Vol.2016-DBS-163 No.14
Vol.2016-IFAT-123 No.14
2016/9/15

学術情報検索における検索熟練度を考慮したユーザ行動の分析

岡崎 伸也[1,a)]　風間 一洋[1,b)]　篠田 孝祐[2,c)]　大向 一輝[3,d)]

概要：本稿では，学術情報検索である CiNii Articles のアクセスログを用いて，検索熟練度を考慮したユーザ行動の分析を行う．従来の情報検索システムの利用履歴の分析の研究においては幾つかの統計的な指標が用いられてきたが，本稿ではこれらをユーザの検索熟練度を判定する 4 つの指標として再構築し，2 つはユーザが入力したクエリの分析に，残りの 2 つはユーザの情報探索行動の分析に使用する．実際にその有効性を，CiNii Articles のアクセスログから同じ 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仮定して抽出したセッションを，書誌情報または著者情報に到達したか否かで成功と失敗に分類した場合に，指標に明らかな差があったことで確認した．CiNii Articles におけるユーザ行動を，検索熟練度に関する 4 つの指標とその相関関係で分析した結果，クエリに関しては短い単語の組み合わせや長い文章を用いる 2 種類のパターンを，ユーザの情報探索行動に関してはクエリ選択指向と検索結果閲覧指向の 2 種類のパターンがあることを確認した．

User Behavior Analysis Considering Search Proficiency in Academic Search

Shinya Okazaki[1,a)]　Kazuhiro Kazama[1,b)]　Kosuke Shinoda[2,c)]　Ikki Ohmukai[3,d)]

1. はじめに

近年，Web 上で電子化された論文や書籍を見つける仕組みとして，CiNii Articles や Google Scholar のような学術情報検索システムが広く普及している．こうしたシステムの利用者には，学術情報データベースが，特に専門性の高い文書で構成されていることや，同じ研究分野において類似論文が多数存在するなどの特性を持つことを考慮して，自らの求める学術情報に合致した専門用語を含むクエリの構築や，効果的な情報探索のスキルが要求される．しかし，普段 Web 検索システムに慣れ親しんでいるユーザであっても，このような検索スキルを十分に習得できていないことから，学術情報検索システムで有効な検索が行えず，目的の情報にたどり着くことができないことが多い．そこで，ユーザの検索熟練度を考慮して，アクセス履歴中のユーザの情報探索行動を分析することは，システムの問題点を把握・改善するために特に重要である．

本稿では，学術情報検索である CiNii Articles において，検索熟練度を考慮したユーザ行動の分析を行う．その際に，従来情報検索システムの利用履歴の分析のために用いられてきた幾つかの統計的な指標を参考にして，ユーザの検索熟練度の異なる側面を測るための 4 つの指標として再構築して使用する．このうち 2 つはユーザが入力したクエリに関する指標で，残りの 2 つはユーザの情報探索行動を反映した指標である．

実際に，それら 4 つの指標の妥当性を，CiNii Articles のアクセスログを用いて分析する．まず，IP アドレスと

[1] 和歌山大学大学院システム工学研究科 / Graduate School of Systems Engineering, Wakayama University
[2] 電気通信大学大学院情報理工学研究科 / Graduate School of Informatics and Engineering, The University of Electro-Communications
[3] 国立情報学研究所コンテンツ科学研究系 / Digital Content and Media Sciences Research Division, National Institute of Informatics
[a)] s171013@sys.wakayama-u.ac.jp
[b)] kazama@sys.wakayama-u.ac.jp
[c)] kosuke.shinoda@uec.ac.jp
[d)] i2k@nii.ac.jp

© 2016 Information Processing Society of Japan

ユーザエージェントの組をユーザ識別子と仮定して抽出したセッションを書誌情報または著者情報に到達したか否かで成功と失敗に分類し，それぞれのセッション集合の 4 指標に差が現れることを確認する．次に，CiNii Articles のユーザ行動を，4 指標の累積確率分布や，互いに関連する指標の間の関係を用いて分析する．

2. ユーザの検索熟練度の定量化

2.1 検索ログの分析

従来，Web 検索システムの検索履歴を用いてユーザ分析を行う際には，入力されたクエリ長（クエリあたりの単語数）などのクエリやセッション（一つの検索タスクを達成するための一連の操作）に関する様々な指標が用いられてきた．

例えば，Jansen らは Excite のクエリログを用いて，クエリ長や演算子，Unique，Modified，Identical という 3 種類のクエリの使用頻度，ユーザあたりのクエリ数・閲覧ページ数などを分析して，従来の情報検索システムに関する研究と比較することで，Web 検索エンジンのユーザは既存の情報検索システムのユーザよりも熟練度が低いことを示した [1]．Silverstein らは AltaVista のクエリログを用いて，クエリ長やセッション当たりのクエリ数，クエリあたりの閲覧画面数の割合や最大値・平均値・標準偏差を分析した [2]．風間らは，AltaVista の平均クエリ長が 2.35 語であるのに対して，ODIN の平均クエリ長が 1.42 語と短い原因が，英語の成句が日本語では 1 語の複合語として表されるからだと分析した [3]．Torres らは，AOL のクエリログと DMOZ の kIds & teens ディレクトリのデータを用いて子供の情報検索について，クエリ長・クリック順位・セッション長・セッション時間と頻度の関係を分析し，平均クエリ長が 3.23 語と通常の Web 検索より長くなり，その理由が文章として入力されるからであることを示した [4]．Church らは今まで利用されてきた多くの評価指標に対して，ヨーロッパのモバイル検索を，異なる年の分析結果や既存研究の Web 検索エンジンの分析結果と比較した [5]．この分析では，各クエリ長と一意なクエリ数，セッションあたりの一意なクエリ数とユーザの割合・セッション数などの分布に加えて，初期クエリ・同一クエリ・変更されたクエリ・単語なしの割合を分析した．

学術情報検索においても，同様の指標が利用できると考えられる．例えば，佐藤らは，国立国会図書館サーチのアクセスログを用いて，簡易検索と詳細検索，ファセット検索の利用状況について分析した．既存研究においては，統計値を求めることが目的であることから機械的なプログラムの影響を特に考慮しない場合が多かったが，この研究においてはファセット検索の絞り込みを詳細に分析するために，さまざまな手法を用いて人間の検索行動に絞り込んだ [6]．亀崎らは，Google Analytics のサンプリングされた一部の集計結果を用いて，CiNii Articles において，求める文献を的確に発見できるユーザとできないユーザを独自の定義に基づいて分類し，それぞれの検索行動の違いを，クエリ長やクエリに含まれる単語の長さの平均値，演算子の利用状況，平均セッション時間などの様々な視点から多角的に分析した [7]．

ただし，既存指標は利用履歴の統計を得ることを目的としていることから，ユーザの情報探索行動における特徴を必ずしもうまく把握できているとは言えなかった．そこで本稿では，ユーザの検索熟練度を判定するための要素を直接調べるための指標として再構築する．まず，既存指標の一部は検索が互いに独立していると仮定しているが，常にセッション単位で考える．次に，ある機能をどの程度活用できているかを明確にするために，セッション中の最大値を取る．さらに，例えばクエリにおいては専門用語は英語ではフレーズ，日本語では複合語として現れたり，検索セッションにはクエリの選択と検索結果の閲覧という二つの要素があるが，それぞれをうまく分離できるようにクエリやセッションの前処理方法を変更する．

2.2 検索熟練度の指標

以下に，セッション単位でユーザの学術情報検索の検索熟練度を測る 4 つの指標を示す．

(1) セッション中の最大クエリ長
(2) セッション中のクエリの最大単語長
(3) セッション中のクエリ選択回数
(4) セッション中の最大閲覧ページ数

(1) は絞り込み検索の使いこなし度合を判定する指標であり，複数の単語で AND 検索するようなクエリを用いた場合に大きくなる．(2) は専門用語の使いこなしを判定する指標であり，英語なら二重引用符で指定したフレーズ，日本語なら複合語がクエリに含まれる場合に大きくなる．(3) はセッション中で探索的な検索の実行を判定する指標であり，クエリの選択を繰り返すほど大きくなる．ただし，同一クエリで別の順位の検索結果を閲覧した場合には，新たなクエリの選択は行われていないとみなす．(4) はセッション中で検索結果を見た量を判定する指標であり，より多くの検索結果を調べるほど大きくなる．ただし，表示件数は変更できるので，20 件を 1 ページと見なす．

(1) と (2) ではユーザの入力クエリの良し悪しを，(3) と (4) ではユーザの情報探索の巧拙を異なる視点から評価する．これらの指標を組み合わせて分析することで，ユーザの文献検索能力や検索パターンを推定できると考えられる．

© 2016 Information Processing Society of Japan

Vol.2016-DBS-163 No.14
Vol.2016-IFAT-123 No.14
2016/9/15

3. 検索熟練度指標を用いたユーザの検索行動分析

3.1 CiNii Articles のアクセスログ

CiNii Articles[*1] は，国立情報学研究所が提供している学術論文や図書，雑誌などの学術情報データベースである．論文検索に加えて，著者検索，全文検索などの複数の検索機能が利用でき，ユーザは，興味を持つ論文や著者の概要を閲覧し，実際に論文を取り寄せたい場合には，CiNii Articles が所蔵していれば直接論文ファイルを取得し，そうでなければ機関リポジトリなどの外部リンクから取得することができる．このために，CiNii Articles は通常の Web 検索よりユーザの状態遷移が複雑なことに注意が必要である．評価には 2014 年 4 月 1 日から 2015 年 3 月 31 日までの Web サーバの Combined Log Format 形式のアクセスログを用いた．

まず，人間に限定するために，ユーザエージェントに bot，crawler などのキーワードを含んだエントリや，ユーザエージェントから Windows, OS X, Linux 上の Google Chrome，Firefox，Safari，Internet Explorer，Microsoft Edge 以外と判断したエントリを除去した．さらに，サービスや情報のアクセス以外の，画像やスタイルシートの取得などの副次的なアクセスや，ステータスコードが 200 以外のアクセスのエントリも除外した．

次に，CiNii Articles はユーザ固有の識別子を使わないために，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仮定して，Web サイトのトップページのアクセスから始まり，論文検索に限定してセッションを抽出した．これは，Google などの Web 検索エンジンの検索結果から書誌情報に直接ジャンプするユーザの場合は，情報探索行動の一部が観測できなくなるからである．ただし，異常な数のページ閲覧やクエリ選択を行っている場合は，機械的なプログラムであるとみなして除外した．セッションタイムアウトは，Silverstein らは 5 分 [2]，Torres らは 30 分 [4]，佐藤らは 1 時間 [6] に設定したが，本研究では検索に不慣れなユーザを想定して，30 分とした．この結果得られたユーザ数は 1,440,802 人，セッション数は 3,348,615 であった．

3.2 検索熟練度指標の妥当性評価

検索熟練度に関する 4 つの指標の妥当性を評価するために，セッションで論文情報，著者情報，書誌情報にたどり着いた場合を成功とみなして，セッションを成功と失敗に分類し，指標に差があるかどうかを分析した．なお，成功したセッション数は 2,484,837，失敗したセッション数は 863,778 であった．

2 種類のセッション集合に対する各指標の平均値を**表 1** に示す．全ての指標が検索に慣れているユーザが多いと思われる成功セッション集合の方が高く，検索熟練度を判定できることを確認できた．

表 1: 検索熟練度指標の平均値
Table 1 Means of four search proficiency measures

	最大クエリ長	最大単語長	選択回数	ページ数
成功	1.99876	8.06864	4.46227	3.09458
失敗	1.82867	6.89094	1.40763	2.57657

3.3 ユーザの検索行動分析

まず，ユーザの検索行動の全体的な傾向を知るために，ユーザが何回検索タスクを実行したかについて分析した．ユーザのセッション数の累積確率分布（CPD: cumulative probability distribution）を，**図 1** に示す．横軸はセッション数，縦軸はそのセッション数以上の検索を行ったユーザの存在確率である．この結果，セッション数が 500 回まではべき分布に従うことがわかり，さらにそれ以降の存在確率が急減することから，機械的なプログラムの影響はほぼ排除できていることも確認できた．

次に，ユーザの検索行動の詳細について知るために，各ユーザの検索タスクの熟練度に関係する 4 種類の特徴を各指標を用いて分析した．ユーザの各指標の累積確率分布を，**図 2**，**図 3**，**図 4**，**図 5** に示す．横軸は指標値，縦軸はその指標値以上のユーザの存在確率である．この結果，各指標値は，概ねべき分布に従う傾向があるが，2 つの興味深い特徴が発見された．

一つは，図 3 から，セッション内の最大単語長の平均値は 20 くらいから 80 くらいまでにグラフの傾きが大きくなっている部分があったことであり，このクエリ長の範囲の検索語としては長すぎる単語を使うユーザ数が多いことが示唆された．実際に，セッションごとにこの範囲の単語長を持つ検索の内容を調べると，論文の題名に相当すると思われる助詞を含む体言止めの文字列，論文の一部と思われる文章，そして URL で検索されていることがわかった．特に，最大単語長が 100 よりも大きい場合では，約 47.6%が URL であった．これは，例えば Safari を使って Google で一旦検索した後に，アドレスバーに表示されている検索クエリをコピーしようとすると，クエリではなく URL が取得されるため，誤ってそれを CiNii Articles の検索窓に入力するようなユーザが存在するからだと推測できる．ここで，クエリから URL を取り除いた場合のユーザの最大単語長の累積確率分布を**図 6** に示す．図 3 と比較すると，単語長が 50 以上の存在確率が大きく減少していることから，単語長が大きい領域において URL の存在がある程度の影響を与えていたことがわかる．

もう一つは，図 5 から，ユーザの閲覧ページ数は，10 ページごとに階段状に急減する傾向があることであったことである．ただし，CiNii Articles の表示検索結果数は標

[*1] http://ci.nii.ac.jp

© 2016 Information Processing Society of Japan

表 2: 表示検索結果数別のユーザ数
Table 2 User counts of the display number of search results

表示検索結果数	20	50	100	200
ユーザ数	1,241,686	22,090	33,202	78,878

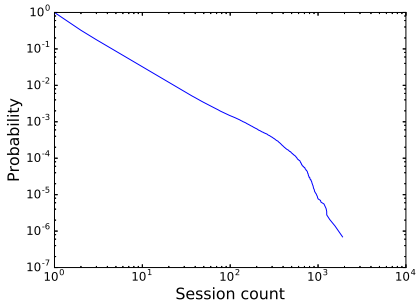

図 1: ユーザの平均セッション数の累積確率分布
Fig. 1 CPD of mean of user session counts

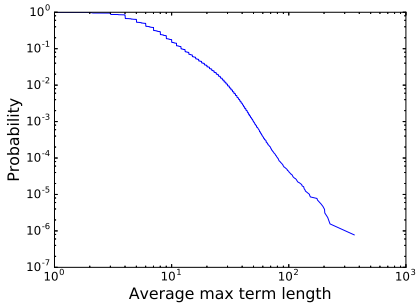

図 6: ユーザの平均最大単語長の累積確率分布 (URL を取り除いた場合)
Fig. 6 CPD of mean of max term lengths (URLs are removed)

準は 20 件だが最大 200 件に変更することもできるために，本稿では 1 ページの検索結果数が同じと仮定して換算したページ数を用いているために，最大の 200 件に変更して閲覧しているユーザが多い可能性が示唆された．そこで，実際にユーザが指定した表示検索結果数を分析した．**表 2** に表示検索結果数ごとのユーザ数を示す．この結果から，大部分のユーザはデフォルトの表示件数のまま使用するが，200 件に指定して網羅的に検索結果を閲覧するユーザも比較的多いことが確認できた．

3.4 検索熟練度指標の関係分析

ユーザの入力クエリの良し悪しやユーザの情報探索の巧拙という同一の目的で使用する異なる指標の間の関係を，ユーザごとの関連指標の平均値の分布をプロットして分析した．

まず，最大クエリ長と最大単語長の関係を**図 7** に示す．横軸がユーザの最大クエリ長の平均値，縦軸がユーザの最大単語長の平均値であり，ピアソンの積率相関係数は 0.13375 とほとんど相関がなかった．最大単語長が比較的

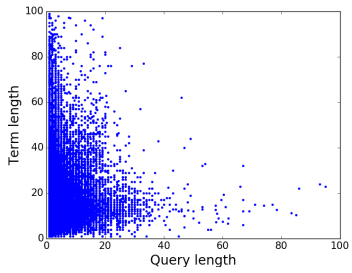

図 7: 最大クエリ長と最大単語長の分布
Fig. 7 Distributions of max query lengths and max term lengths

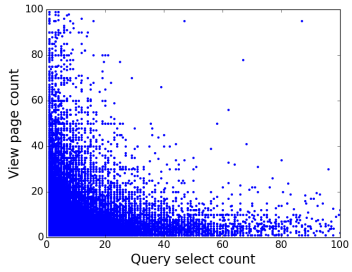

図 8: クエリ選択数と最大閲覧ページ数の分布
Fig. 8 Distribution of query changes and max page views

小さい領域では最大クエリ長との相関があるように見受けられるが，長い領域では最大クエリ長がとても小さい傾向がある．つまり，通常のフレーズや複合語を含む複数の検索語を使いこなして目的の情報が書かれている論文を探す検索パターンの他に，題名や文章に直接該当する論文を探すという検索パターンがあり，前者では専門用語や題名のような長い検索語も使いこなしているが，後者は複雑なクエリを使いこなせない傾向があると推測できる．また，その異なる性質のパターンを持つユーザの混在が混在することで，全体としてほぼ無相関に見えているのだと思われる．なお，実際に入力された長い文章や URL では検索結果が得られなかったので，特に単語長が長い領域は無効なクエリが多いと考えられる．

次に，クエリ変更回数と最大閲覧ページ数の関係を**図 8** に示す．横軸がユーザのクエリ変更回数の平均値，縦軸がユーザの最大閲覧ページ数の平均値であり，ピアソンの積率相関係数は 0.09268 とほとんど相関がなかった．この結果から，ユーザが，クエリを変更して検索して目的の情報を見つけるパターンと多くの検索結果を閲覧して情報を見つけるパターンのどちらを選ぶかは，比較的排他的であることがわかった．

4. おわりに

本稿では，学術情報検索において検索熟練度を考慮したユーザ行動の分析を行うために，従来の情報検索システムの利用履歴の分析の研究において用いられてきた統計的な指標を，ユーザの検索熟練度を判定する 4 つの指標として

© 2016 Information Processing Society of Japan

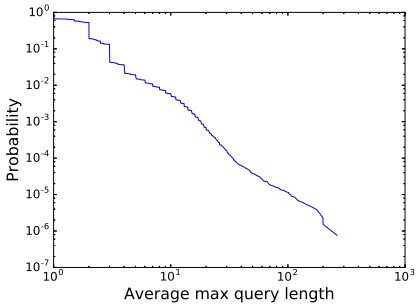

図 2: ユーザの平均最大クエリ長の累積確率分布
Fig. 2 CPD of mean of max query lengths

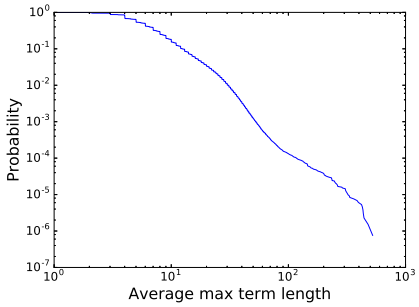

図 3: ユーザの平均最大単語長の累積確率分布
Fig. 3 CPD of mean of max term lengths

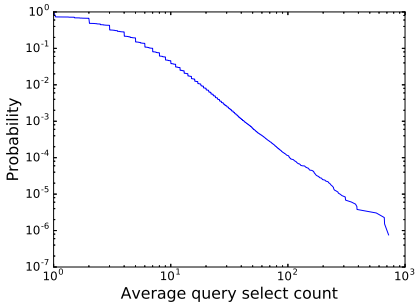

図 4: ユーザの平均クエリ選択数の累積確率分布
Fig. 4 CPD of mean of query selection counts

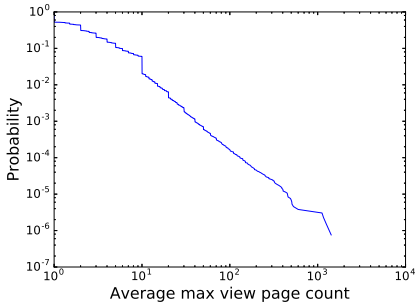

図 5: ユーザの平均最大閲覧ページ数の累積確率分布
Fig. 5 CPD of mean of max page view counts

再構築した．実際に，CiNii Articles のアクセスログから 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仮定して抽出したセッションを，書誌情報または著者情報に到達したか否かで成功・失敗に分類したデータセットを用いて比較し，それらの指標の妥当性を確認した

次に，CiNii Articles におけるユーザの検索行動をそれらの指標値やその相関関係を用いて分析した結果，論文の題名や引用文のような比較的長いクエリを入力するユーザや，一度に多くの検索結果を表示するユーザ，そしてクエリ選択指向と検索結果閲覧指向の異なる検索スタイルを持つユーザが存在することを確認できた．

今後は，アクセスログからユーザの入力ミスなどの不適切なクエリを取り除くさらなる前処理の必要性や，学術情報検索におけるユーザの時期的な検索熟練度変化の分析，また本稿で使用した指標によるユーザの検索熟練度に基づいた分類法の検討などを行う予定である．

謝辞 本研究は，国立情報学研究所公募型共同研究「学術情報サービスのユーザ検索履歴と共著関係を用いた検索支援技術の研究」の助成を受けた．

参考文献

[1] Jansen, B. J., Spink, A., Bateman, J. and Saracevic, T.: Real Life Information Retrieval: A Study of User Queries on the Web, *SIGIR Forum*, Vol. 32, No. 1, pp. 5–17 (1998).
[2] Silverstein, C., Marais, H., Henzinger, M. and Moricz, M.: Analysis of a Very Large Web Search Engine Query Log, *SIGIR Forum*, Vol. 33, No. 1, pp. 6–12 (1999).
[3] 風間一洋, 原田昌紀, 佐藤進也：ハイパーリンクとアンカーテキストを利用した情報検索とランキングの一手法，情報処理学会研究会報告 FI-59-3/DD-24-3，情報処理学会，pp. 17–24 (2000).
[4] Torres, S. D., Hiemstra, D. and Serdyukov, P.: Query Log Analysis in the Context of Information Retrieval for Children, *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*, ACM, pp. 847–848 (2010).
[5] Church, K., Smyth, B., Bradley, K. and Cotter, P.: A Large Scale Study of European Mobile Search Behaviour, *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '08)*, pp. 13–22 (2008).
[6] 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).
[7] 亀崎有紀子，渡邊伸彦：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，`http://www.nii.ac.jp/hrd/ja/ciws/report/h26/h26_gr2_final.pdf` (2014).

© 2016 Information Processing Society of Japan