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 -制約付きクラスタリングの効果の検討-

水野 珠季

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．蓄積された情報の活用法としては，情報推薦やトピックの抽出と追跡など，すでに様々な研究がなされている．

本研究では，論文やブログ・ニュースの記事などの時間とともに増加する静的なコンテンツの集合に着目した．これらのデータは，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すると考えられる．論文の例であれば，研究分野の成長や縮小，分裂，新規分野の出現といった変化があるだろう．このようなデータの時系列変化を把握することが可能となれば，データに対する理解を深めることや今後の変化の予測に役立つと考えられる．そこで本研究では，徐々に増加する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る．

本稿では，制約付きクラスタリング[2]によって徐々に増加する静的コンテンツの集合の時系列変化を把握する方法を提案し．制約付きクラスタリングを用いた場合の効果について通常のクラスタリングと比較し，検討を行う．

2　データの時系列変化の把握

データの時系列変化を把握するには，まずそのデータがそれぞれの時点でどのようなコンテンツから構成されているのか，つまり各時点でのデータの全体像を知る必要がある．なお，ここでいうデータとは前節でも述べた，徐々に増加する静的コンテンツの集合である．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，個体の集合を個体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．そこで，クラスタリングをある時間間隔で繰り返し行い，得られたサブグループの変化からデータの時系列変化を把握することを考えた．

しかしクラスタリングでは，個体間の関連度のみに基づいて分類が行われるため，Fig. 1 のように少量のコンテンツの追加でも全く異なる分類結果が得られてしまうことがあり，このままでは時系列変化を把握するには適さない．この問題を解決する方法として，本研究では制約付きクラスタリングに着目した．

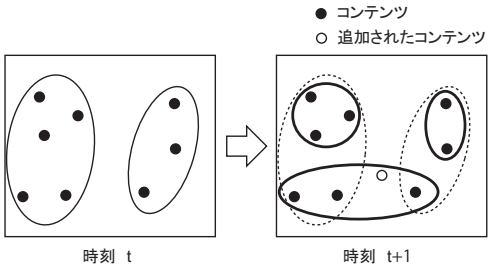

Fig.1　クラスタリングによる時系列変化の把握の問題点

3　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

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，それぞれの問題点を解決して時系列的な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]．制約付きクラスタリングの手順を Fig. 2 に示す．Fig. 2 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て，これをクラスタリングする．

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．$r = 0.0$ は制約の無い通常のクラスタリングとなる．

$$R = (1 - r)S + rC \qquad (0.0 \leq r \leq 1.0) \qquad (1)$$

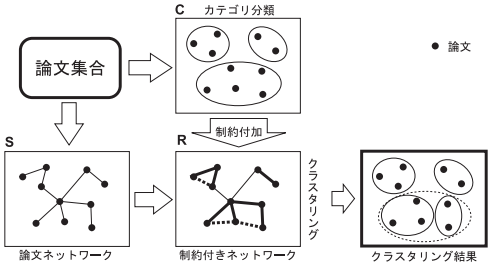

Fig.2　榊らの提案する制約付きクラスタリング

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が，本研究ではカテゴリ分類の変わりに直前の時刻のクラスタリング結果を制約として，連続的にクラスタリングを行うことを考えた．

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

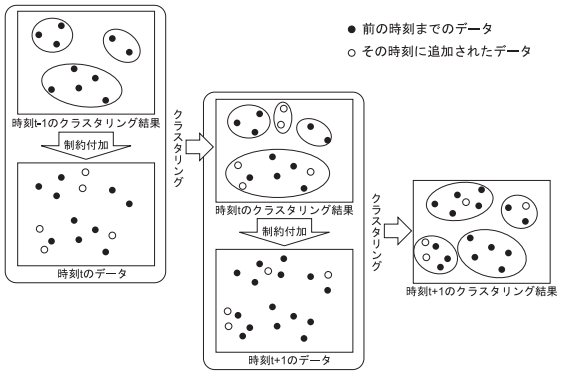

Fig.3　直前の時刻のクラスタリング結果を制約とした制約付きクラスタリング

本研究での制約付きクラスタリングの流れを Fig. 3 に示す．Fig. 3 のように，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．これによって時系列のつながりを考慮に入れたクラスタリング結果が得られるため，データの時系列変化が把握できるのではないかと考えた．なお，Fig. 3 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．このように直前の時刻のクラスタリング結果を制約とすることで，次々と追加されるデータをカテゴリ分類する手間が無くなるだけでなく，カテゴリが未知なデータに対しても制約付きクラスタリングを適用することが可能となる．

4　制約付きクラスタリングの効果の検討

今回，直前の時刻のクラスタリング結果を制約とすることで 2 節で述べたクラスタリングによる時系列変化の把握の問題点を解消することができるかという点について検討を行った．

検討方法としては，ノード数 10，20 の重み付きネットワークそれぞれ 3 つに対して，制約の強さを $0.0 \leq r < 1.0$ の範囲で変化させてクラスタリングを行い，その結果について通常のクラスタリング（$r = 0.0$）と比較した．今回用いたネットワークは GA によって $r = 0.0$ のクラスタリング結果と $r = 0.5$ のクラスタリング結果の違いが大きくなるように最適化したネットワークであり，エッジの重みは 0，0.25，0.5，1 の 4 種類，時刻は最大で 0〜3 の 4 つである．

Table 1 にノード数 10 のネットワークでの通常のクラスタリング（$r = 0.0$）と制約付きクラスタリング（$r = 0.1$）の各時刻のクラスタリング結果を示す．この例では，time0, 1 では両者の結果に差はなく，time2 で初めて異なるクラスタリング結果になっている．この time2 の結果について time1 の結果からの変化を見ると，通常のクラスタリングでは time1 で同じクラスタであったノード 0 とノード 3，ノード 1，4 とノード 2，5 が time2 ではそれぞれ別のクラスタに分類されてしまっている．

一方，制約付きクラスタリングでは time1 でのクラスタを保ったままそれぞれのクラスタに追加されたノードが分類されており，時系列のつながりを考慮に入れたクラスタリング結果が得られたと言える．同様の傾向は他の 5 つのネットワークでも見られた．このことから，制約付きクラスタリングによって前後の時刻で全く異なった分類結果になってしまうというクラスタリングによる時系列変化の把握の問題点を解消することができると考えられる．しかし，今回用いたデータはノード数が 10，20 と小規模なデータであったため，ノード数を増やしたデータでも同様の傾向が見られるか確認する必要がある．

Table1　クラスタリング結果の比較

	r=0.0	r=0.1
time0	[0 3][1 2]	[0 3][1 2]
time1	[0 3][1 2 4 5]	[0 3][1 2 4 5]
time2	[0 1 4 6 7][2 3 5 8 9]	[0 3 6 9][1 2 4 5 7 8]

また，制約の強さを変化させたところ，最終時刻の結果だけを見てもそれぞれのネットワークで 10 種類前後の異なるクラスタリング結果が得られていたことが分かった．クラスタリングの結果には明確な正解というものは無く，これらのクラスタリング結果の優劣を決めたり，最適な結果を決めることは難しい．しかし，明確な正解が存在しないからこそ複数のクラスタリング結果を提供できることで新たな知見が得られる可能性があり，この点でも制約付きクラスタリングは通常のクラスタリングよりも有用であると言える．しかしこの点についても，得られる結果の種類はデータに依存し，通常のクラスタリングとあまり差がでない場合もあることが分かっており，さらに検討する必要がある．

5　今後の課題

前節では GA で作成した小規模なネットワークに対して制約付きクラスタリングを行い，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有用である点について検討した．今後は，大規模なネットワークでの検討や，jaccard 係数などを用いて制約付きクラスタリングと通常のクラスタリングで数値的にどの程度の差が見られるか，ネットワークの特性によって結果にどのような違いがあるかといった多くの点についてさらに検討する必要がある．また，論文やブログ記事などの実データに適用して，そのデータの時系列変化の把握が可能であるか検討する必要がある．

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).