論文・事例研究

グラフ縮約による主要なWeb巡回経路を利用したパターン分析

中原　孝信，森田　裕之，米田　知弘

1. はじめに

情報通信技術の発達は，ブロードバンド接続の利用を促進しており，2000年以降インターネットは急速に普及している．2004年末には，インターネットの利用者数は約8千万人で，人口普及率は62.3%，そして世帯普及率は86.8%にまで達している[3]．また，インターネットの普及とともに，ECサイト[1]の利用者数も急増しており，情報通信総合研究所の報告[2]では，パソコンによるインターネットショッピング利用経験率は91%にまで達している．これにともなって，企業にとってインターネットは，重要なマーケティングツールの1つとなり，SEO（Search Engine Optimization）やSEM（Search Engine Marketing）対策として見られるように，自社サイト内のユーザログを分析することで，HPや広告の改善，そして商品プロモートへ活用しようという試みが高まっている[1]．

本稿では，最初にWebログデータを利用して，ユーザの巡回行動を有向グラフで表現し，そのグラフを縮約することで，主要なサイト構造を抽出する．その後，抽出したサイト構造を利用して，あるコンバージョンイベント[2]に反応するユーザと，反応しないユーザに特徴的な巡回行動のパターンを多目的進化型計算（MOEA）を応用した方法で抽出する．主要なサイト構造を発見することは，ユーザが閲覧する主要なページ間の関係やその内容が特定できるため，より簡潔なサイト構造の構築とページ内容の再設計に対する一助となる．また，あるユーザグループに特徴的な巡回行動を発見することは，効率的なプロモーションを実施するための契機になることが期待される．

2. 分析対象データと基礎分析

本稿で利用するデータは，Webログデータを解析するツールを販売している某企業のサイトで蓄積されたアクセスログデータ[3]である．サイト内のページは約100種類あり，期間は2006年1～6月までの半年間である．ユーザ数は，約18,000人でCookieによりIDが識別されている．分析に利用できる項目は，ユーザIDに加え，参照元と参照先のURL，アクセス日時，検索キーワードなどである．

当該サイトのトップページ，ツールのキャンペーンページ，そしてメインメニューからリンクのある22種類のページに対して基礎分析を行ったところ，トップページの閲覧は全閲覧の約25%であり，最初にトップページを閲覧している割合は，サイト訪問数に対して約50%である．ユーザが最後に閲覧したページ（出口ページ）の割合は，トップページが約35%，ツールのキャンペーンページが約20%，そして，ツールの料金表に対するページが約7%である．料金表に対する出口ページの割合は，閲覧数に対しては最も高いことが確認できている．また，メインメニューの上部からリンクされているページは，下部からリンクされているページに比べ閲覧回数が多く，ユーザはメニューの上部のリンクを多くクリックする傾向にある．

これらの結果からいくつかの考察が得られる．ユーザはツールの料金を見てから退出する傾向があるが，

なかはら　たかのぶ
大阪府立大学 大学院
もりた　ひろゆき
大阪府立大学
〒599-8531 堺市中区学園町1番1号
よねだ　ともひろ
楽天㈱
〒106-6118 港区六本木6-10-1
受付07.7.12　採択07.11.15

[1] 自社の商品やサービスをインターネット上で販売しているサイト．
[2] 当該サイトにおける目的となる事柄を意味する．
[3] 平成18年度データ解析コンペティションで提供していただいたデータを用いている．

単純な基礎集計では，その巡回経路までは確認できない．また，閲覧数の違いからも，多くのユーザに共通する巡回行動や各ユーザで異なる巡回行動が存在するように思われる．本稿では，ユーザの巡回行動を表現したグラフを縮約することで，マイナーなページを閲覧の多いページへと統合し，多くのユーザに共通する主要な巡回経路を発見する方法を提案する．次に，ツールに関する申込経験の有無からユーザを分類し，ツールの課金利用者を増加させるために，主要な巡回経路を用いて申込経験者に特徴的な部分パターンを抽出する．

3. グラフの縮約によるメインストラクチャーの発見

ページ間の閲覧関係を考慮するために，各ページをノード，ページ間のハイパーリンクを有向枝として，各ユーザのセッション[4]ごとの巡回行動を有向グラフで表現する．そして，これを順次行うことで，分析対象ユーザ全体の巡回行動を表す1つの有向グラフを生成する．このグラフは，ユーザが巡回したすべてのページとハイパーリンクを含んでおり，サイトの全体像を把握するうえでは興味深い．しかし巡回されるページには，極端に密な部分と疎な部分が含まれており，効率的に特徴的な巡回パターンを発見することは難しい．本稿では，有向枝に重みを与え，その重みがある閾値以上になる有向枝だけを残すようにグラフを縮約することで，多くのユーザが共通して持つ巡回経路を発見する．

有向枝に与える重みは，ページ間の関係の強さを表す値が望ましい．本稿では，有向枝の重みにサポートを用いることにし，以下の式で定義する．

$$Sup(x, y)=\frac{x\text{から}y\text{への枝を含むセッション数}}{\text{分析対象ユーザの全セッション数}} \quad (1)$$

ここで，x，y はそれぞれノードを表す．

次に，有向枝に与えたサポートを利用して，ある最小サポートを閾値とする頻出経路を抽出するために，以下の方法でグラフの縮約を行う．

1) 有向枝の $Sup(x, y)$ を昇順にソートする
2) 枝を順番に選び，閾値より小さい間，3)の操作を繰り返す
3) If その枝を除去してもトップページを表すルートノードから到達可能ならば除去
   Else If 除去することでルートから到達できなければトップページに近いノードへ縮約

Webページの巡回は，主にトップページから順に閲覧されており，3)でルートノードからの到達可能性を調べることによって，従来のサイト構造ではトップページから到達できるはずのリンクが，枝を除去することで到達できなくなる問題を考慮している．これらの手順で得られた有向グラフを以下ではメインストラクチャーと呼ぶことにする．

3.1 グラフ縮約の適用

ページ間の閲覧関係を考慮する観点から，1セッションで2ページ以上の閲覧があるセッションを対象に上述のグラフ縮約手法を適用する．ユーザ数は6,108人，総セッション数は7,774，ノード数は94，そしてノード間の接続関係を表す有向枝の数は5,722である．グラフを縮約する際に閾値となる最小サポートは，1%，2%，5%の3つの値でそれぞれグラフ縮約を行う[5]．

図1は，最小サポートを1%に設定して，グラフ縮約を行った場合のメインストラクチャーであり，29個のノードから構成されている[6]．図1に出現してい

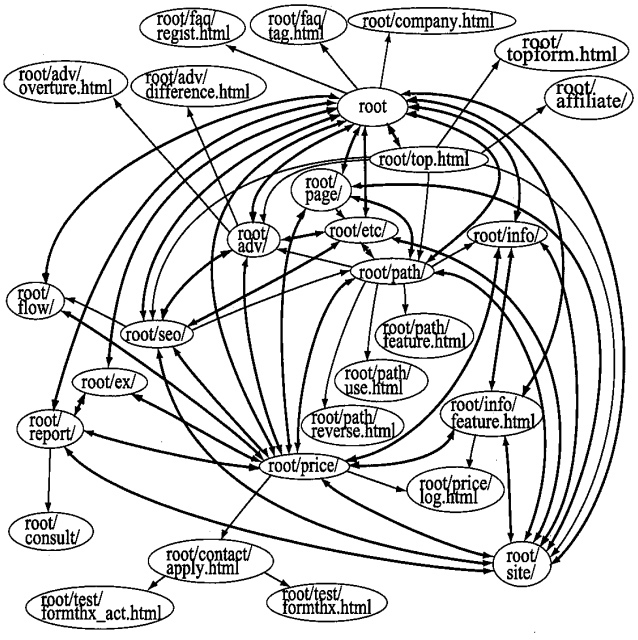

図1　閾値1%以上のメインストラクチャー

[4] 1人のユーザがサイトを訪問してから出ていくまでの間を1セッションと呼ぶ．

[5] 2%，5%を閾値としたグラフ縮約は後述するパターンの抽出で利用している．

[6] ノード内のrootは当該サイトのトップページであるhttp://www.sibulla.comを意味している．

るノードの多くが，トップページのメインメニューからリンクされているノードであり，ツールの機能概要や，導入に関するページ，オプションサービスに関するページなどを表すノードが多く出現している．しかし，同じメインメニューからリンクされているページでも，メニューの下部からリンクされているその他のサービスに関するページは，メインストラクチャーの出現が少なく，この点は基礎分析の結果とも一致している．

最も多くノードを縮約しているのはrootであり，よくあるご質問，お問い合わせ，用語集，そして提携サービスなどを表す36個のノードが縮約されている．次に多いのはroot/report/であり，7つのノードが縮約されている．それらは主にroot/report/からリンクのあるページで，SEO/SEMの効果検証レポート，そしてウェブマイニングに関するレポートなどを表すノードが縮約されている．

上記の点を踏まえ，メインストラクチャーを利用したサイトの再設計に関する考察を行う．rootに縮約されているノードは最も多いが，サイトの再設計の観点からは，ノードが縮約されたからといって，全く異なる内容のページであれば，それらの内容を1つのページに統合することは好ましくないと思われる．例えば，root/report/などは，上述したように，そこからリンクされている多くのページが縮約されており，その内容はそれぞれ関連している．したがって，それらの内容は，独立したページに記載するよりも，root/report/に統合することで，より多くのユーザの閲覧が期待される．一方，root/path/からリンクされているreverse.htmlやfeature.htmlそしてuse.htmlなどは，それぞれ関連した内容であるが，これらのページはメインストラクチャーに出現していることから，ある程度強い接続関係を持っているため，root/path/に内容を統合する必要はないと考えられる．このように，これらの点を手がかりとして検討すれば，より効率的なサイト構造を設計できる．

4. 分析対象ユーザの選択

分析対象となるサイトには，いくつかのコンバージョンイベントがあり，図2はツールの申込に関するコンバージョンイベントと，それらを経験したユーザ数を示している．ここで対象とするユーザは，データ提供期間中に1セッション以上の訪問があり，2ページ以上閲覧したユーザである．まずユーザは外部サイト

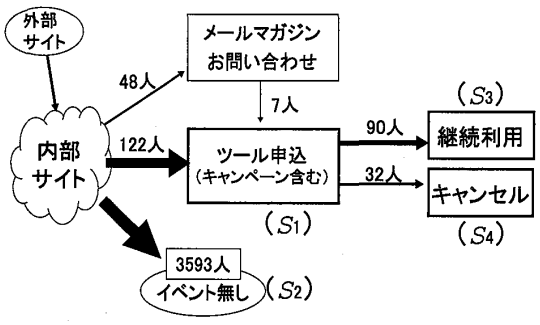

図2 コンバージョンイベントとユーザ数

(検索サイトやお気に入りを含む) から内部サイトへ訪問する．そして，いくつかのページを巡回した後，何らかのイベントを経験するユーザとしないユーザに分かれる．データ提供期間中にイベントの経験がないユーザは約3,500人[7]存在しており，最も多いことが確認できる．一方，メールマガジンの登録・解除，お問い合わせのイベントを経験した後，最終的にツールを申込むユーザは7人しか存在していない．そして内部サイト巡回後，ツールの申込を経験するユーザは，イベント無しに次いで多く，122人存在している．さらにツールを申込んだユーザは，継続してツールを利用するユーザ（90人）とキャンセルするユーザ（32人）に分かれていることが確認できる．ここで，ツール申込は30日間無料で利用可能なトライアルを意味しており，その後継続して利用するユーザだけが課金利用者となることに注意されたい．詳細な分析を行うにあたり，コンバージョンイベントはどれも重要であるが，ユーザ数の観点から，ここではツール申込を経験したユーザに着目して分析を行う．

以下では，ツール課金利用者の増加を目的にツール申込ユーザ集合（S_1）と，イベント経験無しのユーザ集合（S_2）を比較することで，それぞれのユーザ層に特徴的な巡回行動を発見する．その後S_1を，ツールを継続して利用するユーザ（S_3）と，キャンセルするユーザ（S_4）に分解して比較することで，ツールを継続利用するユーザ層に特徴的な巡回行動を発見する．また，各ユーザ層を比較するにあたり，ツール申込時の巡回行動は，ツール申込ユーザ集合に特有の巡回行動として出現することが考えられるため，そのような自明な巡回行動は，前処理によって削除していることに注意されたい．

[7] イベント無しについては，最終アクセスから1カ月以上アクセスのないユーザを対象としている．

5. 進化型計算を利用した巡回パターンの発見

グラフ縮約により得られたメインストラクチャーを利用し，それを巡回したパターンの中から，特定のユーザ層に特徴的な部分パターンを抽出する．前述のように，ユーザ自身の巡回行動も図3のように1つの有向グラフとして表現することができる．複数のセッションを持つユーザの場合は，各セッションで閲覧した最後のページと，次のセッションで閲覧した最初のページを有向枝で接続することでグラフ表現する．また，各セッションの最初のページは，当該サイト以外の外部サイトからのアクセスであることに注意されたい．このときすべてのユーザについて同様の操作をすると，n個のグラフ（ユーザ数をnとする）が得られる．ここで特定のユーザグループを定義し，このグループにのみ共通して出現する特徴的な部分グラフを抽出することができれば，このグループを説明するための有力な説明要因になることが期待される．しかし，このような一般グラフに対する部分グラフ同型判定問題は，*NP*完全であることが知られているため，効率的に厳密解を発見する方法は存在していない．そこで部分グラフの抽出について，近似解法の1つである進化型計算を応用した手法を適用する．

5.1 巡回パターンの遺伝子列化

各ユーザの巡回行動を表現した有向グラフに着目し，Webでの巡回行動に限定すると，ページ間には始点と終点となるページが1ページずつ存在していることがわかる．この特徴を利用すると，各ユーザが巡回した順番にページ（ノード）をたどることで，一連のリスト構造により巡回行動を表現することもできる．このようなリスト構造で表現することは，単に巡回したページデータだけではなく，各ページの閲覧時間や訪問回数など，巡回に関係した各種のデータを併せて表現できる利点がある．図4は，閲覧ページ，閲覧時間，そして各ページの訪問回数を同時に表現した例である[8]．このようなリスト構造とメインストラクチャーに出現したノードを対応させるために，リストの各ページが縮約されたページであるかを調べ，縮約されたページである場合は，縮約したページと置き換えることでメインストラクチャーに出現したノードからなるリスト構造に変換する．そして，そのことによって同一ページが連続する場合は，1つのページとしてまとめることで冗長性を省き，その際の閲覧時間については，まとめた各ページの合計値を利用している．

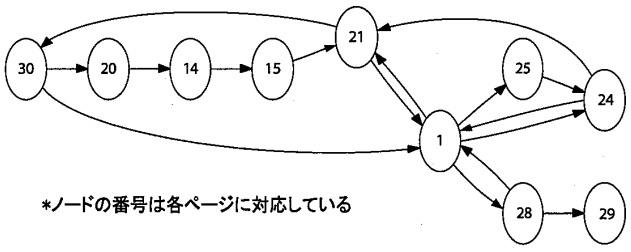

図3 ユーザの巡回行動のグラフ化の例

閲覧ページ	30	20	14	15	…	28	29
閲覧時間	-	10	16	15	…	12	-
訪問回数	1	1	1	1	…	5	5

図4 巡回行動のリスト構造化

今回のデータは，消耗品ではない単一商品の1回の購買に関するログデータであったため，あまり付加的なデータを多く与えても，結果にそれほど強い影響は確認されなかった．そこで以下では，閲覧ページとその閲覧時間を各ユーザの巡回におけるデータとして利用する．また閲覧時間については，実測値は値が多様なため，分布を3等分して3種類の各コード（最小1，最大3）でコード化する方法を用いた．したがって，各ユーザは閲覧ページとその閲覧時間という2重のリスト構造で，その巡回行動が表現されている．

5.2 部分パターンを発見するためのMOEAの適用

いま何らかの特徴的なユーザ層をユーザ集合Aとし，残りをユーザ集合Bとする．このとき，長さlのあるパターンをsp_lとし，各ユーザ集合のユーザがsp_lを保持している割合を$SPT(sp_l, A)$，$SPT(sp_l, B)$としよう．ここで実用的な観点から発見されることが望ましいのは，$SPT(sp_l, A)$が大きく，かつ$SPT(sp_l, B)$が小さいパターンと，その逆のパターンである．そこで式(2)，(3)の2つの2目的最適化問題と考え，MOEAによる部分パターンを求める手法（[4][5]）を利用し，最適なパターンを発見する．

$$\begin{cases}\max & SPT(sp_l, A) \\ \min & SPT(sp_l, B)\end{cases} \tag{2}$$

$$\begin{cases}\max & SPT(sp_l, B) \\ \min & SPT(sp_l, A)\end{cases} \tag{3}$$

[8] このうち，閲覧時間は各ページ間の閲覧開始時間の差を用いて計算しており，各訪問の最初は外部サイトからの訪問で，最後のページはサイトを後にした時間が不明なため，それぞれ閲覧時間を計算することができないので横線が入っている．

基本的な実施方法は文献[4][5]と同じであり，パターン長 l の範囲を設定し，各 l で最適なパターンをMOEAによって抽出し，最後にまとめて出力する．紙幅の関係上，MOEAのプロセスを詳細に記述することはできないが，文献[4][5]と異なる点は2つある．1つは，文献[4][5]では3重のパターンを抽出していたが，以下の計算では，閲覧ページと閲覧時間による2重のパターン，または閲覧ページだけの1列のパターンを抽出している点である．これは単純な修正であり，手法の上では大きな変化はない．もう1つは，エリート解の定義の変更である．文献[4][5]では，エリート解は各世代における近似パレート解集合としていたが，実用上の観点からは，$|SPT(sp_l, A) - SPT(sp_l, B)|$ の値が大きいものも必要となる．これはもちろん近似パレート解集合にも含まれているが，それ以外にも一部存在している．これらの解もなるべくカバーするために，以下のようにエリート解の定義を変更する．

定義（エリート解） 各世代における近似パレート解，およびそれ以外の解で $|SPT(sp_l, A) - SPT(sp_l, B)|$ の大きな値から elt 個の解をエリート解とする

今回の実験においては，$elt=50$ に設定し計算を行った．他のMOEAの設定は，文献[4][5]と同様であるが，任意の1コードと一致するパターン中のワイルドカードの上限数は，0～3でそれぞれ計算し，$l=1, \cdots, 10$ として計算を行った．

5.3 計算結果

以下では，前述の S_1 と S_2 の識別，そして S_3 と S_4 の識別という2種類の問題に対して実験を行う．実験では，グラフ縮約の効果を比較するため，各問題に対して，縮約していないデータ，1%，2%，そして5%の閾値で縮約した4種類のデータを作成した．また，ワイルドカードを導入したMOEAの効果および閲覧ページデータと閲覧時間の効果を比較するため，閲覧ページデータを対象に順序を考慮しないアソシエーションルールによる計算方法（M_1）[9]，閲覧ページデータを対象にワイルドカードを入れたMOEAによる方法（M_2），そして閲覧ページと閲覧時間のデータを対象に，ワイルドカードを入れたMOEAによる方法（M_3）をそれぞれ比較する．どの計算も，$|SPT(sp_l, A) - SPT(sp_l, B)| \geq 10\%$ のルールまたはパターンを最終的に抽出している．

5.3.1 S_1 と S_2 の識別に対する結果

ある長さのパターン sp_l には，$SPT(sp_l, S_1)$ と $SPT(sp_l, S_2)$ の目的関数値が計算される．グラフを縮約する程度によって出てくるパターン，もしくは消えてしまうパターンは異なるが，ここでは，全体的な評価をする観点から，1つの方法として，各手法を各データセットに対して計算した場合の，ルール数，重心 $SPT(sp_l, S_1)$ と $SPT(sp_l, S_2)$ の各平均値），そして共分散値を表1にまとめている．

このデータセットに対しては，両方の目的関数値が小さい領域で比較的多くのパターンが発見される傾向が見られた．解の出現領域は，各2目的問題では $(SPT(sp_l, A), SPT(sp_l, B))$ 空間において各 SPT が0%～100%の範囲で，$|SPT(sp_l, A) - SPT(sp_l, B)| \geq 10\%$ となる右下と左上の三角形の領域となる．エリート解の定義より，$SPT(sp_l, A)=100\%$ かつ $SPT(sp_l, B)=0\%$（またはその逆）に近い解の発見を試みるが，本データに対しては，実際にはあまりその領域付近の解は存在しないか，発見が困難であるため，解の分布は原点から右上がりの放射直線上付近に分布することが多かった（図5，図6参照）．したがって，一般には目的関数値の共分散値の大小で，目的空間上の多様な解を発見したかどうかを判断することは困難である

表1 計算結果の要約（S_1 と S_2 の識別）

アソシエーションルール（M_1）

グラフ縮約	共分散	重心	ルール数
なし	12.98	(15.8,3.4)	379
1%	13.29	(16.0,3.5)	408
2%	18.55	(16.4,4.0)	364
5%	24.34	(17.6,4.2)	149

MOEA:閲覧ページのみ（M_2）

グラフ縮約	共分散	重心	パターン数
なし	95.60	(29.0,10.1)	154
1%	99.70	(27.1,8.9)	201
2%	105.73	(27.7,10.4)	201
5%	109.46	(28.4,11.4)	267

MOEA:閲覧ページと閲覧時間（M_3）

グラフ縮約	共分散	重心	パターン数
なし	60.54	(24.2,6.9)	190
1%	64.95	(23.0,7.1)	34
2%	74.30	(22.7,7.4)	32
5%	69.98	(20.8,6.7)	42

[9] アソシエーションルールによる計算は，データマイニングツールMUSASHI http://musashi.sourceforge.jp/を用いて行っており，最小サポートを0.01に設定しルールを抽出している．

が，このデータについては，この共分散値である程度の多様な解を発見している目安になると考えられる．このように考えると，全体として M_1 よりは，M_2 や M_3 の方が良い傾向を示していることがわかる．また各方法の中では，縮約したデータに対する結果の方が，重心および共分散値もよい傾向を示していることがわかる．図5は，以上の結果を最も顕著に示している例の1つとして，M_1（縮約なし）と M_2（5%の縮約）で発見されたパターンをそれぞれプロットしたものである．ここで1つの点は，1つのルールまたはパターンを表している．図の左下側は両方の方法で共に発見されているため，とても密集している．一方，中央および右上側の領域には黒い点だけが点在しており，これらは M_2 のみが発見できている点である．他の解の分布も同様の傾向を示しており，前述の共分散値の値と併せて考えると M_1 よりは，M_2 や M_3 の方が全体としてはよい結果を示しているといえる．

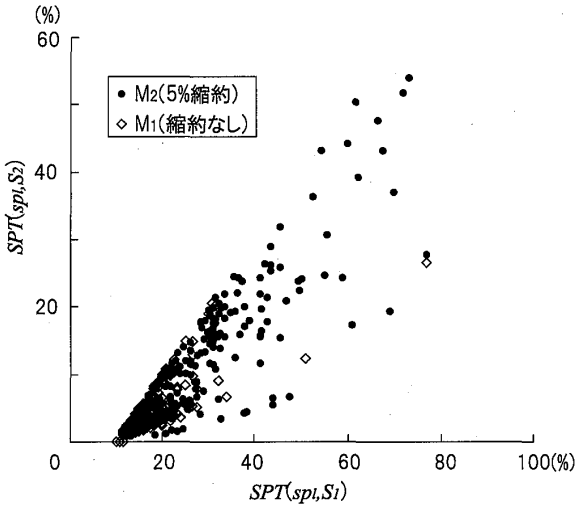

図5　M_1（縮約なし）と M_2（5%の縮約）結果の比較

これらの計算に際して発見されたエリート解のうち，近似パレート解の割合は平均で28.2%であった．これは近似パレート解のみをエリート解と定義していた先行研究[4]と比べて，かなりの数の候補パターンを新たに提示できていることが分かる．

5.3.2　S_3 と S_4 の識別に対する結果

次に S_3 と S_4 の識別問題に対する同様の計算結果を表2に示す．先ほどは図5に示したように，S_2 を最大化し S_1 を最小化する解が出現しにくい傾向にあった．しかし，今回のデータセットに対しては両方の最適化問題に対する解がバランスよく出ていることが，M_1 の共分散が負の値になっていることからも確認できる．ただ全体的な傾向としては同様であり，M_1 よりは，M_2 や M_3 の方が多様な解を発見している．またいずれの手法においても，縮約を行ったほうがよい傾向を示している．

図6は，M_1（縮約なし）と M_3（2%の縮約）でパターンをプロットしたものである．M_1 では両方のパターンを発見してはいるが，全体的に左下側に偏っていることがわかる．一方，M_3 はグラフ縮約やワイルドカードを導入したことによって，かなり広範囲にパターンを発見している．

これらの計算でエリート解に対する，近似パレート解の割合は平均で54.0%であった．これは前節の同じ割合と比較すると増加しているが，やはり近似パレ

表2　計算結果の要約（S_3 と S_4 の識別）

アソシエーションルール（M_1）

グラフ縮約	共分散	重心	ルール数
なし	-10.51	(11.1,16.6)	81
1%	-21.44	(11.5,15.4)	83
2%	-8.94	(12.7,15.5)	62
5%	10.33	(22.7,11.8)	36

MOEA:閲覧ページのみ（M_2）

グラフ縮約	共分散	重心	パターン数
なし	118.32	(14.9,18.7)	157
1%	65.96	(13.7,16.7)	197
2%	129.98	(16.3,17.1)	202
5%	88.26	(16.7,14.9)	291

MOEA:閲覧ページと閲覧時間（M_3）

グラフ縮約	共分散	重心	パターン数
なし	94.47	(16.0,19.1)	179
1%	84.76	(15.9,19.9)	205
2%	148.00	(17.5,20.2)	202
5%	99.03	(17.5,16.3)	266

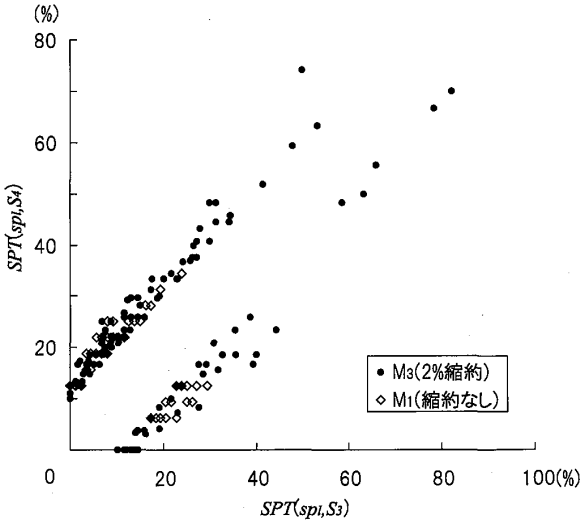

図6　M_1（縮約なし）と M_3（2%の縮約）結果の比較

ート解以外にも，可能性のあるパターンが提示できているといえる．

5.4 パターンの考察

以上のようにして得られたパターンから，まず S_1 と S_2 に対する計算結果の特徴について考察する．抽出されたパターンは多様であるが，S_1 に特徴的なパターンで，類似している主要なものをまとめたところ，表3に示す3つのパターンが確認された．もちろん以下にあげるほかにも，その数は少ないながら，SPT の差が大きなパターンも存在しており，実用上はそれらを個々に解釈することが必要であるが，ここでは紙幅の関係上，比較的共通して出現したパターンに着目して考察を行うことにする．

最初は，申込確認パターンで，これはトップページから申込ページへと向かう巡回である．実際には前処理により申込時点の自明な巡回行動は削除しているため，申込ページへアクセスしたからといって，その時点で申込を行っているわけではないことに注意されたい．このパターンは，すぐにユーザに対するアプローチの必要性を求めるものではないが，ユーザが当該商品に対してどの程度の関心を持っているかを表す1つのバロメータになるのではないかと考えられる．

次に確認されたのは，2番目の価格確認パターンである．ただしこのパターンは，1%で縮約したグラフのデータセットからは抽出されず，5%の縮約グラフにおいて最も顕著にそのパターンが出現している．したがってこれは，グラフ縮約を用いたメインストラクチャーの特定によってもたらされた1つの成果であると考えられる．このパターンは価格表の参照を意味しているが，特徴としては，複数回この価格表を参照していることである．ワイルドカードを表す＊を含んでいることから，価格表と交互に見ているパターンは，トップページであったり，広告効果のページであったりと完全に一致しているわけではないが，価格を複数回確認しているという点で一致している．この点はワイルドカードを導入してパターンを抽出した効果であるといえる．また，解釈としては，当該商品に対するさらなる関心の表れであると考えられる．価格表は商品に対する対価を意味し，単に無料で利用しようとしているだけではないため，コストに関する意識を明確に持った表れであるといえる．また価格表を複数回確認し，特に広告効果と見比べているような行動は，明らかに費用対効果の検討を行っていると考えられる．

最後の機能確認パターンも，縮約の大きなデータに対して顕著に現れており，これはツールに関心を持つユーザが効果面において詳細に検討しようとしていることの表れであるといえる．したがって，このパターンも積極的なアプローチを行うための1つの重要な兆候であると考えられる．

さらに閲覧ページと閲覧時間の両方を考慮したパターン（M_3）についても確認したが，これらのデータに対しても，基本的には同様のパターンが出現していた．閲覧時間の観点から特徴的なポイントとしては，価格や機能などを確認しているページの閲覧時間は，最も長い時間を表すコード3の出現が多いことがわかった．

表3の3つのパターンについて，M_2 と M_3 の出現割合，およびそれぞれの SPT の最大値を以下の表4～表7に示す．

表4と表5から，共通する傾向としては，グラフ縮約の程度が大きくなるにつれ，各パターンに分類され

表3 出現頻度の大きな閲覧ページパターン

申込確認パターン

主な閲覧ページ
/root/
/root/top.html
/root/topform.html

価格確認パターン

主な閲覧ページ
＊
/root/price/
＊
/root/price/

機能確認パターン

主な閲覧ページ
/root/page/
/root/path/
/root/etc/ or /root/price/

表4 M_2 で出現したそれぞれのパターンの割合

グラフ縮約	申込確認パターン	価格確認パターン	機能確認パターン
なし	6.49	0.00	0.00
1%	16.42	0.00	1.99
2%	17.91	15.42	3.98
5%	28.84	35.58	9.74

表5 M_3 で出現したそれぞれのパターンの割合

グラフ縮約	申込確認パターン	価格確認パターン	機能確認パターン
なし	1.58	0.00	0.00
1%	5.88	0.00	0.00
2%	6.25	3.13	3.13
5%	23.81	7.14	2.38

る割合が大きくなっていることがわかる．これはグラフ縮約によって微小な巡回行動の違いが吸収され，共通化が図られた影響が1つにはあるのではないかと考えられる．

また2つの表からM_2とM_3の違いを比較してみると，全体的には，M_2のほうが各パターンの属する割合が大きいことがわかる．この原因の1つは，コード化の違いによる影響であると考えられる．M_3は，巡回パターンに巡回時間を合わせたパターンであるため，同じ巡回パターンであっても巡回時間のパターンが異なれば違うパターンとして識別される．そのためパターンが多様となり，このケースにおいては，別の強力な類似パターンとして出現するよりもエリート解の選択において選択されにくいパターンとなったために，表5の割合も，またパターンの総数も少なくなっていると考えられる．逆にM_2では，グラフ縮約によるパターンの共通化が図られ，かつ，このケースにおいては，本アルゴリズムのエリート解の選択において共存するパターンとして出現したため，表4のパターンの割合はグラフ縮約の程度に伴って増大していると考えられる．

しかしこのような傾向は，必ずしもM_2とM_3に関して見られる共通の傾向とはいえないだろう．データとグラフ縮約の程度によっては，M_3でも新たなエリート解の数が増加することは考えられる．また，M_2においても例えばグラフ縮約の程度が大きすぎるとパターンは共通化しすぎて，その出現数は激減する．Webマイニングにおいて，その巡回時間を考慮することは一般的であるが，今回のデータについて，実験した範囲でいえば，巡回時間のコードはそれほど重要ではなく，M_2による方法で十分なのではないかと考えられる．

表6と表7は，M_2とM_3の各パターンに対するそれぞれの縮約状態に対応したパターンの中でSPTの差が最大の値をそれぞれ記述している．両方に共通して，申込確認パターンは縮約がなくてもある程度顕著なパターンとして発見されている．そのため縮約した場合のほうが若干SPTの差が大きくなっているものの，それほど大きな違いはない．一方，価格確認パターンや機能確認パターンは，縮約によってパターンがより強く識別されている．したがって，このように縮約の効果として，差が現れるかどうかはデータに依存している部分があるといえる．

次にS_3とS_4に対する計算結果の特徴について考察する．これらのユーザは，S_1の中のサブグループであり，先ほどよりも両者の行動は共通している．そのためか，明確に共通して出現した興味深いパターンは発見されなかった．そのうちS_3に特徴的で比較的長いパターンとして，表8のようなパターンが抽出された．このパターンは，M_3には出現せずM_2にのみ出

表6 各パターンのM_2におけるSPT

申込確認パターン

グラフ縮約	差	$SPT(sp_l, S_1)$	$SPT(sp_l, S_2)$
なし	14.47	15.66	1.19
1%	14.87	16.06	1.19
2%	15.55	18.56	3.01
5%	20.9	25.69	4.79

価格確認パターン

グラフ縮約	差	$SPT(sp_l, S_1)$	$SPT(sp_l, S_2)$
なし	0.00	0.00	0.00
1%	0.00	0.00	0.00
2%	14.81	21.65	6.84
5%	21.83	41.56	19.73

機能確認パターン

グラフ縮約	差	$SPT(sp_l, S_1)$	$SPT(sp_l, S_2)$
なし	0.00	0.00	0.00
1%	12.71	24.49	11.78
2%	12.74	24.74	12
5%	13.57	29.87	16.3

表7 各パターンのM_3におけるSPT

申込確認パターン

グラフ縮約	差	$SPT(sp_l, S_1)$	$SPT(sp_l, S_2)$
なし	12.66	13.25	0.59
1%	13.77	15.83	2.07
2%	13.69	15.83	2.07
5%	19.69	22.5	2.81

価格確認パターン

グラフ縮約	差	$SPT(sp_l, S_1)$	$SPT(sp_l, S_2)$
なし	0.00	0.00	0.00
1%	0.00	0.00	0.00
2%	10.64	12.84	2.20
5%	14.05	19.27	5.21

機能確認パターン

グラフ縮約	差	$SPT(sp_l, S_1)$	$SPT(sp_l, S_2)$
なし	0.00	0.00	0.00
1%	0.00	0.00	0.00
2%	10.64	12.84	2.20
5%	13.73	17.43	3.70

表8 S_3とS_4間に違いの現れた閲覧ページパターン

	主な閲覧ページ
全機能確認パターン	/root/site/
	/root/page/
	/root/path/
	/root/etc/
	/root/adv/
	/root/seo/

現した．もっとも多く出現したのはグラフ縮約を5%で実施した場合であり，出現したパターンの6.21%を占めていた．またその中で最大の*SPT*の差は，14.29%であった．これは，当該ホームページのメインメニューに掲載されている機能概要や価格の説明であり，そのほとんどすべてが閲覧されている．これは最初に無料による利用申込だけを行うような，うわべだけの行動ではなく，当該商品をその後も利用し続けようという顧客行動の一端を表しているパターンではないかと考えられる．

6. おわりに

本稿では，課金利用者を増加させるために，ユーザのアクセスログから，特定のユーザ層に特徴的なパターンの発見を行った．分析手法としては，まず初めに各ユーザのセッションごとのページ遷移をグラフで表現し，分析対象ユーザ全体の行動を1つの有向グラフで表現した．そして，グラフ縮約により，多くのユーザに共通する巡回経路をメインストラクチャーとして抽出した．次に，MOEAを用いたパターン抽出を行い，申込ユーザ，および継続利用ユーザに特徴的なパターンを発見した．ユーザのアクセスログをメインストラクチャーに縮約したことによる利点は大きく2点挙げることができる．まず，1点目としては，複雑で情報量の多いユーザの行動から，情報量を圧縮しながら頻出経路を浮き彫りにできるグラフへと変換できたことである．2点目としては，メインストラクチャーを抽出することで，ユーザの関心が高いページとその内容を特定できるため，コンバージョン転換率[10]を高めるサイト構造と，ページ内容を再設計する手がかりを提供できることである．

特徴的なユーザの閲覧パターンからは，申込ユーザは，最終的な申込までに興味シグナルを発していることが確認できた．サイト運営側としては，このシグナルを発するユーザはリードジェネレーションになる可能性が高いので，非常に重要で見逃してはならないものである．また，シグナルを発見したら，発信ユーザに対するアクションをできる限り速やかに行う必要がある．そして，ユーザの興味レベルに応じた適切な方法を使い分けることも重要である．各ユーザの理解レベルを無視して，ただ最終的なコンバージョン[11]だけを促すのは効率的な方策ではない．具体的なプロモーション案としては，ユーザの行動に応じて，ごく自然に，最適なページをカウンターとしてあてる行動ターゲティング[12]が考えられる．例えば，機能ページへのアクセスが規定回数を超えた時点でメルマガの登録を勧誘するバナーを表示したり，料金表ページの二度目のアクセス時には，個別コンサルティングアポイントのパーミッションを取るなどの，コンバージョンを最大化するプッシュコミュニケーション[13]を推進することが考えられる．

今後の課題としては，本稿で扱ったデータは1つの製品に対するデータであり，ショッピングモールのようなデータではないため，そのような複雑なデータにも適用を試み，更なるメインストラクチャーの有効性を確認したいと考えている．

謝辞 本研究の一部は科研費（0196928）の助成を受けたものである．

参考文献

[1] 財団法人インターネット協会，“インターネット白書2006,” 2006.

[2] ㈱ 情報通信総合研究所，“MIN 第44回アンケート ブロードバンド＆インターネットショッピング利用実態調査”, 2004.

[3] 総務省情報通信統計データベース，“通信利用動向調査,” 平成16年調査.

[4] 中原孝信，森田裕之，“ターゲット顧客を識別するためのクレジット購買履歴データを用いたパターン分析,” オペレーションズ・リサーチ，Vol. 51，No. 2，pp. 89-96, 2006.

[5] H. Morita and T. Nakahara, “Pattern Mining for Historical Data Analysis by using MOEA,” *Proceedings of the 7th International Conference on MultiObjective Programming and Goal Programming*, Tours, Loire Valley, 2006.

[6] 横山隆治，“次世代ネット広告テクノロジー究極のターゲティング,” 株式会社宣伝会議, 2006.

[10] サイト来訪者のうちツール申込みに至ったユーザ数の割合．

[11] 目的となるページへの誘導

[12] ネット顧客のWebサイト，ネットワーク上の行動を収集し，把握してターゲティングの精度をあげるテクノロジー[6]．

[13] ユーザの許可を取って広告メールなどを送信すること．