1F4-03

相関ルール導出法によるコールセンター情報からの重要情報の発見

Discovering Knowledge with Association Rules from Customer Inquiry

嶋津恵子 *1
Keiko Shimazu

門馬敦仁 *1
Atsuhito Momma

古川康一*2
Koichi Furukawa

*1 富士ゼロックス(株)
研究本部
Fuji Xerox Co., Ltd.
Corporate Research Group

*1 富士ゼロックス(株)
研究本部
Fuji Xerox Co., Ltd.
Corporate Research Group

*2 慶應義塾大学
大学院政策･メディア研究科
KEIO University
Graduate School of Media and Governance

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をされている[人工知能学会誌 01]。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。

一方データマイニングの研究成果で、既に実用化されている技術に相関ルール(association rule)の導出手法[Agrawal 94]が、あるが、このアルゴリズムをテキストマイニングに応用し有効な結果が得られた報告は少ない。これは対象となるデータが半構造形式をとるため、非均質で多様かつ膨大なデータの集積となっていることが原因である[有村 02]。

今回我々は、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、重要な情報の獲得を試みた。本実験の特徴は、①前処理としてテキストデータを係り受け情報を考慮した系列データに変換したことと、②テキストマイニング･エンジンである相関ルールの絞り込みに、これまでに報告されていないルールの絞込み手法を用いた点である。

また、実験対象としてコールセンターの蓄積情報を採用し、キーワードによる検索や既存のテキスト分類技術では発掘することが困難な重要情報を見つけることを目標にした。

2. 重要情報発見フレームワーク

今回の実験で用いた重要情報発見のフレームワークは図 1 の通りである。つまり、テキスト情報の特徴を反映した前処理部と相関ルールの導出に新手法を採用したテキストマイニング部から成る。

2.1 テキストデータの分かち書きと係り受け情報による系列化

我々は、嶋津ら[嶋津 02]の課題(①意味あるルールの割合の増加と②データ全体に対する網羅性の改善)の達成を目指し、出現語句に文法上の係り受け構造情報を付与した系列データを作成し、それを対象データとした[嶋津 03]。

連絡先：富士ゼロックス(株)研究本部，〒259-0157 神奈川県足柄上郡中井町境 430 グリーンテクなかい，Tel: 0465-80-2321, atsuhito.momma@fujixerox.co.jp

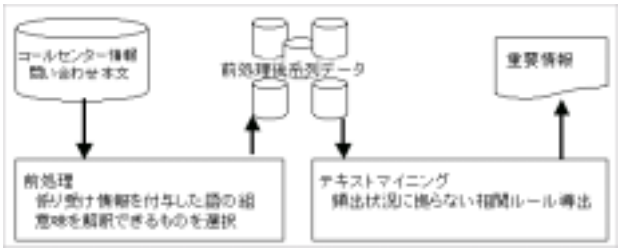

図1 コールセンター情報を対象にした重要情報発見のフレームワーク

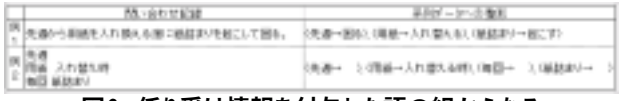

図2 係り受け情報を付与した語の組からなる系列データ化

図 3 意味のとれる系列データの選択

一般に文章の解析結果は 2 種類の木構造を用いて表現される[長尾 96]。一つは、英文の構造解析の研究成果である出現する句の文法関係を表すものであり、もう一つは係り受け(依存)関係を表すものである。前者は、正確な文法に基づいて文が作成されているときに有用である。一方、日本語は本来語順が比較的自由で、また要素の省略が可能である。特に今回対象としたコールセンターでは、メモ程度の文法を無視した形で記録を取られることもあり、これらの傾向が顕著に現われる。ところで後者の係り受け関係は語の並びや記号(矢印など)を用いた記録からも読み取ることが可能である。そこで我々は、問い合わせ記録を係り受け関係にある 2 語の組みをアイテムとする系列データに変換する手法を採用した。

例えば図 2 に示したように、文で記録した(i)と、語の並びで記録した(ii)は、それぞれ図中に示したように系列データに整形できる。これらの用い、パタンマッチングさせることで、同じ意図の問い合わせであるか否かを機械的に把握できる。つまり、この前処理手法を用いることにより、同一内容で表現方法が異なる記録が無数に考えられる場合、解釈の曖昧性を除去し、より明確かつ断定的に意味を決定することが可能になる。

さらに、我々は前処理の段階で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。これは、嶋津[嶋津 02]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。図3において“(こと 使用する),(当社 OS 環境), (変更すること) → ＯS別対応商品の購入希望”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。

これは、より多くの意味あるルールが導出されることを目指したものである。

2.2 頻出傾向に拠らない相関ルールの導出

これまで、従来の方法での相関ルールの導出は、対象データの大多数に対して当てはまる規則性を発見することに用いられてきた。このとき、Agrawal[Agrawal 94]が提案している最小支持度(minimum support)と最小確信度(minimum confidence)を満たすルールのみを出力する方法が利用される。

一方テキストマイニングでは、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。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であるとし、相関ルールを絞り込んだ。これは、松尾ら[松尾 02]の提案を単純化した方法である。

具体的には、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、この差の大きいものだけ抽出する。事前確信度は、生成された相関ルールの前提部を空に置き換えたルールの確信度であり、事後確信度は従来のそれである。[1]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、アイテムの共起に意味があるという前提に立ち、両確信度の差が一定の閾値をこえるものを抽出した。

[1] Apriori4.03[Borgel 02]では、すでにシステムの機能として実装されている。

表1 出現頻度に依存しない意味あるルール

表2 例外ルール
1 行目の結果は我々の手法でも出力済み(表1)

3. コールセンター情報からの重要情報獲得実験.

3.1 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

問い合わせ文を 2.1 に従い、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に整形すると、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[嶋津 02]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数は約7.5 個であった。

3.2 出現頻度に依存しない意味ある重要情報

頻出状況に依存せず重要な情報を発見するために、アイテムの共起に注目し、事前確信度と事後確信度の差が 30 以上ある相関ルールを抽出した(最小支持度は 0.02)。表1は確信度差の大きい順に上位 20 件をみたものであるが、7 件が操作方法･機能仕様に関するものであり、うち 2 件が特に質問(明らかに正答を求めているもの)であった。購入前の情報入手･購入方法に関するものが 3 件あり、OS 別対応に関するもの、性能に関する質問であった。特にこれら 3 件は、相関ルールの導出に従来の高支持度･高確信度を用いた(それぞれ 0.6、40)時には出力されなかった。

3.3 デフォルト規則を使った例外ルール発見手法との比較

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[井上 99]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[鈴木 00]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、高支持度･高確信度の閾値を用いて導出したデータ全体中の頻出傾向を示す相関ルールをデフォルトルールとし、我々の手法と出力結果を比較した。具体的には、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。その結果は表2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、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。

3.4 8月から10月のデータを対象にした追実験

3.2 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 (係り受け情報の付与と意味の取れる系列データの選択) の効用

嶋津ら[嶋津 02]では、動詞に関する係り受け情報を付与し、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取れるものを選択し、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。これに対する実験では、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 テキストマイニング手法 (事前･事後確信度差を利用した相関ルールの絞込み) による重要情報の獲得

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、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[鈴木 00]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[鈴木 00]の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[鈴木 00]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) ⇒ zzzz に関する質問” と”(Ver10.2, 買う), (xxxx, yyyy) ⇒ zzzz に関する質問”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コールセンタ情報に対する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、”文章全体を読むより、重要語の並びを参照したほうが、直感的に重要情報を発見できる可能性がある”[嶋津 02]、の裏づけとなり、今後コールセンタ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[Zaki 02]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、リスクの事前回避に繋が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

[Agrawal 94] Agrawal R.: Fast Algorithms for Data Mining Applications, Proc. of the 20th International Conference on Very Large Databases, pp.487-489, Santiago Chile (1994)

[Borgel 02] Borgel, C. : http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori/

[井上 99] 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)

[人工知能学会誌 01] 特集「テキストマイニング」, 人工知能学会誌, Vol.16, No.2 (2001)

[松尾 02] 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)

[嶋津 02] 嶋津恵子, 山根洋平, 門馬敦仁, 桜井哲志, 古川康一: テキストデータの内容に基づく相関ルールのクラスタリング実験, 人工知能学会研究会, SIG-FAI-A202, pp.55-62 (2002)

[嶋津 03] 嶋津恵子, 山根洋平,古川康一: コールセンタ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)

[鈴木 00] 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)

[Zaki 02] Zaki, M.: Efficiently Mining Frequent Trees in a Forest, In Proc. SIGMOD 2002, ACM (2002)