IR Usage Analysis for the Next Step: Methodology of Comparative Assessment.
これからのサービスのためのIR利用統計: 比較分析の方法論

SAITO Mika[1)], KINTO Tomonari[1)], TANAHASHI Koreyuki[2)], YAMAMOTO Tetsuya [2)], TSUDZUKI Ichirou[3)], MORIISHI Midori[4)], ITSUMURA Hiroshi[5)], TAKEUCHI Hiroya[6)], SATO Yoshinori[7)].

1) University of Tsukuba Library, 2) Information Technology Center, Nagoya University, 3) Kyoto University Library, 4) Nagasaki University Library, 5) Graduate School of Library, Information and Media Studies, University of Tsukuba, 6) Faculty of Letters, Chiba University, 7) Faculty of Letters, Tohoku Gakuin University.

Introduction

Digital technology has brought about new ways to collect assessment evidences, as well as to make information resources open to public. Thanks to this progress, enumerative and/or more detailed analysis of 'information use' is now possible just by tracking system logs. It will bring us in-depth understanding of our customers and help us contrive better strategies.

Comparative analysis is one of the prospective approaches to IR assessment, but there remains a room for discussion, a methodology to be established. For example, different applications and versions show different values even when the same log data are used.

Therefore, standardized ways need to be established. Based on experiments conducted with AWStats and Google Analytics, this study aims to propose one such method.

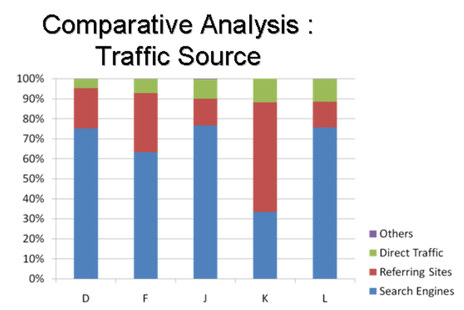
Effectiveness of Access Filter : Log File Size

IR	before (MB)	after (MB)	cut-down rate
A	2,852	2,365	82.9%
B	555	255	46.0%
C	128	72	56.3%
D	507	415	81.7%
E	453	369	81.5%
F	1,412	1,196	84.7%
G	341	313	91.7%
H	787	563	71.6%
I	565	463	81.9%
M	420	363	86.3%

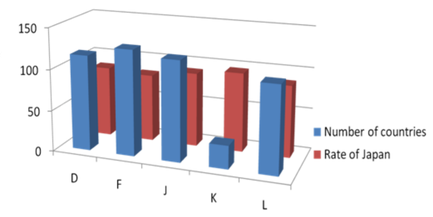
*Period: from July 1 to Dec 31, 2007

First the log files were filtered by a program which extracts the requests having '200 (OK)' or '304 (not modified)' HTTP status codes, and unifies the requests within certain seconds (in this case; 60 seconds). The reduction rates ranged from 46% to 91.7%

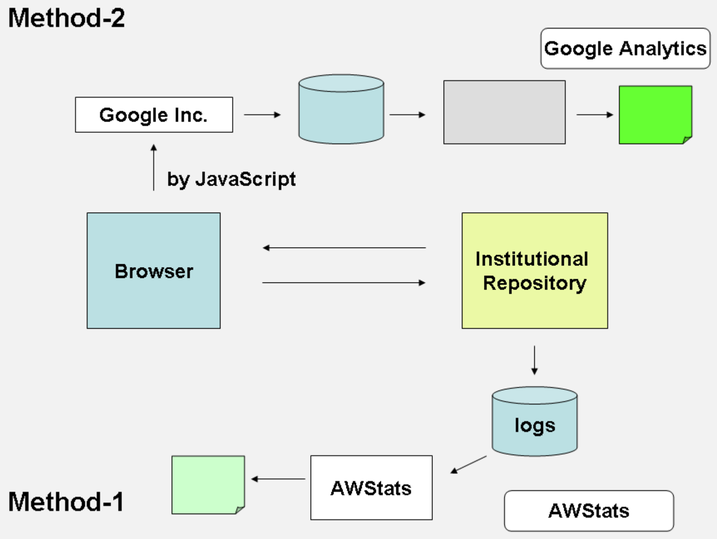
Comparative Analysis : Traffic Source



Comparative Analysis : Traffic
Number of Countries & Rate of visits from Japan



Method-2



	frequency	host name
1	1,352,243	crawl-66-249-73-116.googlebot.com
2	479,149	metasv.nii.ac.jp
3	294,216	local host
4	229,643	shorty.ecs.soton.ac.uk
5	212,839	crawl-66-249-70-122.googlebot.com
6	205,152	Wifi access point in the university library
7	153,142	watchdog.msi.co.jp
8	130,772	peignot5.tulips.tsukuba.ac.jp
9	111,591	spider01.mcm.unisg.ch
10	108,306	spider02.mcm.unisg.ch

As for bots-accesses, it is necessary to eliminate the harvester requests peculiar in IR activities, e.g. 'metasv.nii.ac.jp' 'shorty.ecs.soton.ac.uk' and 'peignot5.tulips.tsukuba.ac.jp,' as well as known and unknown bots. Likewise, the requests from local bots, such as google-mini, should be carefully removed.

*the table represents the ranking of access frequencies to an IR.

AWStats vs Google Analytics

IR – D

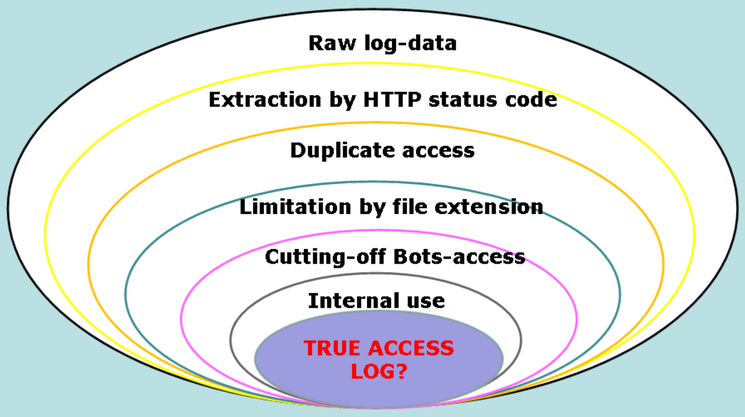
	Jul-07	Aug-07	Sep-07	Oct-07	Nov-07	Dec-07	Total
Google Analytics: Page Views	22,604	20,113	19,273	25,741	25,352	24,873	137,956
AWStats-HTMLs	49,353	36,912	52,252	45,244	74,861	44,312	302,934
AWStats-PDF	12,481	8,457	14,489	11,689	16,630	11,440	75,186

IR – F

	Jul-07	Aug-07	Sep-07	Oct-07	Nov-07	Dec-07	Total
Google Analytics: Page Views	22,525	23,084	38,488	43,475	46,057	42,545	216,174
AWStats-HTMLs	45,078	38,187	104,915	87,434	139,844	81,493	496,951
AWStats-PDF	8,380	6,590	21,065	21,140	32,454	21,659	111,288

As shown in the tables above, the values of Google Analytics are significantly different from the ones of AWStats. One presumable reason for the gap is that some users set the JavaScript code off on their browsers. It is also probable the differences in definition of terms between AWStats and Google Analytics are accountable for the gap.

Method-1: Filtering Weblog



Ref: The Berkeley Electronic Press: the new standard in scholarly Publishing. http://www.bepress.com/download_counts.html

The Interim Result of Method-1

IR	Number of academica staffs (2007.5)	Number of Content (2008.1)	Unique visitors	Number of visits	Hits to HTML or CGI files	Hits to PDF files	Total Hits	Band-width (GB)
A	2,082	21,502	176,966	273,191	1,008,597	230,386	3,536,929	264
B	1,270	20,115	56,738	77,617	286,223	93,756	447,307	93
C	1,032	13,059	3,876	6,867	120,793	14,352	164,663	19
D	1,886	6,612	63,636	91,260	302,934	75,186	523,977	142
E	727	5,445	20,904	26,465	144,931	20,807	252,176	28
F	2,869	22,746	90,132	128,996	496,951	111,288	2,412,809	113
G	2,619	6,922	7,702	10,529	280,993	2,396	489,045	8
H	1,832	12,640	58,916	78,064	387,602	223,481	1,629,261	55
I	2,274	6,202	102,428	135,348	387,496	150,703	694,414	216

time period: Jul 1, 2007 - Dec. 31, 2007

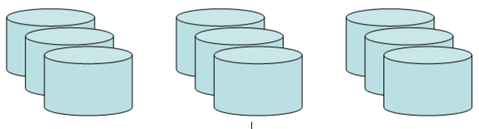
IR	Hits to HTMLs per capita (a staff)	Hits to PDF per capita (a staff)	Hits to HTMLs / Number of Content	Hits to PDF / Number of Content
A	484.44	110.66	46.91	10.71
B	225.37	73.82	14.23	4.66
C	117.05	13.91	9.25	1.10
D	160.62	39.87	45.82	11.37
E	199.35	28.62	26.62	3.82
F	173.21	38.79	21.85	4.89
G	107.29	0.91	40.59	0.35
H	211.57	121.99	30.66	17.68
I	170.40	66.27	62.48	24.30

The tables provide quite a different outlook from the one based on Input-statistics only.

Conclusion

- AWStats is excellent in its usability, functional capability, and graphical representations, but substantial amount of careful adjustment is crucial for getting adequate results. It is therefore recommended that various institutions/ parties cooperate in developing a common tool for IR assessment.
- Google Analytics is also a convenient and excellent tool for fixed-point observation, but is not necessarily suitable for comparable IR statistics. Google Analytics is a tool for identifying the current state of 'IR use' and 'IR users' behavior' rather than for accurate assessment.
- Assuming a common assessment tool is to be developed, the format should be designed to include meta-data of requested materials. OpenURL ContextObject proposed by MESUR Project and supported IRStats Project seems to be preferable.

Method-1: Verification by AWStats



・log file acquisition

Merge & Sort ---logresolvemerge.pl

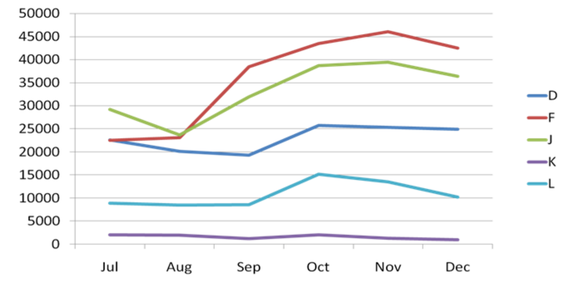
Filtering
・restricting by http status code
・excluding duplicate requests
---access_filter.pl

AWStats ver.6.7J
・limiting by file extension
・skipping robots etc.
・excluding internal use

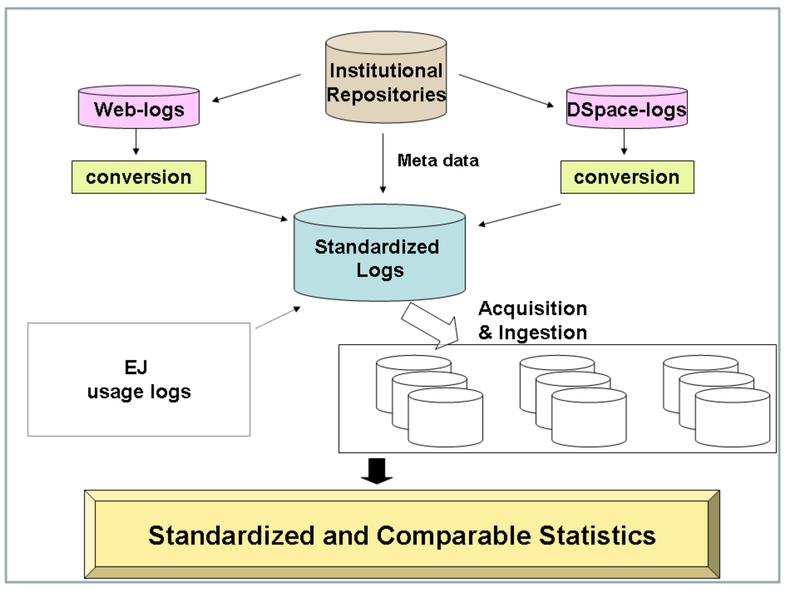
Comparison with tables and graphs
・Usage counts by Institutions, by Countries, etc.
・Percentage of access by the type of materials

Method-2: Google Analytics

Google Analytics is a powerful tool that enables site administrators to easily obtain the statistics. After the sign-up, once a tiny JavaScript code is built in the site, data-gathering starts immediately. The result data will be gained in the form of PDF, XML or CSV format.



Comparative Analysis: Page views



We are grateful to Hokkaido University Library, University of Tsukuba Library, Chiba University Library, Tokyo Institute of Technology, Hitotsubashi University Library, Kanazawa University Library, Nagoya University Library, Mie University Library, Kyoto University Library, Kyoto Institute of Technology Library, Osaka University Library, Hiroshima University Library, Kyushu University Library, and Nagasaki University Library for the cooperation of data provision.