知能と複雑系　132－8
(2003. 3. 13)

動的に変化するコールセンター情報からの重要情報の特徴発見

Experimental Study of Discovering Essential Information from Customer Inquiry with Dynamic Variation

嶋津恵子† 桜井哲志† 門馬敦仁† 山根洋平† 古川康一‡
†富士ゼロックス(株)研究本部 IT メディア研究所
‡慶應義塾大学大学院政策・メディア研究科

This paper reports the results of our experimental study on a new method of applying an association rule miner to discover useful information from inquiry database. It has been claimed that association rule mining is not suited for text mining. To overcome this problem, we propose (1) to generate sequential data set of words with dependency structure from the text database, and (2) to employ a new method for extracting meaningful association rules by applying a new rule selection criterion based on difference between prior and posterior confidences, instead of minimum confidence. This criterion comes from the fact that we put heavier weights to those phenomena with co-occurrence of plural items more than those with single occurrence. Using this method, we succeeded in extracting useful information from the text database, which were not acquired by only simple keywords retrieval.

1　はじめに

最近データマイニングの一研究領域であるテキストマイニングの研究が注目されている[6]。これは Web 技術の浸透に伴い、一般の利用者が扱うことのできる文書量が急増していることが背景となっている[8, 14]。

本論文では、嶋津ら[12]と同様、コールセンターの蓄積データからの重要情報の抽出を狙っている。現在どの企業にとっても、コールセンターは市場や顧客との最も重要な接点である。担当員は、あらゆるリクエストに対し満足度を保ちながら応答するノウハウを持ち、一方、膨大な応答履歴には顧客傾向や市場動向を読み取る情報が潜んでいると考えられる。ところが、問い合わせ記録を再利用する際、予め与えられた情報分類尺度や慣例的に用いられているキーワードに頼っているため、思わぬ発見を見落としていることが多い。また、顧客の生の声を聞くことを専らとしている専門家と、大局的・中長期的観点から経営を司る専門家では、日常接する情報の種類が大きく異なる。従って用意された分類尺度がそのまま利用できないことも多い。これらにより、コールセンターの情報をビジネスチャンスに活かしたり、リスクを事前に回避することに活用できていない。これが、コールセンターを electronic Customer Relation Management (eCRM)のボトルネックと称する所以である。我々は今回の実験で、コールセンターの情報を対象とし、専門家によるキーワード(およびその組み合わせ)の指定や、それらの頻出傾向では獲得できない重要な情報を発見することを目指した。

嶋津ら[12]は、テキストデータから意味ある情報クラスを特定することを目指し、(1)出現する語句に動詞に関する係り受け情報を付与したものをアイテムとし、(2)事前確信度と事後確信度による相関ルールの絞込みをおこなった。これにより、(出現パタンが)頻出はしないが、意味を持つ情報クラスの獲得に成功した。

今回我々は、出力された出現パタンから意味のあるものを獲得する効率を上げるために、テキストデータを動詞だけでなくすべての係り受け構造を付与した語句から成るアイテム集合に整形した。さらに、各アイテム集合から問い合わせ本文と同じ意味を取るのに必要なものだけを選択し、意味の取れる並びを生成した。これらをテキストマイニング用の対象データとし、(i)標準的な相関ルール導出アルゴリズム[1]と、(ii)事前確信度と事後確信度の差を用いるルール絞込み手法[12]と、(iii)例外ルール発見手法[13]をそれぞれ適用し、結果を比較した。

以下に、本論文の構成を示す。2 章では、今回我々が行ったテキストデータからの重要情報の発見のフレームワークを示す。3 章にコールセンターのデータを対象とした実験結果を述べ、4 章でそれに対する考察を与える。5 章にまとめを記す。

2　テキストデータからの意味ある情報を発見するフレームワーク

今回の実験は、嶋津ら[12]のそれの延長と位置付けられる。その全体の流れを図 1 に示す。すなわち、前処理として、各問い合わせデータの問い合わせ文を浅い文法情報を付与した語句からなる系列データに変換し、それらを対象としてテ

キストマイニングをおこなった。

2. 1 テキストデータから意味を特定できる系列データへの整形

2.1.1 テキストデータの分かち書きと係り受け情報付与

各問い合わせ文に対し、事前に開発した専用の辞書に照らし、分かち書き処理をおこなった[1]。このとき、係り受け情報を付与した構造をもつ系列データとして生成した(図2ステップ1)。嶋津ら[12]は、長野ら[11]と小林[7]のコールセンター情報を対象としたテキストマイニングの報告(動詞より名詞が重要、文末表現から意図が読み取れる)に注目し、名詞とそれが係る文末の用言からなる係り受け情報を付与した系列データを生成した。ところが、例えば図 3 に示した問い合わせ文があった場合、従来の手法では系列データから正確な意味を判断することができない。

そこで今回の実験では、すべての係り受け情報を付与した(図 3 最下段)。これにより、嶋津ら[12]で成功した解釈のあいまい性の除去だけでなく、より正確な意味特定が可能になる。[2]

2.1.2 意味の特定可能な系列データの抽出

次に、各アイテム集合から、問い合わせ本文と同じ意味を取るのに必要なものだけを選択し、意味の取れる並びのパタンを生成した。そして、アイテムの並びから内容のクラス分けをおこない結論部とした(図2ステップ2)。これは、嶋津[12]が発見した、系列データを参照することで本文の意味を特定できる性質を利用したものである。例えば、図2において"(こと 使用する),(当社 OS 環境), (変更する こと) → ClassN"のような表現も考えられるが、この場合は前提部から意味が読み取れないので、この段階で排除した。生成した情報クラスは、表1に示すような構成を示している。例えば Class22 に属する問い合わせは、操作方法や機能仕様に関する質問である。問い合わせの傾向分析の際には、それぞれのクラス傾向を見るだけでなく、クラスを列・行でまとめ、例えば Class20 と指定することで"質問"に所属するすべての問い合わせを参照することも可能になる。

2. 2 系列データからの重要パタンの発見

2.2.1 支持度と確信度の閾値によるデータ全体の傾向把握

系列データを対象に相関ルールを導出する分析手法は、一般に、対象データの大多数に対して当てはまるルールを発見することに用いられてきた。特に WebUsage マイニングでは、多くの利用者が何を目的として Web サイトにアクセスしているかの把握に用いられている[4 他に多数]。このとき Agrawal[1]が提案している最小支持度(minimum support)と最小確信度(minimum confidence)を満たすルールの導出手法が利用される。我々は、このアルゴリズムを採用して、テキストデータ全体の傾向を把握することを試みた。

2.2.2 事前確信度と事後確信度の差による重要情報の獲得

テキストマイニングに関する多くの報告によると、特に内容に注目した場合、頻出する語句の重要性が高いとは限らない[大澤 99 他]。そこで我々は、アイテムが単独で存在したときと、別のそれと共起したときの確信度の差が大きいものほど重要性が高いとし、相関ルールを絞り込んだ。これは松尾ら[10]の提案を単純化した方法であり、Apriori4.03[3]では、すでにシステム化されている。これは、従来の相関ルールの導出(2.2.1 節)だけでなく、別の評価法の導入により、より意味のあるルールの抽出を可能としたことが特徴である。これは確信度を、事前確信度(Prior Confidence)と事後確信度(Posterior Confidence)の 2 種類に分けて、算出することに基礎を置いている。具体的には、生成された各相関ルールの結論部に対して前提部が空の場合の確信度を事前確信度として計算し、この値と従来の相関ルールの確信度(事後確信度)との差を利用

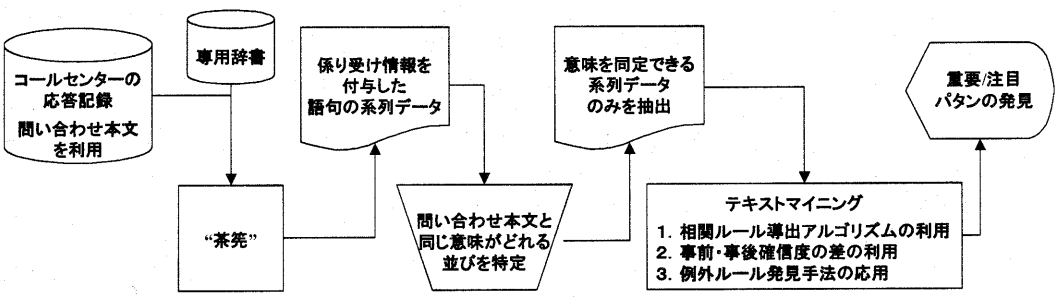

図1 テキストデータからの意味ある情報を発見するフレームワーク

[1]文法情報を付与した分かち書き処理には、茶筅[9]を利用した。
[2]分かち書き処理時に助動詞の意味(要望, 疑問, 否定)を利用し[2]、より正確な意味判断を可能にした。

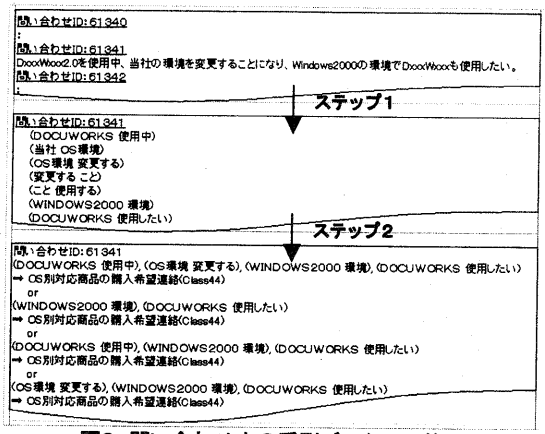

図2 問い合わせ文の系列データへの整形

問い合わせ文
遅いプリンターと早いプリンターを比較したい
↓
従来手法による系列データ化 (プリンター, 比較する), (プリンター, 比較する)
↓
今回の実験採用した手法による系列データ化 (遅い, プリンター), (早い プリンター), (プリンター, 比較する) [要望]

最後尾の"[要望]"は茶筅の助動詞解釈によるもの[1,2]

図3 文章の系列データ化(従来手法との比較)

	大項目クラス	Class01	Class02	Class03	Class04	
大項目クラス	問い合わせ目的 ＼ 問い合わせ対象	購入前の情報入手購入方法	操作方法機能仕様	性能	OS別対応	
Class10	苦情・不満・不安	Class11	Class12	Class13	Class14	
Class20	質問	Class21	Class22	Class23	Class24	
Class30	感謝・おほめ	Class31	Class32	Class33	Class34	

表1 問い合わせ文のクラス分け(部分抜粋)

する。例えば、{cheese, tomato} ⇒{bread}というルールが生成された場合、{ }⇒{bread}の確信度が事前確信度、{cheese, tomato} ⇒{bread}のそれが事後確信度となる。

我々は、語の共起に意味があるという前提に立っている。そこで、事前確信度と事後確信度の差の大きいルールを抽出することで、多くのデータに共通した傾向とは異なるが、意味ある(有用性の高い)ものが獲得できると考えた。

2.2.3 デフォルト規則を使った例外ルールの発見

データマイニングのエンジンとしてもしばしば採用される機械学習のアルゴリズムは、対象とする各データを正例と負例に明示的に区別し、その特徴を抽出する。一方、実世界にはどちらに属するかわからない例が多く存在する。井上ら[5]は、この問題に対し拡張論理プログラミングの形式を用いて、不完全な情報を扱える新しい学習方法を提案している。これにより例外を含むデフォルト規則を学習することが可能である。また、鈴木[13]は、デフォルト規則を支持度と確信度の高いルールとして捉え、例外ルールを同時に発見する手法を提案している。具体的には、Y→X がデフォルトルールとして獲得された場合、関係ルールZ/→X'を特定し、Y, Z→X'を例外ルールとして導出する。ここで、X'はXと属性は同じだが属性値が異なるアトムであり、/→は出現する前提部だけでは結論部が説明できないことを示す。

我々はこの手法によって獲得される例外ルールは、2.2.2 節の手法による獲得ルールの中に既に含まれていると考え、これを検証する実験をおこなった。

3 コールセンターの情報を対象にした実験

3. 1 実験対象テキストデータ

今回の実験の対象として、2002 年 4 月 1 日から同年 7 月 31 日までの特定の商品に関するデータ(総数は 626 件)の問い合わせ本文を採用した。また、同年 8 月 1 日から 10 月 30 日までの同じ商品に関するデータ(総数 725 件)に対しても同じ実験をおこない、傾向の変化を観察した。

3. 2 係り受け情報を付与した意味解釈可能な系列データへの整形

専用辞書を参照し、問い合わせ文をすべての係り受け情報を付与した語句からなる系列データに変換した。この段階で、一つの問い合わせあたりに含まれるアイテム(係り受け関係を持つ 2 語の組)数は平均 14.9 個であった。これに対し、さらに原文と同じ意味が取れるアイテムの並びの整形すると、1つの問い合わせ本文は平均 7.1個のアイテムで構成された。また、総語句数は 9598、語句の種類は 1950、異なるアイテム数は 8157 であった。

一方、嶋津ら[12]の実験では動詞にかかる係り受けのみを付与し系列データを生成したが、このときの一つの問い合わせあたりに含まれるアイテム(係り受け関係を持つ 2 語の組)数は約7.5 個であった。

3. 3 系列データの特徴パタン獲得

3.3.1 対象データの頻出傾向

前節で生成した系列データを対象に、最小支持度 0.6、最小確信度 40 で相関ルールを導出した。表 2 に、これらを支持度の高い順に上位 20 件を示す。11 件が操作方法・機能仕様に関するもの(Class02)であり、7 件は質問(Class20)である。操作方法・機能仕様に関するもので、かつ質問であるものは 4 件ある。操作方法・機能仕様に関するもので質問でないものが 1 件あり、苦情であった(Class12)。また、質問のうち操作方法・機能仕様では無いものは 1 件であり、社内体制・仕組みに関するもの(Class26)であった。残りは、購入前の情報入手に関するものが2件、無償ダウンロードに関するものが 2 件、苦情に関するものが 1 件であった。さらに、各ルールが該当する問い合わせ数は 11 件から 5 件であり、平均 6.7 件である。

3.3.2 出現頻度に依存しない意味ある系列パタン

頻出状況に依存せず、アイテムの共起に重要性のあるルールを獲得するために、事前確信度と事後確信度の差が 30 以上あるもの(最小支持度は 0.02)を抽出した。表 3 にこれらを確信度差の大きい順に上位 20 件を示す。7 件が操作方法・機能仕様に関するもの(Class02)であり、うち 2 件がその質問(Class22)であった。購入前の情報入手・購入方法に関するルールが3件あり、前節のルール抽出では獲得できなかったOS別対応(Class04)に該当する問い合わせのルールと、性能に関する質問(Class03)のそれを獲得している。前節で獲得できているものと重複するルールが4件あった。またアイテム中の語のいずれかが、3.3.1 で獲得したルールにも出現しているものは 9 件である。さらに各ルールが該当する問い合わせ数は 10 件から 2 件であり、平均 4.6 件である。

					支持度	確信度	該当数
⟨操作, 聞く[要望]⟩,			⇒	質問	1.76	91.67	11
⟨ファイル, 開く⟩,	⟨メール, 添付する⟩		⇒	操作方法・機能仕様	1.60	100.00	10
⟨Ver4.1, 利用する⟩			⇒	操作方法・機能仕様	1.60	90.91	10
⟨操作, 聞く[要望]⟩,			⇒	操作方法・機能仕様に関する質問	1.60	83.33	10
⟨メール, 添付する⟩			⇒	操作方法・機能仕様	1.44	100.00	9
⟨方法, おしえる⟩,			⇒	操作方法・機能仕様に関する質問	1.28	100.00	8
⟨ソフト商品XYZ, 試用版⟩,			⇒	購入前の情報入手・購入方法	1.28	57.14	7
⟨スキャナ, 使う⟩,	⟨X社, スキャナ⟩		⇒	操作方法・機能仕様	1.12	87.50	7
⟨ソフト商品XYZ, 試用版⟩			⇒	質問	1.12	50.00	8
⟨ソフト商品XYZ, 試用版⟩			⇒	ダウンロードサイト	1.12	50.00	7
⟨ソフト商品XYZ, 試用版⟩,	⟨試用版, 利用する⟩		⇒	ダウンロードサイト	1.12	50.00	7
⟨試用版, ダウンロードする⟩,			⇒	苦情・不満・不安	0.96	75.00	6
⟨ソフト商品XYZ, カタログ⟩			⇒	購入前の情報入手・購入方法	0.96	75.00	5
⟨サポート窓口, おしえる⟩			⇒	社内体制・仕組みに関する質問	0.84	100.00	4
⟨こと, できる⟩			⇒	操作方法・機能仕様	0.80	83.33	5
⟨ソフト商品XYZ, 取り込む⟩			⇒	操作方法・機能仕様	0.80	83.33	5
⟨Ver4.1, 使う⟩			⇒	操作方法・機能仕様	0.80	71.43	5
⟨この, ソフト⟩			⇒	社内体制・仕組み	0.80	62.50	5
⟨ファイル, 開く⟩,	⟨メール, 添付する⟩		⇒	操作方法・機能仕様に関する質問	0.80	50.00	5
⟨Ver4.1, 利用する⟩			⇒	苦情・不満・不安	0.80	45.45	5

"該当数"は、ルールが該当する問い合わせ件数

表2 対象データの頻出傾向

					確信度差	確信度	該当数
⟨どちら, ダウンロードする⟩			⇒	ダウンロードサイトに関する質問	96.63	100.00	3
⟨パソコン, OS⟩			⇒	OS別対応	95.28	100.00	3
⟨サポート窓口, おしえる⟩			⇒	社内体制・仕組みに関する質問	95.01	100.00	4
⟨HTTP, ある⟩,	⟨FTP, ある⟩,	⟨HTTP, FTP⟩	⇒	ダウンロードサイト	91.51	100.00	4
⟨FTP, ダウンロードする⟩,	⟨HTTP, ダウンロードする⟩,	⟨どちら, ダウンロードする⟩	⇒	操作方法・機能仕様	91.51	100.00	3
⟨電話番号, おしえる⟩			⇒	社内体制・仕組み	89.35	100.00	3
⟨ソフト商品XYZ, カタログ⟩			⇒	購入前の情報入手・購入方法	79.38	100.00	5
⟨ライセンス, 購入する[要望]⟩			⇒	購入前の情報入手・購入方法	79.38	100.00	3
⟨制限, ある[疑問]⟩			⇒	性能に関する質問	73.79	75.00	3
⟨HP, 見る⟩			⇒	ダウンロードサイトを操作中	71.50	75.00	3
⟨OS, する⟩			⇒	OS別対応	70.28	75.00	3
⟨ソフト商品XYZ, 購入する[要望]⟩			⇒	購入方法に関する質問	65.16	75.00	3
⟨操作, 聞く[要望]⟩			⇒	操作方法・機能仕様に関する質問	61.23	83.33	10
⟨使い方, おしえる[要望]⟩			⇒	操作方法・機能仕様に関する質問	52.90	75.00	3
⟨ソフト商品XYZ, Ver3.02⟩			⇒	現使用操作機能の連絡	52.90	75.00	3
⟨ファイル, 開く⟩,	⟨メール, 添付する⟩		⇒	操作方法・機能仕様	46.23	100.00	10
⟨試用版, ダウンロードする⟩,	⟨ソフト商品XYZ, 試用版⟩		⇒	操作方法・機能仕様	41.51	50.00	7
⟨試用版, 利用する⟩,	⟨ソフト商品XYZ, 試用版⟩		⇒	操作方法・機能仕様	41.51	50.00	7
⟨X社, スキャナ⟩	⟨スキャナ, 使う⟩,	⟨Ver10.2, 使う⟩	⇒	苦情・不満・不安	49.60	71.43	2
⟨スキャナ, 使う⟩,	⟨X社, スキャナ⟩		⇒	操作方法・機能仕様	33.73	87.50	7

表3 出現頻度に依存しない意味ある系列パタン

					確信度	該当数
⟨X社, スキャナ⟩	⟨スキャナ, 使う⟩,	⟨Ver10.2, 使う⟩	⇒	苦情・不満・不安	71.43	2
⟨これ, 使う⟩,	⟨こと, できる⟩		⇒	操作方法・機能仕様	87.50	3

表4 例外パタン

3.3.3 例外パタン

3.3.1 節で求めた頻出傾向をデフォルトルールとし、結論部の属性の値(分類クラス)を替え、関係ルールを順に探した。そして、関係ルールを条件部に足すことで確信度が高くなる例外ルールを獲得した。その結果表4に示すような2件の例外ルールを獲得した。1つめのルールは、2.3.2で既に獲得したものであった。それぞれのルール獲得に用いた関係ルールの確信度は、それぞれ45.45、26.08であった。

3.3.4 8月から10月のデータを対象にした追実験

上述と同様の実験を、同年8月から10月のデータ(問い合わせ総数725件)に対し実施した。高支持度と高確信度によって出現パタンを絞り込んだ全体の傾向把握では、"(Ver5.0, 発売開始) ⇒ 購入前の情報入手・購入方法に関する質問"が確信度89.9%、該当する問い合わせ件数13件が特徴的であった。

また、事前・事後確信度差による系列パタン抽出では、"(Ver5.0, 購入), (検索, 可能[疑問]) ⇒ 購入前の情報入手・購入方法" (該当する問い合わせ件数3件)がある。一方、例外ルールは発見されなかった。

4 考察

4.1 前処理(係り受け情報を付与と意味の取れる系列データの選択)の効用

嶋津ら[12]では、動詞に関する係りうけ情報を付与し、実験対象データを生成した。このとき意味ある相関ルールは、出力総件数10333件数741件であった。つまり有用なルールは7%程度である。これに対し今回の実験では、すべての係り受け情報を付与した語句の並びに整形した後、問い合わせ本文の意味がとれるものを選択し、テキストマイニングの対象にしている。これにより、獲得した出現パタンのルールの中で、例えば"(こと, できる) ⇒ 操作方法・機能仕様"のように利用性が無いと判断されたものは、5%であった。また比較実験用に、係り受け情報をすべて付与し、一方意味解釈できるものの抽出処理を行わないデータセットを用意した。これを対する実験では、出力結果380件のルールのうち75件だけが意味を解釈でき利用性が認められた。これらのことから、今回採用した係り受け情報を付与し、さらに意味の取れる系列データだけを選択する前処理は、有効なルールを抽出するのに大きく貢献したと言える。

4.2 意味ある情報の獲得

高支持度と高確信度で導出した結果の結論部を見ると、操作方法・機能仕様に関するルールが半数を占めている。そして、購入前の情報入手・購入方法に関するものと、(インターネットのホームページ上の)ダウンロードサイトに関するものと、社内体制・仕組みに関するものと、苦情・不満・不安に関するものが2件づつである。これが問い合わせ全体のおおまかな傾向だとすると、担当者が作成した月度報告と一致する。つまり、これらの事実は、テキストマイニングを利用するまでも無く、従来の問い合わせ記録データベースの検索機能を用いることで、確認可能である。一方、操作方法・機能仕様に関するルールに、該当製品で作成したファイルをメール添付した場合の利用方法((ファイル, 開く), (メール, 添付する) ⇒ 操作方法・機能仕様)に関するものが多いこと、また一年前に発売した該当製品の新版(Ver10.2)に関する操作方法・機能仕様に関する問い合わせや苦情が発生していることは、今回の実験で明らかになった。つまり、大量の記録に埋もれ見落とされがちなキーワードを獲得することが可能になっていた。

さらに、事前・事後確信度の差を利用してルールを導出した結果では、インターネット上のホームページからのダウンロードに関するものが20件中6件発生しており、特にプロトコル選択に関し迷っていることが確認できた。これは通常の集計時(月度報告作成等)には把握されなかった。また、例外ルール発見手法[13]でも獲得された、特定のスキャナを使用した場合に苦情となるルール("(X社, スキャナ), (スキャナ, 使う), (Ver10.2, 使う) ⇒ 苦情・不満・不安")は、専門家が見逃している問い合わせ傾向であった。

前述した頻出傾向の考察と同様に、一般に大量のテキストデータから注目すべきものを抽出する場合、事前にヒントとなるキーワードが事前に提供されることは少ない。従って、特に出現頻度が低いものに関しては記憶も薄れ、検索されにくくなる。これに対し、今回我々が採用した手法は、問い合わせ件数の頻出状況に依存せず、注目に値する傾向を発見するのに有益であると言える。

4.3 例外ルールの有用性

今回の実験で発見された例外ルールのうち、一つは我々の提案する事前・事後確信度差を用いる方法でも獲得された。また残りの一つは有益性が無く、さらに8月以降のデータでは発見に至らなかった。

例外ルールの発見手法では、事前・事後確信度差による特徴パタン獲得手法における閾値と同様のものを用いている。鈴木ら[13]が応用事例としてあげている髄膜脳炎データからの発見では、最小支持度20%、最小確信度75%を

満たすものをデフォルトルールとして採用している。また例外ルール生成に用いられる関連ルールの確信度は50%以下であり、支持度 3.6%、確信度 80%を満たすものを例外ルールとして特定する。これに対し、我々の今回の実験では、デフォルトルールに相当する高支持度・高確信度による頻出傾向パタンの特定は、この閾値より低い(緩い)ものを用いた。また例外ルールして採用可能な出現パタンを獲得する際にも、事前・事後確信度差を 50%以下に設定した。鈴木ら[13]が応用事例と同じ閾値では該当するルールが出力されなかったためである。この違いは、データの特徴が影響していると考えられる。

我々が取り上げたコールセンターの問い合わせデータは、鈴木ら[13]のそれと異なり、同一の状態を複数の異なる表現で表されていることが多く、異なったアイテムとして処理される。例えば"(Ver10.2, 購入する), (xxxx, yyyy) ⇒ ClassN" と"(Ver10.2, 買う), (xxxx, yyyy) ⇒ ClassN" は、別の系列パタンとして導出されてしまう。このように同じ意味のルールが分散することで、支持度・確信度とも低くなる傾向がある。

これは、テキストマイニング対象データの特徴(総語句数9598、語句の種類1950、異なるアイテム数8157)が原因であり、前処理用の辞書(シソーラス)の精度向上が必要である。これにより、高い閾値を用いることが可能になり有用な傾向が把握できるとも考えられる。しかし"言葉"を扱う以上、表記の揺れを完全に吸収することは不可能である。また、医療データでは、回復するか死に至るかのような絶対的な結論部を想定し、注目すべき出現パタンの発見をおこなう。一方、コールセンターの情報は、記録内容の傾向が推移するにつれ、重要性がダイナミックに変化する。このようにデフォルトルールそのものが変化する対象には、例外ルール発見方法より我々の提案する手法の方が、見落とし無く注目情報を獲得できると考えられる。

5 まとめ

我々は、相関ルール導出アルゴリズムをテキストマイニングに応用し、意味ある情報を特定することを試みた。この際、①出現する語句に係り受け情報を付与し系列データ化したことと、②事前確信度と事後確信度による相関ルールの絞込み手法を採用したことが特徴である。これにより、頻出はしないが、意味のある出現パタンを獲得することに成功した。また、非単調推論に基づく例外ルール発見手法との比較を試みたが、この方法では有効な注目傾向を獲得することが困難であった。これはテキストを対象とした場合、絞込みが強すぎることが原因である。一方、我々の提案した手法(事前・事後の確信度差の利用)では、絞込みにある程度の緩やかさを持たせ、有益なルールを獲得できた。

また、我々はテキストマイニングにおける前処理として、係りうけ情報を付与した系列データに整形する手法が有効であることを 2 つの理由(①元データの意味を損なわないことと、②従来のデータマイニングの手法の利用が可能なこと)から示した。特に理由の①は、専門家によって提示された推測[12]の裏づけとなり、今後コールセンター担当者が問い合わせ内容を整理・報告する際の工数削減にも貢献すると期待できる。一方、Zaki[15]は、前処理に語の出現順序を考慮すると意味を損なわずにテキストマイニングが可能であると主張している。我々は出現順序もルールの導出に考慮すると、分散してしまい傾向を獲得するのが困難になると懸念している。この点に関し、さらに検討が必要である。

また、今回の実験では、リスクの事前開始になる可能性のあるルール(特定のメーカーのスキャナとソフトウエア商品の特定の版との相性で発生する問題)を獲得したが、追実験用データからはこの出現パタンは見られない。このことから、事前予知や予測として利用できるものを特定するにはさらなる分析手法の開発が必要だと考えている。

参考文献

[1] Agrawal R.: Fast Algorithms for Data Mining Applications, Proc. of the 20th International Conference on Very Large Databases, pp.487-489, Santiago Chile (1994)

[2] 浅原正幸, 松本裕治: IPADIC ユーザーマニュアル, version 2.5.1, 30 January 2002 (2002)

[3] Borgel, C.: http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori/

[4] R. Cooley, T. Pang-Ning and R. Cooley: Discovery of Interesting Usage Patterns from Web Data, WERKDD'99, LNAI 1836, pp.163-182 (2000)

[5] 井上克巳, 工藤嘉晃, 羽根田博正：デフォルト規則を含む拡張論理プログラミングの学習, 人工知能学会誌, Vol.14, No.3, pp.437-445 (1999)

[6] 特集「テキストマイニング」, 人工知能学会誌, Vol.16, No.2 (2001)

[7] 小林竜己: 文末表現と内容語に着目した問い合わせメール分析, The 16th Annual Conference of Japanese Society of Artificial Intelligence, 1E4-02 (2002)

[8] Lawrence S. and Giles L.: Searching the World Wide Web, Science, Vol.280, No.5360, pp.98-100 (1998)

[9] 松本裕治: 形態素解析システム「茶筅」, 情報処理, Vol.41, No.11, pp.1208-1214 (2000)

[10] 松尾豊, 石塚満: 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会誌, Vol.17, No.3, pp.217-223 (2002)

[11] 長野徹, 武田浩一, 那須川哲哉: テキストマイニングのための情報抽出, 情報学基礎, 60-5, pp.31-38 (2000)

[12] 嶋津恵子, 山根洋平, 古川康一: コールセンター情報からの重要情報の発見, 人工知能学会研究会, SIG-FAI-A203, pp.43-48 (2003)

[13] 鈴木英之進：共通データからの仮説駆動型例外ルール発見, 人工知能学会誌, Vol.15, No.9, pp.782-789 (2000)

[14] 津田宏: 特集「テキストマイニング」にあたって, 人工知能学会誌, Vol.16, No.2, pp.191 (2001)

[15] Zaki, M.: Efficiently Mining Frequent Trees in a Forest, In Proc. SIGMOD 2002, ACM (2002)