第32回年次大会予稿

CiNii のログから見るユーザーアクセスグラフの計量分析

Statistical analysis of user access graph from CiNii logs

天野晃 [1*], 南山泰之 [1], 大波純一 [2], 遠藤晴義 [1], 長瀬友樹 [1], 山地一禎 [1]
Kou AMANO[1*], Yasuyuki MINAMIYAMA[1], Jun-ichi ONAMI[1], Haruyoshi ENDO[1], Tomoki NAGASE[1], Kazutsuna YAMAJI[1]

1 大学共同利用機関法人情報・システム研究機構 国立情報学研究所
National Institute of Informatics
〒101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: amano@nii.ac.jp
2 国立研究開発法人理化学研究所
RIKEN
〒305-0074 茨城県つくば市高野台 3-1-1
*連絡先著者 Corresponding Author

国立情報学研究所では同研究所が提供する NII Research Data Cloud のログを横断的に分析するための専用環境「統合ログ基盤」を整備してきた. 本発表では特に学術情報検索サービスの CiNii のログを対象にした当該基盤の機能を使った分析結果を報告する. 当該基盤の特徴は, (1)各アクセスログ(宛先 URI)に対してユーザー行動を端的に示すラベル(ボキャブラリ)を付与する機能, (2)アクセスログを意味のある単位でまとめる(「アクセスグラフ」を生成する)機能を持つことであり, 本研究ではこれらの機能を用いることにより個々のユーザーの行動をモデル化できる可能性を示す.

The National Institute of Informatics has been developing an environment called "Integrated Log Platform" for cross-sectional analysis of logs from the Research Data Cloud provided by the Institute. In this presentation, we will introduce some of the results of log analysis using the functions of this platform, especially for CiNii logs. The features of this platform are (1) creating a user action label (vocabulary) for a log (URI), and (2) organizing access logs into meaningful units ("access graph"). This research shows the possibility of these functions to model the behavior of individual users.

キーワード:アクセスログ, リファラーグラフ, ユーザーストーリー
Keywords: Access log, Referrer graph, User story

1 はじめに

国立情報学研究所では,同研究所が提供するNII Research Data Cloud（NII RDC） の各サービスのログを横断的に分析するための専用環境「統合ログ基盤」を整備してきた.当該基盤はユーザー行動をモデル化するための機能として,(1)各アクセスログ（宛先URI）に対してユーザー行動を端的に示すラベル（ボキャブラリ）を付与する機能,(2)アクセスログを意味のある単位でまとめる機能(後述する「アクセスグラフ」生成機能)を持つ.この機能を最大限に活用するためには,ボキャブラリとアクセスグラフの対応関係を整理することが課題となる.我々は,先行研究においてResearch Data Management（RDM）における行動を表現するための語彙であるRDMオントロジーを開発してきた.そこで本研究では,開発されたRDMオントロジーの語彙を参考にボキャブラリの設計を行う．さらに，これを応用したアクセスグラフの計量分析を実施する.ユーザー行動をボキャブラリでラベリングし,アクセスグラフを観察することで,ウェブサービスの全容を見据えた改善のための指標を得ることができる.

2 関連研究

ユーザーストーリーとは,システムやソフトウェアのユーザーが求める機能やケースを説明するものである[1]．ユーザーストーリーはシステム開発者,サービス対象となるユーザーの両者によって作成され,「誰が」「何のために」「何を実現したいか」の3要素を1つの文章として表現する.システム開発者は,収集したユーザーストーリーをもとにサービスの具体的なシステム機能要件を決定する.

先行研究において,著者らはNII RDCを対象にユーザーストーリーの収集,ペルソナ手法によるユーザー像の設定,ユーザーストーリーマッピングによる機能要件の抽出を実施してきた[2]．さらに，複数の分散されたサービスドメイン間で機能要件を共有するため,関心領域に含まれるデータの標準や仕様であるアプリケーションプロファイルの開発を進めている[3]．NII RDCアプリケーションプロファイルは，Dublin Core Metadata Initiative（DCMI）が2008年に発表した「The Singapore Framework for Dublin Core Application Profiles」に準拠しており[4]，構成要素として，機能要件，ドメインモデル，記述セットプロファイル,利用ガイドライン,符号化構文ガイドラインが作成されている．これらのアプローチは,複数の情報システム間で知識を共有するための手法として有用である一方で,その蓋然性の評価が次の課題となる.アプリケーションプロファイル作成時に想定したドメインモデルと実際のユーザー行動が一致しなければ,開発者はユーザーからのフィードバックを適切に理解することができず,ひいてはアプリケーション自体が使われなくなる恐れが生じる.

この課題を解決するためには,実際のユーザー行動に由来するデータを分析し結果とモデルを比較する必要がある.ユーザー行動分析によく用いられる方法としてログを利用するものがある[5]．UI操作の詳細ログを記録するようなシステムも開発されているが[6],現実的な方法としてはhttpサーバー

のアクセスログを対象とする方法が最もコストが低く主流である.アクセスログ分析の手法は多数開発されているが,中でも,ログに含まれるリファラーとアクセス先 URI を連鎖的に関連づける Referrer Graph(RG)アルゴリズム[7]（後述）が実現性の観点から有用である.本アルゴリズムとアクセス元情報を組み合わせて用いることにより,単一のアクセスログでは把握できない個別ユーザーを推定し,ユーザーごとの URI の訪問リストを得ることが可能となる.

3 対象データ

NII RDC サービスの検索基盤 CiNii，解析機能，人材育成基盤，JDcat（以下対象サービスと呼ぶ）のログのうち，2023 年 10 月 1 日から 12 月 25 日（ただし 10 月 8 日，11 月 13 日については予期しないログのため分析できなかった.）のアクセスであり，コードが 200 でないもの,画面部品の読み込みでないものおよびボットアクセスでないものを対象とした. CiNii は NII が提供する学術情報検索サービスであり，NII RDC の中で最大のアクセスがあるため,対象データ(元ログ)の 99%が CiNii へのアクセスである.

4 分析

4.1 分析方法

4.1.1 Referrer Graph/アクセスグラフ

RG アルゴリズムは,Web システムにおけるプリフェッチ用に設計された手法である[7].ユーザーのアクセスから学習することで,RG アルゴリズムは各リクエストの URI とリファラーを活用し,ウェブページ上のオブジェクト間の依存関係を解析するマルコフモデルを構築する.

本研究では RG アルゴリズムをウェブページ上のオブジェクト間の依存関係を分析するためでなく,ログを個別ユーザーごとの一連のアクセス単位に分割するために利用する.そのためマルコフモデルではなく単純な有向グラフが構築される.本研究ではこのアクセス単位をアクセスグラフと呼ぶこととする.

アクセスグラフの構成方法は以下の通りである.(1)対象データの 1 日間のログに対して，アクセス元 IP アドレス・ユーザーエージェントの組み合わせでログを分離する.(2)分離された各ロググループに対して RG アルゴリズムによる巨大グラフを生成する.(3)各グラフをエッジによりリンクしていないグラフに分ける.(4)各グラフに対してリファラーとアクセス先のアクセス時刻を比べ,アクセス先のアクセス時刻が後となっていないエッジを削除し,再度エッジによりリンクしていないグラフに分ける.(5)全てのグラフに対してノード数が 2 以上のグラフを選択する.

4.1.2 ボキャブラリ

対象サービスへのアクセス URI パターンに対応するユーザー行動のボキャブラリとして,ユーザーストーリーのボキャブラリより最適なものを選択して付与,さらにサービス（または基盤）のタグを追加したものとして定義した. この定義は,「サービス-主語-述語-目的語」の 4 要素の組み合わせとなる.CiNii では不特定多数のユーザーに利用許可をしているため主語についてはほとんど定義できず,URI の設計により述語と目的語については結びつきが強いため,基本的に

は CiNii の詳細ページ閲覧および CiNii の検索が主となる．なお，ボキャブラリが適用されない対象サービス外部（Google Scholar など）からのアクセスも存在する．ボキャブラリは全部で 58 件を定義した．表 1 にボキャブラリの一部例を示す．

表 1 ボキャブラリの例の一部．＊（アスタリスク）は主語が特定できないことを示す．

ボキャブラリ	概要	ユーザーストーリー上で対応する語彙
ci:*,view,top	検索基盤のトップページへのアクセス	アクセスする（https://purl.org/rdm/ontology/Access）
ci:*,explore,all	タイプを指定しない検索	調査する（https://purl.org/rdm/ontology/Survey）
ci:*,search,books	CiNii Books の検索	検索する（https://purl.org/rdm/ontology/Search）
ci:*,search,data	データ検索	検索する（https://purl.org/rdm/ontology/Search）
ci:*,confirm,detail	詳細ページへのアクセス	アクセスする（https://purl.org/rdm/ontology/Access）

RG アルゴリズムにより得られたアクセスグラフに含まれる URI にユーザー行動のボキャブラリを付与することにより，ユーザーの一連のアクセスに対してボキャブラリの組み合わせという合成的な属性が付与され，論理的には十分細やかな属性分けが可能となる．

4.2 指標

4.2.1 Zipf 分布における係数

この研究での Zipf 分布における係数とは，生産性（アクセス数）クラスのメンバー数と順位の関係を Zipf の分布$N(r) = Cr^{-a}$ （式 1）に従うとした場合のaを指す．ただし，$N(r)$は順位rにおけるメンバー数，メンバー数は同一アクセス数を持つユーザー数とする．Cとaがフィッティング係数である．

4.2.2 被覆比

被覆比とは本研究に特有の定義であり複数 URI への訪問の効率を示す．ユーザーが一連の URI に訪問する際，同一 URI には 1 回しか訪れないことが最も効率が良いと考えることができる．一方多くのユーザーは同一 URI に複数回訪問することがあり，「実際の訪問数」対「訪問したユニーク URI 数」の比が小さいほど効率的な訪問と言える．

5 結果

5.1 アクセス数

対象期間における全アクセス数は約 1.5 億件，1 日の平均アクセス数は約 164 万件であった．アクセス数の推移を見ると，日による差が大きくかつ不定期にアクセス数が急増する日が見られる（図 1）．

アクセスグラフを構成するアクセスのみを見ると，曜日ごとのアクセス数では土日にアクセスが落ち込み週央にアクセスが上がる傾向が見られた（図 2）．またアクセスグラフにおける平均アクセス数は 3.95，平均 URI 数は 3.83 であった．なお 10 月 8 日，11 月

13 日の値については前後日の平均値より補間した.

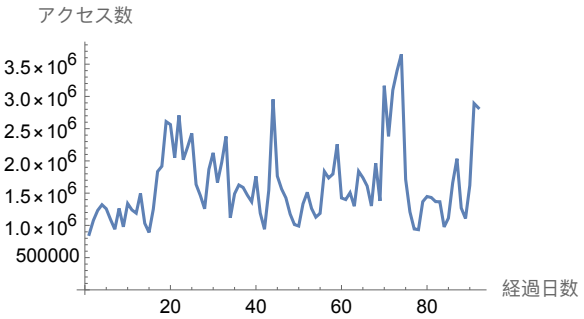

図 1 対象期間におけるアクセス数の推移. アクセスグラフを構成する前のボットアクセス等を除いたアクセス数.

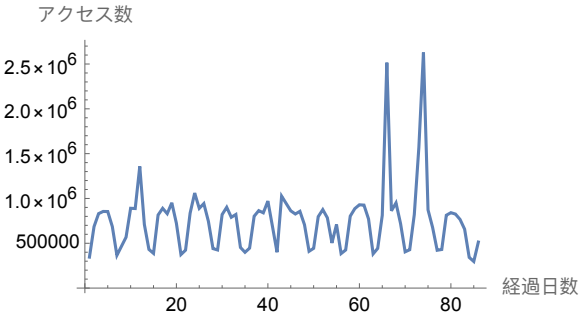

図 2 アクセスグラフを構成するアクセス数の推移.

5.2 Zip 分布における係数

各アクセスグラフにおけるアクセス数をクラスとした, メンバー数（アクセスグラフ数）と順位の式 1 による関係は, 対象期間全体では係数$a = 2.056$となった（図 3 a）. 一方期間を 1 日間とした場合の係数aの値は 3 前後となった. 10 月 1 日から 10 月 7 日までの係数の変化を表 2 に示す.

アクセスグラフにおけるユニーク URI 数に対して同様の分析を行ったところ対象期間全体では係数$a = 2.116$（図 3 b）, 1 日間における係数aの値は 3 前後となりアクセス数の場合と同様に長期間の観測よりも高い値が見られた（表 2）.

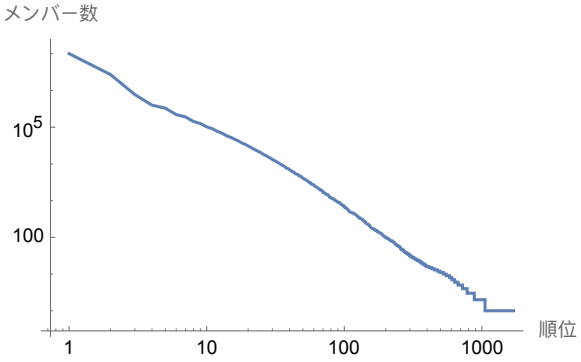

図 3 a. アクセス数をクラスとしたメンバー数と順位の関係の両対数プロット. Zipf 分布における係数は 2.056 であった.

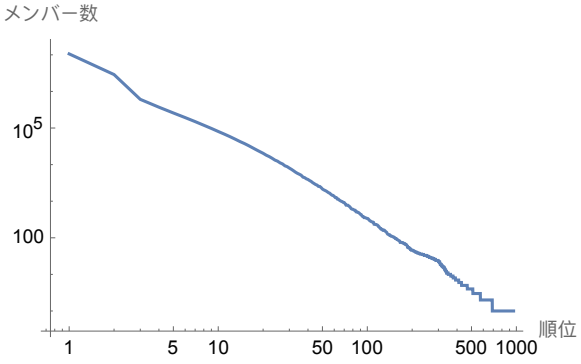

図 3 b. ユニーク URI 数におけるメンバー数と順位の関係の両対数プロット. Zipf 分布における係数は 2.116 であった.

5.3 被覆比

被覆比の対象期間全体での平均は 1.17 であり, 1 日の平均は概ね 1.15 から 1.2 となり, まれに値の落ち込む日が見られた(図 4). 週による周期が見られ, 土日に値が落ち込み週央に値が上がる傾向が見られた. なお 10 月 8 日, 11 月 13 日の値については前後日の平均値より補間した. 表 2 に調査開始 1 週間の 1 日の平均被覆比を示す.

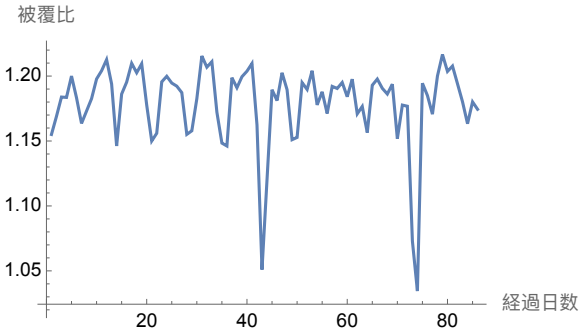

図 4　対象期間における被覆比の 1 日間の平均．週による周期が見られ，週央に値が上がる傾向がある．

5.4 ボキャブラリによるアクセスグラフの類型化

アクセスグラフに付与されたボキャブラリの組み合わせの数は，対象期間全体で 1039 であった．1 日間での組み合わせ数は 200 から 300 程度であった（表 2）．ボキャブラリの組み合わせごとのメンバー数とランクの式 1 による関係において係数$a = 3.221$であった（図 5）．メンバー数上位の 50 クラス（全体の 99%以上を占める）において階層化クラスタリングを行ったところ，組み合わせパターン間の類似度は低くすべてのクラスが大きい距離でクラスターを構成した（図 6）．

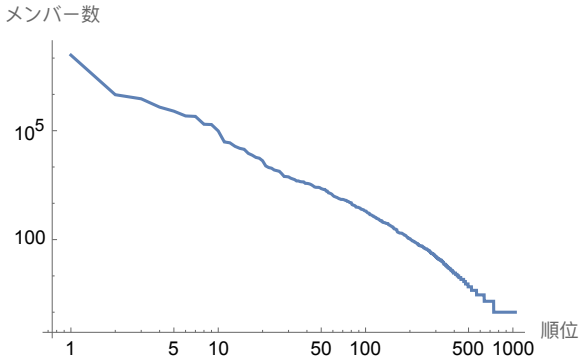

図 5 各ボキャブラリの組み合わせのメンバー数対順位の両対数プロット．Zipf 分布における係数は 3.221 であった．

ボキャブラリの組み合わせとして最も多かったものは詳細ページのみのアクセス，すなわち対象サービス外のページから CiNii の詳細ページへの連鎖（その後離脱）であった．いくつかのケースを目視で確認したところ，そのほとんどが Google 等の外部からのアクセスと CiNii の内部の更新データ間を繋ぐリダイレクトであった．

6 考察

アクセスグラフは個別ユーザーの一連のアクセス連鎖であると推定できるが，(1)同一ユーザーのアクセスであるがリファラーの連鎖が捉えられずに別ユーザーと推定された，(2)別ユーザーであるがアクセス元情報が同じ，かつ訪問先が同じであることによって，誤同定によって同一ユーザーと推定された可能性がある．ユーザーを特定可能なログを用いて分析したところ，同一ユーザーのアクセスから複数のアクセスグラフが生成される場合が多いことから，(1)の状況が起こりやすいと思われる．

アクセス数に関する Zipf 係数は長期には 2 前後，短期（1 日間）では 3 前後の値を得た．この差の理由については，アクセスグラフの構成を 1 日間のデータにとどめており，大きなアクセスグラフへの成長がないためである．学術生産性に係る Zipf 係数は 1.5 から 3 程度と言われており，この知見とも一致する結果を得た．

ボキャブラリ組み合わせの最頻は{詳細ページへのアクセス}であり，この場合に限ったアクセスグラフの最頻と思われるパターンは「対象サービス外部」->「CiNii の詳細

ページ」であった（図 7 a）．ボキャブラリ組み合わせの最頻 2 位は{タイプを指定しない検索，詳細ページへのアクセス}でありこの場合に限ったアクセスグラフの最頻と思われるパターンは「CiNii の検索」->「CiNii の詳細ページ」となるパスであるが，検索開始から詳細ページの閲覧に至るまでに直進的ではない行動が見られた（図 7 b），ボキャブラリの出現だけではユーザー行動の複雑な様相を類型化することは難しく，これについてはグラフ構造のインデックス等を用いた分析が有効となるかもしれない．

図 7 a．ボキャブラリ組み合わせの最も多いアクセスパターン．対象サービス外から直接詳細ページにアクセス．

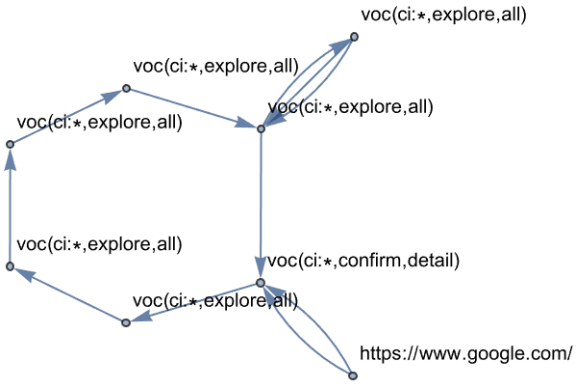

図 7 b．ボキャブラリ組み合わせ 2 位に対応するアクセスパターンの一例．ボキャブラリの組み合わせは 2 つであるが，多くの異なるページを訪れている．

ページ訪問の効率性を示す被覆比は 1.1 から 1.2 程度の安定した値を得た．一方，CiNii の GUI にはタブの切り替えによりページ遷移を効率化する機能があり，これが被覆比を押し上げた可能性がある．GUI の利用と被覆比に関しては一般的にも研究が少ないため，今後は NII RDC サービス内外の比較といったより詳細な分析を必要とする．

7 データ利用可能性宣言

本研究の結果を裏付けるデータは本研究のためにライセンスに基づいて使用されたためその利用には制限があり，一般には公開されていない．しかし，合理的な理由があれば国立情報学研究所の許諾を得て，著者から入手することが可能である．

謝辞

国立情報学研究所 クラウド基盤研究開発センターの皆様には統合ログ基盤の構築に多大な協力をいただきました．感謝申し上げます．

参考文献

[1] Cohn, M. (2004). User stories applied: For agile software development. Addison-Wesley Professional.
[2] 常川真央，朝岡誠，大波純一，河合将志，林正治，南山泰之，藤原一毅，込山悠介．研究行動に沿ったリサーチデータマネージメントサービスのシステム機能要件に関する検討，情報処理学会研究報告，Vol. 2020-IOT-51，No. 10，pp. 1-11，2020.
[3] 南山泰之，林正治，藤原一毅，大波純一，横山重俊，込山悠介，山地一禎．オントロジー技術を用いた NII RDC アプリケーションプロファイル開発に向けて．情報知識学会誌，Vol. 33，No. 2，pp.212-220，2023.
[4] Mikael Nilsson, Tom Baker, Pete

Johnston. The Singapore Framework for Dublin Core Application Profiles. https://dublincore.org/specifications/dublin-core/singapore-framework/, (2024年4月12日参照).
[5] 佐藤翔. リポジトリログ分析による学術情報流通の諸側面. 金沢大学創基150年記念「講演・シンポジウム」シリーズ(特別回). 2010.
[6] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, Ben Y. Zhao. Unsupervised Clickstream Clustering for User Behavior Analysis. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 225-236. 2016.
[7] B. de la Ossa, A. Pont, J. Sahuquillo, J. A. Gil. Referrer graph: a low-cost web prediction algorithm. Proceedings of the 2010 ACM Symposium on Applied Computing. 2010. pp. 831-838. https://doi.org/10.1145/1774088.1774260.

表2 調査期間のうち最初の1週間の各属性値.

日付	2023.10.1	2023.10.2	2023.10.3	2023.10.4	2023.10.5	2023.10.6	2023.10.7
アクセスグラフ数	121359	183802	190096	189729	174420	167499	116212
Zip係数(アクセス/グラフ)	3.205	3.005	2.862	2.846	2.661	2.838	3.163
Zip係数(ユニークURI/グラフ)	3.212	3.003	2.860	2.847	2.690	2.845	3.159
平均被覆比	1.155	1.169	1.184	1.184	1.200	1.183	1.164
ボキャブラリの組み合わせ数	198	277	297	295	302	250	216

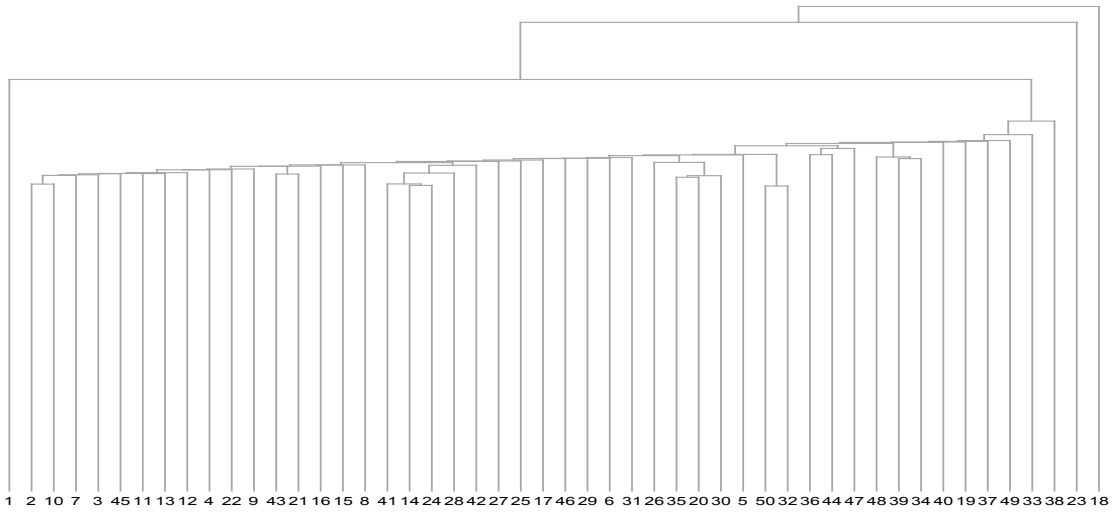

図6 2023.10.01-2023.12.25におけるアクセスグラフのボキャブラリ組み合わせ出現数上位50の階層化クラスタリング. ユークリッド距離による最近隣法を使用. 数字は出現数順位.