The 24th Symposium on Information Theory
And Its Applications (SITA2001)
Kobe,Hyogo,Japan,Dec.4 - 7,2001

大規模データベースにおける定型項目と記述項目を含むデータからの相関ルール抽出法

Generation Method of Association Rules for Combined Categorical Data and Text Data on Very Large Databases

熊谷貴禎* 小林学* 平澤茂一*
Kiyoshi Kumagai Manabu Kobayashi Shigeichi Hirasawa

Abstract Recently, very large databases are easily obtained, and it gets much attention to efficiently use these databases. We consider generating association rules on databases with a viewpoint of data mining. The Apriori algorithm that generates association rules from Categorical Data is widely known.

In this paper, we propose new methods for extending this algorithm to generate association rules from both Categorical Data and Text Data. We show that the proposed methods generate more reliable rules using entropy which is well known in information theory.

Keywords Datamining, Association Rule, Text Data, Apriori Algorithm

1 はじめに

近年企業間の競争は激化を極め，マーケティング戦略も多様化し，柔軟でかつ独自の意思決定に生き残りを左右される時代を迎えている．このため，情報化社会の発展により作られた非常に大規模なデータベースを有効に活用しようという要請がある．例えば CRM(Customer Relationship Management) ではコンビニエンスストアの POS データから，購買パターンの発見や顧客のセグメント化などが行われる．

そのための手法の一つとしてデータマイニング (Data Mining : DM) では，相関ルールの抽出方法が盛んに研究されている[2]．データマイニングとは，大量のデータから表面上には現れない，隠れた規則性や関連性を人工知能の技術を応用して発見しようというものである．また，その対象とするデータ形式は，関係データベースが扱うような定型項目のデータ(Categorical Data)からテキスト形式である記述項目のデータ(Text Data)にまで広がってきている．

一方，アイテム間の共起情報を抽出するための相関ルール抽出手法として，アプリオリアルゴリズム (Apriori Algorithm : 以降アプリオリと呼ぶ) [1]が提案されており，これを応用してテキストデータ内の単語から相関ルールを抽出する手法が提案されている[3][4]．しかし，アンケートデータなど，定型項目と記述項目が混在するようなデータからの相関ルール抽出は困難を要する．

本研究では，定型項目と記述項目の混在するようなデータに対して，従来のアプリオリにエントロピーを用いて興味深い相関ルールを抽出する方法を提案する．また提案手法の評価として，実データのアンケートを用いたシミュレーションを行い，有効性を明らかにする．

2 準備

2.1 本研究の対象分野

本稿では，アンケートデータのような，定型項目と記述項目が混在するデータを対象とする．表 1 のようなアンケートデータにおいては，作成者があらかじめ用意した定型項目よりも，被験者が自由に記述できる記述項目の方がニーズを的確に反映していることも多い．

このようなデータにおいて，記述項目のテキストを単語毎にアイテムとして扱い，定型・記述の両項目から，アイテム間の共起情報の抽出を目的とする．

表 1 アンケートデータの例

住所	定型項目 1 : {東京都}
年齢	定型項目 2 : {20 歳}
性別	定型項目 3 : {男}
職業	定型項目 4 : {大学生}
商品の感想	記述項目 1 : {ご飯の量が少ない}
御意見	記述項目 2 : {夕方に品切れが多い}

2.2 相関ルール抽出

2.2.1 相関ルールの定義[2]

相関ルールとは X , Y をアイテムの部分集合として，各トランザクションにおいて条件 X が成立する時に条件 Y が高い確率で成立する規則を言う．POS データを例にとった場合，レシートがトランザクション，商品がアイテムである．

[定義]相関ルール

全アイテム集合を U と表し $X,Y \subset U$, $X \cap Y = \phi$ なる X , Y に，$X \rightarrow Y$ を相関ルールと定義する．

(例) X ={男,大学生}→Y ={ご飯, 少ない}

相関ルールの抽出は，全トランザクションに対して X と Y を含むトランザクションがどのくらいの割合を占めるかを表す「支持度」(support) $P(X,Y)$, X を含むトランザクションに対して X と Y の両方を含むトランザクションがどのくらいの割合を占めるかを表す「確信度」(confidence) $P(Y|X)$ を用いて行う．

2.2.2 抽出アルゴリズム[1]

本研究では，相関ルールの抽出アルゴリズムとして，2 値属性を対象とした，広く引用されている逐次アルゴリズムであるアプリオリを用いる．

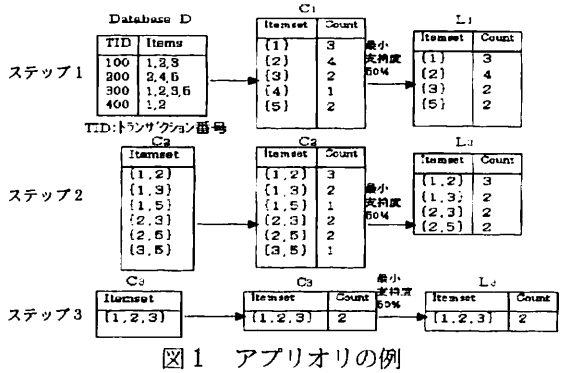

図 1 アプリオリの例

全アイテム集合 U の任意の部分集合 $Z \subset U$ をアイテム集合と呼ぶ．アプリオリでは，ステップ k において，候補となる k 個の要素を持つアイテム集合の集合を C_k とする．即ち，$C_k \subset \{Z \subset U \| Z| = k\}$, ただし $|Z|$ は Z の要素数を表す．この C_k についてデ

*連絡先：早稲田大学理工学部経営システム工学科,〒169-8555
東京都新宿区大久保 3-4-1,School of Science and Engineering, Waseda University,3-4-1 Ohkubo Shinjyuku-ku,Tokyo japan

ータベース上の全トランザクションを検索した結果，最小支持度以上の支持度を持つアイテム集合の集合をラージアイテム集合$L_k \subset C_k$とする．ここで，$Z \in C_k$に対しZから任意の1つの要素を取り除いた集合Z'が，$Z' \in L_{k-1}$を満たすようにC_kが作成される．

図1にアプリオリとC_k, L_kの例を示す．さらに，各ステップのラージアイテム集合について最小確信度を用いてルールの抽出を行う．

2.2.3 アプリオリの多値属性への適用

アプリオリでは，各トランザクションにアイテムが出現する(1)，出現しない(0)の2値でアイテムを扱う．このため，一般的な多値属性を扱うために寺渡らは，アプリオリを多値属性に適用できるように拡張した[5]．

ここでの多値属性は，内部に複数個の属性値を持つものとし，属性値のどれか1つに排他的に1が立つものを考えている．図2のように属性を「方角」としたとき，多値属性のそれぞれの属性値（例：東，西，南，北）を新たに属性とみなし，独立したアイテムと考える．それぞれには{0,1}の要素を割り当てる．

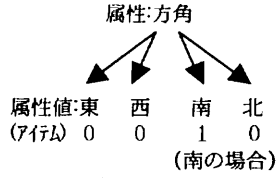

図2 多値属性の扱い

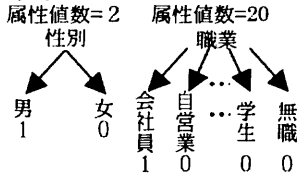

図3 属性値数の違い

2.2.4 多値属性の問題点

アプリオリに多値属性を適用した場合，図3のように属性値数が少ない属性域の出現頻度が高くなる．例えば図3において，「会社員」という属性値は「職業」という属性内では出現数が多く，有用と思われるにもかかわらず，「性別」より属性値数が多いためルールとして抽出されない傾向がある．これにより抽出されるルールの属性に偏りが生じる(属性値数が小さいものが多くなる)問題がある．

2.3 従来のテキストデータからの相関ルール抽出

テキストデータからの相関ルール抽出としては，コールセンターにおけるテキストマイニング[3]がある．ユーザが指定した定型項目(4項目)を固定して，記述項目の相関ルールを抽出するものである．しかし，定型項目数が非常に多い場合にユーザが予期せぬ隠れたルールの抽出には困難を要する．

例：{機種：T，CALL：セットアップ，問題：ソフト，解答：情報提供} → {WIN2000－導入できる？}

3 提案

3.1 定型項目に対する提案手法

本節では2.2.4節の問題点を踏まえ，ルール抽出の際に，アイテムの所属する属性の情報量を考慮する手法を提案する．

3.1.1 多値属性に対する修正最小支持度の導入

各アイテムを$a_1,\cdots,a_j,\cdots,a_m$，各属性を$A_1,\cdots,A_i,\cdots,A_n$，$A_i$の属性値数を$n(A_i)$とするとアイテムと属性の関係は図4のようになる．

従来はルール抽出の際に，この新たに作成したアイテムに対しアプリオリをそのまま適用した．しかし，アイテムにはそれぞれが所属する属性が存在し，これを考慮していくことで，ルール抽出時に生じる属性値数の大小による弊害を減少させることが可能である．

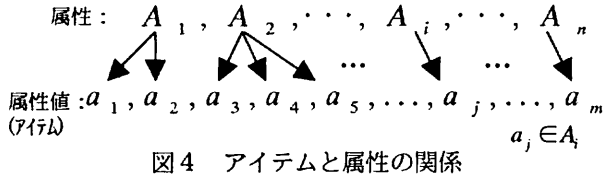

図4 アイテムと属性の関係

本研究では，アイテムが所属する属性内の属性値の出現頻度からエントロピーを計算し，値が高いほど複雑な属性(情報源)と考えルールの絞込み基準を調整することを考える．

例として，図5のように①と②のような属性内の分布を持つ2つの属性を考える．これを情報源と考えると，この2つの情報源のもつ「複雑さ」は異なる．ここではアイテムと属性値は同義である．

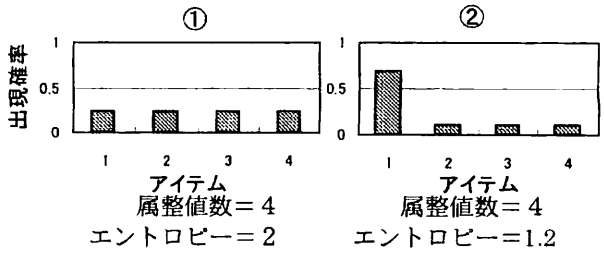

図5 情報源の複雑さ

＜エントロピー＞

属性A_iのエントロピー$H(A_i)$は式(1)で計算される．

$$H(A_i) = -\sum_{a_j \in A_i} P(a_j)\log_2 P(a_j) \quad \text{—(1)}$$

$P(a_j)$はアイテムa_jの出現確率の推定量で，アプリオリのステップ1の支持度と同一であるため，余分にデータベースの検索を必要としないで計算できる．

＜修正最小支持度＞

アイテムa_jの修正最小支持度を次式で定義する．

$$minsupp'_j = minsupp / 2^{H(A_i)}, \quad a_j \in A_i \quad \text{—(2)}$$

エントロピーは，アイテムの出現が一様分布に近づくほど，$\log_2 n(A_i)$に近づく．このことから式(2)の分母に$2^{H(A_i)}$を適用することが妥当と思われる．即ち，$1 \le 2^{H(A_i)} \le n(A_i)$である．アプリオリにおいて，この修正最小支持度を最小支持度に代えて用いる．つまり，修正最小支持度を閾値として，これを満たすアイテム集合を抽出していくことにする．

さらに，式(2)をアイテム集合に拡張していく．ステップkの候補アイテム集合$X \in C_k$の修正最小支持度$minsupp'_{k,X}$を次式で定義する．

$$minsupp'_{k,X} = \underset{j|a_j \in X}{Min}(minsupp_j) \quad \text{—(3)}$$

3.2 記述項目に対する提案手法

テキストデータから相関ルールを抽出するためにテキスト内の各単語をアイテムとして考える．すなわち記述項目iの単語全体をW_iとし，各記述項目を属性と考えて単語w_jをそのアイテムとする．さらに属性の修正最小支持度を用いて相関ルールを抽出する．

3.2.1 テキストデータから単語の抽出(前処理)

記述項目内のテキストデータから単語を抽出するために，前処理として以下のような処理を行う．

1) 形態素解析を行い，テキストを単語に分割する．
2) 同意語辞書を用い，単語内の同意語を統一する．

3) 複合語辞書を用い，単語内の複合語を抽出する．
4) 不要語辞書を用い，単語内の不要語を削除する．
5) 単語の出現を{0,1}のベクトル表現で表す(表2)．

この処理に用いた3種類の辞書は，出現する単語が対象とする分野によって大きく異なるため，各自で作成する必要がある．

表2 テキストデータのベクトル例

	記述項目1			…		記述項目2	…	
単語	a	b	c	…	z	A	…	Z
ID1	1	0	1	…	0	1	…	1
ID2	0	1	0	…	0	0	…	1
ID3	1	0	1	…	1	1	…	0
ID4	0	1	1	…	0	1	…	0
合計	2	2	3	…	1	3	…	2

3.2.2 記述項目に対する修正最小支持度の問題点

各記述項目を属性と考えて，単語をその属性値とする．また，定型項目と同様に修正最小支持度の概念を導入する．

記述項目に対するエントロピーの定義は定型項目と変える必要がある．なぜなら，定型項目は属性内の1つのアイテムにのみ1が立つが，表2のように記述項目は各単語をアイテムと考えるため多くの場合，属性内に1が複数立つからである．また，各単語の相関も考慮してエントロピーを計算する必要がある．

3.2.3 アプリオリによる単語のグルーピング(前処理)

3.2.2節の議論を踏まえて，同時に出現する単語をグルーピングして1つのアイテムと考える．このグルーピングを行うために以下の手順でアプリオリを用いる．表3にその例を示す．

① 各記述項目i毎に，アプリオリでグルーピング用の最小支持度を用いてラージアイテム集合$L_1 \sim L_{k_{max}}$を抽出する．ただし$kmax$は最大のステップ数とする．また$k = k\max$，記述項目iの単語グループの集合$T_i = \phi$にセットする．

② 抽出されたラージアイテム集合L_kの中で，支持度が最大のアイテム集合(単語の集合)$g_j \in L_k$を単語グループとして採用する．すなわち

$$g_j = \underset{Z \in L_k}{\operatorname{argmax}} \{P(Z)\} \quad \text{—(4)}$$

また$T_i = T_i \cup \{g_j\}$とする．

③ 採用された単語グループに含まれている単語を含むアイテム集合を削除する．すなわち

$$L_l = L_l \setminus \{Z \in L_l \mid Z \cap g_j \neq \phi\}, l = 1,2,\ldots,k \quad \text{—(5)}$$

④ もし$L_k \neq \phi$ならば②へ．

⑤ もし$k = 1$ならば終了する．そうでなければ$k = k - 1$として②へ．

アプリオリグルーピングによって採用されたグループを用いてエントロピーを次のように定義する．

記述項目iのエントロピー$H(T_i)$はグループg_jの出現確率の推定量$P(g_j)$を用いて式(6)で計算する．ただし$P(g_j)$はグループg_jの支持度と等しい．

$$H(T_i) = -\sum_{g_j \in T_i} \{P(g_j)\log_2 P(g_j) + (1 - P(g_j))\log_2(1 - P(g_j))\} \quad \text{—(6)}$$

$P(g_j)$はg_jに含まれる全ての単語が出現するトランザクションを有効解答数(記述解答が有効であったトランザクション数)で割ったものである．

このエントロピーの値は，各グループの出現確率の推定量がすべて0.5となるとき，グループ数に一致する．

表3 提案アプリオリグルーピング

L	記述項目(=支持度)			単語グループ
L1	{情報理論}=0.3 {人工知能}=0.1 {理工}=0.3 {経営}=0.1	{大学}=0.4 {学生}=0.3 {研究}=0.3	→	{情報理論} {人工知能}
L2	{理工,経営}=0.1 {理工,学生}=0.1 {経営,学生}=0.2 {大学,学生}=0.3 {学生,研究}=0.1 {大学,研究}=0.2		→	{理工,経営}
L3	{大学,学生,研究}=0.1 {理工,経営,学生}=0.05		→	{大学,学生,研究}

ここで各記述項目毎に定義したエントロピーをアプリオリに用いる．

3.2.4 記述項目の修正最小支持度

＜記述項目の修正最小支持度＞

単語$w_j \in W_i$の修正最小支持度を式(7)のように定義する．

$$minsupp'_j = minsupp / H(T_i) \quad , \quad w_j \in W_i \quad \text{—(7)}$$

記述項目においては，修正最小支持度算出式が定型項目の場合と異なる．これは，記述項目ではエントロピーの和を用いているためである．

さらに，式(3)を用いて，ステップkに拡張する．

4 提案手法の評価

4.1 相関ルールの評価方法

相関ルール抽出手法によって求められた相関ルールは，一般に大量である．この中から有用なルールを抽出するための基準を考える．

相関ルール$X \rightarrow Y$において，Yの出現確率の推定量$P(Y)$，Xを含むトランザクションの中でXとYの両方を含むトランザクションの割合(確信度)を$P(Y|X)$とすると確信度上昇率$confrate(X \rightarrow Y)$を式(8)で定義する．

$$confrate(X \rightarrow Y) = (P(Y|X) - P(Y)) \Big/ P(Y|X) \quad \text{—(8)}$$

確信度上昇率が高いほど，有用なルールと考えられる．これにより，確信度上昇率が0以上の相関ルールのみを抽出する．また，抽出された全相関ルールについて，確信度上昇率の平均をとったものを平均確信度上昇率とする．

4.2 実データによるシミュレーション

提案手法の評価は2種類の実データ(いずれも学生の大学生活意識調査である)を用いたシミュレーションによって行う．本研究で作成した辞書の各単語数は，同意語=374語，複合語=942語，不要語=120語となった．また従来手法として，記述項目に含まれる単語をアイテムとし，定型項目と同様に扱った場合を考える．

4.2.1 小規模アンケートデータ

予備実験として，研究室に関する意識調査のアンケートデータを用いる．トランザクション数=45，定型項目数=21，記述項目数=5．

提案手法と従来手法で最小支持度を変動させて，同一ラージアイテム数のもとで平均確信度上昇率を比較したものを図 6 に示す(提案手法にはグルーピング用の最小支持度を 0.03 で与えた)．また，最小確信度は 0 に固定した．

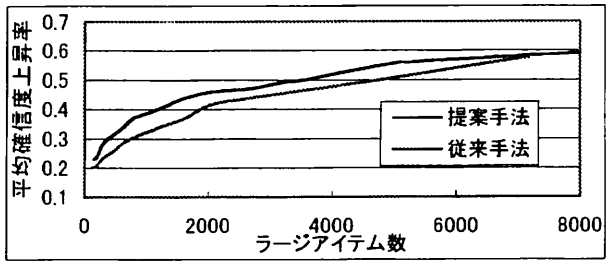

図6　小規模アンケートデータによる平均確信度上昇率

さらに抽出されたルールについて興味深いものをいくつか挙げる．提案手法と従来手法で抽出されるラージアイテム数を 2000 に揃えた時に抽出されたルール(1_は研究室を選んだ理由，2_は研究室で満足している点)である．

＜抽出された興味深いルール例＞

{ 1_教授,1_人柄 }→{ NET5 時間以下,進学希望あり}
＝[支持度:0.0889(4/45),確信度:1.0000(4/4),確信度上昇率:0.7778]

解釈：研究室を選んだ理由が教授の人柄の人は全体の中で 4 人いて，その全員がインターネットを週に 5 時間以下で大学院への進学希望がある．

＜提案手法でのみ抽出された興味深いルール例＞

{ 平澤研,M1 }→{ 研究室登校 4 日,2_楽しい }
＝[支持度:0.0889(4/45),確信度:0.5000(4/8),確信度上昇率:0.7778]

解釈：平澤研究室の修士 1 年の人は全体の中で 8 人いて，その半数の 4 人が研究室に週 4 日登校し研究室が楽しいと感じている．

4.2.2　大規模アンケートデータ

4.2.1 節のデータはトランザクション数が少ないため，より大規模なデータを用いる．

シミュレーションの条件については，4.2.1 と同様である(提案手法にはグルーピング用の最小支持度を 0.008 で与えた)．

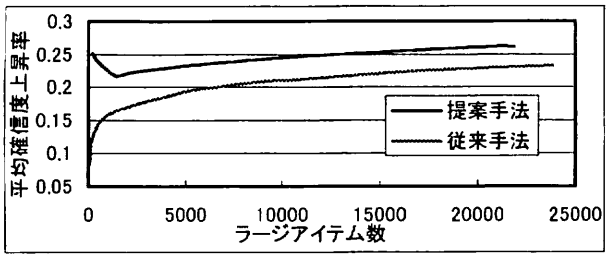

図7　大規模アンケートデータによる平均確信度上昇率

4.3　結果と考察

4.3.1　結果

平均確信度上昇率については，小規模アンケートデータ・大規模アンケートデータ共に従来手法より向上することが確認された．

- 小規模アンケートデータでは，データ件数(トランザクション数)が少ないため，ラージアイテム数 7000 以上ではほぼ全探索を行ってしまい，従来手法と提案手法での差が見られなくなった．
- 大規模アンケートデータでは，ラージアイテム数が非常に少ない場合には特殊なアイテムが抽出されているため提案手法の変動が激しいが，その他ではデータ件数が多いため安定した動きになる．

4.3.2　考察

- ルールを構成するアイテム数の多い相関ルールは確信度が高くなり確信度上昇率も高くなる傾向があることが分かった．従来手法と提案手法ともに最小支持度を下げるにつれてラージアイテム数が多くなるが，同時にステップ数(アイテム数)の多いルールが抽出され易くなる．ラージアイテム数が多くなるにつれて平均確信度上昇率が増加するのはこのためだと考えられる．
- ルールに記述項目が多く含まれているほど確信度が高くなる傾向が見られた．これは定型項目だけでなく，記述項目が含まれたルールのほうが面白いという人間の直感と本研究の目的にも合致している．
- 実際に抽出されたルールについては大変興味深いものが多数含まれていることが確認された．またルールだけでなく，抽出された単語を見ることで被験者がどのようなニーズを持っているか確認できた．

5　むすび

相関ルール抽出アルゴリズムであるアプリオリにエントロピーを導入することで，多値属性と記述項目，あるいはこれらの混合データからもルール抽出が可能となった．また，抽出されたルールは，信頼性が高いことをシミュレーションによって示した．

記述項目に対してはアプリオリを用いて単語のグルーピングを行ったが，この際にグルーピング用の最小支持度を与える必要がある．この最小支持度は，解析対象となる記述項目(テキストデータ)に依存している．

今回は，実際のグルーピングされた結果を見てこの最小支持度を決定したが，妥当な決定方法は今後の課題である．また，この種の手法には客観的，定量的に評価する尺度を示すことが難しい．適用分野の固有の知見を用いて定性的に評価することが多く，評価方法の確立も今後の課題である．

なお，定型項目については属性値に排他的に 1 が立つ問題について提案を行ったが，アンケートの複数回答項目のように，属性内に複数 1 が立つような場合にもグルーピングを行うことで適用できると考えている．

謝辞：著者の 1 人である熊谷は，有益な助言を頂きました平澤研究室・松嶋研究室・大野研究室の皆様に深謝致します．なお，本研究の一部は 2001 年度早稲田大学特定課題研究助成費(課題番号 2001A-556)の助成による．

参考文献

[1] R.Agrawal,R.Srikant:Fast Algorithms for Mining Association Rules,1994,Proceeding of the 20th VLDB Conference
[2] 喜連川優：データマイニングにおける相関ルール抽出技法,1997.7,人工知能学会誌,Vol.12,No.4, PP.513-520
[3] 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,2001.3,人工知能学会,Vol16. 2
[4] 松澤祐史：テキストデータからの頻出パターンのマイニング,1999,11,NLP Symposium
[5] 寺邊正大ら：相関ルールにもとづく属性生成手法，2000.1,人工知能学会誌,Vol.15 No.1 PP187-197
[6] 福田剛志：相関ルールの可視化について,1995.5　電子情報通信学会信学技報,DE95-6