Vol.2010-DBS-151 No.27
2010/11/13

時間グラフパターンを用いたWeb解析

押野泰平[†1]　浅野泰仁[†1]　吉川正俊[†1]

Web の構造的特徴に基づくグラフパターンは，web 解析や情報検索に重要な役割を果たしている．我々は web の構造情報だけでなく時間情報をも用いた新たなグラフパターンとして，時間グラフパターンというものを提案し研究を行ってきた．これまでの研究では，頻出グラフパターンを列挙することによってサンプルグラフ集合から時間グラフパターンをマイニングする手法を提案した．本稿ではマイニングされたパターンを用いた web 解析として，web サイトを話題の盛り上がり時における三つの役割として一次情報源，盛り上げ役，まとめ役に分類することを試みる．新たなグラフマイニング手法を用いることで，これまで達成できなかった三つの分類が可能になったことを示す．

Web Analysis using Time Graph Patterns

Taihei OSHINO,[†1] Yasuhito ASANO[†1]
and Masatoshi YOSHIKAWA[†1]

Graph patterns based on structural characteristics of the web have played important roles in web analysis and information retrieval. We have studied a new type of graph pattern named *time graph pattern* to estimate not only structural information but also temporal information of the web. In our previous work, we proposed a method for mining time graph patterns which enumerates frequent graph patterns from a sample graph set. In this paper, we performed a web analysis using mined patterns as an experiment. We tried to separate web sites related to extensively discussed topics into 3 classes, primary sources, triggers, and summarizers. propose a method for mining time graph patterns which enumerates significant graph patterns. As a result, we found that a new graph mining method help us perform the separation.

†1 京都大学大学院 情報学研究科 社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University

1. はじめに

Web の解析や情報検索のために web グラフを用いた研究は多く行なわれている．そのような研究の中で，web グラフの特徴的な部分構造であるグラフパターンは重要な役割を担ってきた．例えば，Kleinberg らの HITS アルゴリズム[1] では web ページの重要度を計算するためにハブ・オーソリティの関係を利用したグラフパターンを用いている．また，Kumar らの Trawling[2] では web コミュニティを抽出するために完全二部グラフをパターンとして列挙するという手法を用いている．このように，web を解析する上でリンク構造の特徴を利用するためにグラフパターンを用いることは有効である．一方，web の時間的な情報を利用することも web を解析するためには重要なことである．例えば，web ページやリンクの生成された日時を時系列的に追跡することで web 上での情報伝播や話題の盛り上がりなどを解析することができると考えられる．しかし，既存のグラフパターンではこのような時間的な情報を考慮できないため，web の構造的特徴と時間的特徴を同時に扱うことは難しい．

我々はこの問題を解決するために，時間グラフパターンという新たなグラフパターンの提案とそのマイニングを行なっている[3]．時間グラフパターンとは各ノードやエッジの持つそれらの生成日時といった時間情報をラベルとして組み込んだグラフパターンのことである．時間グラフパターンのマイニングとそれを用いた Web 解析等を通して，時間グラフパターンの有用性についての実証を行なってきた．そのために，web 上での話題の盛り上がりについて検証する実験を行った[3]．特に，ブログやニュースサイトなどのページが時間とともにどのように増加し，どのようなリンク構造として成長していくのかを時間グラフパターンを用いて解析を行なった．本研究の概要を図 1 に示す．まず，過去に盛り上がった話題のサンプルをいくつか用意し，それぞれの話題を扱ったページからなる web グラフ集合を作成する．次にグラフ中の各ノード (=ページ) の生成日時が，盛り上がりの前なのか最中なのか後なのか，といった時間的な特徴が区別できるようにノードにラベリングを行なう．そして gSpan[4] などの頻出パターン列挙アルゴリズムを用いてこの中から共通部分構造を時間グラフパターンとしてマイニングする．さらに得られたパターンによる web 解析として，主成分分析を用いた web サイト分類実験を行うことで，時間グラフパターンの有用性を明らかにした．事前に我々は話題の盛り上がりにおける web ページの主要な役割として一次情報源・盛り上げ役・まとめ役の三つがあることを確認した．一次情報源はその話題を最初に報道したページであり，盛り上げ役は盛り上がりのきっかけとなるページであり，まとめ役はその話題に関連する多くのページにリンクするページのことである．あるデータセット

© 2010 Information Processing Society of Japan

Vol.2010-DBS-151 No.27
2010/11/13

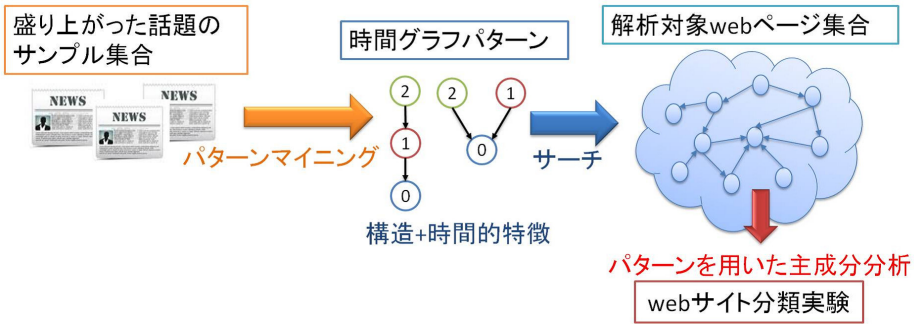

図 1 本研究の概要
Fig. 1 Overview of our study

中の web サイトがそのいずれのページを多く持つサイトであるかを時間グラフパターンによって分類するという実験を行なった．これまでの実験ではまとめ役とそれ以外のサイトを分類することができたが，一次情報源と盛り上げ役については分類できなかった．

本稿では，頻出パターンマイニング以外の新たなグラフマイニングアルゴリズムを用いることでこれらの問題への解決を試みた．具体的には GraphSig[5] という大規模なデータセットから重要なグラフパターンをマイニングする手法を用いた．この手法はデータセットに特徴的かつ一般的でない構造のグラフを重要なものとしてマイニングする手法である．またデータセットを類似したもののグループに分割することで，大規模なデータに対しても効率よく計算できるという特徴を持つ．この GraphSig はこれまで用いていた gSpan と比べてより大きなデータセットから時間グラフパターンがマイニングするのに適していることが分かった．さらに，得られたパターンを用いて新たな web サイト分類の実験を行なったところ，これまでの手法ではできなかった一次情報源と盛り上げ役の分類に成功した．これにより，時間グラフパターンを用いることで話題の盛り上がりにおける web サイトが一次情報源、盛り上げ役、まとめ役の三つの全ての役割について分類することができ，パターンの有用性が明らかになった．

以下，本論ではまず 2 節で関連研究について紹介する．3 節で時間グラフパターンマイニングの方法について説明し，4 節でマイニングされたパターンを用いた web 解析と考察を示す．5 節では gSpan と GraphSig のそれぞれを用いた場合の比較について述べる．最後に 6 節で結論を述べる．

2. 関 連 研 究

本節では，まず時間情報または構造情報を用いた web 解析に関する関連研究を紹介する．次に，本研究で時間グラフパターンのマイニングのために用いる二つのグラフマイニングアルゴリズムについてそれぞれ説明する．

2.1 Web 解析

Web 解析の既存研究として，時間情報を利用するものと構造情報を利用するものがある．

近年，時間情報を利用した多くの web 解析の研究がなされている．例えば，ソーシャルネットワークや web サイトのコミュニティやソーシャルネットワークの成長シミュレーションのためにいくつかのモデルが提案されている[6)–8)]．Leskovec らが提案した “forest fire” モデル[6)] では，web のべき乗則に従って成長し時間とともにグラフの直径が小さくなるという性質[9)] をモデル化できる．Kumar らは “burst” と呼ばれるブログコミュニティ内のブログ記事数の急激な増加現象に着目し，コミュニティの爆発的な成長を観測する研究[10)] を行なった．またブログ空間での情報伝播ネットワークを解析し，情報伝播の特性をリンクの情報を用いて数値化するという研究も行なわれている[11)]．しかし，本研究は時間経過に伴うグラフ構造の変化の過程そのものをグラフパターンとして抽出し，それを用いた応用を目的としている点で異なる．既存研究の多くはグラフ構造に着目した上で時間的な情報を組み込むということまでは行っているが，構造と時間の情報を独立に考えているものが多い．グラフが単調な成長なのか急激な成長なのかといった時間的な特徴と，グラフ構造のパターン抽出といった構造的な特徴を同時に考慮したものは少ない．また本研究で提案する時間グラフパターンは web に限らず，ノードやエッジに時間情報を持つようなネットワークでの応用が可能であると考えている．

構造情報を利用するためにグラフパターンを用いた研究として以下のようなものがある．Kleinberg の HITS アルゴリズム[1)] はハブとオーソリティという概念を用いている．よいハブはよいオーソリティに多くリンクし，よいオーソリティはよいハブに多くリンクされているというパターンを利用することによって，特定の話題に関する良質なページ発見を行っている．Kumar らの Trawling[2)] という手法では完全 2 部グラフと呼ばれるグラフパターンを列挙することによってコミュニティ発見を行っている．しかし，これらのパターンは時間情報を考慮することができない．また，ブログ空間における話題の伝播を解析した研究としては Leskovec らの研究[12)] がある．リンクを情報伝播とみなし，そのリンク構造の広がりを cascade と呼ばれるグラフパターンで表現する．cascade とは自身はリンクを持たな

© 2010 Information Processing Society of Japan

いページから被リンクを逆にたどって到達可能な全ページとそれらのリンクからなる非巡回グラフのことであり，このグラフパターンの特徴に関しての調査がなされている．さらにこの cascade を用いてブログ空間の解析[13] も行われている．ブログサイトに含まれる全 cascade の出現頻度を数えることで，ブログサイトを行，cascade グラフパターンを列とした行列が得られる．その行列に対し主成分分析を行なうことで，ブログサイト集合を政治的に保守派な人のブログサイトと噂好きなブログサイトとに分類することに成功している．この cascade グラフパターンも時間の情報を十分に考慮しているとは言えない．そのため，同じ構造の cascade があったとしてもその情報が伝わる早さの違いを区別することができない．4 節で詳しく述べるが，本研究では cascade の代わりに時間グラフパターンを用いることで同様の解析を行ない，時間的な情報を得ることを目指している．特に，話題の盛り上がりについて扱い，web サイトがどの時期にその話題に言及するかという特徴で分類するという実験を行なう．

2.2 グラフマイニングアルゴリズム

後述する我々の手法では複数のグラフから特徴的な部分構造を時間グラフパターンとしてマイニングする．そのときに利用する二つの既存のグラフマイニングアルゴリズムについて説明する．

我々のこれまでの手法[3] では Yan と Han によって提案された gSpan アルゴリズム[4] を採用していた．これはグラフの集合から頻出する部分構造を列挙するアルゴリズムである．gSpan は入力としてグラフの集合 D と最小支持度 $minSup$ を与えると，頻出パターン，すなわち D 中の $minSup$ 個以上のグラフに部分構造として現れる全てのパターンを効率よく列挙する．図 2 のグラフ集合を入力とした場合の出力を図 3 に示す．最小支持度を $minSup = 3$ とすると出力として G_1，G_2，G_3 のうち三つ全てのグラフに現れる p_1，p_2 が得られる．また $minSup = 2$ ならば三つのうち二つ以上のグラフに現れる p_3 パターンのみが得られる．

gSpan のような頻出パターン列挙の問題点として，以下の二つが挙げられる．

(1) 非常に計算コストがかかる．これは部分グラフ同型判定という NP 完全問題を含むためである．

(2) 多くのパターンが列挙されるがそれら全てが有用とは限らない．入力グラフ集合中で頻出するものの中には一般的な部分構造も多く含まれてしまう可能性があるからである．

このような問題を解決するために，Sayan らは GraphSig アルゴリズム[5] を提案している．GraphSig では，まずグラフ集合中の各グラフごとに，各ノードを始点とした RWR(Random Walk with Restarts) を実行することで一つのグラフを複数の特徴ベクトルに変換する．RWR とは各ステップごとに一定確率で始点ノードに戻るランダムウォークのことである．そしてこれらの特徴ベクトルから “重要” なものをマイニングし，その結果を類似したベクトルが同じグループになるように分類する．最後に，特徴ベクトルのグループをもとに，共通した部分構造を持つグラフ同士が同じグループに含まれるようにもとのグラフ集合を分割できる．これにより，それぞれの分割した集合に対して gSpan のような頻出パターンマイニングアルゴリズムを大きな最小支持度で適用すれば特徴的な部分グラフがマイニングできる，という手法である．この手法は従来のようにグラフ集合から直接頻出パターンをマイニングするのではなく，小さなグループに分割し，それぞれのグループに対して頻出パターンマイニングを適用しているため，(1) の問題を軽減している．また，(2) の問題に対しては，グラフや特徴ベクトルに重要度を表す $p\text{-}value$ という尺度を与えている．p-value とは，その

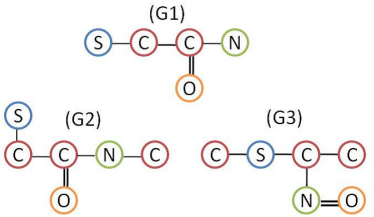

図 2 入力サンプルグラフ集合
Fig. 2 Input sample graphs

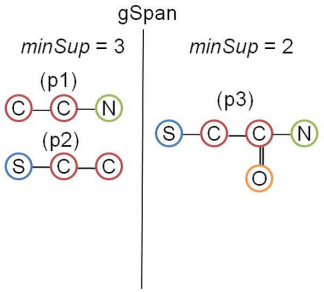

図 3 gSpan による出力
Fig. 3 Output of gSpan

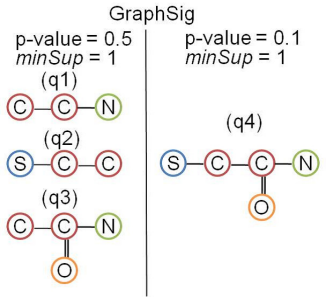

図 4 GraphSig による出力
Fig. 4 Output of GraphSig

© 2010 Information Processing Society of Japan

グラフ (特徴ベクトル) が入力データ集合での出現頻度がランダムグラフ (ベクトル) 集合での出現頻度を下回る確率である．通常，入力データ集合では頻度が高く，ランダムデータ集合では頻度が低いようなデータはその入力データ集合の特徴を大きくとらえていると考えられる．つまり低い p-value を持つものが重要であるという観点から，ユーザが指定した閾値以下の p-value を持つパターンのマイニングを行なっている．またそのときに gSpan と同様に最小支持度 *minSup* を与えることができる．p-value によって重要度を評価することで，与える *minSup* は小さな値であっても重要なパターンが得られることが GraphSig の特徴である．図 2 のグラフ集合を入力とした場合の出力を図 4 に示す．GraphSig の場合は得られたパターンの数は p-value=0.5 のときで $q1$，$q2$，$q3$ の三つ，p-value=0.1 のときで q_4 一つのみとなっている．これらは $minSup = 1$ を指定したものであるが，gSpan の結果とよく似たグラフが重要なものとして得られている．なお，入力グラフのサイズや数が多くなると gSpan の場合は数多くのパターンが列挙されるが，GraphSig の場合は出力パターン数が急激に増加しないという特徴を持つ．

3. 時間グラフパターンマイニング

本節では我々が提案する web グラフからの時間グラフパターンのマイニング方法について説明する．時間グラフパターンとはノードやエッジにその生成日時の情報をラベルとして組み込んだグラフパターンのことである．この時間グラフパターンを用いて，web 上での話題の盛り上がりを解析の対象とする．そのため，web ニュースを報道するサイトや，それを紹介するブログやニュースまとめサイト，そしてその周辺のページをデータとして用いる．全体的な処理の流れは，以下のようになる．

(1) web ページを収集し，web グラフ集合を作成する
(2) 各ページの日付情報を解析し，ノードにラベリングを行なう
(3) グラフパターンマイニングアルゴリズムを適用し，時間グラフパターンを得る

今回は (3) において GraphSig を適用する．

3.1 Web グラフ集合の作成

まず過去に盛り上がった話題をいくつか選定しておく．盛り上がった話題とは，ある一定期間の間にその話題を扱うページ数が急激に増加する話題のことである．選定した各話題ごとにそれを扱った web ページ集合からなる一つの web グラフを作成する．その話題を扱ったページを効率よく収集するために Google 検索エンジン[⋆1]を利用した．話題に特徴的なキーワードと盛り上がった期間を指定して検索することで，検索結果のトップ k(例えば 300) からなる web ページ集合とそのページの日付情報 (Google によって取得された日付を用いる) を得ることができる．このページ集合のリンク構造を解析することで一つの時間情報付きの web グラフを作る．

3.2 日付情報解析とラベリング

Web ページ集合は特定の期間を指定し，その話題に関するページを収集している．Web における話題の盛り上がりとは概ね図 5 のような傾向でページ数が推移していくことが多い．左上の web グラフはある話題を扱ったページから構成されており，各ノードの日付はそのページの生成日時を表している．この web グラフの時間経過と累積ページ数の関係を表したものが図 5 の右のグラフである．まず初期にはその話題を扱ったページはそれほど多くはないが，話題が広まっていくにつれページ数は増加していく．そして成長期になると急に増加していき，盛り上がりはピークに達する．その後しばらくするとその話題に言及する新たなページが増えなくなっていき，安定期へと移行していく．ページ数の増減を解析し，急激な変化があればそれは各期間の境界だと見なすことができる．このようなページ数の変曲点は，web における話題の盛り上がりの時間的特徴が変化した点である．そこでこのような時間的特徴を組み込むため，そのページが生成された時期に応じて，初期は 0，成長期は 1，安定期は 2 のようにノードに番号を付けたラベルを付与する．図 5 の場合だと Apr 4 と Apr 6 にページ数が大きく変化していることが分かる．そのためこれらの日付をラベリングの境界とし，Apr 1 から Apr 3 に作成されたページにはラベル 0 を，Apr 4 と Apr 5 に作成されたページにはラベル 1 を，Apr 6 以降に作成されたページにはラベル 2 を付与する．これにより，一つの web グラフから一つのラベル付きグラフに変換することができる．つまり図 5 において左上の web グラフは左下のようなラベル付きグラフになる．

3.3 特徴的なグラフパターンのマイニング

3.1 節で得た web グラフ集合に 3.2 節の方法によって変換したラベル付きグラフ集合を入力として，グラフマイニングアルゴリズムを適用することで，時間情報を表すノードラベルが付与された時間グラフパターンが得られる．実際に過去に盛り上がった話題を七つ選定し，その話題に関する web ページ集合に対し $minSup = 4$ で gSpan を適用した場合は 107 個のパターンを得ることができた[3]．その一部を図 6 に示す．gSpan では大きな入

⋆1 http://www.google.co.jp/

© 2010 Information Processing Society of Japan

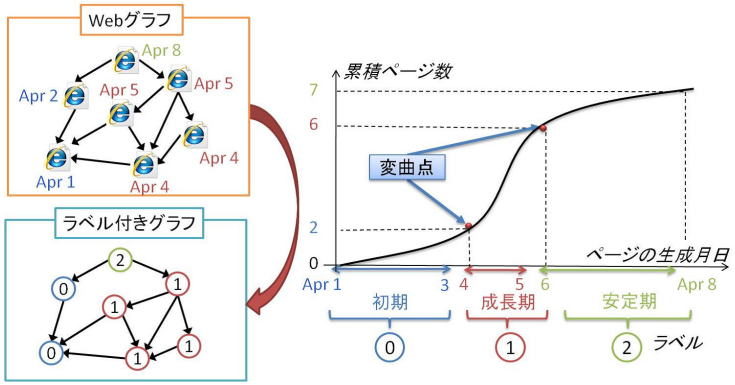

図 5 話題の盛り上がりにおけるラベリング
Fig. 5 Labeling for an extensively discussed topic

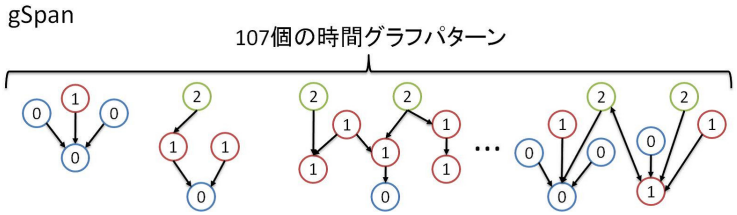

図 6 gSpan によってマイニングされた 107 個の時間グラフパターンの一部
Fig. 6 A part of 107 mined time graph patterns by gSpan

力データ集合に対しては実用的な時間で解を得ることは難しく，また *minSup* の値が小さくなるほど指数的に計算量が増加する．実際に話題を八つ以上集めて得られたグラフ集合に対し，$minSup \leq 5$ として gSpan を適用したところ，実用的な時間で解を得ることはできなかった．

今回は 15 個のグラフ集合に対し，$minSup = 8(50\%)$, p-value=0.1 で GraphSig を適用した．その結果，55 個のパターンを得ることができた．その一部を図 7 に示す．gSpan と比較して，約二倍のサイズのグラフ集合を対象としているにも関わらず得られるパターンの数は半分程度であった．

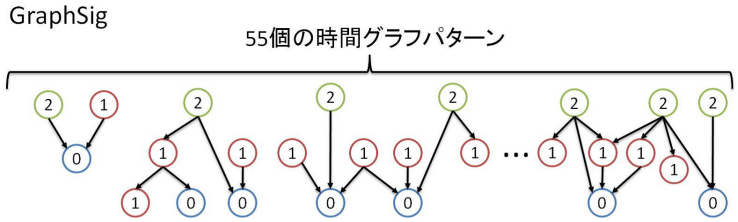

図 7 GraphSig によってマイニングされた 55 個の時間グラフパターンの一部
Fig. 7 A part of 55 mined time graph patterns by GraphSig

4. 時間グラフパターンを用いた web 解析

マイニングされた時間グラフパターンを用いた web 解析の手法について説明する．我々は得られたパターンを利用することで web の構造と時間の二つの特徴を同時に解析することができると考えている．解析対象として web サイトを事前に収集しておき，それらのサイトが話題の盛り上がりに関してどのような役割を果たしているか，という観点での解析を行なった．我々は主成分分析を適用することで，時間グラフパターンを用いて web サイトの性質による分類を行なうことができた．本節では web 解析の方法と結果について説明する．

4.1 データセットの作成

解析対象となる web ページを収集するために，まず流行の話題に敏感で影響力のあるブロガーや，日々いろいろなニュースを紹介するブログ，そしてそのようなページのまとめサイトなどをいくつか基点とし，それらのサイトの全ページからリンクと被リンクを 1 ホップずつたどりページを取得する．こうして得られたページから孤立したノードを除いた 850 のページ，732 のリンクからなる一つの web グラフを作成し，解析対象データセットとした．

4.2 主成分分析

McGlohon ら[13] は，cascade と呼ばれるグラフパターンが情報伝播を特徴づける性質を持つと考えた．ブログサイトに含まれる全 cascade の出現頻度を数えることで，ブログサイトを行，cascade グラフパターンを列とした行列を得，その行列に対し主成分分析を行なうことで，ブログサイト集合を政治的に保守派な人のブログサイトと噂好きなブログサイトとに分類することに成功している．

我々は 3.3 節で得られた時間グラフパターンには話題の盛り上がりを特徴づける性質が含まれていると考えている．そこで，我々は解析対象のデータセット中の各サイトに含まれる

© 2010 Information Processing Society of Japan

Vol.2010-DBS-151 No.27
2010/11/13

全時間グラフパターンの各ラベルごとの出現頻度を数えることで，web サイトを行，時間グラフパターンとラベルの対を列として行列を得て，同様の解析を試みた．つまりある時間グラフパターン p_i においてラベル ℓ と対応するページをサイト s_j が持つ数を $e_{(i,l)j}$ とすると，(i,l) 行目の j 列目の要素が $e_{(i,l)j}$ となる行列である．例を図 8 に示す．まず解析対象データセット中に三つのサイト A，B，C があり，また図中に示すような二つの時間グラフパターン 1 と 2 がマイニングされていたとする．このとき，まずこれらのパターンと同じ部分構造がデータセット中に存在するか確認する．この場合，データセットにはパターン 1 のような構造は一つ，またパターン 2 のような構造は二つ存在することが確認できる．次に，マッチした部分構造の web ページの生成順とパターンのラベリングの順が整合しているかどうかを確認する．例えばパターン 1 の場合，ラベル 2→ ラベル 1→ ラベル 0 の順にエッジが張られているが，上のデータセット中の対応する部分グラフのページ生成日時は May 5→ May 4→ Apr 1 となっていることから整合性がとれていることと言える．パターン 1 におけるラベル 0 のノードに対応するページはサイト B に，ラベル 1 とラベル 2 のノードに対応するページはサイト A にそれぞれ一つずつあるため，$e_{(1,0)B} = e_{(1,1)A} = e_{(1,2)A} = 1$ となる．同様にパターン 2 も数えると，図 8 の右下のような行列が得られる．この行列に対して主成分分析を適用する．

4.3 Web の時間的特徴と構造的特徴

マイニングされたパターンが web の時間的特徴と構造的特徴を解析できるかを確かめる．まず，盛り上がった話題に関する web ページを観察し，以下の三つのタイプの役割を確認した．それぞれのタイプは以下のような時間的，構造的な特徴を持つ．

(1) 一次情報源: ある話題を最初に報道したページ．
(2) 盛り上げ役: 一次情報源を発見し，リンクを張り話題を紹介するページ．著名なブロガーなどがとりあげたことをきっかけとして話題が広まっていくことも多い．
(3) まとめ役: これまでの議論をまとめたり，話題に関連する多くのページにリンクすることで紹介するページ．

我々はデータセット中の各 web サイトが上のいずれの役割を果たすページを多く含むかをあらかじめ人手で検証し，サイトを役割ごとに分類した．これは web サイト分類実験の正解セットとなる．これまでの実験[3] では，gSpan によってマイニングされたパターンを用いた分析結果の第三主成分までの主成分値を図 9 のようにプロットした．図中の “○”は一次情報源サイトを表し，“×” は盛り上げ役サイトを，“△”はまとめ役サイトを表している．三次元主成分空間上で「一次情報源または盛り上げ役」サイトと「まとめ役」サイトに分離

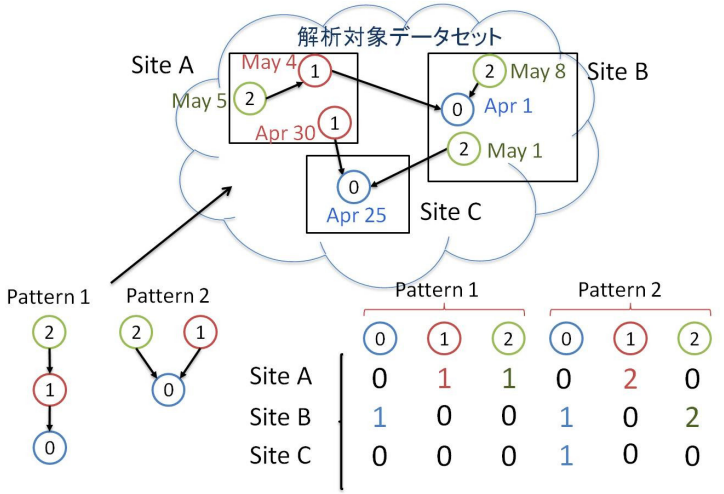

図 8 (時間グラフパターン, ラベル) サイト 行列
Fig. 8 (time graph pattern, label) site – matrix

できることを確認した．これにより時間グラフパターンが時間と構造の解析に有用であることを示している．なお，第三主成分までの累積寄与率は 76.3%であり，もとのパターンの情報を十分に保持していると考えられる．

今回は，同様の実験を GraphSig によってマイニングされたパターンを用いて行う．第三主成分までの主成分値をプロットした散布図を図 10 に示す．累積寄与率は第二主成分までで 89.2%，第三主成分までで 94.8%であるが gSpan との比較のため第三主成分までを用いた．図から分かるように一次情報源サイトと盛り上げ役サイトが分離されていることが確認できる．ただし，今回はまとめサイトに関しては特定のパターンのみに大量にマッチングしたことによって異常な外れ値であったり，まったくマッチングしなかったサイトも多く，gSpan での場合と比べてよい結果を得ることはできなかった．しかし，一次情報源と盛り上げ役という時間・構造的な特徴を持つこれらのサイトを区別できたことから，GraphSig で得られたパターンによっても時間と構造の情報をとらえることができることが分かった．この結果とこれまでの実験結果[3] から，話題の盛り上がりに関する一次情報源，盛り上げ役，まとめ役という三つの役割を時間グラフパターンによって分離できることが分かった．この結果から我々の提案した時間グラフパターンは時間と構造の解析に有用であると言える．

© 2010 Information Processing Society of Japan

Vol.2010-DBS-151 No.27
2010/11/13

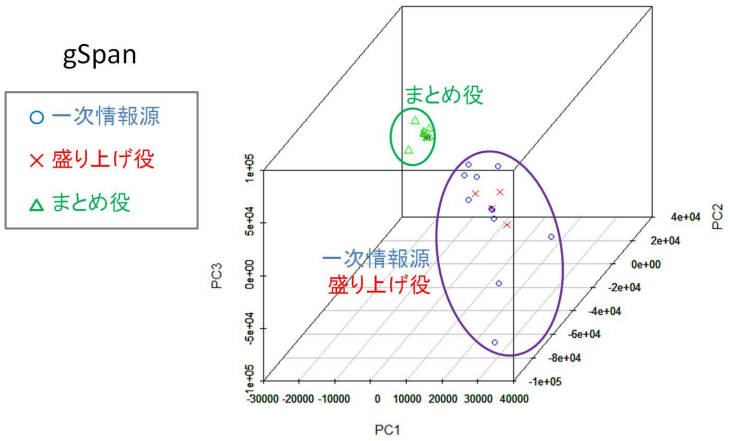

図 9 gSpan での第三主成分値の散布図
Fig. 9 Scatter diagram of first three principal component scores using gSpan

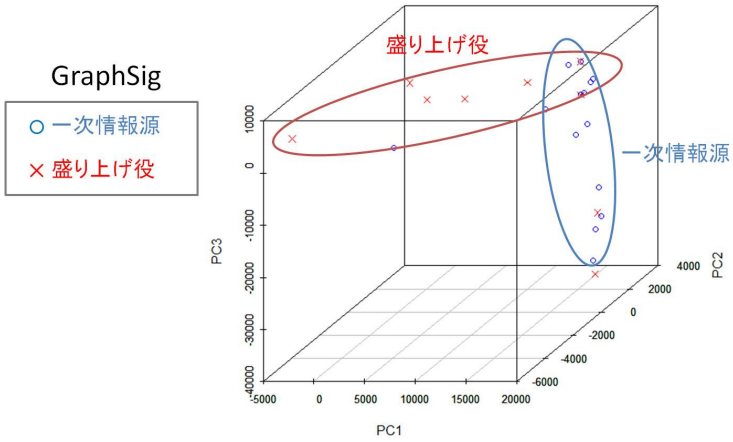

図 10 GraphSig での第三主成分値の散布図
Fig. 10 Scatter diagram of first three principal component scores using GraphSig

5. GraphSig と gSpan の比較

本節ではこれまでの研究と今回の研究によって得られた結果を比較し，gSpan と GraphSig の特徴や性質について考察する．

(1) 対象グラフセット数

ノード数を 50 から 70 程度，ラベル数を 0, 1, 2 の三種類とした場合，gSpan では 7 個のグラフからは数分程度で解を得られたが，8 個以上では結果を得ることができなかった．GraphSig の場合，現在 15 個程度のグラフ集合に対しても結果が得られることを確認している．ただし，グラフセットによっては数時間かかることもある．それ以上大規模なグラフセット (例えばグラフ数 100 個以上) に対しても，データセットによっては数分程度で解を得られることもある．Sayan らによって行なわれたデータセットの大きさに対する gSpan と GraphSig の比較実験[5] によると，gSpan はデータセットの大きさに対して指数的に計算時間が増加するのに対し，GraphSig では線形時間で処理できるという結果が得られている．

(2) マイニングされたパターン数

gSpan では 7 個の入力グラフから 107 個の頻出パターンが列挙された．GraphSig は 15 個の入力グラフから 55 個の重要パターンが得られた．GraphSig は入力グラフが大規模であっても得られるパターン数が gSpan と比較するとかなり少ないことが分かる．これは p-value による制限により，ノード数の小さい一般的なパターンの多くが解に含まれないためであると考えられる．少ないパターンでも web 解析に十分な結果を得ることができるという点からも，GraphSig の方がより重要度の高いパターンのマイニングに焦点を当てていることが分かる．

(3) マイニングされたパターンの特徴

gSpan では全てのパターンのノード数は 10 以下であったが，GraphSig ではノード数が 11 個以上のパターンも 10 個程度含まれていた．これはノード数が大きいと頻度は下がるが特殊な構造となるため，p-value の値が低くなりやすいためと考えられる．

(4) web 解析に適用した場合の特徴

gSpan で得られたパターンはまとめ役サイトとそれ以外を分離する上で有効であった．GraphSig は一次情報源と盛り上げ役を区別する上で有効であった．また gSpan では，各主成分に大きく寄与するパターンの特徴を観察し，その結果第三主成分の軸では早い時期に作られたページ (つまりラベル 0 のページ) を持つサイトほど負方向

© 2010 Information Processing Society of Japan

に大きな値となり，遅い時期に作られたページ (つまりラベル 2 のページ) を持つサイトほど正方向に大きな値を持っていた[3]．このような結果は主成分が時間情報を含んでいることを反映していると考えられるが，今回 GraphSig を用いたパターンの主成分ではそのような特徴は見られなかった．この点に関しては gSpan を用いた方がより時間情報を考慮できるとも言える．

上のような差異を比較すると，gSpan と GraphSig を用いて得られるパターンは異なる特徴を持つことが分かる．扱えるデータセット数は GraphSig の方が優れているものの，得られたパターンは必ずしも GraphSig が gSpan より優れているというわけではなかった．しかし，いずれの手法を用いた場合でも時間と構造の情報を解析する上では役立つことが今回の実験により明らかになった．今後はこれらの長所のみを上手く組み合わせるようなグラフマイニング手法を構築することができれば，さらに有用な時間グラフパターンのマイニングができると期待できる．

6. ま と め

時間と構造の二つの特徴を同時に扱うことのできる時間グラフパターンを提案し，そのマイニングとそれを用いた web 解析を行なった．マイニング手法として gSpan と GraphSig の二つのアルゴリズムを用いた．得られたパターンを用いることで，web の話題の盛り上がりについて重要な役割を持つ，一次情報源，盛り上げ役，まとめ役の三役にサイトを分類することができた．その結果から，時間グラフパターンが構造的な特徴と時間的な特徴を解析するために有用であることを示した．

今後の課題として，gSpan と GraphSig の長所を組み合わせたようなグラフマイニング手法を構築することで，より有用な時間グラフパターンのマイニングを行う．さらに，より大規模なデータセットから時間グラフパターンを効率よく検索する手法を構築したいと考えている．これにより，時間グラフパターンを用いたさらなる応用ができると考えている．例えば，今回は話題の盛り上がりに関して特徴的なパターンをマイニングしたが，得られたパターンを直接検索することで未知の話題の盛り上がりを検出することなどが期待できる．時間グラフパターンを用いてグラフの時間と構造的特徴をとらえることで様々な応用を考えていきたい．

参 考 文 献

1) Kleinberg, J.M.: Authoritative sources in a hyperlinked environment, *Journal of the ACM*, Vol.46, No.5, pp.604–632 (1999).
2) Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A.: Trawling the web for emerging cyber-communities, *Computer Networks*, Vol.31, pp.1481–1493 (1999).
3) Oshino, T., Asano, Y. and Yoshikawa, M.: Time Graph Pattern Mining for Web Analysis and Information Retrieval, *WAIM 2010: Proceedings of the 11th International Conference on Web-Age Information Management*, pp.40–46 (2010).
4) Yan, X. and Han, J.: gSpan: Graph-Based Substructure Pattern Mining, *Proceedings of the 2002 IEEE International Conference on Data Mining*, pp.721–724 (2002).
5) Ranu, S. and K.Singh, A.: GraphSig: A Scalable Approach to Mining Significant Subgraphs in Large Graph Databases, *ICDE '09: Proceedings of the 25th International Conference on Data Engineering*, Washington, DC, USA, IEEE Computer Society, pp.844–855 (2009).
6) Leskovec, J., Kleinberg, J. and Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations, *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 177–187 (2005).
7) Tawde, V.B., Oates, T. and Glover, E.: Generating Web Graphs with Embedded Communities, *Algorithms and Models for the Web-Graph*, pp.80–91 (2004).
8) Pennock, D.M., Flake, G.W., Lawrence, S., Glover, E.J. and Giles, C.L.: Winners don't take all: Characterizing the competition for links on the web, *Proceedings of the National Academy of Sciences of the United States of America*, National Acad Sciences, p.5207 (2002).
9) Huberman, B. and Adamic, L.: Growth dynamics of the World-Wide Web, *Nature*, Vol.401, p.131 (1999).
10) Kumar, R., Novak, J., Raghavan, P. and Tomkins, A.: On the Bursty Evolution of Blogspace, *WWW '03: Proceedings of the 12th International World Wide Web Conference*, pp.159–178 (2003).
11) 風間一洋，今田美幸，柏木啓一郎：ブログ空間の情報伝播特性を用いた情報源の多面的ランキング，*WebDB Forum 2009* (2009).
12) Leskovec, J., McGlohon, M. and Faloutsos, C.: Cascading Behavior in Large Blog Graphs, *SDM '07: Proceedings of Society of Applied and Industrial Mathematics: Data Mining* (2007).
13) McGlohon, M., Leskovec, J., Faloutsos, C., Hurst, M. and Glance, N.: Finding patterns in blog shapes and blog evolution, *Proceedings of ICWSM* (2007).

© 2010 Information Processing Society of Japan