生存分析からみた学術論文 PDF ファイルのクローリング

石田栄美(駿河台大学)*　宮田洋輔(慶應義塾大学大学院)　池内 淳(筑波大学)
安形 輝(亜細亜大学)　野末道子(鉄道総合研究所)　上田修一(慶應義塾大学)
*e-mail: emi@surugadai.ac.jp

【抄録】 学術論文専門の検索エンジンにとってPDFファイルのクローリングは重要な課題である。しかし、ウェブページと異なりPDFファイルでは生存分析が行われていないため、クローリング頻度、収集範囲などの戦略を策定することが困難である。本研究ではクローラーによるアクセス調査、人手によるファイル追跡調査から構成される生存分析を行った。論文ファイルは非論文よりも生存率が高く、たとえ移動していても移動先を再発見しやすいことが明らかとなった。最後に調査結果に基づき論文ファイルの効率的なクローリング手法を検討した。

1. はじめに

学術情報のウェブでの提供が一般的になってきたが、Googleなどの汎用検索エンジンでは検索結果に学術情報が埋没してしまうため、学術情報のみを探すことが困難である。そこで本研究グループでは学術情報に特化した検索エンジン「ＡΛＨΘＥＩＡ（以下、アレセイア）」の開発を進めている。これは公開されたアルゴリズムにより学術論文を自動識別、検索できるシステムである。Google Scholar BETAなど、他の学術情報専門のエンジンと比較し、評価した結果、一定の有用性を示すことができた[1)]。

なお、ウェブ上に公開されている学術情報の多くはPDFファイル形式であることから[2)]、アレセイアは収集・検索対象をPDFファイルとしている。

アレセイアは試作段階であり、現在のところ、定期的なクローリングではなく一括して収集したPDFファイル群を対象としている。実用化のためにはPDFファイルの効率的なクローリング手法の開発が課題となる。これは、PDFファイルは一般的にHTMLで書かれたウェブページと比較してファイルサイズが大きく、一般的なクローリング手法を適用するとウェブサーバやネットワークトラフィックに対する負荷が高くなってしまうためである。

効率的なクローリング戦略を検討するためには生存率、更新率などの情報が必要となる。しかしながら、従来、ウェブページを対象とした生存分析調査は行われてきたが、PDFファイルを対象としたものは行われてこなかった。

そこで本研究では、クローリング戦略を検討するためにまずPDFファイルの生存分析調査を行う。学術論文専門の検索エンジンを前提とした調査であるため、対象が論文か否かを軸とした分析を行った。

2. ウェブページの生存分析調査

ウェブページの生存分析の既往調査での手法を参考に、本研究ではPDFファイルの生存分析調査の設計や結果の分析を行った。

従来、ウェブページの生存分析は主としてウェブダイナミクス的観点から行われてきた。ウェブがスタティックなものではないという事実は早くから知られており、プロキシサーバのキャッシュや、クローリングの効率化などへの応用を中心に、ウェブの時系列変化を捕捉しようとする研究がなされてきた。

例えば、Douglas ら[3)]はAT&T研究所のクライアントがアクセスした474，000URLについて、ファイル属性、アクセス頻度、更新頻度等を調査し、アクセス頻度と更新頻度との間に強い相関のあることを示している。また、Koehler[4)5)]は Web Crawler random URL generator によって選択された361URLを対象に、四年間にわたってその変化についての調査を行い、ウェブページの半数がおよそ2年で消滅すること等を明らかにした。同じく、Brewingston ら[6)]は計200GBのウェブページの時系列変化を確認し、モ

デル化を行うとともに、検索エンジンによるインデクスの更新戦略に対する応用可能性について検討している。さらに、Cho[7]らは270の有名なウェブサーバから4ヶ月間毎日720，000ページをダウンロードし、全てのウェブページの40％が一週間以内に変化すること、50日間に全体の50％が変化すること、comドメインの25％は一日で変化すること、comドメインのページが半分変化するための期間は11日間であるが、govドメインの場合は4ヶ月間を要すること等を明らかにした。Fetterly ら[8]は、Choらの200倍のサンプル集合を収集し、その後、10週間に亘って、計11回のクローリングを行い、その変化を観測し、ドメインごとの変化とアクセス可能性の推移を確認しているものの、Choらとは異なり、govドメインの方が変化のスピードが早いことを報告している。その一方で、Shiら[9]は幾つかの著名サイトにおいて動的に生成されるページの変化を確認している。

3. 生存分析調査手法

本研究では2年前に収集した日本語PDFファイル集合を対象としてPDFファイルの生存分析調査を行った。調査はクローラーを用いた機械的な調査と人手によるファイル追跡調査から構成される。

3.1 調査対象PDFファイル集合

調査対象PDFファイル集合は、2005年5月と半年後の2005年11月との2度にわたって収集された。IPAdicの6つの名詞辞書ファイル（計213，020語）から、それぞれ、9，750語（第1回目）、10，250語（第2回目）を無作為抽出し、各々の語について、検索エンジンYahoo！Japanを用いて検索を行った。その際、検索対象を「PDFファイル」＋「日本語」に限定するとともに、各検索語の最大収集件数は上位100件までとした。次に、各々のURLに対してPDFファイルのダウンロードを試みた。ダウンロードが不可能であったもの、及び、0バイト・ファイル、破損ファイル、暗号化ファイル、PDFファイルでないもの等を除去した結果、584，973件となった。

3.2 学術論文の判定

学術論文か否かを軸とした分析を行うため、上記の PDF ファイル集合に対して既往研究[10]で行った学術論文判定結果の情報を利用した。

集合中のファイル数が膨大であり人手で行うことが困難であるため、判定は機械学習手法を用いて大きくスクリーニングしたのちに人手で判定する形で行った。機械学習手法を適用するために最初に無作為に選択した2万件に対して6人の判定者が学術論文を判定し、それを学習集合とした。その集合を機械学習に基づく16分類器に学習させ、全集合を再現率重視で判定させた。1つ以上の分類器が論文と判定した36，857件に対して再び人手で判定を行ったところ、全集合から学術論文PDFファイル13，446件が得られた。

3.3 クローラーを用いた生存調査

生存分析のためにPDFファイル集合の全ファイル（のURL）を対象にし、クローラーを用いた機械的なクローリングを行った。調査は2007年12月から2008年1月にかけて行った。なお、リダイレクト(転送)には追従する設定でクローリングを行い、ファイルそのものを収集するとともにHTTPレスポンスコードも保存した。

3.4 ファイル追跡調査

今回のクローラーによる調査でアクセスできなかったファイルはすべてが消滅したのではなく、何らかの理由で元のURLとは異なる場所に移動した場合もある。アクセスできない理由としては、HTTP規格では転送先をウェブサーバ側で設定できるが、それが設定されていないこと、ファイルの存在したサーバ自体にアクセスできないことなど様々なものが考えられる。

元のURLからは論文のPDFファイルにアクセスできなくてもインターネット上の他の場所で公開されているならば、そのファイルは生存しているといえる。そこで、アクセスできなかったファイルに対して検索エンジンを用いたファイル追跡調査を行った。

PDFファイルの追跡調査は、キャッシュされた現物ファイルを確認した上で、調査者が的確な結果が出ると考えた検索式により検索エンジンを用い

て行った。検索エンジンとしては代表的なGoogle、Yahoo！Japanと学術情報専門検索エンジンのGoogle Scholar BETAを用い、すべての検索結果について上位20件まで確認した。なお、調査は2008年2月に発表者らが分担して行った。

クローラーでアクセスできなかったファイルすべてを調査することは困難であるため、無作為に抽出したファイルを論文1，328件、非論文351件を対象とした。本研究では学術論文の生存調査に重点を置くため、論文の調査数を増やしている。

4. 調査結果

4.1 クローラーによる生存調査結果

クローラーによりアクセスできたかどうかという生存調査の結果を○はアクセス可、×はアクセス不可として表1に示す。

表1 クローラーによる生存調査

	論文		非論文		全体	
○	9,579	71.2%	313,463	54.8%	323,042	55.2%
×	3,867	28.8%	258,064	45.2%	261,931	44.8%
計	13,446	100.0%	571,527	100.0%	584,973	100.0%

表からは、ウェブページの生存調査と同様にPDFファイルもおよそ2年間で半数がアクセスできなくなることがわかる。また、論文と非論文を比較すると、明らかに論文の方がより生存しやすいといえる。

PDFファイル集合は日本語ファイルで構成されておりJPドメインのURLが全体の85．2%を占める。そこで、アクセス可能性をファイル数が多いセカンドレベルドメイン順に表2に示した。

表2 JPセカンドレベルドメイン別の生存調査

	論文		非論文		全体
co	50,865	52.1%	46,821	47.9%	97,686
or	36,326	59.1%	25,149	40.9%	61,475
ac	38,336	63.1%	22,466	36.9%	60,802
go	41,827	69.0%	18,757	31.0%	60,584
ne	18,947	52.5%	17,142	47.5%	36,089
ed	4,645	55.1%	3,792	44.9%	8,437
tokyo	3,689	47.4%	4,097	52.6%	7,786
hokkaido	3,260	45.6%	3,896	54.4%	7,156
gr	4,041	60.2%	2,670	39.8%	6,711
osaka	2,700	44.6%	3,352	55.4%	6,052
その他	68,802	47.3%	76,620	52.7%	145,422
計	273,438	54.9%	224,762	45.1%	498,200

表からは政府系（go.jp）、大学(ac.jp)は他のドメインと比べ、以前と同じURLでのアクセス可能性が高く比較的変化が少ないサイトであることがわかる。日米間での違いはあるが、ウェブページ調査で政府系サイトは更新が少ないとしたChoらの結果を支持するものといえる。

4.2 ファイル追跡調査結果

クローラーでアクセスできなったファイルを対象とした人手によるファイル追跡調査の結果を表3に示す。表3において「ブラウザでアクセス可能」とはクローラーではアクセスできなかったがファイル追跡調査時にはウェブブラウザでアクセスできたものを示している。この原因はネットワークやクライアント環境の違いによるものであるが、元URLのままでアクセスできることを意味するため、それらを除いたファイルをファイル追跡調査の対象とした。

表3 ファイル追跡調査結果

	論文		非論文	
ブラウザでアクセス可能	48		7	
調査対象数	1280	100.0%	344	100.0%
移動先URL発見数	704	55.0%	84	24.4%
移動先URL不明数	576	45.0%	260	75.6%
計	1328		351	

表3からは以前のURLでアクセスできなくなっていたファイルが、論文の場合には検索エンジンを用いることで半分以上のファイルを再発見できていることがわかる。一方で非論文の場合には再発見できる割合は1／4程度であり、論文はURLが変わったとしてもインターネット上で公開され続ける傾向が強いといえる。

ファイルを再発見できた場合に用いた検索エンジンを表4に示した。表で各エンジンの発見数は排他的でなく、GoogleとYahoo！Japanの両方で再発見できた場合には両方にカウントしている。

表4 検索エンジン別移動先URL発見数

		論文		非論文	
エンジン	Google	632	89.8%	72	85.7%
	Yahoo! Japan	446	63.4%	53	63.1%
	GoogleScholar	260	36.9%	1	1.2%
	その他	42	6.0%	9	10.7%
移動先URL発見数		704	100.0%	84	100.0%

表4からは論文は9割近くがGoogleで再発見可能であること、非論文はGoogle Scholarで再発見できないことが明らかである。さらに表では直接示されているわけではないが、Google Scholarでは検索

結果にリンク先が含まれていても、実は元URLのままでリンク切れとなる事例が多くみられた。GoogleとGoogle Scholarについてこのような違いがみられる原因としては、同じ Google 社による検索エンジンでもクローリング間隔に大きな差があることが考えられる。

5. 学術論文に対するクローリング

PDFファイルの生存分析の結果に基づいて、頻度と対象範囲から学術論文PDFファイルに対するクローリングの戦略について検討を行う。前述のように、一般的なウェブページよりもPDFファイルのクローリングはネットワークやサーバに負荷をかけるため、できるだけ効率化することが望ましい。

一般的なウェブページに対するクローリングでは内容が更新されることがあるため、特にニュースサイトやブログに対しては、その頻度を高くする必要がある。しかし、学術論文PDFファイルでは、その性質上、内容の更新はほとんど行われないため、同じファイルを再度クローリングする必要性は少なく、クローリングを行う回数は一度でよいと考えられる。

また、新規に追加されるファイルを確実に収集するためには、PDFファイルそのものではなく、ファイルへのリンクを張ったウェブページのクローリングが必要である。このようなリンクページに対しては、一見、頻繁なクローリングが望ましいように思われる。しかし、今回の生存分析の結果から論文PDFファイルは非論文と比べて生存率が高いことが判明した。また、ファイル追跡調査では、論文PDFファイルは場所が変わったとしてもインターネット上から消滅することは少なく、非論文ファイルと比較して、移動先を検索エンジンで探せる可能性が高かった。これらの結果は、リンクページに対するクローリングを比較的長い間隔で行ったとしても新規追加の論文 PDF ファイルを逃す可能性は低いことを示唆している。

次にクローリング範囲について検討を行う。機関リポジトリなど、学術論文 PDF ファイルが集中するサイトを優先的にクローリング範囲に含めるべきことは明らかであるが、それ以降の優先順位については日本語ファイルの場合にはドメイン名で設定していくことが考えられる。理由は、筆者らの既往研究[11]でPDF ファイルを収集した際のドメイン名分析からは日本語学術論文 PDF ファイルの多くは大学系(ac.jp)や政府系(go.jp)サイトから収集されたためである。これらのドメインを重点的に収集対象とすることで論文に対する効率的なクローリングが可能になるはずである。

6. まとめ

学術情報専門の検索エンジンにおけるクローリング戦略を検討するためPDFファイルの生存分析調査を行った。論文ファイルは同じURLで生存する割合が高く、さらに移動先URLも再発見しやすいことが明らかとなった。

【注・参考文献】

1) 安形輝ほか 5 名. 学術論文 PDF 検索システムの開発と評価. 2007 年度日本図書館情報学会研究大会発表要綱 2007, 鶴見大学, 2007-10-13/14, p.57-60

2) 三根慎二. “オープンアクセス資料のファイル形式”. http://www.openaccessjapan.com/archives/2006/05/oa_1.html

3) Douglis, Fred., Feldmann, Anja., Krishnamurthy, Balachander., Mogul, Jeffrey. "Rate of Change and Other Metrics: A Live Study of the World Wide Web," Proceedings of the USENIX Symposium on Internetworking Technologies and Systems, p.147-158(1997)

4) Koehler, Wallace. "An Analysis of Web Page and Web Site Constancy and Permanence." Journal of the American Society for Information Science. Vol.50, No.2, p.162-180(1999)

5) Koehler, Wallace. "Web Page Change and Persistence-A Four-Year Longitudinal Study," Journal of the American Society for Information Science and Technology, vol.53, no.2, p162-171(2002)

6) Brewington, Brian E., Cybenko, George. "How Dynamic is the Web?," Proceedings of the 9th International World Wide Web Conference. also available Computer Networks, Vol.33, 1-6,p.257-276(2000)

7) Cho, J., Garcia-Molina, H. The Evolution of the Web and Implications for an Incremental Crawler. Proceedings of the 26th International Conference on Very Large Databases, 2000.

8) Fetterly, Dennis., Manasse, Mark., Najork, Marc. Wiener, Janet. “A Large-Scale Study of the Evolution of Web Pages”, Proceedings of the 12th International Conference on World Wide Web. p.669-678(2003)

9) Shi, Weisong., Collins, Eli. Karamcheti, Vijay. "Modeling Object Characteristics of Dynamic Web Content," Proceedings of the IEEE Globecom 2002 conference in Taipei.

10) 具体的な手順については以下の文献に詳しい。池内淳ほか 5 名. プーリング手法を用いた学術論文の自動判別実験. 情処研報 Vol.2007 No.34, p.33-40

11) 安形輝ほか 5 名. “日本語学術論文 PDF ファイルの自動判定”. Library and Information Science, No.56,p.43-63(2006)