ウェブの動的変化に関する調査

池内淳（大東文化大学）　安形輝（亜細亜大学）
野末道子（鉄道総合技術研究所）　久野高志（作新学院大学）
石田栄美（駿河台大学）　上田修一（慶應義塾大学）

Ⅰ．はじめに：問題意識

かつて指数関数的と言われたウェブの急速な成長は、単にその量的な増加に止まるものではなく、数多くのウェブページが次々と生産、更新、消滅、再起を繰り返し、動的に変化している。

ウェブに対する社会学的関心は、その構造、あるいは、量的な変化だけでなく、その内容的な変化に対しても及んでおり、これまで幾つかの調査が行われてきた。

これらウェブの時系列変化を捉えようとする調査のインセンティヴは、HTTP サーバによる効率的なキャッシング、あるいは、サーチエンジンによる最適なクローリング戦略にとって有益な情報をもたらすことにあり、それぞれ一定の成果を挙げている。

例えば、Douglas ら[1]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、そのファイルの属性、アクセスの頻度、ファイルの更新頻度等を調査し、コンテント・タイプとアクセス頻度が更新頻度に大きく関係していることを示している。同じく、Brewingston ら[2]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、サーチエンジンによるインデッ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。また、Koehler[3, 4]は Web Crawler random URL generator によって選択された 361URL を対象に、四年間に亘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凡そ 2 年で消滅すること等を明らかにしている。一方、Shi ら[5]は幾つかの著名な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。

ところで、こうした時系列調査に限らず、ウェブにおいて標本調査を行う場合に常に問題となるのは、ウェブではランダム・サンプリングが極めて困難であり、その統計学的妥当性を保証されないという点にある。既往調査においても、この問題を克服したものはほとんどないと言える。

そこで、本研究では、数多くの標本ページ集合を用いるとともに、ウェブ全体ではなく、特定のドメイン（jp ドメイン）のページ群のみを対象とした調査とすることで、標本調査に関するサンプリングの妥当性問題をできるかぎり解消した。

また、ここで jp ドメインを対象としたのは、単にわれわれにとって近接性の高いドメインであるというだけではなく、(1) 既に収集された相当数の URL の集合が利用可能であったこと、かつまた、(2) 定期的にそのページ総数を推定した調査結果[6]が公表されていたためである。

すなわち、ウェブの動的な変化はページの生存と消滅のみを捕捉するのではなく、その量的な変化と併せて考察することが肝要であると考えられる。ここでは、2 年前に収集された大規模ウェブ集合を再び調査し、その時系列変化（生存・消滅・更新状況）を分析し、その結果に基づき生成・更新数の推定を行っている。

Ⅱ．方法論：データの収集と解析

本調査では、NTCIR-3 Web タスクのために収集・構築された 100GB の文書集合（NW100G-01）を用いて、これらが一定期間（約 2 年間）を経た後に、どのように変化したのかを確認した。NW100G-01 は、NII のホームページ（http://www.nii.ac.jp/）を起点として、2001 年 8 月 29 日～11 月 12 日にかけて、主に jp ドメインのサーバを対象として収集された文書群の部分集合であり、1100 万以上のページによって構成されている[7]。

ほぼ同時期（2001 年 7 月～10 月）にクローリングを行った内田ら[6]の調査によれば、jp ドメインの推定ページ数は 6507 万ページであり、この文書集合は当時の jp ドメイン全体の 6 分の 1 を占めていたものと推定されることから、上述の意図に適ったものであると言えよう。

本調査では、100GB の文書集合のうち jp ドメインのページのみ 1000 万ページを無作為に抽出した。ちなみに、この標本ページ集合の平均ページ・サイズは 8,231.5 バイト、平均タグ数は 146.6、うちリンク数は 10.2、画像数は 8.1 であった。既存のウェブの統計調査における平均

ページ･サイズは概ね 10,000 バイト前後であり、ここでもそれらの調査結果と符合している。

本調査における基本的な判別方法は、各々の URL について、HTTP レスポンスのヘッダ･フィールドである ETag、 Content-Length、Last-Modified を取得し、元データと照合することによって、各ページの変化を認識している。ETag はウェブページの同一性を識別するためのID、Content-Length はサーバ側で認識するページのサイズ、Last-Modified は最終更新時刻をそれぞれ意味している。ヘッダ情報が利用できないものについては、直接ファイルを読み込んでデータ･サイズを比較した。当然、サーバ自体が存在しない場合や何らかのエラー値が返される場合も多く存在する。

実際の調査手続きは以下の通りである。

①URL に対してヘッダ情報をリクエストする（HEAD メソッド）。
②ウェブ･サーバが見つからない場合にはサーバ（発見）エラーとし、終了する。
③HTTP レスポンス中のレスポンスコードを取得し保存する。
④コードが OK でない場合、終了する。
⑤ヘッダ情報を解析し、ETag、Last-Modified、Content-Length の情報によって更新を判別する。
⑥更新判別ができた場合、終了する。
⑦内容情報のリクエストを行う(GET メソッド)。
⑧サイズ情報を取得し更新判別をし終了する。

調査は 2003 年 9 月 28 日～10 月 1 日にかけて行った。この際、調査期間が延長されることによる観測誤差を最小限に抑えるため、一定の資源の制約の元で、可能な限り迅速な調査が行えるよう、複数の HTTP クライアントを協調動作させてデータを収集した。

III. 結果と考察

A. 更新調査結果（ページ単位）

調査結果の概観を表 1 に示す。各々の状態の判別方法は以下の通りである。

･タイムアウト…タイムアウト（コネクション 5 秒＋データ受信 5 秒）
･ホストエラー…サーバにコネクトできなかった場合
･サーバエラー…ステータス･コード 500～505
･認証失敗…ステータス･コード 401～403, 407
･ファイル移動…ステータス･コード 301, 303, 307
･ファイルなし…ステータス･コード 404
･他のエラー…その他のエラーコード
･更新なし…ヘッダ及びページ･サイズから判定
･更新あり…ヘッダ及びページ･サイズから判定

表 1. 標本ページ集合の変化

状態		ページ数	比率
アクセス不能	タイムアウト	84,255	0.84%
	ホストエラー	878,467	8.78%
	サーバエラー	17,254	0.17%
	認証失敗	74,172	0.74%
	ファイル移動	386,620	3.87%
	ファイルなし	3,219,881	32.20%
	他のエラー	3,252	0.03%
生存	更新なし	2,687,710	26.88%
	更新あり	2,648,389	26.48%
合計		10,000,000	100%

1000 万ページのうち、2 年を経て現在も生存するページの比率は 53%（5,336,099 ページ）、消滅あるいはアクセス不能となったページの比率は 47%（4,663,901 ページ）であり、上述の Koehler[3), 4)]による結果を支持している。

また、現存するページのうち「更新なし」と「更新あり」はほぼ同数であることから、半数のページは 2 年間全く更新されていないことが分かる。

次に、表 2 はサーバが ETag と Last-Modified の値を返したページの比率について、元データ（2001 年）と今回の調査（2003 年）の比較を行ったものである。時系列変化からは、サーバアプリケーションのバージョンアップなどの理由で、より充実したヘッダ情報を返すサーバが増加していることが分かる。「更新なし」であったページと「更新あり」であったページとの比較では、前者の方が高い値を示している。

また、元データのページ･サイズの平均値は 8,231.5 バイトであったが、そのうち、今回の調

表 2. ヘッダ･フィールドの利用可能率

	ETag		Last-Modified	
	2001 年	2003 年	2001 年	2003 年
更新なし	79.6%	86.9%	90.0%	92.7%
更新あり	62.8%	70.5%	73.0%	76.2%
合計	71.2%	78.8%	81.6%	84.5%

査においても生存していたページ群の 2001 年時点での平均サイズは 9033.5 バイトであった。これに対して、今回の調査の結果得られた生存ページ群の平均サイズは 10969.9 バイトと増加

表 3. 更新ページのサイズの変化

	ページ数	比率
増加	1,753,478	66.2%
変化なし	52,247	2.0%
減少	842,664	31.8%
合計	2,648,389	100%

している。いずれの既往調査 [1), 2), 3), 4)]においても、より最近のサイズの方が大きくなっており、ここでもまた、同様の結果が得られた。

表 4. 更新時期の分布

期間	ページ数	割合
～Jul-01 (**異常値**)	103,501	5.1%
Aug-01～Oct-01	67,287	3.3%
Nov-01～Apr-02	257,195	12.7%
May-02～Oct-02	314,094	15.6%
Nov-02～Mar-03	585,515	29.0%
Apr-03～Oct-03	689,888	34.2%
Nov-03～ (**異常値**)	716	0.0%
合計	2,018,196	100%

これと関連し、更新されたページについて、2001 年と 2003 年でファイルサイズがどのように異なっているかを確認したのが表 3 である。

この表によれば、全体の三分の二のファイルが増加し、三分の一弱が減少し、2%程度が変化なしとなっている。

ここで、変化なしであったものについては、偶然、更新されたサイズが以前のものと同じであったというよりはむしろ、サーバのバージョンアップ等によって ETag や Last-Modified の値だけが更新されたと考えることが妥当であろう。

表 4 は更新されたページのうち、ヘッダ情報である Last-Modified が返されたものについて、その更新日時を、6 ヶ月ごとの分布によって確認したものである。したがって、合計値は更新ページ全体よりも若干少ないものとなっている。

これによれば、調査以前、あるいは、調査以降のタイムスタンプを示すものが全体の 5.1%程度存在するなど、サーバの返す値の精度に若干の問題のあることが看取されるものの、一定の傾向を掴むことはできる。

更新頻度は時間が経過すればするほど減少しており、直近の 6 ヶ月間に、全体の 3 分の 1 以上（34.2%）が更新されている。

次に、図 1 は、同じく Last-Modified 値が返されたページについて、更新日時に依らず、その更新曜日と更新時間の推移を示したものである。

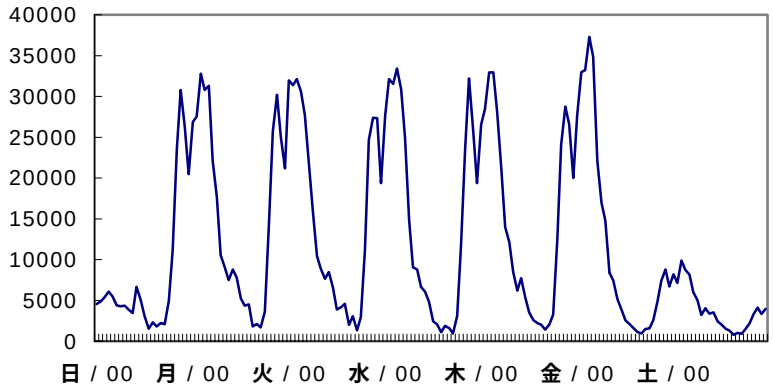
まず、曜日については、平日と土日とで明確な差異が認められ、ウィークデーの更新頻度が高い。とくに、金曜日の午後が一週間のうちで更新のピークとなっていることが分かる。また、時間については、やはりデイタイムに更新されることが多く、午前中の 11～12 時頃にやや落ち込むものの（食事/お昼休みのため）、高い値を示している。

これについては、元データの Last-Modified 情報を用いて確認した場合も、上述の Brewingston ら [2)]による調査においても全く同様の傾向が見受けられた。

B. 更新調査結果（サーバ単位）

サーバの生存と消滅についても触れておく。1000 万ページの元データについて、ドメイン

図1. ウェブ・ページの更新時間・更新曜日の分布



名により判別された異なりサーバ数は 95,372 サーバであり、1 サーバあたりのページ数の平均値は 104.9 ページであった。

これら 95,372 のサーバのうち、リクエストの 4 分の 3 以上に対して何らかのエラーを返した、あるいは、サーバそのものが存在しなかったものは 27,729 サーバであり、全体の 29.1%を占めている。このことから、サーバの安定性がページのそれよりも高いことが確認された。

C. ファイルの形態的属性

ページの生存・消滅と、ファイルの形態的属性との間に関連性があるか否か調べるために、元データにおけるページのページ・サイズ、タグ数、リンク数、画像数の平均値を、2 年後の調査による状態別に集計した（→表 5）。

最も顕著な傾向としては「更新なし」のサイズ、タグ数等の平均値が他の状態と比較して著しく低い値を示していることであろう。反対に、「更新あり」については、高い値を示している。

D. ウェブページの増加率の推計

ページの変化と jp ドメインのページ数の増加との関係について言及する。件の内田ら[6]の調査は 1998 年 2 月～2002 年 2 月までの 4 ヶ年間（計 9 回）に及んでいるが、調査期間後期には、開始当初と比較して、ページ数の伸長率が鈍化していることを指摘している。

そこで、内田らの調査結果におけるページ数の時系列変化を回帰式に当てはめ、筆者らが調査を行った時期（2003 年 10 月）における jp ドメインのページ数を推定した。その結果、線形の関数（$y = a + bx$）に当てはめた場合は 9818 万ページ（R^2=0.96）、指数関数（$y = a \cdot x^b$）に当てはめた場合は 7717 万ページとなった（R^2=0.95）。

調査時点での jp ドメインのページ数は 6507 万ページであるから、この 2 年間におけるページの単純な増加数は、それぞれ 3312 万ページ（線形）と 1210 万ページ（指数）であるが、ウェブの消滅を考慮した場合、新しく生産されたページは 6347 万ページ（線形）と 4302 万ページ（指数）となる。さらに、更新ページを考慮すれば、新たに生産、もしくは、更新されたページは、少なくとも 8070 万ページ（線形）と 6025 万ページ（指数）であると推定される。

【注・引用文献】

1) 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)
2) 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)
3) 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)
4) 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)
5) Shi, Weisong., Collins, Eli. Karamcheti, Vijay. "Modeling Object Characteristics of Dynamic Web Content," Proceedings of the IEEE Globecom 2002 conference in Taipei
6) 内田斉. “メディアとしての Web の成長を測る：サーチロボットを使った Web コンテンツ統計調査の試み, ” http://www.a-brain.com/HP/rep/rep15/ (最終確認 2003/10/15)
7) K. Eguchi, K. Oyama, E. Ishida, N. Kando, K. Kuriyama. "Evaluation Methods for Web Retrieval Tasks Considering Hyperlink Structure", IEICE Transactions on Information and Systems, Vol.E86-D, No.9, p.1804-1813 (Sep. 2003)

表 5．ページの状態とその形態的属性との関係

状態		ページ数	サイズ	タグ数	リンク数	画像数
アクセス不能	タイムアウト	84,255	9,064	155.0	11.1	8.2
	ホストエラー	878,467	8,105	142.2	10.9	8.0
	サーバエラー	17,254	8,169	143.1	13.5	7.3
	認証失敗	74,172	8,081	157.2	14.2	8.3
	ファイル移動	386,620	9,093	169.6	15.4	12.1
	ファイルなし	3,219,881	8,317	146.3	8.9	8.2
	他のエラー	3,252	8,780	191.1	14.7	7.6
生存	更新なし	2,687,710	7,234	109.9	6.6	4.9
	更新あり	2,648,389	9,034	181.6	14.2	10.7
合計		10,000,000	8,232	146.6	10.2	8.1