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

吉田 誠[†] 小林隆志[††] 横田治夫[†,††]

インターネットの普及により，電子的に入手可能な論文数が増大している．それにともない，研究者がそれらの中から求めている情報を見つけ出すことの難易度が上がっている．このため，目的の情報を探し出すコストを減らすための機能が必要である．本研究の目的は論文のメタ情報，論文間の関係を解析することにより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である．我々は以前より研究のマクロな発展経緯を表現するリサーチマイニング手法を提案している．これまでに同一研究室内という閉じた環境の論文に対してリサーチマイニング手法の有効性を確認してきた．本稿では，公開されている電子化された論文 DB に対しリサーチマイニング手法の適用を試み，さらに同じデータセットに対し既存の手法である書誌結合，共引用分析を適用して比較することにより提案手法の特徴を明確化する．

Comparison of the Research Mining and the Other Methods for Retrieving Macro-information from an Open Research-paper DB

Makoto Yoshida,[†] Takashi Kobayashi[††] and Haruo Yokota[†,††]

By progress of the Internet, the number of research papers that can electronically be derived is increasing. However, the cost of searching them for the required information is still high. Therefore, some functions to reduce the cost is required. Our research goal is to provide an advanced retrieval method for the papers using macro-information about the progress flows of related researches. To derive the flows, relations between papers are analyzed by using the meta-information about papers. We call the method research mining since the association rule mining method is applied to derive the relations. So far, we showed the effectiveness of our research mining method by applying it to research papers in our laboratory. In this paper, we apply the method to papers stored in an open research-paper database to demonstrate the applicability of our approach. We then compare the results of the research mining with bibliographic coupling and co-citation analysis methods applied to the same set of papers. The comparison results make the feature of our approach clear.

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，目的とする論文をただちに得られることはあまり多くない．

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)] 等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．これは，参照，被参照関係にある論文は同じ主題を扱っているという理論であり，論文間に類似している要素があることが分かる．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

† 東京工業大学大学院情報理工学研究科計算工学専攻
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
†† 東京工業大学学術国際情報センター
Global Scientific Information and Computing Center, Tokyo Institute of Technology

引用分析は，2 論文が他の論文にとも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，結局目的の論文を検索するコストはあまり小さくならない．我々は研究の発展した過程を研究の発展経緯と呼んでいるが，検索コストを抑えるためにはこの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するためのアプローチとして，研究のマクロな流れを表現するリサーチマイニング手法を提案し，同一研究室内という閉じた環境の論文に対して提案手法であるリサーチマイニング手法を適用し有用性を確認した[4]．

本稿では，公開されている電子化された論文 DB に対してリサーチマイニング手法を適用する方法を提案し，書誌結合，共引用分析と比較することでリサーチマイニング手法の特徴を明確にすることを目的とする．

本稿ではまず，キーワード検索と参照関係のトレースを組み合わせた論文収集を行うことで，特定のテーマに関する論文から研究の発展経緯を抽出する方法を詳しく説明し，実際に NEC Research Institute が作成，提供している公開論文 DB である CiteSeer[5] を利用してリサーチマイニングの実験を行う．さらに同じデータに対し既存の手法である書誌結合，共引用分析を適用し，リサーチマイニング手法と比較を行う．

2. リサーチマイニング手法

2.1 概　　要

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．まず最初に参照関係のアソシエーションルールを発見し，そのアソシエーションルールから重み付き有向グラフを形成し，参照関係と比較することにより，研究の発展経緯を抽出する．さらに，その有向グラフを用いて，論文をクラスタリングする．クラスタリングを行うことにより研究の発展経緯をマクロな視点から見ることが可能となる．以下においてそれぞれについて詳しく述べる．

2.2 論文間の発展経緯の抽出

ここではデータマイニングのアプローチの 1 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[6] を利用する．アソシエーションルールとは発生割合が高い複数要素間の出現ルールのことである．トランザクションログの解析等に用いられることが多いが，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，とも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見なす．また，論文がとも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．

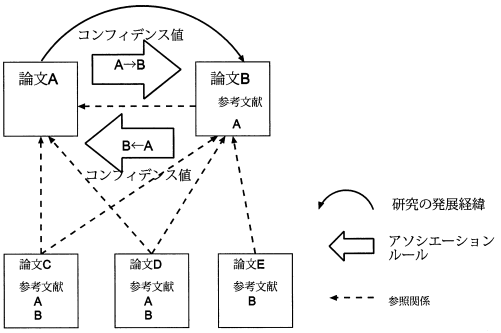

図 1 論文間のアソシエーションルール
Fig. 1 Association rules between research papers.

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発展経緯とする．

- B $\Longrightarrow$ A という参照関係が存在．
- A $\rightarrow$ 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ときに同時に新しい論文も参照していることが多いことに基づいている．たとえば**図 1** のように論文 A が論文 B の関連論文でかつ論文 B より前に発表されていた場合は，論文 A は論文 B より参照される回数が多くなり，その結果として論文 B が参照された際に論文 A が参照されている割合，すなわち B→A のコンフィデンス値は通常は大きくなる．逆に論文 A を参照して

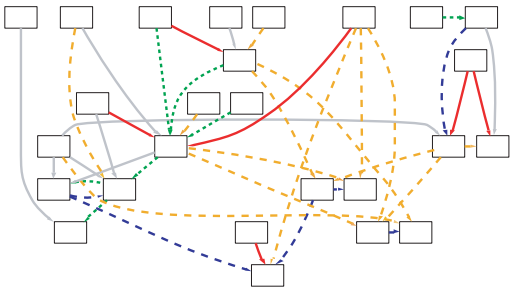

図 2 抽出した研究の発展経緯の一部
Fig. 2 A part of the progress flows of related papers.

いるときに論文 B を参照している割合である A→B のコンフィデンス値が大きい場合，新しい論文 B が多く参照されていることを示しており，研究が論文 A から論文 B へ推移していると考えることができる．

発展経緯抽出において，参照関係は論文間の強い関連と一種の時系列を示すものとして機能している．発展経緯抽出のためにあらかじめ定めた閾値のことを以降では，発展経緯抽出の重みの閾値と呼ぶ．

2.3 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，論文数が多くなると，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．**図 2** に示している抽出された研究の発展経緯を見ても，個々の研究の発展経緯が複雑に関係しあっていることが分かる．

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どうしは参照，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どうし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．たとえば論文 A，B，X が存在し，論文 A から論文 X への発展経緯の重みが 0.5，論文 B から論文 X への発展経緯の重みが 0.3 であり，論文 A，B が同一クラスタ C になる場合，クラスタ C から論文 X への発展経緯の重みは 0.5 となる．

リサーチマイニング手法で得られる上述のクラスタと共引用分析で得られるクラスタは，単に関連のある論文をまとめるか，論文の発展経緯を考慮し研究の発展経緯としてまとめるかという点が異なり，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．

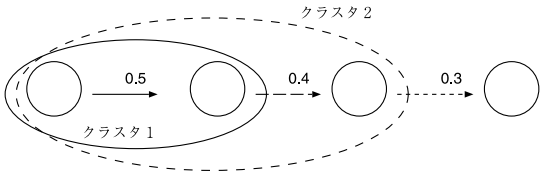

図 3 クラスタリング閾値によるクラスタ粒度の変更
Fig. 3 Cluster granularity with varying clustering threshold.

さらにリサーチマイニング手法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 3** 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．

2.4 関 連 研 究

2.2 節で示した図 1 の論文に対し，共引用分析[2] を適用した場合は論文 A，B というペアが得られるだけであり，方向を持つルールを検出することはできず，研究の発展経緯を抽出できない．また，書誌結合[1] は，対象が論文 C，D，E 側に対するクラスタリングであり，これも同様に研究の発展経緯の抽出はできない．しかし，リサーチマイニング手法では A→B，B→A という方向を考慮しているアソシエーションルールを検出し，コンフィデンス値と参照関係を考慮することで研究の発展経緯を抽出することができる．具体的な比較は，5.2 節と 5.3 節で行う．

共引用分析，書誌結合の研究の発展としては，論文の参照関係の分析を論文の検索，分類以外の目的のために利用している研究として論文誌の重要度を測るインパクト・ファクタ関連の研究[7),8)]，異なる種類の文献の関係をみる研究[9)] 等がある．しかし，これらはいずれも本研究には直接は関係しない．

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，アソシエーション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[10)]．しかしこの方法は，アイテムのグループ化を目的としているものであり，マルチレベルハイパーグラフパーティショニング[11)] を用いて，そのグラフを分割するこ

とでクラスタを形成するため，結果として得られるものはアイテムのグループのみである．そのため，この方法ではクラスタリングを行うことは可能であるが，クラスタ間の関係を抽出することはできない．リサーチマイニング手法では，アソシエーションルールと参照関係からグラフを形成し，その後に論文間の関連度の高いものを同一クラスタとしてまとめるために，結果として得られるものからクラスタ間の関係を知ることができる．

3. 適用対象論文の取得

我々は文献 4) において 1 研究室の論文を対象に，リサーチマイニング手法の有効性を確認した．リサーチマイニング手法において計算コストが最も必要となる部分は，アプリオリアルゴリズムにより 2 アイテムのアソシエーションルールを発見するところであり，V を総論文数とすると，そのオーダは $O(V^2)$ であるため，1 研究室の論文情報の量であれば計算コストはあまり大きくなく，また研究分野も絞られるため，すべてのデータを対象とすることができた．しかしながら CiteSeer[5]，DBLP[12] 等の，公開されている電子化された論文 DB への適用として，DB 上のすべてのデータを手法適応の対象とするのは計算コストの面から非現実的である．また，利用者の求める研究の発展経緯は，DB 全体ではなく，自分が注目する研究分野であると考える．そのため，必要な情報のみを適用対象とするべきであり，対象となる論文情報の取得方法が重要である．以下に，我々が採用した対象となる論文の段階的な取得方法を述べる．

まずキーワード検索により類似分野の論文情報を取得する．そして，得られた論文の参照関係を利用し，参照している論文の情報を取得する．さらに参照を繰り返したどることにより得られた論文に対しても同様に，参照先の論文情報を順次取得する（**図 4**）．この際，論文参照では広く関連論文をあげることが多いことから，一般に発散する傾向にあることを考慮し，繰返し回数を限定して論文情報を収集する．

4. 公開 DB を利用した実験

本稿では，公開されている論文 DB の 1 つである CiteSeer[5] を利用して，1) リサーチマイニング手法の適用対象を取得し，2) 得られた論文に対し リサーチマイニング手法を適用し評価を行うとともに，共引用分析および書誌結合と比較する．

4.1 対象論文取得

CiteSeer からキーワード検索により特定分野の論文を取得する．今回，キーワードは text，clustering とした．それらキーワード検索により得られた論文は 65 編であった．以降これを S1 と呼ぶ．

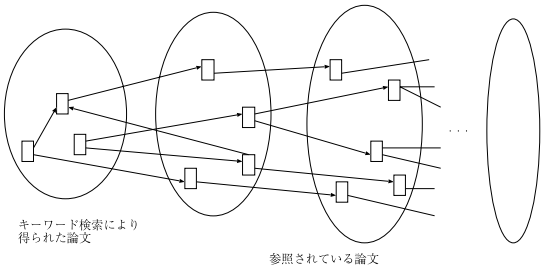

図 4 対象論文取得
Fig. 4 Collecting research papers for research mining.

表 1 各作業で得られた論文数と参照先の論文数（参照回数）
Table 1 The number of collected and cited papers in each step (the number of citations).

from \ to	S1	S2	S3
S1（65 編）	8 (8)	826 (1,072)	—
S2（243 編）	8 (8)	335 (993)	3131 (3768)

この S1 の 65 編の論文から S1 の論文自体への参照は 8 回，8 編であり，新しく出現する論文への参照回数は 1,072 回，重複を除くと 826 編の論文への参照であった．このうち CiteSeer に格納され参照情報を取得可能な論文は 243 編であった．これを S2 とする．

S2 の論文 243 編に対し，同様に S1 の論文への参照は 8 回，8 編，S2 の論文への参照は 993 回，335 編，新しく出現する論文への参照回数はこのうち参照情報が取得可能なものは 627 編であった．これを S3 とする．以上をまとめると**表 1** のようになる．

ここまでのステップで，参照関係が分かるのは，65 編 + 243 編の 308 編で，その中でののべ参照回数は 810 回，また，それらから参照情報が取得不可能な論文の参照回数は 5,039 回であった．

4.2 リサーチマイニング手法適用

4.1 節で得られた論文に対してリサーチマイニング手法を適用する実験を行った．各閾値として論文がともに引用されている最小回数は 2，3，発展経緯抽出の重みの閾値は，0.1，0.2，0.3 の場合を調べた．結果を**表 2** に示す．上段は論文がともに引用されている最小回数と発展経緯抽出の重みの閾値から得られた発展経緯の数（グラフの枝数）を表し，下段は論文がともに引用されている最小回数と発展経緯抽出の重みの閾値から研究の発展経緯が生じる論文数（グラフのノード数）を表す．

ここでは参照回数の閾値が 2 回，重みの閾値が 0.2 のものを以下の実験に用いる．対象となる論文がすで

表 2 共引用されている最小回数，発展経緯抽出の重みの閾値と得られた発展経緯の数
Table 2 Threshold of co-citation times, threshold of progress flow weight and the number of progress flows.

	0	0.1	0.2	0.3
2 回（発展経緯）	260	254	208	154
2 回（論文数）	184	183	175	152
3 回（発展経緯）	87	87	74	53
3 回（論文数）	76	76	73	59

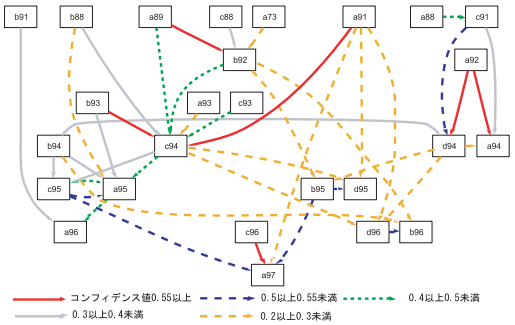

図 5 抽出した研究の発展経緯（最小回数 2，発展経緯抽出の重みの閾値 0.2）
Fig. 5 Progress flows of related researches (threshold of co-citation times is 2, threshold of progress flow weight is 0.2).

に類似分野の研究論文に絞り込まれていること，2 論文間の直接の参照関係も利用していることから，最小参照回数の閾値が 2 回でも十分であると判断した．なお，3 回とした場合は 2 回の場合のような大きなグラフは形成されなかった．また，発展経緯抽出の重みの閾値が 0.2 のとき，得られた発展経緯に出現している論文数が十分多く，抽出される発展経緯の数が絞り込まれている．発展経緯抽出の重みの閾値を変更することで利用者が希望する強度の発展経緯を抽出することが可能である．これは発展経緯抽出の重みの閾値により，強い発展経緯のみを抽出したり比較的関連の弱いものも抽出したりすることが可能ということである．しかし，論文間の関連度が小さい，すなわち重みが小さい発展経緯を結果として含んでしまうと，グラフが繁雑になり論文の位置付け，関連状況を把握することが難しくなる．

リサーチマイニング手法の適用により複数の有向グラフが得られたが，比較的大きいグラフを形成したものについて説明する．**図 5** が適用結果をグラフにしたものの一部である．この適用結果に対し，クラスタリングを行った．

クラスタリング閾値を 0.55 以上としてクラスタリングを行った結果，この部分のグラフから “Induction

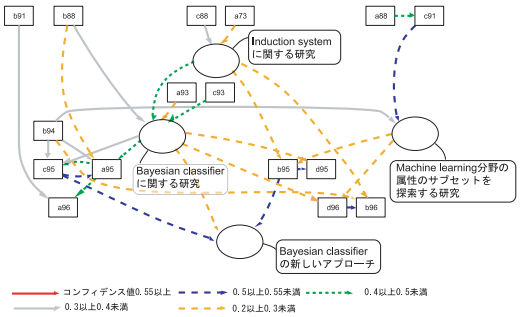

図 6 研究の発展経緯のクラスタ（クラスタリング閾値 0.55）
Fig. 6 Progress flows of related researches (clustering threshold is 0.55).

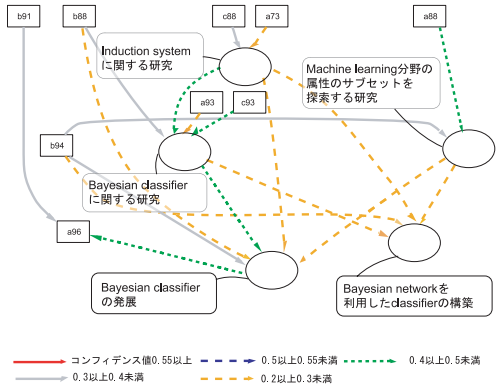

図 7 研究の発展経緯のクラスタ（クラスタリング閾値 0.5）
Fig. 7 Progress flows of related researches (clustering threshold is 0.5).

system に関する研究”（a89，b92）から “Bayesian classifier に関する研究”（a91，b93，c94）への発展経緯を得，さらに “Bayesian classifier に関する研究”と，“machine learning 分野での特徴，属性のサブセットを探索する研究”（a92，d94，a94）から “Bayesian classifier をふまえたうえでの新しいアプローチの研究”（a97，c96）という発展経緯を得た．クラスタリング閾値を 0.5 以上に変更した場合には，大まかな発展経緯は同じであったが，“新しいアプローチ” のクラスタがさらに c95，d95，a95，b95 とともに “Bayesian classifier の効率的な計算等の改良等の Bayesian classifier の発展” を表しているクラスタ（a97，c96，c95，d95，a95，b95）を形成した（**図 6**，**図 7**）．

また，初めに検索した結果である 65 編のうち，研究の発展経緯の枝が張られた論文は 1 編のみであった．これは検索を行った際にヒットした論文が比較的新しいものであったことが原因であると考えている．

4.3 実験に関する考察

4.3.1 情報収集について

発展経緯の枝を得るためには直接的な参照関係が必

要であるが，今回の評価実験では，データを手動取得しているために参照情報が少なく，その結果，研究の発展経緯の枝をあまり多く得ることができなかった．これは，繰返し参照をたどる回数を増やすこと，そしてキーワード検索結果で得られた論文を参照している論文の情報を収集することにより改善できると考えている．

また，キーワード検索の結果として得られた論文数が 65 編であるのに対し，参照関係をたどることにより 4022 編もの論文が出現した理由は，参照関係をたどることで検索の際に用いたキーワードを含まないが，内容が類似している論文が得られたことによると考えている．実際,「text，clustering」というキーワードで検索を行ったが，リサーチマイニング手法で得られた論文の中には classifier，classification という類義語が主題で扱われている論文が含まれていた．このように類義語を主題として扱っている論文が結果として出力されることは，類義語検索と同等の機能であり，類義語が既知でない場合にも結果として得られることは本手法の利点であると考えている．

4.3.2 リサーチマイニング手法適用結果について

リサーチマイニング手法を適用した結果，複数のグラフを生成した．その中で，複数のクラスタを形成する程度のまとまった量の論文で形成されている有向グラフの数は限られていたが，まとまった部分では実際の研究の発展経緯と合致していた．小さな部分でも研究の発展経緯の枝で結ばれていた論文どうしの内容は類似していた．

4.2 節においてはクラスタリング閾値を 0.55，0.5 とした例を示した．これは，1 つのクラスタ内に多数の論文が含まれる場合，発展経緯の理解が難しくなってしまうことから，1 つのクラスタに，2〜5 編程度の論文が含まれるようなクラスタリングを見るためである．クラスタリング閾値を下げるとクラスタ内に含まれる論文数が増え，逆に上げると減ることになる．たとえば，0.3 にした場合，小さなグラフの中でも 10 編以上の論文を含むクラスタを形成してしまう．ただし，クラスタ間の再帰的な発展経緯を見る場合には，クラスタリング閾値を下げることは有効である．クラスタリング閾値の適切な値の調査は今後の課題である．

また，このクラスタリング閾値は以前に行った研究室の論文に対するクラスタリング閾値と比べて高い閾値であった．これは初めにキーワード検索で内容の類似した論文からの参照関係をたどった際に得られた論文であることにより内容が類似しており，その結果，同一グラフ内の論文は重みが大きい枝で結ばれているものが多くなっていることが要因であると考えている．

表 3 各手法の適用結果
Table 3 Result of each method.

手法 \ 結果	得られた関係数	出現論文数
リサーチマイニング	208 本	175 編
共引用分析	4,119 組	675 編
書誌結合	3,455 組	283 編

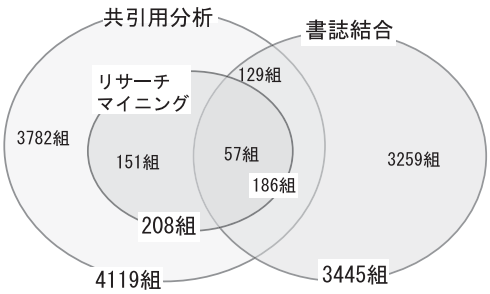

図 8 各手法により得られた関係の包含図
Fig. 8 Relation of each result.

5. 共引用分析と書誌結合との比較

5.1 共引用分析，書誌結合の結果

4 章の実験の対象とした同一の論文情報に対し，共引用分析と書誌結合を行った．結果を**表 3** に示す．なお，各手法における閾値は，リサーチマイニング手法ではミニマムサポート値 2，発展経緯抽出の重み 0.2，共引用分析では共引用されている回数 2，書誌結合では参照先の論文が重複している回数 1 である．

リサーチマイニングにより得られた研究の発展経緯の方向は無視し，各手法で得られた関係の包含図を**図 8** に示す．

5.2 共引用分析との比較

リサーチマイニング手法の結果は共引用分析結果のサブセットである．これは，リサーチマイニング手法ではミニマムサポート値を 2 で固定しており，アプリオリアルゴリズムからルールとして得られるものは，共引用されている最低回数が 2 になる．一方，共引用分析の閾値も 2 としているため，両手法の閾値が一致しており，その結果このような包含関係が存在する．

まず，リサーチマイニング手法および共引用分析の両手法で得られた関係について考察する．リサーチマイニング手法で得られた重みと共引用分析の共引用回数の間に相関関係は見られなかった（**図 9**）．

これは，各々の論文が参照される回数に幅があり，共引用分析の場合は多数の論文から参照されている論文ほど他の論文と関連しているという結果になるのに対し，リサーチマイニング手法では割合を結果として

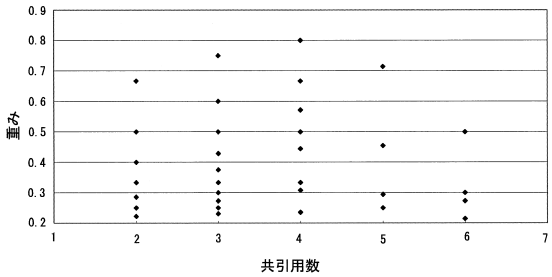

図 9 共引用分析との相関図
Fig. 9 Correlation between co-citation analysis and research mining method.

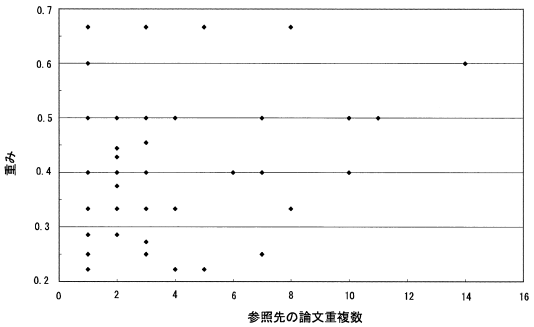

図 10 書誌結合との相関図
Fig. 10 Correlation between bibliographic coupling and research mining method.

いることに起因するものである．つまり，リサーチマイニング手法では，被参照回数にばらつきがある場合でも，そのことに影響されずに論文間の類似度を評価することができることを示している．

たとえば図 5 の a93 → c94 という研究の発展経緯の重みは 0.27 であり相対的に小さい値となっているのに対し，a93 と c94 の共引用数は 6 という比較的大きな値であった．この場合，実際の論文の内容は a93 は machine learning に関するものであり c94 は Bayesian classifier に関するものであることから比較的類似度は低く，我々はリサーチマイニング手法の結果の方が類似度の評価としてはより適していると考えている．

次に共引用分析では得られるが，リサーチマイニング手法では得られない関係について考察する．このとき，a) リサーチマイニング手法では直接の関係としては抽出されない場合，b) 研究の発展経緯として抽出される可能性はあるが重み閾値に満たない場合，c) 元々リサーチマイニング手法では得ることができない場合，の 3 つに分けられる．以下それぞれの場合を考察する．

a) リサーチマイング手法では 2 論文間の研究の発展経緯を表す枝が得られないが，共引用分析では両方の論文を参照しているという関係を得られた場合．
この場合の多くは，研究の発展経緯をたどることでその 2 論文の研究の推移という形で関連を知ることができる．たとえば，図 5 で，b95 と c95 の間には研究の発展経緯の枝がないが，共引用分析では抽出される．しかし，リサーチマイニング手法では発展経緯を a97 までたどることで研究の推移が分かる．
共引用分析では 2 論文間に関連があることは分かるが，それ以上のことは知ることができない．それに対し，リサーチマイニングでは論文の位置付けを知ることができるため，論文を検索する際に検索対象をしぼることが可能となり，この点でリサーチマイニングの方が優れている．このような関係のものは共引用分析の結果の値が大きいものが多かった．

b) 共引用分析では関係が得られたがリサーチマイニング手法では閾値により発展経緯とは判断されない場合．
この場合，それら論文間の関係が弱いことを示しているのでありリサーチマイニング手法の結果が妥当であったと考えている．しかしながら，現在は主観により閾値を定めているため，この値の適切な定め方を調査する必要がある．

c) 共引用分析では得られたがリサーチマイニングでは得られる可能性がない場合．
この場合，論文情報の不足が原因だと思われるものであった．つまり，リサーチマイニングでは参照している論文だけでなく，参照されている論文の参照関係を取得して初めて解析対象となるのに対し，共引用分析は参照されている論文の参照関係情報は利用しないことによる．論文情報のより良い取得方法を考えるのは今後の課題である．

5.3 書誌結合との比較

リサーチマイニング手法，書誌結合の両手法で得られた関係では，共引用分析との比較時と同様，リサーチマイニングの結果得られた研究の発展経緯を表す枝の重みと，引用先の重複数に相関関係は見られなかった（**図 10**）．これは書誌結合の強度は参照先の論文が重複数であるため，共引用分析のときと同様に絶対値で関連度を計っていること，そして書誌結合では対象となる論文が多くの論文を参照している場合に関連度が大きくなりやすい傾向にあることが一因であると考

えている．

もしサーベイ論文のような論文が対象に含まれると書誌結合の値が著しく大きくなってしまう．実際，実験においてそのような論文の組が確認された．リサーチマイニングでは関連の強度を割合で表すため，研究の発展経緯を抽出する際にはサーベイ論文を除外でき，また逆に，サーベイ論文が参照している論文のクラスタへの分散している割合からサーベイ論文を抽出することも可能であり，サーベイ論文を他の論文と区別できるためこの点で書誌結合より優れている．さらに書誌結合には，参照している論文数が少ない論文ではどんなに関連が大きい論文間であっても結果の値は低くなってしまう問題もある．

リサーチマイニング手法で得られた関係の中で，書誌結合に含まれない関係には，内容の関連が大きいものも存在した．しかしながら，書誌結合では抽出されない論文に傾向，関連は見られなかった．

書誌結合で得られた組のうち，リサーチマイニングで得られていないものは，以下の 4 つに分けることができる．

a) 新しい論文が含まれる関係の場合．
この場合，書誌結合では関連が大きいものから小さいものまでも結果として得られたが，リサーチマイニングでは関連度が大きいものであっても，発展経緯を抽出することはできない．これは，リサーチマイニングでは研究の発展経緯を抽出するために，他の論文から参照されることが必要であるためである．この問題は新しい論文のみでなく，他の論文から参照されている回数が少ない論文にも存在する．これらの論文を含むことができるようにすることは今後の課題である．

b) 共引用分析との比較時と同様に，書誌結合では 2 論文間に直接の関係が得られたが，リサーチマイニングでは研究の発展経緯をたどることで，2 論文間の関連を知ることができている場合．
この場合は，リサーチマイニング手法の方が研究の発展経緯を知ることができるため優れている．

c) 書誌結合で得られた関係が古い論文どうしのものであっても，リサーチマイニングでは抽出できなかった場合．
得られた関係が関連度が大きい場合，これはリサーチマイニング対象となる論文の収集，およびそのコストにかかわる問題である．参照を 2 回のみしかたどっていないため，得られた論文情報が少なく，またキーワード検索した論文を起点としているため，出現する論文や参照している論文に偏りが発生していることが原因であると考えている．しかしながら，関連のある論文の組をもれなく抽出することを試みるとノイズとなる結果が多く含まれてしまったり，計算コストが大きくなってしまったりするトレードオフが存在するため，適切な論文取得方法を考える必要がある．これは今後の課題である．

d) 書誌結合で得られた関係がリサーチマイングでは発展経緯の枝の重みが低いため，切られている場合．
この場合，リサーチマイニング手法および書誌結合の閾値の問題に帰着する．

以上から，書誌結合による研究の発展経緯の抽出の精度が低いことが分かった．しかしながら，リサーチマイニングの結果に新しい論文を補完するため等に利用可能であると考えている．また，参照の理由（参照の仕方）を考慮する手法[3] を用いることで，発展経緯の抽出精度を向上させることが可能であると考えているが，その手法との比較は今後の課題である．

6. おわりに

本稿では，特定テーマに関する論文の研究の発展経緯を抽出する方法として，キーワード検索と参照関係のトレースを組み合わせた論文収集により，公開されている電子化された論文 DB から論文を取得しリサーチマイニング手法を適用する手法を提案し，その有効性を確認した．さらに，同一のデータに対し共引用分析，書誌結合を適用し，比較を行うことでリサーチマイニング手法の特徴を明確化した．

リサーチマイニング手法の長所としては，

- 共引用分析と書誌結合に比べ，参照回数のばらつきに影響されない，
- 論文の発展経緯ならびに論文の集合としての研究の発展経緯を抽出可能，

といった点をあげることができる．一方，短所としては，

- 他手法と比べ多くの参照情報が必要，
- 最新の論文の発展経緯が得られない，

といった点がある．

今後の課題は，新しい論文もリサーチマイニング手法の結果として得られるようにすることである．それに関連して，論文収集の際に参照をたどるべき回数の調査，参照タイプを考慮するか等の参照のたどり方の調査を行う必要がある．また，それに加えてリサーチ

マイニング手法においても，参照タイプを考慮すること，そして他のキーワードを用いた実験を行うことや，参照情報以外のメタデータの利用やクラスタのラベルの自動生成を行えるようにすること，さらにクラスタリング閾値の適切な定め方の調査をすることも今後の課題である．

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（15017233）の助成により行われた．

参 考 文 献

1) Kessler, M.: Bibliographic Coupling between Scientific Papers, *American Documentation*, Vol.14, No.1, pp.10–25 (1963).
2) Small, H.: Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents, *J. Am. Soc. Inf. Sci.*, 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，ISSN 1347-4413, DEWS2003, 7-p, 電気情報通信学会 データ工学ワークショップ (2003).
5) CiteSeer. http://citeseer.ist.psu.edu/cs
6) Agrawal and Srikant: Fast Algorithms for Mining Association Rules, *Proc. 20th VLDB Conference* (1994).
7) Garfield, E.: Citation Analysis as a Tool in Journal Evaluation, *Science* (1972).
8) Garfield, E.: The Impact of Cumulative Impact Factors, *Proc. 8th IFSE Conference* (1995).
9) Narin, F., Olivastro, D. and Stevens, K.: Bibliometrics/Theory, Practice and Problems, *Evaluation Review*, Vol.18, No.1, pp.65–76 (1994).
10) Han, E.H., Karypis, G., Kumar, V. and Mobasher, B.: Clustering Based On Association Rule Hypergraphs, *Proc. SIGMOD '97 Workshop on Research Issues on Data Mining and Knowledge Discovery* (1997).
11) Karypis, G., Aggarwal, R., Vipin and Shekhar., S.: Multilevel Hypergraph Partitioning: Applications in VLSI Domain, *Proc. ACM/IEEE Design Automation Conference* (1997).
12) Digital Bibliography & Library Project: DBLP. http://dblp.uni-trier.de

(平成 15 年 9 月 25 日受付)
(平成 16 年 1 月 19 日採録)

（担当編集委員 石川 博，市川 哲彦，原 隆浩，佐藤 聡，土田 正士）

吉田 誠

2003 年東京工業大学工学部電気・電子工学科卒業．同年同大学大学院情報理工学研究科計算工学専攻修士課程入学，現在に至る．データ工学の研究に従事．

小林 隆志（正会員）

1997 年東京工業大学工学部情報工学科卒業．1999 年同大学大学院情報理工学研究科計算工学専攻修士課程修了．2002 年同専攻博士課程単位取得満期退学．同年より同大学学術国際情報センター助手，現在に至る．工学博士．ソフトウェア設計方法論，ソフトウェアパターン，ソフトウェアアーキテクチャ，方法論工学，データ工学等の研究に従事．日本ソフトウェア科学会，日本データベース学会各会員．

横田 治夫（正会員）

1980 年東京工業大学工学部電子物理工学科卒業．1982 年同大学大学院理工学研究科情報工学専攻修士課程修了．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（ICOT）．1986 年（株）富士通研究所．1992 年北陸先端科学技術大学院大学情報科学研究科助教授．1998 年東京工業大学大学院情報理工学研究科助教授．2001 年東京工業大学学術国際情報センター教授．工学博士．主として分散インデキシング，データ工学向けアーキテクチャ，高機能ストレージシステム，ディペンダブルシステム等に関する研究に従事．日本データベース学会理事．ACM SIGMOD 日本支部評議委員．電子情報通信学会データ工学研究専門委員会委員長．人工知能学会，IEEE，ACM 各会員．