2C1-03

インターネットユーザ間の長期にわたる興味遷移パターンの抽出と比較

Extracting and Comparing The Transition Patterns of The Internet User Interests From The Web Access Logs Recorded for A Long Period

山田和明[*1] 中小路久美代[*1] 上田完次[*2]
Kazuaki Yamada Kumiyo Nakakoji Kanji Ueda

[*1]東京大学先端科学技術研究センター RCAST, University of Tokyo
[*2]東京大学人工物工学センター RACE, University of Tokyo

This paper proposes the analysis method of Web access log data recorded for long time period in order to understand how interests of internet users alter with time. The goal of our research is to model the transition processes of user interests and find out the invisible relationships among users. In this study, we made WebPAC Viewer and Time Clip Viewer to extract the transition pattern of the user interests, and we evaluated these viewers through extracting the transition patterns of an internet user interests.

1. はじめに

インターネットユーザの増加にともない Weblog[Takeda04], インターネットオークション，ソーシャルネットワークサービス [Mori04] など，インターネットを利用した新しいサービスが生まれている．現在，このようなサービスを利用しているユーザがどのようなことに興味を持っているのか，また，ユーザ間にどのようなコミュニティが形成されているのか，という隠れたニーズや関係を発見するための研究が行われている [Taniguchi04][Numa04]．これらの研究では，ある特定のサービスを利用しているユーザを対象とした研究が多い．しかし，一般にユーザは内容の異なる複数の Web サイトにアクセスしているため，インターネットユーザの日常の活動を正しく把握できていないのが現状である．

今後，インターネットユーザがどのようなことに興味を持ち，どのように Web サイトを活用しているのか，また，Web の活用方法とユーザ属性の相関関係を調べることはインターネットサービスの新しいモデルを構築する上で重要になると考えられる．そこで本稿では，長期間にわたり記録された個々のインターネットユーザのウェブ・アクセスログを解析することで，各ユーザの興味がどのように変化しているかモデル化する方法を提案する．

2. 関連研究：複雑なシステム挙動の解析

社会システムに見られる複雑な挙動（例えば，経済活動，交通システム，オープンソースソフトウェア開発コミュニティなど）は，ユーザ間の相互作用によって発生する複雑なダイナミクスに起因する．このような複雑な現象を解析する研究が数多く行われ，これらの研究は解析対象をマクロな視点から，あるいは，ミクロな視点から捉えるのか，という視点の位置により以下の３つに大別できる．

(1) システム全体の挙動を統計的手法により数値化し表現する方法．経済指標や交通量，Web アクセスログ解析などがあげられる．

(2) システムの構成要素間の関係からシステム挙動を表現する方法．Web サイト間のリンク構造から関連するコミュニティを抽出する研究 [Numa04][Otsuka03] などがあげられる．

(3) マルチエージェントアプローチ [Namatame98] により個々の構成要素をモデル化しボトムアップ的にシステム挙動をシミュレートする方法．公共財問題におけるゲーム理論，ソーシャルネットワークの成長過程のモデル化 [Madey03] などがあげられる．

統計的手法による表現では，現在のシステムダイナミクスの状態を表すことができるが，ダイナミクスの発生メカニズムを解明することはできない．一方，マルチエージェントアプローチでは，個々の行動主体の意思決定過程をシミュレートし将来起こり得る確率の高い事象を模擬することができるが，現在のシステムダイナミクスの状態を表すことはできない．そのため，インターネットユーザの興味が，時間とともにどのように遷移しているかを解析するには，ユーザがアクセスした Web サイト間の関係から行動パターンを記述する方法が有用であると考えられる．

要素間の関係を表現する方法としてグラフによる可視化手法が用いられる．例えば，Fisher ら [Fisher04] はメールの送受信関係からコミュニティに参加しているユーザ間の関係を抽出し，グラフ表現を用いてソーシャルネットワークを可視化することでユーザ間の隠れた関係を抽出している．しかし，この手法では 2 次元平面に配置されたユーザの座標は固定ではなく，ユーザ間の関係により座標が変化する．そのため，時間とともに変わる要素間の関係を可視化するには不向きである．

本研究では，解析対象とする各要素を 2 次元平面に射影し，要素間の関係が時間とともにどのように変化するのか観察する可視化方法を提案する．また，2 次元平面に射影された要素をクラスタリングすることで解析対象の特徴的な集合を抽出する方法を提案する．

3. WebPAC

本研究では，一般的なユーザのインターネット利用状況を解析するためにインターネットユーザの Web へのアクセス状況を記録した WebPAC データを用いる．WebPAC データは，

連絡先: 山田 和明, 東京大学先端科学技術研究センター, 153-8904 東京都目黒区駒場 4-6-1, Tel/Fax:03-5452-5288, e-mail:yamada@kid.rcast.u-tokyo.ac.jp

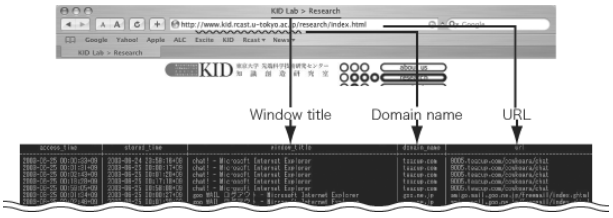


図 1: WebPAC database

ビデオリサーチ社が 2002 年 11 月に 5946 人のユーザを対象に実施したインターネット利用者の特性調査データであり，ユーザの「Web 視聴データ（web browsing data)」と「ユーザ特性調査データ（user profile data)」から構成されている．

「Web 視聴データ」は約 1 ヶ月間の各ユーザが閲覧した Web 履歴を記録したデータで，アクセス時間，閲覧時間，ウィンドウタイトル，URL，ドメインネームなどを記録している．一方,「ユーザ特性調査データ」は，各ユーザのプロファイルデータで，年齢，性別，職業，PC の利用状況，情報源，興味のあるブランドなどを記録している．本稿では，WebPAC データの「Web 視聴データ」の解析を行う．

図 1 に示すように,「Web 視聴データ」は，Web へのアクセス時間，閲覧時間，ウィンドウタイトル，URL，ドメインネームなどから構成されており，ここからユーザがインターネットを利用する，目的，頻度，時間帯などを解析することができる．なお，1 ヶ月の間に実際に Web にアクセスしたユーザ数は 5946 人中 3667 人であった．

4. Web アクセスログからの特徴抽出

ここでは，Web サイト間の関係を 2 次元平面に射影し，時間とともにその関係がどのように変化するのか可視化する．そして，2 次元平面に射影された Web サイトを類似サイトごとにクラスタリングすることでユーザの興味を抽出する方法について説明する．

4.1 単語ベクトル

Web サイトの特徴を記述するために単語ベクトルを作成する．単語ベクトルは，あるユーザが閲覧した Web サイトのウィンドウタイトルから単語を抽出して作成する．例えば，あるユーザが Yahoo auction で iPod を閲覧した場合，ウィンドウタイトルは”Apple iPod - Yahoo! Auction”である．これを形態素解析ツール MeCab[MeCab] を用いて「Apple」,「iPod」,「Yahoo」,「Auction」を単語として抽出する．この操作をユーザが閲覧した全ての Web サイトに対してに行い，単語ベクトル $V = (\mathrm{Yahoo}, \mathrm{Auction}, \mathrm{Apple}, \mathrm{iPod}, \ldots)$ を作成する．単語ベクトルの次元はウィンドウタイトルから抽出された全単語数である．

次に，各単語に重みを付けるために各単語の TF-IDF 値 $w(t,d)$ を次式により計算する．

$$w(t,d) = tf(t,d) \cdot idf(t) \quad (1)$$

$$idf(t) = log\left(\frac{N}{df(t)}\right) \quad (2)$$

ただし，$tf(t,d)$ は文書 d における単語 t の頻度であり，N は全文書数，$df(t)$ は単語 t が 1 回以上出現する文書数である．本稿における文書とは，Web サイトの閲覧間隔が 30 分以上離れている場合を 1 つの文書とする．以上の操作により個々の Web サイトの特徴を記述する．

4.2 x-means 法による自動クラスタリング

ここでは，ユーザが閲覧した Web サイトをクラスタリング手法によって分類し，Web サイトの特徴抽出を行う．しかし，Web サイトを表現する単語ベクトルは多次元のため，各 Web サイトは空間に点在する．そのため，適切なクラスタ数を決定し，類似 Web サイトの分類を行うことは極めて困難である．そこで，本稿では多次元の単語ベクトルを主成分分析により 2 次元まで次元圧縮し，自動的にクラスタ数を決定することができる x-menas 法 [Ishioka00] によりクラスタリングを行う．x-means 法は，情報量規準の一つである BIC（Bayesian Information Criterion）を用いることで，各サブクラスタにおいて分割が妥当と判断されるまで二分割を繰り返すアルゴリズムである．以下に x-mean 法の手順を示す．

1 初期クラスタ数 k_0 を決定.

2 $k = k_0$ として k-means を適用し，クラスタ $C_1, C_2, \ldots, C_{k_0}$ を生成.

3 $i = 1, 2, \ldots, k_0$ とし，手順 4-9 を繰り返す.

4 クラスタ C_i を k-means により 2 つのクラスタ C_i^1, C_i^2 に分割.

5 C_i に含まれるデータ x_i の多変量正規分布（(3) 式）を仮定し，そのときの BIC を (5) 式から求める.

6 各 C_i^1, C_i^2 をパラメータ θ_i^1, θ_i^2 を持つ多変量正規分布で表す．二分割モデルにおいてデータは (4-a) 式の確率密度に従うと仮定し，二分割モデルの BIC を (6) 式により計算.

7 BIC > BIC' のとき，二分割を継続．$C_i \leftarrow C_i^1$ とし，C_i^2 はデータ，クラスタ重心，対数尤度をスタックに記録し，手順 4 へ戻る.

8 BIC $\leq$ BIC' のとき，C_i^1 の二分割を止め，スタックからデータを取り出し，$C_i \leftarrow C_i^2$ とし手順 4 へ戻る．スタックが空きなら次の手順へ進む.

9 C_i における二分割が全て終了．上記の手順で作成されたクラスタが C_i 内で一意になるようにデータの属するクラスタ番号を振り直す.

10 初めに k_0 分割したクラスタ全てについて二分割が終了．全データに対してそれらの属するクラスタ番号が一意になるように番号を振り直す.

11 全データの属するクラスタ番号，各クラスタの重心を出力.

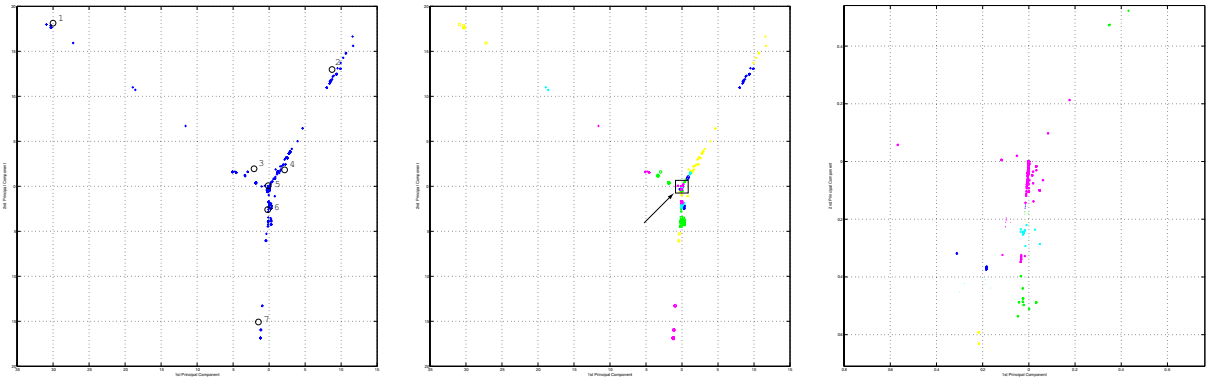
以下にクラスタリングで用いた式を示す.

$$f(\theta_i; x) = (2\pi)^{-\frac{p}{2}} |V_i|^{-\frac{1}{2}} \exp\left[-\frac{(x-\mu_i)^t V_i^{-1}(x-\mu_i)}{2}\right] \quad (3)$$

ただし，$\hat{\theta}_i = [\hat{\mu}_i, \hat{V}_i]$ は，多変量正規分布の最尤推定値であり，μ_i は平均値ベクトル，V_i は分散共分散行列である.

$$g(\theta'; x) = \alpha_i \left[f(\theta_i^1; x)\right]^{\delta_i} \left[f(\theta_i^2; x)\right]^{1-\delta_i} \quad (4\text{-a})$$

$$\alpha_i = 0.5/K(\beta_i) \quad (4\text{-b})$$



(a) Web access logs mapped to 2D space

(b) Web access logs classified by x-means

(c) Web access logs classified by x-means

図 2: Web access logs on 2D space

$$\beta_i = \sqrt{\frac{\|\mu_1 - \mu_2\|^2}{|V_1| + |V_2|}} \quad \text{(4-c)}$$

$$\delta_i = \begin{cases} 1 & x_i \in C_i^1 \\ 0 & x_i \in C_i^2 \end{cases} \quad \text{(4-d)}$$

ここでは，計算を簡単にするため α_i を (4-b) 式で近似している．また，$K(\cdot)$ は標準正規分布の下側確率である．

$$\mathrm{BIC} = -2\log L(\hat{\theta}_i; x \in C_i) + q \log n_i \quad (5)$$

$$\mathrm{BIC'} = -2\log L(\hat{\theta}'_i; x \in C_i) + q' \log n_i \quad (6)$$

ただし，q はパラメータ空間の次元数で，V_i の共分散を無視した場合 $q = 2p$（p はデータの次元数），無視しない場合 $q = p(p+3)/2$ である．また，$\hat{\theta}'_i = [\hat{\theta}_i^1, \hat{\theta}_i^2]$ は，2 つの多変量正規分布の最尤推定値である．q' は共分散を無視した場合 $q' = 2 \times 2p = 4p$（p はデータの次元数），無視しない場合 $q' = 2q = p(p+3)$ である．なお，n_i は C_i に含まれるデータ数，$L(\cdot)$ は尤度関数である．

5. 実験結果

実験には約 1 日分の Web アクセスログデータを用いた．この間，ユーザは 1004 個の Web サイトにアクセスしている．図 2(a) はユーザが閲覧した Web サイトを主成分分析により 2 次元平面に射影した Web アクセスログを示す．この図から主成分分析によって特徴のある単語を含む Web サイトが分散して再配置されていることが分かる．しかし，この図の情報だけでは，ユーザの興味がどのように遷移したのか理解することは困難である．そのため，クラスタリング手法を用いて類似したウィンドウタイトルを持つ Web サイトを分類する．そして，ユーザが Web サイトを閲覧した軌跡を点から点への遷移として捉えるのではなく，類似 Web サイトの集合から集合への遷移として捉え直す．

本実験では，クラスタリング手法として自動的にクラスタ数を決定することができる x-means 法を用いた．実験では初期のクラスタ数 k_0 を 7 として計算を行った．初期のクラスタ中心を図 2(a) 上に○で示す．初期のクラスタ中心は 2 次元平面上に射影された Web サイトの分布から決定した．

Web サイトは x-means 法により最終的に 29 個のクラスタに分類された（図 2(b)）．図 2(c) は，図 2(b) の□で囲まれた領域を拡大したものである．x-means 法により詳細に分類されていることが分かる．表 1 に各クラスタに分類された Web サイトの内容とサイト数を示す．クラスタリング結果を見ると，「カントリー雑貨」や「鉢植え」など，同じ分類が見られるが，これは商品や植物の種類が異なっている．クラスタ No.14 では，サイト数が 524 個と多く，類似していない Web サイトが含まれていた．これは，各 Web サイトのウィンドウタイトルに含まれる単語が少ない，あるいは，形態素解析で適切に単語が切り出せなかったために起こったと考えられる．

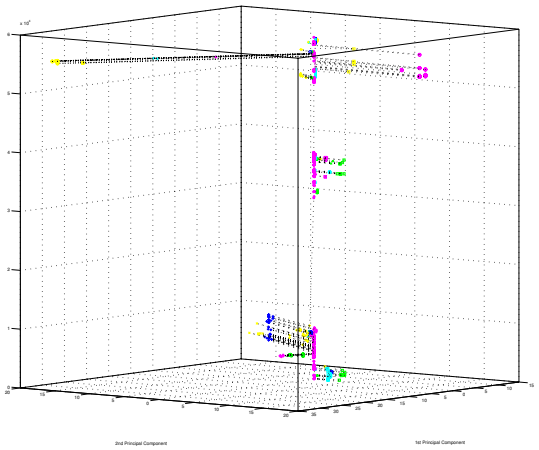


図 3: Clustering result

次に，ユーザの興味が時間とともにどのように推移しているかを観察するため，図 3 に示すように 2 次元平面上に射影された Web アクセスログをアクセスした時間にあわせて z 軸方向に配置する．図 3 に見られるように，時間とともにユーザが異なるクラスタ集合間を遷移している過程が観察できる．

このように，解析者が一目見ただけでは意味の抽出が困難である Web アクセスログをデータマイニング手法により加工することで，インターネットユーザの興味の遷移を抽出し，要約することができる．

表 1: Clustering results

クラス番号	サイト数	内容	クラスタ番号	サイト数	内容
C1	2	コンサート情報	C16	7	鉢植え
C2	12	アミューズメントパーク	C17	32	無料会員制ネット
C3	48	カントリー雑貨	C18	4	ネットショップ
C4	6	アンティーク雑貨	C19	10	子供服
C5	43	ホーローバケツ	C20	19	ガーデニング
C6	10	植物の育て方	C21	58	ネットショップ
C7	43	ネットショップ	C22	25	ネットショップ
C8	10	コンサートチケット	C23	6	カントリー雑貨
C9	43	アミューズメントパークチケット	C24	2	デニム
C10	16	ネットショップ	C25	9	観葉植物
C11	18	アンティーク雑貨	C26	15	植物の育て方
C12	10	ガーデニング	C27	12	サンダル
C13	3	鉢植え	C28	7	tukaeru.net
C14	524	ブーツ，子供用品，ガーデニング関連	C29	19	ガーデニング
C15	11	鉢植え			

6. おわりに

本稿では，インターネットユーザの興味が時間とともにどのように変わるのか理解するために，長期間にわたり記録された個々のユーザの Web アクセスログを解析する方法を提案した．

まず，ユーザが閲覧した Web サイトの特徴を単語ベクトルにより記述し，TF-IDF により各単語の重み付けを行った．そして，主成分分析により 2 次元平面に射影することで次元圧縮を行い，クラスタリング手法の一つである x-means 法により類似 Web サイトの分類を行った．以上の操作によりユーザの興味の遷移パターンが抽出できることを実験を通して確認した．

今後，各 Web サイト間の特徴をより明確に表現するために，単語の重みの付け方，SOM[Kohonen00] や FarstMap[Faloutsos95] などの次元圧縮手法の利用，などが考えられる．

7. 謝辞

研究の一部は，2003-2005 年度，科学研究費補助金（若手研究 (B) 課題番号 15700504），2003-2005 年度，科学研究費補助金（基盤研究 (A) 課題番号 15200012），および，2004-2007 年度，科学研究費補助金（基盤研究 (A) 課題番号 16200008）の助成を受けて実施された．また，数値データを提供して頂いた（株）電通メディアマーケティング局メディアリサーチ 1 部松永久氏，および森田喜文氏に感謝の意を表する．

参考文献

[Takeda04] 武田 英明, Weblog 研究の現状, 人工知能学会：第 7 回セマンティックウェブとオントロジー研究会資料, SIG-SWO-A402-06，(2004).

[Mori04] 森純一郎, 松尾豊, 石塚満, 語の共起情報を用いた Web 上からの個人メタデータ抽出, 第 7 回セマンティックウェブとオントロジー研究会資料, SIG-SWO-A403-01, (2004).

[Taniguchi04] 谷口 智哉, 松尾 豊, 石塚 満, Blog コミュニティの抽出と分析, 人工知能学会：第 6 回セマンティックウェブとオントロジー研究会資料，SIG-SWO-A401-08, (2004).

[Numa04] 沼 晃介, 大向一輝, 濱崎雅弘, 武田英明, Weblog におけるエゴセントリック検索の提案と実装，人工知能学会：第 6 回セマンティックウェブとオントロジー研究会資料, SIG-SWO-A401-06, (2004).

[Otsuka03] 大塚 真吾, 豊田 正史, 喜連川 優, ウェブコミュニティを用いた大域 Web アクセスログ解析法の一提案, 情報処理学会論文誌：データベース, Vol.44, No.SIG18 (TOD 20), pp.32-44, 92003).

[Namatame98] 生天目 章, マルチエージェントと複雑系, 森北出版，(1998).

[Madey03] Greg Madey, Vincent Freeh, Renee Tynan, Yongqin Gao, and Chris Hoffman. Agent-based Modeling and Simulation of Collaborative Social Networks, In Proc. of AMCIS 2003, Tampa, FL, (2003).

[Fisher04] D. Fisher and P. Dourish, Social and Temporal Structures in Everyday Collaboration, Proceedings of the 2004 ACM Conference on Human Factors in Computing Systems (CHI04), pp.551-558, (2004).

[MeCab] http://chasen.org/ taku/software/mecab/

[Kohonen00] T. Kohonen, Self-Organizing Maps (3rd ed), Springer, (2000).

[Faloutsos95] C. Faloutsos and K.I. Lin, FastMap : A Fast Algorithm for Indexing, DataMining and Visualization for Traditional and MultiMedia Datasets, Proceedings of the 1995 ACM SIGMOD international conference on Management of data, pp.163-174, (1995).

[Ishioka00] 石岡 恒憲, クラスター数を自動決定する k-means アルゴリズムの拡張について, 応用統計学, Vol.29, No.3, pp.141-149, (2000).