Home

Arroyo

Bar-Ilan

Barjak

Berkowitz

Bjorneborn

Caldas

Choudhury

Fry

Harabi

Hine

Kim

Mayr

Nentwich

Noruzi

Park

Rall

Scharnhorst

Shaw

Thelwall

Uberti

Website entries from a web log file perspective – a new log file measure[1]

Philipp Mayr, M.A. philippmayr @ web.de, Berlin 26-Jun-2004

Introduction

Web log files record user transactions on webservers and offer due to their extent, their properties and potential an excellent investigation field for contemporary information and online behaviour studies [see also *Nicholas et al.*, 1999]. Web log files actually offer the possibility to extract information about user access pattern, site visibility and site interlinking [see also *Thelwall*, 2001]. Furthermore web log file are excellent sources for informational investigations such as tracking spider behaviour, search engine query analysis or cognitive ergonomics. A drill down analysis to the smallest website entity (a specific web page) and also to other web entities, like directories or sites [see also *Thelwall*, 2003] can be performed with log data. This facilitates counting information usage frequencies on different levels of a website and enables new forms of information studies (e.g. finding regularities). Practical website insights for site access optimisation/evaluation are additional guaranteed.

Navigation on the web occurs in three separable types. The majority of online navigation is realized over hyperlinks which are set-up manually (e.g. directory entries, other intellectually build backlinks) or automatically (e.g. search engines, other web-based information systems). Direct navigation (e.g. over bookmarks, browser history) coexists between the “link-based” navigation and can be seen as an indication of well-known and perhaps authoritative websites. The three distinguishable web navigation types "navigation about search engine queries", "navigation about backlinks" and "direct navigation" can be separately identified in log data if the webserver provides the extended log file field “referer” [see also *Thelwall*, 2001].

The study and the WEF measure

Empirical and exploratory information science study from the field introduces new aspects and analysis possibilities for web log data focussing on an academic webserver. The study develops and tests a quantitative, non-reactive measure method for standard log files, the log file measure Web Entry Factor (WEF) that aggregates common usage frequencies for web pages. The WEF provides detailed numbers about the accessibility, visibility and interlinking of highly used entry points of a website. The concept of this study is tested with a 2-year log file sample from an academic website (the Institute of Library Science website at Humboldt-University, Berlin, http://www.ib.hu-berlin.de/) as well as the 100 most frequently accessed entry pages of this site. The log file study compares the results of two years (2000, 2002) and integrates a classification scheme for web pages (*Haas & Grams*, 2000) as well as a page size correlation and correlations with the prominent external parameter PageRank from the search engine Google. The study shows and discusses some very surprising results which are mainly caused by the multitude and importance of text based entry pages (e.g. articles, papers, documentations, …) for this webserver. The results additionally show the dominance of commercial search engines as information gatekeepers and traffic guarantors. The WEF concept is concentrated on “entries”. An entry means a website visit with an identifiable entry pattern (navigation type) from a logfile perspective. Entry or website entry stands for the request on the first web page/start page and gives an idea about the importance of the entity. All other requests relating to a visit will be identified as navigation requests and will not be considered for the log analysis.

WEF values are simple proportional rates for entry requests which can be calculated for web entities like sites or directories and broken down to pages. The listing below shows the counting of the three entry values and the calculation of WEF values for a web entity. The sum of all WEF values of an entity is

always 1, or 100% of all entries. A web entity always has three WEF values (one for each navigation type: 1. WEF_SearchEngine, 2. WEF_Backlink, 3. WEF_Direct). The single WEF values can vary between 0 and 1 (see example for two pages in Tab.1 below).

WEF_SearchEngine + WEF_Backlink + WEF_Direct = 1

count entries for each navigation type for an entity (e.g. n URL)

For each URL = 1 to n

count Entries_SearchEngine

count Entries_Direct

count Entries_Backlink

Entries_total = Entries_Direct + Entries_SearchEngine + Entries_Backlink

Endfor

compute WEF for an entity

For each URL = 1 to n

WEF_SeachEngine = Entries_SearchEngine / Entries_total

WEF_Direct = Entries_Direct / Entries_total

WEF_Backlink = Entries_Backlink / Entries_total

Endfor

Listing: counting entry types and WEF calculation (pseudocode)

The following section of a results table (see Table 1) shows two pages. For instance the page with rank 1, the Institute's homepage has a WEF_Direct = 0.81 which means that 81% of the total page entries came from direct navigation. Backlink navigation is 10% or WEF_Backlink = 0.10. The second page ascii.htm, an ascii documentation gets it's entry traffic mainly over search engines (88%). Backlink entries account for only 1%.

rank	url	description	class	size	PR	Entry Search engine	Entry Direct	Entry Backlink	WEF Search engine	WEF Direct	WEF Backlink	Entry total
1	/	Institute's Homepage	Home	av	6	6345	54369	6558	0.09	0.81	0.10	67272
2	/~mh/ascii.htm	ASCII-Page	Docu	av	4	19248	2399	187	0.88	0.11	0.01	21834

Table.1: a cut out from the results table with detailed information about the rank, url, description, class, PageRank, Entries counts, WEF values and total entries of a specific page (entity)

The WEF values of an entity display a measure of high validity (real usage). They display aggregates of

the usage of it's external link structures (backlinks and queries) and indicators of authority. They enable the numbering of open information access of web entities from an entity perspective. The established link metrics WIF (*Ingwersen*, 1998) and WUF (*Thelwall*, 2003) deliver aggregated views of the existence of link structure. WEF delivers views of the usage of these structures. A combination of log file based and link based measures would be best of advanced information studies [see also *Thelwall*, 2001].

Results

The following results refer to general usage frequencies, the analysis of the three different navigation / access types and the aggregated WEF values for web pages which come up with a detailed picture of the distribution of web traffic to this specific website.

- The webserver got 59% entries over search engine queries, 34% entries over direct navigation, 7% entries over backlinks in the year 2002. The year 2000 shows similar values. It is surprising that search engines play such a dominant role and backlinks deliver only a fraction of the traffic.
- The top 100 frequented entry pages in 2002 consist of 58 text pages (text), 21 organisational pages (orga), 4 documentation (docu), 6 database entries (db_entry), 10 homepages (home), and1 page was in 2003 not available.
- The median of the WEF values of the top 100 web pages was in 2002:
    - median WEF_SearchEngine = 0.81
    - median WEF_Direct = 0.15
    - median WEF_Backlink = 0.02

Table 2 shows average WEF values and average PageRank values for 100 pages classified in five content clusters (c.f. *Haas & Grams*, 2000).

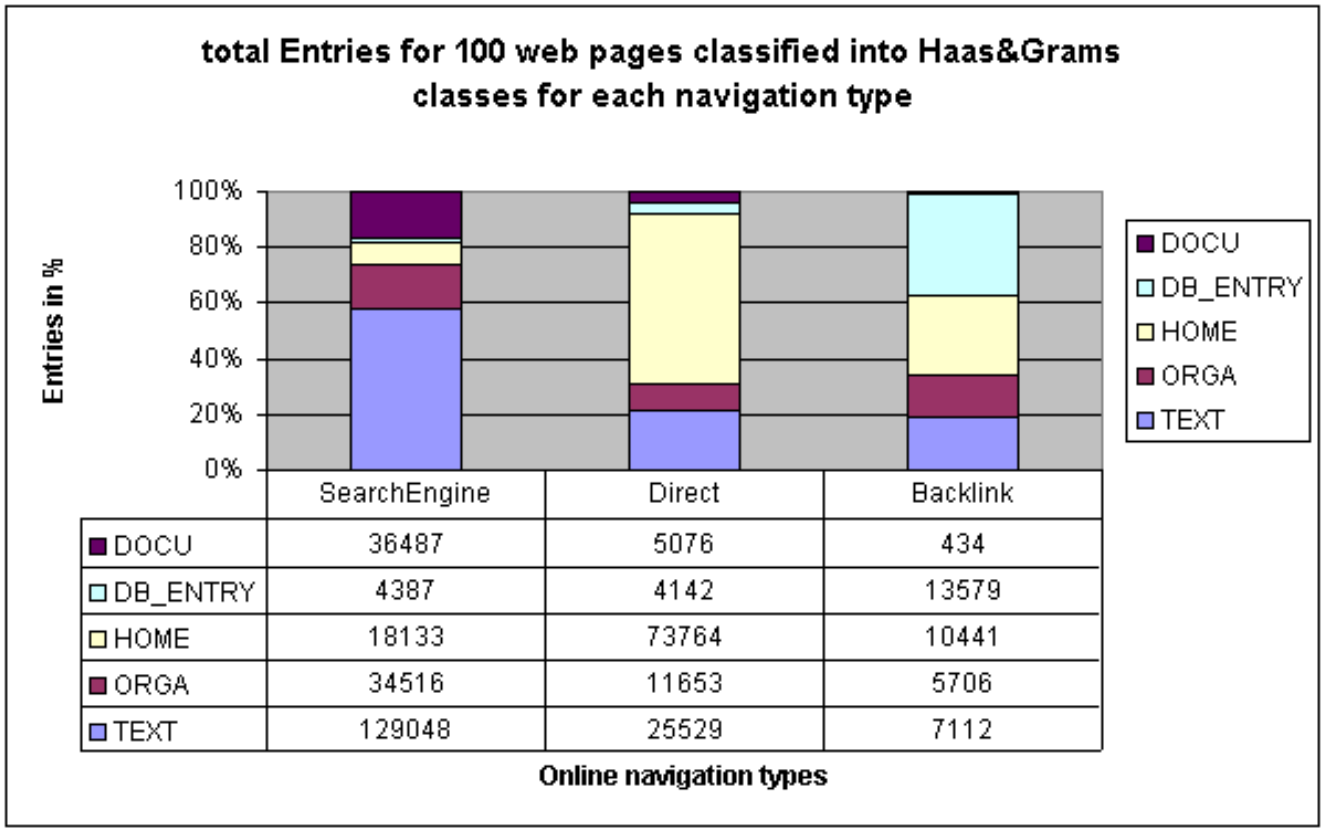
Content clusters	PageRank (average)	WEF_Backlink (av.)	WEF_Direct (av.)	WEF_SearchEngine (av.)
DB_ENTRY (n=2)	5.25	0.57	0.22	0.21
HOME (n=10)	4.90	0.14	0.55 (0.52 without root-Homepage)	0.32
ORGA (n=21)	3.90	0.10	0.26	0.65
TEXT (n=58)	3.71	0.04	0.16	0.80
DOCU (n=4)	3.17	0.02	0.12	0.86

Table .2: average PageRanks and WEF values for the 100 top used entry page (2002)

Further results:

- the log data show power laws in the distribution of the total entries of the 100 most frequented web pages (rsq for the year 2002 = 0.961).
- the log data show positive correlation (Spearman = 1.0) between the average entries measured in WEF and PageRank values (Google Toolbar value) of a web page cluster.
- the log data show power laws in the distribution of search engine query phrases (query strings in the URL of the search engine entry) and frequency of appearance in the log file. The majority of query phrases used in the recorded search engine queries appear only very few times in the log file (see *Thelwall*, 2001).
- the log data show power laws in the distribution of backlinking websites and traffic they transfer to the analysed website. Only a few external website transfer large amounts of traffic (see also *Thelwall*, 2001).

The following figures show some other results from the study.



	SearchEngine	Direct	Backlink
DOCU	36487	5076	434
DB_ENTRY	4387	4142	13579
HOME	18133	73764	10441
ORGA	34516	11653	5706
TEXT	129048	25529	7112

Figure 1: total entry counts for the 100 most used entry points classified into Haas & Grams clusters (2002)

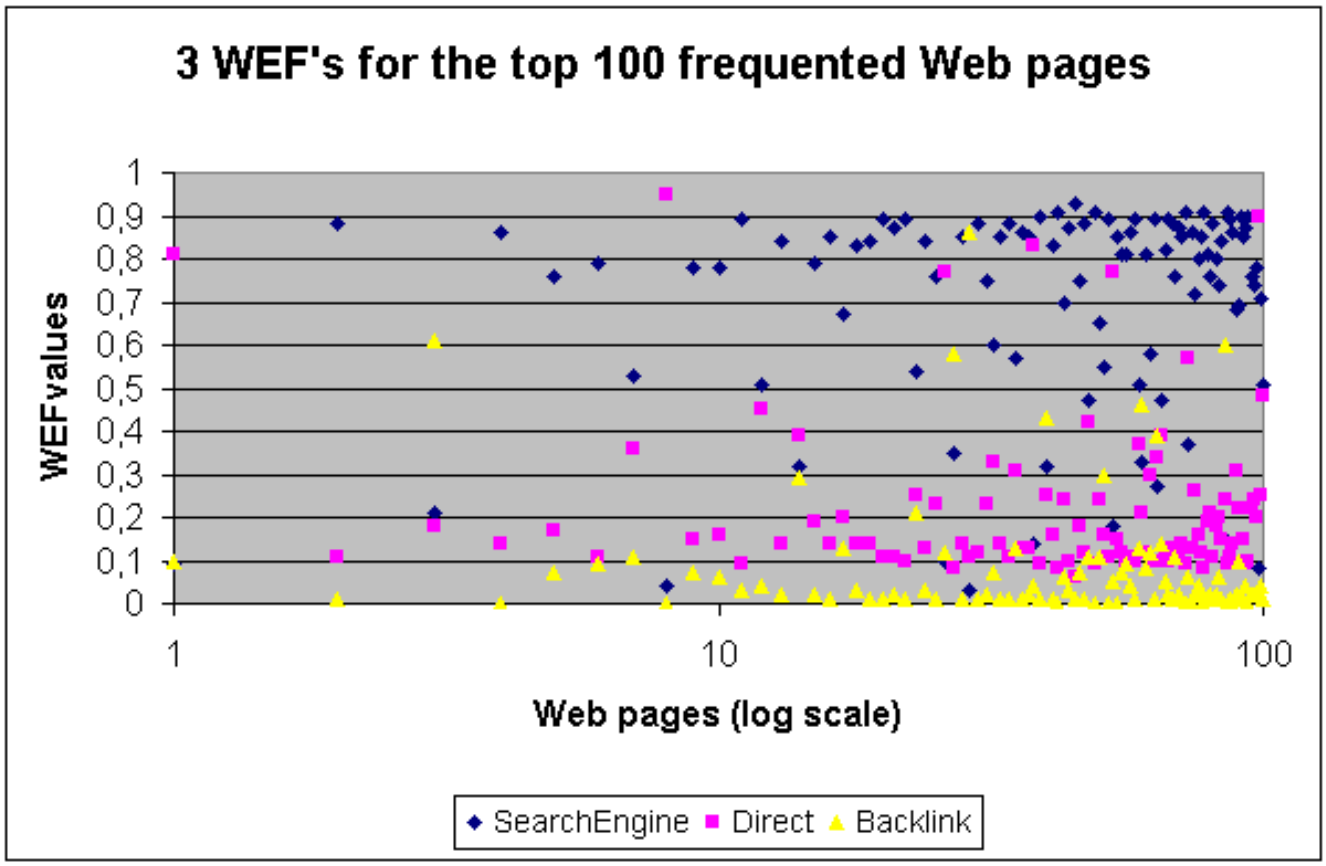
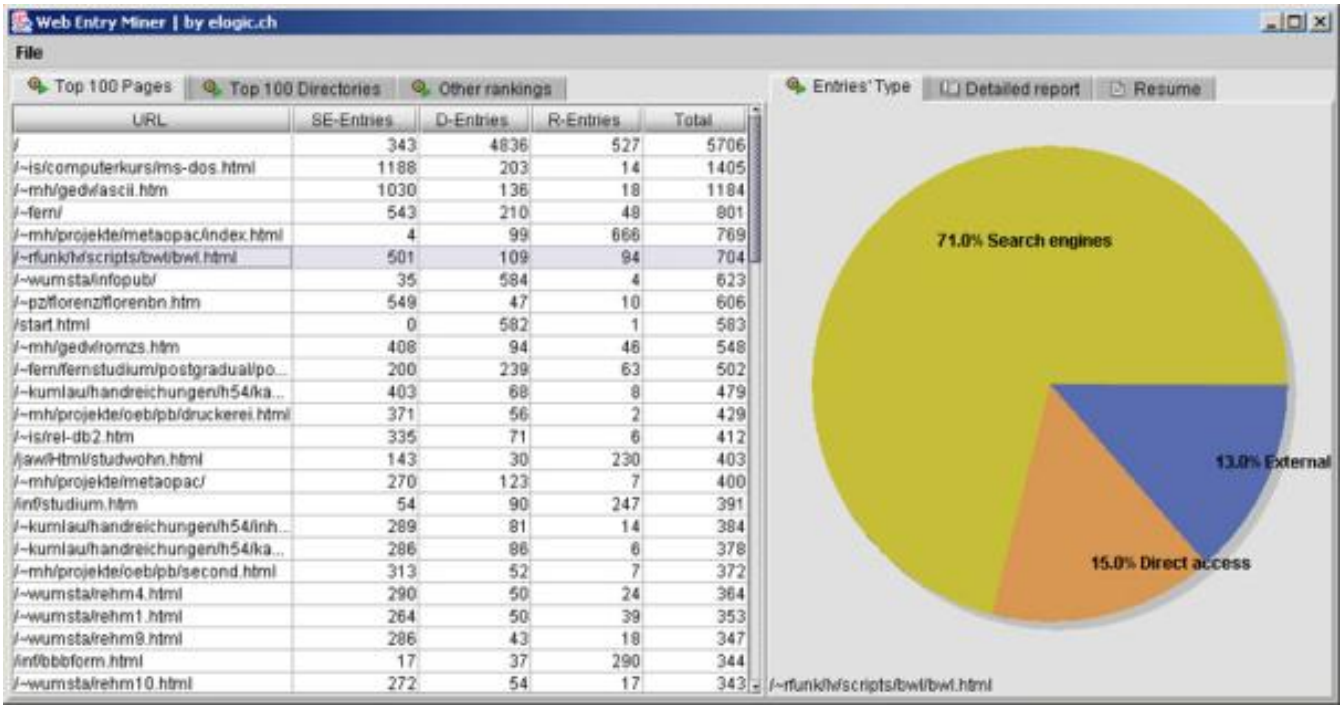


Figure 2: a picture which shows the distribution of all WEF values for the 100 most used entry points (2002). Note: logarithmic scale

WEM

Under http://www.ib.hu-berlin.de/~mayr/wem/ you will find a small Java application, called WEM (Web Entry Miner), which has implemented the WEF concept. WEM displays the top 100 pages and their entries distinguished in search engines (SE-Entries), backlinks (R-Entries) and direct (D-Entries). The only restriction is that the log files have to be from an Apache webserver and the referer field has to be recorded.



URL	SE-Entries	D-Entries	R-Entries	Total
/	343	4836	527	5706
/~is/computerkurs/ms-dos.html	1188	203	14	1405
/~mh/gedv/ascii.htm	1030	136	18	1184
/~fern/	543	210	48	801
/~mh/projekte/metaopac/index.html	4	99	666	769
/~rfunk/hl/scripts/bwl/bwl.html	501	109	94	704
/~wumsta/infoputz/	35	584	4	623
/~pz/florenz/florenbn.htm	549	47	10	606
/start.html	0	582	1	583
/~mh/gedv/romzs.htm	408	94	46	548
/~fern/fernstudium/postgradual/po...	200	239	63	502
/~kumlau/handreichungen/h54/ka...	403	68	8	479
/~mh/projekte/oeb/pb/druckerei.html	371	56	2	429
/~is/rel-db2.htm	335	71	6	412
/jaw/Html/studwohn.html	143	30	230	403
/~mh/projekte/metaopac/	270	123	7	400
/inf/studium.htm	54	90	247	391
/~kumlau/handreichungen/h54/inh...	289	81	14	384
/~kumlau/handreichungen/h54/ka...	286	86	6	378
/~mh/projekte/oeb/pb/second.html	313	52	7	372
/~wumsta/rehm4.html	290	50	24	364
/~wumsta/rehm1.html	264	50	39	353
/~wumsta/rehm9.html	286	43	18	347
/inf/bbbform.html	17	37	290	344
/~wumsta/rehm10.html	272	54	17	343

Pict. 3: Screenshot of the WEM – Web Entry Miner

see http://www.ib.hu-berlin.de/~mayr/wem/

References

Haas, Stephanie; Grams, Erika: Readers, Authors, and Page Structure: A Discussion of Four Questions Arising from a Content Analysis of Web Pages. In: Journal of the American Society for Information Science and Technology, Vol. 51, 2000, S. 181–192.

Mayr, Philipp: Entwicklung und Test einer logfilebasierten Metrik zur Analyse von Website Entries am Beispiel einer akademischen Universitäts-Website / von Philipp Mayr. Berlin : Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin, 2004, 106 S. - (Berliner Handreichungen zur Bibliothekswissenschaft und Bibliothekarsausbildung ; 129) available: http://www.ib.hu-berlin.de/~kumlau/handreichungen/h129/ (04/2004)

Nicholas, David et al.: Developing and testing methods to determine the use of web sites: case study newspapers. In: Aslib Proceedings, Vol. 51, 1999, S. 144-154.

Nicholas, David, et al.: Cracking the code: web log analysis. In: Online & CD-ROM Review, Vol. 23, 1999, S. 263-269.

Thelwall, Mike: Web log file analysis: Backlinks and Queries. In: Aslib Proceedings, Vol. 53, 2001, S. 217-223.

Thelwall, Mike: Methods for reporting on the targets of links from national systems of university Web

sites. In: Information Processing and Management, to appear 2003.

Thelwall, Mike; Vaughan, Liwen ; Björneborn, Lennart: Webometrics. In: ARIST, Vol. 39, 2004, preprint.

This paper abstract goes back to my master thesis “Entwicklung und Test einer logfilebasierten Metrik zur Analyse von Website Entries am Beispiel einer akademischen Universitäts-Website” in Library Science (M.A.) at Humboldt-University, Berlin, Institute of Library Science in December 2003. Supervisers: Prof. Dr. W. Umstätter (email: h0228kdm@rz.hu-berlin.de), PD Dr. R. Wagner-Döbler (email: rfw-d@t-online.de)

Direct navigation can be tracked by the missing entry (e.g. “-“) in referrer-field.

The term factor is misleading because the measure displays proportional rates between 0 and 1.

The study shows e.g. that homepages and typical start pages of large websites are not necessarily the most frequented entry points.

Classification after *Haas & Grams* 2000

This means the average PageRank value (shown with the Google Toolbar) for a page in the Haas & Grams content class.