Poster by Philipp Mayr & Christian Nançoz (3. November 2005)

Macro- & Micro-Mining

Web server log file examples

Introduction

- Current paradigm change in the scholarly publication system (from print to online)
- Growing share and importance of Open Access (OA) documents in the scholarly communication process; few evaluation criteria (e.g. link analysis, usage data)
- No robust web-based methods (indicators)

Log file basics

- Log files are an excellent data source for studying the accessibility, visibility & interlinking of OA content
- Log data is used for website analysis, user modelling & analysis of information behavior (e.g. search engine usage)
- Log files are structured enough to extract pattern and can be used for measuring web impact of a certain entity

```
141.20.20.10 - - [20/Jul/2002:22:50:55 +0200] "GET /~wumsta/ubach/fuss.htm HTTP/1.1" 200 54988 "http://www.backlinks.com/backlink1.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
141.20.20.10 - - [20/Jul/2002:22:50:55 +0200] "GET /~wumsta/ubach/IMG00056.GIF HTTP/1.1" 404 307 "http://www.ib.hu-berlin.de/~wumsta/ubach/fuss.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
141.20.20.10 - - [20/Jul/2002:22:50:55 +0200] "GET /~wumsta/ubach/IMG00057.GIF HTTP/1.1" 404 307 "http://www.ib.hu-berlin.de/~wumsta/ubach/fuss.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
141.20.20.10 - - [20/Jul/2002:22:51:55 +0200] "GET /~wumsta/ubach/index.html HTTP/1.1" 200 19797 "http://www.ib.hu-berlin.de/~wumsta/ubach/fuss.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
200.109.102.199 - - [20/Jul/2002:22:53:27 +0200] "GET /robots.txt HTTP/1.0" 200 279 "-" "BlitzBOT@tricus.net (Mozilla compatible)"
203.122.23.145 - - [20/Jul/2002:23:14:37 +0200] "GET /~pbruhn/gruppe04.htm HTTP/1.1" 200 62766 "http://www.google.de/search?q=%2B%22russische+Frauen%22" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)"
203.122.23.145 - - [20/Jul/2002:23:14:37 +0200] "GET /~pbruhn/photo.jpg HTTP/1.1" 200 62766 "http://www.ib.hu-berlin.de /~pbruhn/gruppe04.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)"
200.109.102.199 - - [20/Jul/2002:23:14:38 +0200] "GET /index.html HTTP/1.0" 200 279 "-" "BlitzBOT@tricus.net (Mozilla compatible)"
203.122.23.145 - - [20/Jul/2002:23:14:39 +0200] "GET /~pbruhn/index.htm HTTP/1.1" 200 62766 " http://www.ib.hu-berlin.de /~pbruhn/gruppe04.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)"
141.20.20.10 - - [20/Jul/2002:23:55:55 +0200] "GET /~wumsta/index.html HTTP/1.1" 200 19797 "http://www.ib.hu-berlin.de/~wumsta/ubach/fuss.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
```

Fig. 1: A log file sample showing three virtual users requesting content via different access pattern (Web Entry pattern)

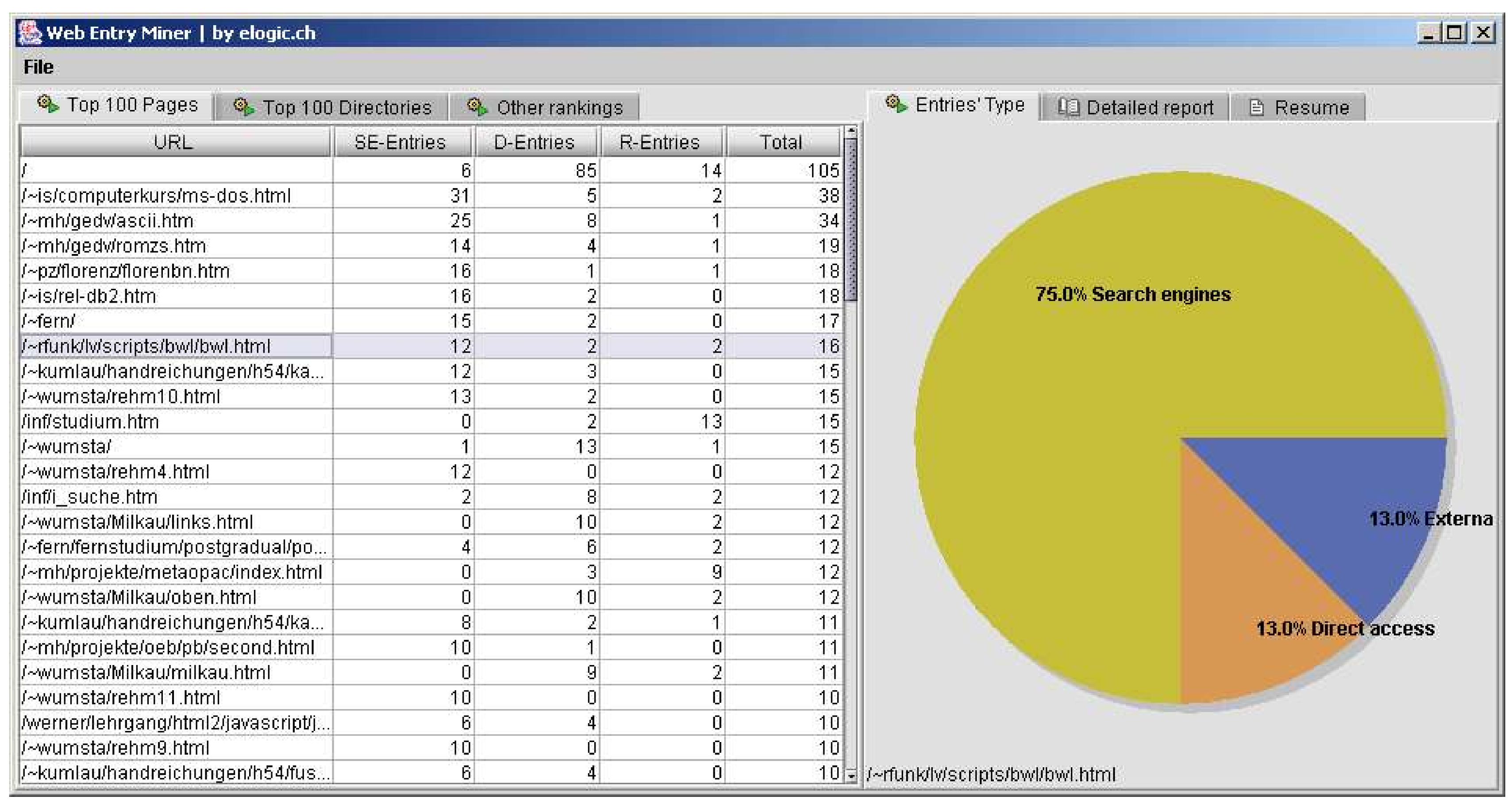
Macro-Mining approach

Macro analysis is state of the art in popular log analysers: they aggregate usage data to common measures (benchmarks) like visits or views to create a macro view on a website

- How much traffic does an entity receive?
- Where are the most important entry pages?
- Where are the users coming from?
- How often do users return?

Advanced macro-mining approach: Web Entry analysis

- Is based on a heuristic that distinguishes three website entry types: 1) “search engines”, 2) “backlinks” or 3) “direct access” (see Fig. 1)
- Web Entry analysis measures the entry ratios for a certain document or entity (see Fig. 2)
- These Web Entry ratios show detailed insights on the accessibility, visibility and interlinking of different levels of a website and can be the basis of deeper analysis (e.g. Micro-Mining)



URL	SE-Entries	D-Entries	R-Entries	Total
/	6	85	14	105
/~is/computerkurs/ms-dos.html	31	5	2	38
/~mh/gedv/ascii.htm	25	8	1	34
/~mh/gedv/romzs.htm	14	4	1	19
/~pz/florenz/florenbn.htm	16	1	1	18
/~is/rel-db2.htm	16	2	0	18
/~fern/	15	2	0	17
/~rfunk/lv/scripts/bwl/bwl.html	12	2	2	16
/~kumlau/handreichungen/h54/ka...	12	3	0	15
/~wumsta/rehm10.html	13	2	0	15
/inf/studium.htm	0	2	13	15
/~wumsta/	1	13	1	15
/~wumsta/rehm4.html	12	0	0	12
/inf/i_suche.htm	2	8	2	12
/~wumsta/Milkau/links.html	0	10	2	12
/~fern/fernstudium/postgradual/po...	4	6	2	12
/~mh/projekte/metaopac/index.html	0	3	9	12
/~wumsta/Milkau/oben.html	0	10	2	12
/~kumlau/handreichungen/h54/ka...	8	2	1	11
/~mh/projekte/oeb/pb/second.html	10	1	0	11
/~wumsta/Milkau/milkau.html	0	9	2	11
/~wumsta/rehm11.html	10	0	0	10
/werner/lehrgang/html2/javascript/j...	6	4	0	10
/~wumsta/rehm9.html	10	0	0	10
/~kumlau/handreichungen/h54/fus...	6	4	0	10

Fig. 2: A screenshot of a macro analysis via the Web Entry Miner which implements a heuristic based on Web Entry diversification

IP Adresse	Webseite	Referrer	Browser	Zeit
18 Oktober 2003				
128.xxx.xxx.xxx	/	"http://www.google.com/search?q=disinfo"	"Mozilla/3.01 [de] (Win16; I)"	01:03:52
128.xxx.xxx.xxx	/content.htm	"http://www.disinfojournal.net"	"Mozilla/3.01 [de] (Win16; I)"	01:11:01
128.xxx.xxx.xxx	/issue1_1.htm	"http://www.disinfojournal.net/content.htm"	"Mozilla/3.01 [de] (Win16; I)"	01:11:01
128.xxx.xxx.xxx	/free.htm	"http://www.disinfojournal.net/issue1_1.htm"	"Mozilla/3.01 [de] (Win16; I)"	05:23:34
128.xxx.xxx.xxx	/hilights.htm	"http://www.disinfojournal.net/free.htm"	"Mozilla/3.01 [de] (Win16; I)"	02:21.56
128.xxx.xxx.xxx	/authors.htm	"http://www.disinfojournal.net/hilights.htm"	"Mozilla/3.01 [de] (Win16; I)"	12:54:05
128.xxx.xxx.xxx	/about-us.htm	"http://www.disinfojournal.net/authors.htm"	"Mozilla/3.01 [de] (Win16; I)"	03:34:41
128.xxx.xxx.xxx	/index.html	"http://www.disinfojournal.net/about-us.htm"	"Mozilla/3.01 [de] (Win16; I)"	00:30:31
19 Oktober 2003				
128.xxx.xxx.xxx	/hilights.htm		"Mozilla/3.01 [de] (Win16; I)"	24:09.36
128.xxx.xxx.xxx	/authors.htm	"http://www.disinfojournal.net/hilights.htm"	"Mozilla/3.01 [de] (Win16; I)"	03:44:02

Fig. 3: A user tracking protocol for one virtual user identified on an E-journal website

Micro-Mining approach

- Nicholas & Huntington (2003) proposed a study which shows how micro analysis techniques can enhance current log analysis
  - The construction and analysis of a user subgroup
  - Tracking and reporting of individual usage
- Potential to extract specific online behaviour and usage trends for a identifiable group (e.g. academic users)
- The results are more detailed and robust, but are based on a much smaller user group (see Fig. 3)

Combined analysis

Scenario with a combined macro-micro-analysis (see Fig. 4)

- Macro analysis of a log file sample (e.g. Web Entry analysis)
- Drilling down the aggregated macro data
- Adoption of different user groups which were identified by a previous micro-analysis

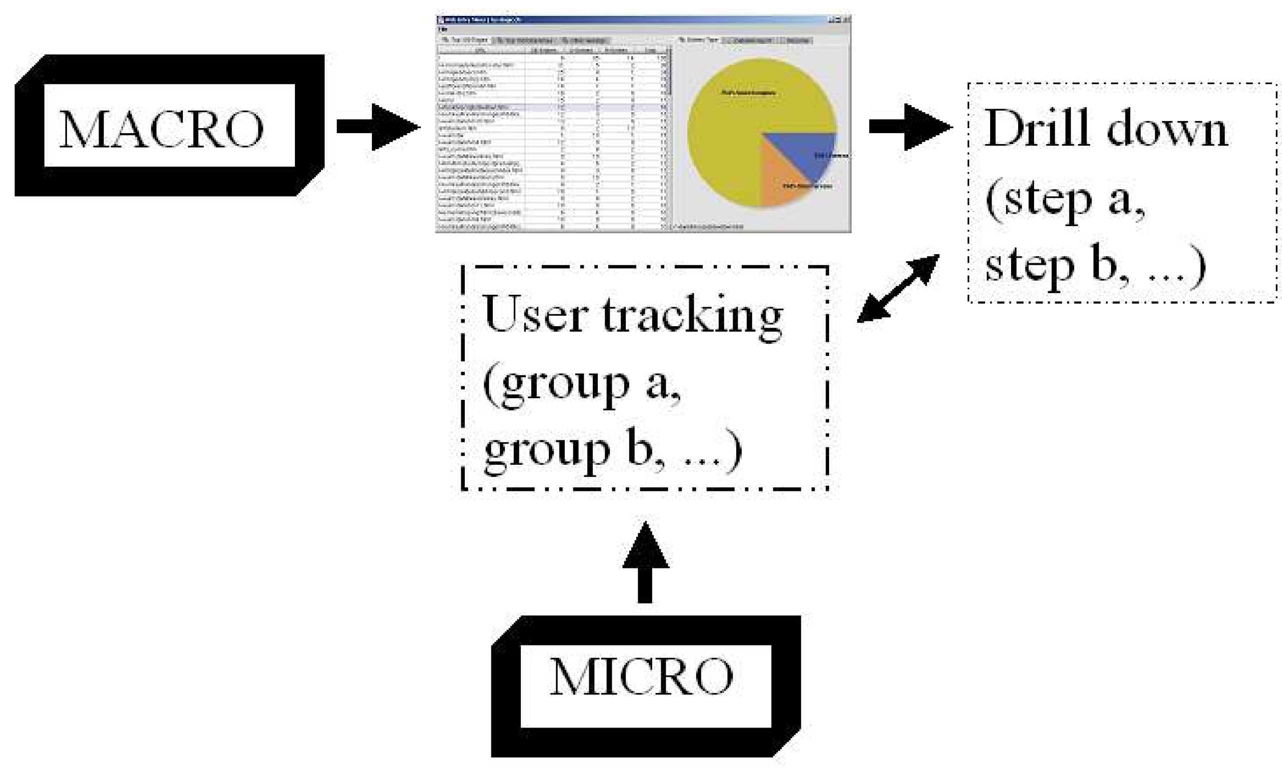


Fig. 4: Scenario of a combined macro-micro-analysis

Conclusion

A combination of macro- & micro log analysis could be a new way:

- to construct more reliable and sophisticated measures
- to test and discover usage trends or problems
- to focus on segmented website and user entities
- to enhance static user information (e.g. geographical or institutional) with dynamic information (specific user behavior)

Future goals in log analysis: Enhance the comfort of web users (e.g. minimizing search

Further research

- Implementation of fuzzy logic especially in micro-mining
- Constructing log file based Web indicators for scholarly information systems
- Describe more combined analysis scenarios
- Eliminating errors in log analysis

Contact

Philipp Mayr, M.A.
email: philippmayr@web.de
www: http://www.ib.hu-berlin.de/~mayr

Christian Nançoz, Msc
Christian.nancoz@elogic.ch

"Transaction log files allow us to look at the behaviour of millions of people, but the aggregation misses the detail and the detail can add to the impressions and thoughts about user behaviour." (Nicholas & Huntington 2003)