6N-02

引用関係に基づくグラフの付与による文献検索エンジンの高機能化

正元 修平† 清水 敏之† 吉川 正俊†

京都大学情報学研究科社会情報学専攻†

1 はじめに

研究者による学術論文のサーベイを省力化し，支援することは，研究者が自らの研究をより高度で客観性の高いものとするために重要である．本稿では，「文献検索エンジン」と呼ばれる，学術論文の検索に特化した検索エンジンの検索結果ページに，引用・被引用関係に基づくグラフを付与することで，論文のサーベイを効率化する手法を提案する．

一般的に，文献検索エンジンを利用する際，ユーザは，各検索結果のスニペット（要約文）を閲覧することで，その検索結果論文が自らの研究に関連するかどうかを判断している．しかし，検索結果ページのスニペットを表示するスペースには限りがあることから，スニペットを閲覧するだけでは，ユーザがどの論文をサーベイの対象とすべきかわからない場合がある．この問題点を解決するために，我々は，論文間の引用・被引用関係に着目する．引用・被引用関係にある 2 つの論文については，一方の論文を調査したユーザは，もう一方の論文を調査対象としやすい傾向があると考えられる．このような論文間の関連を可視化し，文献検索エンジンの検索結果ページ上に理解しやすい形で提示することによって，ユーザが検索結果ページを閲覧する際に，自らの研究に関連すると考えられ，調査の対象となるような論文と，そうでない論文を容易に見分けることができるようになり，結果として，サーベイの効率を向上させることができると考えられる．我々は，文献検索エンジンの検索結果論文と，それらの論文と引用・被引用関係にある論文（以下，これを「検索結果論文周辺の引用・被引用論文」と呼ぶ）をノードとし，それら論文ノード間の引用・被引用関係をエッジとするグラフを，検索結果ページに付与し，理解しやすい形でユーザに提示する手法を提案する．

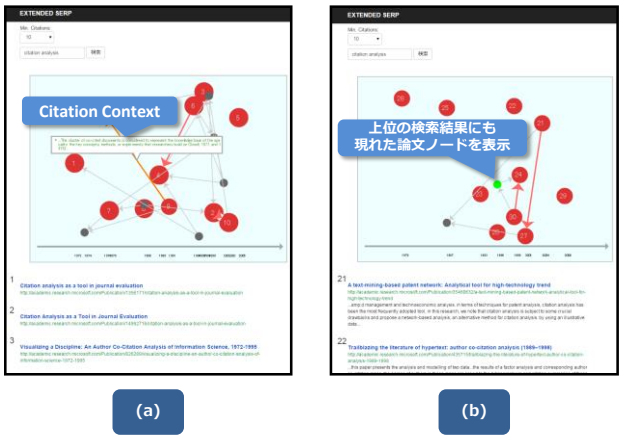

図 1 (a) Citation Context を付与したグラフと，(b)検索結果下位におけるグラフ

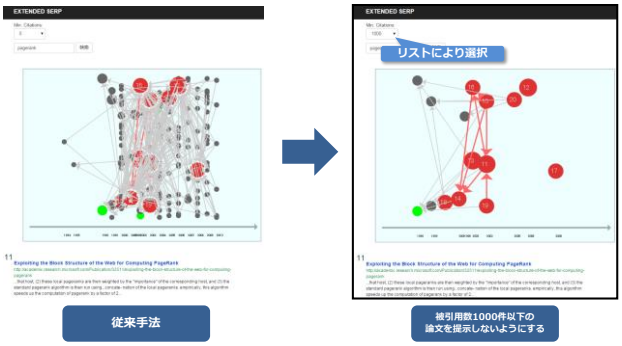

図 2 検索結果論文周辺の引用・被引用論文のノード数削減

2 提案手法

我々は，1 節で述べた研究目的について，以前に[1]に示す手法に基づいてインタフェースを実装し，ユーザ調査を行った．ここでは，以前の手法からの差分について述べることとする．

以前の手法からの差分は，以下の 4 点である．

1. エッジが持つ属性として，「Citation Context」を導入する
2. 検索結果下位の検索結果ページ（例えば，上位 11～20 件目の検索結果ページ）に，グラフをどのように表示すべきかを考慮する
3. 検索結果論文の被引用数に応じて，グラフ上に表示されるノードの大きさを変更するようにする
4. 検索結果論文周辺の引用・被引用論文のうち，ユーザが設定するしきい値以下の被引用数のものについては提示しないようにする

1 について，同じ引用・被引用関係にある 2 論

Enhancement of Academic Search Engines using a Graph Based on Citation

†Shuhei Shogen, †Toshiyuki Shimizu and †Masatoshi Yoshikawa

†Department of Social Informatics, Graduate School of Informatics, Kyoto University

文であっても，例えば，「引用している論文における提案手法を改善する手法を考案した」という理由での引用と，「引用している論文の評価手法と同じ評価手法を用いた」という理由での引用では，それらの論文間の関連の種類は異なると考えられる[2]．今回は，そのような関連の種類の違いをユーザに知らせるために，「Citation Context」と呼ばれる記述を，グラフにおけるエッジが持つ属性として導入する．ここで，ある論文 P が別の論文 Q に引用されているとき，論文 P の論文 Q における「Citation Context」とは，論文 Q 中に存在する，論文 P について言及している記述のことを指す．例えば，ある論文中の「Nanba ら[a]は，論文間の引用・被引用関係に加え，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．」といったような記述が，Citation Context に相当する．このような記述をグラフに付与することで，どのような理由で引用を行ったかが明確になり，ユーザがより論文間の関連を理解しやすくなると考えられる．今回は，図 1 (a) に示すように，ユーザがグラフ上のエッジをマウスホバーした際に，Citation Context が表示されるように実装した．

また，2 について，[1]で実装したインタフェースでは，ある入力クエリに対して提示される検索結果は，すべてのインタフェースについて，上位 10 件までとなっていた．これは「数多くの論文中から自らの研究に関連する論文を探し出す」という論文のサーベイの目的を考えると，不十分であると考えられる．そこで，今回は，入力クエリに対する検索結果を，10 件ずつ，最大上位 100 件まで提示するように拡張し，各々の検索結果ページ（例えば，ある入力クエリに対する検索結果の 21〜30 件目に相当するページ）については，それらの検索結果論文集合に対するグラフを付与することとした．また，これらの検索結果論文集合について，その周辺の引用・被引用論文を提示する際に，それらの論文の中に，その検索結果論文より上位の検索結果に現れる論文が現れた場合（例えば，検索結果の 31〜40 件目に相当するページで，その周辺の引用・被引用論文の中に，検索結果上位 15 件目の論文が現れた場合）については，これを別の種類のノードとして表示するようにしている．これにより，検索結果の上位に現れる論文と下位に現れる論文の関連を理解することができ，結果として，論文間の関連の見落としを減らすことができると考えられる．今回は，図 1 (b) に示すように，この種類のノードを，異なる色で表示するようにした．

3 について，グラフに表示するノードの大きさを変更することは，そのノードに対応する検索結果論文を強調し，ユーザに注目させる効果があると考えられる．また，論文の被引用数は，その論文の重要度を表すと考えられる．これらのことから，今回は，論文の被引用数に応じて，表示するノード（円）の半径の大きさを変更するという手法を取ることとした．これにより，論文間の関連だけでなく，単一の論文の重要度についてもグラフ上で表現できるようになり，例えば，「重要度の高い 2 つの論文間に引用・被引用関係があるので，この論文間の関連は重要である可能性が高い」といったことがわかるようになるのではないかと考えられる．

最後に，4 について，図 2 左に示すように，[1]の提案手法に基づいて実装したインタフェースでは，入力されるクエリによっては，グラフに提示されるノードの数が非常に多くなり，グラフが見づらくなることがあった．そこで，今回は，検索結果論文周辺の引用・被引用論文について，その論文の被引用数がしきい値以下だった場合については，その論文ノードをグラフに取り入れないような手法を取ることとした．ここで，しきい値はユーザが動的に設定できるようにした．図 2 左と同じ入力クエリによる検索結果において，被引用数 1000 件をしきい値とした場合のグラフを図 2 右に示す．これにより，ユーザは，検索結果論文間の大まかな関連を知りたい場合には，しきい値を大きく設定し，そこで見つけた関連についてより詳細な情報が知りたい場合には，しきい値を小さく設定することが可能になると考えられる．

5 まとめ

我々は，文献検索エンジンに，論文間の関連を示すグラフを付与することによって，論文のサーベイの効率を向上する手法を提案した．今後の課題として，実際にユーザに論文のサーベイを行ってもらう形での評価実験を行い，提案手法の妥当性を検証することが挙げられる．

参考文献

[1] 正元修平，清水敏之，吉川正俊．文献検索エンジン結果ページへの引用関係に基づくグラフの付与．第 7 回 Web とデータベースに関するフォーラム．2014.

[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，42(11):2640–2649，2001.