交グラフと意味的解析を利用したコミュニティ発見手法のSNSネットワークへの適用

Applying to SNS Networks of a Community Detection Method using Intersection Graph and Semantic Anlysis

岡田 直樹 Naoki Okada　谷川 恭平 Kyouhei Tanikawa　土方 嘉徳 Yoshinori Hijikata　西田 正吾 Shogo Nishida

大阪大学基礎工学研究科
Graduate School of Engineering Science, Osaka University

There is an increasing number of researches of complex networks such as World Wide Web, social networks and biological networks. One of the hot topics in this area is community detection. Nodes belonging to a community are likely to have common properties. For instance, in the World Wide Web, a community may be a set of pages which belong to a same topic. Community structure is undoubtedly a key characteristic of complex networks. In this paper, we present a new framework for finding communities in complex networks and evaluate detecting the community. This framework uses the idea of intersection graph and uses semantic information such as texts and attributes which appear in networks.

1. はじめに

SNS ネットワークや World Wide Web，タンパク質の相互作用といった複雑ネットワークに対する研究が増えており，スケールフリー性やスモールワールド性，クラスター性などの性質が発見されている [Albert 02]．近年では，複雑ネットワーク内のコミュニティ構造とその発見手法が注目されている．コミュニティとは，ノードの部分集合の中で，その部分集合に含まれるノードが，その集合に含まれないノードよりもその集合に含まれるノードとより密につながっているものと定義されることが多い．この定義に基づき，多くのコミュニティ発見手法が提案され，様々な複雑ネットワークに対してコミュニティ構造の分析が行われている [Danon 05]．

このコミュニティ発見問題において，コミュニティ間の重複を抽出できるかどうかという点が重要視され始めている [Zhang 07]．コミュニティ間の重複とは，あるノードが複数のコミュニティに属している状態を指す．例えば SNS ネットワークにおいては，ある人物が複数のコミュニティ(同じ大学のグループと同じ職場のグループなど) に属するのは自然であるし，World Wide Web においても 1 つのページが複数のトピックに分類されることは十分に考えられる．また，複雑ネットワークにおけるコミュニティ発見手法では，ネットワークの均一性が前提となっていることが多い．つまり，ノード間のエッジがすべて同質であるネットワークを仮定している場合が多い．しかし，実際のネットワーク内のエッジは同質でないことが多い．例えば，SNS ネットワークであれば，大学つながりの友人やサークルつながりの友人など，様々なつながりが存在する．World Wide Web であれば，他のサイトを参照するリンクもあれば，広告へのリンクも存在する．

我々が提案する手法では，交グラフの概念を用いることでコミュニティ間の重複の抽出を可能にしている．また，ネットワーク内に現れる意味的な情報の類似度を用いて，エッジに重み付けを行うことで，ネットワーク内のエッジの不均一性の問題に取り組む．

連絡先: 岡田直樹，大阪大学大学院基礎工学研究科，〒 560-8531 大阪府豊中市待兼山町 1-3，TEL:06-6850-6382，FAX:06-6850-6341，e-mail:okada@nishilab.sys.es.osaka-u.ac.jp

2. 提案手法

入力となるのは，グラフ $G = (V, E)$ である．V はノードの集合，E はエッジの集合である．またエッジには意味的な情報が付与されているものとする．例えば，SNS ネットワークであれば，各個人の友人紹介文に相当する．このグラフに対し，以下の 4 つのステップを適用し，最後にコミュニティへの抽出結果を出力する．

- **Step 1. 密な部分グラフの列挙：**グラフ $G = (V, E)$ から密な部分グラフ (一般にクリークと呼ばれ，様々な定義が存在する) を列挙する．
- **Step 2. 交グラフへの変換：**Step 1. で列挙した各部分グラフを 1 つのノードとする交グラフへと変換する．交グラフとは，ある全体集合の中に複数の部分集合が存在するとき，各々の部分集合を 1 つのノードと見なし，ある 2 つの部分集合の間に共通要素があれば，それらに対応する 2 つのノード間にエッジを張ることで得られるグラフのことである．
- **Step 3. エッジの重みの算出：**交グラフ内のエッジに対して集合の重なり度合いと意味的な情報の類似度を用いて重みを算出する．
- **Step 4. モジュール性に基づくクラスタリング：**モジュール性に基づくクラスタリング [Newman 04] により交グラフ内のノードを複数のクラスタに分割する．この手法を簡単に述べると，まず各ノードを 1 つのクラスタとみなす．あるクラスタ i とクラスタ j を併合する際のモジュール関数 Q の増加分 ΔQ を最大にするクラスタの組み合わせを見つけ，これらのクラスタを併合していくというステップを，クラスタが 1 つになるまで繰り返す．各ステップにおいて Q の値を計算しておき，Q の最大値を与えるステップでの分割結果を出力する．

以降はこの提案手法を「テキスト解析あり」の方法と呼ぶ．また，意味的な情報の類似度の影響がどの程度貢献しているのかを調べるために，Step 3. におけるエッジの重みとして重なり度合いのみを用いた方法も実装し，意味的な情報の類似度

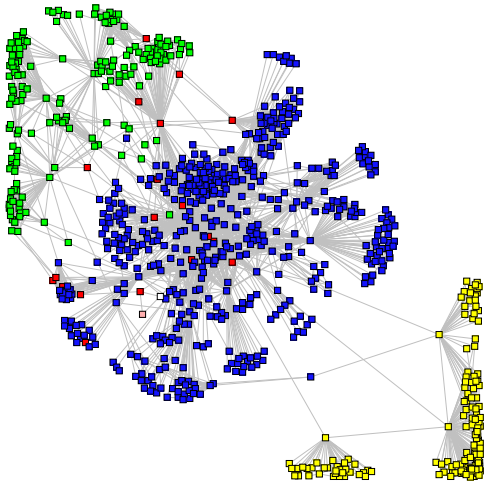

図 1: テキスト解析ありでの抽出結果

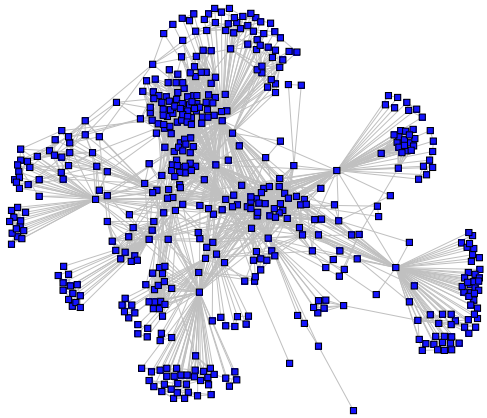

図 2: テキスト解析なしでの抽出結果

を用いる場合と比較した．以降はこの方法を「テキスト解析なし」の方法と呼ぶ．

3. SNS ネットワークへの応用

本研究が対象とするネットワークは，ノード間の関係性が多様であり，この関係性を表す意味的な情報が利用できるネットワークである．よって我々は日本の代表的な SNS ネットワークである mixi[*1] を選択した．mixi では利用者が自己紹介文を書き，さらに自分の友人に対して友人紹介文を書くので意味的な情報を獲得しやすい．また利用者間の関係性は大学つながりや職場つながり，趣味のつながりなど様々である．mixi のネットワークにおけるノード間のつながりのことをマイミクの関係と呼ぶ．データセットとして，被験者を中心ユーザとし，この中心ユーザから半径 2 までのリンク構造を抽出した．さらに，意味的な情報としてデータセット内のユーザ間の友人紹介文を収集した．ある被験者のネットワークに対してテキスト解析ありの方法で適用した抽出結果を図 1 に示す．このように，ネットワークから複数のクラスタを抽出できていることがわかる．また，図 2 に，同じ被験者のネットワークに対してテキスト解析なしの方法で適用した抽出結果も図 2 に示す．ここでは，1 つの大きなクラスタを抽出している．実験の結果，テキスト解析を行うほうがより多くのクラスタを抽出することがわかった．

4. 評価実験

我々の提案手法により抽出された各クラスタがどの程度正確なものかを測定するために，評価実験を行った．被験者には事前に被験者のマイミクに対して，どのような関係であるかのタグ付けを行ってもらい，それを正解データとした．そして，同一のタグを付けられた被験者のマイミクをまとめて 1 つの正解グループとした．この正解グループを用いて抽出したクラスタの精度，再現率を求め，その平均を F 値とし，F 値を抽出したクラスタの評価値とした．結果としては，テキスト解析ありとテキスト解析なしで評価値の F 値は変わらなかった．

しかし，この評価方法では被験者のマイミクだけしか評価に入っていないので，評価する標本数が少なく，真の精度と異なる可能性があると考えられる．よって，今後は被験者のマイミクのマイミクまでの評価を行う必要があると考える．

5. おわりに

本研究では，複雑ネットワークにおけるコミュニティを抽出するための手法を提案し，意味的解析の影響を確かめるための評価を行った．特に，多くのネットワークが持つコミュニティ同士の重なりと，ノード間の関係の不均一性の性質に着目した．また，交グラフの概念を利用することでコミュニティ間の重複を抽出可能にし，ベクトル空間モデルによってネットワーク内の意味的な情報の解析を可能にした方法を提案した．最後に，この提案手法をインターネット上のコミュニティサイトから収集したデータセットに対して適用し，抽出したクラスタの評価を行った．その結果，意味的解析を行うほうが多くのクラスタを抽出することがわかった．また，今回の評価実験では意味的解析の有無で評価値に差が出なかった．

今後の展望としては，抽出したクラスタの適当な評価方法を吟味する．また，SNS ネットワークと異なる複雑ネットワークの World Wide Web ネットワークへの適用も試みる．

参考文献

[Albert 02] R. Albert, A.-L. Barabasi: Statistical mechanics of complex networks, Review of Modern Physics, vol.74, pp.47-97(2002)

[Danon 05] L. Danon, J. Duch, A. D. Guilera and A. Arenas: Comparing community structure identification, J. Stat. Mech, P09008 (2005)

[Zhang 07] S. Zhang, R. Wang and X. Zhang: Identification of overlapping community structure in complex networks using fuzzy c-means clustering, Physica A 374 483-490 (2007)

[Newman 04] M.J.Newman and M.Girvan: Finding and evaluating community structure in networks,Phys.Rev.E 69,026113(2004)

∗1 http://mixi.jp/