SNS におけるコミュニティの関係抽出に関する研究

樫山武浩† 田中成典‡ 三善健太†

関西大学大学院† 関西大学総合情報学部‡

1. はじめに

近年，インターネットの普及に伴い，Web 上でコミュニケーションを行う機会が増加しており，知人関係を Web 上に形成し相互交流を行う SNS(Social Networking Service)が注目されている．しかし，SNS の普及により SNS 内のコミュニティ数が急増し，ユーザが興味をもつコミュニティの発見が困難になっている．そのため，コミュニティの分類が望まれている．

Web における情報の分類では，リンク関係を基にして同じ目的や内容の情報をグループ化して抽出する手法[1][2][3]が多く利用される．しかし，SNS では，コミュニティ間のリンクが少ないため，この手法を適用するためには，コミュニティの関係の有無を抽出する必要がある．コミュニティの関係の有無を抽出する方法として，両コミュニティの参加者の共起が利用した研究[4][5]がある．しかし，コミュニティの内容を考慮して関係を生成していないため，共通のテーマによる関連や内容の類似や話題の細分化などのさまざまな種類の関係が同一の関係と判断されている．これにより，複数のグループを生成できるコミュニティ群において，単一のグループだけしか生成できないことがある．

そこで，本研究では，コミュニティ間のリンクを自動生成し，コミュニティの内容と構成に基づいて生成したリンクの種類を判定することで従来手法よりも詳細なコミュニティの分類を目指す．

2. システムの概要

本研究では，コミュニティ間の関係の有無を示すリンクを自動生成し，コミュニティを分類することを目的とする．本システムの概要を図 1 に示す．本システムは，「リンク生成処理」「関係抽出処理」「可視化処理」で構成される．

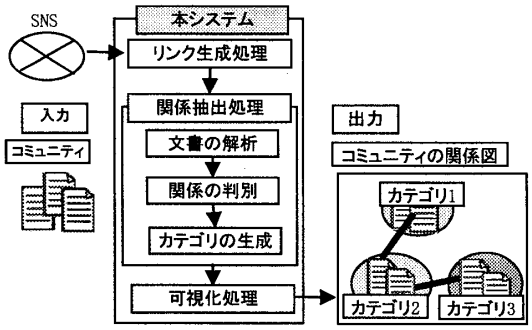

図 1　システムの概要

まず，「リンク生成処理」では，コミュニティの参加者の共起を基に関係の有無を判定する．既存研究では，コミュニティの共起の強さを測る指標として Jaccard 係数を利用している．Jaccard 係数は，規模の差が評価値に大きく影響し，同規模のコミュニティしか抽出できない．しかし，SNS では，異なる規模でも関連が強いコミュニティが多数存在する．そのため，本研究では，Cosine 係数と Simpson 係数をコミュニティの規模の比を基に切り替えて利用している．Simpson 係数では，コミュニティの規模の差の影響が少ない．しかし，規模の差が大きい場合や規模が小さい場合には，全体的に評価値が高くなる欠点がある．そのため，本システムでは，規模の差が大きいコミュニティに限定して利用し，閾値を高くすることで，この問題を解消している．次に，「関係抽出処理」では，リンク生成処理で抽出した関連性の種類をコミュニティの文書を解析することで判別し，グループ化する．文書の解析では，各タイトル，紹介文，トピック，コメントごとに名詞と未知語を抽出し，TF・IDF 法を利用して重要語を算出する．コミュニティの構成とコミュニティ間の重要語の共起関係の分布を基に上下関係のコミュニティ，同内容のコミュニティと連想関係にあるコミュニティの 3 種類の関係に分類する．分類した各リンクを分析することで，カテゴリを生成する．最後に，「可視化処理」では，抽出した関係とグループを描画する．

Research on Relation Visualize of Community in SNS
†Takehiro Kashiyama, Kenta Miyoshi
Graduate School of Informatics, Kansai University, 2-1-1 Ryouzenji-cho Takatsuki-shi, Osaka 569-1095, Japan
‡Shigenori Tanaka
Faculty of Informatics, Kansai University, 2-1-1 Ryouzenji-cho , Takatsuki-shi, Osaka, 569-1095, Japan

3. システムの実証実験と考察

本提案手法の有効性を実証するために，関連コミュニティの抽出精度とカテゴリの生成精度に対して，提案手法と既存手法との比較実験を行う．

3.1 実証実験

本実験では，2004 年 4 月 1 日から 2004 年 10 月 31 日までに作成された mixi のスポーツカテゴリの中で参加者数が 100 名以上のコミュニティ（893 件）を対象とした．関連コミュニティの抽出精度の検証では，10 倍までの規模のコミュニティを同規模とし，抽出リンク数と適合率を比較する．カテゴリの生成は，本手法で抽出したリンクに対して各手法を適応する．カテゴリの生成精度の検証では，抽出カテゴリ数，カテゴリに属するコミュニティの数の平均と適合率を比較する．精度評価における適合率は，抽出された結果を人為作業により正誤判定することで算出する．

3.2 結果と考察

関連コミュニティの抽出精度の評価結果を表 1，カテゴリの生成精度の評価結果を表 2 に示す．

表 1 の結果から，Jaccard 係数よりも Cosine 係数が高い水準の結果が得られることがわかる．また，Cosine 係数と本手法の結果からコミュニティの規模の比に応じて指標を切り替える方法が有効であることがわかる．

表 2 の結果から，既存手法と比較して提案手法が次の 2 点で優れていることがわかる．1 つ目は，カテゴリの正答率の向上である．既存手法の適合率が 0.68 に対して，提案手法では，0.82 であり，カテゴリの正答率が向上していることがわかる．2 つ目は，カテゴリの詳細化である．既存手法では，カテゴリ数が 18 個，カテゴリに所属するコミュニティ数の平均が 14 個に対して，提案手法では，カテゴリ数が 34 個，カテゴリに所属するコミュニティ数の平均が 8 個になっている．このことから，提案手法が従来手法よりも詳細なカテゴリの生成ができたことがわかる．

表 1　関連コミュニティの抽出精度

評価項目	規模	リンク数	適合率
Jaccard 係数	同規模	8,205	0.88
	全体	8,619	0.91
Cosine 係数	同規模	14,108	0.92
	全体	16,505	0.94
Simpson 係数	同規模	4,272	0.87
	全体	7,968	0.82
本手法	同規模	14,108	0.93
	全体	19,126	0.95

表 2　カテゴリの生成精度

評価項目	カテゴリ数	平均規模	適合率
既存手法	24	29	0.73
提案手法	45	16	0.81

4. おわりに

本研究では，コミュニティの共通参加者に基づいて生成したリンクの種類を考慮してグループ化することで，より詳細なコミュニティの関係を抽出した．実証実験の結果より，提案手法の有効性を示すことができた．本研究では，「自己紹介」や「マイミク募集」などのコミュニティの内容と関連の薄いトピックとイベントの告知や試合の実況応援トピックなどの一過性のトピックを除外するために数個のキーワードを設定した．しかし，設定したキーワード以外にも多数除外した方が好ましいキーワードが存在した．そのため，除外した方が良いトピックに共通するキーワードを抽出により辞書を作成し，解析対象の文書を精査することで，より高精度なコミュニティの関係抽出の実現を目指す．

参考文献

[1] 野村早恵子，小山聡，早水哲雄，石田亨：Web コミュニティ発見のための HITS アルゴリズムの分析と改善，電子情報通信学会論文誌，電子情報通信学会，Vol.J85-D-1，No.8，pp.741-750，2002.8.

[2] 加藤一民，松尾啓志：Markov Cluster Algorithm を用いた Web コミュニティ群の発見手法，情報処理学会自然言語処理研究会研究報告，情報処理学会，Vol.2005，No.22，pp.87-93，2005.3.

[3] Jon M. Kleinberg : Authoritative Sources in A Hyperlinked Environment , Journal of Assoc Comput Mach, Assoc Comput Mach, Vol.46, No.5, pp.604-632, 1999.9

[4] 松尾豊，安田雪：SNS における関係形成原理，人工知能学会論文誌，人工知能学会，Vol.22，No.5，pp.531-541，2007.9.

[5] Elen Spertus , Mehran Sahami , Orkut Buyukkokten : Evaluating Similarity Measures:A Large-Scale Study in the Orkut Social Network, Proceeding of the eleventh Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining international conference on Knowledge discovery in data mining , Association for Computing Machinery, pp.678-684, 2005.8.