リサーチマイニング手法を用いた研究の発展経緯可視化ツール

吉田　誠[†]　小林　隆志[††]　横田　治夫[††]

† 東京工業大学大学院 情報理工学研究科計算工学専攻　〒 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
E-mail: †yoshida@de.cs.titech.ac.jp, ††tkobaya@gsic.titech.ac.jp,yokota@cs.titech.ac.jp

あらまし　ネットワーク技術の発達等により、電子的に利用可能な研究論文数が増大してきている．それに伴い，研究者が求めている情報を見つけ出すコストが増大している．このため，目的の情報を探し出すコストを減らす必要がある．本研究の目的は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．そのための手法として我々はリサーチマイニング手法を提案している．しかしながら，本手法により得られた研究の発展経緯を表すグラフは、多数のサブグラフから構成されるため，得られたグラフから各論文間の研究の発展経緯を直感的に把握することが難しいという問題があった．そこで本稿では、得られた発展経緯をユーザに見やすく提供するツールを提案，実装し，その有効性を確認する．

キーワード　マイニング一般，テキストマイニング，データの可視化

A Visualization Tool for Macro-Flow of Research by Mining Research Papers

Makoto YOSHIDA[†], Takashi KOBAYASHI[††], and Haruo YOKOTA[††]

† Department of Computer Science, Graduate School of Information Science and 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
E-mail: †yoshida@de.cs.titech.ac.jp, ††tkobaya@gsic.titech.ac.jp,yokota@cs.titech.ac.jp

Abstract　By progress of the Internet, the number of research papers that can electronically be derived is increasing. However, the cost of searching them for 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 research papers. We have proposed a method of mining research papers to find macro-flow of research. Because it consists of many sub graph,it is still difficult to grasp research macro-flow from the derived result graph. In this paper, a visualization tool providing users visible research macro-flow is proposed and implemented, and the validated.

Key words　General mining, Text mining, Visualization of data

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伴い，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わかる．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をあまり小さくすることはできてない．本研究の大きな目的は，論文を検索するためのコストを低減することであり，そのためには研究の発展した過程を抽出し，利用することが必要である．本研究ではこの “研究の発展した過程” を研究の発展経緯と呼ぶ．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．

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ている [4]．さらに公開論文 DB から，キーワード検索と参照関係を用いた方法により論文情報を収集しリサーチマイニング手法を適用した．そして得られた発展経緯と共引用分析や書誌結合の結果を比較することで本手法の有用性を確認してきた [5]．また本手法では，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，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の考察も行った [6].

しかしながら，これまで本手法により得られた論文の発展経緯を表すグラフは、多数のサブグラフから構成されるため，得られたグラフから各論文間の研究の発展経緯を直感的に把握することが難しいという問題があった．そこで本稿では、得られた発展経緯をユーザに見やすく提供するツールを提案，実装し，その有効性を確認する．

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，リサーチマイニング手法適用の際に得られた発展経緯の重みの分布に応じた，クラスタリング閾値の定め方の指針について説明する．そして，研究の発展経緯確認ツールについて説明をし，実際に論文情報に適用し，有効性を確認する．

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，リサーチマイニング手法の詳細は [5] を参照されたい．

2.1 論文間の発展経緯抽出

論文間の発展経緯の抽出には，データマイニングでのアプローチのひとつであり，アソシエーションルールを効率良く発見する方法である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．

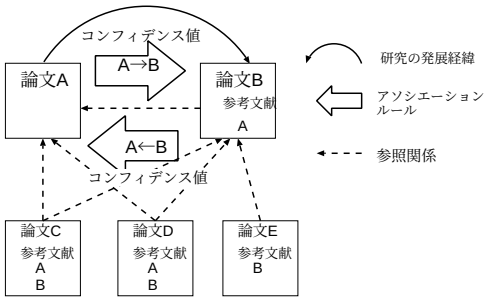

図 1　論文間のアソシエーションルール
Fig. 1　Association rules between research papers

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．

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．

2.2 クラスタリング

2.2.1 クラスタリング方法

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こと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．

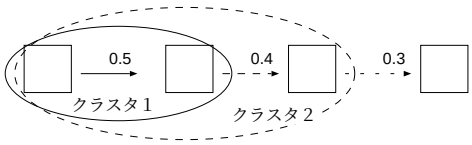

図 2　クラスタリング閾値によるクラスタ粒度の変更

Fig. 2　Cluster granularity with varying clustering threshold

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．なお，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．

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．

2.2.2　クラスタリング閾値設定指針

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できるため，マクロな発展経緯を見たい場合には大きい粒度のクラスタを，詳細な発展経緯を見たい場合には小さい粒度のクラスタを提示することができ，利用者のニーズに合わせてさまざまな粒度の発展経緯を見ることが可能である．しかしながら，目的とするクラスタの粒度に対して，クラスタリング閾値をどのような値に設定するべきかを決定することは難しい．ここではクラスタリングに際し，マクロな研究の発展経緯の理解を助けるようなクラスタリング閾値の設定指針について説明する．詳細は [6] を参照されたい．

一般に，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．

我々はこれまでの研究 [6] により，クラスタリング閾値の増加に伴い，クラスタ内の平均論文数は，図 3 に示すように，若干振動してはいるものの，ほぼ単調に緩やかに減少することがわかっている．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，二分探索を用いることで，指定されたクラスタ内の平均論文数から，対応するクラスタリング閾値を求めることが出来る．

また，クラスタ内の平均論文数の変化率はほぼ一定であるのに対し，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存在する．

このように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ている．

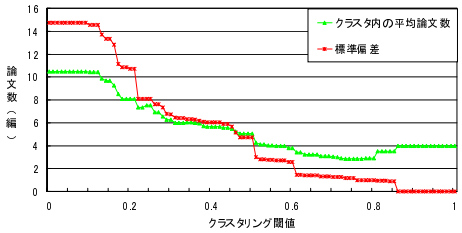

図 3　クラスタリング閾値とクラスタ内の平均論文数，標準偏差の変化の例

Fig. 3　Example of average cluster granularity and standard deviation with varying clustering threshold

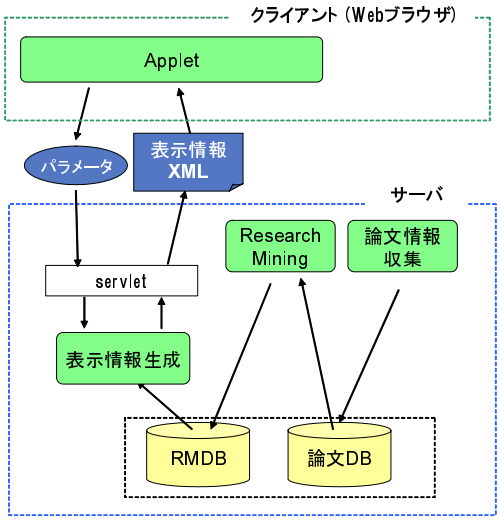

図 4　システム構成

Fig. 4　Details of system

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．

そこで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している．

- まず初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．

3.　システムの概要

システムの全体の構成を図 4 に示す．円柱はデータベースを表している．本システムは，論文情報収集部分 (論文情報収集) と論文の参照関係へのリサーチマイニング手法適用の部分 (Research Mining)，得られた発展経緯をクラスタリングして表

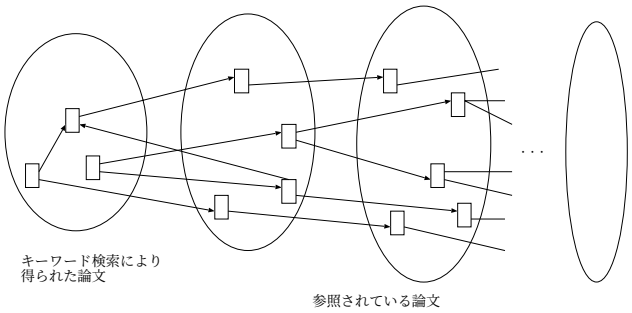

図 5　対象論文取得

Fig. 5　Collecting research papers for research mining

示する部分 (表示情報生成) から成る．

本システムの動作は，まず，論文情報収集部分において論文情報収集を行う．収集方法については後述する．次に，Research Mining 部分において論文情報に対しリサーチマイニング手法を適用し，発展経緯を抽出，テーブルに保持しておく．その後，利用者のリクエストに対し，発展経緯を描画する．グラフ表示する際には，ブラウザ側からサーブレットを介し，得られた発展経緯が格納されているテーブル名とクラスタリング閾値，発展経緯抽出の重みの閾値をパラメータとして指定する．

論文情報の収集方法はいろいろ存在するが, [5] においては以下の論文情報収集方法を用いた．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 5）．この際，論文参照では広く関連論文を挙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，繰り返し回数を限定して論文情報を収集する．

本稿では，抽出した発展経緯を利用者に表示する部分について，詳しく扱う．

4.　研究の発展経緯確認に必要な機能

研究の発展経緯確認ツールに必要であると考えている機能について述べる．

4.1　論文情報表示機能

論文タイトルや著者名，論文発行年，論文への URL 等，論文情報を表示する機能である．これは，発展経緯を確認するために必要な機能である．

4.2　クラスタリング閾値候補表示機能

クラスタリング閾値の候補をユーザに提示する機能である．本研究ではクラスタリングの結果の参考値として，クラスタ内の平均論文数を用いている．クラスタリングに際して，希望のクラスタ内の平均論文数が既知の場合には，2.2.2 の方法でクラスタリング閾値を求めればよいが，希望のクラスタ内の平均論文数が定まってない場合もある．その場合は，クラスタリング閾値をいくつに設定するべきかわからない．そこで，2.2.2 の方法を利用し，あらかじめ利用者にクラスタリング閾値の候補を提示し，その中から選択するような機能が有用である．

4.3　クラスタ粒度把握機能

クラスタに含まれる論文数により，クラスタを表すノードの色を変化させ，クラスタの粒度を一目で知ることができるような機能である．

一般に，リサーチマイニング手法を適用するための論文情報収集に際して，3. に記した方法により論文情報を収集した場合，論文 DB に格納されている全ての情報を利用する場合，どちらの場合であっても，異なるテーマの論文の情報が含まれる．これは，論文 DB に格納されている全ての情報を利用する場合は，当然，多くの異なる分野の論文が含まれる．一方，上記の方法で情報収集した場合であっても，各々の論文が参照してる論文が必ずしも参照元の論文のテーマと一致してはいないため，リサーチマイニング手法により得られた論文の発展経緯を表すグラフは、多数のサブグラフから構成される．そのため，どのサブグラフに注目するべきかわからないという状態に陥る可能性がある．本機能によりこれを解消できる．

4.4　グラフ選択機能

選択したサブグラフ以外のグラフを非表示にする機能である．ある論文やクラスタを表すノードに注目し，そのノードと関係がある部分のみを見たいと思った場合であっても，そのノードと関連がない他のサブグラフのノードや発展経緯を表す枝が存在してしまう場合には，グラフがノード間の関係の把握が困難になってしまう可能性がある．そのため，本機能が必要である．

4.5　グラフ操作機能

グラフを動かすことを可能とする機能である．この機能により，グラフで表されたリサーチマイニング結果を多角的な視点から視覚的に観察でき，研究の発展経緯理解や目的の論文を探す手助けとなる．

5.　研究の発展経緯確認ツールの実装，評価

5.1　ツールの実装

上記の機能を研究の発展経緯確認ツールとして実装した．実装には [8] においてウェブコミュニティの描画に用いられている TouchGraph [9] を利用した．ここでは，本ツールで利用している TouchGraph に関して簡単な説明を行い，その後，各機能の実装の詳細や利点について述べる．

5.1.1　TouchGraph

TouchGraph [9] は，グラフの情報に基づき，グラフを視覚化するツールである．グラフの情報としては，ノードやそのラベル，形や色等のデータ，エッジやその色や長さ等を XML として指定することができる．描画されるグラフの各ノードとエッジの配置は，TouchGraph が行っている．

本ツールでは TouchGraph の LinkBrowser を使用し，発展経緯を描画している．出力するグラフにおいて，ノードは論文もしくはクラスタを表し，エッジは研究の発展経緯を表している．また，クラスタの粒度，発展経緯の重みの違いはグラデーションにより表している．

5.1.2　論文情報表示機能

論文タイトルや著者名，論文発行年，論文への URL 等，発展経緯の把握に役立つ論文情報を表示する．ノードのラベルに

情報を載せた場合，ポップアップメニューを用いた場合について，それぞれ試みた．

グラフから発展経緯を直感的に理解するためには，そのノードの内容を端的に表現するタイトル等の情報をノードのラベルに記載する必要がある．しかしノードにタイトルのみであっても全文を表示してしまった場合，長いタイトルの論文が含まれているため，個々のノードが大きくなってしまった．そのため，ノードやエッジが少ない場合であってもグラフが煩雑に見えてしまい，発展経緯の理解が難しくなることがわかった．よって，論文の詳細情報は別の部分に表示するべきであるという結論に達した．

また，ラベルに著者名も表示する機能は，利用者がグラフ表示された分野の多くの著者，もしくは見たい分野の論文の著者を知っている場合には有効であるが，そうでない場合には，ラベルに著者名を表示した場合であっても意味はない．当研究室の論文に対し適用した場合には，著者名のラベルへの表示が有効であった．それ以外の場合にも，論文検索等の場合には有効である．しかしながら，発展経緯を見る場合には，ノードラベルに著者名を載せることにより，得られる情報と，それに伴うノードサイズの増大によるグラフの煩雑性の増加とのトレードオフが存在する．

更に，論文の発行年をノードのラベルに表示した場合は，ただちに各ノードが表している論文の新旧を知ることができるため，有効であった．

それに加え，ポップアップメニューから直接，論文を閲覧可能であれば，より詳細な情報を知ることができるため，さらに発展経緯把握が容易となる．

以上から，本ツールは，ノードのラベルには，タイトルから文字数を限定して抽出した文字列と論文の発行年を表示し，論文の詳細情報として，ポップアップメニューに論文タイトルの全文，著者名，発行年，論文ファイルのハイパーリンクを表示するようにしている．

5.1.3 クラスタリング閾値候補表示機能

4.2 において，本機能の必要性を述べた．ここでは本機能の動作を説明する．0 から 1 の間で 0.01 刻みに変化させた各クラスタリング閾値によりクラスタリングを行い，クラスタ内に含まれる論文数の標準偏差の変化を調べ，(平均の変化率) × (定数) より変化が大きい部分を，大幅に変化している部分とみなし，その部分のクラスタリング閾値が大きい方をピックアップし，利用者に表示する．

本機能により，研究のマクロな発展経緯を見る場合，利用者がどのようなクラスタリング閾値でクラスタリングを行うべきかを知ることができるため，クラスタリング結果を容易に知ることができ，発展経緯の把握の手助けになる．

5.1.4 クラスタ粒度把握機能

本ツールではクラスタ内の論文数に応じて，クラスタの色を変化させて表示することにより，クラスタの粒度を一目で知ることを可能にしている．クラスタの粒度が大きい部分は，密に関連しているため，キーワード検索から論文情報を収集した場合には，その部分が目的であることが多い．よって，最初にその部分から見始めることで，目的となる研究分野に関連があるサブグラフを早く発見できる可能性が高いと考えている．また，全ての論文データに対して，リサーチマイニング手法を適用した場合であっても，論文数が多いことから，粒度が大きいクラスタ周辺を探すことにより，自分の興味がある分野を発見しやすくなると考えている．

5.1.5 グラフ選択機能

この機能により，注目したサブグラフ以外の不要なサブグラフのノードや発展経緯の枝を表示させないため，グラフの煩雑性を軽減させることができる．その結果，発展経緯を理解しやすくなる．5.1.4 の機能により，見るべきサブグラフを探した後に，本機能を用いることが効果的である．

5.1.6 グラフ操作機能

本機能により，グラフを多角的に見ることができ，発展経緯の理解を助けている．

5.2 有効性の評価

本ツールを実装し，論文情報に対し適用実験を行った．適用対象とした論文情報は，“text clustering”，“software configuration management”，“design pattern observer” をキーワードとして検索を行い，参照関係から収集した論文集合及び当研究室の論文である．“text clustering” の場合はキーワード検索により得られた論文，及び参照方向を順方向でたどって得られた論文集合であるが，これでは集めた論文が古いものが多くなってしまうため，“software configuration management”，“design pattern observer” の場合は，参照方向の順方向だけではなく，逆方向のリンクもたどって集めた．論文情報の収集には，CiteSeer [10] を利用した．

実装したツールの出力例を図 6 に示す．このグラフは当研究室の論文情報に対してリサーチマイニング手法を適用し，描画したものの一部である．クラスタ内の論文が多いほど，ノードの色が赤色に近く，少なくなるにつれ，黄色に近づくようにしている．クラスタを形成していない論文単体は灰色で表している．また，発展経緯の枝は，重みが大きいものは緑色に近く，小さくなるにつれ黄色に近づくようにしている．

図 6 のポップアップメニューは左下の「分散ディレクトリ探索コストを考慮した並列データアクセス偏り制御」クラスタを表すノード内に含まれている論文を示している．それぞれ，論文のタイトル，発行年，著者，被参照回数を表示している．これらの情報から，それぞれのノードが何を表しているかを容易に知ることができ，発展経緯の把握を助けている．

図 7 は図 6 と同じサブグラフをクラスタリング閾値を上げてクラスタリングを行った結果である．これより，クラスタリング閾値の増減により発展経緯の詳細さが変更していることがわかる．これにより，利用者の希望の粒度で発展経緯を見ることができる．また，図 8 は，“text clustering” をキーワードとして論文情報を取得し，リサーチマイニング手法を適用した結果，出現した全てのサブグラフである．ここでは，ズームアウトにより多数のサブグラフを表示し，発展経緯の枝は見えない状態である．クラスタ内の論文数の増加に伴い，黄色から赤茶色に着色されている．これにより，ノードの色の違いによりクラス

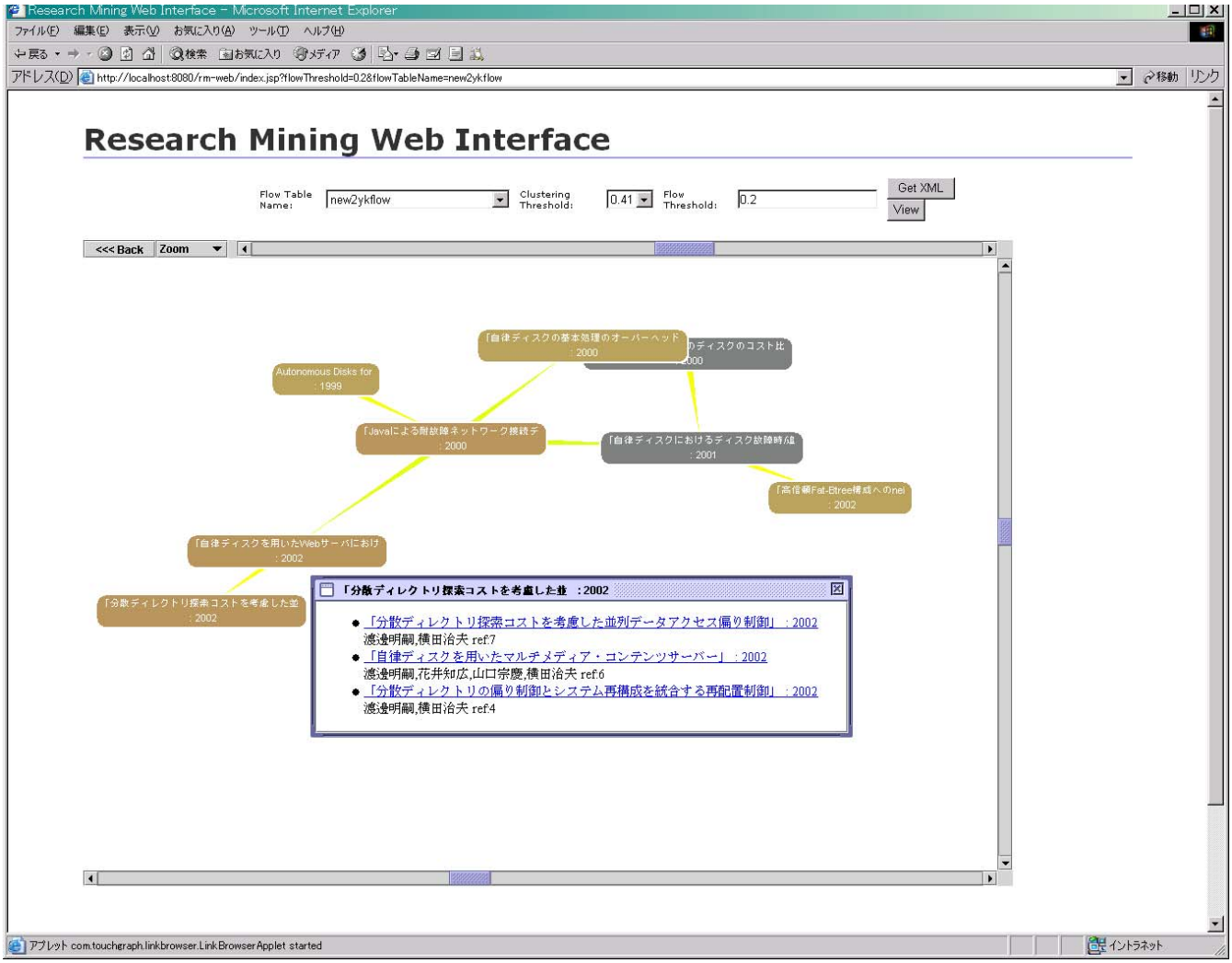

図 6　発展経緯のグラフ

Fig. 6　Graph of research macro-flow

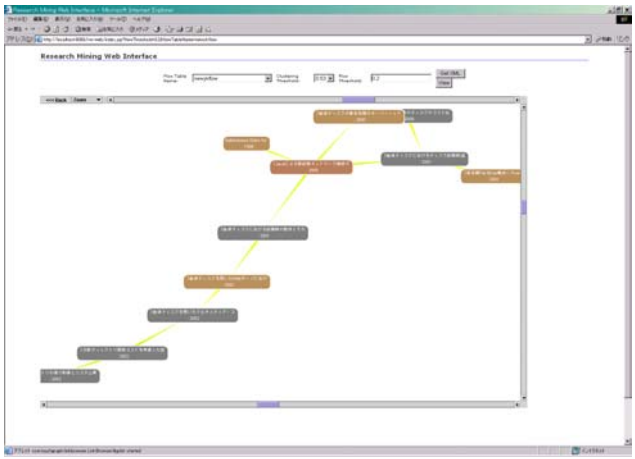

図 7　発展経緯のグラフ

Fig. 7　Graph of research macro-flow

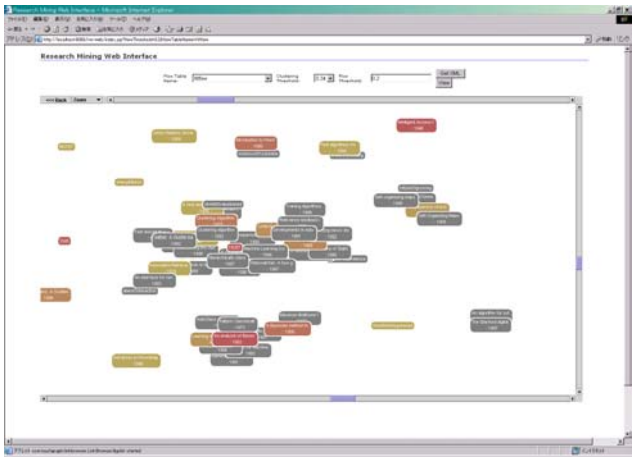

図 8　クラスタ粒度による色の違い

Fig. 8　Color difference of cluster size

タの粒度の差が一目でわかるため，この状態から赤色に近いノードを含むサブグラフを優先的に選択，もしくはズームすることで利用者が望む部分の発展経緯をなるべく早く発見できると考えている．なお，ここでは，論文情報が取得できていないものは，論文 ID で表示している．

本ツールを用いることにより，次のような利点が存在した．

- 論文ノードのラベルにタイトル，発行年を含ませることにより，発展経緯を見やすくなった．

- ポップアップメニューにより，論文ノードやクラスタノードに含まれる論文のタイトルや発行年，論文ファイルを容易に見ることが可能となったため，発展経緯を把握しやすくなった．

また，このツールの問題点は次のようなものが挙げられる．

- クラスタのラベルがクラスタ内の全ての論文の内容を表していない場合がある．

クラスタのラベルは，クラスタ内の被参照回数がもっとも多い論文のタイトルを用いている．しかし，これでは表されている言葉の範囲が狭いため，不適当な場合が存在した．マクロな研究の発展経緯を把握するためには，これを改善することにより良い結果となる可能性がある．

また，論文検索として利用する際には，新しい論文を探すことも多いため，被参照回数がもっとも多い論文ではなく，発行年が最も新しい論文を代表論文とすると，有用となる場合もあると考えている．

6. まとめと今後の課題

リサーチマイニング手法の結果として得られた研究の発展経緯を確認するツールを提案，実装し，本ツールの有効性を確認した．

今後の課題として，リサーチマイニング手法適用結果の評価がある．リサーチマイニング手法により得られる発展経緯でつながれている論文同士は参照，被参照にあり，共に参照される割合が多いため，ほとんどの場合その 2 論文間には関連がある．そのため，発展経緯の評価が難しいが，リサーチマイニング手法自体の評価や用途を考えるためにこれを行う必要がある．

また，論文検索への利用を目的として，キーワードにマッチする論文をハイライトすることにより，注目部分をより早く探すことを可能にし，検索コストを低減するようにシステムを改善することも今後の課題である．

それに加え，リサーチマイニング手法では，発展経緯に出現する論文数が，情報が入力されている論文数より少なくなる．しかしながら，検索への利用に際しては，注目する論文やクラスタを選択した時に，その注目論文が参照している論文や，その注目論文を参照している論文の情報を利用者に提供する機能も有効であり，この機能を持つように本システムを改善することも今後の課題である．その際, [3] で提案されている参照タイプ情報も考慮することにより，より一層利用者の論文検索の手助けとなると考えている．

更に，検索に有用な機能として，この機能以外にも，グラフの注目部分（クラスタ）のみのクラスタリング閾値を上げる機能がある．これにより，その部分のクラスタ粒度が下がり，注目したサブグラフの注目部分のみの発展経緯の詳細情報を知ることができるようになる．

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)，若手研究 (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*, 14(1):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*, 24:265–269, 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**, 42(11):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**, 45(SIG 7(TOD 22)):24–32, 6 2004.

[6] 吉田 誠, 小林 隆志, 横田 治夫. 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. **日本データベース学会** *Letters*, 2(3):73–76, 9 2004.

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

[8] Masashi Toyoda and Masaru Kitsuregawa. Extracting Evolution of Web Communities from a Series of Web Archives. In *Conference Proceedings of Hypertext 2003*, pages 28–37, 2003.

[9] TouchGraph LLC. http://www.touchgraph.com/.

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