ソーシャルネットワークにおける リンク構造を用いた重複クラスタリング手法の提案

金川　元信† 　大豆生田利章††

† 群馬工業高等専門学校専攻科　〒 371-8530 群馬県前橋市鳥羽町 580 番地
†† 群馬工業高等専門学校　〒 371-8530 群馬県前橋市鳥羽町 580 番地
E-mail: †ap09807@ipc.gunma-ct.ac.jp, ††mame@ice.gunma-ct.ac.jp

あらまし　ソーシャルネットワーキングサービス (SNS) におけるユーザは, 趣味や職業などの特徴によって複数のクラスタに属しうる. したがって, SNS のユーザをクラスタリングする際には, 一つのデータに対して複数のクラスタに所属することを許容する重複クラスタリング手法を適用することが有効であると考えられる. しかし, 既存の重複クラスタリング手法は個々のデータの独立した特徴のみに基づくため, ユーザ間のリンク構造のみに現れるユーザの特徴をとらえることができない. 本研究では, ユーザの特徴のみならずユーザ間のリンク構造を用いることで, 従来の手法における問題を解決する重複クラスタリング手法を提案する. 提案手法は, リンク元ユーザがリンク先ユーザに対して抱く「関心」を反映するようにリンクそのものに特徴ベクトルを定義し, 定義されたリンクの特徴ベクトルを用いて重複クラスタリングを実現する. 提案手法と既存手法を Twitter から作成した複数のデータセットに対して適用して比較実験を行うことにより, 提案手法の妥当性を示した.

キーワード　重複クラスタリング, リンクマイニング, ソーシャルネットワーク, Twitter

An Overlapping Clustering Method for Social Networks Using Link Structure

Motonobu KANAGAWA† and Toshiaki OHMAMEUDA††

† Advanced Engineering Course, Gunma National College of Technology
580 Toriba, Maebashi, Gunma, 371-8530 Japan
†† Gunma National College of Technology
580 Toriba, Maebashi, Gunma, 371-8530 Japan
E-mail: †ap09807@ipc.gunma-ct.ac.jp, ††mame@ice.gunma-ct.ac.jp

1. はじめに

Twitter(注1)や facebook(注2)などに代表されるソーシャルネットワーキングサービス (SNS) のユーザは, 職業や趣味などの特徴によって複数のクラスタに属しうる. 例えば, 趣味がアニメ鑑賞で職業は研究者であるユーザは「アニメファン」というクラスタと「研究者」というクラスタに所属するべきである. したがって, SNS のユーザをクラスタリングする場合は, ユーザに一つ以上のクラスタに属すことを許容するクラスタリング手法が有効であると考えられる. 一般に, 個々のデータ点に一つ以上のクラスタに属すことを許容するクラスタリング手法を重複クラスタリング (Overlapping Clustering) と呼ぶ.

重複クラスタリングを実現するための方法として, fuzzy c-means 法 [9] や EM アルゴリズム [8] 等のソフトクラスタリング手法の応用を考えることができる. ソフトクラスタリング手法はデータのクラスタへの帰属度を求めるため, 帰属度に対して一定の閾値を設定することにより個々のデータに対して複数のクラスタを割り当てることが可能である. しかし, この方法はもともとハードクラスタリングの枠組みでの重複クラスタリングを目的として設計されたわけではないため, 適切な閾値の基準が明確でないといった問題が存在する. これに対して, 重

(注1)：http://twitter.com
(注2)：http://www.facebook.com

複クラスタリングをハードクラスタリングの枠組みで行うための手法として Banerjee らによって Model-based Overlapping Clustering(MOC) [1] が提案されている．Banerjee らは MOC と閾値を設定したソフトクラスタリング手法を文書データに適用することで, MOC が重複クラスタリングを実行するうえでソフトクラスタリング手法よりも有効な手法であることを示した．したがって, 既存の重複クラスタリングを行うための手法としては, MOC が最も有望な手法であるといえる．

MOC は個々のデータの独立した特徴に基づく重複クラスタリング手法である．したがって, MOC を SNS のユーザの重複クラスタリングに用いるためには, 個々のユーザをデータ点とみなし, 発言などの特徴を用いて特徴ベクトルを定義すればよい．しかし, このようにして MOC を適用する場合, ユーザの独立した特徴には現れず, ユーザ間のリンク構造のみに現れるユーザの特徴をとらえることができない．例えば, 趣味がアニメ鑑賞で職業が研究者であるユーザがアニメに関する発言しかしない場合, ユーザの特徴ベクトルにはアニメに関する特徴は反映されるが研究者としての特徴は反映されない．したがって, MOC をこのユーザを含むユーザ集合に適用した場合, このユーザは「アニメファン」というクラスタにのみ所属することになる．しかし, このユーザとリンクによって関係づけられる別のユーザの中には, 同僚の研究者といった, このユーザの「研究者」としての特徴を示唆するユーザが存在するはずである．したがって, リンク構造を用いてこれらの同僚の研究者ユーザの特徴を用いることができれば, このユーザを「アニメファン」だけでなく「研究者」というクラスタに所属させることができる．

本研究では, 以上の MOC を SNS のユーザの重複クラスタリングに適用する際に起こる問題を解決するために, ユーザの特徴ベクトルとユーザ間のリンク構造を組み合わせた重複クラスタリング手法を提案する．具体的にはまず, リンク元ユーザがリンク先ユーザに対して抱く「関心」を反映するように, リンクそのものに対して特徴ベクトルを定義する．これにより, リンク先ユーザの一つの側面を表す特徴をリンクによって表すことができる．次に, 定義された特徴ベクトルを用いてユーザ集合中の全てのリンクに対して K-means 法 [10] を実行する．得られたクラスタは各ユーザの特徴のうちの一つの側面が要素であり, このクラスタに対して簡単な後処理を行うことによって重複クラスタリングを実現する．

本研究では, 提案手法の有効性を検証するために, Twitter から作成した複数のデータセットに対して実験を行い, 提案手法と MOC の性能を比較した．その結果, 提案手法は Twitter 上のユーザの重複クラスタリングを行う上で妥当な方法であることが示された．

本稿の構成は次のとおりである．2 節で提案手法の詳細を説明する．3 節で実験方法と実験結果について述べる．4 節で関連研究を紹介する．5 節で本研究のまとめと今後の課題を述べる．

2. 提 案 手 法

提案手法の対象とするデータは, ユーザに特徴ベクトルが定義でき, ユーザ間にリンクが存在する SNS 上のユーザ集合である．提案手法はリンクが有向である SNS, 特に Twitter を念頭に置いているが, facebook の「友達関係」といった無向リンクの SNS であっても, 無向リンクを相互リンクと考えることによって同様に提案手法を適用することができる．

以下で用いる用語を次のように定義する．u_i $(i = 1, \ldots, N) \in U$ をユーザ, U をユーザ集合, N をユーザ数, $\phi(u_i)$ を u_i の特徴ベクトルとする．$l_{ij} \in L$ を u_i から u_j へのリンク, L をリンク集合, $\Gamma(u_i)$ を u_i からリンクされているユーザ集合, $\tilde{\Gamma}(u_i)$ を u_i に対するリンクを有するユーザ集合とする．u_i から u_j へのリンクの特徴ベクトルを $\boldsymbol{\phi}(l_{ij})$ と書く．

2.1 提案手法の考え方

図 1 において, ユーザ a はアニメに関する発言しかしない工学の研究者であるとする．a は発言内容のみに着目すればクラスタリングの結果「アニメファン」というクラスタにのみ所属することになるが, 実際には「アニメファン」と「研究者」という二つのクラスタに所属するべきである．a の「研究者」という特徴をとらえるためには, a に向けられるリンクの違いを考慮にいれる必要がある．例えば, 研究者であるユーザ b が a に対してリンクを行うのは, a が「研究者」であるためであり「アニメファン」であるからではない．一方, アニメファンであるユーザ c が a に対してリンクを行うのは, a が「アニメファン」であるためであり「研究者」であるからではない．このリンクの違いは, リンク元のユーザがどのような「関心」をもってリンクを行っているかに起因する．

提案手法は, リンクをリンク元のユーザの「関心」を表すものであるととらえ, ユーザの「関心」を表現するようにリンクそのものに対して特徴ベクトルを定義する．そして, ユーザに対して抱かれる「関心」は, そのユーザの一つの側面を表す特徴であると考えられる．例えば, 図 1 における a は, b からは「研究者」としての「関心」を, c からは「アニメファン」としての「関心」を抱かれている．この「研究者」と「アニメファン」はそれぞれ a の一つの側面を表す特徴である．図 1 中の他のユーザ間におけるそれぞれのリンクについても, 同様にリンク先ユーザの一つの側面を表す特徴に対応している．したがって, ユーザ集合中の全てのリンクについて, その特徴ベクトルに基づいて K-means 法等の通常の分割クラスタリング手法を適用すれば, 得られるクラスタは各ユーザそれぞれの一つの側面を表す特徴ベクトル（リンクの特徴ベクトル）によって構成される．したがって, 各ユーザは最大で被リンク数と同じだけ異なるクラスタに属すことが可能となる．ただし, 同一ユーザに対する複数のリンクの特徴ベクトルが同一のクラスタに所属すれば, それらが表すユーザの側面は同じ種類のものであると解釈する．リンクの特徴ベクトルに対するクラスタリングについては, 2.5 で詳しく述べる．

2.2 共有リンク先ユーザ集合によるリンク元ユーザの「関心」の表現

リンク元のユーザがリンク先のユーザに対して抱く「関心」は, 両者の共有するリンク先のユーザによって表されると考えられる．なお, 二人のユーザの共有するリンク先のユーザの集合を共有リンク先ユーザ集合と呼ぶことにする．例えば, 図 2

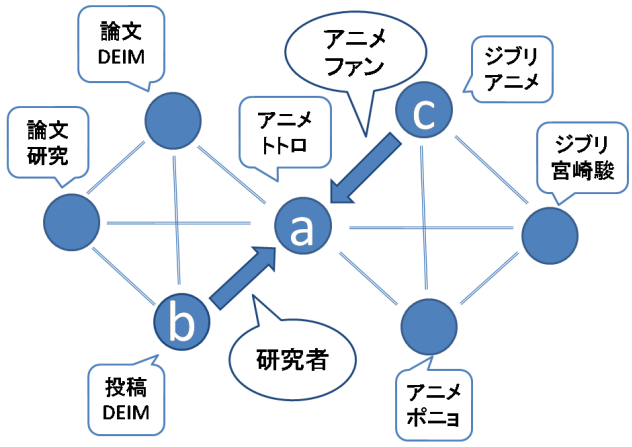

図 1　リンクによるリンク元ユーザの「関心」の表現．b は a を「研究者」という「関心」をもって，c は a を「アニメファン」という「関心」をもってリンクしている．これらの異なる「関心」は，それぞれ a の一つの側面を表す特徴であると考えることができる．

では，a は研究者であるため a のリンク先には研究者が存在し，研究者である b と a の共有リンク先ユーザ集合は研究者で構成される．これは b が a に対してもつ研究者としての「関心」を反映している．また，a はアニメファンであるため，アニメ情報を発信するユーザに対するリンクを有する．したがって，別のアニメファンである c と a の共有リンク先ユーザ集合はアニメに関係するユーザで構成され，これは c が a に対してもつアニメファンとしての「関心」を反映している．

一般に，ユーザ u_i とユーザ u_j の共有リンク先ユーザ集合は $\Gamma(u_i) \cap \Gamma(u_j)$ と書ける．以後，u_i から u_j リンクの特徴ベクトル $\phi(l_{ij})$ を $\Gamma(u_i) \cap \Gamma(u_j)$ を用いて定義することを目標として議論を進める．

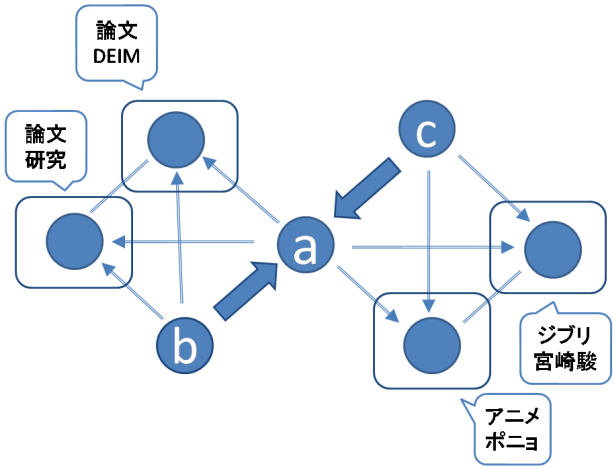

図 2　共有リンク先ユーザ集合による，リンク元ユーザのリンク先ユーザに対して抱く「関心」の表現．b から a の「関心」は両者の共有リンク先ユーザである研究者ユーザによって，c から a の「関心」は両者の共有リンク先ユーザであるアニメ関係ユーザに表される．

ただし，単に共有リンク先ユーザ集合によってリンク元ユーザの「関心」を表そうとすると，リンク先ユーザ自体の特徴が反映されなくなってしまう．また，図 3 のように，リンク元ユーザとリンク先ユーザが共有するリンク先を持たない場合もありうる．

これらの問題を解決するために，提案手法では一般に各ユーザは自分自身に対するリンクを有していると仮想的に考えることにする．これはユーザは各々が自分自身に対して「関心」を抱いていると考えることに相当する．これにより，共有リンク先ユーザ集合には常にリンク先ユーザが存在することになり，その結果としてリンク先ユーザの特徴がリンクの特徴ベクトルに反映されるようになる．図 3 の場合では，b から a のリンクの特徴ベクトルはただひとつの共有リンク先ユーザ a の特徴によって構成される．なお，このような状況は a が有名人であり b が a のファンであるような場合に頻繁に起こる．

また，図 4 のように，b から a へのリンクのみならず a から b へのリンクが存在する場合，つまり相互リンクが存在する場合，a と b の共有リンク先ノード集合には a と b の両方が存在する．したがって，a から b のリンクおよび b から a のリンクの特徴ベクトルは，a, b, c, d から計算される同一のものとなる．このように，一般に相互リンクの特徴ベクトルは向きにかかわらず同一の特徴ベクトルが定義されることになる．このような状況は，a と b が友人同士であったり同僚同士である場合に頻繁に起こり，そのような場合には両者が互いに抱く「関心」は両者の所属するコミュニティの特徴を反映したものとなる．

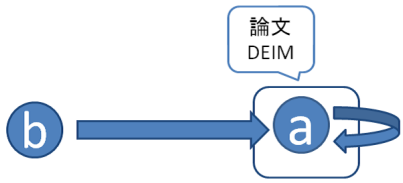

図 3　共有するリンク先が存在しない場合．一般に各ユーザは自身に対するリンクを有していると仮想的に考える．その結果，b と a の共有リンク先ユーザ集合は，a のみにより構成される．これは，b が a に対して抱く「関心」が a の特徴のみに依存していることを示している．

2.3　共有リンク先ユーザ集合の重み

図 5 においてユーザ f は有名な政治家であるとする．f は知名度が高いので多くの人からリンクされており，a と b の両方からもリンクされている．したがって，a と b の共有リンク先ユーザ集合には研究者だけでなく政治家も存在することになる．しかし，b が a に対して抱く「関心」は政治に関するものではない．このように，共有リンク先ユーザ集合中に存在するユーザの全てが等しい重みをもってリンクの特徴ベクトルの計算に寄与すると，ユーザの「関心」を正確に反映できなくなる可能性がある．

このことに対処するために，共有リンク先ユーザ集合中のユーザの重みをその被リンク数を用いて決めることにする．共有リ

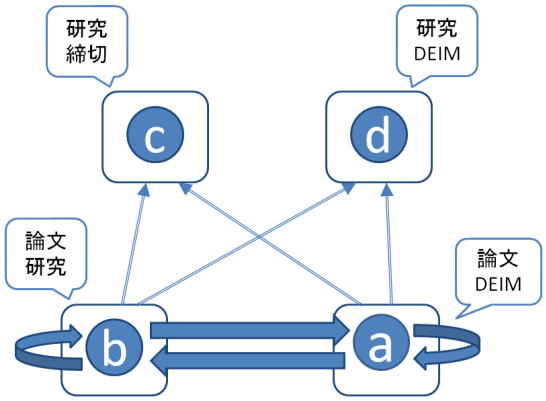

図 4　相互リンクが存在する場合．どちらの向きのリンクも同一の共有リンク先ユーザ集合によって特徴ベクトルが定義される．a と b の共有リンク先ユーザ集合は a, b, c, d により構成され，a から b へのリンクと b から a へのリンクには同一の特徴ベクトルが定義される．

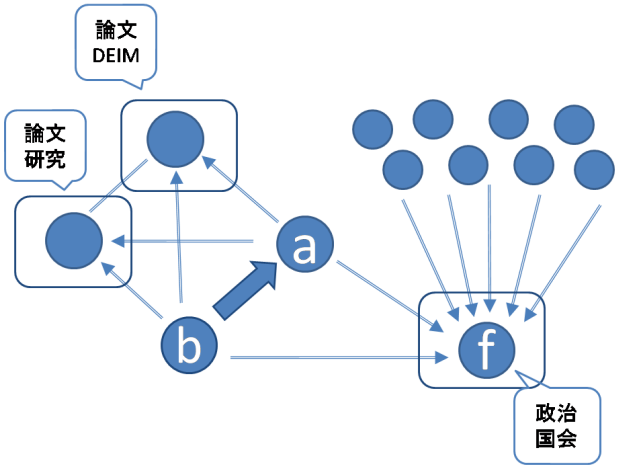

図 5　著名人ユーザは共有リンク先ユーザ集合中に現れやすい．b と a の共有リンク先ユーザとして有名政治家 f が存在するが，b から a への「関心」は政治とはあまり関係がない．b から a への「関心」を強く反映している共有リンク先ユーザは被リンク数の少ない研究者ユーザである．

ンク先ユーザ集合の中で，被リンク数の少ないユーザは被リンク数の多いユーザよりも忠実にユーザの「関心」を表していると考えられる．例えば，図 5 において，b が a に対して抱く「関心」は共有リンク先ユーザ集合中の被リンク数の少ない研究者ユーザが表している．一般に被リンク数の少ないユーザに対するリンクはユーザの強い「関心」を表しており，被リンク数の多いユーザに対するリンクはユーザの弱い「関心」を表していると考えられる．したがって，共有リンク先ユーザの重みは，被リンク数が大きければ小さくなり，被リンク数が小さければ大きくなるべきである．以上の議論より，共有リンク先ユーザ u_k の重みを $1/\log|\tilde{\Gamma}(u_k)|$ と定義し，被リンク数の対数に反比例するようにする．なお，この重みづけは Adamic らによるリンク予測手法における共有隣接ノード指標 [6] の考え方（被リンク数の少ない共有隣接ノードを多く持つノードのペアの間にはリンクが存在する可能性が高い）を取り入れたものである．

重みをここで述べたように定義することによって，発言には表れないユーザの一つの側面を表す特徴を明らかにすることができる．例えば，図 6 において，有名な政治家 f は元研究者であり研究者 a と研究者 b のかつての同僚であるとしよう．f は a と b に対してリンクを行っているため，b と f の共有リンク先ユーザ集合は a と b と f で構成される．b から f のリンクの特徴ベクトルは，被リンクが少ないために大きな重みをもつ a と b によって代表され，被リンク数の多い f からの寄与は少ない．このことから，b から f への「関心」は a と b の特徴によって代表される「研究者」としてのものであるとわかる．したがって，f は発言から推定される「政治家」という特徴だけでなく「研究者」という特徴も持つことがリンクの特徴ベクトルによって明らかになる．これにより，f は「政治家」というクラスタのみならず「研究者」というクラスタにも所属することが可能となる．

なお，図 5 のように，b が f に対して一方的に「関心」を寄せている場合は，b から f のリンクの特徴ベクトルはただ一つの共有リンク先ユーザ f の「政治家」としての特徴から構成される．これは b が f に対して「政治家」としての「関心」を抱いていることを表現している．

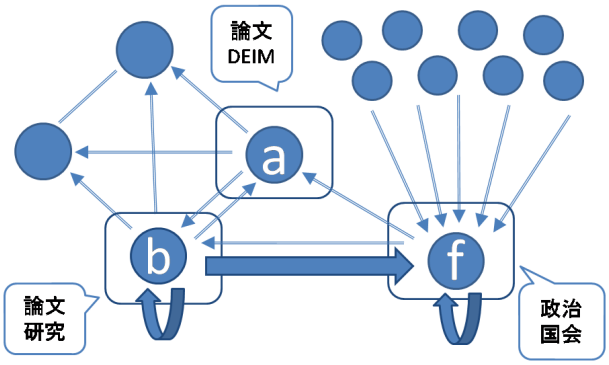

図 6　f は元研究者の政治家であり，a と b はその元同僚である．b と f の共有リンク先ユーザ集合は a, b, f によって構成されるが，特に被リンク数の少ない a, b の「研究者」としての特徴が b から f のリンクの特徴ベクトルに強く反映される．これは，f の発言内容だけからは推定されない「研究者」としての特徴を表している．

2.4　リンクの特徴ベクトル

以上の議論に基づき，ユーザ u_i からユーザ u_j へのリンクの特徴ベクトル $\boldsymbol{\phi}(l_{ij})$ を，u_i と u_j の共有リンク先ユーザ集合 $\Gamma(u_i)\cap\Gamma(u_j)$ を用いて式 (1) のように定義する（なお，2.2 で述べたように，各ユーザは自分自身に対してリンクを有しているとする）．

$$\boldsymbol{\phi}(l_{ij}) = \frac{\sum_{u_k\in\Gamma(u_i)\cap\Gamma(u_j)} \phi(u_k)/\log|\tilde{\Gamma}(u_k)|}{\sum_{u_k\in\Gamma(u_i)\cap\Gamma(u_j)} 1/\log|\tilde{\Gamma}(u_k)|} \qquad (1)$$

分母は共有リンク先ユーザ集合中のユーザの重みの総和であり，

$\boldsymbol{\phi}(l_{ij})$ は共有リンク先ユーザの特徴ベクトルの重みつき平均である．被リンク数 $|\tilde{\Gamma}(u_k)|$ を用いて共有リンク先ユーザの特徴ベクトルの重みを $1/\log|\tilde{\Gamma}(u_k)|$ と定義することにより，被リンク数の少ない共有リンク先ユーザの特徴がリンクの特徴ベクトルに強く反映される．

2.5 重複クラスタリングアルゴリズム

リンクの特徴ベクトルを用いて重複クラスタリングを実現する方法を，図 7 を用いて説明する．図 7 は，リンクの特徴ベクトルを計算した後の状態を示しており，色が同じリンクは類似した特徴ベクトルを持つとする．

2.1 で述べたように，リンクの特徴ベクトルが表現するリンク元ユーザの「関心」は，リンク先ユーザの一つの側面を表す特徴であると考えられる．与えられたユーザ集合中の全てのリンクについて特徴ベクトルを計算すれば，それぞれのユーザに対して異なる側面を表す複数の特徴ベクトルが定義されることになる．例えば，図 7 では，a については b, d, f からのリンクが a の「研究者」としての特徴を表し，c, h からのリンクが a の「アニメファン」としての特徴を表す．また，f については a, d からのリンクが f の「研究者」としての特徴を表し，e からのリンクが f の「政治家」としての特徴を表す．図中の他のユーザにも同様に異なる側面を表す複数の特徴ベクトルが計算される．

したがって，ユーザ集合中のリンク全てに対して K-means 法等の通常の分割クラスタリング手法を適用してクラスタリングすれば，各ユーザそれぞれの一つの側面を表す特徴ベクトル（リンクの特徴ベクトル）によって構成されるクラスタが得られる．図 7 では，同じ色のリンクが同一のクラスタに割り当てられる．「研究者」の特徴を有するリンクは一つのクラスタを構成し，その結果としてこれらのリンク先のユーザ a, b, d, f が「研究者」クラスタを構成する（なお，例えば a をリンク先として持つリンクは複数存在するが，これらは a の同じ側面を表していると考えれるので，同一のものとみなす．クラスタ内の他のリンクについても，リンク先が同じであれば同一のものとみなす）．「アニメ」の特徴を有するリンクも一つのクラスタを構成し，リンク先ユーザ a, c, h, k は「アニメファン」クラスタを構成する．したがって，a は「研究者」と「アニメファン」の 2 つのクラスタに所属する．同様に，f は e から「政治家」の特徴をもつリンクを受けているので，「政治家」クラスタに所属する．したがって，f は「研究者」と「政治家」の 2 つのクラスタに所属することになる．他のユーザも同様にして被リンクの特徴によって一つ以上のクラスタに所属する．

以上の議論をまとめると，提案手法は Algorithm 1 の手順で重複クラスタリングを実現する．

なお，提案手法で用いる分割クラスタリング手法として K-means 法を用いる理由は，K-means 法が高速で実行できるためである．K-means 法は繰り返し回数を定数と考えると，データ数に対して線形の計算量で実行することができる [2]．したがって，提案手法の計算量は K-means 法を用いることによってリンク数に対して線形となる．

Algorithm 1 リンク特徴ベクトルによる重複クラスタリング

1: 与えられたユーザ集合におけるユーザ間の全てのリンク $l_{ij} \in L$ について，リンクの特徴ベクトル $\boldsymbol{\phi}(l_{ij})$ を式 (1) にしたがって計算する．
2: $\{\phi(l_{ij}) \mid l_{ij} \in L\}$ に対して K-means 法を実行する．なお，クラスタ数 K は人手で与える．
3: 得られた K 個のクラスタ $C_k (k = 1, \ldots, \mathrm{K})$ について，$\phi(l_{ij}) \in C_k$ を u_j で置き換える．C_k 内で同一ユーザが複数存在すれば一つにまとめる．

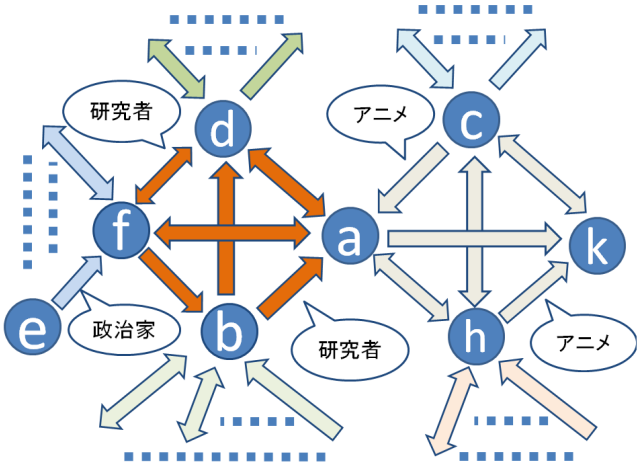

図 7　提案手法のアルゴリズム

3. 実　　験

ユーザに特徴ベクトルを定義することができ，リンク構造を有する重複クラスタリング用のベンチマークデータセットは存在しないため，Twitter から実験用のデータセットを作成することにする．Twitter はユーザ間にフォローと呼ばれる有向リンクが存在し，個々のユーザがツイートと呼ばれる発言を自らのページ上に書き込む SNS である．

3.1 実験用データセットの作成手順

実験で用いるデータセットは Twitter のユーザに対するタグ付けを行うサイト Wefollow(注3)のタグを用いて作成する．具体的には，ベースとなるタグをいくつか決め（例えば「celebrity」，「music」，「socialmedia」，「entrepreneur」），それぞれの上位 25 のユーザを Twitter からダウンロードする．ダウンロードされた各々のユーザに付されたベースのタグとその他のタグを集計し，3 人以上が共有するタグをクラスタとみなし，タグを共有するユーザがそのタグのクラスタを構成するものとする．Wefollow では一人のユーザに対して一つ以上のタグを付しているので，このように作成された正解クラスタはユーザが一つ以上のクラスタに属する重複クラスタとなっている．

ユーザの特徴としてユーザの最新 150 件のツイート（発言）を用いる．発言中の単語を Bag of Words としてベクトル空間モデルを構成し，tf-idf を用いて特徴ベクトルの各次元 (単語) の重みづけを行う．特徴ベクトルを計算するうえで用いる単語

(注3)：http://wefollow.com

は, ノイズとなりうる「in」や「this」といったストップワードや「@」や「#」で始まる文字列（前者は返信先ユーザ名, 後者はハッシュタグ）, および URL を除去したものを用いる. また, リンクとして用いる情報はフォロー関係のみであり, 「@」による返信のリンクは用いない. なお, リンクの特徴ベクトルを計算するうえで用いる共有リンク先ユーザ集合は, ダウンロードされたデータセット内に存在するユーザのみを用いる.

tf-idf は, 式 (2) のように発言中に現れる単語を重みづけして各単語を特徴ベクトルの一つの次元に対応させる手法である.

$$\phi(u_k)_d = \frac{n_w(u_k)}{n(u_k)} \times \log \frac{N}{DF_w} \tag{2}$$

ただし, $\phi(u_k)_d$ はユーザ u_k の特徴ベクトルの単語 w に対応する次元の要素, $n_w(u_k)$ は u_k の発言中に単語 w が現れた回数, $n(u_k)$ は u_k の発言中の総単語数である. N はデータセット中の総ユーザ数, DF_w は単語 w が発言中に現れたデータセット中のユーザ数である.

3.2 作成した実験用データセットの説明

以上の方法で表 1 に示す 8 つのデータを作成した. なお, データ 5 からデータ 8 は各々データ 1 からデータ 4 で用いたベースのタグと同じタグを 2 つ用いたものとなっており, ほぼ同一のユーザ構成からなるより小規模なデータセットを構成するために作成した.

データ 1, データ 2, データ 5, データ 6 は, データ作成における恣意性を排除するために Wefollow のタグのランキング付けで最も高いランクのタグを用いている. このランキング付けはタグを有するユーザのフォロワー数の総計の多寡によって決まるため, これらのデータセットは結果的に, 俳優, 歌手, スポーツ選手, 企業家, 作家と言った著名人およびニュースサイトのアカウント等で構成されている. そのため, これらのユーザは一般人とは異なるリンク構造と特徴（発言内容）を有すると考えれる.

データ 3, データ 4, データ 7, データ 8 は, 一般人ユーザに近いデータセットを構成する目的で作成した. データ 3 とデータ 7 は有名 IT 企業のタグをベースとしている. これらのデータセット中には, IT 企業の情報発信アカウントも含まれているが, 基本的にはそれらの IT 企業にかかわるユーザによって構成されている. データ 4 とデータ 8 は情報工学における隣接する分野のタグをベースとしており, これらのデータセットを構成するユーザはそれらの情報工学分野に携わる人々である. データ 3, データ 4, データ 7, データ 8 を構成するユーザはデータ 1, データ 2, データ 5, データ 6 を構成する著名人ユーザと比べ知名度が低く, より一般人ユーザに近いリンク構造と特徴（発言内容）を有すると考えられる.

3.3 実験方法

提案手法と Banerjee らによる Model-based Overlapping Clustering(MOC) [1] との比較を行う. MOC は Bregman divergence [7] を用いた重複クラスタリング手法である. 先に述べたように, MOC はデータの特徴ベクトルのみに基づく手法であるため, 作成したデータセットにおけるリンク情報は用いない. また, 本研究では MOC の Bregman divergence としてユークリッド距離を用いる.

重複クラスタリンクの性能評価には, Banerjee ら [1] にしたがってクラスタリングの標準的な評価尺度である精度, 再現率, F 値 [2] を用いる. 精度, 再現率, F 値は以下のように定義される. なお, F 値は精度と再現率の調和平均である.

精度 = 同一の出力クラスタに属するデータのペアのうち, 同一の正解クラスタに属するペアの割合

再現率 = 同一の正解クラスタに属するデータのペアのうち, 同一の出力クラスタに属するペアの割合

$$\text{F 値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}}$$

なお, クラスタ数は表 1 にしたがって各データセットについてあらかじめ与えられるものとする.

3.4 実験結果

各データセットについて提案手法と MOC をそれぞれ 5 回ずつ適用し, F 値が最高値を記録した時の結果を表 2 に示す.

精度に関して全てのデータについて提案手法が MOC を上回っている. これは提案手法の方が MOC よりも, 個別の出力クラスタにおける正確性が高いことを示している. 特に, データ 3, データ 4, データ 7 については 30 ポイント以上提案手法の方が高いが, これらは全て一般人ユーザに近いデータセットである. したがって, 提案手法は一般人ユーザに適用した際に個別の出力クラスタにおいてより高い正確性を発揮することがわかる.

一方, 再現率に関してはデータ 2, データ 3, データ 4, データ 6, データ 8 について MOC の方が提案手法よりも高い. これは MOC が提案手法よりも, 一人のユーザ当たりに多くのクラスタを割り振っていることに起因すると考えられる. そのため, MOC の再現率は高い傾向にあるが, それにともなって精度は低下している. この傾向は特にデータ 2, データ 3, データ 4 において顕著であり, これはデータ 2, データ 3, データ 4 のクラスタ数が他のデータセットよりも特に高いことに起因していると考えられる. その証拠に, 同一の構成法（4 つのベースのタグ）で作成したデータ 1 からデータ 4 については, クラスタ数の最も少ないデータ 1 について MOC の再現率は提案手法よりも低く, 2 つのベースのタグから作成したデータ 5 からデータ 8 については, クラスタ数の少ないデータ 5 とデータ 7 について MOC の再現率は提案手法よりも低い. このことより翻って, 提案手法はクラスタ数が比較的小さい場合に高い再現率を与えることがわかる.

F 値に関してはデータ 1, データ 3, データ 4, データ 5, データ 7 について提案手法の方が MOC よりも高い. また, データ 8 については提案手法は MOC とほぼ同等の値であるといえる. したがって, 一般人ユーザに近いデータセットであるデータ 3, データ 4, データ 7, データ 8 について提案手法は MOC と同等以上の性能を有するといえる. これは, F 値は精度と再現率の調和平均であるため, 提案手法の高い精度が MOC の高い再現率よりも F 値に対して高い寄与を与えた結果である. また, デー

表 1　実験に用いるデータセット (正解クラスタ)

	ベースのタグ	ユーザ数 (N)	リンク数	クラスタ数 (K)
データ 1	celebrity, music, socialmedia, entrepreneur	77	665	16
データ 2	news, blogger, tech, tv	86	884	23
データ 3	google, yahoo, microsoft, apple	99	695	20
データ 4	datamining, machinelearning, nlp, bioinformatics	95	901	22
データ 5	celebrity, music	40	203	6
データ 6	news, blogger	48	277	13
データ 7	google, yahoo	49	364	9
データ 8	datamining, machinelearning	48	386	11

表 2　実験結果 (単位は%)

	F 値		精度		再現率	
	提案手法	MOC	提案手法	MOC	提案手法	MOC
データ 1	**58.97**	52.76	**56.26**	47.53	**61.95**	59.27
データ 2	54.72	**58.98**	**54.31**	43.66	55.13	**90.85**
データ 3	**54.45**	45.20	**61.01**	30.32	49.17	**88.74**
データ 4	**60.37**	42.10	**65.19**	28.66	56.21	**79.23**
データ 5	**71.31**	52.84	**72.86**	65.68	**69.82**	44.20
データ 6	56.00	**65.16**	**65.81**	64.39	48.72	**65.94**
データ 7	**75.13**	56.22	**86.56**	56.24	**66.37**	56.20
データ 8	69.06	**69.14**	**80.57**	69.01	60.43	**69.28**

タ 1 とデータ 5 について提案手法の方が MOC よりも高い F 値であるのは, クラスタ数が少ないことにより提案手法の再現率が MOC より高いためであると考えられる. 一方, データ 2 とデータ 6 については提案手法の F 値は MOC よりも低い. これは, MOC の高い再現率に対して提案手法の精度が伸び悩んだことによる. データ 2 とデータ 6 についての提案手法の精度が他のデータセットに対して適用した時よりも低い理由は, データ 2 とデータ 6 を構成するユーザが一般人ユーザとは大きく異なるためであると考えられる. 事実, データ 2 とデータ 6 を作成するために用いたベースのタグは「news」「tv」といったマスメディアに関係するタグであり, これらのユーザ中には情報発信アカウントが多く含まれる. したがって, データ 2 とデータ 6 については提案手法が想定しているようなリンク構造をもたず, 提案手法が効果的に機能しなかったのだと考えられる.

以上の実験結果をまとめる. 提案手法は MOC よりも全般的に高い精度を有し, 特に一般人ユーザに近いデータセットについてはその傾向は顕著である. また, クラスタ数が少ない場合は再現率についても MOC よりも高い. これらの結果より F 値について提案手法は一般ユーザに近いデータセットについて MOC と同等以上の性能を有し, クラスタ数が少ない場合は著名人のデータセットについても MOC よりも高い性能を有する. したがって, 提案手法は Twitter 上のユーザの重複クラスタリングに適用するうえで, 十分妥当な方法であるといえる.

4. 関 連 研 究

関連研究として, Girvan ら [12] によるソーシャルネットワーク上のコミュニティ発見手法を拡張した Gregory [4] の重複コミュニティ発見手法が挙げられる. その他にも Palla ら [3] や Zhang ら [5] によって重複コミュニティ発見手法が提案されている. 本研究で扱った重複クラスタリング手法とこれらの重複コミュニティ発見手法の違いを図 8 を用いて説明する. 重複コミュニティ発見手法はユーザ間のリンク構造のみを用いるため, 個々のユーザの特徴とは無関係に純粋にコミュニティを発見することを目的とする. 図 8 では, 重複コミュニティ発見手法は左側のアニメコミュニティ1 と右側のアニメコミュニティ2 を異なるコミュニティ(クラスタ) とみなす. 一方, 重複クラスタリング手法ではユーザの特徴に従ってクラスタを形成する. 図 8 では, アニメコミュニティ1 とアニメコミュニティ2 に属するユーザは「アニメファン」という特徴を有するため, 重複クラスタリング手法はこれらのコミュニティを同一のクラスタに割り当てる.

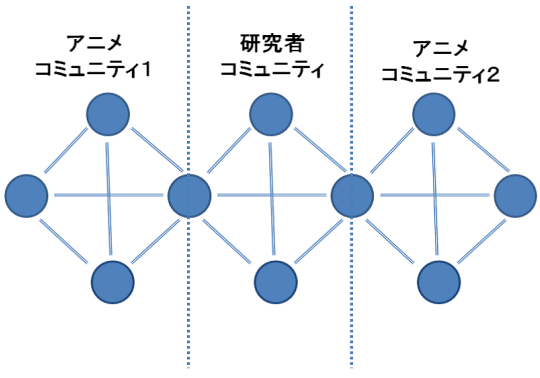

図 8　重複コミュニティ発見手法の適用例

5. まとめと今後の課題

本研究では, 共有リンク先ユーザ集合を用いてリンクの特徴

ベクトルを定義し, リンクの特徴ベクトルをクラスタリングすることによって重複クラスタリングを実現する手法を提案した. また, 提案手法を Twitter のデータに対して適用し, 既存手法との比較を行うことによって提案手法の妥当性を示した.

今後は, 提案手法の一般性を検証するために, Twitter 以外の SNS やウェブネットワークに対して提案手法を適用して評価実験を行うことを考えている. また, K-means 法以外の分割クラスタリング手法を提案手法で用いることで, クラスタリングの性能の向上を図れるかを検証したい.

本研究では, 重複クラスタリングの実行結果の妥当性により間接的にリンクの特徴ベクトルの定義の妥当性を示したが, 今後はより詳細な実証実験を行うことによりリンクの特徴ベクトルが実際にリンク元ユーザの「関心」を表現しているかを検証したい. また, 関連研究で述べた重複コミュニティ発見手法と比較実験を行うことにより, ユーザの特徴とリンク構造を用いた提案手法とリンク構造のみを用いた重複コミュニティ発見手法の違いを明らかにしたい.

文　　献

[1] A. Banerjee, C. Krumpelman, S. Basu, R. Mooney, and J. Ghosh. Model-based overlapping clustering. In KDD, 2005.

[2] C. D. Manning, P. Raghavan, and H. Schutze. Introduction to information retrieval, Cambridge University Press. 2008

[3] G. Palla, I. Derény, I. Farkas and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814-818, 2005.

[4] S. Gregory. An Algorithm to find overlapping community structure in networks. In PKDD, 91-102, 2007.

[5] S. Zhang, R.-S. Wang, and X.-S. Zhang. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A, 374-483, 2007.

[6] L. A. Adamic, and E. Adar. Friends and neighbors on the Web. Social Networks, 25(3), 211-230, July 2003.

[7] A. Banerjee, S. Merugu, I. S. Dhillon and J. Ghosh. Clustering with Bregman divergence. In Proc. SDM-04, 2004.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelifood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B, 39: 1-33, 1977.

[9] J. C. Bezdek and S. Pal. Fuzzy Models for Pattern Recognition. IEEE Press, Piscataway, NJ, 1992.

[10] J. McQueen. Some methods for classification and analysis of multivariate observations, In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281-297, 1967.

[11] D. Pelleg, and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In Proc. ICML, 727-734, 2000.

[12] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821-7826, 2002.