ISDL レポートの分類のためのクラスタリング手法の検討

水野 珠季

1　はじめに

現在，我々の研究室では，学生が研究報告や文献調査の結果を ISDL レポートと呼ばれる HTML 形式のレポートにまとめ，Web 上で公開している．これまでに 1300 以上のレポートが公開されており，これらのレポートは研究室内外で有用な研究資料として活用されている．本研究では，これらのレポートをクラスタリングして得たキーワードを可視化することで，我々の研究室での研究の全体像や内容，それらの時間による変化などの理解をめざす．これらを実現し，論文やブログなどに応用することで，個人やコミュニティにおける研究テーマや興味・関心の移り変わりを捉えることが可能になると考えられる．本報告では，ISDL レポートの分類に用いるクラスタリング手法について検討する．

2　研究理解支援システム

現在の ISDL レポートを集めたポータルサイトは，作成年度と著者による分類しかなされていない．そのため，どのような内容のレポートがどれだけ存在するのか，研究テーマがどのように変化してきたかなど，研究の全体像やその変化を把握することが困難である．本研究では，ISDL レポートに時間変化を考慮したクラスタリングを適用し，得られたキーワードを用いて我々の研究室が扱う研究テーマの変遷を可視化することによって，研究の全体像の把握や内容理解を支援するシステムの構築を目指す．

提案システムでは，ISDL レポートの集合に対して，一定の時間間隔で文書クラスタリングを行う．このとき，研究テーマの変遷を追うためには，ある時点でのクラスタリング結果がそれより過去の時点でのクラスタリング結果を考慮したものでなければならない．そのため，本研究では時間変化を考慮したクラスタリングの手法として，榊ら [1] が提案する制約付きクラスタリングを用いる．

3　文書クラスタリング

提案システムでは，文書クラスタリングが重要な位置を占める．文書クラスタリングとは，図書や雑誌論文などの文書の集合を内容が均質ないくつかの群に分類することである [2]．提案システムでは，ISDL レポートの文書クラスタリングを以下のような手順で行う．

1) 文書に対して形態素解析を行う
2) 文書の特徴語（キーワード）を抽出する
3) キーワードを用いて，文書間の関連度を計算する
4) 関連度を重みとして，文書集合のネットワークを作成する
5) 作成したネットワークと過去のクラスタリング結果を用いて制約付きクラスタリングを行う
6) 1)〜5) を一定の時間間隔で繰り返す

4　制約付きクラスタリング

制約付きクラスタリングとは，文書集合をクラスタリングする際に，過去の時点でのクラスタリング結果を制約として付加することで，時間変化を考慮したクラスタリングを行うというものである．単純に時間ごとに独立にクラスタリングを行うと，Fig. 1 のように時間ごとに全く異なるクラスタリング結果となる可能性があり，過去の時点からの変化を追うことが難しい．そこで，文書集合のネットワークに Fig. 2 のように過去の時点でのクラスタリング結果を制約として付加した制約付きネットワークを作成し，そのネットワークに対してクラスタリングを行う．本報告では，制約付きネットワークに適用するクラスタリングの手法について検討する．

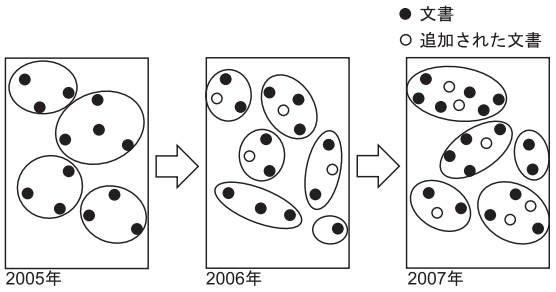


Fig.1　時間ごとの文書クラスタリングにおける問題（出典：文献 [1] を参考に自作）

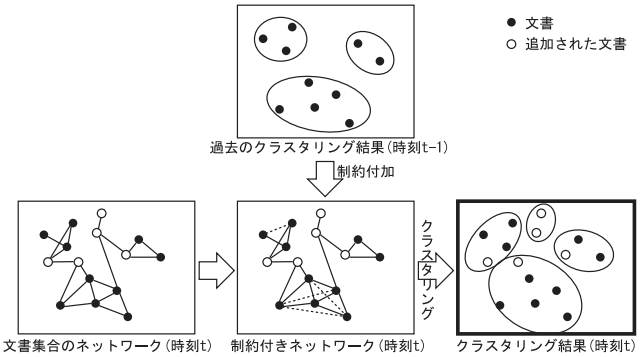


Fig.2　制約付きクラスタリング（出典：文献 [1] を参考に自作）

5　クラスタリング手法の検討

クラスタリング手法としては，階層的クラスタリングや非階層的クラスタリングのほか，主成分分析などの次元縮約法や確率モデルに基づく方法などが考えられる．本研究で対象とする ISDL レポートのクラスタリングで

は，分類されるべきクラスタ数が未知である．また，レポートの数が 1300 以上と多いうえ，一定の時間間隔で何度もクラスタリングする必要がある．そのため，クラスタ数を自動的に決定するメカニズムを持ち，計算量も小さい，Newman[3] が提案するネットワークからのコミュニティ抽出の手法（以下，Newman 法と呼ぶ）を用いる．この手法は，榊ら [1] の制約付きクラスタリングでも利用されている．以下では，Newman 法の概要とそのアルゴリズムについて説明し，実装した結果について述べる．

5.1 Newman 法の概要

Newman 法は，凝集型階層的クラスタリングの一手法である．評価関数 modularity Q を最大化することによって，最適なクラスタリング結果を得る．最悪のケースでも，全体の計算量は $O((m+n)n)$ または $O(n^2)(n$ はノード数，m はエッジ数) と高速で，クラスタ数が自動的に決定されるため，近年，SNS やブログのネットワーク分析などに多く応用されている．

5.2 modularity Q

modularity Q とは，ネットワークのモジュール性を評価する指標である．Newman 法では，この Q の値が大きくなるようなクラスタ構造が最もらしいとされている．Q の値が大きいとは，クラスタ内に存在するエッジの本数が多く，クラスタ間をつなぐエッジの本数は少ないという状態を指しており，modularity Q は式 (1) のように定義される．

$$Q = \sum_i (e_{ii} - a_i^2) \tag{1}$$

式 (1) において，e_{ii} はクラスタの隣接行列 e の対角要素である．行列 e の要素 e_{ij} は，クラスタ i とクラスタ j を結ぶエッジ数の総エッジ数に対する割合を表す．つまり，e_{ii} はクラスタ i 内のエッジ数の割合である．a_i は行列 e の i 行の和，つまり，クラスタ i の持つ全てのエッジ数の割合である．

また，統合するクラスタの決定には，Q の増加量 ΔQ が用いられる．クラスタ i とクラスタ j を統合した際の Q の増加量 ΔQ は式 (2) のように定義される．

$$\Delta Q = 2(e_{ij} - a_i a_j) \tag{2}$$

5.3 アルゴリズム

Newman 法のアルゴリズムを以下に示す．また，Fig. 3 に Newman 法の進行の様子を示す．

1) 初期状態として，ネットワーク内のすべてのノードをそれぞれ構成要素が 1 のクラスタとする
2) すべての 2 つのクラスタの組合せについて，統合した場合の ΔQ を計算する
3) ΔQ が最大となるクラスタの組合せを統合する
4) 得られたクラスタ構造について Q の値を計算する
5) クラスタ数が 1 となるまで 2)〜4) を繰り返す
6) Q の値が最大となるクラスタ構造を結果とする

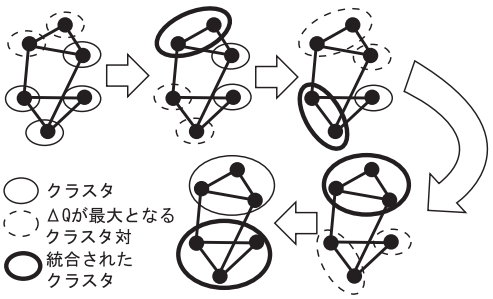
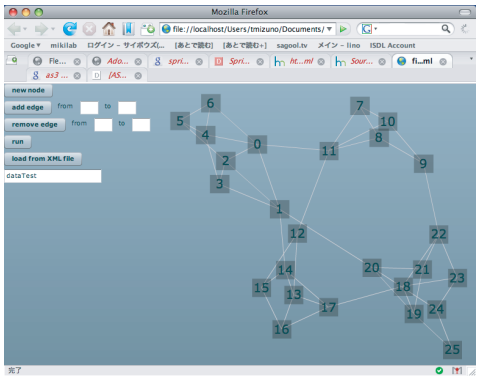


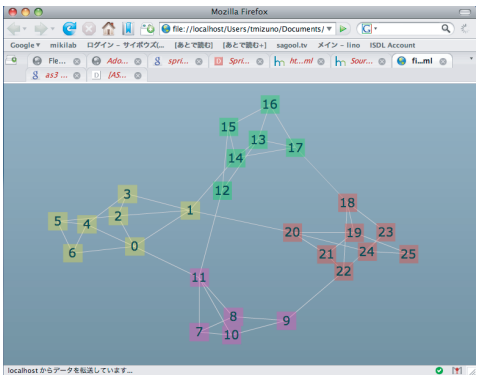
Fig.3　Newman 法の進行イメージ（出典：自作）

5.4 実装結果

Newman 法を Python で実装し，テストデータで動作を確認した．また，Flex を用いてクラスタリング結果の可視化を行った．可視化システムのインタフェースを Fig. 4 に示す．Fig. 4（a）の画面において，ボタンをクリックしてノードやエッジを入力し，ネットワークを作成することができる．また，XML ファイルからネットワークを読み込むことも可能である．



(a) 入力画面

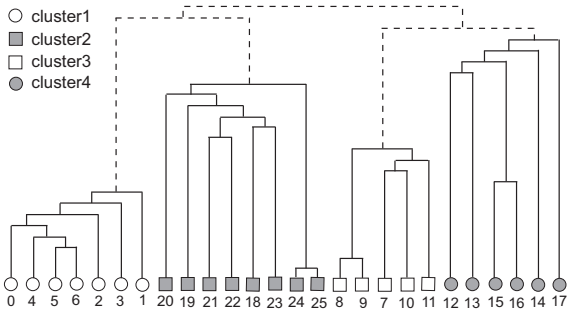


(b) 結果画面

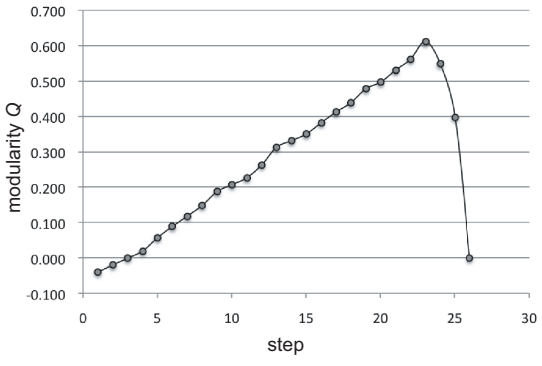
Fig.4　クラスタリング結果の可視化システム（出典：自作）

テストデータは自作のものと，Newman[3] が評価に用いていた Zachary's karate club のものを使用した．それぞれ，クラスタリングの進行の様子を樹形図で表したものと，Q の値の遷移を Fig. 5, Fig. 6 に示す．

Fig. 5, Fig. 6 ともに（a）において，樹形図の末端の記号は理想的なクラスタリング結果を示しており，樹形図の実線部分が実際のクラスタリング結果を示している．また（b）のグラフは，縦軸が Q の値，横軸はクラスタ統

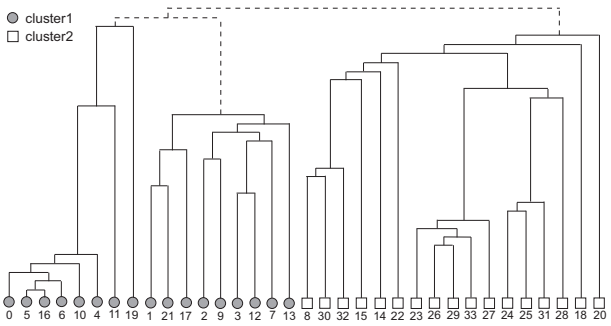


(a) クラスタリング結果

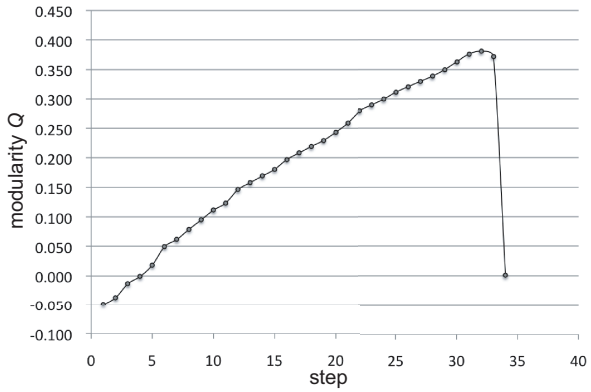


(b) Q の推移

Fig.5　自作テストデータによる結果（出典：自作）



(a) クラスタリング結果



(b) Q の推移

Fig.6　Zachary's karate club による結果（出典：自作）

合のステップ数を表す．自作のテストデータでは，想定した通りのクラスタリング結果が得られている．しかし，Zachary's karate club では，Newman[3] の文献では 2 つのクラスタに分かれていたが，3 つのクラスタに分かれた段階で Q の値が最大となってしまっている．これは，本来分かれるべきクラスタ内での結合の度合いが弱い場合のプログラムの精度に問題があるためだと考えられる．

今後は，Newman[3] の文献にあるように，コンピュータで自動生成されたネットワークデータを用いて以下のようにクラスタリング精度の評価を行う予定である．

1) ノード数は 128 とし，あらかじめ 4 つのグループ（各 32 ノード）に分けておく
2) あるノードから同じグループ内の他のノードへのエッジ数の平均を z_{in}，他のグループ内のノードへのエッジ数の平均を z_{out} とする（ただし，$z_{in} + z_{out} = 16$ とする）
3) z_{out} の値を 0 から徐々に増やし，グループ内での結合の度合いを弱くしていく
4) グループ内のエッジ数とグループ間のエッジ数の割合によって，クラスタリングの精度がどのように変化するのかを観察する

6　まとめと今後の課題

本報告では，研究の全体像や研究テーマの変遷の理解の支援を目指し，ISDL レポートのクラスタリング手法として，制約付きクラスタリング，Newman 法について検討した．制約付きクラスタリングは時間によるクラスタリング結果の変化を考慮しており，ISDL レポートに適用することで研究テーマの変遷を把握できる可能性がある．また，Newman 法は modularity Q を最大化することでネットワークのクラスタリングを行う，凝集的階層的クラスタリングの手法である．Newman 法を実装した結果，クラスタ内での結合の度合いが強いネットワークの場合，問題なくクラスタリングされていることが確認できた．しかし，クラスタ内での結合の度合いが弱いネットワークの場合のクラスタリング精度について調査する必要があることがわかった．

今後は，ISDL レポートに文書クラスタリングを適用し，適切なキーワードやクラスタリング結果が得られるか検討する．その後，Newman 法に基づいて制約付きクラスタリングを実装し，可視化システムを構築して，ISDL レポートのポータルサイトに組み込む．

参考文献

1) 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. 人工知能学会第 20 回全国大会論文集,1A1-1(2006)
2) 岸田 和明, 文書クラスタリングの技法：文献レビュー
Techniques of Document Clustering: A Review
http://wwwsoc.nii.ac.jp/mslis/pdf/LIS49033.pdf
3) M.E.J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E 69, 066133(2004)
4) 大西 祥代, 廣安 知之, 三木 光範, 文書クラスタリングの基礎
http://mikilab.doshisha.ac.jp/dia/research/report/2007/0913/004/report20070913004.html