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

吉田　誠[†]　小林　隆志[††]　難波　英嗣[†††]　奥村　学[††††]　横田　治夫[††]

† 東京工業大学 工学部電気電子工学科集積システムコース　〒 152-8552 東京都目黒区大岡山 2-12-1
†† 東京工業大学 学術国際情報センター　〒 152-8550 東京都目黒区大岡山 2-12-1
††† 広島市立大学 情報科学部　〒 731-3194 広島市安佐南区大塚東 3-4-1
†††† 東京工業大学 精密工学研究所　〒 226-8503 横浜市緑区長津田町 4259
E-mail: †yoshida@de.cs.titech.ac.jp, ††tkobaya@gsic.titech.ac.jp,yokota@cs.titech.ac.jp,
†††nanba@its.hiroshima-cu.ac.jp, ††††oku@pi.titech.ac.jp

あらまし　インターネットの普及により電子的な形で数多くの論文を得ることが非常に容易になった。しかしながら、それらの中から研究者が本当に読みたい論文を選択することは困難であり、支援を行うツールが必要である。このような状況では、論文のクラスタリングを行い、研究や研究グループの動向というマクロな流れを解析するような、リサーチマイニングのためのツールが有用である。参照関係や共引用を解析するツールはすでに存在するが、それらは研究のマクロな流れを見ることができない。本稿では論文の参照関係にアソシエーションルールを発見するためのアプリオリアルゴリズムを適用することにより、研究のマクロな流れを得るための手法を提案する。
キーワード　論文検索、データマイニング、クラスタリング、情報分析

Research Mining: Discovery of Research Macro-Flow from a Research-Paper Database

Makoto YOSHIDA[†], Takashi KOBAYASHI[††], Hidetsugu NANBA[†††], Manabu OKUMURA[††††], and Haruo YOKOTA[††]

† Faculty of Engineering Electrical and Electronic Engineering, Tokyo Institute of Technology.
†† Global Scientific Information and Computing Center, Tokyo Institute of Technology.
Ookayama 2-12-1, Meguro-ku, Tokyo,152-8550 Japan
††† Faculty of Information Sciences, Hiroshima City University.
Ozuka-higashi 3-4-1,Asami-ku,Hiroshima,731-3194 Japan
†††† Precision and Intelligence Laboratory, Tokyo Institute of Technology.
Nagatsudacho4259,Midori-ku,Yokohama-shi,Kanagawa,226-8503Japan
E-mail: †yoshida@de.cs.titech.ac.jp, ††tkobaya@gsic.titech.ac.jp,yokota@cs.titech.ac.jp,
†††nanba@its.hiroshima-cu.ac.jp, ††††oku@pi.titech.ac.jp

Abstract　It has become very easy to derive a great number of research papers in electronic form by progress of the Internet. However, researchers still need useful tools to select truly related papers to be read from them. A tool of research mining, which clusters research papers and analyses trends of research or researchers, must be effective in such situation. There are already tools for analyzing bibliographical relationship and co-citation, but they cannot illustrate macro-flow of research. In this paper, we try to discover the macro-flow of research by applying the apriori algorithm for mining association rules to cited papers.
Key words　Paper retrieval, Data mining, Clustering, Information analysis

1. はじめに

インターネットの発達などの要因により電子的に入手、利用可能な研究論文の数が増大してきている。一方、情報量が増大しても、個々の研究者が情報入手に費やす時間はあまり変わらないため、研究に関連する全ての情報を入手し、利用することが困難となっている。また、仮に全ての情報が入手できても処理能力に限界がある。

従来、自分が興味を感じる内容の論文や文献を検索する際には、タイトルや著者名等の限られた特定の検索キーを利用して検索作業を進めていた。これによって、利用者が求めている情報とは異なる多くの不要情報が利用者に知らされてしまうばかりか、目的とする内容の論文が見つけられない等の不都合があった。

我々はこれらの問題に対し、研究に関連した情報のクラスタリング、統合、簡略化等の負荷を減らすための情報処理が必要であると考えている。本研究の目的は、研究論文を研究内容によってクラスタリングし、そこから研究のプロジェクト間の関係、著者間の関係、研究の流れ等のマクロな情報を利用した高度な検索方法を提案することにある。

論文をクラスタリングするために必要となる、論文間の関係を分析する方法は、書誌結合 [2]、共引用分析 [3] などが提案されているが、これらの手法による分析結果だけでクラスタリングすることには限界がある。そこで、本研究では論文間の関係を分析するために、大量のデータの中から相関関係、規則性、パターンを取り出すデータマイニングの手法である アソシエーションルール発見アルゴリズムを適用することにより論文間の関係を分析し、研究の流れを抽出し、得られた関係を利用して論文をクラスタリングすることにより、研究のマクロな流れを表現する手法を提案する。

マイニングの対象となる論文の情報には、著者、タイトル、キーワード、発行年、出典、参考文献などがあるが、本稿では参照関係に対してアソシエーションルール発見アルゴリズムの 1 つであるアプリオリアルゴリズム [1] を利用してマイニングすることにより論文間の関係を分析し、クラスタリングする手法を提案する。また、本稿では実際のデータに対して本手法を適用しクラスタリングを行う実験を行い、その評価を行う。

2. 関連研究

論文間の関係を分析する方法に、書誌結合 (bibliographic coupling) [2]、共引用分析 (co-citation analysis) [3] が古くから知られている。

書誌結合とは、論文の関連度を測る時に、2 論文間でどれだけ同じ論文を引用しているか、という基準に基づいている手法である。参照、被参照の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わかる。

書誌結合を改良した研究を難波らが行っている。個々の研究論文に関する参照タイプを考慮し、参照のタイプを論説根拠型、問題点指摘型、その他型の 3 種類に分類することにより論文の類似度をはかる研究 [4]、それを利用し論文を組織化する研究 [5] であるが、クラスタリングに関しては十分な成果は得られていない。

また、 共引用分析とは、2 つの論文がどれだけ他の論文に共に引用されているか、という基準に基づいた手法である。ある論文に引用された複数の論文は互いに主題を扱っているという理論であり、これも論文間の類似している度合いを知ることが可能である。この値が高いほど 2 つの論文が類似しているという尺度になる。

書誌結合、共引用分析等は論文間の関係の存在を調べるためには有用であるがクラスタリングするためには情報が不足している。

共引用分析と参照関係のアソシエーションルールを用いる本手法との違いは、適用項目間の関係の方向性を分析できる点にある。

この方向性は、共引用分析が単純に 2 つの論文がともに同一の論文から参照されている回数を調べる点に対し、アソシエーションルールは、どちらの論文を主体として計算するかという 2 通りの計算結果が生じる。例えばある論文 A、B に対し、論文 A を参照している論文で論文 B が参照されている割合や、その逆の場合の値を知ることができる。本研究と共引用分析、書誌結合との違いに関しては 3.6 節で詳しく述べる。

また、WWW 上でハイパーリンクを解析し、コミュニティ群を発見する研究 [6] を豊田らが行っている。この研究では、あるweb ページからリンク構造により関連ページを発見し、その中で authority ウェイトの高いものを取り出し、その authority ウェイトが高いものの各ページをシードとして同様に関連ページを発見することにより上位 N 個の authority、hub を取り出し、その取り出した authority と hub の重複数よりはじめに発見した関連ページをグループに分類する。そしてそのグループの各 URL に対して関連ページ発見アルゴリズムを適用してそのグループのシードと関連ページ発見アルゴリズムの上位 authority を併せて合計で N 個取り出し、コミュニティとして出力する。しかし、対象を論文とした場合には authority、hub という概念をそのまま適用することはできない。

3. 研究の流れを抽出する手法

3.1 概　　要

本稿で提案する研究の流れを取り出す手法は、論文のデータベースから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、研究の流れを抽出するという方法である。本手法には論文間の流れの抽出、論文のクラスタリングという 2 つのフェーズがある。クラスタリングを行うことにより、研究の流れをマクロな視点から見ることを可能とする。以下でそれぞれについて詳しく述べる。

3.2 論文から得られる情報

論文をクラスタリングする際に利用できる、論文から得られる情報は以下のものである。

- 著者名

- キーワード、タイトル中の用語
- 参考文献
- 発行年
- 論文の長さ
- 参照タイプ
- 出典、学会名、ワークショップ名

本稿では参考文献に関して、後述するアプリオリアルゴリズムを適用することにより論文間のアソシエーションルールを発見、数値化する。他の情報に対してアプリアルゴリズムを適用し、その結果を利用する事に関しては今後の課題である。

3.3 アプリオリアルゴリズム

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 [1] が知られている。このアルゴリズムを利用することにより、発生割合が高い、複数要素の出現ルールを効率よく取り出すことが可能である。

以下にアプリオリアルゴリズムの手順を示す。

（1） 各アイテムの出現回数をカウントし、各アイテムについて出現確率 (サポート値) を計算し、そのサポート値が指定したミニマムサポート値以上のものをラージアイテムセット L_1 に追加する。

（2） 各 $k = 2, ...$ について L_{k-1} が空集合で無いかぎり以下を繰り返す。

（a） L_{k-1} を利用してアイテム数が k 個のすべての組合わせ (キャンディィデイトアイテムセット) を作る。

（b） キャンディィデイトアイテムセットの中で各サポート値を計算し、ミニマムサポート値以上のサポート値を持つアイテムセットをラージアイテムセット（L_k）とする。

（3） アイテムセットが示すルールが真になる確率 (コンフィデンス値) が指定したミニマムコンフィデンス値以上のものをアソシエーションルールとする。

本研究では、1 つの論文を 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、方向付けを行う。

3.4 流れの抽出

論文をノード、アプリオリアルゴリズムの結果として得られたルールをコンフィデンス値を重みとした有向枝とすることにより、重み付き有向グラフを作成する。

上記のルールを考えた場合、通常は起源の論文に近い論文のほうが参照される回数が多くなる。しかし、後に発表された関連論文であっても、前に発表された論文の内容を包含していたり、影響力が大きい論文は参照される回数が多くなり、その結果、“ 関連研究の起源に近い論文を参照しているならばその後の論文を参照している ”、という割合が高くなる。例えば論文 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 A が参照された際に論文 B が参照されている割合（A→B のコンフィデンス値）は通常は低い。しかし、論文 B が影響力がある場合や研究として優れている場合はこの値が低くならない。

このことから本研究では、参照関係の方向と比べて逆向きの枝で、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流れを表す枝として扱う。

3.5 クラスタリング

論文単位での研究の流れを追うには、前述した研究の流れを抽出するという処理を行うのみでも十分だが、論文数が多い場合はそれだけでは研究の流れを知ることが困難である。

対象の論文数が多い場合には、よりマクロな視点として研究分野単位での流れを知ることが有用であり、本研究ではこのマクロな流れを表現するために、上述のグラフに対してクラスタリングを行う。クラスタリング手法は、研究の流れ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研究内容が近いことから、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ようにする。本研究ではこの閾値をクラスタリング閾値と呼ぶ。

この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、研究のマクロな流れを柔軟に見ることを可能にする。図 1 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、円は論文、枝は研究の流れ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 2 は重みが 0.4 以上のものを同一クラスタとしたものである。クラスタ 2 はクラスタ 1 より大きい粒度になっていることがわかる。

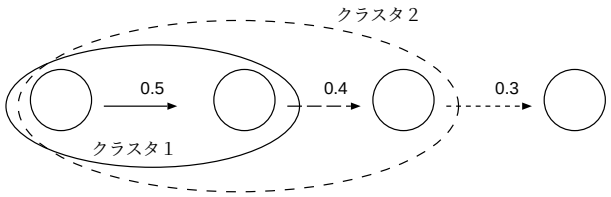

図 1　閾値によるクラスタ粒度の変更

3.6 共引用分析および書誌結合との違い

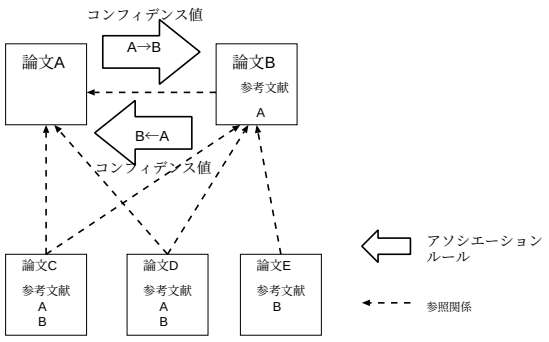

図 2　論文参照の共引用分析、書誌結合、アソシエーションルールの違い

例として図 2 のような参照関係がある論文を考える。これらの論文に対し、共引用分析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、研究のマクロな流れを抽出できない。

また、書誌結合は、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マクロな流れの抽出はできない。

しかし、本研究では 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流れを抽出することができる。さらに、コンフィデンス値に対してクラスタリング閾値を設定することでクラスタリングを行い、マクロな流れを知ることを可能としている。

4. 実　　験

本研究では提案手法を評価するために評価実験を行った。実験対象は当研究室にある 1994 年から 2002 年に外部に発表された 123 本の論文情報を利用した。この情報は 123 本中 87 本の論文の当研究室内での参考文献が入力してあり、その総数は 373 である。論文の参照関係にアプリオリアルゴリズムを適用する。2 論文間の関係を利用するため、今回はアプリオリアルゴリズムをラージ 2 アイテムセットまで適用する。また、研究の流れの枝を抽出する際に、重みを閾値により制限しているため、ミニマムコンフィデンス値は 0 としている。そこで得られたルールを利用して論文間の研究の流れを抽出し、重みがクラスタリング閾値より大きい枝でつながっている論文を同一のクラスタとしてまとめて扱う。今回はミニマムサポート値が 0、0.003 の 2 通り、研究の流れとして扱う最小のコンフィデンス値を 0、0.1、0.2、0.3 の 4 通り、クラスタリング閾値が 0.5、0.4、0.3 の 3 通りを調べた。

4.1 実験結果

表 1　ミニマムサポート値、流れ抽出の重みの閾値と得られた流れの数

ミニマムサポート値 ＼ 流れ抽出の重みの閾値	0	0.1	0.2	0.3
0	170	134	89	59
0.003	109	101	70	49

表 2　ミニマムサポート値、流れ抽出の重みの閾値と研究の流れが生じる論文数

ミニマムサポート値 ＼ 流れ抽出の重みの閾値	0	0.1	0.2	0.3
0	67	59	52	46
0.003	47	44	39	35

ミニマムサポート値が 0、0.003 である 2 つを比べた場合、0.003 である場合は 0 である場合に得られた流れとほぼ同じ流れが抽出可能であったが、新しい論文や、被参照回数が少ない論文に枝が張られないという違いがあった。しかし、ミニマムサポート値が 0 である時には存在していた枝が 0.003 である時には存在しない場合があり、その枝が実際の研究の流れを表していたため、その分、細かい流れ（論文単位など）を見る場合には結果は悪かった。しかし、マクロな流れを抽出する際には抽出可能であった。

ミニマムサポート値が 0.003 の場合は 0 の場合の枝の数が少なくなったものであるため、0 の場合についてのみ説明し、0.003 に関する図、説明は省略する。ミニマムサポート値、流れ抽出の重みの閾値、得られた研究の流れの枝の本数を表 1、また、研究の流れの枝でつながれた総論文数を表 2 に示す。

本稿ではミニマムサポート値を 0、研究の流れ抽出の重みの閾値を 0.2 のものを対象にして説明する。

論文間の研究の流れを示したものが図 3 である。この図で緑色の細かい破線 (白黒の場合は黒色の細かい破線) の枝は重みが 0.5 以上のものを表し、赤色 (白黒の場合は黒色) は 0.4 以上 0.5 未満、青色の破線 (同じく黒の破線) は 0.3 以上 0.4 未満、灰色 (灰色) は 0.25 以上 0.3 未満、黄色の破線 (灰色の破線) は 0.2 以上 0.25 未満を表している。これ以降に出現する図も同様である。

前述のように、クラスタリング閾値を設定し、重みが閾値より大きいものに対してその枝に付随しているノードを同一のカテゴリに属する論文とみなすことによりクラスタを形成する。

図 3 を元に、閾値を 0.3、0.4、0.5 としてクラスタリングを行った。その図が図 4、図 5、図 6 である。楕円はクラスタ、四角形は単一の論文を表している。

なお、これらの図において、研究の流れを表す枝でつながれていない論文に関しては、図が見にくくなるためこの図では省略している。

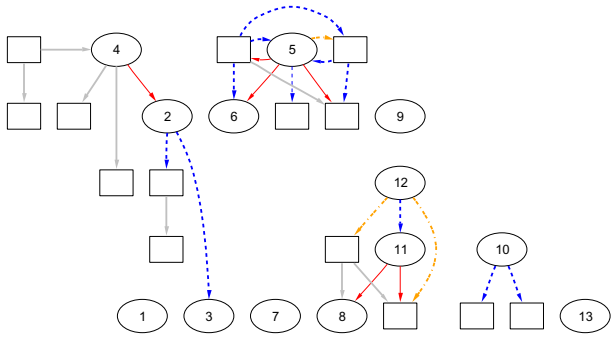

図 4　クラスタリング閾値 0.5 のクラスタ

閾値を 0.5 としてクラスタリングを行ったものが図 4 である。各クラスタがどのような研究内容になっているかを説明する。図 4 における、各クラスタの内容とそのクラスタに含まれる論文数を表したものが表 3 である。

この図 4 からは 4→ 2 → 3 という研究の流れがわかる。この流れは「並列アクティブデータベース」→「並列データベースディレクトリ構成」→「ディレクトリ構成 Fat-Btree」 という流れである。この流れは実際の研究の流れと一致していた。

他には、クラスタ 5 の付近では、5→ 論文単体 → 6、5→ 論文単体などの複数の流れが抽出できるが、研究の流れの内容としてはほぼ同一と考えることができ、「RAID のネットワーク上への展開、DR-net に関する研究」→「並列ディスクに関する研究」という流れである。これも実際の研究の流れと合致して

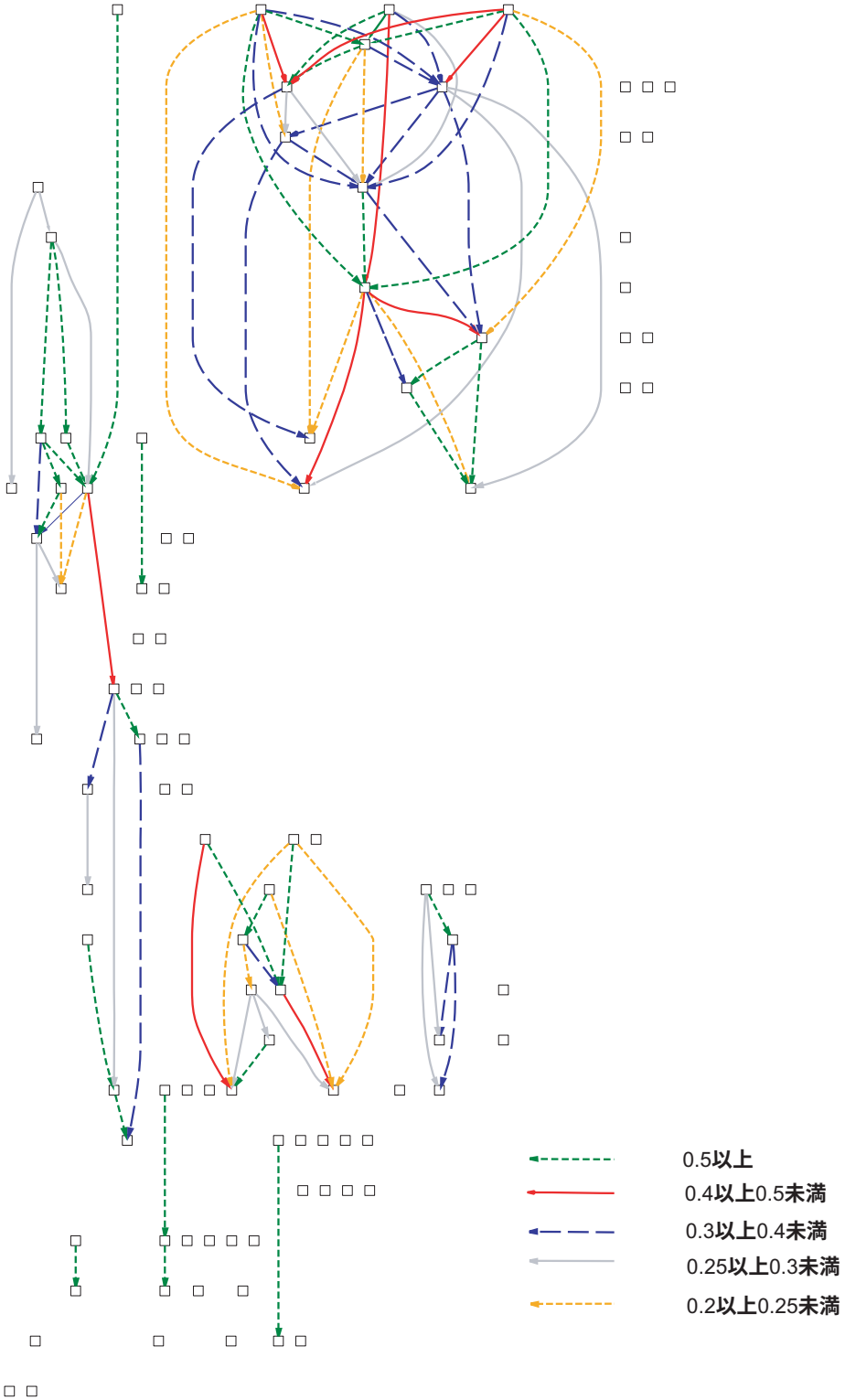

図 3　抽出した研究の流れ、ミニマムサポート値 0、研究の流れ抽出の重みの閾値 0.2

いた。

また、12→ 11→ 8、という流れから、「Fat-Btree と自律ディスクに関する研究」→「自律ディスクに関する研究」という流れがわかる。これは Fat-Btree に関する研究が自律ディスクに関する研究に影響を与えていることを示している。クラスタ 11、8 は内容的にも類似しているがクラスタ 8 は比較的に新しいものであるためにクラスタ 11 と分かれてしまっていると考えている。

10→ 論文単体という流れも実際の研究の流れと合致していた。そのほかクラスタではない論文単体への枝も同様に実際の研究の流れを表していた。

次にクラスタリング閾値を 0.4 としたものを図 5 に示す。また、各クラスタの内容とそのクラスタに含まれる論文数を表 4 に示す。

表 3　クラスタリング閾値を 0.5 としたクラスタ及び研究内容

クラスタ	研究内容	論文数
クラスタ 1	自律ディスクに関する研究 (新しい)	3
クラスタ 2	並列データベース更新を考慮したディレクトリ構成に関する研究	2
クラスタ 3	Fat-Btree に関する研究	3
クラスタ 4	並列アクティブデータベースに関する研究	7
クラスタ 5	RAID、ネットワーク上への展開、並列ディスク、	
	DR-net に関する研究	7
クラスタ 6	並列ディスクシステムに関する研究	3
クラスタ 7	ログに関する研究	2
クラスタ 8	自律ディスクに関する研究 (比較的新しい)	2
クラスタ 9	フォールトトレラントソフトウェアに関する研究	2
クラスタ 10	入れ子トランザクションに関する研究	2
クラスタ 11	自律ディスクに関する研究	3
クラスタ 12	自律ディスクと Fat-Btree に関する研究の起源である論文	2
クラスタ 13	コンテンツ統合に関する研究	2

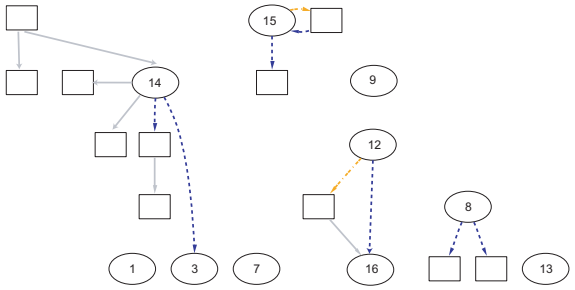

図 5　クラスタリング閾値 0.4 のクラスタ

表 4　クラスタリング閾値を 0.4 としたクラスタ及び研究内容

クラスタ	表 3 との対応	論文数
クラスタ 14	クラスタ 2、クラスタ 4	9
クラスタ 15	クラスタ 5、クラスタ 6、論文単体 × 2	12
クラスタ 16	クラスタ 11、クラスタ 8、論文単体	6

この結果、クラスタ 15 付近で 15→ 論文単体、論文単体 → 15 という双方向に枝を持つものが得られた。クラスタ 15、その隣の論文共に DR-net に関する内容である。これは時間的に差がある複数の論文がクラスタ 15 に属していることによる。この場合は、この 2 つは類似しており、クラスタから両方向の研究の流れを示す枝で論文単体がつながれている場合には、その論文はそのクラスタに属するべきであると考えているが、事例が少ないために確認できていない。

クラスタリング閾値が 0.5 のときに別れていた「並列アクティブデータベース」、「並列データベースディレクトリ」という図 4 のクラスタ 4、2 が融合し、「並列データベース」というクラスタ 14 を構成している。同様に図 4 のクラスタ 5 と 6、論文単体が融合しクラスタ 15 に、クラスタ 11 と 8、論文単体が融合してクラスタ 16 になっている。

次にクラスタリング閾値を 0.3 とした場合の結果を図 6 に、各クラスタの内容と論文数を表 5 に示す。

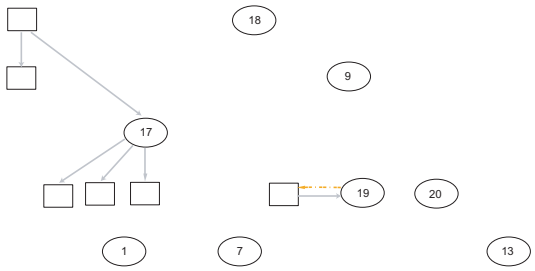

図 6　クラスタリング閾値 0.3 のクラスタ

表 5　クラスタリング閾値を 0.3 としたクラスタ及び研究内容

クラスタ	表 3 との対応	論文数
クラスタ 17	クラスタ 2、3、4、論文単体	13
クラスタ 18	クラスタ 5、6、論文単体 ×4	14
クラスタ 19	クラスタ 8、11、12、論文単体	8
クラスタ 20	クラスタ 10、論文単体 ×2	4

クラスタリング閾値に対して対象論文の発表されたスパンがあまり長くはないため、流れというよりも大きなクラスタが生成される結果が生じた。

以上のクラスタリング閾値に対する、クラスタ内の平均論文数を表 6 に示す。

表 6　クラスタリング内の平均論文数

クラスタリング閾値	0.5	0.4	0.3
クラスタ内の平均論文数	3.1	4.3	6

5. 考　　察

5.1 研究の流れ、ミニマムサポート値

得られた結果を判断するには主観的な部分が入ってしまうが、実験で得られた研究の流れは、実際の研究の流れと合致していた。しかしながら、新しい論文に関しては参照されている回数が少ないためにアソシエーションルールが発見されず、研究の流れが抽出されなかった。これは共引用分析を用いても同様の問題がある。

また、今回は実験対象が同一研究室の論文であったためにミニマムサポート値が小さい方が良い結果が得られた。しかし、実験対象を多数の論文に拡大した場合、ノイズが入ると予想しているため、ある程度ミニマムサポート値を上げる必要があると考えている。アプリオリアルゴリズムを適用する際のミニマムサポート値をあげることにより、抽出される結果にノイズが入りにくく、その上計算量が減るという利点が生じるが、比較的に新しい研究論文に関するアソシエーションルールも発見されない。逆にミニマムサポート値を下げると、ある程度新しい論文からも研究の流れを抽出できるがノイズが入る可能性が高くなる。

5.2 クラスタリング閾値

クラスタリング閾値の増減により、大きい時は比較的に細かい粒度で研究の流れを追うことが可能であり、小さい時は荒い粒度で追うことが可能であった。また、研究の流れとして扱うルールの閾値を上げると、大きな流れのみを抽出することが可能であり、下げると比較的弱い流れも抽出することが可能であった。

5.3 そ の 他

参照、被参照の関係にない場合にも、コンフィデンス値が高いルールが存在する場合がある。両方向のルールの重みが大きいものは同じ分野に属する論文であり、直接参照、被参照の関係にない論文の研究の流れの抽出、クラスタリングを補助することができると考えている。

また、研究の起点、もしくは分岐点となる研究論文は参照される回数が多くなるため、参照回数も考慮した計算を行うことで起点となる論文を特定することができ、クラスタのラベル付け等として利用できると考えているが、具体的な利用方法は今後の課題である。

6. ま と め

本研究では、論文のメタ情報から研究のマクロな流れを抽出する手法を提案した。本手法では、論文の参照関係にアプリオリアルゴリズムを適用することによって参照されている論文間の関係を重み付き有向グラフとして表現し、そのグラフと参照関係のグラフから論文間の研究の流れを取り出す。また、研究の流れのグラフを利用してクラスタリングを行い、マクロな研究の流れを抽出する手法を提案した。

また、提案手法を実際の研究論文に対して適用し、研究の流れを抽出し、クラスタリング、本手法で得られた研究の流れと実際の研究の流れを比べることにより、本手法がクラスタリング及び研究の流れを抽出することに有効であることを示した。

アプリオリアルゴリズムを適用する際のミニマムサポート値、クラスタリング閾値を変更することにより、異なる粒度で柔軟に研究の流れを知ることが可能であることを示した。

7. 今後の課題

本手法では、新しい論文を含むアソシエーションルールを発見できないといった問題点がある。これは新しい論文は参照されている回数が少ないことによる問題である。この新しい論文をクラスタリングできないという問題は書誌結合等と組み合わせるか、他の要素をデータマイニングすることにより解決したいと考えている。現在考えている方法は、あらかじめこの手法で生じたクラスタの第一著者が、どのクラスタにどのような割合で含まれているかを計算し、クラスタリングしたい新しい論文の第一著者だけでなく著者全てに対してに対して、その著者がどこのクラスタにはいっているかを計算し、その論文が属する可能性の最も高いクラスタに属させるという方法である。

また、今回は同一研究室内の論文のみという閉じた環境での論文について適用したが、それだけではなく多数の論文に適用することにより有効性を示すことも今後の課題である。その際には、文献[4][5]等で研究されている参照タイプを考慮することも重要になると考えている。

それに加えて、生成したクラスタのラベルを自動生成することも課題である。この問題は論文のタイトル、キーワード等の情報の利用により解決可能ではないかと考えている。

さらに、表示ツールの実装も課題のひとつである。

謝 辞

本研究の一部は、文部科学省科学研究費補助金基盤研究(14019035)の助成を受け、日本データベース学会・マイクロソフト株式会社 共催 2002 年度データベース研究支援プログラムの一環として行なわれた。

文 献

[1] Rakesh Agrawal and Ramakrishnan Srikant. ”Fast Algorithms for Mining Association Rules.” Proceedings of the 20th VLDB Conference,pages 487-499, 1994

[2] Kessler,M.M. :Bibliographic Coupling between Scientific Papers,American Documentation,Vol.14,No.1,pp.10-25,(1963)

[3] Small,H :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)

[4] 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

[5] 難波英嗣. 論文間の参照情報の抽出と利用に関する研究. 北陸先端科学技術大学院大学博士論文

[6] 豊田正史. WWW における関連コミュニティ群の発見. 情報処理学会研究報告, Vol. 2000, No. 69 (データベースシステム研究報告 No.122), pp. 307-314, 2000.