# 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

Detection of the time series variation of data using constrained clustering

水野　珠季[*1]　廣安　知之[*2]　三木　光範[*3]　伊藤　冬子[*1*4]　横内　久猛[*2]
Tamaki Mizuno　Tomoyuki Hiroyasu　Mitsunori Miki　Fuyuko Ito　Hisatake Yokouchi

[*1]同志社大学 大学院工学研究科
Graduate School of Engineering, Doshisha University

[*2]同志社大学 生命医科学部
Department of Life and Medical Sciences, Doshisha University

[*3]同志社大学 理工学部
Department of Science and Engineering, Doshisha University

[*4]日本学術振興会
Japan Society for the Promotion of Science

In recent years, a bunch of various data is accumulated on the web. We focused on the contents that change over time, such as papers and news articles in this study. Therefore, a method to figure out the time series variation of data using constrained clustering is proposed in this paper. The proposed method was applied to a set of research reports on the web, and its validity was examined. As a result, it is confirmed that the proposed method is more effective to figure out the time series variation than the clustering that is not constrained.

## 1. はじめに

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，論文やニュース記事などの静的なコンテンツの集合に着目した．このようなデータの長期的な時系列変化を把握することが可能となれば，現状理解や今後の予測に役立つと考えられる．そこで本研究では，静的コンテンツの集合の時系列変化を把握する手法について検討していく．

本稿では，制約付きクラスタリング [2] によってデータの時系列変化を把握する方法を提案する．また，提案手法を実データに適用し，その有効性について検討を行う．

## 2. 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

### 2.1 時系列変化の把握

静的コンテンツの集合が時系列で変化していく様子を把握するには，まずその集合がどのようなコンテンツから構成されているのか，つまりデータの全体像を知る必要がある．データの全体像を把握する手法としては，クラスタリングがよく利用されている．クラスタリングとは，データの集合をデータ間に定義された関連度に基づいて，内容的に同質ないくつかのサブグループに分類する手法である．

時系列での変化を把握するには，このクラスタリングによって得られるサブグループが時刻ごとにどのように変化していくかを分析すれば良い．しかし，関連度のみに基づいて分類されるクラスタリングでは，時刻ごとに分類の基準に揺れが生じてしまう可能性がある．分類基準の揺れとは，論文の集合を例に説明すると，ある時刻では論文で扱う手法ごとに分類されていたものが，別のある時刻では，対象問題ごとに分類されてしまうといったものである．このように分類の基準自体が変わってしまっては，時間の経過によって起こった変化を把握することが困難である．本稿では，この問題を解決するために，制約付きクラスタリングを用いる．

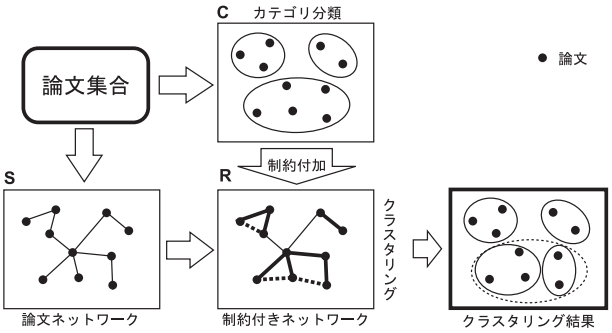


図 1: 制約付きクラスタリング

### 2.2 制約付きクラスタリング

制約付きクラスタリングとはカテゴリによる分類とクラスタリングを組み合わせたもので，論文分類の新しい手法として榊らによって提案された．制約付きクラスタリングの手順を図 1 に示す．図 1 のように，論文集合の関連度によるネットワークを作成し，これにカテゴリ分類の結果を制約として付加することで制約付きネットワークを得る．この制約付きネットワークをクラスタリングすることで，カテゴリ分類との整合性を保ちつつ，現時点の状態を反映した分類結果が得られる．

論文ネットワークの隣接行列を $S$，カテゴリ分類による制約を表した行列を $C$ とすると，制約付きネットワーク $R$ は式 1 のように得られる．式 1 において，r は制約の強さを表すパラメータであり，0 以上 1 以下の実数値をとる．$r = 0$ は制約の無い通常のクラスタリングとなる．

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

連絡先: 水野珠季，同志社大学大学院工学研究科，
〒 610-0394 京都府京田辺市多々羅都谷 1-3
医心館 IN223N，0774(65)6130，
tmizuno@mikilab.doshisha.ac.jp

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

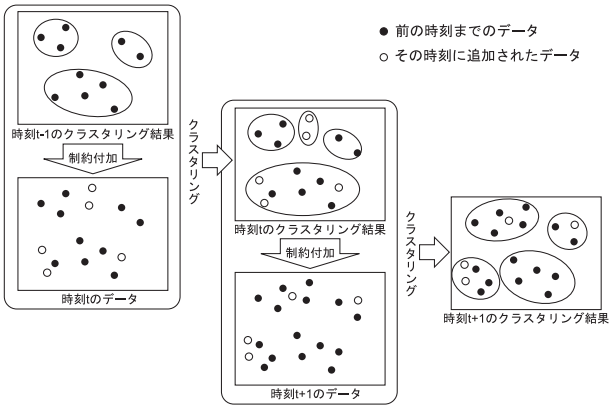


図 2: 過去の時点でのクラスタリング結果を制約とした制約付きクラスタリング

### 2.3 過去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提案

榊らによる制約付きクラスタリングではカテゴリ分類を制約として用いている．本研究では，時刻ごとのクラスタリングにおける分類基準の揺れを抑えるために制約付きクラスタリングを用いる．この場合，あらかじめ分類基準となるカテゴリを作成し，これを制約とすれば良い．しかしこの方法では，どのようなカテゴリに分類するのが適切であるかが未知であるデータには適用できない．また，膨大な量のデータを適当なカテゴリに分類することは大きな手間となる．そこで本稿では，カテゴリ分類の代わりに過去の時点で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を提案する．

提案手法の流れを図 2 に示す．図 2 のように，提案手法では時刻 $t-1$ のクラスタリング結果を制約として時刻 $t$ のクラスタリングを行い，さらにその結果を制約として時刻 $t+1$ のクラスタリングを行う．これによって，分類の基準を維持しつつ各時刻の状態を反映したクラスタリング結果を得る．なお，図 2 からも分かるように時刻 $t$ でのクラスタリングの対象となるデータは，時刻 $t-1$ から $t$ の間に追加されたデータだけでなく，クラスタリング開始時刻から時刻 $t$ までの全てのデータであり，対象データは徐々に増加していくことになる．このようにカテゴリ分類のコストを除くことで，様々なデータへの制約付きクラスタリングの適用を可能にする．

提案手法では，時刻 $t$ のクラスタリングを行う場合，時刻 $t$ の関連度によるネットワークの隣接行列を $S$，時刻 $t-1$ のクラスタリング結果による制約を表した行列を $C$ として，式 1 の制約付きネットワーク $R$ を得る．

## 3. 実データへの適用実験

### 3.1 実験概要

本実験では，提案手法を実データに適用することでその有効性を検証する．実データとして，我々の研究室で研究成果の報告を目的に作成し，公開している研究レポート*2を用いる．

研究レポートの内容は，最適化アルゴリズムに関するものやハイパフォーマンスコンピューティング分野のもの，最近の IT 用語について調査したものなど多岐にわたる．これらのレポートは 2002 年から公開を開始し，2009 年 4 月現在で合計 1,621 本が公開されている．これを提案手法と通常のクラスタリングによって分類し，その結果を比較する．

*2 http://mikilab.doshisha.ac.jp/dia/research/report/2008/

### 3.2 実験目的

提案手法がデータの時系列変化を把握する手法として有効であるかを検証する．具体的には，提案手法が通常のクラスタリングと比較して，各時刻の分類の基準が一定に保たれているか，前後の時刻でレポートが異なるクラスタに分類された場合，その移動がどのように起こっているかといった点について検討する．

### 3.3 実験方法

1. クラスタリング

   2002 年度から 2007 年度にかけて公開された研究レポート計 1,235 本に対し，提案手法を適用する．各年度別の実験に用いたレポート数を表 1 に示す．また，研究レポートに提案手法を適用する手順を図 3 に示す．クラスタリングの間隔は年度ごととし，クラスタリング対象となるレポートは開始年度から徐々に追加されて 2007 年度では 1,235 本全てが対象となる．制約の強さは $r=0.03$ と $r=0.00$，つまり制約のない通常のクラスタリングとを比較する．なお，r=0.03 は文献 [2] で最も適切であるとされた制約の強さである．

表 1: 各年度のレポート数

| 年度 | レポート数 |
|---|---|
| 2002 | 297 |
| 2003 | 274 |
| 2004 | 169 |
| 2005 | 201 |
| 2006 | 197 |
| 2007 | 97 |
| 合計 | 1,235 |

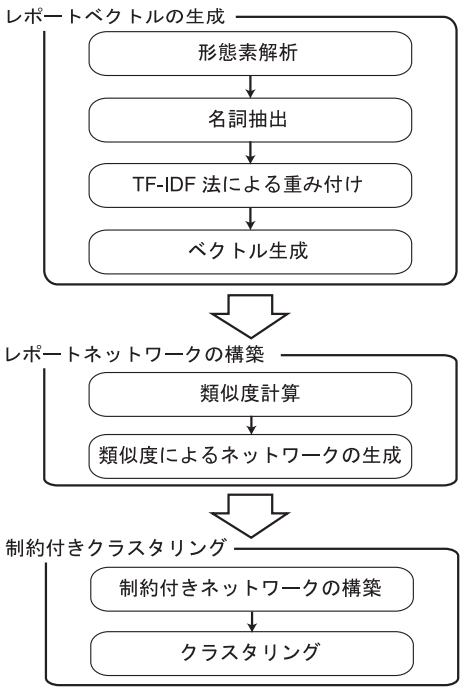


図 3: 研究レポートの制約付きクラスタリング

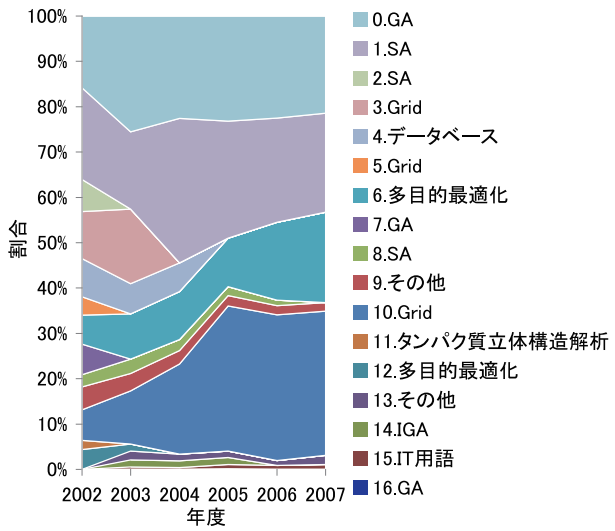


図 4: 制約付きクラスタリング（$r = 0.03$）

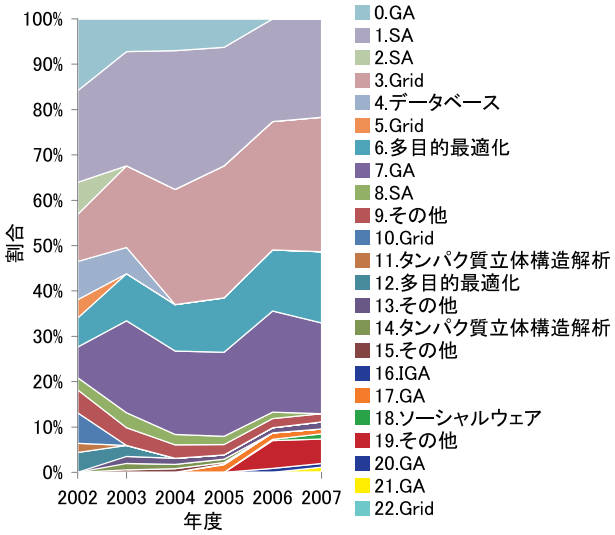


図 5: 通常のクラスタリング（$r = 0.00$）

2. クラスタ間の対応関係の同定
クラスタの変化の流れを追うためには，前後の年度間でクラスタを対応付ける必要がある．そこで，式 2 を用いてクラスタの同定を行う．式 2 は二つのクラスタに含まれるデータ（レポート）のうち，一致しているデータの割合を表しており，これをクラスタの類似度と定義する．年度 $y$ と $y+1$ の全てのクラスタの組合せに対してクラスタの類似度を求め，この類似度が最も大きいものから順に同一のクラスタとする．

$$sim(cluster_1, cluster_2) = \frac{|cluster_1 \cap cluster_2|}{|cluster_1 \cup cluster_2|} \quad (2)$$

3. クラスタのラベル付け
クラスタリング結果の考察を容易にするため，各クラスタがどのような内容のレポートから構成されているかを表すラベルを人手で付与する．今回は我々の研究室の研究グループの構成を元に，以下の 10 種類のラベルを使用する．複数のラベルに対応するレポートが含まれる場合，最も多く含まれるものをそのクラスタのラベルとして付与する．

- SA
- GA
- Grid
- 多目的最適化
- IGA
- タンパク質立体構造解析
- データベース
- ソーシャルウェア
- IT 用語
- その他

### 3.4 実験結果と考察

2002 年度から 2007 年度までのクラスタリング結果の変化の様子を図 4 と図 5 に示す．図 4 は制約付きクラスタリングによる結果，図 5 は通常のクラスタリングによる結果である．横軸は年度，縦軸はレポート数の全体に対する割合である．また，凡例にはクラスタの番号とラベルを示した．一つの帯はクラスタの同定によって得られた一連のクラスタを表す．例えば図 4 のクラスタ 10 は Grid に関するレポートを中心としたクラスタであり，年度が進むにつれレポート数が徐々に増加していったということが読み取れる．

- 分類基準の揺れ
図 5 では，2006 年度にクラスタ 19 が新たに出現した．このクラスタは，文献リストという研究テーマごとの参考文献をリスト形式でまとめたレポートから構成されるクラスタであった．このクラスタの出現は，2005 年度まではレポートの内容によって分類されていたものが文献リストの数が増加したことでレポートの形式によって分類されてしまった結果である．このように制約が無い通常のクラスタリングでは年度ごとに分類の基準に揺れが生まれ，データの時系列変化を把握するには適さないことが分かる．

- クラスタ間の移動
クラスタの同定の際に求めた類似度の平均値は，表 2 のようになった．この平均値について Wilcoxon の順位和検定を行った結果，有意水準 5%で平均値に有意な差があった．つまり，制約付きクラスタリングは通常のクラスタリングに比べてレポートのクラスタ間の移動が少ないといえる．通常のクラスタリングの場合にクラスタ間の移動が起こったレポートについて，制約付きクラスタリングではどのような結果になったのかを，例を挙げて見ていく．通常のクラスタリングの場合にクラスタ間の移動が起こったレポートの例を表 3 に示す．

表 2: クラスタの類似度の平均値

| | |
|---|---|
| 制約付きクラスタリング（$r = 0.03$） | 0.44 |
| 通常のクラスタリング（$r = 0.00$） | 0.27 |

表 3: クラスタ間の移動が起こったレポートの例

| | タイトル |
|---|---|
| 1 | 遺伝的アルゴリズムによるノード数を増やした複雑ネットワーク問題の検討（追加実験）<br>遺伝的アルゴリズムによるノード数を増やした複雑ネットワーク問題の検討<br>遺伝的アルゴリズムによる複雑ネットワーク問題の解法における交叉方法の検討【追加実験】<br>遺伝的アルゴリズムによる複雑ネットワーク問題の解法における交叉方法の検討<br>遺伝的アルゴリズムによる複雑ネットワーク問題の解法における初期個体の発生方法の検討 |
| 2 | PSA/AN(GA)2 における交叉率およびプロセッサ数の検討<br>逐次 SA の性能比較〜対象問題：Rastrigin 関数，Rosenbrock 関数〜<br>Adaptive Simulated Annealing の性能検証 |

– 制約が有効に働いた例
表 3 の 1 に示す 5 本のレポートはいずれも，遺伝的アルゴリズム（GA）を用いて複雑ネットワーク問題を解くことに関するものである．これらのレポートは通常のクラスタリングの場合に，2006 年度までは GA のクラスタ（図 5 のクラスタ 7）に属していたが，2007 年度に Grid のクラスタ（図 5 のクラスタ 3）に移動していた．これは，2006 年度までは GA という手法によって分類されていたものが 2007 年度では複雑ネットワーク問題という対象問題によって分類された結果 Grid のクラスタに移動したと考えられる．しかし制約付きクラスタリングの場合にはクラスタ間の移動は無く，一貫して GA のクラスタ（図 4 のクラスタ 0）に属していた．このように，過去のクラスタリング結果を制約として用いることでクラスタ間の移動が抑制され，分類基準の揺れが生じにくくなったと考えられる．

– 制約が有効に働かなかった例
表 3 の 2 の場合は，SA に属すべきレポートが GA のクラスタに属してしまった例である．通常のクラスタリングの場合は 2006 年度に GA のクラスタ（図 5 のクラスタ 7）から SA のクラスタ（図 5 のクラスタ 1）への移動が起こったことで適当なクラスタに属したが，制約付きクラスタリングの場合にはこの移動が起こらなかった．

以上のように，制約付きクラスタリングでは通常のクラスタリングに比べてレポートの移動が少なく，分類の基準が保たれていた．この点において提案手法はデータの時系列変化を把握するために有効であると言える．ただし，不適切なクラスタに属してしまった状態が維持される場合があることも分かった．この問題への対処法としては，まずクラスタリングの精度を向上させて不適切なクラスタに属すことを防ぐ必要があるが，そのほか，制約の強さを調整することによっても解決される可能性がある．

## 4. 関連研究

データの時系列変化を分析する手法としては，福井らによる Sb-SOM[3] や山村らによるスパイダーアルゴリズム [4] などすでに様々な研究が行われている．しかしこれらの手法が扱うデータは，コンテンツ自体が時系列変化するものであり，本研究の目的は文書のような静的なコンテンツが集合として時系列変化する様子を把握することであるため，対象とするデータが異なっている．

また，ニュースなどの時系列文書からトピックを抽出し，その推移を捉えるといった研究も盛んに行われている [5][6][7][8]．しかしこれらは，大量のデータから短期的なトピックの移り変わりや最新のトピックを容易に把握することを目的としており，長期的な時系列変化の把握によって現状理解や今後の予測に役立てることを目指す本研究とは，目的が異なっている．

## 5. 結論

本研究の目的は，現状理解や今後の予測に役立てるため，論文やニュース記事といった静的なコンテンツの集合の長期的な時系列変化を把握することである．本稿では，制約付き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手法を提案した．提案手法を研究レポートに適用した結果，制約をかけることで，通常のクラスタリングと比較して分類の基準を一定に保つことができ，データの時系列変化を把握する手法として有効であることが分かった．しかし，制約の強さを変化させた場合の効果については，さらに検討が必要である．また，今後はより長期のデータを用いて，クラスタの消滅や出現，分裂，融合がどのように起こるか，それらの現象を把握し，現状理解や今後の予測に役立てることが可能かといった検討を行っていく．

## 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, 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, JSAI2006, pp.1-4(2006).

[3] 福井 健一, 斉藤 和巳, 木村 昌弘, 沼尾 正行. 自己組織化ネットワークによる動的クラスタの可視化編纂. 人工知能学会論文誌, 23(5), pp.319-329(2008).

[4] 山村 雅幸, 亀田 祥平. 時系列クラスタリングのためのスパイダーアルゴリズム (機械学習によるバイオデータマインニング). 電子情報通信学会技術研究報告. NC, ニューロコンピューティング, 106(101), pp.61-64(2006).

[5] James Allan, Jaime. Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic Detection and Tracking Pilot Study : Final Report. Proc. DARPA Broadcast News Transcription * Understanding Workshop, 1998, pp.194-218(1998).

[6] 森 正輝, 三浦 孝夫, 塩谷 勇. 時制クラスタのトピック追跡. データ工学ワークショップ DEWS, (2006).

[7] 平田 紀史, 大囿 忠親, 新谷 虎松. ユーザの選好に基づくトピック分析システムの試作. 人工知能学会全国大会論文集, JSAI2008, (2008).

[8] 菊池 匡晃, 岡本 昌之, 山崎 智弘. 階層型クラスタリングを用いた時系列テキスト集合からの話題推移抽出. データ工学ワークショップ DEWS, (2008).