

# Webstatistik

---

Version 1.0  
Stand Juli 2007

Autor:

Detlef Müller-Solger

durato Ltd.  
Pastor-Klein-Str. 17E  
56073 Koblenz

Tel: (02 61) 2 96 34 39 - 0  
Fax: (02 61) 2 96 34 39 - 199

www.durato.de • info@durato.de

## Inhaltsverzeichnis

Inhaltsverzeichnis .....	2
1. Ausgangspunkt .....	3
2. Begriffsabgrenzung .....	3
3. Seitenstatistik .....	4
3.1. Datenbasis .....	4
3.1.1. Abweichungen der Datenbasis von der tatsächlichen Nutzung .....	4
3.1.2. Lösungsansätze .....	5
3.2. Serverstatistik .....	6
3.3. Kontaktmessung .....	7
3.4. Web-Mining .....	8
3.5. Zusammenhänge der einzelnen Techniken .....	8

## 1. Ausgangspunkt

Im Bereich der Web-Statistik hat sich im Laufe der letzten Jahre ein multidisziplinäres Arbeits- und Forschungsumfeld entwickelt. Wie rasant die Entwicklung fortschreitet, ist erkennbar an der Auflösung des Gesamtbegriffs Web-Statistik und am Aufkommen neuer Begrifflichkeiten. Die Begriffe "Web-Experiment", "Online Research", "Online-Marktforschung" oder "WWW-Kontaktmessung" seien hier als Beispiele genannt. Unabhängig von dieser ausufernden Diskussion soll hier nach einer einführenden Abgrenzung noch einmal der Fokus auf den Ausgangspunkt gerichtet werden. Der Statistik der Internetbenutzer einer Seite.

## 2. Begriffsabgrenzung

Aus heutiger Sicht können unter dem Dach Web-Statistik vier thematische Schwerpunkte unterschieden werden, wobei zwischen den einzelnen Bereichen oftmals Überlappungen und begriffliche Unschärfen herrschen:

1. Seitenstatistik  
Gegenstand: Die Erfassung von Seitenausgaben, Benutzern und die Analyse des Benutzerverhaltens.
2. Netzstatistik  
Gegenstand: Analyse der Web-Benutzer insgesamt, der im Netz verfügbaren Inhalte und der eingesetzten Techniken.
3. Online Marktforschung  
Gegenstand: Klassische Marktforschung mit Hilfe von Online-Medien.
4. Statistik als Thema im Internet  
Gegenstand: Diskussion und Forschung zum Thema Statistik.

Die folgende Grafik ordnet die genannten Schwerpunkte grafisch:

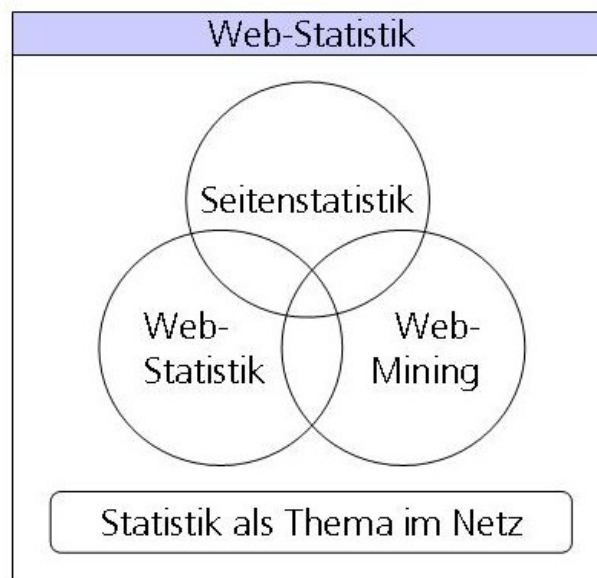


Abbildung 1: Thematische Schwerpunkte

Wie dargestellt ist der Grad der Überlappung bzw. die Unschärfe zwischen den einzelnen Themenfeldern unterschiedlich. Während Punkt 4 "Statistik als Thema im Netz" letztlich als autarkes Thema definiert werden kann, bestehen zwischen den anderen Feldern Abhängigkeiten und Interdependenzen. Dabei dient die Seitenstatistik den Themenkreisen Netzstatistik und Online-Marktforschung sowohl als Datenbasis, als auch als Bezugspunkt der eigenen Untersuchungen und Analysen. Dem im folgenden dargestellten Themenkreis kommt somit eine übergeordnete Bedeutung zu.

### 3. Seitenstatistik

Der Themenkreis Seitenstatistik setzt sich zusammen aus den Teiluntersuchungen Serverstatistik, Kontaktmessung und Web-Mining. Sie stehen in einem engen Bezug zueinander und spiegeln das Ergebnis unterschiedlicher Fragestellungen. Ausgangspunkt und Basis aller drei Größen ist die Auswertung von Daten die mit so genannten nicht-reaktiven Messverfahren gewonnen werden. Bei diesen Verfahren bemerkt der Benutzer nicht, dass seine „Aktionen“ und damit sein "Verhalten" automatisch protokolliert werden.

#### 3.1. Datenbasis

Ursprung und Kern der mit nicht reaktiven Messverfahren gewonnenen Datenbasis ist die Protokollierung der Daten, die durch das Hyper-Text-Transfer-Protocol übertragen werden. Bei diesen so genannten HTTP-Logs handelt es sich in der Regel um Textdateien im ASCII-Format, die unabhängig von der Computerplattform, des Betriebssystems und dem verwendeten Webservern aufgezeichnet werden. Da im HTTP-Log grundsätzlich alle Anfragen und Datenübertragungen einzeln aufgezeichnet werden, lassen sich aus ihnen Zugriffsaktivitäten ableiten.

##### 3.1.1. Abweichungen der Datenbasis von der tatsächlichen Nutzung

Aufgrund einiger grundlegender Internet-Techniken weist das HTTP-Log im Vergleich zu dem tatsächlichen Seitenaufruf durch unterschiedliche Benutzer die folgenden Abweichungen und Fehler auf.

###### *Proxy-Server*

Mit Blick auf die letztlich beschränkte Bandbreite des Internets werden Proxy-Server eingesetzt, die zur Minimierung des Datentransfervolumens dienen, indem sie häufig benutzte Informationen lokal zwischenspeichern. Ausgehend von seiner Netzanbindung ruft ein Benutzer deshalb das gewünschte Angebot nicht immer vom anbietenden Web-Server ab. In der Konsequenz erhält der Server trotz eines tatsächlichen Aufrufs seiner Seite keine Seitenanforderung. Wie unten schematisch dargestellt erfolgt daher auch kein neuer Eintrag im HTTP-Log. Demgegenüber taucht ein weiterer regulärer Eintrag auf, wenn im Rahmen einer Seitennutzung via Proxyserver eine einzelne Seite des gesamten Angebotes aufgerufen wird, die noch nicht lokal gespeichert ist. In Relation zur tatsächlichen Nutzung des gesamten Angebotes wird in diesem Fall die protokollierte Nutzung nicht nur unvollständig erfasst, sondern auch der falsche Benutzer registriert, denn der Abruf der einen fehlenden Seite erfolgt hier durch den Proxyserver. Insgesamt erfasst ein HTTP-Log somit weniger als die tatsächliche Nutzung, enthält Fehler bezüglich der Benutzeranzahl und generiert teilweise unvollständige Angaben.

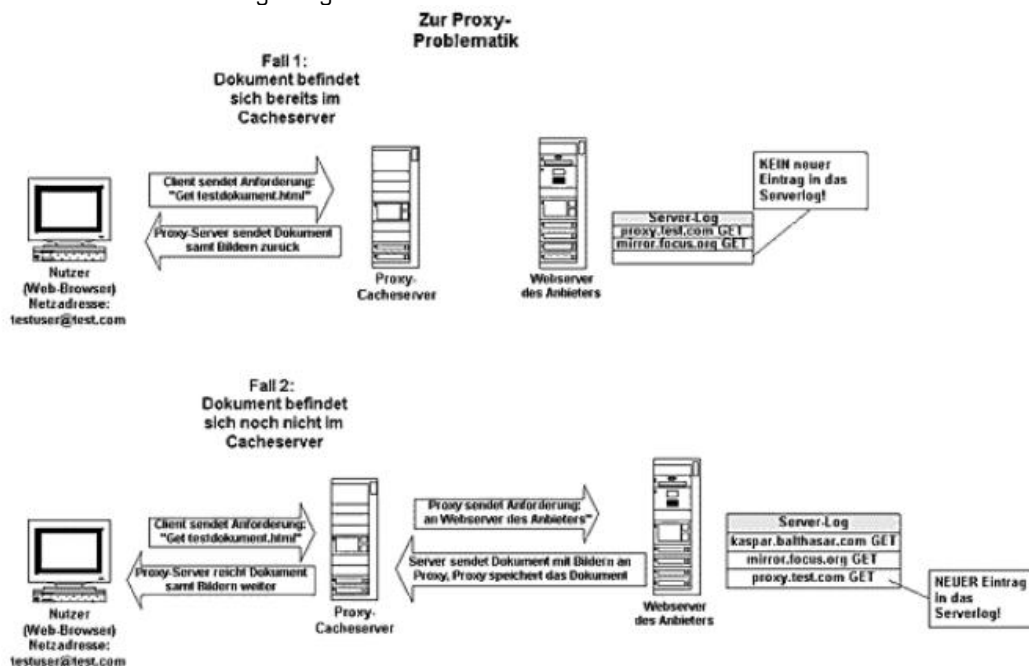


Abbildung 2: Proxy-Server-Problematik

### ***Browser Cache***

Nach dem gleichen Prinzip wie der Proxy-Server arbeiten die lokalen Caches der Browser. Abhängig von den Einstellungen des Internetbenutzers speichern auch sie Seiten auf der lokalen Festplatte und vermeiden deshalb die erneute Server-Abfrage. Der daraus resultierende Fehler im HTTP-Log als Datenbasis für die Benutzung einer Internetseite entspricht dem des Proxy Servers mit Ausnahme der falschen Benutzererfassung.

### ***Dynamische IP-Adressen***

Wegen des begrenzten Vorrats an IP-Adressen vergeben Online-Dienste wie z.B. AOL oder T-online ihre IP-Adressen dynamisch. So kann mit wenigen Adressen eine große Zahl von Web-Nutzern bedient werden können. Im Ergebnis verbergen sich hinter einer dynamisch vergebenen IP Adresse, die im HTTP-Log als Absender der Anfrage erfasst wird, eine unbekannte Anzahl von Benutzern. Und anders herum verbirgt sich hinter mehreren IP-Adressen teilweise nur ein Benutzer, da dieser im Falle wiederholter Seitenbenutzungen mit mehreren IP-Adressen auftritt.

### ***Firewall-Rechner***

Fast alle Unternehmen schützen ihren Internetauftritt aus Sicherheitsgründen mit Firewall-Rechnern. Firewalls setzen interne IP-Adressen auf eine einzige externe IP-Adresse um. Sie wird im Log-File als einzelner Benutzer aufgezeichnet, auch wenn unterschiedliche Personen aus dem Unternehmen auf das Angebot zugegriffen haben. Kommt der Einsatz von Proxy-Servern hinzu, werden die Log-Files hinsichtlich der Besuchszahlen stark verzerrt und korrekte Rückschlüsse auf Benutzer und die Seitennutzung werden unmöglich.

## **3.1.2. Lösungsansätze**

Für die genannten Probleme wurden in den vergangenen Jahren die unterschiedlichsten Lösungsansätze entwickelt. Alle zielen darauf die der Netztechnik des WWW geschuldeten Lücken zu minimieren bzw. zu schließen. Dabei geht es einerseits um die Erfassung der tatsächlichen Seitenaufrufe und andererseits um die Identifizierung eines einzelnen Benutzers. Ausgehend vom derzeitigen Stand der Technik sind hier die folgenden Vorgehen ohne Anspruch auf Vollständigkeit hervorzuheben:

### ***1. Cookies***

Cookies sind spezielle Textdateien, die ein Webserver zu einem Browser sendet, um dem zustandslosen HTTP-Protokoll die Möglichkeit zu geben, Information zwischen Aufrufen zu speichern. Man kann zwischen persistenten Cookies und Session-Cookies unterscheiden. Erstere werden dauerhaft gespeichert, während letztere nur für die Länge einer Sitzung gespeichert werden. Mit Blick auf diese technische Charakteristik bietet es sich zunächst an, einen Cookie durch das abspeichern und wiederholte Auslesen einer Session ID als „Ankerpunkt“ für die Identifizierung eines zusammenhängenden Benutzungsvorgangs zu nutzen. Wird ein persistenter Cookie eingesetzt, kann durch eine Abfrage seiner Existenz zugleich der wiederholte Besuch registriert werden.

Von der Konsistenz der erfassten Daten bis zur Identifizierung eines wiederkehrenden Benutzers erhöhen Cookies die Qualität und Aussagekraft der Datenbasis. Demgegenüber ist deutlich darauf hinzuweisen, dass auch dieses Verfahren lückenhaft ist, denn abhängig von den Benutzereinstellungen akzeptiert nicht jeder Rechner Cookies, und außerdem können sie problemlos gelöscht werden. Weiterhin ist zu beachten, dass Webseitenbetreiber den Einsatz von Cookies aufgrund der potentiellen Gefahren und Risiken oft verbieten. In diesem Fall kann eine Verbesserung der Datenqualität nur über die nachfolgend genannten Techniken erreicht werden.

### ***2. Zählpixel***

Grundsätzlich kann ein Web-Server jeder Datei und jedem Element die Information mitgeben, dass sie nicht auf einem Proxy-Server zwischengespeichert werden soll. Eine grundsätzliche Aktivierung für alle Dateien ist nicht ratsam, da die Netzbelastung dadurch exponentiell steigen würde. Daraus ergibt sich die Idee, jeder

Seite mit Inhalt lediglich ein kleines und schnell übertragbares Element beizufügen, welches die Info „Bitte nicht auf dem Webserver speichern“ enthält. Dieses wird dann immer vom anbietenden Web-Server „abgeholt“. Im Regelfall wird das kleine Element als 1 X 1 großes transparentes GIF realisiert. Aufgrund seines erzwungenen Abrufs vom Webserver löst das Pixel das Problem der Proxyserver, und die entsprechenden Anfrageeinträge im HTTP-log können so als Basis einer validen Statistik dienen. Derart wird das Pixel zum Zählpixel.

Durch die Kombination des Bildaufrufes mit verschiedenen Skripten kann die Methode noch verfeinert werden. So können Serverseitige Skripte dem Abruf des Pixels zum Beispiel noch die Information mitgeben, auf welcher Seite (Startseite, Rubrik 1 etc.) sich das Pixel befindet. Dies ist zum Beispiel immer dann wichtig, wenn der zu analysierende Webauftritt keine „sprechenden URLs“ aufweist. Außerdem ist es möglich, die Abfragen in ein eigenes Log zu schreiben. Das erleichtert später die Analyse. Auf der Seite des Clients kann der Aufruf per Java Skript zusätzlich um auslesbare Parameter wie Bildschirmauflösung, Browserversion etc. ergänzt werden. Bei der Übergabe dieser Parameter an den Server ist darauf zu achten, dass die notwendige Zeichenfolge nicht länger als 255 Zeichen ist.

Die Zählpixel-Methode wird heute auch im Rahmen der offiziellen Messung von Seitenaufrufen genutzt (<http://www.iwonline.de/messverfahren/szm-tag.php#grundsatzlich>). Jenseits dieses Angebotes bieten einige Anbieter fertige Technologien als Mietlösung an. Neben den teilweise erheblichen Kosten liegt ihre Grenze in der mangelhaften Flexibilität. So lassen Sie sich nur in den seltensten Fällen an „Besonderheiten“ der jeweiligen Webseite anpassen.

### ***3. Session Reconstruction***

Die Session Reconstruction ist keine eigenständige Technologie. Sie nutzt die um Systemparameter wie Browserversion, Bildschirmauflösung erweiterte Datenbasis des Zählpixels aus und versucht über die zusätzlich gewonnenen Daten, einen Benutzer nachhaltig zu identifizieren. Dieses Verfahren ist immer dann notwendig, wenn die Webseite z.B. per CMS System selber keine auslesbare Session generiert, oder der Einsatz von Cookies verboten wird. Letztlich betreibt dieses Verfahren, das sich gegen das Problem der Firewalls wendet, jedoch nur Ergebniskosmetik, da in großen Firmen meist eine homogene IT-Landschaft existiert und somit der ursprüngliche Fehler nur marginal korrigiert wird.

### ***4. CMS Logs***

Werden Webseiten mit Hilfe eines Content-Management-Systems erzeugt, so können im Rahmen der dynamischen Erzeugung der einzelnen Webseiten weitere interessante Daten auf „direktem“ Weg erzeugt werden. Im Zentrum steht dabei zunächst die Ausgabe einer Session-ID zur Identifikation eines zusammenhängenden Benutzervorgangs. Zusätzlich können „Bereichsanalysen“ und andere Statistik-Features in das CMS System integriert werden. Die Unterstützung der Auswertung durch dieses Verfahren hat nicht nur Vorteile bezüglich der Performanz, sondern kann die Datenbasis wieder um weitere interessante Werte erweitern.

## **3.2. Serverstatistik**

Ausgehend von der nach den zuvor beschriebenen Verfahren generierten Datenbasis umfasst die Serverstatistik nach allgemeinem Verständnis die quantitative Analyse. Dabei steht zunächst die Serverauslastung im Vordergrund. Hier geht es um die übertragene Datenmengen und die Anzahl der Zugriffe auf dem Server. Im weiteren geht es dann um die Seitenutzung und dabei um die folgende Frage:

- Welche Seiten wurden wie oft abgerufen?

Bei dieser Auswertung wird letztlich immer nach den absoluten Mengen gefragt. Der Bezug auf den Benutzer, der Vergleich der Daten, die Analyse der Benutzungsvorgänge und die mögliche Korrelation mit Daten anderer Statistiken bleiben aus.

### 3.3. Kontaktmessung

Der Begriff der Kontaktmessung stammt aus dem Bereich Printmedien und Fernsehen. Während im Printbereich die Auflage als maßgebliche Grundlage für die Berechnung der Anzeigen zugrunde liegt, wird für den Fernsehbereich die so genannte "Quote", d.h. die Mediennutzung, über das GfK-Panel ermittelt. Auch für Online-Medien, bzw. für das WWW wird daher eine Kontaktmessung benötigt, die als verlässliche Basis für Anzeigenpreise dienen kann und die Werbeträgerleistung in Online-Systemen dezidiert ermittelt.

In der Anfangszeit des Internets wurden oftmals die Anzahl der "hits" als Maß für die Popularität einer Web-Site angegeben. Da jedoch sowohl die HTML-Seite selber, wie auch die in sie integrierten Bilder im HTTP-Log als eigene Einträge verzeichnet werden (Hits) folgte hieraus, dass eine HTML Seite mehrmals als Kontakt gezählt wurde. Dies widersprach dem in den Printmedien und im Rundfunk geprägtem Verständnis.

In Antwort darauf entwickelte die Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e.V. (IWW) zwei Indikatoren zur Kontaktmessung. Beide stellen einen Bezug zum Benutzer her und lassen Layout und Grafik einer HTML-Site unberücksichtigt.

#### ***Page Impressions (früher auch Page Views)***

*Page Impressions bezeichnen die Anzahl der Sichtkontakte beliebiger Benutzer mit einer potentiell werbeführenden HTML-Seite. Sie liefern ein Maß für die Nutzung einzelner Seiten eines Angebots.*

Da bei HTML Seiten die mit einem Frameset arbeiten mit einem "Click" mehrere Seitenabrufe veranlasst werden können, musste die Definition um folgenden Passus erweitert werden:

*Enthält ein Angebot Bildschirmseiten, die sich aus mehreren Frames zusammensetzen (Frameset), so gilt jeweils nur der Inhalt eines Frames als Content. Der Erstabruf eines Framesets zählt daher nur als ein Page Impression, ebenso wie jede weitere nutzerinduzierte Veränderung des entsprechenden Content-Frames. Demnach wird pro Nutzeraktion nur eine Page-Impression gezählt. Zur definitionsgerechten Erfassung der Page Impressions verpflichtet sich der Anbieter, gekennzeichneten Content jeweils nur in einen Frame pro Frameset zu laden.*

Mit der Definition des Begriffs Page Impression ist die Basis der Kontaktmessung klar festgelegt. Darüber hinaus ist nun die Anzahl der tatsächlichen Benutzer einzugrenzen. Welche Schwierigkeiten mit deren Erfassung verbunden sind, wurde bereits bei der Problematik der dynamischen IP-Adressen und der Firewall-Technologie angedeutet. Entsprechend vorsichtig lautet die Definition des IWW:

#### ***Visits (Besuche, Nutzungsvorgänge)***

*Ein Visit bezeichnet einen zusammenhängenden Nutzungsvorgang (Besuch) eines WWW-Angebots. Er definiert den Werbeträgerkontakt. Als Nutzungsvorgang zählt ein technisch erfolgreicher Seitenzugriff eines Internet-Browsers auf das aktuelle Angebot, wenn er von außen erfolgt.*

Als Antwort auf das Problem, dass "echte Personen" sich kaum den einzelnen Nutzungsvorgängen zuordnen lassen, hat die IWW auf die Erwähnung des Begriffs „Nutzer“ in der Definition konsequent verzichtet. Insofern darf dieser Bezug auch in der Auswertung nicht hergestellt werden, denn tatsächlich könnte dieser nur über eine Passwort-Anmeldungen und Registrierungen gezählt werden. Als Nutzungsbarriere würde dies zu einem starken Rückgang der Besucherzahlen führen. Angeführt von der Internet-Site "Hotwired" gibt es dafür eine Vielzahl von Belegen im Internet.

Jenseits der Page Impressions und der Visits kann auch die Reihenfolge der Seitenaufrufe, die innerhalb eines WWW-Angebots während eines Visits abgerufen wird, erfasst werden. Diesen Ablauf des Benutzungsvorgangs bezeichnet man als Clickstream. Er ist eine wertvolle Basis für die weitere Analyse der Seite. An dieser Stelle setzt das Web-Mining an.

### 3.4. Web-Mining

Der Begriff Web-Mining wurde abgeleitet vom allgemeinen Begriff des Data-Mining. Data-Mining-Werkzeuge sind darauf spezialisiert, noch unbekannte Zusammenhänge innerhalb von Unternehmensdaten zu finden. Im Gegensatz zu den klassischen Abfragewerkzeugen muss hier der Anwender nicht vorher wissen, wonach er sucht. Vielmehr wird der Anwender zu den interessanten Informationen geführt. Web-Mining bezeichnet den auf die speziellen Erfordernisse des Internets angepassten Forschungszweig. Ziel des Web-Minings ist die Analyse der Datenbasis auf Regelmäßigkeiten und Muster im Benutzerverhalten. Im Zentrum steht hier der Clickstream als Analysegröße.

Die Methoden, die im Data-Mining und auch im Web-Mining Anwendung finden, sind klassische statistische Verfahren, die auch in der Marktforschung eingesetzt werden. Als wichtigste Beispiele sind hier z.B. Clusteranalysen, CHAID, Assoziationsmaße oder neuronale Netze zu nennen. Ziel ist der gläserne Benutzer. Triebfeder ist der Wunsch die Kundenbeziehung und die Ergonomie der Seite zu optimieren. Die sich anschließende Gefahr ist der Datenmissbrauch. Datenschutzgesetze verbieten deshalb Data-Mining mit personenbezogenen Daten. Jenseits dessen werden keine Grenzen gesetzt.

Da im Rahmen des Web-Mining tatsächlich Ergebnisse erzielt werden können, die erheblich den Mehrwert einzelner Seiten steigern, sind diese Instrumente ohne falsch verstandene Pietät bzgl. der Sensibilität persönlicher Daten nachhaltig einzusetzen. Käufer des Online-Buchhandels AMAZON merken das regelmäßig, wenn ihnen beim Kauf eines Buches weitere Bücher angeboten werden, die entsprechend dem Kaufverhalten anderer Benutzer als dazu passend eingestuft werden.

### 3.5. Zusammenhänge der einzelnen Techniken

Allen drei Auswertungen gemeinsam ist die Auswertung der aufgrund nicht-reaktiver Messverfahren mit den oben genannten Techniken generierte Datenbasis. In einem ersten Schritt lassen sich hier unter dem Titel Serverstatistik quantitative Statistiken ableiten, die über Transfervolumen und Seitenausgaben berichten. Darauf aufbauend geht es um eine Kontaktanalyse, welche hinterfragt, wann wer wie oft eine bestimmte Seite gesehen hat. Daran schließt sich das Web-Mining an, das mit statistischen Methoden ein komplexes Bild von der Nutzung der Seite entwickelt und so Hinweise zum weiteren Auf- und Ausbau der Seite gibt.

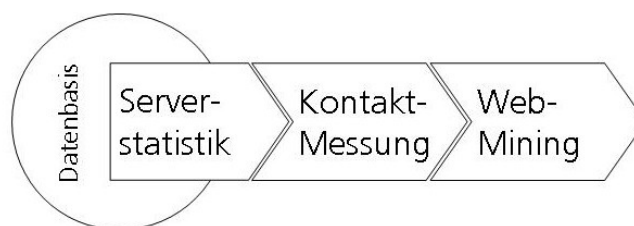


Abbildung: Die Auswertungen im Rahmen der Seitenstatistik

Mit dem aufkommen der Websites als Informationsplattform und Kommunikationsforum setzt auch ein Euphorie bezüglich der Transparenz des Konsumentenverhaltens ein. Das Wort Web-Statistik und der gläserne Kunde war in aller Munde. Heute ist festzuhalten, dass das Benutzerverhalten im Gegensatz zu den Print-Medien und dem Rundfunk sicher besser erfassbar ist, den Online-Medien hingegen aber auch klare Grenzen gesetzt sind. Mit der persönlichen Anmeldung als Seitenzugang lässt sich jedoch der gläserne Kunde mit allen Vor- und Nachteilen verwirklichen.