様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年６月１８日現在

機関番号：６２６０８
研究種目：挑戦的萌芽研究
研究期間：2009～2011
課題番号：２１６５００５２
研究課題名（和文）　ソーシャルネットワークを利用した書誌マイニングに関する研究

研究課題名（英文）　Discovering bibliographic records with Social Network

研究代表者
野本　忠司（NOMOTO TADASHI）
国文学研究資料館・研究部・准教授
研究者番号：20321557

研究成果の概要（和文）：本研究では、共著者ネットワークを使ってOPACのランキング精度を改善する手法について述べる。ウェブページと異なり書誌データは検索に利用できる情報が極めて限られるため、技術的進歩から取り残されてきた。本研究は、OPACに内在する情報、日本十進分類体系と学術コミュニティーなどのソーシャルネットワークを併用することで、書誌ランキングの精度を改善できることを示す。

研究成果の概要（英文）：This work will introduce a new approach to ranking bibliographic records in library search, which is currently dominated by an OPAC style search paradigm, where results are typically not ranked by relevance. The approach we propose in the work provides the user with the ability to access bibliographic records in a way responsive to his or her preferences, which is essentially done by looking at a community or a group of people who share interests with the user and making use of their publication records to re-rank search results. The experiment found that the present approach gives a clear edge over conventional search methods.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	900,000	0	900,000
２０１０年度	600,000	0	600,000
２０１１年度	600,000	180,000	780,000
年度			
年度			
総　計	2,100,000	180,000	2,280,000

研究分野：情報学
科研費の分科・細目：情報学、図書館情報学・人文社会情報学
キーワード：OPAC / テキスト自動分類 / 図書分類法 / 図書館情報 / 情報検索 / 機械学習 / 関連性ランキング

１．研究開始当初の背景

本件代表者が所属する国文学研究資料館では、国文学に関わる、画像、本文(テキスト)、書誌関係、約３０本のデータベースを現在運用しているが、検索数から見たデータベースの利用数は、その９割以上が書誌関連データベースに集中している(平成２０年計画申請当時)。このことから国文学における書誌情報に対する強い検索需要があることが分かる。

他方、国立国会図書館や国立情報学研究所では、それぞれ NDL-OPAC, Webcat などを通して国文学を含む書誌情報の統合化を進めて

いる。その結果、近年ではインターネットで簡単に目録情報を入手できるようになってきた。

その一方で、多くの地方図書館は、インターネットで情報を提供しているにも関わらず、全国レベルの統合化の動きから外れているため、インターネットからのアクセスが極めて難しくなっている。地方図書館には国文学や歴史関連の文献情報が多々蓄積されていると見られるが、その全容は明らかではなく現実的にこれらを網羅的、横断的に調査する方法はいまのところ存在しない。

２．研究の目的

本研究では、国文学に注目した書誌の発掘と集約に向けた試みとして、人物名に基づくソーシャルネットワークを活用した書誌収集手法を提案し、その有効性を、OPAC を使って検証する。

３．研究の方法

本件課題を解決するには、

(1) 人物名の収集
(2) OPAC による書誌検索と国文学との関連性ランキングによる検索結果のフィルタリング

が必要となる。今回の研究では、特に(2)に重点を置いて、研究を実施した。(1)については、国会図書館の全国書誌データ、ウィキペディア、国文学研究資料館の典拠データベースなどからの直接採取、あるいは米カーネギーメロン大の NELL (Never Ending Language Learner)の自動採取法を利用して、数年かけてインターネットからしらみつぶしに採取するという方法が考えられるが、今回の研究では、資金的制約、また３年間という時間的な制約を考慮して、(1)については将来の課題とすることにした。

前述の(2)について説明する。まず、書誌検索、特に OPAC における関連性ランキングの歴史的背景について述べる。

オンライン蔵書目(OPAC) は 1980 年代に本格実用化され、現在では全国のほとんどの公共・大学図書館に導入されている。しかし、30 年近く経た現在においても、ウェブサーチでは当然のように備わっている関連性ランキングが未だに欠落しているという大きな問題を抱えている。例えば、図 1 は国会図書館 OPAC での検索の実例を表しているが国文学関連書誌が最下位に現れている。

	書誌タイトル（検索語:解釈 総数 200 件）	国文研究への関連（1:ある 0:なし ?:不明）
2	アインシュタインとの論争	0
3	赤松俊秀教授退官記念国史論集	0
4	阿部次郎全集. 第4巻	0
5	アメリカ大陸の奴隷制：南北アメリカの比較論争	0
6	石原謙著作集. 第2巻	0
7	市河博士還暦祝賀論文集. 第3輯	0
8	市邨学園大学開学記念論集	0
9	岩波講座現代法. 第15	0
10	岩波講座国語教育. 第2巻	0
11	大阪府会資料. 第7巻	0
12	沖縄久米島の総合的研究	0
13	小倉進平博士著作集. 2	0
14	親教育と家族心理学	0
16	川端康成の文学：『解釈』所収論文集. その1	1
17	川端康成の文学：『解釈』所収論文集. その2	1

160	法学概論	0
161	法学協会五十周年記念論文集. 第1,2部	0
162	法学・政治学の動向：北海学園大学法学部20周年	0
163	封建制と資本制：野村博士還暦記念論文集	0
164	法と政治：21世紀への胎動 九州大学法学部創立七	0
165	法の理論. 10	0
166	法理学の諸問題：加藤新平教授退官記念	0
167	法律評論創刊十週年記念論文集	0
168	法華経の中国的展開	0
169	穂積先生還暦祝賀論文集	0
170	穂積先生還暦祝賀論文集	0
171	毎日宗教講座. 第1	0
173	万葉集講座. 第2巻	1
174	万葉集大成. 第2-3巻	1
175	万葉集大成. 第3巻	1
176	万葉集の伝承と創造	1

図1 国会図書館 OPAC をキーワード「叙説」で検索した結果。ユーザは国文学関連資料を期待。

このような中、ウェブサーチの社会への急速な浸透から、OPAC に同等の機能を望む気運か高まっており、TFIDF に基づく関連性ランキングを取り込んだ OPAC システムも登場してきている。他方 OPAC の使い勝手の悪さは、変化を望まない図書館司書の価値観に原因があると指摘する声もある。

書誌検索は、文書検索、ウェブ検索とは異なり付随する情報が極めて少ないところに大きな特徴かつ問題がある。しかし、これは書誌検索に固有というわけではなく、曲目検索、ビデオ検索、商品検索など、いわゆるメタデータ検索一般に当てはまる現象である。

メタデータ検索では、協調フィルタリングが有効であることが知られている。メタデータ自体に使える情報がなくても、アクセス回数や購入行動のパターン、ソーシャルネットワークなど、データ外の情報を参照することで対象を精度よくランクできることがある。

このような背景のもと、本件では共著者ネットワークを利用して、ユーザの関心を図書分類体系の分布として表して、検索結果を再ランクすることで、国文学など特定分野に合致した書誌を優先的に提示する手法を提案し

た。本手法は OPAC が出力した書誌結果の分類コードを手がかりに、ランク付けしようというものである。

具体的には、以下のモデルを使って、書誌の関連度を計算する。

$$\mathcal{R}(r;\lambda,Z) = \lambda P(L(r)|\mathrm{COP}(\mathbf{C})) + (1-\lambda)P(L(r))|\mathrm{PUP}(Z))$$

上式の右辺第二項をユーザモデル、第一項を、コミュニティーモデルと呼ぶ。それぞれ、ユーザの関心の指向、コミュニティーの関心の指向をモデル化したものである。

ユーザモデルは、ユーザ自身の発表論文の題目を使って、以下の手順で構成する。(1) 論文の題目から､1から3単語グラムを抽出し、それぞれを検索キーワードとして NDL-OPAC で検索する。(2) 検索結果リストにある書誌から日本十進分類コード（以下、NDC）を取り出す。(3) 検索キーワードを取り出した NDC 集合のまとまりの良さ（エントロピーの小さい）順にランク付けをして上位キーワードに現れた NDC の出現頻度のベクトルを構成する。

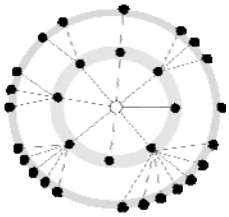

図2 共著ネットワーク. ノードは著者を表す。ノード間の長さは関係の強弱を表す。

この頻度ベクトルをもとにディレクレ分布で平滑化した多項分布がユーザモデルとなる。

ちなみに、日本十進分類法（大分類）は、総記（000)、哲学（100)、 歴史（200)、 社会科学（300)、自然科学（400)、 技術・工学（500)、 産業（600)、芸術・美術（700)、言語（800)、文学（900)で構成されている. 本稿では、上位三桁までのコードを用いた。

一方、コミュニティーモデルは、ユーザモデルを補完（バックオフ）するために導入したものであり、以下のように構成する。ウェブ上の学会、研究組織・機関のサイトから役員・職員名簿を抽出し、名簿に現れる氏名を検索キーにして NDL-OPAC で検索する。さらに役員・職員氏名から直接得られる書誌情報だけではなく、共著関係にある著者を再度検索キーにして、書誌検索を行い、書誌リストを拡張する。このプロセスを何回か繰り返したのち、得られた書誌リスト中の NDC の頻度を調べ、ユーザモデルと同様に頻度ベクトルを作る。本件では、エッジ距離１までの共著者の著作リストを考慮する（図２参照)。このようにして得られたヒストグラムの例を以下に示す（図３)。横軸は日本十進分類コード。縦軸は頻度の割合。左上から時計回りに、国際日本文化研究センター、日本近世文学会、日本言語学会、国立民族学博物館の頻度分布を表す。この図は、コミュニティーごとに扱うトピックが顕著に異なることを示している。文学系コミュニティー（日本文化、日本近世）は共に 200 番台 900 番台に大きなピークを持つ。ユーザモデルと同様、これらヒストグラムから構成した多項分布が、コミュニティーモデルとなる。

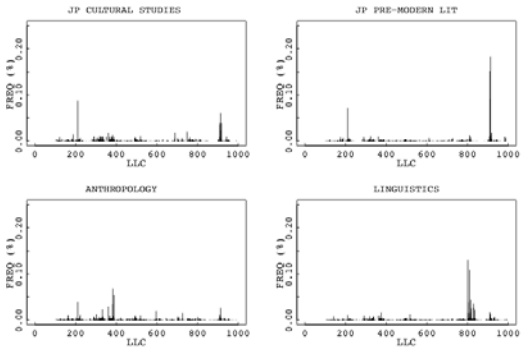

図3 分類コードの分布

４．研究成果

以下は、国文学関連書誌のランキングの精度を被験者毎に MAP（mean average precision）で比較したものである。PUP/COP は本件提案手法、その他はベースライン、特に NDL は国会図書館ＯＰＡＣを示す。 SG, EZ, OO, YZ は被験者を指す。KER, LM は当ランキング法が参照するコミュニティーモデルの選択法を指す。KER は多項式カーネル、LM は言語モデルを示す。

表1 提案手法の精度

SG						
	PUP/COP	PUP	COP	NDL	COS	ESA
KER	0.4377	0.3950	0.3754	0.1919	0.1901	0.2602
LM	0.4380	0.3784	0.3355	0.1919	0.1901	0.2602

EZ						
	PUP/COP	PUP	COP	NDL	COS	ESA
KER	0.4432	0.4060	0.4224	0.2512	0.2873	0.2745
LM	0.4108	0.4062	0.4066	0.2512	0.2873	0.2745

OO						
	PUP/COP	PUP	COP	NDL	COS	ESA
KER	0.5081	0.4108	0.4279	0.1840	0.2324	0.1674
LM	0.5051	0.4216	0.4292	0.1840	0.2324	0.1674

YZ						
	PUP/COP	PUP	COP	NDL	COS	ESA
KER	0.7554	0.7486	0.7263	0.4741	0.4780	0.4975
LM	0.7554	0.7490	0.7263	0.4741	0.4780	0.4975

いずれの被験者においても、本件手法が現在の国立図書館ＯＰＡＣに比べて精度が 2 倍近く向上していることが分かる。このことは共著ネットワークによる書誌分類が可能であることを示しており前述(2)に対するソリューションとなっている。

関連性ランキングにおける将来の方向としては、文学以外のドメインでの検証、学術コミュニティーの自動発見、ユーザの探索行動、購買行動、ソーシャルメディアにおける人間関係をもとにした個人的な興味、関心、性癖の同定とそれらを利用したランキングの高精度化などが考えられる。

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３件）

Tadashi Nomoto (2012) Re-ranking Bibliographic Records for Personalized Library. In Proceedings of the ACM / IEEE-CS Joint Conference of Digital Libaries (JCDL 2012), pp 125-128.

野本忠司、（2011）、共著者ネットワークによる書誌検索の高度化、第 17 回言語処理学会年次大会論文集（電子出版のためページ番号なし）

Tadashi Nomoto (2009) Classifying Library Catalogue by Author Profiling. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 644-645.

６．研究組織
(1)研究代表者
野本　忠司（NOMOTO TADASHI）
国文学研究資料館・研究部・准教授
研究者番号：20321557

(2)研究分担者
相田　満（AIDA MITSURU）
国文学研究資料館・研究部・准教授
研究者番号：00249921

(3)連携研究者
なし