FACT-Graphと逐次確率比検定を用いたWebアクセスログの分析

佐賀　亮介，Mauricio Letelier，開作　直樹，高山　幸大，辻　洋

本論文では，FACT-Graphと逐次確率比検定を用いてWebアクセスログの分析を行う．本論文では，商品の情報が結びつけられたアクセスログに対して，安定したトレンドを持つ期間に注目し，その期間の訪問頻度のトレンドとページ間の関係に注目しながらFACT-Graphで可視化を行っている．この安定したトレンドを認識するために，逐次確率比検定を用いてトレンドの変化点を認識しそれを分析期間として用いている．また，商品のカテゴリをより効果的に把握し分析するために階層的FACT-Graphにより可視化を行っている．実際にゴルフサイトのWebアクセスログを用いて，13個のトレンドの変化の認識と2つの期間におけるアクセス状況の差を比較分析し，そこから得られた知見をまとめている．

キーワード：マーケティング，データマイニング，FACT-Graph，逐次確率比検定，可視化

1. はじめに

Webサイトは昨今では重要な販売チャネルの一つであり，多くの訪問者を集め収益を上げるべく，バナー広告や各種サイトとのコラボレーションなどさまざまな試みがなされている．そして，サイトの管理者は訪問者がどのページから訪れ，どのページから去っていくのか，また時期によってどのような商品群に興味を持つかなど，訪問者の振る舞いを把握し，ユーザビリティを改善させ売上につなげようとしている [1]．

この試みを行うために用いる主な情報源として，Webアクセスログ（以降，アクセスログ）がある．アクセスログから得られる属性は標準化されているため，さまざまな分析ソフトウェアが提供されている．統計手法を用いたAnalog [2]，AWStats [3] といったソフトウェアや，昨今ではWeb上で解析結果が閲覧できるGoogle Analytics [4] など現在もそのソフトウェアの進化は続いている．これらのソフトウェアを用いることで，ページビューやページ閲覧時間，リンク解析など基本的な集計情報が提供されている．

一方，より学術的な視点からWeb利用マイニングという概念が提供されている [5]．Web利用マイニングはWebマイニングの分野の中の一つとされており，その目的は訪問者の振る舞いの認識・予測，そして支援とされている．実際にWeb利用マイニングではアクセスログからアクセスパターンを抽出し，訪問者が訪問しそうなページを予測してナビゲーションを行ったり，また同一のカテゴリにあると見なされるページや動向が似ている訪問者をクラスタリングし，訪問者に提示するなどさまざまな試みがなされている [6]．

可視化もまたWebの利用状況を認識するために有用な手法の一つである．多くの商用ソフトやさまざまな研究においてアクセスログの可視化方法が既に使われている．例えば，先ほど述べたGoogle Analyticsなどでは折れ線グラフや棒グラフなど従来から多用されているグラフによりページアクセスのトレンドを効果的に表している．

また，グラフ構造を元にした可視化も，ページ構造や訪問経路を分析するのに有用な手法として昨今注目を浴びている．その一例として，佐賀らは，FACT-Graphと呼ばれる共起グラフを元にした可視化手法を提案している [7]．先述したとおり，従来の手法でもページの訪問頻度のトレンドやページ間の関係を可視化することができたが，それぞれ独立した可視化結果として表されていた．FACT-Graphはこれらの情報を一つに統合し，トレンドと関係性の可視化を実現している．実際に，大学のWebアクセスログや新聞記事のトレンド可視化などが行われている．しかし，FACT-Graphでは，複雑なトレンド変化を分析期間に内包するとき，分析結果が適切に表示できない可能性がある．そのため，従来は全体のトレンドの変化状況を鑑み，それを元に分析期間を分割し，可視化を行ってきた．ただしこの方法は分析者によって分析期間が異なることが多々あり，それゆえに可視化結果も異なることがあった．そ

さが　りょうすけ，まうりしお　れていえる，かいさく　なおき，たかやま　ゆきひろ，つじ　ひろし
大阪府立大学工学部
〒599–8231 大阪府堺市中区学園町1–1

のため，安定した結果が得られないという問題を生じてしまっている．

そこで，本論文では，商品の情報が結びつけられたアクセスログに対して，安定したトレンドを持つ期間に注目し，その期間の訪問頻度のトレンドとページ間の関係に注目しながら FACT-Graph で可視化を行う．トレンドの変化を認識するために，逐次確率比検定 [8], [9] を用いて，トレンドの変化点を発見する．そして，検出された複数の変化点の間において，関心のあるトレンドに対して商品の情報を可視化し，どのような傾向があるのかを FACT-Graph にて可視化する．逐次確率比検定を用いることで，トレンドの変化を客観的に得ることができるため，出力結果が安定するという利点も得られる．本論文は以下の構成を採る．第 2 章では，まず対象データについて述べ，大まかに分析プロセスについて述べる．第 3 章では，逐次確率比検定と FACT-Graph，そして対象データに対して効率的に分析するために FACT-Graph の階層化を提案する．その後，第 4 章にてゴルフダイジェストオンラインから提供された実データに対して可視化を行い，そこから得られた知見を述べた後，第 5 章にて本論文を閉じる．

1.1 本論文の貢献

本論文は，以下の点で貢献している．1 点目は，逐次確率比検定と FACT-Graph の組合せによる可視化分析の有用性を示したことである．トレンド可視化手法として使われてきた FACT-Graph だが，トレンドの変化が多く含まれる場合にはうまく可視化できない可能性がある．逐次確率比検定を用いることで，そのトレンドの変化を客観的に得られるため，可視化にふさわしい期間を得られるだけでなく，可視化結果を安定して得られるというメリットを得られる．2 点目は，現実の商品の状況を知るために，FACT-Graph の階層化が挙げられる．この階層化は今回商品のカテゴリの範疇で行ったが，オントロジなどの意味的階層化にも利用できる．

2. 対象データと分析プロセス

本論文で扱うアクセスログは，ゴルフダイジェストオンライン [10] より提供された 2010 年 7 月から 2011 年 6 月までの 1,561,193 件のデータであり，すでにいくらか加工された状態にある．具体的に，一般的なデータであるアクセスページやリファラーページのほか，セッション情報や Web と商品の対応情報もすでに記入されている状態にある．また，このデータには，シューズやクラブといった商品カテゴリも記載されており，その他，商品ページとは関係ないトップページや商品検索結果，セール情報などのページに関するアクセスログも含まれている．

このデータに対して，まず日単位で Web サイトのアクセス頻度を算出する．その後，そのアクセス頻度に対して逐次確率比検定を行い，それからトレンドの変化点を抽出する．そして，その変化点ごとにデータを区分けし，それぞれについて FACT-Graph を出力し，分析する一連の流れが本論文での分析プロセスとなる．

3. 適用手法

3.1 逐次確率比検定

逐次確率比検定（Sequential Probability Ratio Test：以下，SPRT）は，トレンドの変化点を抽出するためのキーとなる手法である [8]．SPRT は，品質管理などの分野で使われてきた統計的仮説検定であり，Chow 検定などより速く検出ができ，また構造変化後の時系列の発生分布を考慮する必要がないという特性がある [9]．一般的に，SPRT は統計的仮説検定に従い，帰無仮説と対立仮説を用いて，各データが得られる度に検定を行う．このとき，ある観測データはある母数 θ における確率密度関数 $f(y|A)$ に従って分布が生成されているとすると，各仮説は，$f(y|A)$ において帰無仮説 $H_0 : \theta = A_0$，対立仮説 $H_1 : \theta = A_1$ となる．SPRT は尤度比 λ_i を計算し，そして以下の式のように検定対象データ Z_i に対して，累積的に λ_i の値を使って計算を行っていく．

$$\begin{aligned}\lambda_i &= \frac{P(Z_1|H_1)P(Z_2|H_1)P(Z_3|H_1)\cdots P(Z_i|H_1)}{P(Z_1|H_0)P(Z_2|H_0)P(Z_3|H_0)\cdots P(Z_i|H_0)} \\ &= \lambda_{i-1}\frac{P(Z_i|H_1)}{P(Z_i|H_0)} \qquad (3.1)\end{aligned}$$

ここで，$P(Z_i|H_0)$ は帰無仮説 H_0 における Z_i の発生確率であり，$P(Z_i|H_1)$ は 対立仮説 H_1 における Z_i の発生確率である．今回，それらの値をそれぞれ，そして，SPRT は次のような停止条件を持っている．

1. $\lambda_i > C_2 \rightarrow H_1$ を採択.
2. $\lambda_i < C_1 \rightarrow H_0$ を採択.
3. $C_1 \leq \lambda_i \leq C_2 \rightarrow$ 観測の継続.

ここで，$C_1 = \beta/(1-\alpha), C_2 = (1-\beta)/\alpha$ であり，α と β は第 1 種・第 2 種の誤りをそれぞれ示す．

この SPRT を用いて時系列データからトレンドの変化点を発見するためには，トレンドの傾向を表す予測モデルを求め，そのモデルから外れているかどうかという許容区間が必要である．つまり，構造の

変化をトレンドの変化と見なす．この予測モデルにはARMA モデルなどさまざまな時系列モデルが使用できるが，今回は学習期間 L の間の観測データを用いて算出した単回帰モデルを当てはめる．具体的に今，$(x,y)=(x_1,y_1),(x_2,y_2)\cdots(x_L,y_L)$ と，ある時間 x_i において観測データ y_i，つまり今回の実験データでは，1 日単位で集計されたアクセス頻度が得られたとき，単回帰モデル $y=ax+b$ により観測データ y_i は次式のように表される．

$$y_i = ax_i + b + \epsilon_i \tag{3.2}$$

ただし，$a=\dfrac{n\sum_{k=1}^{L}x_k y_k-\sum_{k=1}^{L}x_k\sum_{k=1}^{L}y_k}{n\sum_{k}^{L}x_k^2-(\sum_{k=1}^{L}x_k)^2}$

$$b=\frac{\sum_{k=1}^{L}x_k^2\sum_{k=1}^{L}y_k-\sum_{k=1}^{L}x_k y_k\sum_{k=1}^{L}x_k}{n\sum_{k=1}^{L}x_k^2-(\sum_{k=1}^{L}x_k)^2}$$

このとき，検定対象データ Z_i は，平均 0，分散 σ^2 の正規分布に従う推定誤差 ϵ_i となる．

この SPRT を用いた構造変化点検出のステップは以下のようになる．最初に，L の期間の観測データ y_i を用いてモデルを学習し，予測モデルを生成し，許容区間を設定する．続いて，初期値として $\lambda_1=1$ と設定し，観測データ y_i から λ_i を計算していく．このとき，Z_i（つまり ϵ_i）が許容区間内にある場合には $P(Z_i|H_0)=\theta_0, P(Z_i|H_1)=\theta_1$ を与え，許容区間を超えたとき $P(Z_i|H_0)=1-\theta_0, P(Z_i|H_1)=1-\theta_1$ を与える．そして，上記の停止条件において H_0 が採択されたとき，$\lambda_i=1$ と設定し直し検出を続け，H_1 が採択されたとき，その時点においてトレンドが変化したと見なす（図 1）．その後，再び予測モデルの構築を行いトレンドの検出を続ける．ただ，モデルの再学習方法についてはさまざまな方法が考えられる．今回は，信頼区間を外れだしてから変化点を検出するまでいくらか遅れがあることから，誤差が信頼区間を外れだしたとき，つまり λ の値が 1 より変化し始めた点から再学習をするという方法を用いて行っている．詳しくは，参考文献 [8], [9] を参考にされたい．

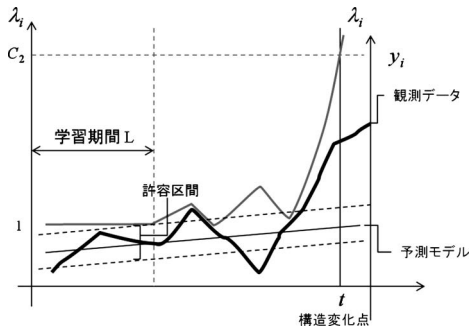

図 1　逐次確率比検定を用いたトレンド変化点検出

3.2　FACT-Graph

FACT-Graph はグラフ構造により表現されるデータ可視化手法の一つである．データ可視化手法は情報を圧縮し一覧可能にすることで，分析者に分析の仮説やきっかけを与えるものとして，昨今，データマイニングの手法自体が進んできた結果，注目を浴びてきている．この FACT-Graph は，キーワードのトレンド可視化のために使用され，新聞記事やアクセスログへの適用とその分析において実績がある．FACT-Graph は分析期間においてノードやリンクがどのように変化しているか？という大域的な視点での可視化を目指している．

FACT-Graph の特徴は，ノードとリンクにトレンドや属性に関する情報を組み込んでいる点にある．FACT-Graph はクラス遷移分析と共起遷移分析の 2 つからなっている．クラス遷移分析は各ノードの状態をクラスという大きな枠に当てはめ，時間が経つにつれてどのようにクラスが変化しているか，という推移に注目している．このとき，クラス間に大まかな大小関係を設定することで，クラスの推移によりトレンドの変化を把握できるようにしている．共起遷移分析は，共起グラフの状態に注目したものである．共起グラフは共起関係にて形成されたクリークを把握することで，文章でいうトピックやクラスタなどが把握できる．この共起の状態が時間に応じてどのように変化しているか，という情報を赤・青・黒という色情報（今回は，紙面の都合上，濃淡にて表している）にてまとめ，その情報を見ることで，トピックの推移が確認できる．

FACT-Graph はトレンドの大域的な可視化，つまりトレンドが上昇したか，下降したか，現状を維持しているかの可視化を目指している．そのため，分析期間において，トレンドが大まかに単調増加，単調減少といった状態が最も可視化しやすい状況である．しかしながら，分析期間において必ずしも単調増加，単調減少になっているとは限らず，増加と減少を大きく繰り返したりすることがありうる．ここで，SPRT を用いることで，その増加トレンドにある期間，減少トレンドのある期間などを検出することができ，FACT-Graph にとって，より表しやすい分析期間を提示することができる．

3.3　FACT-Graph の改良

FACT-Graph では新聞記事やアクセスログなどを対象に，また共起の様子からトピックやクラスタなど

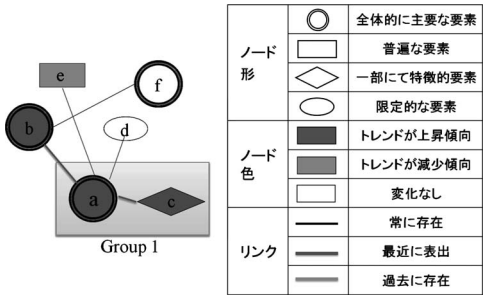

図 2　FACT-Graph

を暗黙的にだが識別できた．しかしながら，今回取り扱うデータのように，商品にカテゴリが明確に存在し，アクセスログと商品の対応などが明確にわかっている場合，商品のカテゴリなどで大きく分類することで，分析者に可視化を通して効果的な情報の概要把握や分析がよりしやすくなる．そこで，カテゴリ情報を各商品の上位階層とみなし，同一階層下にある商品をグループ化する．こうしてできた FACT-Graph の概要図を図 2 に示す．

4. 実験

SPRT と改良した FACT-Graph を使用し，2 章で述べたデータセットを可視化する．まず，SPRT を実施する．この時のパラメータとして α, β をそれぞれ 0.05, 0.05 とし，その α に対応して信頼区間 95%に対応するように許容区間を 2 σ に設定した（正規分布において許容区間 2 σ は全体の約 95%のデータを内包する）．学習期間としては，1 カ月単位では学習期間が長く，学習中にトレンドを見逃す可能性があり，また 1 週間では十分な学習ができない可能性もあることから，今回は 2 週間分のデータ，つまり $L = 14$ と定めて学習を行った．また，データに急激な変化（バースト状態）などがないことから，比較的緩やかに検出することを考え，文献 [9] を参考に θ_0, θ_1 を 0.20,0.80 として実行した．

その結果，13 点の変化点が見られた．これらの変化点をデータ上に表したものを図 3 に示す．図 3 を見たところ，これらの変化点はもっともらしく見えるため，これらを採択し FACT-Graph を生成する．今回，これらの期間のうち，上昇トレンドにあった 2010 年 10 月 14 日〜10 月 29 日（期間 1）と 2011 年 4 月 6 日〜5 月 27 日（期間 2）の 2 つに注目して可視化を行った．

期間 1 と期間 2 を可視化した結果を，図 4，図 5 に示す．図 4，図 5 において，それぞれノードは Web ページを表しており，そのページが商品を指す場合，その商品は所属するグループ内に描画されている．ここで，期間 1 を表した図 4 を見てみると，そこからさまざまなことがわかる．例えば，ハーフショートパンツは濃い二重丸，つまりクラス A の重要度の高い商品だということが推測できる．また，アイアン・キャディバッグ・アンダーウェアと同時に見られることが多くなっているため，この期間において推薦すべき商品の一つであると考えられる．一方，クラブ関連の商品であるドライバ，ウェッジなどはほかの商品とともに見られる傾向にあるが，同一カテゴリ内でのリンクが少ないから，同一カテゴリの商品間を比較することは少なかったと考えられる．

一方，期間 2（図 5）においても次のようなことが FACT-Graph からわかる．

- 一部の人がスパイク鋲とシューズに関するページを一緒に訪問している．また，シューズのページは全体的に注目を浴びているページである．
- セール情報に関するページはアクセスが増加していることがわかる．例えば，あるページ（図 5 内，(a) 枠）はゴルフウェアのセールに関するページであり，そのページと共起を持つパンツに関する

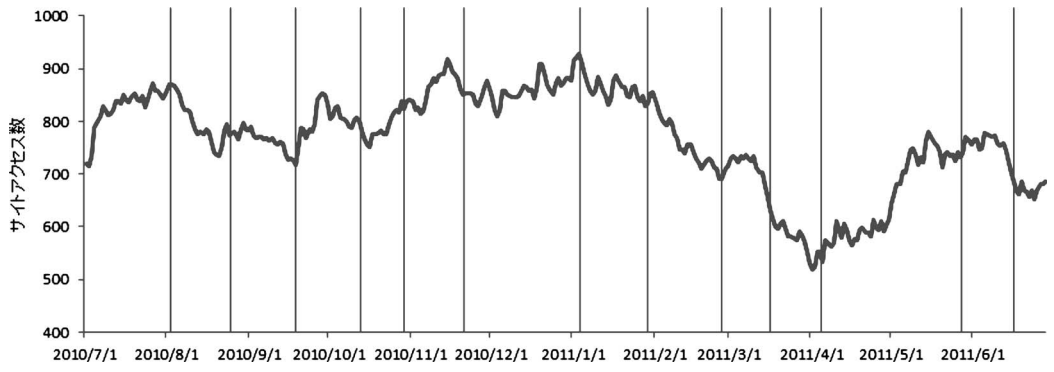

図 3　Web アクセスログへの SPRT の実行結果

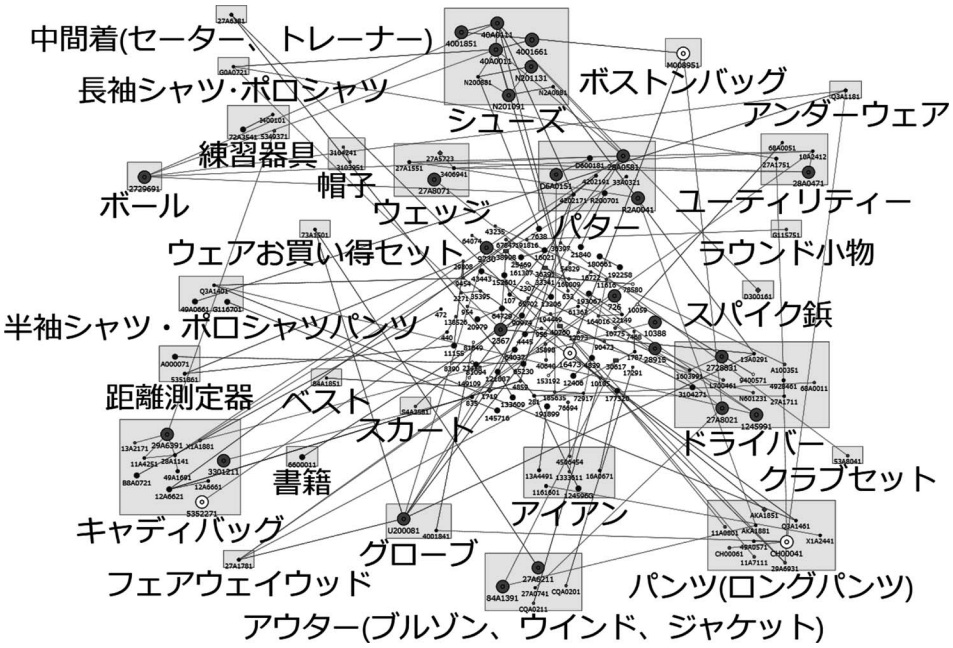

図 4　2010 年 10 月 14 日～10 月 29 日における FACT-Graph

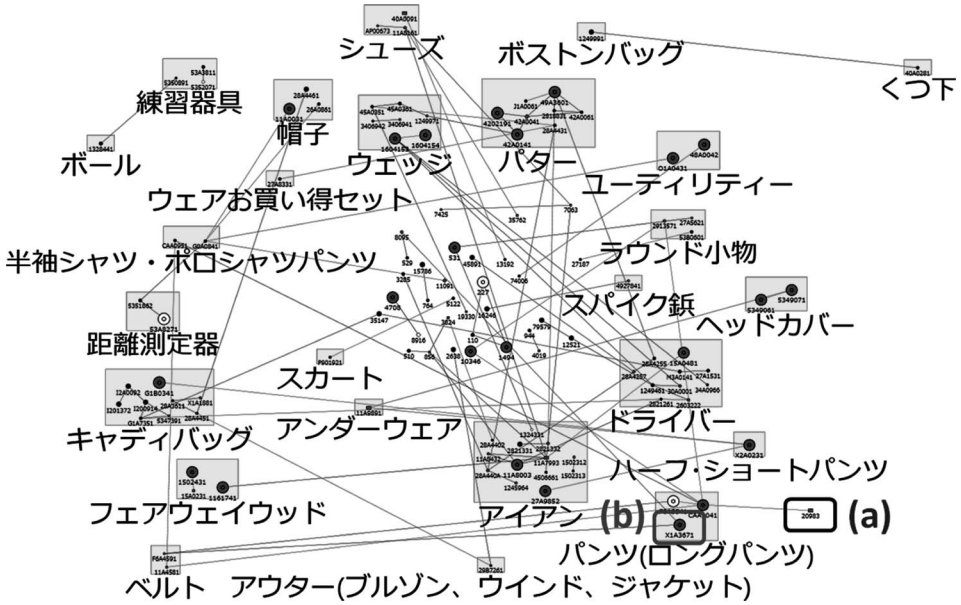

図 5　2011 年 4 月 6 日～5 月 27 日における FACT-Grpah

ページはアクセスが増加している．

- パンツやベルトと違って，ドライバー内ではリンクが数多くあるが，次第に複数のドライバー間のリンクが消滅している傾向が見られる．すなわち，次第に一部だけ注目を浴びており，多くのドライバーは注目を浴びていないことがわかる．
- パンツに関するページに訪問するときはベルトにも訪問している傾向が，FACT-Grpah に表されているリンクからわかる．特に，このパンツカテゴリ内には興味深い動向が見られる．例えば，ある商品（図 5 内，(b) 枠）は，ほかのものとは異なり，ノードが二重円で濃く，そしてリンクの色が

薄いことから，昨今流行ってきている一方，リンクは消滅してきている傾向にあることがFACT-Graphからわかる．このことから，この商品は過去には他の商品と一緒に見られていたが，現在ではその商品単体で訪問されていると考えられる特徴的な商品だと推測することができる．

そして，期間1（図4）と期間2（図5）を比べたとき，さまざまなことが把握できる．例えば，共通した傾向としては，

- 商品自体は異なっているがパターに分類される商品は多くが注目を集める傾向にある．
- スパイク鋲に関してはアクセス頻度は小さいが，共に注目を集める傾向にある．

といったことがわかる．一方，異なった傾向としては次のようなことがわかる．

- アウターは期間2では重要でない商品であったが，期間1では閲覧回数とその関係性が増加傾向にあることがFACT-Graphから確認できる．このことから，期間1は秋にあたるため，アウターが重要視するべき商品になったと言える．
- FACT-Graphから期間1では，同じ種類のクラブ間ではあまり一緒に見られていないが，期間2では一緒に見られる傾向にあった．このクラブについてのWebページを実際にみてみると，期間2ではクラブのセールが行われていることが確認できた．つまり，この時期2において，クラブのセールにより興味を持った訪問者は，実際にそのページを訪問し，（実際に購入しているかどうかは別として）複数の商品を比較するという傾向にあったことがわかった．
- また，ボールに関しては期間1では商品のクラス・関連性が重要だと示しているが，期間2ではアクセスされておらず注目を浴びてないことがFACT-Graphからわかる．これは，寒くなる（期間1）と見失ったボールをあまり探さなくなり，ボールが重要とされたのではないか，または，新しいシーズンであるためにボールを新調したのではないかなどさまざまな観点での考察ができる．

これらから，この2期間においてを商品ページへのアクセスを比較すると，両期間において同じように上昇傾向にあるページもあるが，全体的に各商品のアクセス動向は異なっていることがわかる．このように，今回の可視化からアイテム単体だけでなく，カテゴリ単位により分析が可能になったため，より抽象的かつさまざまな観点による分析への手掛かりを得られることができ，可視化としては一定の成果が得られたと考えられる．

5. おわりに

本論文では，実在のWebアクセスログに対してSPRTを用いて分析期間を客観的に求め，そしてFACT-Graphを用いて可視化を行った．本論文により主観的に設定していた分析期間が客観的に求まることで可視化結果が安定して得られるようになり，またFACT-Graphにおいて階層的な可視化をサポートすることにより，同一カテゴリにある商品の傾向を効率的に表せることができた．ゴルフダイジェストオンラインから提供されたデータを分析した結果，商品の注目度の傾向や共に考慮されている商品などが把握でき，一定の有用性が示すことができた．

ただ，本論文の制約として，SPRTは観測データが大きく外れたような外れ値やある複数区間が外れるような異常部位変化といったものを対象としていない点や，FACT-Graphもこれらに対応できない点が挙げられる．また訪問者がどこからどこに移動しているか，という動向の可視化まで至っていない．これらの点を解決することが今後の課題として挙げられる．

参考文献

[1] J. Nielsen, Web Usability, *Peachpit Press*, 1999.
[2] Analog. http://www.analog.cx (2012)
[3] AWStats official web site: Free real-time logfile analyzer to get advanced statistics (GNU GPL). http://awstats.sourceforge.net (2012)
[4] Google Analytics. www.google.com/intl/en/analytics/ (2012)
[5] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, *Springer-Verlag*, Berilin, 2007.
[6] S. Gunduz–Oguducu, and M. T. Ozsu, Web Page Recommendation Models: Theory and Algorithms, *Morgan and Claypool Publishers*, 2010.
[7] 佐賀亮介，寺地雅弘，辻洋，FACT-Graph：頻度と共起度を用いたトレンド可視化，電気学会論文誌C, **129**(12), 545–552, 2009.
[8] A. Wald, Sequential Analysis, *John Wiley & Sons*, 1947.
[9] K. Takeda, T. Hattori, T. Izumi, and H. Kawano, Extended SPRT for Structural Change Detection of Time Series Based on a Multiple Regression Model, *Artificial Life and Robotics*, **15**(4), 417–420, 2010.
[10] ゴルフダイジェストオンライン, http://www.golfdigest.co.jp/ (Accessed in 2012)

