

文献の参照関係の可視化

渡部秀文[†] 北川晴香^{†*} 齋藤隆文[†]

本報告では、多数の文献の参照関係を可視化した北川らの手法の改良手法として、ノード位置の自動補正の手法を提案する。北川らの手法は、文献をノード、その参照関係をリンクとしてグラフ形式で可視化する。ノード配置法については、横方向を文献の発表年にあてている。このことにより、先行研究の手法では把握が難しかった、文献発表年の前後関係が一目で把握できるようになった。しかし、縦方向については、同年の文献については単純に下から上に積み上げているだけであるため、ノード数が多くなるとリンクが乱雑になるため画面全体が塗りつぶされてしまい、概要情報が把握できなくなる問題がある。そこで本報告では、グラフィックレイアウト技術の1つであるばねモデルを応用し、ノードの縦方向位置を自動的に補正することで、リンクで接続されたノードを近くに配置する手法を提案する。また、北川らの手法ではノードの縦方向が固定されている上リンクの色が同一である。そのため、多数の文献から参照される文献の把握はできても、どのようなトレンドを持っているか、概要情報からの把握は困難であった。そこで、リンクの色を塗り分けることで、その問題を解決する手法を提案する。

Visualization of reference relations in documents

Hidefumi Watanabe[†] Haruka Kitagawa^{†*} and
Takafumi Saito[†]

We propose methods to adjust the arrangement of nodes automatically for improving the method of Kitagawa et al which visualizes the relationship of references among the vast of papers. In Kitagawa's method, information of papers is visualized by a type of graph assigning papers to nodes, and relationships of references to links. In the theory of arrangement of the nodes, horizontal direction is assigned as published year. By this theory, we can know the order of publishing by a glance. But there is a problem that if the number of papers is vast, links are drawn messy and the display would be painted out because there is no special theory in the arrangement of the nodes by the vertical direction. As the result of this problem, we can't get any information from the overview. Therefore, we propose a method to adjust the vertical arrangement of the nodes automatically by applying "Spring model". On the other hand, in the Kitagawa's method, it is difficult to know the trend of reference from the overview because all the links are not only messy but also drawn by one color. Then, we propose a method to solve this problem drawing links by different colors.

1. 緒言

本研究では多数の文献の参照関係を手軽に描画する可視化手法を提案する。

ある研究分野の文献を調査する際は、一般にはある文献が参照している文献を順に辿っていくことで多くの研究の基礎となる文献を知ることや、その分野で注目度の高い文献を知ることができる。多くの文献から参照される文献は、その分野の研究成果をまとめたサーベイ論文であることや、その分野の重要な文献であることが多い。そのため、文献の参照関係は文献調査における文献探索の重要な指標となりうる。

また、一般に文献から必要な文献を探す際には、Web上で検索エンジンを用いたり、文献の参照関係から調べたりする。しかし、膨大な数の文献の中から必要な文献を探し出すことは困難である。そこで、効率的に検索するための方法として、文献情報によって選択した文献データの中で多くの文献に参照されている文献や、多くの文献を参照している文献などの代表的な文献に注目することが有効であると考えられる。

先行研究では、文献とその関係を2次元平面上にグラフ形式で可視化するものが多くみられる。一般には文献の発表年をノードの位置情報にあてずにノードの色やテキストで可視化することが多い。このような可視化は、文献の参照関係を把握できても発表順を一目で判断できず、調査に十分な情報を与えるものではないと考えられる。また、ノードサイズにも情報を与えることがあるが、このような場合は可視化できる文献数が限られる問題もある。

そこで北川らによって、多数の文献の参照関係を高速かつ手軽に、整然と描画する可視化手法が提案された[1]。可視化は多くの先行研究同様2次元平面上にグラフ形式で行うが、ノードは一定の大きさで格子点上に固定して配置する。横方向は、可視化対象文献の発表年ごとにソートして時系列順に並ぶように固定する。こうすることで発表年ごとに文献情報がまとまり、発表順の前後関係が把握できるようになった。また、参照関係に着目したフィルタリングを用い、ある文献と近い関係にある文献だけを提示する選択的表示も提案された。この手法によって必要な文献の参照関係に着目した情報を対話的に表示し、観察することが可能になった。文献1)の手法を用いることで、ある分野の研究をこれから始める人や、新しく注目する文献を必要とする人など、ユーザの目的に応じた文献の検索が可能となった。しかし、文献1)の手法では、横方向を固定しているため、文献数が多くなるとノードとノードのリンクが煩雑になり、各々のリンクを把握できなくなる問題がある。

その解決のためには、関連性の高いノード同士を近くに配置することで、リンクの

[†] 東京農工大学
Tokyo University of Agriculture and Technology

^{*} 現在、東芝ソリューション
Presently, Toshiba Solutions Corporation

描画総延長を短くすることが考えられる。具体的には、表示の前にクラスタリングなどで関連性の高いノードをグループ化する手法も考えられる。

本研究では、関連性の高いノードを近づけ、そうでないノードを遠ざけることで可視化の概要表示段階での全体像を見やすくすることを目的とする。その手法として、ばねモデルを応用したノードの位置補正手法を提案する。本報告では、ばねモデルを北川らの手法に応用する手法について述べ、現段階での結果を紹介し、問題点や今後の方針について考察する。

2. 文献データ可視化手法

ここでは、文献データの可視化手法の先行研究例について述べる。多くの先行研究では、文献とそれらの関係をグラフ形式で可視化する手法がみられる[2]。多数の文献情報をノードとして1画面内に配置し、関係を持つノード同士をリンクで結ぶ。文献情報やリンクを結ぶ関係性は研究によってさまざまである。

2.1 知識分野可視化技術と Pajek を適用した文献の参照関係の可視化

知識分野可視化技術では、引用分析を通じてデータセットの意味的な空間のマッピングに適用されている。W.Keらは、主要な文献(50程度)の引用分析の結果をPajekというネットワーク描画システムを用いて可視化した(図1)[3]。それぞれの円(ノード)は各文献を、サイズは参照された回数を、色は発表された年をそれぞれ表している。また、円周の色は参照された年の平均を表していて、矢印は文献間の参照関係を示している。図1より、参照回数が多い文献とそれらの関連性を読み取ることができる。また、発表年と引用年の差は円周と円の内部の色の差で表れるため、文献の参照関係の流行の傾向も知ることができる。

しかしこの手法では、参照された文献数の情報をノードサイズにあてているため、多数の文献に適用することができない。そのため、可視化を行う前にある程度重要な文献を選ぶ必要がある。また、文献の発表年をノードの色で表し、参照されている文献の発表年の平均値を円周の色で表すことによって文献の参照関係の傾向を知ることができるが、色が目立つノードが重要な文献であるとは限らない。さらに、全体の概観を調べる点では有用であるが、すでに重要な文献を選んで可視化を行っているために、ユーザが注目している文献についての検索を行うことができない。

2.2 文献の共著者の可視化

文献の共著者の可視化[3]はW.Keらによって提案された手法である。情報可視化の分野においてACMに発表された文献の著者の中から主要な138人を選び可視化を行う。図2において、各ノードは文献の著者を表しており、そのサイズは発表されている文献数に依存している。また、ノードの色についてはその著者の文献が参照された

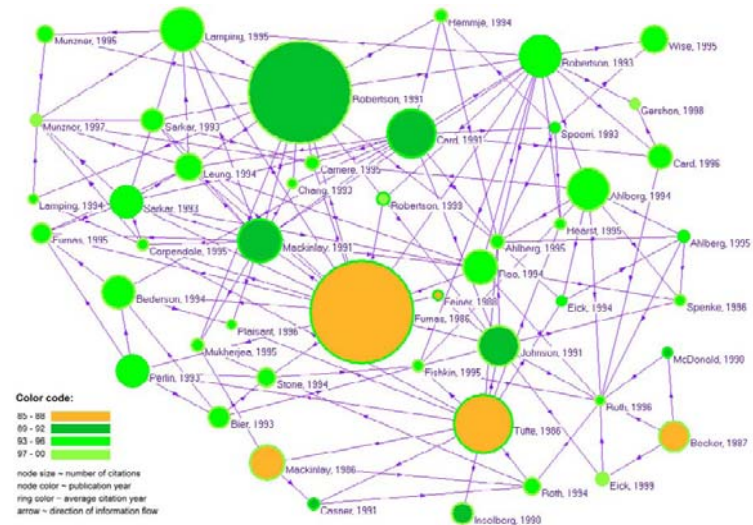


図1 Pajekによる文献引用分析結果可視化例

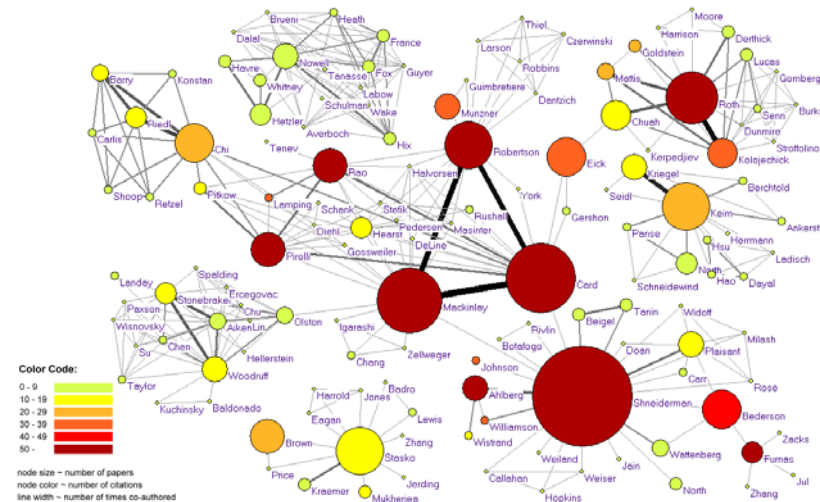


図2 文献共著者の可視化例

回数の合計を示している。著者同士をつなぐリンクは共著の文献が存在することを表して、太さは共著した回数に依存する。この図から重要な著者と大まかな研究グループを読み取ることができる。

しかし、この手法は各分野で著名な人物を知り、また研究団体を調べるためには有用であるが、文献そのものに対する重要度を知ることができない。1つの優れた研究が評価されている著者や、多くの研究が評価されている著者がいると考えられるが、文献ではなく著者について参照されている回数が可視化されているため、注目する著者の文献のなかで重要な文献などの詳細な情報を読み取ることができない。また、ユーザが注目している文献が著名な著者のものでなかったときに、その参照関係を検索することができない。

2.3 北川らの手法による文献の参照関係の可視化

北川らの手法では、各文献について、文献 ID、それが参照している文献の ID、それを参照している文献の ID、文献名、著者名、掲載雑誌名または学会名、発表年、文献の詳細情報の URL、所属団体のドメイン名、概要を使用する。

可視化は図3のように行う。ノードは文献を表し、ノードの色はその文献が参照された回数を擬似カラーで示す。赤に近いほど参照された回数が多く、青に近いほど参照された回数が少ないことを表している。ノードは左から年代順に配置する。データの中には発表年が明記されていないものもあるが、参照されている文献の年代から年代を推定し、おおまかに年代順に並ぶようにする。年情報をもとにして配置してい

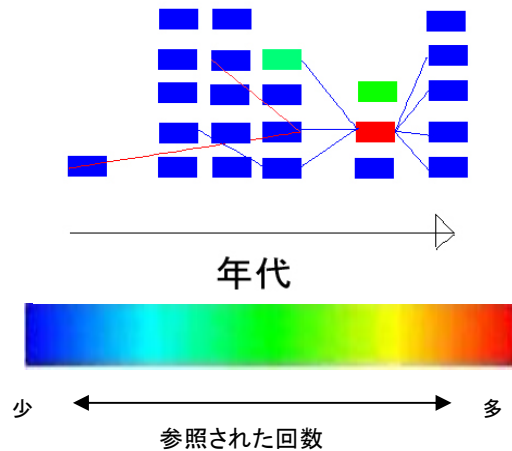


図3:北川らの手法による文献情報の可視化

るため、リンクは基本的に右側にある文献が左側にある文献を参照していることになるが、中には年代の古いものが新しいものを参照していたり、同じ年内で参照してい



図4 北川らの手法例：リンクのフィルタリング表示

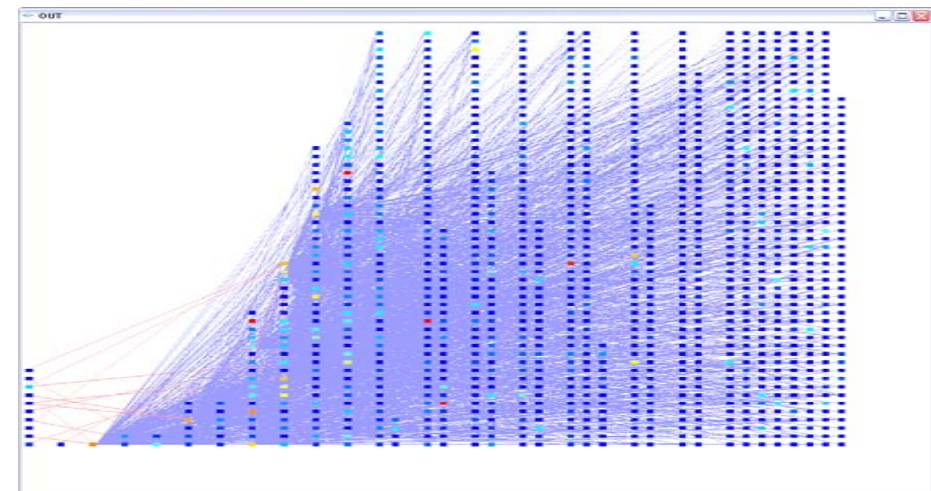


図5 北川らの手法例：概要表示段階の可視化画像

たりするものもある。そのようなリンクは赤色で表示することによって特異性を発見しやすくする。参照している文献のリンクはノードの右側に、参照されている文献のリンクはノードの左側に接続されている。なお、画面の簡略化のために参照関係以外の情報は省略している。

文献の参照関係の詳細を知る際には、注目している文献複数（たとえば図4の囲みのついでに2ノード）を選択することによってそれに繋がっているリンクとその先のノードを表示できる。リンクが繋がっていないノードは、全体の形状を維持しながら必要な情報を強調したいので明度を下げて表示する。この手法によって接続関係と個々の文献の情報を同時に見せ、双方の情報を選択して表示できる。この図から文献Aと文献Bは直接の参照関係はないが、2つの文献は共通して3つの同じ文献からのリンクが繋がっているため、関連性の強い研究であることが読み取れる。

2.4 北川らの手法の問題点

図5は図4の文献情報の、フィルタリング前の概要表示の様子である。北川らの手法では、1000を超える文献が一度に可視化できるが、全体的にノードとリンクの数が増えた分見目が乱雑になってしまっている。

このように、北川らの手法の問題は、可視化する文献情報の数が増えると、描画されるリンクも多く、乱雑になるため、画面全体が塗りつぶされてしまうことにある。このような場合、個々のリンクはもとより、重要な文献のグループも判別できなくなる。前項のように大量の文献情報から必要なものを選択してフィルタリングする手法も提案されているものの、乱雑なリンクで画面全体が塗りつぶされては全体の概要がつかめず、フィルタリングのための最初の文献を選択することは困難である。

この問題の原因として、ノードの配置法のうち横方向は時系列と決められているものの、縦方向が明確に決められていないことがあげられる。北川らの手法では、読み込んだ文献情報を発表年ごとに仕分けし、それぞれID等でソートする。しかし、それ以降はソートした順番に画面下から格子点上に単純に積み上げているだけである。このことがリンクの描画を乱雑にしている要因の1つであると考えられる。文献1)においてもその点は指摘されており、解決策の1つとして文献の関連を用いる手法が示唆されている。また、従来法ではリンクの色はある文献が数年後の文献を参照するような不自然な参照のときを除けばすべて同色である。このことも概要の把握を阻害する一因となっていると思われ、可視化手法に改良の余地があると考えられる。

そこで、次節以降でばねモデルを用いて関連する文献情報をまとめて配置する手法と、可視化手法の改良について提案する。

3. ノードの自動配置法と可視化手法の改良

ばねモデルは、グラフィックアウト技術の1つで、Eadesが提案した比較的単純なアルゴリズムである[4]。最近Webコンテンツ等を可視化した例で見られるばねモデルでは、ノードに次のような条件を与えて各ノードに働く力を物理シミュレーションする。ノードに働く力からノードの速度を計算し、ノードを再配置する。これを繰り返す、すべてのノードに働く力が収束したところでシミュレーションを終了する。

- ・各ノードは質点とする
- ・リンクはある長さを持つばねとし、リンクに接続されたノードは、ノード間の距離によってフックの法則に基づいた弾性力を受ける
- ・任意のノード同士には斥力がはたらき、ノード間同士が近づきすぎることを防ぐ
- ・シミュレーション収束のために、各ノードには摩擦力をはたらかせる

ばねモデルは、このような比較的単純な物理シミュレーションにより接続関係にあるノードを近づけ、接続関係のないノードを遠ざける。本研究では、このばねモデルを北川らの手法に合わせて応用する。

3.1 北川らの手法へのばねモデルの応用

北川らの手法では、文献の発表年を横軸に設定しているため、一般のばねモデルでは起こりうるノードの横方向の移動は好ましくない。そのため、本報告においては、ばねモデルに次のような制約を設けて収束の様子について考察することとする。

(1) ノードの移動方向

ノードの移動方向は縦方向だけとし、実際にノードに働く力は縦方向だけを考える。

(2) ノードの入れ替え

上記と、ノードが格子点上に配置されることから、隣接する2ノードが次のような場合、ノード間に働く斥力を無視してノード位置を入れ替える必要がある。

- ・上のノードに下向きの、下のノードに上向きの力が働く場合
- ・2ノードの移動の方向は同じだが、後ろのノードの力が強い場合

(3) 弾性力の重みづけ

重要な文献ほど他の多くの文献に参照されると考えられるため、多くの文献に参照されるノードがそれを参照するノードに引かれる力は強くするほうが、より関連する文献が近づきやすくなると思われる。そこで、各ノードに働く弾性力は、対向ノードが持つリンクの数に比例して強くなるよう重みづけする。

(4) 発表年が離れた文献情報にかかる弾性力

ノードの横方向の移動を制限したために、すべてのリンクに同一の長さのばねを用いると、発表年が離れた文献情報ほど弾性力が強くなり不自然となる。また、物理シミュレーションの収束も遅くなる。そこで、ばねの自然長は2ノード間の発表年の間隔に比例させる。

3.2 可視化手法の改良

ばねモデルの応用により、年をまたいで横方向にグループが存在する場合には概要が見やすくなる。しかし、各ノードを厳密な格子点に配置するわけではないので、従来手法では見やすかった発表年の判別が難しくなっている。そこで、可視化画像の背景色を、発表年ごとに交互に塗り分けることで判別を容易にした。また、リンクの色については、ノードの右側から描画されるリンクについて発表年ごとに交互に塗り分けることとした。こうすることで、ある文献が後年の多数の文献から参照されている

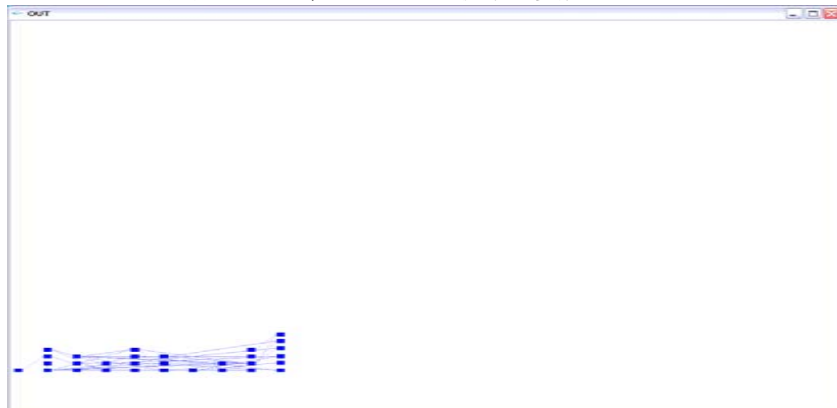


図6 従来手法による可視化結果・人工データ

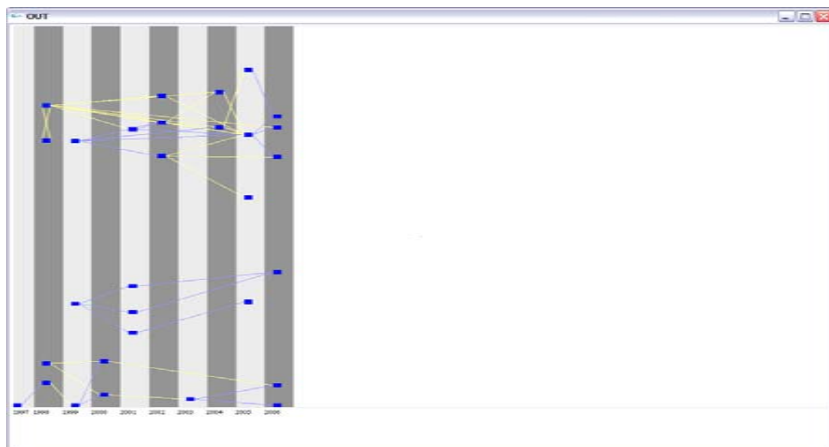


図7 提案手法による可視化結果・人工データ

場合に、参照している年が異なってもリンクが同じ色で1点に集中するように描画される。このため、ある文献が参照される時代のトレンドの把握もしやすくなると考えられる。可視化結果について次節で示す。

4. 可視化結果の評価

前節で述べたばねモデルの応用手法により、文献情報の可視化と評価を行う。まず手始めに、少数の人工データを用いた評価を行う。このデータは、人為的に3つのグループになるように参照関係を作成した30ノードからなる文献情報である。これを可視化して、ノードが正しく分割表示され、3つのグループが把握できるか評価する。

次に、多数の文献情報の例として、CG関連の文献データで同様の実験を行う。このデータは、420のノードからなる情報である。このデータの可視化結果を用いて、従来手法からどの程度見やすくなったかについて考察する。

4.1 人工データでの評価

まず、人工データでの可視化結果を示す。従来の北川の手法による可視化結果を図6に、提案手法の可視化結果を図7に示す。両者を比較すると、従来手法では3つのグループになっていることが把握できない。一方、提案手法では明らかに文献情報が3つのグループに分かれている様子が把握できる。

4.2 実際の文献データでの評価

次に、実際の文献データでの可視化結果を示す。従来手法による可視化結果を図8に、提案手法の可視化結果を図9に示す。

こちらについても、従来手法では文献情報グループが把握できない。一方提案手法では、完全に大規模な集団から離れてはいないものの、独立したノードのかたまりが把握できる(図9の囲み)。このように、概要のレベルで大量の文献情報の中から小規模な独立した集団も把握できることが分かる。また、ノードの持つリンク数によりばねの弾性力に重みづけし、より重要なノードを参照するノードが近づくよう意図したが、必ずしもそうになっていない。このことは、ノード数が多くなることで、多数のノードを対象としてからの斥力計算が複雑となることが原因であると考えられる。また、ばねモデルではあるノードに働く斥力を、他のすべてのノードとの距離から計算するため、ノード数を N とすると、計算量は $O(N^2)$ となる。一般にWebを用いた文献検索結果は数千~数万になることも珍しくない。また、単純なばねモデルでは、複数のリンクが交差してしまう問題がある[5]。提案手法においてもリンクがいたるところで交差し、乱雑さの残るノード配置となっている(図9)。このことから、斥力計算法をはじめとしたノード位置補正法については改良の必要がある。

一方、リンクの色分けについては、3.2項で述べたとおり、ある文献が、他の多数の

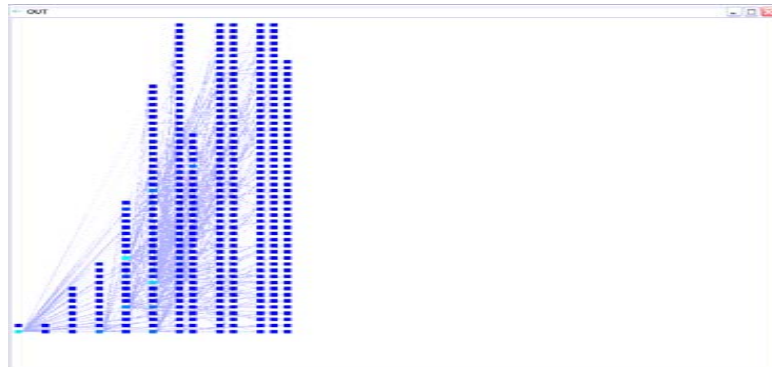


図8 従来手法による可視化結果・検索結果データ

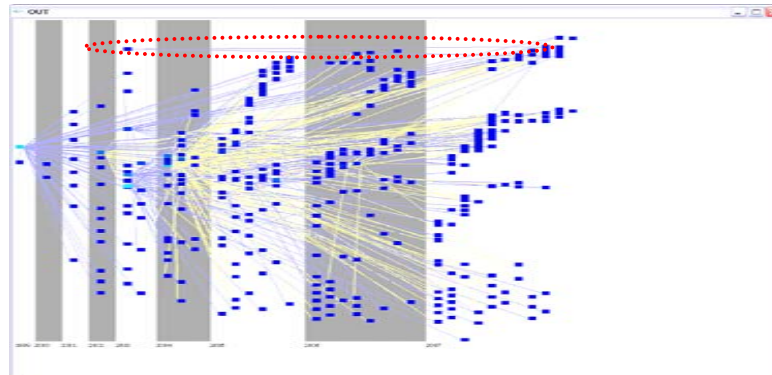


図9 提案手法による可視化結果・検索結果データ

文献から参照される様子を把握するためにはある程度効果があると考えられる。また、従来は把握できなかった文献参照のトレンドについては、ノードの縦方向の位置補正と併用したことによりある程度把握できるようになった。たとえば、1999年発表の、2つの文献情報(画面左端)のうち、多数のリンクが伸びているノードに着目すると、後年のノードにまんべんなく伸びている様子が把握できる。このことから、この文献は多数の文献から長期にわたって参照されている文献であることが分かり、重要なものであると考えられる。また、ノード終点は、注目ノードから伸びるリンクの色が他と異なることや、従来法と比べてリンクが他のノードやリンクに重なる頻度が軽減されたことにより把握しやすくなっている。終点が把握しやすくなったことにより、注目している文献が参照される年代のトレンドも把握しやすくなる。ところが、単純に

2色を交互に塗り分けるだけではノード数がさらに増えたときにやはり従来と同様の問題が起きるものと考えられ、リンクの色分け法についても改良の必要がある。

5. 結言

本報告では、文献の参照情報を可視化した北川らの手法の概要表示段階の改良を目的とし、ノードを自動配置する手法を提案した。ばねモデルを応用した提案手法により、従来把握できなかった文献のグループをある程度把握できるようになった。また、ノード位置の補正と併せてリンクの色を塗り分けることで、従来は把握が難しかった文献が参照されるトレンドがある程度把握できるようになった。

しかし、現段階の計算モデルでは計算量がノード数の2乗のオーダーであるため、計算量を軽減するモデルに改良する必要がある。また、ばねモデルが持っている、リンク交差の問題についても改良の必要がある。改良法については、文献5)で伊藤らが提案した手法などがヒントになると考えられる。また、本報告の手法は複雑な前処理を用いない自動補正を目指し、重みづけも単純なものであった。しかし、クラスタリング等を併用すれば、より適切に重みづけできると考えられる。

リンクの色については、より多く参照されるノードにリンクが集中することと、着目ノードの参照元ノードがある程度把握できるようになったと考えられる。しかし、リンク数が多くなると表示が乱雑になり概要情報が把握できなくなる問題は本質的に残っている。今後はこの点についても改良を行っていく。

参考文献

- 1) 北川 晴香, 宮村 (中村) 浩子, 古谷 雅理, 斎藤 隆文, “文献における参照関係の可視化,” 情報処理学会第70回全国大会 講演論文集 2ZE-5, 2008.
- 2) K. Börner, C.Chen, K.Boyack, “Visualizing Knowledge Domains,” Annual Review of Information Science & Technology, Vol.37, pp.179-255, 2003.
- 3) W. Ke, K. Borner, L. Viswanath, “Major Information Visualization Authors, Papers and Topics in the ACM Library,” IEEE Symposium on Information Visualization (INFOVIS'04), pp. r1, 2004.
- 4) P. Eades, “A Heuristic for Graph Drawing,” Congressus Numerantium, vol. 42, pp. 149-160, 1984.
- 5)伊藤 貴之, 井上 恵介, 土井 淳, 梶永 泰正, 池端 裕子, “力学モデルを用いたグラフデータの画面配置手法の改良,” 情報処理学会研究報告, 2001-CG-103, Vol.2001, No.35, pp.7-12, 2001.