論文 DB からのマクロ情報抽出のためのクラスタリング閾値設定指針

Guide for Deciding a Clustering Threshold to Retrieve Macro-information from a Research-paper Database

吉田 誠♡ 小林 隆志♢ 横田 治夫♢

Makoto YOSHIDA Takashi KOBAYASHI Haruo YOKOTA

電子的に利用可能な研究論文数の増加に伴い，求める情報の検索コストも増大している．目的の情報を探し出すコストを減らすため，我々は論文データベースから研究の発展経緯等のマクロな情報を抽出し，それらを利用した高度な検索を行うことを目的として，リサーチマイニング手法を提案している．この手法は，論文間の発展経緯の抽出と，経緯のコンフィデンス値を用いた論文のクラスタリングという 2 つのフェーズからなる．クラスタリングにおいては，研究の発展経緯の理解し易すさに影響するクラスタの粒度を定めるコンフィデンス値のクラスタリング閾値が重要となるが，これまでその基準は明確ではなかった．そこで本稿では，実験結果を基に，クラスタリング閾値を設定するためのアプローチを提案する．

The number of research papers on electronic publishing is increasing. It raises the cost of searching them for the desired information. To reduce the search cost, we have proposed the research mining method extracting macro information, such as research evolution flows, from a research-paper database. There are two phases in the method: resolving paper progress flows and clustering papers using confidence values for the flows. To cluster papers, a threshold of the confidence values is influential because it determines the granule of clusters, which affects understandability of the macro information. However, the criterion for the threshold was unclear so far. In this paper, we propose an approach to decide the threshold, based on experimental results.

1. はじめに

ネットワーク技術の発達，情報インフラの普及にともない，電子的に利用可能な研究論文の数が増大してきている．これにより必要とする文献を電子的に入手することが可能となったが，目的の論文を探すコスト，論文の位置付け，関連状況を知るコストが大きくなってきている．これまでは検索手段として，キーワード検索が多く用いられてきた．しかしながらキーワード検索だけでは，目的とする論文を直ちに得られることがあまり多くない．

このため，論文間の関係を利用するアプローチが研究されている．引用関係を利用し，論文間の類似度を知る手法として書誌結合（bibliographic coupling）[1]，共引用分析（co-citation analysis）[2] などが古くから提案されている．書誌結合とは 2 つの論文間の関連度を知るために，その 2 論文が参照している論文の重複数を考慮するものである．この書誌結合を改良した研究として，難波らによって参照の仕方を考慮した研究もなされている [3]．この手法では，被参照論文の参照の理由を考慮し，参照構造を用いて論文間の類似度を測ることを行っている．また共引用分析は，2 論文が他の論文に共に引用されている回数を基準としている手法である．

これらの方法では何らかの関係にある論文の集合を発見することは可能であるが，新しい研究が古い研究を包含している，複数の研究が融合して新しい研究になっているといった研究の発展した過程等に関するマクロな情報を抽出することはできない．そのため，目的の論文を検索するコストは小さくならない．

我々は検索コストを抑えるために，この研究の発展した過程 (研究の発展経緯) を考慮する必要があると考えている．そこで，我々はこれまでに，研究の発展経緯を抽出し，さらにそれらのマクロな流れを表現することができるリサーチマイニング手法を提案し [4]，さらに実際の論文に対して提案手法を適用し有用性を確認してきた [5]．本手法では，マクロな流れを表現するためにクラスタリングを行うが，適切なクラスタリング閾値を定めるコストが大きく，また得られたクラスタが理解し易くしているかどうかの根拠が明確ではなかった．そこで本稿では，研究の発展経緯の把握を容易にする論文クラスタを形成するためのクラスタリングの指針について考察する．

本稿ではまず，次節において論文から研究の発展経緯を抽出するリサーチマイニング手法を説明する．次に，公開論文 DB である CiteSeer[6] を利用して得た，いくつかの論文情報集合に対しリサーチマイニング手法を適用し，得られた結果のクラスタ内の論文数とクラスタリング閾値の対応関係を分析を行う．そして，その特徴に基づきクラスタリング閾値を適切に定める方法を提案する．さらに 4. において，提案した方法の有効性を実験によって確認する．

2. リサーチマイニング手法

リサーチマイニング手法は論文間の発展経緯の抽出，論文のクラスタリングという 2 つのフェーズからなる．以下ではそれぞれについて説明を行うが，詳細は [5] を参照されたい．

2.1 論文間の発展経緯抽出

ここではデータマイニングのアプローチの一つであるアソシエーションルールを発見する方法としてアプリオリアルゴリズム [7] を利用する．本研究では，1 つの論文が持つ参照を 1 つのトランザクションと考え，共に参照されている論文の関連度を数値化し，方向付けを行う．つまり「論文 A を参照しているならば論文 B も参照している」というルールをアソシエーションルール，ルールの条件付き確率をコンフィデンス値とみなす．また，論文が共に引用されている回数の閾値をミニマムサポート値とする．さらに，論文をノード，結果として得られたアソシエーションルールを有向枝，コンフィデンス値を重みとすることにより，重み付き有向グラフを作成する．

本研究では，参照関係の方向と比べて逆向きの枝があり，コンフィデンス値があらかじめ定めた閾値より大きいものを研究の発展経緯を表す枝として扱う．すなわち，ある 2 論文 A（古い論文），B（新しい論文）を考えた場合，以下の 3 つの条件を満たす場合のアソシエーションルールを研究の

♡ 学生会員　東京工業大学 大学院 情報理工学研究科 計算工学専攻
yoshidada@de.cs.titech.ac.jp

♢ 正会員　東京工業大学 学術国際情報センター
tkobaya@gsic.titech.ac.jp, yokota@cs.titech.ac.jp

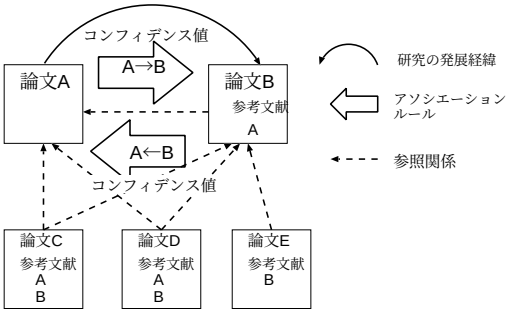

図 1: 論文間のアソシエーションルール
Fig.1 Association rules between research papers

発展経緯とする．

- B ⟹ A という参照関係が存在
- A → B というアソシエーションルールが存在
- そのアソシエーションルールのコンフィデンス値があらかじめ定めた閾値より大きい

このような論文間のアソシエーションルールを考えた場合，参照関係がある 2 論文では通常はその分野の起源の論文に近い古い論文のほうが参照される回数が多くなる．しかし研究が古い論文から新しい論文に発展している場合には，古い論文を参照している時に同時に新しい論文も参照していることが多い．発展経緯抽出はこの事に基づいている手法である．例えば図 1 を考えた場合，論文 A から論文 B へのアソシエーションルールを研究の発展経緯とする．

2.2 クラスタリング

論文単位での研究の発展経緯を追うためには，前述した研究の発展経緯を抽出するだけでも十分であるが，論文数が多い場合には，それのみでは研究の発展経緯を把握するのが容易ではなくなる．対象の論文数が増えた場合には，よりマクロな視点として研究分野単位での発展経緯を知ることが有用である．本研究ではこのマクロな発展経緯を表現するために，上述のグラフに対してクラスタリングを行う．

研究の発展経緯を表す枝でつながれている論文同士は参照，被参照という直接的な関係があり，その中でも重みが大きい枝でつながれている論文同士は他の多くの論文から関連が強いと判断されていることを意味する．そこで，重みが閾値より大きい枝である場合は，その枝で結ばれている論文を同一のクラスタに属すると扱う．本研究ではこの閾値をクラスタリング閾値と呼ぶ．なお，クラスタを形成する際，同一の論文やクラスタへの発展経緯が複数存在する場合，その論文やクラスタへの発展経緯はその中で最も重みが大きいものとする．

本手法ではマクロな視点として，クラスタ間の発展経緯を抽出することができるが，さらにクラスタリング閾値を変化させることにより，クラスタの粒度を変化させることが可能であり，研究のマクロな発展経緯を柔軟に見ることを可能にする．図 2 は，閾値によるクラスタ粒度の変化を表現したものである．この図で，四角は論文，枝は研究の発展経緯を表している枝である．クラスタ 1 は重みが 0.5 以上のものを同一クラスタとしたものである．クラスタリング閾値を下げることで大きい粒度のクラスタ 2 を得ることが可能となる．本稿ではこのクラスタリング閾値を適切に定める指針を考察する．また，本稿では論文 2 編以上がまとまっているものをクラスタとして扱う．

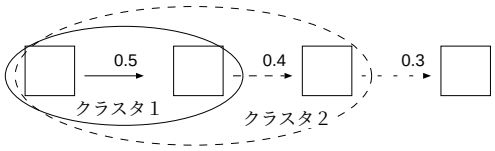

図 2: クラスタリング閾値によるクラスタ粒度の変更
Fig.2 Cluster granularity with varying clustering threshold

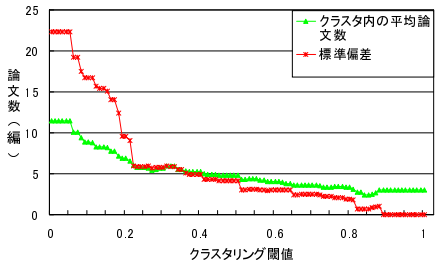

図 3: “software configuration management” の場合
Fig.3 Case of “software configuration managemen”

また，クラスタの主題を知るために，同一クラスタ内の論文のうち，被参照回数がもっとも多い論文をそのクラスタの代表的な論文として扱う．

3. クラスタリング閾値の設定指針

リサーチマイニングでは，クラスタリング閾値を変化させることにより，クラスタの粒度を変化させることができるため，研究の発展経緯を柔軟に見ることが可能であるが，クラスタリング閾値を定める際，マクロな研究の発展経緯の理解を助けるような指針が必要である．

クラスタの見やすさの観点として，総ノード数や総エッジ数等のグラフ全体の大きさに着目した観点と，グラフ全体の大きさには着目せず，グラフ中の任意の部分に着目し，その部分の発展経緯が理解しやすいという観点がある．本稿では，グラフの大小には着目せず，グラフの部分に着目し，クラスタ間の関係を理解しやすいクラスタを形成するという観点からクラスタリング閾値の設定方法に関して実験し，考察を行い，クラスタリング閾値設定の指針を示す．

一般には，利用者の目的によって適切な粒度は異なるため，クラスタリング閾値を一様に定めることは難しい．しかし，我々はこれまでの研究により，理解し易いクラスタを形成する指標として，クラスタ内の平均論文数が関係しているという知見を得ている．そこで本稿では，クラスタ内の平均論文数が利用者が望む値に近くなるようなクラスタリング閾値を発見することを目標とする．

本来であれば，結果として出力される発展経緯のグラフの形状も，適切なクラスタリング閾値を定めるパラメータとして考慮するべきではあるが，グラフの形状は様々であり，扱いが難しく，各グラフについて議論することは一般性を大きく損なう可能性がある．そのため，以降ではクラスタリング閾値の評価をクラスタ内の平均論文数とし，クラスタリング閾値の変化とクラスタ内の平均論文数の関係に絞って調査を行う．

3.1 クラスタリング閾値に関する実験

リサーチマイニング手法により得られた論文間の研究の発展経緯に対し，クラスタリング閾値を 0.01 毎に定め，得られたクラスタ内の平均論文数，および標準偏差を調査した．リサーチマイニング手法適用対象の論文の収集は，まず

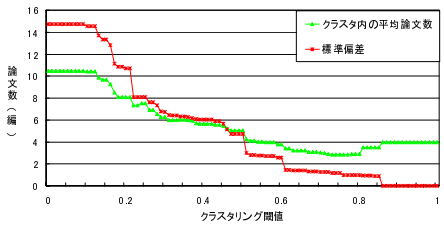

図 4: “web service composition” の場合
Fig.4 Case of “web service composition”

CiteSeer によりキーワード検索を行い，検索により得られた論文の参照関係をたどることにより収集した [5]．図 3 は「software, configuration, management」をキーワードとしてリサーチマイニング手法により得た研究の発展経緯を各クラスタリング閾値によりクラスタリングした結果である．同様に図 4 は「web, service, composition」をキーワードとした結果であり，上記の場合と比較し，収集した論文の内容が分散したため，ミニマムサポート値を小さくした場合である．これら以外にも異なるキーワードの組み合わせにより，他の複数種類の場合についても調査を行ったが，ここでは紙面の都合上省略する．

グラフの形状等の影響もあるため，我々はクラスタリング閾値とクラスタ内の平均論文数の関係は複雑になると考えていたが，図 3，4 および他のキーワードの調査結果から，クラスタ内の平均論文数は若干振動しているが，クラスタリング閾値の増加に伴い，ほぼ単調に緩やかに減少することがわかった．このことから，クラスタ内の平均論文数とクラスタリング閾値はほぼ 1 対 1 に対応しており，任意のクラスタリング閾値を求める場合には，二分探索を用いることが可能である．

また，図 3，4 において，クラスタ内の平均論文数はほぼ一定の変化率であるのに対し，標準偏差は部分的に急激に変化する部分と，ほとんど値が変化しない部分があることがわかった．急激に変化する部分とは，例えば，図 3 においてクラスタリング閾値が 0.5 と 0.51 の部分や 0.21 と 0.22 の部分等が該当する．該当部分はその他の部分，例えば 0.41 から 0.5 の部分等と比べ明らかに異なる変化率で値が変化している．このようなグラフの形状は他のキーワードによる結果においても同様であった．

このような標準偏差が急激に変化する主な要因は，その部分のクラスタリング閾値付近において複数のクラスタの融合，または一つのクラスタが複数のクラスタに分裂しているためであると考えられる．そのため，標準偏差が大きく変化する部分 の平均論文数が小さい側に対応するクラスタリング閾値は，クラスタ同士の融合を発生させずに，より多くの論文をクラスタ化することができるクラスタリング閾値となると考える．

3.2 クラスタリング閾値設定の指針

3.1 より，標準偏差が急激に変化している部分までクラスタリング閾値を変更しても，クラスタ内の論文数はあまり変化しない．一方，クラスタに含まれる論文が多ければ多いほどグラフに現れるノードが減り，直感的に理解しやすくなるため，指針として，クラスタリング閾値をクラスタ内の平均論文数は望む値に近く，クラスタに属さない論文ができるだけ少なくなるように設定するべきである．

一方，3.1 の実験結果によりクラスタリング閾値とクラスタ内の平均論文数は 1 対 1 に対応していること，そして，クラスタ内の論文数の標準偏差が急激に変化している部分ではクラスタ同士が融合していることから，本研究では適切なクラスタリング閾値の定め方を以下のように提案する．

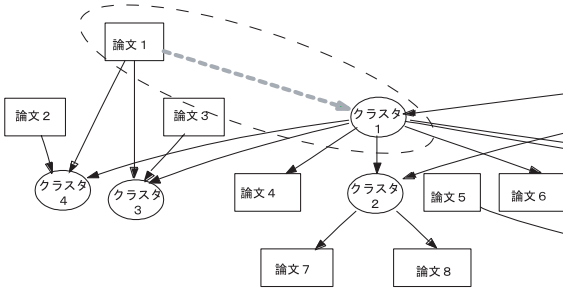

図 5: software configuration management の発展経緯の一部 (クラスタリング閾値 0.5〜0.41)
Fig.5 A part of ”software configuration management” graph

- まずはじめに二分探索によりクラスタ内の平均論文数が利用者が望むクラスタ内の平均論文数にもっとも近い部分を発見する．
- 次に発見した値からクラスタリング閾値を下げていき，クラスタ内の論文数の標準偏差が急激に変化している部分を発見，その直前のクラスタリング閾値を採用する．

4. クラスタリング閾値設定指針の有効性確認

クラスタリング閾値設定指針の有効性を確かめるために，クラスタ内の論文数の標準偏差が急激に変化する部分の前後のクラスタリング閾値，および標準偏差の変化がとても緩やかな部分のクラスタリング閾値を用いてクラスタリングし，比較する．ここでは「software, configuration, management」というキーワードによる発展経緯について調べる．

クラスタリング閾値が 0.41 から 0.5 の間はクラスタ内の論文数の標準偏差が緩やかに変化している (図 3)．図 5 は，クラスタリング閾値 0.5 を用いクラスタリングし，図示したものの一部である．なお，長方形は論文，楕円はクラスタ，矢印は発展経緯を表す．灰色の破線の矢印はクラスタリング閾値を 0.41 とした場合にクラスタを形成する発展経緯である．0.41 から 0.5 の間でクラスタリング閾値を変化させた場合，論文がクラスタに吸収される，もしくは論文同士により新たなクラスタを形成しているのみであり，クラスタ同士がより大きなクラスタを形成することはなかった．そのため，クラスタの粒度はほぼ同じであり，クラスタに論文が多く含まれているクラスタリング閾値 0.41 のほうが 0.5 の場合と比べ，論文がまとまっており，理解し易いクラスタを形成している．

次に，クラスタ内の論文数の標準偏差が急激に変化してる部分であるクラスタリング閾値 0.22 と 0.21 の場合を比較する．図 6 はクラスタリング閾値 0.22 とした場合の発展経緯の一部である．灰色の破線の矢印はクラスタリング閾値 0.21 とした場合に融合し，クラスタを形成している部分である．この場合それぞれ 2 編，7 編，3 編の論文により構成される 3 つのクラスタが融合し，12 編の論文から構成される粒度が荒いクラスタを形成してしまっているため，この閾値で粒度が急激に変化している．また，他のキーワード検索により行ったものも同様の結果が得られた．

以上から，提案手法によりクラスタリング閾値を求めることによって，望む粒度であり，同時に余分なノードが少な

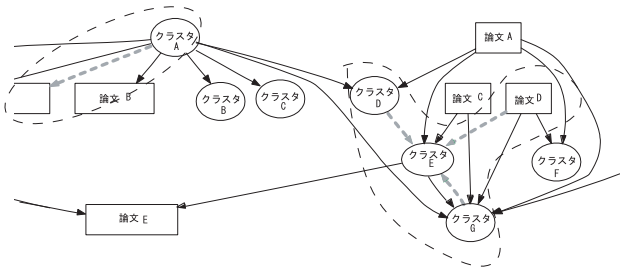

図 6: software configuration management の発展経緯の一部 (クラスタリング閾値 0.22〜0.21)

Fig.6 A part of "software configuration management" graph

くなるようなクラスタリング閾値を求めることができ， 提案手法が有効であることがわかる．

5. 考察

クラスタ内の論文数の標準偏差をみてもわかるように，クラスタリングにより得られたクラスタは，一つのクラスタに含まれる論文数にばらつきが生じていた．これは対象とする論文と引用関係が偏っていることが原因であり，偏ってしまう原因が，CiteSeer に入っていない，入っていても引用関係が取れていない．そもそも研究分野間で論文数や引用数が偏っている等であると考えられる．

また，クラスタリング閾値設定方法を利用し，クラスタ内の平均論文数を基準とすることにより，クラスタリング閾値を段階的に変化させ，クラスタの包含関係構造を表示できる機能を持たせる機能なども提供可能である．

6. 関連研究

アソシエーションルールを用いたクラスタリングとしては，ルールのハイパーグラフを基にクラスタリングする方法が研究されている [8]．この手法は，予め分割数を与える必要があり，ボトムアップに分割数を自動で定めることは今後の課題となっていた [9]．しかし，その後の研究では Nearest Neighbor を用いる研究に移っており，自動的に定める研究は行われていない．本研究では分割数ではなく，クラスタ内の平均論文数に着目し，独立したノードを減らすような方法を提案した．

7. まとめと今後の課題

リサーチマイニング手法において，クラスタリング閾値とクラスタ内の平均論文数の関係を調査し，さらに，クラスタ内の論文数の標準偏差とクラスタ内の平均論文数，クラスタリング閾値の関係から，クラスタの規模の変化を知ることができるという特徴を発見した．また，その特徴を利用することにより，クラスタリング閾値を定める方法の提案を行った．

今後の課題としては，本指針をより多くの対象に適用し，有効性を確認することが考えられる．また，今回は発展経緯として得られたもの全てに対しまとめて計算を行い，クラスタリング閾値設定指針を適用したが，リサーチマイニングで得られるグラフは独立した複数のグラフであるため，独立したグラフを個別にクラスタリングする手法を考えることも今後の課題である．さらに，グラフの形も考慮したクラスタリング手法を考案することもその一つである．

それに加え，発展経緯の重みはリサーチマイニングの適用対象により大きく異なるため，リサーチマイニング適用対象となる論文情報のより良い収集法も考える必要がある．

[謝辞]

本研究の一部は，文部科学省科学研究費補助金特定領域研究 (16016232, 16700023)，東京工業大学 21 世紀 COE プログラム「大規模知的資源の体系化と活用基盤構築」および科学技術振興事業団戦略的創造研究推進事業 CREST の助成により行なわれた．

[文献]

[1] M.M. Kessler. Bibliographic Coupling between Scientific Papers. *American Documentation*, Vol. 14, No. 1, pp. 10–25, 1963.

[2] H Small. Co-citation in the Scientific Literature:A New Measure of the Relationship between Two Documents. *Journal of the American Society for Information Science*, Vol. 24, pp. 265–269, 1973.

[3] 難波英嗣, 神門典子, 奥村学. 論文間の参照情報を考慮した関連論文の組織化. 情報処理学会論文誌, Vol. 42, No. 11, pp. 2640–2649, 2001.

[4] 吉田誠, 小林隆志, 難波英嗣, 奥村学, 横田治夫. Research Mining:研究論文データベースからの研究のマクロな流れの抽出. DEWS2003, 7-p, DEWS2003, 3 2003.

[5] 吉田誠, 小林隆志, 横田治夫. 公開されている論文 DB からのマクロ情報抽出に対するリサーチマイニング手法と他手法の比較. 情報処理学会論文誌データベース, Vol. 45, No. SIG 7(TOD 22), pp. 24–32, 6 2004.

[6] CiteSeer. http:// citeseer.ist.psu. edu/.

[7] Agrawal and Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, 1994.

[8] E.H.Han, George Karypis, Vipin Kumar, and Bamshad Mobasher. Clustering Based On Association Rule Hypergraphs. Proceedings of SIGMOD '97 Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.

[9] E.H.Han, George Karypis, Vipin Kumar, and B. Mobasher. Hypergraph based clustering in high-dimensional data sets: A summary of results. Technical Report 1, Bulletin of TCDE, 3 1998.

吉田 誠 Makoto YOSHIDA

平 15 東工大・工・電電卒．同大大学院・情報理工・計算工・修士課程在学中．日本データベース学会学生会員．

小林 隆志 Takashi KOBAYASHI

平 9 東工大・工・情報工学卒. 平 11 同大大学院・情報理工・計算工学・修士課程了. 平 16 同大大学院・同専攻・博士課程了. 平 14 同大学術国際情報センター・助手. 工博. 情報処理学会, 日本ソフトウェア科学会, 日本データベース学会, ACM 各会員.

横田 治夫 Haruo YOKOTA

昭 55 東工大・工・電物卒．昭 57 同大大学院・情報・修士課程了. 同年富士通 (株)．同年 6 月 (財) 新世代コンピュータ技術開発機構研究所．昭 61(株) 富士通研究所．平 4 北陸先端大・情報・助教授. 平 10 東工大・情報理工・助教授．平 13 東工大・学術国際情報センター・教授．工博．日本データベース学会, 電子情報通信学会, 情報処理学会，人工知能学会，IEEE，ACM 各会員.