Makro- und Mikro-Mining am Beispiel von Webserver Logfiles

von Philipp Mayr & Christian Nançoz

Abstract (deutsch)

Webserver Logfiles sind eine hochinteressante Informationsquelle zur Untersuchung der Zugänglichkeit, Sichtbarkeit und Verlinkung von beliebigen Webinhalten. Dieser Beitrag stellt zwei neuere Ansätze der Logfile Analyse bzw. des Web Mining vor (Makro-Mining & Mikro-Mining). Der weitverbreiteten Methode der Makro-Analyse, die hauptsächlich allgemeine Zugriffszahlen aggregiert (z. B. Anzahl der Downloads eines Dokuments), wird die bislang weniger bekannte Methode der Mikro-Analyse gegenübergestellt. Die Mikro-Analyse konzentriert sich auf schmale Segmente des Logfiles, die bis auf Transaktionen einzelner User zurückgehen. Beide Analysemethoden werden anhand eines Beispiels erklärt. Weiterhin wird versucht neue Einsatzbereiche der beiden Web-Mining Verfahren zu identifizieren und Formen der kombinierten Nutzung der beiden Methoden zu skizzieren.

Abstract (englisch)

Webserver log files are a very interesting data source for analysing the accessibility, visibility and interlinking of any web content. This paper proposes two recent log file or web mining approaches (macro-mining & micro-mining of webserver log files). We try to bring together the popular method called macro analysis which aggregates common server request counts (e.g. number of downloads of a certain document) with the micro analysis method which is less known in log analysis. The micro-mining approach focuses on segmented log files which can be drilled down to transactions of single users. Both analysis methods will be explained by an example. Furthermore we try to identify new use cases and try to sketch ways of combined analysis for both web mining methods.

1. Einleitung

Als Folge des Paradigmenwechsels im wissenschaftlichen Publikationswesen werden immer häufiger Forderungen nach Bewertungskriterien deutlich, die den spezifischen Eigenschaften der neuen Publikationstypen Rechnung tragen. Die zunehmende Anzahl und wissenschaftliche Bedeutung von Online-Publikationen z.B. aus dem Open Access Bereich, macht es daher notwendig, innovative webbasierte S&T-Indikatoren zu entwickeln. Diese Web-Indikatoren sollen helfen, die neuen Publikationen einzuordnen und ggf. zu bewerten. Web-Indikatoren werden laut WISER charakterisiert als:
„A policy relevant measure that quantifies aspects of the creation, dissemination and application of science and technology in so far as they are represented on the internet or the World Wide Web.” (vgl. dazu WISER Projekt[1])

[1] WISER steht für Web Indicators for Science, Technology & Innovation Research, siehe http://www.wiserweb.org

Das Fehlen von robusten webbasierten Messmethoden, beispielsweise zur Impactbestimmung wissenschaftlicher Literatur, wirkt sich für alle Akteure, die in den wissenschaftlichen Prozess involviert sind (Wissenschaftler, Bibliothekare, Gutachter, …), als sehr störend aus. Das Hauptproblem bestellt darin, dass bislang noch keine tragfähigen Indikatoren in Sicht sind, die die Einschränkungen aber auch Potenziale des neuen Publikationsmediums Internet genügend berücksichtigen.
Die Webometrie (Björneborn & Ingwersen 2001, Thelwall, Vaughan & Björneborn 2003), die wie die Zitationsanalyse (vgl. Bibliometrie) auf quantitativen Methoden basiert, verspricht, hier ansatzweise Hilfestellung, aber noch keine Lösungen anzubieten. Neben der Analyse der Hyperlinkstrukturen im Web, bieten sich zur Analyse auch die Nutzungsinformationen an, die auf Webservern durch die Protokollierung der Onlinebenutzung entstehen. Dieses Analyseverfahren nennt sich Logfile Analyse.

Die Logfile Analyse, z.T. auch Web-Mining genannt, soll im Mittelpunkt des folgenden Beitrags stehen.

2. Logfile Analyse

Die Logfile Analyse ist als Verfahren zur nichtreaktiven Nutzungsmessung von Webnutzern allgemein anerkannt und erfreut sich trotz einiger systematischer Einschränkungen (Nicholas et al. 1999) großer Beliebtheit. Beispiele für Anwendungsgebiete der Logfile Analyse sind neben der weit verbreiteten Websiteanalyse bzw. -statistik auch das User Modelling oder die quantitative Untersuchung des Informationsverhalten auf Websites, sowie die Nutzung von Suchmaschinen (Thelwall, Vaughan & Björneborn 2003). Logfile Analysen liefern darüber hinaus wertvolle Informationen zur Zugänglichkeit (accessibility), Sichtbarkeit (visibility) und Verlinkung (interlinking) von Webinhalten eines spezifischen Webservers (Thelwall 2001, Mayr 2004a).
„With the web being such a universally popular medium, accounting forever more of people’s information seeking behaviour, and with every move a person makes on the web being routinely monitored, web logs offer a treasure trove of data. This data is breathtaking in its sheer volume, detail and potential ... Unfortunately the logs turn out to be good on volume and (certain) detail but bad at precision and attribution.” (Nicolas et al. 1999)
Dem verwandten Konzept Web-Mining liegt die Annahme zugrunde, dass die im Web verfügbaren Daten (Webdaten) ausreichend strukturiert sind, um mit Algorithmen nach Mustern zu suchen. Unter Web-Mining werden allgemein Data Mining-Techniken[2] verstanden, die zum automatischen Auffinden und Extrahieren

[2] Data Mining wird allgemein als explorative Verdichtung von Daten verstanden, um neue Erkenntnisse aus den Daten zu gewinnen.

von Informationen aus Webdokumenten und -services dienen. Siehe dazu die Definition von Kosala & Blockeel:
„The Web mining research is at the cross road of research from several research communities, such as database, information retrieval, and within AI, especially the sub-areas of machine learning and natural language processing.” (Kosala & Blockeel 2000)
Das Verfahren der automatischen Zitationsextraktion und -analyse, dass von Lawrence, Giles und Bollocker beschrieben und im Citeseer-System eingesetzt wird, (Lawrence, Giles & Bollocker 1999) kann beispielsweise als Web-Mining Verfahren bezeichnet werden.
Sowohl in der Webometrie (Thelwall 2001, Thelwall, Vaughan & Björneborn 2005) als auch in anderen informationswissenschaftlichen Disziplinen haben sich Logfiles z. B. ansatzweise zur Indikatorenbildung (Brody & Harnad 2005, Mayr 2004b) und Messung des Nutzerverhaltens (Nicholas et al. 1999, Koch, Golub & Ardö 2004) bewährt.

Im nachfolgenden Kapitel soll sehr knapp auf die Spezifika der Webserver Logfiles eingegangen werden.

2. Webserver Logfiles

Webserver Logfiles sind aufgrund ihrer Struktur, Größe und Verfügbarkeit exzellente Datenquellen für Untersuchungen des Benutzungs- und Kommunikationsverhaltens im Web. Die Webserver Logfiles stellen große Informationsmengen (i.d.R. ohne Unterbrechung) über alle Zugriffe auf einen bestimmten Webserver bereit. Die Qualität der enthaltenen Daten ist ein wichtiger Aspekt der Logfile Analyse, der in vielen Standardanalysen leider wenig Beachtung findet. Es ist aus vielen Untersuchungen bekannt, dass einige Störfaktoren die Ergebnisse der Analyse stark verfälschen können und die erste Reinigungsphase (“Cleaning”) ein außerordentlich wichtiger Schritt des Web-Mining darstellt.
Einer der bekanntesten Störfaktoren in Logdaten öffentlicher Webserver sind die Einträge der Suchroboter, die sogenannten “Spider”[3]. Diese Programme generieren im Logfile Einträge, die keine Hinweise über das Verhalten des User geben und müssen vor der Analyse unbedingt entfernt werden. Teilweise bietet schon die Analyse-Software diese Funktionalität an, indem sie die bekanntesten Such-Roboter innerhalb der Logfiles verfolgen und herausfiltern. Die Anwesenheit von unbekannten Robotern kann mit letzter Gewissheit nicht ausgeschlossen werden. Weitere Faktoren sind noch schwieriger zu beseitigen: u.a. die breite Verwendung der Rück-Navigation und die Optimierung durch die Zwischenablage der neusten Seiten

[3] Ein Programm, das automatisch Webseiten herunter lädt und der Suchmaschine zuführt. (Quelle Google: http://www.google.ch/intl/de/ads/glossary.html)

inmsogenannten “Cache Memory” auf der User Seite. Dazu gehören auch die “proxy caching servers”, die Bestandteil der Website Architektur sind, und die Geschwindigkeit der Website erhöhen sollen. Alle Abfragen der Benutzer, die zwischengelagert werden, kommen nicht im Logfile vor und können daher auch nicht untersucht werden. Laut Gutzman (Gutzman 1999) verursachen die “Spider” und Such-Roboter, die von Suchmaschinen wie Google verwendet werden, ein Drittel der Einträge der Logfiles (siehe dazu auch Nicholas & Huntington 2003). Neben diesen Störfaktoren für die Logfile Analyse existieren noch weitere, wie z.B. die immer mehr für das E-Business verwendeten dynamisch gebildeten Webseiten. Wegen der großen Informationsmenge, der zugrundliegenden Ungenauigkeit und der Schwierigkeit Störfaktoren zu identifizieren, würden Logfiles einen idealen Einsatzbereich für die unscharfe Logik darstellen (Nançoz 2004).
Diese Aspekte führen zur Behauptung, dass Logfiles per se nicht vollständig sind und ein verfälschtes Bild des Verhalten des Websitebesuchers abgeben können.
Neben den zahlreichen Einschränkungen der Webserver Logfiles existieren eine Reihe von Potenzialen, die hier aber nur ansatzweise aufgezählt werden können:

- Vergleichsweise einfach zugängliche Datenquelle zur Untersuchung einer umfangreichen und heterogenen Nutzergruppe.
- Zeitnahe und nahezu vollständige Analyse des Informationsaufnahme-verhaltens einzelner Websitebesucher oder Gruppen.
- Hinweise zur Optimierung, Evaluation und Adaption von Webinhalten und -services.
- Potenzial zur Messung von Web Impact (vgl. Brody & Harnad 2005) und zur Entwicklung weiterer webbasierter Indikatoren.

Im folgenden Abschnitt wird erläutert inwiefern Webserver Logfiles zur Makro- bzw. Mikro-Analyse eingesetzt werden können und ob eine kombinierte Analyse der beiden methodisch unterschiedlichen Verfahren denkbar und sinnvoll ist.

3. Makro- und Mikro-Mining

In diesem Kapitel werden zwei neuere und bereits publizierte Methoden vorgestellt mit denen Webserver Logfiles analysiert werden können. Die erste Untersuchungsmethode lässt sich als Makro-Mining-Methode beschreiben. Makro-Mining von Webserver Logfiles meint hier die Extraktion von Daten, die sich zu einfachen Maßen wie Anzahl der Visits oder Downloads zusammenfassen lassen. Neben diesen einfachen Standardmaßen wird in Kapitel 3.1 eine erweiterte Makro-Methode beschrieben, die auf der Unterscheidung verschiedener Einstiegszugriffe einer Website basiert.
Nicholas & Huntington haben 2003 erstmals eine Methode vorgestellt, die als Mikro-Mining-Methode bezeichnet wird.

“The aim of the study is to show how microanalysis can enhance current log analyses techniques. In particular the paper seeks to demonstrate three potential ‘micro’ techniques”

1) *the construction of a subgroup of users for which we can feel confident in regard to their geographical origin;*
2) *the analysis of a subgroup of users whose Internet Protocol (IP) addresses are more likely to reflect the use of individuals, and the same individuals;*
3) *the tracking and reporting of the use made by individuals rather than groups.”*

(Nicholas & Huntington 2003)

Eine Idee dieses Beitrags ist es die Potenziale der beiden unterschiedlichen Analyseformen zu kombinieren und die Stärken der jeweiligen Methode zu nutzen, um sie künftig zu aussagekräftigeren Kennzahlen zu verdichten bzw. spezifischere Analysen auf ihnen aufzubauen. Des weiteren soll motiviert werden, warum es für bestimmte Analysen notwendig ist, über die Standard Logfile-Auswertungen hinaus eine erweiterte Logfile Analyse durchzuführen. Der Beitrag beschränkt sich hier auf die Vorstellung der beiden Verfahren und skizziert lediglich mögliche Einsatzszenarien. Die Kennzahlenbildung bzw. empirische Prüfung der Analysen steht nicht im Mittelpunkt dieses Papers.

3.1 Makro-Mining Ansatz

Der erste Ansatz betrachtet Logfile-Einträge aus einer Makro-Perspektive. Die unten aufgezählten Makro-Analysen können weitgehend als Standardverfahren angesehen werden und werden in vielen Logfile Analyseprogrammen angeboten. Ihre Kennzahlen (z.B. Hit, View, Visit) sind zwar verbreitet, geben aber lediglich ein sehr grobes und eingeschränktes Bild des eigentlichen Onlineverhaltens. Typische Makro-Mining Auswertungen beantworten z.B. folgende Fragen:

- Wie viele Zugriffe erhalten bestimmte Bereiche bzw. Entitäten (Directory, Page) einer Website?
- Welche sind die wichtigsten Einstiegsseiten einer Website? (siehe Abb. 1)
- Über welche Suchmaschinen bzw. Suchbegriffe finden die Nutzer ein bestimmtes Webseitenangebot?

Die aus den Logdaten extrahierten Informationen liefern bei dieser Methode Hinweise auf einer relativ abstrakten Makroebene und sind i.d.R. nur hilfreich, wenn Vergleichswerte (Benchmarks) z. B. Werte vom vorherigen Quartal vorliegen.

Top Entry Pages			
	Page	% of Total	Visits
1	http://www.ib.hu-berlin.de/	5.35%	229
2	http://www.ib.hu-berlin.de/~ hab/arnd/Start.html	3.74%	160
3	http://www.ib.hu-berlin.de/~ mh/gedv/ascii.htm	2.26%	97
4	http://www.ib.hu-berlin.de/~ mh/css/css2/fonts.html	1.96%	84
5	http://www.ib.hu-berlin.de/~ mh/projekte/metaopac/	1.75%	75
6	http://www.ib.hu-berlin.de/ jaw/Html/studwohn.html	1.26%	54
7	http://www.ib.hu-berlin.de/~ hab/christine/gaudi1.html	1.16%	50
8	http://www.ib.hu-berlin.de/~ pbruhn/russgus.htm	1.14%	49
9	http://www.ib.hu-berlin.de/~ wumsta/rehm8.html	1.14%	49
10	http://www.ib.hu-berlin.de/~ wumsta/rehm4.html	1%	43

Abb. 1: Beispiel einer typischen Makro-Analyse. Die wichtigsten Einstiegsseiten (Top Entry Pages) einer Website gemessen an der absoluten Anzahl der Besuche (Visits).

Die Abbildung 1 zeigt einen Ausschnitt einer Liste mit Einstiegs-Webseiten einer Website. Die Webseiten sind nach der Häufigkeit der Websitebesuche (visits) geordnet, die auf den jeweiligen Seiten begonnen haben. Die Seite mit dem Rang 1 http://www.ib.hu-berlin.de ist für den Untersuchungszeitraum die wichtigste Einstiegsseite (229 visits, d.h. 5.35% aller visits, haben auf dieser Seite begonnen).

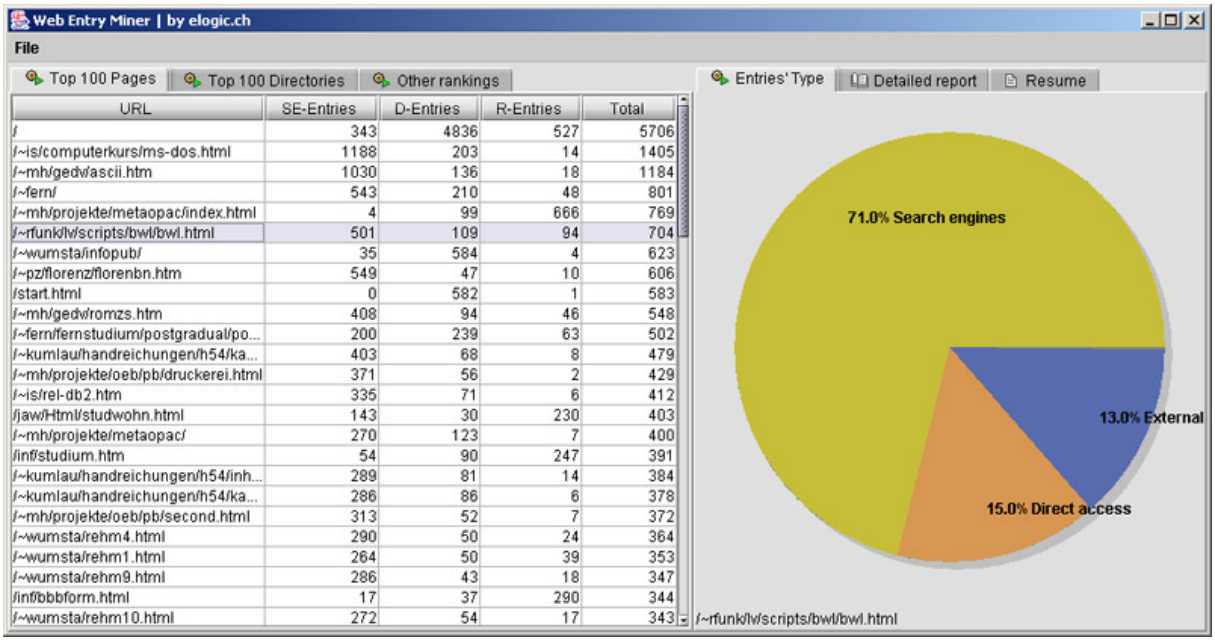

Abb. 2: Screenshot einer Makro-Analyse (Web Entry Analyse) mit dem Prototypen Web Entry Miner[4]

Abbildung 2 zeigt die Analyse und Visualisierung von Logdaten über eine neuere Makro-Mining-Methode (Mayr 2004b, siehe Abb. 2). Die Web Entry Analyse basiert auf der eindeutigen Unterscheidung von Einstiegszugriffen (Web Entries) aller Besucher einer Website, die sich auf die drei Einstiegsarten Suchmaschinen,

[4] Web Entry Miner, siehe http://www.ib.hu-berlin.de/~mayr/wem/

Backlinks[5] oder direkte Zugriffe zurückführen lassen. Diese drei sehr unterschiedlichen Zugriffs- bzw. Einstiegsarten, die sich zu drei Anteilswerten (Suchmaschinen-Einstieg, Backlink-Einstieg und direkter Einstieg) pro Einstiegsseite aggregieren lassen (siehe linker Tabellenbereich in Abb. 2), geben detaillierte Einsichten über die Bedeutung und Qualität einer konkreten Seite als Einstiegsseite innerhalb der Website aus. Die in Abbildung 2 ausgewählte Seite (siehe Spalte URL in Abb. 2) erhält z.B. 71% der Einstiegszugriffe über die Einstiegsart Suchmaschine. Die übrigen Einstiegszugriffe verteilen sich auf die beiden anderen Einstiegsarten. Aus dieser Auswertung lässt sich schließen, dass die entsprechende Seite zum Zeitpunkt der Analyse sehr gut bei einer oder mehreren Suchmaschinen positioniert ist. Eine tiefere Analyse der Zusammensetzung des Zugangsverhalten kann diese Makrosicht aber vorerst nicht liefern. An dieser Stelle wird es notwendig das nachfolgend beschriebene Mikro-Mining an die Analyse anzuschließen.

3.2 Micro-Mining Ansatz

Als ein Anwendungsbeispiel der Mikro-Analyse wird das detaillierte Online-Verhalten (User Tracking) einzelner akademischer Benutzer nachvollzogen. Die grundsätzliche Idee der Mikro-Analyse (Nicholas & Huntington 2003) besteht darin, eine Untergruppe von Besuchern einer Webseite zu analysieren und sich an Benutzer-Sessions zu orientieren. Innerhalb dieser Sessions wird die Abfolge der gesichteten Seiten extrahiert, um daraufhin Benutzer-Tendenzen zu identifizieren. Die Untergruppe kann so ausgewählt werden, dass sie repräsentativ ist oder die Merkmale einer größeren identifizierbaren Gruppe trägt. Zum Beispiel wählt Gutzman eine Untergruppe, (eine akademische Benutzergruppe), die über ihre geographische Zuordnung zuverlässig extrahiert werden kann (Gutzman 1999). Die Mikro-Analyse erlaubt gewisse Tendenzen im Verhalten der Untergruppe zu identifizieren, indem der Fokus auf einzelne Benutzer gesetzt wird. Es wird weiterhin versucht die Benutzer-Sessions, die von Störfaktoren sehr wahrscheinlich beeinflusst wurden, auszuschließen.

[5] Hyperlinks, die sich außerhalb der analysierten Website befinden und auf eine Seite der analysierten Website verweisen, werden als Backlinks bezeichnet.

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

Abb. 3 : Beispiel mehrerer Sessions eines anonymisierten Users (User Tracking Protokoll)

Das User Tracking erfolgt durch die Identifikation der einzelnen Sessions des Benutzers. Durch die Einbeziehung sämtlicher in Logfiles enthaltenen Informationen wird es möglich den Weg einer IP-Adresse durch ein Logfile zu verfolgen. Folgende Logfile-Felder werden dazu benötigt: IP-Adresse, Referrer[6] und verwendeter Browser-Typ. Die Zugriffszeit ermöglicht die Zugriffsdauer zu berechnen und mehrere Sessions logisch zu unterscheiden, indem ein gewisser Time-out definiert wird. Das Beispiel der Abbildung 3 illustriert die Sessions eines Users auf der disinfojournal.net Website. Die Benutzung des gleichen Browsers weist darauf hin, dass sich höchstwahrscheinlich ein einziger User hinter dieser IP-Adresse verbirgt. Ein anderer Hinweis, dass es sich bei den Sessions um den gleichen User handelt, ist die gemeinsame Thematik die sich aus den angesehenen Seiten und vielleicht sogar aus der kurzen Zeitspanne zwischen den Sessions ableitet. In unserem Beispiel war der User offensichtlich auf der Suche nach allgemeinen Informationen über die online Zeitschrift „Disinfojournal“. Einen Tag später, am 19. Oktober, hat der User direkt auf die „Hilights“ Seite zugegriffen oder hat die Seite vielleicht in seinen Favoriten gespeichert und hat wiederum nach den gleichen Informationen gesucht. Die genauere Uhrzeit und vor allem die Dauer der Sessions geben zudem Erläuterungen über die Verweildauer von bestimmten Benutzern auf der Website.

Durch diese feine und bewusst eingeschränkte Analyse kann die Navigationsmethodik des einzelnen Benutzers nachvollzogen werden. Damit wird auch der Inhalt der Logfiles mit allen Details ausgenutzt und der Verlust großer Informationsmengen vermieden. Im Gegensatz zur Makro-Analyse sind diese Resultate und Aussagen viel genauer und zuverlässiger, beschränken sich aber auf eine viel kleinere Benutzergruppe.

[6] URL der Webseite von der der Besucher die Seite aufgerufen hat.

Auf der Basis der komplementären Eigenschaften beider Analysen stellt das nächste Kapitel ihre sinnvolle Kombination anhand verschiedener Szenarios vor.

4. Einsatzszenario – kombinierte Analysen

Zum Abschluss dieses Beitrags wird zur Verdeutlichung der Möglichkeiten der beiden beschriebenen Analysemethoden ein potenzielles Einsatzszenario einer kombinierten Analyse skizziert. Das folgende Beispiel, bei dem Makro- und Mikro-Analyse ineinander greifen, basiert auf folgenden abgestuften Analyseschritten:

Schritt 1: Zunächst wird ein Logfileausschnitt über die in Kap. 3.1 beschriebene Makro-Methode Web Entry Mining (Mayr 2004b) analysiert. Nach der Analyse steht eine typische Makrosicht (vgl. Abb. 2) zur Verfügung, die einzelne Webseiten und ihre Zugänglichkeit über die drei Zugangsarten (Suchmaschinen, Backlinks, Direkte Zugriffe) darstellt. Diese Makrosicht lässt sich über den folgenden Schritt weiter vertiefen.

Schritt 2: Der Betrachter will sich im zweiten Schritt die genaue Zusammensetzung der Einstiege z.B. über Suchmaschinen einer beliebigen Webseite ansehen. Dazu startet er ein sogenanntes „Drill down“[7], indem er sich die genaue Zusammensetzung der Gesamtzahl der Suchmaschinen-Einstiege anzeigen lässt. Beispielsweise gehen die 71% der Suchmaschinen-Einstiege einer Seite auf mehrere Suchmaschinen und verschiedene Suchbegriffe zurück. Über die Drill down-Analyse werden z.B. alle Suchmaschinen sichtbar, die zu den Suchmaschinen-Einstiegen der Webseite geführt haben. Über einen weiteren Drill down-Schritt werden z.B. die Suchbegriffe einer konkreten Suchmaschine sichtbar. An dieser Stelle würde die Makro-Analyse enden und die Mikro-Analyse ansetzen.

Schritt 3: Das User Tracking des Mikro-Mining Ansatzes lässt sich in diesem Szenario über einen konkreten Suchmaschinen-Suchbegriff aktivieren. Die angeschlossene Mikro-Analyse versucht, alle Websitebesucher, die mit einem konkreten Suchbegriff auf eine Seite gefunden haben, inkl. der weiteren Session-Informationen anzuzeigen. Die User, die über ihre IP-Adresse oder andere eindeutige Merkmale authentifiziert werden konnten, werden mit allen ihren weiteren Transaktionen (vgl. User-Tracking-Protokoll in Abb. 3) angezeigt. Damit endet das kombinierte Analyse-Szenario.

[7] Drill down (engl.: tiefer bohren oder graben) meint im Zusammenhang mit Logfile Analysen, dass dem Nutzer über ein Funktionsmenü ein weiterer tieferer Analyseschritt zur Verfügung steht.

Die Kombination der beiden Analysemethoden erlaubt, die mit der Makro-Analyse abgeleiteten Tendenzen mit einer Testgruppe zu prüfen und damit noch genauer auf das Verhalten des Users zu fokussieren.
Auf der anderen Seite kann die Mikro-Analyse verwendet werden, um „Micro-Trends“ zu entdecken. In einem zweiten Schritt würde eine erweiterte Makro-Analyse das Ausmaß des Trends abschätzen. Vorteil dieses Szenario ist die Fähigkeit der Mikro-Analyse kleine aber aussagekräftige Benutzersegmente zu identifizieren, die mit einer Makro-Methode unsichtbar bleiben.

5. Ausblick

„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)

Ziel des vorliegenden Papers war es zwei methodisch sehr unterschiedliche Webserver Analyseansätze zu beschreiben und die Kombination abgestufter Analysemethoden für künftige Untersuchungen vorzuschlagen. Durch die Kombination beider Analysen (Makro- und Mikro-Mining) wird beispielsweise ein völlig unterschätzter Aspekt des „User Information Retrieval“ aufgedeckt: zu den „klassischen“ und statischen Informationen, wie der geographischen Herkunft der Benutzer und den Einstiegsseiten, kann das Benutzer-Verhalten als neue dynamische Komponente hinzugefügt werden.
Neue Ansätze zur Analyse von Webserver Logfiles werden immer notwendiger um genauere und stabilere Maße des Websitegebrauchs zu entwickeln. Dies gilt für wissenschaftliche Logfile-Untersuchungen und kommerziell orientierte Verfahren gleichermaßen. Gerade die kombinierten Analysen versprechen hier neue Ergebnisse und tiefere Einsichten in das Benutzungsverhalten zu liefern.
Neben der Reduzierung der Fehleranfälligkeit von Logfile Analysen, sind der gesteigerte Komfort der Websitebesucher, die Reduzierung der Suchzeit und damit letztlich die Besucher- bzw. Kundenzufriedenheit die wichtigsten Ziele künftiger Entwicklungen.

Literatur

1. Björneborn, Lennart; Ingwersen, Peter (2001): Perspectices of webometrics. In: Scientometrics, Vol. 50, pp. 65-82.
2. Brody, Tim; Harnad, Stevan (2005): Earlier Web Usage Statistics as Predictors of Later Citation Impact. Technical report. URL: http://eprints.ecs.soton.ac.uk/10647/ (access date 14 August 2005)
3. Gutzman, A. (1999): Analysing Traffic on Your E-commerce Site. URL: http://ecommerce.internet.com/solutions/tech_advisor/article/0,,9561_186011,00.html (access date 14 August 2005)
4. Koch, Traugott; Golub, Koraljka; Ardö, Anders (2004): Log Analysis of User Behaviour in the Renardus Web Service. URL: www.it.lth.se/knowlib/publ/LIDA2004_final.doc (access date 14 August 2005)
5. Kosala, Raymond; Bockeel, Hendrik (2000): Web mining research: A survey. In: SIGKDD Explorations, Vol. 2, pp. 1-15.
6. Lawrence, Steve; Giles, C. Lee; Bollacker, Kurt (1999): Digital Libraries and Autonomous Citation Indexing. In: IEEE Computer, Vol. 32 (6), pp. 67-71. URL: http://citeseer.ist.psu.edu/aci-computer/aci-computer99.html (access date 14 August 2005)
7. Mayr, Philipp (2004a): Entwicklung und Test einer logfilebasierten Metrik zur Analyse von Website Entries am Beispiel einer akademischen Universitäts-Website. (Berliner Handreichungen zur Bibliothekswissenschaft und Bibliothekarsausbildung ; 129). URL: http://www.ib.hu-berlin.de/~kumlau/handreichungen/h129/ (access date 14 August 2005)
8. Mayr, Philipp (2004b): Website entries from a web log file perspective - a new log file measure. Proceedings of the AoIR-ASIST 2004 Workshop on Web Science Research Methods. URL: http://cybermetrics.wlv.ac.uk/AoIRASIST/mayr.html (access date 14 August 2005)
9. Nançoz, Christian (2004): mEdit – membership function editor for fCQL-based architecture. Master Thesis, URL : http://diuf.unifr.ch/is/studentprojects/pdf/M-2004_Christian_Nancoz.pdf (access date 14 August 2005)
10. Nicholas, David, et al. (1999): Cracking the code: web log analysis. In: Online & CD-ROM Review, Vol. 23, pp. 263-269.
11. Nicholas, David; Huntington Paul. (2003): Micro-Mining and Segmented Log File Analysis: A Method for Enriching the Data Yield from Internet Log Files. In: Journal of Information Science, Vol. 29 (5), pp. 391-404.
12. Thelwall, Mike (2001): Web log file analysis: Backlinks and Queries. In: Aslib Proceedings, Vol. 53, pp. 217-223.
13. Thelwall, Mike; Vaughan, Liwen; Björneborn, Lennart (2003): Webometrics. In: ARIST, Vol. 39, preprint. URL: http://www.db.dk/lb/2003preprint_ARIST.doc