D-019

技術コラムにおける Web リンクの寿命
Web Decay in a Series of Technical Columns

飯尾 淳 †
Jun Iio

1. 背景と目的

World Wide Web（WWW）は，今やインターネットによる情報公開・情報流通の仕組みとして確固たる地位を確立し，社会基盤としても無くてはならないものとして位置付けられている．WWW の技術が急激に普及した理由は，マルチメディアによる表現の多様性とハイパーリンクによる直感的な操作がユーザに受け入れられやすかった点にあろう．検索システムの発展や各種アプリケーションの開発，表現の更なる多様化などシステム自体が複雑化し，さらにはセマンティック・ウェブ [1] や Linked Data[2] といった仕組みが導入され WWW を大規模なデータ基盤として取り扱う研究も盛んに行われるようになった．

一方で，WWW におけるハイパーリンクの仕組みはしごく単純なものとして，当初提案されたままの方式が未だに活用されている．すなわち，ハイパーリンクは HTML 文書中に記述された単なる参照関係として実装されており，リンクの永続性や可用性は保証されていない．この単純さと柔軟性が WWW の普及を加速した一因であることは否めないが，一方で，ハイパーリンクを辿ったときに「404 Not Found」のようなエラーメッセージが出てアクセスしたいページを参照できない，というケースに遭遇する機会も日常的である．

本論文は，このような参照エラー，いわゆるダングリング・ポインタとなっているハイパーリンクの割合を明らかにし，その寿命に関する定式化を行うことで，ハイパーリンクの有効性に関する定量的な評価を行うことを目的とする．

2. 参照の永続性確保と参照可能性に関する分析

WWW におけるハイパーリンクの永続性は担保されにくいという性質を踏まえ，当初，学術論文においては WWW のみで提供されている文書の引用は控える，引用する場合にはアクセス日時を明記するなど，暗黙のルールが考えられてきた．また文書を永続的に参照できるような文書番号体系 [3] も提唱されている．文献の参照可能性を重視する学術文書においては，このような試みが重ねられており，参照の永続性を実現する動きが進んでいる．

また関連研究の節で述べるように学術文書を対象にしたものでは，文書の到達可能性やハイパーリンクの寿命に関する研究が，これまで様々な分野を対象として行われてきた．ただし一般の文書に関しては，文書数があまりにも膨大であることから，参照の永続性に関する考察が加えられた例はほとんどない．

なお個々の Web サイトにおけるリンクについては，Search Engine Optimization（SEO）の観点から「きちんと管理してリンク切れとなっているハイパーリンクを削除すべし」という助言が浸透しており，リンクの有効性を管理するツールが多数，提案されている．本研究においても，リンク先の状態をチェックするために，既存のツールである `linkchecker`‡を用いた．

図 1. 「週刊 Take IT Easy」のスクリーンショット

3. 技術コラムにおけるハイパーリンクの分析

本研究は，筆者らが 1998 年に開始し，現在まで 12 年以上続けてきた技術コラムを対象に分析を試みたものである．対象とした Web サイトは「週刊 Take IT Easy」§と題された Web サイトである（図 1）．

「週刊 Take IT Easy」においては，盆と正月，大型連休時等を除いてほぼ毎週，記事の更新が維持されてきた．2011 年 3 月 8 日の時点で 596 本の記事が蓄積されている．記事の執筆は筆者が所属する組織の若手メンバを中心に行われているが，公開前に関係者の査読を経ることで記事の品質を一定以上のレベルに保証する仕組みが用意されている点で，単なるブログ記事とは異なる．なお本コラムにおける過去記事 10 年分をテキストマイニングで分析した結果，10 年間における情報技術トレンドの推移を明示することに成功した報告もなされている [4] ので，そちらも参考にされたい．

本 Web サイトにおける記事の特徴として，記事中にキーとなるハイパーリンクが埋め込まれていること，および関連するリンク集を記事末尾に示していることを挙げることができる（図 2）．これらのリンクは執筆者本人が自ら指定し，手動で埋め込んでいる点に注意されたい．さらに，ミスを防ぐため公開前に別の作業者が確認する手順を踏むことが定められている．

本研究では，1998 年の公開開始から 3 月 8 日に公開された記事までの 596 本を対象に，各記事に含まれて

† （株）三菱総合研究所, MRI
東京都千代田区永田町 2-10-3

‡ `http://linkchecker.sourceforge.net/`
§ `http://easy.mri.co.jp/`

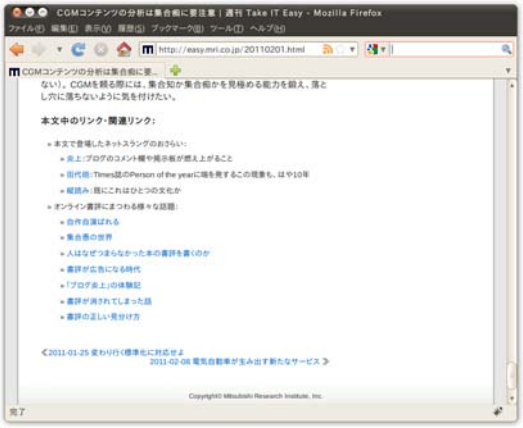

図 2. 記事末尾に記載されているリンク集の例

いるハイパーリンクを抽出，それらのリンク先にアクセスできるかどうかの実験を実施した．なお 1998 年の開始当初，各記事は静的な HTML 文書として実現していたが，現在，同 Web サイトは Content Management System（CMS）によって管理されており，記事はデータベースに格納されている．

実験の具体的な手順を以下に示す．

1. 記事データベースから全記事を取り出し，さらに記事中のハイパーリンクを抽出．
2. 抽出したリンクを，外部参照と内部参照に分離．
3. 外部参照のみを抽出とした Web サイトのミラーを作成し，同サイトに対してチェックツールを適用．
4. ツールのログを解析，傾向分布について考察．

なおここで外部参照とは外部の Web サイトに向けたリンクを指し，内部参照とは「週刊 Take IT Easy」サイトに含まれる記事および画像への参照のことをいう．今回は外部参照のみを対象とした．その理由は，リンクの循環を排除し手順を簡略化することと，そもそも内部参照は完全性が確保されているという前提に立っているためである．

4. 実験結果

本節では，実験の結果を 1. 外部参照リンク数の傾向, 2. 外部参照先の状況, 3. モデル式の当てはめの順に示す．

4.1. 外部リンクの傾向

まず各記事に含まれる外部参照リンクの数がどのような分布を示すかをまとめた．平均と標準偏差を表 1 に示す．

また，外部参照リンク数のヒストグラムを図 3 に示す．本分布は著者の恣意性が強く影響し，ポアソン分布と比較すると同分布よりやや裾の厚い分布を示している．

表 1. 記事中に含まれる外部リンク数の平均と標準偏差

平均	6.64	標準偏差	4.23

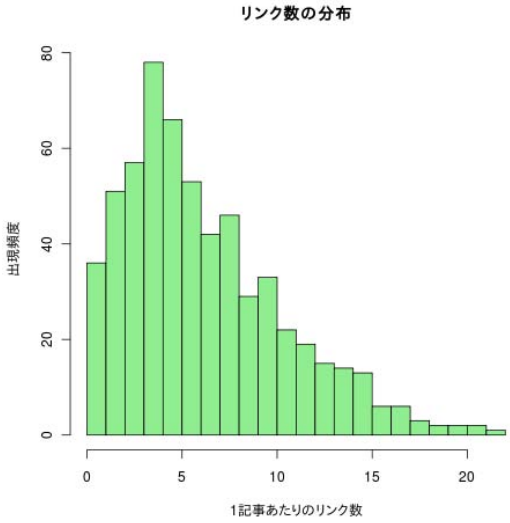

図 3. 記事中に含まれる外部リンク数のヒストグラム

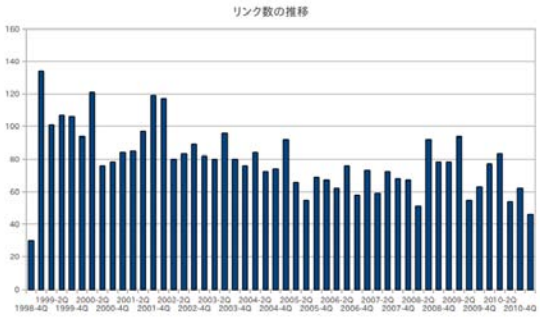

図 4. 記事中に含まれる外部リンク数の推移

続いて，図 4 は同リンク数を四半期ごとに集計し，その推移をグラフ化したものである．左端（1998 年 4Q）および右端（2011 年 1Q）は，期間を満たしていないため少なくなっている．全体を通じて，開始当初はリンク数がやや多めであったが途中から一定の数値に落ち着いている傾向が見られる．これは，1 つの記事執筆にかけられる時間的コストが次第に最適化されていったことや，執筆陣が途中から拡大され，発起メンバー以外の執筆者が中心となったことが影響していると考えられる．

4.2. 外部参照先の状況

次に，本記事から参照されている外部 Web サイトの状況が現在どのようになっているかを確認した結果について述べる．外部 Web サイトの確認はツールを用いた自動アクセスにより実施した．レスポンス確認の制限時間として，リクエストメッセージ送信後 10 秒以内にレスポンスがなかった場合には「反応なし」とした（この種のエラー、サーバエラー等を Err2 とした）．なおアクセス先の検証実験は 2011 年 3 月 9 日に実施した．

検証した 596 本の記事に含まれていた全ての外部参照 3,958 件を辿った結果の状況を表 2 に示す．なお今回の実験では、`ftp:`や`mailto:`など、HTTP 以外のプロトコルに関しては参照先調査の対象外とした．

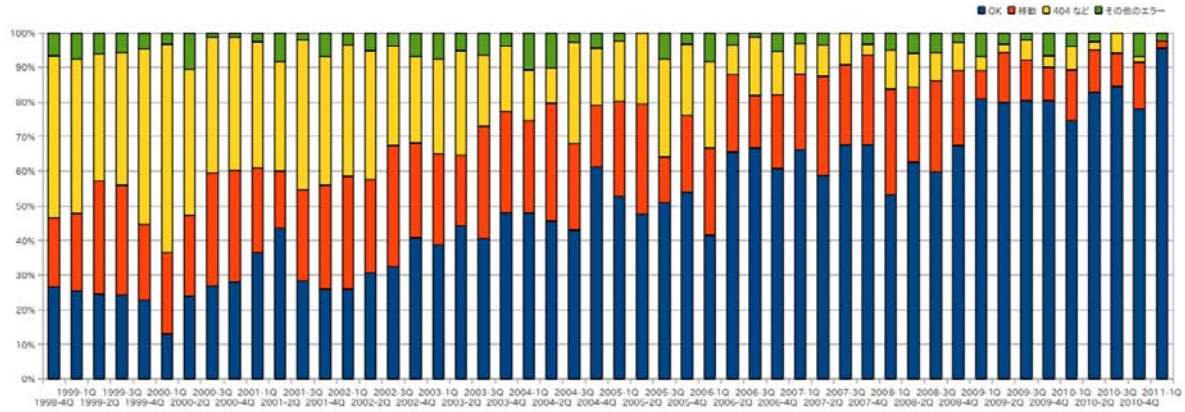

図 5. 外部参照先からの反応状況（四半期ごと）

1998 年 4Q から 2011 年 1Q まで，各記事に含まれる参照先からのレスポンスを四半期ごとにまとめた状況を表 3 および図 5 に示す（図のグラフは対象外とされたリンクを除外して作成した）．

4.3. 半減期の推定

Spinellis[5] は Web サイトの寿命に関する半減期をおよそ 4 年と論じている．半減期 $t_{1/2}$ は減衰定数 λ を用いて下記で定義される指数である．

$$t_{1/2} = \lambda^{-1} \log 2$$

表 3 に示すデータを用い，最小二乗近似により λ を求めた結果，$\lambda = 0.118$ の値を得た．このことから本件に関する半減期は 6 年弱 ($t_{1/2} = 5.87$) と推定される（図 5 と対比されたい）．なおここで参照先の有効性は「正常」なレスポンスを返したサイトの数のみを有効として計算し，対象外のデータは除外している．

5. 考察

本研究が対象とした Web コラム「週刊 Take IT Easy」において，各記事に含まれている外部参照は 6 年弱の半減期を持つと推定された．このことは，長期的な参照を維持したい場合について，1 つの目安を提示すると考えられる．すなわち，リンクの有効性が推定された半減期以内であれば，単に参照するだけでよく，それ以上の永続性を維持したいのであれば，参照先のコピーを保存すべきという指針である．なおここでは半減期を示しているので，リンクとして参照するだけでよいか参照先を保存するかは，50%の確率として論じている．もしそれよりも高い確率で参照可能性を維持したいのであれば，参照先を保存しなければならない期間は短くなる．

表 2. 外部参照先からの反応状況

レスポンス		件数	比率 (%)
正常	[OK]	1,818	45.9
リダイレクト	[Trans]	910	23.0
参照先エラー	[Err1]	920	23.2
その他のエラー	[Err2]	183	4.6
対象外	[N/A]	127	3.2

本研究における分析に際して注意すべき点を 2 点挙げておく．1 つは，本分析において参照先の有効性を「200 OK」というレスポンスが返されたこととしたが，その内容については吟味していない点である．Web サイトの内容は随時書き換えられている可能性があり，「200 OK」というレスポンスが返されたといって当初参照していた内容がそのまま保存されているとは限らないという点には注意しなければならない．

もう 1 点は，「301 Moved Permanently」あるいは「302 Moved Temporary」等のレスポンスコードで返される，今回「有効ではない」として取り扱ったページの扱いである．多くの場合，これらは既に有効ではないコンテンツに対する参照を適切なページにリダイレクトするために利用されるが，場合によってはコンテンツは有効でありながらもサイトの移転等で URL が変更になった場合にも利用される．今回の分析では，この状況は考慮していない点を留意しておく．

6. 関連研究

本稿の冒頭で述べたように，学術文献を対象とした Web リンクの有効性に関する研究は数多く行われている．Wren[6, 7] は医学文献データベースの MEDLINE を対象として長年に渡る調査を実施している．またコンピュータ科学に関する同様の研究が，Lawrence ら [8] によって報告されている．

McCown ら [9] は，デジタルライブラリ研究に関する電子出版である D-Lib における記事中の Web 参照を分析した．Ortega ら [10] らは，欧州において公開されている科学技術に関する Web 記事を分析した．この分析においては Web 記事の半減期が国別に集計されており興味深い結果が得られている．

7. まとめ

本研究では，10 年以上継続している Web コラムにおける外部参照を題材として Web におけるハイパーリンクの有効期限についての分析を行った．分析の結果，情報系技術コラムにおける有効な Web リンクの半減期

は 6 年弱と推定された．

長期的な参照を保証するためには，自らの責任においてキャッシュを保存することが必要である．そのためにはコストがかかるため，本研究で明かになった有効期限とのトレードオフを考慮してキャッシュの保存を行うか否かを判断すべきである．

参照の半減期は Web サイトの性質で異なる．ニュースサイトのような速報性を重視する Web サイトにおいては，リンクの有効期限は非常に短いものと考えられる．一方で学術文献に関する Web サイトではリンクの有効期限は比較的長期のものとなろう．今後，Web サイトの分野，性質と参照の有効期限がどのような相関を示すかについて，分析を加えることが課題として残されている．

参考文献

[1] T. Berners-Lee, J. Hendler and O. Lassila, “The semantic web”, *Scientific American*, Vol.284, No.5, pp.34–43, 2001.

[2] C. Bizer, T. Heath, T. Berners-Lee, “Linked Data - The Story So Far, *International Journal on Semantic Web and Information Systems* Vol.5, Issue 3, pp.1–22, 2009.

[3] N. Paskin, “Digital Object Identifiers for scientific data,” *Data Science Journal*, Vol. 4 pp.12–20, 2005.

[4] J. Iio, S. Udoguchi, and Y. Shirai, “Information Technology Trends in a Decade Revealed by Text-mining,” *Poster Proceedings of the 3rd IEEE Pacific Visualization Symposium (PacificVis2010)*, pp.17–18, 2010.

[5] D. Spinellis,“The Decay and Failures of Web References,” *Communications of the ACM*, Vol.46, No.1, pp.71–77, 2003.

[6] J.D. Wren, “404 not found: the stability and persistence of URLs published in MEDLINE,” *Bioinformatics,* Vol.20, No.5, pp.668–672, 2004.

[7] J.D. Wren, “URL decay in MEDLINE — a 4-year follow-up study,” *Bioinformatics,* Vol.24, No.11, pp.1381–1385, 2008.

[8] S. Lawrence, F. Coetzee, E. Glover, D. Pennock, G. Flake, F. Nielsen, R. Krovetz, A. Kruger, and C.L. Giles, “Persistence of Web References in Scientific Research,” *IEEE Computer*, Vol.34, No.2, pp.26–31, 2001.

[9] F. McCown, S. Chan, M.L. Nelson, and J. Bollen, “The Availability and Persistence of Web References in D-Lib Magazine,” *Proceedings of the 5th International Workshop on Web Archiving and Digital Preservation (in conjunction with ECDL2005)*, 2005.

[10] J. Ortega, V. Cothey,and I.F. Aguillo, “How old is the Web? Characterizing the age and the currency of the European scientific Web,” *Scientometrics*, Vol.81, No.1, pp.295-309, 2009.

表 3. 外部参照先からの反応状況（四半期ごと）

期間	OK	Trans	Err1	Err2	N/A
1998 年 4Q	8	6	14	2	0
1999 年 1Q	34	30	60	10	0
1999 年 2Q	24	32	36	6	3
1999 年 3Q	26	34	41	6	0
1999 年 4Q	24	23	53	5	1
2000 年 1Q	12	22	56	3	1
2000 年 2Q	28	27	49	12	5
2000 年 3Q	20	24	29	1	1
2000 年 4Q	22	25	30	1	0
2001 年 1Q	30	20	30	2	2
2001 年 2Q	37	14	27	7	0
2001 年 3Q	27	25	41	2	2
2001 年 4Q	30	35	43	8	3
2002 年 1Q	30	38	44	4	1
2002 年 2Q	24	21	29	4	2
2002 年 3Q	26	28	23	3	3
2002 年 4Q	36	24	22	6	1
2003 年 1Q	31	21	22	6	2
2003 年 2Q	35	16	24	4	1
2003 年 3Q	39	31	20	6	0
2003 年 4Q	38	23	15	3	1
2004 年 1Q	36	20	11	8	1
2004 年 2Q	36	27	8	8	5
2004 年 3Q	31	18	21	2	0
2004 年 4Q	41	12	11	3	7
2005 年 1Q	48	25	16	2	1
2005 年 2Q	30	20	13	0	3
2005 年 3Q	27	7	15	4	2
2005 年 4Q	34	14	13	2	6
2006 年 1Q	25	15	15	5	7
2006 年 2Q	38	13	5	2	4
2006 年 3Q	48	11	12	1	4
2006 年 4Q	34	12	7	3	2
2007 年 1Q	45	15	6	2	5
2007 年 2Q	33	16	5	2	3
2007 年 3Q	44	15	6	0	7
2007 年 4Q	42	16	2	2	6
2008 年 1Q	33	19	7	3	5
2008 年 2Q	32	11	5	3	0
2008 年 3Q	52	23	7	5	5
2008 年 4Q	50	16	6	2	4
2009 年 1Q	59	6	3	5	5
2009 年 2Q	72	13	2	3	4
2009 年 3Q	41	6	3	1	3
2009 年 4Q	49	6	2	4	2
2010 年 1Q	56	11	5	3	2
2010 年 2Q	68	10	2	2	1
2010 年 3Q	44	5	3	0	2
2010 年 4Q	46	8	1	4	2
2011 年 1Q	43	1	0	1	0