2007－ICS－148 (8)
2007／7／4

ページ閲覧時間を考慮したWebログマイニング手法の提案

三原宏一朗[1*] 寺邊正大[1*] 橋本和夫[1*]

[1]東北大学大学院情報科学研究科

概要：近年，Webサイトの複雑化・大規模化に伴い，Webサイト管理者がユーザのニーズに合わせてWebサイトを適宜改善していく必要が高まっている．このユーザのニーズを知る手がかりの一つとして有用なのがWebアクセスログである．Webアクセスログを解析することで，ユーザのWebサイト内での行動の様子を知ることができる．本稿では，Webアクセスログ解析にデータマイニングを応用するWebログマイニングに関連して，ユーザが各ページを閲覧した時間を考慮したアクセスパターンを抽出する手法を提案する．次に，実際のWebアクセスログに対して本手法を適用し，ページ閲覧時間を考慮することで，より詳細なユーザ行動分析が可能となることを示す．

A Proposal of Web Log Mining Method Considering Page Browsing Time

Koichiro Mihara[1*], Masahiro Terabe[1*], and Kazuo Hashimoto[1*]

[1]Graduate School of Information Sciences, TOHOKU University

Abstract : It is a major responsibility for Web Site administrators to organize Web contents reflecting user's needs and demands. Considering the increasing hugeness and complexity of today's Web Sites, it is necessary to develop a novel method to facilitate Web Site maintenance. Web Access Log Mining is a means to estimate user's potential needs from their access behavior. This paper treats a browsing time as an index of user's interest, and proposes a method to extract user's access patterns considering browsing time. The experiment shows that the proposed method is capable of analyzing user's behavior in finer granularity.

1 はじめに

通信技術やWeb関連技術の進歩により，インターネット上で公開されているWebサイトの複雑化・大規模化が進んでいる．一方で，Webサイトは今や人々が日常的に利用する重要な情報獲得元の1つとなっている．したがって，複雑なWebサイトほどユーザにとって使いやすいものにする必要があるため，Webサイト管理者はユーザのニーズに合わせてWebサイトを適宜改善していかねばならない．このユーザのニーズを知る手がかりの1つとしてWebアクセスログが注目されている．Webアクセスログとは，ユーザがWebページ上のリンクをクリックした際にWebサーバに送られるリクエストを時系列順に保存したものである．これを解析することでユーザのWebサイト内での行動の様子を知ることができ，Webサイト管理者がWebサイトを改善する上での足がかりとなる[1].

最近では，Webアクセスログ解析にデータマイニング技術を応用するWebログマイニングの研究が進められている．Webログマイニングで用いられる手法は，アイテム集合マイニングや順序パターンマイニング，グラフマイニングなどが提案されている[2]．特にグラフマイニングは，Webサイトの持つリンク構造に沿った形でユーザのページ遷移の様子を表現できる上，ユーザが複数のページを同時閲覧しているような状況も含めてパターン化して抽出できる．タブ機能付きブラウザの普及も始まり，今後ユーザが複数ページを同時に閲覧するような利用状況が増えると予想される．ユーザの行動解析をより詳細に行うには，グラフマイニングが有用である．

また，Webアクセスログ解析においてはページ遷移だけでなくページ閲覧時間も重要な要素とされる．同じページでもユーザ毎に，あるいは経

*連絡先：東北大学大学院情報科学研究科先端情報交換技術論（KDDI）寄付講座
〒980-8579 宮城県仙台市青葉区6-6-11-304
E-mail: mihara@aiet.ecei.tohoku.ac.jp

表1：Web利用マイニングで用いられる手法例（パターン中の各ノードはWebページ，エッジは移動経路）

パターン	抽出手法	特徴
a b c d	アイテム集合マイニング	相関ルールとして扱うこともでき，各ページ間の関連性の理解に有効．
a → b → c → d	順序パターンマイニング	リクエスト順序から得られるページ遷移のみを対象としており，複数ページを同時に閲覧している場合などの分岐遷移は考慮できない．
a b d c	グラフマイニング	分岐遷移も考慮し，Webサイトのリンク構造に近い形でパターンが抽出できる．サイト内でよく利用される経路の特定がし易く，Webサイトのリンク構造の再構成などに有効．

路毎にそのページの持つ重みは異なり，その違いは閲覧時間の長さから推定できる．ここでいう重みとは，ユーザがあるページに対して興味を持った度合いのことをいう．セッション中の閲覧時間が長いほど，ユーザはそのページに強く興味を持ったと推定される．

本稿では，Webサイト管理者がユーザの興味を反映しながらWebサイトの改善を行うための情報提供を目的とし，グラフマイニングを用いて，閲覧時間による各ページの重みを考慮したWebアクセスパターンを抽出するWebログマイニング手法を提案する．

2 関連研究

2.1 Web利用マイニング

Webログマイニングとは，Web利用マイニングとも呼ばれ，データマイニング技術を応用してWebアクセスログなどのページリクエスト履歴からユーザのWebサイトへのアクセスパターンを抽出し，Webサイト内での行動の分析や予測を行うものである．

例えばあるWebサイトにおいて，ユーザがページaを閲覧した後bに移動し，cを別ウィンドウで開いた上でbからdに移動して，cとdを同時閲覧するという利用のされ方が多い場合を考える．この場合，Web利用マイニングで抽出されるアクセスパターンは，表1のようにまとめられる．

近年急増しているECサイトなどでは複数の商品ページを同時に開いて比較・検討するといった利用方法も考えられ，またタブ機能付ブラウザの普及も進みつつある．

このため，ユーザのページ遷移に分岐遷移が含まれる状況が増加している．Webサイトの改善を行う管理者の立場からは，このような分岐遷移も含めたユーザの行動を把握することが望まれる．そこで，提案手法ではこの分岐遷移をグラフ構造で抽出できるグラフマイニングを採用する．

2.2 グラフマイニングアルゴリズム

提案手法において，アクセスパターンの抽出にはグラフマイニングを用いるが，そのアルゴリズムは既存のものを使用することが可能である．代表的な既存アルゴリズムとしては，猪口ら[3]のAGM，Kuramochiら[4]のFSG，Yanら[5]のgSpan，Nijssenら[6]のGASTONなどがある．

AGMは最初のグラフマイニングアルゴリズムとして提案された手法である．グラフ構造を隣接行列で表し，ノードの追加による隣接行列の拡張により，Apriori風のアルゴリズムに基づいて頻出部分グラフを抽出する．

FSGもAGMと同様に隣接行列を用いたアルゴリズムだが，エッジの追加によって隣接行列の拡張を行う点で異なる．

gSpanはこれらとは違い，グラフをDFS（Depth First Search）木で表し，その最右拡張により頻出部分グラフを抽出する．

GASTONはパス，自由木，閉路を含むグラフと徐々に複雑な構造のパターンを抽出していく．アルゴリズムである．疎なデータを対象としており，非常に高速に頻出部分グラフを抽出できる．

2.3 時間を考慮したパターン抽出手法

平手ら[7]は，順序パターンマイニングに関して，ページリクエストの時刻差により入口ページから各ページに到達するまでの時間を計算し，パターン抽出の際に考慮する手法を提案している．

この手法は時間要素をページ間の距離として扱っており，マーケティングにおけるコンバージョンの達成度を評価する場合などに適している．

一方，本研究は各ページとユーザの興味の関係を考慮したWebサイトの改善を行う際の参考と

なるようなパターンの抽出を行うことを目的としている．

そこで，時間要素をページ間の距離ではなく，ユーザの興味の強さに応じた各ページの重みとして扱う．

3 提案手法

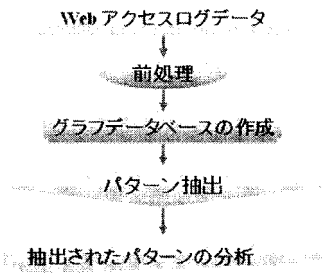

図1：提案手法の処理の流れ

提案手法の処理の流れを図1に示す．

まず，前処理では，不要な情報を除去し，ユーザセッションの識別を行う[8]．

次に，グラフデータベース作成の際は，各ページの閲覧時間と重みを求め，セッション毎にグラフ構造に落とし込む．

あるページPの閲覧時間（t_b）は，次にPがリクエストされるまで（再びリクエストされない場合はセッションの最後のリクエストまで）の間で，Pをリファラ（Ref，あるページをリクエストしたときユーザが閲覧していたページ）として含むリクエストのリクエスト時刻（t_r）と，Pをリクエストページ（Req）として含むリクエストのt_rの差の内，最長のものとする．ただしPがリクエストされて以後Pをリファラとするリクエストが行われていない場合，t_bは空（null）とする．

続いて，計算された閲覧時間に応じて，各ページにユーザの興味の強さを示す重みを付与する．重みの付加には重み付け関数$w(P, t_b)$を導入する．

$$w(P,t_b)=\begin{cases}0 & where\ t_b \neq null\ and\ f(t_b,T_b)=true\\ 1 & where\ t_b \neq null\ and\ f(t_b,T_b)=false\\ 2 & where\ t_b = null\end{cases} \quad (1)$$

重み付け関数により，各ページに0，1，2のいずれかの重みを付与する．$f(t_b, T_b)$はt_bを変数とする不等式であり，t_bの値がWebサイト管理者によって任意に定められた最小閲覧時間T_b以下の場合，そのページはユーザの興味を引かなかったと判断して重み0を付与する．重み1は$f(t_b, T_b)$を満たさず，ユーザの興味を強く引いたと思われるページに付ける．重み2はセッション内でRefとして存在しないページ（離脱ページ）に付ける．離脱ページに関しては閲覧時間が計算できない上，そのページから先に進んでいないことから他のページと区別して扱えるように重みを付ける．また，$f(t_b, T_b)$はWebサイト管理者が任意に設定でき，重み0，1の区別を行う際の基準を与えるものとする．

各セッションをグラフ化する際は，各ノードをページと重みの組とし，同じページかつ同じ重みの場合は同一ノードとする．エッジはリクエスト毎にRefとReqのノードを結ぶものとする．同じノードを結ぶリクエストが複数存在する場合は一本のエッジにまとめて扱う．こうしてできる各セッションのグラフは閉路を含む無向グラフとなる．

そして，作成されたグラフデータベースに既存のグラフマイニングアルゴリズムを適用して頻出部分グラフを抽出し，結果を出力して分析する．

なお，提案手法において抽出されたパターンが“頻出である”とは，事前に設定した最小支持度（支持度とは，全セッション中，各アクセスパターンが出現するセッションの割合）以上の支持度を持つことを意味する．

2.1節の例において，提案手法により得られる結果は例えば図2のようになる．

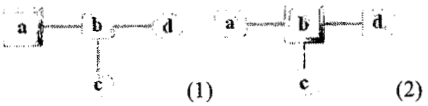

図2：提案手法によって抽出されるアクセスパターンのイメージ（各ノードはWebページ，エッジはユーザの移動経路を表す．また，四角，二重四角，丸はそれぞれ重み0，1，2を意味する．）

提案手法では，抽出されたパターンの一部に含まれるページとその部分グラフ構造が同じでも，ユーザの興味を引いたページの違いによって異なるパターンとして抽出できる．そのため，ユーザの興味とページ遷移の変化を比較でき，より詳

細なWebサイト利用状況の解析が可能となる.

4 評価実験

4.1 実験項目・目的

提案手法の評価を行うため，実際の Web アクセスログを用いた評価実験を行った.

各ページに重み付けを行った場合と，行わなかった場合について，次の 2 種類の実験を行った.

1. 抽出されたパターン数を比較し，重み付けによって各ページの扱いを区別することで，元々同一のパターンとして扱われていたものが区別できるようになったことを確認する.
2. パターン抽出にかかった時間を計算，比較し，重み付けによるマイニング時間への影響を検証する.

4.2 実験環境・設定

実験を行った PC マシン環境を表 2 に示す.

表 2：実験環境

OS	Microsoft Windows XP Professional SP2
CPU	Intel Core 2 Duo 1.80GHz
Memory	1.99GB

実験において，セッションの識別は 30 分のタイムアウトにより行った.

また，重み付け関数については，$f(t_b, T_b)$ を次のように設定した.

$$f(t_b, T_b) : t_b \le T_b = 30[\text{sec}] \quad (2)$$

これらの条件の下，各 Web アクセスログに対して，最小支持度を変えて実験した.

パターンの抽出にかかった時間の比較においては，各最小支持度において各々10 回ずつ測定し，その平均をとった.

なお，グラフマイニングアルゴリズムには GASTON を使用した.

4.3 実験データ

実験では，仙台市産業振興事業団，及び KDDI 研究所の協力により，各々から提供していただいた Web アクセスログを用いた.

各データの概要を表 3 に示す.

表 3：実験用 Web アクセスログ

仙台市産業振興事業団 http://www.siip.city.sendai.jp/	
収集期間	2006 年 9 月 1 日 ～2007 年 3 月 1 日
元データサイズ	約 1.94GB
前処理後のサイズ	約 92.7MB
総リクエスト数	9,440,870
セッション数	96196

KDDI 研究所 http://www.kddilabs.jp/	
収集期間	2005 年 10 月 1 日 ～2007 年 2 月 27 日
元データサイズ	約 827MB
前処理後のサイズ	約 252MB
総リクエスト数	5,057,390
セッション数	211849

4.4 実験結果・考察

4.4.1 実験 1：抽出されたパターン数の比較

実験 1 では抽出されたパターン数の比較を行った.

まず，仙台市産業振興事業団の Web アクセスログについて見てみると（図 3），最小支持度が 1%以下では，重み付けを行った場合の方が多くのパターンが抽出されている．これは，重み付けを行わない場合には同一ノードして扱われるページが重み付けを行ったことで区別され，異なるパターンとして抽出されたためである．また，最小支持度が大きくなると重み付けした方の抽出パターン数が減っているのは，重み付けによって区別された各ページは支持度が下がってしまい，大きな最小支持度は満たせなくなったためである.

一方，KDDI 研究所の Web アクセスログのマイニング結果（図 4）は，総じて重み付けを行った方が抽出パターン数は少なくなった．これは，重みによって区別された各ページがいずれも最小支持度を満たせなくなってしまうことが多かっ

たためである．

実際に抽出されたパターンの例を図5，6，7に示す．図5，6が重み付けによって区別された結果であり，図7が重み付けを行わなかった場合である．図5，6，7に示したのはKDDI研究所のWebアクセスログからのパターン抽出結果の例である．重み付け関数により重み付けされた結果，“/menu”ページについて異なるパターンとして抽出されている．これによって，図5のパターンでは“/menu”ページから先へは進まなかったが，図6のパターンではさらにサイト内のどこかのページに移動したことが読み取れる．なお，“-”は外部から入ってきたことを意味し，閲覧時間の計算ができないため自動的に重み2が付与されている．

このように，閲覧時間に応じた重みを各ページに付与することで，パターンの区別ができるようになったことが確認できた．

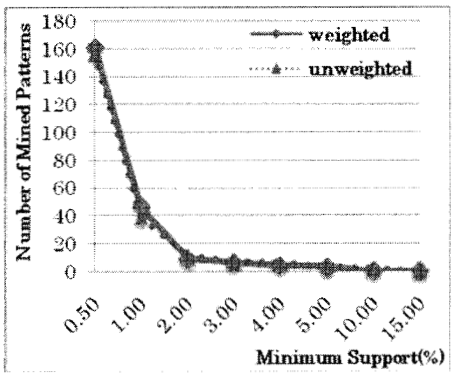

図3：抽出されたパターン数の比較（仙台市産業振興事業団）

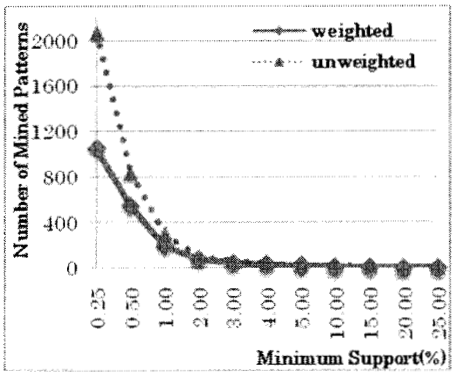

図4：抽出されたパターン数の比較（KDDI研究所）

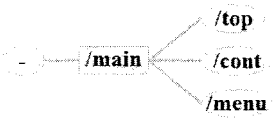

図5：重み付けした場合の抽出パターン例1（ノードはWebページ，エッジはユーザの移動経路．四角は重み0，丸は重み2）

図6：重み付けした場合の抽出パターン例2（ノードはWebページ，エッジはユーザの移動経路．四角は重み1，二重四角は重み1，丸は重み2）

図7：重み付けしなかった場合の抽出パターン例（ノードはWebページ，エッジはユーザの移動経路．）

4.4.2 実験2：抽出にかかった時間の比較

実験2では重み付けによるマイニング時間への影響を検証した．

仙台市産業振興事業団の場合（図8），特に最小支持度1%以下では重み付けした方がマイニングに時間を要している．これは重みによって各ページが区別されたことでノードの種類が増え，支持度の計算により多くの時間が必要となったためである．

逆にKDDI研究所の場合（図9）は抽出に要する時間が全体的に短くなった．これは最小支持度を満たすページが少なくなったことで，支持度計算の時間が削減されたためである．

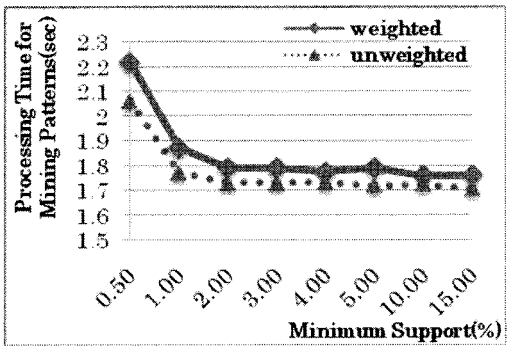

図 8：パターンの抽出に要した時間の比較（仙台市産業振興事業団）

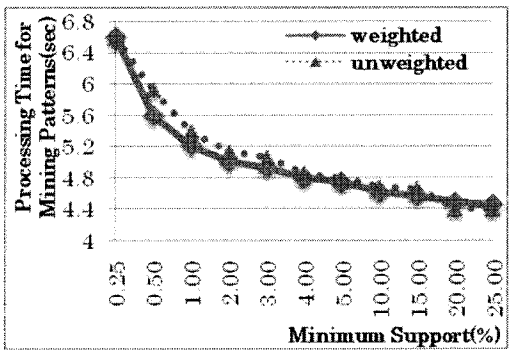

図 9：パターンの抽出に要した時間の比較（KDDI 研究所）

5　結論・今後の課題

本稿では，各ページに閲覧時間の長さに応じた重みを付け，グラフマイニングによりアクセスパターンを抽出する Web ログマイニング手法を提案した．そして実際の Web アクセスログに対して本手法を適用し，閲覧時間に基づくユーザの興味の度合いを踏まえたパターン抽出が可能となることを示した．

本手法の最終目標は Web サイト管理者が Web サイトを改善する際の支援を行うことであるため，抽出したアクセスパターンの中から Web サイト管理者にとってより有意義なものを選び，理解しやすい形で示すことが望ましい．従って，これを実現するための要素技術の研究が必要である．

このための今後の課題としては，まずユーザの興味と閲覧時間の関係について詳細な調査を行い，より的確な重み付け関数を導出する手法の研究が必要である．

重みについても，今回は 3 つの値だけを用いたが，より細かく区別できる可能性もあり，重みの値とパターンの関係の検証も必要である．

また，提案手法では各セッションのグラフは無向グラフとしたが，各リンクを移動した方向もわかるよう有向グラフによるマイニングを行うことで，より詳細な分析が可能となる．ただし，この場合は有向グラフに対応したグラフマイニングアルゴリズムを選択する必要がある．

謝辞

提案手法の評価のため，仙台市産業振興事業団及び KDDI 研究所より Web アクセスログを実験用データとしてご提供いただいた．ここに深謝する．

参考文献

[1] 石井研二：ホームページ アクセスログ解析の教科書，翔泳社 (2004)

[2] Iváncsy,R. and Vajk,I.: Frequent Pattern Mining in Web Log Data, Acta Polytechnica Hungaria, Journal of Applied Science at Budapest Tech Hungary, Special Issue on Computational Intelligence, Vol.3, No.1, pp.77-90 (2006)

[3] Inokuchi,A. Washio,T. and Motoda,H.: An Apriori -Based Algorithm for Mining Frequent Substructures from Graph Data, In Proc. PKDD 2000, pp.12--13, LNAI 1970, Springer-Verlag (2000).

[4] Kuramochi, M. and Karypis, G.: Frequent Subgraph Discovery, In Proc. IEEE ICDM'01, pp.313-320 (2001).

[5] Yan,X. and Han,H.: gSpan:Graph-Based Substructure Pattern Mining, In Proc. IEEE ICDM'02, pp.721-724 (2002).

[6] Nijssen,S. and Kok,J.N.: A Quickstart in Frequent Structure Mining Can Make a Difference, Proc. 10th. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp.647-652 (2004).

[7] 平手勇宇，山名早人.: 時間情報を含むシーケンシャルパターンマイニングの一般化，DEWS2006 (2006).

[8] Cooley,R. Mobasher,B. and Srivastava,J.: Data Preparation for Mining World Wide Web Browsing Patterns, Knowledge and Information Systems, Vol.1, No.1, pp.5-32 (1999).