相関ルールの活用によるコールセンター業務への貢献

嶋津恵子† 山根洋平† 門馬敦仁† 桜井哲志† 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

1 はじめに

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として捉えられている。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、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ている。

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。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)の指定では選択されない重要な情報を発見することを試みた。

2 係り受けを考慮した系列データの生成と相関ルールの効果的絞込み

2.1 問い合わせ文の分かち書きと係り受け情報による系列化

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、英文の構文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(図 1 参照)。我々は、問い合わせ記録が日本語のメモレベルの表記であることに注目し、語と係り受け関係用いた系列データに変換する手法を採用した。

この手法により、同一内容や意図の記録でも表現方法のパタンが無数考えられる場合、解釈のあいまい性を除去し、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。

2.2 相関ルールの導出

作成された系列データを対象に相関ルールを導出する。この手法は、これまで対象データの大多数に対して当て嵌るルールを発見することに用いられてきた。このとき、一般的には最小指示度(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性があるとし、系列データから制せ去れた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り、意味ある(有用性の高い)ものが導出される。

2.3 Apriori4.03

Apriori[Agrawal 94]は、現在最も引用されている相関ルール導出用アルゴリズムの一つであり、Apriori4.03[Borgel 02]はその改良版である。このシステムは、従来どおりの指示度と確信度の高い相関ルールの導出をするだけでなく、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く。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}の確信度を事前確信度、{cheese, tomato} ⇒{bread}のそれを事後確信度とする。

我々は、語句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した。

3 コールセンターの情報を対象にした実験

3.1 対象データと前処理

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するデータ(総数は 602 件)の問い合わせ本文を採用した。これらに対し、語句を細かく分けすぎることを避けるために(“Windows2000”を一つの語句としてとりたい)、イント

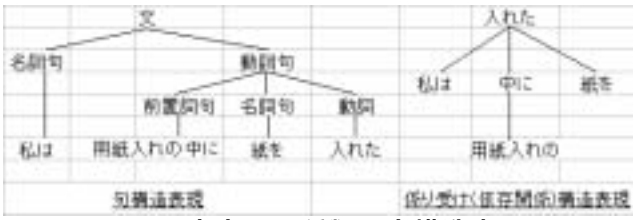

図 1 文章の 2 種類の木構造表現

ラネット上の IT 用語リストと商品リストを辞書として参照させた上で、茶筅[松本 00]を利用して分かち書き処理をおこなった。さらに語句の共起リスト上に現れてもほとんど意味のとれない(もしくは必要ない)、助詞や指示代名詞を削除した。

3．2 係り受け情報を付与した系列データへの整形

出力された語句リストの動詞に注目し、これに係る名詞との組み表現を用いて系列データに整形した。例えば、”拡大コピーを取る度に、いつも紙詰まりが発生する”は、3.1 節の前処理で”拡大コピー，取る，紙詰まり，発生する”となり、さらに本節の処理で”(拡大コピー → 取る), (紙詰まり → 発生する)”となる。

これらの結果作成されたソースデータは、一つの問い合わせあたりに含まれる平均語句数は約 12.5 個、総語句数は 7517、語句の種類は 1772 となった。2つの語句による異なる係り受け数は、3116 であった。

3．3 相関ルール抽出によるクラス生成実験

前節で生成した系列データを対象に、Apriori4.03 による相関ルールの導出をおこなった。ルールの絞込み基準はソースデータ中に頻出するものではなく、最小指示度 0.33 かつ事前確信度と事後確信度の最小差 15%を満たすものとした。

①出力相関ルール：数と構成

10333 件の相関ルールが出力された。アイテム数 2 のルール 7608 件、以下同様に3のもの 2633 件、4 のもの 107 件であった。すべての相関ルールを構成するアイテム(係り受け情報による語句の組)は、23543 個であり、異なり数は 2496 種類であった。これらの相関ルールを人手により意味が取れるものを特定すると 751 件であり、内訳はアイテム数 2 のもの475件、3 のもの 221 件、4 のもの 65 件であった。

②生成された情報クラスの有用性

出力された相関ルールを、意味が同じであるものをまとめ 1 クラスとして分類した。その結果、23 のクラスが生成された。このうち、内容から有用性が高いと判断できるものを特に有用性”A”として 11 クラスを特定した。この判断基準は、”現稼動中の記録システムで属性名の選択やキーワードの組み合わせ指定では(困難を伴わず)同じ結果を入手できない”ものである。表 1 は、これら 11 クラスの概観を示している。問い合わせデータのうち 54 件がこれらのクラスに分類され、11 件が複数に所属した。また、係り受け情報を付与した相関ルールから解釈した各クラスの意味を、同表の 1 列目(「所属する情報の意味」)に示した。オペレーティング･システムの違いによって発生する問い合わせが半分近くを占めている。

③生成された情報クラスの網羅性

意味の解釈できる相関ルール 751 件は、全問い合わせデータ 602 件(3.1 節)の 430 から生成されていた。つまり全データの約 71%を網羅していることになる。

3．4 係り受け情報を利用した系列データの有効性検証

今回の実験で用いたものと同じデータを対象に、係り受け情報を付与しない系列データを用い同様の実験をおこなった[嶋津 02]。同じく 10000 件のルール導出を目指したところ、最小指示度 0.53、事前確信度と事後確信度の最小差 28%で、11054 件を出力した。この中で、意味を解釈できるものは 127 件であり、これらは全問い合わせのうちの 170 件から抽出されたものであった。さらに、これらの相関ルールから生成された情報クラスは 21 種類であった。このうち有用性が”A”であるものは、9クラスであった(表 2)。

3．5 事前確信度と事後確信度の差を用いた絞込みの有効性検証絞込みの有効性検証

今回の実験対象と同じ係り受け情報を付与した系列データを用い、最小指示度と最小確信度で絞り込む実験をおこなった。4.2 説と同様に最小指示度を 0.33 に設定し、最小確信度を 1 から 0.05 まで下げ、抽出できるルールの数を確認した。最小確信度 0.05 で、抽出ルールは 306 件であり、意味を解釈できるものは 12 件、さらに有用性が”A”であるものは 4 件であった。さらにこれらの値は、最小確信度の値よる大きな変化は見られなかった(表3参照)。

3．6 生成された情報クラスの精度･再現率

今回の実験で生成した有用性”A”であるクラスに含まれるデータの、全データに対する精度と再現率を確認した(表 5)。例えば、クラス C02 に含まれるものは係り受け情報を付与した相関ルールから内容を解釈すると”バージョンアップ版の購入に関するもの”であり、該当するルールで特定される(ルール生成に用いた)元データは 6 件である(表 1)。これらの問い合わせ内容を(相関ルールではなく)本文そのものから解釈したとき、クラス C02 に含まれるものではない(正解データではない)と専門家が判断した場合、精度は下がる。一方、相関ルール出力に用いられなかった文書の中に、これと同じ内容のものがあれば再現率が下がる。この考え方に基づく、

クラスID	所属する情報の意味	所属する元データ列ID
C01	ファイルダウンロードのHTTPとFTPの違い	62099, 62267, **67005**, 67445, 68759, 69524, 71775
C02	バージョンアップ版購入	61162, 61553, 66041, 67960, **72295**, 72363
C03	ライセンスの購入	**61499**, 63755, **73974**
C05	異なるバージョン同士の互換性	61553, 62095, 62198, 63541, 63556, **64390**, 66295, **66568**, 66618, 67960, 70163, 70487, 73170
C06	WindowsXP上での現保有版の動作	**64262**, 66813, 66828, **67904**, 69961, **72295**
C07	操作方法に関する問い合わせ窓口	63064, 63071, 63556, **64262**, 66665, 67842, **67904**, 68542, 69268, 70216, 72822, 74483, 75595, **75992**
C08	インストール方法	63701, 64267, 66026, 66210, 66353, 66947, **71674**, 72083, 72878, 74560, 74565, **75992**
C10	体験版の入手	66295, **71674**
C12	ファイルのダウンロード	**61499**, **66568**, **73974**
C13	Windows2000上での現保有版の動作	61973, 63413, **64390**, 65529, 65563, 66025, 66188, 69686, **76064**
C14	WindowsMAC上での現保有版の動作	66155, 66385, **67005**, **76064**

下線付き太字で記載されているものは複数のクラスに所属するデータ

表1　有用性が高いクラスの意味と所属するデータ

	最小指示度	事前/事後確信度差	導出ルール数	意味が解釈できるルール数	導出全ルール数に対する意味解釈できるルール数の割合	意味が解釈できるルール抽出に用いた元データ件数	意味別分類クラス数	有用性"A"クラス数	有用性"A"であるクラスに属する平均元データ数
係り受け情報を付与しない系列データ	0.53	28%	11054	127	0.011	170	21	9	18.6
係り受け情報を付与した系列データ	0.33	15%	10333	741	0.072	430	23	10	6.4

表2　係り受け情報の付与による効果

事前確信度と事後確信度の差	出力ルール数（累計）	意味を解釈できるルール数（累計）	有用性"A"であるルール数（累計）
10	10348	741	59
15	10333	741	59
20	8976	600	53
25	7936	519	43
30	7839	519	43
35	6004	374	36
40	5987	374	36
45	5886	374	36
50	109	10	5
55	109	10	5
60	105	10	5
65	93	10	5
70	74	9	5
75	69	9	5
80	69	9	5
85	69	9	5
90	69	9	5
95	65	9	5

表3　事前･事後確信度の操作による獲得ルール数の変化

最小確信度	出力ルール数（累計）	意味を解釈できるルール数（累計）	有用性"A"であるルール数（累計）
0.05	306	12	4
0.10	289	12	4
0.15	249	12	4
0.20	227	12	4
0.25	199	11	4
0.30	194	11	4
0.35	169	11	4
0.40	150	10	4
0.45	134	10	4
0.50	131	10	4
0.55	109	9	4
0.60	109	9	4
0.65	107	9	4
0.70	107	9	4
0.75	74	8	4
0.80	69	8	4
0.85	69	8	4
0.90	69	8	4
0.95	69	8	4
1.00	69	8	4

表4　従来の相関ルール絞込み(高指示度･確信度)による獲得ルール数の変化

クラス別精度と再現率を求めた(表 5)。精度はクラスC12 を除きいずれも高いが、再現率はC01 とC06 を除いて低い値を示している。

4　考察

4.1 既存の手法では獲得できない意味ある情報クラスの獲得率と網羅性の向上

今回の実験で、我々は、問い合わせ登録時に振られた属性の選択や本文に対するキーワード検索では獲得できない有用性の高い情報クラス(有用性"A")を 11 件生成した。このことから、問い合わせ本文を参照することなく語句の系列データを用いることで内容を把握することがある程度可能であると言える。また意味の解釈できるルール 751 件は、元データの 71%を使って生成されたものであり、係り受け情報の付与をおこなわなかった場合の 18%を大きく上回る。対象データ全体から網羅的に、有用な(意味の解釈できる)ルールを抽出できるように改善されたと言える。

テキスト情報の分類手法の代表として用いられるものに TF/IDF 値の利用がある[Salton 83]。これは重要性が高い語ほどその文書内に頻発に出現し、抽象度が高い語ほど多くの文書に

クラスID	C01	C02	C03	C05	C06	C07	C08	C10	C12	C13	C14
精度	100%	100%	100%	83%	100%	92%	100%	89%	33%	100%	100%
再現率	88%	30%	18%	38%	70%	35%	7%	67%	7%	11%	29%

表 5　　情報クラスの精度と再現率

出現するという考えを基にしている。この手法を用いた時の課題は、①出現する語が文書の特徴を示さない場合の対処と、②意味的に複数のクラスに所属するデータへの対応(TF/IDF では排他的に分類)であった[嶋津 02]。今回の実験では、キーワードによる検索が不可能な有用性”A”である情報クラスを 11 生成でき、またデータを複数のクラスに所属させることに成功した。このことから、TF/IDF 値による手法の課題を解決し、実用上利便性の高い、内容依存型の分類に近づいたと言える。

一方、相関ルールから生成されたクラスに含まれるデータ(ルール生成の基データ)の再現率が、低い。これは同じ意味を持つデータでも、係り受け情報を付与した語句の系列データだけでは拾いきれないことを示している。今回は、那須川ら[那須川 01]の提案を採用し、係り受け構造のうち動詞のみに注目した。今後、形容詞等も採用することで、精密なルール表現が可能になり再現率を改善させることができると考えられる。

4．2 係り受け情報を付与した系列データの効用

我々は係り受け情報に注目し、テキストデータを語句とこれを付与した系列データに整形する手法を採用した。係り受け情報を付与しないデータと比較し、意味を解釈することのできるルールの割合(対出力全ルール)が約 6 倍に増えている(表 2)。有効なルールの獲得効率に貢献したと言える。

また 4.1 節で報告した”多数存在はしないが重要であるような情報の獲得に成功した”理由は、基本的には、相関ルールの絞込みの際に事前確信度と事後確信度の差を採用したことによると考えられる[嶋津 02]。ただし、同じ絞込み方法を採用しても、係り受け情報を付与しなかった実験では有用なクラスに含まれるデータ数が平均約 18 件である(表２)。ことから、係り受け情報の付与も、出現数は少なくとも意味ある情報を特定するという効用をより高めるに貢献したと言える。

4．3 事前･事後確信度を利用した相関ルールの絞込みの効用

我々は、相関ルールの絞込みに従来の支持度と確信度の高いものを選択する方法でなく、事前確信度と事後確信度の差の大きいものを選択する方法を採用した。この結果、10000 件のルール出力を目指し、700 件以上の意味を解釈できるルールを獲得した。ところが、同じ系列データを用い、従来の手法で絞り込むと、計算爆発をおこさない程度まで確信度を下げていっても全出力ルール数は 300 件ほどであり、そのうち意味を解釈できるルールは 12 件であった。これは、テキストマイニングには従来の相関ルールの絞込み手法が不向きである[有村 02]ことを裏付け、一方、我々が採用した手法が効果的に作用することを示すものである。

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、意味ある情報のクラス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長である。これにより、頻出はしないが、意味を持つ情報クラスを獲得することに成功した。また、従来手法に対し、内容に依存した複数のクラスへの分類という課題への解決も図った。

係り受け情報を付与しない同じデータを対象とした先行実験では、出力されたすべての相関ルールから意味あるものだけ出力する際の膨大な手作業が課題として報告されていた。これに対し、係り受け情報を付与することで、大きく改善することを示した。今回は、動詞に注目した係り受けのみ付与したが、小林[小林 02]と同様に文末情報全体に注目し形容詞等の係り受けも採用することで、より改善効果が高まると予想される。

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)

[有村 02] 有村博紀, テキストマイニング:ウェブデータからの知識発見を目指して,(招待論文) , 第25回情報化学討論会概要集, J13, 日本科学会情報科学部会 (2002)

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

[松本 00] 松本裕治: 形態素解析システム「茶筅」, 情報処理, 41-11, pp.1208-1214 (2000)

[那須川 01] 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)

[Salton 83] Salton G. and McgGill M.: Introduction to Modern Information Retrieval, McGraw-Hill Book Company (1983)

[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)