

学術情報検索における閲覧論文の文献種別による分析

小林 和央^{1,a)} 風間 一洋² 吉田 光男³ 大向 一輝⁴ 佐藤 翔⁵

概要：インターネットの普及と文献の電子化に伴い、CiNii Articlesなどの学術論文検索が日常的な学習や研究で重要な役割を果たすようになった。ただし、研究者以外の多彩なユーザに、文献がどのように利用されているかは明らかではない。本稿では、学術情報検索サービスのユーザの閲覧行動を、何人が閲覧したかで重要度を示す閲覧人数と、ある論文を閲覧したユーザが他にどのくらい論文を閲覧したかで内容の普遍性を示す閲覧普遍度の2つの指標と、学術および一般出版物の種別を用いて分析する。まず、CiNii Articlesのサービスの利用履歴から、ユーザが閲覧した各文献の閲覧人数と閲覧普遍度を求める。次に、計算機科学分野のクエリの検索結果を閲覧人数と閲覧普遍度で順位付けした後に、該当分野における主要3学会が発行・開催する論文誌、研究会、大会・シンポジウム、そして一般的な出版社の雑誌などの異なる特性を表す種別を付与することで、各指標の順位に伴う文献種別の傾向や特性の変化を分析する。

1. はじめに

研究者にとって、文献を読むことは重要なタスクである。文献や検索手段が紙媒体だった時代は、購入や図書館の貸し出しなどに手間や時間が掛かったが、インターネットの普及と学術出版物の電子化と共に、CiNii Articlesのような学術論文検索サービスが日常的な学習や研究活動に重要な役割を果たすようになった。同時に、研究者だけでなく、学生、開発者などの多彩なユーザに使われるようになったが、学術情報検索では時間順、被引用数順など順位付け方法が限られており、すべてのユーザにとって必ずしも使いやすいとは限らない。そこで、学術情報検索の利用履歴から、実際に学術出版物をどのように利用しているかを明らかにすると共に、その知見に基づいて、特定少数の研究者や一般読者のようなユーザの特性に合った指標や特徴量を順位付けに用いる必要がある。

本稿では、学術情報検索サービスのユーザの閲覧行動を、何人が閲覧したかで重要度を示す閲覧人数と、ある論文を閲覧したユーザが他にどのくらい論文を閲覧したかで内容の普遍性を示す閲覧普遍度の2つの指標と、学術および一般出版物の種別を用いて分析する。

2. 関連研究

学術情報検索サービスのユーザに関する特徴分析として、佐藤らは、国立国会図書館サーチ（NDLサーチ）のアクセスログに基づき、利用者の簡易検索、詳細検索、ファセット検索の利用状況を分析した。NDLサーチのファセット検索では、本や記事・論文などの「資料種別」や「データベース」「所蔵館」「出版年」「分類」「分野」「国・地域」の7観点や特徴語による絞り込みが可能で、多くのユーザは簡易検索を行った後、ファセット検索等の絞り込みのための機能を利用していると述べている。[1]。

また、学術情報の評価指標の多様化に関して、Hristakevaらは論文出版数から研究経験を5段階（Unpublished, Postgraduate, Postdoc, Lecturer, Professor）に分類して4種類の論文推薦システムで評価し、研究経験によって異なる推薦方法が良いことを示した[2]。また、Thelwallは文献利用者を12属性（Student, Lecturer, Professor, Librarianなど）に分類し、Mendeleyのユーザが保存した文献の種類を調査して、属性ごとに異なる傾向を持つことを示した[3]。つまり、ユーザの経歴や現在の仕事などの属性の違いによって、異なる指標を使い分ける必要がある。

3. 文献の評価指標

文献の重要度を、内容ではなくユーザの論文閲覧履歴から求めるために、文献 $p_i (i = 1, \dots, M)$ とユーザ $u_j (j = 1, \dots, N)$ の2部グラフ（bipartite graph） G としてモデル化し、式(1)に示す行が文献、列がユーザを表す

¹ 和歌山大学大学院システム工学研究科

² 和歌山大学システム工学部

³ 豊橋技術科学大学情報・知能工学系

⁴ 国立情報学研究所コンテンツ科学研究系

⁵ 同志社大学 免許資格課程センター

a) s206099@wakayama-u.ac.jp

M 行 N 列の接続行列 (incidence matrix) $R_{p \times u}$ で表す.

$$R_{p \times u} = \begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,N} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{M,1} & r_{M,2} & \cdots & r_{M,N} \end{bmatrix} \quad (1)$$

なお, 関係があれば $r_{i,j} = 1$, なければ $r_{i,j} = 0$ である.

文献間の関係の有無は, G 上で 2 ホップで p_i から p_j に到達可能かに相当し, 接続行列 $R_{p \times u}$ から式 (2) のように求めた N 次の隣接行列 (adjacency matrix) R_p で表す.

$$R_p = R_{p \times u} \times R_{p \times u}^T \quad (2)$$

$R_{p \times u}^T$ は $R_{p \times u}$ の転置行列で, R_p の要素 $r'_{i,j}$ は p_i から p_j に到達可能なら $r'_{i,j} > 0$, 不可なら $r'_{i,j} = 0$ である.

同様に, ユーザ間の関係の有無は, 接続行列 $R_{p \times u}$ から式 (3) のように求めた M 次の隣接行列 R_u で表す.

$$R_u = R_{p \times u}^T \times R_{p \times u} \quad (3)$$

ここで, R_u の要素 $r''_{i,j}$ は u_i から u_j に到達可能なら $r''_{i,j} > 0$, 不可なら $r''_{i,j} = 0$ である.

3.1 閲覧人数

文献 $p_i (i = 1, \dots, M)$ の次数 $\deg(p_i)$ は接続行列 $R_{p \times u}$ から式 (4) で計算する.

$$\deg(p_i) = \sum_{j=1}^N r_{i,j} \quad (4)$$

以降では, 次数 $\deg(p_i)$ をより一般的に**閲覧人数**と呼ぶ.

3.2 閲覧普遍度

文献の閲覧人数が同じでも, 必ずしもその内容の特性も同じとは限らない. 例えば, 最新技術やトップデータを提案する論文以外にも, 初学者向けのやさしい解説文献があれば, 一般人の興味も惹くような一般的な文献もあり, ユーザによって求める文献の特性は異なるはずである.

そこで, 文献の特性は閲覧ユーザ集合から推定できると考えて, ある文献に 2 部グラフ上でエッジで接続しているユーザ群に, さらにエッジで接続している文献の和集合が全体に占める割合を求めることで文献の普遍性を定量化する**普遍度** (general degree) [4] を用いる. 文献 p_i の普遍度 $S(p_i)$ は, G 上で 2 ホップで到達可能な (式 (2) の R_p で $r'_{i,j} > 0$) 文献数を全文献数 M で割って求める.

$$S(p_i) = \frac{|\{r'_{i,j} | r'_{i,j} > 0\}|}{M} (j = 1, \dots, M) \quad (5)$$

例えば, 文献が特定の分野のユーザにしか閲覧されない場合は和集合が小さいので閲覧人数が多くても本指標は小さく, 逆に異なる分野に興味を持つユーザに広く閲覧された場合は和集合は大きくなり, 本指標は大きくなる. ただし, この普遍性は, 対象とするユーザ集合に大きく影響されることに注意する必要がある.

表 1 文献種別の特性

文献種別	難易度	ページ数	査読	読者数
論文誌	高	8	有	中
研究会	中	6	無	少
大会・シンポジウム	低～中	2～4	有 or 無	中
一般雑誌	低	5	無	多

4. 閲覧文献の分析手法

学術情報検索サービスとして, CiNii Articles を対象に, ユーザの閲覧行動を, 閲覧人数と閲覧普遍度の 2 つの指標と, 学術および一般出版物の種別を用いて分析する.

まず, CiNii Articles のアクセスログを分析し, アクセス先の URL に含まれる NAID (NII 論文 ID) を論文識別子, IP アドレスとユーザエージェントの組をユーザ識別子として, 文献の閲覧人数と閲覧普遍度を計算する.

次に, 調査対象のクエリを用いて, 使用したアクセスログの期間に合うように出版年を指定して, 文献を CiNii Articles の OpenSearch API^{*1} で N 件取得し, 閲覧人数と閲覧普遍度の 2 つの指標で文献を再順位付けする.

最後に, 閲覧人数・閲覧普遍度で再順位付けした場合の文献の特性の順位に伴う変化を調べるために, 各指標の順位付け結果に文献情報を表示すると共に, 文献に種別情報を付与し, 各指標の文献種別の累積確率と順位の間隔をグラフとして可視化する. 文献種別としては, 学会関連の論文誌, 研究会, 大会・シンポジウムと, 一般出版社関連の一般雑誌の 4 種類を用いた. 文献種別は文献の特性の違いを反映すると考えられ, 難易度, ページ数, 査読の有無, 想定読者数の 4 つの属性の一般的な傾向を表 1 に示す.

5. 分析

5.1 データセット

本稿では, CiNii Articles の 2014 年 4 月 1 日～2016 年 3 月 31 日の 2 年間の Apache Web サーバの Combined Log Format のアクセスログを用いた. なお, クローラ等のプログラムのアクセスはユーザの分析に悪影響を与えるので, 佐藤の手法 [5] を参考に除去した. この結果, 閲覧論文数は 8,826,591 件, ユーザ数は 12,460,893 人となった.

文献種別は, 今回対象とする計算機科学分野の主要 3 学会である情報処理学会と電子情報通信学会, 人工知能学会の出版者名と刊行物名, 一般出版社の出版者名で分類した. これは, 主要 3 学会だけでもかなりの文献をカバーできることと, 規模の小さな学会では今回用いた種別が必ずしも明確に区別して運営されていないからである. なお, 学会誌は検索結果中の出現数が少なく, 除外した.

^{*1} https://support.nii.ac.jp/ja/cia/api/a_opensearch

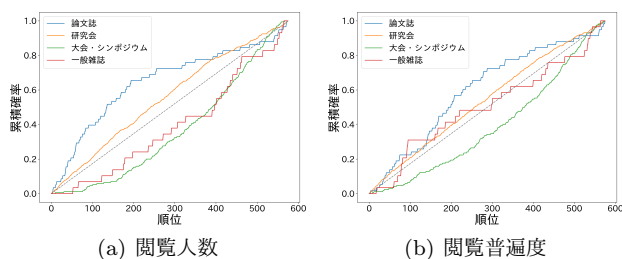


図 1 「twitter」の文献種別の累積確率分布

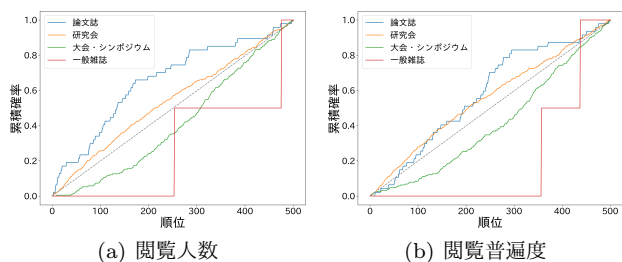


図 2 「推薦システム」の文献種別の累積確率分布

5.2 文献種別と順位の関係分析

まず、閲覧人数や閲覧普遍度の順位付け結果を分析するために、各指標の各文献種別と順位の間を分析した。計算機科学分野で検索結果数が1,000件以上の「twitter」を一般的、それ以下の「推薦システム」を専門的なトピックとし、各順位付け結果に対し、横軸を順位、縦軸を文献種別の累積確率をプロットして図1と図2に示す。補助として、閲覧確率が順位によらず一定な場合を破線で示す。

閲覧人数の場合は、上位に論文誌が、下位に大会・シンポジウムが来る傾向があり、研究会はその中間であることがわかる。すなわち、CiNii Articlesのユーザは、検索結果から技術性や信頼性を考慮して閲覧論文を選んでいると考えられる。ただし、一般雑誌はクエリによって文献数が大きく異なり、クエリのトピックが一般的か、専門的に大きく左右されていると思われる。

これに対して、閲覧普遍度の場合は、論文誌は全体的により低い順位にシフトし、逆に一般雑誌類は「twitter」で全体的に高い順位にシフトしていることがわかる。これは論文誌は同じ分野のユーザに、一般雑誌は様々な分野のユーザに読まれているからだと思われる。逆に、大会・シンポジウムは順位が高い部分でのみ、より高順位にシフトしている。これは、大会・シンポジウムは玉石混交で多くの人に読まれる文献は限られるが、そのような一部の文献は比較的広い分野の人に読まれているからだと思われる。

5.3 閲覧人数・閲覧普遍度による再順位付け結果の分析

次に、実際にどのような文献が上位かを分析した。「twitter」と「推薦システム」をクエリとして閲覧人数と閲覧普遍度で順位付けした上位12件の論文を閲覧人数と閲覧普遍度の順位、文献の題名、出版者名、文献種別と共に表2と表3に示す。出版者名は略記形式で示し、対象3学会と

対象文献種別以外の文献種別は括弧で囲んだ。

閲覧人数より閲覧普遍度で大きく順位が下がる文献としては位置情報分析や感情分析などの専門的な技術に関する論文が、逆に大きく順位が上がる文献としては不登校問題や情報モラル、メンタルトレーニングの音楽や観光イベントのように、他分野にも関わる論文が見られた。

なお、閲覧人数の上位に、今回の対象3学会以外の大学紀要やJSET（日本教育工学会）、NAL（日本図書館研究会）の文献が出現し、閲覧普遍度ではさらに順位が高くなる傾向が見られた。これらの出版者における計算機科学分野の関連文献数は少ないものの、CiNii Articlesのユーザは人文系・図書館系のユーザが多いことから閲覧人数が多くなり、またそれらの文献の内容の分野横断性から閲覧普遍度も高くなったと考えられる。

6. おわりに

CiNii Articlesのユーザの閲覧行動を分析するために、検索結果を閲覧人数と閲覧普遍度の2つの指標で再順位付けした結果を、異なるユーザ特性を示す文献種別情報を付与し、各指標の特性の違いと変化を分析した。その結果、文献種別を手掛かりに閲覧論文を選択しているユーザが多いこと、閲覧普遍度はそのユーザ集合で一般的な論文を上位にする傾向があることを確認できた。

今後の課題として、普遍度以外の文献の異なる特性を示す他の特徴量を見つけ出すと共に、それらの特徴量をLearning to Rankのような手法で組み合わせてユーザの特性に応じた文献の検索・推薦手法を検討する予定である。

謝辞 本研究はJSPS 科研費19H04421と国立情報学研究所公募型共同研究「学術情報検索における論文閲覧行動とソーシャルメディアにおける論文言及行動の関連性に関する研究」の助成を受けた。

参考文献

- [1] 佐藤 翔, 安藤大輝, 川瀬直人, 北島顕正, 塩崎 亮, 那珂 元, 原田隆史: ディスカバリサービスにおける絞り込みプロセス: 国立国会図書館サーチのアクセスログ分析, 図書館界, Vol. 67, No. 4, pp. 244-261 (2015).
- [2] Hristakeva, M., Kershaw, D., Rossetti, M., Knoth, P., Pettit, B., Vargas, S. and Jack, K.: Building Recommender Systems for Scholarly Information, *Proceedings of the 1st Workshop on Scholarly Web Mining (SWM '17)*, pp. 25-32 (2017).
- [3] Thelwall, M.: Does Female-authored Research have More Educational Impact than Male-authored Research? Evidence from Mendeley, *Journal of Altmetrics*, Vol. 1, No. 1, p. 3 (2018).
- [4] 小林和央, 風間一洋, 吉田光男, 大向一輝: インターネット上の論文の閲覧行動と言及行動の関係の分析, 第3回計算社会科学ワークショップ (CSSJ2019) (2019).
- [5] 佐藤 翔: コンテンツ入手元として機関リポジトリが果たしている役割, 博士論文, 筑波大学 (2013).

表 2 「twitter」の上位 12 件
(a) 閲覧人数

人数	普遍度	題名	出版者名	文献種別
1	1	ソーシャルメディアにおけるプライバシーリスクの盲点：リスク軽減に向けた論点整理	東京情報大学	(大学紀要)
2	2	若者における SNS 利用行動およびリスク認知の検討：LINE と Twitter を中心に	プール学院大学	(大学紀要)
3	4	Twitter における影響力の分析手法	IPSJ	研究会
4	6	SNS の投稿を用いた感情記録ライフログシステム ～Emote～	IPSJ	研究会
5	12	ソーシャルネットワークによる Web からの情報収集（＜特集＞『検索』のゆくえ）	INFOSTA	(学会誌)
6	30	商品に関する Twitter 上のコミュニケーションと販売実績の関連性分析	ORSJ	(学会誌)
7	136	位置情報付きツイートを利用した観光ルート推薦	IPSJ	研究会
8	3	朝ドラ『あまちゃん』はどう見られたか：4 つの調査を通して探る視聴のひろがり視聴熱	NHK 文研	(研究会)
9	14	ソーシャルメディア活用による個人・組織のブランディング戦略（若手カンファレンス報告）	SSI	(学会誌)
10	10	一般ユーザの観点に基づく Twitter からの人物関係の可視化と事例の考察	IPSJ	論文誌
11	48	Twitter におけるユーザの興味と話題の時間発展を考慮したオンライン学習可能なトピック...	IPSJ	論文誌
12	38	ソーシャルメディアの普及がファッションの学習と情報流通に与えた影響に関する一考察	光葉会	(大学紀要)

(b) 閲覧普遍度

人数	普遍度	題名	出版者名	文献種別
1	1	ソーシャルメディアにおけるプライバシーリスクの盲点：リスク軽減に向けた論点整理	東京情報大学	(大学紀要)
2	2	若者における SNS 利用行動およびリスク認知の検討：LINE と Twitter を中心に	プール学院大学	(大学紀要)
8	3	朝ドラ『あまちゃん』はどう見られたか：4 つの調査を通して探る視聴のひろがり視聴熱	NHK 文研	(研究会)
3	4	Twitter における影響力の分析手法	IPSJ	研究会
15	5	Twitter 上でのシャイなユーザーの自己開示（思考と言語）	IEICE	研究会
4	6	SNS の投稿を用いた感情記録ライフログシステム ～Emote～	IPSJ	研究会
57	7	Twitter による不登校の母親援助 -140 字支援による可能性とその限界-	京都文教大学	(大学紀要)
44	8	Twitter に見る若者たちのコミュニケーション考（特集 ネット化社会）	青少年問題研究会	(学会誌)
75	9	大学生の情報モラル醸成における SNS の利用：「twitter かくれんぼ」を事例として	畿央大学	(大学紀要)
10	10	一般ユーザの観点に基づく Twitter からの人物関係の可視化と事例の考察	IPSJ	論文誌
36	11	Twitter からのキャラクター印象表現の抽出（言語理解とコミュニケーション）	IEICE	研究会
5	12	ソーシャルネットワークによる Web からの情報収集（＜特集＞『検索』のゆくえ）	INFOSTA	(学会誌)

表 3 「推薦システム」の上位 12 件
(a) 閲覧人数

人数	普遍度	題名	出版者名	文献種別
1	12	推薦システムのアルゴリズム (1)	JSAI	(学会誌)
2	1	絵本の読み聞かせにおける子どもの好みと絵本の主題との関係性	JSET	(論文誌)
3	56	協調フィルタリングとコンテンツ分析を利用した観光地推薦手法の検討	上武大学	(大学紀要)
4	3	図書館の貸出履歴と書誌情報を用いた図書推薦システムの有効性	NAL	(学会誌)
5	11	オノマトペロリ：オノマトペを利用した料理推薦システム	IPSJ	研究会
6	5	図書館の貸出履歴を用いた図書推薦システムの有効性検証	NAL	(学会誌)
7	112	Twitter 感情分析を用いた感情値可視化とユーザ推薦システム	IPSJ	研究会
8	50	状況依存型ユーザ嗜好モデリングに基づく Context-Aware 情報推薦システム	IPSJ	論文誌
9	2	読み聞かせ時の反応に着目した絵本に対する子どもの好みの取得方法に関する検討	JSET	(論文誌)
10	23	Web 上に混在する観光情報を活用した観光地推薦システム（言語理解とコミュニケーション・第 4...	IEICE	研究会
11	6	脳波利用の音楽推薦システムに向けた前額部導出電位を利用した心理状態分類の基礎検討（2014 年...	IPSJ	研究会
12	27	推薦システム-情報過多時代をのりきる（＜特集＞情報のフィルタリング）	INFOSTA	(学会誌)

(b) 閲覧普遍度

人数	普遍度	題名	出版者名	文献種別
2	1	絵本の読み聞かせにおける子どもの好みと絵本の主題との関係性	JSET	(論文誌)
9	2	読み聞かせ時の反応に着目した絵本に対する子どもの好みの取得方法に関する検討	JSET	(論文誌)
4	3	図書館の貸出履歴と書誌情報を用いた図書推薦システムの有効性	NAL	(学会誌)
23	4	複数人での旅行における嗜好分析による観光地推薦システムの提案	IPSJ	研究会
6	5	図書館の貸出履歴を用いた図書推薦システムの有効性検証	NAL	(学会誌)
11	6	脳波利用の音楽推薦システムに向けた前額部導出電位を利用した心理状態分類の基礎検討（2014 年...	IPSJ	研究会
71	7	スポーツメンタルトレーニングへの応用を目指した脳波利用の音楽推薦システム	IPSJ	大会・シンポジウム
37	8	映像コンテキストに基づく多視点映像の視点列推薦	IPSJ	研究会
60	9	類似性に基づく観光イベント情報推薦システムの提案	IPSJ	大会・シンポジウム
128	10	F-004 タグによる嗜好抽出を用いた観光イベント推薦システムの提案（F 分野:人工知能・ゲー...	FIT	大会・シンポジウム
5	11	オノマトペロリ：オノマトペを利用した料理推薦システム	IPSJ	研究会
1	12	推薦システムのアルゴリズム (1)	JSAI	(学会誌)