修士論文

# データの時系列変化を把握するための制約付きクラスタリング

同志社大学大学院 工学研究科 情報工学専攻
博士前期課程 2009 年度 737 番

水野　珠季

指導教授　三木 光範 教授

2011 年 1 月 21 日

# Abstract

In this thesis, restricted clustering technique was applied into detecting time series variation of subgroups which occurs on group of text contents such as scientific papers and blog entries. In the conventional clustering methods, data is classified only by similarity of static data information and data is not classified with along to time information. Restricted clustering method was introduced to categorize data and content preliminarily was used as restriction. In the proposed method, previous clusters are used as restrictions to consider association of clusters through time. Using this approach, data which is not categorized is also classified. In restricted clustering, there is a parameter called the "restriction strength". This parameter controls the degree of influence of the restriction on the data and the parameter value affects to the results of clustering. In the former studies of restricted clustering, how to decide the value of this parameter was not discussed. In this thesis, the approach to determine the restriction value was introduced. In the proposed approach, the relation between restriction value and Jaccard index is prepared and restriction value is derived by Jaccard index. This relation can be derived by preliminary experiments using various value of the parameter on test data. Through numerical experiments, the proposed restricted clustering was discussed. As a result, it was found that restricted clustering using previous cluster as a restriction is effective for detecting time series variation of clusters. The transition of clusters where some clusters are merged of split was observed. It was also found that the Jaccard index is useful to decide the value of " restriction strength ".

# 目 次

# 1　序論

情報通信技術の発展に伴い，インターネット上では文書，画像，動画など，多種多様な情報が公開され，入手可能となっている．そのため近年では，経済産業省が行う情報大航海プロジェクト[1]や加藤らの提唱する情報編纂（Information Compilation）[1)] のように，蓄積された情報を解析し，活用する動きが活発化している．本研究では，蓄積された情報の中でも，論文やブログ記事といったテキストコンテンツの集合に着目した．このようなテキストコンテンツの集合は，個々のコンテンツの内容は不変であるが，次々と新たなコンテンツが追加されることによって集合としての全体像が変化していくと考えられる．論文の例であれば，年代ごとに盛んに研究される分野が変化していくことで研究分野ごとの論文のサブグループが成長・縮小し，さらには既存分野の分裂や統合，新規分野の出現といった変化が起こるだろう．ブログ記事であればもっと短期的に，著者の興味・関心の移り変わりを反映して記事のサブグループが変化していくと予想される．

テキストコンテンツの集合を時系列で俯瞰し，このような変化を捉えることによって，その集合に対する新たな知見が得られる可能性がある．また，同種の集合を複数比較して分析し，分析結果をマーケティングや情報推薦などに利用することも考えられる．そこで本研究では，こうした変化を捉えるための手法について検討している．

本論文では，制約付きクラスタリング[2)] を用いてこの時系列変化を把握することを考え，小規模なテストデータを用いた実験によって通常のクラスタリングよりも制約付きクラスタリングがデータの時系列変化の把握に有効であることを確認した．また，制約付きクラスタリングのパラメータである制約の強さを決定する指標として Jaccard 係数を利用することを提案し，実験によりこれが有用であることを示した．

本論文の構成を以下に示す．第 2 章では，本研究で対象とするデータと把握したい時系列変化について説明し，これをクラスタリングで実現する際の問題点について述べる．第 3 章では，2 章で述べたクラスタリングの問題点を制約付きクラスタリングによって解決する方法を述べる．第 4 章では，制約付きクラスタリングの有効性を確認するための実験を行い，第 5 章で Jaccard 係数を用いた制約の強さの決定方法を提案し，これが有用であるか確認するための実験を行う．最後に第 6 章で結論を述べる．

[1] http://www.meti.go.jp/policy/it_policy/daikoukai/index.htm

# 2　クラスタリングによる時系列変化の把握

## 2.1　データの時系列変化

本研究では，データの時系列変化を把握することを目的としている．ここでいうデータとは学術論文やブログ記事などのテキストコンテンツの集合であり，特定の学会や分野の学術論文の集合，特定のブログの記事の集合といったコンテンツ提供者を単位とした集合だけでなく，ある研究者が収集した文献の集合，あるユーザがブックマークしたブログやニュースの記事の集合といったようにコンテンツ利用者を単位とした集合も考えられる．これらのデータは，各コンテンツの内容が時間によって変化することは無い．しかし，集合内のコンテンツは内容によっていくつかのグループに分類でき，次々と新たなコンテンツが追加されていくことによって，このグループが変化していくと考えられる．本論文ではこのような集合内のグループの構成が時間によって変化していく様子をデータの時系列変化と定義する．

Fig. 2.1 はあるユーザがブックマークしたコンテンツの集合を想定したデータの時系列変化の例である．この例では，初期時刻にはコンテンツが野球，ゲーム，映画という 3 つのグループから構成されており，時間が進むにつれてゲームのグループは成長していき，逆に野球のグループは縮小して最終時刻には消滅している．また，映画のグループが分裂して韓国映画が独立したグループになったり，犬のグループが新たに出現したりという変化が起こっている．映画のグループの分裂からは，この時期にユーザが映画の中でも特に韓国映画を好んでチェックしていたということが読み取れる．犬のグループの出現はユーザが犬を飼い始めた時期と重なっているのかもしれないし，野球のグループの消滅はプロ野球のシーズンが終わったことでこのユーザが野球に関係する話題に興味を持たなくなり，次のシーズンの開始が近づくとまた野球のグループが出現するのかもしれない．

Fig. 2.1 に見られるように，データの時系列変化としては，グループの成長，縮小，出現，消滅，分裂，また，Fig. 2.1 中には無いが，分裂とは逆に 2 つのグループが 1 つになる統合の 6 種類の変化が考えられる．このような時系列変化を把握することは，データについて客観的に俯瞰することを可能にし，自身の興味・関心や研究対象などに対する新たな知見をもたらす可能性がある．また，複数の集合の時系列変化を比較・分析することで今後起こりうる変化の予測やマーケティング，ユーザへの情報推薦などにも利用できるのではないかと考えられる．

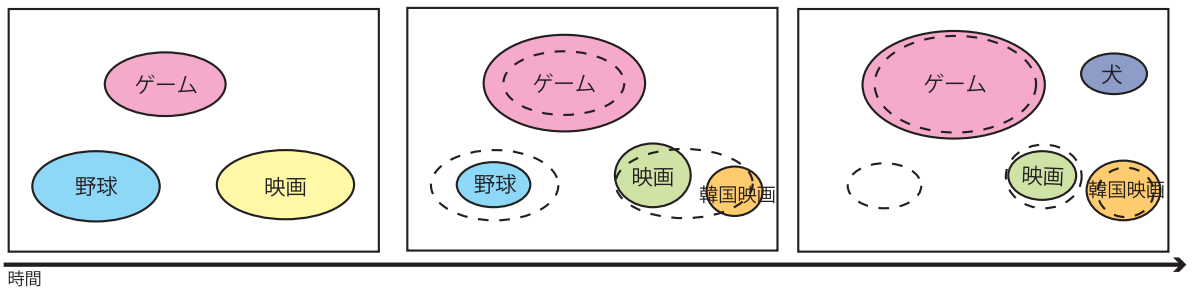



Fig. 2.1　データの時系列変化の例

## 2.2　クラスタリングによる時系列変化の把握とその問題点

前節で述べたようなデータの時系列変化を把握する方法として，任意の時間間隔で繰り返しクラスタリングを行い，前後の時刻間でクラスタの対応関係を同定するというものが考えられる．クラスタリングとは，分類すべき個体群を個体間に定義された関連度に基づいていくつかのサブグループに分類する手法である[3]．このサブグループはクラスタと呼ばれ，同じクラスタ内においては個体間の関連度が大きく，異なるクラスタにおいては関連度が小さくなるように分類される．Fig. 2.2 のように各時刻のクラスタとその対応関係を図示することで，前節で説明したようなグループの分裂や統合などの変化を把握することが可能となる．

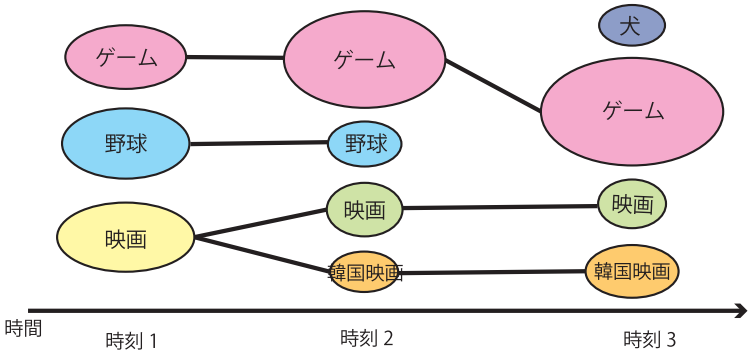



Fig. 2.2　目標とするクラスタリング結果

しかし，単一の話題について簡潔に記述されたニュース記事のようなコンテンツであればともかく，学術論文のような複数の要素から構成されるコンテンツに対してこのような方法を用いると，Fig. 2.3 のように少量のコンテンツの追加によってクラスタリング結果が大きく変化してしまう可能性がある[4]．テキストコンテンツをクラスタリングするには，単語の出現頻度などをもとにテキストコンテンツを特徴ベクトルとして表し，その特徴ベクトル同士のコサイン距離などをコンテンツ間の関連度とする．この際，例えば「多目的遺伝的アルゴリズムによる SVM 学習データ選択手法[5]」という論文の場合,「多目的遺伝的アルゴリズム」と「SVM」という少なくとも 2 つの要素が含まれているが，これをクラスタリングすると他のコンテンツとの関係によって多目的遺伝的アルゴリズムに関連するクラスタに分類されることもあれば SVM に関連するクラスタに分類されることもある．

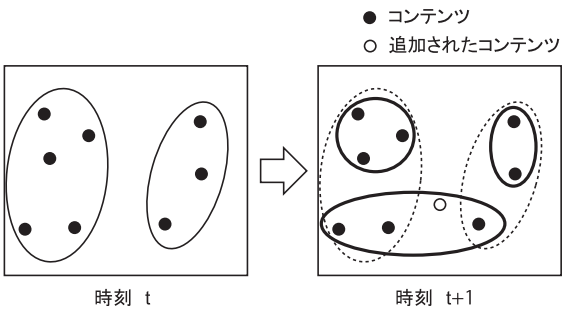



Fig. 2.3　分類される要素の変化によるクラスタの変化

コンテンツの追加によって隣接する時刻間で分類される要素が変わってしまうと，Fig. 2.3 のようにクラスタが大きく変化してしまう．その結果，Fig. 2.4 のようにクラスタとその対応関係を図示しても，クラスタの対応関係が複雑で分裂，統合などのデータの時系列変化を読み取ることが困難な状態になってしまう．クラスタリングは通常，個体間の関連度のみに基づいて分類を行い，時系列でのクラスタの関連は考慮されないのでこのような問題が起こる．

本論文ではこの問題の解決法として制約付きクラスタリングを用いている．制約付きクラスタリングについては次章で詳述する．

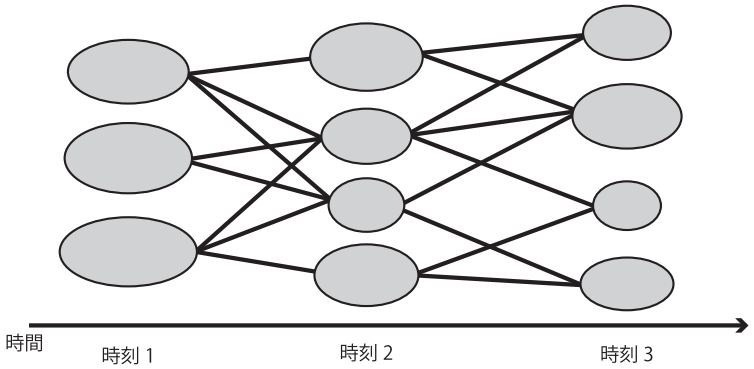



Fig. 2.4　クラスタリングでデータの時系列変化を把握する際の問題点

# 3　制約付きクラスタリング

## 3.1　制約付きクラスタリングの概要

制約付きクラスタリングは時間の経過によるカテゴリの変化を考慮した論文分類の手法として榊らによって提案された[2)]．榊らは，現時点でのカテゴリに分類された論文を過去の時点のカテゴリに再分類したデータを用いて実験を行った．その結果，再分類した過去のカテゴリを制約として制約付きクラスタリングを行うことで，通常のクラスタリングよりも現在のカテゴリを高い精度で再現できることを示した．

Fig. 3.1 に文献[2)] で提案された制約付きクラスタリングの概要を示す．まず，論文集合をカテゴリに分類する．また同時に，各論文間の関連度を求め，これを重みとして論文集合を関連度による論文ネットワークとして表現する．次に，カテゴリ分類の結果を論文ネットワークに制約として付加する．つまり，2 つの論文が同じカテゴリに属す場合は関連を強め，別のカテゴリに属す場合は関連を弱める．最後に，制約を付加した論文ネットワーク（制約付きネットワーク）をクラスタリングする．

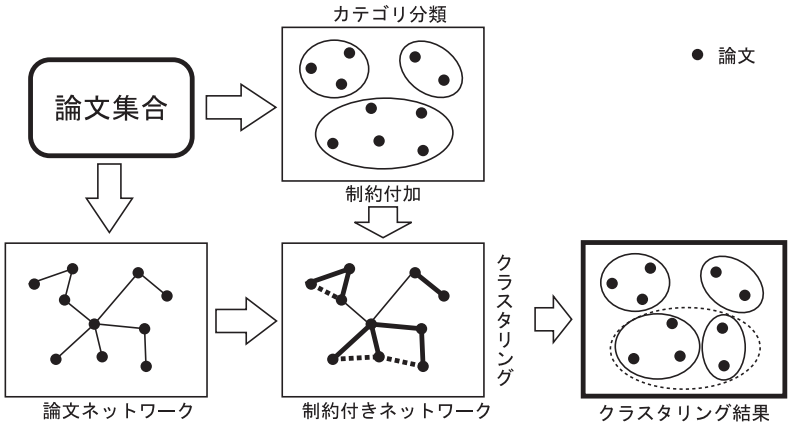



Fig. 3.1　榊らの提案した制約付きクラスタリング

上述のように，文献[2)] で提案された制約付きクラスタリングではクラスタリング対象となるデータを予め用意されたカテゴリに分類しておく必要があった．しかし，本研究では様々なテキストコンテンツの集合を対象としている．そのため，どのようなカテゴリが存在するのかが未知であるデータにも適用できるよう，カテゴリ分類の代わりに直前の時刻のクラスタリング結果を制約として使用する方法を考えた．次節では，本研究での制約付きクラスタリングの手順について詳述する．

## 3.2　制約付きクラスタリングの手順

ここでは本研究で用いている，直前のクラスタリング結果を制約とした制約付きクラスタリングの手順について述べる．概要を Fig. 3.2 に示す．

前提として，クラスタリングの対象となるデータはクラスタリングを行う時刻以前の全てのコンテンツである．例えば 2000 年から年単位でクラスタリングを行う際に，2005 年のクラスタリングで対象となるのは 2005 年の一年間に追加されたコンテンツのみではなく，2000 年から 2005 年までの 5

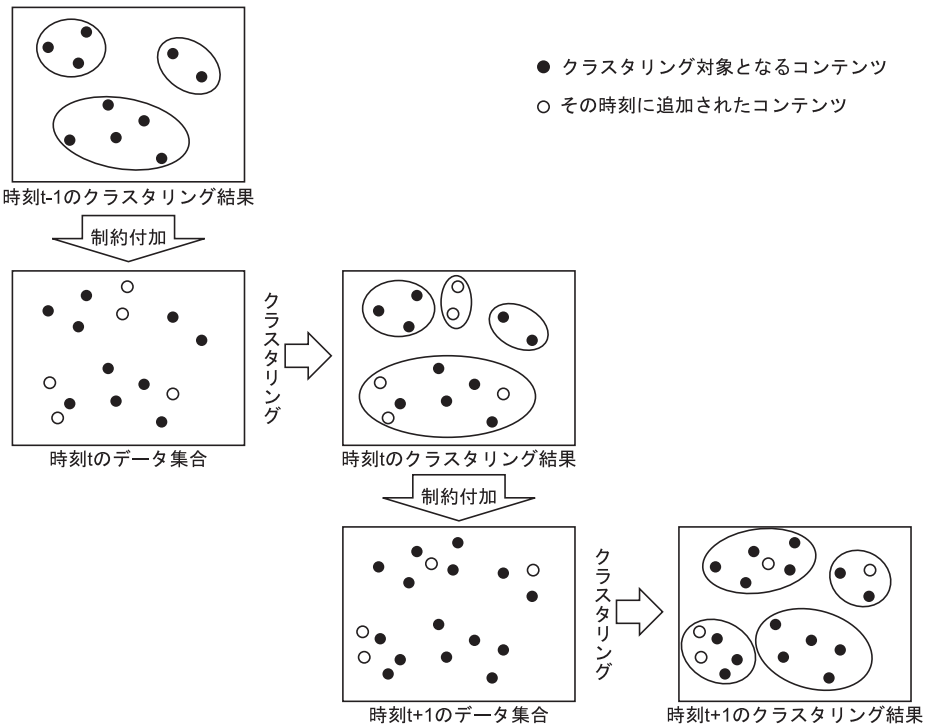



Fig. 3.2　直前の時刻のクラスタリング結果を制約とした制約付きクラスタリング

年分のコンテンツである．

以下に手順を示す．

**初期時刻**

（1）各コンテンツ間の関連度を求め，関連度ネットワーク（データを関連度を重みとしたネットワークの形式で表現したもの）を作成する．

（2）関連度ネットワークをクラスタリングし，初期時刻のクラスタリング結果を得る．

**時刻 $t$**

（3）追加されたコンテンツを加えて，関連度ネットワークを更新する．

（4）時刻 $t-1$ のクラスタリング結果から，制約行列 $C$ を作成する．制約行列 $C$ の各成分 $c_{ij}$ はコンテンツ $i$ とコンテンツ $j$ が同じクラスタに属すときに 1，別々のクラスタに属すときに 0 となる．

（5）式 (3.1) を用いて制約付きネットワーク（関連度ネットワークに時刻 $t-1$ のクラスタリング結果を制約として付加したもの）を求める．なお，式 (3.1) において $S$ は関連度ネットワークの隣接行列，$r$ は制約行列 $C$ による制約の強さを表すパラメータである．

$$R = (1-r)S + rC \qquad (0 \leq r \leq 1) \tag{3.1}$$

（6）制約付きネットワークをクラスタリングして時刻 $t$ のクラスタリング結果を得る．

手順（3）から（6）を最終時刻まで繰り返す．

このように，直前の時刻に同じクラスタに属していたか否かによってコンテンツ間の関連度を強めたり弱めたりすることで，時系列でのクラスタの関連を考慮したクラスタリングを行う．なお，手順（5）で制約を付加する範囲は時刻 $t-1$ 以前のコンテンツ間の関連度に対してのみで，時刻 $t$ に追加されたコンテンツとの関連度は弱めないようにしている．これは，制約によって把握したい変化まで打ち消されてしまうことを防ぐためである．

上述の手順のほか，コンテンツ間の関連度を求める方法を定義する必要があるが，これは 2.2 節でも触れたように，単語の出現頻度をもとに特徴ベクトルを作成する方法などが考えられる．また，クラスタリングの手法については任意のものが使用できるが，本論文では文献[2] と同様に Newman 法[6] を用いている．Newman 法は併合型の階層的クラスタリング手法であり，式 (3.2) のような *modularity* と呼ばれる評価関数 $Q$ を最適化することによってクラスタ数が自動的に決定される．なお，式 (3.2) において $e_{ij}$ はクラスタ $i$ からクラスタ $j$ へのエッジの重みの和を全エッジの重みの和で割った値であり，$a_i = \sum_j e_{ij}$ である．

$$Q = \sum_i (e_{ii} - a_i^2) \tag{3.2}$$

# 4　制約付きクラスタリングの有効性の検証

## 4.1　実験概要

前章で述べた方法がデータの時系列変化を把握するために有効であるかをテストデータを用いた実験により検証した．実験では，通常のクラスタリングの場合はFig. 2.4のように隣接する時刻間でクラスタが大きく変化してクラスタの対応関係が複雑になってしまうデータが，制約付きクラスタリングによってFig. 2.2のようにクラスタの分裂や統合がそれぞれ独立して起こる状態になるかを確認した．

実験に使用したテストデータはノード数20の重みつきネットワークで，各ノードは自身がネットワークに追加された時刻の情報を持っており，エッジはノード間の関連度を重みとして持っている．ネットワークの平均次数の違いによって結果に差が出るのではないかと考え，平均次数が1.0，1.5，2.0の3種類のネットワークをそれぞれ5個，計15個のテストデータを作成した．各ノードが持つ時刻の情報は15個全てのデータにおいて共通で，0から14の15ノードは時刻1，つまり初期時刻から存在するノードとし，15から19の5ノードが時刻2に新たに追加されたノードとした．エッジの重み，つまりノード間の関連度は0.0，0.25，0.5，1.0の4種類から遺伝的アルゴリズム（GA）を用いて指定した平均次数を満たし，かつ制約の強さが0.00のときと0.90のときでクラスタリング結果の差が大きくなるように選択した．

上述の15個のテストデータそれぞれに対し，3.2節の方法で制約の強さを0.00から1.00まで0.01づつ変化させて制約付きクラスタリングを行い，制約の強さが0.00，つまり通常のクラスタリングの場合の結果と初めてクラスタの分裂，統合がそれぞれ完全に独立して起こる状態になった制約の強さでの結果（以降これを制約付きクラスタリングの結果と呼ぶ）とを比較する．

## 4.2　実験結果

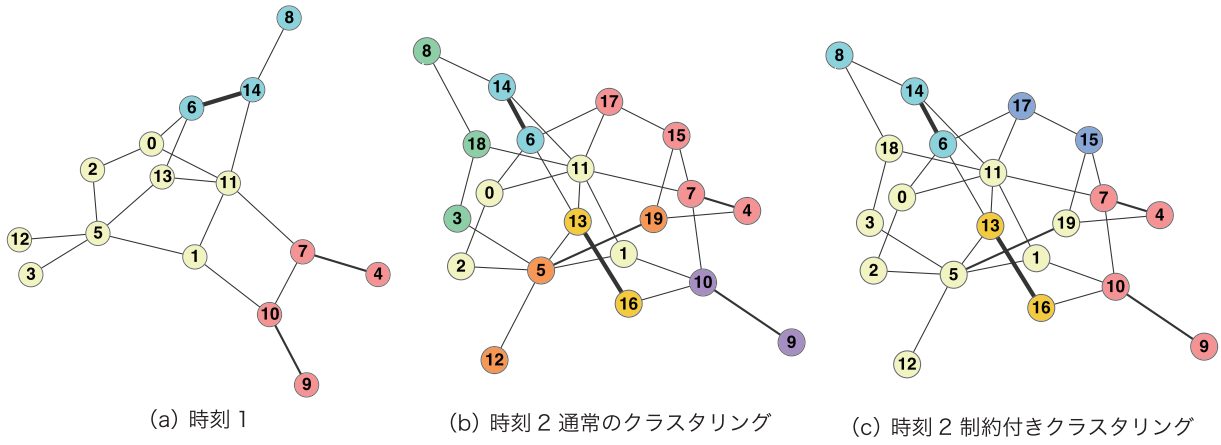
平均次数1.0，1.5，2.0のデータの結果をそれぞれ1つ例として示す．

Fig. 4.1，Fig. 4.3，Fig. 4.5はそれぞれ平均次数1.0，1.5，2.0のデータであり，ノードの色は属するクラスタを示している．また，エッジの太さは関連度の強さを表している．いずれも，(a）は時刻1の結果，(b）は通常のクラスタリングの場合の時刻2の結果，(c）は制約付きクラスタリングの場合の時刻2の結果である．また，Fig. 4.2，Fig. 4.4，Fig. 4.6はそれぞれFig. 4.1，Fig. 4.3，Fig. 4.5のデータの2時刻間のクラスタの対応関係を図示したもので，2時刻間のクラスタをつなぐ線の太さは両クラスタで共通しているコンテンツ数を表している．いずれも，(a）は通常のクラスタリングの結果，(b）は制約付きクラスタリングの結果である．

Fig. 4.2，Fig. 4.4，Fig. 4.6を見ると，いずれも（a）の通常のクラスタリングの場合には，たとえばFig. 4.2のc1とc2から分裂したノードが統合してc7になるというように，時刻1の複数のクラスタから分裂したノードが統合されて時刻2のクラスタになっている箇所があり，クラスタの分裂と統合が独立せず混ざり合って起こっている．これに対して（b）の制約付きクラスタリングの場合には，Fig. 4.2では分裂（c1 → c11，c12）と出現（c15），Fig. 4.4では成長（c1 → c12とc2 → c13）と統合（c3，c4 → c14），Fig. 4.6では分裂（c1 → c10，c11とc2 → c12，c13）というようにそれぞ

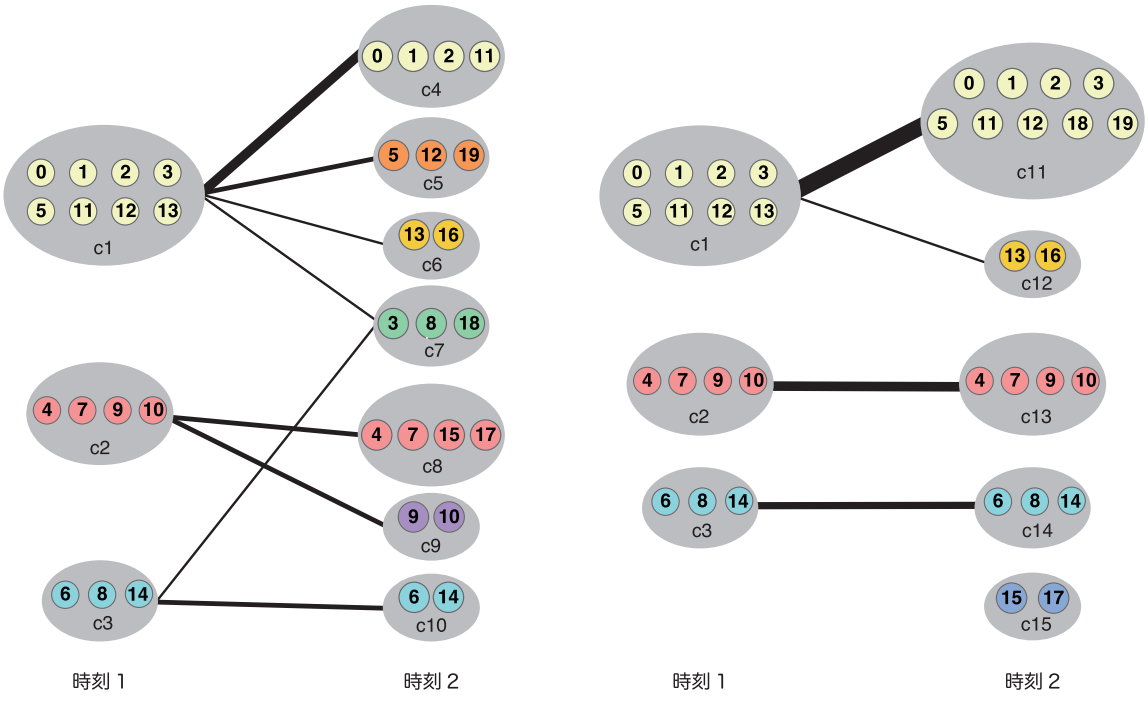
れの変化が独立して起こっている．15 個すべてのデータにおいて，同様の結果が確認できた．

Table 4.2 には各テストデータにおいて初めて目標とする結果が得られた時点の制約の強さを示した．この表を見ると分かるように，目標とする結果を得るために必要な制約の強さはデータによって様々な値になっている．なお，目標とする結果とはクラスタの分裂，統合がそれぞれ完全に独立して起こる状態になる結果であり，Fig. 4.1 から Fig. 4.6 で制約付きクラスタリングの結果として示したものである．



(a) 時刻 1　(b) 時刻 2 通常のクラスタリング　(c) 時刻 2 制約付きクラスタリング

Fig. 4.1　平均次数 1.0 のデータのクラスタリング結果




(a) 通常のクラスタリング　(b) 制約付きクラスタリング

Fig. 4.2　平均次数 1.0 のデータのクラスタの対応関係

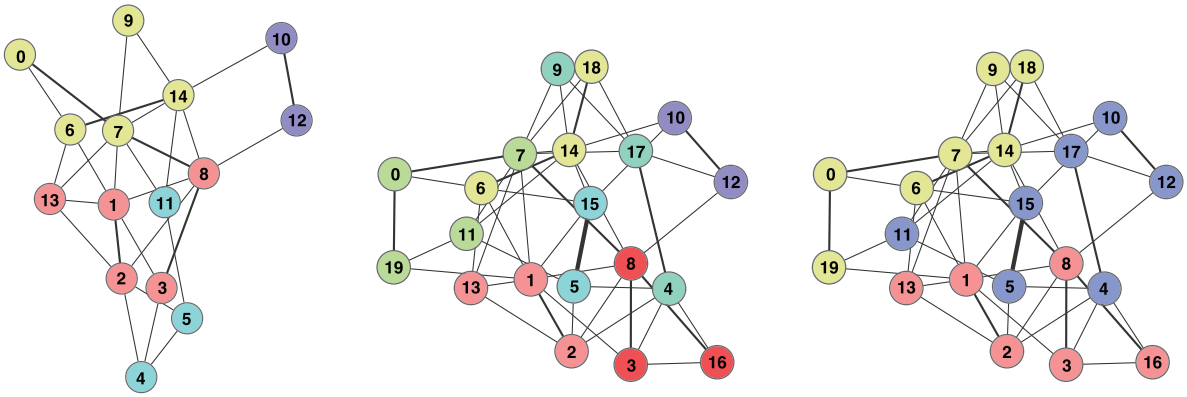



Fig. 4.3　平均次数 1.5 のデータのクラスタリング結果

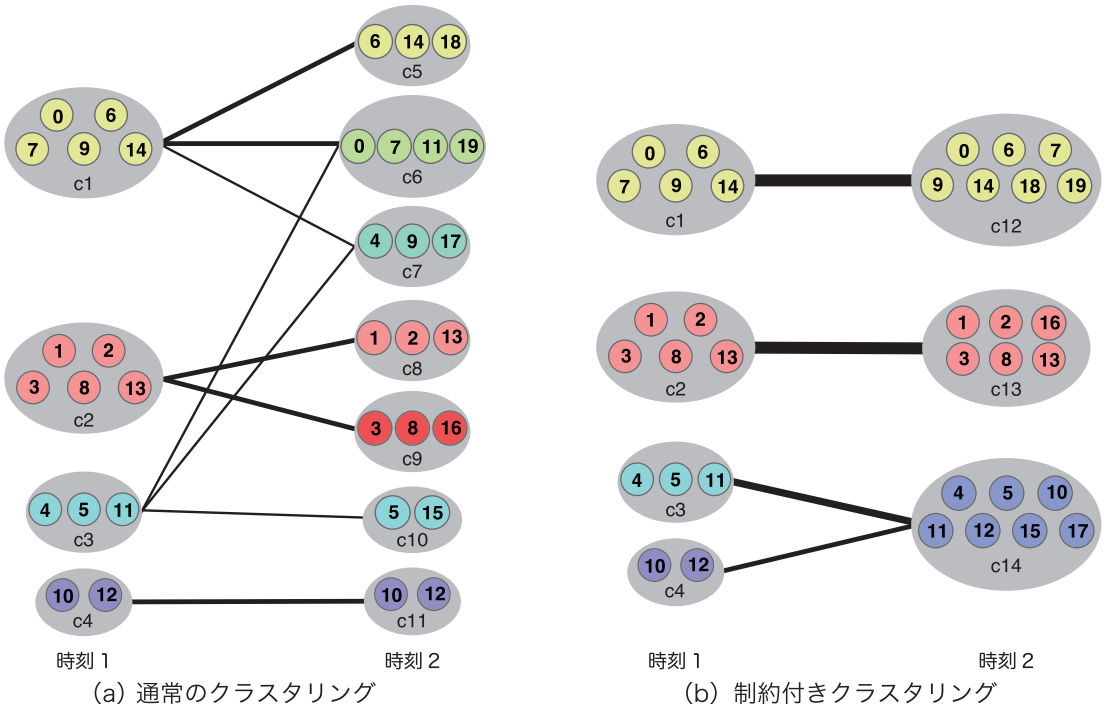



Fig. 4.4　平均次数 1.5 のデータのクラスタの対応関係

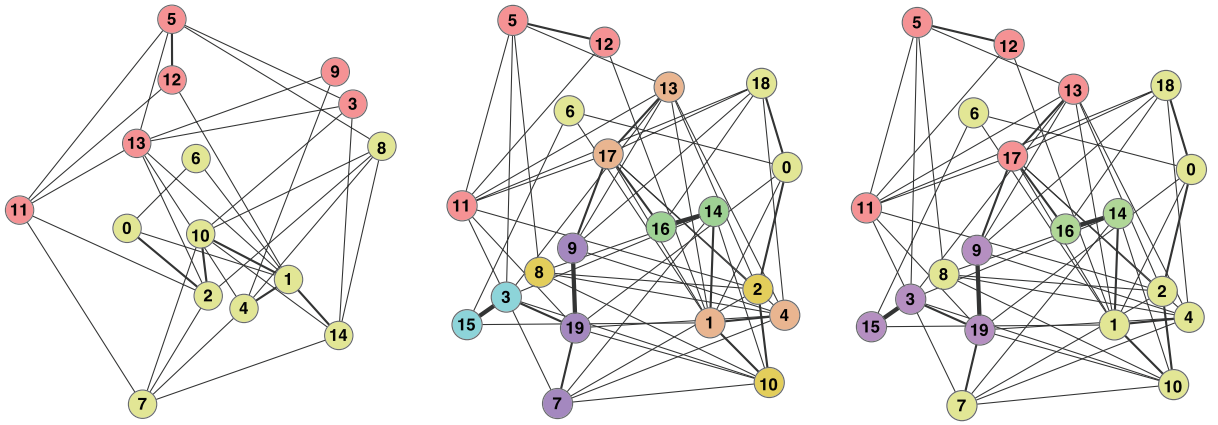



Fig. 4.5　平均次数 2.0 のデータのクラスタリング結果

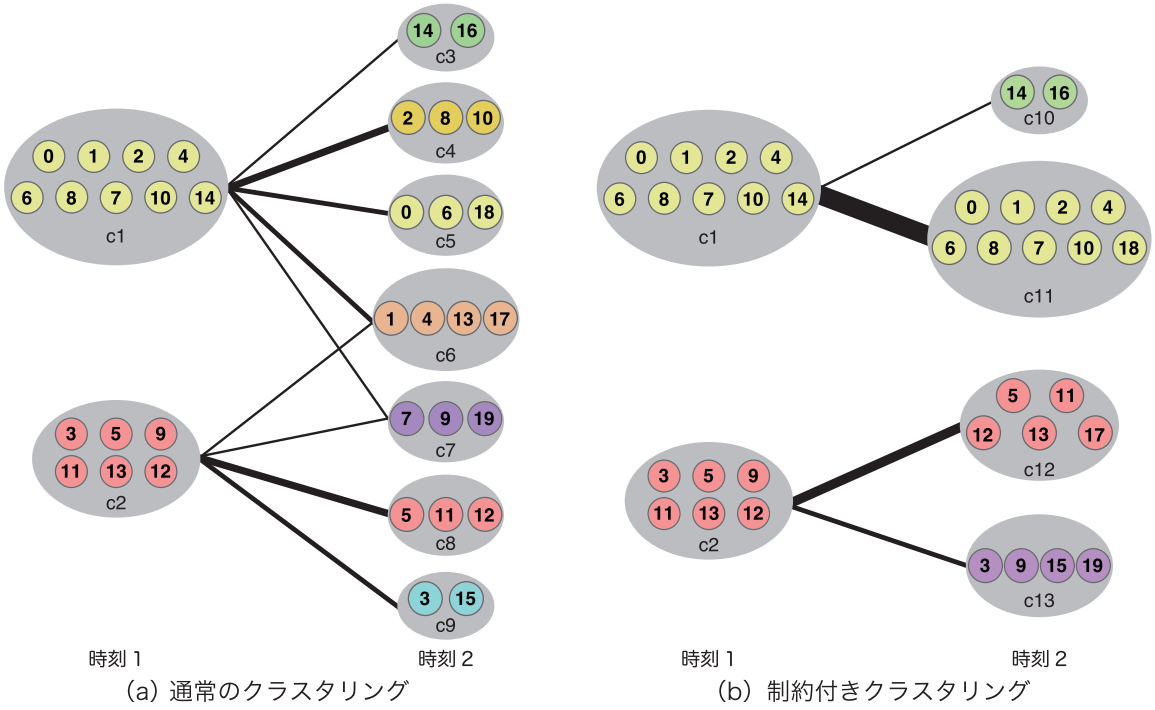



Fig. 4.6　平均次数 2.0 のデータのクラスタの対応関係

Table 4.1　目標とする結果が得られた時点の制約の強さ

| データ | 平均次数 | 制約の強さ |
|---|---|---|
| 1 | 1.0 | 0.02 |
| 2 | 1.0 | 0.04 |
| 3 | 1.0 | 0.12 |
| 4 | 1.0 | 0.05 |
| 5 | 1.0 | 0.01 |
| 6 | 1.5 | 0.16 |
| 7 | 1.5 | 0.04 |
| 8 | 1.5 | 0.14 |
| 9 | 1.5 | 0.07 |
| 10 | 1.5 | 0.07 |
| 11 | 2.0 | 0.14 |
| 12 | 2.0 | 0.07 |
| 13 | 2.0 | 0.08 |
| 14 | 2.0 | 0.05 |
| 15 | 2.0 | 0.06 |

### 4.3　考察

実験結果より，通常のクラスタリングでは隣接する時刻間でクラスタが大きく変化してクラスタの対応関係が複雑になってしまうデータが，制約付きクラスタリングによってクラスタの分裂や統合がそれぞれ独立して起こる状態になることが確認できた．

ここで，制約付きクラスタリングによって起こらなくなった変化がどのような変化であったか確認するために Fig. 4.1，Fig. 4.2 のノード 8 に注目する．まず Fig. 4.2 を見ると，ノード 8 は時刻 1 ではノード 6，14 とともにクラスタ c3 に属している．制約付きクラスタリングでは時刻 2 でも同じくノード 6，14 とともにクラスタ c14 に属している．しかし通常クラスタリングの場合には，時刻 2 で c3 のうちノード 8 だけがノード 3，18 が属すクラスタ c7 に移動している．ここで Fig. 4.1 を見ると，時刻 2 においてノード 8 はノード 14，18 に同じ関連度で繋がっているため，c7，c10 のどちらのクラスタに属しても不自然ではないと考えられる．これは，2.2 節で述べたコンテンツに複数の要素が含まれている状態に相当する．このようにどちらのクラスタに属しても不自然ではない状況において，データの時系列変化を把握するという観点からは，できるだけ時刻 1 のクラスタの構成に従って分類されることが望ましい．しかし，通常のクラスタリングでは時刻 1 でノード 8 がどのクラスタに属していたかということは考慮されていないため，c7 に移動してしまった．制約付きクラスタリングでは時刻 1 のクラスタリング結果を制約としているので，ノード 8 とノード 6，14 との関連が強められてクラスタが維持された．

また，Fig. 4.2，Fig. 4.4，Fig. 4.6 を見ると通常のクラスタリングでは全体的に時刻 1 の時点よりも細かく分類されていることが分かる．データを分類する際のグループは，例えば映画，音楽というような大まかな分類もあればアクション，アニメ，ドキュメンタリーやジャズ，クラシックといったジャンルごとの分類など様々なレベルでの分類が考えられる．よって，通常のクラスタリングで細かく分類されていることは間違った分類結果であるというわけではないが，時系列変化を把握するという観点からは時刻ごとに分類のレベルが大きく変わってしまうことは望ましくない．制約付きクラスタリングの結果では，ほとんどの場合は時刻 1 のクラスタをそのまま維持しており，Fig. 4.2 の c12，Fig. 4.4 の c14，Fig. 4.6 の c10，c13 のように追加されたノードと強い関連を持った部分だけが分裂したり，統合したりしている．この分裂や統合が，把握したいデータの時系列変化に相当すると考えられる．

以上のように，実験結果から制約付きクラスタリングは通常のクラスタリングと比較してデータの時系列変化の把握に有効であることが確認できた．

しかし同時に，制約がかかった状態にするために必要な制約の強さはデータに依存しており，一様に決めることは出来ないということも明らかになった（Table 4.2）．実験で用いた 20 ノードのデータでは，制約の強さを 0.00 から 1.00 まで変化させる間に多いもので 12 種類，少ないものでも 5 種類の異なるクラスタリング結果が得られた．今回の実験ではこの 5〜12 種類のクラスタリング結果を全て確認してどこで制約がかかった状態になっているかを調べた．しかし実際にテキストコンテンツ集合に適用することを考えると，コンテンツ数が多くなると分裂や統合が完全には独立せずに多少の雑音が入ると考えられるし，クラスタリング結果の種類もさらに増加すると考えられるので，全ての結果を確認することは現実的ではない．そのため，どの制約の強さのクラスタリング結果に注目すべきか判断するための何らかの指標が必要となる．

次章では，この指標として Jaccard 係数を利用することを提案する．

# 5　Jaccard 係数の平均値による制約の強さの決定

## 5.1　Jaccard 係数の平均値による制約の強さの決定

本節では，どの制約の強さの結果に注目すべきかを判断するための指標として Jaccard 係数の平均値を用いる方法を提案する．Jaccard 係数とは，2 つの集合の要素がどの程度一致しているかを示す指標で，クラスタの対応関係を同定するためにもよく用いられている．2 つのクラスタ $C_1$ と $C_2$ の Jaccard 係数は式 (5.1) のように求められる．

$$J(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} \tag{5.1}$$

Jaccard 係数の平均値を算出する手順を Fig. 5.3 に示す．図中の番号は以下の説明と対応している．

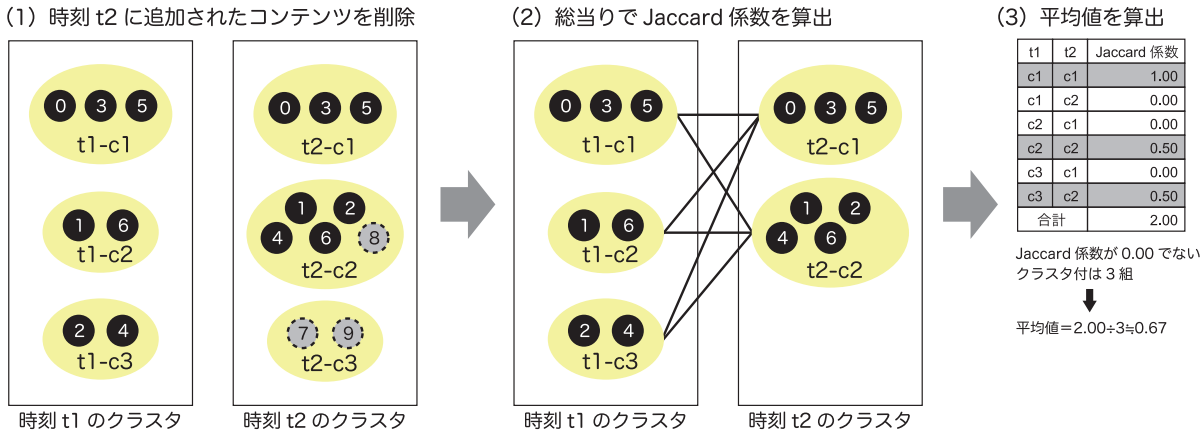



Fig. 5.1　Jaccard 係数を用いた指標の算出手順

(1) 隣接する 2 時刻のクラスタリング結果において，後の時刻のクラスタリング結果からその時刻に追加されたコンテンツを削除する．この処理によって追加されたコンテンツの数に影響されることが無くなる．

(2) 2 時刻間の全てのクラスタの組み合わせで Jaccard 係数を算出する．

(3) (2) で算出した値の総和を Jaccard 係数の値が 0.00 でないクラスタの組み合わせの数で割って平均値を算出する．この値を指標として用いる．

なお，3 時刻以上の場合には全ての隣接する時刻間において個別に制約の強さを決定する必要がある．

## 5.2　実験概要

本実験では，前節の方法で算出した Jaccard 係数の平均値が制約の強さを決定する指標として有用であるか確認した．

実験には前章と同じ 15 個のテストデータを使用した.

15 個のテストデータそれぞれで，各制約の強さのクラスタリング結果について前節の方法で Jaccard 係数の平均値を算出し，Jaccard 係数の平均値と制約の強さの関係について調査した.

## 5.3 実験結果と考察

Table 5.3 に 15 個のデータそれぞれの目標とする（分裂や統合が独立して起こる）結果が得られた時点の制約の強さとその時の Jaccard 係数の平均値，その直前の Jaccard 係数の平均値を示す．Table 5.3 を見ると分かるように，15 個中 12 個のデータにおいて Jaccard 係数の平均値が 0.5 を超えた時点で目標とする結果が得られていた．残り 3 個のデータでは，0.6 を超えた時点で目標とする結果が得られていた．この結果から，Jaccard 係数の平均値 0.5 という値が注目するクラスタリング結果を決める目安として使用できることが分かった.

Table 5.1　目標とする結果が得られる前後の Jaccard 係数

| データ | 平均次数 | 目標とする結果が<br>得られた時点の制約の強さ | その時点の<br>Jaccard 係数の平均値 | その直前の<br>Jaccard 係数の平均値 |
|---|---|---|---|---|
| 1 | 1.0 | 0.02 | 0.75 | 0.35 |
| 2 | 1.0 | 0.04 | 0.75 | 0.48 |
| 3 | 1.0 | 0.12 | 1.00 | 0.59 |
| 4 | 1.0 | 0.05 | 0.67 | 0.39 |
| 5 | 1.0 | 0.01 | 0.67 | 0.39 |
| 6 | 1.5 | 0.16 | 0.75 | 0.57 |
| 7 | 1.5 | 0.04 | 0.50 | 0.33 |
| 8 | 1.5 | 0.14 | 1.00 | 0.49 |
| 9 | 1.5 | 0.07 | 0.75 | 0.49 |
| 10 | 1.5 | 0.07 | 1.00 | 0.48 |
| 11 | 2.0 | 0.14 | 0.50 | 0.32 |
| 12 | 2.0 | 0.07 | 0.60 | 0.37 |
| 13 | 2.0 | 0.08 | 0.50 | 0.31 |
| 14 | 2.0 | 0.05 | 0.60 | 0.58 |
| 15 | 2.0 | 0.06 | 0.50 | 0.32 |

Fig. 5.2 のグラフは 5 番のデータの Jaccard 係数の平均値の推移を表したもので，横軸が制約の強さ，縦軸が Jaccard 係数の平均値となっている．また，グラフ上の点はクラスタリング結果が変化した点を表している．このデータでは，8 種類の異なるクラスタリング結果が得られており，Jaccard 係数の平均値が 0.5 を超えるのは 0.67 の時点であり，この値で 4 種類の結果が得られているので，まずこの 4 種類の結果に注目して分析を行えば良い.

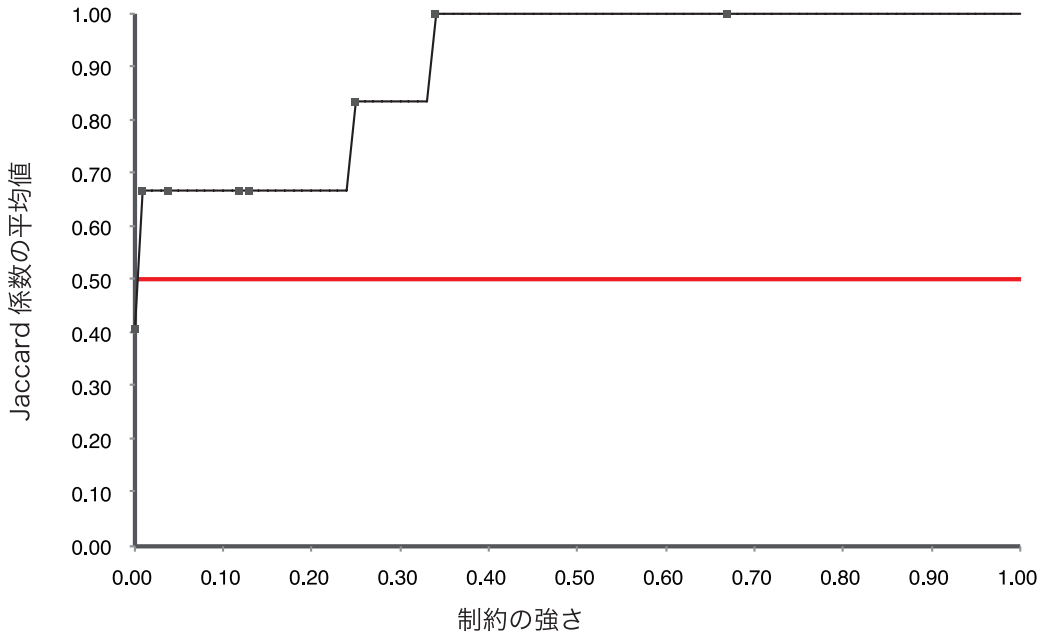



Fig. 5.2　Jaccard 係数による制約の強さの決定

# 6　結論

本研究では，テキストコンテンツの集合を対象としてその時系列変化を把握する方法について検討している．

本論文ではクラスタリングによってデータの時系列変化を把握することを考え，その際の問題点を整理し，この解決法として制約付きクラスタリングを利用することを提案し，小規模なテストデータを用いた実験によってその有効性を検証した．実験では，すべてのテストデータにおいて制約をかけることで分裂や統合が独立して起こるクラスタリング結果が得られ，制約付きクラスタリングが通常のクラスタリングと比較してデータの時系列変化の把握に有効であることが確認できた．また同時に，分裂や統合が独立して起こるクラスタリング結果が得られる制約の強さはデータによって様々であり，これを決定するための指標が必要であることも分かった．そこで第 5 章では Jaccard 係数の平均値を制約の強さを決定する指標とする方法を提案し，実験によってこの指標が有用であることを確認した．

今後は，実データを用いて手法の有効性を検討していく．また，今回のコンテンツを次々と追加していく方法では，6 種類のデータの時系列変化のうち縮小と消滅を捉えることが出来ない．そのため，コンテンツを削除する方策についても検討する必要がある．

# 謝辞

本研究を遂行するにあたり，多大なるご指導，ご協力を頂きました同志社大学生命医科学部の廣安知之教授に心より感謝いたします．また，研究生活を送る上で素晴らしい環境を与えて下さり，様々なご指摘を下さいました同志社大学理工学部の三木光範教授，本研究に様々なアドバイスを頂きました吉見真聡助教に心より感謝いたします．

さらに，同志社大学理工学部の松村冬子さんには，研究を進める上で適切なアドバイスを頂き，大変丁寧なご指導をして頂きました．本当にありがとうございます．また，本論文の執筆の際に，お忙しい中時間を割いて多くの助言を下さった田中美里さん，宮部洋太くんには大変感謝しております．

研究に対して貴重なご意見を下さった知的システムデザイン研究室および医療情報システム研究室の皆さまにも心より感謝いたします．

最後に，私をこれまで精神的，経済的に支え見守ってくれた両親に心より感謝いたします．本当にありがとうございました．

# 参考文献


1) 加藤恒昭, 松下光範. 情報編纂（information compilation）の基盤技術. 人工知能学会全国大会論文集, Vol. JSAI2006, pp. 51–54, 2006.

2) 榊剛史, 松尾豊, 石塚満. 制約付きクラスタリングを用いた論文分類. 人工知能学会全国大会論文集, Vol. JSAI2006, pp. 1–4, 2006.

3) 宮本定明. クラスター分析入門 - ファジィクラスタリングの理論と応用 -. 森北出版, 1999.

4) 水野珠季, 廣安知之, 三木光範, 伊藤冬子, 横内久猛. 制約付きクラスタリングによるデータの時系列変化の把握. 2009 年度人工知能学会全国大会 (第 23 回) 論文集, 2009.

5) 廣安知之, 西岡雅史, 三木光範, 横内久猛. 多目的遺伝的アルゴリズムによる svm 学習データ選択手法. MPS, 数理モデル化と問題解決研究報告, Vol. 2008, No. 126, pp. 77–80, 2008.

6) M.E.J.Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, Vol. 69, p. 066133, 2004.