時系列データを用いたWebグラフマイニング

押野　泰平† 　浅野　泰仁†† 　吉川　正俊††

† 京都大学工学部情報学科　〒606-8501 京都市左京区吉田本町
†† 京都大学情報学研究科社会情報学専攻　〒606-8501 京都市左京区吉田本町
E-mail: †oshino@db.soc.i.kyoto-u.ac.jp, ††{asano,yoshikawa}@i.kyoto-u.ac.jp

あらまし　現在，グラフマイニングは多くの分野で盛んに研究され，グラフの特徴的な構造を用いて有益な情報を抽出する試みも広く行われている．Webグラフマイニングを行う手法も数多く提案され，例えば，完全2部グラフという構造を利用することでWebコミュニティの発見に役立てる研究もある．多くのWebグラフマイニング手法は主にページをノード，リンクをエッジと見なした静的なグラフ構造を解析するものだが，それらはページやリンクの生成日時といった時系列データを利用することができない．Webは動的なものであるが，Webにおけるグラフ構造は時間とともに動的に変化する．そのため，ページやリンクの時系列データを用いることでより有益な情報を得ることが期待できる．そこで我々は，時系列データを考慮したWebグラフのマイニング手法を提案する．具体的には，各ページの生成日時を反映したノードラベルを導入する．この手法を用いたWebグラフマイニングの応用例として，ブログやニュースサイトに適用し，そのグラフ構造から特徴的なパターンをいくつか抽出することができた．
キーワード　グラフマイニング，Webグラフ，時系列データ

Web Graph Mining using Time-Series Data

Taihei OSHINO†, Yasuhito ASANO††, and Masatoshi YOSHIKAWA††

† School of Informatics and Mathematical Science Faculty of Engineering,　Kyoto University
Yoshida Honmachi,　Sakyo-ku,　Kyoto,　606–8591 Japan
†† Department of Social Informatics,　Kyoto University
Yoshida Honmachi,　Sakyo–ku,　Kyoto,　606–8591 Japan
E-mail: †oshino@db.soc.i.kyoto-u.ac.jp, ††{asano,yoshikawa}@i.kyoto-u.ac.jp

Abstract　Graph mining has been studied in many fields for obtaining useful information from a graph by finding its characteristic substructures. For example, several methods for Web graph mining has been proposed. Some of these methods identify communities on the Web by finding complete bipartite graphs. Most methods for Web graph mining deal with only static graph structures whose nodes are pages and edges are links, that is, they utilize no time-series data of pages and links, such as the creation date of each page or link. The Web is dynamic, however, and the graph structure of the Web is changed dynamically over time. Therefore, it is expected that we can obtain more useful information by utilizing time-series data of pages and links. We propose a new technique for Web graph mining to utilize time-series data of pages and links, by introducing a node label reflecting the creation date of the page corresponding to the node. By applying our technique to graphs among blogs and news sites, we succeeded in identifying some characteristic patterns in them.
Key words　Graph Mining,　Web Graph, Time-series data

1. はじめに

現在，グラフマイニングは様々な分野で盛んに研究され，応用されている．化学の領域では分子構造の解明のために，原子をノード，原子間の化学的な結合をエッジとしたグラフのマイニングが行われている．例えばHIV(Human Immunodeficiency Virus:ヒト免疫不全ウィルス) データから分子構造マイニングを行った医学的な研究 [1] がある．また社会学ではソーシャルネットワーク解析のため様々なグラフマイニングの手法が用いられ，人と人とのつながりをグラフと見なすことで多くの社会的な知見を得ることができる [2]．実社会での応用として，病院での医療業務にグラフマイニングを利用する試みもある [3].

そして Web 上のページをノード，リンクをエッジとした Web グラフを解析する研究も非常に多くなされている．Kleinberg は Web ページの評価のためにリンク構造を用いて重要なページを探し出す HITS アルゴリズム [4] を提案した．Kumar らは Web コミュニティを抽出するために，グラフから全ての完全 2 部グラフを列挙する Trawling [5] という手法を用いた．また，巨大な疎グラフに対して実用的な時間で極大クリークを列挙するアルゴリズム [6] も宇野によって提案された．これら Web 上の特徴的なリンク構造を見つけ，情報発見に役立てる手法を確立するためには人間が Web を調査して何らかの特徴的な構造を発見し，その構造を持つ部分グラフを列挙するアルゴリズムを構築するという手順を踏んでいた．しかし，グラフマイニングを利用して頻出パターンを見つければ自動的に特徴的な構造を発見することができるようになる可能性がある．

また，Web における時系列データの分析も近年研究されている．例としては，あるブログコミュニティ内での話題の盛り上がり方と，ブロガーの記事のタイムスタンプに注目して重要なブロガーを発見しようとする研究 [7] やソーシャルブックマークにおける利用者とタグ付けの時系列データから新たなページの注目度を予測する研究 [8] などがある．

既存の Web グラフマイニングの研究ではリンクとページのグラフ構造のみに注目したものが多く，時系列データを考慮したものはそれほど多くない．しかし，Web 上のページやリンクは時間の変化とともに増減する動的なものであり，変化の仕方も様々である．Web グラフマイニングに時系列データを取り入れることで，今までに発見されていない，情報発見に役立つ新たなパターンが見つかることも期待できる．

そこで本研究では時系列データを Web グラフマイニングに取り入れる手法を提案する．具体的には，各ページの生成日時を反映したノードラベルを導入する．ブログ記事とそれを参照するまとめサイトの間にできる Web グラフに適用し，特徴的なパターンを抽出することができた．

以下， 本論ではまず 2 章で関連研究を紹介する．3 章で今回提案する手法について述べ，4 章で実験として実際に手法を適用した応用例と考察を示し，5 章で課題を挙げ，最後に 6 章で結論を述べる．

2. 関連研究と基本事項

本節ではまずグラフマイニング一般に関して，実世界での応用例について述べる．次に Web の時系列データを利用した研究を紹介する．さらに Web コミュニティの抽出にグラフパターンを用いた研究について概要を述べる．最後に Yan と Han によって提案された既存のグラフマイニング手法であり今回用いた頻出パターン抽出アルゴリズム gSpan について説明する．

2.1 グラフマイニングの応用例

グラフマイニングは様々な分野に応用できる．医学の例では化合物の分子構造をグラフ化し，さらに HIV に対する活性の度合いのデータを用いてグラフマイニングを行い，特徴的なパターンを抽出している [1]．また社会学ではソーシャルネットワーク解析のため様々なグラフマイニングの手法が用いられ，

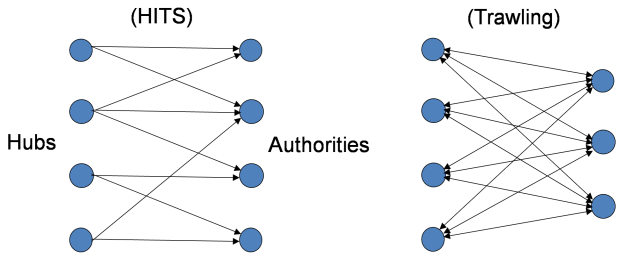

図 1　パターンの例

人と人とのつながりをグラフと見なすことで多くの社会的な知見を得ることができる [2]．実社会での応用例としては，病院で行う医療行為のプロセスを効率化するために行為の組み合わせパターンで頻出しているものを見つけるために時間とグラフを利用してマイニングを行っている [3]．このようにグラフを用いることで様々なデータをモデル化することができ，そこから有用な知見を得ることができる．

2.2 Web 時系列分析

近年，Web の時系列データ分析の研究が多くなされている．時系列データを用いたブログ解析サービスとして以下のようなものがある．blogWatcher [9] はブログの収集とテキストマイニングを行い話題の注目度や評判情報の抽出などを行うものである．既存の Web サービスである kizasi.jp [10] はブログを対象に，出現する単語を時系列で解析することで話題の変化やトレンドを知ることができるサイトである．

Web の時系列的な成長を扱った研究も多く存在する．例えばブログコミュニティにおけるページや記事数がある期間内にどのくらいの数が増加しているかに着目し，話題の急速な盛り上がりについて解析した研究として [7] [11] などが知られている．中島ら [7] はさらにコミュニティの盛り上がりに重要な影響を及ぼすブロガーの発見を行っている．またソーシャルブックマークにおける利用者とタグ付けの時系列データから新たなページの注目度を予測する研究 [8] などがある．これらの研究はコミュニティ内のページ数の変化を解析したものであり，本研究はグラフ構造の時系列的な変化を解析している点で異なる．リンク構造の成長パターンを抽出するための手法を構築することを本研究の目的とする．

2.3 グラフパターンを用いた Web グラフマイニング

Web グラフマイニングにおいてグラフパターンを利用して情報を得ようとする研究は数多くなされている．例えば Web コミュニティ発見のためにグラフパターンを利用する手法として以下のようなものがある．

Kleinberg の HITS アルゴリズム [4] はハブとオーソリティという概念を用いている．これらは相互に強めあう関係で，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している．Kumar らの Trawling [5] という手法では完全 2 部グラフと呼ばれるパターンを探すことでコミュニティ抽出を行っている．これらの手法で用いられているパターンの例を図 1 に示す．

頻出パターン抽出に関する研究は数多くある [12] [13] が，こ

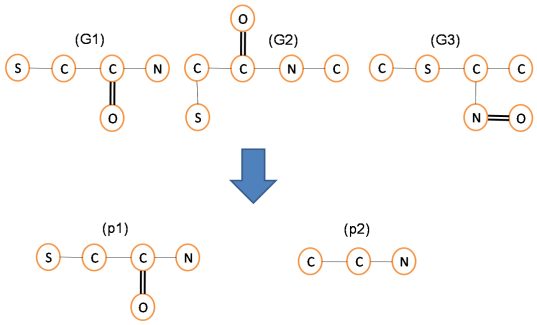

図 2　頻出パターン

こでは Yan と Han によって提案された，複数のグラフの集合から頻出するパターンを抽出するアルゴリズム gSpan [14] [15] について説明する．頻出パターン抽出アルゴリズムの概要を図 2 に示す．まず対象とするグラフはノードとエッジのそれぞれにラベルが与えられている無向単純グラフである．例えば分子構造をグラフで表現するにはノードに元素記号を表すラベルを振り，エッジには単結合または二重結合を表すラベルを振ればよい．G_1, G_2, G_3 のようなグラフの集合を考えると，その中に現れるパターンは数多く存在する．g_1 のようなパターンは G_1 と G_2 の二つのグラフに現れ，g_2 のようなパターンは三つ全てのグラフに現れることが確認できる．このように多くのグラフに現れるパターンだけを列挙するアルゴリズムが gSpan 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，D 中に含まれる数が $minSup$ 以上である全てのパターンを列挙する．つまり図 2 の場合，入力を

$$D = \{G_1, G_2, G_3\},\ minSup = 2$$

とすると g_1, g_2 のように二つ以上のグラフに現れるパターンが全て出力される．まず頻度は以下のように定義する．

[定義 1]（frequency） グラフ集合 $D = \{G_1, G_2, \cdots, G_n\}$ が与えられたとき，パターン g の **頻度(frequency)** を $frequency(g)$ と表し，

$$frequency(g) = \left|\{G_i | G_i \in D, g \subset G_i\}\right|$$

で定義する．また，g が D において**頻出** (**frequent**) であるとは，ある最小支持度 $minSup$ に対し，

$$frequency(g) \geqq minSup$$

であることを言い，このときの g を**頻出パターン**と呼ぶ．

頻出パターンを抽出するには，候補グラフを一つずつ生成し，それが D の各グラフに部分グラフとして含まれるかどうかを判定することで各候補グラフの頻度を求めればよい．最初の候補グラフは単一ノードのみからなるグラフであり，それが頻出がどうかを調べる．もし頻出であればそれにノードやエッジを追加し次の候補グラフを作りまた頻出がどうか調べる．このような処理を繰り返しながら候補グラフを一つずつ生成する．

あるグラフに候補グラフが部分グラフとして含まれるかどうか判定する部分グラフ同型判定問題は NP 完全であるため，候補グラフが多すぎると実用的な時間でそれらの頻度を求めることは難しい．上のような単純方法では同じ候補グラフを重複して生成してしまい，結果として候補グラフの数が非常に多くなってしまうため，候補グラフの生成数を抑える工夫が必要となる．そこで候補グラフを深さ優先探索することで一意に決まる **DFS(Depth First Search) コード**で表現し，無駄な候補グラフを生成しないように工夫されたのが gSpan アルゴリズムである．詳しい説明は省略するが，DFS コードに大小関係を定義することで，一つのグラフに対して最小 DFS コードがただ一つ定められる．候補グラフ生成の際に DFS コードの最小性を判定することで重複する候補グラフの生成を防ぐことができる．なお，今回はこの gSpan アルゴリズムを実現したライブラリ [16] を使用した．

3. 時系列データを用いた Web グラフマイニング手法

本節では時系列データを考慮した Web グラフのマイニング手法を提案する．処理の流れは以下のようになる．

（1） クローラを用いて Web ページ収集を行いグラフ集合を作成する

（2） ページからリンクと時系列データを抽出する

（3） ノードとエッジに時系列データを反映したラベリングを行う

（4） gSpan アルゴリズムを適用し頻出パターンを抽出する

（5） グラフ集合の各グラフからそれらのパターンに対応する部分グラフを列挙する

与える時系列データはページとリンクの生成日時とする．上の流れに従って以下では提案手法の詳細について述べる．

3.1 Web グラフ集合の作成

提案手法の応用例として Web 上で盛り上がった話題を上手く見つけ出すことができると考えているため，今回はブログやニュースサイトとその周辺の構造に焦点を当てる．まず gSpan の入力として与えるグラフ集合作成のために，過去に Web 上で盛り上がった話題をいくつか選び，その各話題ごとに [17] の検索エンジンやソーシャルブックマークなどを用いてページを 5，6 ページ程度収集する．それらを起点としてクローリングを行うことでその周辺の 5,000 ページ程度を取得する．この中にはその話題とまったく無関係なページも数多く含まれているため，今回はその話題について記述されたページを 10–15 ページ程度を人手で選別した．こうして各話題ごとに Web グラフを作成することができる．ここでクローラは estwaver [18] を用いた．

3.2 リンクと時系列データの抽出

収集したページを解析し，時系列データを取得する．ブログやニュースサイトなどの場合は図 3 のように，各エントリごとに記事のタイトル・日付・本文の構成になっていて，それを複数表示している，というのが典型的である．つまり，同じページ内に存在するリンクが同じ日に生成されたとは限らず，記事作成日をチェックし，リンクの生成日時を確認しなければならな

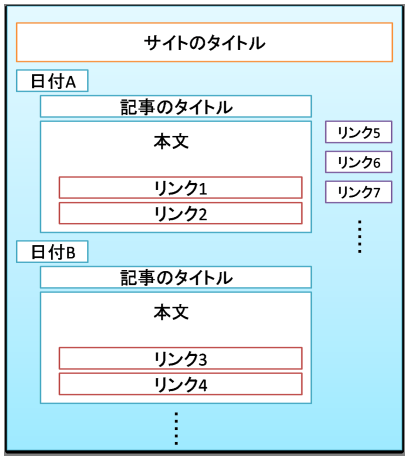

図 3　ブログやニュースサイトの構成

い．記事に関連するリンクは記事の中にあり，このようなリンクは主に他の記事の紹介や批評などそのサイトの運営者によって取り上げられる情報の参照元を表していることが多い．このようなリンクの生成日時は，そのリンクが存在する記事の日付であると考えられる．例えば図 3 のリンク 1 とリンク 2 には日付 A の日に生成されたと見なし，リンク 3 とリンク 4 は日付 B と見なす．またリンク 5，6，7 のような位置には広告や恒久的な相互リンクである場合が多く，今回の目的にはそれほど重要性はないと思われる．

抽出の方法として，まず HTML の div などのノードの class 属性がもし “entrydate”や “date”などの日付を表すと思われる属性値である場合，そのノードが日付を表しているかを正規表現を用いて確認する．そのように明示的に日付を表す属性値が定義されていないサイトの場合もある．その場合でも H タグや B タグなどで日付を表現していることがあるため同様に日付表現がないかチェックを行う．日付は様々な方法で表現が可能である．今回抽出できる代表的な日付表現の例を表 1 に示す．日付表現が発見できたら次はリンクを探し出す．見つけた日付情報を持つノードの兄弟およびその子孫すべてのノードから A タグ href 属性を探し，そのリンクをその日付のものとして抽出する．この方法により図 3 のリンク 5，6，7 のような不要なリンクも取り除くことができる．

日付表現	
Thr, 1 Jan 2009 12:12:30	1 Jan 2009
2009.1.1	2009/1/1
2009-1-1	2009.1.1 (Thursday)
2009.1.1 (Thur)	2009.1.1 (Thr)
2009.1.1 (Thursday) 12:12:30	2009.1.1 12:12
2009.1.1 (Thursday)	2009 年 1 月 1 日 (Thursday)
2009 年 1 月 1 日 (木曜日)	1 月 1 日 (木) 12 時 12 分 30 秒
2009/1/1 (木)	2009 年 1/1
etc...	

表 1　様々な日付表現

3.3　ラベリング

gSpan アルゴリズムを適用するためにはノードとエッジにラベリングをする必要がある．そこでノードに上で取得した時系列データを用いてラベル付けを行う．Web 上で話題が盛り上がった場合でも，それが 1,2 週間程度かけて緩やかに広まっていく場合や，3 日程度で急激に周知され，その後大きな動きもなく収束していく場合など，グラフの成長の仕方もさまざまである．このような複数の時間の変化に対応できるようなラベル付けが必要となる．そこで今回は，各話題ごとに各ページの記事の生成日時がグラフ内でどのくらい新しいものかを判定し，それに応じてラベルを付けることにした．そのために，まずグラフ内のページを日付の古い順にソーティングし，各ページのグラフ全体の中での新しさでランキングを付ける．ページをその順位に応じて以下の四つのグループに分け，ラベルを付ける．具体的には次のようになる．

- 最も新しいページ

最も新しいページはその話題を一番最初に報道したページである．一般的なニュースの場合はそれを最初に報道した新聞社やマスコミのサイトなどのニュース記事や議論の発端となるページである．このようなページを一次情報源と呼び，ノードにはラベル N を振る．

- 上位 30%

一次情報源をすばやく発見し参照したページである．ニュースをいちはやく発見し，自分の記事にするアルファブロガーによるページは主にここに含まれる．また議論の場合は問題となった一次情報源に最初にリンクし，自分の意見を述べているページであることもある．傾向として，最初のページが作成された日と同じ日か翌日程度に作成されたページであることが多い．このようなノードにはラベル N_0 を振る．

- 中位 40%

最初のページから 3–5 日後程度に書かれた記事であることが多い．この時期に話題が広がらない場合はそのまま盛り上がることなく収束することもある．このグループにはラベル N_1 を振る．

- 下位 30%

グラフの中では最も生成日時が遅い部類のページである．これまでその話題を扱った多くのページへのリンクを含んでいることが多く，そのような場合，包括的な意見や総合的なまとめ記事を書いていることが多い．このグループにはラベル N_2 を振る．

ノードへのラベリングの例を表 2 に示す．これは後述する実験で実際に取得したグラフのノードへのラベリング例である (図 4 の G_2 に対応)．“生成日時”の列は各ページが生成された日付を表し，“ラベル”の列は与えられたラベルを表している．

なお，エッジに対しては簡単のため全て単一のラベル E を割り振った．

3.4　頻出パターンの抽出

3.3 で説明したラベリング方法を用いた Web グラフの集合を作成する．この集合を gSpan アルゴリズムへの入力として与えることで頻出パターンの抽出を行う．入力として与える $minSup$ の値にもよるが，gSpan によって列挙された頻出パターンの数は非常に膨大になってしまった．その中から特徴的

ページ	生成日時	ラベル
1	2008/12/21	N
2	2008/12/22	N_0
3	2008/12/22	N_0
4	2008/12/22	N_0
5	2008/12/23	N_1
6	2008/12/23	N_1
7	2008/12/24	N_1
8	2008/12/24	N_1
9	2008/12/24	N_1
10	2008/12/25	N_2
11	2008/12/25	N_2
12	2008/12/28	N_2

表 2　ノードへのラベリング

と思われるパターンの選別を行い，有用な構造だけを取り出す必要がある．今回は人手で特徴的なパターンの判断を行った．

3.5　パターンに対応する部分グラフの列挙

このようにして特徴的なパターンが抽出できたら，新たにそのような構造を Web から探し出すことで有用な情報を発見できるのではないかと考えている．しかし，得られたパターンを Web 上から列挙することはパターンを発見することとは別の大きな課題であるため，本論文ではこの課題については扱わない．

4.　実　　験

4.1　ブログやニュースサイトに関する予備実験

提案手法を用いて実際にブログやニュースサイトの周辺のマイニングを行った．そのようなサイトの管理者は自分でまずニュースを探し，有益であったり興味深い内容であれば情報源となったページにリンクを張った上で紹介や批評を掲載することが多い．そしてそのサイトを参照した別のブログサイトの管理者がさらにリンクを張り話題を伝播させていく．有益なニュースが Web 上で盛り上がる場合，リンクを張られる時間がある短い期間に集中し，時系列も考慮したリンク構造は特徴的なパターンで成長していくのではないかと考えている．パターンが結果として抽出できれば，そのような構造をした部分グラフを Web から実際に探し出すことで新たなニュース発見につながることも期待できる．そこで，実際にそのようなパターンが抽出できないか実験を行った．今回扱ったトピックはマスコミにも報道され Web でも盛り上がった一般的なニュースや，ブログを中心に Web 上で盛り上がった議論など五つの話題を選んだ．その詳細は以下の通り．

（1）　Web の製作費に関する議論

- 期間　2008/2/20 – 2008/3/31
- 一次情報源　http://web-tan.forum.impressrd.jp/e/2008/02/20/2456
- 概要

Web サイト制作を制作会社に委託した場合，どのくらいの相場でどのような作業が行われるかの目安を示したところ，その値段について多くの反響があった．

（2）　勉強論に関する議論

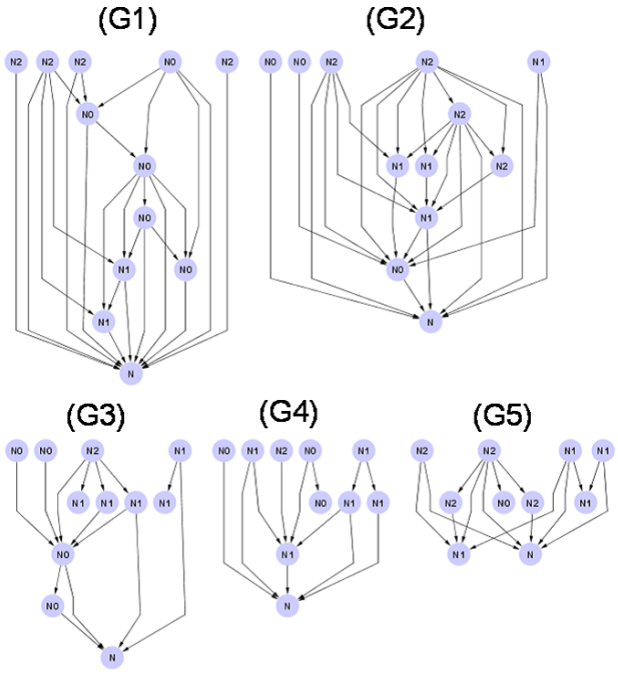

図 4　入力グラフ集合

- 期間　2008/12/21 – 2008/12/28
- 一次情報源　http://anond.hatelabo.jp/20081221200806
- 概要

はてなダイアリー [19] にて “勉強ができる=頭がいい?” という記事が匿名で掲載された．それを見て自分のブログに意見を述べる人が増えてくると，支持される意見，あまり共感されなかった意見など多くのページへの参照を示したうえでさらに自分の意見を述べる，というように議論が活発に行われた．なお表 2 のページセットはこの話題から作成されたグラフのノードに対応している．

（3）　派遣村に関するニュース

- 期間　2009/1/4 – 2009/1/15
- 一次情報源　http://blog.livedoor.jp/dqnplus/archives/1206317.html
- 概要

年末から年始にかけての派遣村に関するニュース．派遣村に関して，マスコミや政府が肯定的な意見なのに対し，Web 上では批判的な意見も目立ったニュース．

（4）　韓国向航海ボイコットに関するニュース

- 期間　2008/12/29 – 2009/1/1
- 一次情報源　http://media.daum.net/
- 概要

韓国のニュースサイトが一次情報源となったニュース．世界の主要船員組合が韓国向け航海をボイコットした事件．日本のマスコミにはあまり報道されていないが Web 上ではこのようなニュースをいちはやく報道するブロガーも存在する．

（5）　ニュースサイトの在り方に関する議論

- 期間　2008/12/14 – 2008/12/20
- 一次情報源　http://d.hatena.ne.jp/BABYPEENATS/

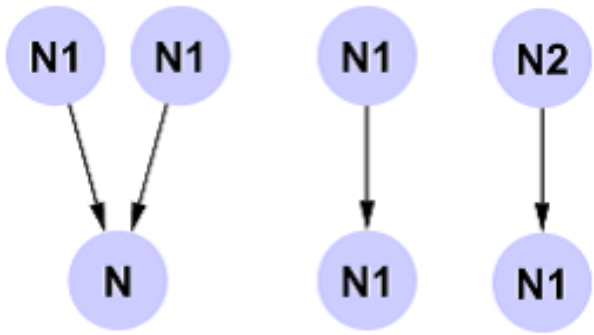

図 5　最小支持度 $minSup = 5$

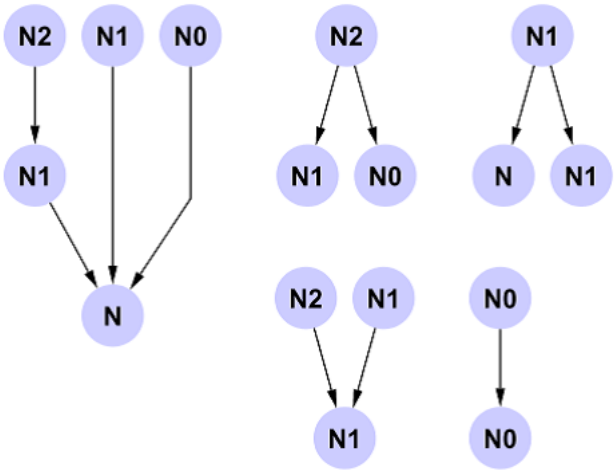

図 6　最小支持度 $minSup = 4$

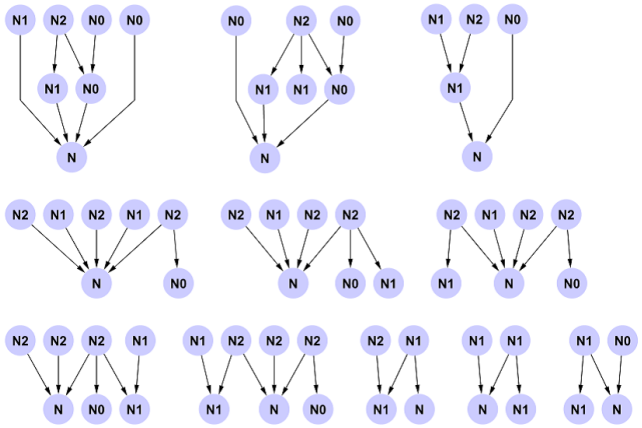

図 7　最小支持度 $minSup = 3$

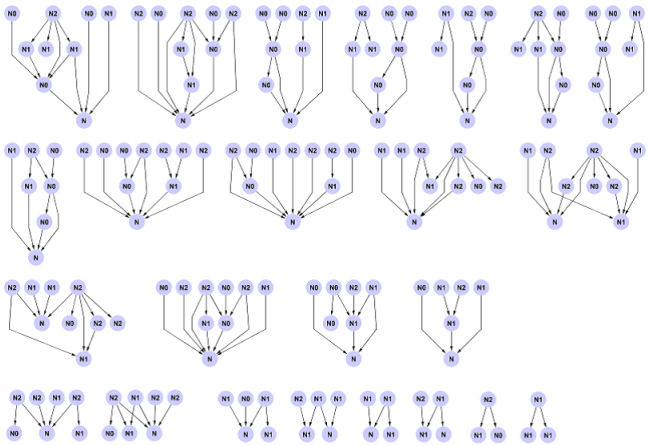

図 8　最小支持度 $minSup = 2$

20081214/1229196334

- 概要

ニュースサイトの管理人が紹介する記事に対してどのような立場であるべきかの議論がニュースサイトの運営者やブロガー中心に行われた．

この話題 (1)–(5) をもとに 3.1 で説明した方法で収集したグラフを図 4 の G_1–G_5 に示す．これらのグラフを入力として gSpan アルゴリズムを適用する．与える最小支持度は $minSup = 5,\ 4,\ 3,\ 2$ の 4 通りで実験を行った．その結果として出力された頻出パターンを以下に示していく．ただし，ここではより大きい最小支持度でも出力されたパターンは省略している (つまり $minSup = n$ で発見されたパターンは $minSup < n$ なる最小支持度でも出力結果としては得られるが，図からは除外した)．

まず，$minSup = 5$ の結果を図 5 に，$minSup = 4$ の結果を図 6 に示す．ここで抽出されたものはグラフの構造としては小さく，一般的な構造のものばかりである．つまりブログやニュースに関係なく，Web グラフ上のあらゆるところでよく見られるものばかりであるため，このようなパターンから得られるものは少ないと考えられる．

続いて $minSup = 3$ の結果を図 7 に，$minSup = 2$ の結果を図 8 に示す．$minSup = 2$ の場合は，抽出されたパターンがかなり入力グラフに依存している構造が多いと思われる．そのためその後パターンを利用して新たな情報の発見に役立つ可能性という意味では大きな期待はできない．一方，$minSup = 3$ は全体の 6 割に含まれる構造であり，そしてそれほど自明でない汎用的な性質を含んだパターンが含まれていると考えられる．そこで，今回の実験結果から得られたパターンとしては主に $minSup = 3$ であるものを中心に考察を行うことにする．

4.2　考　　察

実験の結果得られたパターンがどのような意味を持つかの考察，そして今回適用したラベリング手法に関する考察について述べる．

4.2.1　抽出したパターンに関する考察

$minSup = 3$ のパターンである図 7 から特徴的であると思われるパターンを二つ選び，それを図 9 の p_1，p_2 に示す．この二つのパターンから読み取れる共通する性質として

- 一次情報源となるページ N をすばやく発見しリンクを張るページが，さらに他の複数のページからリンクされる．(例: a，b)
- 時間経過につれ，リンク構造が広がっていく．
- 数日すると，多くのページにリンクを張るハブの役割と

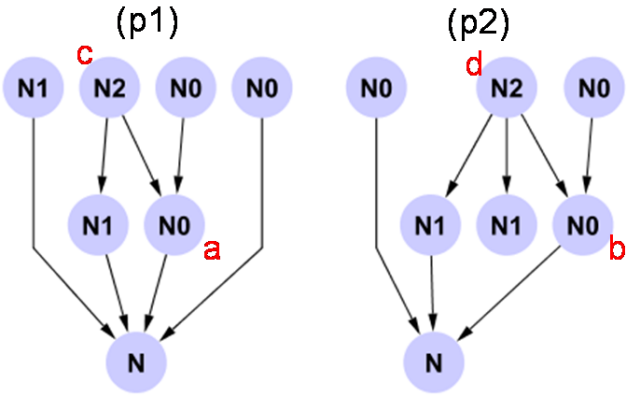

図 9　$minSup = 3$ から抽出したパターン

なるページが現れる．(例: c，d)

これらの性質が実際のグラフではどのようなページに相当しているのか検証を行った．まず a，b のようにすばやくリンクを張るページは，いわゆるアルファブロガーと呼ばれるブロガーのページであることがあった．つまり今回の話題だけでなく，これまでもいろいろな情報をすばやく発見したことのあるサイトのページであるものが含まれていた．また，特定の分野に興味を示し，専門的で独自のニュースを配信するブログの場合もあった．例えば，韓国向航海ボイコットの話題から作成されたグラフにおいてこのような役割を果たすページは常に韓国系のニュースを多く取り上げているブログであった．

c，d のようなページは盛り上がる話題の種類によって性質が異なる．議論の場合は，時間がたつにつれ，多くの人が自分の意見を自分のブログなどに書くようになる．それらの多くの意見を参照しながら自論を展開する場合が多かった．ニュースの場合も，参照しただけでなく管理人の意見などが書き加えられることがあるため多くの記事を参照するページも現れる．それ以外にもニュースの場合はページごとに取り上げている事実が部分的であったり細部が異なったりすることもあるので，補完する役目のページとしてソーシャルブックマークなどがこのノードに相当することもあった．

4.2.2　ラベリングに関する考察

3.3 で説明したラベリングの手法はノードに日付情報を与え，エッジには単一のラベルを振るというものであったが，ラベリングの手法はこの他にもいろいろ考えられる．例えば全ノードに単一のラベル N を与え，エッジに日付に応じたラベル E_0，E_1，E_2 を振る方法は 3.3 で説明した方法のノードとエッジの関係を逆にしたものである．この方法を使う利点は，ページ内に異なる日付の記事があり，それぞれが別のページを参照していた場合に有効であると考えられる．また Web 上での議論は，あるブログ a が何か記事を書くと，そのページにブログ b がリンクを張って意見を述べ，さらにブログ a がブログ b にリンクを張って反論するということも起こりうる．このような場合，記事が同一のページに含まれていたとすると，それぞれのリンクに時系列データを持たせる方が有効である．しかし，gSpan アルゴリズムはエッジラベル数を増やすよりノードラベル数を増やす方がより高速にパターン抽出が行える．実際，ノード数が大きく，ノードラベル数が少ないと gSpan では結果が実用的な時間で結果が得られないことも多くなるという問題が生じた．

ラベリングの方法はこれ以外にもいろいろと試してはいるが，意味のあるパターンがとれないラベリングもあった．3.3 の手法はその中では比較的計算時間が早く結果が得られたものである．しかし，ページ数が多くなるとノードラベルが 4 種類程度では実用的な時間で結果が得られないと思われる．大規模なサイズのグラフにも対応でき，かつ意味のあるパターンが抽出できるようなラベリングの方法も検討する必要がある．

5. 課　　題

本論文では時系列データを利用した Web グラフマイニングの手法を提案し，予備実験としてブログやニュースサイトを中心とした Web 上で盛り上がった話題の周辺のパターン抽出を行った．さらに得られたパターンの中から特徴的と思われるものを二つ選別し，それらの考察を行った．実験を通していくつかの課題も得られたので以下に示す．

（1）　グラフ集合作成の自動化

今回行った実験手法は，クローラによって収集したページの集合の中からその話題について書かれたかどうかを人手で判定した上でグラフ集合を作成している．今後は入力グラフの作成も自動化した上でのデータセットに対する実験も行いたいと考えている．

（2）　大規模グラフへの適用

現在行っていることとして，4.1 のような小規模なグラフではなく，クローラによって収集された大量のページに対してグラフを作成し，頻出パターンを発見できないか，という実験を行っている．ノードを取得したページ単位ではなくドメイン単位とすることでリンクやページの集約を行っているが，それでも入力として与える一つのグラフのノード数が 5,000–10,000 程度と非常に大きいものである．このようなグラフに対してはノードのラベル数をより増やしたり，何らかの方法で不要なノードを判別しグラフから取り除くなどの工夫を行う必要がある．なぜなら，gSpan アルゴリズムを適用している段階で候補グラフの数が爆発的に増えてしまい，頻出パターンが得られないためである．例えば，ノード数 300 程度のグラフ 10 個からなる入力を与えると，10GB のヒープを食いつくしたり，Java の int の最大値を超える配列を必要としたりして，結果を得ることはできなかった．そのため，このような実験を行うためにはクローラによるページ収集方法やグラフ作成方法，ラベルの与え方を工夫する必要がある．このように，より大規模なグラフに対してマイニングが行えるような手法を構築することは今後の重要な課題である．

（3）　得られた頻出パターンのフィルタリング

今回の実験ではパターン抽出の結果から特徴的なものとして二つのグラフを選別した．これらは 4.2.1 で述べたような共通する特徴を持つと思われるものを人手で選んでいる．このように抽出されたパターンから意味のあるものとないものを自動的に区別するための方法が必要になる．そのための手段として考え

られるものとして以下のようなものがある．

- グラフパターンをまとめる方法を構築する．グラフ間に類似度を定義し，抽出された大量の頻出パターンを分類する．
- 性質の異なる入力グラフ集合を用意し，それぞれの頻出パターンを抽出する．これらの全てのセットから得られたようなパターンは特徴的でないと見なし除外する．

（4） 得られたパターンの利用方法

今回抽出したパターンを Web 上の中から効率よく列挙する手法が確立できたならば，Web 上での盛り上がっている話題を発見するのに利用できると考えている．またそうしたことができるようになったならば，今回見つけたパターンとその意味づけに対する評価も可能である．しかし実際に Web 上からこれらのパターンに対応した部分グラフを発見することはまた別の大きな問題である．

（5） 追加実験

まず，今回は一つの話題セットから実験を行ったが，他にも多くの話題を用いて類似のパターンが取得できるか確認する必要がある．さらに今回は時系列データを加えた Web グラフを実験に用いたが，提案した手法は Web グラフに限らず時系列データとグラフ構造で表現できるような様々なデータに応用できると考えている．Web グラフ以外の時系列データを含んだ実例を用いてさらなる実験を行うことで，今回見つけられなかった他のパターンの発見も期待できる．

6. 結　　論

今回用いた手法は既存の頻出パターン抽出アルゴリズムを応用し時系列データをグラフのノードのラベルとして組み込んだものである．本研究により時間情報を単純にラベリングに利用しパターンを抽出しただけでも有用な情報が抽出できることが分かった．時間間隔や Web 構造の変化をより柔軟に利用するにはラベリングの工夫だけでは足りないこともあると思われる．今後は gSpan を用いる以外にも独自の方法でグラフを扱い時系列データのさらなる活用法も考えていく予定である．

文　　献

[1] S. Kramer, L.D. Raedt, and C. Helma, “Molecular feature mining in hiv data,” KDD ’01: Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.136–143, New York, NY, USA, 2001, ACM.

[2] T. Coffman, and S. Marcus, “Pattern classification in social network analysis: a case study,” Proceedings of 2004 IEEE Aerospace Conference, vol.5, pp.3162-3175, 2004.

[3] F. Lin, S. Chou, S. Pan, and Y. Chen, “Mining time dependency patterns in clinical pathways,” International Journal of Medical Informatics, vol.62, no.1, pp.11–25, 2001.

[4] J.M. Kleinberg, “Authoritative sources in a hyperlinked environment,” J. ACM, vol.46, no.5, pp.604–632, 1999.

[5] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, “Trawling the web for emerging cyber-communities,” Computer Networks, pp.1481–1493, 1999.

[6] 宇野毅明，“大規模グラフに対する高速クリーク列挙アルゴリズム，” 電子情報通信学会技術研究報告，vol.103，no.31，pp.55-62，2003.

[7] 中島伸介，舘村純一，原良憲，田中克己，植村俊亮，“重要な blogger 発見を目的とした blog スレッド解析手法，” 知能と情報 (日本知能情報ファジィ学会誌)：Journal of Japan Society for Fuzzy Theory and Intelligent Informatics，vol.19，no.2，pp.156-166，2007.

[8] 毛受崇，吉川正俊，“ブックマークの時系列情報を利用したソーシャルブックマークにおける注目度予測，” DEWS B9-5，2008.

[9] 奥村学，南野朋之，藤木稔明，鈴木泰裕，“blog ページの自動収集と監視に基づくテキストマイニング，” 第 6 回人工知能学会セマンティック Web とオントロジー研究会：SIG-SWO-A401-01，2004.

[10] kizasi.jp. http://kizasi.jp/.

[11] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, “On the Bursty Evolution of Blogspace,” World Wide Web: Internet and Web Information Systems, vol.8, no.2, pp.159–178, 2005.

[12] A. Inokuchi, T. Washio, and H. Motoda, “An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data,” PKDD ’00: Proceedings of 2000 European Symposium Principle of Data Mining and Knowledge Discovery, pp.13–23, 2000.

[13] M. Kuramochi, and G. Karypis, “Frequent subgraph discovery,” ICDM ’01: Proceedings of the 2001 IEEE International Conference on Data Mining, pp.313–320, 2001.

[14] X. Yan, and J. Han, “gSpan: Graph-based substructure pattern mining,” ICDM ’02: Proceedings of the 2002 IEEE International Conference on Data Mining, pp.721–724, IEEE Computer Society, 2002.

[15] X. Yan, and J. Han, “Closegraph: mining closed frequent graph patterns,” KDD ’03: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.286–295, ACM, 2003.

[16] ParSeMiS. http://www2.informatik.uni-erlangen.de/.

[17] Google ブログ検索 http://blogsearch.google.co.jp/.

[18] 全文検索エンジン HyperEstraier http://hyperestraier.sourceforge.net/index.ja.html.

[19] はてなダイアリー http://d.hatena.ne.jp/.

[20] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graphs over time: densification laws, shrinking diameters and possible explanations,” KDD ’05: Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp.177–187, ACM, 2005.

[21] J. Han, and M. Kamber, Data Mining: Concepts and Techniques (Morgan Kaufmann Series in Data Management Systems), Morgan Kaufmann Pub, 2nd edition, 2006.

[22] D.J. Cook, and L.B. Holder, eds., Mining Graph Data, Wiley-Interscience, 2005.