共引用情報を利用した論文関係グラフの提示

中野　優† 清水　敏之†† 吉川　正俊††

† 京都大学工学部情報学科　〒 606–8501 京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8501 京都市左京区吉田本町
E-mail: †ynakano@db.soc.i.kyoto-u.ac.jp, ††{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

あらまし　電子的に利用可能な学術情報が増大しており，研究者にとっては文献検索エンジンの利用が一般的になっている．多くの論文が取得できる現在，研究サーベイを行う際には大量の論文をどのように整理して閲覧するかが課題となる．その際，引用・被引用関係に代表されるように論文間には関連がある場合があるため，その関係が把握できると理解の補助になると考えられる．そのため，サーベイを支援する手法の一つとして，論文間の関係を可視化して提示することが考えられる．我々は，論文間の関係として引用関係だけではなく，共引用関係も利用し，さらに共引用の位置も考慮して複数論文間の関係をグラフで表示することを考えた．さらに，引用および共引用のそれぞれに関して関係の強さの定量化を行い，引用関係と共引用関係をグラフ上で同時に提示することを考えた．また，実際にデータベース分野の論文集合を用いて論文関係グラフを構築し，そのグラフから取得できる情報などについて議論した．

キーワード　共引用分析，関係抽出，引用グラフ

1. はじめに

研究者は研究を行う際に，自身の研究対象に対する知識を得るためや，関連研究を知るために論文のサーベイを行う．その際，Google Scholar(注1) や Microsoft Academic Search(注2) といった文献検索エンジンなどの Web 上の情報を用いることが一般的になっており，研究者はそれらの情報を使って，自らの研究分野に関連がある論文を収集する．しかし，一人の研究者が処理することができる論文数には限界があることに対し，Web 上に存在する文献の量は年々増え続けていおり膨大であるという現状において，いかに大量の文献を整理して効率よく理解するかが問題である．

文献の効率的な整理や理解を補助するためのアプローチとして，論文間の関係を解析し，利用者に提示することが考えられている [1] [2]．論文間の関係を把握することができれば，個々の論文における主張や位置付けをより明確に理解することができる上，関連分野全体に対する理解も得ることができると考えられる．論文間にどのような関係が存在するかを詳しく解析して提示することは，論文間の関係理解を促進させ，論文のサーベイの省力化を手助けすることに繋がると考えられる．本研究では，論文間の関係理解を支援し，論文サーベイを促進させることを目的として，論文関係グラフの構築を行う．

論文間の関係を考える上で，引用・被引用関係をもとに解析を行う手法はこれまでにも多数の研究が行われている [3] [4] [5]．その背景として，研究者が論文を執筆する際には，自身に関連のある研究について引用を行うことによって，自己の研究との類似性や差異について説明し，提案手法の優位性やその研究の位置付けを明確にする，ということがある．本研究においても，ある論文が別の論文を引用している場合，二つの論文の間には何らかの関係性があると考え，論文間の引用関係に着目することとする．

本研究では論文間の関係として，引用関係に加えて共引用の情報にも着目する．共引用とは，ある二つの文献が同時に他の一つの文献から引用されている状況のことであり，共引用を用いた分析は，Small [4] 以来数多く行われてきた [6] [7] [8]．

論文間の関係を分析した後には，利用者に対してどのようにして理解しやすい形で論文間の関係を提示するか，ということが問題となると考えられる．この問題に対して，論文間の関係を可視化してグラフの形で提示することは，これまで多くの研究で行われてきた [1] [2] [9]．我々は，論文間の引用関係や共引用関係，共引用位置など，論文間の関係を可視化するために利用できる情報を整理し，これらを考慮した上で，引用関係を表す枝と共引用関係を表す枝を同時に提示するグラフによって複数論文間の関係を可視化することを考えた．このことによって，ある論文集合において，引用関係から得られる情報と共引用関係から得られる情報を同時に得ることができるようになり，その論文集合全体に対する理解を促進させることができると考えられる．

本稿の構成は以下の通りである．2 節で関連文献について述べる．3 節で論文間の関係をグラフとして提示する際に利用できる情報について整理する．4 節で提案手法について実験を行い，得られた論文関係グラフについての議論を行う．最後に，5 節でまとめと今後の課題について述べる．

2. 関連研究

2.1 論文の引用解析

論文間の関係を得る代表的な手法として，書誌結合 [3] と共引用分析 [4] がある．これら二つの手法はともに，論文間の類似度を測るための指標である．書誌結合とは，二つの文献のそ

(注1)：https://scholar.google.co.jp
(注2)：http://academic.research.microsoft.com

れぞれが，一つ以上の同一の文献を引用していることであり，同一の文献を引用する数が多ければ多いほど，二つの文献の関連が強いと考えられる．共引用分析とは，二つの文献が同時に他の一つの文献から引用されている状況である共引用を用いた分析であり，共引用されている数が多ければ多いほど，二つの文献の関連が強いと考えられる．

文献 [3] および文献 [4] においては，全ての引用が等価に扱われているが，Moravcsik ら [10] が述べているように，それぞれの引用には様々な理由が存在するので，全ての引用を等価に扱うことは適切ではないと考えられる．江藤 [6] は，従来の共引用分析が引用している論文の文脈を無視していることを問題視し，共引用を「非同一段落共引用」「同一段落共引用」「同一文共引用」「列挙共引用」というように，段落や文章などの論文の構成単位ごとに論文の構造から引用箇所の間隔を四つに分類した．さらにこれら四つの分類ごとに，共引用されている二論文間の類似度を計算した結果,「非同一段落共引用」「同一段落共引用」「同一文共引用」「列挙共引用」の順で，類似度が大きくなることを実験によって示している．これは，共引用されている二論文間の間隔が近いほど，類似度が大きくなることを示している．これらの研究は，論文間の関係の強さの判定や，論文間にどのような関係があるかの解析を行ったものであり，本研究はこれらの情報を利用して，論文間の関係の可視化を行った．

2.2 論文間の関係の解析と抽出

難波ら [5] は，論文中において別の論文が引用されている部分の周辺のテキスト（これを引用部周辺テキストと呼ぶ）に対して，手がかり語を用いて，各引用を，type B（他の研究の成果を利用している場合），type C（関連研究との比較，あるいは既存研究の問題点の指摘を行っている場合），type O（type B，type C 以外）の三つに分類を行った上で，その分類を利用して類似度を計算する手法を提案している．難波らはこの手法の有効性を確かめるために，書誌結合や語の共出現などを用いて類似度を計算する手法と比較した結果，難波らが提案した引用分類を用いた類似度計算手法は，精度において最も良い結果を出し，計算コストの面でも十分な結果が出ることを示した．

Teufel ら [11] は，引用の種類を 12 種類に分類する枠組みを提案し，その枠組みの妥当性を検討した上で，その分類を手がかり語などの素性を用いた機械学習で自動的に分類する手法を提案している．さらに，Valenzuela ら [12] は，引用されている論文が，引用している論文に対して大きく影響を与えたかどうかを判別するタスクを行っている．この研究では，12 個の素性を用いて SVM とランダムフォレストのそれぞれで学習させた場合に，SVM では 90%の再現率に対して，65%の適合率を達成している．Semantic Scholar(注3) という文献検索エンジンでは，実際にこの手法を用いて，検索結果論文に対して，その論文に大きく影響を与えた論文と，その論文に大きく影響を与えられた論文の両方を利用者に対して提示している．

正元ら [7] は，共引用テキストという概念を提案している．共引用テキストとは，共引用されている論文が，一つの文中で引用されているような文のことである．また，正元らは，20 本の論文について，論文中にどの程度共引用テキストが出現するかについての予備調査を行っており，どの論文にもある程度多くの共引用テキストをが存在するということを確認している．

これらの研究は，論文間の関係を提示しているという点において本研究と関連しており，これらの情報を論文関係グラフにおいて提示することも考えられる．

引用を用いて，マクロな視点から論文間の関係を解析している研究として，He ら [13] は大規模な文献集合に対して，トピックの変遷を分析するモデルと分析手法を提案している．また，吉田ら [8] は共引用関係をもとにした相関ルールを抽出後，論文のクラスタリングを行うことによって，研究のマクロな流れを提示している．これらの研究が，複数の論文から生成されたトピックやクラスタを一つの節点として扱っていることに対して，本研究は，一つの論文を一つの節点として扱うという点において異なる．

2.3 引用の可視化

論文間の関係を可視化する研究としては以下のような研究がある．Nanba ら [2] は論文間の引用関係に加え，引用論文に記載されている，被引用論文を引用した理由についての記述を可視化してユーザに提示するシステムを提案している．正元ら [1] は文献検索エンジンにおいて，検索結果論文に対して，各論文を節点，引用関係を枝とするグラフを構成し，各節点を時系列順に配置した上で，検索結果論文とともに利用者に提示することで，サーベイの支援を行っている．Shahaf ら [9] は，あるトピックを地下鉄の路線図のように可視化する手法を提案し，その手法を学術論文に適用している．Shahaf らは,「良い路線図」には何が必要かを考え，それをトピックの可視化の最適化問題として定義している．本研究は，引用関係と共引用関係の両者を利用し，共引用の間隔まで利用したグラフの提示を行っているという点において，これらの研究とは異なる．

3. 論文間の関係の提示

本研究では，与えられたある論文集合に対して，有向グラフを作成し提示を行う．このグラフのことを本論文では**論文関係グラフ**と呼ぶ．与えられる論文集合は，例えば論文検索エンジンに対するあるクエリの検索結果論文集合といったような，互いに関連のある論文集合であると仮定する．また，論文関係グラフにおける節点（node）とは，論文集合に含まれる一つ一つの論文を表し，枝（edge）は，各論文間の関係を表すものとする．つまり，論文集合に含まれる二論文間に関係がある場合には，それらの論文を表す節点の間に枝が存在すると考える．

本研究で利用する論文間の関係は，引用関係と共引用関係である．つまり本研究では，これら両者の関係を考慮し，両方の関係を同時に表示するグラフを提示する．これにより，ある論文集合において，ある論文が論文集合に含まれる他の論文からどれだけ引用されているかといった引用関係に関する情報と，どの二論文が他の論文からどれだけ共引用されているか，つまり二論文間にどれだけ関連があるかといった共引用関係に関する情報を，我々が提示する論文関係グラフならば同時に読み取

(注3)：https://www.semanticscholar.org/

ることができると考えている．

本節では，ある関連する論文集合に対して，それらの論文間にどのように枝を張るべきか，どのように節点を配置すべきかについて考える．また，枝に対して，その枝が持つ関係の種類(引用関係か共引用関係か) や，関係の強さについても考え，それらを考慮した論文間の関係を論文関係グラフとして提示する手法を提案する．

論文集合から，論文関係グラフを構成する手法の概要は以下の通りである．

（1） 与えられた論文集合内の論文について，引用関係と共引用関係についての情報を抽出する

（2） 得られた引用関係と共引用関係のそれぞれについて，関係の強さを計算する

（3） 関係の強さに基いて，出力すべき枝を決定する

（4） 節点を適切な位置に配置する

以下の小節で，どのような枝を出力するか，どのように節点を配置するかを決定する際に考えるべき情報について整理する．

3.1 枝の決定手法

ここでは，論文集合から得られた引用・被引用関係に基づいて，どの枝を論文関係グラフの枝として提示するか，提示する枝をどのように表示するかについて説明する．本研究で，二論文間に関係があると考える場合は，以下の 2 通りの場合である．

（1） 二論文間が引用関係にある場合

（2） 二論文間が共引用関係にある場合

以下では，上記 (1)(2) に対して，どのように枝を提示するかについて述べる．

3.1.1 二論文間に引用関係がある場合

研究者は論文を引用する際，どの論文が自らの研究に関連があるかを判断して引用するため，引用関係にある二論文間には強い関係があると考えられる．

しかしながら研究者は，一つの論文において数多くの論文を引用するため，全ての引用関係を表す枝を提示してしまうと，利用者に提示する情報量が多くなりすぎる可能性がある．そこで，引用論文と強い関係にある被引用論文の間の枝に絞って提示する手法を考える．例えば，一つの論文内で何度も引き合いに出されて引用されている論文は，引用論文との関係がより強い論文であると考えられるため，そのような引用論文と被引用論文の間には枝を張る価値が高いと考えられる．さらに提示する関係に関しても，関係の強さを考慮して提示を行うべきであると考えられる．例えば，提示する枝の中でも，より関係が強い枝をより太い枝で表現するといった方法が考えられる．

本研究では，引用関係にある二論文間の関係の強さについて一つの定式化を試みた．行列 M^{cite} の (i,j)-成分 m_{ij}^{cite} を，論文 i と論文 j の二論文間の引用関係の強さとしたときに，m_{ij}^{cite} を以下で定義する．

$$m_{ij}^{cite} = \frac{\text{論文 } i \text{ における論文 } j \text{ の引用回数}}{\text{論文 } i \text{ における引用の総回数}} \tag{1}$$

そして、m_{ij}^{cite} が閾値 $\alpha(i)$ 以上の場合に論文 i を表す節点から論文 j を表す節点への枝を構築することを考えた．つまり，$\alpha(i)$ の値が大きければ大きいほど，構築される枝の数は少なくなる．ただし，α は一つの論文を引数にとって数値を返す関数であり，

$$\alpha(i) = M^{cite} \text{の } i \text{ 行の値の上位 } a\% \tag{2}$$

とする．a はパラメータであり，$a = 0$ のとき，全ての引用関係について枝が構築され，a が大きくなるに従って構築される枝が減少していき，$a = 100$ のときは，引用関係に関する枝は一つも構築されない．

3.1.2 二論文間に共引用関係がある場合

2 節でも述べたように，共引用関係にある二つの論文には何らかの関係があると考えられる．しかしながら，二論文間に引用関係がある場合と同様に，全ての共引用関係を提示するべきではないと思われる．そこで，共引用されている回数をもとに類似度を計算し，類似度が高いもののみを提示することとする．

また，従来の共引用分析には回数のみが考慮されていた問題点を克服する際に使うことができる情報としては，共引用の間隔を用いることが考えられる．例えば，一つの文で列挙して共引用されている二論文と，異なる章で引用されている二論文では，前者のの二論文の方がより強い関係を持つと考えられる．2 節でも説明したように，江藤 [6] は共引用の間隔を用いて，定義した間隔ごとに二論文間の類似度を計算した結果，共引用の間隔が短ければ短いほど，二論文間の類似度が高くなることを示している．

本研究では，共引用関係にある二論文間の関係の強さについて一つの定式化を試みた．行列 M^{cocite} の (i,j)-成分 m_{ij}^{cocite} を論文 i と論文 j の二つの論文間の共引用関係の強さとしたときに，行列 m_{ij}^{cocite} を以下で定義する．

$$m_{ij}^{cocite} = year\_coef_{ij} \times \sum_{x \in papers} \text{論文 } x \text{ での } cocite\_prox(i,j) \tag{3}$$

ただし $papers$ は，論文 i と論文 j が共引用されている論文集合全体を指す．ここで，論文 x での $cocite\_prox(i,j)$ は共引用の間隔を表しており、

$$cocite\_prox(i,j) = \begin{cases} 1 & \text{(列挙共引用)} \\ 0.75 & \text{(同一文共引用)} \\ \big(0.5 & \text{(同一段落共引用)}\big) \\ 0.25 & \text{(同一節・章共引用)} \\ 0 & \text{(それ以外)} \end{cases} \tag{4}$$

とする．江藤 [6] の分類では,「非同一段落共引用」「同一段落共引用」「同一文共引用」「列挙共引用」の四つであったが，本研究では実装の都合上，段落を取得することが容易ではなかったので，共引用の分類から「非同一段落共引用」を除外し,「同一節・章共引用」と「それ以外（非同一節・章共引用）」を導入した．また，ある共引用されている二論文に対して，同一論文内で複数の共引用が存在する場合は，共引用の間隔が一番近いものを共引用の間隔とする．

$year\_coef_{ij}$ は以下で定義される値である．

$$year\_coef_{ij} = \left(\frac{year_i + year_j - (start - 1) * 2}{interval * 2} \right)^2 \tag{5}$$

ここで，$year_i, year_j$ は論文 i, j の発表年であり，また，$start$ は論文集合内における初年，$interval$ は論文集合の期間である．これは，「発表年が新しい割に共引用数が多い論文間には比較的強い関係がある」という考えに基づいた値となっている．

そして，m_{ij}^{cocite} が閾値 $\beta(i,j)$ 以上の場合に，論文 i と論文 j の間に枝を構築することを考えた．つまり，$\beta(i,j)$ の値が大きければ大きいほど，構築される枝の数は少なくなる．ただし，β は共引用関係にある二つの論文を引数にとり，数値を返す関数であるとし，ここでは，

$$\beta(i,j) = M^{cocite}\text{の非ゼロ要素の値の上位 } b\% \qquad (6)$$

とした．b はパラメータであり，$b = 0$ のとき，全ての共引用関係について枝が構築され，b が大きくなるに従って構築される枝が減少していき，$b = 100$ のときは，共引用関係にある枝は一つも構築されない．

3.2 節点の配置手法

利用者が提示されたグラフを見た際に，表示される節点の位置に一貫性がなければ，そのグラフから情報を得ることが難しくなると考えられる．そこで本研究では，発表年を横軸として固定し，各論文をその論文が発表された年に節点として配置した．このことによって，論文集合における研究の時系列的な流れを理解することができる可能性があると考えられる．論文を発表年ごとに配置する手法は，2.3 節で紹介した引用の可視化に関する全ての既存研究で用いられている．

4. 実　　験

3 節において説明した手法を実際の論文集合に適用して，実験を行った．

4.1 実 験 手 順

実験の大まかな流れは以下の通りである．

（1） ある論文集合を，全体論文集合と定める

（2） 提示したいある一つのトピックについて適当な単語を一つ決めて，その語を用いて Google Scholar で検索を行う

（3） 検索クエリの結果から，先程の全体論文集合に含まれる論文を抽出したものを関連論文集合とする

（4） 関連論文集合から被引用数上位 k 本の論文に対して，論文関係グラフを作成する

本実験では，全体論文集合を，ACM SIGMOD International Conference on Management of Data (SIGMOD)(注4)，International Conference on Very Large Data Bases (VLDB)(注5)，IEEE International Conference on Data Engineering (ICDE)(注6) の三つの会議において，2000 年から 2015 年に発表された 16 年分の論文と定めた．これら三つの国際会議を選んだ理由としては，三会議ともデータベース分野における著名な国際会議であり，類似したトピックの論文が発表されているため，関係を得るのに適した集合であると考えたからである．

表 1　引用と共引用の個数

	全体	論文集合内
引用	201,404 個	47,716 個
共引用	1,664,014 個	100,355 個

次に実験に使用するデータの収集について説明する．まず先程述べた三つの国際会議の論文 PDF を収集した．その後 Poppler(注7) の `pdftotext` を用いて PDF からテキストを抽出した．そこから，引用関係と共引用関係を含む引用・被引用関係を抽出した．引用・被引用関係の取得には ParsCit(注8) [14] を用いた．ParsCit は，論文中のどの部分が本文でどの部分が図や数式であるかや章の始まりがどこかなどの論文の論理的な構造，参考文献リストに含まれる各参考文献のタイトルや著者などの書誌情報，各参考文献の本文中における引用箇所などの情報を XML 形式で取得することができる．最後に，収集した論文のタイトルなどの書誌情報を持つテーブルと各論文から抽出した引用情報を持つテーブルが含まれるデータベースを作成した．

収集したデータの概要が表 1 である．収集の結果，6,977 個の論文が得られ，これらの論文から 201,404 個の引用，1,664,014 個の共引用を取得できた．引用関係や共引用関係にある二つの論文がともに 6,977 個の論文集合に含まれる場合は，引用は 47,716 個で，共引用は 100,355 個であった．ここでは，一つの論文内で同じ論文が複数回引用されている場合も，それぞれの引用を一回として数えた．また，共引用の個数は，ある論文が N 個の論文を引用しているとき，N 個から 2 個選ぶ組み合わせの数の合計として計算した．

これらの収集したデータを用いて，“skyline”，“top-k queries”，“uncertain data” の三つのクエリのそれぞれに対して論文関係グラフを作成した．具体的には，各クエリに対して，$k = 10, 15, 20$ と $a = 0, 5, 10, 20, 30, 100$ と $b = 0, 10, 20, 30, 40, 100$ を組 (k, a, b) として組み合わせた $3 \times 6 \times 6 = 98$ 通りについて，グラフを作成した．ただし，k は上記の手順で説明した，Google Scholar でから抽出した関連論文集合の被引用数上位 k 本のことを指し，a, b はそれぞれ 3 節の式 (2)，式 (6) におけるパラメータ a, b である．

また，グラフの図示には Graphviz(注9) [15] というオープンソースのグラフ可視化ソフトウェアを用いた．Graphviz では，DOT 言語という形式でグラフにどのような節点や枝があるかを記述するが，各節点や枝に対して，円形や四角などの節点の形や枝の太さなどを属性として指定することができる．

4.2 結　　果

クエリ “skyline” で $(k, a, b) = (15, 0, 20)$ とした論文関係グラフが図 1 である．図の各要素について以下で説明する．図において，グラフの各節点は論文を表しており，節点には論文の ID，会議名，発表年が書かれている．黒色の枝は引用関係を示

(注4)：http://www.sigmod.org/
(注5)：http://www.vldb.org/
(注6)：http://www.icde.org/
(注7)：http://poppler.freedesktop.org/
(注8)：http://aye.comp.nus.edu.sg/parsCit/
(注9)：http://www.graphviz.org/

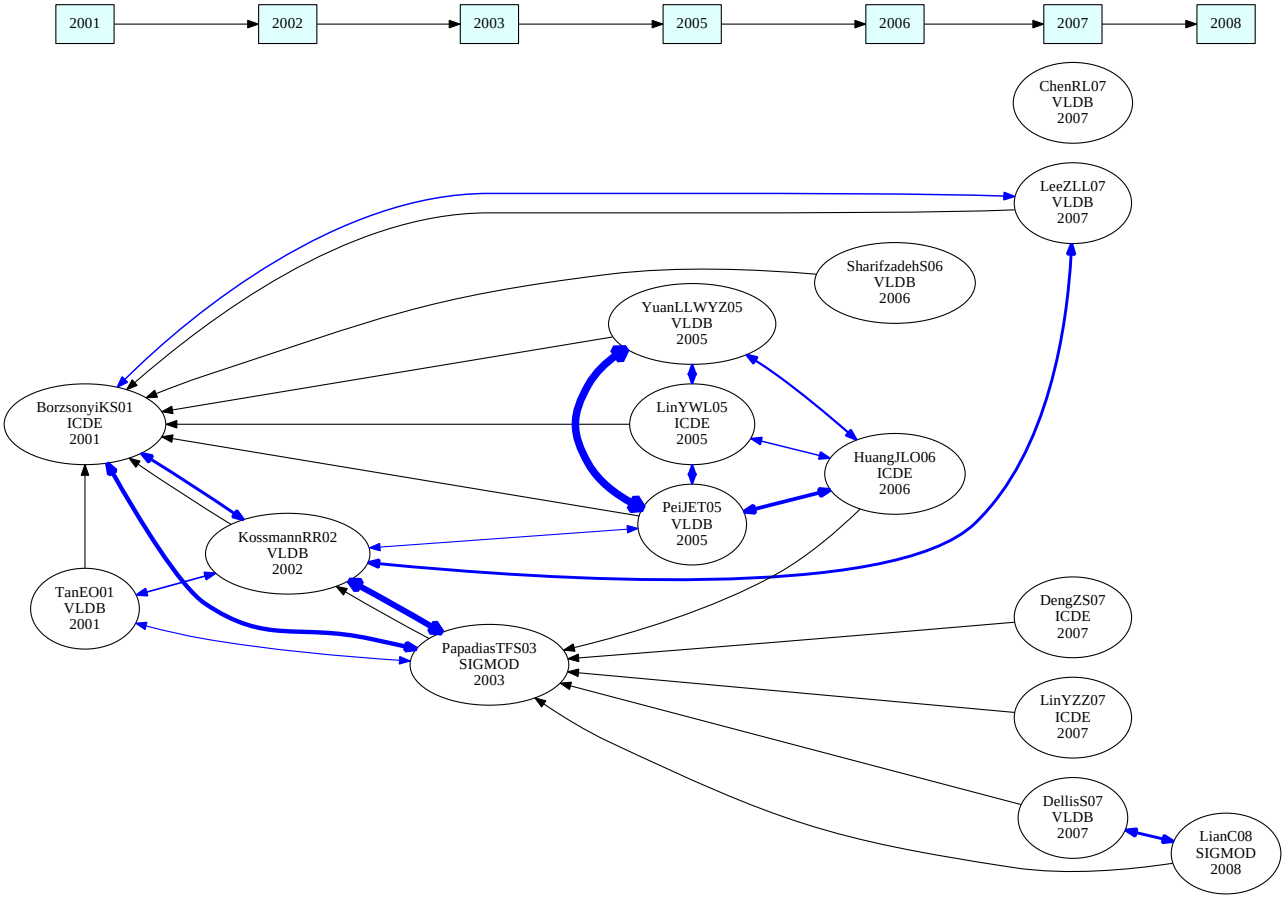

図 1　クエリ:skyline, $k = 15, a = 5, b = 20$

しており，節点 A から節点 B を指す黒色の矢印は，節点 A を表す論文 A が節点 B を表す論文 B を引用していることを意味する．青色の枝は共引用関係を表している．また，枝の太さはその関係の強さを表している．

図 1 から読み取れることとしては，例えば，論文 BorzsonyiKS01 や PapadiasTFS03 は多くの引用関係の枝が接続しているので，この二論文は他の論文に大きな影響を与えた重要な論文であるのだろうといったことが推測できる．これは，単に引用関係のみを可視化したグラフ，つまり $(k, a, b) = (15, 5, 0)$ として図 1 から引用関係の枝 (黒色の枝) のみを残して提示するグラフからも推測できることである．また，共引用関係を表す枝に着目することによって，2005 年から 2006 年の論文 YuanLLWYZ05，LinYWL05，PeiJET05，HuangJLO06 の四つの論文や，2001 年から 2003 年の BorzsonyiKS01，KossmannRR02，PapadiasTFS03 の三つの論文については，それぞれ一つのクラスタのようなものを構成するのではないか，といったような情報を得ることができると考えられる．これも同様に，共引用関係のみを可視化したグラフ，つまり $(k, a, b) = (15, 0, 20)$ として図 1 から引用関係の枝 (黒色の枝) のみを残して提示するグラフからも推測できることである．しかし，引用関係を表す枝と共引用関係を表す枝の両方を同時に表示した図 1 においては，共引用関係を表す枝によって作られると推測できる複数のクラスタ間の関係にどのような関係があるかといった情報や，あるクラスタがどの論文から強い影響を受けたかという情報などを得ることができる可能性があるという利点を持っている．これは，引用関係や共引用関係をそれぞれ別個に表示した場合にはない利点である．なお，この利点は，クエリ “top-k queries” および “uncertain data” に関して作成した論文関係グラフにおいても観察された．

図 2 は，図 1 の a の値を 5 から 10 に変化させたときの図である．図 1 と比較した場合，引用関係を表す黒色の枝が増えていることが確認できる．図 2 を詳しく見ると，論文 BorzsonyiKS01 と PapadiasTFS03 の他に，論文 TanEO01 や KossmannRR02 が多く引用されていることがわかるが，その中でも前者の二論文を表す節点には，多くの太い枝が接続していることから，前者の二論文とより強い引用関係を持つ論文が多いことがわかる．また，図 2 は，ある一つの論文節点に着目した際に，その節点から出る枝の太さを見ることによって，着目した論文とその論文が引用している論文の間の関係の強弱について，図 1 よりも詳細に読み取ることができるという利点がある．その一方で，枝が多くなったことによって図が煩雑になり，全体の関係把握に負担がかかるといった欠点が挙げられる．よって，一つ一つの論文について被引用論文との引用関係の強さを知りたい場合には a を大きくすればよく，また，論文集合全体として論文間にどのような引用関係があるかを知りたい場合には a を小さくすればよいと考えられる．

図 3 は，図 1 の b の値を 20 から 40 に変化させたときの図である．図 1 と比較した場合，共引用関係を表す青色の枝が増

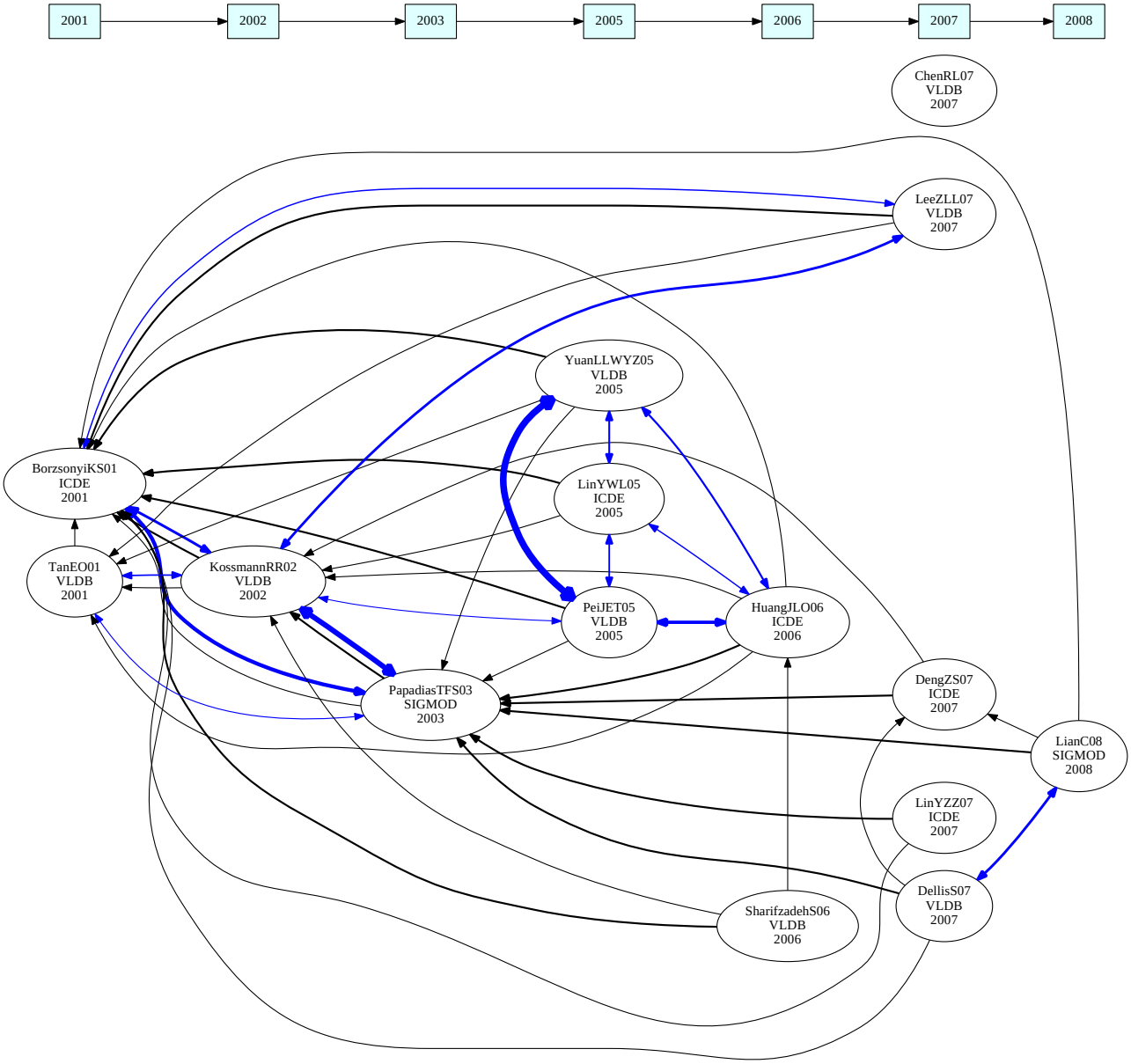

図 2　クエリ:skyline, $k = 15, a = 10, b = 20$

えていることが確認できる．図 3 を見ると，出力した論文を表す節点間に共引用関係があり，多くの論文間に互いに何らかの関係があることがわかる．しかしながら図 3 は，図 1 のように上記で述べたような関係は推測できるものの，図 1 と比べるといささか共引用を表す枝が煩雑であるように思われる．クエリ “skyline” のように，出力する論文を表す節点間で多くの共引用関係がある場合，図 3 のように枝が煩雑になりすぎる可能性がある．

クエリ “skyline” の論文集合において特徴的である点として，論文 BorzsonyiKS01 を他のほぼ全ての論文が引用しているという点が挙げられる，これは，SQL における skyline 問合せという操作がこの論文において初めて提案された概念であるという理由が考えられる．また，論文間に非常に共引用が多いということがある．これは，関連論文集合に含まれてる論文が $k = 20$（引用数上位 20 本）であっても，全ての論文の引用数が 100 以上の非常に著名な論文であるという理由が考えられる．また，論文 ChenRL07 を表す節点が図 1，2，3 のどの図においても引用関係の枝も共引用関係の枝も全く接続されていないことが確認できる．実際，この論文の節点は，本論文で示した図以外のパラメータのグラフにおいても他の論文節点と接続する枝が存在しないことを確認した．この理由として，論文 ChenRL07 における skyline 問合せという概念自体は論文 BorzsonyiKS01 と同様の操作であるが，論文 ChenRL07 の扱っている分野がプライバシーに関する分野であり，引用されている論文もプライバシーに関する分野の論文ばかりであるという理由が考えられる．

4.3 議　　論

引用を行う回数は会議や著者によって大きく異なることが多いと思われるため，「論文 i が論文 j を引用している」といったような引用関係を枝として提示する際に，本実験では，式 (1) および式 (2) に基づき，論文 i での論文 j の引用回数の割合によって枝を出力するかどうかを決定した．例えば，ある二つの

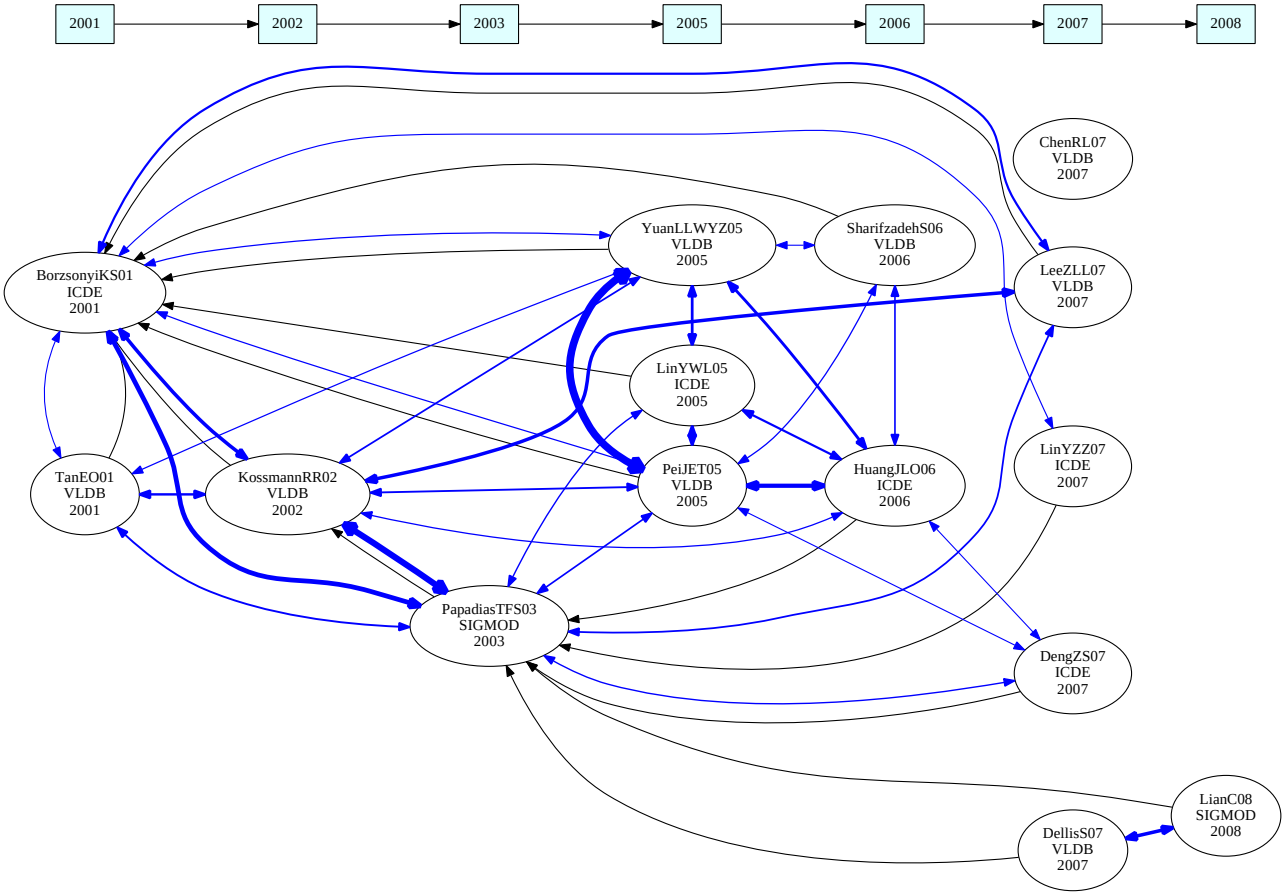

図 3　クエリ:skyline, $k = 15, a = 5, b = 40$

論文 i_1, i_2 の両方において，論文 j が c 回引用されており，論文 j 以外の論文が全て 1 回ずつ引用されているという状況を考える．このとき，論文 i_1, i_2 における引用の総回数がそれぞれ 20 個，50 個である場合，つまり式 (1) の分母が大きく異なる場合，これらの論文間の引用関係の強さは大きく異なることとなる．上記の状況で関係の強さを等価と考える場合は，単に引用回数を用いるということも考えられる．

5. おわりに

5.1 まとめ

本研究では，論文間の関係理解を支援し，論文サーベイを促進させることを目的として，引用関係と共引用関係の両方を利用した論文関係グラフを構築するために，引用関係や共引用関係，共引用の間隔などの用いることができる引用情報について整理し，論文間の関係の強さを表す式の一つの定式化を行った．提案した式に基づいて，実際に論文関係グラフを構築するにあたり，データベース関連の国際会議である SIGMOD，VLDB，ICDE の三つの会議について，2000 年から 2015 年の 16 年分の論文を全体論文集合と定めて論文 PDF から引用情報を収集した．そこから “skyline”，“top-k queries”，“uncertain data” という三つのクエリそれぞれを Google Scholar で検索した結果から全体論文集合に含まれる論文を抽出し，論文関係グラフを構築した．提案する論文関係グラフは，引用関係と共引用関係の枝をグラフ中に同時に提示したものであり，得られたグラフに対して議論と考察を行った．

5.2 今後の課題

今後の課題としてまず考えうることは，提案手法の評価に関する課題である．つまり，提案した手法をシステム化し，そのシステムにおいて提案手法の評価を行うということが考えられる．評価としては，例えば，利用者に対して「サーベイの効率が向上したか」といった主観的な項目をアンケートなどを用いて問う実験や，ある論文集合を可視化した際に，その論文集合からあるトピックに関連がある論文を利用者に選択させた場合の精度や再現度などを用いて評価を行う実験などが考えられる．

次に，論文集合に関する課題が挙げられる．本研究では全体の論文集合をデータベース系の三つの国際会議のみについて関係を調べたが，より広い論文集合に対して提案した手法を適用することが考えられる．これにより，より多くの引用関係や共引用関係を取得することができ，より広い範囲の論文集合に対して論文関係グラフの提示を行うことができるようになると考えられる，

最後に，論文関係グラフに関する課題が挙げられる．本研究では，引用や共引用の回数を数えることや，共引用の間隔を用いて論文関係グラフを可視化したが，これら以外にも用いることができると考えられる情報として，暗黙の引用，引用の位置，引用部周辺テキスト，共引用テキストなどが挙げられる．

- 暗黙の引用

提案手法において引用は，発表される国際会議等で定められた

形式での明示的な引用のみを引用として用いた．これに対して，論文中には，単にアルゴリズム名や（主）著者名のみを文中で述べるといったような場合が考えられ，このようなものは暗黙の引用として考えることができる．暗黙の引用については，Valenzuela ら [12] の研究において抽出が試みられている．本実験では考慮しなかったが，より詳細な関係の提示を考えていく上では，暗黙の引用も考慮に入れるべきであると考えられる．

- 引用の位置

提案手法では，引用関係にある二論文間の関係の強さを，引用論文において被引用論文が引用されている割合で定式化したが，これは 2 節において述べた「全ての引用を等価に扱っている」という従来の書誌結合の問題点を克服できていないという課題がある．この課題を解決するために利用できる情報として考えられるものとして，引用関係の位置が挙げられる．例えば，関連研究の章で引用されている論文と，実験の章で引用されている論文を比較すると，関連研究の章では単に自らに関連する研究についていくつか列挙しただけという可能性があるが，実験の章で引用される場合は，その論文の手法などと密接な関係があると推測され，その関連度は異なると考えられる．引用関係の位置を考慮した研究としては，Valenzuela ら [12] の研究が挙げられる．Valenzuela らは，ある論文が被引用論文から大きく影響を与えられたかを判定するために SVM とランダムフォレストを用いた機械学習を用いているが，その素性の一つとして，被引用論文の章ごとの被引用数を数えたものを用いている．

- 引用部周辺テキスト

引用部周辺テキストは，2 節でも触れた概念である．ある論文 A が別の論文 B を引用する際，論文 A は論文 B が行った手法や実験に加えて，論文 A の筆者からみた論文 A との関連性や問題点などを説明する．よって論文 B が引用されている箇所周辺のテキスト，つまり引用部周辺テキストを自然言語処理によって解析を行うことで，二論文間の関係の強さをより正確に算出することができると考えられる．さらには，単に引用関係や共引用関係にとどまらず，論文間の並列関係や発展関係などといった，論文関係グラフの枝の属性のタイプ分けを行うことができれば，論文間の関係をより詳細に提示できると考えられる．

- 共引用テキスト

共引用テキストは，2 節でも触れた概念である．共引用テキストを用いることで，引用部周辺テキストの場合と同様に，論文関係グラフにおける枝の属性のタイプ分けを行うことができると考えられる．引用部周辺テキストの場合，一つの論文中において同一論文が引用されることは高々数回しかないため，それだけの情報で論文間の関係を明らかにできるかどうかには疑問が残ると思われるが，共引用テキストを追加の情報として用いることで，論文間の関係をより明確にすることができる可能性があると考えられる．さらに，引用部周辺テキストにはないが共引用テキストにある利点として，引用関係がない場合であっても，文脈に踏み込んだ論文間の関係が解析できる可能性があるという点が挙げられる．

また，利用者が論文間の関係をより具体的に解釈することを可能にするために，引用部周辺テキストや共引用テキストの分かりやすい要約を枝の情報として提示することもできると考えられる．これらの情報を用いて，より分かりやすい論文関係グラフを提示することが，今後の課題である．

文　献

[1] S. Shogen, T. Shimizu, and M. Yoshikawa. Enrichment of academic search engine results pages by citation-based graphs. In *Proceedings of the 11th Asia Information Retrieval Societies Conference (AIRS)*, pp. 56–67, 2015.

[2] H. Nanba, T. Abekawa, M. Okumura, and S. Saito. Bilingual presri-integration of multiple research paper databases. In *Proceedings of 7th International Conference on Recherche d'Information et ses Applications (RIAO)*, pp. 195–211, 2004.

[3] M. M. Kessler. Bibliographic coupling between scientific papers. *American documentation*, Vol. 14, No. 1, pp. 10–25, 1963.

[4] 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*, Vol. 24, No. 4, pp. 265–269, 1973.

[5] 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.

[6] 江藤正己. 論文の構成単位に基づいた共引用関係の尺度. 情報処理学会論文誌, Vol. 49, No. 7, pp. 1–15, 2008.

[7] 正元修平, 清水敏之, 吉川正俊. 共引用テキストを利用した論文間の関係抽出. In *DEIM Forum 2014 C5-1*, 2014.

[8] 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, Vol. 45, No. 7, pp. 24–32, 2004.

[9] D. Shahaf, C. Guestrin, E. Horvitz, and J. Leskovec. Information cartography. *Commun. ACM*, Vol. 58, No. 11, pp. 62–73, 2015.

[10] M. J. Moravcsik and P. Murugesan. Some results on the function and quality of citations. *Social studies of science*, Vol. 5, No. 1, pp. 86–92, 1975.

[11] S. Teufel, A. Siddharthan, and D. Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 103–110, 2006.

[12] M. Valenzuela, V. Ha, and O. Etzioni. Identifying meaningful citations. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[13] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and C. L. Giles. Detecting topic evolution in scientific literature: how can citations help? In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, pp. 957–966, 2009.

[14] I. G. Councill, C. L. Giles, and M. Kan. Parscit: an open-source CRF reference string parsing package. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2008.

[15] J. Ellson, E. R. Gansner, E. Koutsofios, S. C. North, and G. Woodhull. Graphviz and dynagraph - static and dynamic graph drawing tools. In *Graph Drawing Software*, pp. 127–148. Springer, 2004.