情　報　学　基　礎　70－10
デジタル・ドキュメント　38－10
（２００３．３．２８）

技術文書引用情報の可視化

小川 知也　渡部 勇
{ogawa.tomoya, watanabe.isamu}@jp.fujitsu.com
富士通研究所
〒 211-8588 川崎市中原区上小田中 4-1-1

本稿では, 技術文書とその引用情報のフロー (有向グラフ) による可視化において, 文書セットに特徴的な技術の流れを明確化する手法について論ずる. 本手法では, ノード間の連結性を保持したまま情報の損失が少なくなるようにエッジ絞り込みおよびエッジ強調を行うことで, 文書セットに特徴的な技術の流れを適切かつ簡潔に表現する. また, 多数の文書をフローにより可視化するための, 文書間の関係を考慮したエッジ重みによるノード絞り込み方法を提案する. クラスタリング関連の技術文書を対象とする実験を通じ, 本手法の有効性を確認した.

Visualization of Citation Information of Technical Documents

Tomoya OGAWA　Isamu WATANABE

Fujitsu Laboratories Ltd.
4-1-1, Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa, 211-8588 Japan

In this paper, we propose a novel visualization method using directed graphs, for the citation information contained in technical documents, where our method is designed to effectively visualize the flow of characteristic techniques present in the given set of documents. For visualizing the flow of characteristic techniques in a simplified, clarified manner, we have carried out filtering and emphasizing processes to certain edges, so as to reduce information loss while maintaining the connectivity information. For visualizing a large number of documents with directed graph, we propose a node filtering method with edge weighting. Through a set of experiments using technical documents belonging to the technical field of clutering, our method has been proven effective.

1　はじめに

論文, 特許, 技術報告書など, さまざまな電子化されたネット上の技術文書が増加している. それら飛躍的に増大する文書を有効活用するには, 何らかの方法で文書を整理する必要がある. 本稿では文書間の関係を示す文書の引用情報に着目し, 文書セットに特徴的な技術の流れを明確にする手法を提案する.

本手法は, 文書の引用情報に基づき文書セットをフローとして可視化し, ノード間の連結性を保持したまま情報の損失が少なくなるようにフローの簡潔化を行うことで, フローの特徴的な部分を適切かつ簡潔に表現する. 本手法により, 複数文書間の全体構造を容易に把握したり, 調査対象分野のサーベイを効率的に行うことが可能となる.

2　関連研究

文書の引用情報に基づく分析としては, 2 件の論文が同じ文献を引用する度合いを示す書誌結合 (bibliographic coupling), あるいは 2 件の論文が同じ文献から引用される度合いを示す共引用分析 (co-citation analysis) という共起関係に基づく分析が行われる. [1] [2]

共引用分析に基づき著者やキーワードをグラフの形に可視化する手法としては, Small らの研究 [3], White らの研究 [4], Chen らの Pathfinder[5] などがある. これらの手法では, 著者やキーワードなどの静的な類似関係は分かるものの, 技術の流れのような時間的に推移する情報を読み取ることは難しい.

文書間の引用情報を明示的に可視化する手法としては, Hyperbolic Tree[6] が有名である. しかしこの手法は, 本研究のように文書セット全体の特徴を反映する簡潔化は行っていない.

フローを簡潔に表示する手法としては, フローの主な流れとして最長パスなどを木探索により求める Hummon らの研究 [7] がある. しかしこの手法は本研究の手法とは異なりグラフのノード間の連結性が必ずしも保持されないため, 本来ある流れが適切に可視化されない.

3　引用フロー

本節では, 本研究の可視化手法のベースとなる引用フローについて説明する.

3.1　引用フロー

引用フローとは, 文書をノード, 文書間の引用関係をエッジとする有向グラフである.

引用フローでは, 文書 A を文書 B が引用する場合に A から B へエッジを張る (図 1). エッジの向きが引用の向きとは逆になるが, エッジの向きを左から右へと向かう時間の流れに合わせるためにこの向きとする.

図 1: 引用フロー

引用フローにおいては関連の高い文書同士が近くに配置されるよう, エッジが最小交差するようにノードおよびエッジのレイアウティングを行う [8]. これにより, エッジの交差が減り引用フローの可読性も向上する.

被引用数が多い文書はそれだけ注目度の高い文書ということであり, 文書の重要さやその分野を代表する文書かどうかを表すひとつの目安と考えられる. 引用フローでは各文書の被引用数の多さを, 被引用数が多くなるほどノード枠幅が太くなるようにノード枠幅で表す. 被引用数を視覚的に表現することで, 注目すべき文書を容易に知ることができる.

3.2　ビュー

引用フローの見せ方としては, 次のビューがある.

論理ビュー
引用関係を簡潔に表現する (図 2). 引用関係は正確に表現されるが, 時間順序は必ずしも正しく表現されない.

時系列ビュー
引用関係に加え, 時間順序も正しく表現する (図 3). 引用フローが時間軸方向に長くなることがある.

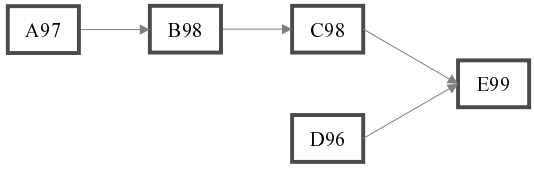

図 2: 論理ビュー

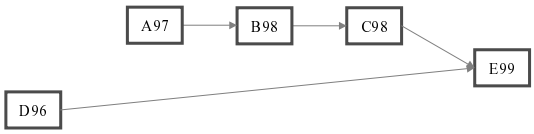

図 3: 時系列ビュー

時系列ビューの方が時間順序が保存されるため情報を読み取る際に無用な混乱を防ぐことが出来る．本稿では基本的に時系列ビューを用い，ノード数が多い場合に論理ビューを用いる．

4 技術文書引用情報の可視化手法

本節では，本研究の提案手法である技術文書引用情報の可視化手法について述べる．

4.1 エッジ絞り込み

フローの形で可視化を行う場合，エッジの数が多いためにフローが煩雑になり，情報を読み取ることが難しくなることがある．

本手法では，図 4 のように 3 件の文書 A，B，C が三角形の関係にある時に，近道となるエッジ AB を削除するようなエッジ絞り込みを行う．エッジ絞り込みは三角形の近道を削除するだけなので，グラフのノード間の連結性を保持したままフローを簡潔にすることが出来る．

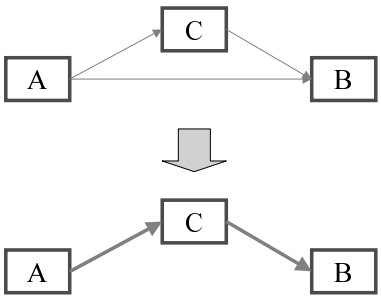

図 4: エッジ絞り込み

エッジ絞り込みは図 5 のように，三角形から n 角形へと拡張できる．

エッジ絞り込みによる情報の損失を抑えるため，代替パス AC … DB の各エッジの重みを増やすことでエッジ強調を行う．これは，B から A への引用があるというこのフローの特徴を反映させる，という考えに基づく．

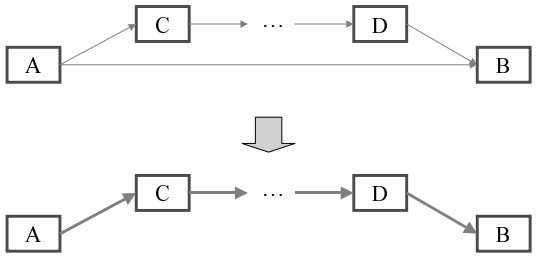

図 5: n 角形のエッジ絞り込み

代替パスのどの範囲のエッジ強調を行うかについては AC … DB 以外に，共引用を反映するように代替パスを強調する (AC … D)，書誌結合を反映するように重みを増やす (C … DB)，などの方法も考えられる．エッジ強調範囲による違いを調べるために行った予備実験より，共引用強調は引用された文書が数多く現れる，書誌結合強調は引用された文書が現れ難い，代替パス全体強調は比較的バランスのとれたフローが得られる，という結果を得た．本稿では以降，エッジ強調の方法として代替パス全体強調を用いる．

代替パスが 2 つ以上存在する場合，すべての代替パスについてエッジ強調を行う．ひとつの文書がフロー全体に与える影響が大きくなりすぎないよう，代替パスの各エッジに加える重みは代替パス数で割った値とする．

ひとつの文書 (図 4，図 5 における B) について複数のエッジ絞り込みが起こることがある． この場合も，ひとつの文書がフロー全体に与える影響が大きくなりすぎないよう代替パスに加える重みは参考文献数で割った値とする．

以上より，エッジ pq 絞り込みにおいて代替パスの各エッジに加える重み w は，q の参考文献数を b_q，代替パス数を a_{pq} とすると次のとおりである。

$$w = \frac{1}{b_q \times a_{pq}}$$

フローとして可視化する際，エッジ重みに応じた線幅で表示することによりエッジ強調を行う．

エッジ絞り込みを三角形の場合のみ行うと短い流れが強調され，n 角形の場合にも行うと長い流れが強調される． 本研究は文書間の技術の流れを明確にすることが目的であるため，以降では n 角形の場合にもエッジ絞り込みを行う．

なおエッジ絞り込みの際，エッジ絞り込みを行わずエッジ強調のみを行う方法も考えられる． その場合，

引用情報を失うことなくエッジを強調することが出来る. 一般にはエッジ絞り込みも行う方がグラフを簡潔化する効果が大きいので, 本稿では以降エッジ絞り込みも行う.

文書同士が相互に引用し合うことによるループなどの強連結成分があると, エッジ絞り込みの際ノード間の連結性が保存されなくなる可能性がある. そこでエッジ絞り込みに先立ち, フローの各強連結成分はそれぞれひとつのノードにまとめておく.

4.2 ノード絞り込み

多数の文書を対象にフローによる可視化を行う場合, ノードの数が多いためにフローが煩雑になることがある.

可視化手法にもよるが, 通常一つの図に表現して意味のあるノード数は数十程度である. それ以上の数のノードを可視化しても, そこから何かを読みとるのはかえって難しくなることが多い.

そのような場合, 文書セットに特徴的な情報に絞り可視化を行うことが必要となる. そのための方法としては, 次のようなものが考えられる.

- 文書セットに特徴的なキーワードを可視化する
- 文書セットに特徴的なクラスタを可視化する
- 文書セットに特徴的な文書を可視化する

特徴的なキーワードを可視化する方法は, キーワードで表される対象範囲は一般に時間的な幅があるため, 時間情報を正確かつ簡潔に表現することは容易ではない. 特徴的なクラスタを可視化する方法は, 表現の粒度が粗くなるという欠点がある. そこで本研究では, 特徴的な文書を可視化する方法を用いる.

特徴的な文書 (ノード) への絞り込み方法として, 本手法では文書間の関係を考慮したエッジ重みによるノード絞り込み方法を提案する. 一般的な被引用数に基づくノード絞り込み方法に対し, 次のような違いがある.

被引用数に基づくノード絞り込み
: 文書の重要度の目安である被引用数の多い順にノード絞り込みを行う. 文書重視の絞り込み方法である.

エッジ重みによるノード絞り込み
: 技術の流れの重要度の目安であるエッジ重みの大きい順にエッジを選択し, そのエッジの始点および終点であるノードへと絞り込みを行う. 流れ重視の絞り込み方法である.

実験では, 両者の比較を行う.

5 実験

本手法の有効性を検証するために, 実際の技術文書データについて行った実験および考察について述べる.

5.1 実験データ

実験データとしては, クラスタリング手法のサーベイ論文 [9], [10] の参考文献として挙げられている文献を用いた. それら文献 76 件から, 本などを除いた 52 件について引用情報の抽出を行った. 引用された参考文献を含む全文献数は 779 件である.

本実験での引用フローのノードには次の書誌情報をノードラベルに表示する.

1 行目 第一著者＆掲載年
2 行目 キーワード (タイトルから冠詞や前置詞などの不要語を除いた先頭語)

引用情報として抽出された文献 (703件 = 779−76) は, 1 行目の掲載年の後ろに「*」を付けて表す.

5.2 ある研究以後の技術の流れを知る

引用フローの利用目的のひとつとして, 注目する研究のその後の研究や技術の流れを知ることが挙げられる. ここでは, 最近の代表的なクラスタリング手法のひとつである DBSCAN について, それ以後の技術の流れを知ることを想定した引用フロー作成実験を行う.

DBSCAN (ノードラベル ester96densitybased. 以下同) および DBSCAN を引用している文献について作成したフロー (時系列ビュー) を図 7 に示す. フローにおける被引用数の多い文献に注目することで, DBSCAN 以後の主な文献を読み取ることは出来る. しかし, エッジ数が多いためフローが煩雑であり, 各文献間の関係は読み取り難い.

図 7 についてエッジ絞り込みを行ったフロー (時系列ビュー) を図 8 に示す. このフローのノード数, エッジ数, 平均次数は表 1 の通りである.

エッジ絞り込みを行うことで主な文献がどれかということだけではなく, それら文献間の関係まで読み取ることが容易となる.

表 1: エッジ絞り込み前後の平均次数

フロー	ノード数	エッジ数	平均次数
絞り込み前	19	48	5.1
絞り込み後	19	31	3.3

5.3 文書セット全体の流れを知る

引用フローの利用目的のひとつとして, 収集した手元の文書セット全体の流れや各文書の位置付けを知ることが挙げられる. ここでは, 今回対象とする文献全体の流れを可視化する引用フロー作成実験を行う.

エッジ絞り込みを行わずに可視化すると, 元の文献 76 件の引用情報を表現するだけでも図 6 (論理ビュー) のようになり, そこから何かを読みとるのは困難となる. このフローのノード数, エッジ数, 平均次数はそれぞれ 76, 163, 4.3 である.

文献全体 779 件について, 被引用数に基づくノード絞り込みにより 40 ノードに絞り込み, エッジ絞り込みを行ったフロー (論理ビュー) を図 9 に示す. ノード数, エッジ数, 平均次数は 40, 37, 1.9 である. フローのノードに関する被引用数の分布を表 2 に示す. この被引用数は文献全体 779 件におけるものであり, フローに出現しない文書からの引用も含む. 被引用数の平均は 6.4 件である. 被引用数に基づくノード絞り込みによるフローはノード選択において文献間の関係を考慮していないため, 他のノードと接続しないノードが 8 ノード存在するなど文献間の関係を見るには必ずしも適しているとは言えない. また被引用数の大きな文書は書かれてから時間の経過した文書が多いため, このフローの最も新しい文書でも 99 年のものとなり, それ以降の技術の流れを知ることは出来ない.

次に同じ 779 件について, エッジ重みによるノード絞り込みにより 40 ノードに絞り込んだフロー (論理ビュー) を図 10 に示す. このフローのノード数, エッジ数, 平均次数は 40, 46, 2.3 である. フローに出現するノード (文書) に関する被引用数の分布を表 3 に示す. 被引用数の平均は 4.7 件である. エッジ重みによるノード絞り込みによるフローではノード選択において文献間の関係を考慮しているため, 他のノードと接続しないノードも存在しなくなるなど文献間の関係を見るのに適していると言える. なお, 被引用数に基づくノード絞り込みの平均被引用数 6.4 件以上の

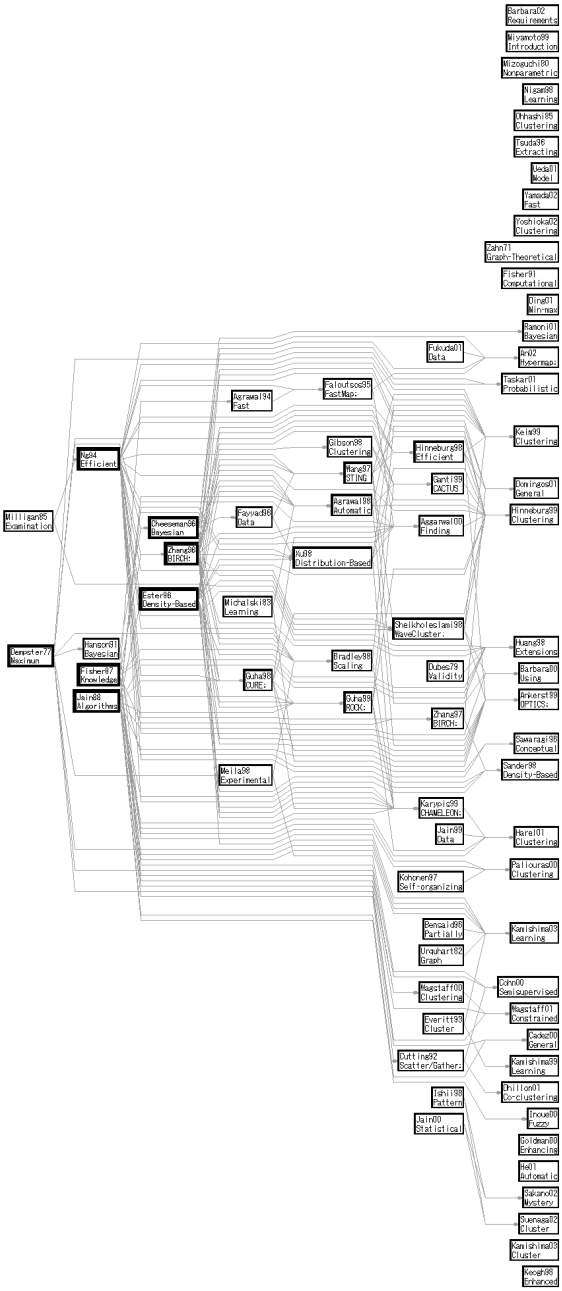

図 6: 元の文献 76 件の引用フロー (論理ビュー)

文書に関しては, 被引用数 11 の fisher87knowledge, 被引用数 7 の macqueen67some の 2 件が出現しなくなる.

このフローからクラスタリング手法のひとつの流れとして, CLARANS (ng94efficient) から BIRCH (zhang96birch), STING (wang97sting), DENCLUE(Hinneburg98), WaveCluster (sheikholeslami98wavecluster) へという流れが読み取れる. これら

の文献を実際に見てみると, 大規模データに対処する手法に関するものであることが分かる. これらは, 元のサーベイ論文でも同じ章「8 章 大規模データへの挑戦」で紹介されている.

もうひとつの流れとして, DBSCAN (ester96 densitybased) から CURE (guha98cure), ROCK (guha99rock), CHAMELEON(Karypis99chameleon) へという流れがある. これらは ROCK を除き, クラスタ毎の形状の異なりに対処する手法である. 元のサーベイ論文でも同じ章「10 章 任意形状のクラスタを抽出する手法」で紹介されている.

元のサーベイ論文の他の章をフローの流れと比較した場合も, 一致している点が概ね見られる.

カテゴリ属性への対応に主眼を置く ROCK が混入する原因としては, 次のことが挙げられる.

- 同種手法の文献間の引用関係があまり多くないこと
- CURE と ROCK は同じ著者による文献であり, 参考文献に重なりが多いこと

対処法としては, 文献の内容や引用の種別に関する情報を用いてエッジを張るといったことが必要と考えられる.

6 まとめ

技術文書の引用情報を用いて, 文書セットに特徴的な技術の流れを明確にし各文書の位置付けを明らかにする手法を提案した. 本手法は, ノード間の連結性を保持したまま情報の損失を少なく抑えるエッジ絞り込みによりフローの簡潔化を行うことで, フローの特徴的な部分を適切かつ簡潔に表現する.

クラスタリング関連の技術文献を対象とする実験を通じ, エッジ絞り込みの有効性を確認した. また, 多数の文書を対象にフローによる可視化を行うためのノード絞り込み方法に関して, エッジ重みによるノード絞り込みが技術の流れを見るには適した方法であることを明らかにした.

今後の課題としては, 大規模データベースでの本手法の有効性の検証が挙げられる.

謝辞

引用情報の可視化にあたり, 富士通研究所の三末主任研究員の開発されたプログラムを利用させていただきました. ここに感謝いたします.

参考文献

[1] Garfield, E.: From Bibliographic Coupling to Co-Citation Analysis via Algorithmic, *A citationist's tribute to Belver C. Griffith* (2001).

[2] 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 (2001).

[3] Small, H.: Visualizing Science by Citation Mapping, *J. Am. Soc. Information Science*, Vol. 50, No. 9, pp. 799–813 (1999).

[4] White, H. D. and McCain, K. W.: Visualizing a Discipline: An Author Co-citation Analysis of Information Science 1972-1995, *J. Am. Soc. Information Science*, Vol. 49, No. 4, pp. 327–356 (1998).

[5] Chen, C. and Paul, R. J.: Visualizing a Knowledge Domain's Intellectual Structure, *IEEE Computer*, Vol. 34, No. 2, pp. 65–71 (2001).

[6] Lamping, J., Rao, R. and Pirolli, P.: A Focus + Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies, *ACM Conference on Human Factors in Computing Systems (CHI '95)* (1995).

[7] Hummon, N. P. and Doreian, P.: Connectivity in a Citation Network, *Social Networks*, Vol. 11, pp. 39–63 (1989).

[8] Sugiyama, K. and Misue, K.: Visualization of structural information: Automatic drawing of compound digraphs, *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 24, No. 4, pp. 876–892 (1991).

[9] 神嶌敏弘：データマイニング分野のクラスタリング手法 (1) －クラスタリングを使ってみよう！ －, 人工知能学会誌, Vol. 18, No. 1, pp. 59–65 (2003).

[10] 神嶌敏弘：データマイニング分野のクラスタリング手法 (2) －大規模データへの挑戦と次元の呪いの克服 －, 人工知能学会誌, Vol. 18, No. 2, pp. 170–176 (2003).

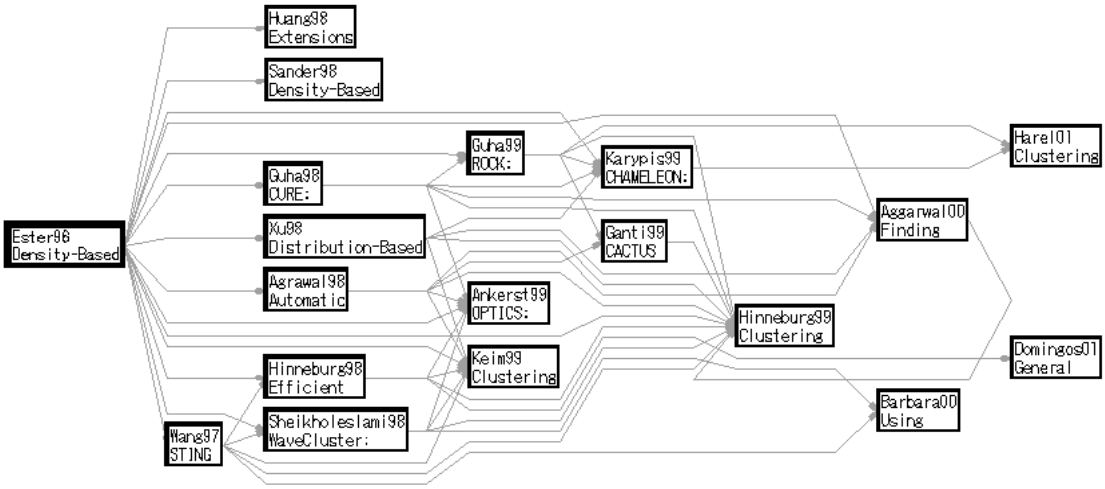

図 7: DBSCAN 以後の引用フロー (時系列ビュー)

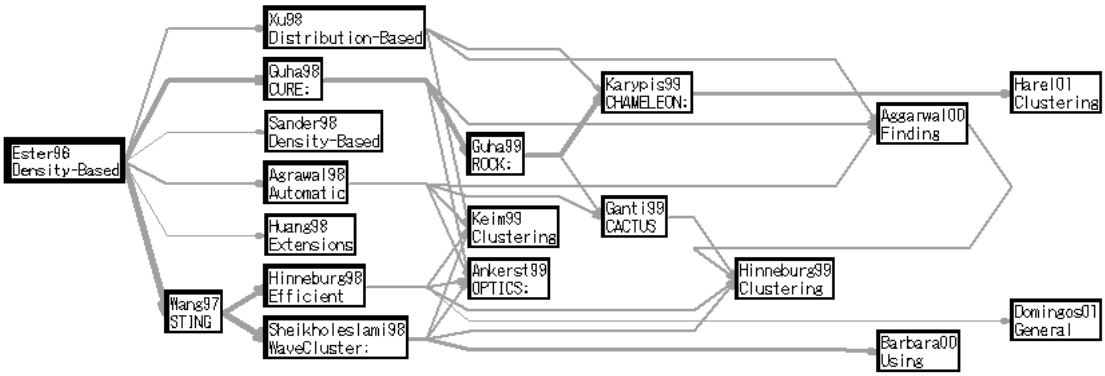

図 8: DBSCAN 以後の引用フロー (エッジ絞り込み&時系列ビュー)

表 2: 被引用数に基づくノード絞り込みの被引用数

被引用数	件数
21	1
16	2
14	1
13	1
11	1
10	2
9	2
8	2
7	1
5	5
4	11
3	11

表 3: エッジ重みによるノード絞り込みの被引用数

被引用数	件数
21	1
16	2
14	1
13	1
10	2
9	2
8	2
5	3
4	5
3	2
2	5
1	4
0	10

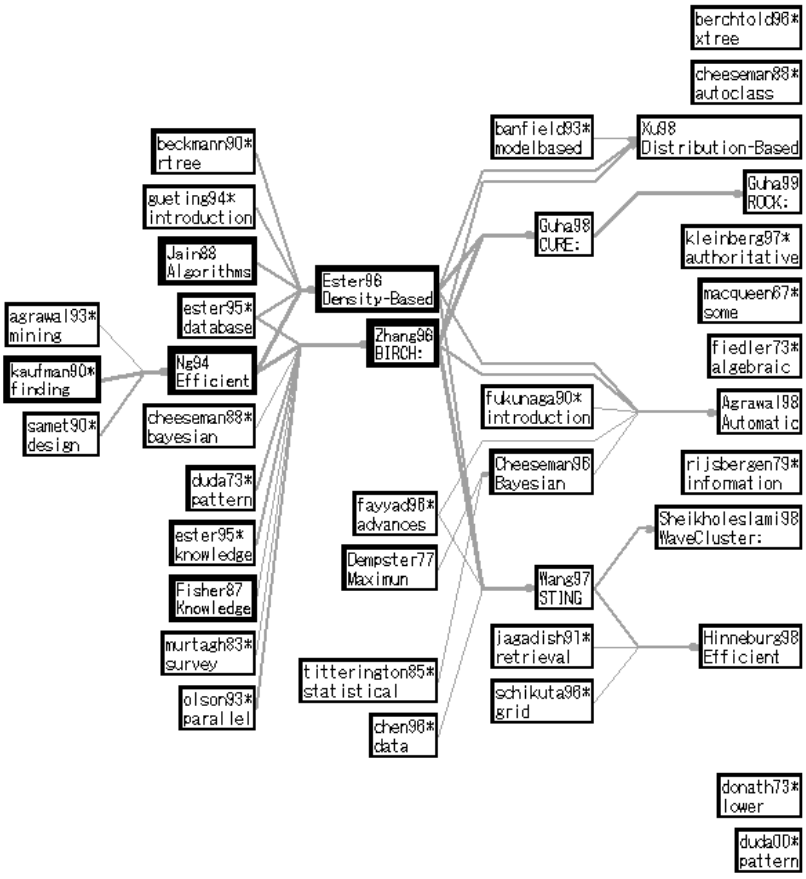

図 9: 文書セット全体の引用フロー (被引用数に基づくノード絞り込み＆論理ビュー)

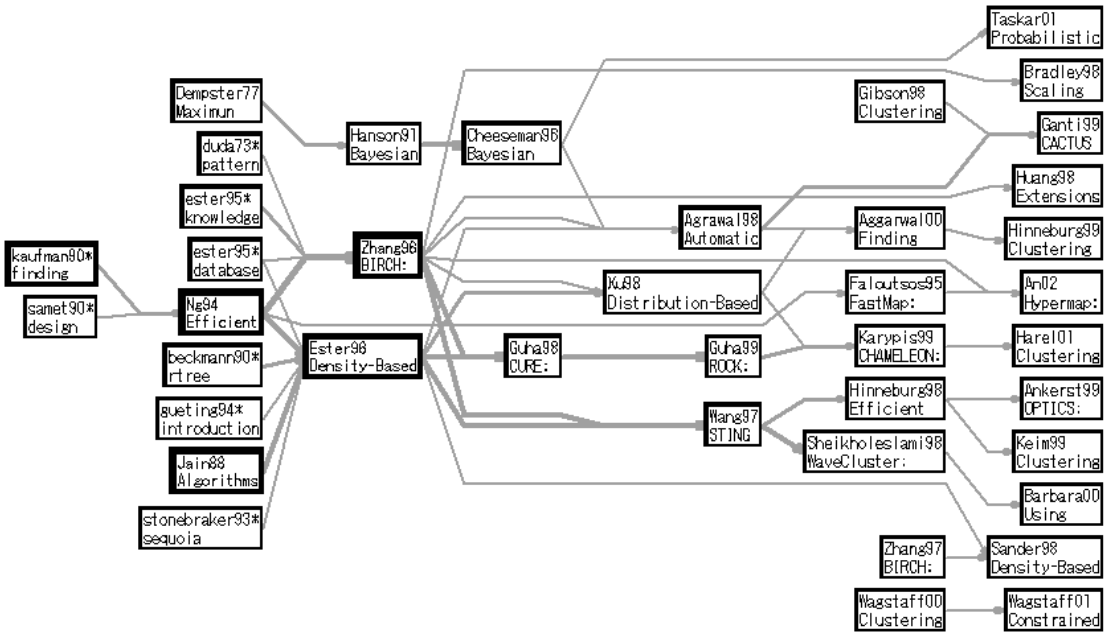

図 10: 文書セット全体の引用フロー (エッジ重みに基づくノード絞り込み＆論理ビュー)