In Laboratory Now

研究室訪問 4

Webデータから新たな知識を

村田 剛志 研究室～計算工学専攻

村田 剛志 准教授

今日、SNSや質問サイトといったさまざまなWebサービスが普及している。このようなサイトでは、サービスの質を高めるためにさまざまな技術が使われている。

その技術の一つに、村田研究室が研究しているWeb構造マイニングと呼ばれるものがある。Webから得られる大量のデータをコンピュータで処理することにより、新たな知識を得ようというものだ。本稿では村田研究室が行ってきた数多くの研究の中から、コミュニティ抽出とリンク予測について紹介しよう。

C++ Python Ruby Java ホットな技術・Web構造マイニング

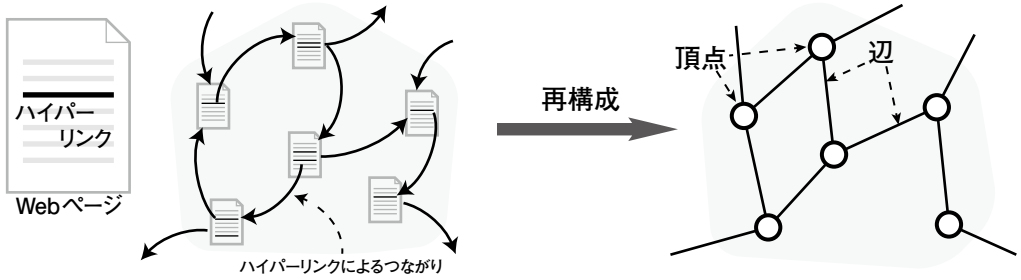
世の中には莫大な量のデータがさまざまな形で存在しており、それらのデータをうまく分析することで有用な知識を得ようとする試みがいたるところで行われている。例えば、スーパーにおいて商品の陳列を改善するために売り上げのデータを分析することで、同時に購入されることが多い商品の傾向を見つけようとする試みがある。その結果、ビールと紙おむつといった、一見関係がないような商品がよく同時に購入されていることが判明することがある。このように、データを分析することで､新たに有用な知識を得ることができる。

しかし、データの量が膨大である現代では人間がすべてを処理することはできない。そこで、コンピュータを用いてデータを分析することで、膨大な量のデータの中から有用な知識を得るという技術が生まれた。これをデータマイニングという｡マイニングとは英語で発掘という意味であり、データの山から貴重な知識を得ようという意味が込められている。

村田研究室では、データマイニングの中でも、Web上のデータを分析することにより有用な知識を得ようとするWeb構造マイニングという技術について研究している。

例えば、Webページ間のつながりを分析することで、それぞれのページの重要度を知ることができる。Webページ同士は、文書内に埋め込まれたハイパーリンクという参照情報によりつながっている。より多くのページからこのハイパーリンクによって参照されているほど、そのページはより重要度が高いと考えられる。しかし、他のページを大量に参照するばかりで特に内容も無いページから多数参照されていても、重要度が高いとはいえない。これらを考慮してWebページ同士のつながりを分析することで、ページの重要度を数値化することができる。この技術はページランクと呼ばれており、検索サイトにおいて、検索結果を表示する順番を決める際に用いられている。

では、Web構造マイニングはどのような手順で行われるのだろうか。まず、Web内の雑多なデータの中から必要な情報を抽出し、グラフと呼ばれる構造にデータを再構成する。ここでいうグラフとは頂点と、二つの頂点間をつなぐ辺の集まりから構成されるデータの構造である。Webページ同



Web上のデータからハイパーリンクによるつながりを抽出し、グラフ構造を再構成する。

図1　ネットワークとグラフ構造

士のつながりをグラフで表現すると、Webページが頂点、それらをつないでいるハイパーリンクが辺になる（**図1**）。

次に、得られたグラフ構造をコンピュータで分析して、有用な知識を得る。後に詳しく説明するが、頂点同士のつながりの構造を調べることで、関連性の高い頂点のグループを見つけることができる。これはコミュニティ抽出と呼ばれる技術である。また、リンク予測という、頂点同士の類似度を計算することにより将来のネットワークの様子を予測する技術もある。

このほかにも、Web構造マイニングに関する研究内容は多数ある。例えば、ネットワークが全体としてどのように成長してきたか、今後どのように変化していくのかといったふるまいをモデル化しようとする研究がある。これにより、あるネットワークが今後どのように発展していくのかを予測することができる。

Web構造マイニングは新しい研究分野であるため、まだ研究が不十分な部分もあるが、マーケティングへの活用やWebサービスの向上などといった幅広い分野で役立つことが期待される。そこで、村田研究室ではWeb構造マイニングという分野そのものを発展させていくために日々研究を行っている。本稿では、村田研究室が行っている数多くの研究の中から、先ほど述べたコミュニティ抽出とリンク予測の二つの技術の研究について紹介する。

n部ネットワークからのコミュニティ抽出

まずは、コミュニティ抽出についての研究を紹介する。先にも述べたが、コミュニティ抽出とは、ネットワークの中から関連性の高い頂点のグループを見つけだす技術のことである。関連性が高いというのは、ネットワークにおいて頂点同士が他の部分よりもより密につながっているということである。

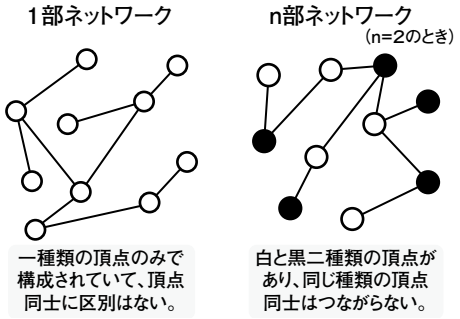


図2　ネットワークの分類

コミュニティ抽出が応用されている例としては、動画投稿サイトにおける、ユーザへの動画推薦が挙げられる。動画投稿サイトに投稿された動画は、投稿者や内容に関するキーワードやタグなどの情報によってグループに分けることができる。同じグループに属する動画を紹介することにより、ユーザが興味を示す可能性の高い動画を推薦することができるようになる。

コミュニティ抽出に関する重要なことの一つとして、ネットワークは、1部ネットワークとn部ネットワークの二種類に分けられるということがある。この場合、nは2以上である。Twitterというsnsでは、ユーザ同士はフォローという関係でつながっており、Facebookというsnsでは、

友達同士という関係を通じてつながっている。これらはユーザという一種類の頂点のみから構成されたネットワークと捉えることができる。このようなネットワークを１部ネットワークという（**図２－左**）。一方、質問サイトにおいては、ネットワークを構成する頂点を、投稿された質問と、質問を投稿・回答するユーザの二種類に分けることができる。

このように、複数の種類の頂点から構成されていて、かつ同じ種類の頂点同士が直接つながっていない、つまり一つの頂点が別の種類の頂点とのみつながっているようなネットワークを、n 部ネットワークという。これから紹介する２部ネットワークとは、n 部ネットワークのうち、頂点が二種類のものである（**図２－右**）。

１部ネットワークからコミュニティを抽出する研究は昔から盛んに行われている。現在では、大量のデータをいかに速く処理できるかが１部ネットワークに関する研究の主な焦点となっている。一方で、n 部ネットワークは１部ネットワークと比べて構造が複雑で、コミュニティの抽出そのものが容易ではないため、研究があまり進んでいないのが現状である。

村田研究室では、n 部ネットワークからコミュニティを抽出する方法について研究している。以下では、村田研究室が主に取り組んでいる２部ネットワークについての研究を紹介する。

２部ネットワークからコミュニティを抽出する方法は、１部ネットワークから抽出する方法を拡張したものである。そこで、まず１部ネットワークからどのようにコミュニティを抽出するのかについて説明する。

はじめに、どのような頂点の集まりが、人間の考えるコミュニティに近い集まりなのかを数値として表さなければならない。与えられた頂点の集まりが、どのくらいコミュニティとしてまとまっているかを表す数値のことをモジュラリティといい、モジュラリティをできるだけ大きくすることを最適化という。コミュニティを抽出するということは、適切にモジュラリティを定義し、それを最適化することだと言える。

モジュラリティの算出は次のように行われる。ある頂点から出ている辺の数を次数という。モジュラリティを計算したい部分ネットワーク上にあるすべての頂点の次数を変えずに、頂点のつながりだけをランダムに組み替えた部分ネットワークを null モデルと呼ぶ。この null モデルと実際のネットワークとを比較する。よりコミュニティがまとまっているほど、そのコミュニティ内では辺のつながりが他の場所より密になっていて、null モデルとの辺の密度の差が大きくなる。その差を数値化して計算することでモジュラリティが求まる。以上のようにして、コミュニティの集まり方を定義し、数値化するのだ。

このようにして定められたモジュラリティを用いて、１部ネットワークからコミュニティを抽出する方法のうち、代表的なものを二つ紹介する。

一つは、CNM と呼ばれる手法である。CNM では、まず各頂点をそれぞれ一つのコミュニティとみなす。そして、隣接したコミュニティのペアを融合することでモジュラリティが増加する場合、融合して一つのコミュニティとする。これを何度もくり返して最大化していくことにより、コミュニティを抽出する（**図３－上**）。

もう一つは、Fast unfolding と呼ばれる手法である。こちらも、まずはすべての頂点をそれぞれ一つのコミュニティとする。次に、モジュラリティが増加するように、すべての頂点を順に他のコミュニティの内部に移動させる。もし、ある頂点を動かしてもモジュラリティが増加しない場合は、その頂点は動かさずにそのままにする。すべての頂点を移動させ終わったら、分割したコミュニティを一つの頂点と考え、新たなネットワークを重み付きで再構成する。この繰り返しによって、コミュニティを抽出するのだ（**図３－下**）。

先生は、これらの１部ネットワークを対象としたコミュニティ抽出方法を、２部ネットワークにも使えるように拡張することを考えた。そのためにはまず、１部ネットワークにおけるモジュラリティを２部ネットワークでも使えるように拡張しなければならない。先生は、それぞれの頂点の種類ごとにモジュラリティを計算し、二つの和をとるという方法でモジュラリティを拡張した。この結果、モジュラリティを２部ネットワークでも使えるようになり、二種類の頂点に関するモジュラリティが互いに影響を及ぼすことにより、二種類の頂点間のつなが

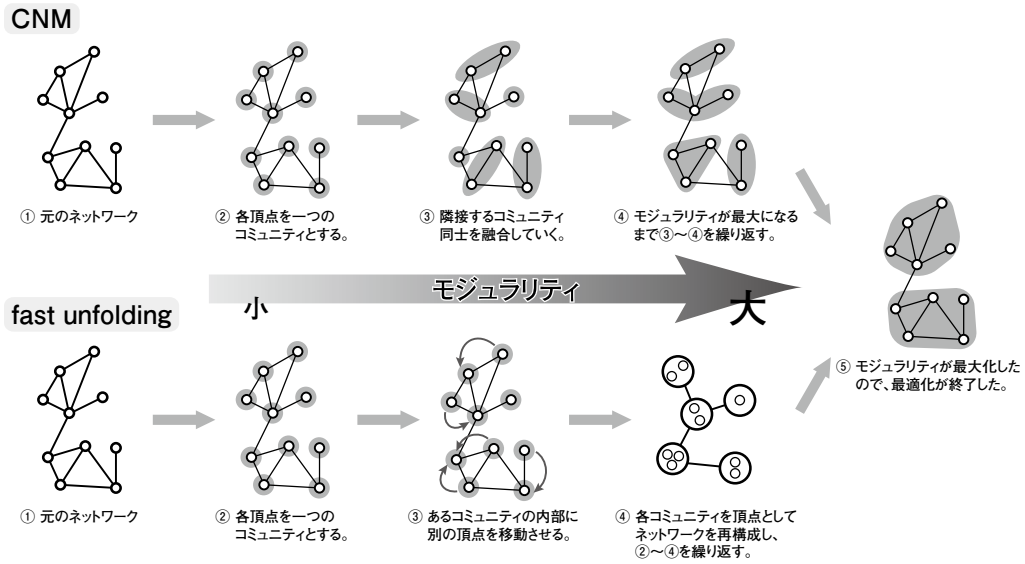


図３　モジュラリティの最適化

りも考慮できるようになった。

しかし、実際に２部ネットワークからコミュニティを抽出するに当たって、先ほど述べた二つの方法をそのまま用いたのでは、問題が生じる。村田先生は、応用のことも考えて一つのコミュニティには一種類の頂点のみを内部に含めるということにしている。コミュニティを抽出する際にはこの制約も考慮しなくてはならない。なぜなら、そのまま隣接したコミュニティ同士を融合したり、頂点を移動したりすることができないからである。

そこで先生は、隣り合う頂点やコミュニティではなく、共通隣接点をもつコミュニティや頂点を融合、移動することにした。共通隣接点とは、同じ種類の頂点二つに共通してつながっている頂点のことである。２部ネットワークの場合、ある共通隣接点をもつ二つの頂点は同じ種類の頂点となる。

このように、先生はモジュラリティの定義や最適化の手法を拡張することによって、２部ネットワークからコミュニティを抽出する方法を考案した。このような研究の積み重ねが、実際のソーシャルメディアへの応用につながっていくのだ。

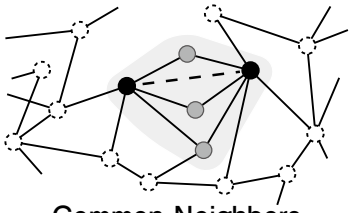
類似度を用いたリンク予測

次に、リンク予測に関する研究について紹介する。リンク予測とは、刻々と変化するネットワークにおいて、頂点同士の現在のつながりから将来の頂点同士のつながりを予測する技術である。これは、あるユーザに、共通点のありそうなユーザを紹介するようなSNS上の機能などに応用されている。リンク予測の手法に関しても、コミュニティ抽出と同様、盛んに研究が行われてきた。

主な予測法として、二つの頂点の類似度を用いる方法がある。類似度とは、ある頂点のペアを考えたときに、その二つの頂点の関係性を示す数値のことである。具体的には、SNSなどにおいて、どのくらい似たような友人関係をもつかということを表すものである。

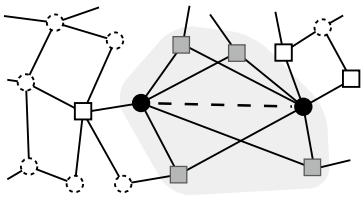
類似度を用いたリンク予測では、まず類似度という量を定義しなければならない。二つの頂点の関係性といってもさまざまなとらえ方が可能であるからだ。そしてその定義にもとづいて頂点のペアの類似度を計算する。この類似度が高いほどリンクが存在する可能性が高いと考える。村田研究室では、複数の種類の類似度を組み合わせることによってリンク予測の精度を向上させた。

類似度の定義の仕方として、大きく分けて二つのアプローチがある。一つは、ネットワークの局



Common Neighbors

二つの黒頂点に対する共通隣接点（灰色頂点）の数が多い方が、リンクが存在する可能性が高い。



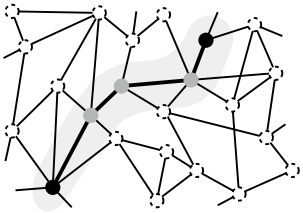
Jaccard's Coefficient

二つの黒頂点それぞれとつながっている頂点（四角頂点）のうち、共通隣接点（灰色四角頂点）の割合が多いほど、リンクが存在する可能性が高い。

図４　リンク予測の手法（その１）

所的な部分だけを計算に用いる方法である。その中の一つにCommon Neighborsという手法がある。これは、ある頂点と別の頂点の間の共通隣接点の数が多いほどリンクが存在する可能性が高いという考え方にもとづいている（**図４**－上）。また、Jaccard's Coefficientという手法もある。この手法は、ある二つの頂点それぞれとつながっている頂点のうち、共通隣接点の割合が大きいほどリンクが存在する可能性が高いという考え方にもとづいている（**図４**－下）。これらの手法は、計算量が少ないので、高速な計算が可能となっている。一方で、用いる情報が局所的なものなので、リンク予測の精度が落ちるという欠点がある。

もう一つのアプローチはネットワーク全体からリンクを予測しようとする方法である。例えば、Shortest Pathという手法がある。これは、ある頂点からもう一つの頂点までたどり着



Shortest Path

黒点間を移動するのに必要な辺の数が少ないほど、リンクが存在する可能性が高い。

図５　リンク予測の手法（その２）

くために通らなければならない辺の数が少ないほど、リンクが存在する可能性が高いという考え方にもとづいている（**図５**）。こういった手法は、多くの情報を用いるために比較的精度は高くなるが、計算時間が長くなるという欠点をもっている。

村田研究室では、これらのさまざまな特徴をもった複数の類似度を適切な係数を掛けた上で足し合わせることで、より精度の高いリンク予測ができるのではないかと考えた。

具体的な予測の方法は次の通りである。まず、ネットワークに現在存在するリンクのうち、一部のリンクを除いたネットワークを用意する。次に、除いたリンクをプログラムに発見させる。これを繰り返しながら、除いたリンクを正確に発見できるよう、各類似度に掛ける係数をプログラムに適宜学習させていく。こうすることで、高い精度でのリンク予測を実現したのだ。

今後の課題としては、計算時間の短縮が挙げられる。分析対象のネットワークが大きくなるほど計算にかかる時間が莫大になってしまうからだ。また、今のところ短期的な予測についての研究が盛んに行われているが、長期的なリンク予測を行うことも課題となっている。村田研究室の今後の研究によりこれらの課題が克服されることに期待したい。

今回の取材は自分がこれから学ぼうと考えている情報工学分野の研究であり、一年生である今のうちからその最先端の研究に触れることができたのはとても幸運でした。

村田研究室ではWeb構造マイニングを発展させるべく、紹介したもののほかにもさまざまな研究を行っており、紙面の都合上そのすべてを紹介できなかったことが残念です。

最後になりますが、大変お忙しい中、度重なる取材や質問に快く応じてくださった先生方に心よりお礼申し上げます。ありがとうございました。

（松山 祐輔）