様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 ２９ 日現在

機関番号：３２６９９
研究種目：若手研究(B)
研究期間：2011～2014
課題番号：２３７００２８９

研究課題名（和文）共引用関係に基づく文献検索手法の高度化

研究課題名（英文）Improving Document Retrieval Using Co-citation Relationships

研究代表者
江藤　正己（ETO, Masaki）

学習院女子大学・国際文化交流学部・講師

研究者番号：１０５８４８０７

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究の目的は，共引用関係を利用した検索において，その長所「キーワード検索では見つけられない適合文献を発見できる」を強化した新たな検索手法を開発することである。本研究では，共引用関係を多段階化することでより多くの文献を見つけ，その見つけた文献群を文脈情報や共引用回数を用いてランク付けする方法を考案した。さらに，検索実験をおこない，従来の共引用関係を利用した手法と比較し，開発手法の検索性能の方が高いことを確認した。

研究成果の概要（英文）：The aim of this research is to extend co-citation searching; a strength of the searching is retrieving relevant documents undetectable by words. This research proposes a retrieval method which detects many documents through spread co-citation relationships and ranks the detected documents by co-citation frequency and co-citation contexts. Results of information retrieval experiments showed that the proposed method outperformed the traditional co-citation searching.

研究分野：図書館情報学

キーワード：情報図書館学　情報システム　情報検索　共引用　引用索引　文脈情報

１．研究開始当初の背景

次世代型の文献検索システムでは，キーワード検索は当然のものとして，「それを補完する検索機能をいかに充実させるか」が重要な要素である。そして，学術文献の検索システムでは，「キーワード検索では見つけられない適合文献を発見できる」という長所を持つ，引用関係を利用した検索をさらに充実させることが有力な発展の方向性の一つといえる。

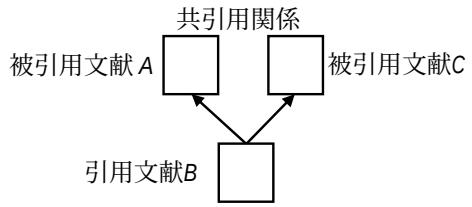

図１ 共引用関係

特に本研究では，引用関係の一形態である，共引用の関係に着目した。共引用とは，一つの文献（文献 B）が二つの文献を同時に引用していた場合に，二つの被引用文献間（文献 A と文献 C）の関係のことを指す（図１）。この共引用関係は，キーワード検索では得られないような関連性を提示できるため，科学技術分野の代表的なデータベースである CiteSeerX 等いくつかのデータベースで，その関係を利用した文献検索機能が提供されている。

２．研究の目的

共引用の関係を用いた従来の検索システムの概要が図２である。図２で示されるように，検索システムの利用者は既知の適合文献をシステムに入力し，検索システムは入力された既知の適合文献と共引用関係にある文献群を出力する。したがって，共引用関係を利用した検索は，類似文書検索の一つととらえることもできる。

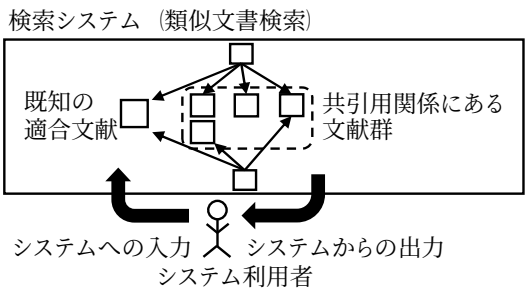

図２　従来の共引用関係に基づく検索

本研究の目的は，共引用関係を利用した検索において，その長所「キーワード検索では見つけられない適合文献を発見できる」を強化した新たな検索手法を開発することである。具体的には，共引用関係を多段階化することで，共引用関係を手がかりにして検索できる文献の範囲を拡大し，かつ拡大した範囲の中から適合文献を選び出す検索手法を開発する。

３．研究の方法

本研究において開発する検索手法は，検索キーとなる既知文献を入力した場合，従来の出力結果である「既知文献と共引用関係にある文献群」だけでなく，さらに「その文献群と共引用関係を持つような文献群」も合わせて出力することで，再現率の向上を図るものである。再現率の向上の際に問題となる検索精度の低下に対しては，共引用関係の文脈情報の利用する手法を適用し解決する。

この検索手法を「共引用関係の多段階化」，「共引用関係の文脈情報の利用」，「見つけた文献群のランク付け」の三つ部分に大別し，テストコレクションを用いた検索実験により実証的に開発をおこなう。

４．研究成果

(1)共引用関係の多段階化

従来の共引用関係では，文献 A をシステムに入力した場合，文献 B を介して文献 C までの範囲しか検索対象にしてこなかった（図３）。本研究では，従来よりも多くの適合文献を探すことができるようにするため，共引用関係の多段階化をおこなった。

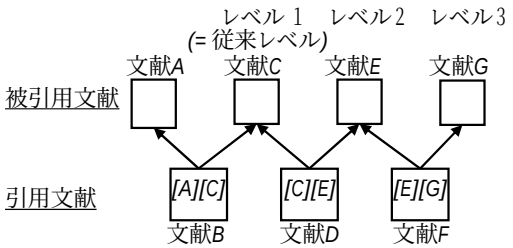

図３　多段階化された共引用関係

この多段階化は，「共引用関係にある論文」を共有する文献同士にも関連性があるという仮説に基づいている。図３の場合，文献 A と文献 C の間には文献 B を介して共引用関係があり，かつ文献 C と文献 E の間にも文献 D を介して共引用関係がある。したがって，文献 A と文献 C に関連性があり，かつ文献 C と文献 E の間にも関連性があるのであれば，文献 A と文献 E の間にも，共引用関係を拡大的に解釈した関連性があると考えることができる。

この多段階化は，図３における文献 E までの範囲にとどまらず，媒介する引用文献の数を増やすことによって，さらに拡大することができる。たとえば，文献 G は文献 F を介して文献 E と共引用関係にあるため，文献 A と文献 G の間にも関連性があると考えられる。本研究では，媒介する引用文献の数に応じたレベルの概念を定義し，共引用関係の多段階化の概念を整理した。文献 E は媒介する引用文献の数が二つ（文献 B と文献 D）であるた

めレベル２の共引用関係となり，文献 G は，媒介する引用文献の数が三つであるため（文献 B，文献 D,文献 F），レベル３の共引用関係となる。この多段階化は，従来の共引用もレベル１として扱うことができるため，自然な共引用関係の拡張といえる。

テストコレクションを使って，検索実験をおこなった結果，レベル２やレベル３の共引用関係で見つけることのできる文献群のなかにも，多くの適合文献が含まれていることが明らかとなった。

(2)共引用関係の文脈情報の利用

図３のように，単純に共引用関係を拡大していくと，多くの適合文献を検索できるようになる一方で，同時に既知文献とは関連性がないノイズ文献も多く検索されてしまう問題がある。

そこで，ノイズ文献が増える問題を軽減するために，共引用関係の文脈情報の解析結果を利用する。この解析は，引用文献の本文から，当該の引用文献の著者が暗示している二つの論文の関係性の強弱を求めるものである。図４の場合，文献 X と文献 Y は同一の段落で用いられているため，強いつながりの共引用関係，文献 Y と文献 X は，離れた箇所で引用されているため，弱いつながりの共引用関係と自動的に判定できる。

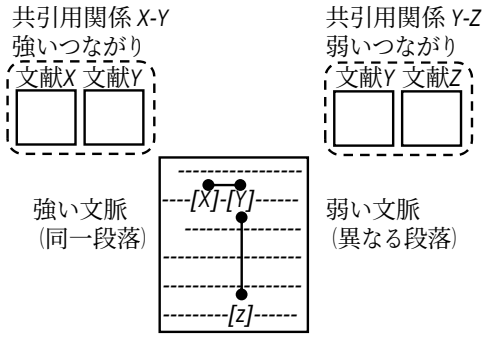

図４　文脈に基づく共引用関係

テストコレクションを用いて，単純に共引用関係を多段階化した場合と，共引用関係の多段階化をその文脈が強い場合にのみ限定した場合（たとえば，図３において，文献 C が文献 A と文献 B を同一段落で引用し，かつ文献 D が文献 C と文献 E を同一段落で引用しているなど）とを検索実験により比較した。その結果，文脈が強い場合にのみ限定して，共引用関係を多段階化した方が，検索結果に含まれるノイズ文献の割合が少なくなる傾向がみられた。したがって，共引用関係を多段階化する際に，共引用関係の文脈情報を利用することで，ノイズ文献が増加してしまう問題を軽減できることが明らかになった。

(3)見つけた文献群のランク付け

①ネットワークモデルの利用の有効性

「共引用関係の多段階化」と「共引用関係の文脈情報の利用」を適切に組み合わせ，文献群を適切にランク付けするにおいて，ネットワークモデルを採用した。図５は，被引用文献をノード，共引用関係をエッジとした無向グラフである。すなわち，文献 A と C_1 は共引用関係にあり，文献 C_1 と文献 E_1 も共引用関係にある。さらにこのネットワークにおけるエッジには共引用関係の強さに基づく重みが設定される。この共引用関係の重みは，当該の二つの文献が強い文脈で共引用された回数と弱い文脈で共引用された回数に基づいて算出される。

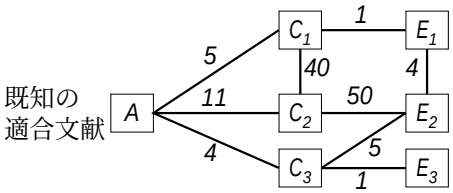

図５　重み付き共引用ネットワーク

ネットワークモデルを用いることで，既知文献を手がかりに，多段階化された共引用関係に基づいて文献を検索する課題は，グラフにおいて，特定ノードとネットワーク上にあるその他のノードとの類似度を導く問題に置き換えることができる。本研究では，この類似度を算出する際に，当該の問題に対する代表的なアルゴリズムである，Random Walk with Restart(RWR)を用いた。

テストコレクションを用いて，従来の共引用関係に基づく検索手法と開発検索手法のそれぞれで，検索をおこなった。その検索結果を検索指標に基づいて比較した結果，従来手法よりも開発手法の検索性能の方が上回った。

②共引用ネットワークに特化したランク付けアルゴリズム

ネットワークモデルの有効性を明らかにする際に用いた RWR アルゴリズムは，汎用的なアルゴリズムである。そこで，本研究の対象である「共引用関係」に特化したランク付けのアルゴリズム(Random Walk with Wait and Restart)を考案した。

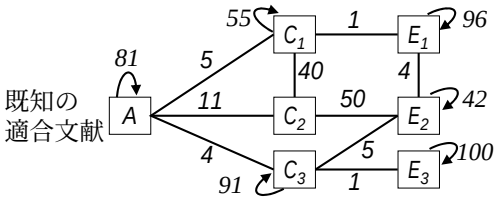

図６　滞留エッジ付きネットワーク

このアルゴリズムは，図６のように，接続しているエッジの重みの合計が最も高いノード（ノード C_2）に基づいて，グラフ上のノ

ードに滞留するエッジを加えることにより，ノードからノードへの遷移確率を補正するものである。

テストコレクションを用いて検索実験をおこなった結果，汎用的なアルゴリズムを用いる場合よりも，考案したアルゴリズムを用いた方が，検索性能が上昇する傾向がみられた。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

① Masaki Eto, Evaluations of context-based co-citation searching, Scientometrics, 94, 2, 651-673, 2013. 査読有, DOI: 10.1007/s11192-012-0756-z

〔学会発表〕（計 5 件）

① Masaki Eto, Document retrieval method using random walk with restart on weighted co-citation network, Proceedings of the American Society for Information Science and Technology Volume 51 (77th ASIS&T Annual Meeting) 2014年10月31日-11月4日, Seattle (USA), http://doi.org/10.1002/meet.201405101126

② Masaki Eto, Random walk with wait and restart on document co-citation network for similar document search, Poster Proceedings of the 8th ACM Conference on Recommender Systems (RecSys 2014), 2014年10月6日-10日, Foster City (USA), http://ceur-ws.org/Vol-1247/recsys14_poster3.pdf

③ Masaki Eto, Document retrieval method using random walk with restart on co-citation network, Workshop on Informetric and Scientometric Research (76th ASIS&T Annual Meeting, METRICS 2013). 2013年11月2日, Montreal(Canada).

④ Masaki Eto, Spread co-citation relationship as a measure for document retrieval, Proceedings of the 5th ACM workshop on Research advances in large digital book repositories and complementary media (CIKM 2012, BooksOnline '12). 2012年10月29日, Maui(USA). http://dx.doi.org/10.1145/2390116.2390121

⑤ 江藤　正己，検索システムの新たな動き－キーワード検索の限界を超えて－，学習院女子大学学会講演会，2011年11月1日，学習院女子大学(東京都・新宿区).

６．研究組織

(1)研究代表者

江藤　正己（ETO, Masaki）
学習院女子大学・国際文化交流学部・講師
研究者番号：10584807

(2)研究分担者

なし

(3)連携研究者

なし