人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第11回)
SIG-AM-11-09

参照関係の可視化による論文サーベイの効率化

The Visualization of Citation Information and Its Application in Literature Survey

井上絢翔[1]　韓 東力[2]

Ayato Inoue[1]　Dongli Han[2]

[1]日本大学大学院 総合基礎科学研究科

[1] Graduate School of Integrated Basic Sciences, Nihon University

[2]日本大学文理学部 情報科学科

[2] Department of Information Science, College of Humanities and Sciences, Nihon University

Abstract: 論文サーベイは学術論文の執筆において重要なタスクの一つである。必要な論文を収集する方法はいくつか存在するが、その中でも我々は論文間の参照関係に注目し、「どのような理由で参照を行っているのか」といった情報を明らかにすることで効率的に論文サーベイを行うことができるのではないかと考えた。本研究では論文間参照関係の可視化を行うことにより論文サーベイの効率化を図った。

1. はじめに

論文を執筆する上で論文サーベイは必要不可欠である。その論文サーベイによく利用されているツールとして、CiNii[1]や Google Scholar[2]、CiteSeerX[3]などの電子図書館が挙げられ、これらを使った検索手法としてはキーワード検索や論文間参照関係に注目した探し方が挙げられる。

キーワード検索はキーワードとの照合により論文を検索する方法である。キーワード検索に関する既存研究は既にいくつもの既存研究が存在する[4][5]。

一方、論文間の関連性に着目して検索を行う方法としては、大きく分けて「共参照を利用した方法」と「過去の論文をたどっていく方法」が挙げられる。共参照とは、気になる論文(以下「起点論文」と呼ぶ)と同じ論文をなるべく多く参照している論文、もしくは同じ論文になるべく多く参照されている論文は関連度が高いのではないかという考え方で収集していく方法で、こちらも既にいくつもの既存研究が存在する[6][7]。

過去の論文をたどっていく方法は起点論文を 1 つ選定し、その論文が参照している論文や、さらに参照論文が参照している論文という順にサーベイの対象を広げていくという方法である。本研究ではこの「過去の論文をたどっていく方法」に焦点を当てていく。

関連度の高い過去の論文は起点論文が直接参照しているものだけとは限らない。例えば、起点論文がある論文Ａの手法を参考にしていたとする。その論文Ａがさらに別の論文Ｂの手法を参考にしていた場合、起点論文は間接的に論文Ｂの手法を参考にしていたといえる可能性がある。既存の電子図書館である CiNii[1]や Google Scholar[2]などでも、このように起点論文が直接参照していない論文も探索対象に含みたい場合は、気になる被参照論文を一つ選んで、それを起点論文として改めておいてさらにその被参照論文をたどっていくことはできる。しかし、この方法ではたどり着いた被参照論文を全て表示すると情報が多くなりすぎてしまう。実際に、論文の関係性に着目して可視化を行った研究として清水ら[8]や渡部ら[9]の研究が挙げられるが、どちらも論文の数が多すぎて見た目が乱雑になりすぎているという問題点を挙げている。そこで我々は論文間の関係性を明らかにし、これを利用して関連論文を絞り込むことで効率的な文献検索が可能になるのではないかと考えた。

論文間の関係性を明らかにすることができれば、「被参照論文の手法を改良して利用している文献が欲しい」や「被参照論文の実験結果と比較している論文を読みたい」などといったような検索が行えるようになり、多くの論文候補から検索目的にそぐわない論文をシャットアウトすることができるように

なる。このような検索が行えるような文献検索システムの構築が本研究の最終目標である。

2. 論文間の参照関係付与

このようなシステムを作るにはもちろん論文間の参照関係の付与が必要になる。その方法を大きく分けて機械的に参照関係の付与を行うものと、手動による付与が挙げられる。前者に関するものとして難波ら[10]、小出ら[11]と Teufel ら[12]の研究がある。難波らは論文間の参照タイプを 3 種類に分類しているが、効率的な論文サーベイを行うのに分類数が不十分と思われる。小出らと Teufel らの研究では、それぞれ 9 種類と 12 種類の参照理由を定義し機械学習を用いて参照理由を付与しているが、精度は最大で 60%～70%台に留まっている。論文間関係の解明を最終目標とするような研究では上記の精度でも一定の有効性があるかもしれないが、自動付与された参照理由を異なる目的で再利用する場合には、連鎖的誤りを回避するためにはより正確な分類結果が必要であろう。

それに対して手動付与ではより高精度のアノテーションを行うことができるが、時間がかかることや論文サーベイに精通している専門家を雇うのに多大なコストがかかることなどが問題点としてあげられる。

そこで我々はクラウドソーシングを利用することによりコストの問題に対処できるのではないかと考えた。クラウドソーシングの既存のサービスとしては yahoo クラウド[13]やランサーズ[14]などが存在している。これらのサービスはインターネット上の不特定多数の作業者に仕事を依頼する雇用形式で、低コストで迅速な作業が可能である。また、クラウドソーシングでは主にアンケート調査やデータ入力などの単純な業務が多い。それに対して論文間参照情報のアノテーションは比較的難易度の高いタスクである。そのため不特定多数の作業者がどの程度遂行できるのか、また専門家と比べるとどのような差があるのかなど大きな不安が挙げられる。

そこで我々は上記の懸念を念頭に、まずは論文間参照情報をアノテーションするためのプロトタイプを構築した。次に、論文サーベイに精通している大学教員を専門家に、大学生をクラウドソーシングで働く一般作業者に見立て、構築されたプロトタイプを利用してアノテーションしてもらった結果を比較した。この過程を通じて論文間参照情報のアノテーションにクラウドソーシングを利用する可能性の検討を行った[15]。

アノテーションのタグに関しては図 1 のようなタグの階層および種類を利用した。既存研究では 9~12 種類の参照理由が定義されていたが、いずれも単一階層で構成されているので、アノテーションやそれを利用した論文検索が容易ではないという問題がある[11][12]。本研究では図 1 のように 3 階層構造にし、少ない選択肢を複数回与えることで検索やアノテーションの負担を減らすことを目指した。

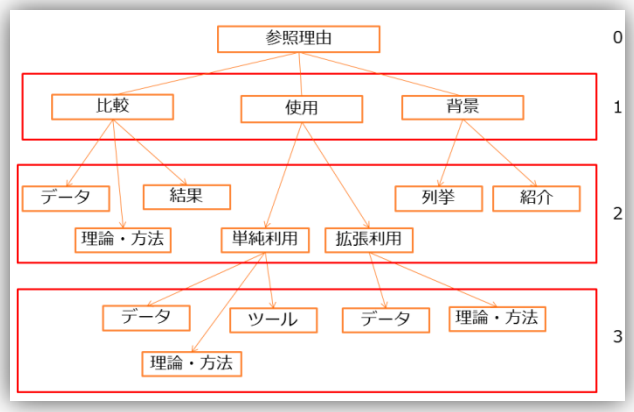

図 1 論文間参照関係の分類

実験に使用したデータは「言語処理学会年次大会発表論文集」に掲載された論文で、本文が日本語で書かれているものに限定した。

実験結果により、いくつかの課題が残ったものの、論文サーベイにそれほど精通していない一般作業者でも専門家に近い、良質なアノテーションを行う可能性が十分あることが示唆された。

3. 可視化システム

今までの研究で論文間参照情報のアノテーションにクラウドソーシングを利用することに関して、良質なアノテーションを行う可能性が十分あることが分かった。そのため本研究では十分なアノテーションが行われたものと仮定して、論文間参照情報を利用した可視化システムを構築する。3.1 では大まかなシステムの流れ、3.2 では論文間参照情報データベース、3.3 ではインターフェース・機能に関してそれぞれ説明していく。

3.1. システムの流れ

システムの流れは図 2 のようになっている。まずユーザが起点論文を選択し、その起点論文が参照している論文の情報(論文タイトル・著者等)や、どのような理由で参照を行っているのか(以下「参照理由」)などの内容を論文間参照情報データベースで検索をする。そして検索でヒットしたデータをもとに可視化を行い、ユーザに提示していく。単純にヒットした物を提示するだけでは論文候補が雑多になっ

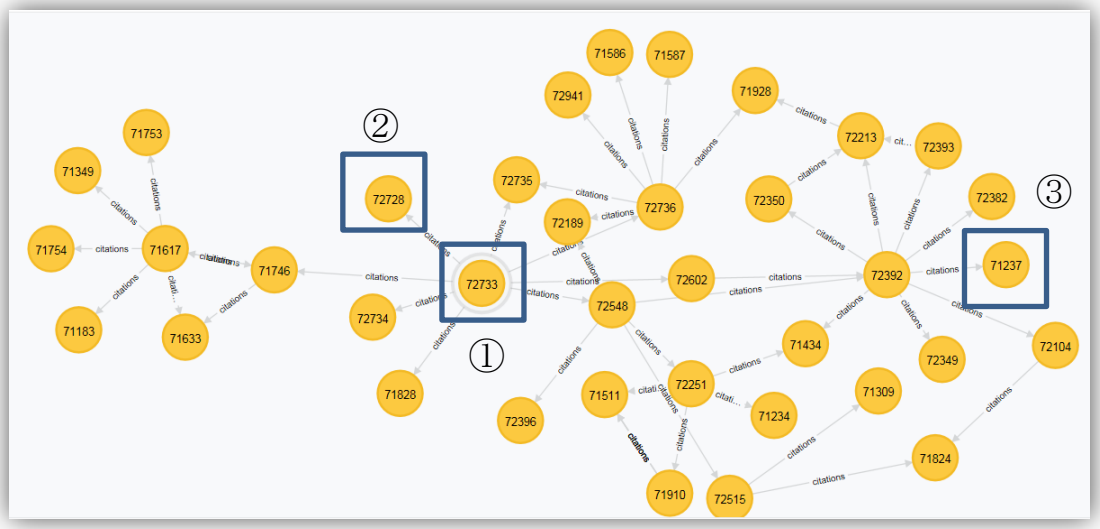

図 3 Neo4j ウェブインターフェース実行画面

てしまうので、その場合はユーザに表示したい参照理由を選択してもらい、フィルタリングをかけることで必要な論文だけを表示する。

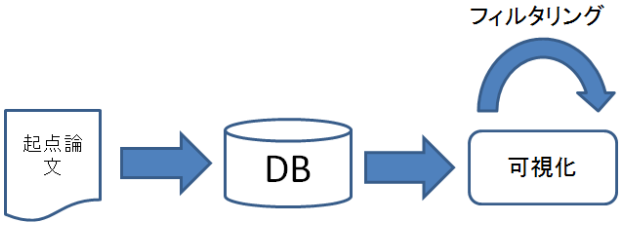

図 2 システム構成図

3.2. 論文間参照情報データベース

本研究では主に論文のタイトルや著者等の「論文情報」と、どの論文がどの論文を参照しているのか、どのような理由で参照を行っているのかといったような「論文間の参照情報」の 2 つの情報を扱う。これらのデータベース作成はグラフデータベースである neo4j[16]を利用して構築した。

neo4j ではグラフ構造のデータを扱うことができ、「ノードとノードがどのような関係性で結ばれているのか」といったような表現でデータを格納でき、本研究のように「論文と論文がどのような関係で結ばれているのか」といった情報を取り扱う場合には最適であると思われる。

図 3 は neo4j に実装されているウェブインターフェースの実行画面で、実際に構築した論文間参照関係のデータベースを使用している。図 3 の①は起点論文を表している。②は距離 1 の論文で起点論文が直接参照しているものである。③は距離 3 の論文で、今回は距離 3 までに限定して論文を表示しているため、終端ノードとなっている。

3.3. インターフェース・機能

図 4 に構築している可視化システムのインターフェースを示す。

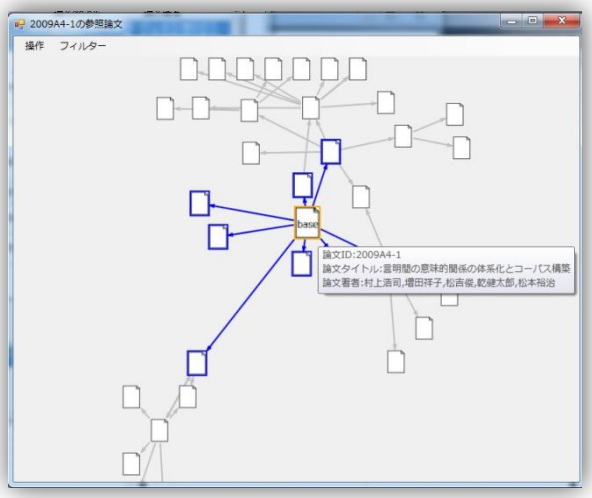

図 4 参照関係の可視化

図 7 参照情報

図 4 では起点論文が参照している論文、そしてさらにその論文が参照している論文、といったように参照をたどり、3 つ先まで参照関係を可視化している。論文やエッジが重なって見づらくなってしまった場合でも、マウス操作で論文の位置を移動することができるようになっている。今回のように表示される論文の数が多い場合には参照理由によるフィルタリングが効果的である。

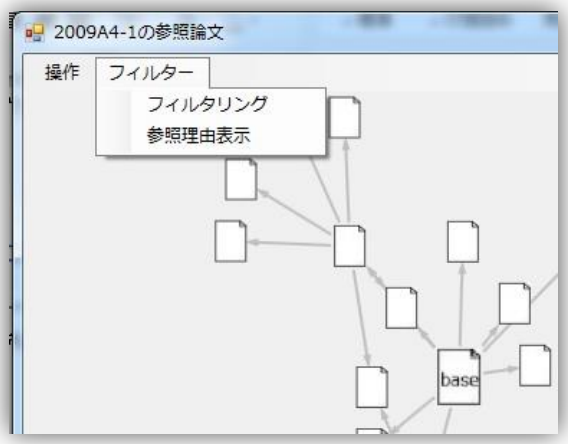

図 5 メニューバー

フィルタリングの実行と参照関係に関する情報の表示は図 5 のようにメニューバーから行えるようにした。そしてフィルタリングの設定画面は図 6 である。フィルタリングの設定方法は図 1 に示した階層型参照理由をもとに考案した。

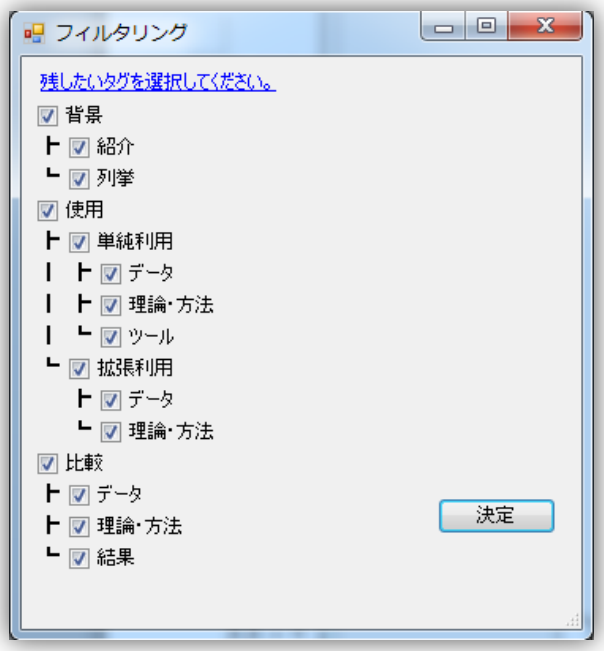

図 6 フィルタリング

このように階層式を導入することで、「背景を全て表示」や「データの比較だけを表示」といったような形でフィルタリングを変化させることにより、改めて再描写することができるようになっている。フィルタリングの対象は起点論文が直接参照している

被参照論文だけでなく、間接的に参照している被参照論文も全て含んでいる。これにより、1 章で述べているような「間接的に手法を参照している論文」等の探索も容易にできるようになっている。

図 7 は参照関係に関する情報で、参照理由や実際に論文の参照を行っている部分の情報などが表示される。これにより参照理由を確認し、最終的な必要な論文の取捨選択がしやすくなるようにした。図 7 の参照理由情報では情報量が多すぎる場合でも、被参照論文を一つ選択し、起点論文との関係性を最短距離で求めることができる。また、図 8 のように起点論文と興味を持った被参照論文との間にある参照理由のみが表示されるため、必要のない情報をシャットアウトすることができる。

図 8 起点論文との関係性表示機能

4. 評価実験

研究手法の有効性を検証するため客観評価、主観評価を行った。それぞれを順に述べる。

4.1. 客観評価

客観評価としては実際にフィルタリングシステムを利用することで検索タスクに対してどれだけ作業効率が上がるのかを評価した。検索タスクは 3 つ用意し、それに該当する解答を手作業で作成した。そして検索タスクに最適であると思われる参照理由でフィルタリングをかけることで検索効率を調査した。サーベイ検索効率は以下の式で評価した。

$$\frac{\textit{検索目的に合致した論文数}}{\textit{被参照論文数}}$$

また、タスクに利用した起点論文は以下の 3 編である。

1. 阿辺川武, 影浦峡:下訳から修正訳への訳文修正要因の分析, 言語処理学会第 14 回年次大会 pp. 253-256. (2008)
2. 大平真一, 山本和英:保険関連文書を対象とした校正支援システム, 言語処理学会第 18 回年次大会 pp. 243-246. (2012)
3. 大野潤一, 柴木優美, 山本和英 : Wikipedia のエントリ-リダイレクト間を対象にした同義関係抽出, 言語処理学会第 17 回年次大会 pp. 296-299. (2011)

探索は起点論文から 3 つ先までの参照関係を利用し、検索された被参照論文の数は論文 1、論文 2 が 14 編で、論文 3 が 24 編である。は論文 1 については「類似研究として翻訳を行っている研究」、論文 2 は「どのような理論や手法が利用されているのか」、論文 3 は「同義語抽出を行っている類似研究」をそれぞれタスクとして定めた。フィルタリング時に選択したタグは表 1 の通りで、1 が「表示」で 0 は「非表示」である。

表 1 フィルタリング時の選択したタグ

フィルタリングタグ			論文 1	論文 2	論文 3
背景		紹介	1	0	1
		列挙	0	0	0
使用	単純利用	データ	0	0	0
		理論･方法	1	1	1
		ツール	1	1	1
	拡張利用	データ	0	0	0
		理論･方法	1	1	1
比較		データ	0	0	0
		理論･方法	1	1	1
		結果	1	1	1

表 2 は客観評価のサーベイ検索効率を表している。フィルタリング前に比べて全体的に検索効率は向上しているのがわかる。特に起点論文 1 では 40%も向上していて効果が大きく表れている。フィルタリング後の正解個数の表示に関しては、起点論文 1 で 6 個中 5 個、起点論文 2 では 3 個中 3 個、起点論文 3 では 14 個中 10 個表示できた。表示個数を減らすのが目的なので論文を多く消しすぎてしまうことも懸念していたが起点論文 1 では 1 つを除いて全て、起点論文 2 では全て表示できていたという結果が得られた。

表 2 客観評価のサーベイ検索効率

	サーベイ検索効率	
	フィルタリング前	フィルタリング後
起点論文 1	43%	83%
起点論文 2	21%	30%
起点論文 3	58%	59%

今後の予定として、今回使用した解答やフィルタリングの内容に対する信憑性に疑問が残るため、今後それらの妥当性を検証していく。

4.2. 主観評価

主観評価では同じ大学で情報科学を専門とする学部生 10 人に本システムを使用してもらい、その使用感をアンケート調査することで評価した。評価はフィルタリングの実行・参照理由の表示が行えない物(以下システム 1)と、フィルタリングの実行・参照理由の表示が行える物(システム 2)との比較で行った。表 3 と表 4 はそれぞれシステム 1、システム 2 のどちらの方が使いやすかったか、またどちらの方が検索の効率が上がったと思うかのアンケート結果である。

表 3 どちらの方が使いやすかったかアンケート

	回答者数
システム 1 の方が使いやすい	1
システム 2 の方が使いやすい	8
変わらない	1

表 4 どちらの方が効率が上がったかアンケート

	回答者数
システム 1 の方が効率が良い	1
システム 2 の方が効率が良い	9
変わらない	0

表からわかるとおり、ほとんどの被験者からフィルタリングシステム・参照理由を利用してのシステムの方が使いやすい、効率が上がったという評価が得られた。

「システム 1 の方が使いやすかった」、または「変わらない」と答えた人の意見としては参照理由のタグの種類がよくわからないというものがあった。またシステム 1 の方が効率が上がったと答えた人に関してはフィルタリングシステムがうまく扱えず、関連性が高い論文までシャットアウトしてしまうという事が起きたのが原因ではないかと考えられる。これらの問題点の対策としてはチュートリアルを充実させることや、システム中の説明を増やす等のことが挙げられる。

その他の意見としては参照理由を可視化画面に表示してほしい、論文のアイコンが起点論文以外すべて同じなのでどの論文を見ていたのかわからなくなった等があった。これらのような UI の問題点は今後システムの改良の際に考慮する予定である。

5. まとめ

本研究では論文サーベイの効率化のための参照情報の可視化システムを構築し、その評価を行った。

実験結果から論文サーベイの検索効率向上を図れた他、システムに関するアンケートでは本システムを利用することで論文サーベイが効率的に行えたという意見を多く得ることができた。しかし、客観評価で用いた解答やフィルタリングの内容が本当に正しいのかどうかという信憑性に対しての疑問が残っているため、今後は解答の妥当性を検証していく予定である。その他、主観評価からの実験で得たアンケートからシステムの改善点などが意見として寄せられているため、これらに対応していくのも今後の課題である。

参考文献

[1] http://ci.nii.ac.jp/
[2] http://scholar.google.co.jp
[3] http://citeseerx.ist.psu.edu/
[4] R. Baeza-Yates, B. Ribeiro-Neto. “Modern Information Retrieval” (1999)
[5] K. Dobashi, H. Yamauchi., R. Tachibana. “Keyword Mining andVisualization from Text Corpus for Knowledge Chain Discovery Support”,Technical Report of IEICE, NLC2003-24, pp.55-60. (2003) (in Japanese)
[6] 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)
[7] M. Kessler. “Bibliographic Coupling between Scientific Literatures”,Journal of the American Documentation, Vol.14, No.1, pp.10-25, (1963)
[8] 清水 成昭,竹中 豊文: 文献の参照関係を視覚化するアプリケーションの提案・実装,電子情報通信学会技術研究報告. IN,情報ネットワーク 109(449), 389-394, (2010)
[9] 渡辺 秀文,北川 晴香,齋藤 隆文: 文献の参照関係の可視化,情報処理学会 研究報告グラフィクスと CAD (CG) 2010-CG-139(6), 1-6,(2010)
[10] 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,情報通信学会論文誌,42(11), pp.2640-2649. (2001)
[11] 小出寛史, 韓東力: 論文間参照情報のデータベース化に基づく参照タイプの同定,自然言語処

理研究会報告 2012-NL-209(2), 1-7(2012)
[１２] Teufel, S.: The Structure of Scientific Articles -Applications to Citation Indexing and Summarization. CSLI Publications. (2010)
[１３] http://www.lancers.jp/
[１４] http://crowdsourcing.yahoo.co.jp/
[１５] 井上 絢翔,韓 東力: 論文間参照情報のアノテーションにおけるクラウドソーシングの利用検討, 言語処理学会 第 21 回年次大会 pp.736-739.(2015)
[１６] http://neo4j.com/