

平成 26 年度 学術情報システム総合ワークショップ 調査報告書

【2 班】 利用ログの分析

福岡大学 亀崎有紀子

京都大学 渡邊伸彦

目次

1	はじめに	1
2	目的	1
3	方法	2
3.1.	調査対象	2
3.2.	調査内容	3
4	検索回数による二分化基準（第一基準）	3
4.1.	方法	3
4.2.	結果	4
4.3.	考察	13
5	演算子使用による二分化基準（第二基準）	15
5.1.	方法	15
5.2.	結果	16
5.3.	考察	18
6	まとめ	19
7	参考文献	24

1 はじめに

インターネット、携帯電話、スマートフォン、タブレット端末等の普及により、我々の情報環境は劇的に変化した。また、Googleなどの検索エンジン等の発達により、専門知識がなくても容易に検索ができるようになった。いまや老若男女誰もが端末を操作し、知識を得るために気軽に検索を行っている。検索行動は生活の一部になったと言っても過言ではない。

しかし、そのような情報環境の中にあって、学術情報データベース（DB）は次のような問題を抱えている。

- 網羅性の高いDBは基本的な検索機能のみで、絞り込みや新しい提案をする機能まで持たない。
- 高い検索機能や絞り込み機能を持つDBは検索対象が極めて限定的であったり、有料のサービスであったりする。
- DB内のコンテンツはそのDB内に限定されてしまっていて、その情報を求める利用者に認知されない。
- Google ScholarなどからDBに誘導されても、使い方がわかりづらい。

このように、学術情報が手軽に検索できない状況が、非学術研究者や学生などの初学者を学術情報から遠ざけてしまっている。

「普段インターネットで手軽に検索しているユーザー」は、学術情報を必要としたとき、日常の検索体験と学術情報データベースでの検索方法の違いに戸惑い、十分な検索を行わず、求める学術情報にたどり着けない、もしくは偶然発見した学術情報で満足し、本来たどり着くべき多くの学術情報と出会えていない可能性がある。このようなユーザー、つまり求める文献を的確に発見できないユーザーが適切な文献検索を行うために、学術情報データベースの利用ログの分析を通じて何らかの提案ができるかもしれないと考え、この調査を行うことにした。

2 目的

適切な文献検索を行う方法を提案するために、求める文献を的確に発見できないユーザーと的確に発見できるユーザーの検索行動や傾向を多角的に把握し、比較する手法を用いることにした。今回の調査では、その手段として、学術情報データベースのアクセスログを利用する。求める文献を的確に発見できないユーザーは、「検索語の選定を疎かにする」

「網羅的な検索を行わない」「検索結果を十分に検討しない」「学術 DB に関する知識を得る機会が少ない」「学術 DB を利用することが少ない」などの特徴を持つと推測される。学術情報データベースのアクセスログを行動や属性で分類することができれば、それらの特徴を明確に掴むことができ、また、文献を的確に発見できるユーザーの特徴も掴むこともできると考えたからである。加えて、彼らが的確に文献を発見できない外的要因についてもログから把握できると考えた。本調査はアクセスログを利用し、ユーザーが求める文献を的確に発見できない要因を内的・外的の両面から明らかにした上で、適切な文献検索を行う方法を提案することを目的とする。

3 方法

3.1. 調査対象

本調査では学術情報データベースのアクセスログの例として、CiNii Articles の 1 年分（2013 年 9 月 1 日～2014 年 8 月 31 日）のアクセスログデータを分析対象とした。

CiNii（サイニイ）は、国立情報学研究所（NII）が運営する国内最大の学術情報データベースである。CiNii は学協会刊行物・大学研究紀要・国立国会図書館の雑誌記事索引データベースなどの学術論文情報を検索できる CiNii Articles と全国の大学図書館等が所蔵する本（図書・雑誌）の情報を検索できる CiNii Books に分かれている。今回の調査は、適切な文献検索を行う方法を提案することが最終目的であるため、調査対象を学術論文情報探索行動の軌跡を追究しやすいと推測される CiNii Articles に限定した。CiNii Articles で検索できる論文数は、引用情報を除き、約 1,800 万件である。（2014 年 9 月現在）

調査対象期間のデータの状況は次の表 1 のとおりである。

表 1 CiNii Articles の状況（2013.9.1-2014.8.31）

セッション数	41,353,793 件
ユーザー数	19,686,834 件
ページビュー数	199,565,613 件

参考までに、同期間の福岡大学（学生数：約 2 万人）OPAC のセッション数は 337,309 件であり、CiNii Articles のセッション数は実にその約 122.6 倍と桁違いの数値となっている。

ログ解析には主に Google Analytics を使用し、補助ツールとして Google Analytics Query Explorer、データの抽出・集計ツールとして、Microsoft Excel、Microsoft Access、Ruby2.1 を使用した。

Google Analytics は Google 社が提供するウェブビーコン型の高機能なアクセス解析ツールで、Web サイトでのヒット数が月間 1,000 万件を超えない限り、誰でも無料で活用でき、初期設定画面を眺めるだけでも様々な解析結果を知ることができる。また、指標とディメンションと呼ばれる集計区分を自由に組み合わせて分類・集計したり、セグメント機能でユーザーやセッションを絞り込んだりすることも可能である。今回の調査では、セッションを特定の条件で絞り込むのにセグメント機能が大いに役立った。ただし、Google Analytics はデータ（訪問回数）が初期設定で 25 万件を超える場合、「このレポートは N 回の訪問を基に作成されています。」というメッセージが表示され、サンプルデータとしてレポートが作成される仕様となっている。そのため、実数を対象とする統計調査には適していない。しかし、今回の調査は割合比較から傾向を分析する調査であるため、使用ツールとして妥当と思われる。

3.2. 調査内容

CiNii Articles のアクセスログデータを特定の基準に基づいて「求める文献を的確に発見できるユーザー」と「求める文献を的確に発見できないユーザー」の二つに分け、様々な視点から比較調査を行うことで、それぞれのユーザーの傾向を分析した。ユーザーの傾向の分析をより正確なものとするため、調査後に二分化基準の評価を行った。

4 検索回数による二分化基準（第一基準）

4.1. 方法

(1) 調査内容

この基準では「求める文献を的確に発見できないユーザー」を「CiNii Articles を 1 回検索し、検索結果一覧を眺めて終了するユーザー」と「CiNii Articles を 1 回検索し、1 回論文詳細画面を見て終了するユーザー」の 2 種類のユーザーを合わせたものとした (1-A)。また、「求める文献を的確に発見できるユーザー」を「CiNii Articles を 1 回以上検索し、その後も論文情報を閲覧し続けたユーザー」とした (1-B)。

これは、「求める文献を的確に発見できないユーザー」とは検索結果を十分に検討してい

ないユーザーのことであり、そのために文献を的確に発見できていないと想定したためである。そして、検索結果を十分に検討していない行動を「1回検索し、検索結果一覧を眺めて終了」と「1回検索し、1回論文詳細画面を見て終了」の2通りと考えた。ユーザーの具体的な設定は、Google Analytics のセグメント機能を利用している。この機能ではデータをセッション単位やユーザー単位で設定することが可能であり、今回はセッション単位としている。詳細な設定条件については別紙の【付録1】を参照いただきたい。

この基準によるデータの抽出状況は表2のとおりである。1-A、1-B、全体と比較すると、本基準が少数の絞られた層の傾向であることがわかる。また、1-Bはセッション数やユーザー数が全体に占める割合に比べると、ページビュー数の割合が全体の約半数となっており、一人あたりのページビュー数が多いことがわかる。

表 2 第一基準及び全体の件数

	1-A	1-B	全体
セッション数	950,198 件 (2.29%)	5,626,031 件 (13.6%)	41,353,793 件
ユーザー数	671,429 件 (3.35%)	2,844,799 件 (14.2%)	19,686,834 件
ページビュー数	2,251,505 件 (1.12%)	93,367,016 件 (46.72%)	199,565,613 件

※下段の()は全体との比率

(2) 傾向比較視点

「検索語」「端末」「滞在時間」「所属」「演算子を使った検索」「検索方法」「参照元」の視点で傾向比較を行った。

各比較視点でのデータの抽出方法については別紙の【付録3】を参照いただきたい。

4.2. 結果

(1) 検索語

ここでは各検索において使用された「検索語」に着目し、その傾向を3つの視点から調査をした。今回の調査では、単語の区切りをスペースで判断したため、記号も1語とカウントされていることがある。また、アルファベット言語は単語をスペースで区切る性質であるため、語数が多くなる傾向にある。この点はあらかじめ断っておきたい。

(1)-1 調査 1

この調査では、検索語の長さや単語数の傾向を知るために、CiNii Articles における簡易検索・全文検索を対象に、使用した検索語全体の単語の長さ、1 単語当たりの単語の長さ、単語数、それぞれの平均値を調査した。表 3 はその結果を示したものである。

表 3 検索語の単語数と単語の長さの平均値

		1-A	1-B
簡易検索	単語数	1.51	1.65
	単語の長さ（全体）	8.59	7.13
	単語の長さ（単語）	5.71	4.33
全文検索	単語数	1.43	1.87
	単語の長さ（全体）	7.25	8.37
	単語の長さ（単語）	5.06	4.49

単語数の平均値は、わずかな差であるが簡易検索・全文検索共に 1-B の方が多くなっている。検索語全体の単語の長さの平均値は、1-B が簡易検索では 1-A より短く、全文検索では 1-A より長くなっている。

(1)-2 調査 2

この調査では、ユーザーが検索で使用する単語数を調べるために、簡易検索を対象に単語数ごとにページビュー数を集計した。図 1 はその割合を表したものである。

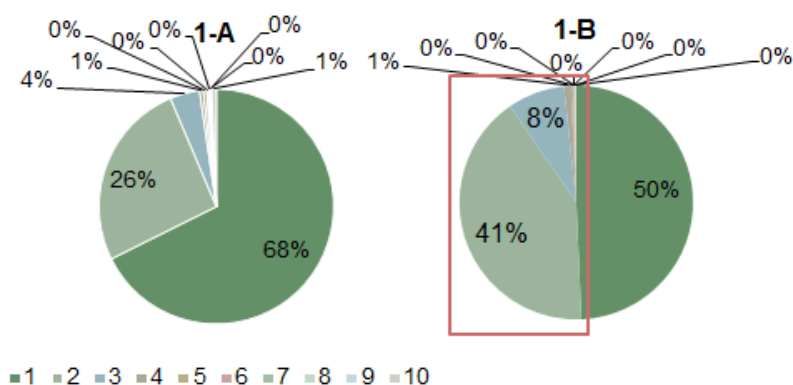


図 1 簡易検索 単語数別ページビュー割合比較

1-A では1単語による検索の割合が7割程度であるのに対し、1-B では2単語以上の検索が約半数となっている。

図2は図1の円グラフの1-Aについて、検索語の単語数別の割合を出し、その比率を1-Bと比較したものである。

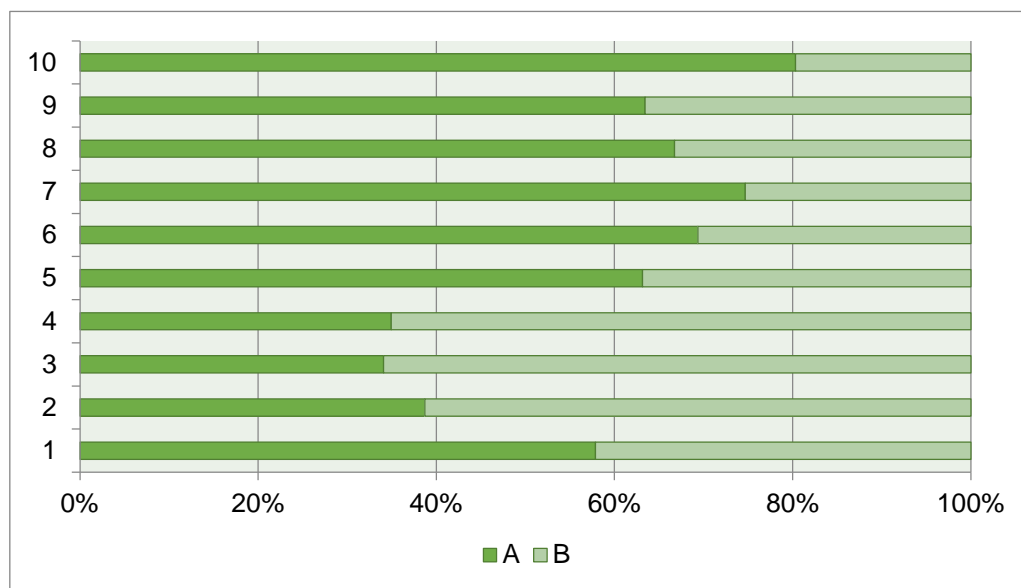


図2 簡易検索 単語数別ページビュー割合

この結果から、1-Bに2~4単語で検索を行うことが多いという特徴が出ていることがわかる。

(1)-3 調査3

1-Bの検索語設定の傾向を知るために、簡易検索を対象にページビュー数の多い検索語5語「アレロパシー」、「サッカー」、「バレーボール」、「児童虐待」、「認知症」について、1回目の検索と2回目の検索の単語数や使用語の違いを比較した。表4は1回目と2回目の検索語の単語数を比較している。

表 4 1-B の 1 回目・2 回目単語数比較

1 回目の単語数	1 回目と比較した 2 回目の単語数	件数
1	同数	80
	増加	112
	減少	0
2	同数	228
	増加	41
	減少	101
3	同数	14
	増加	4
	減少	45
4	同数	6
	増加	1
	減少	6
5	同数	0
	増加	0
	減少	2

この表から、最初の語数が 3 語以上の場合、次の語数を減らす傾向にあり、最初の語数が 1 語の場合は、次の語数を増やす傾向にあることがわかる。

なお、検索語の選択については、あらかじめ資料を特定した検索では傾向を測ることが難しいため、固有名詞、複数の文節が含まれる検索語は除外している。

表 5 は 1 回目の検索語が 2 単語以上の場合について、2 回目の検索の検索語の 1 番目と 2 番目が 1 回目と等しいか否かを調査した結果を示している。

表 5 1-B 1回目と2回目の検索語比較

2回目の1番目、2番目	件数
○、○	24
○、×	187
×、×	60
×、○	10

2語以上で再検索を行う時は1番目の1語は残して、2番目の1語を別の検索語に置き換える傾向にあることがわかる。

(2) 「端末」

この調査では、ユーザーがどのような端末からアクセスしているかを知るために、利用している端末を調べた。図3は1-A、1-B、全体それぞれについてその割合を示したものである。

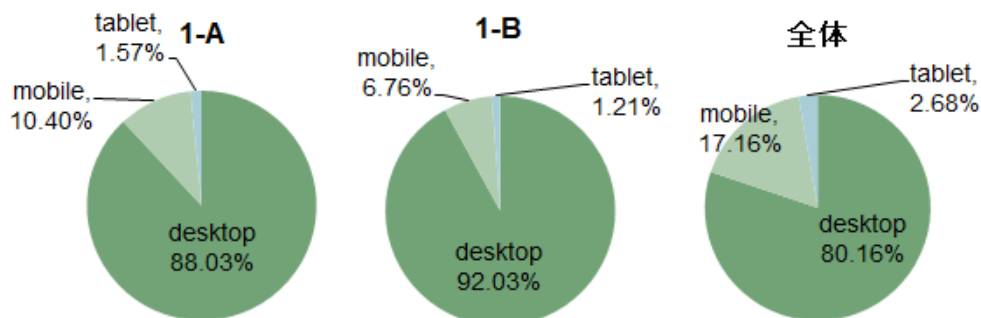


図 3 端末の割合

Desktop はパソコン(PC)、Mobile はスマートフォン、フューチャーフォン、Tablet はタブレット端末を指している。

パソコンからのアクセスは、1-A が 88.03%であるのに比べ、1-B の方は 92.03%と 1-A より割合が高くなっている。全体的にパソコンの占有率が圧倒的に高いと言える。

反対に、スマートフォンやフューチャーフォンからのアクセスは 1-B が 6.76%であるのに比べ、1-A の方は 10.4%と 1-B より割合が高くなっている。

(3) 「滞在時間」

この調査では、1-A、1-B、全体、それぞれが検索行動にどれくらい時間を費やしたのか

を知るために、ユーザーが CiNii Articles で検索に使用した時間の平均値を求めた。表 6 がその結果である。

表 6 平均セッション時間

	1-A	1-B	全体
平均セッション時間	44 秒	13 分 15 秒	3 分 46 秒

1-Bの平均セッション時間が1-Aより長いのは、条件設定から見て当然の結果と言えるが、全体の時間と比べても約 3.5 倍あることは特筆すべきことである。

なお、アクセスログによる平均セッション時間の計算は、各ページにアクセスした時間と次のページにアクセスした時間の差で求めているため、最終ページの滞在時間が合算されない。また、Google Analytics では直帰したセッションは正しい閲覧時間を計測できないため、一律 0 秒を割当てることになっている。このため、厳密に言えば滞在時間とは言えないが、三者の検索に費やした時間の平均値を比較していることから、視点として有効と考える。

(4) 「所属」

この調査では、所属機関を学術情報 DB に接する機会が多いと予測する学術機関とその他に分け、ユーザーの割合を調べた。図 4 は 1-A、1-B、全体のユーザーについて、その割合を示したものである。あくまでアクセス時に使用したネットワークに依存するため、研究者が自宅やモバイルネットワークからアクセスした場合には捕捉できない。つまり、次の結果の学術機関所属とは最低限の割合で、実際はもっと大きい可能性があることに留意したい。

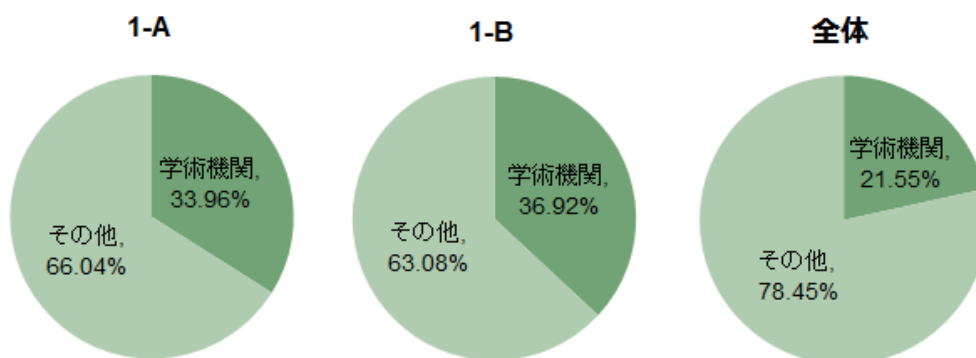


図 4 所属機関の割合

学術機関の割合が、1-A が 33.96%であるのに比べ、1-B は 36.92%と割合が高いが、わずかな差である。

全体ではそのほか 78.45%と学術機関以外からの割合が高いことがわかる。

なお、今回の調査では、学術機関とみなす条件を、アクセス元のネットワークドメインが ac.jp で終わるもの、及びサービスプロバイダの名称から学術機関と判断できるもの（名称に University、College、Institute を含むもの）とし、ユーザーはそのアクセス元の機関に所属していると仮定している。

(5) 「演算子を使った検索」

この調査は、ユーザーがどのくらい演算子を活用して検索を行っているのかを知る目的で行った。図 5 は 1-A、1-B のユーザーについて、簡易検索、全文検索、詳細検索対象に演算子を使用した割合を示したものである。

なお、今回は CiNii Articles のヘルプに記載のある演算子を調査対象とした。

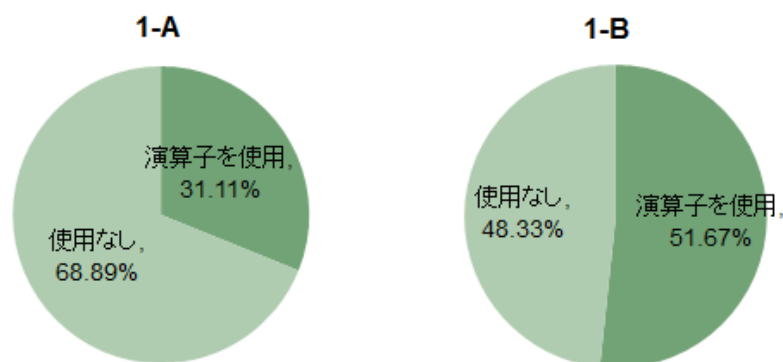


図 5 演算子使用の割合

図から 1-B の方が演算子を活用していることがわかる。

なお、CiNii Articles では、フレーズ検索や前方一致検索は 2 バイト文字では無効であるが、ユーザーが演算子を使う意思があったとみなし、「演算子を使用」に含めている。

表 7 は、図 5 の「演算子を使用」の内訳を示したものである。

表 7 使用演算子内訳

	1-A			1-B		
	件数	ページビュー数	%	件数	ページビュー数	%
完全一致検索	0	0	0.000%	22	4,147	0.011%
フレーズ検索	6	996	0.111%	84	18,406	0.050%
丸っこ検索	2	332	0.037%	228	71,246	0.192%
前方一致検索	0	0	0.000%	66	19,390	0.052%
NOT 検索	4	664	0.074%	120	31,658	0.085%
OR 検索	2	332	0.037%	313	92,129	0.248%
AND 検索	1,663	277,213	30.856%	71,331	18,962,585	51.034%
総数		898,407			37,156,996	

1-B は演算子の中でも特に AND 検索を活用している割合が高いことがわかる。

本調査でデータを参照したところ、演算子の誤用が散見された。次の図 6 は 1-B でフレーズ検索と前方一致検索の誤用が行われた割合を示している。この場合の誤用とは、無効である 2 バイト文字を使ったことを指している。1-A の場合は、フレーズ検索で誤用 100%、前方一致検索は使用が確認されなかった。

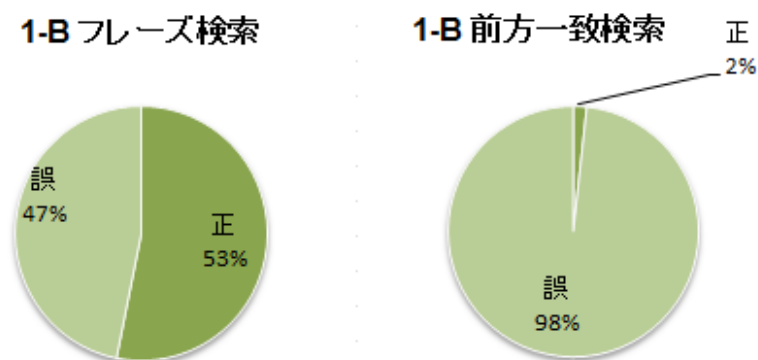


図 6 1-B のフレーズ・前方一致検索の誤用の割合

フレーズ検索のおよそ半数の 47%、前方一致検索の 98%が間違った使い方をされていることがわかる。演算子使用全体から見れば少数ではあるが、特定の演算子で正しく使われていない割合が高いことは確かである。

(6) 「検索方法」

この調査は、ユーザーが様々な検索方法を使い分けているのかを知るために行った。図7は1-A、1-Bのユーザーについて、検索方法の簡易検索、詳細検索、全文検索、著者検索の割合をそれぞれ示している。

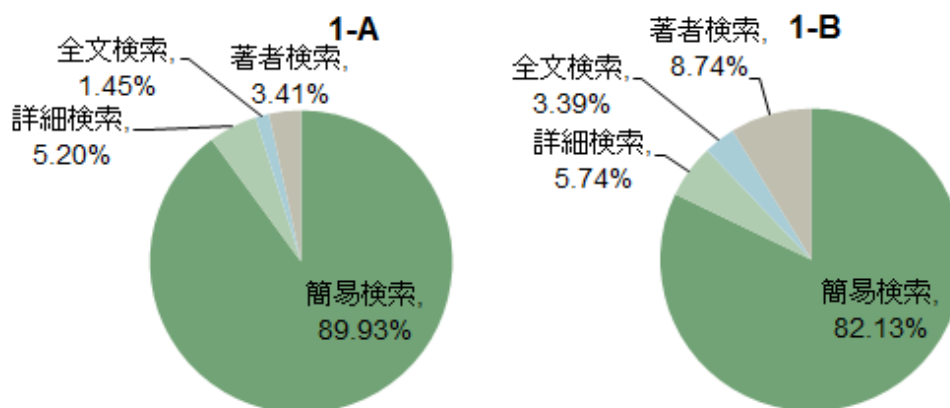


図7 検索方法の割合

簡易検索以外の検索方法の割合は、1-Aが10.06%であるのに比べ、1-Bは17.87%と簡易検索以外の検索方法の割合が高く、特に全文検索や著者検索を活用する割合が高いことがわかる。しかし、8割以上のユーザーが簡易検索を利用していることもわかる。

一方で、条件設定による検索である詳細検索の利用割合は、両者で差が見られなかった。

(7) 「参照元」

この調査は、ユーザーがどのような経路をたどってCiNii Articlesにアクセスしたかを知るために行った。図8は、1-A、1-B、全体のユーザーについて、参照元の割合を示したものである。

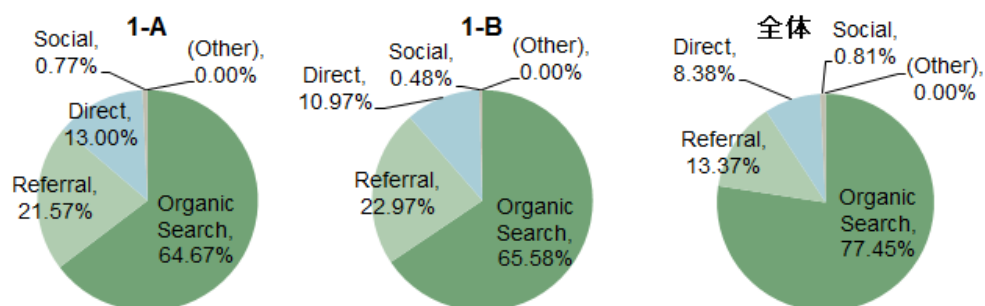


図8 参照元の割合

Organic Search は検索エンジンからの流入、Referral は Web サイトやメルマガの URL リンクからの流入、Direct はブラウザのブックマークや URL の直打ちでの流入、Social は SNS からの流入、(Other) はその他、不明などをそれぞれ示している。

1-B が 10.97% であるのに比べ、1-A は 13% とブラウザのブックマークや URL の直打ちでの流入 (Direct) の割合が高いことがわかる。

また、全体の 8 割近くの訪問が検索エンジン (Organic Search) を経由して来ていることが明らかである。

4.3. 考察

4.2 のそれぞれの結果について考察し、総合的に第一基準の妥当性を検証する。

(1) 「検索語」

調査 1 の結果より、1-B の方が単語数を多めに設定したり、単語の長さを検索方式に合わせて設定したり、工夫をしているように推測されるが、差がわずかであることから、明確にそれぞれの特徴を表しているとは言い難い。

調査 2 や調査 3 の結果からは、1-B が検索語の設定を工夫しているという特徴が読み取れる。

(2) 「端末」

1-A、1-B、全体について比較すると、多少の差はあるものの、類似した傾向を示している。従って、特徴を表しているとは言えない。ただし、全体としてパソコンの割合が圧倒的に高いことは確実である。

(3) 「滞在時間」

1-B が全体の平均を大幅に上回っていることから、時間を掛けて検索を繰り返したり、ページを回遊したりして、欲しい情報を得ようとしていることが推測される。ただし、第一基準の条件設定上、必ず差が生まれる項目なので、この項目の結果をもって妥当性を検証することはできない。

(4) 「所属」

1-A と 1-B の差がわずかであるため、それぞれの特徴を表しているとは言えない。

(5) 「演算子を使った検索」

1-A が 31.11%であるのに比べ、1-B は半数以上が演算子を使用しているため、1-B は演算子を活用しているという特徴を持っていると言える。

(6) 「検索方法」

1-A と 1-B の簡易検索以外の検索方式の使用に顕著な差がないため、それぞれの特徴を表しているとは言えない。ただし、全体的に簡易検索の割合が圧倒的に高いことは明白である。

(7) 「参照元」

1-A、1-B は共に全体に比べ検索エンジン (Organic Search) からの流入の割合が低く、Web サイトやメルマガの URL リンクからの流入 (Referral) の割合が高いが、両者の間にはあまり開きがない。従って、それぞれの特徴が出ているとは言えない。

これらの結果を総合的に分析すると、条件を満たしている比較視点が少なく、第一基準は二分化基準としては適切ではないと結論付けられる。

この基準が条件を満たさなかった要因については次のことが考えられる。

- ユーザーは必ず同じパターンで検索するとは限らない。従って、時に 1-A の行動を、時に 1-B の行動を取っているユーザーが含まれている可能性がある。
- 検索回数や閲覧ページが多いことが必ずしも求める文献を的確に発見できる行動とは限らないのかもしれない。

この問題点を克服できる適切な基準を探し、再調査することにする。再調査の内容については、第 5 章で詳しく述べることにする。

二分化基準としては適切な設定とは言えなかったが、1-A と 1-B を合わせて検索を実行したユーザーと捉えることは可能であり、全体と比較して両者に共通してみられる特徴は DB の改善に活かせる可能性を秘めていると考える。従って、この章の最後にその特徴を記すことにする。

【検索を実行したユーザー】(1-A、1-B)

- パソコンの割合が高く、スマートフォンやフューチャーフォンの割合が少ない。(端末)
- 学術機関の割合が高い。(所属)

- 全体と比べて検索エンジンからの流入の割合が低く、Web サイトやメルマガの URL リンクからの流入の割合が高い。（参照元）

5 演算子使用による二分化基準（第二基準）

5.1. 方法

(1) 調査内容

第一基準の結果により、二分化基準による比較調査の難しさを改めて思い知らされたが、別の視点での二分化基準設定の可能性を追求したところ、「演算子の使用」による二分化基準を考案した。これは CiNii Articles 内で検索を行う際に、論理演算子や比較演算子など技術的な手法を用いることができるユーザーは「求める文献を的確に発見できるユーザー」と呼べるのではないかと考えられたためである。事実、第一基準の比較調査の結果では、AND 検索以外の演算子を使って検索するユーザーはごく少数ではあるが、的確な検索語を設定できている印象があった。そこで、この基準で再度いくつかの視点から比較調査を行い、それぞれのユーザーの傾向についてさらに深い分析を行った。

本基準では「求める文献を的確に発見できるユーザー」を「AND 検索以外の演算子を使って検索できるユーザー」(2-B) とし、「求める文献を的確に発見できないユーザー」を「演算子を使わずに検索したユーザー」(2-A) とした。(実際の設定条件方法は別紙の【付録 2】を参照)

この基準によるデータの抽出状況は表 8 のとおりである。2-A、2-B、全体と比較すると、本基準、特に 2-B がごく少数の絞られた層の傾向であることが理解できる。また、2-A と全体の差については、第一基準同様に CiNii Articles 内で検索を行っていないユーザーによるものだと推量される。

表 8 第二基準及び全体の件数

	2-A	2-B	全体
セッション数	13,355,409 件 (32.30%)	87,074 件 (0.21%)	41,353,793 件
ユーザー数	6,346,790 件 (32.24%)	66,381 件 (0.34%)	19,686,834 件
ページビュー数	151,717,593 件 (76.02%)	1,614,508 件 (0.81%)	199,565,613 件

※下段の () は全体との比率

(2) 傾向比較視点

比較調査は時間的制約もあり第一基準で行った比較視点すべてに対しての実施ができなかった。今回は主に第一基準比較で傾向に差が見られなかった項目を中心に、比較調査を行うこととした。

5.2. 結果

(1) 「端末」

図9はユーザーがCiNii Articlesへアクセスする際にどのような端末を使用してアクセスしているかを集計したものである。

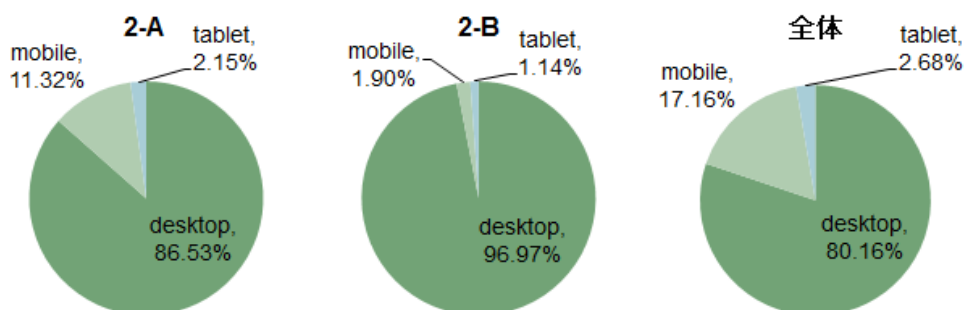


図9 端末の割合

パソコン（desktop）からのアクセスが、2-Aは86.53%であるのに比べ、2-Bは96.97%と2-Aより割合が大きくなっている。

(2) 「滞在時間」

表9はユーザーがCiNii Articlesで行った検索に使用した時間の平均値を求めたものである。

表9 平均セッション時間

	2-A	2-B	全体
平均セッション時間	9分3秒	18分23秒	3分46秒

2-Bでは、検索に使用した時間が2-Aに比べて倍以上の18分23秒となっており、かなり時間をかけていることがわかる。また、全体と比較するとさらに大きな開きが見られた。これは全体がCiNii Articles内で検索を行わず、検索エンジンからアクセスし文献の詳細

情報のみを閲覧したユーザーを多く含むからだと推察できるが、やはり 2-B がずば抜けて時間をかけていることには違いがない。

(3) 「所属」

図 10 は CiNii Articles にアクセスするユーザーのアクセス元（ドメイン）を学術機関とその他に分類し、ユーザーの所属と仮定した場合の結果である。

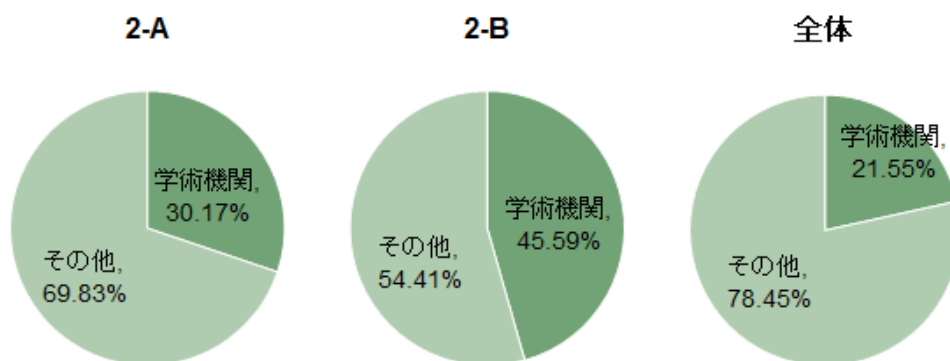


図 10 所属機関の割合

2-A に比べ 2-B の方が学術機関所属の割合は大きくなっていることがわかる。また、全体からするとさらに大きな隔たりが生まれており、学術機関所属者の方の検索技術が高いことが見て取れる。

(4) 「参照元」

図 11 はユーザーがどのような経路で CiNii Articles にアクセスしたかを示したものである。

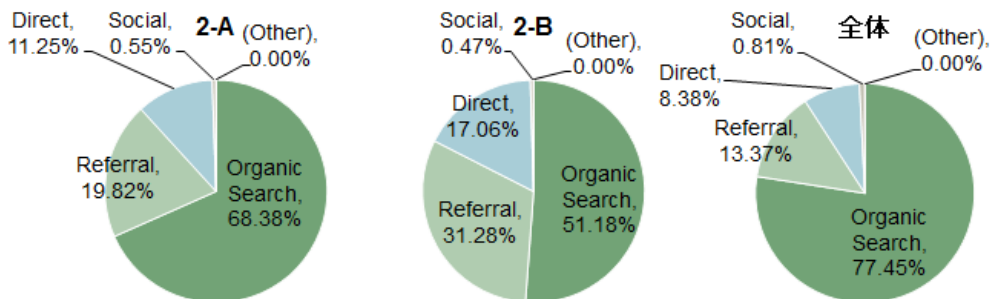


図 11 参照元の割合

2-B では外部参照（Referral）と直接来訪（Direct）を合わせた割合は半分程度を占め、

2-A と比べるとかなり差がついていることがわかる。さらにこの 2-B の外部参照の元となる URL を調べてみると、図 12 のような結果が判明した。

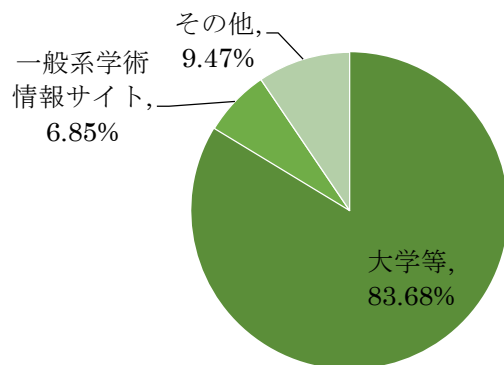


図 12 2-B Referral 内訳

8割以上が大学等の Web サイトであり、さらに一般学術系の Web サイトを加えると全体の 9割以上を占めることがわかる。このように、検索技術の上達者は学術機関の Web サイトをポータルサイト（調査の入り口となるサイト）して活用したり、ブックマークを使用したりするなど独自のルートで学術情報データベースへアクセスしている。

逆に、2-A や全体をみると、ほとんどが検索エンジンからの訪問であり、日常的な使用を感じさせない結果が見て取れる。

5.3. 考察

5.2 のそれぞれの結果について考察し、総合的に第二基準の妥当性を検証する。

(1) 「端末」

2-A、2-B についての比較で言えば、2-A はスマートフォンやタブレットの使用が一定数見られ、2-B ではその傾向はほとんど見られない。このように、大きくではないものの差を見ることができた。

(2) 「滞在時間」

2-A に比べ 2-B は倍以上の時間をかけていたことが判り、また全体からすると 2-B はさらに相当な時間をかけて検索をしていることがわかった。大きな差異を見出すことができる項目であった。

(3) 「所属」

2-B の学術機関所属者は 2-A の 1.5 倍の割合を示していた。また 2-B の中でも学術機関所属者が半分近く存在するなど、2-A との差が見られる結果であった。

(4) 「参照元」

2-B では外部参照と直接来訪を合わせた割合は半分程度を占め、2-A と比べるとかなり差がついていることがわかった。

以上の結果を受けて、第二基準の妥当性を検討すると、それぞれの項目について一定の差異を作り出すことができ、明らかな違和感を覚えるような結果ともならなかったことを踏まえ、適切な基準設定であったと結論付けられる。

本基準で得られた「求める文献を的確に発見できないユーザー」の傾向をまとめると、以下のとおりとなる。

【求める文献を的確に発見できるユーザー】(2-B)

- ほぼ全てのユーザーがパソコンから利用する。(端末)
- かなりの時間をかけて検索を行う。(滞在時間)
- 学術機関に所属するユーザーが一定数存在する。(所属)
- 学術機関の Web サイトやブックマークを使いアクセスするユーザーが一定数存在する。(参照元)

【求める文献を的確に発見できないユーザー】(2-A)

- B に比べると、スマートフォン・タブレット端末からの利用が多い。(端末)

これらが両者の明らかとなった差異であり、「求める文献を的確に発見できないユーザー」に適切な文献検索を行う方法を提案するための基礎と考える。

6 まとめ

「求める文献を的確に発見できるユーザー」と「求める文献を的確に発見できないユーザー」の存在を元に、第一基準・第二基準と調査を行ってきた。第一基準では残念ながら有効な比較基準として条件設定ができず、比較視点としての知見を得ることができなかった。第二基準では時間的制約により想定していた比較視点の全ての調査が実施できず、全

体を見通す比較結果を得ることができなかった。

各調査においては、十分・満足といえる結果とは言いがたいものにはなってしまったが、その分多くの経験と考察を得ることができたとも言える。ここからはそれらをまとめ、次への課題を提起することで、まとめとする。

(1) 調査結果

先にも述べたとおり、「求める文献を的確に発見できるユーザー」と「求める文献を的確に発見できないユーザー」の比較調査は第一基準による結果を採用できず、不完全な第二基準の結果のみをもって考察することとなった。まずは今一度、第二基準で得られた結果を下に挙げる。

【求める文献を的確に発見できるユーザー】(2-B)

- ほぼ全てのユーザーがパソコンから利用する。(端末)
- かなりの時間をかけて検索を行う。(滞在時間)
- 学術機関に所属するユーザーが一定数存在する。(所属)
- 学術機関の Web サイトやブックマークを使いアクセスするユーザーが一定数存在する。(参照元)

【求める文献を的確に発見できないユーザー】(2-A)

- Bに比べると、スマートフォン・タブレット端末からの利用が多い。(端末)

これらの内容を元にユーザーへの適切な文献検索方法の提案を検討したところ、やはり今まで見てきた学術情報データベースを通じて案内や誘導を行うのがよいのではないかという結論に至った。そこで次の3点について提案を行うものとする。

【的確な検索ができるユーザーを増やすための、学術情報データベースに対する提案】

- 1) 外部サイト(特に学術研究機関)に参照されやすいようなデータベース/Web サイト構築を行う。
- 2) パソコン向けのユーザーインターフェースを改良し、使いやすいデータベースを目指す。
- 3) 長時間の検索でもストレスにならないようなデザインを考慮しつつ、検索の継続を促

す示唆機能を備える。

上記3点について、それぞれ解説を行い、本テーマのまとめと考察とする。

- 1) 外部サイト（特に学術研究機関）に参照されやすいようなデータベース/Web サイト構築を行う。

第二基準「参照元」「所属」からの結果にもあるように、熟達したユーザーは独自の来訪ルートを持つようになる。その時、データベースの構造が参照しにくいような構成であれば、それだけでユーザーの利用を阻害してしまい、ひいては上達を妨げる要因にもなりうる。例えば Web サイト内の各ページや検索窓口の機能ごとに独自の URL を持つておけば、そこをピンポイントで必要とするユーザーの入り口を作ることになり、独自のルートを構築するきっかけとなるかもしれない。特に他の学術研究機関が学術情報データベースを参照・引用・紹介する際に、必要となるページの固定 URL がなければ、その研究機関のユーザー全体の利用を損なってしまうという結果にも結び付く。また、一般の検索サイトや SNS などにも開かれたサイトポリシーであれば、より多くのユーザーに利用される機会が増え、的確な検索ができるユーザーもまた増えていくことになる。

- 2) パソコン向けのユーザーインターフェースを改良し、使いやすいデータベースを目指す。

第二基準の「端末」「参照元」にもあるように、現在大半のユーザーはパソコンのブラウザからアクセスしている状況であり、このユーザーインターフェースの構成を検討し、ユーザビリティを高めることが快適な利用に直結することは疑う余地がない。特に様々な切り口での検索方法や、演算子などを使った検索テクニックを簡単に利用できるようなインターフェースデザインとなっていれば、ユーザーはデータベースを利用していっただけで、おのずと考え方やテクニックを身につけていくことにもなる。便利に使わせながら、ユーザーの成長も促すようなデザインを追求するというのも、一つの方向性なのではないかと思われる。

- 3) 長時間の検索でもストレスにならないようなデザインを考慮しつつ、検索の継続を促す示唆機能を備える。

第二基準の「滞在時間」「端末」をみると感じるのだが、実際に検索を行うと、どんなに正確で理想的な検索を行ったとしても時間だけは相応に使用することになってしまう。そこで長時間データベース内を回遊してもストレスにならないような画面遷移・ナビゲー

ション・サイトデザインが重要になってくる。そこには分かりやすさといったユーザビリティの基本となるものが入ってくるのは当然だが、さらにその低ストレス環境で続けて検索を行えるように、様々な示唆機能を盛り込むことも望みたい。具体的には、関連検索語の示唆や、2)でも挙げたような検索テクニックの提案、はたまたもっと別のユーザーのモチベーションを高める内容でも構わない。ユーザーに長時間の検索を体験させ、そのことに抵抗感をなくし、充実感を与えるためにも、データベース側でできることはまだまだあるはずである。

(2) 調査方法

本調査では、学術情報データベースを通じてユーザーの行動を理解し、新たな提案を行うために調査を行った。この調査を進める段階や結果から得た調査方法自体についての考察と課題も以下の通りまとめることとする。

・ ツール

本調査では Google Analytics をツールとして採用し、様々な調査や分析を行った。これは CiNii Articles という学術情報データベースの世界でもかなり巨大な部類となるデータベースで、かつアクセス数も相当な量になっているため、取り扱うことができるソフトウェアの候補がほとんどない状況であったためである。このツールが使えないのであればあとはサーバのアクセスログを使って分析するしかない状況であり、とても助かったのは事実である。しかし、この Google Analytics もあらかじめデータベース側に適切に設定を施しておくことで初めて有用なデータが取得できるようになるもので、この設定ができていなければ使うことができない。加えて、そもそもデータベースの構造的に設定することができない場合もありうるので、データベース/Web サイト側の要件をクリアさせることも大切である。また、Google Analyticsに限った話ではないが、無料で提供されるツールは仕様変更が突然行われるなど、分析項目の一貫性を保つことが難しくなる場合も想定されるので、それを見越した分析計画を立てることも重要である。

サーバのアクセスログを使った分析においても、事前の設定が適切になされていることが重要になる。デフォルトのサーバログでは追求できない情報も Cookie 等のデータが含まれていれば分析可能になる場合もあるので、サーバログの設定が対応できているか確認しておくにこしたことはない。ただし、サーバログによる分析の場合、分析対象となる以外のログデータも大量に存在するため、仕分けのためにプログラム言語による処理が必要に

なる可能性が高いので、そのためのスキルを身に付けることが必要となる。

・限界

データベースの利用ログを元にしてユーザーの行動を把握する場合、ユーザーの動きは記録から追うことができても、その意図を追うことはできない。その点においてこの調査手法は限界があり、ユーザーの意図を追うことを含めるのであれば聞き取り調査など別の手法を組み合わせる必要がある。

Google Analytics で長期間にわたるデータを分析するなど大量データが対象となる場合は、実データではなくサンプリングされたものが使われることになる。例えば実データに近い細かいパターンも分析したいとなると、このツールが使えなくなるので、別のツールを検討する必要がある。

・調査の課題

今回の調査では「検索回数」「演算子の使用」という基準（条件設定）にてユーザー層を仕分けて比較を行ったが、結果は前述のとおりであり、適切な基準設定の難しさを痛感させられたものであった。基準検討時には以下のような候補も挙げられ、今後さらに適切な基準を、データを見ながら検討していく必要があると思われる。

- 一般的で雑駁な1単語検索を行っているユーザーと3~4語で検索しているユーザー
- 何回も検索している（セッション回数が20回以上）ユーザーとあまり検索経験がない（セッション回数が3回以下）ユーザー

また、当たり前のことではあるが、ユーザーは1回の検索だけで検索行動を終了させるわけではなく、時に複雑で難しい検索、時に単純で簡単な検索を織り交ぜながら検索行動を進めるので、このようなユーザーをどうやって条件設定で捕捉するかを検討するのも大きな課題だと思われる。同様に、例えば学术研究機関所属というユーザーの中でも検索初心者もいれば上級者もいるように、ある程度同質的な要素を持っているユーザーをどのように仕分けていくかというのも、難しい問題になると思われる。

これらは実際に分析を進めていく中で徐々に明らかになっていく課題であった。今回は追求するには至らなかったが、ぜひ別の機会にでもこの結果を踏まえた調査が実施されることを期待したい。

本調査において得た様々な知見は紙幅に語りつくせないほどのものがあったが、その一部を報告書とその付録という形で記録にさせていただいた。末筆となるが、このような経験を研修という形で与えて下さった国立情報学研究所の皆様と様々なアドバイスを与えて下さった先生方へ感謝の言葉を申し上げてまとめとさせていただきます。

7 参考文献

1. 国立情報学研究所. “CiNii Articles - 日本の論文をさがす - 国立情報学研究所”. CiNii Articles. (オンライン), <http://ci.nii.ac.jp/>, (参照 2014-12-09).
2. 国立情報学研究所. “CiNii Articles - マニュアル - キーワードによる論文検索方法 - A-3. 複雑な検索方法について”. CiNii Articles. (オンライン), http://support.nii.ac.jp/ja/cia/manual_keyword, (参照 2014-12-09).
3. 国立情報学研究所. “CiNii Books - 大学図書館の本をさがす - 国立情報学研究所”. CiNii Books. (オンライン), <http://ci.nii.ac.jp/books/>, (参照 2014-12-09).
4. Google Inc.. “Google アナリティクス公式サイト - ウェブ解析とレポート機能 - Google アナリティクス”. Google アナリティクス. (オンライン), http://www.google.com/intl/ja_jp/analytics/, (参照 2014-12-09)
5. Google Inc.. “正規表現の概要 - アナリティクス ヘルプ”. Google アナリティクス. (オンライン), <https://support.google.com/analytics/answer/1034324?hl=ja>, (参照 2014-12-09)
6. Google Inc.. “Query Explorer — Google Analytics Demos & Tools”. Google Query Explorer. (オンライン), <https://ga-dev-tools.appspot.com/explorer/>, (参照 2014-12-09)
7. 佐藤翔ほか. アクセスログに基づく CiNii による本文提供とその利用状況の分析. 2012 年日本図書館情報学会春季研究集会発表要綱. 2012, vol. 2012, p. 73-76. <http://hdl.handle.net/2241/117031>, (参照 2014-12-9)

8. 佐藤翔. “日本の学術文献需要に占める機関リポジトリの貢献：CiNii 分析”. 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. 筑波大学, 2013, p.125-182, 博士論文. 入手先, 筑波大学学術機関リポジトリ, <http://hdl.handle.net/2241/118741>, (参照 2014-12-09)

9. 佐藤翔. 電子リソースのアクセスログ分析(ログの活用). 情報の科学と技術. 2013, vol.63, no. 2, p. 51-56. <http://ci.nii.ac.jp/naid/110009578960/>, (参照 2014-12-09)