インターネットコンテンツ容量計測法と次世代インターネット通信モデルの提案

中川晋一

東京都小金井市貫井北町４－２－１
独立行政法人　通信総合研究所

概要

次世代インターネット通信量の予測を目的として、２００１年、現在のわが国インターネットのコンテンツ量を人口 200 万人（都市部と郡部の割合がほぼ１：１）の２県を対象として総当りで教育機関ならびに学術機関の WWW ホームページ容量を実計測した。その結果学校関のホームページ（小中高大学）の合計容量は２県とも２ギガバイト以下、機関別容量は平均２メガバイトの指数分布を示した。また、それぞれのホームページへのインターネットからの到達性についても検討し、アクセス回線容量、ASP（アプリケーションサービスプロバイダ）の比率から今後のデータ通信量を検討した。

Estimation of total Volume of Contents and Next Generation Internet Traffic using the Survey of World Wide Web server on the Current Internet.

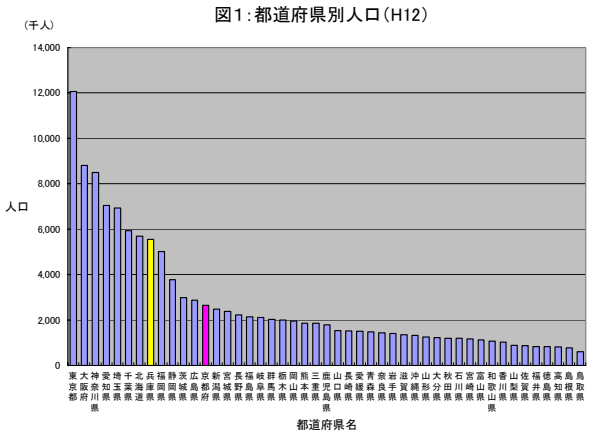
Shin-ichi Nakagawa,
E-mail: snakagaw@crl.go.jp
Communications Research Laboratory, Japan

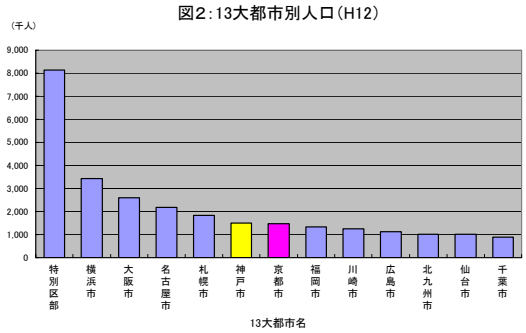
To estimate the future amount of the Next Generation Internet traffic model, we conducted a survey, in 2001, of the World Wide Web homepage contents amount which educational and academic institutions held in two prefectures each with a population of two millions (where the population of its city area equals to that of its rural districts). The result is; the total amounts for homepages at the educational institutions (primary schools, junior high schools, high schools and universities) in both two prefectures are less than 2.5 Giga Bytes; the index number of homepage contents amount by institution distributed over the average of 2 Mb.　Besides, we investigated the attainability of each homepage through the Internet and estimated the future amount of data traffic from the ratios of the circuit capacities and the ASPs.

Key words: Internet, World Wide Web, Volume of Contents, Traffic Model

はじめに

インターネットのコンテンツ情報量の推定は、いわゆる home page の Hosting を行う ASP (Application Service Provider)のコスト算定根拠となるばかりでなく、インターネットサービスプロバイダにとっての回線設計根拠、バックアップセキュリティを考える上においても重要である。例えば 2002 年 3 月 31 日現在で、日本の公立小中高校全体における教育用コンピュータ設置台数は、1 校あたり 32.4 台、インターネット接続率 97.9%、ホームページ保有率 45.7%（文部科学省調べ[1]）であるといわれている。インターネットホームページによる情報提供は１９９５年頃から開始され、「現在、世界では年間 1 から 2 エキサバイトの「ユニークな」情報が生み出されている。これは、地球上の男性、女性、子供一人あたりにするとおおよそ 250 メガバイトの情報となる。」といわれている[2]。中島らはわが国のインターネットコンテンツ統計を行い[3] 、2002 年 2 月における、わが国の Web コンテンツの総量はサーバ数 197000、ページ数 65,550,000、ファイル数 173,880,000、データ量 5,002 ギガバイトであると推定した。また、同報告で近年 1998 年調査開始時サーバ数 36,000、データ量 305 ギガバイトであり、ページ数の増加が 1998 年から 99 年にかけての増加率が約 3 倍、サーバ数が 2 倍に増加した。その後 1999 年から 2002 年にかけサーバ数は約 2 倍、ページ数も約 2 倍に増加したのに対してデータ量の増加は葉ファイル数の増加に比べて大きいこと、その原因の一つに 2000 年頃から一般化した PDF（Portable Document Format）増加の影響も指摘している。これらの調査は、コンテンツの平均値ならびに総量はされているものの、調査のベースが検索エンジンを用いた機械（またはソフトウェア）による推計であること、全数調査が理論的には可能であるにも関わらず、サンプリングデータでいっていることから全体傾向を示唆するデータではあるが検討が必要である。





わが国 Web コンテンツ量推定用母集団統計モデル推定のため、人口 555 万人の H 県と 220 万人 K 府の小中高学校と大学の全ホームページを手作業で調査、ホームページデータ量統計分布を求め、二県の総人口ならびに各学校種別による分布の相似性に関して検討した。同定した URL への到達性を検索することによって、各 URL の運営状態を調査した。

方法

対象として用いた H（人口 550 万人）、K（220 万万人）は、それぞれ県庁所在地人口 100 万人規模の政令指定都市を持つ。わが国全体の分布との対比を図１，２に示した。

1. ホームページ全数調査法

文部科学省学校基本調査[4]に基づいて県下／府下の公立学校数を母数とした。また、個々の学校名称のリストは、県／府教育委員会や各市町村役所や教育委員会のホームページから母数との照合を行いつつ作成した。学校名称を元に、研究目的で K12 サイトのポータルサイトを運営している URL[5]、民間検索エンジン（Yahoo, Infoseek, Google 等）に作成した学校名称リストを直接打ち込んで検索、サーチ結果から一般名称として学校のホームページ以外で用いられている場合を手作業で除外、該当するホームページを探索し URL を同定した。両府県とも大学での開設率は１００％であった。

２．各ホームページ容量の計測と URL への到達性調査

上記にて得られた URL リストを元に、わが国の IPv4 インターネット集約点の一つである NSP-IXP2[6]に対して 45Mbps 以上の帯域で接続される通信総合研究所実験ネットワークから、到達可能な URL ファイルツリーの全ファイルダウンロードと、該当する URL へのインターネット伝送路探索を行った。ホームページファイルダウンロードにはオフラインでのホームページ閲覧用として開発されたホームページデータダウンロード用ソフトウェア[7]を用い、到達できた URL からの関連リンクをダウンロードしない（Bergman[8]の Deep web 条件を配慮、同一ドメイン内まで、教育委員会等へのリンクが張られている場合は手作業で除き）条件で容量を計測した。また、同時に`Traceroute' [9]を用い、該当する URL の所属ドメインの種別を調査した。また、統計解析には Windows 版 SPSS 10.1J を用いた。

表１．H と K の学校数、HP 開設数

学校別	府県別	学校数				HP開設校							
		計	設置者別										
			県立府立	市立	私立	県立府立	%	市立町立*1	%	私立国立*2	%	計	%
小	H	857		845	12			**292**	**35%**	9	75%	301	35%
	K	452		444	8			**294**	**66%**	8	100%	302	67%
中	H	396		362	34			147	41%	31	91%	178	45%
	K	202		179	23			127	71%	23	100%	150	74%
高	H	228	151	25	52	115	76%	20	80%	46	88%	181	79%
	K	105	55	9	41	54	98%	9	100%	40	98%	103	98%
全体	H	**1,481**	151	1,232	98	115	76%	459	37%	86	88%	660	**45%**
	K	**759**	55	632	72	54	98%	430	68%	71	99%	555	**73%**

結果と考察

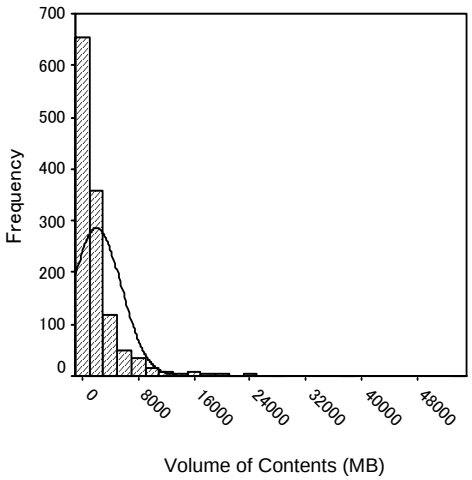
ホームページ開設率

調査は２００２年６月から８月まで上記ネットワークから行った。表１に今回 H,K府県の各教育機関種別のホームページ開設数（到達可能な URL の保有）ならびに率を示した。その結果、HP 開設率は H が 45%に比べ、K は 75%と高率であり、学校開設者別では私立学校が高率であった。両府県のホームページ開設率の差は学校運営母体である市立、町立学校で高率な K の方が高いことが示唆された。また、小学校、中学校、高等学校の順に HP 開設率が増加する傾向があった。これらのことから、今回の調査の母数を表 1 の HP 解説数とした。

Web コンテンツ容量の統計的検討

H,K それぞれの教育機関種別の１URL ツリーの含む Web コンテンツのデータ量の平均標準偏差ならびにそれぞれの種別の合計容量を表２に示す。また、各教育機関種別のコンテンツ容量を比較するため、H,K それぞれにおいて小学校（ES）,中学校（JH）, 高等学校（HS)、大学（Univ.）を独立因子とする１元配置分散分析を行った。その結果、ES,JH,HS の三者では等分散性がなかったのに対して Univ が他の教育機関種に比べ、有意にコンテンツ量が多いことが Tukey、Scheffe 両方で $p<0.05$ で示唆された。また、統計処理を行うにあたり、H の 1 大学だけが合計 URL ツリー容量 2 ギガバイトと統計的極値を示したため、今回の解析からは除外した。採取した 1262 個の URL ツリーのデータ量の頻度分布は図３のように指数分布を示した。府県別、４つの教育機関種別でも同様に全て指数分布であり、各サイト毎の Web コンテンツの情報量がこの対象群では指数関数分布であることが示唆された。また、各教育機関種別の H,K 両府県間のコンテンツ容量の平均値の相異を検定するため、Unpaired-T 検定を行った結果、等分散性、平均値とも $p<0.05$ で棄却された。以上から、各教育機関種別、府県別の統計学的近似性（平均、分散、分布形式）が示唆された。

図３：Frequency of Web Contents (MB)



ネットを用いた情報提供が十分に浸透した対象で行った調査結果ではないことを前提としなければならないが、ランダムサンプリングではなく全数を調査したことから現状を把握するための一手法として今後も検討を加えていく必要性が示唆された。また、対象とした府県の人口は合わせて 770 万人であり、2002 年 770 万人の教育機関の Web コンテンツ容量の実測値は示していると考えると今回得られた 5.6（Gbyte/770 万人）から比例すると仮定すれば、407G バイト

になる。あくまで表層（Surface）の Web コンテンツの容量ではあるが、小学校から大学の合計が 500 ギガバイト程度であるとすれば、セキュリティを含めて新しいコンテンツ流通の可能性も示唆される。例えばサイトセキュリティやネットワークそのもののセキュリティも有効ではあるが、超広域ネットワークに分散していても全コンテンツバックアップを定期的に行うことができる可能性が示唆される。中島[3]もわが国の".jp"の Web データ容量はランダムサンプリング法ではあるが約 4 テラバイトから 5 テラバイトの間であると推定している。インターネット情報量が Web を中心とするサーバクライアントモデルでデータ提供を行う、固定的で"Fixed"なコンテンツ提供を行う様式に限っては全量バックアップが可能､特に ac ドメインを中心とした教育機関の Web サービスに関してはたかだかパーソナルコンピュータで数十万円のコストで全量をバックアップすることが不可能ではないことが示唆される。500 ギガバイトのデータ伝送を帯域１Gbps のネットワーク伝送性能で行えば約 4000 秒＝約 1 時間で終了する可能性も示唆された。

中川らは、ネットワーク社会の情報流通の担い手として重視される遠隔医療の必要とするデータ伝送容量を推計し、わが国の医療機関のうち病院（施設数約 1 万）の必要とするデータ伝送量は 1 施設あたり約 128K ビット/秒（静止画とテキストベースのデータ伝送であると仮定）であると推定した[10]。また、インターネット手順でのデジタルビデオ伝送（DV：１ストリーム 50Mbps）[11]、フルデジタルビデオ伝送（D1：約３００Mbps）が実用化[12]され、今後も SHD-TV 伝送（1.4－5.6Gbps [13],[14]）等の当大容量コンテンツ伝送も予想され、アドレス空間を一戸あたりとして考えることが出来た IPv4 からそれぞれのデバイスに対して個々にアドレスをアサインすることが可能な IPv6 への移行と相乗して 1998 年頃からの静止画＋テキストをサーバクライアントモデルで配布するというコンテンツ流通形態が変化する可能性も示唆される。

まとめ

2002 年におけるわが国のインターネット情報量と次世代インターネットにおける通信量予測のための母数となる基礎データを検討するため、人口 220 万の K、550 万人の H 府県において、小中高大 4 種の教育機関の全数検索と提供されている URL からダウンロード可能な Web コンテンツの全量をダウンロードし、得られた結果を検討した。その結果、

１．Web コンテンツは 2 府県合わせて 5.6 ギガバイトであり、各教育機関ならびにそれぞれの府県において分布、平均値、分散が相似しており、統計学的近似性のある可能性が示唆された。頻度分布は全て指数分布を示した。

２．また、2 府県を合わせ人口 770 万人に対しての容量であると仮定すれば、わが国の教育

機関の Web コンテンツの総容量は５００GB,たかだか 1 テラバイトであり、今後ハードディスクの低価格化とブロードバンドネットワークの条件によってページのバックアップも可能であることも示唆された。

これらから、Web のような従来型サーバクライアントモデルでの固定的（Fixed）コンテンツ提供の今後のあり方、流通形態に関して検討が必要なことが示唆された。

謝辞

本研究を行うにあたり、御指導御助力を頂いた、通信総合研究所主任研究員　島田淳一氏、特別研究員三木まゆみ氏、同土池政司氏、佐分利友木子氏、元特別研究員小巻有子氏、同蒲池光一氏、東京工業大学山岡克式博士、ならびに通信総合研究所諸氏に深謝する。

参考文献

[1] 学校における情報教育の実態等に関する調査結果, 文部科学省, 2001
[2］Peter Lyman, Hal R. Varian, James Dunn、Aleksey Strygin、Kirsten Swearingen , "How much Information", 2001 http://www.sims.berkeley.edu/research/projects/how-much-info
[3]中島睦晴, 島田博也.インターネットコンテンツ統計に関する調査研究、http://www.iptp.go.jp/reserch/monthly/2002/168-h14.09/168-asearch2.PDF, 2002
[4] 学校基本調査, http://www.mext.go.jp/b_menu/toukei/, 2002
[5] 大阪教育大学, 「インターネットと教育」http://www.osaka-kyoiku.ac.jp/educ/, 2002
[6] WIDE Project, http://nspixp.sfc.wide.ad.jp/
[7] BUG, “波乗野郎” Version 3.12, http://www.bug.co.jp/nami-nori/, 2001
[8] M.K. BERGMAN, The Deep Web: Surfacing Hidden Value, http://www.brightplanet.com/deepcontent/tutorials/DeepWeb/index.asp, 2001
[9] G. Kessler, S. Shepard, `A Primer On Internet and TCP/IP Tools and Utilities', RFC 2151, 1997
[10] 中川晋一、田中健二、小峯隆宏、岡沢治夫、久保田文人「次世代医療情報ネットワークへの提案 -次世代インターネット技術開発からの検討-」医療とコンピュータ（日本電子出版） Vol.10 No.10 pp.25-30, 1998
[11] 村井純、小林克志、中村修、小川昭道、杉浦一徳 「DV(Digital Video) Stream on IEEE 1394 Encapsulated into IP between　SC98 and Japan」 10th High performance network and computing conference, 1998
[12] M. Katsumoto, M. Harada, H. Furuse and S. Nakagawa, “Design of the VoD System for High-quality　Video and audio with D1 over IP,” Proceedings of IEEE International Conference on Networking (ICOIN-15), 2001
[13] 田中 健二, 視覚限界に迫る超高精細映像を提示する QHD（Quadruple HD：3840＊2048 画素）プロジェクタ, 映像情報メディア学会技術報告(25-76), pp87-90, 2001
[14] 田中 健二, QHD（Quadruple HD：3840＊2048 画素）映像の主観評価実験と QHD 映像を用いたアプリケーションの提案, 電子情報通信学会, 画像工学研究会 IE2001(124), pp59-63, 2001