DEIM Forum 2015 E6-6

時系列グラフクラスタリングによるトレンド分析

岸田　吉弘† 　塩川　浩昭 †† 　鬼塚　真 †

† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1-5
†† 日本電信電話 (株) NTT ソフトウェアイノベーションセンタ　〒 180–0012 東京都武蔵野市緑町 3-9-11
E-mail: †{kishida.yoshihiro,onizuka}@ist.osaka-u.ac.jp, ††shiokawa.hiroaki@lab.ntt.co.jp

あらまし　大規模なデータの登場に伴い，大量のデータに対する解析は重要な技術となっており，時間変化を把握するためのトレンド分析は有用な解析手法のひとつである．しかし，トレンド分析に必要な分析工程と可視化工程は連携がなされておらず，開発コストがかかってしまうという問題と，クラスタの年代長を事前にユーザが指定する必要があるという問題があり，データを活用する上での障壁となっている．そこで本稿では，分析技術と可視化技術を適切に組み合わせてトレンド分析をするシステムの開発に取り組む．提案システムでは，クラスタリングの分析結果をシームレスに可視化できるよう可視化工程との連携を行い，ユーザがデータを事前に分割せずにクラスタ分割と年代分割を同時に実行するようクラスタリングを行う．本稿では，情報不純度を表すエントロピーの値が最小となり，トレンドに応じて異なる年代長となるようなクラスタの抽出を行った．評価実験から，変遷を分析するためにはクラスタリング後にエントロピーに基づいてクラスタ分割を行う手法が有効であることを示した．さらに，クラスタ内分析とクラスタ間分析が行えるよう 2 種類の手法で可視化を行い，クラスタの分析を行った．

キーワード　グラフクラスタリング，時系列，可視化

1. はじめに

世の中では Data is the new oil （データは新たな石油資源である）という言葉に代表されるように，多量のデータを資源として捉えて新たな知識を発見する技術に注目が集まっている．また近年では，数億ノードから構成される大規模なグラフデータが登場してきた．グラフデータはデータをノードとエッジで表現した基本的なデータ構造であり，情報推薦や情報検索，科学データ分析など様々な分野で利用されており，大規模なグラフデータに対する解析処理技術への需要が高まっている．例えば，Facebook では 2014 年に 1 日当たりのアクティブユーザ数が 8.02 億人であることが報告されており (注1)，Twitter では 1 日当たり平均 5 億件のツイートが投稿されていることがわかっている (注2)．このように，大規模のデータは現実に存在し，今後もその規模をさらに増大させていくことが考えられる．そして，これらのデータに対する解析手法は将来的に必要不可欠な技術となってくると言える．

大量のデータの概要を把握するためには，ある尺度に従ってデータをいくつかのグループに自動分類するクラスタリング技術が重要であり，これまで様々なクラスタリング手法が研究されてきた [1], [2], [4]．その中でも特に，トレンド分析はクラスタリング結果の時間的な変化を分析する手法であり，様々な分野において有用な手法である [5], [6]．例えば，マーケティングや経済データの分析，具体的にはメガバンクの統廃合の分析等に用いられているほか，幕府や政権の存続年数の分析や，各時代における人口の変遷の分析などにも使われる．さらに，技術変遷の分野においてもトレンド分析は重要な分析手法である．例えば，スキーマ統合や分散データベースの研究の前後は研究分野は何があったか，論文数が年代によってどのように変化していったかを分析することが可能である．実際に電子情報通信学会では数年に一度，専門委員会ごとに専門家が集まって技術年表を作成している．クラスタ化と可視化はすべて人手で行われており，作成に必要な論文の数は年々増えていくため，非常に多量の時間を要している．

トレンド分析には，データを分析する工程と分析結果を可視化する工程の 2 つの工程が必要である．しかし，既存技術には，分析工程において，クラスタの年代長を予めユーザが指定する必要があるという問題がある．そのため，本来トレンドが持つ年代長を適切に抽出できないことがある．例えば，技術変遷において，スキーマ統合の研究は 10 年程度の期間なのに対し，XML ストリーム処理の研究は 5 年程度の期間であるため，ユーザが年代長を指定する従来の技術では，このように異なる長さのトレンドを抽出することが難しい．

そこで本稿では，既存技術である分析技術と可視化技術の双方を適切に組み合わせてトレンド分析をするシステムの開発に取り組む．提案システムでは，上記の問題点を解消するため，ユーザがクラスタの年代長を指定することなく，トレンドに応じた年代長のクラスタが抽出でき，かつ情報不純度を表す指標エントロピーが小さくなるクラスタリング手法を提案する．本稿では，提案システムのプロトタイプを実装し，実データを用いて評価実験を行った．この結果，グラフクラスタリングにより抽出されたクラスタが，年代長がトレンドによって異なる値をとり，かつエントロピーの小さい値を示すことを確認した．さらに，抽出したクラスタを，2 種類のグラフで可視化できるシステムを作成し，クラスタの分析を行った．

(注1)：http://investor.fb.com/releasedetail.cfm?ReleaseID=842071
(注2)：https://biz.twitter.com/ja/whos-twitter

本稿の構成は以下のとおりである．まず，2 章で研究の前提となる知識について概説し，3 章で提案システムの説明を行う．4 章で提案システムの評価と分析を行い，5 章で関連研究について述べた後，最後に 6 章にて本稿のまとめと今後の課題について述べる．

2. 事 前 準 備

2.1 クラスタリング指標 Modularity

本稿の提案システムでは，クラスタリング結果の精度を評価するために，Modularity [8] という指標を用いる．クラスタリング指標 Modularity について概説する．Modularity は，直感的にはクラスタ内に含まれたノード間のエッジが密であり，クラスタ間に存在するエッジが疎となる程良い値を示す指標で，コミュニティ構造を抽出して分析する手法として近年注目を集めている手法である．グラフクラスタリング手法により抽出したクラスタの集合を C，クラスタ i からクラスタ j へ接続されているエッジ数を e_{ij}，グラフ全体に含まれる総エッジ数を m とするとき，Modularity Q は以下のように定義される．Modularity が負の値を取る場合は $Q = 0$ とし，常に Modularity Q は正の値を示す．本研究では，この Modularity を用いたクラスタリング手法を対象とし，最終的に Modularity が上昇しなくなるまでクラスタの統合を行う．

$$Q = \sum_{i \in C} \left\{ \frac{e_{ii}}{2m} - \left(\frac{\Sigma_{j \in C} e_{ij}}{2m} \right)^2 \right\} \tag{1}$$

2.2 平均情報量（エントロピー）

本稿の提案システムでは，抽出されたクラスタの評価と，時系列方向でのクラスタの分割に，エントロピー [9] という指標を用いる．エントロピーについて概説する．エントロピーは，情報の無秩序さや曖昧さ，不確実さを表す尺度であり，情報が不規則であればあるほど，平均として多くの情報を運んでいることを意味する．ある観点でデータを分割した時に，情報が最も綺麗に分かれた場合は 0，最も不純である場合には 1 となる．

エントロピーは以下のように求めることができる．属性 A を使うことにより，集合 S が部分集合 $S_1, S_2, ..., S_v$ に分割されたとする．S_i がクラス P の要素を p_i 個，クラス N の要素を n_i 個含んでいたとすると，全体のエントロピー，すなわちすべての部分集合 $S_1, S_2, ..., S_v$ に属する対象を分類するのに必要な情報量は，以下で与えられる．

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i) \tag{2}$$

クラスタのエントロピー I は以下の式で算出する．ただし，c_i はクラスタに含まれるノード P_i に接続されているエッジの本数を表すとする．

$$P_i = \frac{c_i}{c_1 + c_2 + ... + c_n} \tag{3}$$

$$I(P_1 ... P_n) = \sum_{i=1}^{n} -P_i log_2 P_i \tag{4}$$

さらに，属性 A によってクラスタが分割されるとき，分割後のクラスタに含まれているデータを更に分類するのに必要な情報量 $E(A)$ は以下の式で算出する．$E(A)$ は，分割後のクラスタの情報不純度を表している．属性 A で分割することによって得られる情報量ゲインは，以下の式で算出する．

$$gain(A) = I - E(A) \tag{5}$$

3. 関 連 研 究

3.1 時系列クラスタリング

トレンド分析に関する技術としては，時系列クラスタリング [1], [2], [3] や，時系列トピックモデル [4], [5], [6] がある．[1] では，クラスタの連続性を重視するために時間スムーズを取り入れ，[2], [3] では時間スムージングのパラメータを自動調整しているが，いずれもグラフをどう時間分割するかは固定で，利用者がデータを時間方向に手動で分割し，分割した時間区間（年代）毎にクラスタを導出している．さらに，いずれもグラフは単一オブジェクトグラフが対象であるため，複数オブジェクトを持つデータを扱うことができないという問題がある．[5] では論文のトピックを一年単位で抽出し，トピック毎の単語分布と論文ごとのトピック分布の影響度を捉えることで動的なトピックの抽出を行っており，[6] ではトピックモデルを使って技術領域を見つけ，関連するトピックをコサイン類似度を用いて関係づけている．時間の異なるトピックを関連付ければ，時系列変遷として構築することができる．また，[10] では，時系列クラスタリングの中でも特に技術動向の変遷を見るために技術年表の構築を行っている．この手法では，時系列方向でデータを分割後，時系列分割で得られる年代毎にデータを分野方向で分割してクラスタを得て，最後にクラスタ間の変遷（年表のつながり）を作成している．

3.2 可視化技術

データを可視化するライブラリには，prefuse(注3) や protovis(注4)，D3(注5) などが存在する．そして，具体的な可視化技術として，ストリームグラフやサンキーダイアグラムなどがある．ストリームグラフ（Streamgraph）は，時系列の連続的なデータを積み重ねの面グラフによって表現する図表であり，可視化対象の分類が明確な場合に用いられる．例えば，Twitter で，ある単語が日に応じてどれだけ呟かれたかをグラフ化したり，石油や石炭などのエネルギーが年と共にどれだけ使用量が変化しているのかといったことを可視化することができる [12]．

サンキーダイアグラム（Sankey diagram）は，ノード間の流量を表現する図表である．そのため，ノード間の単なるつながりのみでなく，つながりの度合いを把握することができる．さらに，エッジが近くなるように自動でノードを配置するという特徴を持つため，変遷がわかりやすい．化学工学，環境工学，物流管理等の分野でシステム内の各プロセス間の流量を表す為に使われ，エネルギー収支の視覚化に頻繁に使われている [13]．

(注3)：http://prefuse.org/

(注4)：http://mbostock.github.io/protovis/

(注5)：http://d3js.org/

4. 提案システム

本章では，分析技術と可視化技術を適切に組み合わせたトレンド分析システムについて述べる．まず，4.1 節で提案システムに必要な機能などの要求条件を述べたあと，4.2 節で既存技術の課題について概説する．その後，4.3 節で時系列クラスタリング，4.4 節で可視化技術との連携について提案手法をそれぞれ説明する．

4.1 要求条件

トレンド分析には，クラスタ内の変化に関する分析とクラスタ間の変化に関する分析が存在する．クラスタ内の変化に関する分析の場合，クラスタ毎にクラスタ内の時間変化を分析する．また，クラスタ間の変化に関する分析の場合，類似するクラスタの関係に着目してその時間変化を分析する．

4.2 トレンド分析の課題

クラスタ内の変化およびクラスタ間の変化に関する分析のいずれの場合も，既存の時系列クラスタリング技術及び可視化技術に課題がある．以降の項ではそれぞれについて詳しく説明する．

なお，既存の一般的な時系列クラスタリングでは，以下の手順で時系列クラスタリングを行っている．(1) 時系列方向にデータを分割後，分割で得られる年代毎にデータをクラスタリングする．(2) クラスタ間の類似度に基づいて閾値以上のクラスタ間に変遷を作成する．(3) クラスタを表現するため，顕著なキーワードを抽出してクラスタの代表とする．

4.2.1 分析工程の課題

クラスタ内の変化に関する分析の場合，既存の時系列クラスタリングによって抽出されたクラスタが，クラスタに応じて異なる年代長で分割されないという問題があるため，内容が類似するクラスタが年代方向で分断されてしまい，時系列データとして長い期間の変化を分析することが難しい．また，クラスタ間の変化に関する分析の場合も，既存の時系列クラスタリングでは，データの分割年代幅は利用者が事前に手動で行うため，同じ内容のクラスタ同士もクラスタ間の変化として捉えられ，本質的なクラスタの変化を捉えることが困難である．例えば，技術変遷において，スキーマ統合の研究は 10 年程度の期間なのに対し，XML ストリーム処理の研究は 5 年程度の期間ということがある．既存の時系列クラスタリング技術では異なる長さのトレンドを抽出することが難しく，改善する必要がある．

4.2.2 可視化技術の課題

可視化技術が分析工程との連携がされておらず，連携のための開発コストがかかってしまうという問題がある．可視化に関する技術としては流量を可視化するものや連続的なデータを可視化するものなどがあるが，いずれの技術も分析工程との連携はされておらず，分析結果を可視化するための拡張が必要である．また，既存の時系列クラスタリングでもクラスタリング結果の可視化は単純なものしかできていない．例えば I-Scover チャレンジ 2013 で用いられた手法 [10] では，クラスタリングにより技術領域を抽出し，クラスタ間の類似度を判定することでクラスタ間の変遷を特定していたが，可視化については単に線でクラスタをつないでいただけで，手動で結果を生成していた．

可視化技術に関しても，既存のサンキーダイアグラムは，エッジに流量を与えてノード間のフローを可視化するものであるため，クラスタ間の変化を見るのに適しているが，ノードの大きさを決定することができなかったり，エッジの流量が陽に与えられていない場合は流量の定義が難しいといった問題がある．

4.3 時系列クラスタリング

提案システムでは，ユーザが予め年代長を指定することなく，年代長は自動で決定されてクラスタが生成される．これにより，クラスタ内の変化に関する分析の場合には同一になるべきクラスタが細分化されることがなくされることなく，長期の年代に渡る 1 つのクラスタとして抽出することができる．また，クラスタ間の変化に関する分析の場合も，クラスタに応じて年代長が適切に決定されるため，本質的なクラスタのみの変化を分析することが可能となる．

提案システムでは，クラスタ間の変化に関する分析に主に焦点を当てて議論していく(注6)．提案システムの処理の概要は以下のとおりである．

（1） 分析する対象と，それに関する情報をノードとして，データを N 部グラフで表現する．

（2） 技術領域に応じた年代長のクラスタを抽出するため，以下の 2 手法を提案する．

- グラフクラスタリングを行い，トレンドを構成する要素であるクラスタを抽出した後，エントロピーを導入して時系列分割を行う「クラスタリング後エントロピー分割」手法
- 時間スムージングを行うために，年代ノードを追加してクラスタリングをする「年代ノードありクラスタリング」手法

（3） クラスタは時間情報を持つため，年代が異なるクラスタ同士の類似度を判定してクラスタ間の変遷を導出する．

エントロピーを導入する場合は，N 部グラフのグラフクラスタリングが完了した時点で，抽出したクラスタに対しエントロピーが小さくなるよう時系列方向に分割を行う．年代ノードを追加する場合は N 部グラフに年代ノードを追加し，論文ノードとのエッジを構築する．以降の節にて，処理の詳細について述べる．

4.3.1 データのグラフ化

元となるデータから，トレンド分析を行う対象と，それに関連する情報をノードとして，N 部グラフを作成する．例えば，技術年表を作成する場合は，論文データから論文の著者や論文のタイトルを抽出し，グラフ構造を構築する．タイトルは，単語ごとに切り分けてそれぞれの単語をひとつの要素とみなす．具体的にはまず,「論文」をノードとし，論文ノードは属性に論文の発表された年を持つとする．また，論文の「著者」，論文のタイトルに含まれる「単語」をそれぞれノードとする．そして，論文ノードと著者ノードの間、論文ノードと単語ノードの間にそれぞれエッジと重みを設定して 3 部グラフを構築する．

(注6)：なお，クラスタ内に関する分析については，N 部グラフ作成後，クラスタの抽出を行った時点で可視化工程に移行できる．

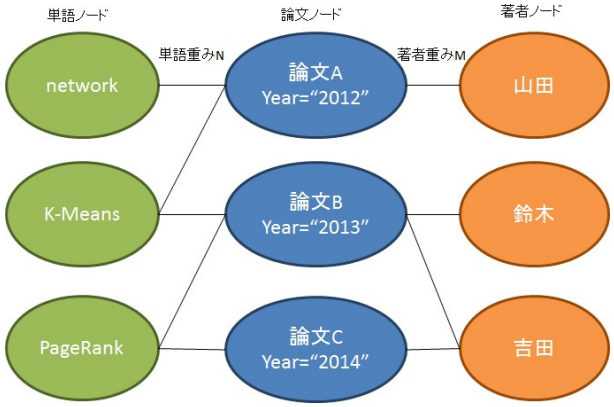

図 1　N 部グラフの例

それぞれのエッジの重みは以下のとおりに定める．

著者ノードと論文ノード間のエッジの重み

$$= \frac{1}{\text{論文に含まれる著者の総数}} \quad (6)$$

単語ノードと論文ノード間のエッジの重み

$$= \frac{1}{\text{論文のタイトルに含まれる単語の総数}} \quad (7)$$

図 1 は，N 部グラフの例で，論文データを元に 3 部グラフを作ったものである．論文ノードを起点とし，その論文を書いた著者ノードと，論文のタイトルに含まれる単語ノードを，それぞれエッジでつなぐ．同一の著者や単語が出てきた場合は，新たにノードを複数生成せずに既存のノードとエッジをつなぐ．

4.3.2　クラスタ化

4.3.1 で作成した N 部グラフに対してグラフクラスタリングを実行する．グラフクラスタリングを行うことで，共通の著者や，論文タイトルに共通の単語が複数含まれる論文は，同一のクラスタとして抽出されることが期待される．同じ著者が書いた論文や，論文のタイトルに同じ単語が複数含まれる論文は，技術領域が近い論文であると考えることができるため，抽出したクラスタは技術領域を表しているとみなす．

さらに，クラスタは属性として年代情報を持つものとする．クラスタに含まれる論文ノードには，論文が書かれた年代を情報として持っており，これをクラスタが持つ年代としてみなす．クラスタ内に特定の 1 年のみの論文しか含まれていない場合は，クラスタが持つ論文の年代長は 1 年とする．クラスタ内に複数の年代にわたる論文が含まれている場合は，クラスタに含まれる論文のうち最も古い年代のものと最も新しいものを抽出し，その期間をクラスタの年代長とする．

クラスタのサイズは，ノード数として定義される．しかし，これは N 部グラフのユースケースであり，特定の種類のノード数をクラスタサイズに割り当てたい場合も考えられるため，ユーザがクラスタサイズを指定することができるものとする．

4.3.3　エントロピーを用いた時系列分割

N 部グラフのグラフクラスタリングでは，全年代の論文を対象にクラスタリングしているため，長期間に渡る広域な技術領域しか抽出することができないという問題がある．この問題を解決し，技術領域に応じた年代長のクラスタを抽出するため，「クラスタリング後エントロピー分割」の手法では，抽出したクラスタに対し，著者や単語に偏りが出る年代で分割を行う．つまり，4.3.2 での N 部グラフのグラフクラスタリング後，抽出したクラスタに対してエントロピーを用いて時系列的な変化点での分割を行い，狭域な技術領域を表すクラスタを抽出する．時系列的な変化点とは，時間方向においてクラスタに含まれる技術領域の分布が偏ることの条件を満たすことと定義する．すなわち，情報不純度を表すエントロピー $E(A)$ が，最小となる年代を変化点とみなす．具体的には，年代で分割することによって得られる情報量ゲイン（相互情報量）$gain(A)$ が最大となる点で分割を行う．クラスタの分割は，時間情報を持つクラスタに対して時系列方向についてクラスタに存在する論文ノードに紐づく著者・単語ノードへのエッジ数を目的変数，年代を説明変数として各年代で分割した時のゲインを算出し，ゲインが最大となる年代でクラスタを分解する．エントロピーに基づくクラスタ分割の流れをアルゴリズム 1 に示す．

Algorithm 1 エントロピーに基づくクラスタの分割

Ensure: クラスタ集合 C，クラスタ c が持つ年代集合 $AllYears(c)$, 論文集合 $AllPapers(c)$, ノード集合 $AllNodes(c)$
エントロピー $entropy = 0$
分割する年代 $divideYear = -1$
for $c \in C$ **do**
  if c の年代長 $= 1$ **then**
    return
  else
    //分割する年代を取得
    for $year \in AllYears(c)$ **do**
      //$year$ で分割した時のエントロピー $newEntropy$ を計算
      if $newEntropy \geqq entropy$ **then**
        $divideYear = newyear$
      end if
    end for
  end if
end for
//クラスタを分割し，再構築
for $paper \in AllPapers(c)$ **do**
  if $year \leqq divideYear$ **then**
    $newCluster1.insert(paper, paper.year)$
  else
    $newCluster2.insert(paper, paper.year)$
  end if
  for $paper$ のエッジの接続先ノード $targetNode$ **do**
    if $targetNode \in AllNodes(c)$ **then**
      if $year \leqq divideYear$ **then**
        $newCluster1.insert(targetNode)$
      else
        $newCluster2.insert(targetNode)$
      end if
    end if
  end for
end for

ゲインは，以下の式を使って導出する．

$$gain(\text{年代}) = I(\text{著者 1，著者 2，}\cdots\text{，単語 1，単語 2，}\cdots) - E(\text{年代}) \quad (8)$$

$$E(\text{年代}) = \sum_{i=1}^{\text{著者と単語の総数}} \frac{\text{クラスタに含まれるノード数}}{\text{全ノード数}} \times I(\text{著者 1，著者 2，}\cdots\text{，単語 1，単語 2，}\cdots) \quad (9)$$

クラスタの分割後，論文ノードを中心に，グラフを再構築する．その際，分割後のクラスタに含まれる論文ノードがどちらも同じ著者・単語ノードを持つ場合はそれらを複製する．これは，固定長で年代分割を行う手法と，クラスタの仕様を統一するためである．そして，分割後のそれぞれのクラスタに著者・単語ノードを含め，論文ノードとのエッジを設定する．

4.3.4 年代ノードありクラスタリング

「論文」,「著者」,「単語」の 3 部グラフのグラフクラスタリングでは全年代の論文を対象にクラスタリングしているため，長期間に渡る広域な技術領域しか抽出することができないという問題がある．この問題を解決し，利用者が予めデータを時系列方向に分割することなく技術領域に応じた年代長のクラスタを抽出するため,「年代ノードありクラスタリング」手法では 4.3.1 でのグラフ化の際，年代ノードと年代ノードへのエッジを導入して，時間スムージング [1] を行う．また，同じ年代や近傍の年代に書かれた論文は領域も同じである可能性が高いと想定し，近傍の年代が同じクラスタに属するようクラスタリングを行うことにする．そこで，同じ年代ノードに接続されている論文がクラスタリングの際に同じクラスタに統合されやすくするために，年代ノードへのエッジは，論文の持つ年代に一致した年代のみへエッジを構築する場合と，論文の持つ年代に一致した年代と更にその前後 1 年の年代へのエッジを構築する場合の 2 種類を用意する．実際の処理としては，論文データかグラフ化する際に，論文の発表された「年代」をノードとして追加し，論文ノードからのエッジを構築してからグラフクラスタリングを行う．クラスタリングの精度を上げるため，論文ノードと年代ノードのエッジの重みはグラフクラスタリング後の Modularity が最大となるように決定する．

4.3.5 クラスタ間の変遷の導出

グラフクラスタリングによってクラスタが抽出された後，クラスタの変遷を導出するためにクラスタ間の関係を求める．変遷は，クラスタ間の類似度を用いて，一定の閾値 ϵ 以上の類似度を持つクラスタ間にエッジを構築することで表現する．グラフクラスタリングによって抽出されたクラスタ集合を C とするとき，クラスタ間の類似度は以下のように定義する．

$$s(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} \quad \text{ただし，} C_1, C_2 \in C \quad (10)$$

$C_1 \cap C_2$ は C_1 と C_2 に含まれるノードの積集合，$C_1 \cup C_2$ は C_1，C_2 の和集合に含まれるノードである．クラスタ間の関係は閾値 ϵ を用いて以下のように定義する．

$$N_\epsilon(C_1, C_2) \quad \text{ただし，} s(C_1, C_2) \geqq \epsilon, C_1, C_2 \in C \quad (11)$$

技術変遷の時系列方向での前後関係を明確にするため，変遷を構築する際は技術領域の開始点を基準とする．クラスタの最も古い論文の年代を技術領域の開始点とみなし，クラスタ間の関係を導出するときは C_1 が C_2 の年代より古い年代となるよう設定する．

4.4 クラスタ分析と可視化の連携

抽出したクラスタを可視化するため，データ可視化ライブラリを利用して可視化を行う．本稿では D3（Data-Driven Document）で提供されているライブラリを用いる．クラスタ内分析を行うため，ストリームグラフを用いて可視化を行う．さらに，クラスタ間分析を行うため，サンキーダイアグラムを拡張してから可視化を行う．

クラスタ内の変化を分析するためには，クラスタ内のデータ量の時間変化を可視化することが必要である．例えば，ある技術領域に関して書かれている論文の量が，年代に応じてどのように変化しているのかを可視化する場合は，ストリームグラフを活用する．ストリームグラフは，トレンド分析においては，クラスタの量を連続的に可視化できるという特徴をもつ．時系列クラスタリングの結果を可視化する際には，可視化させたいクラスタの，時間ごとの論文の量を入力することで可視化することができる．

クラスタ間の関係を可視化する際はサンキーダイアグラムによる可視化を行う．変遷を持つクラスタ同士にエッジを構築することで，ある技術領域がどのような技術領域から影響を受け，そしてどのような技術領域に影響を及ぼしているかを分析することができる．時系列クラスタリングの結果を可視化する際には，N_ϵ のクラスタ間の関係にある C_1, C_2 に対してエッジを構築することで可視化することができる．しかし，前述のように既存のサンキーダイアグラムではトレンド分析の可視化に際し問題があるため，これを解決するために以降に示すような拡張を行う．

4.4.1 サンキーダイアグラムの拡張

（1） エッジの流量設定

クラスタリング結果ではクラスタ間の流量というものが陽に与えられないため，クラスタ内毎に含まれているノードやエッジ数を元に流量を計算する必要がある．そこで，エッジで繋がれているクラスタ同士の関係を流量により把握するため，クラスタ間のエッジの流量は，論文の流量あるいはクラスタ間の類似度を表すように設定する．

（2） 横軸と年代の連係

サンキーダイアグラムでは，クラスタ間の相対的な関係が左から右へのフローとして表現される．そのため，クラスタの変遷はわかりやすいが，あるクラスタと同年代のクラスタがどれであるかが把握しにくい問題がある．これは，サンキーダイアグラムがクラスタ間のエッジ長が最小となるようクラスタの配置を行っているため，同年代のクラスタであっても横軸方向に関して違う箇所に配置されてしまうことが原因である．そこで，同年代のクラスタやその前後の年代のクラスタをひと目で確認

するために，横軸方向を時間軸に割り当て，クラスタが持つ年代に応じた位置に配置するよう変更する．

（3） ノードの縦サイズ

サンキーダイアグラムでは，クラスタを表すノードの縦のサイズは流入してくるエッジ数を元に算出されているが，これはクラスタ自身の要素数を表すサイズではない．そこで，単に技術領域のみならず，著者や単語に関しての時間変化も分析するため，ノードの縦サイズはクラスタサイズを表すよう設定し，利用者が任意にクラスタサイズの属性を指定できるようにする．

5. 評価実験

手法毎に抽出したクラスタの性能を比較するために，評価実験を行った．本実験では特に技術年表に関するデータを用いて評価した．年表作成には DBLP Computer Science Bibliography(注7) から得られる論文のメタデータを用いる．データには 1936 年〜2015 年の論文計 2,668,102 件が含まれている．実験は CPU が Intel Xenon CPU E7-4850，メモリが 512GB の Windows サーバを利用した．

評価は，以下の 5 つの手法に対して行った．

（1） 年代分割ありクラスタリング

論文データを固定長で予め年代分割をしてから，各年代毎にクラスタリングを行う．

（2） クラスタリング

年代分割を行わず，年代ノードも追加しないで全年区間でクラスタリングを行う．

（3） クラスタリング後エントロピー分割

（4） 年代ノード（一致）ありクラスタリング

（5） 年代ノード（一致+前後）ありクラスタリング

なお，固定長で分割する手法では，最終的なクラスタ数が 500 程度であった 10 年区切りで年代分割を行い，年代ノードが含まれている手法については，年代ノードと論文ノードのエッジの重みは，グラフクラスタリング後の Modularity が最もよかった，著者エッジの重みの 1/1000 に設定して実験を行っている．

以降の節では，全手法の性能評価を行ったあと，年代ノードへのエッジを追加したことによる影響を比較する．

5.1 データの統計情報

表 1 は，各手法におけるノード数とエッジ数，そして 1 ノードあたりの平均接続エッジ数を記載したものである．固定長で年代分割をした手法については，分割した年代毎にグラフが作成されるため，各年代のエッジ数を記載している．表 1 より，論文数が年代とともに右肩上がりとなっていることがわかる．2010 年代の論文が減少しているのは，使用した論文データに含まれている論文が，2015 年までしか含まれていないためである．1 ノードあたりの平均接続エッジ数も年代とともに増加しており，1 論文に接続されている著者ノードや単語ノードが増えていることを示している．これは，年代が経つとともに研究内容が多様化し，論文のタイトルが長くなったことが原因として考えられる．

(注7)：http://dblp.uni-trier.de/xml/

表 1 手法毎の統計情報

手法	年代	ノード数	エッジ数	平均接続エッジ数/ノード
年代分割	1930	51	290	5.69
+クラスタリング	1940	138	881	6.38
	1950	1,059	7,110	6.71
	1960	8,016	56,834	7.09
	1970	28,158	213,701	7.59
	1980	92,885	749,275	8.07
	1990	357,826	3,224,432	9.01
	2000	1,241,193	12,812,215	10.32
	2010	938,776	10,506,049	11.19
クラスタリングのみ		4,427,581	27,570,787	6.23
クラスタリング後エントロピー分割		4,427,581	27,570,787	6.23
年代ノード（一致）ありクラスタリング		4,427,661	30,238,889	6.83
年代ノード（一致+前後）ありクラスタリング		4,427,663	35,575,093	8.03

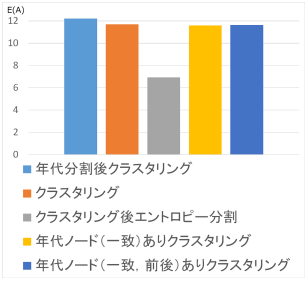

図 2 エントロピー $E(A)$

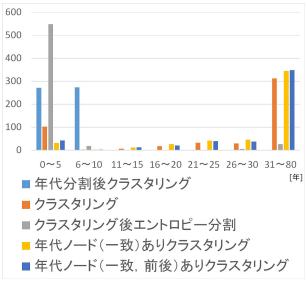

図 3 平均年代長

5.2 評価観点

手法のクラスタ性能を評価するため，以下の項目について比較を行った．まず，クラスタが技術領域の観点で精度よく分割できているか評価するため，エントロピーの値を比較し，クラスタが技術領域に応じた年代長となっているかを評価するため，クラスタの年代長分布を調査した．そして，クラスタリングの精度を評価するために Modurlarity を比較し，さらにクラスタ抽出までの実行時間を比較した．エントロピーにおける評価観点「精度よく分割する」とは，同じ技術領域の論文ノードが同じクラスタに含まれていることを示すものとする．そこで，クラスタに含まれる論文ノードに接続されている著者・単語ノードへのエッジ数を目的変数として情報不純度 $E(A)$ の値を算出し，比較を行う．評価は，分割後のクラスタの情報不純度が小さいほど精度よく分割できているとみなす．クラスタの年代長については，本研究では技術変遷を構築するために，短い年代長の方が望ましいとする．Modularity の値は，最終的に得られたクラスタに対し算出する．本実験では，最終的な出力クラスタ数が 500 程度となるよう指定して実験を行った．

5.3 実験結果

エントロピーと年代長の観点で比較を行うと，「クラスタリング後エントロピー分割」がバランスが良いが，Modularity の

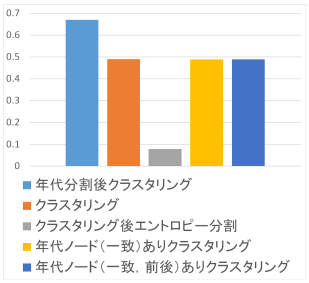

図 4　Modularity

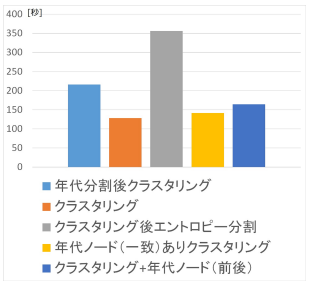

図 5　実 行 時 間

表 2　クラスタの年代幅

手法	最大	最小	平均	中央
年代分割ありクラスタリング	10	1	5.42	6
クラスタリング	80	1	33.34	37.5
クラスタリング後エントロピー分割	51	1	3.52	1
年代ノード（一致）ありクラスタリング	80	0	37.64	40
年代ノード（一致+前後）ありクラスタリング	80	0	37.22	40

値と実行時間が他の手法に比べ悪いという結果を示した．

図 2 より，エントロピーについては，「年代分割ありクラスタリング」に比べ，「クラスタリング後エントロピー分割」手法が 43%，「年代ノード（一致）ありクラスタリング」手法では 5%低い値を示し，提案手法の有効性が確認できた．「クラスタリング後エントロピー分割」ではエントロピーに基づいて年代分割をしているため，妥当な結果といえる．年代ノードを追加した手法においてもクラスタリングによって技術領域方向に精度よく分割できているといえる．

図 3 より，クラスタに含まれる論文の年代長は，「年代分割後クラスタリング」がすべて 0〜10 年の年代長となっているのに対し，「クラスタリング後エントロピー分割」手法では 0〜10 年の他に 26 年以上の年代長が長いクラスタも出力された．年代ノードを追加した手法については，幅広い年代長のクラスタ出力されたが，31 年以上のものも多く出力されている．年代長が長いと，クラスタ間の細かな変遷を分析することができないため，本研究の目的にはそぐわない結果となった．この原因として，著者ノードや単語ノードに比べてグラフクラスタリング時における年代ノードの影響力が小さく，単語や著者といった技術領域に関するクラスタが中心に生成され，そこに長期間の年代の論文が含まれてしまったためだと考えられる．

クラスタの年代長についてさらに分析を行うため，各手法に対し，クラスタの年代長の最大，最小，平均，中央を求めた．結果を表 2 に記載する．「年代ノード（一致）ありクラスタリング」，「年代ノード（一致+前後）ありクラスタリング」の最小値が 0 となっているのは，論文ノードが含まれないクラスタが生成され，年代長を求めることができなかったためである．今回実験に用いた論文データの最大年代長は 80 である．従って，「クラスタリング」，「年代ノード（一致）ありクラスタリング」，「年代ノード（一致+前後）ありクラスタリング」の 3 手法については，すべての年代の論文が含まれているクラスタが

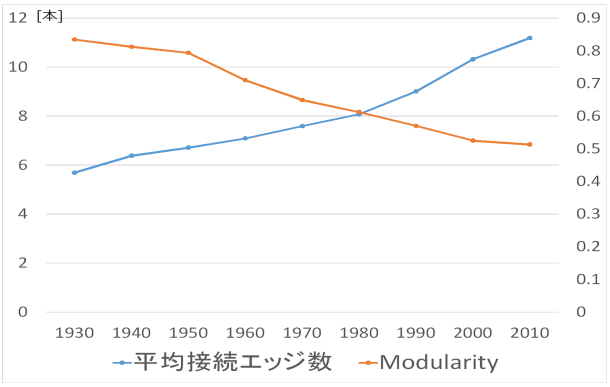

図 6　ノードあたりの平均接続エッジ数

存在していることとなる．これら 3 つの手法については，平均だけでなく中央値も高い値を示しており，広い期間に渡る論文が同クラスタに含まれていることを表している．対して，「年代分割ありクラスタリング」や「クラスタリング後エントロピー分割」といった，年代分割をしている手法は平均値も中央値も短い期間を示している．しかしながら，「年代分割ありクラスタリング」は固定長で年代分割を行っているため，最大値は常に一定である．以上の結果から，クラスタの年代長の分布が最も分散している「クラスタリング後エントロピー分割」が，技術領域に応じた年代長のクラスタを抽出するのに適している．

図 4 より，Modularity については，「年代分割ありクラスタリング」が他の手法に比べ 26%以上高い値を示している．この理由を分析するため，グラフのノード数やエッジ数と，Modularity 値の関連を調べた．図 6 は「年代分割ありクラスタリング」における，ノードあたりの平均接続エッジ数をグラフ化したものである．表 1 からわかるように，年代とともにノード数と接続エッジ数が増加している．そして，接続エッジ数が増えると，Modularity の値は減少している．しかしながら，接続エッジ数が少ない 1930 年代や 1940 年代では高い Modularity の値を示しているため，この影響が強く，全体の平均 Modularity 値も高い値となったと考えられる．

「クラスタリング後エントロピー分割」は他の手法に比べ実行時間がかかってしまうため，これを改善することが課題である．これは，「クラスタリング後エントロピー分割」以外の手法についてはグラフクラスタリングのみであるのに対し，「クラスタリング後エントロピー分割」はグラフクラスタリングを行ったあと，抽出したクラスタすべてに対し，年代毎にエントロピーの *gain* を計算する操作を行っているためである．

5.4　ユースケース

論文データからグラフクラスタリングによって抽出したクラスタが，どのような結果となっているかを分析するため，可視化技術を用いて可視化を行う．可視化は，エントロピーと平均年代長のバランスが最もよかった，「クラスタリング後エントロピー分割」の結果を用いて行う．クラスタに含まれる単語ノードのうち，最も多くの論文ノードと接続している上位数件の単語を，クラスタの代表的な単語とみなす．この単語群によって，クラスタがどんな技術領域を表しているか認識できる．

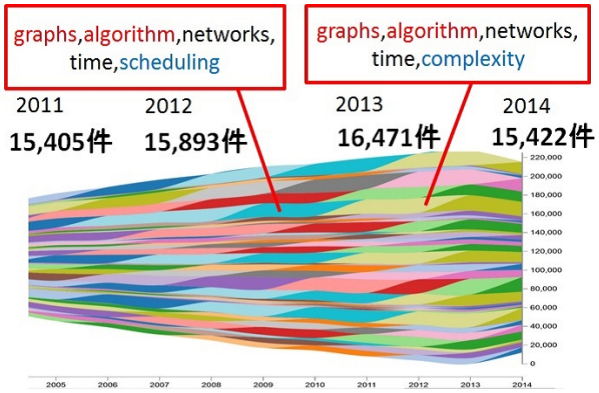

図 7　クラスタ内分析の例

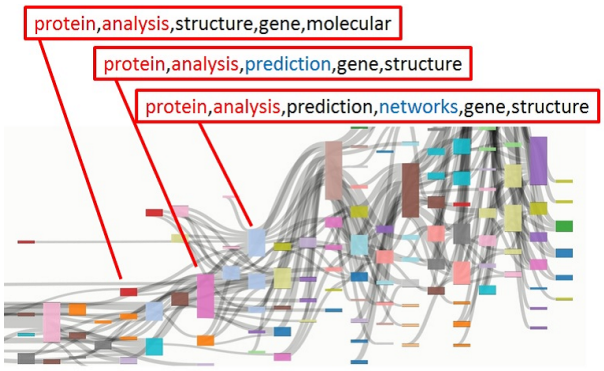

図 8　クラスタ間分析の例

5.4.1　クラスタ内分析

図 7 は，ストリームグラフでクラスタを可視化したものである．色の付いている部分がクラスタを表しており，異なるクラスタには異なる色が割り当てられている．縦軸は論文の数，横軸は時間軸を示している．実際に得られた可視化結果から分析を行う．代表的な単語として，“graph,algorithm,networks”があり，このクラスタがグラフアルゴリズムに関する技術領域であることがわかる．そして，2011 年と 2012 年にまたがるクラスタと，2013 年と 2014 年にまたがるクラスタでは異なる単語が示されており，この 2 つのクラスタが異なる技術領域を表していることがわかる．クラスタに含まれる論文数に着目すると，2011 年から 2013 年まで右肩上がりで増加しており，流行の領域であることがわかる．

5.4.2　クラスタ間分析

図 8 は，拡張したサンキーダイアグラムでクラスタを可視化したものである．色が付いている部分がクラスタを表している．縦軸がクラスタの種類，横軸が時間軸を示しており，各クラスタは自身が示すの年代に応じた横軸部分に位置している．実際に得られた可視化結果から分析を行う．代表的な単語として，“protein,analysis,structure”があり，このクラスタがタンパク質の構造解析に関する技術領域であることがわかる．この領域について，一番古い年代のクラスタでは “gene,molecular” という単語があるため，遺伝子や分子の構造解析の領域であることがわかる．時代が経つと，“prediction”という単語が含まれ，タンパク質構造予測の領域に変化している．さらに時間が経つと，“network”という単語が含まれ，タンパク質間相互作用ネットワークに関する技術領域に変遷していることがわかる．

6.　おわりに

本稿では，分析技術と可視化技術を連携させてトレンド分析をするシステムの開発に取り組んだ．提案システムでは，データを事前に分割せずに時系列クラスタリングを行い，年代長が自動で決定されるためにエントロピーに基いてクラスタを分割する手法と，年代ノードを追加してクラスタリングを行う手法の 2 種類を提案した．さらに，クラスタ内の変化に関する分析とクラスタ間の変化に関する分析ができるよう可視化した．

提案システムのプロトタイプを作成し評価実験で比較することで，エントロピーとクラスタの年代長のバランスがとれている，エントロピーに基づいてクラスタ分割をする手法の有効性が示された．

今後は，次の点について継続して研究を行う予定である．得られたクラスタが正しく領域分割できているか確認するため，評価データを用いて確認を行うことを検討する．可視化技術ではクラスタ間分析での可視化の際，類似したトレンド領域の明示ができるようにするなど更に改良をしていく予定である．

文　献

[1] M. S. Kim, and J. Han, “A Particle and Density Based Evolutionary Clustering Method for Dynamic Networks,” Proceedings of the VLDB Endowment, 2009.
[2] F. Folino and C. Pizzuti, “A Multiobjective and Evolutionary Clustering Method for Dynamic Networks,” Proceedings of ASONAM, 2010.
[3] F. Folino and C. Pizzuti, “An Evolutionary Multiobjective Approach for Community Discovery in Dynamic Networks,” Knowledge and Data Engineering 26(8), 2014.
[4] P. Lee, L. V. S. Lakshmanan, and E. E. Milios, “Incremental Cluster Evolution Tracking from Highly Dynamic Network Data,” Proceedings of ICDE, 2014.
[5] D. M. Blei and J. D. Lafferty, “Dynamic Topic Models,” Proceedings of international conference on Machine learning, 2006.
[6] 芹澤 翠, 小林 一郎 “潜在的ディリクレ配分法に基づくトピック類似度を考慮したトピック追跡,” DEIM Forum, 2011.
[7] Y. Zhou, H. Cheng, and J. X. Yu,“Graph Clustering Based on Structural/Attribute Similarities,” Proceedings of the VLDB Endowment 2(1), 2009.
[8] M. E. J. Newman, and M. Girvan “Finding and evaluating community structure in networks,” Physical review E 69(2), 2004.
[9] マクロ経済データと企業業績 - 20.pdf : 決定木分析, http://www5.atpages.jp/kera/up/log/20.pdf
[10] 審査結果 (I-Scover チャレンジ 2013) — IEICE I-Scover Project:, http://www.ieice.org/ iscover/iscover/?page_id=640
[11] J. Leskovec, J. Kleinberg, and C. Faloutsos “Graph Evolution: Densification and Shrinking Diameters,” ACM Transactions on Knowledge Discovery from Data (TKDD) 1(1), 2007.
[12] L. Byron, and M. Wattenberg, “Stacked graphs?geometry & aesthetics,” Visualization and Computer Graphics, 2008.
[13] M. Schmidt, “The Sankey diagram in energy and material flow management,” Journal of industrial ecology 12(1), 2008.