Group Formation in Large Social Networks: Membership, Growth, and Evolution (2006)
L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan (Cornell Univ.)
KDD2006, http://www.cs.cornell.edu/ lars/kdd06-comm.pdf

1 本論文の目的

本研究では二つの巨大データを用いてソーシャルネットワークの構造などの属性と、コミュニティの変化がどのように関係があるのかを調べる。問いは大きく以下の三つである。

- 参加: どのような構造的特徴が個人が特定のグループに参加するのに影響を与えるのか。
- 成長: どのような構造的特徴が既存のグループの成長に影響を与えるのか。
- 変化: どのグループもある時点において目的というものがある。そのフォーカスがどのように変化するのか、また、その変化にはグループメンバーの移動がどのように関係しているのか。

データは Weblog サービスの LiveJournal と、論文 DB の DBLP を用いる。LiveJournal は友達リンクを持ち、ブログを投稿することでコミュニティに参加できる。DBLP は共著関係を持ち、投稿先のカンファレンスをコミュニティとみなす。

2 背景

複数の集団によって起きるプロセス、例えばメンバーの移動やそれによる集団の変化は、社会科学の中心的な課題である。政治や企業、宗教団体などは全てそのよう集団の基本的な例となる。近年では、オンライン上での人の集まりがコミュニティやソーシャルネットワークサイトの成長により目立って増えてきている。

これらについては既存研究により様々な分析がなされているが、コミュニティの巨大時分割データの集積や分析という点では、グループの成長という基礎的な問題が残されたままである。本研究ではこれらに取り組む。

3 コミュニティへの参加

3.1 手法・アルゴリズム

LiveJournal は、10 日間の間に少なくとも 1 つの投稿があったコミュニティのうち、もっともアクティブだった 700 件とそれ以外からランダムで選択した 300 件の合計 1000 コミュニティを選んだ。しかし、1000 人以上のメンバーがいるコミュニティは正確にデータが収集できないため、それを除いた 875 コミュニティを今回のデータセットとした。DBLP は全てのデータを用いた。

19 個のコミュニティを特徴づける指標を用意する。大きく分けて 2 種類あり、一つはコミュニティそのもの特徴（コミュニティのメンバー数、fringe [1]の数、閉じたトライアドの開いたトライアドに対する比、など)、もう一つは個人とコミュニティ内にいる友人に関する特徴（コミュニティ内にいる友人の数、コミュニティ内にいる友人同士が互いに友人である割合、など）である。

この指標を用いてユーザ u がコミュニティ C に参加するかどうかを推定する決定木を作成する。高い精度を持つ決定木を作るために必要な特徴は何か、また、決定木の上位に来る特徴は何かを調べることで、コミュニティへの参加にとって重要な特徴を見つける。

3.2 評価

友人 (DBLP では共著者) が k 人が入っているコミュニティ C にその人が一ヶ月後以内に入った割合を調べた。コミュニティ内にいる友人の数 k が多いほど参加する割合は増加するが、ある程度で収束する。コミュニティへの参加を推定する決定木を作成したところ、最上位のルールは「コミュニティにいる知り合いの内、その中で互いに知り合いである割合」であった。

3.3 知見

コミュニティへの参加は、その中にいる知り合いの数以上に、その知り合い同士に関係があることが重要であった。コミュニティ内において知り合い同士に関係があることは、信頼の観点からアドバンテージがあると社会的資本の文脈ではいわれており、この結果はその説に沿ったものであるといえる。

4 コミュニティの成長

4.1 手法・アルゴリズム

4ヶ月の期間を空けて二つの LiveJournal のスナップショットを取った。小さいコミュニティは見えにくい変化要因が多いため、100 人以上のメンバーがいるコミュニティに限定する。

そしてコミュニティの成長を調べるために、成長率という指標を設定する。コミュニティのサイズが最初のスナップショットと比べて x%大きければ、成長率 x%とする。そして成長率 9%未満のコミュニティ集合 0 と 18%超とのコミュニティ集合 1 とを推定する問題を用意する。なお、成長率 9〜18%のコミュニティ集合は評価をはっきりとさせるために取り除く。残ったコミュニティは 13570 個で、コミュニティ集合 0 と 1 のサイズはほぼ 1:1 である。

前節同様にコミュニティの成長率を推定する決定木を作成し、成長の要因となる特徴を調べる。

4.2 評価

全コミュニティの成長率の平均は 18.6%、中央値は 12.7%であった。コミュニティ内の閉じたトライアドの割合と成長率の関係を調べたところ、閉じたトライアドの密度が高いとコミュニティの成長に対してネガティブな影響を与える

[1] Fringe とは、自分はそのコミュニティに属していないが友人/共著者がそこに属している人を指す

ことがわかった。成長率を推定する決定木を作成したところ、最上位のルールは「13 人以上の知り合いを持つ Fringe の割合」であった。(1)Fringe、(2) コミュニティのサイズ、(3) コミュニティのサイズに対する Fringe の割合、(4) これら 3 つの指標の組み合わせ、(5) 全指標、の 5 つの指標を用意してどの程度成長を推定できるかどうかを調べた。評価指標には ROC、適合度の平均値、クロスエントロピーを用いた。結果、全指標を用いたものがもっとも良かった。

4.3 知見

コミュニティの成長には、コミュニティのサイズやコミュニティの周辺の人の数などだけでなく、コミュニティ内のメンバー間の関係が影響を与えていることがわかった。閉じたトライアドの割合が増えると成長率が下がるのは、前節の友人が構成するトライアドが閉じている方が参加する割合が高くなるという知見と反する結果だが、これは閉じたトライアドが多すぎる＝派閥化が進んだことにより新規参入者が減るという現象であると考えられる。これについては現在調査中である。

5 コミュニティの変化

5.1 手法・アルゴリズム

データには DBLP を使う。データセットには 15 年の間に行われた 87 カンファレンスの論文が含まれている。そして同じ年に開催されたことのあるカンファレンス間の人・トピックの移動を調べる。

コミュニティ間の移動を計量するために 2 種類の指標を用意する。Term Bursts はカンファレンスの論文のタイトル中の単語の頻度から求める、その単語の盛り上がり度である。Movement Bursts は人の動きを示す指標である。y-1 年にカンファレンス B で論文があり、y 年にカンファレンス C で論文がある場合、その著者は B から C へ動いたと見なす。

5.2 評価

人の移動を伴った論文が、(1) 現在盛り上がっている単語を含む論文か、(2) 将来盛り上がる単語を含む論文か、(3) 過去に盛り上がった単語を含む論文かを調べた。結果、(1) は 40%、(2) は 10%、(3) は 30%程度であった。さらに二つのカンファレンス間での人の動きとトピックの動きとの関係を調べた。組み合わせとしては以下の四つがあり、結果は (a) が 60%、(b) が 10%、(c) が 10%、(d) が 20%程度であった。

a カンファレンス B と C とで w が Term Burst しているが、それは B と C との間で人の移動が起こる前だった場合、B と C とは興味を共有しているとみなす。

b B で w が Term Burst し、その後 B から C への人の移動が起こり、C で w の Term Burst が起こった場合、B から C への入植が起こったとみなす。

c C で w が Term Burst し、その後 B から C への人の移動が起こり、B で w の Term Burst が起こった場合、B から C への調査が起こったとみなす。

d B から C への人の移動が起こり、その後 B と C とで w の Term Burst が起こった場合、B と C とはメンバーを共有しているとみなす。

5.3 知見

多くの学会の組み合わせにおいて、関心が移植されるというよりは、人が移動する前にすでに興味が共有されているということがわかった。

なお、このようなコミュニティの変化については、年ごとのデータを LSI を用いて 2 次元上にマッピングさせその変化を見ることで、トピックにてグルーピングされたコミュニティの変化をよりわかりやすく表現できる可能性がある。

（文責：はまたろう）