P-9

SNS における高速コミュニティ抽出法の提案とその評価に関する検討

Study on High Speed Network Community Extraction Method and its Performance.

○佐藤真弥子[1]，　吉開範章[2]，　栗野俊一[2]
*Mayako Sato[1], *Noriaki Yoshikai[2], Shun-ichi Kurino[2]

Abstract: From the fact that SNS has a big data, it is difficult to extract the valid community in a short period of time and acquire the characteristics of the community by the mathematical treatment. In this study, we have proposed high speed network community extraction method using a two-stage clustering in the SNS. In addition, we examined its effectiveness by the case studies using the Twitter data and computational complexity.

1.　背景

ソーシャルネットワーク分析（以下，SN 分析）を用い，ネットワーク上のコミュニティ構造を明らかにし，ネットワーク上の活動を活性化や効率化することを目的とした研究が行われている[1][2].

近年 Facebook や Twitter などの SNS が手軽さ・話題性などから急速に普及し，多くの人が SNS を利用してネットワーク上でコミュニケーションを取るようになった．SNS の情報 はビッグデータであることから，その中から短時間で有効なコミュニティを抽出することが難しい．さらに，コミュニティの特徴を機械的に獲得することは困難と考えられる．そのため， API が公開されデータの入手が可能な Twitter を対象に，HITS (Hyperlink-Induced Topic Search)[3] を用いて高速処理を行い，ユーザーの活動を情報提供・情報収集・情報共有の 3 タイプに分類する方法[4] や，コミュニティ自体を分析するパラメータとして，「類似性」を使ってコミュニティを分析する方法[5]等が検討されている．

しかし，コミュニティ抽出の高速化に着目をした研究は，あまりなされていなかった．

本稿では，SNS における効率的なコミュニティ抽出手法を提案し， その有効性の検討を，計算量と Twitter データを用いたケーススタディにより行ったので報告する．

2.　提案手法

図 1 に，提案手法の主要処理フローを示す.

使用言語は，スクリプト言語 Python であり，Twitter API に準じて必要なプログラムを作成し，データ収集やデータの出力を行った．また，Excel およびネットワーク 分析ソフト NetMiner[6] を用いてデータの分析を行った.

2.1.　コミュニティ抽出

2.1.1.　対象ユーザーの決定とネットワーク生成

HITS を用いてノードの Hub 値および Authority 値を算出し，ユーザーのランキング付けを行う．その後，閾値を設定（今回は 0.05 ）して足切りを行い，最終的に Hub・Authority の両方に残っているユーザーのみを抽出して，対象ユーザーとする.

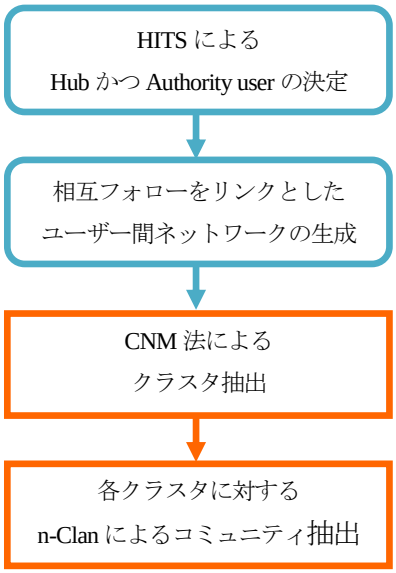

図 1.　ネットワーク生成
コミュニティ抽出フロー

次に，対象ユーザー群に対して，2 ユーザー間に相互フォローが認められる場合にのみリンクを張り，対象ユーザー間ネットワークを生成する.

ここで出てくる Authority とは，情報を発信しているノード，Hub とは，情報の収集・取りまとめを行っているノードを指す．ノードの Hub 値は，出リンク先のノードが持つ Authority 値の合計から算出され，Authority 値も同様に入リンク先のノードが持つ Hub 値によって決まる.

2.1.2.　コミュニティ抽出

対象ユーザー間ネットワークに対し，モジュラリティによってクラスタ抽出処理を行う CNM (Clauset Newman Moore) 法[7] により，クラスタ抽出を行う.

さらに，各クラスタに対して，n-Clan (n=2) によりクラン抽出を行い，そこで抽出されたクランをコミュ

1：日大理工・院（前）　2：日大理工・教員・数学

ニティとする.

なお CNM 法では, 最初それぞれのノードを独立したクラスタとして扱う. 任意のクラスタのペアについて，合併した場合のモジュラリティの増加量 ΔQ_{ij} を計算し，ΔQ_{ij} が最大となるペアの合併を繰り返す. そして，ΔQ_{ij} 負になった時点で終了し，最終クラスタを決定する. a_i を総エッジ本数に対するクラスタ i から他のクラスタに張られているエッジ本数の割合, e_{ij} を総エッジ本数に対するクラスタ i からクラスタ j に張られているエッジ本数の割合とすると，モジュラリティの増加量 ΔQ_{ij} は以下の式で求まる.

$$\Delta Q_{ij} = 2(e_{ij} - a_i a_j)$$

2.2. 高速化

従来，よく知られたコミュニティ抽出手法である n-Clan の計算量は n をノード数とする時, $O(2^n)$である．したがって，分析データのノード数が増えると計算量が大きくなり，計算時間が急激に増大する.

一方，今回提案したコミュニティ抽出手法では，CNM 法と n-Clan の 2 段階クラスタリングを行っている．これにより，n-Clan を適用する際のノード数が小さくなり，そのまま n-Clan を適用する場合と比べて計算量が少なく, 計算時間が短くなることが予想される.

3. データの分析と考察

2012 年 12 月に行われた衆議院議員選挙に関わる Twitter データを対象として分析した．収集した日数は 12 月 9 日から 12 月 15 日までの 7 日間，ツイート件数は 38,205 件, ユーザー数は, 18,794 人である.

3.1. コミュニティ抽出時間と抽出数

コミュニティ抽出の所要時間とデータサイズ，コミュニティ抽出数は, 表 1 の通りである.

表 1. コミュニティ抽出におけるデータサイズと所要時間

	ノード数	リンク数	所要時間	コミュニティ抽出数
12/9	93	1968	46ms	3
12/10	80	1092	78ms	3
12/11	67	845	31ms	2
12/12	107	3104	78ms	3
12/13	91	1978	48ms	5
12/14	106	3298	62ms	2
12/15	94	2104	62ms	7

※分析にかけるまでのデータ編集時間は含まない

分析に使用したデスクトップコンピュータでは, 表 1 に示した程度のデータサイズであっても，n-Clan のみのコミュニティ抽出を試みると，計算量の多さから一日経過しても計算が終了しないケースや, エラーを出力して停止するケースがあった.

しかし，今回提案した CNM 法と n-Clan の 2 段階クラスタリングを行うことで，表 1 に示したデータサイズでは, 1 秒以下でのコミュニティ抽出が可能となった.

4. まとめ

今回は，2 段階クラスタリングを行うことで，短時間でコミュニティを抽出が可能な手法を提案した. またケーススタディでは，選挙に関わる Twitter データを用いて，HITS による対象ユーザーの決定と提案手法が, 短時間でコミュニティ抽出を行うことに有効であることを示した. コミュニティ抽出が出来たことで, 次はコミュニティの特徴把握が必要となる.

そこで, ユーザー属性から得られる類似性を用いたコミュニティの特徴づけについて, 形態素解析を駆使した属性パラメータの生成などを経て, 現在検討を行っている. 今後も検討を続け，結果がまとまり次第報告する予定である.

5. 参考文献

[1] Mohsen Jamali, Hassan Abolhassani: “Different Aspects of Social Network Analysis”, Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference, ISBN: 0-7695-2747-7, pp.66-72, (2007)

[2] 北原, 吉開： “アフィリエーションネットワークを用いた活動評価法の提案と評価”, 信学技法 SITE2011-55 pp.317-322,2012.

[3] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment", Journal of the ACM, 46(5):604632, 1999.

[4] A.Java, X.Song, T.Finin, B.Tseng, “Why we twitter: understanding microblogging usage and communities”, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis Pages 56-65.

[5] Yang ZHANG, Yao WU, Qing YANG, “Community Discovery in Twitter Based on User Interests”, Journal of Computational Information Systems 8: 3 (2012) 9911000.

[6] NetMiner, http://www.netminer.com/index.php

[7] Aaron Clauset, M. E. J. Newman, Cristopher Moore, “Finding community structure in very large networks”, Phys. Rev. E, 70:066111, Dec 2004