パタンを利用した分類予測モデルとグラフ表現

森田　裕之，西口　真央

近年，さまざまなデータの蓄積が可能になるにつれて，マイニングの際にデータ要素間の関連性を考慮することが，より重要になってきている．その関係をグラフで表現することは，最適化モデルを設計するためのヒントになるばかりでなく，パタンなどの出力結果をよりコンパクトに，より多くの情報を集約して表現することが可能となる．本稿では，頻出パタンや頻出系列パタンとグラフとの関係について考察した後，クラス予測に特徴的なパタンを利用した分類予測モデルの1つである Classification by Aggregating Emerging Patterns とその出力結果のグラフ化について，実際のデータを用いた分析例を示しながら考察する．

キーワード：頻出パタン，頻出系列パタン，クリーク，分類予測モデル

1. はじめに

近年，さまざまなデータの蓄積が可能になるにつれて，マイニングの際にデータ要素間の関連性を考慮することが，より重要になってきている．そのための1つの方法としてグラフ表現があるが，これはモデルを設計するために必要であるばかりでなく，出力された結果を上手に表現することで，結果の理解を深め，より深い解釈を加えるためにも重要である．以下では，頻出パタンや頻出系列パタンとグラフとの関係について考察する．また，顕在パタンを利用した分類モデルである Classification by Aggregating Emerging Patterns[1] (以下，CAEP と略す) とその出力結果がどのようにグラフで表現され，解釈することが可能であるか，実用的な観点から実際の分析例を用いて考察する．

2. 頻出パタンとグラフ

あるスーパーマーケットの POS データが存在し，以下のようなレシート番号と購入アイテムが記録されているとする．

例えば，レシート番号：1の買い物カゴには，“ビール”と“紙おむつ”だけが入っており，これがレジでチェックされた際にレシート番号をキーとして，購入されたアイテムが記録されている（表1）．このとき，これら購入アイテムのグラフ表現の方法はいろいろ考えられるが，ここでは，併買の有無をエッジの接続によってグラフで表現すると，図1のように図示される．買い物カゴに3つのアイテムが同時に購入されていれば，レシート番号：2の3アイテムの場合のようになるし，4アイテムが同時購入されれば，同様にしてレシート番号：3のようになる．ここで，レシート番号：3の場合を見てみると，そこに存在する任意のサブアイテム集合は，レシート番号：3のグラフのクリークとなっており，このようなクリークが全体の購入データを見たとき，多くのレシートで出現しているなら，つまりそのパタンのサポート値が事前に決められた最小サポート値を上回るなら頻出パタンと呼ばれる．これらを全体のグラフとして表すとき，レシート番号：1と2のグラフについては，“ビール”と“紙おむつ”は同じアイテムであるため，これら2つのグラフを併合すると，図1の真ん中の3アイテムの場合と同じ接続になり，“紙おむつ”と“ビール”の間のエッジが2本存在する多重辺として記述できる．そのような多重辺が

表1　あるスーパーマーケットの POS データの例

レシート番号	購入アイテム
1	ビール
1	紙おむつ
2	ビール
2	紙おむつ
2	ワイン
3	ビール
3	紙おむつ
3	ワイン
3	粉ミルク
4	ビール
4	紙おむつ
4	ミネラルウォーター
4	粉ミルク
4	チョコレート

もりた　ひろゆき，にしぐち　まお
大阪府立大学大学院 経済学研究科
〒 599–8231 大阪府堺市中区学園町 1 番 1 号

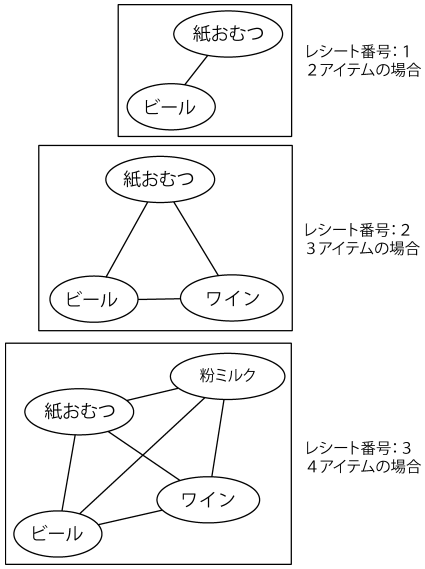

図 1　買い物カゴの中のアイテムの関係

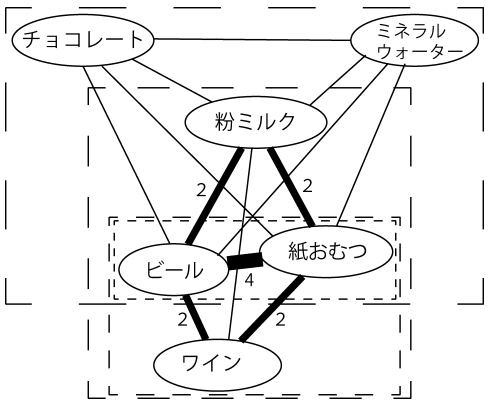

図 2　全体の購入アイテムの関係

存在する場合，1 本のエッジに統合し，その重複本数をエッジの横に記述すると，図 2 のように表現することができる．このとき最小サポートが件数で 3 件以上とすれば，{ ビール，紙おむつ } のみが頻出パタンとなるが，最小サポートの件数が 2 件以上であれば，{ ビール，紙おむつ，ワイン }，{ ビール，紙おむつ，粉ミルク } および，それぞれの任意のサブアイテム集合も頻出パタンとなる．またマーケットバスケット分析における条件部を { ビール，紙おむつ }，結論部を { 粉ミルク } とすると，4 レシート中 2 つのレシートで発生しているアソシエーションルールということになる．このように，POS データから頻出パタンを抽出するということは，個々の取引データから構成される完全グラフを併合し，全体の取引データを表現するグラフとして作成したときに，そのうちの多頻度で出現したクリークを探していると考えることもできる．

表 2　あるウェブログデータの例

セッション ID	順番	巡回ページ
1	1	トップページ
1	2	商品リスト
1	3	商品 A の詳細
2	1	トップページ
2	2	商品リスト
2	3	商品 A の詳細
2	4	商品リスト
2	5	商品 B の詳細

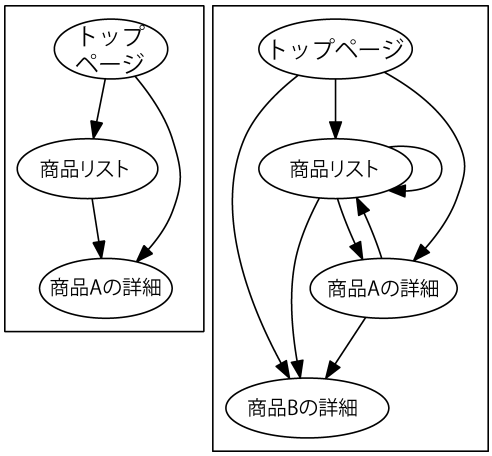

図 3　Web ページの巡回ログデータのページ間の関係

次に，アイテムの発生順序が存在するデータの例として，あるウェブサイト内のページ巡回ログデータを考えてみよう．あるユーザが，トップページからサイトに入り，サイト内の任意のページを巡回した状況が記録されている表 2 のようなデータが存在するとする．この巡回順序を有向枝によって，図 3 のように表現することができる．このときセッション ID：2 には商品リストが 2 回巡回されているので，それを自己ループで描画している．これは同一アイテムの繰り返しの出現を表現するためである．また，セッション ID：1 で，出現した順番だけを追うと，“トップページ ⇒ 商品リスト ⇒ 商品 A の詳細” だけのように思えるが，“トップページ ⇒ 商品 A の詳細” というシークエンスも存在するので，あるアイテムからそれ以降の順序で出現するアイテムへは，すべて有向枝で接続して表現する．このような有向グラフの中から，有向枝で直接接続されている部分グラフを抽出したとき，無向グラフの場

合と同様に，その出現頻度が最小サポート以上大きな値を持っていれば，それは頻出系列パタンとなる．例えば，最少サポートを 2 件以上とすれば，〈トップページ，商品 A の詳細〉などがそれに該当する．

このような単一のクラスに所属する頻出パタンや頻出系列パタンは，前述のようにあるパタンの出現を条件部としたとき，それとは異なるほかのアイテムから構成されるパタンの出現頻度が高ければ，アソシエーションルールとして用いられ，店舗内での商品配置などに活用される．

3. クラス分類とパタン

次に，クラス分類予測を行うための特徴的なパタンである顕在パタン [1]（以下，*ep* と呼ぶ）やコントラストパタン [2]（以下，*cp* と呼ぶ）と，グラフとの関係を考えてみる．いま，クラス 1 とクラス 2 という 2 つのクラスが存在し，すべてのトランザクションは，クラス 1 またはクラス 2 のいずれか一方のクラスに属しているものとする．その際，それぞれに所属するデータベースを D_1, D_2 とする．このとき訓練データの中で，あるパタンの出現頻度が相対的にクラス 1 とクラス 2 で大きく異なるものに着目し，分類予測モデルを作成する．クラス 1 における，あるパタン p のサポートを以下のように定義する．

$$sup_1(p) = \frac{cnt_1(p)}{|D_1|} \tag{1}$$

ここで $cnt_1(p)$ は，D_1 のパタン p を含むレコード件数を示し，$|D_1|$ は D_1 の総レコード件数とする．このとき，1 つのパタン p を特定すると，クラス 1 とクラス 2 のそれぞれのサポート値 $sup_1(p)$，$sup_2(p)$ が得られる．*ep* は，式 2 のような 2 つのサポートの比率を増加率 (Growth Rate) として定義し，その値が大きなものをより強力な説明力のあるパタンとするアイディアである．

$$gr_1(p) = \begin{cases} \frac{sup_1(p)}{sup_2(p)}, & sup_2(p) > 0, \\ \infty, & sup_2(p) = 0. \end{cases} \tag{2}$$

一方，*cp* は，式 3 のように，2 つのサポートの差が大きなものをより強力な説明力のあるパタンとするアイディアである．

$$df_1(p) = sup_1(p) - sup_2(p),\ df_1(p) > 0. \tag{3}$$

これらのパタンにはそれぞれの長短があるので，一概に良し悪しを決めることはできないが，実データに適用した傾向としては，*ep* は，説明力の強いパタンが出現するが，そのパタンがカバーするクラス内の要素数は少なくなる傾向があり，逆に，*cp* は，それほど強力でないパタンであってもカバーするクラス内の要素数が比較的大きいという傾向がある．これらのパタンを分類予測モデルに利用することの有用性としては，次のような例が挙げられる．図 4 の左側は，任意のパタン a と b が存在し，$sup_1(a) = sup_2(a)$，$sup_1(b) = sup_2(b)$，それぞれは 0 より大と考えるが，$sup_1(a) = sup_1(b)$ である必要はない．このとき，各クラスで両方のパタンを保持する要素がないため，そのようなパタンは予測モデルの説明要因とはならない．一方，2 つのパタンの各クラスのサポートは同じ値であるため，それぞれ別々に考えたとしても，予測モデルの説明力としては平凡で強い説明力があるとは思えない．しかしながら，先ほどと同様に，パタン c と d が存在し，各クラスの各サポートについては同じ条件であるような右側の場合を考えてみると，状況は少し異なる．図からわかるように，パタン c と d の両方を持つ要素がクラス 1 には存在するが，クラス 2 には存在していない．したがって，パタン c と d を別々に考えると，それぞれは平凡なパタンであるが，両方のパタンを併せ持つ要素は，クラス 1 にしか存在しないため，クラス 1 の所属を予測するには説明力の強いパタンであると考えられる．これは例えば，一般的な決定木モデルを作成する際には，仮にパタン c と d がカテゴリー属性を持つ変数として入力に与えられていたとしても，それぞれのパタンを単独で分岐候補として計算する際は平凡であるため採用されない．したがって，いずれかの分岐で，どちらかのパタンがたまたま強い説明力を持つようなサポートの偏りが発生しない限り，パタン c と d を併せ持つ分岐ルールが出現する可能性は小さい．も

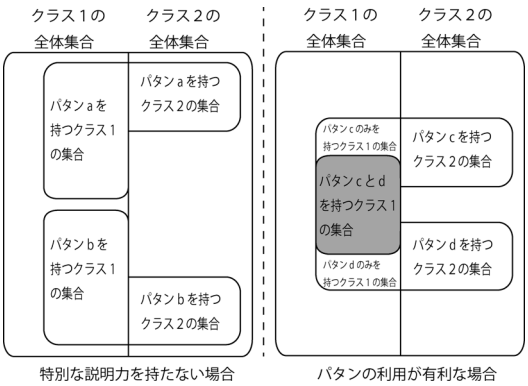

図 4　特徴的パタンと分類予測モデルの関係

ちろんこの例は少し極端であり，実際には右図のクラス 2 にもパタン c と d を併せ持つ要素が存在しているが，その違いがクラス 1 と 2 では大きく，その強力さが，式 2 や 3 で計算されて採用されることになる場合も多い．いずれにしても，決定木モデルのように局所探索を繰り返す方法では採用されない有用なパタンを利用して予測モデルを作成するオプションが存在する点は，大変魅力的な点であると考えられる．そのような 1 つの手法として CAEP が提案され，実用データへの適用結果も報告されている [3, 4, 5, 6]．その際，出力結果としては予測モデル作成に関係した *ep* や顕在系列パタン（以下，*esp* と呼ぶ）が出力されるが，これらの出力パタンも図 2 や図 3 で表現したようなグラフとして視覚化することが可能である．次節では，この CAEP の応用例としてある実用データに計算した結果を用い，工夫したグラフ表現を示す．

4. 計算機実験

4.1 分析対象データの概要と基礎分析

実験で使用するデータ[1]は，ゴルフ用品を取り扱うオンラインショップ会員の顧客属性データ（以下，属性データと呼ぶ），web アクセスログデータ（以下，ログデータと呼ぶ），受注データから構成されている．データ期間は 2010 年 7 月 1 日から 2011 年 6 月 28 日までの約 1 年間で，データ期間中にアクセスのある会員は 4053 人存在している．ある会員がショップにアクセスし，そのアクセスが終了するまでを 1 セッションとし，ある会員があるページから別のページに移動することを 1 イベントとすると，セッション数は約 12 万件，イベント数は約 156 万件存在している．このサイトは，ゴルフ用品のオンライン販売のほか，ゴルフ場の予約，および情報提供サービスを行っており，前二者の売上収入のほか，ネット広告の収入が，収入源であると考えられる．情報収集が目的の会員も存在するため，アクセスした会員の大部分が何かを購入しているとは限らない．また，予約や広告をクリックしたデータなどは提供されていないので，ゴルフ用品の販売と，ページの巡回情報が利用可能なデータとなる．

データ期間中，購入が確認されるユニークな ID 数は 1,343 人であり，全体の 33.14%程度であることがわかる．また各 ID の購入回数を見てみると，その購入頻度分布は，図 5 のようになる．購入顧客のうち，

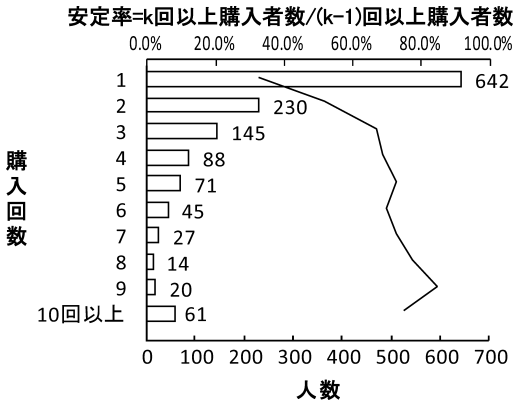

図 5 購入頻度分布

購入が 1 回しか確認できない顧客数は 642 人で，購入者数全体の 47.8%に達しており，ほぼ半数は 1 回しか購入していないことがわかる．また，安定率を見ると，2 回購入すれば 50%を超えているため，その半数以上は 3 回以上購入してくれることがわかる．安定購入顧客層の目安を 50%と考えれば，2 回以上の購入を達成してもらうことが 1 つのハードルになっている．また，その安定購入顧客層を増大するには，1 回購入した顧客をいかに 2 回目の購入につなげるか，さらには，どうやって 1 回目の購入を経験してもらうかということが，分析上重要な問題設定となっていることがわかる．そのため表 3 で示す 2 つの問題を以下では考えることにする．

分類予測モデルを構築するためのセッション数は，比較的クラスの増減人数が安定している早期のセッション数である 5 回とした．購入顧客は，6 回目以降のセッション中に 1 回以上の購入が確認される顧客，未購入顧客は，5 回以上セッション数が存在し，一度も購入していない顧客とし，潜在顧客は，6 回目以降のセッションの中に 1 回だけの購入が確認される顧客，安定顧客は，6 回以上のセッションの中に 2 回以上の購入が確認される顧客とした．

各クラスの人数と平均イベント数は，表 4 のようになっている．モデル構築のためのセッション数はすべて 5 回で統一されているが，1 セッションには複数のイベントがある可能性があるため，必ずしもイベント数が 5 ではない．

表 3 解決すべき分類問題

クラス 1	クラス 2
購入顧客	未購入顧客
安定顧客	潜在顧客

[1] 平成 23 年度データ解析コンペティションにおいて，経営科学系研究部会連合協議会と株式会社ゴルフダイジェスト・オンライン（GDO）より提供されたデータ．

表 4　分類クラスの概要

クラス	人数	平均イベント数
未購入顧客	962	40.4
購入顧客	735	52.9
潜在顧客	306	50.6
安定顧客	429	54.4

4.2　計算結果

使用するデータは，属性データとログデータであり，これらデータは形式が異なる．属性データのキー項目は 1 つであるが，ログデータは，2 つのキー項目が存在する系列データである．これら両方のデータを利用して ep および eps からなる統合化顕在パタンを用いた先行研究として，羽室ら [7] の研究が挙げられる．[7] では，“時間幅による制約”や“タクソノミの導入”の工夫を行いながら，統合化顕在パタンを用いて CAEP を構築する方法であるが，紙幅の都合上，詳細な説明は割愛する．本実験では，データの制約上，“時間幅による制約”や“タクソノミの導入”は実施しないが，ほかはほぼ同じフレームワークで計算している．以下では，グラフ表現の際にそれぞれのパタンが明確になるように，属性データから列挙される ep を ep_1，esp はログデータから列挙され，ep_1 と esp を組み合せて生成された統合化顕在パタンを ep_2 と呼ぶ．

予測モデルの入力として与えるログデータは，1 つの URL を 1 アイテムとしたが，トップページは出現回数があまりにも多いため，カテゴリー変数に変換して用いる．ほかの属性データについては，アイテムとして性別，スコアハンディキャップ，メルマガ購読の有無，会員登録年，年齢を使用した．性別，メルマガ購読の有無は 2 値であるため，そのまま使用し，スコアハンディキャップ，年齢，そして会員登録年は，適切にカテゴリー変数化して利用している．

以上の入力データを用いて，予備実験により適切なパラメータを設定し，テストサンプル法 [2] によって，3 種類のシードを用いて 3 種類の問題例を作成して検証を行った．検証結果の正答率の平均が，問題 1 では 0.606，問題 2 では 0.554 であり，多いクラスの割合が 0.55 前後から開始しているので，それほど良い結果ではないが，モデルを作成するパタンは興味深いものも列挙されている．モデルに利用された ep_2 のうち，スコアの大きい順にそれぞれのクラスから上位約 20 パタンを

[2] 本実験では訓練データ 6 割，テストデータ 4 割の割合でランダムに選択して作成

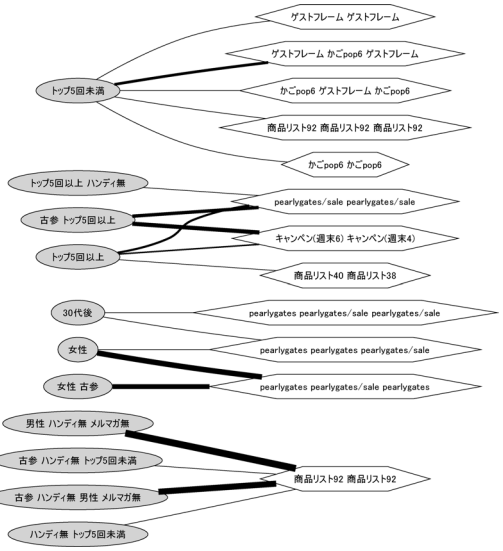

図 6　問題 2 の安定顧客クラスに出現した ep_2 のパス図 1

抽出し，2 つの異なる方法により，パス図として表現する．図 6 と図 7 は，問題 2 の安定顧客クラスに出現したパタンのパス図を表している．図 6 では，ノードは，丸いグレーのノードが ep_1 を，6 角形の白いノードが esp を表している．この 2 種類のノードがエッジで結ばれているものが 1 つの ep_2 であり，エッジの太さはスコアの大きさに比例している．一方，図 7 では，丸いグレーノードは属性データのアイテムを，また 6 角形の白いノードは esp のアイテムを表している．点線で囲まれたアイテム集合のうち，無向枝で接続されている部分が ep_1 を，また有効枝で接続されている部分が esp を表している．そして，ep_1 と esp が無向枝で接続されている部分が，1 つの ep_2 を表している．

図 6 と図 7 は，同じ入力パタンから描画されているが，図 6 は，比較的シンプルであるものの内部のアイテムは冗長に出現している．また図 7 は，そのような冗長性は排除され，ep_1 間，および esp 間のアイテムの関係も表現されているが少し複雑である．例えば，図 7 を見ると，ep_1 の包含関係や，esp のアイテムの繰り返しが表現されており，トップ閲覧 5 回以上や古参（比較的古い登録者）といったアイテムが，複数の異なる行動パタンと結びついていることが視覚的に見てとれる点は，この詳細な表現の長所であると言える．

次に，図 7 と図 8 をもとにモデルに出現した特徴的なパタンの考察を行う．図 8 は，問題 2 の潜在顧客クラスに出現したパタンのうち，図 7 と同様の基準でパタンを抽出し，同様にパス図として表現したものである．まず，安定顧客クラスでは，メルマガの登録は行っ

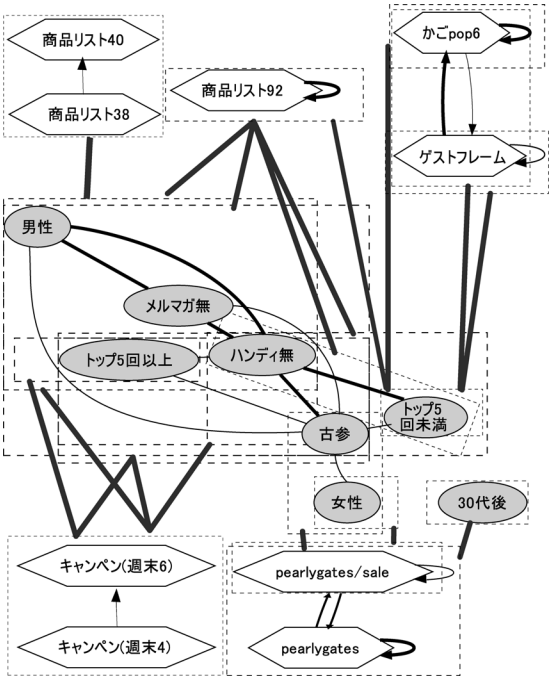

図 7　問題 2 の安定顧客クラスに出現した ep_2 のパス図 2

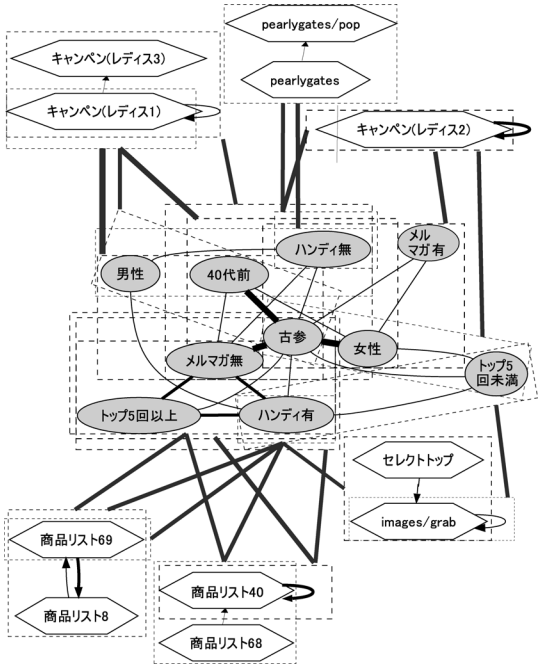

図 8　問題 2 の潜在顧客クラスに出現した ep_2 のパス図

ていないが，2010 年以前登録である顧客セグメントや，スコアハンディキャップのない，おそらくゴルフ初心者であるだろうセグメントが確認できる．*esp* の内容としては，セレクトショップのパーリーゲイツというブランドのページにセールである場合も含めてアクセスしていることがわかる．パーリーゲイツは，デザイン性を重視したゴルフウェアなどであるようだ．そのため，ゴルフ初心者の顧客セグメントで，ゴルフグッズへのこだわりを示すアクションは，ゴルフへの興味を示す複数購入のサインであり，重視する 1 つのポイントの一端が表れたのではないかと考えられる．一方，潜在顧客クラスは，スコアハンディキャップを持つ顧客セグメントや，メルマガ購読を行っている顧客，そして女性であるといったセグメントが確認できる．これらセグメントが，レディースのキャンペーンページに複数回アクセスすることは，あまり良い兆しではないのかもしれない．単に安かったらここで購入するというような顧客の態度であるとすると，単発的な購入で終わってしまう危険性もあり，継続購入を重視するプロモーションを展開するならば，注意すべきポイントであると言える．

5. おわりに

本稿では，頻出パタンとグラフ，そして頻出パタンを活用した分類モデルの関係を考察し，実データの適用例を使って，属性データとログデータが混在するような状況でも CAEP を応用した予測モデルが構築可能であり，その際出現したパタンがどのように表現されるかを示した．表現方法として，1 つのパタンを 1 つのノードとして表現する場合は，2 部グラフとして比較的シンプルに描画が可能であるが，内部のアイテムレベルでは，冗長な出現や，アイテム間の関係が無視されるという問題点が存在した．また，アイテムレベルから出現したパタンを表現すると，冗長性や関係性の無視という問題は，解決されるものの多少複雑になりすぎ，見にくいものになる危険性があることも指摘した．しかし見やすさの問題は，描画スペースとアイテム配置の問題によってかなり改善できる可能性はあり，実用性を考えると，モデルからさまざまな考察を行うという意味においては，図 7 や 8 のようなアイテムレベルでの記述が望ましいのではないかと思われる．まだまだグラフとしての表現方法においても，工夫できる点も残っていると考えており，今後，改善を試みたいと考えている．

謝辞　本研究では経済産業省「情報大航海プロジェクト」の“パーソナル情報保護・解析基盤開発・改良と検証”において開発されたプログラムを一部使用している．また，国立情報学研究所　宇野 毅明先生が開

発されたプログラム [8] も使用させていただいている．ここに深謝の意を表する．

参考文献

[1] Guozhu Dong, Xiuzhen Zhang and Limsoon Wong, “CAEP: Classification by Aggregating Emerging Patterns,” *Proceedings of the 2nd International Conference on Discovery Science*, 30–42, 1999.
[2] Stephen D. Bay and Michael J. Pazzani, “Detecting Change in Categorical Data: Mining Contrast Sets,” KDD, 302–306, 1999.
[3] 高橋宣行，瀧澤重志，加藤直樹，具源龍，“賃貸用オフィスビルのエントランスホールに対する感性評価の CAEP を用いた分析，”日本建築学会計画系論文集，**74**(640)，1403–1410, 2009.6.
[4] Atsushi Takizawa, Wonyong Koo, and Naoki Katoh, “Discovering Distinctive Spatial Patterns of Snatch Theft in Kyoto City with CAEP,” *Journal of Asian Architecture and Building Engineering*, **9**(1), 103–110, May 2010.
[5] 西口真央，森田裕之，“ブランドの価格属性を考慮したシークエンシャルパタンによるブランドスイッチ予測，”オペレーションズ・リサーチ：経営の科学，**57**(2), 79–87, 2012.
[6] Hiroyuki Morita and Yukinobu Hamuro, “A Classification Model Using Emerging Patterns Incorporating Item Taxonomy,” *Proceedings of the International Conference on Data Engineering and Internet Technology* (DEIT 2011), 15–17 March 2011, Indonesia.
[7] 羽室行信，中西正雄，山本昭二，“統合化顕在パターン判別モデルによる Web アクセスログデータの分析，”オペレーションズ・リサーチ：経営の科学 **53**(2), 75–84, 2008.
[8] 国立情報学研究所，宇野毅明先生のページ “http://research.nii.ac.jp/uno/index-j.html”

