Access Log Analysis for Institutional Repository

馬場, 謙介
九州大学附属図書館研究開発室

伊東, 栄典
九州大学情報基盤研究開発センター

廣川, 佐千男
九州大学情報基盤研究開発センター

https://doi.org/10.15017/20104

出版情報：九州大学附属図書館研究開発室年報. 2010/2011, pp.5-8, 2011-08. Kyushu University Library
バージョン：
権利関係：

論文

リポジトリのアクセスログ解析

馬場 謙介[†]，伊東 栄典[‡]，廣川 佐千男[§]

<抄録>

機関リポジトリは，セルフアーカイビングによる研究成果へのオープンアクセス実現のひとつの手段である．しかし，多くの機関リポジトリにおいて，登録されている文献の数は，その機関の研究成果の数に比べて非常に少ない．登録文献数が少ない理由のひとつは，ほとんどの研究者にとって研究成果を登録する動機がない，つまり，機関リポジトリの効果が明らかでないことだと考えられる．本稿の著者らは，機関リポジトリに研究成果を登録した研究者へのフィードバックシステムを開発している．本稿では，機関リポジトリのアクセスログからの知識発見を試みている．「誰が，いつ，どんな文献を利用したか」という情報は，機関リポジトリの利用者への文献の推薦に利用できる．本稿は，より高度な解析へ向けての準備として，九州大学の機関リポジトリのアクセスログについて単純な共起解析を行った結果を示している．

<キーワード> 機関リポジトリ，アクセスログ，共起解析，推薦，可視化

Access Log Analysis for Institutional Repository

Kensuke BABA, Eisuke ITO and Sachio HIROKAWA

1. はじめに

学術情報への「オープンアクセス[21]」とは，学術論文等の研究成果の無料での利用を実現するものである．研究者にオープンアクセスを義務づける研究機関の数は増加している[7]．特に，公的機関からの資金を受けた研究については，その成果を公開するべきであるという考えが一般的になりつつある．例えば，米国国立衛生学研究所は，この機関による資金を受けた研究についてのオープンアクセスを求めるポリシーを示している[8]．オープンアクセスを実現するための方法のひとつは「セルフアーカイビング」であり[16]，研究成果を蓄積・公開するシステムが**リポジトリ**である．ある研究機関における成果のためのリポジトリは**機関リポジトリ**と呼ばれ，特定の研究分野のためのもの（例えば，arXiv [1]）は**主題リポジトリ**と呼ばれる．

研究成果のリポジトリの登録率は非常に少ないと推測される．例えば，九州大学の機関リポジトリ（QIR）[5]については，2011 年 6 月現在，登録文献数が世界で 76 位[6]であるのに対し，登録率は高々30% [11]である．この状況を改善するには，オープンアクセスとリポジトリの考えを広く周知するだけでなく，リポジトリの効果を実際のデータに基づいて明らかにする必要がある．

リポジトリの大きな特徴のひとつは，文献利用の詳細をアクセスログとして観測できることである．アクセスログから意味のある情報を発見することができれば，それはリポジトリの効果のひとつと考えることができる．基本的な解析（例えば，各文献や著者についてのアクセス数の集計やアクセス元の地域の解析等）は DSpace [2]の標準的な機能や Google Analytics [3]によって実現できる．機関リポジトリのアクセスに関する量的解析についての研究はいくつかある[13,19]が，リポジトリの効果を明らかにするためにはより詳細な質的解析が必要である．いくつかのオープンアクセスジャーナル[9,17,18,22]や主題リポジトリ[14,15]については，アクセス数と被引用数との相関が報告されている．しかし，機関リポジトリについての同様の解析では意味のある相関は見つかっていない[20].

機関リポジトリのアクセスログからの知識発見が困難である理由のひとつは，登録文献の数と文献へのアクセスの数が十分でないからであると考えられる．さらに，オープンアクセスジャーナルや主題リポジトリと比べ，機関リポジトリは広い分野についての研究成果を蓄積しており，各文献へのアクセスは少なくなってしまう．よって，我々は同一利用者についてのアクセスの共起に着目した. つまり，「ある利用者がどういう文献へ同時にアクセスしたか」を考えることにより，少ないアクセスからでも意味のある知識の発見が可能ではないかと考えた．我々は，実際に QIR のアクセスログに対して共起解析を適用する．本稿では，その解

[†] ばば けんすけ　九州大学附属図書館研究開発室　　E-mail: baba@lib.kyushu-u.ac.jp
[‡] いとう えいすけ　九州大学情報基盤研究開発センター　　E-mail: itou@cc.kyushu-u.ac.jp
[§] ひろかわ さちお　九州大学情報基盤研究開発センター　　E-mail: hirokawa@cc.kyushu-u.ac.jp

析の準備として，各文献のアクセス数と同一利用者についての共起の数を調べる．これによって，共起解析による知識発見の可能性を調べる．

本研究の主なアイデアは，アクセスログの解析結果を著者に見せることで，研究成果の機関リポジトリへの登録を促すことである．本稿はアクセスログの解析のケーススタディであり，著者にとってどういう解析が効果的かを調べる研究の最初の試みである．この研究に基づいて，研究者への動機付けの観点からアクセスログの様々な解析を評価する事ができる．

2. 研究の動機

本節では九州大学学術情報リポジトリ（QIR）[5]の詳細とアクセスログ解析の目的を示す．

2.1. QIR

QIR はDSpace [2]を基にした機関リポジトリであり，九州大学附属図書館によって 2006 年より管理されている．2011 年 3 月現在の登録文献数は約 16,000 件である．一般に，機関リポジトリは論文の書誌情報に加え，本文を保存するものである．登録文献の内訳を表 1 に示す．最も多い文献の種類は紀要論文であり，その割合は約 72%である．つまり，QIR の登録文献のうち多くは他の論文誌等で公開されていないものであり，被引用数による論文の評価が困難である．

表 1: QIR に登録されている文献の内訳（2011 年 3 月 31 日現在）．

文献の種類	文献の数
学術雑誌論文	1,464
学位論文	121
紀要論文	11,825
会議発表論文	1,024
会議発表資料	255
書籍	135
技術報告	547
研究報告	240
一般雑誌記事	301
プレプリント	160
教材	39
その他	544
合計	16,655

九州大学の研究者データベース[4]に登録されている学術論文（のタイトル）の数は約 70,000 件であり，複数の著者による重複した登録を考慮しても少なくとも 50,000 件の異なる論文が登録されている[11]．つまり，多くの論文が九州大学で執筆されているが QIR には登録されていないということである．もし，QIR への登録の効果が明らかになれば，これらの埋もれた論文が QIR に登録されることが予想できる．我々は，この研究者データベースと QIR とをリンクするシステムを開発した[10,12]．このシステムは，我々が本稿で取り組む問題に対するもうひとつの解法であると言える．

2.2. 解析の目的

本研究の目的は機関リポジトリのアクセスログから意味のある情報を見つけることである．特に，機関リポジトリの利用者が文献から何を知ろうとしているかを推定する．この利用者の趣向についての情報があれば，ある種の推薦が可能になり，これによって機関リポジトリの利便性が高まる．機関リポジトリに研究成果を登録した研究者にとっては，この利用者の趣向を一般的な研究動向を知るための材料として活用することができる．

アクセスログについてのいくつかの基本的な解析は，DSpace の標準的な機能や Google Analytics [3]等によって行うことができる．例えば，DSpace の基本的な機能により，各文献へのアクセス数をカウントし，それについての機関リポジトリ内でのランキングを表示することができる．Google Analytics では，アクセス元の地域や，検索結果からのアクセスであれば検索キーワード等についての統計情報を得ることができる．しかし，利用者の趣向を知るには，より詳細で高度な解析が必要である．特に，アクセス数が少ない機関リポジトリにおいては，少ないサンプルから知識を発見することが求められる．

我々は，アクセス数の単純な合計だけでなく，アクセスの共起に着目した．本稿では，共起解析による知識発見の可能性を確かめるために，実際に QIR のアクセスログに対し共起解析を行った．この際，同日の同一アドレスからのアクセスを共起アクセスとした．最初の取り組みとして，ボット等による機械的なアクセスを除いた上で，解析を行うのに十分な数の共起アクセスがあることを確かめた．

3. 実データの解析

3.1. ログデータの前処理

実験対象のデータは，2008 年 6 月から 2009 年 12 月までの QIR のアクセスログである．総アクセス数は 23,847,393 件であった．

まず前処理として，アクセスログから，Web クローラー等の所謂ボットによるノイズの除去を行った．具体的には，文字列「bots」を含むアクセスを除去した．これによりデータの件数は 14,870,045 件（約 62%）になった．さらに詳しいノイズ除去も考えられるが，機械的なアクセスは次の小節の共起の数についての制限によっても除去される．

3.2. 共起の数

同日の同一アドレスからのアクセスを同一利用者によるアクセスと定義した．表 2 はアクセス頻度と利用者の数の分布を表している．我々が着目したのは，一日のアクセス数が 2 から 50 である 87,628 の利用者（表 2 の*印）である．

表 2: 同一アドレスからのアクセスの数

アクセス数 n	利用者数	a に対する割合	b に対する割合
0<n	[a]852,346	100.0 %	-
n=1	763,882	89.6 %	-
1<n	[b]88,464	10.4 %	100.0 %
1<n<51	[*]87,628	10.3 %	99.1 %
50<n	836	-	0.9 %

アクセス数の多い利用者と 1 回のみの利用者は解析の対象にしなかった．非常に多い利用者はボットであると予想される．人間によるアクセスであるとしても，あまりに頻度の高いアクセスは論文についての興味を推定するという目的に有用ではないと考えた．1 回のみのアクセスは共起解析の情報としては意味が無い．結果として，QIR のアクセスログには，ボットによるノイズを除いた後でも十分な数の共起があることがわかった．

3.3. 共起解析

QIR のアクセスログについての共起解析の結果を図 1 に示す．このグラフで，ノードは文献を，2 つの整数はそれぞれアクセス数と文献の識別子を表している．ここで，丸いノードは文献を四角いノードはリストセットを表している．矢印は，先のノードに対応する文献が元のノードに対応する文献とともに利用されたことを表している．例えば，最も上の部分グラフは，文献 2961 へのアクセス数は 19 であり，このうち 2 人の利用者が文献 10651 も利用していることを表している．このグラフを作成するための初期ノードは，グラフ中で*印が付いたものであり，これらはキーワードによる検索等で決められる．

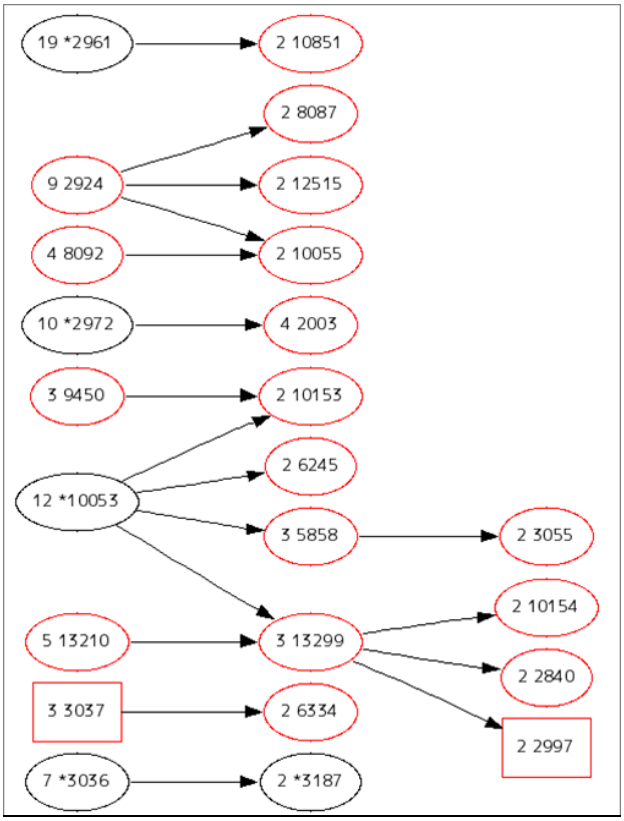

図 1: 共起解析によって作成されたグラフの例

4. おわりに

共起解析の有用性を確かめるために，九州大学の機関リポジトリの実際のアクセスログデータについて解析を行った．アクセス数の単純な合計とともに共起アクセスの数を調べた．結果として，十分な数の共起が存在することがわかった．

今後の研究として，QIR のアクセスログに対してより詳細な共起解析を行う予定である．さらに，他の機関リポジトリのアクセスログについても同様の解析を行う予定である．また，九州大学附属図書館では，購読契約をしている電子ジャーナルについても閲覧履歴を記録しているおり，このデータについても同様の解析を行う予定である．

参考文献

[1] arXiv. http://arxiv.org/, [Accessed Jun. 2, 2011].

[2] DSpace. http://www.dspace.org/, [Accessed Jun. 2, 2011].

[3] Google Analytics. http://www.google.com/intl/en/analytics/, [Accessed Jun. 2, 2011].

[4] 九州大学研究者情報. http://hyoka.ofc.kyushuu-u.ac.jp/search/index_e.html, [Accessed Jun. 2, 2011].

[5] QIR: 九州大学学術情報リポジトリ. https://qir.kyushu-u.ac.jp/dspace/, [Accessed Jun. 2, 2011].

[6] Ranking Web of World Repositories. http://repositories.webometrics.info/, [Accessed Jun. 2, 2011].

[7] ROARMAP: Registry of Open Access Repository Material

Archiving Policies. http://www.eprints.org/openaccess/policysignup/, [Accessed Jun. 2, 2011].

[8] Analysis of comments and implementation of the NIH public access policy. The National Institutes of Health, 2008. http://publicaccess.nih.gov/analysis_of_comments_nih_public_access_policy.pdf, [Accessed Jun. 2, 2011].

[9] Deciphering citation statistics. Nature Neuroscience, 11(6):619, 2008.

[10] K. Baba, M. Mori, and E. Ito. A synergistic system of institutional repository and researcher database. In Proceedings of the Second International Conferences on Advanced Service Computing (SERVICE COMPUTATION 2010), pages 184--188. IAIRA, 2010.

[11] K. Baba, M. Mori, and E. Ito. Identification of Scholarly Papers and Authors. In Proceedings of the Third International Conference on 'Networked Digital Technologies' (NDT 2011), Communications in Computer and Information Science, volume 136, pages 195--202. Springer-Verlag, 2011.

[12] K. Baba, M. Mori, E. Ito, and S. Hirokawa. A Feedback System on Institutional Repository. In Proceedings of the Third International Conference on Resource Intensive Applications and Services (INTENSIVE 2011), pages 37--42. IAIRA, 2011.

[13] A. I. Bonilla-Calero. Scientometric analysis of a sample of physics-related research output held in the institutional repository strathprints (2000--2005). Library Review, 57(9):700--721, 2008.

[14] T. Brody, S. Harnad, and L. Carr. Earlier web usage statistics as predictors of later citation impact. Journal of the American Society for Information Science and Technology, 57(8):1060--1072, 2006.

[15] P. M. Davis and M. J. Fromerth. Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? Scientometrics, 71(2):6203--215, 2007.

[16] S. Harnad, T. Brody, F. Vallieres, L. Carr, S. Hitchcock, Y. Gingras, C. Oppenheim, H. Stamerjohanns, and E. Hilf. The access/impact problem and the green and gold roads to open access. Serials Review, 30(4):310--314, 2004.

[17] D. E. O'Leary. The relationship between citations and number of downloads in decision support systems. Decision Support Systems, 45(4):972--980, 2008.

[18] T. V. Perneger. Relation between online ¥hit counts" and subsequent citations: Prospective study of research papers in the BJM. BMJ, 329:546--547, 2004.

[19] P. Royster. Publishing original content in an institutional repository. Serials Review, 34(1):27--30, 2008.

[20] 佐藤翔, 冨本壽子, 逸村裕. 論文の被引用数と機関リポジトリにおけるダウンロード数の関係, http://www.tulips.tsukuba.ac.jp/dspace/handle/2241/104229, [Accessed Jun. 2, 2011].

[21] P. Suber. Open access overview. Open Access News, 2007. http://www.earlham.edu/~peters/fos/overview.htm, [Accessed Jun. 2, 2011].

[22] B. A. Watson. Comparing citations and downloads for individual articles. Journal of scientific research on biological vision, 9(4):1--4, 2009.