M-006

文書クラスタリングを用いたコミュニティ抽出

Community extraction using text clustering

寺本やえみ† 隈井裕之† 宮田辰彦† 森本康嗣†
Yaemi Teramoto Hiroyuki Kumai Tatsuhiko Miyata Yasutsugu Morimoto

1. はじめに

知識労働における付加価値は、人と人とのコミュニケーションに基づく知識の交換、アイデアの創出によって生み出される。よって、組織内のコミュニケーションの状態を把握し更に活性化することは、今後の知識企業等の組織において重要な課題である。

人間関係のネットワークを分析する技術は、社会ネットワーク分析の分野で古くから議論されており、グラフ理論などの数学的な手法の適用や、社会ネットワークの特徴と生成モデルを議論したワッツ、バラバシらの研究を経て発展した。近年、メール、ブログ、SNS などの電子コミュニケーションツールの普及により、コミュニケーション履歴データを大量に蓄積し解析することが可能となり、現代社会のコミュニケーションネットワークを、社会ネットワーク分析を用いてどう捉えるかの議論が盛んとなっている。我々は、組織内のコミュニケーション状況把握において、継続的な知識交換・情報共有により形成されるコミュニケーションの「場」＝コミュニティが重要であると考える。

本報告では、文書データとして蓄積された人物間のコミュニケーションの履歴からコミュニティを抽出する方法の、提案・実装・実験・評価について述べる。

2．関連技術

人をノード、人物間の関係をノード間のリンクとしたネットワーク構造から、コミュニティを表すサブグラフを発見する従来技術を以下に挙げる。

(1) グラフ理論における n-Clique[2]

n-Clique は、n 本以内のリンクによって相互に結合するノードの集合と定義される。1-Clique は、全てのノード間が直接リンクで結ばれた完全グラフを形成する。

(2) NEGOPY 法[3]

距離の近い（少ないリンクで到達可能な）ノード同士が近くに集まるようにノードを一列に並べ、部分集合の凝集度に基づいて全体を複数のコミュニティに分割する手法である。

(3) スペクトラルグラフ理論に基づく SR 法[4]

平均リンク数が最大となる部分集合の探索問題を固有値問題として解き、コミュニティとして抽出されたリンクを削除して抽出を再帰的に行うことで、ノードの重複を許容したコミュニティを抽出する手法である。

3. 提案手法

3.1 アプローチ

従来技術は、ネットワーク構造の中でリンクが「密である」サブグラフをコミュニティとして抽出している点が共通している。一方我々は、コミュニティを抽出するための尺度は、リンクの「密さ」だけではないと考える。社会学者の G.A.ヒラリーは、著書「コミュニティの定義」(1955)の中で、多くのコミュニティ定義に共通する概念を、「社会的相互作用」、「共通の絆」、「地域」の 3 つとしている。この定義を基に、組織内のコミュニティにとって重要な特徴として、以下の２つを挙げる。

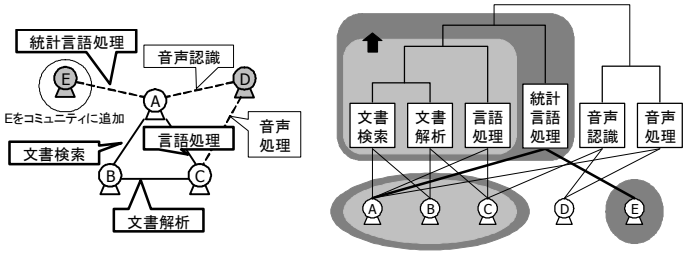
特徴１）相互のコミュニケーションが盛んである。

特徴２）共通の話題や関心を持つ。

従来技術は、上記特徴１のみを基準としたコミュニティ抽出方法であり、組織内のコミュニティの特徴を捉えきれていない。本報告では、上記特徴１および２を基準とし、リンクの密度に加えリンクの意味情報を用いてコミュニティを抽出する方法を提案する。

3.2 コミュニティ抽出方式

提案手法では、リンクが非常に密な、コミュニティの一部である可能性の高いサブグラフからスタートして、周囲の適切な人物を追加してコミュニティを拡張することでコミュニティを形成する。従来技術のアプローチでは、追加メンバの判定において既存のコミュニティと各メンバ間のリンクの数が問題になるのに対し、提案手法ではリンクの数に加え、リンクの意味の類似を考慮する点が特徴である。図 1 (a)において、A,B,C ３名がコミュニティメンバであり、追加する人物として D,E が候補であるとする。従来技術のアプローチでは、A,C 双方とリンクを持つ D が追加される。一方、提案手法では、A のみとリンクを持つがリンクの意味が A,B,C 相互のリンクに類似する E が追加される。リンクの意味の類似は、コミュニケーション文書を用いたリンクのクラスタリングによって表す。図 1 (a)の関係を、リンクのクラスタリングを用いて表したものを図 1 (b)に示す。A,B,C の３名からなるコミュニティに追加する人物を考慮する際、１段階広げたクラスタに含まれるリンク「統計言語処理」によって A とつながる E をコミュニティに追加する。



(a) (b)

図１　コミュニティ拡張方式

コミュニケーションにおける話題や関心は相手や状況に応じて変化するため、リンクの意味を考慮する際には、個々のコミュニケーションを区別する。クラスタリング

†（株）日立製作所　中央研究所

を行う際には、図２に示すように、１回のコミュニケーションの内容を表す文書をクラスタの要素とする。

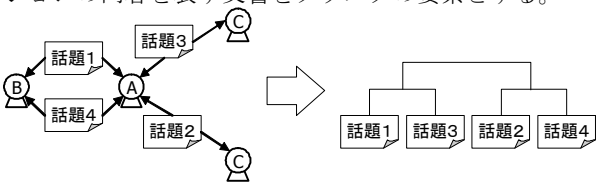


図２　リンクの意味の単位

3.3　アルゴリズムの詳細

具体的な処理の流れを以下に示す。

処理１：コミュニケーション文書をクラスタリングする。
処理２：ネットワークから取り出した 1-Clique を、最もメンバの重複する文書クラスタに対応づける。
処理３：文書クラスタと 1-Clique との重複を、初期コミュニティメンバとする。
処理４：文書クラスタに含まれる文書によって、コミュニティメンバと直接コミュニケーションを取った人物をコミュニティメンバに加える。
処理５：コミュニティ内のリンク密度が閾値以下ならば処理を終了する。
処理６：着目する文書クラスタを１段階広げ、処理４－６を繰り返す。

処理繰り返しの終了判定の閾値には、コミュニティのリンク密度を用いる。グラフ G のリンク密度 d(G)は次式で表される。

$$d(G)=m(G)/m(K_k) \qquad (1)$$

ただし、m(G)はグラフ G のリンク数、K_n は n ノードからなる完全グラフ、k はコミュニティのメンバ数（グラフ G のノード数）を表すものとする。

文書のクラスタリングには、文書間類似度の高い順に文書を纏め上げる階層的クラスタリング手法を用いる。クラスタ間の類似度の算出には、群平均法を用いる。群平均法では、クラスタＡ、Ｂ間の類似度を、クラスタＡに属する文書とクラスタＢに属する文書の間の類似度の平均値とする。

さらに、提案手法では、各コミュニティに対し、メンバを決定するコミュニケーション文書の集合が定まる。この文書集合によって、コミュニティの共通の関心や話題が表される。

4. コミュニティ精度評価

社内の報告書 3415 報を用い、共著がある２者間にはコミュニケーションが存在したものとみなして、筆頭者および連名者の共著関係ネットワークを作成した。報告書の要旨を、コミュニケーションの内容を表す文書とした。このネットワークを用い、提案手法によるコミュニティ抽出を行った。抽出されたコミュニティの構成メンバに対し、評価アンケートを行った。（対象：11 コミュニティ 22 名）

評価アンケートでは、コミュニティメンバとコミュニティの関心を表す 30 個のキーワードを被験者に提示して、コミュニティのメンバの過不足を指摘してもらい、正解であるコミュニティを作成した。提案手法、1-Clique、2-Clique の３手法によって抽出されたコミュニティと正解を比較することにより、漏れの少なさを表す再現率、ノイズの少なさを表す適合率、F 値を算出した。F 値は、一般的にトレードオフの関係にある再現率と適合率を同時に評価するための指標である。評価値の算出式を以下に示す。

再現率＝正判定メンバ数/正解コミュニティメンバ数
適合率＝正判定メンバ数/抽出結果コミュニティメンバ数
F 値＝2×再現率×適合率 / (再現率＋適合率)

評価結果を図３に示す。

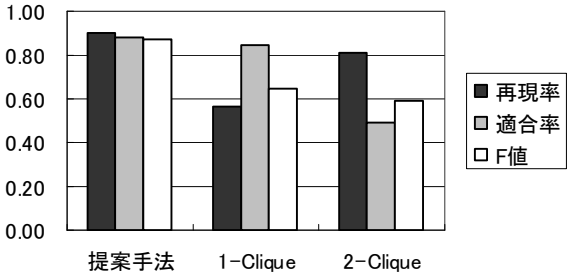


図３　コミュニティ評価結果

1-Clique は制約が強いためノイズは少ないが漏れが多く、2-Clique は制約の緩和が強く効いて漏れは少ないがノイズが多くなっている。一方、提案手法では、漏れもノイズも少なく抑えていることが確認できる。リンクの意味情報を用いることで、コミュニティ抽出精度が向上したと言える。

6. まとめと今後の課題

本報告では、文書クラスタリングを用いたコミュニティ抽出技術を開発し、コミュニティ抽出精度の評価を行った。また、報告者らは、本報告の提案技術を用いたコミュニティ検索機能を実装した、有識者検索システム[1]を開発した。

今後の課題を以下に挙げる。

(1)アルゴリズムのブラッシュアップ

本報告では、リンクの密な部分グラフの抽出に 1-Clique を用いたが、関連技術に述べた、より精緻な手法を検討するべきである。

(2)評価に用いる正解データの整備

本報告では、アンケートを実施して正解データを作成したが、アンケートにかかるコストが大きく、十分な規模で実施できたとは言えない。開発した有識者検索システムにコミュニティ評価アンケート機能を盛り込むなど、正解データを充実させる検討が必要である。

参考文献

[1] 宮田辰彦 他, 人間関係分析技術を活用した有識者検索システムの提案, 第 64 回グループウェアとネットワークサービス研究発表会, 信学技報, 2007
[2] Scott, J., Social Network Analysis A Handbook Second edition, SAGE Publications, 2000
[3] 金光淳 著, 社会ネットワーク分析の基礎, 勁草書房, 2003
[4] 斉藤和巳 他, SR:ネットワークの蜜結合するコア部抽出法, WEIN2005