1C3-OS-12-5

交グラフと意味的解析を利用したコミュニティ抽出手法のTwitterネットワークへの適用

Applying to Twitter Networks of a Community Extraction Method using Intersection Graph and Semantic Anlysis

倉持 俊也 Toshiya KURAMOCHI　岡田 直樹 Naoki OKADA　谷川 恭平 Kyohei TANIKAWA　土方 嘉徳 Yoshinori HIJIKATA　西田 正吾 Shogo NISHIDA

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

多くの研究者によって SNS ネットワークや World Wide Web などの複雑ネットワークの解析が行われてきた．近年ではコミュニティ構造に着目した研究が盛んである．我々は交グラフと意味的な情報を用いた階層的クラスタリングによるコミュニティ抽出手法を提案してきた．本研究では提案する手法を Twitter ネットワークに適用，評価し，その有効性を検証する．

1. はじめに

多くの研究者によって，SNS ネットワークや World Wide Web といった複雑ネットワークに対する解析が行われ，様々な性質が発見されている．中でもコミュニティ構造は複雑ネットワークを分析するのに不可欠となりつつあり，近年ではコミュニティ抽出手法が数多く提案され，様々な複雑ネットワークに対して適用されている [Danon 2005]．

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という点が重要視され始めている．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．社会ネットワークにおいては，ある人が複数のグループに属しているのは自然である．そのため，1 つのノードを 1 つのコミュニティに割り当てるのではなく，複数のコミュニティに割り当てることのできる手法が必要である．

多くのコミュニティ抽出手法では，ネットワークの均一性が前提となっている．つまり，全てのエッジが同質であると仮定している．しかし，実際のネットワークでは，エッジが同質でない場合が多い．World Wide Web では，ページ間のハイパーリンクをエッジと見なすことができるが，他のサイトを参照するリンクやサイト内リンク，広告リンクなど，様々な種類のエッジが存在する．このような異なる性質や意味を持つエッジは，同質として扱うのでなく，区別したり，適切な重み付けを行えることが重要であると考えられる．

ネットワークのクラスタリングでは階層的クラスタリングが広く用いられている．提案されている多くのクラスタリング手法では，予め出力したいクラスタの分割数を入力する必要がある．しかし，実際のネットワークでは，事前に正解コミュニティ数が分からないことが多い．よって，人手で分割数を与えなくとも自動的に適した分割数を判定する方法が求められる．

我々は上記の問題を解決する新しいコミュニティ発見手法を提案する．我々の手法では，交グラフの概念 [Everett 1998] を用いることでコミュニティ間の重複を抽出可能とし，また，意味的情報の解析によりエッジに重み付けを行うことでネットワークの不均一性を表現する．さらに，モジュール性 [Newman 2004] に基づくクラスタリングによって適切なクラスタ数を自動で判別する．

連絡先: 倉持俊也，大阪大学大学院基礎工学研究科，大阪府豊中市待兼山町 1-3 基礎工学研究科，06-6850-6383，kuramochi@nishilab.sys.es.osaka-u.ac.jp

2. 提案手法

我々の提案手法はグラフ $G = (V, E)$ と意味的情報 (テキスト情報など) を入力とし，以下の 4 ステップの処理を行うことでコミュニティを抽出する．

Step 1. 密な部分グラフの列挙

入力されたグラフ $G = (V, E)$ から密な部分グラフとしてクリーク (完全グラフ) を列挙する．これらのクリークの要素数に閾値 (クリークの閾値) を設け，閾値以下のクリークを取り除いた．ただし，クリークは無向グラフに対して定義されている．そこで，入力されるグラフが有向グラフの場合は，エッジの方向を無視し，エッジが存在するかどうかのみに着目して無向グラフ化する．この方法は World Wide Web の解析などに用いられている [Albert and Barabasi 2002]．

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，グラフ $G = (V, E)$ 中の部分グラフ S_i をノードと見なし，集合間に空でない共通集合が存在するときに，これらのノード間にエッジを張ることで生成される無向グラフである．交グラフのノード間にエッジを張る際に，集合の共通要素数に閾値 (重複度の閾値) を設けられる．閾値を高くすることでエッジ数を減らし，クラスタリングの計算時間を削減することができる．また，重複度の小さい集合間は弱いつながりであると考えられ，これらを取り除くことで精度の高いクラスタリングが実現できると考えられる．

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 (交グラフのノード) の重なりの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．集合の重なりの度合いは Jaccard 係数 $d(X, Y) = |X \cap Y|/|X \cup Y|$ を用いて計算する．意味的情報の類似度についてはベクトル空間モデルを利用した．集合内のテキスト中に生起するキーワードに対して TFIDF 値を求め，その各値をベクトルの要素とすることで各集合を 1 つのベクトルと見なす．ベクトル表現された集合間の類似度 $sim(X, Y)$ はコサイン類似度を用いて計算する．集合の重なりの度合い $d(X, Y)$ と意味的情報の類似度 $sim(X, Y)$ から式 (1) により各集合間のエッジの重み付けを行う．

$$w(i, j) = w(X, Y) = \frac{d(X, Y)}{1 + \varepsilon - sim(X, Y)} \quad (1)$$

ただし，ε は分母を 1 としないための定数である．

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により交グラフから複数のコミュニティを抽出する．このクラスタリングは，式 (2) で表されるモジュール関数 Q を最大とするようにノードを併合する凝集型のアルゴリズムである．

$$Q = \sum_{i=1}^{k} (e_{ii} - a_i^2) \tag{2}$$

ここで，e_{ii} は総エッジ数に対するクラスタ i 内に存在するエッジ数の割合を，a_i は総エッジ数に対するクラスタ i から他のクラスタに張られるエッジ数の割合を表す．最適な Q を求める方法として，最も一般的である貪欲法を用いた．

3. Twitter での評価実験

我々は提案手法をマイクロブログサービス Twitter[*1] のネットワークに適用し，評価実験を行った．Twitter では利用者は自由記述でプロフィールを書き，“ツイート” と呼ばれるショートメッセージで交流している．我々はこれらのテキスト情報を意味的情報として用いた．意味的解析を行わない方法 (“None” と記す) と，意味的解析を行う方法を実装した．意味的解析を行う方法の例としては，プロフィールとツイートを用いた方法 (“Profile”，“Tweet” と記す) や，それらを会話 (“メンション”) に含まれるユーザ名 (“Name” と記す) や “ハッシュタグ”(“Hash” と記す) と組み合わせた方法を実装した．比較手法として Everett らの方法 (“Everett” と記す) を選択した．Everett らの方法は，Step 2. まで我々の提案手法と同様の処理を行い，交グラフにおいて重なりの度合いの大きいものから順に併合を繰り返す方法である．

我々は 9 名の被験者を中心とした 2 近傍ネットワークと，その中に記述された意味的情報をデータセットとして用いた．被験者が回答したデータセット内のユーザとのつながり名 (関係性) を正解データとする．

Everett，None，Profile，Tweet の各手法を精度，再現率，F 値で評価した結果を図 1 に示す．クリークの閾値，重複度の閾値が大きいほど精度が高い傾向が見られる．また，重複度の閾値が 2 のときに最も良い再現率を得る．各手法の平均評価値を図 2 に示す．提案手法を従来手法 (“Everett”) と比較すると，精度に関しては従来手法と変わらない値であるが，再現率では提案手法が大きく上回っており，F 値でも大きな向上が見られた．提案手法のうち，意味的解析を行わない方法 (“None”) と意味的解析を行う方法を比較すると，意味的解析によって精度が向上していることが分かる．また，F 値においても意味的解析を行う方法が，わずかではあるが意味的解析を行わない方法を上回る傾向が見られた．プロフィール文のみを用いる手法で最も高い F 値が得られた．意味的解析に用いる情報の組み合わせに関しては一定の傾向は見られなかった．

4. おわりに

本研究では，交グラフと意味的解析を用いたコミュニティ抽出手法を提案し，評価実験を行った．我々の手法はコミュニティ間の重複，エッジの不均一性の表現，クラスタ数の自動判別に着目している．Twitter ネットワークでの評価実験を行い，有用性を検証した．今後はより多様な Twitter 利用者を中心として得たネットワークに提案手法を適用し，評価，分析を行う．

*1 http://twitter.com

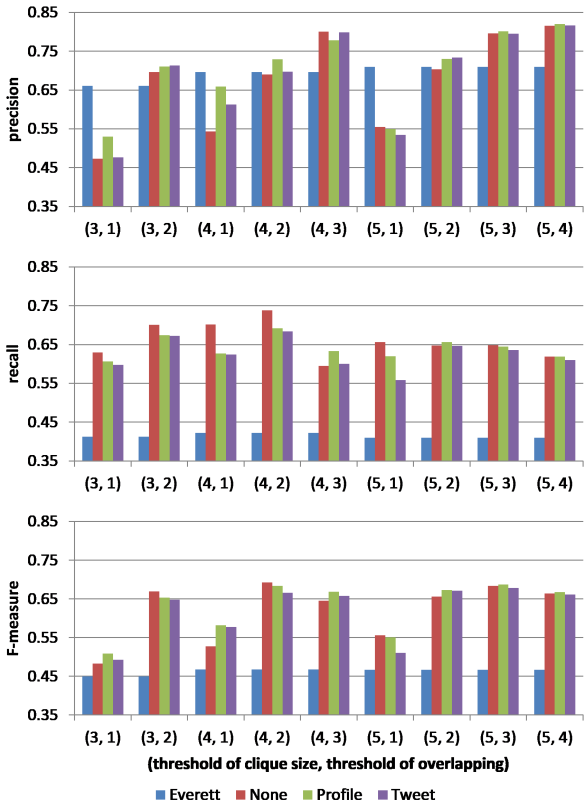

図 1: 抽出したコミュニティの精度，再現率，F 値

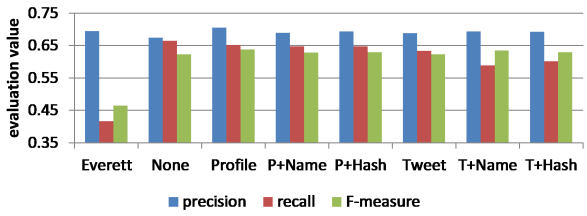

図 2: 各手法の平均評価値

参考文献

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

[Danon 2005] Danon, L., Duch, J., Guilera, A. D., Arenas, A., Comparing community structure identification, J. Stat. Mech, p.09008 (2005).

[Everett 1998] Everett, M. G., Borgatti, S. P., Analyzing Clique Overlap. CONNECTIONS 21(1):pp.49-61. (1998)

[Newman 2004] Newman, M. E. J., Girvan, M., Finding and evaluating community structure in networks, Phys. Rev. E69, p.026113 (2004).