共著者ネットワークによる書誌検索の高度化

野本 忠司
国文学研究資料館
nomoto@acm.org

1　はじめに

オンライン蔵書目録 (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 [1, 3, 4, 5, 6]．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，関連書誌が最下位に現れている．

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が高まっており，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，ないしは問題がある．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 [2].

このような背景のもと，本稿では、共著者ネットワークを利用し，ユーザの関心を図書分類体系の分布として表し OPAC の検索結果を再ランクすることで，その精度を改善する手法を提案する．文学系のドメインについて現在の国立国会図書館 OPAC（以下，NDL-OPAC）との比較を行い，本手法の性能を評価する．

詳細は後述するが，本手法は OPAC が出力した書誌の結果リストの分類コードを手がかりに，基本的に以下の尺度でランク付けしようというものである．

書誌の重要度　=　ユーザの関心との関連度
　　+　ユーザと繋がりのあるコミュニティーの関心との関連度

さらに，ユーザの関心との関連度，コミュニティーとの関連度を定義するため，それぞれユーザ・プロファイル，コミュニティー・プロファイルという概念を導入する．特に本稿では，ユーザが自分の専門分野に関連する書誌を OPAC を使って探す，というシナリオで話を進める．

2　ユーザ・プロファイル

ユーザ・プロファイルは，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 $\mathbf{y} = (c(000), \ldots, c(999))$ を構成する．このベクトルをユーザ・プロファイルと呼ぶ．

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．

3　コミュニティー・プロファイル

コミュニティー・プロファイルは，ユーザ・プロファイルを補完（バックオフ）するために導入する．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

	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）
1	アインシュタインとの論争	0
2	赤松俊秀教授退官記念国史論集	0
3	阿部次郎全集. 第4巻	0
4	アメリカ大陸の奴隷制：南北アメリカの比較論争	0
5	石原謙著作集. 第2巻	0
6	市河博士還暦祝賀論文集. 第3輯	0
7	市邨学園大学開学記念論集	0
8	岩波講座現代法. 第15	0
9	岩波講座国語教育. 第2巻	0
10	大阪府会資料. 第7巻	0
11	沖縄久米島の総合的研究	0
12	小倉進平博士著作集. 2	0
13	親教育と家族心理学	0
14	[illegible]	[illegible]
15	[illegible]	[illegible]
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
172	[illegible]	[illegible]
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「叙説」で検索．ユーザは日本文学関係資料を意図．左の番号が出現順序を表す．

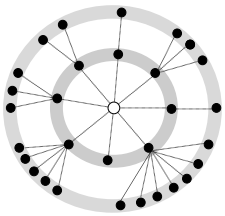

図 2: 共著者ネットワーク．中心のノードが注目著者．エッジが共著関係を表す．

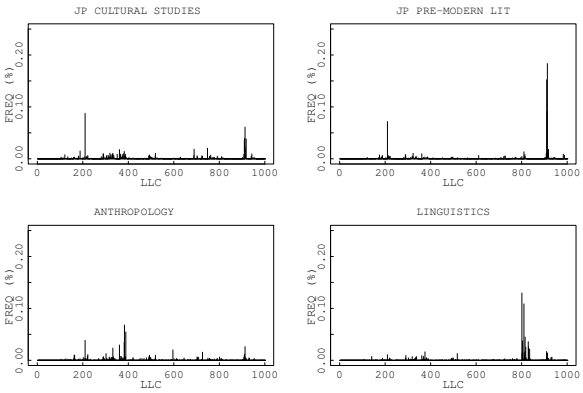

図 3: コミュニティー・プロファイル．左上から時計回りで，「日本文化研究センター」，「日本近世文学会」，「日本民族学博物館」，「日本言語学会」．横軸は NDC．縦軸は頻度の比率．

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，書誌リストを拡大する．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・プロファイルと同様に頻度ベクトルを作る．これを，コミュニティー・プロファイルと呼ぶ．本稿では，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コミュニティー・プロファイルの例を図 3 に示す．コミュニティーごとに扱うトピックが異なることが視覚的に確認することができる．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．

4　関連性モデル

次にプロファイルを用いて書誌レコード r の OPAC 検索後の関連性を以下のように定める．

$$\mathcal{R}(r) = \quad P(L(r)|\mathcal{A}(u)) + (1 \qquad)P(L(r))|\mathcal{A}(\mathbb{C})) \quad (1)$$

但し，

$$P(x|\mathbf{y}) = \mathsf{E}_x[g(\ \ |\mathbf{y})] \quad (2)$$

ここで，$\mathcal{A}(u)$ はユーザ・プロファイル，$\mathcal{A}(\mathbb{C})$ は，コミュニティー・プロファイル，g をディレクレ分布の密度関数，$P(x|\mathbf{y})$ をディレクレ事後確率 ($Dir(\mathbf{y})$) のもとでの p_x の期待値とする．但し，　$= (p_0, \ldots, p_m)$．p_j は分類番号 j (0　j　999) がプロファイルに出現する確率を表す．　はユーザ・プロファイルとコミュニティー・プロファイルへの重みの配分をコントロールする変数を表す．

5　コミュニティーの選択

ウェブ上には，同業でも数多くのコミュニティー存在し，バックオフモデルとしてどれが適切か直ちには判断できない．このため，コミュニティーをなんらかの方法で選択することが必要になる．本稿では，ユーザ・プロファイルとコミュニティー・プロファイルを以下の尺度を用いて計測し，二者間の距離に基づき最適なコミュニティーを選択するアプローチを採用する．

なお，以下で，$D(x||y)$ は，KL ダイバージェンスを

表す．q, r は確率分布で, それぞれ，ユーザモデル，コミュニティーモデルを表す．avg(q, r) は， q, r を平均した分布．

双方向 KL ダイバージェンス (Symmetric KL Divergence).

$$SKL(q, r) = D(q||r) + D(r||q) \tag{3}$$

ジェンセン・シャノン ダイバージェンス (Jensen-Shannon Divergence).

$$JS(q, r) = \frac{1}{2}\left[D(q\|\text{avg}(q, r)) + D(r\|\text{avg}(q, r))\right] \tag{4}$$

L1 ノルム (L1 Norm). LCC は，000 から 999 までの NDC の集合．

$$L1\,(q, r) = \sum_{t \in LCC} \Big|q(t) \quad r(t)\Big| \tag{5}$$

残差平方和 (Residual Sum of Squares).

$$RSS(q, r) = \sum_{t \in LCC} (q(t) \quad r(t))^2 \tag{6}$$

多項式カーネル (Polynomial Kernel)．後述の実験では，$d = 2, c = 0$ とした．

$$KER(q, r) = (\mathbf{q}^T\mathbf{r} + c)^d \tag{7}$$

言語モデル．

$$LM(d_1, d_2) = \prod_{t \in LCC} p(t \mid d_2)^{c(t, d_1)} \tag{8}$$

$c(t, d_1)$ は d_1 における分類コード t の頻度，$p(t|d_2)$ は，d_2 のもとでの t の生起確率（最尤値）を表す．

コミュニティーは，節 2 で述べた手順でユーザの文献リストから生成したユーザ・プロファイルを用いて，上記尺度において，もっとも近いものを選択する．無論，選択されるコミュニティーは，尺度によって異なる場合がある．我々の興味はどの尺度を用いたとき，ランキングの精度が最も高くなるかという点にある．これを以下で確認することにする．

6 実験と結果

実験では，コミュニティーとして以下の機関，学会を用いた．日本言語学会 (141) , 中世文学会 (28)， 和漢比較文学会，国立民族学博物館 (58)，国際日本文化センター (20)，和歌文学会 (68)，国立国語研究所 (34)，国文学研究資料館 (30)．括弧内数字は，収集した人名数．なお，人名は，各機関，学会のホームページに掲載されている役員名簿，職員録から手作業で抽出した．

コミュニティー・プロファイルは，節 3 で述べた方法で構成した．すなわち，ぞれぞれの人名について NDL-OPAC で書誌検索を行い，その著書および共著者の著書の書誌情報を集め，対応する分類コードを集積した．

さらに，大学院生を含む日本文学を専門にする研究者 4 人に，40 から 80 の検索クエリに対して NDL-OPAC が出力した検索結果リストを自身の専門分野との関連性で適合・不適合の判定をしてもらった.[1] 1 人が判定を行った書誌数は，多い場合で 13,369 に上った．また，同じ研究者に業績リストを提出してもらい，ユーザ・プロファイルを構成した．適合性の評価尺度としては，MAP (Mean Average Precision) を採用した．

今回の実験では，特に以下の 3 つのモデルに注目した．

$= 0$	$= 0.9$	$= 1$
$\mathcal{R}($,COP)	$\mathcal{R}_{0.9}$(PUP, COP)	$\mathcal{R}$(PUP,)

$= 0$ のモデルは，ユーザ・プロファイルを全く利用しないケースで，ランキングをすべてコミュニティー・モデルに任せる． $= 0.9$ のモデルは，基本的にユーザ・モデル主導型で，コミュニティー・モデルへのバックオフも許すタイプである．3 番目の $= 1$ は，すべてユーザ・モデルでランキングするコミュニティー非依存型のアプローチである．

表 1 は，MAP による本手法の全体的なパフォーマンスを示したものである．NDL は国会図書館 OPAC の出力をそのままの提示順で評価した．TEXT は，NDC のコードを一切使わず，アノテータの論文タイトルと書誌のタイトルの語彙的類似度のみに基づきランキングした結果を評価したものである．TEXT 法は NDL に比べて，やや優勢であるものの，プロファイルをベースにした本稿提案手法に遠く及ばない結果となった．各アノテータで NDL の精度が異なるのは，アノテータの判定スタイルに差があるからである．例えば，YZ は他のアノテータに比べて，適合性の許容度が広い．

しかし，アノテータの許容度の違いに関わらず，概ね，COP, PUP, PUP/COP の順で精度が向上している．PUP/COP モデルは，一貫して，NDL, TEXT の 2 倍程度の精度をマークしており提案手法の有効性を実証している．

コミュニティー選択については，RSS, KER, LM の有効性が明らかになった．概して PUP モデルが COP

[1] 国会図書館 OPAC は検索クエリに対して表示結果の最大数が 200 件という上限があるため，上限を超えてヒットしたクエリについては，200 件で足切りということにした．

表 1: MAP (Mean Average Precision) による本手法のパフォーマンス．以下 NDL は国会図書館 OPAC，TEXT は書誌題目間の TFIDF をベースにした単語のコサイン類似度によるランキング．PUP/COP, PUP, COP はそれぞれ $\mathcal{R}_{0.9}$(PUP, COP), $\mathcal{R}$(PUP,), $\mathcal{R}$(, COP) を表す．SG, EZ, YZ, OO はアノテータ ID． MAP は，各被験者の業績リストの文献各 1 編から生成したユーザ・プロファイル毎に算出，その平均を示している．RSS, KER, LM が一貫して性能が高い．

SG					
	PUP/COP	PUP	COP	NDL	TITLE
RSS	0.4022	0.3949	0.3698	0.1872	0.1870
KER	0.4019	0.3949	0.3787	0.1872	0.1870
SKL	0.3996	0.3949	0.3537	0.1872	0.1870
LL	0.3990	0.3949	0.3632	0.1872	0.1870
JS	0.3995	0.3949	0.3535	0.1872	0.1870
L1	0.3994	0.3949	0.3527	0.1872	0.1870

EZ					
	PUP/COP	PUP	COP	NDL	TITLE
RSS	0.3830	0.3670	0.4069	0.2230	0.2552
KER	0.3824	0.3670	0.4091	0.2230	0.2552
SKL	0.3718	0.3670	0.3667	0.2230	0.2552
LL	0.3714	0.3670	0.3939	0.2230	0.2552
JS	0.3718	0.3670	0.3667	0.2230	0.2552
L1	0.3718	0.3670	0.3667	0.2230	0.2552

YZ					
	PUP/COP	PUP	COP	NDL	TITLE
RSS	0.7382	0.7348	0.7179	0.4500	0.4477
KER	0.7375	0.7348	0.7111	0.4500	0.4477
SKL	0.7383	0.7348	0.7131	0.4500	0.4477
LL	0.7383	0.7348	0.7131	0.4500	0.4477
JS	0.7380	0.7348	0.7133	0.4500	0.4477
L1	0.7375	0.7348	0.7111	0.4500	0.4477

OO					
	PUP/COP	PUP	COP	NDL	TITLE
RSS	0.4372	0.4187	0.4274	0.1850	0.2168
KER	0.4362	0.4187	0.4235	0.1850	0.2168
SKL	0.4269	0.4187	0.3962	0.1850	0.2168
LL	0.4269	0.4187	0.4162	0.1850	0.2168
JS	0.4259	0.4187	0.3866	0.1850	0.2168
L1	0.4260	0.4187	0.3962	0.1850	0.2168

より良好であるが，EZ では RSS と KER で COP が PUP を顕著に凌いでおり，PUP の精度に引き上げに成功している．また，COP モデルの精度がばらつきが，そのまま PUP/COP モデルの精度に反映しており，いかにコミュニティーを選択するかが，ランキングの精度を決める重要な要因であることが確認された．

7 おわりに

以上，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概要を述べた．書誌データは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と言える．本研究は，OPAC に内在する情報，特に膨大なコストをかけて構築されている国会図書館の分類体系を利用することで，ランキングの精度を改善できることを示した．

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，ユーザの業績リストの自動構築など検討していきたい．また，図書分類体系を用いた語彙の意味記述なども興味深いトピックと言える．

参考文献

[1] Kristin Antelman, Emily Lynema, and Andrew K. Pace. Toward a Twenty-First Century Library Catalogue. *Information Technology and Libraries*, 2006.

[2] Liang Gou, Xiaolong (Luke) Zhang, Hung-Hsuan Chen, Jung-Hyun Kim, and C. Lee Giles. Social network document ranking. In *JCDL '10*, 2010.

[3] Jia Mi and Cathy Weng. Revitallizing the Library OPAC: Interface, Searching, and Display Challenges. *Information Technology and Libraries*, 2008.

[4] Karen G. Schneider. How OPACs Suck, Part 1: Relevance Rank (Or the Lack of It). ALA TechSource Blog, March 2006.

[5] Karen G. Schneider. How OPACs Suck, Part 2: The Checklist of Shame. ALA TechSource Blog, April 2006.

[6] Karen G. Schneider. How OPACs Suck, Part 3: The Big Picture. ALA TechSource Blog, May 2006.