論文・事例研究

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

羽室　行信，中西　正雄，山本　昭二

1. はじめに

近年のWWW（The World Wide Web）の爆発的な発展に伴い，多くの企業で従来の情報システムをWeb上で実現するeコマースの動きが活発化している．この動きに呼応するように，Web上での広告収入を基盤としたビジネスが台頭するなど，Web上のデータやサービスをいかにビジネスに有効活用するかに関心が集まっている．このような状況を背景として，Webマイニングと呼ばれる研究領域が発展してきた．

Webマイニングとは，WWW上のデータやサービスから自動的に有用な情報を発見するためにデータマイニング技術を用いる研究領域の呼称である[4]．Webマイニングは大きく，1)Web文書から有用な情報を発見することを目的としたWeb内容マイニング(Web Content Mining)，2)Webのリンク構造をモデル化することを目的としたWeb構造マイニング(Web Structure Mining)，そして3)Web閲覧者の行動記録からアクセスの規則性を発見することを主目的としたWeb利用マイニング（Web Usage Mining)，の3つの領域に分類される[5]．

本研究の目的は，Web利用マイニングについて，いくつかの既存の手法を組み合わせることでユーザのページ遷移に関する判別予測モデルを構築し，その有効性を示すことにある．

本研究で利用する手法は，相関ルールマイニングを源流とする頻出パターンマイニング（Frequent Pattern Mining）の分野で開発されてきた各種手法（相関ルール[1][6]，系列パターン[2][7]，Emerging Pattern, CAEP[3]）を組み合わせたもので，我々が統合化顕在パターン判別モデル（CAIEP：Classification by Aggregating Integrated Emerging Pattern）と呼ぶ判別モデル構築手法である．

論文の構成は以下の通りである．2節では分析対象とするWebサイトの概要を示す．3節にて構築する判別モデルの概要を述べ，4節で判別モデルの構築手法であるCAIEPについて論じる．そして5節にてデータセットを示した後に，6節で判別モデルの構築結果について考察を行い，最後に7節にて，他の手法との比較において提案手法であるCAIEPの精度評価を行い，その有効性についての考察を行う．

2. 対象Webサイトの概要

本研究では，平成18年度データ解析コンペティションで提供されたWebアクセスデータを利用した．これはログ解析ツール「シビラ」をASPでサービス販売する株式会社環から提供されたデータである．

このシビラのサービスを受けるWebサイトAの管理者は，そのサイトのWebページに特殊なタグを埋め込むことでユーザのアクセスログがシビラのサーバーに転送され，シビラはそのデータを基に各種分析サービスをASPで提供するというものである（図1)．今回対象となるWebサイトAはシビラの販売サイトそのものである．

シビラの販売サイトのページ構成の一部を表1に示

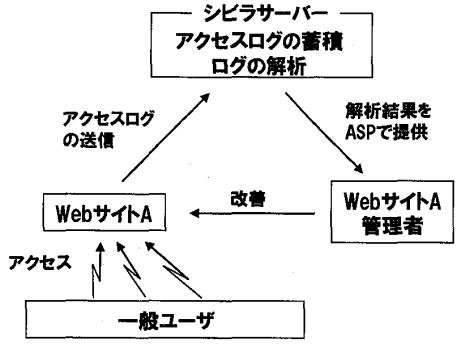

図1　アクセスログの収集方法

はむろ　ゆきのぶ，なかにし　まさお，やまもと　しょうじ
関西学院大学 経営戦略研究科
〒662-8501 西宮市上ケ原一番町1-155
受付07.7.13　採択07.11.15

表1　Webサイトの構成概要

HOMEページ内メニュー分類	
メニュー	ディレクトリ名
サイト紹介	
内容紹介	info
特徴	info
機能概要	
サイト解析	site
ページ解析	page
経路解析	path
その他の機能	etc
広告効果測定	adv
SEO効果測定	seo
導入	
申込の流れ	flow
料金表	price
導入事例	ex
申し込み	contact

表2　Webアクセスログデータ

ユーザID	UNIX時刻	URL	検索語
uid01	1138333340	Yahoo_Overture	解析, ブログ
uid01	1138333340	/top.html	
uid01	1138333469	/page/	
uid02	1138333748	Google_None	アクセスログ
uid02	1138333748	/topform.html	
:	:	:	

している．ホームページ内の各メニューは右列に示されたディレクトリに保存されており，各ディレクトリには複数のHTML文書が格納されている．

本研究で利用したユーザのアクセスログデータの項目は「ユーザID」，「アクセス日時」，「リクエストURL」の3つである．ユーザIDはクッキーの機能を利用して記録しているため，ユーザ環境によってはセッションをまたがるユーザのアクセスを同一と識別できないこともある．また，ログデータとして直前にアクセスしたURL（リファラURL）が記録されており，ユーザが，どの検索エンジン（Yahoo，Google，MSNなど）を利用して訪れたかが記録される．また，検索エンジン連動型広告（リスティング広告）であるOvertureもしくはGoogle Adwordsを利用した場合にはそれも記録される．さらに，ユーザが検索に利用したキーワードもわかるようになっている．表2にデータのサンプルを示す．

3．ページ遷移判別モデル

本研究では，ユーザの対象サイトへの初回訪問からのページ遷移の回数に応じて，以下に示す2つのページ遷移判別モデルを構築する．

直帰・滞在判別モデル　ユーザの初回のアクセスログから，その直後にサイトから出ていく（直帰）か，もしくはそのままサイトに滞在するかを判別するモデルを構築する．利用できるデータは，外部リファラURL（検索語含む）と初回アクセスURLのみである．直帰ユーザの特徴を知ることでリスティング広告の有効性や検索語についての知識発見が期待できる．

申込・未申込判別モデル　初回訪問から10 PV (Page View) の時点で，そのユーザがシビラの申込を行うかどうかを判別するモデルを構築する．直帰・滞在判別モデルで利用するデータに加えてページ遷移について10回分のアクセスログデータを利用できるため，実際にサービスを申し込む（もしくは申し込まない）ユーザのページ巡回行動についての有用な知識発見が期待できる．

4．判別モデル構築手法

本研究では頻出パターンマイニングの分野で展開されているいくつかの手法を組み合わせた判別モデル構築手法CAIEPを提案する．

頻出パターンマイニングとは，データベースにある一定以上の頻度で存在するパターンを列挙し，それらのパターンから有用な知識を発見することを目的としたマイニング手法の総称である．

パターンをどのように定義するかによって様々なバリエーションが存在する．マーケットバスケット分析に見られるように，単なるアイテムの共起を扱う「アイテム集合」は，パターンの最も単純な定義である．その他のパターンとしては「時系列」や「モチーフ」，そして最近では「グラフ」といったより複雑なパターンのマイニングが注目されている．

以下では，CAIEPを構成する要素技術として一般化アイテム集合，一般化系列パターン，顕在パターン，そして顕在パターンを利用した判別モデルCAEPについて論じていく．

4.1　一般化アイテム集合

Srikantらの研究[6]に従って，一般化アイテム集合の定義を以下に示す．

用語　$\mathcal{I}=\{i_1, i_2, \cdots, i_m\}$ を m 種類の**アイテム**と呼ばれるリテラルの集合とする．アイテム $i \in \mathcal{I}$ をノードに持つ有向非循環グラフ $\mathcal{T}$ をアイテムのTaxonomy（分類）と呼ぶ．$\mathcal{T}$ 上でノード p から c へのパスが存在する場合，p は c の先祖（c は p の子孫）と呼ぶ．また $\mathcal{I}$ の部分集合 X を**アイテム集合**と呼ぶ．

アクセスログデータでは，URLや検索語などがアイテムに相当する．図2にアイテムとしてのURLお

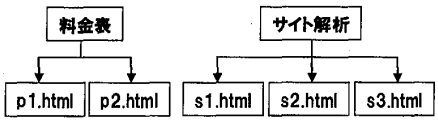

図2　Taxonomyの例

表3　URL閲覧データベース例（アイテム集合）

TID	閲覧したURL(アイテム)
1	p2.html, s2.html
2	p1.html, s1.html, s2.html
3	s1.html, s2.html
4	p1.html, s3.html

よびそのTaxonomyを例示している．5つのURLが2つに分類されており，アイテム「料金表」はアイテムp1. htmlおよびp2. htmlの先祖であり，アイテム「サイト解析」は3つのアイテムs1. html，s2. html，s3. thmlの先祖であることが示されている．

入力データベース　いま複数のアイテム集合から構成されるデータベースD，およびTaxonomy Tについて考える．データベースを構成するアイテム集合のことを特にトランザクションと呼ぶ．アクセスログデータにおけるデータベース例を表3に例示している．TIDで示される4つのトランザクションで構成されており，それぞれのトランザクションにユーザが閲覧したURL（アイテム）が示されている．

支持　あるアイテムxがトランザクションTに含まれるか，もしくはxがトランザクションTのいずれかのアイテムのT上の先祖である場合，「トランザクションTはアイテムxを支持（support）する」という．またトランザクションTがアイテム集合Xの全アイテムを支持しているとき，「トランザクションTはアイテム集合Xを支持する」という．データベースDの全トランザクションのうちσ%のトランザクションがアイテム集合Xを支持するとき，σをXの支持度（*support*）と呼び*Support*(X)で表す．

多頻度一般化アイテム集合　データベースDとTaxonomy Tが与えられたとき，ユーザにより指定された最小支持度σ%以上の支持度を持つアイテム集合を多頻度一般化アイテム集合（Frequent Generalized Itemset）と呼ぶことにする．

例　図2および表3の例を用いて多頻度一般アイテム集合について確認する．いま最小支持度を50%（2つのトランザクションに相当する）とすると，アイテム集合{s1. html，s2. html}はユーザ2，3に支持されており多頻度である．{p1. html，s2. html}はユーザ2のみに支持されているだけであり多頻度ではないが，{料金表，s2. html}はユーザ1，2に支持されており多頻度となる．また{料金表，サイト解析}はユーザ1，2，4に支持されており多頻度である．

表4　URL閲覧データベース例（系列）

系列ID	閲覧したURL(アイテム)
1	$\langle\{p1, s1\}\mathbf{1}\ \{s2\}\mathbf{9}\rangle$
2	$\langle\{p2\}\mathbf{2}\ \{s2\}\mathbf{8}\rangle$
3	$\langle\{s1, s2\}\mathbf{1}\ \{p1, s1\}\mathbf{5}\ \{s2\}\mathbf{9}\rangle$
4	$\langle\{s2\}\mathbf{1}\ \{p1, s1\}\mathbf{2}\{p1\}\mathbf{9}\rangle$

URLの".html"は省略している．

4.2　一般化系列パターン

次に，前節の続きとしてSrikantらが行った研究[7]に従い一般化系列パターンの定義を以下に示す．

用語　系列Sとは，アイテム集合$s \subseteq I$の順序付けられたリストのことで，$S=\langle s_1 s_2 \cdots s_n\rangle$と表す．系列を構成するアイテム集合のことを特に**エレメント**と呼ぶ．

入力データベース　複数の系列から構成されるデータベースDを考える．データベースを構成する系列のことを特にデータ系列と呼ぶ．データ系列を構成するアイテム集合をトランザクションと呼ぶ．同一のトランザクションを構成するアイテムは同一時刻に生起したものとして考える．

アクセスログデータにおけるデータベース例を表4に例示している．4つのデータ系列で構成されており，各データ系列は順序付けられたトランザクションで構成されている．閲覧URLについてトランザクションの後ろに強調表示された数字はトランザクションの時刻を表している．データ系列をユーザに，そしてトランザクションをセッションに対応させると考えやすい．

支持（基本系列）　系列$S=\langle s_1 s_2 \cdots s_n\rangle$とデータ系列$Q=\langle q_1 q_2 \cdots q_m\rangle$についてトランザクション$q_{i_1}$がエレメント$s_1$を支持し，$q_{i_2}$が$s_2$を支持し，…，$q_{i_n}$が$s_n$を支持するような整数$i_1 < i_2 < \cdots < i_n$が存在するとき「データ系列$Q$は，系列$S$を支持する」という．データベース$D$の全データ系列のうち$\sigma$%のデータ系列が系列$S$を支持するとき，$\sigma$を$S$の支持度と呼び，*Support*$(S)$で表す．

例（基本系列）　前節で用いたTaxonomy（図2）および表4を用いて支持の定義を確認する．系列$\langle\{s1\}\{s2\}\rangle$はデータ系列1，3に支持されている[1]．ま

[1] データ系列3については，1つ目と2つ目のトランザクションによる支持と2つ目と3つ目のトランザクションによる支持の2つの対応関係が存在するが，たとえ複数の対応があっても支持度のカウントにおいては一件と扱う．

た，系列 $\langle\{p1\}\{s2\}\rangle$ はデータ系列1，3に支持されているが，$p1$ をその祖先である「料金表」に換えた系列 $\langle\{料金表\}\{s1\}\rangle$ はデータ系列2にも支持される．系列 $\langle\{p1\}\{s1\}\rangle$ を支持するデータ系列は存在しない．

支持（トランザクション時間幅） トランザクションの時間幅の概念を導入することで「支持」を再定義する．これは，異なる時刻のトランザクションであっても，ユーザが指定した時間幅 *win-size* 以下であれば同一のトランザクションと見なそうというものである．

より正確には，ユーザによって *win-size* が与えられたとき，系列 S とデータ系列 Q について，以下の2つの条件を満たすような $l_1 \leq u_1 < l_2 \leq u_2 < \cdots < l_n \leq u_n$ が少なくとも一組存在するとき，「データ系列 Q は系列 S を支持する」という．

1. エレメント（アイテム集合）s_i が $\cup_{k=l_i}^{u_i} q_k$，$1 \leq i \leq n$ に支持される
2. トランザクション時刻 (q_{u_i}) − トランザクション時刻 $(q_{l_i}) \leq$ *win-size*，$1 \leq i \leq n$

なお *win-size*$=0$ とすれば基本系列における支持と同じ意味となる．

例（トランザクション時間幅） 表4を用いて支持の定義を確認する．トランザクション時間幅の概念を導入しなければ，系列 $\langle\{s1, s2\}\{p1\}\rangle$ はデータ系列3にのみ支持されているが，*win-size*$=1$ とすると系列4にも支持されることになる．なぜならば系列4の最初の2つのトランザクションの時間幅が1であるため，$\langle\{s1, s2, p1\}\{p1\}\rangle$ とみなすことができるからである．

支持（トランザクション間時間幅） さらにトランザクション間の時間幅の制約を加えることで「支持」を再定義する．これは，トランザクション間の時間間隔（Gap）が開きすぎていたり逆に狭すぎたりする系列を「支持」の対象から除外しようというものである．

より正確には，ユーザによって *win-size*，*min-gap*，*max-gap* が与えられたとき，系列 S とデータ系列 Q について，前項の2つの条件および以下の2つの条件を満たすような $l_1 \leq u_1 < l_2 \leq u_2 < \cdots < l_n \leq u_n$ が少なくとも一組存在するとき，「データ系列 Q は系列 S を支持する」という．

1. トランザクション時刻 (q_{l_i}) − トランザクション時刻 $(q_{u_{i-1}}) >$ *min-gap*，$2 \leq i \leq n$
2. トランザクション時刻 (q_{u_i}) − トランザクション時刻 $(q_{l_{i-1}}) \leq$ *max-gap*，$2 \leq i \leq n$

なお *win-size*$=0$，*min-gap*$=0$，*max-gap*$=\infty$ とすれば基本系列における支持と同じ意味となる．

例（トランザクション間時間幅） 表4において何の条件もなければ，系列 $\langle\{p1\}\{s2\}\rangle$ は系列1，3に支持されるが，*max-gap*$=6$ の条件を加えることで系列1は支持しなくなる．なぜならば系列1のトランザクション間の時間幅は8であり，*max-gap* より大きくなるからである．また系列 $\langle\{s1, s2\}\{p1\}\rangle$ は *win-size*$=1$ の条件のもとで系列3，4に支持されていたが，*min-gap*$=6$ とすると，系列3の最初の2つのトランザクション間の時間が4のために，この制約条件を満たしておらず，系列3は支持しなくなる．一方で系列4はこの条件を満たしている．

多頻度一般化系列パターン ここで，系列データベース D と Taxonomy T が与えられたとき，ユーザにより指定された最小支持度 σ% 以上の支持度を持つ系列を多頻度一般化系列パターンと呼ぶことにする．

4.3 顕在パターン

データマイニングの分野では，カテゴリ型の値（「クラス」と呼ぶ）をとる目的変数の判別モデルの構築を目的として，相関ルールや多頻度パターンを用いる研究が盛んである．その中でも Don らによって提案された CAEP (Classification by Aggregating Emerging Pattern) [3]は，C 4.5 などの従来の手法に比べ高い精度のモデルを構築できるといわれている．本節では CAEP で用いられる顕在パターンについて述べる．

顕在パターン（Emerging Pattern）とはあるクラスに多頻度で，その他のクラスでは多頻度ではないようなパターンのことである．ここでパターンとは前節にて見てきた一般化アイテム集合，もしくは一般化系列パターンのこととする．

成長率 いま異なる2つのクラスに属するデータベース D_1，D_2 について考える[2]．D におけるあるパターン e の支持度を $Support_D(e)$ で表すと，パターン e の D_2 に対する D_1 の成長率（GR: growth rate）は以下のように定義される．

$$GR_{D_1}(e) = \begin{cases} \dfrac{Support_{D_1}(e)}{Support_{D_2}(e)} & , Support_{D_2}(e) \neq 0 \\ \infty & , Support_{D_2}(e) = 0 \end{cases} \quad (1)$$

例 表5を用いて成長率について確認する．表5は2つのデータベース申込（D_1）と未申込（D_2）が示されている．アイテム集合 $e_1=\{p1, s2\}$ および $e_2=\{s1, s3\}$ の D_1 および D_2 についての成長率の計算が式(2)(3)に

[2] 3つ以上のクラスへの拡張も容易であるが，ここでは簡単のために2つのクラス $c=1,2$ を前提として論じる．

表5　URL閲覧データベース例

申込 ($\mathcal{D}_1$)		未申込 ($\mathcal{D}_2$)	
TID	閲覧したURL	TID	閲覧したURL
1	p1, p2, s2	1	s1, s2, s3
2	p1, s2, s3	2	s1, s3
3	s1, s3	3	p1, s2, s3
4	p1, s2		

それぞれ示されている．

$$GR_{\mathcal{D}_1}(e_1)=\frac{Support_{\mathcal{D}_1}(\{p1, s2\})}{Support_{\mathcal{D}_2}(\{p1, s2\})}=\frac{3/4}{1/3}=2.25 \quad (2)$$

$$GR_{\mathcal{D}_2}(e_2)=\frac{Support_{\mathcal{D}_2}(\{s1, s3\})}{Support_{\mathcal{D}_1}(\{s1, s3\})}=\frac{2/3}{1/4}=2.67 \quad (3)$$

アイテム集合 e_1 は未申込ユーザより申込ユーザに2.25倍出現しやすいパターンであり，e_2 は2.67倍未申込ユーザに出現しやすいパターンであるといえる．

顕在パターン　ここで，異なるクラスに属するデータベース $\mathcal{D}_1$，$\mathcal{D}_2$ が与えられたとき，ユーザにより指定された最小支持度 σ，および最小成長率 ρ について，$Support_{\mathcal{D}_1}(X)\geq\sigma$ かつ $GR_{\mathcal{D}_1}(e)\geq\rho$ を満たすパターン e をクラス1の顕在パターン（Emerging Pattern）と呼ぶ[3]．特にパターンが一般化アイテム集合の場合**一般化顕在パターン**（Generalized Emerging Pattern）とよび，また一般化系列パターンの場合**一般化顕在系列パターン**（Generalized Emerging Sequence Pattern）と呼ぶことにする．さらに，一般化顕在パターンと一般化顕在系列パターンを組み合わせたパターンの場合，**統合化顕在パターン**（Integrated Emerging Pattern）と呼ぶことにする．

4.4　パターンの剪定

以上の方法で顕在パターンを列挙すると多くの冗長なパターンが生成されることがわかっている．祖先－子孫の関係にあるTaxonomyを含むパターン[6][7]，および包含関係にある顕在パターン[3]の2つで，これら冗長なパターンを剪定（pruning）する方法を以下に示す．

R-Interesting パターン　アイテム $\hat{x}$ をアイテム x の先祖としたとき，同数のアイテム数で構成される2つのアイテム集合 X，$\hat{X}$ について，X のいくつかのアイテムをその先祖で置換して $\hat{X}$ が得られるとき，アイテム集合 $\hat{X}$ はアイテム集合 X の先祖であるという．特に先祖が親である場合，近接先祖（close ancestor）という．いま，$\hat{X}$ をアイテム集合 X の近接先祖とするとき，X の生起確率 $P(X)$ が近接先祖のアイテム集合 $\hat{X}$ の生起確率 $P(\hat{X})$ から計算される期待値の R 倍のとき，X はR-interestingであるという．ユーザにより与えられたRの値以下であるR-interestingなパターンは剪定する[4]．系列についても同様の剪定方法が適用できる．

必須顕在パターン　包含関係にある2つの顕在パターン $e\subset x$ を考える．このとき $Support(e)\geq Support(x)$ であるので，もし $GR(e)\geq GR(x)$ であれば e は x よりもカバーする範囲は広くかつ判別能力も高い．よってこのような条件を満たす x は剪定される．さらに $GR(e)<GR(x)$ であっても，e の支持度が相対的に十分高く[5]，かつ e の成長率がある程度高ければ[6]そのような x は剪定される．残った顕在パターンのことを必須顕在パターンと呼ぶ．

4.5　CAEP

顕在パターンは，その定義からみても明らかなように，データベースの一部において高い判別能力を有するルールである．以下ではDongらにより提案された顕在パターンを利用した判別モデルの構築手法CAEP（Classification by Aggregating Emerging Patterns）[3]について見てみる．

CAEPでは，訓練データにより列挙されたクラス $c=1, 2$ についての顕在パターン集合 $E(c)$ を利用して未知のインスタンス u を判別する．このとき，u に支持される各クラスの顕在パターン $e\in E(c)$ による重み付き投票で判別を行う．

集約スコア　未知データ u がクラス c と判別される確信度を集約スコア $score(u, c)$ として定義している（式(4)）．

$$score(u, c)=\sum_{e\subseteq u, e\in E(c)}\frac{GR_{\mathcal{D}_c}(e)}{GR_{\mathcal{D}_c}(e)+1}\times Support_{\mathcal{D}_c}(e) \quad (4)$$

集約スコアは，未知のインスタンス u がパターン e を含むときに u のクラスが c である条件付確率に e の支持度を重みとしてかけ合わせた値の総和になっている．条件付確率について詳しく見てみる．

パターン e を含むときに u のクラスが $c=1$ である条件付確率は，ベイズの定理より以下のように表される．

[3] Dongらによる定義ではサポートによる条件を含んでいないが，ここでは計算時間の観点から最小支持度 σ の条件をつけた．

[4] 本研究では $R=2$ で固定して実験を進めた．

[5] $\frac{Support(e)}{Support(x)}\geq\sigma'$ の条件により判定する．本研究では $\sigma'=30$ とした．

[6] $GR(e)\geq\rho'$ の条件により判定する．本研究では $\rho'=20$ とした．

$$P(\mathcal{D}_1|e)=\frac{P(e|\mathcal{D}_1)P(\mathcal{D}_1)}{P(e|\mathcal{D}_1)P(\mathcal{D}_1)+P(e|\mathcal{D}_2)P(\mathcal{D}_2)} \quad (5)$$

ここで，$GR_{\mathcal{D}_1}(e)=P(e|\mathcal{D}_1)/P(e|\mathcal{D}_2)$ なので上式は，

$$P(\mathcal{D}_1|e)$$

$$=\frac{GR_{\mathcal{D}_1}(e)P(e|\mathcal{D}_2)P(\mathcal{D}_1)}{GR_{\mathcal{D}_1}(e)P(e|\mathcal{D}_2)P(\mathcal{D}_1)+P(e|\mathcal{D}_2)P(\mathcal{D}_2)} \quad (6)$$

$$=\frac{GR_{\mathcal{D}_1}(e)}{GR_{\mathcal{D}_1}(e)+\frac{P(\mathcal{D}_2)}{P(\mathcal{D}_1)}} \quad (7)$$

となる．ここで $P(\mathcal{D}_1)=P(\mathcal{D}_2)$，すなわち2つのクラスの生起確率が同等と仮定すると，式(4)の中の条件付確率（分数の部分）が導かれる．

式(4)に示される *score* 値は，クラスごとの件数に大きく影響を受けるため，式(8)に示される基準化した値（normalized score）を最終的に用いている．

$$norm\_score(u,c)=\frac{score(u,c)}{base\_score(c)} \quad (8)$$

ここで $base\_score(c)$ は，訓練データの全ケースの集約スコアの中央値である．そして未知のインスタンス u について $norm\_score(u,c)$ が最も大きいクラス c を予測クラスと判別する．

文献[3]ではアイテム集合による顕在パターンに基づいたモデルが示されているが，一般顕在パターンや一般化顕在系列パターン，そして統合化顕在パターンに基づいたモデルにも容易に拡張可能であり，特に統合化顕在パターンに基づいた分類モデルを CAIEP (Classification by Aggregating Integrated Emerging Pattern) と呼ぶことにする．

4.6 カバー率

CAIEP にユーザが与えるパラメータは最小支持度 σ および最小成長率 ρ である．両パラメータともに大きな値を指定すると列挙されるパターン数は少なくなり，それに伴い列挙されたいずれのパターンも含まない（支持しない）インスタンス（トランザクションもしくはデータ系列）も増えることとなる．本研究では，このようなインスタンスについては沈黙し判別不能として扱う．そして少なくとも1つのパターンを含む判別可能なインスタンスの全インスタンスに対する割合をモデルの**カバー率**と呼ぶことにする．

5. データセットについて

CAIEP を Web アクセスログデータに適用するに当たって利用したデータセットについて述べる．

5.1 アイテムと Taxonomy

今回の分析で利用したアイテムおよびその Taxonomy の一覧を表6に示す．表中の LV は Taxonomy 階層レベルを示している．レベル0が最小単位のアイテムに相当し，数字が大きい方がより一般的な分類概念となる．

アイテムの最小単位としての URL は，その URL の属するディレクトリ名（DIR）が親で，DIR の親が Web ページ上で分類されたメニュー（MN）となる（表1参照）．外部リファラである FRM は，検索エンジンとリスティングの組み合わせでアイテムを構成しており，その親として検索エンジン（SEG）とリスティング（LST）を同レベルの分類階層として位置づけている．検索語（KW）は，例えば大文字と小文字の違い（SEO と seo など）などの表記上異なるが意味は同じと判断したものをまとめて親の分類階層を作成した（KW_TX）．曜日（DOW）は月から金までを平日として土日を休日として分類した（DOW_TX）．そして最後に時刻（TM）は1時間おきに00時から23時までの24アイテムを用意し，9時から18時までを勤務時間内，それ以外を勤務時間外と分類した（TM_TX）．

5.2 データセット

今回の分析では，URL のみを系列データとして利用し，その他のアイテムはすべてアイテム集合として利用することにした．アイテムは「項目略号＝値」の形式で表現している．表7にデータセット例を示す．

データ系列については，直帰・滞在判別モデルにお

表6 アイテム-Taxonomy 一覧

	項目略号	LV	値 (例)
URL	URL	0	/, /top.html, /contact/apply.html
	DIR	1	ADV, CONTACT, ETC, EX, PATH
	MN	2	機能概要, シビラの導入
外部	FRM	0	Google_adwards, YahooSearch_Non
	SEG	1	GoogleSearch, YahooSearch
	LST	1	Overture, adwords, Non
検索語	KW	0	アクセス解析, SEO, seo
	KW_TX	1	解析関連, SEO 関連
曜日	DOW	0	月, 火, 水, 木, 金, 土, 日
	DOW_TX	1	平日, 休日
時刻	TM	0	00,01,02,…,21,22,23
	TM_TX	1	勤務時間内, 勤務時間外

表7 データセット例

UID	UNIX 時刻	系列	アイテム集合
uid01	1138333340	URL=/top.html	FRM=GoogleSearch_Non DOW=水 TM=14
	1138333560	URL=/page/	
	1138335112	URL=/price/	
	1138335160	URL=/	
	1138335222	URL=/top.html	
uid02	1138333748	URL=/	FRM=YahooSearch_Non KW=ログ解析 KW=シビラ DOW=日 TM=21
	1138333749	URL=/top	
	1138333824	URL=/topform.html	
	1138333991	URL=/price/	
	1138333995	URL=/price/	
	1138334091	URL=/page/	
:	:	:	:

いては，系列のトップ行だけを選択する．申込・未申込判別モデルではトップ10回のデータを選択する．ただし，10回以上のアクセスのないユーザは対象外としている．直帰・滞在判別モデルにおいては，系列データとして初回のURLしか扱わないので，実質的には一般化顕在パターンのみの探索と同じ結果となる．

クラスについてであるが，直帰・滞在判別モデルでは2回目のアクセスがあったかどうかでデータセットを D_1，D_2 にクラス分割している．また申込・未申込判別モデルにおいては，11回目以降のアクセスに申込を完了する一連のURLがあるかないかによりデータセットを D_1 と D_2 にクラス分割した．また申込の系列からは，申込を完了する一連のURLはデータから削除している．

以上のように作成されたデータセットを入力として，ユーザが与えた最小支持度や最小成長率などの条件を満たす顕在パターンが全列挙され，判別モデルの構築が行われる．

6. 結果

最小サポート σ，最小成長率 ρ，*win-size*，*min-gap*，*max-gap* それぞれの値を適当な間隔で設定し，それらの組み合わせをパラメータとしてモデルを構築した．それぞれのモデルについて，モデル精度の指標としてのF-measure（後述），カバー率，ルール数を表したのが図3である．1つのプロットが1つのモデルに対応している．円の大きさはルール数を表している．

分析者は，図3よりカバー率とF-measureについて優劣をつけることができないパレート最適解の中のいずれかを目的に応じて選択し利用することになる．

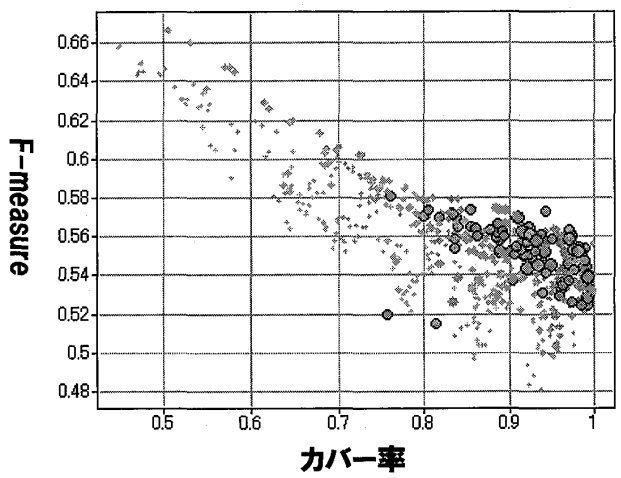

図3　直帰・滞在判別モデルで得られた全モデル

このような方法でユーザのページ遷移判別モデルを構築し，得られたパレート最適なモデルのうち適当に選んだ1つのモデルについて，その結果を以下に示し，得られたパターンの考察を行う．

6.1 直帰・滞在判別モデル

σ=20%，ρ=2.0の条件で得られた直帰・滞在判別モデルについて，全パターンを図4に，そしてConfusion Matrixを表8に示す．

このモデルは時系列情報を含んでいないので，多頻度一般化アイテム集合としてのパターンのみが示されている．1つの行はパターンを構成する1つのアイテムに対応している．1つの列が1つのパターン（多頻度一般化アイテム集合）に対応しており（a～wのパターンIDを付与している），網掛けで示された行に対応するアイテムをもつパターンとして示されている．例えば，アイテム集合a＝{DOW_TX＝休日，FRM＝YahooSearch_Non，KW＝シビラ，URL＝/}である．aからgは「直帰」に特徴的な顕在パターンであり，hからwは滞在に特徴的な顕在パターンである．濃い網掛けは，いずれかのクラスにのみ出現するアイテムであることを表している．例えばアイテム「T

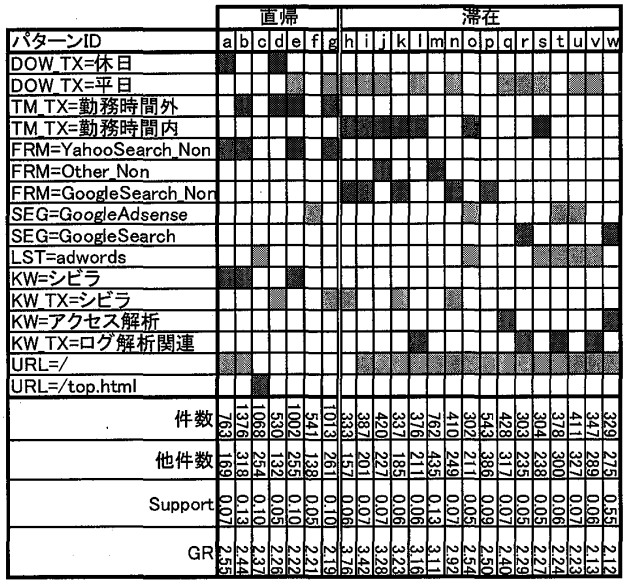

図4　直帰・滞在判別モデル

表8　Confusion Matrix

		判別されたクラス			計
		滞在	直帰	判別不能	
実際のクラス	滞在	2,056	770	3,104	5,930
	直帰	1,370	2,967	6,183	10,520
計		3,426	3,737	9,287	16,450

推定F-measure(滞在)=0.657, 推定カバー率=43.5%
(10-fold cross validation 法による)
滞在ルール数=7, 直帰ルール数=16

M_TX＝勤務時間外」は直帰のルールにのみ出現している．

図4から得られた知見を以下に示す．

知見1　検索語「シビラ」は休日や勤務時間外であれば直帰確率が高く，平日や勤務時間内であれば滞在確率が高くなる（a, b, d, e, g, h, k, l）．これは企業ユーザによる仕事時間内でのアクセスがメインであることがわかる．また他の分野で「シビラ」という名称が存在しており，その検索が休日や勤務時間外に行われることが多いと推察される．

知見2　adwordsでサイトにアクセスするユーザは滞在する確率が高い一方で，adwordsによりtop.html（キャンペーン販売のページ）へアクセスするユーザは(c)直帰確率が高くなる．adwordsによるキャンペーンへの誘導がうまく機能していないことがわかる．

知見3　yahooによる検索でアクセスするユーザは直帰確率が高く（a, b, e, g），googleによる検索でアクセスするユーザは滞在確率が高い（h, i, k, n, p）．

知見4　「ログ解析」関連もしくは「アクセス解析」のキーワードは有効である（l, q, r, t, v, w）ことがわかる．

6.2　申込・未申込判別モデル

σ=20%，ρ=2.0，*win-size* =20，*min-gap* =0の条件で得られた申込・未申込判別モデルについて，「申込」に特徴的な顕在パターンの一部を図5に，「未申込」に特徴的な顕在パターンの一部を図6にそれぞれ示している．そしてConfusion Matrixを表9に示す．

このモデルはパターンとして一般化アイテム集合と一般化系列パターンの両方を含む．例えば，図5のパターンbは，「勤務時間内」に「URL＝/」→「URL＝/top.html」の順で巡回するパターンが11名の申込者に含まれており，未申込者より5.3倍起こりやすいパターンであることを示している．

図5から得られる知見を以下に示す．

知見5　「価格情報」，「申込フォーム」，「申込の流れ」など申込に特有なページへの複数アクセスが申込につながる傾向が強い．

知見6　多くのパターンにおいて，「平日」と「勤務時間内」のアイテムが含まれており，企業からの申込が多いことが推察される．

図6から得られる知見を以下に示す．

知見7　多くのパターンにおいて1つのエレメント

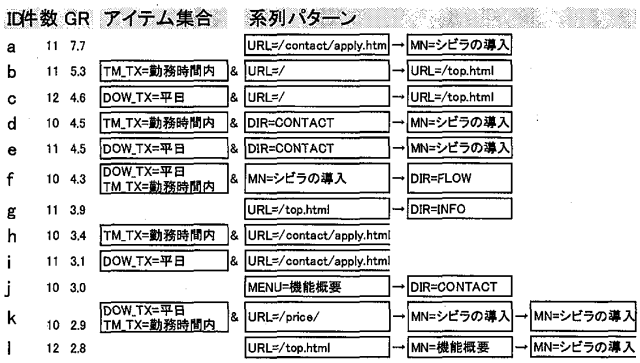

図5　申込・未申込判別モデル（申込）

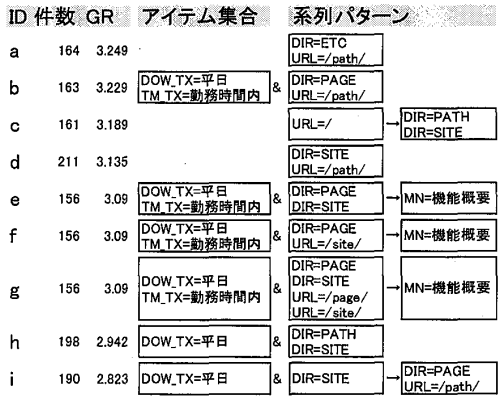

図6　申込・未申込判別モデル（未申込）

表9　Confusion Matrix

		判別されたクラス			計
		申込	未申込	判別不能	
実際のクラス	申込	32	14	0	46
	未申込	159	522	93	774
計		191	536	93	820

推定 F-measure(滞在)=0.270, 推定カバー率=88.8%
(10-fold cross validation 法による)
申込ルール数=25, 未申込ルール数=32

に複数のアイテムが含まれている．win-sizeが20秒であることを考えると，内容をじっくりと読んでいないことの表れであり，平日の勤務時間内のアクセスであっても，そのようなユーザは未申込である傾向が強いといえる．

7.　モデル性能の評価

7.1　他手法との比較

本研究で提案したCAIEPの有効性を見るために，他の判別モデル構築手法との性能比較を行う．構築したモデルは時系列情報を含んだ申込・未申込判別モデルで，比較した手法はC4.5，Naive Bayes，k-Nearest Neighbours法の3つである[7]．CAIEPにつ

[7] データマイニングソフトウェアWeka（Ver. 3.4）を利用した．

いては，前節と同じパラメータによって実行した．

いずれの手法においてもコスト考慮型学習法（cost-sensitive learning）を利用し，正事例と負事例のコスト評価を同等にしてモデル構築を行い，10-fold cross validation 法により精度推定を行った．

精度評価には，再現率（Recall），精度（Precision），F-measure の 3 つの指標を利用する．ここで Recall とは，表 10 に示される Confusion Matrix において，Positive クラスに属するケース全体のうち正しく判別されたケースの割合のことである（式(9)）．一方，Precision とは式(10)で定義される通り，Positive クラスと判別したケースのうち正しく判別されたケースの割合である[8]．一般的に Recall と Precision はトレードオフの関係にあることが知られている．そして，これらトレードオフの関係にある Recall と Precision を α による加重調和平均で評価した指標が F-measure であり，式(11)にその定義が示されている．今回は $\alpha=0.5$ として評価を行った．

$$Recall=\frac{TP}{TP+FN} \tag{9}$$

$$Precision=\frac{TP}{TP+FP} \tag{10}$$

$$F\text{-}measure=\frac{1}{\alpha\frac{1}{Recall}+(1-\alpha)\frac{1}{Precision}} \tag{11}$$

手法間の性能比較にあたっては，CAIEP 以外の手法が直接時系列データを扱うことができないため，単純な比較ができない．そこで 1)時系列データの利用に伴う性能差，および 2)判別におけるルールの集約方法についての性能差，の 2 つについて見ていく．なお，いずれの比較においても「申込」クラスの判別精度についての性能について検証する．

まずは時系列データを扱えることで，どの程度の精度の高いモデル構築が可能かを比較する．CAIEP 以外の手法では，今回利用したすべてのアイテムについて，そのアイテムを含むかどうかをダミー変数に変換して説明変数として利用した．表 11 にその結果が示されている．Recall, Precision, F-Measure いずれにおいても CAIEP が他の手法を上回っており，時系列データを扱うことのできる優位性が示されている．

次に，CAIEP の中で統合化顕在パターンとして列挙されたルールを所与とした場合に，CAIEP で採用されている集約スコアによる方法と他の 3 手法との性能を比較検証する．CAIEP 以外の手法では，CAIEP によって得られた全ルールについて各ケースがそれらのルールを含むかどうかをダミー変数に変換し説明変数として利用した．

表 12 にその結果が示されている．CAIEP による判別の方法が概ね他の手法より優れているが，表 11 に示される結果ほどの差は認められない．特に Naive Bayes は CAIEP に迫る性能を示している．これらのことより，CAIEP は，判別方法というよりむしろ時系列データを扱えることに，その優位性があることが示されたといえよう．

さらにリフトチャートを図 7 に示す．横軸に予測確率の高い順に並べたときの正事例（「申込」クラス）と予測する上位の人数を，縦軸に Recall を示している．全体にわたって CAIEP が他手法を上回っている

表 10　Confusion Matrix

		判別されたクラス	
		Positive	Negative
実際のクラス	Positive	TP	FN
	Negative	FP	TN

表 11　他手法との精度比較

	CAIEP	C4.5	NB	kNN
Recall	0.696	0.217	0.413	0.391
Precision	0.163	0.081	0.089	0.072
F-measure	0.264	0.118	0.147	0.122

表 12　他手法との精度比較

	CAIEP	C4.5	NB	kNN
Recall	0.696	0.348	0.609	0.174
Precision	0.163	0.143	0.163	0.111
F-measure	0.264	0.203	0.257	0.136

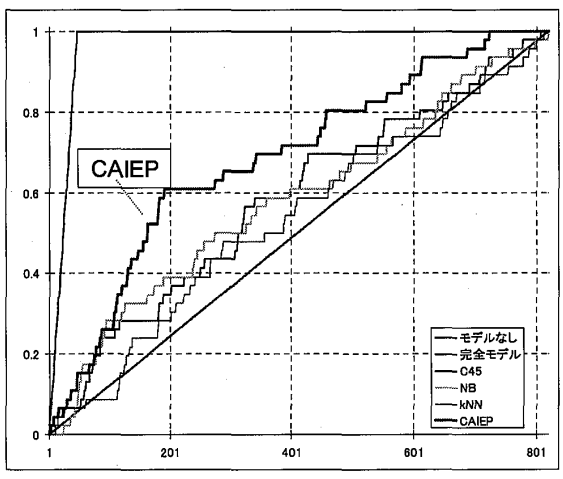

図 7　リフトチャート

[8] 判別不能のケースについては，各クラスの出現割合に従ってカウントしている．

表13　CAIEP機能別の精度比較

		取り除いた機能			
	CAIEP	SQ	IS	TX	TM
Recall	0.696	0.385	0.698	0.533	0.696
Precision	0.163	0.091	0.157	0.148	0.122
F-measure	0.264	0.148	0.256	0.232	0.208

SQ:系列パターン,IS:アイテム集合,TX:Taxonomy,TM:時間制約

ことが確認できる.

7.2　感度分析

CAIEPは，アイテム集合，系列パターン，Taxonomy, 時間制約といった様々な機能の組み合わせで構成されている．そこで申込ユーザ判別モデルにおいて，どの機能が有効に機能しているかを確かめるために感度分析を行った．

方法は，アイテム集合，系列パターン，Taxonomy, 時間制約のそれぞれの機能を省くことで判別モデルを再構築し，Recall, Precision, F-measureの3つの指標についてオリジナルのモデルと比較した（表13）.

F-measureについて見てみると，一般化系列パターンを省いたときに精度の落ち込みが最も大きく，次に時間制約が続いている．一方で，アイテム集合とTaxonomyを除いても比較的精度の落ち込みは小さい．以上のことより，対象とするデータの内容にも大きく依存するであろうが，Webアクセスログについていえば，ユーザの行動はURLの遷移という時系列パターンとして現れてくる傾向が強いということができるであろう.

8.　おわりに

本研究では，Webアクセスログデータを用いて，ユーザのWeb巡回行動についてページ遷移判別モデルを構築した．そして判別モデル構築手法として頻出パターンマイニングの分野で開発されてきた各種手法を組み合わせたCAIEP（統合化顕在パターン判別モデル）を開発した．Webアクセスログのようなカテゴリ値の時系列データに対して有効性が高いことを示した．また構築したモデルを解釈することで，いくつかの興味深い知見も得られ，モデルを構成するルール(パターン）の可読性の高さも示すことができた.

一方で研究を進める中でいくつかの課題もわかってきた．1つは多くの機能を組み合わせているが故に，ルールの視覚化が難しく，特にモデルに多数のルールが含まれる場合には，その傾向が顕著になる．さらにユーザが調整可能なパラメータが多く，推定精度を最も高める最適なパラメータを導出する効率的な方法も求められる．さらに，今回は多頻度パターンの列挙にApriori アルゴリズム[1]を利用したが，近年，パターン列挙について飽和集合に基づくより高速な方法が提案されており[8]，その点についての改良も行っていきたい.

謝辞　本稿で用いたいくつかのグラフは，データ視覚化ソフトSpotfire（Spotfire Japan株式会社より提供）により作成した．本研究の一部は，平成18年度科学研究費補助金（基盤研究(b)）により行った.

参考文献

[1] R. Agrawal and R. Srikant: "Fast algorithms for mining association rules," *Proc. of the 20th Int'l Conference on VLDB*, pp. 487-499 (1994).

[2] R. Agrawal and R. Srikant: "Mining sequential patterns," *Proc. of ICDE-95*, pp. 3-14 (1995).

[3] G. Dong, X. Zhang, L. Wong and J. Li: "CAEP: Classification by aggregating emerging patterns," *Proc. of the 2nd Int'l Conference on Discovery Science*, pp. 30-42 (1999).

[4] O. Etzioni: "The World Wide Web: quagmire or gold mine?," *Communications of the ACM*, 39(11), pp. 65-68 (1996).

[5] R. Kosala and H. Blockeel: "Web mining research: A survey," *SIGKDD Explorations*, 2(1), pp. 1-15 (2000).

[6] R. Srikant and R. Agrawal: "Mining Generalized Association Rules," *Proc. of the 21st Int'l Conference on VLDB*, pp. 407-419 (1995).

[7] R. Srikant and R. Agrawal: "Mining Sequential Patterns: Generalizations and Performance Improvements," *Proc. 5th Int'l Conference on EDBT*, pp. 3-17 (1996).

[8] T. Uno, T. Asai, Y. Uchida and H. Arimura: "An Efficient Algorithm for Enumerating Closed Patterns in Transaction Databases," *Lecture Notes in Artificial Intelligence 3245*, pp. 16-31 (2004).