DEIM Forum 2013 E4-5

# グラフマイニングによるソーシャルグラフの成長分析

山段　裕貴† 　浅野　泰仁† 　吉川　正俊†

† 京都大学大学院情報学研究科　〒 606–8501 京都市左京区吉田本町 36-1
E-mail: †sandan@db.soc.i.kyoto-u.ac.jp, ††{asano,yoshikawa}@i.kyoto-u.ac.jp

**あらまし**　近年のネットワーク分析において，グラフ構造で表せるネットワークの，将来的な成長を予測するリンク予測の研究が盛んに行われている．また，グラフマイニングはグラフの特徴的な構造を用いて有益な情報を抽出するもので，こちらの研究も注目を浴びている．本研究では，情報解析の分野において時系列情報を用いたグラフ構造に着目する．グラフ構造のうち，特にソーシャルグラフに関して，時間的特徴を考慮したグラフマイニングにより得られるパターンからどのような有用な知識が得られるか，またグラフの成長分析にどう活かせるか．グラフマイニングとリンク予測の方法を組み合わせ，グラフ構造からなるネットワークの新たな成長予測の手法を提案する．

**キーワード**　リンク予測 グラフマイニング ソーシャルグラフ 時系列情報

Hiroki SANDAN†, Yasuhito ASANO†, and Masatoshi YOSHIKAWA†

† Graduate School of Informatics, Kyoto University
Yoshida-honmachi 36-1, Sakyo-ku, Kyoto 606–8501 Japan
E-mail: †sandan@db.soc.i.kyoto-u.ac.jp, ††{asano,yoshikawa}@i.kyoto-u.ac.jp

## 1. はじめに

近年のネットワーク分析において，将来的なネットワークの成長を予測する，リンク予測の研究が盛んに行われている．リンク予測ではネットワークをグラフ構造と捉える．グラフ構造とは，ノード (節点) 群とノード間のつながりを表すリンク (エッジ，枝) 群からなるデータ構造である．リンク予測とは，このグラフ構造に対して，ある 2 ノード間にリンクがあるかどうかを予測するものである．

複雑ネットワークにおけるリンク予測の研究は様々な目的で行われている．企業間における新たな協業の発見，人間関係における友好関係の発見，将来購入する商品を推薦するリコメンデーションシステムへの応用，遺伝子やたんぱく質の相互作用ネットワークから新たな発見を予測するなどの生物学的な応用，さらには情報検索，Web ハイパーリンクの自動生成などへの応用などに利用することが考えられる [1] [2] [3]．リンク予測問題は大きく二つに大別される．一つは，静的なネットワーク構造に着目して，既に観測されたリンク構造から未観測のリンク構造を予測すること．もう一つは，動的な成長するネットワーク構造に着目して，現在のリンクから将来のいリンク構造を予測することに分類される．

一方 Web ネットワークにおけるグラフ構造の解析では，Web ページを節点，リンクを枝ととらえ，グラフ構造として扱う方法がある．Web ページやリンクが新しく生成したり消滅したりと，Web の構造は絶えず変化をする動的なグラフネットワークである．Web ネットワークのグラフ構造の大きさは大規模であり，成長を続けるネットワークなため，構造的特徴が明らかにされていない未知なる有用な部分ネットワークが数多く存在すると考えられる．

グラフ構造の研究では，グラフパターンを用いるという手法がとられてきた．グラフパターンとは，特別な部分構造を持つグラフのことである．これをマッチングや列挙に用いることで，ある特徴を持った部分構造を抽出することができる．そして抽出したグラフパターンを元に，グラフの構造解析を行ったり，有用な情報，未知なる知識の発見が実現できる．グラフパターンの解析を利用可能な具体例としては，Web ネットワーク以外にも言語理論，ベイジアンネット，ニューラルネットワーク，更には生物学や化学分野など，非常に多岐にわたる [4]．このようにグラフパターンに着目する手法は, 様々な分野に応用されている．このグラフパターンを抽出する手法をグラフマイニングと呼ぶ．

既存研究には静的なネットワークのグラフパターンを抽出するものが多い．そこから一歩踏み込んで，Web などの動的なネットワークに対して，時系列を考慮したグラフパターンを用いる研究も登場している [5]．例えば Web だと，Web ネットワークの構造的特徴と時間的特徴を同時に解析し，Web 解析や情報検索への応用を目的としている．時間的特徴とは，例えば

Web のページやリンクの生成日時を考慮したものであり，それらの時間情報をグラフの節点や枝のラベルに用いることにより解析に反映させている．

本研究では，情報検索の分野において，時系列情報を用いたグラフ構造に着目する．グラフ構造のうち，特にソーシャルグラフに関して，時間的特徴を考慮したグラフマイニングにより得られるパターンからどのような有用な知識が得られるか，またグラフの成長分析にどう活かせるか．将来のリンク予測に関する手法を考える．

表 1　ネットワーク構造データの例

| ネットワーク | ノード | リンク |
|---|---|---|
| WWW | Web ページ | ハイパーリンク |
| 社会ネットワーク | 人物 | 人物関係 |
| 文献ネットワーク | 論文 | 引用関係 |
| 生体ネットワーク | 遺伝子<br>タンパク質 | 制御関係<br>相互関係 |

## 2. リンク予測手法に関して

成長する社会ネットワークのリンク予測問題に関して，Liben-Nowell と Kleinberg によるノードの”proximity”に基づいたリンク予測法が知られている [3] [6]．現在リンクで結ばれていないノードペアに関して，例えばその二つのノードが共通のノードにリンクを張っているなど，一方のノードから他方のノードに情報が伝わりやすいような状況であれば，その二つのノード間には新たなリンクが張られやすいといった予測が立てられる．

以下に様々な研究がなされている主なリンク予測問題の具体例，リンク予測の既存手法について述べる．

### 2.1 実　用　例

**2.1.1**　人間関係の友好関係の発見

所属するコミュニティ予測．コミュニティ情報８つ，ノード (個人) 情報６つの属性を生成し，リンクに基づく分類．

結果，ユーザがあるコミュニティ属する確率は，そのコミュニティ内に友人が多いほど高くなる傾向が見られた．さらにコミュニティ内にいる友人が互いに知り合いであるほうが，ユーザはそのコミュニティに所属しやすい傾向が見られた (Backstrom ら)．

**2.1.2**　ソーシャルネットワーク

SNS におけるユーザ推薦は社会ネットワークにおけるリンク予測問題として捉えられる．SNS などによる社会ネットワークは，個人や企業などの社会的主体と, 友人関係などのそれらの間の関係によって表現されている．そしてその構造は一般的にネットワーク (グラフ) 構造によって表現が可能である．

現在の人間関係を表した社会ネットワークが与えられたとき，今後この人間関係がどのように変化していくかを予測する問題を考えると，予測をもとにして，SNS などにおける推薦機能を実現できる可能性がある．

この問題は「現在のネットワーク構造が与えられたとき，将来のネットワーク構造を予測する問題」と捉えることができる．場合によっては，これまでどのような過程を経て構造が変化したかという，ネットワーク構造の遷移の履歴がデータとして利用で切るような問題設定も考えられる．

**2.1.3**　企業間における新たな協業関係の発見

友人関係などの比較的静的な関係だけでなく，e メールの送信や，共同作業などの動的な環境をリンクとして表すこともありうる．このような企業内や企業間での仕事上のつながりに着目することにより，将来的に有用な協業関係を見出すヒントとしてリンク予測を利用できる．

**2.1.4**　バイオインフォマティクスへの応用

バイオインフォマティクスとは，生物学の解決すべき課題を情報学の技術により問題解決を試みる学問分野である．実験的に相互作用があることが分かっているタンパク質のペアの間にリンクを張ることで構成されたネットワークの一部分を手がかりに，まだ知られていない相互作用を予測するという「ネットワークの補完・外挿問題」と捉えることができる．

この場合，新たに見込みがあると予測された相互作用の候補を，実際に実験的に確認してみることで，未知の相互作用を効率よく発見することが期待できる．

### 2.2 リンク予測の指標

**2.2.1**　グラフの局所的特徴に着目

(文献 [3])

- Common neighbors

$$common := |\Gamma(i)| \cap |\Gamma(j)| \tag{1}$$

ノード $i$ とノード $j$ が共通の隣接ノードを多く持っているほど２つのノードの間にはリンクが現れやすいとする指標.

- Jaccard 係数

$$Jaccard's := \frac{|\Gamma(i)| \cap |\Gamma(j)|}{|\Gamma(i)| \cup |\Gamma(j)|} \tag{2}$$

common neighbors を正規化したもの．情報検索において類似度として用いられる．少数の隣接ノードをもつノードのリンクほど重宝される．

- Adamic/Adar

$$Adamic/Adar := \sum_{k \in |\Gamma(i)| \cap |\Gamma(j)|} \frac{1}{\log |\Gamma(k)|} \tag{3}$$

同じく正規化された common neighbors 指標．隣接ノードごとに異なった重みを割り当てている．少数の隣接ノードをもつノードに大きな重みが割り当てられる．

- $\mathrm{Katz}_\beta$

$$Katz_\beta := \sum_{l=1}^{\infty} \beta^l |paths_{i,j}^{(l)}| \tag{4}$$

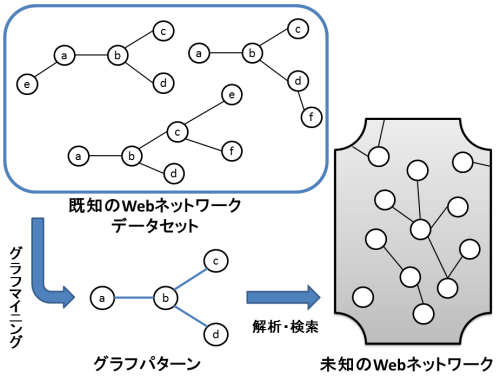


図 1　Web ネットワークのマイニングと解析

common neighbors 指標の一般化．より遠くの関係を考慮する．paths はノード $i$ からノード $j$ への長さ $l$ のパスの数．

$\mathrm{paths}^{\langle l \rangle}_{x,y}$ := ノード $x$ からノード $y$ への長さ $l$ のパス集合

- Preferential attachment

$$preferential := |\Gamma(i)| \cdot |\Gamma(j)| \tag{5}$$

common neighbors とは異なり，隣接ノードが多いノードほど新たなリンクを得やすいという考えに基づいたもの．

#### 2.2.2　グラフ全体を考慮

- RWR によるリンク予測

あるノードを起点に，グラフ上をランダムウォークする．このとき各ノードへの遷移確率は同じとする．一定の確率で起点のノードに戻りランダムウォークを再会する．最終的に各ノードに滞在している確率を類似度とする．

- $\mathrm{SimRank}_\gamma$

$$\begin{cases} 1 & \text{if } x = y \\ \gamma \cdot \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} \mathrm{score}(a,b)}{|\Gamma(x)| \cdot |\Gamma(y)|} & \text{otherwise} \end{cases}$$

SimRank は，”類似するノード同士は類似するノードに関連付けられる” という直感を基に，任意の 2 つのノードに対して共引用 (共参照) を反復的に計算し類似度を求め，類似度を伝搬させるというもの [7].

## 3.　グラフパターンマイニングに関して

### 3.1　グラフパターンによる Web 解析

特別な部分構造を持つグラフパターンを用いた web の解析や情報検索の研究は様々存在する．解析を行う Web グラフのデータセットを用意し，データセットに対してグラフパターンによるマッチングを行う．そして各々のパターンの構造とマッチする全ての部分グラフを求める．得られた部分グラフに対応する Web ページ群を解析することにより特徴を見出す．特徴をもとに有用な分類が可能であれば，さらなる解析や情報検索に応用することができる．

### 3.2　頻出グラフパターンマイニング

膨大なグラフ構造データから，有用な知識を効率良く抽出するためのグラフマイニングアルゴリズムがある．代表的なものに，gSpan があげられる [10]．gSpan は頻出パターン列挙アルゴリズムとして知られる．マイニングのグラフ集合において，出現する頻度が高い部分グラフ構造を，特徴的で有用なグラフパターンと判断し，抽出する．

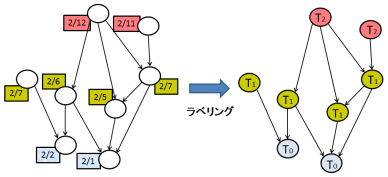


図 2　時間情報をもとにラベリング：時間グラフパターン

グラフ集合内の部分グラフの全てを候補として，各々の部分グラフに対して網羅的に出現頻度を調べることは, 非常に計算時間が大きくなってしまう．gSpan では，グラフパターンの候補になり得ない無駄な部分グラフを生成しない工夫や，候補となる部分グラフからグラフパターンを見つけるときの不要な探索を取り除く工夫がされている．gSpan は，グラフ集合と最小サポート $minSup$ を入力に与える．そしてグラフ集合中の $minSup$ 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見つける．

gSpan の他にも様々なパターン抽出アルゴリズムの研究が取り組まれている．部分グラフの重要度 $p$-$value$ を定義して，重要度の高いパターンを抽出する GraphSig がある [11]．ランダムに生成させて頻繁に出現する部分グラフは重要度が低く，ランダムグラフでは得られにくい部分グラフは重要度が高いと，$p$-$value$ の閾値を設定してパターン抽出をするというものである．

### 3.3　時間ラベルを用いた Web グラフマイニング

グラフパターンを抽出するためには，まずマイニングを行う Web ネットワークのデータセットを作成する．このときデータセット中の全てのグラフの節点と枝にその生成日時の情報をラベルとして付与する．つまり Web ページとリンクの生成日時を追加する．この時間ラベルを各々の節点と枝の特徴とし，グラフマイニングを行う．これにより得られたグラフパターンを時間グラフパターンとして, Web ネットワークの構造的，時間的特徴を同時に解析していく手法が本研究の焦点である．

### 3.4　時間グラフパターン

Web ページやリンクの生成時間を考慮することにより，ページ，リンク構造，つまり部分グラフの成長の時間的特徴を捉えることが可能となる．ページとリンクの構造的特徴だけでなく，成長の特徴も含めてパターンとして見出し，その時間グラフパターンがどのような意味を持つかを考える．

あるトピックについて，Web 上で盛んに情報交換が行われる場合，グラフ構造の規模や成長速度は大きい．特に話題性が大きく，流行のものであれば，成長速度がより大きくなる．一方で，広く知られているが大きく話題として取り上げられないようなトピックでは，グラフ構造の時間的変化が現れにくいもの

となる．

グラフ構造の時間的変化が大きいトピック同士でも，その変化の特徴には様々な違いがある．その中で類似した変化の特徴を持つトピックが多く存在すれば，頻出パターンマイニングにより，時間グラフパターンとして抽出される．

### 3.5 時間ラベル付きグラフ集合の作成

グラフに対するラベリングの工夫次第により，抽出される時間グラフパターンの形も様々である．例えば取得した Web ページ，リンクの時間情報をもとに分類を行い，いくつかの期間に分割する．分割された期間の数だけラベルの種類を用意し，グラフの各節点，枝にラベリングを行うなどの方法があげられる．ここではその一例を用いる．

過去の一定期間で話題性が大きかったトピックについて，既存の検索エンジンにより検索する．そして検索結果の上位から一定数の Web ページの URL とそのページの日付情報を取得する．さらに取得した Web ページ集合のリンク構造を解析することにより，時間情報をもつ Web グラフを得られる．

得られた日付情報を元に，期間をいくつかに切り分ける．例えばそのトピックのある期間で話題性が大きく，関連ページ，リングが急激に増加したとする．その期間を成長期とし，成長期以前の期間を初期，成長期以後の期間を安定期とするなどで期間をわける．この場合だと，3つの期間にわけたのであるから，3種類の時間ラベルを用いて全てのグラフの節点 (Web ページ) にラベリングを行う (図 2).

### 3.6 時間グラフパターンによる解析

得られた時間グラフパターンをもとに解析を行う．まず解析対象の Web ネットワークのデータセットを作成する．様々なニュースを扱うブログや, 話題のまとめサイトなど，情報の集まるサイトのいくつかを基準にして，それぞれからリンク，被リンクを 1 つずつたどり，得られたページをデータセットとする方法が考えられる．

次に作成した解析対象となるデータセットの Web グラフの中から，各々の時間グラフパターンの構造とマッチングを行い，それぞれのパターンの構造とマッチする全部分グラフを求める．

こうして得られた Web ネットワークの部分グラフのデータに対して，サポートベクターや主成分分析などを用いて分類・解析を行う．

### 3.7 グラフマイニングアルゴリズム：gSpan

(文献 [10]) 本研究では,gSpan と呼ばれるグラフマイニングアルゴリズムを使用する．ここではそのアルゴリズムの一部を記述している (Algorithm 1).

### 3.8 サブグラフマイニング

グラフ集合内における同一の部分グラフ構造の出現数を調べることにより，頻出な部分グラフ構造を抽出することができる．その方法を実現したものが, このサブグラフマイニングである．これは gSpan アルゴリズムの中でも用いられている．

**Algorithm 1** GraphSetProjection($\mathbb{GS}, \mathbb{S}$).

グラフ集合 $\mathbb{GS}$ 中のノードとエッジのラベルを出現頻度順にソート;
出現頻度の低いノードとエッジを除去;
残ったノードとエッジをラベリングし直し, 頻度で降順に並べる;
$\mathbb{S}^1 \leftarrow \mathbb{GS}$; 中のエッジが一つのすべての頻出グラフ;
$\mathbb{S}^1$ を DFS コードに関して辞書順に並べる;
$\mathbb{S} \leftarrow \mathbb{S}^1$;
**for each** edge $e \in \mathbb{S}^1$ **do**
  initialize $s$ with $e$, set $s.GS = \{g | \forall g \in \mathbb{GS}, e \in E(g)\}$; (only graph ID is recorded)
  Subgraph Mining($\mathbb{GS}, \mathbb{S}, s$);
  $\mathbb{GS} \leftarrow \mathbb{GS} - e$;
  **if** $|\mathbb{GS}|$ **then**
    break;
  **end if**
**end for**

**Algorithm 2** SubGraphMining($\mathbb{GS}, \mathbb{S}, s$).

**if** $s \neq min(s)$ **then**
  return;
**end if**
$\mathbb{S} \leftarrow \mathbb{S} \cup \{s\}$;
$s$ に一本のエッジを加えることで得られる全ての $s$ の子を生成する;
Enumerate($s$)
**for each**$c$, $c$ is $s$' child **do**
  **if** $support(c) \geqq minSup$ **then**
    $s \leftarrow c$;
    Subgraph Mining($\mathbb{GS}, \mathbb{S}, s$);
  **end if**
**end for**

サブグラフマイニングは，1本の枝によるグラフ $s$ を成長させ，全ての頻出な $s$ の子を見つけるものである．ここで $s$ は DFS コード，または DFS コード木の節点を表す．このサブグラフマイニングを再帰的に行う (Algorithm 2)．

また DFS コードとは，部分グラフから木構造を作成し，深さ優先探索 (Depth-First Search) をすることによって得られる，部分グラフを表現するコードである．

DFS コードは，ノードとエッジのラベルと，探索順を示すラベルで表現される．そしてグラフのすべてのエッジをそのラベルで表現するものである．DFS によるグラフ探索を行うと，始点ノードの選び方やエッジのたどり方により，複数とおりの DFS コードが生成される．それらの DFS コードは同じグラフを示すため，これら全ての DFS コードに対してサブグラフマイニングを行うのは冗長である．そこで一つのグラフに対応する DFS コードをただ一つに絞り，探索の枝刈りを行うことが求められる．

そこで同一グラフ $G$ を表現する複数通りの DFS コードのラベルに関して，辞書順にソートしたもののうち，先頭にくるコードのみを利用する．このコードを Minimun DFS Code と定義し，$min(G)$ と表す．これにより，グラフとコードを一意に対応づけられる．

## 4. ソーシャルグラフの分析

表 2　主な SNS が所有する個人データの特徴 (2009 年) [13]

| 組織名称 | ユーザ ID | 属性情報 | コンテンツ | 行動情報 | 人間関係情報 |
|---|---|---|---|---|---|
| Facebook | ◎ | ◎ | ◎ | △ | ◎ |
| Twitter | ○ | △ | ◎ | △ | ○ |
| Google | ○ | △ | △ | ◎ | × |
| GREE | ◎ | ○ | ◎ | △ | ○ |
| Linkedin | ◎ | ◎ | △ | △ | ◎ |

### 4.1 SNS

SNS(Social Networking Service) は学術的にも注目を集めており，SNS に関する研究は広く行われている．例えば，Mislove らは Flickr,Youtube,LiveJournal,Orkut それぞれのネットワークを比較分析している．また大規模な SNS 上でのユーザのクリックデータを分析することで，各サービスにおけるユーザ行動の違いを示した．Viswanath らは Facebook においてユーザのもつ友人関係数とそのユーザのサービス上におけるアクティビティの関係性を分析した．日本においては，松尾らがコミュニティと呼ばれるグループ機能がネットワークに及ぼす影響について論じている．また So-netSNS を用いた研究として鳥海らは小規模な SNS の比較分析を行い SNS ごとにユーザの行動に差があることを示した．山本らはコミュニケーション構造の時系列変化の分析を行っている [12]．

表 2 において，代表的な SNS が持つ個人データの特徴について記している [13]．各サービスによって，その形態は様々であり，SNS サービスごとに得られる情報が共通する，または異なるものがあることが予想される．

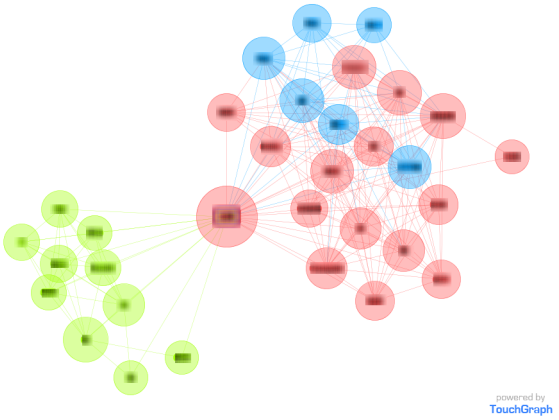


図 3　例：ソーシャルグラフ

### 4.2 Facebook

Facebook はソーシャルネットワーキングサービス (SNS) の一種である．Facebook の特徴は，SNS の中でも，実名で現実の知り合いとインターネット上でつながり，交流を行うところにある．現在世界最大の SNS である Facebook のアクティブユーザ数は，全世界で 8 億人，日本国内においても 1000 万人に達したと発表されている．

こうした SNS の広がりにともない，SNS 機能をベースにした様々なサービスが生まれている．中でもゲームとの組み合わせのサービスを提供する Cithville,Mobage などが注目を集めている他，ビジネスのつながりに特化した linkedin, 動画を介したコミュニケーションを図る Youtube, ニコニコ動画，音楽を中心とした Last.fm など，目的に応じて多種多様な SNS が運営されている [12]．

### 4.3 Twitter

限られた字数制限の中で，ユーザが言い切り型 (つぶやき型) の内容の投稿を主に行う．投稿する内容を広く共有し，ユーザ同士の関係の数が非常に多い．ユーザ同士の関係は，各ユーザの「つぶやき」をフォローという形式により生まれる．自分がフォローしたお気に入りの人物の発言がすべて転送される．これらの特徴が Twitter にはあげられる．

ある人をフォローする人の数が数十万，数百万といった非常に多くのつながりをもつ人気の高い有名人なども数多く利用しており，ブーム的な広がりを見せている．匿名での利用者が多いが，有名人などはあえて実名で登録している．(図 3)

## 5. 提案手法

### 5.1 グラフマイニングによってグラフの成長パターンを発見

まずデータセットから複数のグラフをつくり, グラフ集合を作成する．ここで扱うデータセットは，複数のスナップショットを保有しているもの，すなわち複数の時期のデータを持つものを扱う．これは過去のグラフから予測されたグラフと，それに対応する未来 (現在) の正解グラフを比較するために必要であ

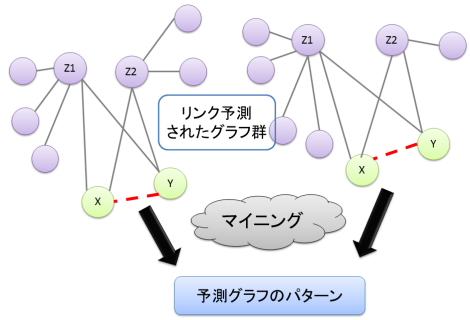


図 4　予測グラフに対するマイニング

る．Facebook や Twitter などの SNS の場合，リンクはユーザのつながり，ノードはユーザ ID となるグラフを作成する．このグラフをクラスタリングでコミュニティごとに分割するなどして複数のグラフ集合を作成し，それぞれのグラフがどのような成長をするかの予測を行う．

グラフパターンマイニングでは，グラフの特徴的なパターンを抽出することができる．ある時点でのグラフのノードについて，ノードが生まれた時間情報をもとにノードの種類を分ける．そしてラベル付けして，頻出パターンマイニングをすることにより，グラフの成長パターンを知ることができる．

この得られた成長パターンを，既存のリンク予測の指標 (common neighboes や,Jaccard's 係数など) と組み合わせることにより，既存手法では得られなかった予測グラフを得ることを考える．

### 5.2　既存のリンク予測手法により得られたグラフの成長パターンを発見

また既存のリンク予測手法を用いて得られた各々の予測グラフと，実際の成長後のグラフである正解グラフと比較し，どのような予測が成功, 失敗したかを調べる．ここでは予測グラフと正解グラフの違いを比較するために，２つのグラフの差分をとる．つまり正解グラフのように予測できなかった部分を，部分グラフとして取り出し，検証を試みるというものである．そしてそれらの部分グラフ集合に対してグラフパターンマイニングを行うことにより，予測手法がどのような事柄に関する予測に強みをもつのかを検証することも考えられる．既存の予測手法には複数の方法があり，それぞれがどのような予測の傾向があるのかを分析することが一つの目標として考えられる (図 4)．

### 5.3　グラフデータの作成方法

グラフパターンマイニングを行うためには，複数のグラフを集めたグラフ集合が必要となる．ある時点のグラフ情報から将来的なリンクの予測に関して，グラフパターンマイニングを利用する場合，例えばそのグラフ情報が大規模なものであれば，それらを分割してグラフ集合にすることが考えられる．このとき分割方法の選び方によって，得られるグラフの成長パターンも異なってくるため，分割方法について考える必要がある．

ソーシャルネットワークでは共通の知り合いの集合によって形成される「コミュニティ(クラスタ)」が形成されていることが多く，このコミュニティに着目してクラスタリングを行う方法がグラフにおいて存在する．

この方法でグラフを分割してグラフパターンマイニングを行う場合，ソーシャルグラフにおけるコミュニティの成長の特徴を捉えることができると予想される．ただこの分割の仕方だけでは, 他の成長の特徴，例えばコミュニティ間のリンクの成長などを調べることに結びつきにくいと考えられる．ランダムに起点となるノードを選び，グラフ上でそのノードからの一定距離を含む部分グラフを作成することや，グラフにする前のデータに対してクラスタリングを行うなど，別の視点で分割を行う方法を用いることも考える必要がある．

### 5.4　実験で用いるデータセット

時系列情報が存在し，かつ膨大な情報量のデータセットがあることが望ましい．ソーシャルグラフの場合，データ項目はユーザ ID, つながり (リンク, エッジ) の情報, いつリンクが張られたか, ユーザの傾向 (アプリケーション利用, 投稿頻度等), ユーザの年齢 (生年月日) などがあれば，特徴的なグラフパターンの抽出が期待できる．

また２つ以上の異なる時期のグラフ構造のスナップショットがあれば，古いデータからリンク予測した成長グラフと，実際の成長グラフを比べ，その精度を測定できる．

実際に取得できるデータセットの中に，各ユーザのフォロー，フォロワー関係が示され，ユーザアカウントがいつ作成されたかが分かる Twitter のデータセットや，各ユーザに対して Facebook のアプリがいつインストールされたかがわかる Facebook のデータセットなどがある [14] [15] [16].

### 5.5　データセットの分割

グラフデータが大規模な場合，グラフを扱いやすい大きさに分割することがある．特にグラフパターンマイニングを行う場合，出現頻度の多いグラフパターンを見つけるために，多くの数のグラフが必要になるため，分割が欠かせない．

### 5.6　グラフにおけるクラスタリング

ノードをクラスタリング対象とする．ノード間に張られている「エッジ」の密度が高いノードの集まりを「コミュニティ(クラスタ)」とする．

コミュニティの内部は，ノード同士が互いに，密に (たくさん) エッジを張り合っている．またコミュニティと他のコミュニティの間は，エッジの密度が疎となる (少ない)．

グラフのクラスタリングには，トップダウンアプローチとボトムアップアプローチの２つが考えられる．

トップダウンアプローチは，グラフ全体を一つの大きなコミュニティとした状態から始め，エッジを一つずつ切り離して，徐々に小さなコミュニティに分割していく．ボトムアップアプローチは，各ノードを個々の独立したコミュニティとした状態から始め，コミュニティ間のエッジの密度を参考に，コミュニティを順次併合していく．

## 6. 実験手順

### 6.1 データセットの準備

表 3 Twitter データセット

| Twitter グラフ | ノード数 | エッジ数 |
|---|---|---|
| 抽出したグラフ | 6,500 | 5,320,806 |
| 入力グラフ集合 (一例) | 108 | 258 |
| (データセット全体) | *** | (1,468.365,182) |

データセットは文献 [14] で扱われるマイクロブログ,Twitter のデータセットを利用する．ここには celebrities とよばれる，フォロワーの多い著名人のアカウントに関する詳細なデータが 6,500 件ある．ここでフォロワーが多いというのは，2010 年の段階で 10,000 件以上フォロワーをもつものに限定している．以下ではこのデータセットを celebrities のデータセット，このデータセット内の多くのフォロワーをもつアカウントを celebrity なアカウントと表記する．

それに加えて，celebrities の人物を含めた，フォロー関係，フォロワー関係のグラフのエッジのデータがある．このグラフのエッジのデータは，エッジの数 15 億近く, データ容量にして約 25GB あり一つのファイルに格納されているため，このままこの全てをグラフデータとして扱うのは難しい．さらにデータセットを検証すると，celebrity なアカウントと関係をもたないエッジも含まれていたため，ファイルを扱える大きさに分割したのちに，今回関係のないデータを取り除く操作を行った (表 3)．またグラフ構造のデータセットに対して，グラフパターンマイニングを行うことを想定する際，複数の部分グラフからなるグラフ集合から頻出のパターンを抽出を行う．その点を考慮してもデータセットファイルを分割する操作が必要となる．ここではまず扱うデータを，celebrity なアカウント同士の相互フォローの部分に絞って分析を進めていく．

celebrities のデータセットでは，時系列データとしてアカウント作成日時を利用する．これは Twitter の API サービスを用いることにより収集することが可能であり，あらかじめ celebrities のデータセットにも含まれている．またこのデータセットのフォローフォロワーグラフは 2010 年時のものである．

このデータセットによりアカウントをノード, フォローフォロワー関係をエッジと捉えて人間関係を表すソーシャルグラフを生成できる．

### 6.2 TwitterAPI による収集

celebrities のデータセットは,2010 年，つまりある過去の時点での各々のアカウントのフォローフォロワー数が収集されたものである．このある一時点のスナップショットであるデータセットをから，ソーシャルグラフの成長を予測するにあたって，それより後のフォローフォロワー数の推移がわかるデータが必要となる．

そこで実際のフォローフォロワー数がどのように推移したかを確かめるために，Twitter のサービスである OAuth 認証を利用し，TwitterAPI を駆使することにより，celebrity な 6,500 件のアカウントの現在のフォローフォロワーの全アカウント ID を収集を行った．

### 6.3 使用するグラフデータ

celebrities のデータセットや, 収集した現在のデータのうち，実験に扱う部分はともに clebrity なアカウント 6,500 件同士が相互にフォローしているエッジに限定する．このエッジだけ改めてとりだしたデータは，エッジ数が約 500 万ほどとなった (表 3)．そしてこれらのエッジデータを比較して差分をとることにより，増減するフォローフォロワー数の推移を調べることが可能である．ここで celebrities のグラフを過去グラフ (2010 年), 新たに収集した現在のグラフを正解グラフ (2013 年) とする．

### 6.4 予測と正解グラフの差分グラフに対するマイニング

#### 6.4.1 予測グラフの作成

過去グラフに対してリンク予測を行い，予測グラフを生成する．ここでリンク予測の指標は，*Adamic/Adar* を選び，用いることとする．

$$Adamic/Adar := \sum_{k \in |\Gamma(i)| \cap |\Gamma(j)|} \frac{1}{\log |\Gamma(k)|}$$

この指標は，つながりを持たないある 2 つのノードから共参照しているノードの数が多いほど，つながりをもたないある 2 つのノードの間に新たなリンクが張られやすいという考えに基づいたものである．*Adamic/Adar* は，さらにリンクの重要度を表す重みを導入している．ここでいう重要度は，多くのリンクを持たないノードから張られるリンクは重要であると言う考えのもと，ノードが持つリンクの数が少ないほど，そのノードの重要度が大きくなるような重みを採用している．そしてこれを正規化することにより, ノードの類似度としても利用することが考えられる．

この指標を用いて，そしてどの程度 (期間) までグラフの成長予測を行うかのしきい値を設定することにより，過去グラフから現在までの成長を目標にして予測グラフを作成する．

#### 6.4.2 予測グラフと正解グラフの比較

*Adamic/Adar* により生成した Twitter のフォローフォロワー関係の予測グラフと，収集した現在の実際のフォローフォロワー関係のグラフである正解グラフとを比較して，どこが正しく予測されていて，どこの予測が不十分であるかを比較したい．そこでこの予測グラフと正解グラフの差分をとることにする．2 つのグラフの差分をとることにより，2 つのグラフの間で異なる部分のノードとエッジの情報を得ることができる．この差分をとる操作により，元の大きなグラフから切り離された複数の部分グラフを得ることができ，これをグラフパターンマイニングアルゴリズム gSpan に入力する部分グラフ集合として利用する．この部分グラフ集合を差分グラフ (集合) とする．

#### 6.4.3 差分グラフに対するグラフパターンマイニング

前節にて予測グラフと正解グラフの差分をとることで差分グラフ集合を取得した．差分グラフは予測が正解と異なった部分を抜き出したということである．この差分グラフ集合に対してグラフパターンマイニングを適用させる．差分グラフ集合にグ

ラフパターンマイニングを適用させることの目的は，グラフの成長予測に使用したリンク予測指標の精度や特徴を抽出できるとする仮説を検証するためである．

### 6.5 リンク予測に対して時間情報を用いたマイニングによる結果の取り入れ

この実験では，まず過去グラフと現在の (正解) グラフに対して，ノードであるアカウントの生成日時という時間情報を用いることによりグラフパターンマイニングを行う．

### 6.6 グラフデータセットに対する時間情報を用いたグラフパターンマイニング

グラフパターンマイニングを行うとき，グラフは複数に分割されたものが必要となる．そのためまずグラフを分割し，gSpanの入力に与えるためのグラフ集合を作成する．分割方法はいくつか考えられるが，ここではアカウントを作成した期間の近いノード同士ごとにまとめることとする．これは，同じ時期に始めたもの同士がフォローフォロワーのつながりを持ちやすく，同時期に作成されたアカウント同士による有意義なグラフ構造が得られやすいという仮説のもと，実行を試みる．

## 7. おわりに

今回はソーシャルグラフのリンク予測問題に対する，グラフパターンマイニングによるアプローチを提案し，実験手順を示し，考察を行った．リンク予測はある時点のネットワークの構造的特徴から，将来的に変化するネットワークのリンク構造を推測するものであるのに対し，時間情報をラベルにもつグラフマイニングは，頻出する成長パターンを抽出して，ネットワークの成長の特徴を見出すものである．

この２つの手法は，方法は異なるが，いずれもネットワークの将来的な成長の動きのヒントを得られるものである．これらの手法を関連付けて，ネットワーク分析を行うことにより得られる情報は未知であるが，何らかの有意義な結果が得られることは期待できる．

本発表までに，提案した手法を軸にしてソーシャルグラフなどの既存のデータセットに対して解析を試みる．またグラフデータを新たにデータを収集して実験を試みるなども行っていく予定である．

## 文　献


[1] Kashima, H.:Survey of Network Structure Prediction Methods:

[2] Kashima, H. and Abe, N.: A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction, *in Proceedings of the Sixth IEEEInternational Conference on Data Mining(ICDM)*, 2006

[3] David Liben-Nowell, and Jon Kleinberg: The Link-Prediction Problem for Social Networks, *in Proceedings of the twelfth international conferenceon Information and knowledge management(CIKM'2003)*, pp. 556 - 559, 2003

[4] 鷲尾 隆, 樋口 知之, 井元 清哉, 玉田 嘉紀, 佐藤 健, 元田 浩: “ グラフマイニングとその統計的モデリングへの応用, ”統計数理研究所, 統計数理 vol.54, no.2, pp.315-331, 2006

[5] Taihei Oshino: “ Time Graph Pattern Mining for Network Analysis and Information Retrieval, ”*Mastar Thesis*, University of Kyoto, 2011

[6] S. TAKIGAWA: 第 76 回情報処理学会数理モデル化と問題解決 (MPS) 研究会に参加して

[7] Glen Jeh, Jennifer Widom: SimRank: A Measure of Structural-Context Similarity, *in proceedings of the ACM SIGKDD International Congerence on Knowledge Discovery and Data Mining*, 2002

[8] 唐門 準, 松尾 豊, 石塚 満: リンクに基づく分類のためのネットワーク構造を用いた属性生成, 情報処理学会論文誌 vol.47 no.4, 2006

[9] Taiki Miyanishi, Kazuhiro Seki, Kuniaki Uehara: Promissing Nodes Detection using Link Prediction, *The 24th Annual conference of the Japanese Society for Artificial Intelligence*, 2010

[10] X. Yan, J. Han: “ gSpan: Graph-Based Substructure Pattern Mining, ”*Proceeding of the 2002 IEEE International Conference on Data Mining*, pp. 721-724, 2002

[11] S. Ranu, A. K. Singh: “ GraphSig: A Scalable Approach to Mining Signigicant 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

[12] 関 善史, 福田 一郎, 松尾 豊: コミュニティ構造を用いた Web サービスにおけるユーザ推薦手法の検討, 人工知能学会全国大会 (第 26 回), 2012

[13] 稲増 文夫, 海部 美知: ソーシャルグラフの活用可能性について, *KDDI RESEARCH INSTITUTE Inc.*, no.2, 2009

[14] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon: What is Twitter, a social Network or a News Media?, *WWW '10 Proceedings of the 19th international conference on World wide web*, Pages 591-600, 2010

[15] Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou, Walking in Facebook: A Case Study of Unbiased Sampling of OSNs, *in proceedings of IEEE Infocom*, 2010

[16] Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou, Practical Recommendations on Crawling Online Social Networks, *in proceedings of IEEE journal on selected areas in communications*, vol.29, no.9, 2011