数	0.0159
Price	家族人数	−0.00165	Price	世代	0.0148
			Product	家族人数	0.0130
			Promotion	世代	0.00748

品をよく知っている消費者に「音」や「処理時間」を懸念事項に挙げている記述が存在した．一般に製品・サービスを改善するにあたって消費者へのインタビューを実施することが行われているが，時間とコストの観点からインタビューの項目数には制限があることが多い．本分析結果から,「におい」だけでなく「音」や「処理時間」を中心に詳細なインタビューをすることで，改善につながる知見を得られる可能性が高いことがわかり，効果的なインタビューができる可能性がある．

4.5.3　テキストマイニング分析結果を元にしたデータマイニングの実践例

顧客アンケート中の定型質問項目への回答に対して決定木分析を適用し，対象製品の購買者と非購買者に関するルールを抽出した．4.2 節で述べたように，リーフのみを持つ末端ノードに注目し，ある特定の顧客層に対して，非購買者から購買者へと遷移できる施策を見つけることを考える．J48 を用いた決定木学習において名義尺度の分割に 2 分分割を用いている．そのため，自治体からの助成金の存在についての認知度 (額まで知っている，存在だけ知っている，全く知らない) という属性からは，条件とその値として，例えば,「助成金の存在=全く知らない」と「助成金の存在=額まで知っている or 存在だけ知っている」が得られる．本実践例では，27 のルールが抽出された．

しかしながら，4.3 節で述べたように，ルールを構成している条件の組み合わせは必ずしも意味ある顧客層を表すわけではなく，専門家が抽出されたルールを精査する必要があ

表 4.7: Product に関する概念への言及と製品への認知度との関係

概念（頻度）	製品認知度 (文書数)		
	知らない/名前を知っている程度 (245)	調べたことがある (122)	欲しいと思っている/購入済 (233)
におい (144)	61	26	57
処理 (42)	7	10	25
匂い (42)	12	4	11
音 (17)	1	3	13
処理時間 (15)	2	3	11
ランニングコスト (9)	6	2	1

る．抽出されたルールの 1 例を表 4.8 に示す．この結果では，条件 1 から 6 で表現される顧客層には製品に関する情報をきちんと提供することが有効である可能性があることが推察される．しかしながら，なぜこの顧客層に有効であるかは自明ではなく解釈が必要となる．その際，本例では，東北地方での販売実績，助成金の有無，心理学属性（他人の評価を気にするか?）について，対象製品に関する領域知識やマーケティング理論の知識が必要であり，専門家が精査する必要がある．表 4.8 は，対象製品のマーケティング専門家の評価で，解釈が困難とされたルールである．このように抽出された結果を専門家が精査した結果，知見とならないルールが多数を占めることもある．

そこで，4.3 節で述べたようにテキストマイニングで得られた傾向分析の結果を元に，抽出されたルールをフィルタリングする．テキストマイニングの結果，マーケティング要素と顧客属性の間では，Promotion と世帯年収，Promotion, Product と製品認知度が関係していることがわかった．こうして得られた傾向分析の結果を元に，データマイニングで得られたルールをフィルタリングすることで，ルール数は 8 に削減された．

抽出されたルールに対して，対象製品のマーケティング担当者による評価を実施した．評価では，各ルールが施策を立てる上で有用かどうかを以下の 2 つの視点で分類した．

1. 新規の知見が得られる，または，現在検討している仮説を（部分的に）裏付ける結果である

2. 役に立たない，解釈が困難

表 4.8: 抽出されたルールの例（専門家の評価では，知見に結びつかないと判定されたルール）

	各ノードでの条件	条件が取る値	判定結果
1	自治体による購入に対する助成金制度	知らない or 名前だけ知っている	
2	生ごみの処理方法	分別・その他	
3	自分が購入した商品に対する他者の評価	気にならない or やや気になる	
4	製品を操作する手間	あまり手間でない or 手間である	
5	居住地域	東北地方	
6	世帯年収	600 万円以上	
7	操作時の臭い	やや気になる or 気になる	購買者
		あまり気にならない	非購買者

27 のルールを評価した結果，上記の評価 1 に相当するルールの数は 3 であり，全てフィルタリングされたルールに含まれていた．本実践結果では，テキストマイニングによる分析結果を利用することで効果的にルールをフィルタリングできることがわかった．

評価が 1 であったルールを表 4.9，4.10，4.11 に示す．

テキストマイニングで関係があると分析された観点の組み合わせについて，表 4.9 のルールは Product と製品認知度（条件 2,3,5,6），Promotion と製品認知度 (条件 1,2,3) を，表 4.10 のルールは Promotion と製品認知度（条件 1,3）を，表 4.11 のルールは Promotion と製品認知度（条件 1,3），Promotion と年間世帯所得（条件 1,6）を含んでいる．

これらの分析結果から対象製品の販売促進につながる仮説として以下を導出することができた．

(i) 自治体による助成を知らない顧客が存在するため，アピールが必要 (ルール (2),(3)).

(ii) 現在の生ごみ処理に問題意識を持っている顧客には，実演販売などを通して機能だけでなく，生ごみ処理費用が削減できることを訴求する (ルール (1)).

(iii) 類似した製品であるディスポーザーを知っている顧客もいるため，実演販売では比較し価値を訴求する (ルール (2)).

表 4.9: フィルタリングで得られたルール (1)

	各ノードでの条件	条件が取る値	判定結果
1	自治体による購入に対する助成金制度	知っている	
2	メーカーによる実演販売	見たことがない	
3	家電量販店などでの製品広告	見たことがある	
4	生ごみ処理で得られる有機肥料の利用法	園芸や菜園の肥料	
5	処理時間が 100 分であることに対する評価	やや長い	
6	生ごみの減量効果 (最大 1/7) に対する評価	価値がある	
7	購入時に価格を重視するか	重視する	
8	生ごみの衛生面	気になる	
9	生ごみの臭い	気になる	
10	(ごみの有料化に伴う) 生ごみ処理にかかるごみ袋費用の負担	やや気になる or 気になる	購買者
		あまり気にならない	非購買者

(iv) 購入時の負担は安いと思われているが低所得家庭での普及は進んでいない．利用することで，ごみ処理費用が削減できることを訴求する (ルール (3)).

これらに対する，対象商品のマーケティング専門家の考察は以下のようなものであった.

(A) 各自治体で助成金の有無・金額に対する情報をすでに提供している (i).

(B) 実演だけでなく，ごみの有料化によって発生するごみ袋料などの処理費用を削減できることを示すことは有効だと考えられる (ii).

(C) ディスポーザーに興味を持っている顧客は潜在顧客層と考えられ，すでに施策を取っている (iii).

表 4.10: フィルタリングで得られたルール (2)

	各ノードでの条件	条件が取る値	判定結果
1	自治体による購入に対する助成金制度	知らない or 名前だけ知っている	
2	生ごみの処理方法	燃えるごみ	
3	ディスポーザーという家電を知っているか？	関心がある or 購入予定 or 購入済み	
4	生ごみ処理機を使う場合，コストは変わるか？	高くなると思う	
5	販売店での実演販売を見たことがあるか	はい	購買者
		いいえ	非購買者

(D) ごみ処理費用の削減を訴えることは重要である．しかしながら生活必需品ではないため，低所得家庭向けに普及するには，それだけでは不十分である (iv).

導出された仮説や施策は，専門家にとって全く新規な気づきとなるものばかりではないが，実施している施策の重要性の裏付けや，現在の課題点を解決するための糸口として利用できることがわかった．

4.6　考察

本研究では，テキストマイニングを用いて得られた傾向分析の結果を用いて，データマイニングで得られるルールの分析を支援することを検討した．本研究ではデータマイニングで得られる大量の結果に対して，テキストマイニングの結果に関連するもののみをフィルタリングすることを行った．そして，生ごみ処理機の市場分析という実践例では，実際に専門家の作業を軽減できることが確認された．データマイニングで得られたルールのフィルタリングでは，ルールを構成する条件の記述とテキストマイニングで得られた結果の関連性を調べたが，ルールが持つ定量的な情報は用いなかった．例えば，各ルールについて末端ノードまでの条件で表現されるデータ数などが定量的な情報であり，ルールの重要度を測る一つの指標となる．このようなデータマイニングの結果が持つ，定量的な情報を検討することは今後の課題の１つである．

また，テキストマイニングによる分析において，顧客属性が順序属性を持つことを利用

表 4.11: フィルタリングで得られたルール (3)

	各ノードでの条件	条件が取る値	判定結果
1	自治体による購入に対する助成金制度	知らない or 名前だけ知っている	
2	生ごみの処理方法	燃えるごみ	
3	ディスポーザーという家電を知っているか？	知らない	
4	作動中の臭いの有無を購入時に検討する (した) か？	する (した)	
5	購入時の負担額が 3 万円であることをどう思うか？	安いと思う or やや安いと思う	
6	世帯年収	やや多い or 多い	購買者
		やや少ない or 少ない	非購買者

し，特徴的な傾向を示す分析観点の同定方法を提案した．本手法によって,「○○（顧客属性）が大きくなるほど，△△（マーケティングミックス）に対する言及が多くなる」といった傾向を得ることができる．生ごみ処理機の市場分析での実践例では，製品が持つ潜在的課題を示す傾向が得られ，詳細な顧客インタビューへの糸口となることがわかった．また，データマイニングの結果とあわせることで，助成金の周知徹底だけでなく，ごみの有料化に伴って発生しているごみ袋費用を削減できることを示すことが有用であることがわかった．テキストマイニングをデータマイニングに結果を合わせることで，具体的なルールを導出できたことから，本手法の実践は有効であると考えられる．

本研究では，実践例として，生ごみ処理機という未普及製品の販売促進を対象とした市場分析を行った．この場合では，顧客は新規のサービスや製品に十分な経験と知識を持っていないため，マーケティング要素 (7P) と顧客属性を元に知見の抽出を試みた．

一方で，4.2 節で述べたように，顧客が知識や経験を持っている既存の製品やサービスに対する市場分析では，サービスや製品の品質に対して顧客が持つ感覚尺度 (5D) との関係も考慮しなければならない．ここでは，少数の収集データに対して分析手法を適用する．データとして，最近利用したホテルに対する簡易アンケートを 93 人の被験者に行った結果を用いた．アンケートデータには，定型的な質問回答として，性別，年代，ホテルの価格帯，ホテルを利用する目的（重視する点=顧客が抱く期待品質）が含まれている．また，テキストデータとして，利用したホテルの良かった点，悪かった点についてのコメ

ントが含まれている．このデータをテキストマイニングの対象とした．分析前に辞書として，テキストデータ中に 2 回以上出現する名詞 (164 語) から 7P に関係する名詞（92 語）を選択して分類した．

分析では，分析観点を用いて，ホテルサービスの各サービス構成要素が持つ特性，顧客が抱く期待品質，顧客属性らの間の関係から，ホテルサービスの改善につながる知見を得ることを試みる．

1. ホテルサービスの各サービス構成要素が持つ特性は？

2. 顧客が抱く期待品質ごとに各サービス構成要素の特性は傾向を持つのか？

3. 顧客属性と期待品質の間の関係は？

まずサービスのマーケティング要素が持つ特性として，各構成要素の品質分類を考える．品質の分類として魅力的品質・当たり前品質・一元的品質があり [91]，それぞれ以下の特徴を持つ．

1. 魅力的品質：充足されると満足するが，不充足でも仕方がないと感じる

2. 当たり前品質：不充足であれば不満を感じ，充足でも当たり前と感じる

3. 一元的品質：充足されると満足し，不充足であれば不満を感じる

図 4.10 に良かった点と悪かった点のそれぞれのコメントにおけるマーケティング要素の言及頻度を示す．

各顧客のコメントには顧客が重視する点が定型項目として付与されており，それらは分析観点の顧客が抱く期待品質と結びつけられる．これにより，顧客が抱く期待品質ごとにマーケティング要素が良かった点，悪かった点に関するコメントでの言及頻度を求めることが可能になる (図 4.11)．

図 4.10 と図 4.11 に示した言及頻度の差から品質分類に関して以下の仮説が立てられる．

1. 確実性重視の顧客にとって Product は魅力的品質である

2. 確実性重視の顧客にとって Place は一元的な品質である

3. 共感性重視の顧客にとって Product は当たり前品質である

4. 有形性重視の顧客にとって Product は魅力的品質である

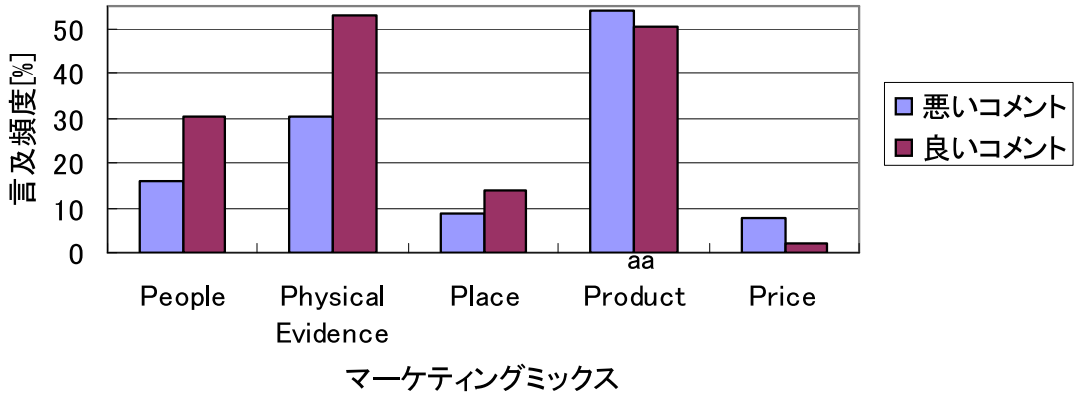

図 4.10: 各コメント欄におけるマーケティング要素の言及頻度

表 4.12: 分析観点の特徴量 (S_{seg})

	世代	価格帯
保証性	−0.0114	−0.00448
共感性	0.0391	−0.0465
有形性	−0.0144	−0.0801

しかし，サービス提供者側がこれらの仮説をサービス改善に結びつけるのが難しい．なぜならば，サービス提供者側が顧客が 5D のうち何を重視しているかどうかを事前に知ることが難しいからである．サービス提供者側が得られる情報は顧客属性（世代，性別，人数）や提供している価格などである．

これらの顧客属性のうち順序性を持つものに対して，顧客の期待品質の重要視の増減傾向を持つかどうかを提案手法を用いて導出した．得られた特徴量 S_{seg} を表 4.12 に示す．ここから,「年代が低いほど共感性を重視する」,「価格帯が高いほど有形性を重視する」,「価格帯が高いほど共感性は重視しなくなる」という傾向が顕著であることが推測できる．これらの傾向は実際の分析結果 (図 4.12) で確認できる．この結果と品質分類から得られた結果を組み合わせることでホテルの形態にあった，サービス改善のための仮説を立てることができる．例えば，若い世代の利用者が多いホテルの場合，共感性重視の顧客が多いことが予想できる．共感性重視の顧客には Product が当たり前品質になっているため，設

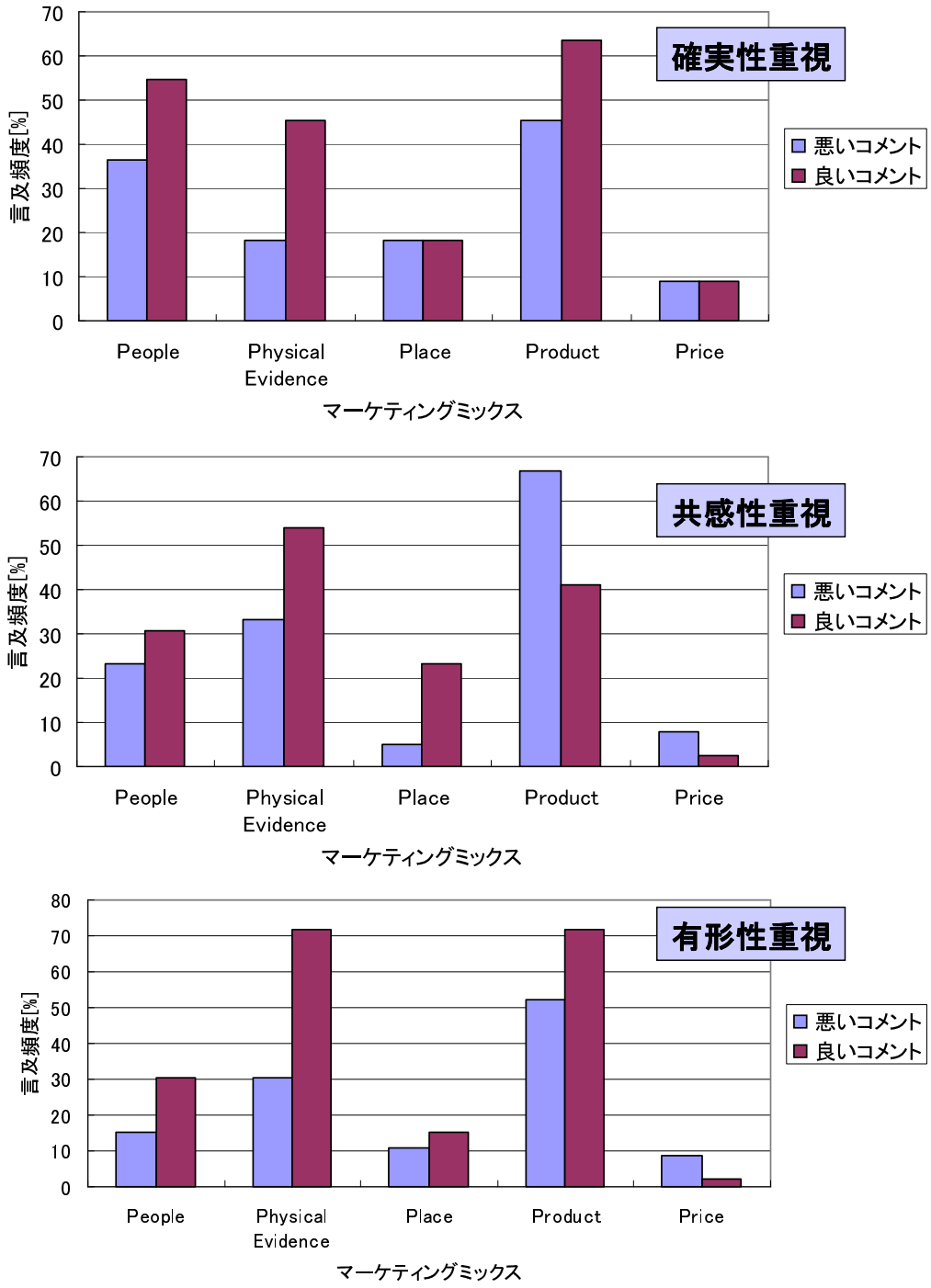

図 4.11: 確実性 (上)，共感性 (中)，有形性 (下) を重視する顧客コメントにおけるマーケティング要素の言及頻度

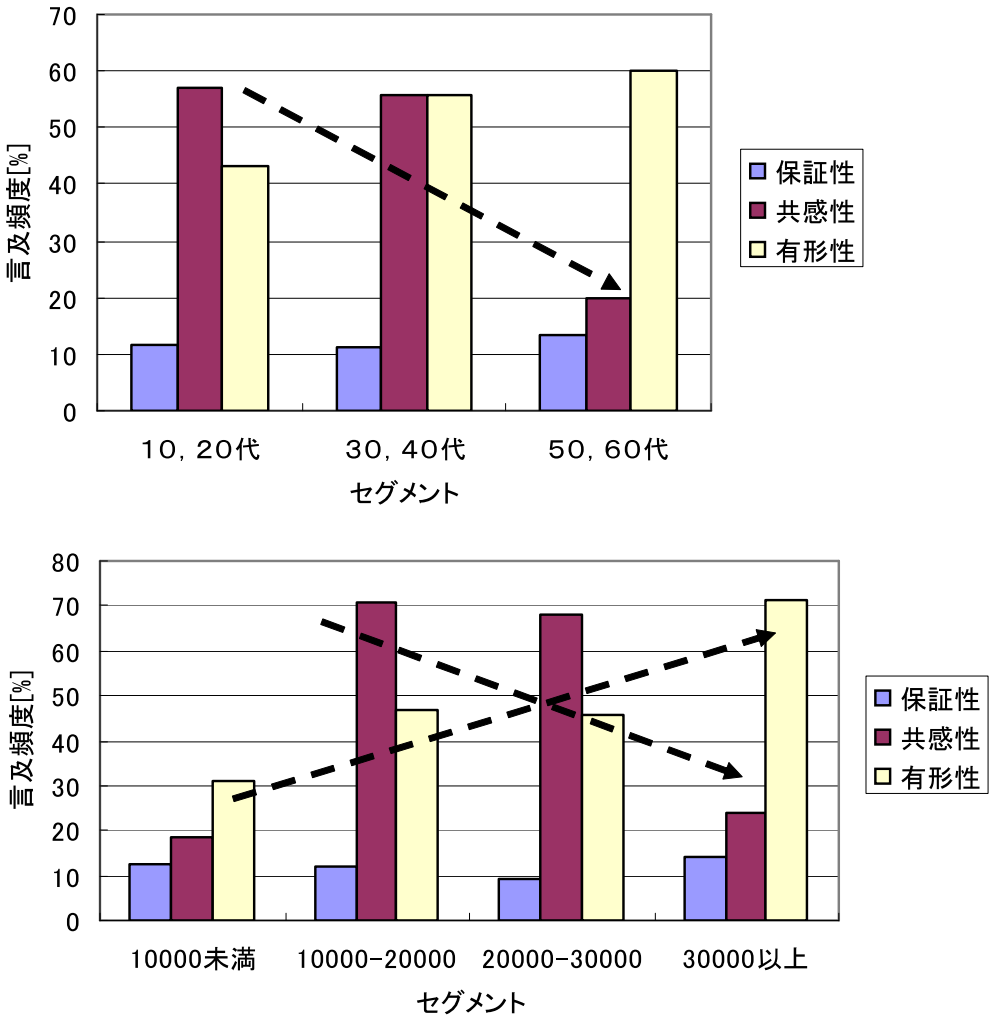

図 4.12: セグメント軸を年代 (上) と価格帯 (下) の時の期待品質 (5D) の言及頻度

備が最低限の基準を満たしているかどうかを検証する必要があると考えられる．また，価格帯が高いホテルは有形性重視の顧客が多いことが予想される．有形性重視の顧客にとって Product は魅力的品質であり，Physical Evidence（眺望など）は一元的品質である．したがって，最新設備や他にない設備の考案が重要であり，眺望が良ければ積極的にアピールする価値があると考えられる．

この簡易データの分析では，定型項目データが少ないため，4.5 節で行ったようなデータマイニングを用いたルール発見が行えない．そのため，テキストデータからの傾向分析のみとなり，具体的なアクションにつながる分析結果は得られなかった．既存の製品やサービスに対する市場分析での深い実践を通して，本手法の有用性を検証することが今後の課題である．

4.7　本章のまとめ

本研究では，市場分析におけるデータマイニング実践においてテキストマイニングを活用する手法について検討した．テキストマイニングによる傾向分析では，概念や観点の間に関係があるかどうかを効果的に同定する必要がある．そのためには，文書に付与されている定型項目やキーワードを分類して作成した分析観点の組み合わせ方が重要となる．本研究では，対象にデータに付与されている定型データが順序特性を持つ場合にその性質を利用した分析手法を提案した．そして，収集した顧客コメントに対して，提案した分析観点と分析手法を適用し，傾向分析結果を効果的に得られることを示した．またテキストマイニングによる傾向分析とデータマイニングによって抽出されるアクションにつながるルールを組み合わせることで，意思決定の支援ができることを実践例を通して示した．本研究では，未普及製品の販売促進という市場分析の一つを検討したが，他の市場分析への展開などが今後の課題である．

第5章　結論

本論文では，テキストマイニングの実践に関する研究を行った．通常，テキストマイニングでは，分析を行う前に分析観点を定義し，各観点に関係するキーワードや表現を辞書として登録する前処理が行われる．しかしながら，分析観点や辞書といった分析モデルの初期設定は分析者の対象分野に対する知識に依存する．また，キーワードや表現の出現頻度を分析するための分析観点の選択や結果の解釈といった後処理も，分析者の経験や勘に依存することが多い．その結果，分析観点や辞書の再設定につながるフィードバックをし，分析ループを回すことができず，知見につながる分析結果が得られない場合が存在している．

そこで，本論文では，テキストマイニングの実践において，有効な分析観点や辞書の初期設定，そして分析時における分析観点の選択といった課題を対象とし，局所化手法の適用を行った．局所化手法として，分析目的に応じて，各テキストデータにおいて分析する範囲を限定する手法を用いて，前処理を効果的に行う分析手法を提案した．また，情報抽出で得られた様々な分析結果を局所化することで絞り込み，専門家が精査すべき分析結果を削減することで，後処理を効果的に行う分析手法を提案した．本論文では，市場分析および会話分析において，これら2つの局所化手法を利用した分析手法を実践し，実践例を通してその有用性を検証した．

3章では，会話分析を対象とし，タスクを持った会話からタスクの成功につながる発言パターンの抽出を行った．例えば営業活動や問題解決のようにタスクとその結果を伴う会話において，何が成功に寄与しているかといった要因分析は，生産性の向上への活用が期待できることから，テキストマイニングの魅力的なアプリケーションである．しかしながら会話データの場合，各会話のデータサイズが大きくなり冗長な表現も多く含まれる．そのため，この要因分析においては，冗長性の高い会話の一体どこに着目すれば有益な知見の獲得につながるかの判断が重要である．しかしながら，分析者の勘に依存しながら試行錯誤していては効率が悪いうえ，たとえ要因が存在しても，そこに気づけるとは限らない．そのため分析者の知識や経験に依存しない分析手法が必要となっている．そこで，局所化手法として，冗長な発言を含む会話データからタスクの成功に寄与する重要発言区間

を同定する手法を提案した．タスクを持った会話は話の流れが事前に決まっている，という性質に注目し，各会話データの最初の発言から特定の発言までを集めた時系列累積データを定義した．そして，タスクの結果を分類する学習器を時系列累積データを用いて作成し，その精度の算出し，その推移を元に重要発言区間を同定する手法を提案した．また，同定した重要発言区間からタスクの成功に関連するキーワードを偏在性と新規性の観点で抽出する指標を提案した．提案手法の実践例としてコンタクトセンター受託企業で収集されたレンタカーの予約会話データを対象とした．そして，顧客が予約した車を取りに来る/来ないと結果が異なる予約会話間の差異分析を行った．提案手法を適用した結果，長い会話の中から結果に影響を与える重要発言区間として，顧客の最初の発言および提案時の発言を同定した．そして，その中から結果に関連する発言パターンを抽出し，顧客の最初の発言には車を借りる意思を明確にする発言と，値段の問い合わせを主目的とした発言があり，前者の発言をした顧客ほど予約した車を取りに来る可能性が高いという知見が得られた．また，提案時の発言として，ディスカウントに関連する表現や提案内容が良いことを訴求する表現が結果に影響を与えることを抽出した．そして抽出した発言パターンから得られた知見を元にオペレーターへの教育を実施した．教育受けたオペレーターグループを他のグループと比較した結果，予約された車の利用率を約３％向上することができることがわかった．提案した局所化手法は，話しの流れが事前に定義されているという性質に基づいている．このようなタスクを持ったビジネス会話と同様の性質を持つ会話以外のデータへの適用拡大が今後の課題となっている．

４章では，市場分析を対象とし，自由回答および選択回答形式のアンケートデータから次期購買層の発見につながるルールの抽出を行った．テキストマイニングでは通常，キーワードの出現頻度を分析することが行われ，市場分析においては特定の顧客層のテキストデータに多く出現するキーワードを同定し，知見を導出するすることが試みられている．このような分析では，単に多く出現するキーワードがわかるのみであり，具体的なアクションにつながる結果が得られないことが多い．また，キーワードの出現傾向を調べる際，顧客属性など様々な分析観点の洗濯が考えられる．そのため効果的な分析結果を得るためには，分析観点を試行錯誤してしながら選択する必要がある．一方，データマイニングを用いたアンケート分析として次期購買者につながるルールを抽出することが考えられる．通常，データマイニングによるルール発見では，結果としてルールが多数抽出されるが，そのほとんどが対象分野の専門家によって解釈できないことがある．そのため，データマイニングの結果を有効に活用できないことが実践上の課題であった．そこで，まずテキストマイニングにおいて，分析するキーワードを限定する局所化手法として，順序関係を持った顧客属性に対して頻度が増加・減少する傾向を持つキーワードをランキングする

手法を提案した．そして，データマイニングによるルール発見の結果から，テキストマイニングで関連があると分析したキーワードと顧客属性の組を含むルールをフィルタリングする手法を提案した．実践例として，生ごみ処理機の市場分析を目的とした購買者・非購買者へのアンケートデータから次期購買層の発見につながるルールの抽出を行った．テキストマイニング分析において，提案手法を用いて，顧客属性に対してテキスト中のキーワードの出現頻度が増加・減少の傾向示す組み合わせを抽出した．その結果，Promotion に関するキーワードは世帯年収が高くなるにつれて言及が減少するという傾向があると抽出できた．また，製品認知度に対して，製品を知っているほど Promotion に関するキーワードの言及頻度が下がり，Product に関するキーワードの言及頻度が上がる傾向を抽出できた．この分析結果を用いてデータマイニングで得られた購買者・非購買者を決定付けるルールのフィルタリングを行った．そして，生ごみ処理機の市場分析例では，提案手法によりマーケティング専門家が解釈・評価を行うルール数を，精度を保ちながら約 1/3 に削減することができた．分析例として，生ごみ処理機という普及が進んでいない製品の販売促進を対象とした市場分析を行った．この場合では，顧客は新規のサービスや製品に十分な経験と知識を持っていないため，マーケティング要素 (7P) と顧客属性を元に知見の抽出を試みた．一方で，顧客が知識や経験を持っている既存の製品やサービスに対する市場分析では，サービスや製品の品質に対して顧客が持つ感覚尺度 (5D) との関係も考慮しなければならない．このような既存製品やサービスの市場分析への適用が今後の課題である．

3 章で扱った会話データのように各データのサイズが大きい場合，各データにおいて特定の範囲に限定する局所化手法が有効であることがわかった．会話データに限らず，報告書データなど，各文書のサイズが大きいデータを分析対象とすることは，今後増大すると考えられる．そのような場合，全データを対象にするのではなく，分析目的に応じて，積極的に各データから分析範囲を限定し効果的な分析を行う手法の開発が今後必要になってくると考えられる．

参考文献

[1] S. Agarwal, S. Godbole, D. Punjani, and S. Roy. How much noise is too much: A study in automatic text classification. In *7th IEEE International Conference on Data Mining*, pp. 3–12, 2007.

[2] F. Beil, M. Ester, and X. Xu. Frequent term-based text clustering. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 436–442, 2002.

[3] CaboCha. http://code.google.com/p/cabocha/.

[4] S. Challa, S. Roy, and L. V. Subramaniam. Analysis of agents from call transcriptions of a car rental process. In *Proceedings of the Language, Artificial Intelligence and Computer Science for Natural Language Processing applications (LAICS-NLP)*, 2006.

[5] K. Chantola. Surveys on inverted index updating and semistructured data indexing and aggregation for takmi. *IBM Research Report*, No. RT0816, 2008.

[6] ChaSen. http://chasen-legacy.sourceforge.jp/.

[7] M.-C. Chen, L.-S. Chen, C.-C. Hsu, and W.-R. Zeng. An information granulation based data mining approach for classifying imbalanced data. *Information Sciences*, Vol. 178, No. 16, pp. 3214–3227, 2008.

[8] Y. Chen, F. S. Tsai, and K. L. Chan. Machine learning techniques for business blog search and mining. *Expert Systems with Applications*, Vol. 35, No. 3, pp. 581–590, 2008.

[9] L. Chiticariu, R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, and S. Vaithyanathan. SystemT: An algebraic approach to declarative information extraction. In *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 128–137, 2010.

[10] A. Cockburn. *Writing Effective Use Cases.* Addison-Wesley, 2000. (邦訳 : ユースケース実践ガイド, ウルシステム株式会社 監訳, 山岸 耕二, 矢崎 博英, 水谷 雅宏, 篠原 明子 訳, 翔泳社, (2001)).

[11] A. M. Cohen and W. R. Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, Vol. 6, No. 1, pp. 57–71, 2004.

[12] I. Donaldson, J. Martin, B. Brijin, C. Wolting, V. Lay, B. Tuekam, S.Zhang, B.Baskin, GD. Bader, K. Michalickova, T. Pawson, and CW. Hogue. Prebind and textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, Vol. 4, No. 11, 2003.

[13] S. Douglas, D. Agarwal, T. Alonso, R. M. Bell, M. Gilbert, D. F. Swayne, and C. Volinsky. Mining customer care dialogs for "daily news". *IEEE Transaction on Speech and Audio Processing*, Vol. 13, No. 5, pp. 652–660, 2005.

[14] W. Fan, L. Wallace, S. Rich, and Z Zhang. Tapping the power of text mining. *Communication of the ACM*, Vol. 49, No. 9, pp. 77–82, 2006.

[15] R. Feldman and J. Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press, New York, 2007.

[16] D. Ferrucci and A. Lally. Accelerating corporate research in the development, application and deployment of human language technologies. In *Proceedings of the HLT-NAACL workshop on software engineering and architecture of language technology system*, pp. 67–74, 2003.

[17] G. Feuerlicht. Database trends and directions: Current challenges and opportunities. In *Proceedings of the Database, Texts, Specifications, and Objects (DATESO)*, pp. 163–174, 2010.

[18] T. Finin, et al. National science foundation symposium on next generation of data mining and cyver-enabled discovery for innovation: Final report. 2007.

[19] R. Gacitua, P. Sawyer, and V. Gervasi. On the effectiveness of abstraction identification in requirements engineering. In *Proceedings of the 18th IEEE International Requirements Engineering Conference*, pp. 5–14, 2010.

[20] M. Golfarelli, S. Rizzi, and I. Cella. Beyond data warehousing: what's next in business intelligence? In *ACM 7th International Workshop on Data Warehousing and OLAP(DOLAP)*, pp. 1–6, 2004.

[21] B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, Vol. 12, No. 3, pp. 175–204, 1986.

[22] P. Haffner, G. Tur, and J. H. Wright. Optimizing svms for complex call classification. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 632–635, 2003.

[23] H. W. Hastie, R. Prasad, and M. A. Walker. What's the trouble: Automatically identifying problematic dialogues in darpa communicator dialogue systems. In *Proceedings of the 40th Annual Meeting of the ACL*, pp. 384–391, 2002.

[24] T. Hisamitsu and Y. Niwa. A measure of term representativeness based on the number of co-occurring sailent words. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pp. 1–7, 2002.

[25] H.-L. Hu and Y.-L. Chen. Mining typical patterns from databases. *Information Sciences*, Vol. 178, No. 19, pp. 3683–3696, 2008.

[26] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 133–142, 2002.

[27] Z. Jourdan, R. K. Rainer, and T. E. Marshall. Business intelligence: An analysis of the literature. *Information Systems Management*, Vol. 25, No. 2, pp. 121–131, 2008.

[28] JUMAN. http://nlp.ist.i.kyoto-u.ac.jp/index.php?juman.

[29] KNP. http://nlp.ist.i.kyoto-u.ac.jp/index.php?knp.

[30] R. Kohavi, N. J. Rothleder, and E. Simoudis. Emerging trends in business analytics. *Communication of the ACM*, Vol. 45, No. 8, pp. 45–48, 2002.

[31] P. Kotler, T. Hay, and P. N. Bloom. *Marketing Professional Services*. Pearson Education, 2002.

[32] H.-K J. Kuo and C.-H. Lee. Discriminative training of natural language call routers. *IEEE Transaction on Speech and Audio Processing*, Vol. 11, No. 1, pp. 24–35, 2003.

[33] A. Kusiak. Data mining: manufacturing and service applications. *International Journal of Production Research*, Vol. 44, No. 18-19, pp. 4175–4191, 2006.

[34] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pp. 591–600, 2010.

[35] Y. Li, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, and H. V. Jagadish. Regular expression learning for information extraction. In *In Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 21–30, 2008.

[36] C. D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[37] H. Matsuzawa and T. Fukuda. Mining structured association patterns from databases. In *Proceedings of 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2000)*, pp. 233–244, 2000.

[38] P. Mazzoleni, S. Goh, R. Goodwin, M. Bhandar, S. K. Chen, J. Lee, V. S. Sinha, S. Mani, D. Mukherjee, B. Srivastava, P. Dhoolia, E. Fein, and N Razinkov. Consultant assistant: a tool for collaborative requirements gathering and business process documentation. In *Proceedings of ACM SIGPAN*, pp. 807–808, 2009.

[39] G. Mishne, D. Carmel, R. Hoory, A. Roytman, and A. Soffer. Automatic analysis of call-center conversations. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pp. 453–459, 2005.

[40] A. Murakami and T. Nasukawa. Term aggregation: mining synonymous expressions using personal stylistic variations. In *Proceedings of the 20th international conference on Computational Linguistics*, pp. 77–82, 2004.

[41] K. Murakami, E. Nichols, J. Mizuno, Y. Watanabe, S. Masuda, H. Goto, M. Ohki, C. Sao, S. Matsuyoshi, K. Inui, and Y. Matsumotp. Statement map: Reducing web information credibility noise through opinion classification. In *Proceedings of the*

Fourth Workshop on Analytics for Noisy Unstructured Text Data (AND 2010), pp. 59–66, 2010.

[42] T. Nanno, T. Fujiki, Y. Suzuki, and M. Okumura. Automatically collecting, monitoring, and mining japanese weblogs. In *Proceedings of the 13th International World Wide Web Conference*, pp. 320–321, 2004.

[43] T. Nasukawa and T. Nagano. Text analysis and knowledge mining system. *IBM Systems Journal*, Vol. 40, No. 4, pp. 967–984, 2001.

[44] R. L. Oliver. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, Vol. 17, No. 3, pp. 460–469, 1980.

[45] D. Padmanabhan and K. Kummamuru. Mining conversational text for procedures with applications in contact centers. *International Journal on Document Analysis and Recognition*, Vol. 10, No. 3-4, pp. 227–238, 2007.

[46] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrievel*, Vol. 2, No. 1-2, pp. 1–135, 2008.

[47] A. Parasuraman, A. Zeithaml, and V. A. Berry. SERVQUAL: A multiple-item scale for measuring customer perceptions of service quality. *Journal of Retailing*, Vol. 64, No. 1, pp. 12–40, 1988.

[48] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*, pp. 1–6, 2000.

[49] UIMA Component Repository. http://uima.lti.cs.cmu.edu/ucr/.

[50] Information Retrieval and Extraction Exercise(IREX). http://nlp.cs.nyu.edu/irex/.

[51] S. Roy and L. V. Subramaniam. Automatic generation of domain models for call centers from noisy transcriptions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (COLING/ACL)*, pp. 737–744, 2006.

[52] A. Rzhetsky, I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P. A. Duboué, W. Weng, and W. J. Wilbur. Geneways: a system for extracting, ana-

lyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, Vol. 37, No. 1, pp. 43–53, 2004.

[53] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pp. 851–860, 2010.

[54] P. Sawyer, P. Rayson, and K. Cosh. Shallow knowledge as an aid to deep understanding in early phase requirements engineering. *IEEE Transactions on Software Engineering*, Vol. 31, No. 11, pp. 969–981, 2005.

[55] A. Seufert and J. Schiefer. Enhanced business intelligence - supporting business processes with real-time business analytics. In *IEEE 16th International Workshop on Database and Expert Systems Applications (DEXA)*, pp. 919–925, 2005.

[56] A. Simitsis, A. Baid, Y. Sismanis, and B. Reinwald. Multidimensional content exploration. In *Proceedings of the Very Large Data Bases (VLDB) Conference*, pp. 660–671, 2008.

[57] T. Simons. Speech patterns and the concept of utility in cognitive maps: The case of integrative bargaining. *The Academy of Management Journal*, Vol. 36, No. 1, pp. 139–156, 1993.

[58] A. Sinha, A. Paradkar, P. Kumanan, and B. Boguraev. A linguistic analysis engine for natural language use case description and its application to dependability analysis in industrial use cases. In *Proceedings of IEEE/ACM DSN*, pp. 327–336, 2009.

[59] M. Sokolova, V. Nastase, and S. Szpakowicz. The telling tail: Signals of success in electronic negotiation texts. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pp. 257–264, 2008.

[60] P. Srinivasan. Text mining: Generating hypotheses from medline. *Journal of the American Society for Information Science and Technology*, Vol. 55, No. 2, pp. 396–413, 2004.

[61] D. R. Swanson and N. R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, Vol. 91, No. 2, pp. 183–203, 1997.

[62] H. Takeuchi, T. Nakamura, and T. Yamaguchi. Predicate argument stracture analysis for use case description modeling. *IEICE Transactions on Information and System*, Vol. E95-D, No. 7, pp. 1959–1968, 2012.

[63] H. Takeuchi, S. Ogino, H. Watanabe, and Y. Shirata. Context-based text mining for insishts in long documents. In *7th International Conference Practical Aspacets of Knowledge Management (PAKM 2008)*, pp. 123–134, 2008.

[64] H. Takeuchi, L. V. Subramaniam, T. Nasukawa, and S. Roy. Automatic identification of important segments and expressions for mining of bussiness-oriented conversations at contact centers. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 458–467, 2007.

[65] L. Tanabe, U. Scherf, L. H. Smith, J. K. Lee, L. Hunter, and J. N. Weinstein. Prebind and textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, Vol. 4, No. 11, 2003.

[66] M. Tang, B. Pellom, and K. Hacioglu. Call-type classification and unsupervised training for the call center domain. In *Proceesings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 204–208, 2003.

[67] Apache UIMA. http://uima.apache.org/.

[68] N. Uramoto, H. Matsuzawa, T. Nagano, A. Murakami, H Takeuchi, and K. Takeda. A text-mining system for knowledge discovery from biomedical documents. *IBM Systems Journal*, Vol. 43, No. 3, pp. 516–533, 2004.

[69] M. A. Walker, I. Langkilde-Geary, H. W. Hastie, J. Wright, and A. Gorin. Automatically training a problematic dialogue predictor for a spoken dialogue system. *Journal of Artificial Intelligence Research*, Vol. 16, pp. 393–319, 2002.

[70] WEKA. http://www.cs.waikato.ac.nz/ml/weka/.

[71] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 42–49, 1999.

[72] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML)*, pp. 412–420, 1997.

[73] K. Zechner and A. Waibel. Minimizing word error rate in textual summaries of spoken language. In *Proceedings of 1st North American chapter of the Association for Computational Lunguistics (NAACL 2000)*, pp. 186–193, 2000.

[74] V. Zeithaml, M. J. Bitner, and D. Gremier. *Service Marketing: Integrating Customer Focus Across the Firm*. Mcgraw-Hill, 2009.

[75] G. Zweig, O. Shiohan, G. Saon, B. Ramabhadran, D. Povey, L. Mangu, and B. Kingsbury. Automatic analysis of call-center conversations. In *Proceedings of IEEE Internatinal Conference of Acoustics, Speech and Signal Processing (ICASSP)*, pp. 589–592, 2006.

[76] 元田浩, 津本周作, 山口高平, 沼尾正行. データマイニングの基礎. 共立出版, 2006.

[77] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, pp. 201–241, 2006.

[78] 竹内広宜, 杉山喜昭, 太田千景, 山口高平. マーケティングミックスとテキストマイニングを用いた市場分析支援. 第 24 回人工知能学会全国大会予稿集. 3B3-01, 2010.

[79] 竹内広宜, 中村大賀, 山口高平. テキスト分析技術を用いたユースケース分析. 信学技法 KBSE2010-32, pp. 55–60, 2010.

[80] 竹内広宜, 金山博, 武田浩一, 渡辺日出雄. UIMA（非構造情報処理アーキテクチャー）. 人工知能学会誌, Vol. 22, No. 6, pp. 808–813, 2007.

[81] 山田寛康, 工藤拓, 松本裕治. Support vector machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, 2002.

[82] 安部潤一郎, 藤野亮一, 下薗真一, 有村博紀, 有川節夫. テキストデータからの高速データマイニング. 人工知能学会誌, Vol. 15, No. 4, pp. 618–628, 2000.

[83] 矢田勝俊. スーパーマーケットにおける顧客動線分析と文字列解析. 統計数理, Vol. 56, No. 2, pp. 199–213, 2008.

[84] 寺田昭, 吉田稔, 中川裕志. 同義語辞書作成支援システム. 自然言語処理, Vol. 15, No. 2, pp. 39–58, 2008.

[85] 堀聡, 瀧寛和, 鷲尾隆, 元田浩. データマイニングを用いた市場品質監視システム. 電気学会論文誌, Vol. 121-C, No. 8, pp. 1289–1295, 2001.

[86] 那須川哲哉. コールセンターにおけるテキストマイニング. 人工知能学会誌, Vol. 16, No. 2, pp. 219–225, 2001.

[87] 那須川哲哉. テキストマイニングを使う技術/作る技術－基礎技術と適用事例から導く本質と活用法. 東京電機大学出版局, 2006.

[88] 那須川哲哉, 宅間大介, 竹内広宜, 荻野紫穂. コールセンターにおける会話マイニング. 言語処理学会第 13 回年次大会予稿集, pp. 590–593, 2007.

[89] 長尾 真編. 自然言語処理. 岩波書店, 1996.

[90] 大田朋子. 自然言語処理技術に基づく意味構造を利用した情報検索と情報抽出. 情報管理, Vol. 49, No. 10, pp. 555–563, 2007.

[91] 具本瑛, 中條武志. 魅力的品質・当たり前品質を中心とする消費者品質要求のモデル化. 品質, Vol. 31, No. 4, pp. 105–118, 2001.

[92] 市村由美, 酢山明弘, 櫻井茂明, 折原良平. 知識辞書構築支援ツールの開発. 情報処理学会 自然言語処理研究会, Vol. 143, No. 4, pp. 25–31, 2001.

[93] 西山莉紗, 竹内広宜, 渡辺日出雄, 那須川哲哉. 新技術が持つ特長に注目した技術調査支援ツール. 人工知能学会論文誌, Vol. 24, No. 6, pp. 201–241, 2009.

学位論文に関連する論文および口頭発表

学会誌論文

1. Hironori Takeuchi, Taiga Nakamura, Takahira Yamaguchi, “Predicate Argument Structure Analysis for Use Case Description Modeling,” IEICE Transactions on Information and Systems, Vol. E95-D, No. 7, pp. 1959–1968, 2012

2. 竹内広宜, 杉山喜昭, 山口高平, “市場分析におけるテキストマイニングを活用したデータマイニングの実践 - 生ごみ処理機の市場分析を例として –,” 日本知能情報ファジイ学会誌 (知能と情報), Vol. 24, No. 3, pp. 728–741, 2012

3. Hironori Takeuchi, L Venkata Subramaniam, Tetsuya Nasukawa, Shourya Roy, “Getting insights from the voices of customers: Conversation mining at a contact center,” Information Sciences, Vol. 179, No. 11, pp. 1584-1591, 2009

4. 竹内広宜, 那須川哲哉, 渡辺日出雄, “コールセンターにおける目的をもったビジネス会話のマイニング,” 人工知能学会論文誌, Vol. 23, No. 6, pp. 384–391, 2008

国際会議論文

1. Hironori Takeuchi, L Venkata Subramaniam, Tetsuya Nasukawa, Shourya Roy, Sreeram Balakrishan, “A Conversation-Mining System for Gathering Insights to Improve Agent Productivity”, Proceedings of IEEE Joint Conference on E-Commerce Technology (CEC ’07) and Enterprise Computing, E-Commerce and E-Services (EEE ’07), pp. 465–468, 2007

2. Hironori Takeuchi, L Venkata Subramaniam, Tetsuya Nasukawa, Shourya Roy, “Automatic Identification of Important Segments and Expressions for Mining of Business-Oriented Conversations at Contact Centers”, Proceedings of the 2007 Joint

Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 458–467, 2007

国内学会発表

1. 竹内広宜, 杉山喜昭, 大田千景, 山口高平, “マーケティングミックスとテキストマイニングを用いた市場分析支援,” 第 24 回 人工知能学会全国大会論文集, 3A3-01, 2010
2. 竹内広宜, 那須川哲哉, 渡辺日出雄, “コールセンターにおける目的を持ったビジネス会話のモデリングと会話マイニングへの応用,” 人工知能学会言語・音声理解と対話処理研究会 第 52 回研究会, SIG-SLUD-A703, pp. 15-20, 2008
3. 竹内広宜, 那須川哲哉, 渡辺日出雄, “会話データを対象にした有効な分析視点の設定と対象概念の自動抽出,” 第 21 回 人工知能学会全国大会論文集, 2H4-4, 2007

謝辞

本研究は、著者が慶應義塾大学大学院理工学研究科後期博士課程在学中に、同大学理工学部 山口高平 教授の指導のもとに行ったものです．博士論文を執筆するにあたり，多くの方々から多大なるご指導およびご助言を賜りました．まず，本研究を行う契機と環境を与えて下さり，研究の全過程を通じて，常に温かく適切な御指導を頂いた，山口高平 教授に心から感謝いたします．博士論文の副査を快諾していただき，本論文の執筆にあたり，有益な御助言および御指導を頂いた，櫻井彰人 教授，鈴木秀男 教授，萩原将文 教授に厚く御礼申し上げます．

山口研究室 森田武史 助教（現在，青山学院大学），玉川奨 氏には研究室の環境整備などで多大なる支援を頂きました．ここに感謝いたします．山口研究室 杉山嘉昭 氏には 4 章の研究を進める上で分析実験などで協力していただきました．ここに感謝いたします．研究活動全般に渡って，援助を下さった山口研究室の兼田浩明 氏，中村尚広 氏をはじめ学生諸氏に感謝いたします．

日本アイ・ビー・エム株式会社 東京基礎研究所の鎌田真由美 氏，中村大賀 氏には研究と仕事を進める際，多くの助言や支援をいただきました．ここに感謝します．また，東京基礎研究所の那須川哲哉 氏，インド研究所の L Venkata Subramaniam 氏，Shourya Roy 氏 (現在，Xerox 社) には 3 章の研究を進める上で有益なコメントを頂きました．ここに感謝します．そして，研究活動を進めていく上で様々な支援をいただいた同僚の研究員諸氏に感謝いたします．

最後に，常に精神的に支えてくれた妻と 2 人の娘，そして両親へ感謝の言葉を送りたいと思います．

2012 年 8 月

竹内 広宜