機関リポジトリと外部情報源を連携した関連論文探索手法

NGUYEN MANH CUONG† 渡辺 陽介‡ 横田 治夫‡§

†東京工業大学工学部情報工学科 〒152-8552 東京都目黒区大岡山2-12-1

‡東京工業大学学術国際情報センター

§東京工業大学院 情報理工学研究科 計算工学専攻

1. はじめに

近年，大学，研究所などの機関リポジトリが構築され、多数の論文が蓄積されるようになってきた．蓄積された論文の有効利用のため、引用・被引用関係を利用して、関係のある論文同士を推薦することは重要である．

論文の引用情報を利用して，関係のある論文を抽出する手法として書誌結合（bibliographic coupling）[1]，共引用分析（co-citation analysis）[2]，リサーチマイニング[3]などが提案されている．これらの手法は，引用・被引用関係にある論文は関連する主題を扱っているということを前提にしている．しかし，機関リポジトリに蓄積される論文は，論文の引用・被引用情報が十分にないため，的確に関連する論文を抽出することが困難である．

本研究では，引用・被引用関係を取得するために，機関リポジトリ以外の外部情報源を使用することで，機関リポジトリに蓄積された論文の関係を発見する手法を提案し，それを実装する．ここで，外部情報源とはウェブ上にある公開論文データベースシステムを指す．関連論文の探索に関連する引用情報から探索可能な論文をすべて対象とする全探索と，枝刈りを行う引用数優先探索の2つの手法を用いる．全探索手法では可能なすべての引用論文を探索対象とするため，結果の精度が高いがコストも高い．一方，引用数優先探索手法では論文の引用数を優先度として扱い，優先度が高いもののみに対して検索を行うことにより，処理時間を短縮できる．

2. 関連論文発見

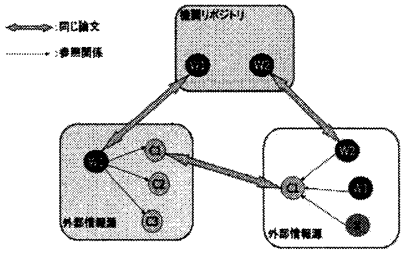

図1：アプローチ

図1を例として，関連論文を検索する手順を説明する．まず，機関リポジトリ内の対象とする論文（W1）に対して，外部情報源の引用情報を基に，その対象論文が参照している引用論文(C1，C2，C3,…)の情報を取得する．次に，外部情報源から被引用情報を利用して各引用論文(C1，C2，C3,...)に対し，それぞれの論文を引用している候補論文(X，W1，W2,...)を取得する．この候補論文の中で，対象論文と異なるものでかつ機関リポジトリにも蓄積されているもの(W2)があれば，それが対象論文と関連しているものとする．それらの関連論文に対して，書誌結合の関連度計算に基づいて対象論文との関連度を算出し，関連度が高いもののみ結果として出力する．こうすることで，機関リポジトリ内の論文間の関連性を抽出できる．

2.1.1 全探索手法

全探索手法は取得可能な論文の引用・被引用情報をすべて解析に用いる．全探索手法の検索手順を図2に示す．

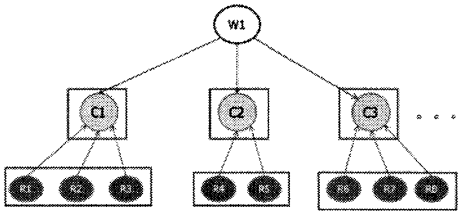

図2：全探索手法

まず，対象論文W1に対して，外部情報源から，引用情報を利用してW1が引用している論文（C1，C2，C3,...）の情報を取得する．次に，それらの引用論文の各々に対して，別外部情報源の被引用情報を利用してそれを引用している論文を検索する．図2ではC1を引用している論文のR1，R2，R3の情報を検索し，次にC2，C3,...すべてに対して同様に検索する．この手法は単純であるが，1つの対象論文に対して，外部情報源に問合せする回数が多くため，処理時間が長くなる．

2.1.2 引用数優先探索手法

全探索手法のように，すべての引用論文に対して検索を行うと，コストが高い．そこで論文の重要度を考慮して，1つのアプローチとして優先度が高いもののみに対して検索を行う手法を提案する．論文の引用数を優先度として扱うことを考える．つまり，引用されている論文の数が多いものには優先度が高い．検索した引用論文に優先度を付け，上位k件だけに対して検索を行う．

引用数優先探索手法の検索手順を図3に示す．まず，対象となった論文W1に対して，外部情報源から，引用情報を利用してW1が引用している論文（C1，C2，C3,...）の情報を取得する．次に，検索結果の引用論文に対して，優先度を付ける．優先度が高いもののみに対して，外部情報源から被引用情報を利用してその論文を引用している論文を検索する．ここでは優先度が高い論文C1とC3に対して検索を行う．優先度が低い論文

「Method of searching related papers from institution's repository using external information sources」

† 「Manh Cuong NGUYEN・Department of Computer Science, Tokyo Institute of Technology」

‡ 「Yousuke WATANABE, Haruo YOKOTA・Global Scientific Information and Computing Center, Tokyo Institute of Technology」

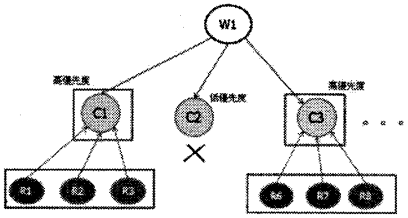

図 3：引用数優先探索手法

C2 には検索を行わない．このように，関連する可能性が高いもののみを検索し，検索論文の数を制限し，探索コストを減少する．

2.2 論文間関連度評価

全探索手法または引用数優先探索手法を用いて候補論文を取得した後，それらの論文と対象論文の関連度を計算し，関連度が閾値 t 以上のものだけを出力する．論文間の関連度を評価するには書誌結合（Bibliographic coupling）[4]を使う．2 つの論文が同じ文献を引用しているとき，関連度を 1 点追加する．2 つの論文に対して，同じ参照文献の数が多いとき関連度が高い．

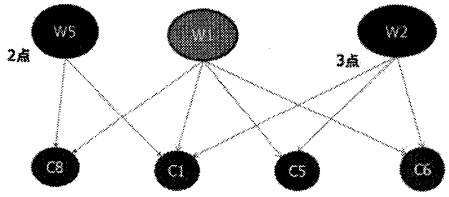

図 4：論文間関連度評価

例えば，図 4 では対象論文が W1 で，W1 と W2 が両方とも C1，C5，C6 を引用し W1 と W5 は C8 と C1 を引用している．このとき，W1 と W2 の関連度が 3 点，W1 と W5 の関連度が 2 点となる．

3. 実験

提案手法の有効性検証のために，プロトタイプシステムを実装した．プロトタイプシステムからアクセスする機関リポジトリとして東京工業大学の T2R2[5]を対象とした．外部情報源は CiNii[6]と Google Scholar[7]を用いた．プロトタイプの詳細については文献[8]を参照されたい．

今回は作成したプロトタイプシステムで全探索手法を使った検索について実験を行った．T2R2 にある 4 件の論文を実験の対象とした．各論文のタイトルで検索し，返された結果を解析した．この 4 つの論文に対して実験結果を表 1 に示す．

ここで，結果の関連論文数というのは対象論文に対して，出力されたものである．対象論文と関係しているものを T2R2 に入っているかどうかを問わず，全部出力した．この中で，実際に関連しているかどうかは目で見て判断した．関連度の閾値は t=2 とした．候補論文総合数は対象論文の引用論文の各々に対して，それを引用している候補論文の総合数（例えば図 2 では R1，R2，R3，...の総合数）である．枝刈を行うことにより，候補論文の総合数を減少できると考える．時間(s)は対象論文に対して，引用・被引用論文情報の可能なものをすべて取得し，関連度を算出して関連論文群を出力するまでの実行時間である．

表 1：実験結果

論文番号	結果の関連論文数	関連あり	正解率	T2R2 内あり	実際の引用論文数	CiNii から取った引用論文数	候補論文総合数	時間(s)
1	14	14	100%	10	35	35	293	148
2	7	6	85.7%	0	18	18	90	78
3	6	5	83.3%	3	24	21	97	66
4	3	2	66.6%	1	7	7	40	25
合計	30	27	90%	14	84	81	520	317

論文番号 3 において，CiNii から取得した引用論文数は実際の引用論文数より少ないのは，CiNii の登録内容によるものである．

4 回の実験で関連論文として合計 30 件の論文が出力された．この中で，実際に対象論文と関係するものが 27 件であり，関係しないものが 3 件であった．これにより，本研究では対象とした論文に対して，関連している文献を見つけることができた．関連文献検索結果の平均 precision が 0.9 となっている．

4. まとめと今後の課題

本研究では，引用・被引用関係を取得するために外部情報源を使用して，機関リポジトリ内の論文の関係を発見する全探索と引用数優先探索の 2 つの手法を提案した．T2R2 と CiNii，Google Scholar を利用してプロトタイプシステムを実装し，全探索手法を実現した．精度が高い関連論文検索の実現ができた．

今後の課題として全探索手法と引用数優先探索手法を用いた評価実験を行い，検索結果の精度，外部情報源問合せ回数，時間などの尺度で 2 つの手法を比較する予定である．

謝辞

本研究の一部は文部科学省科学研究費補助金特定領域研究(#21013017)の助成により行われた．

文献

[1] M. Kessler, Bibliographic Coupling between Scientific Papers, *American Documentation*, Vol. 14, No. 1, pp.10–25 (1963)

[2] 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, pp. 265–269 (1973).

[3] 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データベース, Vol. 45, No. SIG 7(TOD 22), pp. 24-32 (2004).

[4] Bing Liu, Web Data Mining, pp.245, Springer, New York, 2008.

[5] T2R2, http://t2r2.star.titech.ac.jp/

[6] CiNii, http://ci.nii.ac.jp/

[7] Google Scholar, http://scholar.google.com/

[8] Nguyen Manh Cuong, 渡辺　陽介, 横田　治夫, “機関リポジトリと外部情報源を連携した関連論文探索の実現”, DEIM 2010