Böhm, Peter

Ermittlung von Nutzungsweisen auf dem Deutschen Bildungsserver mittels Webanalyse-Verfahren

2010 - (Zugl.: Darmstadt, Hochschule, Masterarbeit, 2010)

Quellenangabe/ Reference:
Böhm, Peter: Ermittlung von Nutzungsweisen auf dem Deutschen Bildungsserver mittels Webanalyse-Verfahren. 2010 - (Zugl.: Darmstadt, Hochschule, Masterarbeit, 2010) - URN: urn:nbn:de:0111-dipf-38006 - DOI: 10.25657/02:3800

https://nbn-resolving.org/urn:nbn:de:0111-dipf-38006
https://doi.org/10.25657/02:3800

Kontakt / Contact:

DIPF | Leibniz-Institut für
Bildungsforschung und Bildungsinformation
Frankfurter Forschungsbibliothek
publikationen@dipf.de
www.dipfdocs.de

Hochschule Darmstadt
Fachbereich Media
Studiengang Information Science and Engineering / Informationswissenschaft

Masterarbeit

Ermittlung von Nutzungsweisen auf dem Deutschen Bildungsserver mittels Webanalyse-Verfahren

Bearbeiter: Peter Böhm

Referent: Prof. Dr. Marc Rittberger

Korreferent: Prof. Dr. Reginald Ferber

Betreuerin: Carola Carstens

Eingereicht am 14. Januar 2010

Lizenz

- Rechte anderer Personen, entweder am Lizenzgegenstand selber oder bezüglich seiner Verwendung, zum Beispiel Persönlichkeitsrechte abgebildeter Personen.

Hinweis – Im Falle einer Verbreitung müssen Sie anderen alle Lizenzbedingungen mitteilen, die für dieses Werk gelten. Am einfachsten ist es, an entsprechender Stelle einen Link auf diese Seite einzubinden.

`http://creativecommons.org/licenses/by-sa/3.0/de/`

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Akronymverzeichnis

AG	Amtsgericht
BLK	Bund-Länder-Kommission
CMS	Content Management System
CSS	Cascading Stylesheet
CSV	Comma-Separated Values
DBIS	Datenbank-Infosystem
DBS	Deutscher Bildungsserver
DIPF	Deutsches Institut für Internationale Pädagogische Forschung
DNS	Domain Name System
ELIXIER	Elaborated Lists in XML for Internet Educational Ressources
ELP	Event-Locality-Pair
FPC	First-Party Cookie
FIS	Fachinformationssystem
GIF	Graphics Interchange Format
GPL	GNU General Public License
HDSG	Hessisches Datenschutzgesetz
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
ID	Identifikator
IETF	Internet Engineering Task Force
IP	Internet Protocol

ISP	Internet Service Provider
IZ	Informationszentrum
IZB	Informationszentrum Bildung
JS	JavaScript
KIA	Key Insights Analysis
KDD	Knowledge Discovery in Databases
LG	Landgericht
NAT	Network Address Translation
OLAP	Online Analytical Processing
PASW	Predictive Analytics Software
PDF	Portable Document Format
PHP	PHP: Hypertext Preprocessor
PRM	Pattern Restore Method
RDBMS	Relationales Datenbankmanagementsystem
RDF	Resource Description Framework
RFC	Request for Comments
RSS	Really Simple Syndication
SaaS	Software as a Service
TMG	Telemediengesetz
TPC	Third-Party Cookie
URL	Uniform Resource Locator
UTM	Urchin Traffic Monitor
WAA	Web Analytics Association
WWW	World Wide Web
XML	Extensible Markup Language

1. Einleitung

Der Deutsche Bildungsserver (DBS) ist ein breit gefächertes Themenportal im Bildungsbereich. In seiner Grundfunktion weist er bildungsrelevante Internetangebote im deutschsprachigen Internet nach, es handelt sich somit um einen so genannten Verweisserver. Der inhaltliche Zugang erfolgt über die nach Themen und Zielgruppen ausgerichtete Navigationshierarchie oder über die Suchfunktion.

Eine Auswertung der Nutzungsdaten der Monate Juli und August 2009 zeigt, dass 68% der Besuchseinstiege über eine Suchmaschine erfolgten. Nur 12% der Besuche wurden über eine sonstige Verweisseite initiiert und 20% direkt gestartet, d.h. durch Eingabe der URL oder den Aufruf eines Lesezeichens.

Für die DBS-Redaktion stellt sich die Frage, ob sich die Suchmaschineneinsteiger hinsichtlich ihrer Nutzungsweise des DBS von den anderen Besucherarten unterscheiden und ob eventuell Änderungen am Webangebot erforderlich sind, um der besonderen Nutzungsweise gerecht zu werden.

In einem zweiten Auswertungsschritt wurden daher drei gängige Sitzungsmerkmale ermittelt, um erste Hinweise auf etwaige Unterschiede zu erhalten. So variiert die durchschnittliche Anzahl der Aktionen je Sitzung leicht (3,5 bei Suchmaschinen; 3,9 bei sonstigen Verweisen und 3,7 bei Direkteinstiegen). Die Absprungrate, also jener Teil der Sitzungen, die nach einem Seitenaufruf wieder beendet werden, ist bei Direkteinstiegen am höchsten (52%) und liegt bei Suchmaschineneinstiegen und sonstigen Verweisen deutlich darunter (40% bzw. 39%). Der Anteil der Besuche, die mit dem Klick auf einen externen Link enden, unterscheidet sich kaum zwischen Suchmaschinen- und Direkteinstiegen (30% zu 30,9%), liegt aber bei sonstigen Einstiegen darunter (25,7%).

Diese einfache Auswertung deutet auf Unterschiede im Nutzungsverhalten der DBS-Besucher in Abhängigkeit von der Art ihres Einstiegs in den DBS hin. Sie ist jedoch nicht umfassend genug, um zweifelsfreie Rückschlüsse auf die Art und Ausprägung jener Unterschiede zu ziehen.

In der vorliegenden Arbeit soll daher in einer ausführlicheren Datenanalyse eines längeren Nutzungszeitraums überprüft werden, ob diese Unterschiede in statistisch

signifikanter Form bestehen und welcher Art sie sind. Auf Grundlage dieser Analysen sollen, wenn möglich, unmittelbar Verbesserungsvorschläge für den Deutschen Bildungsserver abgeleitet werden.

Angewendet werden Methoden und Metriken der Webanalyse. Diese Metriken werden nach der Einstiegsart segmentiert – ein Auswertungsschritt, den gängige Webanalyse-Programme nicht anbieten. Weiterhin wird mit Web Usage Mining ein elaborierteres Verfahren vorgestellt und in einzelnen Aspekten umgesetzt.

Das folgende zweite Kapitel stellt den Deutschen Bildungsserver als Untersuchungsgegenstand der Arbeit mit seiner Struktur und seinen Inhalten dar. Eine Beschreibung von Information-Retrieval-Systemen und Hypertext sowie die Einordnung des DBS in diese Begriffsfelder wird in Kapitel drei vorgenommen. Kapitel vier umfasst eine Abgrenzung der Begriffe Webanalyse und Web Usage Mining sowie eine Vorstellung von Datenerhebungsmethoden einschließlich ihrer Vor- und Nachteile. Webanalyse und Web Usage Mining werden in den beiden darauf folgenden Kapiteln fünf und sechs jeweils ausführlicher vorgestellt. Dabei wird in Kapitel fünf besonders auf die verschiedenen Webanalyse-Metriken und die beim DBS eingesetzten Auswertungsprogramme eingegangen. Kapitel sechs beinhaltet neben einer Beschreibung der Vorgehensweise beim Web Usage Mining eine Vorstellung von fünf Ansätzen des Web Usage Mining und eine Besprechung ihrer Anwendbarkeit beim DBS.

Die Konzeption der praktischen Umsetzung der zuvor dargestellten Auswertungsansätze wird in Kapitel sieben beschrieben. Gegenstand des Kapitels ist außerdem die Auswahl und Vorstellung der durchzuführenden statistischen Auswertungen. Die Vorgehensweise und technische Umsetzung der Auswertungen wird in Kapitel acht dargestellt. Kapitel neun umfasst die Vorstellung und Interpretation der Auswertungsergebnisse. In Kapitel zehn werden die sich aus dem vorhergehenden Kapitel ergebenden Empfehlungen für den DBS-Betreiber vorgestellt. Kapitel elf schließlich beinhaltet das Fazit.

Konvention

Zur Vermeidung von Doppelnennungen bei männlichen und weiblichen Bezeichnungen wird durchgehend die maskuline Form verwendet. Dies ist keine Wertung und schließt, wenn nicht anders angegeben, die feminine Form stets mit ein.

2. Der Deutsche Bildungsserver als Untersuchungsgegenstand

Der Deutsche Bildungsserver (DBS)[1] ist ein deutsches Internetportal im Bildungsbereich. Er startete 1996 im Rahmen der Initiative „Schulen ans Netz“ und wird seit Oktober 1999 beim Informationszentrum Bildung (IZB) des Deutschen Instituts für Internationale Pädagogische Forschung (DIPF) inhaltlich und technisch betreut [vgl. Botte 2005, S. 5]. Sein Zweck ist nicht die Bereitstellung von Inhalten im Sinne von redaktionellen Beiträgen oder Volltextdokumenten, sondern die Bereitstellung von Verweisen zu gesichteten und geprüften bildungsrelevanten Internetseiten. Dazu findet eine redaktionelle Erfassung und Strukturierung zugehöriger Metadaten statt. Der DBS ist somit als Verweis- oder Metaserver einzuordnen [Diepold 2000].

2.1. Inhalte und Struktur des DBS

Die Eintragung der Datensätze wird von der DBS-Redaktion vorgenommen. Darüber hinaus können auch Benutzer Inhalte zur Aufnahme vorschlagen. Die Vorschläge werden nach einer redaktionellen Prüfung und ggf. Überarbeitung freigeschaltet.

Innerhalb des DBS existieren mehrere Datenbanken, deren Datensätze zentral durchsuchbar sind und zum Teil auf Überssichtsseiten verlinkt werden. Den umfangreichsten Teil machen mit etwa 27.000 Datensätzen die so genannten Onlineressourcen aus. Jede Onlineressource beschreibt eine Internetseite oder einen im Internet abrufbaren Volltext mit Bildungsbezug. Eine weitere Datenbank bildet derzeit etwa 13.000 Veranstaltungen ab. Außerdem gibt es eine Institutionen-Datenbank mit knapp 8.000 Datensätzen. Weitere Datenbanken decken Wettbewerbe, Stellenangebote und -gesuche sowie Projekte der Bund-Länder-Kommission (BLK) ab. Zusammengeführt werden all diese Datentypen in über 5.500 redaktionell gestalteten „Seiten“. Jede dieser Seiten wird dabei nach dem Baukastenprinzip aus Kurzbeschreibungen der einzelnen Datensätze zusammensetzt. Enthalten kann eine Seite daneben

[1] `http://www.bildungsserver.de`

auch HTML-formatierten Freitext, Verweise auf andere vom IZB betriebene Angebote wie das Bildungswiki oder das Fachinformationssystem (FIS) Bildung[2] sowie auf externe Datenbanken wie den „ZDF-Katalog für Unterricht und Bildungsmedien“. Durch die Möglichkeit, auch andere Seiten einzubinden, lässt sich eine Hierarchie aufbauen.

Aus der Sicht eines Nutzers entsteht auf diese Weise ein thematisches Linkverzeichnis. Es bietet drei Zugänge zu den Inhalten: In der Navigationsspalte sind inhaltlich zusammengehörende Themenbereiche, sogenannte „Informationssäulen“, aufgelistet, wie beispielsweise „Schule“, „Berufliche Bildung“ oder „Hochschulbildung“ (vgl. Bildschirmfoto in Abb. 12.1, S. A1). Ebenfalls in der Navigation gibt es einen nach Zielgruppen geordneten Zugang, der unter anderem SchülerInnen, LehrerInnen und Wissenschaftler bedient.

Ein weiterer, dritter Zugang zu den Datensätzen des DBS besteht in der Suche. Bei der einfachen Suche, deren Eingabetextfeld auf jeder Seite angezeigt wird, kann der Nutzer eine Stichwortsuche über alle Datenfelder durchführen. Die stets neben diesem Eingabefeld verlinkte erweiterte Suche bietet eine feldspezifische Suche und erlaubt zudem die Auswahl einzelner Datenbanken bzw. Datenbankgruppen.

Über die DBS-eigenen Datenbanken hinaus wurden und werden eine Reihe von Unter- und Schwesterportalen aufgebaut. Das Innovationsportal[3] beispielsweise ist ein Subportal des DBS, welches auf bildungspolitisch innovative Maßnahmen abzielende Projekte von Bund und Ländern verzeichnet. Darüber hinaus bietet es in Form der oben beschriebenen DBS-„Seiten“ strukturierte Informationen und Links auf die Internetseiten einer Reihe bildungswissenschaftlicher Projekte.

Im Rahmen des „Bildungsserver 2.0“-Projektes[4] wurden weitere Angebote auf Basis von Opensource-Content-Management-Systemen (CMS) aufgebaut. Dazu zählen fünf Wikis[5], wovon das Bildungslexikon[6], eine Wikipedia-Abspaltung, das umfangreichste ist. Daneben wurden zwei Blogs eingerichtet: Der Bildungsserver-Blog[7] enthält als Blog im klassischen Sinn [vgl. Wikipedia 2009a] Meinungen und Meldungen aus dem Bildungswesen. Der InfoBlog[8] dagegen erlaubt es seinen Nutzern, Fragen zu stellen, die dann von anderen Nutzern oder DBS-Redakteuren beantwortet werden.

Eine Übersicht sämtlicher Datenbanken und Portale des DBS ist in Abbildung 2.1 dargestellt. Abbildung 2.2 zeigt den Aufbau und die interne Verlinkung des DBS. Die

[2] `http://www.fachportal-paedagogik.de/`
[3] `http://www.bildungsserver.de/innovationsportal/`
[4] `http://www.dipf.de/projekte/bildungsserver-2.0`
[5] Übersicht: `http://www.bildungsserver.de/zeigen.html?seite=5277`
[6] `http://wiki.bildungsserver.de`
[7] `http://blog.bildungsserver.de`
[8] `http://blog.bildungsserver.de/infoboerse/`

beiden äußeren Rahmen visualisieren die technische Umsetzung des DBS. Während für die Web 2.0-Angebote eine eigene Plattform verwendet wird, liegt den DBS-spezifischen Datenbanken ein selbstentwickeltes CMS zugrunde („Kern-CMS“ in der Abbildung). Die DBS-eigene Suchfunktion deckt sämtliche Inhalte dieses CMS sowie die Web 2.0-Angebote des DBS und weitere, vom IZB oder externen Partnern betriebene Datenbanken ab. Neben dem Zugang über die Suche besteht ein Navigationszugang in Form von hierarchisch organisierten, nach Themenbereichen und Zielgruppen strukturierten Redaktionsseiten.

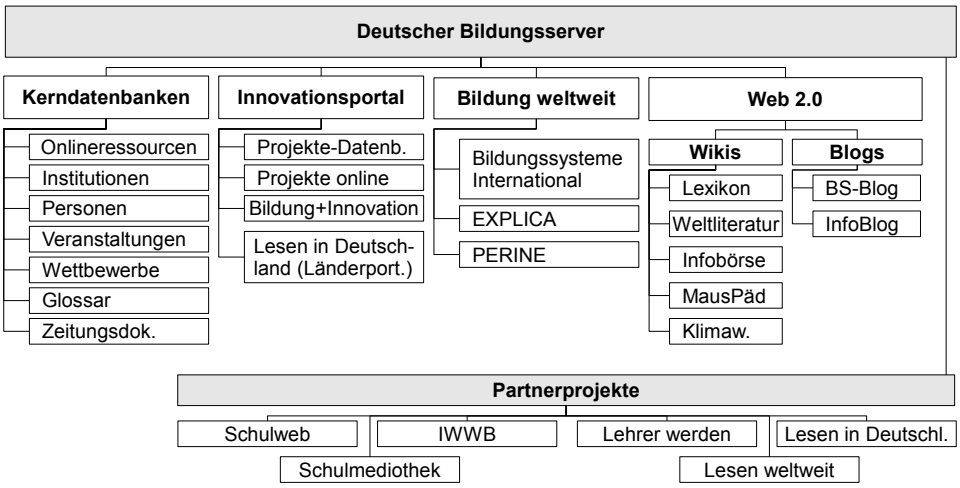


Abbildung 2.1.: Übersicht der Datenbanken, Portale und Partnerportale des DBS [eigene Darst.]

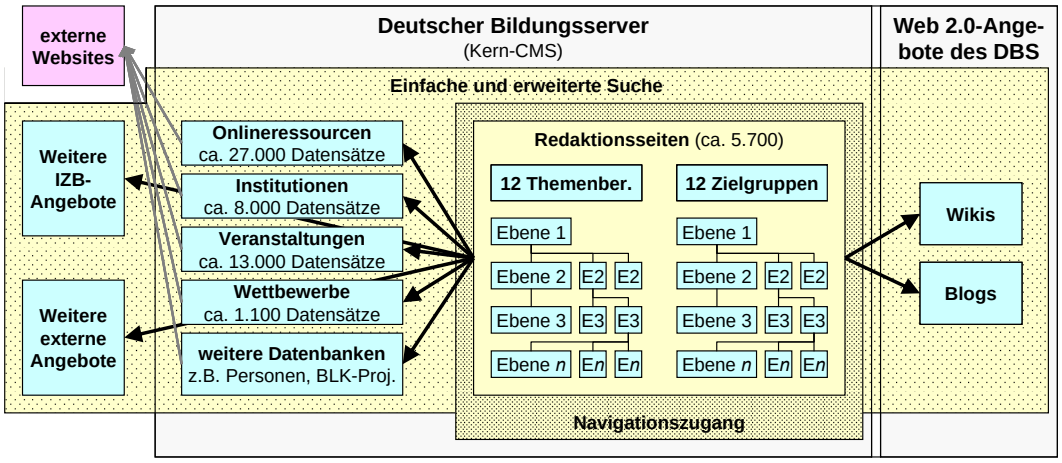


Abbildung 2.2.: Schematische Darstellung DBS-Struktur [eigene Darst.]

2.2. Ziele des Bildungsservers

Ziel des DBS ist es, ein zentraler Einstieg für Internetquellen jeder Art im Bildungswesen zu sein. Thematisch soll die gesamte Breite des Bildungswesens abgedeckt werden. Dazu zählen praxis-, forschungs- und politikbezogene Inhalte. Die nachgewiesenen Inhalte müssen frei zugänglich sein. Die Nutzergruppe selbst ist nicht näher umrissen, vielmehr sollen alle interessierten Nutzer angesprochen werden [vgl. Botte und Hirschmann 2000, S. 1].

Darüber hinaus richtet sich der DBS auch an ein internationales Publikum, das Informationen zum deutschen Bildungswesen sucht. Zu diesem Zweck wird ein inhaltlich an das deutschsprachige Angebot angelehnter, aber mit anderen Schwerpunkten versehener englischer Auftritt gepflegt.

Im Unterschied zu von Suchmaschinen oder klassischen Webkatalogen wie Yahoo[9] bereitgestellten Links werden die beim DBS bereitgestellten Ressourcen von wissenschaftlich ausgebildeten Fachkräften aufwändig gesichtet und dokumentarisch erschlossen [vgl. Botte und Maier 1999, S. 2].

2.3. Technische Grundlagen des DBS

Softwareseitig basiert der DBS auf einem selbstentwickelten System zur Inhaltsverwaltung (CMS). Die Skripte zur Seitengenerierung sind in der Web-Skriptsprache PHP verfasst. Die Inhalte selbst werden in einer MySQL-Datenbank abgelegt.

Hardwareseitig läuft der DBS zusammen mit FIS Bildung auf einer hochverfügbaren Server-Gruppe bestehend aus vier Einzelrechnern. Jeder einzelne Zugriff wird über einen Lastverteiler (sog. Load Balancer) zufällig auf einen der Server verteilt. Die Lastverteilung und die Sicherstellung der Verfügbarkeit wird dabei mittels der freien Software Heartbeat[10] realisiert.

Die Web 2.0-Inhalte sind auf einem gesonderten Server eingerichtet. MediaWiki[11] dient als Grundlage der Wikis, die Blogs verwenden WordPress[12].

[9] `http://de.dir.yahoo.com`
[10] `http://www.linux-ha.org/TechnicalOverview`
[11] `http://www.mediawiki.org`
[12] `http://de.wordpress.org`

3. Grundlagen von Web-Informationssystemen

Zur Einordnung des DBS in die verschiedenen Typen von Informationssystemen werden in diesem Kapitel Grundbegriffe des Web-Retrievals definiert.

3.1. Informationen und Informationsbedarf

Die Abgrenzung der Begriffe Daten, Wissen und Information wird in der Literatur nicht einheitlich vorgenommen[1]. In dieser Arbeit werden die an Kuhlens Informationsbegriff angelehnten Definitionen verwendet.

Daten Bei Daten handelt es sich um Einträge, „deren Typ oder syntaktische Struktur bekannt ist“ [Ferber 2003, S. 27]. Das können Zahlen, Wörter oder Bitfolgen sein. Ein weiteres Merkmal ist die Verwendung von Zeichen aus einem festgelegten Zeichenvorrat [vgl. Probst u. a. 1999, S. 36]. Daten sind auf ein bestimmtes Format, eine Syntax, beispielsweise auf Ganzzahlen oder Zahlenformate wie ZZZZ-ZZ-ZZ, festgelegt. Ihre semantische Bedeutung spielt auf dieser Betrachtungsebene keine Rolle.

Wissen Wissen entsteht Kuhlen zufolge „durch Umsetzung von Daten über Informationen in interne Wissensstrukturen“ [Kuhlen 2004, S. 13]. Wissen sei dabei dem aktuellen Nutzungskontext entzogen und somit nachhaltig.

Information Kuhlen zufolge werden Daten durch gezielten Abruf (beispielsweise aus einer Datenbank) und Wahrnehmung in einem bestimmten Kontext oder zu einem bestimmten Zweck zu Information [vgl. Kuhlen 2004, S. 12]. Zusammenfassend definiert er Information so:

[1] Für eine Übersicht der jeweiligen Begriffsdefinitionen siehe `http://beat.doebe.li/bibliothek/w00600.html` für Daten, `http://beat.doebe.li/bibliothek/w00050.html` für Wissen und `http://beat.doebe.li/bibliothek/w00021.html` für Information [Stand: 01. Dezember 2009].

> Informationen gibt es nicht als Objekt für sich. Information ist eine Referenzfunktion. Information kann nur über eine repräsentierte/kodierte Form von Wissen aufgenommen werden. Informationen referenzieren nicht nur auf repräsentiertes Wissen, sondern entfalten diese Bedeutung nur mit Berücksichtigung der aktuellen Benutzungssituation. Informationen bedeuten etwas, aber – und das macht das pragmatische Grundverständnis aus – sie existieren nicht losgelöst von ihrer Nutzung. Von Information kann man nur im aktuellen Kontext ihrer Verwendung sprechen, unter Berücksichtigung der verschiedenen Rahmenbedingungen ihrer Bedeutung. [Kuhlen 2004, S. 14]

Von diesem pragmatischen Informationsbegriff soll im Folgenden ausgegangen werden. Bezogen auf den DBS bedeutet dies, dass die Einträge Daten entsprechen. Findet ein Nutzer im DBS einen Eintrag, der sein Informationsbedürfnis befriedigt, wird der Datensatz zur Information. Merkt sich der Nutzer die URL der Zielseite oder den Namen ihres Betreibers, entsteht Wissen.

Informationsbedarf Beim Informationsbedarf kann zwischen einem konkreten und einem problemorientierten Bedarf unterschieden werden [vgl. Stock 2007, S. 51ff.]. Beim konkreten Informationsbedarf lässt sich durch die klaren thematischen Grenzen die Suchanfrage in exakten Termen formulieren. Zur Deckung des Informationsbedarfs reicht meist eine einzige Fakteninformation aus. Der problemorientierte Bedarf bildet hierzu einen Gegensatz: Die thematischen Grenzen sind nicht präzise bestimmbar und bei der Formulierung der Suchfrage sind mehrere terminologische Varianten möglich. Eine weitere Unterscheidung lässt sich zwischen Informationsbedarf und -bedürfnis vornehmen: Bedarf bezeichnet dabei einen objektiven, von einzelnen Personen abstrahierbaren Sachverhalt. Betrachtet man hingegen auch individuelle Präferenzen, beispielsweise für bestimmte Autoren, spricht man von einem Bedürfnis. Zur Befriedigung von Informationsbedürfnissen lassen sich Information-Retrieval-Systeme nutzen. Sie werden im nächsten Abschnitt beschrieben.

3.2. Information Retrieval

Information Retrieval bezeichnet ein Teilgebiet der Informationswissenschaft, das sich mit dem Wiederauffinden gespeicherter Informationen beschäftigt [vgl. Stock 2007, S. 38]. Ähnlich wie die beiden oben beschriebenen Bedarfstypen voneinander, lässt sich Information Retrieval von Data Retrieval abgrenzen [vgl. Rijsbergen 1979, S. 1f.]. Während beim Data Retrieval Suchanfragen in einer künstlichen Anfragesprache gestellt werden, mit dem Ziel, exakt passende Elemente zu finden, wird beim

Information Retrieval die natürliche Sprache zur Anfrageformulierung verwendet. Es werden nicht nur exakte Übereinstimmungen, sondern auch Teiltreffer ausgegeben, wobei der Fokus auf ihrer Relevanz liegt.

Die in einem informationspraktischen Arbeitsprozess automatisch oder intellektuell erschlossenen Gegenstände des jeweiligen Systems werden als Dokumentationseinheiten bezeichnet [vgl. Stock 2007, S. 54f.].

Das Information-Retrieval-System fungiert als Schnittstelle zwischen dem Informationsbedürfnis der Nutzer und den im System repräsentierten Dokumentationseinheiten [vgl. Ferber 2003, S. 24f.]. Abbildung 3.1 zeigt das Grundprinzip eines Information-Retrieval-Systems.

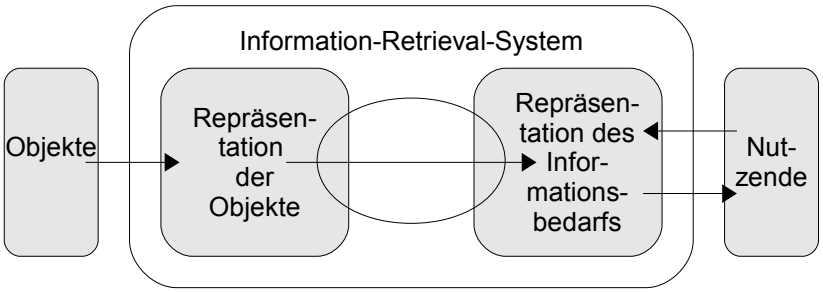


Abbildung 3.1.: Schema eines Information-Retrieval-Systems [Ferber 2003, S. 25]

Der Interaktionsmechanismus, dargestellt durch den Kreis als Schnittmenge zwischen beiden Repräsentationen, kann über unterschiedliche Verfahren umgesetzt werden. Ein weit verbreitetes Verfahren ist das Boolesche Retrieval, bei dem mittels der Operatoren `UND`, `ODER` und `UND NICHT` eine Schnittmenge zwischen der Suchanfrage und der Dokumentmenge ermittelt wird [vgl. Ferber 2003, S. 33f.]. Da die Dokumente lediglich auf ihre Übereinstimmung mit der Suchanfrage überprüft werden, ist das Ergebnis eine ungeordnete Treffermenge. Diesen Nachteil behebt das Vektorraummodell. Es nimmt eine Gewichtung der Dokumentterme vor, beispielsweise auf Grundlage von Art und Häufigkeit ihres Auftretens in den Dokumenten [vgl. Ferber 2003, S. 61f.]. Suchanfragen und Dokumente werden durch Vektoren repräsentiert, deren Übereinstimmung mittels einer Ähnlichkeitsfunktion berechnet wird. Aus den unterschiedlichen Ähnlichkeitswerten der verschiedenen Dokumente ergibt sich eine Rangfolge, die sich für eine entsprechend sortierte Ergebnisausgabe verwenden lässt. In Hypertexträumen wie dem World Wide Web (WWW) nutzen linktopologische Modelle wie der von Google verwendete PageRank-Algorithmus die Verlinkung der Dokumente untereinander, um auf Basis der Art und Anzahl ausgehender und eingehender Links eine Gewichtung zur Ergebnissortierung vorzunehmen [vgl. Ferber 2003, S. 303f.]. Neben den bisher genannten Verfahren können auch Klassifikationen,

Thesauri und semantische Netze (z.B. Ontologien) dazu verwendet werden, Begriffsbeziehungen und Bezeichner festzulegen [vgl. Ferber 2003, S. 59ff.]. Mit ihrer Hilfe lassen sich beispielsweise synonyme Suchworte auf einen im Information-Retrieval-System einheitlich verwendeten Deskriptor überführen.

3.3. Hypertext

Hypertext bezeichnet in Blöcke zerlegte Texte. Diese Blöcke sind durch Verweise untereinander verknüpft [vgl. Hammwöhner 2004, S. 419f.]. Dadurch muss der Leser den Text nicht linear wahrnehmen, sondern kann ihn manipulieren [vgl. Kuhlen 1991, S. 13ff.]. Manipulation meint in diesem Zusammenhang, dass Benutzer durch individuelle Nutzung der Verknüpfungen die Textblöcke in andere oder neue Kontexte stellen können. Das Ergebnis der Manipulation ist sofort sichtbar und erlaubt somit schnelle und inkrementelle Aktionen.

Die Textblöcke werden auch als informationelle Einheiten bezeichnet. Informationelle Einheiten werden bei der Gestaltung des Hypertextsystems festgelegt [vgl. Kuhlen 1991, S. 80f.]. Sie beziehen sich entsprechend dem oben beschriebenen Grundsatz von „Wissen in Aktion“ auf die aktuelle Dialogsituation. Sie sollten so umfangreich sein, dass sie als solche verstanden werden können, aber auch Verweise auf ihren informationellen Kontext sind von Bedeutung. Kontext bei der Darstellung einer informationellen Einheit kann auch über die Anzeige von ergänzenden Metainformationen hergestellt werden.

Verknüpfungen zwischen Hypertextelementen lassen sich anhand ihres Strukturierungsgrades und ihres Radius' unterscheiden [vgl. Kuhlen 1991, S. 105ff.]. Verknüpfungen, die sich keinem Strukturierungstyp zuordnen lassen, bezeichnet Kuhlen als referentielle Verknüpfungen, welche den assoziativen Relationen bei Thesauri entsprechen. Beschreibt eine Verknüpfung dagegen eine semantische Relation (z.B. Ober-/Unterbegriffe) oder eine Kohärenz, handelt es sich um eine typisierte Relation. Neben dieser Unterscheidung lässt sich auch eine Einteilung in intra-, inter- und extrahypertextuelle Verknüpfungen vornehmen. Bei ersteren liegen Ausgangs- und Zielpunkt innerhalb der selben informationellen Einheit. Interhypertextuelle Verknüpfungen verbinden verschiedene informationelle Einheiten. Extrahypertextuelle Verknüpfungen dagegen verweisen auf externe Websites oder Datenbanken.

Otter und Johnson beschreiben eine andere Klassifikation von Hyperlinks, die in Abbildung 3.2 dargestellt ist. Danach wird zunächst in Anlehnung an die Linguistik zwischen intensionalen und extensionalen Links unterschieden, wobei Letztere als permanente Gegenstände einer Hypertext-Struktur verstanden werden [vgl. Otter

und Johnson 2000, S. 8f.]. Sie werden wiederum in relationale, d.h. nur zu einem einzigen Knoten verbindende, und inklusive Links, die zu mehr als einem Knoten verbinden, untergliedert. Relationale Links lassen sich unterscheiden in assoziative Links, die beliebige Dokumente verlinken, und erläuternde Links, die beispielsweise zu einer Informationsseite über einen Text verlinken. Auch bei inklusiven Links wird eine weitere Unterteilung vorgenommen in sequentielle und taxonomische Links. Erstere verlinken zu mehreren Dokumenten in einer festgelegten Reihenfolge, Letztere ebenso, allerdings ohne die Festlegung einer Sequenz.

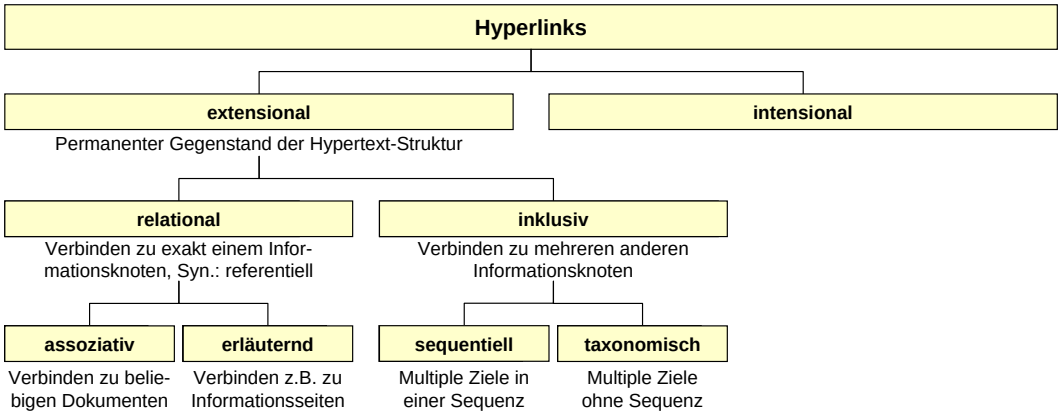


Abbildung 3.2.: Hyperlink-Klassifikation [eigene Darst. nach Otter und Johnson 2000, S. 8f.]

Das auf Grundlage der Auszeichnungssprache Hypertext Markup Language (HTML) realisierte WWW ist heute das umfassendste und bekannteste Hypertextsystem [vgl. Wikipedia 2009d].

3.4. Information Retrieval und Hypertext

In umfangreichen oder thematisch breit angelegten Hypertextsystemen sind je nach Einstiegspunkt und Ziel sehr viele Navigationsschritte erforderlich, um dieses Ziel anhand von Verknüpfungen zu erreichen. Daraus lässt sich die Erforderlichkeit gezielter Suchmöglichkeiten ableiten [vgl. Bekavac 1999, S. 53ff.]. Dies ermöglicht dem Nutzer die Wahl zwischen einem navigatorischen, Hypertext-basierten Zugang und einer Information-Retrieval-Komponente. Verschiedene Autoren betrachten den Hypertextansatz als ein Modell des Information Retrieval [vgl. u.a. Kuhlen 1995, S. 279].

3.5. Fachinformation und Webangebote

Während Information-Retrieval-Systeme und Hypertext-Strukturen technische Aspekte von Informationssystemen beschreiben, sollen im Folgenden einige Anwendungsbeispiele von Web-Informationssystemen beschrieben werden.

Fachdatenbanken zeichnen sich aus durch eine gute Strukturierung ihrer Felder und eine hohe Qualität durch fachliche Auswahl und Bearbeitung [vgl. Kind 2004, S. 389]. Diese professionellen Fachdatenbanken sind fast immer kostenpflichtig. Zur Ermöglichung eines Wiederauffindens der gespeicherten Informationen werden Information-Retrieval-Methoden eingesetzt.

Webkataloge verzeichnen Websites in einer thematischen Übersicht [vgl. Bekavac 2004, S. 400f.]. Die Einträge werden redaktionell zusammengestellt. Häufig besteht für Nutzer die Möglichkeit, neue Einträge zu ergänzen. Nutzer können durch Browsing, also der Nutzung von Hypertext-Strukturen, oder durch eine Suchfunktion auf die Kataloginhalte zugreifen.

3.6. Bezug zum DBS

Der DBS lässt sich in den oben beschriebenen Kategorien und Dimensionen mehrfach verorten. Mit seiner Funktion als Verweisserver und mit der themen- bzw. zielgruppenorientierten Kategorisierung der Einträge ist der DBS ein Webkatalog. Durch die umfangreiche Suchfunktion, die weitreichende inhaltliche Erschließung und die festgelegten Aufnahmekriterien werden zudem Kriterien einer Fachdatenbank erfüllt.

Die umfangreiche Suchfunktion des DBS ist eine Information-Retrieval-Komponente. Mit seinen Navigationsstrukturen und den zahlreichen internen und externen Verknüpfungen ist er zugleich ein Hypertext-System. Die einzelnen Datenbankeinträge und redaktionellen Seiten bilden jeweils informationelle Einheiten. Eine Ausnahme hiervon sind Themendossiers, die meist aus mehreren redaktionellen Seiten bestehen und somit je Dossier eine informationelle Einheit bilden.

Die Verknüpfungen innerhalb einer Themensäule oder Zielgruppe bilden Hierarchien ab oder stellen anderweitige Zusammenhänge her und entsprechen somit typisierten Relationen nach Kuhlen. Weitere Links, beispielsweise zur „Über uns“-Seite, haben referenziellen Charakter.

4. Nutzungsanalyse von Webangeboten

In diesem Kapitel wird zunächst eine Abgrenzung zwischen Webanalyse und Web Usage Mining vorgenommen. Im Anschluss daran werden die grundlegenden Methoden und ihr Vorliegen beim DBS der Erhebung von Nutzungsdaten beschrieben.

4.1. Begriffliche Einordnung

Eine grundsätzliche Unterscheidung lässt sich zwischen Webanalyse und Web Usage Mining vornehmen. Während die Vorgehensweise bei der Webanalyse stark praxisorientiert ist und große Bedeutung insbesondere bei kommerziell betriebenen Webauftritten hat, ist Web Usage Mining tendenziell stärker wissenschaftlich-experimentell orientiert.

4.1.1. Webanalyse

Der Begriff „Webanalyse“ bzw. die englische Entsprechung „Web Analytics“ ist weitgefasst und soll daher zunächst eingegrenzt werden.

Der Branchenverband Web Analytics Association (WAA) definiert Web Analytics wie folgt: „Web Analytics is the measurement, collection, analysis and reporting of Internet data for the purposes of understanding and optimizing Web usage.“ [Web Analytics Association 2009]

Ferrini und Mohr setzen „Web analytics“ mit „Logfile analysis“ gleich und verstehen darunter dementsprechend konkret die Auswertung von Logfiles eines bestimmten Webauftritts mit dem Ziel, die Performanz der Website zu bewerten [vgl. Ferrini und Mohr 2009, S. 125].

Diesen traditionellen Ansatz hält Kaushik für nicht mehr zeitgemäß („Traditional web analytics is dead.“) [vgl. Kaushik 2007b, S. 8]. Er propagiert die Definition individueller, strategiebezogener Kennzahlen (sog. Key Insights Analysis (KIA)), wie z.B. die Rate erfolgreich abgeschlossener Anwendungsszenarien durch die Nutzer oder eine weitreichende Datensegmentierung nach Nutzergruppen [vgl. Kaushik 2007b,

S. 10ff.]. Dazu seien auch weitere Datenquellen wie Befragungen oder Nutzertests heranzuziehen.

Trotz dieser begrifflichen Weiterentwicklung kann die Auswertung von Nutzungsdaten weiterhin als der Kern- und Hauptbestandteil von Webanalyse angesehen werden [vgl. Hassler 2009, S. 26ff.]. Von diesem Begriffsverständnis wird in dieser Arbeit ausgegangen.

Die im Rahmen der Webanalyse typischerweise erhobenen Metriken werden in Kapitel 5.1 beschrieben.

4.1.2. Web Mining

Neben dem Begriff der „Webanalyse“ gibt es den Begriff „Web Mining“. Darunter wird die Anwendung von Data-Mining-Techniken zur Entdeckung und Extraktion von Informationen aus Web-Dokumenten und -Diensten verstanden [Kosala und Blockeel 2000, S. 2]. Data Mining und die häufig synonym verwendete Bezeichnung Knowledge Discovery in Databases (KDD) sind „automatisierte Verfahren, mit denen Regelmäßigkeiten in Mengen von Datensätzen gefunden und in eine für Nutzende verständliche Form gebracht werden“ [Ferber 2003, S. 14].

Das Vorgehen beim Web Mining umfasst generell vier Schritte:

1. Das Finden gewünschter Web-Ressourcen als Analyseobjekte,
2. die automatische Auswahl und eine Vorverarbeitung der abgerufenen Web-Ressourcen,
3. die Identifikation allgemeiner Muster auf Ebene einer Website oder Website-übergreifend und
4. die Analyse und Interpretation der gefundenen Muster.

Dadurch ergeben sich Überschneidungen mit den Themenfeldern „Data Mining“, „Maschinelles Lernen“ und „Datenanalyse“. Drei Teilgebiete von Web Mining lassen sich unterschieden: „Web Structure Mining“, „Web Content Mining“ und „Web Usage Mining“ [vgl. Kosala und Blockeel 2000, S. 3f.].

Web Content Mining beschreibt das Entdecken von Informationen in Web-Inhalten. Das Spektrum solcher Inhalte ist dabei sehr breit. Zu den Analyseobjekten zählen verschiedenartige Inhaltstypen wie Texte, Bilder, Videos, Musik, Metadaten und Hyperlinks. Unterschieden wird beim Web Content Mining zwischen unstrukturieren Inhalten wie Freitext und strukturierten Inhalten wie Datenbank-Inhalten, wobei semistrukturierte Inhalte wie HTML-Seiten einen Übergangsbereich kennzeichnen.

Web Structure Mining hingegen widmet sich der Hyperlink-Topologie im Web [vgl. Kosala und Blockeel 2000, S. 4]. Ziel ist es, das zugrundeliegende Modell der Linkstrukturen im Web zu identifizieren. Auf dessen Grundlage ist es dann möglich, Internetseiten zu kategorisieren sowie Ähnlichkeiten und Beziehungen zwischen Seiten zu bestimmen. Ein Anwendungsbeispiel hierfür ist der von Google verwendete PageRank-Algorithmus [vgl. Kosala und Blockeel 2000, S. 9].

Das dritte Teilgebiet von Web Mining bildet das Web Usage Mining. Gegenstand ist die automatische Erkennung und Analyse von Mustern in Klickstromdaten und sonstigen Datenquellen mit dem Ziel, Bewegungsmuster und Profile der Nutzer zu identifizieren [vgl. Mobasher 2007, S. 449]. Zwei Ansätze der Datenverarbeitung können beim Web Usage Mining unterschieden werden [vgl. Kosala und Blockeel 2000, S. 9]. Bei dem ersten Ansatz werden die Nutzungsdaten zunächst in eine relationale Datenbank geschrieben und anschließend entsprechende Data-Mining-Methoden angewendet. Beim zweiten Ansatz hingegen werden die Protokolldaten des Webservers direkt unter Anwendung spezieller Vorverarbeitungstechniken verwendet. Beim Einsatz von Web Usage Mining können weiterhin zwei Anwendungsgebiete unterschieden werden. Gegenstand der Datenanalyse kann entweder die Ermittlung von personalisierten Nutzerprofilen oder von nicht personalisierten Navigationsmustern sein.

Abbildung 4.1 zeigt die Unterteilungen des Begriffs Web Mining.

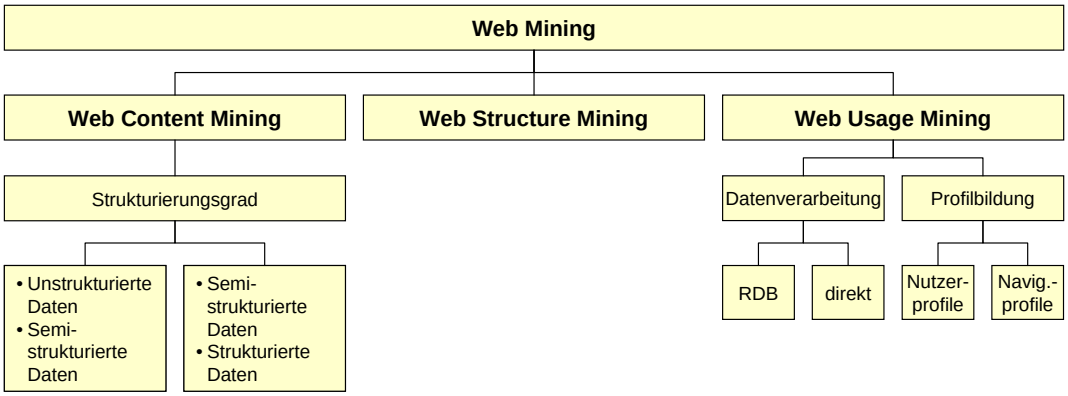


Abbildung 4.1.: Unterbegriffe und Unterscheidungen bei Web Mining [eigene Darst. nach Kosala und Blockeel 2000]

Auf Grundlage dieser Definitionen von Webanalyse und Web Mining bzw. im Speziellen Web Usage Mining lässt sich folgende Abgrenzung vornehmen: Bei der Anwendung von Webanalyse kommen vorwiegend statistische Metriken zum Einsatz (vgl. Kap. 4.2.4), während beim Web Usage Mining komplexere Verfahren wie Assoziative Regeln, Clustering oder Automatische Klassifkation angewendet werden.

Die allgemeine Vorgehensweise beim Web Usage Mining wird in Abschnitt 6.1 dargestellt. Eine Beschreibung und Bewertung von konkreten Auswertungsansätzen folgt in Abschnitt 6.2.

4.2. Methoden der Datenerhebung

Bei der Erfassung von Nutzungsdaten eines Webangebots kann zwischen server- und clientbasierten Verfahren unterschieden werden. Die am weitesteten verbreitete serverbasierte Art stellen Logfiles dar. Clientseitig werden meist JavaScript-Tags verwendet. Cookies dienen der Besucheridentifikation und können mit Logfiles oder JavaScript-Tags verwendet werden. Diese drei am weitesten verbreiteten Verfahren werden im Folgenden beschrieben.

4.2.1. Logfiles

Logfiles sind Zugriffsprotokolle, die direkt vom Webserver generiert werden. Protokolldateien sind die älteste Datenquelle für Webanalyse, allerdings wurden sie nicht für diesen Zweck konzipiert. Die Intention war vielmehr technischer Natur; der primäre Zweck war die Überprüfung der Funktionsfähigkeit des Webservers [vgl. Kaushik 2007b, S. 26f.].

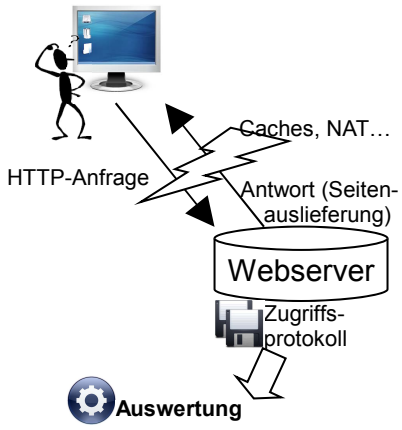


Abbildung 4.2.: Funktionsweise der Logfile-basierten Zugriffserfassung

Bei jedem Seitenaufruf durch den Benutzer stellt der Browser eine HTTP-Verbindung zum Webserver her und ruft zunächst die zur URL gehörende HTML-Seite ab. Während der anschließenden Browser-seitigen Verarbeitung der Seite wird für jede darin referenzierte Datei (beispielsweise Bilder, JavaScript-Bibliotheken oder

Formatvorlagen) eine neue HTTP-Verbindung zum jeweils angegebenen Server hergestellt und die Datei abgerufen. HTTP ist somit ein zustandloses Protokoll, bei dem die Verbindung nach jeder Dateiübertragung wieder abgebaut wird. Zugleich wird für jede seitens des Webservers übermittelte Datei ein Eintrag im Logfile erstellt. Die Funktionsweise wird in Abbildung 4.2 dargestellt.

Je nach verwendetem Webserver und dessen Konfiguration variiert das Format der Protokolldateien, also Art, Umfang, Reihenfolge und das Auszeichnungsformat der Daten. Am weitesten verbreitet ist der Apache-Webserver[1] [vgl. Netcraft 2009]. Er kommt auch bei allen vom IZB betriebenen Webangeboten zum Einsatz. Apache führt neben einem nicht beeinflussbaren Fehlerprotokoll ein Zugriffsprotokoll, dessen Format in einer Konfigurationsdatei festgelegt werden kann [vgl. Apache Software Foundation 2009]. Neben dem sog. „Common Log Format“ ist auch das im Vergleich dazu um zwei Einträge erweiterte „Combined Log Format“ gebräuchlich. Es soll am folgenden Beispiel erklärt werden:

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326
"http://www.xyz.de/start" "Mozilla/4.08 [en] (Win98; I ;Nav)"
```

Die einzelnen Bestandteile werden in Tabelle 4.1 vorgestellt. Beispielsweise ist ersichtlich, dass die abgerufene Datei (Nr. 6) stets als relativer Pfad ausgehend vom Hauptverzeichnis des Webservers, und nicht als vollständige URL gespeichert wird. Im Gegensatz dazu ist dies beim Referrer (Nr. 10) der Fall.

Nr.	Protokoll	Bedeutung
1	`127.0.0.1`	IP-Adresse des Clients
2	-	Nutzeridentifikation nach IETF-RFC 1413[2], wird in der Praxis kaum verwendet
3	`frank`	Benutzername bei HTTP-Identifikation
4	`[10/Oct/2000:13:55:36 -0700]`	Datum, Uhrzeit des Zugriffs aus Serversicht
5	`"GET`	Erster Abfrageteil: Zugriffsmethode
6	`/apache_pb.gif`	Zweiter Abfrageteil: Abgerufene Datei
7	`HTTP/1.0"`	Dritter Abfrageteil: Protokoll und Protokollversion
8	`200`	Zugriffscode
9	`2326`	Größe der Datei in Byte
10	`"http://www.xyz.de/start"`	Verweisende Adresse, sog. „Referer“[3] bzw. Referrer (vgl. Abschn. 5.1.7)

[1] `http://httpd.apache.org/`

[2] Weitere Informationen siehe `http://de.wikipedia.org/w/index.php?title=Ident&oldid=45930966`; Stand: 10. Dezember 2009

[3] „Die korrekte englische Schreibweise lautet Referrer. Der ursprüngliche Request for Comments (RFC 2068) enthielt jedoch versehentlich die falsche Schreibweise Referer und erhebt diesen Wortlaut damit zum Standard innerhalb von HTTP. In anderen Standards wie im DOM wird die korrekte Schreibweise verwendet [...].“ `http://de.wikipedia.org/w/index.php?title=Referrer&oldid=66554185`

11	`"Mozilla/4.08 [en] (Win98; I ;Nav)"`	Endsystem des Benutzers

Tabelle 4.1.: Bestandteile eines Logfile-Eintrags nach dem „Combined Log Format"

Vorteile der Logfile-Erhebung

Die Verwendung von Webserver-Protokolldateien für die Webanalyse bietet eine Reihe von Vorteilen gegenüber JavaScript-Tags. Die Daten werden in der Apache-Standardkonfiguration ohnehin gesammelt [vgl. Kaushik 2007b, S. 26f.]. Werden sie archiviert, steht somit auch eine Datenquelle für historische Analysen zur Verfügung. Im Gegensatz zu JavaScript-basierten Verfahren werden auch die Abrufe von Binärdateien wie zum Beispiel PDF-Dokumenten erfasst. Ferner werden auch Robot-Zugriffe protokolliert, sodass Umfang und Häufigkeit von Indexierungsvorgängen überwacht werden können [vgl. Hassler 2009, S. 56]. Die durch die Generierung der Protokolldateien erzeugte Systemlast ist zu vernachlässigen. Die rechenintensive Verarbeitung und Auswertung der Protokolle kann auf anderen Systemen oder – sofern sie auf dem Webserver selbst stattfinden – zu Zeiten geringer Frequentierung, also meist nachts, erfolgen.

Nachteile der Logfile-Erhebung

Den Vorteilen der Logfile-Erhebung stehen eine Reihe von Nachteilen gegenüber. Nicht alle Zugriffe können protokolliert werden. Zwischenspeicher (sog. „Caches") sowohl bei den Zugangsdienstleistern (Internet Service Provider (ISP)) als auch in Browsern speichern bereits abgerufene Seiten und liefern sie bei einem erneuten Abruf direkt aus, ohne den Originalserver zu kontaktieren. Die Zahl der dadurch entgangenen Protokolleinträgen von Zugriffen ist nur schwer abschätzbar. Kaushik geht von einer Größenordnung von $\geq 10\%$ aus [vgl. Kaushik 2007b, S. 27]. Des Weiteren ist die IP-Adresse als alleiniges Identifizierungsmerkmal einzelner Sitzungen oder gar Besucher ungeeignet. Der Einsatz von Network Address Translation (NAT) erlaubt es, mehrere Rechner (und damit Personen) über eine IP-Adresse mit dem Internet zu verbinden. NAT wird in Unternehmens- oder Universitätsnetzen, aber auch von Privathaushalten genutzt. Zur näherungsweisen Identifikation von Nutzern wird daher meist eine Kombination aus IP-Adresse und Browserkennung verwendet [vgl. Hassler 2009, S. 100].

In Protokolldateien werden zunächst alle Zugriffe erfasst. Nebensächliche Zugriffe (z.B. Grafiken, CSS-Dateien, JavaScript-Bibliotheken) müssen daher gefiltert wer-

den. Dies gilt auch für Robot-Zugriffe, da bei den meisten Auswertungen nur auf die Zugriffe durch Personen gezielt wird [vgl. Kaushik 2007b, S. 27].

Ein weiterer Nachteil ergibt sich aus dynamischen IP-Adressen. Wählt sich ein Privatnutzer während eines Besuchs erneut ins Internet ein, erhält er eine neue IP-Adresse. Da die IP-Adresse ein Unterscheidungsmerkmal bei der Sitzungserkennung ist, wird der eigentlich zusammenhängende Besuch doppelt gezählt.

Da für jeden Zugriff eine Zeile im Protokoll angelegt wird, nimmt der Speicherplatzbedarf der Logfiles proportional zur Nutzungshäufigkeit des Webangebotes zu. Das DBS-Logfile vom 27.04.2009 beispielsweise umfasst rund 623.600 Zeilen und belegt 150 Megabyte, welche durch gzip-Kompression auf 10,2 Megabyte reduziert werden können. Abschließend ist zu vermerken, dass die Erhebungsmethode funktional ausgereizt ist und nicht weiterentwickelt wird [vgl. Kaushik 2007b, S. 27].

4.2.2. JavaScript-Tags

Bei dieser Erhebungsmethode wird jeder einzelnen Seite der gesamten Website ein HTML- und JavaScript-Codeschnipsel hinzugefügt. Dazu genügt je nach Aufbau des CMS ein Eingriff in die zentrale Seitenvorlage. Der ergänzte Code bewirkt, dass jeder Browser, bei dem die Ausführung von JavaScript aktiviert ist, die angegebene URL aufruft [vgl. Kaushik 2007b, S. 30ff.]. Der Ablauf ist in Abbildung 4.3 dargestellt. Die Code-Einbettung ist für den Nutzer mit keinerlei Einschränkungen verbunden, selbst wenn die Ausführung von JavaScript im Browser deaktiviert ist. Die Position der Seitenauszeichnung innerhalb des HTML-Codes wird je nach Literaturquelle hinter dem öffnenden oder vor dem schließenden body-Element vorgenommen[4]. Der Aufruf wird von der Gegenstelle gezählt. Die Gegenstelle kann bei einem externen Dienstleister (Software as a Service (SaaS)-Lösung) oder in derselben Umgebung wie der Webserver oder sogar auf dem Webserver selbst installiert sein. Bei hohen Zugriffszahlen empfiehlt sich aus Performanzgründen bei einer hausinternen Datenerhebung der Einsatz eines eigenen Servers für die Webanalyse-Software.

Beim Aufruf werden zusätzliche Daten von der Bildschirmauflösung bis – je nach Analysepaket – zur Bildschirmposition des Mausklicks erfasst. Diese Daten können nur mittels JavaScript gesammelt werden. Robots führen bislang in der Regel kein JavaScript aus und werden daher nicht erfasst [vgl. Hassler 2009, S. 59].

[4] Peterson geht von einer Implementierung beim öffnenden body-Element aus [vgl. Peterson 2005, S. 42], Kaushik dagegen plädiert im Sinne der Benutzerfreundlichkeit für eine Positionierung möglichst nahe dem schließenden body-Element, damit im Falle einer langsamen oder ausbleibenden Antwort des Auswertungsservers keine Beeinträchtigungen bei der Seitendarstellung auftreten [vgl. Kaushik 2007a].

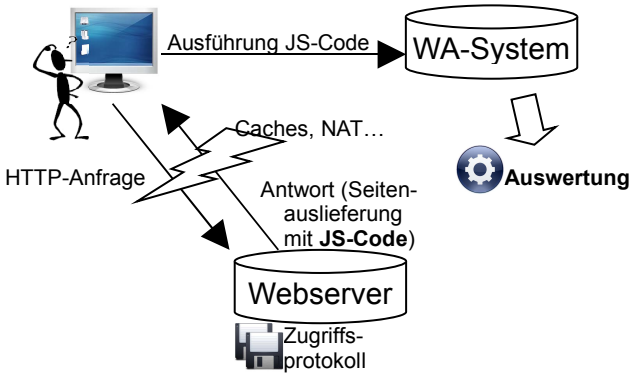


Abbildung 4.3.: Funktionsweise der JavaScript-basierten Zugriffsfassung

Vorteile der JavaScript-Tags

Ein Vorteil bei der Verwendung von JavaScript-Auszeichnung liegt in der Flexibilität bei der Erhebung der Daten. Der Umfang der erhobenen Daten kann frei festgelegt werden. So lassen sich beispielsweise ausgehende Links mit einem Auslöser ausstatten, der beim Klick auf den Link eine Erfassung dieser Aktion auf dem Auswertungsserver gestattet. Darüber hinaus werden Auslieferung der Internetseiten und Datenerhebung vollständig entkoppelt, so dass die Sammlung der Klickstromdaten auf einem eigenen Server erfolgen kann. Weiterhin werden Zwischenspeicher umgangen, da der JavaScript-Code stets neu ausgeführt wird [vgl. Kaushik 2007b, S. 32]. Im Gegensatz zu technisch ausgereizten und nicht mehr weiterentwickelten Serverprotokollen wird die JavaScript-basierte Klickstromerfassung aktiv weiterentwickelt.

Nachteile der JavaScript-Tags

Nachteilig ist die Abhängigkeit vom Nutzersystem. Ist die Ausführung von JavaScript im Browser deaktiviert, erfolgt keine Erfassung, oder – bei Angabe eines `<noscript>`-Elementes – eine im Umfang eingeschränkte Erfassung des Zugriffs. Die Höhe des Anteils der Nutzer, die JavaScript deaktiviert haben, ist schwierig zu messen. Kaushik [2007b, S. 33] geht von einem Anteil von zwei bis sechs Prozent aus. Eine DBS-Erhebung ergab 0,78% weniger Hits bei JavaScript-basierter Erhebung. Dabei wurden in eine gewisse Anzahl von DBS-Seiten jeweils zwei transparente, 1 mal 1 Pixel große GIF-Grafiken eingebunden. Eine der beiden Dateien wurde über ein einfaches img-Element eingebunden (HTML-Standard), die andere per JavaScript. Nach einer gewissen Zeitspanne lassen sich die Zugriffe auf die Grafikdateien über die Zugriffsprotokolle vergleichen. Die Differenz ist der Saldo aus einer höheren Zu-

griffszahl durch die Umgehung von Zwischenspeichern und einer Verringerung der Zugriffszahl durch deaktiviertes JavaScript.

Da es für das Auswertungssystem von Bedeutung ist, welche Internetadresse die mit JavaScript ausgezeichnete Seite hat, muss bei mehrfachen, inhaltsgleichen Domains ein entsprechender Hinweis übergeben werden, sofern das Webanalyse-System dies unterstützt. Weiterhin sind serverseitige Weiterleitungen und Abrufe von Binärdateien schwierig zu erfassen, da diese nur indirekt per JavaScript registriert werden können, indem für alle Links mit einer bestimmten Dateiendung (z.B. PDF) ein sog. JavaScript-Event registriert wird, das bei einem Klick auf den Link eine Benachrichtigung darüber an den Webanalyse-Server absetzt.

Zwecks vollständiger Erfassung aller Zugriffe auf ein Webangebot muss sichergestellt werden, dass jede Seite innerhalb des Angebots mit dem JavaScript-Code ausgezeichnet wird. Je nach technischer Struktur des Webangebots kann dies mit einem erheblichen Aufwand verbunden sein [vgl. Hassler 2009, S. 55]. Bei der üblichen Verteilung von Seitenauslieferung und Zugriffserfassung auf verschiedene Server muss die Erreichbarkeit des Erfassungsservers sichergestellt werden. Ist der Webanalyseserver bei funktionierender Seitenauslieferung nicht erreichbar, werden die in dieser Zeit stattfindenden Seitenabrufe nicht gezählt.

Fast immer werden bei der JavaScript-basierten Datenerhebung auch Cookies eingesetzt, die im folgenden Abschnitt beschrieben werden.

4.2.3. Cookies

Bei einem Cookie[5] handelt es sich um eine kleine Textdatei, die vom Webserver an den Browser gesendet, dort gespeichert und beim Aufruf der Herausgeber-Domain (und nur dieser) wieder an den Webserver übermittelt wird [vgl. Kaushik 2007b, S. 129]. Die Funktionsweise ist in Abbildung 4.4 dargestellt. Hinsichtlich der Speicherdauer werden Sitzungs-Cookies und persistente Cookies unterschieden. Erstere bleiben nur bis zum Schließen des Browsers erhalten, letztere dagegen bis zu einem frei definierbaren Ablaufdatum.

Eine weitere Unterscheidung lässt sich anhand des ausgebenden Servers vornehmen. Wird das Cookie von dem besuchten Server (definiert per Domainnamen) ausgegeben, spricht man von einem Erstpartei-Cookie (First-Party Cookie (FPC)). Wird das

[5] Definitionsgemäß handelt es sich bei einem Cookie allgemein um einen Datenbankeintrag oder eine Datei zum Datenaustausch zwischen Computerprogrammen [vgl. Wikipedia 2009b]. Im Web-Kontext ist mit Cookie jedoch fast immer das genauer spezifizierte HTTP-Cookie gemeint [vgl. Wikipedia 2009c]. Diesem Sprachgebrauch folgend wird im weiteren Verlauf stets die Bezeichnung „Cookie“ statt „HTTP-Cookie“ verwendet.

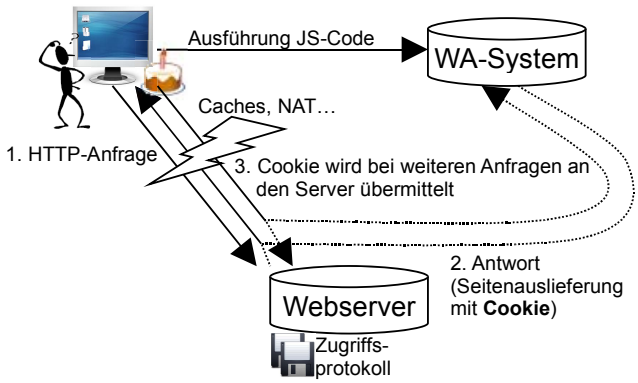


Abbildung 4.4.: Funktionsweise von Cookies

Cookie von einer anderen als der besuchten Domain gesetzt, handelt es sich um ein Drittpartei-Cookie (Third-Party Cookie (TPC)) [vgl. Peterson 2005, S. 56f.]. Letztere haben bei vielen Internetnutzern einen schlechten Ruf, da sie häufig von Werbeanbietern wie z.B. Doubleclick verwendet werden [vgl. Kaushik 2007b, S. 30]. Einer Studie zufolge liegt die Löschrate für Drittpartei-Cookies bei etwa 15%, die für Erstpartei-Cookies bei nur etwa 2% [vgl. Enge 2007a]. Eine Erhebung der Webtrekk GmbH weist für das dritte Quartal 2009 für deutsche Nutzer dagegen eine Ablehnungsrate von lediglich ca. 7,3% [vgl. Webtrekk GmbH 2009, S. 2]. Fortgeschrittene externe Webanalysesysteme wie Google Analytics verwenden Erstpartei-Cookies, auch wenn die Webanalyseanwendung selbst unter einem anderen Domainnamen läuft. Dies kann auf zwei verschiedene Arten geschehen: Entweder ändert der Seitenbetreiber die Domain Name System (DNS)-Einstellungen seiner Domain, so dass eine Weiterleitung von der eigenen Domain auf die des Webanalyse-Anbieters entsteht, oder die Webanalyse-System benutzt JavaScript zum Setzen des Cookies [vgl. Enge 2007b].

Vorteil von Cookies

Die Verwendung persistenter Erstpartei-Cookies erlaubt durch das Setzen einer anonymen, zufällig generierten Nutzerkennung im Cookie eine wesentlich zuverlässigere Identifikation zusammenhängender Seitenabrufe (und damit auch Sitzungen) und die Zählung eindeutiger Besucher als mit der bei der Logfile-Auswertung meist eingesetzten Kombination aus IP-Adresse und Benutzersystem [vgl. Peterson 2005, S. 62f.].

Nachteile von Cookies

Die Ablehnung oder Löschung von Cookies bringt für den Nutzer keinerlei Nachteile oder Funktionseinschränkungen, verringert aber die Genauigkeit der Messung

von Sitzungen, Besuchern und Eindeutigen Besuchern (vgl. Kap. 5.1). Während das Ablehnen von Cookies vom Server registriert werden kann, ist dies beim Löschen von Cookies, z.B. beim Beenden des Browsers, nicht möglich [vgl. Kaushik 2007b, S. 131].

Für sich genommen ermöglichen Cookies keine Ermittlung von Besuchszahlen, sie dienen lediglich der Unterscheidung von Browsern und davon abgeleitet Nutzern. Cookies werden also stets in Verbindung mit einer anderen Erhebungsmethode eingesetzt.

Problematisch ist, dass Cookies bei vielen Nutzern einen schlechten Ruf haben und diese durch eine manuelle Konfiguration des Browsers oder den Einsatz von Sicherheitsprogrammen auf ihrem Benutzersystem die Annahme von Cookies unterbinden können. Da Cookies stets im lokalen Speicher des Browsers gespeichert werden, können Besuche von Nutzern, die mehrere Browser einsetzen, nicht zusammengeführt werden. Die bekannte Erweiterung „CookieSafe“[6] für den Browser Firefox erlaubt die selektive Annahme von Cookies je nach Domain, wobei die Ablehnung von Cookies voreingestellt ist.

4.2.4. Weitere Erhebungsmethoden

Zählpixel (Web beacon, web bug) funktionieren ähnlich wie JavaScript-Tags, allerdings wird bei dieser Methode lediglich eine meist 1 mal 1 Pixel große, transparente (und somit unsichtbare) Grafik eingebaut, die auf einem beliebigen Server abgelegt werden kann [vgl. Kaushik 2007b, S. 28f.]. Mit Auslieferung der Grafik kann der Server versuchen, ein Cookie zu setzen. Der Umfang der erhobenen Daten ist nicht größer als der bei Logfiles, die Erhebungsmethode bietet jedoch eine größere Flexibilität hinsichtlich der Serverauswahl.

Bei so genannten Packet Sniffern (Packet Analyzer, Sniffer) handelt es sich um Server, die – ähnlich einer Firewall – zwischen die Webserver eines Angebotes und das Internet geschaltet werden und anschließend jeglichen Netzwerkverkehr protokollieren und auswerten [vgl. Hassler 2009, S. 55f.]. Bei dieser serverseitigen Erhebungsmethode können nicht mehr Daten als bei Zugriffsprotokollen erhoben werden. Dennoch ist der Aufwand für die Integration der Packet Sniffer in die Netzwerkstruktur hoch und es entsteht eine Gefahr der Nichterreichbarkeit der eigenen Webserver im Falle technischer Probleme bei den Packet Sniffern. Aus diesen Gründen wird die Methode kaum eingesetzt.

[6] `https://addons.mozilla.org/de/firefox/addon/2497`; Stand: 18. Oktober 2009

4.2.5. Datenerhebung beim DBS

Beim DBS liegen Logfiles seit dem Start des Webangebots im Jahr 1996 vor. Sie weichen in ihrem Aufbau vom oben vorgestellten Combined Log Format ab. Ein Eintrag ist wie folgt aufgebaut:

```
67p4fd556db.dip.t-dialin.net - - [01/Aug/2008:00:00:04 +0200]
"GET /zeigen.html?seite=3011 HTTP/1.1" 200 16073
Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.9.0.1) Gecko/2008070208 Firefox/3.0.1
http://www.bildungsserver.de/zeigen.html?seite=1828
```

Es bestehen folgende Unterschiede zum Combined Log Format:

1. Die IP-Adresse (Feld 1) wird über einen so genannten „Reverse Lookup“ schon vom Apache-Webserver in einen Domainnamen aufgelöst[7].
2. Die Felder der verweisenden Seite und des Benutzersystems (Felder 10 und 11) sind vertauscht.
3. Die Werte dieser beiden Felder sind nicht in Anführungszeichen gesetzt.

Da die Bezeichnung des Benutzersystems beliebig viele Leerzeichen enthalten kann, fehlt eine eindeutige Trennung der Datenfelder. Sofern das gewünschte Auswertungsprogramm dieses benutzerdefinierte Format nicht unterstützt, müssen die Protokolldateien vor der Weiterverarbeitung umgeformt werden. Zusätzlich kommt seit Ende Juni 2009 das JavaScript- und Cookie-basierte Webanalyse-System „Piwik“ zum Einsatz (vgl. Kap. 5.2.2).

In diesem Kapitel wurden gängige Methoden der Erhebung von Nutzungsdaten bei Webangeboten vorgestellt. Diese Daten können auf vielfältige Weise ausgewertet werden. Die sich im Bereich der Webanalyse ergebenden Auswertugsmöglichkeiten werden im folgenden Kapitel besprochen.

[7] Dabei wird der Apache-Server über die „HostnameLookups“-Konfigurationsdirektive angewiesen, über eine Anfrage beim Domain Name System (DNS) zu versuchen, herauszufinden, welche Domain zur aktuellen IP-Adresse gehört. Diese Anfrage kann unter Umständen fehlschlagen oder fehlerhafte Ergebnisse liefern. Beispielsweise sind die IZB-Angebote DBS, Schulweb und Fachportal Pädagogik unter derselben IP-Adresse erreichbar (193.175.194.33). Ein Reverse Lookup liefert jedoch nur die Domain schulweb.de.

5. Webanalyse

Nach der bereits im vorangegangenen Kapitel erfolgten Definition von Webanalyse werden im Folgenden verbreitete Webanalyse-Metriken vorgestellt und die beim IZB eingesetzten Softwareprodukte zur Webanalyse in einer Klassifikation eingeordnet und besprochen.

5.1. Metriken der Webanalyse

Webanalyse-Metriken umfassen Nutzungshäufigkeiten auf verschiedenen Abstraktionsebenen (Hits, Seitenabrufe, Sitzungen, Besucher), eine Charakterisierung besonders kurzer Sitzungen, sowie Seitenmerkmale, weitere Metriken (z.B. Sitzungsdauer) und eine Unterscheidung der Referrer-Arten.

5.1.1. Seitenabrufe

Als Seitenabruf (Page View, Page Impression) bezeichnet man die Anzahl der erfolgreichen Abrufe von Internetseiten. Im Gegensatz zur grundlegensten aber auch am wenigsten aussagekräftigen Metrik der Zugriffe (Hits) [vgl. Hassler 2009, S. 88f.] wird hierbei die Seite als Gesamteinheit einschließlich aller eingebetteter Dateien wie CSS-Stylesheets oder Bildern betrachtet [vgl. Kaushik 2007b, S. 140]. Bei der Verwendung von Logfiles müssen dementsprechend nebensächliche Dateiabrufe sowie fehlerhafte und umgeleitete Zugriffe (Statuscode 4xx für fehlende Dateien, 5xx für Serverfehler und 3xx für Weiterleitungen) entfernt werden. Berücksichtigt werden bei der Zählung grundsätzlich nur Abrufe durch menschliche Betrachter, Zugriffe durch Webcrawler werden gesondert erfasst.

5.1.2. Sitzungen

Eine Sitzung (Session, Visit, Besuch) bezeichnet die zusammenhängende Interaktion (bestehend aus einem oder mehreren Seitenabrufen) eines Individuums innerhalb eines gewissen Zeitraums mit der Website [vgl. Hassler 2009, S. 93]. Bei ei-

ner rein Logfile-basierten Auswertung wird der Zusammenhang zwischen einzelnen Seitenabrufen wie bereits beschrieben aus verschiedenen Merkmalen eines Logfile-Eintrags hergestellt. JavaScript-Tags hingegen werden ausschließlich in Verbindung mit Cookies verwendet. Dabei dient die über das Cookie vergebene zufällige Kennung als Identifikationsmerkmal eines Besuchs. Die typische Leerlaufzeit, nach der ein Visit bei ausbleibenden Seitenabrufen auswertungstechnisch abgebrochen wird, beträgt 30 Minuten [vgl. Burby u. a. 2007, S. 8]. Ruft ein Besucher nach dieser Leerlaufzeit erneut eine Seite ab, zählt dies als neuer Besuch.

5.1.3. Besucher

Unter der Bezeichnung „Eindeutige Besucher“ (Unique Visitors) versteht man die Anzahl der Individuen, die innerhalb des definierten Betrachtungszeitraums (z.B. Tag, Woche oder Monat) eine oder mehrere Sitzungen erzeugten [vgl. Hassler 2009, S. 96]. Die Zuordnung von Sitzungen zu Personen ist schwierig und mit Unzuverlässigkeiten verbunden. Die zuverlässigste Methode ist die Zuordnung anhand einer persönlichen Anmeldung bei der Website [vgl. Burby u. a. 2007, S. 9]. Da diese Möglichkeit auf den meisten Websites nicht gegeben ist, werden in der Regel persistente Cookies zur anonymen Nutzeridentifikation verwendet. Hierbei besteht die Gefahr, dass Cookies vom Browser gar nicht angenommen oder nach einem Besuch gelöscht werden (vgl. Abschnitt 4.2.3). Bei einer Auswertung von Logfiles ohne Verwendung von Cookies wird zur Besucheridentifikation das gleiche Verfahren wie zur Sitzungsidentifikation herangezogen.

Eine weitere Unterteilung der Besucher ist möglich [vgl. Burby u. a. 2007, S. 10ff.]:

- Neue Besucher (New Visitors) besuchen die Website innerhalb des Betrachtungszeitraums zum allerersten Mal. Ein Neuer Besucher kann innerhalb dieses Zeitraums nicht gleichzeitig auch ein Wiederkehrender Besucher sein.
- Wiederholungsbesucher (Repeat Visitors) haben innerhalb des Betrachtungszeitraums zwei oder mehr Visits auf der Website. Daher ist ein Wiederholungsbesucher gleichzeitig auch ein Neuer oder Wiederkehrender Besucher.
- Wiederkehrende Besucher (Return Visitors) besuchten die Website bereits vor dem Betrachtungszeitraum mindestens einmal. Sie können daher nicht zugleich als Neue Besucher, wohl aber ggf. als Wiederholungsbesucher gezählt werden.

Die drei Metriken Seitenabrufe, Sitzungen und Besucher bauen – in dieser Reihenfolge – aufeinander auf. Dabei nimmt die Aussagekraft, aber auch die Komplexität der Metrikberechnung zu. Ein Besucher ist, ähnlich dem Leser einer Zeitschrift, geistig besser zu vergegenwärtigen als ein abstrakter Seitenabruf [vgl. Hassler 2009, S. 97].

5.1.4. Absprünge

Zur Bewertung eines Internetangebotes sind auch auffällig kurze Sitzungen von Bedeutung:

- Ein-Seiten-Besuche sind Besuche, bei denen nur eine einzige Seite einmal oder mehrmals abgerufen wurde [vgl. Burby u. a. 2007, S. 29].
- Ein Absprung (Bounce) kann auf verschiedene Weise definiert werden. Kaushik plädiert dafür, Sitzungen mit einer Dauer von weniger als zehn Sekunden als Absprung zu werten [Kaushik 2007b, S. 143]. Der WAA zufolge ist ein Absprung eine Sitzung, in der nur eine Seite aufgerufen wurde und bei der somit Einstiegs- und Ausstiegsseite identisch sind [vgl. Burby u. a. 2007, S. 30]. Entsprechend dieser Definition wird ein Absprung auch als „Single Page View Visit“ bezeichnet. Als problematisch an der Definition von Kaushik erscheint vor allem die Schwierigkeit, die Verweildauer auf der letzten Seite innerhalb einer Sitzung zu messen (vgl. Abschnitt 5.1.6). Daher wird im Folgenden die Definition der WAA verwendet.
- Setzt man die Absprünge auf einer Seite ins Verhältnis zu den Sitzungen, innerhalb der die Seite aufgerufen wurde, erhält man die Absprungrate (Bounce Rate) [vgl. Burby u. a. 2007, S. 31].

5.1.5. Seitenmerkmale

Die Einstiegsseite ist die innerhalb einer Sitzung zuerst aufgerufene Seite [vgl. Burby u. a. 2007, S. 14]. Da jeder Besuch mit einer Seite beginnt, ist die Summe der Einstiegsseiten und Besuche identisch. Eine Landeseite ist eine Seite, die speziell für den Einstieg auf der Website optimiert wurde [vgl. Burby u. a. 2007, S. 15]. Sie ist daher nicht notwendigerweise auch eine Einstiegsseite. Die Optimierung kann u.a. im Hinblick auf Schlagworte, Zielgruppen oder Handlungsaufforderungen erfolgen.

Die Ausstiegsseite ist die innerhalb einer Sitzung zuletzt aufgerufene Seite [vgl. Burby u. a. 2007, S. 16]. Da die meisten Browser mehrere Fenster oder Tabs darstellen können, lässt sich nicht eindeutig sagen, welche Seite tatsächlich zuletzt angesehen wurde. Da jeder Besuch mit irgendeiner Seite endet, ist die Summe der Ausstiegsseiten und Besuche identisch.

Das Ausstiegsverhältnis gibt das Verhältnis der Ausstiege auf einer Seite zur Häufigkeit ihres Aufrufs (Seitenabrufe) an.

5.1.6. Weitere Metriken

Eine weitere wichtige Metrik ist die Konversionsrate (Conversion rate). Gemeint ist damit das Verhältnis zwischen der Anzahl der Resultate (Outcomes) und der Anzahl der Eindeutigen Besucher [vgl. Kaushik 2007b, S. 172]. Resultat ist dabei in Abhängigkeit von den Zielen des Webangebots beispielsweise als Kauf, Newsletter-Abonnement oder Kontaktaufnahme definiert.

Die Sitzungsdauer (Besuchsdauer, Visit Duration) ergibt sich durch Subtraktion des Zeitstempels der ersten Aktivität in einer Sitzung vom Zeitstempel der letzten Aktivität. Die Ansichtsdauer der letzten Seite kann nur in bestimmten Fällen ermittelt werden, beispielsweise wenn der Aufruf externer Links per JavaScript erfasst wird und ein solcher Aufruf erfolgt [vgl. Kaushik 2007b, S. 136f.]. Wird bei einer Sitzung nur eine Seite aufgerufen, wird die Sitzungsdauer in der Regel nicht angezeigt.

Auch auf Seitenebene kann die Betrachtungsdauer ermittelt werden [vgl. Hassler 2009, S. 177]. Wie auch bei der Sitzungsdauer ist statt einer Durchschnittsangabe die Verteilung auf bestimmte Zeitklassen (z.B. 0-10, 11-30 Sekunden usw.) interessant.

Die Anzahl der Klicks auf einen bestimmten Link wird als Durchklick (Click-through) bezeichnet. Setzt man diesen Wert ins Verhältnis zu den Abrufen der Seite, die den Link aufweist, erhält man die Durchklickrate (Click-through rate).

Ein weiteres häufig verwendetes Verhältnis ist das der Seitenabrufe pro Besuch, auch als Besuchstiefe bezeichnet [vgl. Hassler 2009, S. 180f.].

5.1.7. Verweisende Seiten

Wird eine URL über einen Link auf einer Internetseite aufgerufen, übermittelt der Browser die URL der verweisenden Seite an den Webserver der aufgerufenen Seite. Der Server speichert die URL der verweisenden Seite, auch als „Referrer“ bezeichnet, meist im Abrufprotokoll im Eintrag der aufgerufenen Seite (vgl. Abschnitt 4.2.1). Wird die Seite durch direkte Eingabe der URL in die Adresszeile, über ein Lesezeichen oder über ein anderes Anwendungsprogramm, beispielsweise eine E-Mail-Software, aufgerufen, liegt keine Referrer-Angabe vor [vgl. Hassler 2009, S. 107f.].

Referrer lassen sich zunächst anhand ihrer Zugehörigkeit zur betrachteten Website unterscheiden [vgl. Burby u. a. 2007, S. 18–21]. Gehört die Referrer-Domain zur betrachteten Website, handelt es sich um einen internen Verweis. Bei externen Referrern wird unterschieden zwischen Suchmaschinen und sonstigen Websites.

Aus dem Vorliegen und Nichtvorliegen eines Referrers beim ersten Seitenabruf einer Sitzung lassen können drei Arten unterschieden werden, über die der Besucher zur Website gelangte:

1. Direkteinstiege, wenn keine Referrer-Angabe vorliegt
2. Suchmaschineneinstiege, wenn der externe Referrer einer Websuchmaschine zugeordnet werden kann
3. Einstiege über sonstige Verweise, wenn ein externer Referrer vorliegt, der keiner Suchmaschine entspricht

5.1.8. Anwendung der Metriken zur Schwachstellenanalyse

Die vorgestellten Metriken sind interpretationsbedürftig und stets mit den Zielen der Website-Betreiber abzugleichen. Auch ist die Charakteristik des Angebotes zu berücksichtigen. So unterscheidet sich beispielsweise die Seitenstruktur eines Blogs deutlich von der eines breit angelegten Portals wie dem DBS dahingehend, dass bei Ersterem alle aktuellen Beiträge bereits auf der Startseite enthalten sind und weitere Seiten meist lediglich das Archiv oder ergänzende Informationen bereitstellen [vgl. Wikipedia 2009a].

Die Häufigkeit von Seitenabrufen, Sitzungen und Besuchern gibt einen Eindruck von der Beliebtheit einzelner Seiten oder Seitenbereiche. Über die Zeit hinweg abnehmende Nutzungszahlen können dabei ein Ansatzpunkt für Verbesserungen sein, ebenso wie im Verhältnis zum Gesamtangebot schwach genutzte Seiten [vgl. Hassler 2009, S. 93]. Zu beachten ist allerdings, dass diese Metriken stark von externen Einflüssen wie Werbekampagnen oder dem Rangplatz in Suchmaschinenergebnisseiten geprägt sind [vgl. Hassler 2009, S. 89].

Ein Seitenausstieg ist nicht grundsätzlich negativ zu deuten, schließlich endet jede Sitzung zwangsläufig an einer Stelle innerhalb der Website. Auch sind unterschiedliche Nutzerintentionen zu berücksichtigen, denn nicht jeder Besucher eines Einkaufportals möchte tatsächlich etwas kaufen [vgl. Kaushik 2007b, S. 153]. Negativ sind dennoch Ausstiege auf Seiten, für die dies nicht beabsichtigt ist, z.B. Seiten des Bestellprozesses bei Internetkaufhäusern. Im konkreten Beispiel könnte dies auf eine schlechte Benutzerführung durch den Bestellprozess sprechen.

Auf Grundlage der Website-internen Referrer lassen sich Pfadanalysen erstellen, bei denen Seitenabfolgen gruppiert werden [vgl. Hassler 2009, S. 189f.]. Auch hier lässt sich durch einen Abgleich mit der intendierten Nutzung feststellen, ob die Benutzer im Webangebot so navigieren, wie geplant oder erwünscht.

Nicht direkt eine Metrik, aber auch URL-Parameter können zur Identifikation von Schwachstellen herangezogen werden. Ein Beispiel hierfür sind Suchphrasen, mit denen Besucher über externe Suchdienste auf das eigene Webangebot gelangen, aber auch der angebotsinternen Suche [vgl. Hassler 2009, S. 196ff.]. Voraussetzung hierfür ist die Verwendung von GET-Parametern[1]. Derartige Suchphrasen liefern einen direkten Hinweis auf das Informationsbedürfnis des Nutzers. Anzahl, Art und Aussagekraft solcher Parameter hängen von der individuellen technischen Umsetzung einer Website ab.

Zur Durchführung von Webanalysen wird meist Standardsoftware eingesetzt. Eigenschaften derartiger Programme werden im nächsten Abschnitt beschrieben.

5.2. Webanalyse-Software

Zahlreiche Softwareprodukte sind im Bereich Webanalyse angesiedelt. Zunächst soll daher eine Klassifikation zu deren Einordnung vorgestellt werden. Anschließend werden zwei beim IZB eingesetzte Produkte vorgestellt.

5.2.1. Klassifikation von Webanalyse-Software

Als Facetten einer Klassifikation von Webanalyse-Software eignen sich der Speicherort der Daten, die unterstützten Datenerhebungsmethoden, der Funktionsumfang und schließlich der Softwarehersteller.

Speicherort der Daten

Die Datenanalyse und -speicherung kann entweder bei der Institution, die die Webanalyse-Software einsetzt, oder bei einem Anbieter von Webanalyse-Dienstleistungen erfolgen [vgl. Hassler 2009, S. 64]. Die erste Variante wird auch als „In-house“-Variante, die Zweite als SaaS- oder „Hosted“-Lösung bezeichnet.

Ein Vorteil der hausinternen Speicherung liegt in der rechtlichen und physischen Verfügungsgewalt über die Daten [vgl. Hassler 2009, S. 64]. Im Gegensatz dazu begibt sich der Nutzer eines SaaS-Produktes in eine Abhängigkeit vom Anbieter [vgl. Hassler

[1] Zur Datenübertragung vom Browser an den Server sieht das Übertragungsprotokoll HTTP zwei Wege vor: Daten können, abgetrennt durch ein Fragezeichen, als Bestandteil der URL (GET-Methode) oder in den Kopfdaten des Protokolls (POST-Methode) übertragen werden [vgl. Wikipedia 2009d]. GET-Daten werden als URL-Bestandteil in der Protokolldatei erfasst.

2009, S. 67]. Serverausfälle oder gar die Zahlungsunfähigkeit des Anbieters können den Nutzen einer Mietlösung enorm beeinträchtigen.

Nachteilig bei einer hausinternen Lösung sind die Kosten für Hard- und Software [vgl. Hassler 2009, S. 65f.]. Die Hardwareanforderungen hängen von der Datenerhebungsmethode und dem Funktionsumfang der Software ab. So muss bei einer JavaScript-basierten Erhebung der Auswertungsserver stets verfügbar sein, während die rechenintensive Logfile-Auswertung in ansonsten lastarme Nachtstunden verlegt werden kann. Auch ist der Aufwand für Installation und Implementierung zu berücksichtigen. Ein In-house-System muss zeitintensiv installiert und konfiguriert werden, während bei der Nutzung eines Dienstleisters meist lediglich der entsprechende JavaScript-Code in das eigene Webangebot eingebettet werden muss [vgl. Hassler 2009, S. 67f.]. Bei der Softwarelizenzierung entstehen meist hohe Einmalkosten für den internen Einsatz gegenüber deutlich geringeren monatlichen Kosten für die Anmietung einer SaaS-Lösung. Abhängig vom Besucheraufkommen kann die hausinterne Lösung längerfristig jedoch günstiger sein, da die laufenden Kosten für Mietprodukte von der Anzahl der Seitenabrufe abhängen und bei vielbesuchten Webangeboten entsprechend steigen.

Datenerhebungsmethoden

Als Datenerhebungsmethoden werden, wie bereits in Abschnitt 4.2 beschrieben, die beiden Wichtigsten – Zugriffsprotokolle und JavaScript-Tags – berücksichtigt. Software zur Logfileauswertung wird ausschließlich zur lokalen hausinternen Installation angeboten [vgl. Hassler 2009, S. 38]. JavaScript-basierte Auswertungsprogramme werden dagegen überwiegend als SaaS-Produkt angeboten; nur wenige, meist im oberen Preisbereich angesiedelte Produkte sind zusätzlich oder ausschließlich für lokale Installationen verfügbar. Einige Programme bieten beide Erhebungsmethoden an, dies jedoch stets nur alternativ, da eine zuverlässige Zusammenführung von Tag- und Protokolldatei-basierten Zugriffsaufzeichnungen aufgrund fehlender eindeutig übereinstimmender Merkmale nicht möglich ist.

Funktionsumfang

Hinsichtlich des Funktionsumfangs gibt es keine klare und etablierte Unterscheidung zur Klassifizierung von verschiedenen Lösungen. Mögliche Merkmale für eine Einordnung sind die Anzahl und Komplexität der bereitgestellten Auswertungen, die Unterstützung einer getrennten Auswertung mehrerer Webangebote, die Möglichkeit

einer verteilten Nutzung (Client/Server-Architektur) und der Grad der Anpassbarkeit der Berichte. Diese Kriterien zieht beispielsweise die Produktübersicht von Ideal Observer[2] zum Vergleich heran.

Als objektives Unterscheidungskriterium können die Lizenzierungskosten verwendet werden. Ideal Observer nimmt eine Aufteilung in drei Preisgruppen vor: Bis zu 100 Euro pro Jahr für Einsteigerprodukte, bis zu 1.000 Euro pro Monat für Produkte der Mittelklasse und monatliche Kosten von über 1.000 Euro für Produkte der Premiumklasse. Ein eindeutiger Zusammenhang zwischen Kosten und Funktionsumfang besteht allerdings nicht, da beispielsweise das kostenlose Google Analytics umfangreiche Funktionen bietet und bei o.g. Vergleich dementsprechend der Mittelklasse zuzuordnen ist.

Softwarehersteller

Als vierte Klassifikationsfacette kann der Softwarehersteller angesehen werden. Hier ist lediglich zwischen einer aus Anwendersicht externen und internen Erstellung zu unterscheiden [vgl. Peterson 2005, S. 13f.]. Aufgrund der Vielfalt bereits auf dem Markt erhältlicher Produkte ist das Programmieren einer neuen Auswertungslösung nur bei hochspeziellen Anforderungen sinnvoll. Dennoch bietet z.B. Peterson eine JavaScript- und Perl-basierte Lösung zur Erstellung eigener Protokolldateien [Peterson 2005, S. 40–45].

5.2.2. Webanalyse-Software im IZB

Im Rahmen der Webanalyse der verschiedenen vom IZB betriebenen Webangebote kommen mit Urchin und Piwik zwei Softwareprodukte zum Einsatz, die nachfolgend vorgestellt werden.

Urchin

Zur Auswertung der Logfiles wird beim IZB die Software Urchin verwendet.

Urchin wird lokal beim Website-Betreiber installiert. Die von der Software erzeugten Berichte werden im Installationsverzeichnis gespeichert, die Datenhaltung erfolgt also rein lokal.

Als Datenerhebungsverfahren können reine Logfiles oder eine Urchin-spezifische Kombination aus JavaScript, Cookies und Logfiles, dem so genannten Urchin Traffic

[2] `http://idealobserver.com/web-analytics-tools/` [Stand: 13.10.2009]

Monitor (UTM), eingesetzt werden. Beim UTM-Verfahren wird eine JavaScript-Datei in jede Seite des Webangebots eingebettet. Durch Ausführung dieses Java Scripts wird bei jedem Seitenabruf der Abruf eines Zählpixels erzeugt und eine Cookie-ID generiert. Der Zählpixel-Abruf und der Cookie-Wert wiederum werden im Logfile eingetragen und von Urchin ausgewertet. Beim IZB werden jedoch aufgrund ihrer historischen Verfügbarkeit für den DBS ausschließlich konventionelle Logfiles verwendet.

Mit der Möglichkeit, verschiedene Websites in eigenen Berichtsprofilen auszuwerten, einem großen Umfang an Berichten und einem Nettopreis von über 2.000 Euro ist Urchin hinsichtlich des Funktionsumfangs im mittleren bis gehobenen Bereich anzusiedeln.

Urchin ist ein proprietäres Produkt von Google[3], das über autorisierte Vertriebspartner verkauft wird. Anpassungen am Quellcode kann der Anwender nicht vornehmen.

Piwik

Zusätzlich zu Urchin wird seit Ende Juni 2009 mit Piwik ein Webanalyse-System im IZB eingesetzt.

Einordnung in der Klassifikation

Wie Urchin ist Piwik kein SaaS-Produkt, sondern wird lokal beim Betreiber des Webangebots installiert. Die Datenerfassung erfolgt ausschließlich über die Einbindung von JavaScript-Tags in jede Seite des Webangebots und die Verwendung von Cookies [vgl. Piwik Developers 2009b]. Logfiles können weder direkt verarbeitet noch konvertiert werden.

Piwik ist kostenlos verfügbar. Der Berichtsumfang ist gegenüber Urchin deutlich geringer. Demnach ist der Funktionsumfang im unteren bis mittleren Bereich einzuordnen.

Die Auswertungssoftware ist unter der GNU General Public License (GPL) veröffentlicht und somit quelloffen. Ziel der vier Hauptentwickler ist die Schaffung einer offenen und lokal installierbaren Alternative zu Google Analytics.

[3] Das Unternehmen Urchin wurde 2005 von Google aufgekauft, das gleichnamige Softwareprodukt diente anschließend als Grundlage für das SaaS- und JavaScript-basierte Google Analytics, wird aber nach wie vor auch als eigenständiges Programm zur lokalen Installation vertrieben [vgl. heise online 2005].

Funktionsweise

Die verwendeten Basistechnologien sind PHP und MySQL für das Auswertungssystem sowie HTML und JavaScript für Code zur Zugriffserfassung („Tracking-Code“). Des weiteren macht Piwik vom Zend Framework, einer Rahmenumgebung für PHP-Webapplikationen, Gebrauch. Zur Zeit befindet sich Piwik offiziell noch im Beta-Entwicklungsstadium, es wird jedoch bereits verbreitet produktiv eingesetzt.

Der beim DBS implementierte Tracking-Code ist im Anhang auf Seite A3 dargestellt. Er besteht aus einem JavaScript- und einem HTML-Teil. Der JavaScript-Code innerhalb der beiden `<script>`-Elemente (Zeilen 3–4 und 6–15) weist den Browser an, die Funktionsbibliothek `piwik.php` aus der Piwik-Installation auf dem Server `http://analyse.dipf.de/` zu laden. Anschließend wird die Funktion `trackPageView()` ausgeführt, die über eine Reihe weiterer interner Funktionen in der Funktionsbibliothek für eine Erfassung des aktuellen Seitenaufrufs in der Piwik-Installation sorgt. Die zweite Funktion `enableLinkTracking()` ergänzt bestimmte Links (`<a>`-Elemente) in der aktuellen Seite um ein JavaScript-Ereignis, um Klicks auf diese Links ebenfalls zu erfassen. Manipuliert werden in der Standardkonfiguration jene Links, die von der aktuellen Website wegführen, und solche, die auf Binärdateien führen. Die Funktionsbibliothek enthält eine Liste von rund 50 Dateiendungen, die als Downloads erfasst werden. Diese Liste kann im Tracking-Code angepasst werden. Eine Besonderheit dieser Piwik-Implementierung ist die Ausführung eines zweiten Tracking-Vorgangs (Zeilen 11–13). Auf den Grund hierfür wird später eingegangen.

Der HTML-Code im `<noscript>`-Element wird vom Browser nur dann verarbeitet, wenn die Ausführung von JavaScript deaktiviert ist. Es handelt sich somit um eine Rückfalloption. Sie stellt sicher, dass in jedem Fall eine Erfassung des Seitenabrufs stattfindet. Der Browser wird über eine Grafikeinbindung angewiesen, die PHP-Datei mit den Tracking-Funktionen abzurufen. Rückgabewert ist eine 1 mal 1 Pixel große Grafikdatei, die die Seitendarstellung nicht beeinträchtigt. Eine weitergehende Datenerfassung, beispielsweise detaillierterer Angaben zur Browser-Konfiguration oder der Klicks auf Links, ist in diesem Fall nicht möglich.

Das Datenbankmodell von Piwik ist im Anhang auf Seite A4 abgebildet. Die Klickstromdaten werden in den Tabellen im Bereich „Statistics logger“ abgelegt. Für jede von Piwik erfasste Aktion wird in der Tabelle `piwik_log_action` einmalig ein Eintrag angelegt. Drei Typen von Aktionen werden unterschieden: interne Seitenabrufe, Klicks auf zu externen Websites führende Links und Klicks auf Binärdateien („Downloads“). Neben der fortlaufenden Eintragsnummer `idaction` enthält die Tabelle die Felder `name` für den Aktionsnamen, wobei alternativ zur standardmäßig verwendeten URL im Tracking-Code eine andere Bezeichnung angegeben werden kann, und die

Art der Aktion `type`, die die Werte 1 für interne Links, 2 für externe Links und 3 für Datei-Downloads annehmen kann. Die Tabelle `piwik_log_visit` enthält zahlreiche Angaben zur Sitzung, darunter eine fortlaufende Sitzungsnummer, Zeitangaben, Angaben zur ersten und letzten sowie zur Anzahl der durchgeführten Aktionen, zur verweisenden Seite, zur Browser- und Betriebssystemkonfiguration und zur Herkunft der Zugriffe. In der Tabelle `piwik_log_action` schließlich werden die Datensätze der beiden zuvor beschriebenen Tabellen zusammengeführt: Für jede Aktion wird ein Eintrag angelegt, der die Sitzungsnummer (`idvisit`), die Aktionsnummer (`idaction`) und, falls vorhanden, die Nummer der verweisenden Aktion (`idaction_ref`) sowie die mit der verweisenden Aktion verbrachte Zeit (`time_spent_ref_action`) enthält. Jeder Eintrag wird von der fortlaufenden Identifikationsnummer `idlink_va` eineindeutig identifiziert. Die Tabellen 12.1, 12.2 und 12.3 im Anhang ab Seite A5 enthalten Beispieldatensätze der drei MySQL-Tabellen zur Abbildung einer Nutzersitzung.

Die Berichte werden in der Piwik-Standardkonfiguration ad hoc beim Aufruf einer Auswertungsseite generiert. Da dies bei zugriffsstarken Websites zu langen Wartezeiten führt, steht alternativ ein PHP-Skript zur Verfügung, das automatisiert in einem frei definierbaren Rhythmus ausgeführt werden kann und dann die Berichtskompilierung vornimmt. Die Berichte werden in der Datenbank in den beiden Tabellen im Bereich „Archived data“ abgelegt. Um die zukünftige Verfügbarkeit der Primärdaten in den „Statistics logger“-Tabellen auch für den Fall einer partiellen Löschung durch das Archiv-Skript sicherzustellen, wird die Datenerfassung mittels der beiden Tracking-Anweisungen im Tracking-Code in zwei Piwik-Installationen (in den Ordnern `/wasystem1/` und `/wasystem2/`) gespiegelt. Bei der zweiten Installation ist die Datenarchivierung deaktiviert.

5.3. Datenschutzaspekte

Beim DIPF ist das Hessische Datenschutzgesetz (HDSG) anzuwenden (§ 3 I S. 1 HDSG). Personenbezogene Daten sind „Einzelangaben über persönliche und sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person.“ (§ 2 I HDSG). Umstritten ist, ob auch die bei der Webanalyse zwangsläufig erhobene IP-Adresse dazu zählt. Das Amtsgericht Berlin-Mitte untersagte in einem vielbeachteten Urteil dem Bundesjustizministerium zunächst die Speicherung sämtlicher typischerweise in Logfiles erfassten Daten über den Nutzungszeitraum hinaus (AG Berlin, Az. 5 C 314/06)[4]. In der Begründung heißt es „Nach zutreffender Ansicht sind IP-

[4] `http://www.gerichtsentscheidungen.berlin-brandenburg.de/jportal/?quelle=jlink&docid=JURE070116976&psml=sammlung.psml&max=true&bs=10`; Stand: 13. Oktober 2009

Adressen personenbezogene Daten [...]“. In einem Berufungsurteil wurde das Speicherverbot auf die IP-Adresse beschränkt (LG Berlin, Az. 23 S 3/07)[5]. Insgesamt betrachtet gibt es einen Trend, die IP-Adresse als personenbezogenes Datum anzusehen [vgl. Ulbricht 2007].

Dadurch ergeben sich für das Erheben, Speichern und Verarbeiten von IP-Adressen besondere Einschränkungen nach HDSG und Telemediengesetz (TMG). Die bei der Bereitstellung der Internetseiten erhobenen Daten dürfen nur verarbeitet werden, wenn dies per Gesetz vorgesehen ist oder der Nutzer eingewilligt hat (§ 12 II TMG). Die Einwilligung muss „bewusst und eindeutig“ erteilt und protokolliert werden; sie ist außerdem widerrufbar (§ 13 II TMG).

Bei der Einbindung eines externen Webanalyse-Dienstes, sei er kostenlos wie Google Analytics oder kostenpflichtig, würde in jedem Fall mit der Anfrage des Browsers an den Dienst die IP-Adresse übermittelt. Dies könnte als Übermittlung personenbezogener Daten an Dritte angesehen werden, auch wenn die Anfrage beim Webanalyse-Server HTTP-technisch vom Nutzer selbst ausgeht. Bei amerikanischen Anbietern liegt außerdem eine Übermittlung personenbezogener Daten an Stellen außerhalb des EU-Datenschutzraums vor[6]. Auch wenn Google IP-Adressen nicht für personenbezogen hält [vgl. Whitten 2008] und die Verwendung von Webanalyse-Paketen in SaaS-Form weitverbreitet ist, liegt dieses Vorgehen juristisch allerdings in einer Grauzone.

Durch die hausinterne Datenspeicherung und -verarbeitung findet beim DIPF keine Weitergabe potentiell personenbezogener Nutzungsdaten an Dritte statt. Die IP-Adresse ist Bestandteil der Datenerhebung durch Piwik. Sie wird nächtlich automatisiert bei den Sitzungen des Vortags gelöscht. Diese Verzögerung ist notwendig, um Piwik im Fall eine Cookie-Ablehnung den Rückgriff auf die Merkmalskombination von IP-Adresse und Browserkennung zur Sitzungserkennung zu ermöglichen.

[5] `http://www.gerichtsentscheidungen.berlin-brandenburg.de/jportal/?quelle=jlink&docid=JURE070115765&psml=sammlung.psml&max=true&bs=10`; Stand: 13. Oktober 2009

[6] Viele Anbieter sind allerdings dem „Safe Harbor“-Abkommen beigetreten, nach dem sie sich zum Datenschutz gemäß EU-Recht verpflichten.

6. Web Usage Mining

Nachdem die Definition von Web Usage Mining bereits unter 4.1.2 erfolgte, werden in diesem Kapitel zunächst die allgemeine Vorgehensweise beim Web Usage Mining und anschließend fünf Auswertungsansätze beschrieben.

6.1. Vorgehensweise beim Web Usage Mining

Abbildung 6.1 zeigt die allgemeine Vorgehensweise beim Web Usage Mining. Die ersten beiden Bereiche werden in diesem Abschnitt beschrieben.

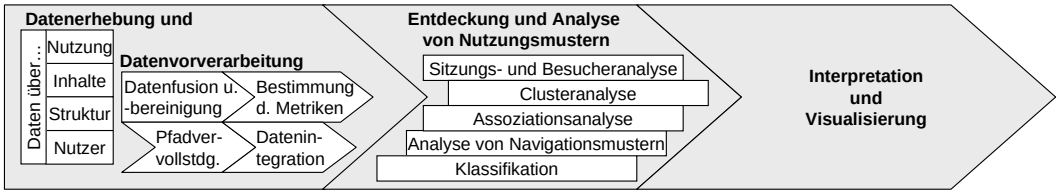


Abbildung 6.1.: Ablauf des Web Usage Mining [eigene Darst. nach Mobasher 2007]

6.1.1. Datenerhebung und -vorverarbeitung

Der erste Schritt beim Web Usage Mining besteht in der Datenerhebung und Datenvorverarbeitung [vgl. Mobasher 2007, S. 450f.].

Wichtigste Datenquelle für die Datenerhebung sind Nutzungsdaten (vgl. Kap. 4.2). Aus diesen Daten können je nach gewünschtem Aggregationsgrad Seitenabrufe oder Sitzungen extrahiert werden. Je nach Auswertungsziel können die Nutzungsdaten um weitere Datenquellen ergänzt werden [vgl. Mobasher 2007, S. 454]. Eine mögliche Datenquelle sind Inhaltsdaten. Bei Inhaltsdaten handelt es sich um die an die Nutzer ausgelieferten Texte bzw. Medienobjekte und dazugehörige semantische oder strukturelle Metadaten. Beispiele für solche Metadaten sind Schlagwörter oder Dokumentattribute. Ontologien können beispielsweise in Form von RDF-Tripeln[1] Aussagen

[1] Mit dem Resource Description Framework (RDF) lassen sich semantische Aussagen als gerichtete Graphen bestehend aus zwei Knoten und einer Verbindung beschreiben [vgl. Hitzler u. a. 2008,

über das zugrunde liegende Themengebiet liefern. Eine weitere Datenquelle sind Strukturdaten, die die Sicht des Angebotsbetreibers auf die Organisation des Webangebots abbilden. Betrachtungsebenen können die Verlinkungen innerhalb einer Seite oder die seitenübergreifenden Verlinkungen innerhalb der Website sein. Solche Baumstrukturen können in HTML oder XML repräsentiert werden. Neben Inhalts- und Strukturdaten können die Nutzungsdaten auch um Nutzerdaten ergänzt werden. Dazu zählen demographische Daten über registrierte Nutzer, explizite Interessensbekundungen oder Bewertungen und vergangene Sitzungsverläufe oder Produktkäufe. Einige dieser Daten können anonym erfasst werden, sofern beispielsweise mittels Cookies zwischen Nutzern unterschieden werden kann.

Die Datenfusion und -bereinigung ist der erste Arbeitsschritt bei der Datenvorverarbeitung [vgl. Mobasher 2007, S. 455f.]. Ein Zusammenführen von Nutzungsdaten ist häufig bei der Verwendung von Logfiles zugriffsstarker Webangebote notwendig, da der Betrieb solcher Websites zwecks Lastverteilung meist auf mehrere Server verteilt ist. Im Zuge der Datenbereinigung werden Zugriffe auf nebensächliche Dateien und Robot-Zugriffe entfernt. Da sich Robots als gewöhnliche Browser ausgeben können, ist ggf. die Anwendung von Heuristiken zur Erkennung maschineller Abrufe notwendig.

Im zweiten Schritt werden die Analyseobjekte in den Daten identifiziert. Je nach Auswertungsziel kann es sich dabei um Seitenabrufe, Sitzungen oder Besucher handeln. Definition und Bestimmung unterscheiden sich nicht von denen bei der Webanalyse (vgl. Abschn. 5.1.1 bis 5.1.3).

Aus den in Kapitel 4.2 geschilderten Unzulänglichkeiten der unterschiedlichen Erhebungsmethoden resultieren unvollständige Nutzungsdaten. Im Rahmen der Datenvorverarbeitung kann daher versucht werden, Nutzungspfade zu vervollständigen [vgl. Mobasher 2007, S. 460f.]. Dazu können Referrer- und Seitenstruktur-Informationen verwendet werden. Häufig ergeben sich mehrere Seitenkandidaten für eine Vervollständigung. In diesem Fall wird normalerweise der kürzeste Pfad eingesetzt. Als Beispiel für eine Methode zur Pfadvervollständigung soll hier der Pattern-Restore-Method-Algorithmus (PRM-Algorithmus) von Ting u. a. beschrieben werden. Dabei bildet $C_i = [U_i, T_i, L_i, R_i, Mark]$ einen Seitenabruf ab. U_i entspricht der IP-Adresse, T_i dem Zeitstempel des Aufrufs, L_i der abgerufenen URL, R_i der Referrer-URL und $Mark$ einem Unterscheidungsmerkmal zwischen unveränderten und ergänzten Klickstrom-Einträgen. Zunächst wird überprüft, ob zum ersten Seitenaufruf in einer Sitzung eine Referrer-Angabe vorhanden ist (vgl. Abbildung 6.2). Ist dies der Fall, wird die Referrer-URL auf ihre Zugehörigkeit zur untersuchten Website hin

S. 36–40]. RDF-Tripel sind verschriftlichte Graphen in der Form Subjekt, Prädikat, Objekt. Ein Beispiel lautet: *BuchtitelX verlegt_bei VerlagY*. Die Notation kann in XML erfolgen.

untersucht. Handelt es sich um einen internen Referrer, muss eine Seitenansicht vor dem ersten protokollierten Seitenabruf stattgefunden haben. Dieser Eintrag wird ergänzt mit einer angenommenen Zeitdifferenz von fünf Sekunden vor dem ersten protokollierten Seitenabruf und der Referrer-URL als abgerufener URL. Für jeden Seitenabruf wird weiterhin überprüft, ob es sich um den zweiten oder späteren Abruf innerhalb der jeweiligen Sitzung handelt. Ist diese Bedingung erfüllt, wird die Referrer-URL des überprüften Seitenabrufs mit der Seiten-URL des vorhergehenden Eintrags verglichen. Stimmen sie nicht überein, wird die Referrer-URL als Seitenabruf ergänzt. Als Zeitstempel dient der Mittelwert der beiden umgebenden Seitenabrufe.

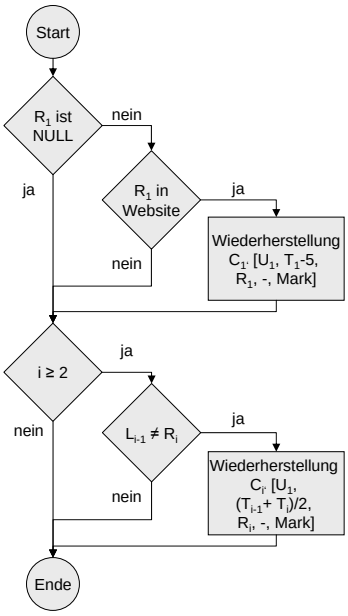


Abbildung 6.2.: PRM-Algorithmus [eigene Darst. nach Ting u. a. 2005, S. 506]

Ein weiterer Schritt der Datenvorverarbeitung kann die Integration von Daten aus den unterschiedlichen, oben beschriebenen Quellen sein [vgl. Mobasher 2007, S. 461f.]. Besonders wichtig ist eine solche Datenintegration im elektronischen Handel. Dazu werden die Daten meist in einem Data Warehouse abgelegt und per Online Analytical Processing (OLAP) analysiert und visualisiert. OLAP ist eine Methode, die Datenanalytiker bei der Abfrage großer Datenmengen in operationalen Datenbanken oder Data Warehouses unterstützt [vgl. Wikipedia 2009e].

6.1.2. Entdeckung und Analyse von Nutzungsmustern

Die Sitzungs- und Besucheranalyse umfasst eine Ermittlung der unter 5.1 beschriebenen Webanalyse-Metriken [vgl. Mobasher 2007, S. 466f.]. Erweiterte Auswertungs-

möglichkeiten entstehen durch die Nutzung von OLAP, beispielsweise in Form von OLAP-Würfeln.

Beim Clustern werden Gruppen von Objekten mit ähnlichen Eigenschaften gebildet. Gruppiert werden können Besucher und Seiten [vgl. Mobasher 2007, S. 467]. Besucher werden dabei nach der Ähnlichkeit ihrer Nutzungsmuster zusammengefasst. Die Ergebnisse eines solchen Clusterings können u.a. zur Personalisierung einer Website verwendet werden. Seiten können ebenfalls auf Grundlage der Nutzungsdaten oder unter Verwendung von Metadaten gruppiert werden. Die resultierenden Seitengruppen fassen je nach Angebotstyp Seiten zusammen, die beispielsweise von denselben Produkten oder Themen handeln.

Durch die Assoziationsanalyse lassen sich Seiten oder Produkte identifizieren, die häufig gemeinsam besucht oder gekauft werden. Solche Zusammenhänge lassen sich durch assoziative Regeln ausdrücken, deren allgemeine Form $W \Rightarrow B(\gamma, \sigma)$ lautet [vgl. Ferber 2003, S. 163ff.]. W und B stehen für diffenzierte Untermengen einer Menge von binären Attributen, σ für einen support threshold und γ für einen confidence threshold. Die Beispielregel

$$\text{Besuch von Seite a.html} \Rightarrow \text{Besuch von Seite d.html } (0,8,\ 0,5)$$

besagt, dass 80% der Besucher von Seite a.html auch Seite d.html besucht haben und insgesamt 50% aller Besucher beide Seiten besucht haben. Um die Bestimmung von Regeln in großen Datenmengen zu vereinfachen, kann Wissen über das Themengebiet in Form von Teilmengen eingebracht werden. Ein Online-Händler könnte z.B. Teilmengen für Produktseiten für Bücher und für Tonträger bilden und die Nutzungszusammenhänge zwischen diesen Teilmengen untersuchen. Verwendet werden können die so gefundenen Regeln für Empfehlungen.

Ein weiterer Analysegegenstand beim Web Usage Mining kann die Untersuchung sequentieller Bewegungsmuster in Webangeboten sein. Ziel hierbei ist die Identifizierung sitzungsübergreifend gültiger, aufeinanderfolgender Seitenaufrufe [vgl. Mobasher 2007, S. 475f.]. Solche Muster können z.B. im Marketing dazu verwendet werden, das Nutzerverhalten vorherzusagen und passende Werbeanzeigen einzublenden.

Als fünfter und letzter Teilbereich der Datenanalyse ist die Klassifikation von Seitenabrufen oder Sitzungen zu nennen. Voraussetzung hierfür sind bestehende Kategorien und ihre Definition mittels Eigenschaften [vgl. Mobasher 2007, S. 479].

6.2. Auswertungsansätze

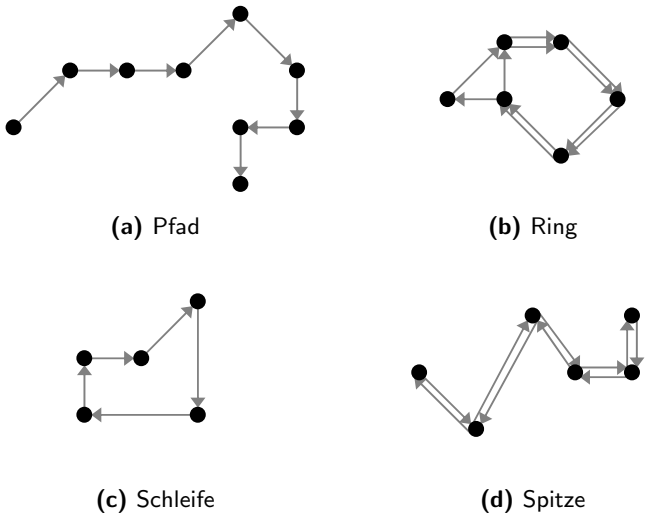
Im Rahmen einer Literaturrecherche zu Ansätzen des Web Usage Mining wurden verschiedene Anwendungskonzepte ermittelt, die Web Usage Mining-Methoden zur Identifizierung von Nutzungsmustern verwenden.

6.2.1. Charakterisierung von Nutzernavigation in komplexen Datenstrukturen nach Canter u. a.

Canter u. a. beschreiben vier Bewegungsmuster von Nutzern in Webangeboten und ihre Bedeutungen. Ausgangspunkt für die Identifikation dieser Muster waren Handlungen des täglichen Lebens, wie das morgendliche Verlassen und abendliche Zurückkehren in die eigene Wohnung oder das Verbinden mehrerer Erledigungen zu einer Wegstrecke [vgl. Canter u. a. 1985, S. 93f.]. Die Anwendbarkeit dieser Muster auf Hypertexte wurde von Canter u. a. in Experimenten nachgewiesen. Die vier Muster werden nachfolgend beschrieben und in Abbildung 6.3 grafisch dargestellt.

- Als Pfad („path“) wird der Weg durch ein Webangebot beschrieben, bei dem kein Knoten (Seite) zwei Mal aufgerufen wird. Auf reale Fortbewegungen bezogen lässt sich dieses Muster als eine Erkundung oder Reise bezeichnen, die endet, wenn ein zuvor bereits besuchter Punkt erreicht wird.
- Bei einem Ring („ring“) entspricht der Endpunkt dem Startpunkt. Ein Ring kann dabei weitere Ringe enthalten. Als Beispiel hierfür seien Hin- und Rückfahrt zwischen Wohnung und Arbeitsplatz genannt, auf denen Besorgungen erledigt werden und als innerer Ring ein mittaglicher Hin- und Rückweg zwischen Arbeitsplatz und einem Supermarkt.
- Eine Schleife („loop“) entspricht einem Ring, der keine weiteren Ringe enthält.
- Eine Spitze („spike“) ist eine Route durch das Webangebot mit einem Rückweg, der exakt dem Einstieg folgt. Als reales Beispiel hierfür kann eine Autofahrt ohne Navigationsgerät in eine dem Fahrer unbekannte Stadt angesehen werden, wobei auf der Rückfahrt die Einfahrtsstrecke zurückverfolgt wird.

Unter Verwendung von Pfadalgebra bilden die Autoren für jedes der vier Muster einen mathematisch nicht näher definierten Index (Pathiness, Ringiness, Loopiness und Spikiness). Zwei weitere Indizes zur Charakterisierung des Nutzerverhaltens werden gebildet: NV/NT gibt das Verhältnis wieder zwischen der Anzahl der besuchten Knoten (NV) und der Gesamtzahl aller im Hypertextsystem vorhandenen Knoten (NT). Ein hoher Wert deutet dabei auf eine umfassendere Exploration des Systems



(a) Pfad **(b)** Ring **(c)** Schleife **(d)** Spitze

Abbildung 6.3.: Bewegungsmuster in Hypertexten [nach Canter u. a. 1985, S. 96]

durch die Nutzer hin. NV/NS beschreibt das Verhältnis zwischen der Anzahl unterschiedlicher, besuchter Knoten NV und der Gesamtzahl der besuchten Knoten NS. Dieser Index drückt damit den Grad von redundanten Seitenaufrufen innerhalb einer Sitzung aus.

Unter Verwendung dieser grundlegenden Indizes identifizieren die Autoren in einer Untersuchung fünf Suchstrategien, die sich aus bestimmten Kombinationen und Ausprägungen der einzelnen Indizes ergeben [vgl. Canter u. a. 1985, S. 100].

1. Scannen – Eine Mischung aus Spitzen und kurzen Schleifen, während der Nutzer versucht, einen großen Bereich mit nur geringer Tiefe zu überfliegen (vgl. Abb. 6.4a) wird als Scannen bezeichnet. Charakteristisch sind hohe Werte für NV/NT und Spikiness und mittelhohe Werte für Loopiness.

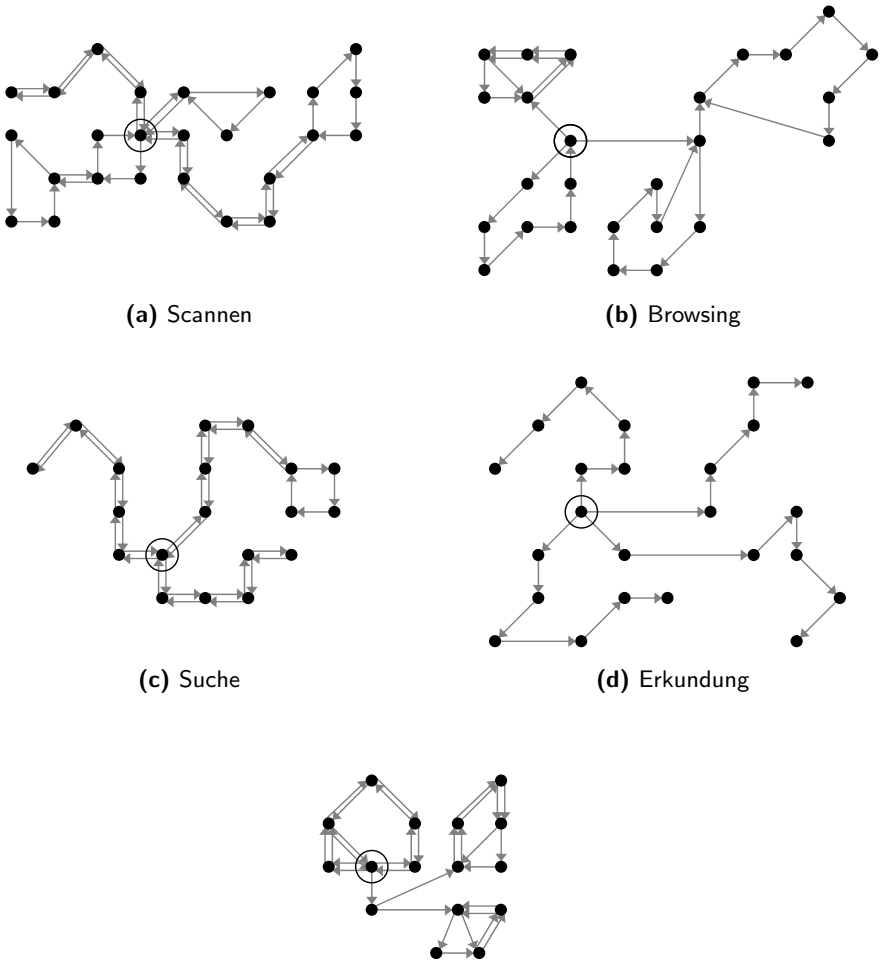
2. Browsing – Ein Durchbrowsen der Website ist gekennzeichnet durch zahlreiche lange Schleifen und einige wenige große Ringe (vgl. Abb. 6.4b). Besucher „lassen sich treiben“ bis ihr Interesse für ein bestimmtes Thema und damit für eine bestimmte Website-Sektion geweckt ist. Typisch sind mittelhohe Werte für Loopiness, Ringiness und NV/NS.

3. Suche – Zunehmende Spitzen und einige wenige Schleifen deuten auf ein Interesse des Nutzers für ein bestimmtes Ziel und damit auf eine Suche hin (vgl. Abb. 6.4c). Kennzeichen sind ein hoher Wert für Spikiness, ein mittelhoher Wert für Loopiness und ein niedriger Wert für NV/NS.

4. Erkundung – Eine explorative Nutzungsstrategie ist gekennzeichnet durch zahlreiche unterschiedliche Pfade, die auf eine Suche nach dem Umfang eines The-

menfeldes hindeuten (vgl. Abb. 6.4d). Typisch sind ein hoher Pathiness- und NV/NT-Wert.

5. Wanderung – Zahlreiche mittelgroße Ringe und mehrmals besuchte Knoten deuten auf eine unstrukturierte Vorgehensweise des Nutzers hin (vgl. Abb. 6.4e). Bei den Index-Werten sind ein hoher Wert für Ringiness, ein niedriger Wert für LV/NS und ein mittelhoher Wert für NV/NT charakteristisch für dieses Verhalten.



(a) Scannen **(b)** Browsing

(c) Suche **(d)** Erkundung

(e) Wanderung

Abbildung 6.4.: Suchstrategien in Hypertexten [nach Canter u. a. 1985, S. 100]

Diese Bewegungsmuster und Nutzungsstrategien können beispielsweise dazu herangezogen werden, das Verhalten erfahrener und unerfahrener oder regelmäßiger und unregelmäßiger Nutzer zu vergleichen.

6.2.2. Messung der Verlorenheit nach Otter und Johnson

Otter und Johnson stellen einen Ansatz vor, um das bei umfangreichen, nicht optimal strukturierten Hypertext-Systemen möglicherweise auftretende Gefühl der Verlorenheit („lostness“) zu messen. Verlorenheit wird als eine Situation verstanden, in der der Nutzer keine klare Vorstellung von der Konzeption der Beziehungen innerhalb eines Systems hat und nicht weiß, an welcher Stelle er sich innerhalb des Systems befindet [vgl. Otter und Johnson 2000, S. 4f.]. Vorhergehende Untersuchungen ergaben, dass das Auftreten von Verlorenheit bei Nutzern unabhängig von ihrem Wissensgrad in der jeweiligen Informationsdomäne ist. Drei Formen der Verlorenheit können differenziert werden:

1. Nicht zu wissen, wohin man als nächstes gehen soll.
2. Zu wissen, wohin man als nächstes gehen soll, aber nicht, wie man dorthin gelangt.
3. Nicht zu wissen, an welcher Stelle man sich in der übergreifenden Struktur befindet.

Bei auftretenden Problemen hinsichtlich des Verstehens der Dokumentstruktur wenden Nutzer vorhandenes Wissen aus der Nutzung von gedruckten Medien an. Dort liefern Seitenmengen, Kapitelnummerierung und Schlagwortverzeichnis eine Hilfestellung beim Auffinden einer bestimmten Information. Diese Hinweise fehlen bei Hypertextdokumenten häufig. Vier Ursachen für eine Desorientierung werden genannt:

1. Das Ankommen an einem bestimmten Punkt im Dokument und das gleichzeitige Vergessen des Grundes für den Dokumentaufruf.
2. Nach einem Exkurs wird vergessen zurückzukehren.
3. Es wird vergessen, einen vormals geplanten Exkurs zu machen.
4. Der Nutzer weiß nicht mehr, welche Bereiche er bereits besucht hat.

Desweiteren können Orientierungsprobleme in drei Gruppen eingeordnet werden:

1. Kognitive Überforderung entsteht aus der geistigen Belastung durch die Nutzung von Hypertextdokumenten. Auch Desorientierung selbst führt zu einer kognitiven Belastung.
2. Eine zweite Gruppe besteht aus der mangelnden Vertrautheit mit der Struktur bzw. des Organisationskonzeptes des Hypertextsystems.

3. Die dritte Problemgruppe kann der generellen Unerfahrenheit mit dem Browsing in Hypertextsystemen zugeschrieben werden.

Das erste von Otter und Johnson vorgestellte Maß für die Verlorenheit in Hypertextsystemen basiert auf einer Arbeit von [Smith 1996], die im Folgenden nach [Otter und Johnson 2000, S. 7f.] zitiert wird. Grundkonzept dieser Arbeit ist die Ermittlung von Abweichungen gegenüber einem optimalen Lösungsweg bei der Bearbeitung von Retrievalaufgaben. Bestandteil der Berechnung des Verlorenheitsmaßes sind folgende Werte:

T = Gesamtzahl der besuchten Knoten
D = Anzahl der unterschiedlichen besuchten Knoten
R = Anzahl der Knoten, die zur erfolgreichen Aufgabenerfüllung besucht werden müssen
S = Gesamtzahl der Knoten, die während der Suche besucht wurden
N = Anzahl der unterschiedlichen Knoten, die während der Suche besucht wurden

Es wird davon ausgegangen, dass der Nutzer bei einem perfekten Suchvorgang exakt die Anzahl von Knoten besuchen wird, die zum Lösen der Retrieval-Aufgabe erforderlich sind. Daher gilt:

$$T = D = R; \qquad S = T; \qquad \frac{N}{S} = 1; \qquad \frac{R}{N} = 1$$

Das von Smith entwickelte Maß zur Bestimmung der Verlorenheit wird wie in Formel 6.1 dargestellt berechnet. L steigt mit zunehmender Verlorenheit an, bei einer perfekten Suche gilt $L = 0$.

$$L = \left(\frac{N}{S} - 1\right)^2 + \left(\frac{R}{N} - 1\right)^2 \tag{6.1}$$

Otter und Johnson ergänzen in ihrer Erweiterung eine Berücksichtigung der Hyperlinks gemäß der Typologie in Abbildung 3.2 auf Seite 11. Grundlage hierfür ist die Vermutung, dass unterschiedliche Arten von Hyperlinks das Gefühl von Verlorenheit unterschiedlich beeinflussen [vgl. Otter und Johnson 2000, S. 8]. Gemeinsam mit Experten der Mensch-Computer-Interaktion wurde für die vier Arten extensionaler Hyperlinks die jeweilige Wahrscheinlichkeit festgelegt, mit der ein Gefühl der Verlorenheit von diesem Linktyp verursacht wird [vgl. Otter und Johnson 2000, S. 10f.]. Erläuternden Links wird die geringste Wahrscheinlichkeit zugesprochen, da sie lediglich auf einen Knoten verweisen und das Ziel explizit in der Linkbeschriftung beschrieben ist. Das Gewicht dieser Links, das später als Divisor dient, wird auf 4 festgelegt. Die nächsthöhere Wahrscheinlichkeit wird sequentiellen Links zugeordnet

(Gewicht: 3), da die von ihnen verlinkten Sequenzabschnitte vom Nutzer geistig leicht zu einer Einheit zusammengefügt werden können. Bei taxonomischen Links wird die Wahrscheinlichkeit als mittelhoch eingeschätzt, da sie gegenüber sequentiellen Links keine Einheit implizieren (Gewicht: 2). Eine hohe Wahrscheinlichkeit schließlich wird den assoziativen Links zugeordnet, da dieser Verknüpfungstyp für Nutzer nicht vorhersehbar ist (Gewicht: 1). Bei der Durchführung eines Retrieval-Tests muss für jeden Teilnehmer eine Klassifikation der von ihm besuchten Hyperlinks erfolgen und daraus eine Gesamt-Gewichtung LW („Link weighting“) gebildet werden. L entspricht dem oben vorgestellten Maß nach Smith. Das erweiterte Maß $LWLM$ wird wie in Formel 6.2 dargestellt berechnet.

$$LWLM = \frac{L}{\left(\frac{LW}{4}\right)} \tag{6.2}$$

Zur Überprüfung der beiden Maße führten Otter und Johnson einen Retrievaltest mit zwölf Teilnehmern durch, bei dem die Werte für L und $LWLM$ berechnet wurden und die Teilnehmer in einem Fragebogen anschließend Auskunft zu ihrem Eindruck einer Verlorenheit im Hypertextsystem gaben. Dabei entsprachen die Werte des $LWLM$-Maßes stärker der Selbsteinschätzung der Teilnehmer als die Werte für L [vgl. Otter und Johnson 2000, S. 30f.]. Sowohl für L als auch für $LWLM$ schlagen die Autoren einen Wertebereich von 0,4 bis 0,5 vor, oberhalb dessen wahrscheinlich eine Verlorenheit vorliegt. Über den Fragebogen brachten die Teilnehmer ferner zum Ausdruck, dass schlechte oder mehrdeutige Linkbeschriftungen Hauptursache für ein Gefühl der Verlorenheit sind.

6.2.3. Charakterisierung typischer und atypischer Sitzungen nach Sadagopan und Li

Ziel von Sadagopan und Li ist es, die Genauigkeit der Einschätzung der Nutzererfahrung bei Suchmaschinen mittels Klickstrom-Daten zu verbessern, indem atypische Sitzungen identifiziert und anschließend näher untersucht werden [vgl. Sadagopan und Li 2008, S. 885]. Dazu wird die Seltenheit einer Sitzung ins Verhältnis zur Wahrscheinlichkeit ihrer Klickstrom-Eigenschaften gesetzt. Sitzungen mit einer geringen Rarität werden als typisch betrachtet, jene mit einer hohen Rarität als untypisch. Bei typischen Sitzungen folgt der Klickstrom einer logischen Ereignissequenz, ohne dass Grund zur Annahme einer maschinell generierten oder missbräuchlichen Nutzung besteht.

Ausgangspunkt für diesen Ansatz sind Klickstromdaten einer Suchmaschine. Im ersten Schritt werden daher die sich dabei ergebenden Klickstrom-Ereignisse typisiert:

- Abruf einer Suchergebnisseite (P)
- Klick auf einer Suchergebnisseite, spezifiziert nach:
  - Klick auf Suchergebnis im organischen Ranking (W)
  - Klick auf einen werbefinanzierten Link (O)
  - Aufruf einer mittelbar oder unmittelbar folgenden Suchergebnisseite (N)
  - Sonstiger Klick, z.B. auf die Such-Schaltfläche (A)

Für jede Suchanfrage lässt sich nun ein Ereignis-Lokalität-Paar (Event-Locality-Pair (ELP)) bilden, bestehend aus einem $Ereignis \in \{P, W, O, N, A\}$ und einer $Seitennummer \in \mathbb{N}$ der Suchergebnisseite [vgl. Sadagopan und Li 2008, S. 886f.]. Klickt ein Nutzer nach dem Abruf der ersten Suchergebnisseite dort auf ein Ergebnis im organischen Ranking, ruft anschließend die zweite Suchergebnisseite auf und klickt dort auf eine Anzeige, ergibt sich folgendes ELP: $(P, 1)$,$(W, 1)$, $(N, 1)$, $(P, 2)$, $(O, 2)$.

Um einen Punktwert für die Seltenheit einer Sitzung zu errechnen, wird sie als Markow-Kette modelliert. Der Beginn einer Sitzung wird zusätzlich zu den oben beschriebenen Variablen als S repräsentiert. Der Zustand der Markow-Kette wird durch $\{S\} \cup \{\{P, W, N, O, A\} \times \mathbb{N}\}$ definiert. Die Übergangswahrscheinlichkeit $Pr(i, j)$ vom Zustand i zu j wird wie folgt berechnet:

$$Pr(i, j) = \frac{Q_{i,j}}{\sum_j Q_{i,j}} \tag{6.3}$$

Dabei entspricht $Q_{i,j}$ der Häufigkeit der Instanzen, bei denen der Zustand i vom Zustand j in allen Sitzungen gefolgt wird. Zustandsveränderungen mit einer geringen Wahrscheinlichkeit können mit einem seltenen Verhalten assoziiert werden. So hat in der vorliegenden Konstellation die Transition $S \to (P, 1)$ eine Wahrscheinlichkeit von nahezu 1, da fast alle Sitzungen mit dem Aufruf der ersten Suchergebnisseite beginnen. Für die oben genannte ELP-Sequenz ergibt sich eine Markow-Kette wie in Abbildung 6.5 dargestellt.

Im nächsten Schritt wird für die gesamte ELP-Sequenz ein Wahrscheinlichkeitswert ϕ errechnet:

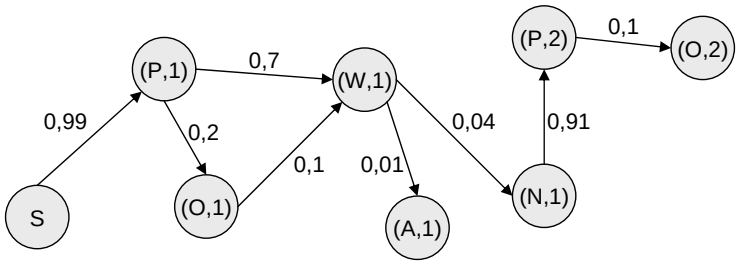


Abbildung 6.5.: Beispiel einer Markow-Kette [Sadagopan und Li 2008, S. 887]

$$\begin{aligned}\phi &= Pr((P,1)|S) \times Pr((W,1)|(P,1)) \times Pr((N,1)|(W,1)) \times \\ &\quad Pr((P,2)|(N,1)) \times Pr((O,2)|(P,2)) \\ &= 0,99 \times 0,7 \times 0,04 \times 0,91 \times 0,1 \\ &= 0,00252\end{aligned} \tag{6.4}$$

Um umfangreichere Sitzungen nicht durch das damit einhergehende niedrige Produkt zu benachteiligen, wird der Logarithmus des Wahrscheinlichkeitswertes gebildet und durch die Anzahl der Ereignisse (Transitionen) geteilt:

$$\begin{aligned}MLH_{avg} &= \frac{\ln(\phi)}{|events|} \\ &= \frac{\ln(0,00252)}{5} = -1,2\end{aligned} \tag{6.5}$$

MLH_{avg} ist somit ein Maß für die Konformität der ELPs einer Sitzung mit der populären Nutzung der Suchmaschine [vgl. Sadagopan und Li 2008, S. 888]. Ein hoher Wert zeigt, dass die Mehrzahl der Transitionen in der betrachteten Sitzung populär sind. Allerdings können qualitativ sehr unterschiedliche Sitzungen ähnliche Werte für MLH_{avg} erreichen. Daher werden weitere Klickstrom-Daten in einem multidimensionalen Modell herangezogen:

- P_t: Anzahl der Seitenabrufe
- W_t: Anzahl der Klicks auf Suchergebnisse
- O_t: Anzahl der Klicks auf Werbelinks
- N_t: Anzahl der Aufrufe einer weiteren Suchergebnisseite
- A_t: Anzahl der sonstigen Klicks

- $E = P_t + W_t + O_t + N_t + A_t$: Gesamtzahl der Ereignisse

In dem Modell wird jede ELP-Sequenz wie folgt repräsentiert:

$$q \equiv (MLH_{avg}, E, P_f, W_f, O_f, N_f, A_f) \tag{6.6}$$

Es gelten $P_f = \frac{P_t}{E}$, $W_f = \frac{W_t}{E}$, $O_f = \frac{O_t}{E}$, $N_f = \frac{N_t}{E}$ und $A_f = \frac{A_t}{E}$. Zur Bestimmung der Seltenheit einer Sitzung für den Punkt q, der eine Nutzersitzung repräsentiert, wird die Mahalanobis-Distanz verwendet.

$$d = \sqrt{(q-\mu)\sum^{-1}(q-\mu)^T} \tag{6.7}$$

μ ist der mittlere Zeilenvektor und $\sum$ die Kovarianzmatrix der Klickstrom-Charakteristiken im siebendimensionalen Raum. Je höher die Distanz d, desto höher die Seltenheit von q.

Auf der Grundlage dieses Distanzmaßes lassen sich nun weitere statistische Auswertungen durchführen. So identifizieren Sadagopan und Li nach einer Analyse der Ausreißer-Sitzungen vier Kategorien solcher Sitzungen, beispielsweise Sitzungen, in denen wiederholt auf die „Nächste Seite“-Schaltfläche geklickt wird. Derartige Verhaltensweisen können unter Umständen Hinweise auf Schwachstellen in der Oberflächengestaltung des Webangebots liefern, beispielsweise mehrdeutige Funktionsbezeichnungen.

6.2.4. Ableitung von Nutzerprofilen aus Logfile-Daten nach Stermsek u. a.

Stermsek u. a. haben untersucht, inwiefern sich Nutzerprofile durch die Auswertung von Klickstrom-Daten identifizieren lassen [vgl. Stermsek u. a. 2007, S 1f.]. Es wird davon ausgegangen, dass Nutzer verschiedenartige Informationsinteressen haben, die sich in Kategorien modellieren lassen. Zur Durchführung des Verfahrens sind eine Reihe von Daten erforderlich:

- Nutzer-ID als Identifikator individueller Nutzer
- Abgerufene Seiten
- Inhaltsbezogene Metadaten, beispielsweise Meta-Tags oder RDF-Tripel

- Sitzungsinformationen
- Strukturelle Informationen zur untersuchten Website (optional)
- Zeit (t), die der Nutzer auf einer Seite verbrachte
- Frequenz (f), mit der ein Nutzer eine Seite abrief
- Zentralität (c), eine Aussage über die Kürze der Pfade zu anderen Knoten
- Prestige (p), eine Angabe zur Anzahl der Links von anderen Knoten

Die Vorverarbeitung der Logfiles umfasst die Extraktion der Seitenzugriffe (und somit das Entfernen sonstiger Zugriffe, z.B. auf Bilder), das Herausfiltern unerwünschter Zugriffe, z.B. auf HTTP-Fehler oder Robot-Zugriffe, und die Identifikation von Nutzersitzungen anhand von Cookies [vgl. Stermsek u. a. 2007, S 3f.].

Grundlage der Auswertung ist die Graphentheorie. Zur Auswertung der Klickstrom-Daten werden diese in Adjazenzmatrizen überführt [vgl. Stermsek u. a. 2007, S 4]. Für jeden Nutzer wird eine eigene Matrix erstellt. Bei der weiteren Analyse werden s aus n Seiten einer jeden Sitzung berücksichtigt. Stermsek u. a. gehen davon aus, dass bei einer kleinen, inhaltlich speziellen Website fast alle Seiten von Interesse für die Auswertung sind, diese Relevanz aber mit zunehmender Website-Größe abnimmt. Für die Bestimmung von s wird daher folgende Berechnung verwendet: $s = 1 + 2 \times \ln(n)$. Besteht die Website beispielsweise aus 250 Seiten, werden gemäß $s = 1 + 2 \times \ln(250) = 12,04$ zwölf Seiten für die Ableitung der Nutzerprofile herangezogen.

Alle folgenden Schritte werden für jede Nutzersitzung durchgeführt. Zunächst sind die signifikantesten zu berücksichtigenden Seiten zu bestimmen. Dazu werden die oben aufgezählten Daten in Abhängigkeit von ihrer Verfügbarkeit und dem Inhalt der Website gewichtet und addiert. Die generische Formel hierzu lautet $total = \alpha \times t + \beta \times f + \gamma \times c + \delta \times p$, wobei die griechischen Symbole für die Gewichtungsfaktoren und die lateinischen Buchstaben für normalisierte Werte entsprechend ihrer oben genannten Bedeutung stehen. Das Beispiel eines Ergebnisses dieses Vorgehens ist in Tabelle 6.1 dargestellt.

	Zeit (t)		**Frequenz** (f)		**Zentralität** (c)		**Prestige** (p)		**Total** (g)	
Gewichtung	**0,7**		**0,2**		**0,05**		**0,05**			
	Wert	Rang	Wert	Rang	Wert	Rang	Wert	Rang	Wert	Rang
/ordner-a/seite1/	0,005	4	0,042	3	0,020	4	0,000	4	0,013	4
/ordner-a/seite2/	0,057	2	0,167	1	0,208	1	0,174	1	0,092	2
/ordner-b/seite1/	0,010	3	0,125	2	0,157	2	0,130	2	0,046	3
/ordner-b/seite2/	0,112	1	0,033	4	0,077	3	0,087	3	0,093	1

Tabelle 6.1.: Beispielhafte Ergebnisse der Klickstrom-Analyse nach Stermsek u. a. [vgl. 2007, S. 5]

Die Autoren ziehen dabei die Besuchszeit einer Seite als Indikator für das Nutzerinteresse an dieser Seite heran. In einem iterativen Prozess wird ein Verfahren zur Normalisierung entwickelt. Dabei werden Werte, die über dem dreifachen Median liegen, durch den Mittelwert der übrigen Werte ersetzt. Die weiteren Werte werden auf Grundlage einer Adjazenzmatrix von $n \times n$ berechnet, wobei n der Anzahl der Seiten der Website entspricht. Nach Sicht der Autoren ist die auf einer Seite verbrachte Zeit der wichtigste Indikator für das Nutzerinteresse und wird daher mit $0,7$ am stärksten gewichtet. Die weiteren Werte werden entsprechend Tabelle 6.1 schwächer gewichtet.

Zur Bildung der Nutzerprofile werden die s Seiten mit dem höchsten Gesamtwert g verwendet. Jeder dieser Seiten werden semantische Informationen zugeordnet. Diese können seitenspezifischen Metatags (beispielsweise nach Dublin Core) entnommen, per Textextraktion gefunden oder aus dem Pfad- und Dateinamen gewonnen werden. Anschließend werden je Seite zunächst alle redundanten, dann die insgesamt häufigsten Attribute entfernt, bis ein Attribut verbleibt. Die verbleibenden Attribute sind in Abbildung 6.6 fett hervorgehoben. Die im vorherigen Schritt als signifikant identifizierten Seiten sind in der Abbildung dick umrandet, die anderen Seiten werden entfernt. Ergebnis ist das im rechten Teil von Abbildung 6.6 dargestellte Nutzerprofil. Solche Nutzerprofile geben einen Überblick über die Interessenschwerpunkte der Website-Besucher und können als Grundlage für die thematische Weiterentwicklung eines Angebotes dienen.

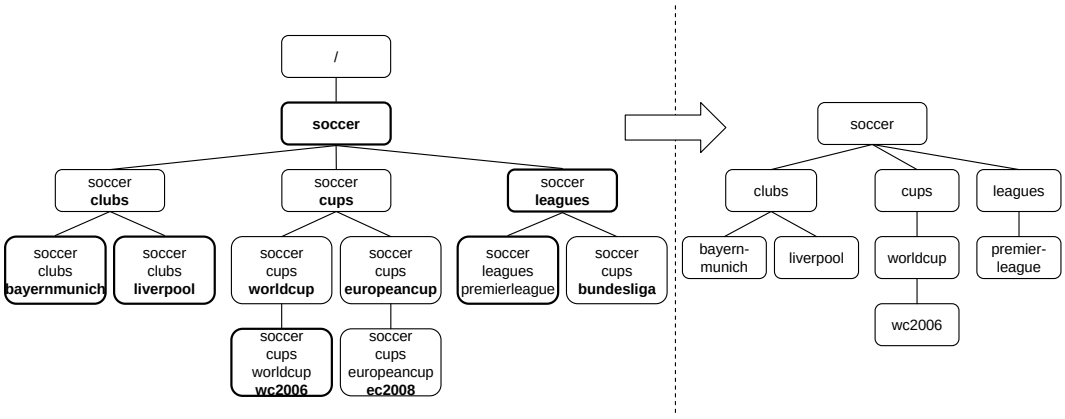


Abbildung 6.6.: Nutzerprofil-Generierung [Stermsek u. a. 2007, S. 7]

6.2.5. Kombination ethnographischer Daten und Klickstromdaten zur Identifikation von Browsing-Strategien nach Clark u. a.

Bei diesem Ansatz werden aus Klickstrom-Daten ermittelte Bewegungsmuster mit qualitativen, in Befragungen ermittelten Aussagen über Nutzungsgewohnheiten kom-

biniert [Clark u. a. 2006, vgl.]. Die Untersuchung wurde mit Studenten in einem neunwöchigen Seminar durchgeführt. Grundlage der Klickstrom-Auswertung sind Logfiles und die Bewegungsmuster nach Canter u. a. (vgl. Abschnitt 6.2.1).

Zur Verbesserung der Datenqualität der Logfiles wird eine Pattern Restore Method (PRM) nach [Ting u. a. 2005] angewendet. Ziel hierbei ist die Ergänzung fehlender Seitenzugriffe in den Logfiles [vgl. Ting u. a. 2005, S.503f.]. Dieser Effekt entsteht aus der Bedienung erneuter Seitenaufrufe aus dem lokalen Zwischenspeicher des Browsers und aus der Nutzung mehrerer Browser-Fenster oder -Tabs. Nach dieser Vorverarbeitung der Logfiles wird für jede Sitzung ein Nutzerschritte-Diagramm erzeugt. Dazu werden auf der y-Achse die besuchten Seiten und auf der x-Achse der Zeitverlauf eingetragen. Tabelle 6.2 zeigt als Beispiel die einzelnen Schritte einer Nutzersitzung beim DBS. Die mit jedem Seitenaufruf insgesamt verbrachte Zeit ist in Spalte drei addiert. „Seitennummer“ bezeichnet eine individuell für diese Sitzung generierte Identifikationsnummer einer jeden Seite.

Zeit vorh. Seite	Seite	Zeit gesamt	Seitennummer
0	innovationsportal/bildungplus.html	0	0
48	innovationsportal/zeigen.html?seite=1584	48	1
37	innovationsportal/zeigen.html?seite=2147	85	2
51	innovationsportal/zeigen.html?seite=1584	136	1
5	innovationsportal/zeigen.html?seite=2145	141	3
34	innovationsportal/zeigen.html?seite=1584	175	1
2	innovationsportal/bildungplus.html	177	0

Tabelle 6.2.: Klickstrom-Sequenz

Aus diesen Daten lässt sich das in Abbildung 6.7 dargestellte Nutzerschritte-Diagramm erzeugen. Da für die letzte Seite keine Ansichtsdauer bekannt ist, ist sie im Diagramm durch lediglich einen x-Wert repräsentiert.

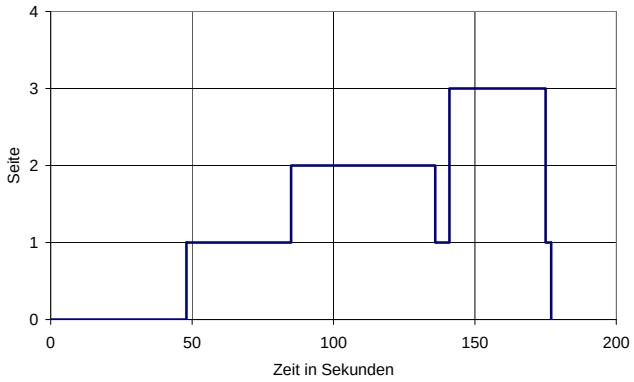
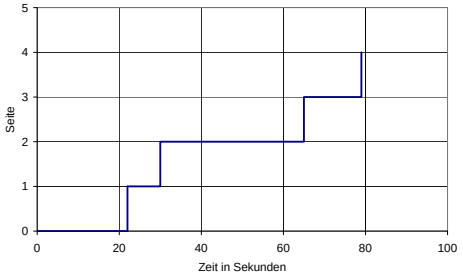


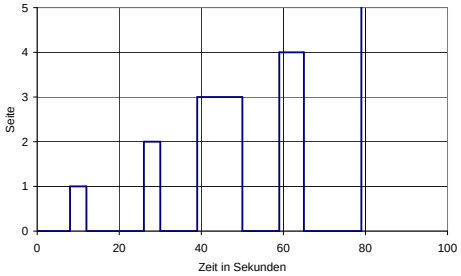
Abbildung 6.7.: Beispiel für ein Nutzerschritte-Diagramm nach Clark u. a.

Clark u. a. identifizieren unterschiedliche Typen von Nutzerschritte-Diagrammen. Bei

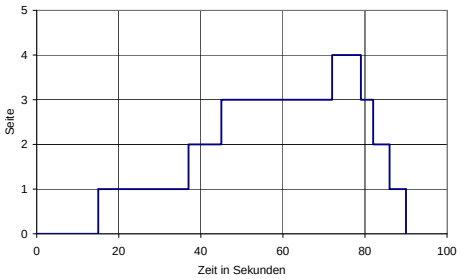
einer „Treppe“ bewegt sich der Nutzer vorwärts durch die Website, ohne eine zuvor besuchte Seite noch einmal aufzurufen (Abbildung 6.8a). Beim „Finger“-Muster kehrt der Nutzer nach jedem Seitenaufruf innerhalb der Website stets wieder auf die Ausgangsseite zurück (Abbildung 6.8b). Bei einem „Berg“-Muster ruft der Nutzer mehrere Seiten in einer Vorwärtsbewegung auf, um schließlich eine bestimmte Seite zu erreichen oder zu einer bestimmten Seite zurückzukehren (Abbildung 6.8c).



(a) Treppe



(b) Finger



(c) Berg

Abbildung 6.8.: Typen von Nutzerschritte-Diagrammen [nach Clark u. a. 2006]

Diese Diagrammtypen lassen sich den von Canter u. a. eingeführten Bewegungsmustern zuordnen. Das Treppen-Muster entspricht dabei dem Pfad, das Berg-Muster der Schleife und das Finger-Muster der Spitze.

Diese Musterzuordnung wurde im Rahmen eines Hochschulseminars an einer für dieses Seminar konzipierten hierarchisch strukturierten Website überprüft. Dabei wurden für 476 Sitzungen Nutzerschritte-Diagramme erstellt und kategorisiert. Um die manuelle Kategorisierung zu erleichtern, wurden die Diagramme geglättet, indem für jeden Seitenaufruf eine Standardzeitspanne statt der tatsächlichen Besuchszeit verwendet wurde. Dabei konnte eine Verschiebung der Muster weg von komplexen Bewegungen, die zwei oder mehr der oben vorgestellten Muster enthalten, hin zu den drei einzelnen Mustern festgestellt werden. Zusätzlich zur Untersuchung der Klickstrom-Daten wurden 86 Studenten in der zweiten und achten Woche per Fragebogen befragt sowie sechs Studenten interviewt und unter Laborbedingungen beobachtet. Befragung und Beobachtung dienten einer Ergänzung der quantitativen

Klickstrom-Daten durch qualitative, ethnographische Daten. Clark u. a. zeigten dabei, dass eine derartige Kombination quantitativer und qualitativer Daten zusätzliche Erkenntnisse liefern kann. Beispielsweise wird anhand der Abweichung zwischen dem erklärten Ziel eines Teilnehmers für die Website-Nutzung und dem entsprechenden Nutzerschritte-Diagramm eine Schwachstelle des PRM-Algorithmus aufgezeigt.

6.3. Bewertung der Ansätze

Die vorgestellten Auswertungsansätze werden im Folgenden im Hinblick auf ihre Umsetzbarkeit beim DBS im Allgemeinen und mit den vorliegenden Daten im Speziellen bewertet.

Canter u. a. beschreiben kein direkt praktisch umsetzbares Modell, sondern eine Typologie von Bewegungsmustern und eine theoretische Vorgehensweise zur Identifizierung dieser Muster. Andere Autoren greifen auf diese Muster für weiterführende Analysen zurück.

Mit dem von Otter und Johnson verwendeten Konzept der Verlorenheit könnte überprüft werden, ob bei der DBS-Nutzung ein Gefühl der Verlorenheit auftritt. Allerdings lässt sich das Verfahren aus drei Gründen nicht automatisiert auf den vorliegenden, quantitativen Nutzungsdaten anwenden. Erstens geht aus den Nutzungsdaten nicht eindeutig hervor, welches Informationsbedürfnis und damit welche Retrieval-Fragestellung vorliegt. Zweitens lässt sich kein optimaler Lösungsweg der Retrieval-Fragestellung bestimmen, da aufgrund der verschiedenen Facetten, nach denen die DBS-Datensätze organisiert sind, häufig unterschiedliche Pfade zu einer Zielseite existieren. Drittens ist die Kategorisierung der Links sehr aufwändig, auch wenn hier aufgrund gleichförmiger Datenstrukturen möglicherweise Generalisierungen möglich sind.

Der Ansatz von Sadagopan und Li scheint gut auf den DBS übertragbar zu sein. So ist beispielsweise eine Anpassung der von ihnen definierten Seitentypen auf die Seitentypen des DBS denkbar. Statt Suchergebnisseiten und Aufrufe von Ergebnisseiten könnten die unterschiedlichen Seitenklassen des DBS verwendet werden. Vor einer Umsetzung muss jedoch noch festgelegt werden, wie und nach welchen Kriterien die resultierenden untypischen und typischen Sitzungen analysiert werden sollen.

Das Verfahren von Stermsek u. a. zur Ermittlung von Nutzerprofilen ist grundsätzlich beim DBS anwendbar. Die benötigten Nutzungsdaten sind sowohl über die Logfiles als auch über die Piwik-Datenbank verfügbar. Problematisch ist hier, dass die verschiedenen DBS-Datentypen über keine gemeinsame Kategorisierung o.ä. verfügen.

Eine Umsetzung insbesondere der Adjazenzmatrizen mit einem relationalen Datenbanksystem wie MySQL erscheint aufgrund der dazu erforderlichen großen Tabellen unrealistisch. Ein Export der Daten in ein Statistikprogramm ist demnach notwendig. Doch auch dort muss überprüft werden, wie der Ansatz von Stermsek u. a. bei großen Datenmengen skaliert. Ein weiterer Nachteil ist in der Fokussierung auf einzelne Sitzungen zu sehen. Die daraus generierten Nutzerprofile müssten in einem weiteren Schritt aggregiert werden, um generelle Aussagen zum Nutzerverhalten treffen zu können.

Ein Vorteil des Ansatzes von Clark u. a. liegt in der anschaulichen Visualisierung der Bewegungsmuster. Diese Art der Darstellung ist jedoch bei umfangreichen Websites mit vielen Nutzersitzungen in der geschilderten Vorgehensweise mit einer manuellen Kategorisierung nicht mehr handhabbar. Hierzu wäre eine Ergänzung um eine automatische Klassifkation der Sitzungen notwendig.

Keiner dieser Ansätze ist im Rahmen dieser Arbeit unmittelbar umsetzbar. Stattdessen wird zunächst eine Auswahl klassischer Webmetriken ermittelt, anschließend werden Auswertungen mit einzelnen Aspekten des Web Usage Mining durchgeführt. Die Konzeption dieser Untersuchungen wird im folgenden Kapitel beschrieben.

7. Konzeption der Datenanalyse

In diesem Kapitel wird die Planung der Nutzungsdatenanalyse beschrieben. Grundlage für die Konzeption ist der von Diekmann beschriebene Forschungsablauf [Diekmann 2004, S. 30ff.].

7.1. Forschungsablauf

Im ersten Schritt ist das Untersuchungsziel zu definieren. Im Rahmen dieser Arbeit soll mittels quantitativer Nutzungsdaten untersucht werden, ob, und wenn ja, inwiefern, sich das Nutzungsverhalten der DBS-Besucher in Abhängigkeit von der Art ihres Einstiegs unterscheidet. Unterschieden wird zwischen Direkteinstiegen (keine Referrer-Angabe beim ersten Seitenaufruf einer Sitzung), Suchmaschineneinstiegen (Referrer-URL, die einer Suchmaschine zugeordnet werden kann) und Einstiegen über sonstige Verweise (sonstiger Referrer).

Aufbauend auf dem Untersuchungsziel sind anschließend Hypothesen zu bilden. Sie stellen Zusammenhänge zwischen bereits vorhandenen theoretischen Grundlagen und der durchzuführenden Untersuchung her. Während der Umgang von Nutzern mit Suchmaschinen häufiges Forschungsthema ist [vgl. z.B. Erlhofer 2007], konnten in einer Literaturrecherche keine Forschungsergebnisse zu besonderen Charakteristiken in der Nutzung allgemeiner Webangebote durch Suchmaschinennutzer ermittelt werden. Aufgrund des in der Voruntersuchung nachgewiesenen hohen Anteils von Suchmaschineneinstiegen sind etwaige Unterschiede in der Nutzungsweise durch jene Besucher von großer Bedeutung und sollen daher untersucht werden. Da die in der Einleitung vorgestellte Vorabuntersuchung nicht umfangreich war und die Ergebnisse keine eindeutigen Hinweise auf eine unterschiedliche Nutzungsweise des DBS hinsichtlich der unterschiedlichen Einstiegsarten liefern, wird in der zu falsifizierenden Grundannahme davon ausgegangen, dass keine Unterschiede bestehen. Wegen des geringen Anteils der sonstigen Referrer mit knapp 12% wird zudem der Schwerpunkt auf die beiden häufigeren Einstiegsarten (Direkt- und Suchmaschineneinstiege) gelegt.

Die allgemeine Nullhypothese, die später für die unterschiedlichen Untersuchungsmerkmale spezifiziert wird, lautet:

H_0 : Das Nutzungsverhalten ist von der Art des DBS-Einstiegs **unabhängig**.

Die zu verifizierende Alternativhypothese lautet dementsprechend:

H_1 : Das Nutzungsverhalten ist von der Art des DBS-Einstiegs **abhängig**.

Gegenstand der Konzeptualisierungsphase ist eine Operationalisierung, also die Überführung der theoretischen Begriffe in messbare Merkmale und die Festlegung der Variablen [vgl. Raithel 2008, S. 36].

Als Hauptbetrachtungsebene wird die Sitzung gewählt. Sie erlaubt anders als Seitenabrufe die Betrachtung von Ein- und Ausstiegspunkten sowie wichtigen Webanalyse-Metriken wie der Besuchstiefe. Im Gegensatz zur Betrachtungsebene der Besucher ist bei Sitzungen die Akzeptanz persistenter Cookies durch die Benutzer nicht erforderlich, wodurch das Fehlerrisiko bei der Sitzungsbestimmung reduziert wird.

Dazu wird von den drei grundlegenden Phasen einer Websession ausgegangen: dem Einstieg, dem Verhalten auf der Website und dem Ausstieg. Für den DBS-Einstieg werden die absoluten und relativen Anteile der drei Einstiegsarten berechnet. Das Nutzerverhalten auf dem DBS wird an einer Auswahl der unter 5.1 vorgestellten Webanalyse-Metriken festgemacht. Folgende Metriken werden berechnet und nach der Einstiegsart segmentiert:

- Zeitverteilung der Sitzungen auf Wochentag und Stunde

  Hierbei soll überprüft werden, ob die verschiedenen Einstiegsarten zu unterschiedlichen Zeiten verstärkt auftreten, um so ggf. Rückschlüsse auf unterschiedliche Nutzergruppen oder Nutzungsgewohnheiten zu ziehen.

- Anzahl der Aktionen je Sitzung (Besuchstiefe)

  Hierbei handelt es sich um einen Indikator für die Nutzungsintensität. Er eignet sich gut zur Identifizierung unterschiedlicher Nutzungsweisen, auch wenn ohne zusätzliche qualitative Angaben zum Nutzerbedürfnis nicht beurteilt werden kann, ob ein hoher Wert positiv oder negativ zu beurteilen ist.

- Sitzungsdauer

  Für sie gilt Ähnliches wie für die Besuchstiefe: Die Sitzungsdauer eignet sich gut zur Charakterisierung einer Sitzung, hohe oder niedrige Werte können aber nicht allgemeingültig beurteilt werden.

- Absprungrate

  Eine hohe Absprungrate impliziert eine hohe Diskrepanz zwischen dem Informationsbedürfnis bzw. den Erwartungen der Nutzer und ihrer Wahrnehmung der Einstiegsseite. Die Absprungrate ist damit ein weiterer gut geeigneter Bestandteil einer Nutzercharakterisierung.

- Anteil wiederkehrender Besucher

  Ein hoher Anteil wiederkehrender Besucher spricht für eine gewisse Zufriedenheit und Loyalität gegenüber dem DBS. Sie soll ebenfalls für die verschiedenen Einstiegsarten ermittelt werden.

- Ausstiegsart

  Bei der Ausstiegsart kann beim Typ der Ausstiegsseite unterschieden werden zwischen interner und externer Seite. Das Verlassen über einen externen Link deutet auf eine Erfüllung der Verweisserverfunktion des DBS hin.

Für Suchmaschineneinstiege werden zusätzlich die ranghöchsten Suchmaschinen und -phrasen und für Einstiege über sonstige Verweise die ranghöchsten Verweisseiten hinsichtlich ihrer absoluten Häufigkeit, Besuchstiefe und Besuchsdauer ermittelt.

Neben diesen reinen Webanalyse-Metriken werden weitere Auswertungen durchgeführt, die auf die inhaltliche und funktionelle Struktur des DBS eingehen. Zwar handelt es sich dabei noch nicht um ein vollumfängliches Web Usage Mining, da diese Auswertungen jedoch über klassische Webanalyse-Metriken hinausgehen, werden sie gesondert dargestellt. Jeweils segmentiert nach Einstiegsart und ausgewertet werden:

- Einstiege auf und Absprünge von thematisch und funktionell kategorisierten Seiten
- Einstiege in die Hierarchieebene der DBS-Redaktionsseiten
- Auf- und Abwärtsbewegungen innerhalb der Hierarchieebenen
- Bewegungen innerhalb der Navigationsstruktur zwischen Themen- und Zielgruppenbereichen sowie zwischen der Navigationsstruktur insgesamt und der Suche

Prinzipiell ließe sich jede Webmetrik wie z.B. die Anzahl der Aktionen oder die Besuchsdauer auf inhaltlich oder funktionell kategorisierte Seiten beziehen. Die Beschränkung dieser Auswertungen auf die Absprungrate wird vorgenommen, da sich an ihr die unmittelbare Wirkung der zuerst aufgerufenen Seite auf den Besucher vergleichsweise gut ablesen lässt. Die Häufigkeit der Absprünge bzw. die Absprungrate

ist daher für eine erste Teilumsetzung des Web Usage Mining die interessanteste Webmetrik.

Variablen werden unterschieden in abhängige und in unabhängige Variablen [vgl. Raithel 2008, S. 33]. Abhängige Variablen sollen untersucht werden. Von unabhängigen Variablen wird angenommen, dass sie die abhängigen Variablen beeinflussen. Ihr Einfluss wird untersucht. Somit sind die oben genannten Metriken im Rahmen der statistischen Analyse die abhängigen Variablen, während die Einstiegsart die unabhängige Variable mit den drei möglichen Ausprägungen direkt, Suchmaschine und sonstige Verweise bildet.

Abbildung 7.1 stellt die Operationalisierung grafisch dar.

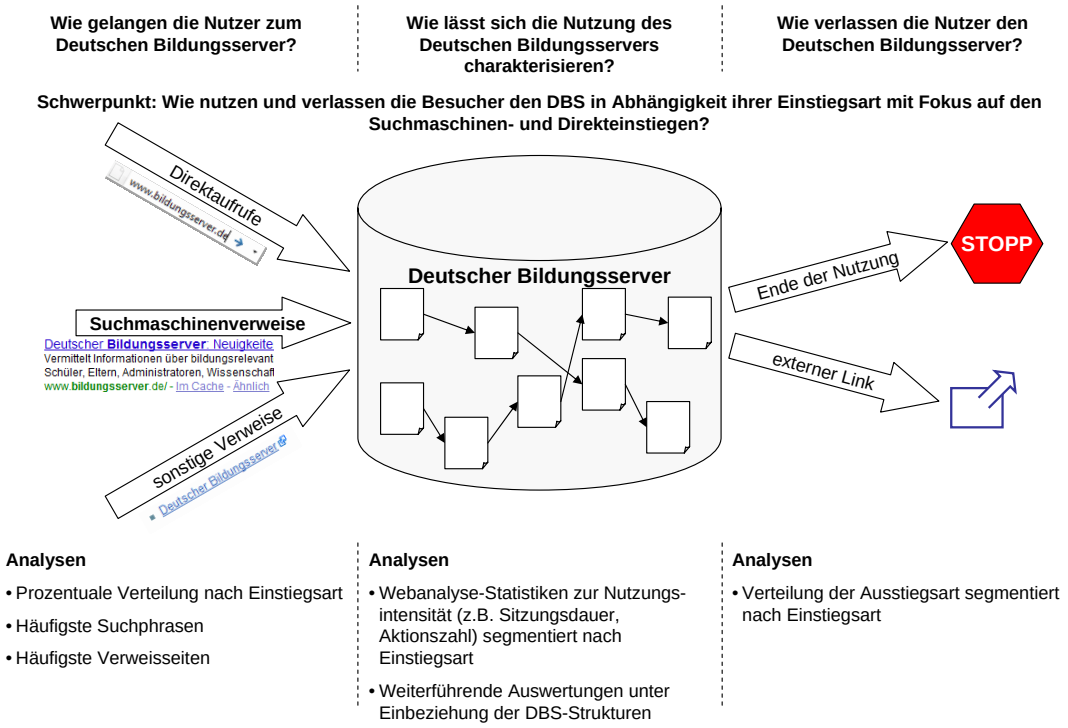


Abbildung 7.1.: Operationalisierung der Forschungshypothesen

7.2. Grundlagen der statistischen Auswertung

Die Nutzungsdaten sollen mittels statistischer Methoden untersucht werden. Das Gebiet der Statistik lässt sich in die Teilbereiche beschreibende oder deskriptive Statistik, schließende oder induktive Statistik und Wahrscheinlichkeitsrechnung untergliedern [vgl. Bourier 2009, S. 1ff.]. Ziel der deskriptiven Statistik ist eine Beschreibung des Datenmaterials durch die Berechnung von Kennzahlen oder die Bestimmung von Abhängigkeiten zwischen Messgrößen. Bei der induktiven Statistik geht es darum,

von Daten, die nur für einen Teil des interessierenden Untersuchungsobjektes vorliegen, auf das gesamte Objekt zu schließen. Das Fehlerrisiko, mit dem ein solcher Rückschluss verbunden ist, wird quantifiziert. Gegenstand der Wahrscheinlichkeitsrechnung ist die Bestimmung des Ausmaßes, mit dem ein möglicher Ausgang eines Vorgangs eintritt. Wikipedia nennt statt der Wahrscheinlichkeitsrechnung die explorative Statistik als drittes Teilgebiet [vgl. Wikipedia 2009f]. Dabei handelt es sich um eine an eigenständiger Bedeutung gewinnende Mischform der deskriptiven und induktiven Statistik, welche die beim Data Mining benötigten statistischen Methoden umfasst.

Wichtige Maßzahlen der deskriptiven Statistik sind Mittelwerte, hierunter Modus, Median und arithmetisches Mittel, sowie Streuungsmaße und Assoziationsmaße.

Die Standardabweichung ist das gebräuchlichste Streuungsmaß und gibt die Streuung einer Variablen um ihren Mittelwert in der Maßeinheit der Variablen an [vgl. Benninghaus 2007, S. 58ff.]. Das Maß wird wie folgt berechnet:

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N}} \tag{7.1}$$

Weiterhin wird die statistische Beziehung (Assoziation) zwischen Variablen betrachtet. „Zwischen zwei Variablen besteht eine statistische Beziehung, wenn die konditionalen Verteilungen verschieden sind; zwei Variablen stehen nicht miteinander in Beziehung, wenn die konditionalen Verteilungen gleich sind.“ [Benninghaus 2007, S. 83] Es geht also um die Überprüfung der Frage, ob zwei Variablen abhängig oder unabhängig voneinander variieren. Um die Beziehung zwischen Variablen in einer Kennzahl beschreiben zu können, werden Assoziationsmaße gebildet (bivariate Zusammenhangmaße). Voraussetzung für die Wahl eines Assoziationsmaßes ist die Bestimmung des Skalenniveaus der vorliegenden Daten. Unterschieden werden hauptsächlich vier Skalenniveaus, die hinsichtlich ihrer Anforderungen aufeinander aufbauen [vgl. Benninghaus 2007, S. 22f.]. Bei Nominalskalen wird lediglich zwischen der Gleichheit und Ungleichheit von Elementen unterschieden, z.B. männlich/weiblich. Lassen sich Variablen zwei Kategorien zuordnen, spricht man von Dichotomie. Kann eine Rangordnung gebildet werden, ohne dass sich Abstände zwischen den Kategorien spezifizieren lassen, liegt eine Ordinalskala vor [vgl. Ferber 2003, S. 122f.]. Beispiele hierfür sind Schulnoten oder Kleidungsgrößen (S bis XXL). Lassen sich Werte addieren und somit Abstände messen, z.B. bei der Celsius-Temperaturskala, liegt eine Intervallskala vor. Beim vierten Niveau, der Rationalskala, sind zusätzlich noch Multiplikation und ein absoluter Nullpunkt definiert, beispielsweise bei der Kelvin-Temperaturskala.

Für nominale Skalen sind Chi-Quadrat-basierte Maße wie Phi oder Cramers V verbreitet [vgl. Cleff 2008, S. 22f.]. Letzteres kann auch verwendet werden, wenn eine Variable nominal- und die andere ordinalskaliert ist. Als Zusammenhangmaße zwischen ordinalskalierten Variablen kommen Spearman (ρ) und Kendalls Tau (τ) infrage. Zwischen zwei rationalskalierten Variablen kann die Korrelation nach Pearson (r) verwendet werden. Da es sich bei den ausgewählten Metriken mit Ausnahme der Besuchszeit um nominalskalierte Werte handelt, wird im Folgenden vor allem auf dafür geeignete Assoziationsmaße eingegangen.

Ausgehend von einer Kontingenztabelle, in der die Werte für beide zu untersuchenden Variablen samt der Zeilen- und Spaltenssummen eingetragen werden, wird eine Indifferenztabelle mit den erwarteten Häufigkeiten berechnet, die auftreten würden, wenn kein Zusammenhang zwischen beiden Variablen besteht [vgl. Cleff 2008, S. 82ff.]. Die Formel hierfür lautet:

$$n_{ij}^{e} = \frac{\text{Zeilensumme} \times \text{Spaltensumme}}{\text{Gesamtsumme}} = \frac{n_i \times n_j}{n} \tag{7.2}$$

Die Chi-Quadrat-Werte der einzelnen Zellen ergeben sich nun aus der quadrierten Differenz zwischen Häufigkeit und erwarteter Häufigkeit geteilt durch die jeweils erwartete Häufigkeit [vgl. Cleff 2008, S. 87f.]. Die Summe dieser Einzelwerte ergibt Chi Quadrat.

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(n_{ij} - n_{ij}^{e}\right)^2}{n_{ij}^{e}} \tag{7.3}$$

Ein Wert von 0 bedeutet, dass zwischen den Variablen kein Zusammenhang besteht. Die unabhängige Variable hat keinen Einfluss auf die abhängige Variable. Ein Nachteil von Chi Quadrat ist, dass der Wert mit zunehmender Zeilen- und Spaltenzahl der Kontingenztabelle zunimmt. Cramers V ist eine auf Chi Quadrat basierende Weiterentwicklung und liefert stets Werte zwischen 0 und 1 [vgl. Cleff 2008, S. 92].

$$\text{Cramers V} = \sqrt{\frac{\chi^2}{n \times (\min(k, m) - 1)}} \in [0; 1] \tag{7.4}$$

Für eine Einordnung von Cramers V wird häufig folgende Einteilung verwendet:

$$V \in [0,0;0,1] \rightarrow \text{kein Zusammenhang}$$
$$V \in [0,1;0,3] \rightarrow \text{schwacher Zusammenhang}$$
$$V \in [0,3;0,6] \rightarrow \text{mittlerer Zusammenhang}$$
$$V \in [0,6;1,0] \rightarrow \text{starker Zusammenhang}$$

Aufgrund dieser Normierung ist Cramers V eines der gebräuchlichsten Zusammenhangmaße und soll daher auch in dieser Arbeit zur Berechnung der Zusammenhangstärke zwischen abhängiger und unabhängiger Variable eingesetzt werden.

Zur Überprüfung der Hypothesen wird außerdem der Chi-Quadrat-Unabhängigkeitstest aus dem Gebiet der induktiven Statistik durchgeführt. Anhand einer Stichprobe wird dabei geprüft, ob die Behauptung, zwei Merkmale in der Grundgesamtheit seien voneinander unabhängig (Nullhypothese), aufrecht erhalten werden kann oder verworfen werden muss [vgl. Bourier 2009, S. 303f.]. Grundgesamtheit ist die Menge aller DBS-Besucher. Als Stichprobe liegt für einen Zeitraum von fünf Monaten eine Vollerhebung des Nutzerverhaltens vor.

Beim Chi-Quadrat-Unabhängigkeitstest wird ein Beibehaltungs- und ein Ablehnungsbereich definiert [vgl. Bourier 2009, S. 303f.]. Beide Bereiche sind durch den sog. kritischen Wert getrennt. Der wie oben dargestellt berechnete Wert für Chi-Quadrat (χ^2) wird anschließend in einen der beiden Bereiche eingeordnet. Liegt χ^2 im Beibehaltungsbereich, wird die Nullhypothese (H_0) beibehalten und die Alternativhypothese (H_1) verworfen.

Das Risiko, die Nullhypothese irrtümlich abzulehnen, kann mit der Wahrscheinlichkeit α (auch Signifikanzniveau genannt) nach oben begrenzt werden [vgl. Bourier 2009, S. 288]. Fällt χ^2 in den Ablehnungsbereich, wird die Abweichung vom Erwartungswert als signifikant angesehen. Bei Vorliegen einer signifikanten Abweichung wird davon ausgegangen, dass die Abweichung zu groß ist, um als zufallsbedingtes Abweichen erklärt werden zu können. Gebräuchliche Werte für das Signifikanzniveau sind 1%, 5% und 10%, meist wird eine Fehlerwahrscheinlichkeit von 5% verwendet.

Der kritische Wert ist das Quantil der χ^2-Verteilung für das gewählte Signifikanzniveau und der Anzahl der Freiheitsgrade für die Kontingenztabelle:

$$\chi^2_{1-\alpha,\ k=(i-1)\cdot(j-1)} \tag{7.5}$$

Wird beispielsweise eine Fehlerwahrscheinlichkeit von 5% angenommen und ist die Kontingenztabelle, für die Chi-Quadrat berechnet wurde, drei Spalten mal vier Zeilen groß, ergibt sich ein kritischer Wert $\chi^2_{1-0,05,k=(3-1)\cdot(4-1)} = \chi^2_{0,95,\ 6} = 12,592$[1]. Für Chi-Quadrat ergibt sich ein Beibehaltungsbereich von $[0;\ 12,592]$, in dem die Nullhypothese beibehalten wird [vgl. Bourier 2009, S. 304].

Mit dem Unabhängigkeitstest wird also überprüft, ob die Hypothese einer bestehenden Unabhängigkeit zweier Variablen bei einem zu bestimmenden Fehlerrisiko auch für die Grundgesamtheit beibehalten werden kann. Die Stärke des Zusammenhangs beider Variablen wird durch Cramers V quantifiziert [vgl. Raithel 2008, S. 139].

Die weiteren Schritte im Forschungsablauf entsprechen den Schritten beim Web Usage Mining (vgl. 6.1). Sie werden ab dem folgenden Abschnitt beschrieben.

[1] Das Quantil kann in Tabellen der χ^2-Verteilung abgelesen oder mittels einer Tabellenkalkulation abgerufen werden.

8. Durchführung der Datenanalyse

Nach erfolgter Konzeption wird in diesem Kapitel die praktische Umsetzung der Datenanalyse dargestellt.

8.1. Art und Umfang der vorhandenen Daten

Als Datengrundlage für die Auswertung werden die mittels Piwik erhobenen Daten verwendet. Sie liegen im Gegensatz zu Urchins proprietärer Datenhaltung in einer frei manipulierbaren und in ihrer Struktur dokumentierten MySQL-Datenbank vor. Zudem bestimmt Piwik durch die Verwendung von Cookies und die Erfassung von Klicks auf externe Links Sitzungen präziser als das Logfile-basiert arbeitende Urchin. Ein weiterer Vorteil gegenüber Urchin liegt in der Beschränkung bei der Datenerfassung auf relevante Nutzeraktionen. Bei der neben Piwik und Urchin dritten Möglichkeit eines Datenzugangs in Form einer direkten Verarbeitung der Logfiles müssten große Datenmengen verarbeitet und viele Hits herausgefiltert werden, weshalb auf diese Möglichkeit verzichtet wird. Für die Auswertungen steht der Erhebungszeitraum Juli bis November 2009 zur Verfügung. Die Verallgemeinerbarkeit der aus diesen Daten gewonnenen Aussagen wird mittels statistischer Verfahren überprüft.

8.2. Datenvorverarbeitung

In einem ersten Schritt wurden Anfang Dezember die Tabellen `piwik_log_action`, `piwik_log_link_visit_action` und `piwik_log_visit` exportiert und auf dem Auswertungsrechner wieder importiert. Die Datenbank enthielt rund 9,1 Millionen Datensätze und war etwa 1 Gigabyte groß.

Entsprechend der unter 6.1.1 dargestellten Vorgehensweise wird eine Datenbereinigung durchgeführt. Dabei werden zunächst Sitzungen und damit verbundene Aktionen außerhalb des Betrachtungszeitraums vom 1. Juli bis 30. November 2009 aus den

Tabellen `piwik_log_visit` und `piwik_log_link_visit_action` entfernt[1]. Betroffen waren insgesamt ca. 163.500 Datensätze.

Während des Betrachtungszeitraums wurden vier Piwik-Updates eingespielt (Versionen 0.4.2 bis 0.4.5). Eine Überprüfung des Änderungsprotokolls[2] ergab keine grundlegenden Änderungen des Tracking-Prozesses. Die Änderungen in diesem Bereich beschränkten sich auf die Suchmaschinen-Erkennung bei den Referrern und die Robot-Erkennung.

Zwischen dem 19.11.2009 und dem Morgen des 23.11.2009 fiel die Datenerfassung in Piwik aufgrund eines Datenbankfehlers aus. Unter der Annahme, dass es an diesen Tagen keine bedeutenden Abweichungen im Nutzungsverhalten gab, wird keine Interpolation der Daten vorgenommen.

In einem zweiten Schritt wird überprüft, ob bestimmte Nutzergruppen aufgrund ihrer Besuchstiefe von der Auswertung auszuschließen sind. Grund hierfür ist der allgemeine Trend zur Ausführung von JavaScript durch Robots, so dass deren Nichterfassung durch das JavaScript-basierte Piwik nicht mehr vorausgesetzt werden kann. Dazu wird die durchschnittliche Anzahl der Aktionen je Sitzung und deren Standardabweichung ermittelt, die 3,47 respektive 6,21 betragen. Anschließend werden diese Werte nach dem Feld `location_provider` segmentiert. Dieses Feld enthält den Hostnamen, über den der DBS-Nutzer mit dem Internet verbunden war[3]. Die Abfrage wird gefiltert nach einer durchschnittlichen Anzahl von Aktionen größer als zehn, um nur deutliche Abweichungen vom Mittelwert zu erfassen. Absteigend nach Anzahl der Sitzungen sortiert ergibt sich das in Tabelle 8.1 ausschnittsweise dargestellte Ergebnis.

Eine genauere Betrachtung der von `googlebot.com` generierten Sitzungen zeigt, dass diese mit einem Piwik-Update am 24. August enden, die fehlende Filterung des Google-Robots also mit einer neuen Piwik-Version ergänzt wurde. Die hohen Besuchstiefen von `dipf.de` und `fwu.de` lassen sich mit der Tätigkeit der DBS-Redakteure beim DIPF und dessen Partnerinstitut FWU[4], das Teile des DBS redaktionell betreut, erklären. Zwar ist nicht auszuschließen, dass Mitarbeiter beider Institute den DBS auch für Recherchen nutzen, doch aufgrund dieser weit überdurchschnittlichen Besuchstiefe werden ihre Besuche gemeinsam mit den von `googlebot.com` ausgehenden entfernt. Bei den weiteren Hostnamen aus Tabelle 8.1 ist nach einer intellektuellen Sichtung nicht von einer automatisierten Nutzung auszugehen. Nach

[1] Piwik wurde beim DBS am 29. Juni 2009 in Betrieb genommen.

[2] `http://piwik.org/changelog/`; Stand: 14. Dezember 2009

[3] Piwik benutzt dazu die PHP-Funktion `gethostbyaddr()`. Gespeichert werden nur die ersten beiden Ebenen des Hostnamens, beispielsweise `t-ipconnect.de` statt `p4FE998DC.dip0.t-ipconnect.de`. Lässt sich der Hostname nicht ermitteln, wird „Ip“ eingetragen.

[4] FWU Institut für Film und Bild in Wissenschaft und Unterricht, München

Durchschn. Besuchstiefe	Standardabw.	Anz. Sitzungen	Hostname
20,42	53,39	5509	dipf.de
15,12	29,52	468	googlebot.com
15,99	30,40	345	fwu.de
10,16	41,34	68	caritas.de
11,54	12,86	24	weimar-klassik.de
11,32	25,64	22	rwi-essen.de
14,74	14,21	19	bow.de
18,89	58,92	18	school-scout.de
11,88	25,07	17	outermedia.de
10,57	12,72	14	lear.com

Tabelle 8.1.: Häufigste Hostnamen mit einer überdurchschnittlicher Besuchstiefe

dieser Bereinigung verbleiben rund 1,9 Mio. Sitzungen und rund 6,5 Mio. Aktionen zur weiteren Auswertung.

Im Zuge einer Datenfusion werden anschließend Kategorie und Seitentitel der Redaktionsseiten als semantisches Datum ergänzt. Datengrundlage ist eine von den DBS-Webmastern bereitgestellte Textdatei mit einer Zuordnung zwischen Seitennummer und einem virtuellen Seitenpfad, welcher sich aus der mit dem Redaktionssystem erzeugten Verschachtelung der Redaktionsseiten ergibt. Das folgende Listing enthält zwei Beispielzeilen.

```
71      /Uebergreifende_Informationen/Bildungsrecht/↘
→Grundgesetz_und_Landesverfassungen.html
5715    /Schule/Unterricht/Online-↘
→Angebote_fuer_faecheruebergreifende_Themenbereiche/Berufsorientierung.html
```

Listing 8.1: Auszug aus der Datei zur Pfadzuordnung

Demnach ist die Redaktionsseite mit dem Titel „Grundgesetz und Landesverfassungen“ eine Unterseite der Seite „Bildungsrecht“, die wiederum direkt dem Themenbereich „Übergreifende Informationen“ zugeordnet ist. Zu beachten ist, dass Redaktionsseiten in mehr als eine andere Redaktionsseite als Unterseite „eingehängt“ werden können und so eine Polyhierarchie entsteht. In der zur Datenfusion verwendeten Liste wird jedoch nur die primäre Einordnung abgebildet. Der Pfad wird in die neue Spalte „path“ der Tabelle `piwik_log_action` geschrieben. Um die spätere Auswertung des Pfades zu vereinfachen, werden zwei weitere Spalten mit Angaben zur Tiefe und zum Bereich ergänzt. Die Spalte „depth“ enthält die Hierarchieebene als Zahlenwert, wobei 1 für die oberste Ebene der Startseite steht. Themen- und Zielgruppenseiten werden als zweite Ebene betrachtet. Dementsprechend ergibt sich die Tiefe aus der Anzahl der Trennzeichen (Schrägstrich) der oben beschriebenen Pfade plus eins. Desweiteren wird die erste Pfadebene in einer neuen Spalte „cat“ kodifiziert, um Wechsel zwischen Themenbereichen besser erfassen zu können (siehe

Tab. 12.4 auf Seite A7). Diese Kodifizierung wird im Folgenden als „Inhaltskategorie“ bezeichnet. Seiten der DBS-internen Suchfunktionen werden mit der Kategorie 400 und der Tiefe 2 kodifiziert, um Sprünge zwischen Themen- bzw. Zielgruppenzugang und Suche identifizieren zu können.

Zusätzlich wird eine Kodifizierung der Seitenklassen vorgenommen. „Seitenklasse“ lässt sich weitgehend mit „Datenbank“ gleichsetzen, da für jede Unterdatenbank des DBS mindestens eine Seitenklasse zur Darstellung der jeweiligen Einträge existiert. `/db/mlesen.html` beispielsweise repräsentiert die Seitenklasse der Onlineressourcen, während durch Anfügen eines Parameters in `/db/mlesen.html?seite=123` eine einzelne Seite dargestellt wird. Da einer Datenbank häufig mehrere Seitenklassen zugeordnet werden können, erfolgt zur Vereinfachung der Auswertung eine nummerische Kodifizierung mittels Regulärer Ausdrücke. Dabei handelt es sich um ein quasi-standardisiertes und von einer großen Zahl von Skriptsprachen und Anwendungsprogrammen unterstütztes System zur Festlegung von Such- und Ersetzungsmustern unter Verwendung von Zeichen- und Symbolklassen und Platzhaltern [vgl. Friedl 2006, S. 4f.]. Die Aufteilung der Kodes, Regulären Ausdrücke und Funktionstypen ist in Tabelle 12.5 (S. A8) dargestellt. Die Regulären Ausdrücke werden auf die Spalte „name“ der Tabelle `piwik_log_action` angewendet und die Kodes in eine neue Spalte „idpagetype“ derselben Tabelle geschrieben. Zur Unterscheidung von der Inhaltskategorie wird diese zweite Kodifizierung als „Funktionskategorie“ bezeichnet.

Weitere Datenarten, die eine Verknüpfung mit den Nutzungsdaten erlauben würden, sind nicht verfügbar, weswegen keine Datenintegration durchgeführt wird.

8.3. Ermittlung der Webanalyse-Metriken

Ziel der Auswertung ist es, festzustellen, ob sich das Verhalten jener DBS-Nutzer, die den DBS über Suchmaschinen erreichen, von dem Besucherverhalten der übrigen Nutzer unterscheidet. Dazu werden zunächst die Metriken der Webanalyse ermittelt und anschließend Methoden eingesetzt, die eher dem Web Usage Mining zuzuordnen sind. In diesem Abschnitt wird die Durchführung des ersten Auswertungsteils dargestellt. Er umfasst die Ermittlung der Zeitverteilung, Anzahl der Aktionen, Sitzungsdauer, Absprünge, wiederkehrender Besucher sowie Ausstiegsarten und die Segmentierung dieser Metriken nach der Einstiegsart.

Gemäß der unter 5.2.2 dargestellten Funktionsweise von Piwik zählen als Aktion auch Klicks auf externe Links sowie auf Binärdateien. Zwar handelt es sich dabei nicht um Seitenabrufe im klassischen Sinn, jedoch um bewusste Handlungen der Nutzer. Da

eine Erfassung dieser Klicks zudem sehr gut dem Verweisserver-Charakter des DBS entspricht, werden sie im Rahmen dieser Auswertung nicht gesondert berücksichtigt, sondern gemeinsam mit den klassischen Seitenabrufen als „Aktion“ bezeichnet.

Häufigkeiten, prozentuale Verteilungen, Mittelwerte und Streuungsmaße der Metriken können mittels SELECT-Abfragen in MySQL ausgelesen oder berechnet werden. Zur Berechnung der Assoziationsmaße jedoch werden eine Reihe von Datensets aus der Datenbank ausgelesen und über das Austauschformat CSV in die Statistiksoftware PASW importiert. Die einzelnen Abfragen beeinhalten jeweils:

1. Die Tabelle `piwik_log_visit` mit Ausnahme von `visitor_idcookie`, `location_ip` und den `config_`-Spalten.
2. Angaben zur ersten Aktion einer Sitzung, darunter URL, Seitentyp, Inhaltskategorie, Funktionskategorie, Hierarchieebene und Daten zur dazugehörigen Sitzung wie Einstiegstyp, Anzahl der Aktionen und Sitzungsdauer.
3. Die gleichen Angaben wie unter 2., jedoch bezogen auf die letzte Aktion einer Sitzung.

In PASW wird für univariate Auswertungen die Funktion „Häufigkeiten“ aus dem Menü „Deskriptive Statistiken“ verwendet. Sie erlaubt auch die Erzeugung von Histogrammen mit Darstellung der Normalverteilung. Die Funktion „Kreuztabellen“ aus demselben Menü findet bei der Untersuchung bivariater Zusammenhänge Anwendung. Spaltenweise wird stets die Variable Einstiegsart, zeilenweise die jeweils abhängige Variable angeordnet. Diese PASW-Funktion berechnet zugleich auch Cramers V und Chi-Quadrat, nicht jedoch den kritischen Wert für den Chi-Quadrat-Unabhängigkeitstest. Weiterhin werden einige Variablen, darunter die Sitzungsdauer, klassiert, d.h. die Vielzahl der Einzelwerte durch Zusammenfassung zu Wertegruppen reduziert. Dies geschieht in PASW mittels der Funktion „Visuelles Klassieren“ aus dem Menü „Transformieren“. Die sekundengenaue Angabe des Sitzungsbeginns beispielsweise lässt sich auf diese Weise auf einen Stundenbereich abstrahieren.

Die Abfrage- und Auswertungsergebnisse aus MySQL und PASW werden in der Tabellenkalkulation Excel zwecks Weiterverarbeitung und Visualisierung zusammengeführt. Das Programm ermöglicht außerdem die Ermittlung des kritischen Wertes für den Chi-Quadrat-Unabhängigkeitstest mittels der Funktion `CHIINV`.

8.4. Auswertungen mit Elementen des Web Usage Mining

Grundlage dieser erweiterten Auswertungen sind die im Zuge der Datenvorbereitung ergänzten Daten über Pfad, Hierarchieebene, Inhaltskategorie und Funktionskategorie der Redaktionsseiten des DBS. Da diese Daten bereits mit der im vorherigen Abschnitt beschriebenen Datenübertragung von MySQL nach PASW portiert wurden, wird die Statistiksoftware auch in diesem zweiten Auswertungsteil zur Generierung der Kreuztabellen verwendet. Mit diesen Werkzeugen lassen sich Einstiege und Absprünge bezogen auf Inhalts- und Funktionskategorien ermitteln.

8.4.1. Auswertung der Bewegung durch die Hierarchieebenen

Im Gegensatz zu den bisher beschriebenen Metriken ist zur Ermittlung von Bewegungsmustern nicht nur eine Betrachtung der Ein- und Ausstiegspunkte einer Sitzung erforderlich, stattdessen sind auch einzelne Aktionen innerhalb der Sitzungen zu

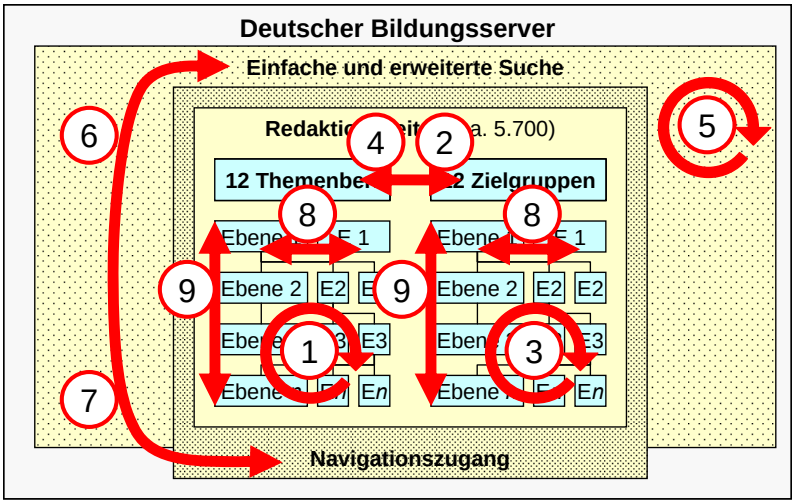


Abbildung 8.1.: Untersuchungsgegenstände der DBS-internen Navigation

Zur Ermittlung der Auf-, Ab- und Seitwärtsbewegungen innerhalb der Hierarchie der Redaktionsseiten (Punkt 9 in Abb. 8.1) wird ein PHP-Skript erstellt, dessen Ablauf im Folgenden beschrieben wird. Für jeden Einstiegstyp werden die über ihn eingeleiteten Sitzungen mit einer Gesamtaktionszahl von mindestens zwei ermittelt. Nur bei Sitzungen mit zwei oder mehr Aktionen kann eine Navigationsbewegung innerhalb des DBS stattgefunden haben. Für jede der selektierten Sitzungen wird im zweiten Schritt überprüft, ob ihr mindestens zwei Aktionen mit einer Angabe der Hierarchieebene und einer Inhaltskategorie kleiner als 400 zugeordnet sind. Diese

Einschränkungen sind erforderlich, da nur zu bestimmten Seiten eine Hierarchieebene vorliegt und die mit Werten ab 400 kategorisierten Inhaltsbereiche nicht zu den hierarchisch strukturierten Themen- und Zielgruppenseiten zählen. Sind diese Bedingungen erfüllt, kann ab der zeitlich zweiten Aktion einer Sitzung innerhalb einer Inhaltskategorie die Differenz der Hierarchieebene zur jeweils unmittelbar oder mittelbar vorhergehenden Aktion ermittelt werden. Durch das Erlauben einer Mittelbarkeit der Aktionen werden Hierarchiebewegungen auch dann erfasst, wenn beispielsweise zwischen zwei Redaktionsseiten (mit Hierarchie- und Kategoriezuordnung) die Detailansicht eines Datensatzes (ohne Hierarchie- und Kategoriezuordnung) erfolgt.

Zu jeder derart ermittelten Sitzung wird ein Datenbankeintrag mit folgenden Angaben angelegt:

- Der Einstiegstyp mit den Werten 1 (Direkteinstieg), 2 (Suchmaschineneinstieg) oder 3 (sonstiger Verweis).
- Eine kodifizierte Beschreibung des Musters. Bestandteile sind:
  - Die Art der Bewegung, wobei Aufwärts-, Abwärts- und Seitwärtsbewegungen respektive mit `u` (up), `d` (down) und `e` (equal) kodiert werden.
  - Die Differenz der Ebenen als nummerischer Wert

  Die Musterbeschreibung `u2e1d1` steht demnach für eine Aufwärtsbewegung um zwei Ebenen, gefolgt von einer Seitwärtsbewegung und einer Abwärtsbewegung um eine Ebene.
- Eine vereinfachte kodifizierte Musterbeschreibung ohne Angabe der Ebenendifferenz, z.B. `ued`.
- Je einem Zahlenwert für die Anzahl der Auf-, Ab- und Seitwärtsbewegungen.

Zur Bestimmung der häufigsten Muster können die auf diese Weise erstellten Datenbankeinträge nach einfachen oder komplexen Mustern gruppiert werden.

8.4.2. Auswertung der Bewegungen zwischen Navigation und Suche

Für die Ermittlung der häufigsten Bewegungsmuster innerhalb der Navigationsstruktur und Suchfunktion werden drei Bereiche identifiziert: Zielgruppenseiten, Themenseiten und Suchfunktion. Wie in Abbildung 8.1 (S. 69) dargestellt, werden folgende Bewegungen ausgewertet:

- Zielgruppenseite zu Themenseite und umgekehrt (Punkte 4 und 2 in der Abbildung)

- Navigation innerhalb eines Bereiches
  - Zielgruppen- zu Zielgruppenseite (Punkt 3)
  - Themen- zu Themenseite (Punkt 1)
  - Such- zu Suchseite (Punkt 5)
- Navigationsbereich (Zielgruppen- und Themenseite) zu Suchseite und umgekehrt (Punkte 6 und 7)

Zunächst wird ein PHP-Skript erstellt, das die oben beschriebenen Bewegungsmuster nach Einstiegsart erfasst und zählt. Dabei werden analog zur Vorgehensweise bei den Hierarchieebenen auch indirekte Abfolgen berücksichtigt, d.h., dass zwischen dem Abruf inhaltskategorisierter Seiten auch andere Aktionen stattfinden können.

Ein zweites Skript dient der exakteren Abgrenzung von Bewegungen innerhalb der Navigationsstruktur. Dabei werden nicht die Bereiche Zielgruppen und Themen jeweils als Ganzes betrachtet, sondern Wechsel zwischen einzelnen Kategorien innerhalb der Bereiche (Punkt 8 in Abb. 8.1) von bereichsübergreifenden Bewegungen (Punkte 2 und 4) abgegrenzt.

9. Darstellung und Interpretation der Ergebnisse

In diesem Kapitel werden die Auswertungsergebnisse dargestellt und interpretiert. Zunächst erfolgt dies für die Ergebnisse der Webmetrik-Ermittlung, anschließend für die Nutzungsmuster-Ermittlung.

9.1. Ergebnisse der Webmetrik-Ermittlung

374.843 Sitzungen – 19,9% – im Betrachtungszeitraum Juli bis November 2009 beginnen mit einem direkten Aufruf des DBS. Suchmaschinenverweise erzeugen 1.291.313 Sitzungen bzw. 68,7%. Den geringsten Anteil mit 214.601 Sitzungen bzw. 11,4% haben sonstige Verweise. Insgesamt fanden im Betrachtungszeitraum 1.880.757 Sitzungen statt.

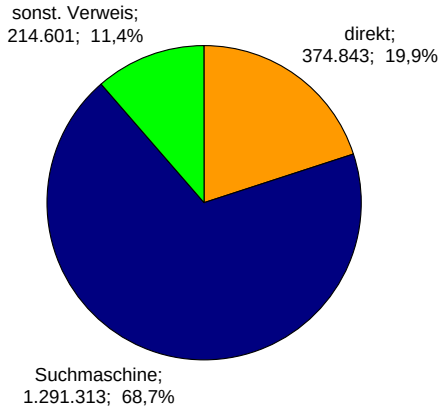


Abbildung 9.1.: Sitzungen nach Einstiegsart

Das Besucheraufkommen des DBS hängt in hohem Maße von Suchmaschinen ab. Der Webstatistik-Dienstleister Hitwise weist in einer branchenbezogenen Darstellung die Anteile von Suchmaschinen am gesamten Besuchsaufkommen für April 2009 aus [vgl. Dougherty 2009]. Den höchsten Anteil haben Suchmaschinen mit 50% aller Sitzungseinstiege in der Bildungsbranche. Für den Regierungssektor, dessen Websites

aufgrund ihres tendenziell informationsvermittelnden Charakters potentiell ebenfalls für einen Vergleich mit dem DBS relevant sind, liegt der Suchmaschinenanteil am Gesamtaufkommen bei 31%. Auch wenn die Zusammensetzung der Branchen nicht aus dem Bericht hervorgeht und sich auch leichte Abweichungen aus eventl. unterschiedlichen Bezugseinheiten ergeben können, ist doch deutlich, dass der Suchmaschinenanteil beim DBS mit fast 69% weit über diesen Durchschnittswerten liegt. Daraus ergibt sich eine überdurchschnittliche Abhängigkeit von den Suchdiensten, da sich das von ihnen generierte Besucheraufkommen nur sehr begrenzt vom Website-Betreiber beeinflussen lässt und sich beispielsweise durch Änderungen an der Ergebnisrangfolge oder -darstellung des Suchmaschinenbetreibers verringern kann.

Nach dieser Bestimmung der Anteile der drei Einstiegsarten an den DBS-Sitzungen werden im weiteren Verlauf die einzelnen Webmetriken ermittelt und nach den Einstiegsarten segmentiert. Dem Ablauf einer Websitzung folgend wird zunächst der Einstieg in den DBS beschrieben.

9.1.1. Einstieg in den DBS

Suchmaschinen

Eine Betrachtung der Anteile einzelner Suchmaschinen unter den Suchmaschineneinstiegen zeigt, dass Google mit 94% den weitaus höchsten Anteil hat (vgl. Abb. 12.4, S. A9). Auf Platz zwei folgt mit 1,6% die ebenfalls von Google betriebene Websuche des Zugangsanbieters T-Online. Den dritten Rang nimmt der im Juni 2009 von Microsoft gestartete Suchdienst Bing ein. Die eigenständig ausgewiesene Bildersuche von Google ist Ausgangspunkt für 0,7% der Suchmaschineneinstiege beim DBS (Rang fünf). Auf Rang sechs folgt mit 0,5% Yahoo. Der Suchdienst Ask liegt mit 0,2% auf Rang acht. Die bisher nicht genannten der zehn häufigsten Suchdienste greifen wie T-Online auf Google zurück. Eine Ausnahme hiervon ist Conduit auf Rang zehn, das zwar ebenfalls Google nutzt, jedoch eine Browser-Funktionsleiste mit Suchfunktion anbietet. Weitere 52 Suchdienste mit einer jeweils geringeren Anzahl von Verweisen als Conduit machen insgesamt 0,3% der Suchmaschineneinstiege beim DBS aus. Einer vom Webanalyse-Dienstleister Webtrekk GmbH durchgeführten Erhebung[1] zufolge liegt der Marktanteil von Google in Deutschland bei 87,2%, gefolgt von T-Online (3,1%), Bing (2,1%) und Yahoo (1,2%) [vgl. Webtrekk GmbH 2009, S. 2]. Verglichen mit dieser Verteilung liegen die Anteile beim DBS für Google über

[1] Datengrundlage sind Zugriffe auf Websites von Webtrekk-Kunden mit insgesamt 105 Mio. Besuchern. Erhebungszeitraum ist das dritte Quartal 2009.

dem Durchschnitt und für die anderen genannten Anbieter darunter. Dies führt zu einer höheren relativen Abhängigkeit vom ohnehin stark dominanten Marktführer Google. Ein Verbesserungspotential ergibt sich hieraus hinsichtlich der Optimierung des DBS für jene Google-Konkurrenten mit eigener Datenbasis, insbesondere Bing[2].

Tabelle 12.6 auf Seite A9 zeigt die absoluten Häufigkeiten der Suchmaschineneinstiege sowie die durchschnittliche Anzahl der Aktionen und die durchschnittliche Dauer dieser Sitzungen jeweils mit ihrem Mittel, ihrer Standardabweichung und ihrem Maximalwert. Besonders auffällig sind die von der Google-Bildersuche (Google Images) ausgelösten Sitzungen mit einer sehr geringen durchschnittlichen Besuchstiefe und -dauer. Eine Auswertung der Einstiegsseiten der über Google Images gestarteten Sitzungen zeigt einen besonderen Effekt. Die Google-Bildersuche führt ausschließlich auf Seiten aus dem Bereich ELIXIER[3]. Dabei handelt es sich um ein Projekt zum Datenaustausch zwischen dem Deutschen Bildungsserver und den Landesbildungsservern, dessen bisheriger Arbeitsstand zwar noch nicht gegenüber Nutzern beworben wird, aber bereits über das Internet erreichbar und auch für Suchmaschinen nicht gesperrt ist. In Googles Bildersuche sind Grafiken aus dem Bestand einiger Kooperationspartner indexiert, die über `www.bildungsserver.de/elixier/` erreichbar sind. Da es sich bei ELIXIER im Wesentlichen um eine Suchmaske handelt und es keine Verknüpfungen zum originären DBS gibt, sind die Sitzungen in diesem Bereich entsprechend kurz.

Auffällig ist weiterhin, dass Sitzungen, die über die nicht-Google-basierten Suchdienste Bing, Yahoo und Ask eingeleitet werden, durchschnittlich mehr Aktionen ($\overline{a}$) aufweisen als die Google-basierten unter den nach Anzahl der Einstiege zehn ranghöchsten Suchdiensten, dies allerdings mit einer höheren Standardabweichung (vgl. Abb. 12.5, S. A11). Somit lässt sich sagen, dass ein gewisser Teil der Nutzer dieser Suchmaschinen deutlich mehr Aktionen ausführt als Google-Nutzer. Die durchschnittliche Sitzungsdauer ($\overline{t}$) ist bei Einstiegen über die drei genannten Google-Konkurrenten nur geringfügig höher als bei Google-Einstiegen, die höchste Dauer weisen die Einstiege via Ask auf. Unter den Suchmaschinen mit einer geringeren Zahl von Verweisen erreichen Sitzungen via AltaVista und Meinestadt.de hohe Werte für ($\overline{a}$) und ($\overline{t}$) auf. Die durchschnittliche Betrachtungsdauer $\left(\overline{t/a}\right)$ je Aktion variiert kaum unter den zehn häufigsten verweisenden Suchmaschinen mit Ausnahme der Google-Bildersuche. In dieser Auswahl von Suchdiensten ist sie bei den Yahoo-Einstiegen am geringsten.

[2] Aufgrund einer im Dezember 2009 getroffenen Kooperationsvereinbarung zwischen Microsoft und Yahoo wird die Suchtechnologie von Bing nach wettbewerbsrechtlicher Freigabe zukünftig auch bei Yahoo zum Einsatz kommen [vgl. heise online 2009].

[3] Elaborated Lists in XML for Internet Educational Ressources

Die geschilderten Daten lassen sich so deuten, dass die unmittelbaren Google-Konkurrenten Bing, Yahoo und Ask zwar deutlich weniger DBS-Sitzungen generieren, diese aber eine verglichen mit Google etwas höhere Nutzungsintensität aufweisen.

Suchphrasen bei Suchmaschineneinstiegen

Die gleichartige Auswertung der Suchphrasen (Tab. 12.7 auf S. A12) zeigt, dass die meisten Suchmaschineneinstiege über die schon sehr spezifische Suchphrase „bildungsserver“ erfolgen. Die Anzahl der Aktionen und die Besuchszeit dieser Sitzungen liegen weit über dem Gesamtdurchschnitt. Noch höher liegen diese Werte bei der noch spezifischeren Suchphrase „deutscher bildungsserver“ (Rang vier). Dies kann so interpretiert werden, dass jene Besucher das Konzept eines Bildungsservers oder den DBS im Speziellen kennen und den DBS statt über seine URL gezielt über eine Suchmaschine aufrufen. Insgesamt 2.200 Sitzungen begannen sogar mit „www.bildungsserver.de“ bzw. „bildungsserver.de“ als Suchphrase (Rang 43 bzw. 50). Sofern sich den Einstiegsseiten eine Hierarchieebene zuordnen lässt, was hauptsächlich auf Startseite und Redaktionsseiten zutrifft, wird dies in der drittletzten Tabellenspalte gezählt und die durchschnittliche Hiearchieebene in der vorletzten Spalte als Durchschnittswert angegeben. Die durchschnittliche Einstiegsebene der vier genannten Suchphrasen ist mit 1,1 bis 1,3 sehr hoch, was mit der Beobachtung konform geht, dass diese Suchphrasen bei Google auf die DBS-Startseite (Ebene 1) führen. Sämtliche weiteren Suchphrasen führen zu tieferen Einstiegen in die DBS-Struktur. Bei den Suchphrasen „hamburger bildungsserver“ und „bildungsserver hamburg“ (Ränge 8 und 10) muss davon ausgegangen werden, dass der DBS nicht das anvisierte Ziel der Suchmaschinenbenutzung war. Dennoch liegen die durchschnittlichen Seitenabrufe bei Sitzungen mit diesen beiden Suchphrasen über dem Gesamtdurchschnitt (4,2 bzw. 5,2 zu 3,4). Interessant hierbei ist, dass beide Suchanfragen bei Google an zweiter bzw. erster Stelle auf dieselbe DBS-Redaktionsseite verweisen, die einen direkten Link zum Hamburger Bildungsserver[4] enthält. Ein direktes Verlassen über diesen Link würde nur zwei Aktionen auslösen. Viele Besucher, die den DBS auf diese Weise erreichen, führen also noch weitere Aktionen dort aus.

Besonders viele Aktionen erfolgen durchschnittlich bei Sitzungen via Suchphrasen, die den inhaltlichen Themen „Unterrichtsmaterial“ (Ränge 12 und 40), „Referendariat“ (Ränge 19 und 37), „Erwachsenenbildung“ (Rang 14), „Stellenangebote“ (Rang 45) und „Wettbewerbe“ (Rang 46) zuzuordnen sind. Bei den Themen „Referendariat“ und „Wettbewerbe“ liegt auch die durchschnittliche Ansichtsdauer je Aktion deutlich über dem Gesamtdurchschnitt, bei den anderen Themen liegt sie darunter, am deutlichsten bei „Unterrichtsmaterial“.

[4] `http://www.hamburger-bildungsserver.de/`

Für eine eingehendere Auswertung erschiene es sinnvoll, zunächst eine intellektuelle Zusammenfassung der unterschiedlichen Suchphrasen zu Themenblöcken vorzunehmen, um beispielsweise (Quasi-)Synonyme wie „Referendariat“ und „Vorbereitungsdienst“ (Ränge 19 und 37) zusammenzufassen. Eine solche Auswertung könnte in einer Folgearbeit vorgenommen werden.

Verweisende Seiten

Die häufigsten Verweisseiten sind analog zu den Suchphrasen in Tabelle 12.8 auf Seite A14 aufgeführt. Trotz des gewählten Einstiegstyps „verweisende Seiten“ liegt die Suchmaschine Google auf Rang 4 mit 5.900 Sitzungen. Eine Auswertung der vollständigen Referrer-URLs zeigt, dass ein Teil dieser Verweise fälschlicherweise von Piwik den Verweisseiten zugeordnet wurde, obwohl es sich um Suchanfragen an Google handelt. Ursache hierfür dürfte eine abweichende Variablenbezeichnung der Suchphrase seitens Google sein[5]. 28,6% der Google-Verweise sind hiervon betroffen. Ein Grund für diese Abweichung ist nicht erkennbar, Funktionsunterschiede zwischen den beiden Variablenbezeichnungen treten bei der Suchmaschine nicht auf[6]. Weitere Google-Verweise lassen sich der Startseite des Suchdienstes, der personalisierbaren Suchoberfläche („iGoogle“) sowie dem Online-RSS-Aggregator „Google Reader“ zuordnen.

Eine weitere Betrachtung der Tabelle zeigt, dass viele Einstiege beim DBS über verwandte Portale des IZB bzw. DIPF erfolgen. Dazu zählen die Wikis und Blogs des DBS sowie die eigenständigen Websites Fachportal Pädagogik, Schulweb, Bildungsbericht, Lesen in Deutschland und die DIPF-Website. Die durchschnittliche Aktionszahl bei den über das Fachportal Pädagogik oder den DIPF-Internetauftritt eingeleiteten Sitzungen liegt mit 5,0 bzw. 7,8 über dem Gesamtdurchschnitt von 3,9. Auch die DBS-Besuchsdauer ist bei den beiden Verweisseiten vergleichsweise hoch. Bei den anderen IZB-Portalen liegt die Aktionszahl nahe dem Gesamtdurchschnitt. Insbesondere das Fachportal Pädagogik und die DBS-Wikis sind somit aufgrund der Vielzahl der über sie erfolgten DBS-Einstiege wichtige Verweisseiten.

Auffällig ist weiterhin, dass Sitzungen mit Verweisen aus dem Bereich Arbeitsmarkt/Stellensuche (Ränge 22, 28, 46 und 50), darunter der Internetauftritt der Arbeitsagentur, eine hohe Anzahl von Seitenabrufen enthalten. Eine Ausnahme hiervor

[5] Statt `http://www.google.de/search?q=suchphrase` lauten die betroffenen Referrer-URLs `http://www.google.de/search?as_q=suchphrase`

[6] Eine Korrektur der vorliegenden Daten wird nicht vorgenommen, bei zukünftigen Analysen sollte dieser Effekt jedoch bei der Datenvorverarbeitung berücksichtigt werden.

stellt die „Jobsuchmaschine“ „Kimeta“ (Rang 5) dar, die unter anderem auf die Stellenbörse des DBS zurückgreift und auf ihre Datensätze verweist. Die durchschnittliche Aktionszahl beträgt hier nur 2,4. Es kann also davon ausgegangen werden, dass Kimeta-Nutzer überwiegend lediglich die verlinkten Stellenangebote betrachten und keine weiteren Aktionen ausführen. Weitere häufige Verweisseiten mit weit überdurchschnittlichen Aktionszahlen und Sitzungsdauern sind die des Ministeriums für Schule und Weiterbildung in Nordrhein-Westfalen (Rang 10 mit durchschnittlich 9,2 Aktionen) und das Datenbank-Infosystem (DBIS) der Universität Regensburg (Rang 15; 7,2 Aktionen).

Tabelle 12.8 kann, ggf. ergänzt um weitere Zugriffsseiten und die aus Kapazitätsgründen nicht in dieser Arbeit aufgeführten vollständigen Referrer, z.B. bei den Verweisen von anderen DBS- und IZB-Portalen sowie Wikipedia, von der DBS-Redaktion dazu verwendet werden, nach eigenem Empfinden fehlende oder unterrepräsentierte Verweisseiten zu identifizieren, um anschließend – soweit möglich – Einfluss auf die Verlinkung zu nehmen.

Zeitverteilung der Einstiege

Eine Auswertung der nach Einstiegsarten segmentierten tagesbezogenen Sitzungshäufigkeiten ergibt für den Auswertungszeitraum insgesamt steigende Nutzungszahlen (vgl. Abb. 12.6, S. A16). Erkennbar ist ein Tiefpunkt Anfang August, eine anschließende tendenzielle Zunahme der Sitzungszahlen bis Ende November mit einem leichten zwischenzeitlichen Rückgang im Oktober. Der Tiefpunkt im August lässt sich zeitlich den Sommerferien, der Rückgang im Oktober den Herbstferien zuordnen. Mit fünf Monaten ist der Betrachtungszeitraum zu kurz, um einen allgemeinen Trend bezogen auf die Zunahme der Sitzungen abzuleiten. Vielmehr muss aufgrund der allgemeinen Urlaubs- und Ferienzeit zu Beginn des Zeitraums von einem „Sommerloch“ ausgegangen werden. Zwischen Werktagen und Wochenenden sind ausgeprägte Auf- und Abwärtsbewegungen erkennbar mit einem deutlichen Rückgang der Sitzungszahlen am Wochenende. Auf diese Schwankung wird später genauer eingegangen.

Die relativen Anteile der drei Einstiegsarten bleiben im Betrachtungszeitraum weitgehend unverändert (vgl. Abb. 12.7, S. A17). Dies gilt besonders für den Anteil der sonstigen Verweisseiten, der bis auf zwei Ausschläge im Juli und Oktober fast gleich bleibt. Stärkere Schwankungen gibt es zwischen den Direkt- und Suchmaschineneinstiegen. Absolut wie relativ betrachtet fallen die zahlreichen über direkte Zugriffe eingeleiteten Sitzungen am 28. Juli auf. Diese Zunahme kann auch nach Rückfrage

bei der Redaktion keinem besonderen Ereignis, z.B. einer Pressemitteilung, zugeordnet werden und bleibt somit ungeklärt.

Bezogen auf die fünf Monate des Betrachtungszeitraums nimmt der Anteil der Suchmaschineneinstiege von 68% im Juli auf 70,8% im November zu, während der Anteil der Direkteinstiege von 11,9% auf 10,8% abnimmt (vgl. Abb. 12.8, S. A17). Auch der Anteil der sonstigen Einstiege nimmt leicht ab. Allerdings ist auch hier die Kürze des Betrachtungszeitraums zu beachten. So variieren die monatsbezogenen Anteile der Einstiegsarten zwischen Juli und November, so dass nicht von einem bestimmbaren Trend ausgegangen werden kann. Aussagekräftig erscheint erst ein monatsweiser Vergleich über mindestens zwei Jahre.

Am Beispiel der DBS-Einstiege je Einstiegstyp und Wochentag soll die Berechnung von Chi Quadrat und dem kritischen Wert gezeigt werden. Tabelle 12.9a (S. A18) ist die Kontingenztabelle mit den absoluten Häufigkeiten sowie Spalten- und Zeilensummen. Gemäß Formel 7.2 (S. 61) ergeben sich die erwarteten Häufigkeiten wie in Tabelle 12.9b dargestellt. Aus den Differenzen zwischen tatsächlichen und erwarteten Häufigkeiten wird nach Formel 7.3 (S. 61) Chi Quadrat berechnet (Tabelle 12.9c). Somit lässt sich Cramers V berechnen:

$$\begin{aligned}\text{Cramers V} &= \sqrt{\frac{\chi^2}{n \times (\min(k,m) - 1)}} \in [0;1] \\ &= \sqrt{\frac{844,4}{1.880.384 \times (3-1)}} \in [0;1] \\ &= 0,015\end{aligned}$$

Zur Durchführung des Chi-Quadrat-Unabhängigkeitstests werden die zu überprüfenden Hypothesen formuliert:

H_{1_0}: Die Verteilung der Sitzungen auf die Wochentage ist von der Einstiegsart unabhängig.
H_{1_1}: Die Verteilung der Sitzungen auf die Wochentage ist von der Einstiegsart abhängig.

Der kritische Wert für die Annahme der Nullhypothese lässt sich unter Annahme eines Fehlerrisikos von 5% und unter Verwendung der Spalten- und Zeilenzahl von Tabelle 12.9a in der χ^2-Verteilung ablesen:

$$\begin{aligned}\text{kritischer Wert} &= \chi^2_{1-\alpha,\ k=(i-1)\cdot(j-1)} \\ &= \chi^2_{1-0{,}05,\ k=(3-1)\cdot(7-1)} \\ &= \chi^2_{0{,}95,\ 12} \\ &= 21{,}026\end{aligned}$$

Nach dem Unabhängigkeitstest liegt mit $844{,}4 > 21{,}026$ eine statistisch signifikante Abhängigkeit zwischen beiden Variablen vor, die zur Ablehnung der Nullhypothese und zur Annahme der Alternativhypothese führt. Mit einem Wert von 0,015 für das Zusammenhangsmaß Cramers V für die Variablen Wochentag und Einstiegsart liegt kein bedeutsamer Zusammenhang zwischen den beiden Variablen vor. Dies ist so zu interpretieren, dass für die Grundgesamtheit eine signifikante Abhängigkeit besteht (die Abweichungen zwischen den Variablen also nicht zufallsbedingt sind), dieser Zusammenhang jedoch sehr gering ist.

Eine Betrachtung der Kontingenztabelle (12.9a, S. A18 und Abb. 12.9, S. A19) zeigt, dass die Häufigkeit der Sitzungen von Montag bis Samstag abnimmt und sonntags wieder zunimmt. Diese grundsätzliche Differenz zwischen Werktagen und Wochenende spricht dafür, dass der DBS sowohl privat als auch beruflich genutzt wird. Der bereits donnerstags stattfindende Rückgang und die Zunahme am Sonntag deuten darauf hin, dass Lehrer den DBS zur Unterrichtsvorbereitung nutzen. Dafür spricht auch die oben beschriebene Häufigkeit entsprechender Suchphrasen zum Thema Unterrichtsmaterialien (vgl. Abschn. 9.1.1).

Die Anteile der drei Einstiegsarten an den Wochentagen variieren dagegen kaum (vgl. Abb. 12.10, S. A19). Dies entspricht dem geringen Wert für Cramers V. Lediglich der Anteil der Direktzugriffe nimmt leicht zu zwischen Mittwoch und Samstag, wenn also gleichzeitig der absolute Gesamtwert abnimmt. Der Rückgang geht somit überproportional stark auf die Suchmaschineneinstiege zurück. Dies wird besonders deutlich in einer variierten Betrachtung, bei der die Anteile der Wochentage an den Gesamtzahlen der drei Einstiegsarten ermittelt werden (vgl. Abb. 12.11, S. A20). Suchmaschineneinstiege bilden demnach das Extrem unter den Einstiegsarten. Sie weisen zugleich den höchsten und den niedrigsten Tagesanteil (Montag bzw. Samstag) auf. Der größte Abstand zwischen den Einstiegsarten fällt mit 1,1 Prozentpunkten auf den Freitag. Eine Berechnung des Mittelwertes der wochentäglichen prozentualen Schwankungen (vgl. Abb. 12.12, S. A20) ergibt für Direkteinstiege den Wert 2,6, für Suchmaschinen 2,9 und für sonstige Verweise 2,8. Neben dem sehr hohen Gesamtanteil der Suchmaschineneinstiege beim DBS verstärkt diese erhöhte Schwankung den Eindruck der Abhängigkeit des DBS von Suchmaschinen. Gleichzeitig kann der relativ niedrige Wert bei den Direkteinstiegen als Indiz für die Charakterisierung dieser

Nutzer als Stammnutzer angesehen werden. Generell ist bei dieser wochentagsbezogenen Betrachtung eine leichte Verzerrung zu Lasten der Tage Donnerstag bis Sonntag aufgrund des Piwik-Ausfalls zu beachten.

Die stundenbezogene Auswertung der absoluten Sitzungszahlen ergibt, dass die meisten Sitzungen zwischen 14 und 15 Uhr beginnen (vgl. Abb. 12.13, S. A21). Zuvor steigt die Zahl der Sitzungen von 4 bis 11 Uhr an, bevor es zu einem leichten Rückgang in der Mittagszeit zwischen 12 und 14 Uhr kommt. Nach 15 Uhr nimmt die Anzahl der Sitzungen bis 4 Uhr ab, mit stärkeren Rückgängen ab 18 und ab 22 Uhr. Damit entspricht die Verteilung in etwa gängigen Büroarbeitszeiten, bleibt allerdings bis 21 Uhr über dem Niveau von 9 Uhr. Für den Unabhängigkeitstest werden folgende Hypothesen formuliert:

H_{2_0}: Die Verteilung der Sitzungen auf die Tagesstunden ist von der Einstiegsart unabhängig.
H_{2_1}: Die Verteilung der Sitzungen auf die Tagesstunden ist von der Einstiegsart abhängig.

Der Wert für χ^2 lautet 7.113,3. Der kritische Wert liegt bei 62,830 und jener für Cramers V bei 0,043. Das Ergebnis entspricht somit jenem bei der Verteilung nach Wochentagen: Die Nullhypothese ist abzulehnen, gleichzeitig ist die Ausprägung des Zusammenhangs sehr gering. Das heißt, dass in der Grundgesamtheit eine signifikante Abhängigkeit zwischen Einstiegsart und Wochentagsverteilung besteht, der Zusammenhang, mit dem sich beide Variablen gemeinsam verändern, jedoch unbedeutend gering ist.

Die relativen Anteile der drei Einstiegsarten an den stundenbezogenen Sitzungen bleiben tagsüber zwischen 10 und 0 Uhr stabil, während sich in den Nachtstunden eine Verschiebung zu Lasten der Suchmaschineneinstiege feststellen lässt, die zwischen 4 und 5 Uhr ihren Höhepunkt erreicht (vgl. Abb. 12.14, S. A21). Der Anteil der Suchmaschineneinstiege verläuft somit antiproportional zur absoluten Gesamtzahl der Sitzungen. Auch die stundenweise Verteilung spricht, wie die wochentagsbezogene, für eine gewisse Stabilität bei den Direkteinstiegen.

9.1.2. Nutzung des DBS

Aktionen je Sitzung (Besuchstiefe)

In einem ersten Schritt werden die Häufigkeiten der Aktionszahlen je Sitzung unabhängig von der Einstiegsart ermittelt (vgl. Tab. 12.10, S. A22). Es zeigt sich, dass über 44% aller Sitzungen lediglich eine Aktion umfassen, es sich also um Absprünge

handelt. Weitere 17,4% beinhalten zwei Aktionen, 11,9% drei Aktionen. 16,8% entfallen auf Sitzungen mit vier bis sieben Aktionen. Aufsummiert entfallen 90,25% aller Sitzungen auf jene mit bis zu sieben und 94,4% auf solche mit bis zu zehn Aktionen. Das Mittel liegt bei 3,45, die Standardabweichung bei 5,46. In Anbetracht eines Maximums von 513 Aktionen entspricht die geringe Standardabweichung der oben beschriebenen Konzentration bis etwa zehn Aktionen.

Der Chi-Quadrat-Unabhängigkeitstest setzt erwartete Häufigkeiten in Höhe von mindestens fünf voraus. Um dieser Bedingung zu entsprechen, werden Sitzungen mit 71 und mehr Aktionen zu einer einzigen Klasse zusammengefasst. Dadurch reduziert sich Skalenniveau von einem Rational- auf ein Nominalniveau. Aus der nun praktikablen Darstellung der Häufigkeitsverteilung als Histogramm (vgl. Abb. 12.15a, S. A24) wird deutlich, dass die Häufigkeiten von Sitzungen mit ein bis drei Aktionen verglichen mit der Normalverteilung deutlich überrepräsentiert und jene mit vier bis 16 Aktionen unterrepräsentiert sind. Auch nach Ausschluss der Ein-Aktionen-Sitzungen bleiben Sitzungen mit zwei bis fünf Aktionen gegenüber der Normalverteilung überrepräsentiert (vgl. Abb. 12.15b, S. A24). Das arithmetische Mittel steigt von 3,45 auf 5,38, die Standardabweichung von 5,46 auf 6,69 Aktionen. Unabhängig von der Einstiegsart werden auf dem DBS also tendenziell wenig Aktionen je Sitzung durchgeführt.

In einem zweiten Schritt wird die Aktionszahl nach Einstiegsarten segmentiert (vgl. Tab. 12.11, S. A25). Dabei bestätigt sich die Rangfolge der Einstiegsarten aus der Voruntersuchung, wenn auch mit anderen Werten. Die durchschnittlich wenigsten Aktionen finden bei Suchmaschineneinstiegen statt (3,39 Aktionen). Bei direkt gestarteten Sitzungen finden im Durchschnitt nur unwesentlich mehr Aktionen statt (3,41). Deutlich darüber liegen über sonstige Verweise initiierte Sitzungen mit 3,88 Aktionen. Die Standardabweichung ist mit 4,9 Aktionen bei Suchmaschineneinstiegen am geringsten und mit 6,8 Aktionen bei Direkteinstiegen am höchsten.

Ergänzend dazu liefert die Betrachtung der Anteile der Einstiegsarten an den Sitzungen mit den jeweiligen Gesamtaktionszahlen weitere Hinweise auf Unterschiede in der DBS-Nutzung (vgl. Abb. 12.16, S. A26). Der Anteil der Direkteinstiege ist bei Ein-Aktionen-Sitzungen mit 24,9% überdurchschnittlich hoch (vgl. 19,9%), fällt ab zwei Aktionen deutlich ab und erreicht ab 14 Aktionen wieder durchgängig überdurchschnittliche Werte. Dies deutet auf eine Zweiteilung der Direkteinsteiger in eine Gruppe von Nutzern, die lediglich eine Seite abrufen und eine zweite Gruppe, die den DBS besonders intensiv nutzen. Ausgangspunkte für die Sitzungen der ersten Gruppe könnten einerseits ein abonniertes Newsfeed oder der zweiwöchentlich erscheinende Newsletter sein[7] oder andererseits der bewusste regelmäßige Abruf der

[7] Sofern die Feed-Einträge bzw. E-Mails in einer Desktopanwendung wie Mozilla Thunderbird

Startseite oder bestimmter Redaktionsseiten, um sich über Neuigkeiten zu informieren. Umgekehrt verhält es sich bei Suchmaschineneinstiegen: Ihr Anteil ist bei Ein-Aktionen-Sitzungen unterdurchschnittlich gering (64,9% zu 68,7%), erreicht bei drei Aktionen einen Höhepunkt von 75,8% und nimmt mit steigender Aktionszahl wieder kontinuierlich ab. Der Anteil der sonstigen Verweise nimmt von 10,3% bei einer Aktion (Gesamtdurchschnitt 11,4%) bis auf 15,3% bei 27 Aktionen zu und variiert bei höheren Aktionszahlen stark. Ab etwa 40 Aktionen nimmt die Schwankung unter allen drei Einstiegsarten deutlich zu, was auf die geringere Fallzahl zurückzuführen sein dürfte. Dennoch ist erkennbar, dass Sitzungen mit zahlreichen Aktionen tendenziell häufiger über Direktzugriffe und sonstige Verweisseiten als über Suchmaschinen eingeleitet werden.

Zum Test auf statistische Signifikanz lassen sich folgende Hypothesen formulieren:

H_{3_0}: Die Anzahl der Aktionen je Sitzung ist von der Einstiegsart unabhängig.
H_{3_1}: Die Anzahl der Aktionen je Sitzung ist von der Einstiegsart abhängig.

Der Wert für Chi Quadrat für Tabelle 12.11 liegt bei 30.831,6. Bei einem kritischen Wert von $\chi^2_{0{,}95,\ 70\cdot 2} = \chi^2_{0{,}95,\ 140} = 168,61$ ist H_{3_0} signifikant abzulehnen und H_{3_1} anzunehmen. Wie schon bei der Zeitverteilung deutet jedoch auch hier ein Wert für Cramers V von 0,091 auf einen sehr schwachen Zusammenhang hin.

Sitzungsdauer

Da für die Sitzungsdauer 5.777 Einzelwerte vorliegen, wird im Rahmen dieser Arbeit auf den Abdruck einer Häufigkeitstabelle verzichtet. Die Betrachtung der Häufigkeitsverteilung zeigt, dass mit einer Häufigkeit von 841.169 44,73% aller Sitzungen eine Länge von zehn Sekunden aufweisen (vgl. Abb. 12.17a, S. A28). Dieser außergewöhnlich hohe Wert lässt sich darauf zurückführen, dass Piwik allen Sitzungen, die lediglich eine Aktion aufweisen, einen Standardwert von zehn Sekunden für die Sitzungdauer zuweist. Dies trifft auf 829.503 und somit 98,6% der Zehn-Sekunden-Sitzungen zu. Dieser hohe Anteil führt zu einer starken Verschiebung bei der Häufigkeitsverteilung. Nach einem Ausschluss der Sitzungen mit einer Aktion liegt der Median der Sitzungsdauer bei acht Sekunden (vgl. Abb. 12.17b, S. A28). Das arithmetische Mittel verschiebt sich von 172,6 auf 300,8 Sekunden.

Zur besseren Veranschaulichung wird die Sitzungsdauer in Anlehnung an Hassler in sieben Klassen[8] zusammengefasst [vgl. Hassler 2009, S. 177]. Auch nach der Klassierung zeigt ein Vergleich zwischen der Berücksichtigung und Nichtberücksichtigung

betrachtet werden, wird beim Aufruf einer darin enthaltenen DBS-URL kein Referrer übermittelt – die Sitzung erscheint somit als Direkteinstieg.

[8] Die Aufteilung lautet: ≤10, 11–30, 31–40, 61–180, 181–600, 601–1800 und ≥1801 Sekunden

der Ein-Aktionen-Sitzungen eine deutliche Verschiebung weg vom ersten Segment (vgl. Abb. 12.18, S. A29). Nach Ausschluss der Absprünge entspricht die Klassenverteilung ungefähr der Normalverteilung. Der größte Anteil der Sitzungen entfällt nach diesem Ausschluss mit 23% auf den Bereich zwischen 61 und 180 Sekunden, der zweitgrößte auf jenen zwischen 181 und 600 Sekunden (vgl. Abb. 12.19, S. A30). Über 56% der Sitzungen haben eine Länge von über einer Minute, unter Einbeziehung der Absprünge trifft dies nur auf gut 30% der Sitzungen zu. Aufgrund dieser starken Verzerrung durch die fiktive Sitzungsdauer von zehn Sekunden bei Absprüngen beziehen sich die nachfolgenden Ausführungen auf die Absprung-bereinigten Daten.

Bereits bei den absoluten, nach Einstiegsarten segmentierten Häufigkeiten der Sitzungsdauerklassen ist erkennbar, dass sich der Anteil der Direkteinstiege bei hohen Sitzungsdauern weniger stark verringert gegenüber dem Anteil der Suchmaschineinstiege (vgl. Abb. 12.20a, S. A31). Der Anteil der Direkteinstiege verringert sich von 17,9% bei Sitzungsdauern bis zehn Sekunden auf 11,7% in der nächsten Klasse von elf bis 30 Sekunden und steigt in höheren Klassen wieder an bis auf 31,2% bei Sitzungen ab 30 Minuten (vgl. Abb. 12.20b, S. A31). Insgesamt beträgt ihr Anteil 16%. Mit einem Minimum von 10,6% und einem Maximum von 13,4% variiert der Anteil der sonstigen Verweise nur gering. Somit ergeben sich für die Suchmaschineneinstiege ähnliche Schwankungen wie für die Direkteinstiege. Auf Erstere entfallen 76,2% der Sitzungen von elf bis 30 Sekunden, jedoch nur 58,2% der Sitzungen mit einer Länge von über 1.800 Sekunden.

Somit ergibt sich eine Verteilung ähnlich jener der Aktionszahl: Direkteinstiege sind bei sehr kurzen und sehr langen Sitzungen überproportional stark vertreten. Dies spricht für einen naheliegenden Zusammenhang zwischen Aktionszahl und Sitzungsdauer. Die zahlreichen Aktionen, die bei Direkteinstiegen verstärkt ausgeführt werden, nehmen mehr Zeit in Anspruch. Ein von der Einstiegsart unabhängiger Zusammenhang lässt sich über ein Punktdiagramm visualisieren (vgl. Abb. 12.21, S. A32). Sitzungen mit 50 Aktionen beispielsweise dauern demnach mindestens 61 Sekunden.

Zur Überprüfung der statistischen Unabhängigkeit der Sitzungsdauer von der Einstiegsart werden folgende Hypothesen formuliert:

H_{4_0}: Die Sitzungsdauer ist von der Einstiegsart unabhängig.
H_{4_1}: Die Sitzungsdauer ist von der Einstiegsart abhängig.

Der Wert für Chi Quadrat für eine auf Tab. 12.13 (S. A27) basierende Kontingenztabelle beträgt 15.946,8. Die Nullhypothese ist bei einem kritischen Wert von $\chi^2_{0,95,\ 12} = 21,026$ abzulehnen. Somit besteht eine signifikante Abhängigkeit zwischen

beiden Variablen. Die Stärke des Zusammenhangs ist mit Cramers V = 0,087 als sehr schwach einzustufen.

Absprünge

Bei Ein-Aktionen-Sitzungen kann, mit Ausnahme einer deaktivierten JavaScript-Ausführung, ausgeschlossen werden, dass der DBS über einen externen Link verlassen wurde, da dies bei der Datenerfassung in Piwik eine zweite Aktion auslösen würde.

Folgende Gründe für Absprünge sind denkbar:

1. Ein (eventl. regelmäßiger) DBS-Nutzer ruft gezielt die Startseite oder eine andere Redaktionsseite auf, um sich über Neuigkeiten zu informieren.
2. Ein Abonnent des DBS-Newsletters oder eines Newsfeeds klickt auf einen darin enthaltenen Link und betrachtet nur diese Seite.
3. Ein Quereinsteiger (über Suchmaschinen oder sonstige Verweisseiten) ruft eine Seite des DBS auf, führt jedoch keine weitere Aktion aus. Ein Grund hierfür könnte sein, dass die aufgerufene DBS-Seite das Informationsbedürfnis des Nutzers nicht befriedigt.
4. Nach einem ungewollten Klick auf einen Weblink oder Favoriteneintrag bemerkt der Nutzer seinen Fehler und schließt das Fenster oder klickt auf die Zurück-Schaltfläche.

Demnach kann es sowohl negative wie auch positive Gründe für einen Absprung geben, wobei sich die positiven Gründe eher Direkteinsteigern und die negativen Gründe eher Quereinsteigern zuordnen lassen.

Bei Direkteinstiegen beträgt die Absprungrate 55%, bei Suchmaschineneinstiegen 41,7% und bei sonstigen Verweisen 39,7% (vgl. Tab. 12.14 u. Abb. 12.22, S. A33). Unabhängig von der Einstiegsart liegt sie bei 44,1%. Bei mehr als der Hälfte aller Direkteinstiege wird also lediglich eine Seite abgerufen.

Die Testhypothesen zur Durchführung des Chi-Quadrat-Unabhängigkeitstest zwischen den Absprüngen und der Einstiegsart lauten:

H_{5_0}: Die Zahl der Absprünge ist von der Einstiegsart unabhängig.
H_{5_1}: Die Zahl der Absprünge ist von der Einstiegsart abhängig.

Der kritische Wert für die zwei mal zwei Zellen große Kontingenztabelle beträgt 5,99 und wird von Chi Quadrat mit einem Wert von 23.024,7 überschritten. Auch hier ist die Nullhypothese abzulehnen. Ein schwacher Zusammenhang zwischen Absprüngen

und Einstiegsart besteht gemäß einem Wert von 0,111 für Cramers V. Es besteht eine statistisch signifikante Abhängigkeit zwischen beiden Variablen, die Ausprägung des Zusammenhangs ist jedoch nur schwach.

Eine detaillierte Betrachtung der Absprungrate nach Inhalts- und Funktionskategorien wird in Abschnitt 9.2.1f. vorgenommen.

Wiederkehrende Besucher

Ein erneuter Besuch des DBS nach dem Ende einer Sitzung kann insbesondere bei den Direkteinstiegen als Indiz für die Loyalität der Nutzer betrachtet werden. Bei den anderen Einstiegsarten kann, mit Ausnahme gezielter Suchen nach dem DBS, davon ausgegangen werden, dass der DBS-Aufruf weniger zielgerichtet erfolgt. So ergibt sich denn auch ein Anteil wiederkehrender Besucher an den Direkteinstiegen von 47,5% (vgl. Tab. 12.15 u. Abb. 12.23 ab S. A34). Er liegt damit weit über den Anteilen bei Suchmaschineneinstiegen (16%), sonstigen Verweisen (14,9%) und auch über dem Gesamtwert (22,2%). Dieser hohe Anteil wiederkehrender Besucher bei Direkteinstiegen ist ein starkes Indiz dafür, dass jene Besucher als Stammnutzer angesehen werden können.

Bei den Suchmaschineneinstiegen variiert der Wert je nach Suchphrase. So liegt er beispielsweise bei den 1.022 Einstiegen über die Suchphrase „bildungsserver.de“ bei 22,4%, für „bildungsserver“ bei 28,7% und bei „deutscher bildungsserver“ gar bei 36,8% und damit jeweils deutlich über dem oben genannten Gesamtwert dieser Einstiegsart.

Die entsprechenden Testhypothesen lauten:

H_{6_0}: Die Anzahl wiederkehrender Besucher ist von der Einstiegsart unabhängig.
H_{6_1}: Die Anzahl wiederkehrender Besucher ist von der Einstiegsart abhängig.

Chi Quadrat ist mit einem Wert von 173.795,7 größer als der kritische Wert 3,841, somit ist die Nullhypothese erneut abzulehnen. Die hohe Abweichung bei den Direkteinsteigern spiegelt sich in einem Wert für Cramers V von 0,304 wieder, was auf einen mittleren Zusammenhang hindeutet.

9.1.3. Verlassen des DBS

Das Art des Ausstiegs kann anhand des Typs der letzten Aktion in einer Sitzung unterschieden werden:

1. Eine Sitzung endet auf einer Seite des DBS. Dies kann mehrere Ursachen haben:

a) Der Nutzer schließt das Browserfenster oder verlässt den DBS über die Zurück-Schaltfläche des Browsers.

b) Der Nutzer lässt das Fenster geöffnet, widmet sich jedoch einer gänzlichen anderen Tätigkeit, so dass nach 30 Minuten die Sitzung auswertungstechnisch beendet wird.

c) JavaScript ist im Browser deaktiviert und die Erfassung der anderen beiden Aktionstypen ist somit unmöglich.

2. Die Sitzung endet mit dem Klick auf einen Link, der auf ein externes Ziel verweist. Als Verweisserver enthält der DBS zahlreiche derartige Links.

3. Letzte Aktion einer Sitzung ist der Klick auf eine Binärdatei, z.B. im Portable Document Format (PDF).

Eine Segmentierung dieser drei Ausstiegsarten nach den drei Einstiegsarten zeigt, dass die dritte Ausstiegsart mit stets 0,0% praktisch nicht auftritt (vgl. Abb. 12.24, S. A35). Hinsichtlich der anderen beiden Ausstiegsarten gibt es keine nennenswerten Unterschiede zwischen Direkt- und Suchmaschineneinstiegen. 72,0% bzw. 71,8% der Sitzungen enden auf einer DBS-Seite. Die restlichen 28,0% bzw. 28,2% entfallen auf externe URLs. Bei Einstiegen über sonstige Verweise ist der Anteil der auf einer DBS-Seite endenden Sitzungen mit 75,1% noch etwas höher.

Die Testhypothesen für den Chi-Quadrat-Unabhängigkeitstest lassen sich wie folgt formulieren:

H_{7_0}: Die Häufigkeit der Ausstiege auf internen bzw. externen Seiten ist von der Einstiegsart unabhängig.
H_{7_1}: Die Häufigkeit der Ausstiege auf internen bzw. externen Seiten ist von der Einstiegsart abhängig.

Die Nullhypothese ist bei einem Chi-Quadrat-Wert von 1.087,1 und einem kritischen Wert von 9,488 abzulehnen. Es besteht demnach eine statistisch signifikante Abhängigkeit zwischen der Ausstiegs- und Einstiegsart. Die Stärke des Zusammenhangs zwischen beiden Variablen ist nach Cramers V mit einem Wert von 0,017 jedoch sehr schwach.

9.2. Ergebnisse der Nutzungsmuster-Ermittlung

In diesem Abschnitt werden die Ergebnisse der fünf in Anlehnung an das Web Usage Mining durchgeführten Untersuchungen beschrieben.

9.2.1. Einstiege und Absprünge nach Inhaltskategorien

Einstiege nach Inhaltskategorien

Inhaltskategorien spiegeln primär die einzelnen Zielgruppen- und Themenbereiche des DBS wieder (vgl. Tab 12.4, S. A7). Die Auswertung wird auf die Bereiche „Zielgruppen“ (1xx) und „Themen“ (2xx) beschränkt, da sie den Navigationszugang beim DBS abbilden.

Eine Betrachtung der absoluten Häufigkeiten der Einstiege in die einzelnen Kategorien zeigt, dass die Kategorie „Schule“ aus dem Themenbereich die weitaus meisten Einstiege verzeichnet (vgl. Abb. 12.25, S. A36). Mit großem Abstand folgen weitere Themen wie „Übergreifende Informationen“, „Elementarpädagogik“ und „Behindertenpädagogik“. Deutlich wird außerdem, dass weitaus mehr Einstiege im Themenals im Zielgruppenbereich erfolgen. Weiterhin gibt es in beiden Bereichen Kategorien, die nur sehr wenige Einstiege erfahren, darunter „Ausbilder“ und „Behindertenpädagogen“ bei den Zielgruppen sowie „Medien und Bildung“ und „Interkulturelle Bildung“ bei den Themen.

Unterschiede zeigen sich auch bei den relativen Anteilen der drei Einstiegsarten an den einzelnen Kategorien. Auffällig ist, dass drei Kategorien, die absolut betrachtet sehr wenige Einstiege erfahren, einen Anteil der Direkteinstiege von über 60% aufweisen (vgl. Abb. 12.26, S. A37). Dies betrifft die Zielgruppen „Ausbilder“ und „Behindertenpädagogen“ sowie das Thema „Bildung weltweit“. Das heißt, es gibt bei diesen Kategorien verhältnismäßig viele gezielte Einstiege. Entsprechend gering sind die Anteile der anderen Einstiegsarten: Innerhalb der Kategorie „Behindertenpädagogen“ machen Suchmaschinenverweise nur 1,8% der Einstiege aus. Unwahrscheinlich erscheint die Erklärung, dass bei Suchmaschinen die DBS-Seiten dieser Kategorien gezielt schlechter gerankt werden, zumal die Kategorie nicht aus der URL hervorgeht. Denkbar ist vielmehr, dass diese Themen bei Suchmaschinen schlichtweg weniger häufig nachgefragt werden oder die beim DBS verwendeten Bezeichnungen nicht mit den real verwendeten Suchphrasen übereinstimmen und die DBS-Seiten daher nicht bei den Suchergebnissen berücksichtigt werden. Weiterhin fällt auf, dass bei den meisten Kategorien der Anteil der Suchmaschineneinstiege über und der Anteil der anderen beiden Einstiegsarten unter ihrem jeweiligen Durchschnitt für den gesamten DBS liegen. Dies deutet darauf hin, dass Redaktionsseiten verglichen mit anderen Seitentypen des DBS von Suchmaschinen stärker gewichtet und häufiger als Treffer ausgegeben werden.

Zur Durchführung des statistischen Tests auf Unabhängigkeit zwischen den Einstiegen in den Inhaltskategorien und der Einstiegsart werden folgende Hypothesen

aufgestellt:

H_{8_0}: Die Häufigkeit der Einstiege in den verschiedenen Inhaltskategorien ist von der Einstiegsart unabhängig.
H_{8_1}: Die Häufigkeit der Einstiege in den verschiedenen Inhaltskategorien ist von der Einstiegsart abhängig.

Der Wert für Chi Quadrat beträgt 72.743,3, der kritische Wert 81,381. Auch diese Nullhypothese ist somit abzulehnen. Ein schwacher Zusammenhang besteht gemäß einem Wert von 0,177 für Cramers V. Demnach besteht eine statistisch signifikante Abhängigkeit zwischen der Verteilung der Einstiege auf die Inhaltskategorien und der Einstiegsart mit einer schwachen Ausprägung des Zusammenhangs.

Absprünge nach Inhaltskategorien

Eine Betrachtung der nach Inhaltskategorien und Einstiegsart aufgeteilten Absprungrate und deren Vergleich mit den Gesamtwerten liefert Hinweise auf Verbesserungspotentiale in einzelnen Kategorien (vgl. Abb. 12.27, S. A37). Eine Priorisierung kann dabei anhand der absoluten Einstiegshäufigkeiten vorgenommen werden (vgl. Abb. 12.25, S. A36). So liegt die Absprungrate bei der Kategorie „Bildungswissenschaftler“ deutlich über dem DBS-Gesamtschnitt, allerdings verzeichnet die Kategorie eine der geringsten Einstiegshäufigkeiten im Bereich der Zielgruppen, so dass eine Optimierung hier nicht lohnend oder zumindest nicht vordringlich erscheint. Die Absprungrate für die Kategorie „Schule“ mit den mit Abstand zahlreichsten Einstiegen liegt beim Gesamtwert aller drei Einstiegsarten deutlich unter dem DBS-Schnitt, lediglich die Absprungrate bei den Direkteinstiegen liegt um 0,6 Prozentpunkte über ihrem DBS-Schnitt. Ähnliches gilt für die anderen fünf Kategorien mit den häufigsten Einsteigen aus dem Bereich Themen. Eine Ausnahme hiervon ist lediglich die Kategorie „Behindertenpädagogik“, deren Gesamt-Absprungrate um 6,7 Prozentpunkte über dem DBS-Gesamtschnitt liegt.

Zur Identifizierung konkreter Verbesserungspotentiale lässt sich die Absprungrate auch für einzelne Suchphrasen und Verweisseiten bestimmen. Dies soll am Beispiel der Kategorie „Studierende und Studieninteressierte“ dargestellt werden. Bei dieser Kategorie liegt die Absprungrate für Einstiege über sonstige Verweise über jener der Suchmaschineneinstiege und zugleich auch leicht über dem Gesamtschnitt dieser Einstiegsart für den DBS (40,7% zu 39,7%). Anhand einer Auswertung der Verweis-URLs mit den meisten Absprüngen in dieser Kategorie lässt sich eine Verweis-URL auf der Website der Hochschule Neubrandenburg mit einer Absprungrate von 70% identifizieren[9] (vgl. Tab. 12.16, S. A38). Die absolut meisten Absprünge gehen mit

[9] URL: `http://www.hs-nb.de/hochschule-nb/studium/studienberatung/`

einer Absprungrate von 50,3% indes von einer anderen Verweisseite aus[10]. Die Redaktion des DBS könnte nun die konkreten Einstiegsseiten zwecks Verbesserung des inhaltlichen Kontextes anpassen oder eventl. die Betreiber der Verweisseiten kontaktieren, um eine Anpassung des Linktextes zu erreichen.

Auch die Häufigkeit der Absprünge soll auf ihre Abhängigkeit von der Einstiegsart hin überprüft werden, zu welchem Zweck folgende Hypothesen formuliert werden:

H_{9_0}: Die Häufigkeit der Absprünge in den verschiedenen Inhaltskategorien ist von der Einstiegsart unabhängig.
H_{9_1}: Die Häufigkeit der Absprünge in den verschiedenen Inhaltskategorien ist von der Einstiegsart abhängig.

Chi Quadrat beträgt 31.535,1, der kritische Wert 81,381. Somit ist die Alternativhypothese anzunehmen. Der bivariate Zusammenhang ist mit einem Wert für Cramers V von 0,364 als mittelstark einzustufen.

9.2.2. Einstiege und Absprünge nach Funktionskategorien

Einstiege nach Funktionskategorien

Die Funktionskategorien orientieren sich an den verschiedenen Seitenklassen und bilden somit weitgehend die unterschiedlichen Datenbanken und Funktionen des DBS ab. Die Redaktionsseiten verzeichnen mit über 1,1 Mio. Einstiegen mehr als sieben Mal so viele wie die Startseite (vgl. Abb. 12.28, S. A39). Weitere häufig genutzte Einstiegspunkte sind die „Datenbank Bildungsmedien“ und der nach Fachgebieten strukturierte Zugang zu den Onlineressourcen. Jeweils ca. 35.000 bis 50.000 Einstiege sind in den Kategorien Onlineressourcen, Zeitungsdokumentation und Stellenbörse zu verzeichnen.

Betrachtet man, wie viele Aktionen insgesamt im DBS nach Einstiegen in diesen fünf Funktionskategorien erfolgen, so ergibt sich für Einstiege in die Stellenbörse ein besonders hoher Multiplikatoreneffekt (vgl. Abb. 12.29, S. A40): Auf jeden Einstieg in der Kategorie „Stellenbörse“ folgen durchschnittlich 6,1 weitere Aktionen im DBS – weit mehr als der Gesamtdurchschnitt von 2,5 Aktionen. Für die Startseite beträgt der Wert 3,9. Die entsprechenden Werte der Kategorien „Redaktionsseite“ und „Onlineressourcen nach Fach-/Sachgebiet“ liegen etwa im Gesamtdurchschnitt, jener für die „Datenbank Bildungsmedien“ mit 1,5 darunter. Bereits auf diesem Betrachtungsniveau wird unabhängig von der Einstiegsart wird die Bedeutung der einzelnen

[10] `http://eu.daad.de/eu/sokrates/infos-studierende/05393.html`

Funktionskategorien, insbesondere der Redaktionsseiten, für die Generierung von Nutzungsaktionen auf dem DBS deutlich.

Eine weitergehende Differenzierung der Multiplikatoreneffekte nach Einstiegsart zeigt, dass bei Einstiegen über sonstige Verweise relativ die meisten Aktionen auf dem DBS generiert werden (vgl. Abb. 12.30, S. A40). Der Multiplikator beträgt durchschnittlich für alle Funktionskategorien 3,9, d.h. je Einstieg erfolgen im Schnitt insgesamt 3,9 Aktionen. Für Suchmaschineneinstiege liegt der Wert bei 3,4, für Direkteinstiege bei 3,6. Zu berücksichtigen ist hierbei, dass absolut betrachtet deutlich mehr Einstiege über Suchmaschinen als über die anderen Einstiegsarten erfolgen. Auch innerhalb einzelner Funktionskategorien sind zum Teil große Unterschiede zwischen Aktions-Multiplikatoren der Einstiegsarten festzustellen. So werden nach Suchmaschineneinstiegen auf der Startseite im Schnitt 8,2 Aktionen ausgeführt, was dem 2,5-fachen des Gesamtdurchschnitts entspricht.

Bei einer Betrachtung der relativen Anteile der einzelnen Einstiegsarten and allen Einstiegen lassen sich große Unterschiede zwischen den einzelnen Funktionskategorien identifizieren (vgl. Abb. 12.31, S. A41). So werden bei Redaktionsseiten als der Funktionskategorie mit den häufigsten Einstiegen mit 79% überdurchschnittlich viele Einstiege über Suchmaschinen eingeleitet. Dies deutet darauf hin, dass Suchmaschinen die Strukturierung und Zusammenstellung von ausgehenden Links bei Redaktionsseiten über ihre Ranking-Algorithmen honorieren. Regelmäßigkeiten zwischen vergleichbaren Seitentypen sind nicht erkennbar. So handelt es sich bei fünf Inhaltskategorien, deren Seiten jeweils Einzeldatensätze einer Datenbank abbilden (Onlineressourcen, Personen-, Veranstaltungs-, Wettbewerbs- und Institutionendatenbank) um gleichförmig aufgebaute Seitentypen. Die Anteile der Einstiegstypen unterscheiden sich innerhalb dieser fünf Kategorien jedoch stark. Ohne Herstellung einer Verbindung zwischen dem Anteil der Suchmaschineneinstiege und den verwendeten Suchphrasen samt dem entsprechenden Ranking der DBS-Seiten lässt sich keine Aussage darüber treffen, ob bestimmte Funktionskategorien tatsächlich von Suchmaschinen bevorzugt oder benachteiligt werden oder ob bestimmte Funktionskategorien schlichtweg häufiger als Lesezeichen abgelegt werden.

Zu Durchführung des Unabhängigkeitstests zwischen der Einstiegsverteilung auf die Funktionskategorien und der Einstiegsart lassen sich die Testhypothesen wie folgt formulieren:

H_{10_0}: Die Häufigkeit der Einstiege in den verschiedenen Funktionskategorien ist von der Einstiegsart unabhängig.
H_{10_1}: Die Häufigkeit der Einstiege in den verschiedenen Funktionskategorien ist von der Einstiegsart abhängig.

Bei einem Wert für Chi Quadrat von 480.706 und einem kritischen Wert von 65,171 ist die Nullhypothese abzulehnen. Gemäß Cramers V mit 0,368 besteht ein mittlerer Zusammenhang zwischen der Einstiegsart und der Einstiegshäufigkeit in den Funktionskategorien.

Absprünge nach Funktionskategorien

Bei fünf Funktionskategorien den häufigsten Einstiegen liegt, mit Ausnahme der „Datenbank Bildungsmedien“, die für alle Einstiegsarten gemittelte Absprungrate unter dem DBS-Gesamtschnitt (vgl. Abb. 12.32, S. A41). Liegt die Absprungrate hingegen für alle drei Einstiegsarten über dem Gesamtdurchschnitt, was bei der Personen- und der Veranstaltungsdatenbank, dem Glossar und der Datenbank Bildungsmedien der Fall ist, deutet dies auf eine generelle Diskrepanz zwischen den Erwartungen bzw. dem Informationsbedürfnis der Nutzer und den präsentierten Inhalten hin. Weichen die Absprungraten bei Suchmaschinen- oder sonstigen Verweisen nach oben ab, können die bei den Inhaltskategorien vorgestellten Detailanalysen und Optimierungen vorgenommen werden (vgl. Kap. 8, S. 88). Angebracht erscheint dies bei den Funktionskategorien „Zeitungsdokumentation“, „Eintragsfunktion“ und „Mein Bildungsserver“, wobei die beiden letztgenannten Kategorien nur sehr wenige Seiten umfassen und entsprechend wenige Einstiege auf ihnen stattfinden. Über die Gründe für die in einigen Fällen verglichen mit den anderen Einstiegsarten hohe Absprungrate bei Direkteinstiegen kann nur spekuliert werden. Bei der Startseite und Redaktionsseiten könnten Direktaufrufe der gezielten Überprüfung auf Aktualisierungen dienen.

Zur Durchführung des Unabhängigkeitstests werden folgende Hypothesen formuliert:

H_{11_0}: Die Häufigkeit der Absprünge in den verschiedenen Funktionskategorien ist von der Einstiegsart unabhängig.
H_{11_1}: Die Häufigkeit der Absprünge in den verschiedenen Funktionskategorien ist von der Einstiegsart abhängig.

Die Berechnung von Chi Quadrat ergibt den Wert 226.405,1. Bei einem kritischen Wert von 65,17 ist die Nullhypothese daher abzulehnen. Ein mittlerer Zusammenhang besteht gemäß Cramers V mit 0,383.

9.2.3. Einstiegstiefe

Unter Verwendung der aus der Pfadangabe ermittelten Hierarchieebene der einzelnen Redaktionsseiten innerhalb der Navigationssstruktur kann die Tiefe des DBS-

Einstiegs ermittelt und ausgewertet werden. Eine derartige Zuordnung war bei 70,5% aller Einstiege möglich (vgl. Kap. 8.2). Dabei ist ausgehend von der ersten Ebene (der Startseite) eine Ab- und anschließende Zunahme der Einstiege zu verzeichnen (vgl. Abb. 12.33, S. A42). Finden auf der Startseite vergleichsweise viele Direkteinstiege statt, nimmt die Gesamtzahl der Einstiege bei Ebene zwei ab und von Ebene drei bis sechs wieder zu und anschließend erneut ab. Die tiefsten Einstiege erfolgten auf Ebene zwölf. Hinsichtlich der relativen Anteile der Einstiegsarten ergibt sich eine klare Unterteilung: Haben auf der ersten Ebene sonstige Verweise und Direkteinstiege und auf der zweiten Ebene noch Direkteinstiege deutlich überdurchschnittliche Anteile, dominieren ab der dritten Ebene klar die Suchmaschineneinstiege mit Anteilen von bis zu 89% (vgl. Abb. 12.34, S. A43). Lediglich bei Ebene zwölf ergibt sich ein anderes Bild, allerdings umfasst diese Ebene nur sieben Seiten und rund 290 Einstiege, so dass hier von einer Verzerrung ausgegangen werden kann. Somit lässt sich eindeutig sagen, dass Suschmaschineneinstiege innerhalb der DBS-Navigationsstruktur tendenziell tiefer, am häufigsten auf den Ebenen vier bis sechs, stattfinden (vgl. Abb. 12.35, S. A43). Es ist daher anzunehmen, dass die inhaltliche Granularität dieser Ebenen am ehesten jener der Suchanfragen entspricht.

Zur Überprüfung, ob die Tiefe des Einstiegs von der Einstiegsart abhängig ist, werden folgende Hypothesen formuliert:

H_{12_0}: Die Tiefe der Einstiege in die Navigationsstruktur ist von der Einstiegsart unabhängig.
H_{12_1}: Die Tiefe der Einstiege in die Navigationsstruktur ist von der Einstiegsart abhängig.

Die Berechnung von Chi Quadrat ergibt einen Wert von 368.259,4. Bei einem kritischen Wert von 33,924 ist auch bei diesem Test die Nullhypothese abzulehnen. Cramers V deutet mit eine Wert von 0,372 auf einen mittleren Zusammenhang hin.

9.2.4. Bewegung durch die Hierarchieebenen

Obwohl nur Bewegungen innerhalb jeweils einer Inhaltskategorie berücksichtigt wurden, treten Seitwärtsbewegungen auf (vgl. Kap. 8.4.2). Gründe hierfür können ein erneutes Laden der Seite oder der zwischenzeitliche Aufruf beispielsweise einer Detailseite zu einer Onlineressource sein. Eine Einschränkung muss außerdem hinsichtlich der Querverlinkungen im DBS vorgenommen werden. Die Einschränkung äußert sich in einer Nichterfassung eines Wechsels zwischen Hierarchieebenen, wenn dieser Wechsel durch eine Querverlinkung aus dem Zielgruppen- in den Themenbereich ausgelöst wurde. So führt beispielsweise ein Klick auf den Link *Studierende und Studieninteressierte* (Zielgruppe) → *Studieninteressierte* → *Studien- und Hochschulführer* auf

die Themenseite *Hochschulbildung → Studium → Studienwahl, Studien- und Studierendenberatung → Studien- und Hochschulführer*. Folgt ein Nutzer dieser Verlinkung, wird diese Hierarchiebewegung nicht erfasst, da mit ihr ein (unbewusster) Kategoriewechsel einhergeht.

Die Zählung der Hierarchiebewegungen nach Einstiegsart ergibt keine deutlichen Unterschiede zwischen den Einstiegsarten (vgl. Abb. 12.36, S. A44). Zwar liegt der Anteil der Aufwärtsbewegungen bei Suchmaschineneinstiegen mit 28,3% etwas höher als bei den anderen Einstiegsarten, in Anbetracht der generell tiefen Einstiegsebene bei Suchmaschinenverweisen wird jedoch deutlich, dass diese Nutzer kaum mehr Navigationsschritte auf allgemeinere Themen vollziehen. Dies kann positiv wie negativ gedeutet werden: entweder entspricht die Einstiegstiefe dem Informationsbedürfnis der Suchmaschineneinsteiger oder sie sind sich der Hierarchiestruktur und den übergeordneten Seiten nicht bewusst. Auffällig ist der hohe Anteil von Seitwärtsbewegungen von durchschnittlich 27,8%. Den höchsten Anteil jedoch haben bei allen Einstiegsarten die Abwärtsbewegungen mit durchschnittlich 44,9%.

Die Überprüfung der statistischen Abhängigkeit zwischen Hierarchiebewegungen und Einstiegsart wird anhand folgender Hypothesen überprüft:

H_{13_0}: Die Häufigkeit der Auf-, Seit- und Abwärtsbewegungen in der Navigationsstruktur ist von der Einstiegsart unabhängig.
H_{13_1}: Die Häufigkeit der Auf-, Seit- und Abwärtsbewegungen in der Navigationsstruktur ist von der Einstiegsart abhängig.

Auch diese Nullhypothese ist bei einem Wert für Chi Quadrat von 4.502 und einem kritischen Wert von 9,488 zu verwerfen. Somit lieg eine statistisch signifikante Abhängigkeit vor. Gemäß Cramers V besteht kein bedeutender Zusammenhang zwischen den Hierarchiebewegungen und der Einstiegsart (Wert 0,046).

Wie in Kap. 8.4.1 beschrieben, erfasste das Auswertungsskript nicht nur die Häufigkeit der drei Bewegungsarten, sondern auch deren individuelle Abfolge. Insgesamt 1.381 einfache und 2.282 komplexe Muster konnten identifiziert werden. Für weitere Auswertungen erscheinen vor allem Erstere aufgrund der geringeren Fallzahl und der damit besseren Vergleichbarkeit geeignet zu sein. „Muster“ bezeichnet daher im Folgenden einfache Muster. Ein Vergleich der 30 häufigsten Muster zeigt, dass nur zwei der drei möglichen einschrittigen Muster die ersten Ränge belegen (vgl. Abb. 12.37, S. A45). Auffällig ist außerdem, dass nur sechs der 30 Muster mit einer Aufwärtsbewegung beginnen, verglichen mit acht Seitwärts- und 16 Abwärtsbewegungen. Die Ursache hierfür dürfte in der katalogartigen Seitengestaltung liegen, bei der die Unterbereiche den Hauptteil einnehmen (vgl. Abb. 12.38, S. A46). Die Hyperlinks, die eine Seit- oder Aufwärtsbewegung abbilden, sind weniger zahlreich und unauffälliger

positioniert. Insgesamt lässt sich feststellen, dass der Anteil der Direkteinstiege bei zunehmender Zahl der Hierarchiebewegungen leicht zunimmt und der Anteil der sonstigen Verweise leicht abnimmt (vgl. Abb. 12.40, S. A47).

9.2.5. Bewegung zwischen Inhaltsbereichen und Suche

Ziel dieser Auswertung ist es, etwaige Unterschiede bezogen auf die Einstiegsart in der relativen Häufigkeit von Wechseln zwischen den Navigations- und Suchinstrumenten des DBS zu identifizieren. Denkbare Gründe für einen Wechsel des Bereichs sind:

- Eine Unzufriedenheit mit dem inhaltlichen Angebot des aktuellen Bereichs
- Der Quereinstieg in die Navigationsstruktur oder die Suche bei gleichzeitiger individueller Präferenz des jeweils anderen Zugangs
- Eine Unkenntnis der DBS-Struktur verbunden mit einem Ausprobieren der Bereiche
- Ein bewusstes Durchstöbern des DBS

Ebenso wie bei den Absprüngen ist somit eine positive wie negative Interpretation von Bereichswechseln möglich. Die Betrachtung der absoluten Häufigkeiten der Bewegungen jeweils innerhalb der und zwischen den Bereichen zeigt, dass bereichsinterne Bewegungen bei den Themen und Zielgruppen die häufigsten Bewegungsarten sind (vgl. Tab. 12.18, S. A48). Die nach Gesamthäufigkeit zweite Gruppe bilden die gegenseitigen Wechsel zwischen dem Themen- und dem Zielgruppenbereich. Im Bereich Suche finden die wenigsten Bewegungen statt. Dies gilt sowohl für Bewegungen innerhalb der Suchfunktion als auch für Wechsel aus dem und in den Navigationsbereich.

An den Gesamthäufigkeiten lässt sich festmachen, dass der Themenbereich generell dem Zielgruppenbereich vorgezogen wird. Erstes Indiz hierfür ist die höhere Anzahl von Thema-zu-Thema-Bewegungen verglichen mit den Bewegungen innerhalb der Zielgruppen. Hierin ist allerdings noch nicht zwangsläufig eine Intention zu vermuten, vielmehr dürfte der Hauptgrund in der oben beschriebenen größeren Häufigkeit der Einstiege in Themenbereiche liegen. Ein zweites, stärkeres Indiz liefern die Wechselhäufigkeiten zwischen beiden Bereichen. Die Häufigkeit der Zielgruppe-zu-Thema-Wechsel liegt über jener der Gegenrichtung, obwohl die Anzahl der Zielgruppen-internen Bewegungen geringer ist. Prozentual ausgedrückt liegt das Verhältnis der Zielgruppe-zu-Thema-Wechsel zu den Zielgruppen-internen Bewegungen bei 45%. In der Gegenrichtung beträgt der Anteil nur 23%. Diese Beobachtung wird verstärkt,

berücksichtigt man die Tatsache, dass der Zielgruppenbereich im Hauptmenü vertikal über dem Themenbereich angeordnet und somit hinsichtlich der Nutzeraufmerksamkeit dominant ist. Abgeschwächt wird die Beobachtung durch Links aus dem Zielgruppen- in den Themenbereich. Dieser Effekt lässt sich nicht quantifizieren.

Auch bei der Häufigkeit der Bewegungen innerhalb und zwischen der Navigationsstruktur und der Suche kann mittels folgender Hypothesen einem Chi-Quadrat-Unabhängigkeitstest gegenüber der Einstiegsart unterzogen werden:

H_{14_0}: Die Häufigkeit der Bewegungen innerhalb und zwischen Themen-, Zielgruppen- und Suchbereich ist von der Einstiegsart unabhängig.
H_{14_1}: Die Häufigkeit der Bewegungen innerhalb und zwischen Themen-, Zielgruppen- und Suchbereich ist von der Einstiegsart abhängig.

Der Wert für Chi Quadrat liegt bei 20.843,4 und damit über dem kritischen Wert von 21,026. Damit ist die Nullhypothese abzulehnen. Kein bedeutender Zusammenhang besteht gemäß Cramers V mit einem Wert von 0,075.

Setzt man die nach Einstiegsarten segmentierten Häufigkeiten ins Verhältnis zur Gesamtzahl der Aktionen der jeweiligen Einstiegsart, ergibt sich ein ausgeglichenes Bild (vgl. Abb. 12.41, S. A49). Bei den Direkteinstiegen ist der Wechsel Thema-zu-Zielgruppe etwas unterrepräsentiert, dafür lassen sich dieser Einstiegsart relativ betrachtet mehr Such-interne Bewegungen zuordnen. Insgesamt betrachtet wird bei Direkteinstiegen die Suche etwas häufiger genutzt, was sowohl für Such-interne Bewegungen als auch für beide Wechselrichtungen von und zur Suche gilt.

Dem zweiten Auswertungsschritt liegt die Annahme zugrunde, dass die einzelnen Kategorien innerhalb des Themen- und Zielgruppenbereichs jeweils als weitgehend disparat zu erachten sind und dass Wechsel zwischen verschiedenen Kategorien innerhalb eines Bereiches als Indiz für eine Irritation der Nutzer hinsichtlich der DBS-Navigationsstruktur betrachtet werden können. Dazu werden die Kategoriewechsel innerhalb eines Bereiches, die bereichsübergreifenden Wechsel und die Bewegungen innerhalb einer Kategorie erfasst (vgl. Tab. 12.19, S. A48). Der Kategoriewechsel innerhalb eines Bereiches weist dabei die geringste Häufigkeit auf, obwohl diese Art von Navigationsbewegung durch die Polyhierarchie der Redaktionsseiten begünstigt wird. Bei Betrachtung der relativen Anteile an der Gesamtzahl der Bewegungen der einzelnen Einstiegsarten lassen sich zwischen den Einstiegsarten keine interpretationsfähigen Unterschiede ausmachen (vgl. Abb. 12.42, S. A50). Der Anteil der Kategoriewechsel innerhalb eines Bereiches liegt zwischen 13,2% bei Suchmaschineneinstiegen und 14,2% bei Einstiegen über sonstige Verweise. Den höchsten Anteil mit durchschnittlich 62,7% haben Bewegungen innerhalb derselben Kategorie. Der Anteil der bereichsinternen Kategoriewechsel ist somit vergleichsweise gering, auch

lässt sich keine Einstiegsart mit einem besonders hohen Anteil dieser Wechsel ausmachen.

Die geringen relativen Unterschiede zwischen den Einstiegsarten spiegeln sich auch in einem Wert von 0,020 für Cramers V wieder.

H_{15_0}: Die Häufigkeit der Bewegungen innerhalb der Navigationsbereiche ist von der Einstiegsart unabhängig.
H_{15_1}: Die Häufigkeit der Bewegungen innerhalb der Navigationsbereiche ist von der Einstiegsart abhängig.

Dennoch ist die Nullhypothese mit $\chi^2 = 1.321,4 > 9,488$ abzulehnen.

Nach der erfolgten Beschreibung sämtlicher Auswertungen und Tests werden die Ergebnisse im nächsten Abschnitt zusammenfassend betrachtet.

9.3. Bewertung der Ergebnisse

Deutlich wurde die Dominanz der Suchmaschinen unter den drei Einstiegsarten. Obwohl die Anteile der Einstiegsarten je nach Webangebot variieren und daher nur eine eingeschränkte Vergleichbarkeit gegeben ist, gibt es Hinweise darauf, dass der Suchmaschinenanteil bei den Einstiegen in den DBS überdurchschnittlich hoch ist. Innerhalb der Suchmaschineneinstiege erfolgen verglichen mit einer Website-übergreifenden Statistik überdurchschnittlich viele Einstiege über den Marktführer Google.

Unter den sonstigen Verweisen konnten das Bildungswiki und weitere IZB-Angebote wie das Fachportal Pädagogik als hinsichtlich ihrer Häufigkeit besonders wichtige Einstiegsquellen identifiziert werden. Die wichtigste DIPF-externe Verweisquelle ist Wikipedia.

Bei sämtlichen durchgeführten statistischen Tests wurde eine statistisch signifikante Abhängigkeit zwischen der Einstiegsart und der jeweiligen abhängigen Variablen nachgewiesen. Das Ausmaß des Zusammenhangs zwischen beiden Variablen ist jedoch meist sehr gering. Tabelle 9.1 zeigt eine Übersicht der durchgeführten statistischen Tests unter Angabe der abhängigen Variable mit den jeweiligen Testergebnissen.

Bei allen 15 Tests liegt eine statistisch signifikante Abhängigkeit zwischen der abhängigen Variable und der Einstiegsart vor. Zufallseffekte können somit ausgeschlossen werden. Aufgrund des mit fünf Monaten recht lang bemessenen Erhebungszeitraums kann von einer Verallgemeinerbarkeit der Ergebnisse ausgegangen werden.

Nr	abhängige Variable	χ^2	Krit. Wert	Cramers V	Zusammenh.
1	Sitzungen nach Wochentag	844	21,026	0,015	sehr schwach
2	Sitzungen nach Stunde	7.113	62,830	0,043	sehr schwach
3	Aktionen je Sitzung	30.832	168,610	0,091	sehr schwach
4	klassierte Sitzungsdauern	15.947	21,026	0,087	sehr schwach
5	Absprünge	21.025	5,990	0,111	schwach
6	wiederkehrende Besucher	173.796	3,841	0,304	mittel
7	Ausstiegsart	1.087	9,488	0,017	sehr schwach
8	Einstieg nach Inhaltskategorie	72.743	81,381	0,177	schwach
9	Absprünge nach Inhaltskategorie	31.535	81,381	0,364	mittel
10	Einstieg nach Funktionskategorie	480.706	65,171	0,368	mittel
11	Absprünge nach Funktionskategorie	226.405	35,171	0,383	mittel
12	Einstiegstiefe	368.259	33,924	0,372	mittel
13	Hierarchiebewegung	4.502	9,488	0,046	sehr schwach
14	Bewegung zw. Navigation und Suche	20.843	21,026	0,075	sehr schwach
15	Bewegung innerhalb Navigation	1.321	9,488	0,020	sehr schwach

Tabelle 9.1.: Ergebnisübersicht der statistischen Tests

Die Ausprägung des Zusammenhangs, mit dem sich die abhängigen Variablen gemeinsam mit den Häufigkeiten der drei Einstiegsarten verändern, ist generell gering. Die stärksten, als mittelstark charakterisierbaren Zusammenhänge treten im Bereich der Ansätze des Web Usage Mining auf, hier bei den Absprüngen nach Inhalts- und nach Funktionskategorien, den Einstiegen nach Funktionskategorien und der Einstiegstiefe. Bei den Einstiegen nach Inhaltskategorie besteht ein schwacher Zusammenhang. Im Bereich der Webmetriken besteht ein mittelstarker Zusammenhang bei den wiederkehrenden Besuchern, ein schwacher Zusammenhang bei den Absprüngen. Bei allen weiteren abhängigen Variablen ist der Zusammenhang mit der Einstiegsart sehr schwach oder unbedeutend. Insgesamt lässt sich demnach ein erhöhter Zusammenhang bei den Absprüngen feststellen. Bei allen drei diesbezüglichen Tests weisen Direkteinstiege die mit Abstand höchste Absprungrate auf, gefolgt von Suchmaschineneinstiegen und, in geringem Abstand, Einstiegen über sonstige Verweise.

Trotz der geringen Unterschiede zwischen Direkt- und Suchmaschineneinsteigern lässt sich in eingeschränktem Umfang eine Charakterisierung beider Nutzertypen vornehmen. Direkteinsteiger bilden dabei die Stammnutzer des DBS. Bei ihnen sind die geringsten Schwankungen bei der Zeitverteilung der Einstiege zu verzeichnen. Dies gilt sowohl bei den Wochentags- als auch bei den Tagesstunden-bezogenen Auswertungen. Sie haben einen überdurchschnittlich hohen Anteil an den nach Aktionsanzahl umfangreichen und an den langen Sitzungen. Die dem vermeintlich widersprechende hohe Absprungrate bei den Direkteinstiegen kann zumindest teilweise mit Einstiegen über Newsletter- und Newsfeed-Links erklärt werden. Deutlichstes

Indiz für eine Beschreibung der Direkteinsteiger als Stammnutzer ist der gegenüber den anderen Einstiegsarten etwa drei Mal so hohe Anteil an wiederkehrenden Besuchern.

Suchmaschineneinsteiger bilden zwar nicht den Gegenpol zu Direkteinsteigern, unterscheiden sich jedoch in gewissen Merkmalen von ihnen. Der Anteil von Suchmaschineneinsteigern ist bei der Aktionszahl und der Sitzungsdauer jeweils im mittleren Wertebereich höher. Auch die zeitliche Verteilung der Einstiege schwankt stärker.

Hinsichtlich der Navigationsbewegungen im DBS, sowohl zwischen Hierarchieebenen als auch zwischen der Navigationsstruktur und der Suchfunktion, sind keine bedeutenden Abweichungen zwischen den Einstiegsarten feststellbar.

9.4. Bewertung der Vorgehensweise

Die mittels Piwik erhobenen Webanalysedaten sind eine gute Ausgangsbasis für die in dieser Arbeit durchgeführten Auswertungen. Abfragen der etwa ein Gigabyte großen MySQL-Datenbank konnten auf einem marktüblichen Computer mit guter Performanz durchgeführt werden. Auch der Datenexport nach PASW und dessen Verwendung für statistische Auswertungen und die anschließende Weiterverarbeitung und Datenvisualisierung in einer Tabellenkalkulation haben sich bewährt. Aufgrund des damit verbundenen Aufwands sollten bei zukünftigen Auswertungen mögliche Vereinfachungen oder eine Teilautomatisierung geprüft werden (vgl. Kap. 10).

Die Durchführung der Chi-Quadrat-Unabhängigkeitstests ergab sehr hohe Werte für Chi Quadrat, die den kritischen Wert häufig um das mehrere Hundert- bis Tausendfache übersteigen. Dies lässt sich zum einen auf die teilweise umfangreichen Kontingenztabellen zurückführen, mit deren Spalten- und Zeilenzahl Chi Quadrat zunimmt. Eine weitere Ursache ist in den meist hohen Häufigkeiten zu sehen, bei denen selbst prozentual geringe Abweichungen zwischen Kontingenz- und Indifferenztabelle zu hohen absoluten Differenzen und letztlich einem hohen Wert für Chi Quadrat führen. Das deutliche Überschreiten des kritischen Wertes bei den Unabhängigkeitstests belegt die statistische Signifikanz der Abhängigkeit der jeweiligen Variablen. Trotz der grundsätzlichen Eignung des Testverfahrens für die durchgeführten Auswertungen kann auf dessen Anwendung bei zukünftigen Untersuchungen mit ähnlichen Häufigkeiten ggf. verzichtet werden oder nur eine exemplarische Durchführung vollzogen werden, da hier bereits die statistische Signifikanz für diese Häufigkeiten nachgewiesen wurde.

Cramers V hat sich im Rahmen der statistischen Auswertung als geeignetes Zusammenhangsmaß erwiesen. Wegen seiner Normierung auf den Bereich von Null bis Eins

ist es leicht interpretierbar. Außerdem lässt es sich gut anhand der Visualisierungen nachvollziehen. Deutlich wird dies beispielsweise bei einem Vergleich zwischen Abbildung 12.22 (S. A33; Cramers V = 0,111) und Abbildung 12.23 (S. A34; 0,304).

10. Empfehlungen

Nachdem gezeigt werden konnte, dass sich Direkteinsteiger als Stammnutzer umschreiben lassen können, sollte versucht werden, deren Anteil an den Einstiegen zu erhöhen, auch um gleichzeitig den Anteil der Suchmaschineneinstiege etwas zu reduzieren. Folgende Maßnahmen hierzu sind denkbar:

- Die Referrer- und HTTP-Angaben können mittels PHP zur Laufzeit, d.h. während der Seitengenerierung durch das CMS, ausgewertet werden. Damit ließe sich bei Suchmaschineneinstiegen über wichtige Suchanbieter (also vor allem Google, T-Online und Yahoo/Bing) und Suchphrasen, aus denen eindeutig ein gezieltes Nutzerinteresse am DBS hervorgeht (z.B. „(deutscher) bildungsserver“, „bildungsserver.de“), ein Hinweis auf die Lesezeichen-Funktion des Browsers einblenden.
- Mittel- bis langfristig sollte ein Ausbau der Community-Funktionen des DBS geprüft werden. Denkbar ist ein Ausbau der vorhandenen Personalisierungsfunktion „Mein Bildungsserver“ und eine stärkere funktionale Integration der Web 2.0-Angebote. Ein gewisse Nutzerakzeptanz vorausgesetzt, ist davon auszugehen, dass derartige Personalisierungs- und Austauschplattformen zu einer stärkeren Nutzerbindung und einer Steigerung des Anteils der Direkteinsteiger führen.

Zur Reduzierung der übermäßigen Abhängigkeit vom Suchanbieter Google sollte versucht werden, die Anteile der anderen Anbieter, insbesondere von Yahoo/Bing mittels einer Suchmaschinenoptimierung zu erhöhen. Zwei Aspekte sind dabei zu berücksichtigen: Erstens sollte überprüft werden, ob technische Maßnahme ergriffen werden können, um eine Indexierung für diese Suchmaschinen zu vereinfachen, beispielsweise durch Bereitstellung von Informationen zur DBS-Struktur („Sitemap“) in einer maschinenlesbaren XML-Datei. Zweitens ist zu prüfen, ob Änderungen auf der Ebene der Seitenklasse zur Erhöhung des DBS-Rankings möglich sind, beispielsweise durch eine Anpassung von Metatags. Aufgrund der insgesamt niedrigen Marktanteile der Google-Konkurrenten ist dabei das Aufwand-/Nutzen-Verhältnis zu beachten.

Der Portalverbund innerhalb des IZB ist aus DBS-Sicht ein wichtiger Besuchsgenerator und sollte in jedem Fall beibehalten werden. Da hier die Verlinkungen unmittelbar verändert werden können, sollte hier eine Auswertung der Webmetriken, insbesondere der Absprungrate, auf der Ebene einzelner URLs erfolgen, um im Kontext der

Verweisseiten weniger relevante Verlinkungen verbessern zu können. Ähnliches gilt für Wikipedia-Artikel, da auch hier Verlinkungen und Linkbeschriftungen verändert werden können. Speziell bei Wikipedia bietet sich auch eine systematische Prüfung bildungsrelevanter Artikel an, um ggf. zusätzliche Verlinkungen zum DBS zu generieren.

Hinsichtlich einer durchschnittlich sehr tiefen Einstiegstiefe, aber eines kaum höheren Anteils von Aufwärtsbewegungen in der DBS-Hierarchie, sollte überprüft werden, ob Suchmaschineneinsteiger bewusst auf die Möglichkeit einer Aufwärtsbewegung verzichten oder die Möglichkeit nicht wahrnehmen. Für eine derartige Prüfung erscheinen Vergleichstests (sog. A/B-Tests) als geeignetes Instrument.

11. Fazit

Ziel der Arbeit war es, unter Anwendung von Verfahren der Webanalyse und des Web Usage Mining zu prüfen, ob Unterschiede im Nutzungsverhalten zwischen den verschiedenen Einstiegsarten bestehen.

Im theoretischen Teil der Arbeit wurde der Deutsche Bildungsserver mit seinen verschiedenen inhaltlichen und technischen Bestandteilen vorgestellt. Daneben wurde eine informationswissenschaftliche Einordnung des DBS als eine Mischform aus Webkatalog, Fachdatenbank, Information-Retrieval-System und Hypertextsystem vorgenommen. Durch die Verwendung von Webtechnologie waren beim DBS die Auswertungsverfahren Webanalyse und Web Usage Mining anwendbar. Deren Abgrenzung voneinander wurde anhand des Umfangs der Datenvorverarbeitung und der Komplexität der zugrunde liegenden mathematischen Verfahren vorgenommen. Als wichtigste Datenquellen wurden Logfiles und die Kombination aus JavaScript und Cookies vorgestellt. Es bleibt festzuhalten, dass letztere Erhebungsmethode aufgrund des größeren Datenumfangs und der präziseren Erfassung von Nutzeraktionen gegenüber Logfiles vorzuziehen ist. Weiterhin wurden verschiedene Metriken der Webanalyse vorgestellt. Bei der Vorstellung der beiden beim IZB eingesetzten Softwareprodukte zur Webanalyse wurde der Schwerpunkt auf Piwik gelegt, dessen Datenbestand später als Grundlage der Auswertungen diente. Dabei wurden die Funktionsweise des Tracking-Codes von Piwik und die Aufteilung der Nutzungsdaten auf drei Tabellen einer Relationalen Datenbank beschrieben. Außerdem wurde die typische Vorgehensweise beim Web Usage Mining bestehend aus Datenerhebung und -vorverarbeitung, Entdeckung von Nutzungsmustern sowie der Interpretation und Visualisierung dargestellt. Zur Veranschaulichung der Möglichkeiten des Web Usage Mining wurden fünf konkrete Vorgehensweisen vorgestellt, darunter ein Verfahren zur Charakterisierung typischer und atypischer Nutzersitzungen und eines zur Ableitung von Nutzerprofilen.

Bei der praktischen Umsetzung der Auswertungen wurde zunächst der Forschungsablauf festgelegt. Dabei wurde die Einstiegsart als unabhängige Variable definiert. Als abhängige Variablen dienten aus dem Bereich der Webmetriken die Zeitverteilung der Sitzungen, die Anzahl der Aktionen je Sitzung, die Sitzungsdauer, die Absprungrate, der Anteil wiederkehrender Besucher und die Ausstiegsart. Für die Durchführung

von Auswertungen nach Ansätzen des Web Usage Mining wurden mit der Einstiegshäufigkeit und den Absprüngen zwei Webmetriken auf Seiten bezogen, deren Repräsentation in der Datenbank mit Methoden des Web Usage Mining um semantische Angaben ergänzt wurde. Weitere abhängige Variablen waren die Häufigkeit von Bewegungsmustern durch die DBS-Hierarchie und die Häufigkeit vordefinierter Bewegungsmuster innerhalb und zwischen Navigationsbereich und Suchfunktion. Für die Durchführung der statistischen Auswertungen wurde der Chi-Quadrat-Unabhängigkeitstest als Test auf die statistisch signifikante Unabhängigkeit bzw. Abhängigkeit zweier Variablen und Cramers V als bivariates Zusammenhangsmaß zur Messung der Stärke des Zusammenhangs zwischen unabhängiger und abhängiger Variable gewählt. Der Vorgehensweise beim Web Usage Mining folgend, wurde zunächst eine Datenvorverarbeitung durchgeführt. Dabei wurden Zugriffe aus den Netzen des DIPF, des FWU und des Googlebots aufgrund der von ihnen ausgehenden Verzerrungen des Datenmaterials entfernt. Anschließend wurden die oben bereits angesprochenen semantischen Daten ergänzt, darunter eine die DBS-Hierarchie abbildende Pfadangabe. Schließlich wurden die Auswertungen mittels MySQL und PASW durchgeführt.

Dabei konnte gezeigt werden, dass statistisch signifikante, wenn auch tendenziell nur schwach ausgeprägte Unterschiede im Nutzungsverhalten zwischen den drei Einstiegsarten Direkteinstiege, Suchmaschineneinstiege und Einstiege über sonstige Verweise bestehen. Ein sehr schwacher Zusammenhang besteht beispielsweise bei der Häufigkeit von Hierarchiebewegungen (auf-, ab-, seitwärts) innerhalb der Redaktionsseiten. Ein hoher Zusammenhang lässt sich zwischen der Einstiegsart und der Häufigkeit wiederkehrender Besucher nachweisen.

In dieser Arbeit wurde die grundsätzliche Eignung der von Piwik erhobenen Nutzungsdaten, der gewählten Auswertungsverfahren sowie der statistischen Untersuchungen für die gewählte Zielsetzung gezeigt.

Ein Nachteil der gewählten Vorgehensweise ergibt sich aus der Beschränkung auf quantitative Nutzungsdaten, bei denen Abweichungen vom Durchschnitt kaum eindeutig positiv oder negativ beurteilt werden können, da unter anderem das Informationsbedürfnis, aber auch weitere Angaben zum Nutzer wie z.B. die individuelle Kompetenz im Umgang mit Webangeboten, nicht bekannt sind. Dennoch konnten in dieser Arbeit einige grundsätzliche Nutzercharakteristika insbesondere der Direkteinsteiger identifiziert werden wie beispielsweise ihr häufiges Wiederkehren zum DBS oder ihre Aufteilung in Kurz- und Langbesucher (anhand der Anzahl der Aktionen und der Sitzungsdauer). Auch konnten Empfehlungen für Verbesserungen und den Ausbau des DBS erarbeitet werden.

Es bleibt festzuhalten, dass die Kombination aus Piwik-Daten und Ansätzen des Web Usage Mining ein hohes Potential für weitergehende, vertiefende Auswertungen aufweist.

Literaturverzeichnis

[Apache Software Foundation 2009] Apache Software Foundation: *Log Files – Apache HTTP Server*. 2009. – URL `http://apache.org/docs/2.2/logs.html`. – [Online; Stand 21. Oktober 2009]

[Bekavac 1999] Bekavac, Bernard: *Suche und Orientierung im WWW : Verbesserung bisheriger Verfahren durch Einbindung hypertextspezifischer Informationen*. Konstanz : UVK Universitätsverlag Konstanz, 1999 (Schriften zur Informationswissenschaft 37)

[Bekavac 2004] Bekavac, Bernard: Metainformationsdienste des Internet. In: [Kuhlen u. a. 2004], S. 399–407

[Benninghaus 2007] Benninghaus, Hans: *Deskriptive Statistik*. 11. Aufl. Wiesbaden : VS Verlag für Sozialwissenschaften, 2007

[Botte 2005] Botte, Alexander: Der Deutsche Bildungsserver als Antwort auf aktuelle Anforderungen an die Fachinformation Bildung. In: *dipf informiert* 2005 (2005), Nr. Sonderausgabe 1, S. 5–8

[Botte und Hirschmann 2000] Botte, Alexander ; Hirschmann, Doris: Informationsplattform für Nutzer, Anbieter, Forschende. Deutscher Bildungsserver als Support für die Weiterbildung. In: *DIE - Zeitschrift für Erwachsenenbildung* (2000), Nr. 3, S. 32–34

[Botte und Maier 1999] Botte, Alexander ; Maier, Heike: Was ist der Deutsche Bildungsserver (DBS) und wozu dient er? In: *Zeitschrift für Bildungsverwaltung* Jg. 14 (1999), Nr. 1/2, S. 79–89

[Bourier 2009] Bourier, Günther: *Wahrscheinlichkeitsrechnung und schließende Statistik*. 6. Aufl. Wiesbaden : Gabler, 2009

[Burby u. a. 2007] Burby, Jason ; Brown, Angie ; WAA Standards Committee: *Web Analytics Definitions*. 2007. – URL `http://www.webanalyticsassociation.org/attachments/committees/5/WAA-Standards-Analytics-Definitions-Volume-I-20070816.pdf`

[Canter u. a. 1985] Canter, David ; Rivers, Rod ; Storrs, Graham: Characterizing user navigation through complex data structures. In: *Behaviour & Information Technology* 4 (1985), April, Nr. 2, S. 93–102

[Clark u. a. 2006] Clark, Lillian ; Ting, I-Hsien ; Kimble, Chris ; Wright, Peter ; Kudendko, Daniel: Combining ethnographic and clickstream data to identify user browsing strategies. In: *Information Research* 11 (2006), Nr. 2. – URL `http://informationr.net/ir/11-2/paper249.html`

[Cleff 2008] Cleff, Thomas: *Deskriptive Statistik und moderne Datenanalyse*. Wiesbaden : Gabler, 2008

[Diekmann 2004] Diekmann, Andreas: *Empirische Sozialforschung : Grundlagen, Methoden, Anwendungen*. 11. Aufl. Reinbek bei Hamburg : Rowohlt-Taschenbuch-Verlag, 2004

[Diepold 2000] Diepold, Peter: *Das Konzept des Deutschen Bildungsservers*. September 2000. – URL `http://www.bildungsserver.de/zeigen.html?seite=494`

[Dougherty 2009] Dougherty, Heather: *Search Referrals and Organic Traffic Rising*. 2009. – URL `http://weblogs.hitwise.com/heather-dougherty/2009/05/search_referrals_and_organic_t_1.html`. – [Online, Stand 17. Dezember 2009]

[Enge 2007a] Enge, Eric: *2007 Web Analytics Shootout - Final Report*. 2007. – URL `http://www.stonetemple.com/articles/analytics-report-august-2007.shtml`. – [Online; Stand 21. Oktober 2009]

[Enge 2007b] Enge, Eric: *Web Analytics and Cookies*. 2007. – URL `http://www.stonetemple.com/articles/analytics-and-cookies.shtml`. – [Online; Stand 21. Oktober 2009]

[Erlhofer 2007] Erlhofer, Sebastian: *Informationssuche im World Wide Web*. Berlin : Wissenschaftlicher Verlag Berlin, 2007

[Ferber 2003] Ferber, Reginald: *Information Retrieval : Suchmodelle Und Data-Mining-Verfahren Für Textsammlungen Und Das Web*. Heidelberg : dpunkt-Verl, 2003

[Ferrini und Mohr 2009] Ferrini, Anthony ; Mohr, Jakki J.: Uses, Limitations and Trends in Web Analytics. In: *Handbook of Research on Web Log Analysis*. [Jansen u. a. 2009], S. 124–142

[Friedl 2006] Friedl, Jeffrey: *Mastering Regular Expressions*. 3. Aufl. Sebastopol : O'Reilly Media, August 2006

[Hammwöhner 2004] Hammwöhner, Rainer: Hypertext. In: [Kuhlen u. a. 2004], S. 419–428

[Hassler 2009] Hassler, Marco: *Web Analytics : Metriken auswerten, Besucherverhalten verstehen, Website optimieren*. Heidelberg : mitp, 2009

[heise online 2005] heise online: *Google hilft bei Web-Traffic-Analyse*. November 2005. – URL `http://www.heise.de/newsticker/Google-hilft-bei-Web-Traffic-Analyse-Korrektur--/meldung/66128`. – [Online; Stand 18. September 2009]

[heise online 2009] heise online: *Microsoft und Yahoo schließen Verhandlungen über Kooperation ab*. 2009. – URL `http://www.heise.de/newsticker/meldung/Microsoft-und-Yahoo-schliessen-Verhandlungen-ueber-Kooperation-ab-878224.html`. – [Online; Stand 15. Dezember 2009]

[Hitzler u. a. 2008] Hitzler, Pascal ; Krötzsch, Markus ; Rudolph, Sebastian ; Sure, York: *Semantic Web : Grundlagen*. Berlin : Springer, 2008 (eXamen.press)

[Jansen u. a. 2009] Jansen, Bernard J. ; Spink, Amanda ; Taksa, Isak: *Handbook of Research on Web Log Analysis*. Hershey, London : Information Science Reference, 2009

[Kaushik 2007a] Kaushik, Avinash: *Technical Implementation Best Practices. (JavaScript Tags)*. 2007. – URL `http://www.kaushik.net/avinash/2007/01/web-analytics-technical-implementation-best-practices-javascript-tags.html`. – [Online; Stand 21. Oktober 2009]

[Kaushik 2007b] Kaushik, Avinash: *Web Analytics: An Hour a Day*. Indianapolis : Sybex, Juni 2007

[Kind 2004] Kind, Joachim: Praxis des Information Retrieval. In: [Kuhlen u. a. 2004], S. 389–398

[Kosala und Blockeel 2000] Kosala, Raymond ; Blockeel, Hendrik: Web mining research: a survey. In: *SIGKDD Explorations Newsletter* 2 (2000), 7, Nr. 1, S. 1–15

[Kuhlen 1991] Kuhlen, Rainer: *Hypertext : Ein nicht-lineares Medium zwischen Buch und Wissensbank*. Berlin : Springer, 1991

[Kuhlen 1995] Kuhlen, Rainer: *Informationsmarkt : Chancen und Risiken der Kommerzialisierung von Wissen*. Konstanz : UVK, 1995 (Schriften zur Informationswissenschaft)

[Kuhlen 2004] Kuhlen, Rainer: Information. In: [Kuhlen u. a. 2004], Kap. Information, S. 3–20

[Kuhlen u. a. 2004] Kuhlen, Rainer (Hrsg.) ; Seeger, Thomas (Hrsg.) ; Strauch, Dietmar (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. 5. Aufl. München : Saur, 2004

[Mobasher 2007] Mobasher, Bamshad: Web Usage Mining. In: *Web Data Mining*. Berlin : Springer, 2007, S. 449–483

[Netcraft 2009] Netcraft: *March 2009 Web Server Survey*. 2009. – URL `http://news.netcraft.com/archives/2009/03/15/march_2009_web_server_survey.html`. – [Online; Stand 21. Oktober 2009]

[Otter und Johnson 2000] Otter, M. ; Johnson, H.: Lost in hyperspace: metrics and mental models. In: *Interacting with Computers* 13 (2000), Nr. 1, S. 1–40

[Peterson 2005] Peterson, Eric: *Web Site Measurement Hacks: Tips & Tools to Help Optimize Your Online Business*. Sebastopol : O'Reilly, 2005

[Piwik Developers 2009a] Piwik Developers: *Piwik - Developer Zone*. 2009. – URL `http://dev.piwik.org/trac`. – [Online; Stand 21. Oktober 2009]

[Piwik Developers 2009b] Piwik Developers: *Piwik Homepage*. 2009. – URL `http://piwik.org`. – [Online; Stand 21. Oktober 2009]

[Probst u. a. 1999] Probst, Gilbert J. B. ; Raub, Steffen ; Romhardt, Kai: *Wissen Managen: Wie Unternehmen Ihre Wertvollste Ressource Optimal Nutzen*. 3. Aufl. Frankfurt am Main : Frankfurter Allgemeine Zeitung, 1999

[Raithel 2008] Raithel, Jürgen: *Quantitative Forschung : Ein Praxiskurs*. 2. Aufl. Wiesbaden : VS Verlag für Sozialwissenschaften, 2008

[Rijsbergen 1979] Rijsbergen, C. J.: *Information Retrieval*. 2. Aufl. London [u.a.] : Butterworths, 1979. – URL `http://www.dcs.glasgow.ac.uk/Keith/Preface.html`. – [Digitale Neuausgabe, online verfügbar; Stand 15. Oktober 2009]

[Sadagopan und Li 2008] Sadagopan, Narayanan ; Li, Jie: Characterizing typical and atypical user sessions in clickstreams. In: *Proceeding of the 17th international conference on World Wide Web*. Peking : ACM, April 2008, S. 885–894

[Smith 1996] Smith, Pauline: Towards a practical measure of hypertext usability. In: *Interacting with Computers* 8 (1996), Nr. 4, S. 365–381

[Stermsek u. a. 2007] Stermsek, Gerald ; Strembeck, Mark ; Neumann, Gustaf: A User Profile Derivation Approach based on Log-File Analysis. In: *International Conference on Information and Knowledge Engineering (IKE)*, 2007

[Stock 2007] Stock, Wolfgang G.: *Einführung in die Informationswissenschaft.* Bd. 1: *Information retrieval : Informationen suchen und finden.* München : Oldenbourg, 2007

[Ting u. a. 2005] Ting, I-Hsien ; Kimble, Chris ; Kudenko, Daniel: A pattern restore method for restoring missing patterns in server side clickstream data. In: Zhang, Yanchun (Hrsg.) ; Tanaka, Katsumi (Hrsg.) ; Yu, Jeffrey X. (Hrsg.) ; Wang, Shan (Hrsg.) ; Li, Minglu (Hrsg.): *Web Technologies Research and Development – APWeb 2005: 7th Asia-Pacific Web Conference, Shanghai, China, March 29–April 1, 2005.* Berlin : Springer, 2005, S. 501–512

[Ulbricht 2007] Ulbricht, Carsten: *Ist die Nutzung von Google Analytics und Co rechtswidrig?* November 2007. – URL `http://www.rechtzweinull.de//index.php?/archives/50-Ist-die-Nutzung-von-Google-Analytics-und-Co-rechtswidrig.html`. – [Online; Stand 21. November 2009]

[Web Analytics Association 2009] Web Analytics Association: *About Web Analytics Association.* 2009. – URL `http://www.webanalyticsassociation.org/aboutus/`. – [Online; Stand 21. Oktober 2009]

[Webtrekk GmbH 2009] Webtrekk GmbH: *Webtrekk Kunden-Langzeitstudie für das 3. Quartal 2009.* 2009. – URL `http://www.webtrekk.de/fileadmin/pdf/pm/2009/PM_langzeitstudie_3_Quartal_09.pdf`. – [Online; Pressemitteilung; Stand 15. Dezember 2009]

[Whitten 2008] Whitten, Alma: *Google Public Policy Blog: Are IP addresses personal?* Februar 2008. – URL `http://googlepublicpolicy.blogspot.com/2008/02/are-ip-addresses-personal.html`. – [Online; Stand 21. November 2009]

[Wikipedia 2009a] Wikipedia: *Blog.* 2009. – URL `http://de.wikipedia.org/w/index.php?title=Blog&oldid=65707717`. – [Online; Stand 18. Oktober 2009]

[Wikipedia 2009b] Wikipedia: *Cookie.* 2009. – URL `http://de.wikipedia.org/w/index.php?title=Cookie&oldid=64835738`. – [Online; Stand 23. September 2009]

[Wikipedia 2009c] Wikipedia: *HTTP-Cookie.* 2009. – URL `http://de.wikipedia.org/w/index.php?title=HTTP-Cookie&oldid=65822266`. – [Online; Stand 21. Oktober 2009]

[Wikipedia 2009d] Wikipedia: *Hypertext.* 2009. – URL `http://de.wikipedia.org/w/index.php?title=Hypertext&oldid=67461136`. – [Online; Stand 30. November 2009]

[Wikipedia 2009e] Wikipedia: *Online Analytical Processing*. 2009. – URL `http://de.wikipedia.org/w/index.php?title=Online_Analytical_Processing&oldid=65534348`. – [Online; Stand 12. Dezember 2009]

[Wikipedia 2009f] Wikipedia: *Statistik*. 2009. – URL `http://de.wikipedia.org/w/index.php?title=Statistik&oldid=67973641`. – [Online; Stand 15. Dezember 2009]

Anhang

A. Der DBS

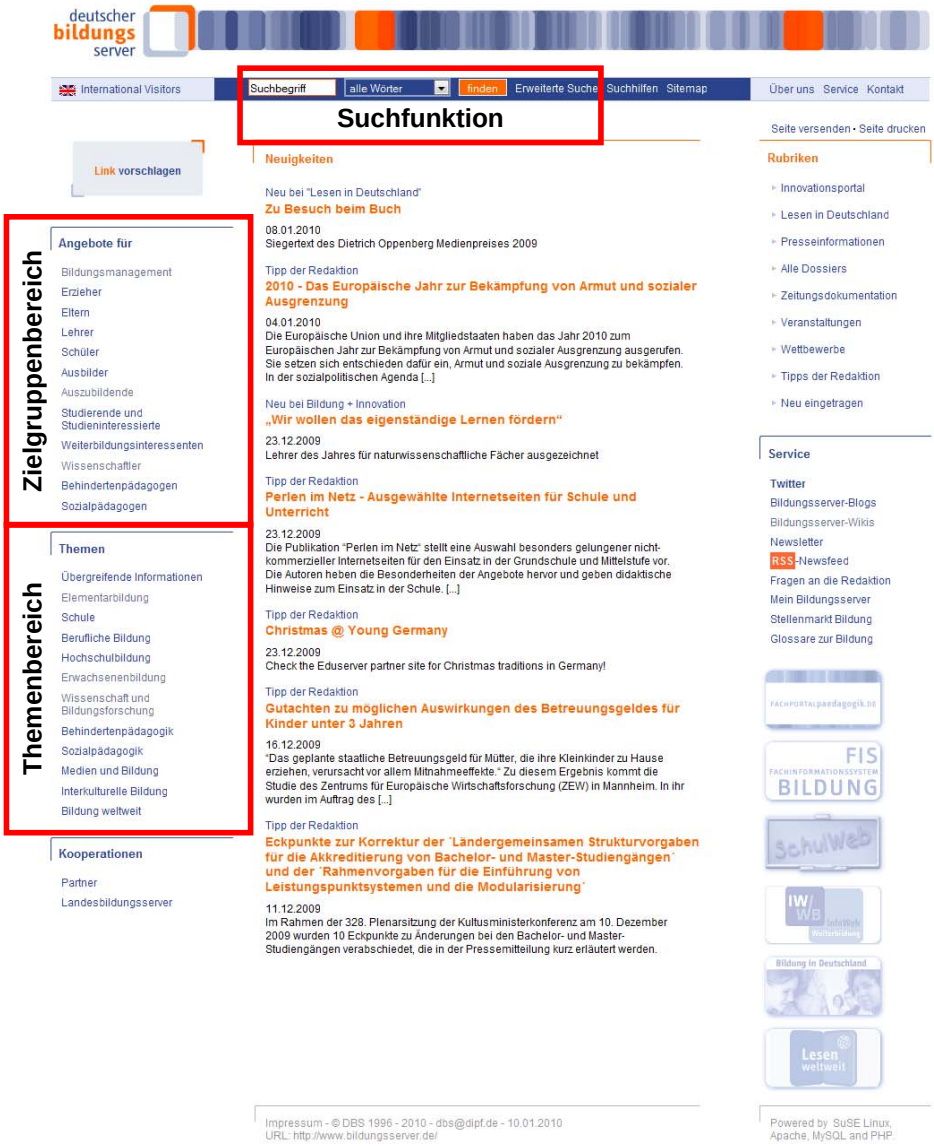


Abbildung 12.1.: Bildschirmfoto der DBS-Startseite mit Hervorhebung auswertungsrelevanter Bereiche

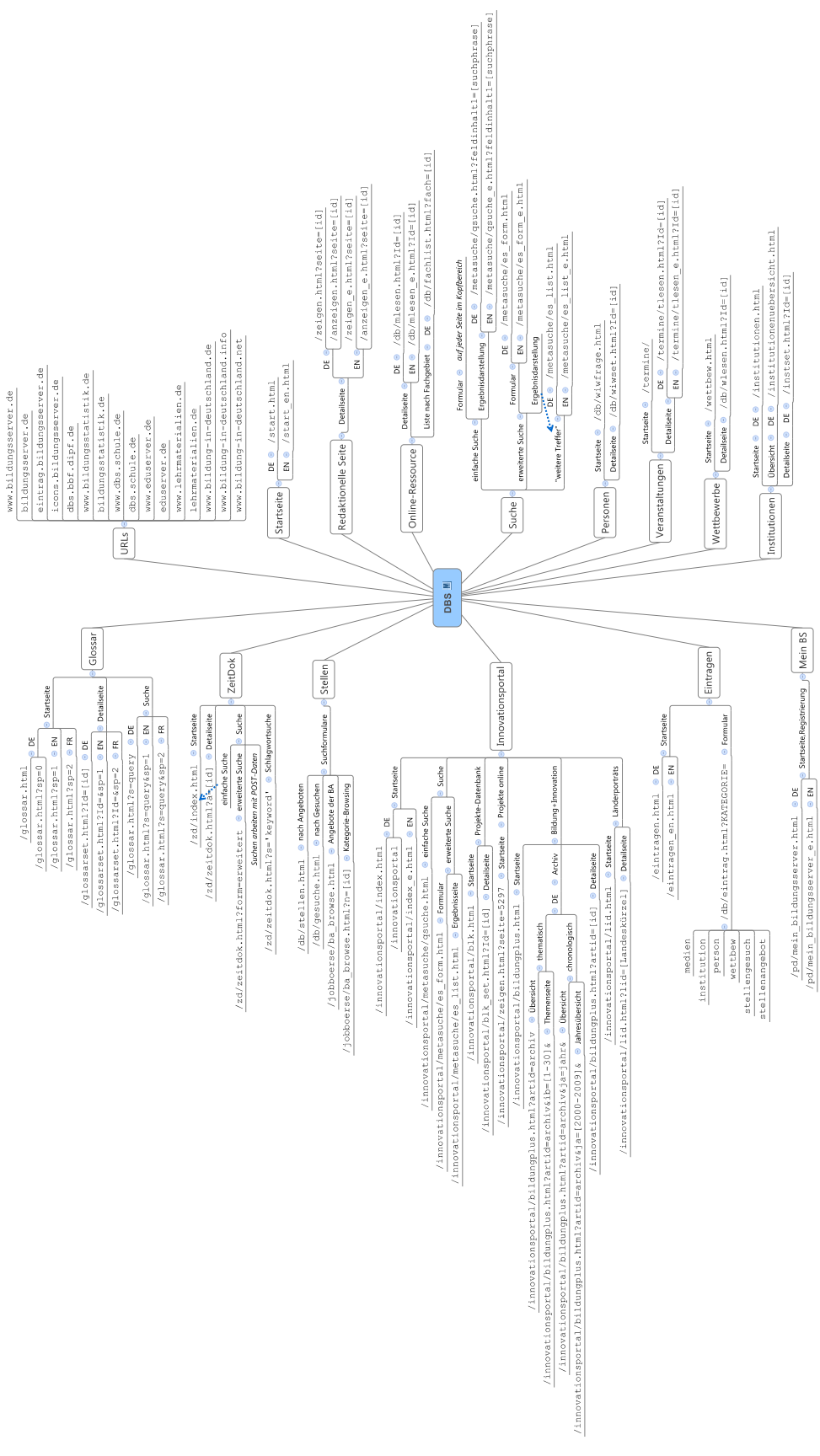


Abbildung 12.2.: Strukturierte Darstellung der Seitenklassen

B. Piwik

```
<!-- Piwik -->
<script type="text/javascript">
var pkBaseURL = (("https:" == document.location.protocol) ? "https://analyse.↘
→dipf.de/wasystem1/" : "http://analyse.dipf.de/wasystem1/");
document.write(unescape("%3Cscript src='" + pkBaseURL + "piwik.js' type='text↘
→/javascript'%3E%3C/script%3E"));
</script><script type="text/javascript">
try {
var piwikTracker = Piwik.getTracker(pkBaseURL + "piwik.php", 1);
piwikTracker.trackPageView();
piwikTracker.enableLinkTracking();

var piwik2 = Piwik.getTracker("http://analyse.dipf.de/wasystem2/piwik.php", ↘
→1);
piwik2.trackPageView();
piwik2.enableLinkTracking();
} catch( err ) {}
</script><noscript><p><img src="http://analyse.dipf.de/wasystem1/piwik.php?↘
→idsite=1" style="border:0" alt=""/><img src="http://analyse.dipf.de/↘
→wasystem2/piwik.php?idsite=1" style="border:0" alt=""/></p></noscript>
<!-- End Piwik Tag -->
```

Listing 12.1: Code-Abschnitt zur Zugriffserfassung in Piwik

Abbildung 12.3.: Datenbankmodell von Piwik [Piwik Developers 2009a]

idvisit	200014
idsite	1
visitor_localtime	10:17:27
visitor_idcookie	37cc9ed73984677b2dbb52d77a6f3ebb
visitor_returning	0
visit_first_action_time	19.07.09 10:18
visit_last_action_time	19.07.09 10:24
visit_server_date	19.07.09
visit_exit_idaction	72940
visit_entry_idaction	18
visit_total_actions	12
visit_total_time	369
visit_goal_converted	0
referer_type	3
referer_name	www.schule-bw.de
referer_url	http://www.schule-bw.de/aktuelles/neu/
referer_keyword	
config_md5config	c14cf104ff36b462f8687f8b02cd196c
config_os	WXP
config_browser_name	IE
config_browser_version	8.0
config_resolution	1024x768
config_pdf	1
config_flash	1
config_java	0
config_director	0
config_quicktime	0
config_realplayer	1
config_windowsmedia	1
config_gears	0
config_silverlight	0
config_cookie	1
location_ip	0
location_browser_lang	de
location_country	de
location_continent	eur
location_provider	t-dialin.net

Tabelle 12.1.: Beispieldatensatz aus der Datenbanktabelle piwik_log_visit

idlink_va	idvisit	idaction	idaction_ref	time_spent_ref_action
724506	200014	18	0	0
724516	200014	15738	18	21
724531	200014	6	15738	33
724548	200014	34	6	32
724564	200014	2879	34	44
724571	200014	34	2879	17
724573	200014	6	34	1
724592	200014	1846	6	69
724611	200014	6	1846	34
724613	200014	1552	6	7
724640	200014	72936	1552	67
724649	200014	72940	72936	44

Tabelle 12.2.: Datensätze aus der Datenbanktabelle piwik_log_link_visit_action zur idvisit 200014 aus Tabelle 12.1

idaction	name	type
18	index	1
15738	zeigen.html?seite=3942	1
6	zeigen.html?seite=5	1
34	zeigen.html?seite=702	1
2879	zeigen.html?seite=3772	1
34	zeigen.html?seite=702	1
6	zeigen.html?seite=5	1
1846	zeigen.html?seite=1801	1
6	zeigen.html?seite=5	1
1552	zeigen.html?seite=703	1
72936	metasuche/qsuche.html?feldinhalt1=Literaturempfehl...	1
72940	http://www.fachportal-paedagogik.de/fis_bildung/su...	2

Tabelle 12.3.: Datensätze aus der Datenbanktabelle piwik_log_action zu den Einträgen in Tabelle 12.2

C. Datenanalyse

100	Bildungsmanagement
101	Erzieher
102	Eltern
103	Lehrer
104	Schueler
105	Ausbilder
106	Auszubildende
107	Studierende und Studieninteressierte
108	Weiterbildungsinteressenten
109	BildungsWissenschaftler/innen
110	Behindertenpaedagogen
111	Sozialpaedagogen
200	Uebergreifende Informationen
201	Elementarbildung – Bildung und Erziehung in Kindertagesbetreuung
202	Schule
203	Berufliche Bildung
204	Hochschulbildung
205	Weiterbildung und Erwachsenenbildung
206	Wissenschaft und Bildungsforschung
207	Behindertenpaedagogik
208	Sozialpaedagogik
209	Medien und Bildung
210	Interkulturelle Bildung
211	Bildung weltweit
300	Innovationsportal
400	Suchfunktionen
700	Lesen weltweit
701	Dossiers fuer internationale Nutzer
750	Rahmenprogramm zur Foerderung der empirischen Bildungsforschung
751	Bildungsforschung Online
752	Bildungsforschung Online – Alternativ
753	Inhalt
800	Service
801	Hilfe bei der Benutzung des Deutschen Bildungsservers
802	Ueber den Deutschen Bildungsserver
803	Copyright
900	Internes, Sonstiges

Tabelle 12.4.: Inhaltskategorien

	Regulärer Ausdruck	**Bezeichnung**
1	^start(_e)?.html	DBS: Startseite
2	^(an)?zeigen(_e)?.html	DBS: Redaktionsseite
3	^db/(fwu/)?(mlesen(_e)?|fach).html	DBS: Onlineressource
4	^db/(wiwfrage|wiwset).html	DBS: Personendatenbank
5	^termine2?/.*	DBS: Veranstaltungsdatenbank
6	^(wettbew.html|db/wlesen.html)	DBS: Wettbewerbsdatenbank
7	^(institutionen(uebersicht)?|instset(_e)? |instliste).html	DBS: Institutionendatenbank
8	^glossar(set)?.html	DBS: Glossar
9	^zd/.*	DBS: Zeitungsdokumentation
10	^(db/(stellen|gesuche|stgset2|staset2) .html|jobboerse/).*	DBS: Stellenboerse
11	^(eintragen(_e)?.html|db/eintrag).*	DBS: Eintragsfunktion
12	^pd/mein_bildungsserver.*	DBS: Mein Bildungsserver
13	^(od_suche.html|metasuche/).*	DBS: Suchfunktionen
14	^innovationsportal/index(_e)?.html	Innovationsportal: Startseite
15	^innovationsportal/metasuche/	Innovationsportal: Suchfunktion
16	^innovationsportal/blk(_set)?.html	Innovationsportal: Projektedatenbank
17	^innovationsportal/zeigen(_e)?.html.*	Innovationsportal: Redaktionsseiten und Projekte Online
18	^innovationsportal/(bildungplus| kommentar).*	Innovationsportal: Bildung plus Innovation
19	^innovationsportal/lid.html	Innovationsportal: Lesen in Deutschland
20	^db/(bw/|fwu/)?fachlist.html	DBS: Onlineressourcen nach Fach-/Sachgebiet
21	^zdf/	DBS: ZDF Katalog fuer Bildungsmedien
22	^elixier/	ELIXIER
24	^db/(fwulesen|flisten2|ffach2?).html	DBS: Datenbank Bildungsmedien
25	^sek2.*html.*	DBS: Abschluesse im Sekundarbereich

Tabelle 12.5.: Funktionskategorien (vgl. Abb. 12.2, S. A2)

D. Auswertungsergebnisse zum DBS-Einstieg

D.1. Suchmaschineneinstiege nach Suchmaschinen

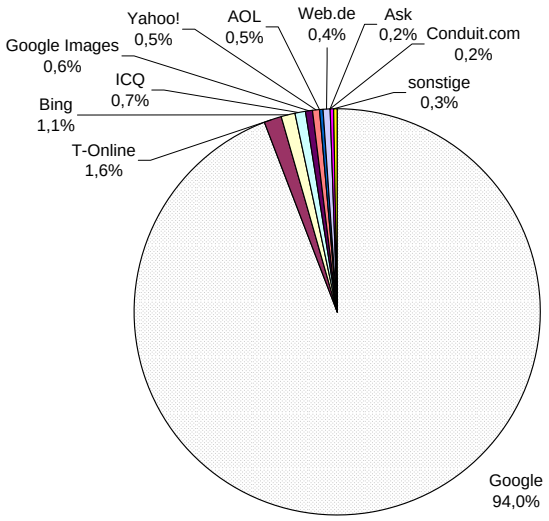


Abbildung 12.4.: Verteilung der Suchmaschineneinstiege nach Suchmaschinen

Tabelle 12.6.: Suchmaschinen

			Aktionen (a)			**Dauer** (t, **in Sek.**)			
	Suchmaschine	**Einstiege**	$\overline{a}$	$s(a)$	$\max(a)$	$\overline{t}$	$s(t)$	$\max(t)$	$\overline{t}/\overline{a}$
-	*gesamt*	1.291.313	3,4	4,9	495	162,8	412,5	15.685	47,9
1	Google	1.214.247	3,4	4,9	416	163,4	414,1	15.685	48,2
2	T-Online	20.141	3,3	4,6	192	161,5	380,8	4.996	48,5
3	Bing	14.665	3,7	6,4	408	167,5	426,5	8.321	45,3
4	ICQ	8.665	3,3	4,3	101	162,2	391,8	6.353	49,2
5	Google Images	8.314	1,9	2,6	87	52,1	199,9	3.754	27,3
6	Yahoo!	6.312	4,0	8,3	495	165,4	393,5	7.166	41,7
7	AOL	6.247	3,5	5,1	119	173,8	405,9	6.002	49,1
8	Web.de (Websuche)	4.693	3,6	4,2	59	177,6	399,4	5.788	49,7
9	Ask	2.144	3,8	5,7	93	170,4	395,3	5.356	44,5
10	Conduit.com	1.975	3,9	7,6	229	179,3	447,7	5.197	46,5
11	GMX	870	3,8	4,9	60	171,7	382,3	3.414	45,7
12	Freenet	581	3,7	4,9	53	192,0	407,7	3.087	51,5
13	123people	513	1,8	1,7	18	42,7	133,3	1.690	23,6
14	Babylon (Powered by Google)	475	3,9	5,9	86	205,0	441,0	3.727	53,0
15	Metager	241	3,6	5,2	50	197,9	481,2	4.986	55,6
16	AltaVista	210	4,3	5,6	33	229,9	463,3	3.034	53,2
17	Meinestadt.de	159	5,2	6,9	43	272,7	537,1	3.178	52,3
18	Yasni	140	1,8	2,0	14	57,1	202,2	2.026	31,0
19	Google Blogsearch	115	2,4	3,6	28	96,4	260,2	1.974	40,0
20	Bluewin	109	3,3	4,0	21	125,9	259,6	1.285	37,8
21	Metager2	80	4,1	4,9	24	193,8	422,5	2.593	47,2
22	Yandex	48	3,3	4,2	21	215,3	414,6	2.028	65,0
23	MyWebSearch	41	2,6	2,5	11	85,0	153,0	827	32,3
24	Seznam	40	3,3	3,1	14	191,0	383,1	1.602	58,2

weiter auf der nächsten Seite

Fortsetzung

			Aktionen (a)			**Dauer** (t, **in Sek.**)			
	Suchmaschine	**Einstiege**	$\overline{a}$	$s(a)$	$\max(a)$	$\overline{t}$	$s(t)$	$\max(t)$	$\overline{t}/\overline{a}$
25	Ixquick	37	2,8	2,6	13	156,5	372,5	1.714	55,7
26	AllTheWeb	28	3,4	3,7	19	89,4	193,6	833	26,1
27	Yahoo! Images	25	1,6	1,8	10	26,4	47,8	217	16,9
28	Mailru	21	1,8	1,4	6	51,0	83,9	277	28,9
29	Startpagina (Google)	19	3,1	2,8	9	157,4	387,3	1.751	51,6
30	Allesklar	18	6,1	14,3	63	311,2	863,7	3.640	51,3
31	Onet.pl	15	5,8	4,9	16	148,1	204,6	790	25,5
32	Clusty	14	3,4	2,6	9	147,1	270,4	1.027	43,8
33	Lycos	12	2,8	2,0	7	205,0	316,0	917	74,5
34	Exalead	10	3,0	2,0	7	208,5	412,8	1.401	69,5
35	Crossbot	10	9,1	12,3	44	353,7	619,3	2.107	38,9
36	Apollo7	7	2,0	1,4	5	54,4	57,9	158	27,2
37	Google Translations	7	4,0	4,5	14	468,7	683,6	1.871	117,2
38	Vivisimo	7	1,0	0,0	1	10,0	0,0	10	10,0
39	Daemon search	6	5,3	4,6	14	746,5	962,4	2.353	140,1
40	Nigma	6	5,0	5,9	16	368,3	720,5	1.972	73,7
41	Mister Wong	6	3,2	1,8	6	137,5	134,4	402	43,4
42	Virgilio	4	1,8	0,8	3	208,3	328,5	777	119,0
43	qip.ru	4	2,8	2,1	6	388,0	599,5	1.424	141,1
44	Baidu	4	8,0	5,6	15	986,5	864,1	2.389	123,3
45	MySpace	4	5,0	3,1	8	305,8	239,3	678	61,2
46	Earthlink	3	3,0	2,2	6	114,0	123,2	287	38,0
47	Bloglines	3	1,3	0,5	2	13,7	5,2	21	10,3
48	Comcast	2	1,5	0,5	2	61,0	51,0	112	40,7
49	Netscape	2	3,0	1,0	4	374,5	352,5	727	124,8
50	Cuil	2	4,5	2,5	7	46,0	43,0	89	10,2
51	Lo.st (Powered by Google)	2	4,5	1,5	6	104,0	72,0	176	23,1
52	Search.com	1	3,0	0,0	3	54,0	0,0	54	18,0
53	Hooseek	1	2,0	0,0	2	29,0	0,0	29	14,5
54	Voila	1	6,0	0,0	6	147,0	0,0	147	24,5
55	Alice Adsl	1	3,0	0,0	3	1385,0	0,0	1.385	461,7
56	Soso	1	1,0	0,0	1	10,0	0,0	10	10,0
57	Najdi.si	1	3,0	0,0	3	149,0	0,0	149	49,7
58	Biglobe	1	9,0	0,0	9	683,0	0,0	683	75,9
59	Abacho	1	6,0	0,0	6	276,0	0,0	276	46,0
60	Trusted Search	1	3,0	0,0	3	21,0	0,0	21	7,0
61	dmoz	1	2,0	0,0	2	218,0	0,0	218	109,0

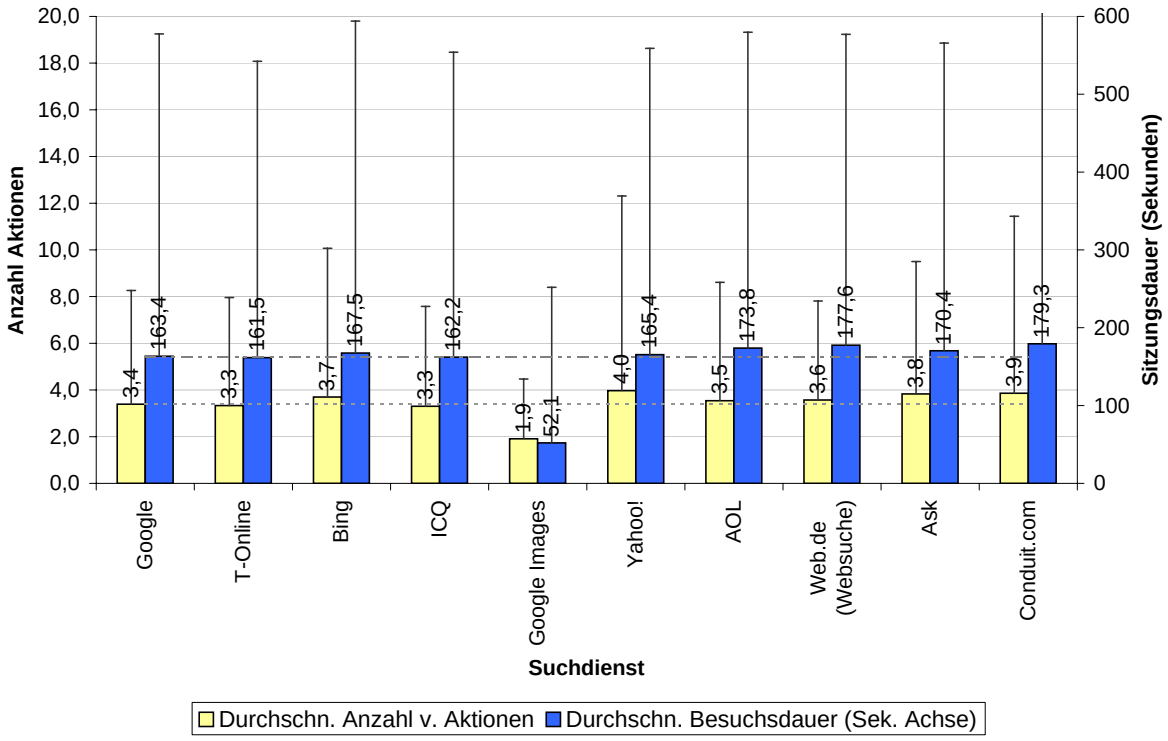


Abbildung 12.5.: Durchschnittliche Anzahl der Aktionen und Besuchsdauer mit Standardabweichung je Suchmaschine

D.2. Suchmaschinen-Einstiege nach Suchphrasen

Tabelle 12.7.: Häufigste Suchphrasen bei Suchmaschineneinstiegen

			Aktionen[1]			**Dauer ((t), in Sek.)**				**Einst. Red.seiten[2]**		
	Suchphrase	$\sum$	$\overline{a}$	$s(a)$	max	$\overline{t}$	$s(t)$	max	$\overline{t/a}$	$\sum$	$\overline{h}$	$s(h)$
-	*gesamt*	1291313	3,4	4,9	495	162,8	412,5	15.685	47,9	939418	5,4	1,8
1	bildungsserver	19.451	8,5	10,6	192	361,3	681,1	11.174	42,6	18.067	1,1	0,4
2	studienberatung	9.455	3,3	3,0	44	135,2	330,8	5.618	41,6	9.444	4,0	0,0
3	deutscher bildungsserver	4.833	10,0	11,3	140	467,1	782,2	10.946	46,6	4.312	1,2	0,4
4	freiwilliges soziales jahr	4.743	3,6	3,1	28	158,6	366,3	5.762	44,7	4.739	4,0	0,0
5	erzieherausbildung	4.590	3,4	4,2	91	147,2	368,2	4.687	43,9	4.582	4,1	0,4
6	hochschulranking	4.153	2,0	1,6	26	58,5	185,8	2.982	29,4	4.145	6,0	0,0
7	referendariat niedersachsen	4.149	3,2	4,3	62	232,9	524,1	6.766	73,2	4.144	5,0	0,0
8	hamburger bildungsserver	4.098	4,2	5,3	131	146,4	359,5	5.971	35,3	4.086	5,0	0,1
9	studieren ohne abitur	3.891	2,7	2,9	54	157,9	348,4	4.197	58,0	3.883	6,0	0,0
10	bildungsserver hamburg	3.450	5,2	6,0	62	203,1	465,9	7.463	39,2	3.421	5,1	0,5
11	eignungstests	3.145	2,5	2,3	32	128,4	362,9	4.428	51,0	3.137	3,0	0,0
12	unterrichtsmaterial	3.016	6,5	6,5	73	187,4	394,2	8.881	28,7	3.012	4,0	0,1
13	soziales jahr	2.824	3,7	3,2	46	143,1	320,2	5.656	39,2	2.820	4,0	0,0
14	erwachsenenbildung	2.541	5,6	6,8	98	251,2	505,1	5.447	44,7	2.536	2,0	0,2
15	sozialpädagogik	2.351	3,6	4,9	71	127,2	330,4	3.628	35,3	2.349	2,0	0,1
16	ferienkalender	2.211	2,0	1,9	53	46,7	143,3	1.830	23,2	2.207	3,0	0,2
17	fsj im ausland	2.193	2,6	2,5	28	148,7	358,3	3.757	57,2	2.193	5,0	0,0
18	erweiterter realschulabschluss	2.159	2,0	2,1	30	53,1	156,5	2.270	26,3	0	-	-
19	referendariat	2.092	7,4	9,0	86	476,3	907,6	10.542	64,5	2.090	4,0	0,0
20	fsj	2.040	3,4	2,8	28	147,2	346,8	4.556	43,1	2.037	4,0	0,1
21	studium ohne abitur	2.019	2,9	3,3	66	170,2	390,7	5.682	58,5	2.013	6,0	0,0
22	masterstudiengänge	2.013	2,7	2,5	26	121,0	308,5	3.592	45,1	2.011	7,0	0,0
23	gewaltprävention	1.956	4,4	4,5	67	236,8	467,5	4.750	54,4	1.951	5,2	0,4
24	e-learning	1.938	4,9	5,2	48	222,2	486,2	7.563	45,4	1.934	3,0	0,0
25	schüleraustausch	1.930	3,6	3,4	38	177,8	385,3	4.332	49,5	1.924	5,0	0,1
26	inklusion	1.850	2,8	3,5	45	172,5	421,6	4.289	62,3	1.839	5,0	0,0
27	dyskalkulie	1.846	2,6	2,5	24	149,2	352,1	4.104	58,5	1.841	6,0	0,0
28	ausbildungsberufe	1.717	2,9	3,2	54	98,2	291,0	5.348	34,0	1.712	4,0	0,0
29	schulfähigkeit	1.590	3,9	3,8	39	284,8	512,9	4.159	72,5	1.586	6,0	0,0
30	medienkompetenz	1.585	3,5	4,4	57	201,6	434,7	3.787	57,3	1.581	3,0	0,1
31	bildungsstandards	1.580	4,1	3,9	53	185,1	406,6	4.692	44,7	1.577	6,0	0,1
32	sprachförderung	1.572	4,7	5,8	75	203,8	471,3	7.863	43,2	1.568	6,0	0,1
33	quereinstieg lehramt	1.455	3,2	3,3	37	257,0	504,2	7.352	79,8	1.449	6,0	0,0
34	schulkonferenz	1.433	1,7	2,1	22	60,0	185,6	2.297	35,5	1.426	5,0	0,0
35	entwicklungspsychologie	1.414	2,5	2,7	29	151,5	353,1	3.737	59,6	1.411	5,0	0,0
36	referendariat baden württemberg	1.404	5,0	5,2	58	274,8	570,4	5.958	55,3	1.402	4,1	0,4
37	vorbereitungsdienst	1.338	7,1	9,3	97	456,6	925,3	8.897	64,1	1.335	4,0	0,0
38	interkulturelle kommunikation	1.318	2,9	3,6	30	202,0	449,1	4.293	68,9	1.312	8,0	0,0
39	ganztagsschule	1.269	3,3	3,6	39	176,4	406,3	4.551	53,8	1.259	4,0	0,0

weiter auf der nächsten Seite

Fortsetzung

			Aktionen			**Dauer ((t), in Sek.)**				**Einst. Red.seiten**		
	Suchphrase	$\sum$	$\overline{a}$	$s(a)$	max	$\overline{t}$	$s(t)$	max	$\overline{t/a)}$	$\sum$	$\overline{h}$	$s(h)$
40	unterrichtsmaterialien	1.264	6,7	6,8	72	176,6	329,8	3.516	26,4	1.231	4,0	0,1
41	behindertenrecht	1.198	3,0	2,7	23	177,2	359,0	3.524	59,5	1.180	5,0	0,0
42	gesunde ernährung	1.186	1,7	1,8	16	66,5	227,0	2.685	39,8	1.180	6,0	0,0
43	www.bildungsserver.de	1.180	7,7	9,3	121	365,9	661,4	8.150	47,3	1.054	1,3	0,6
44	vokabeltrainer französisch	1.154	2,2	1,8	17	110,4	289,4	2.591	49,3	1.150	8,0	0,0
45	stellenangebote lehrer	1.130	7,3	10,0	119	287,2	478,4	4.593	39,2	1.002	3,3	1,0
46	wettbewerbe	1.067	6,8	8,6	113	343,9	607,3	4.781	50,3	0	-	-
47	lehramt quereinstieg	1.055	3,2	3,8	44	228,7	475,9	5.347	70,6	1.051	6,0	0,0
48	lehrereinstellung hessen	1.038	3,5	3,6	42	227,2	477,8	4.577	64,9	1.037	3,0	0,1
49	kinderarmut in deutschland	1.031	3,3	3,1	31	166,4	377,3	4.014	51,2	1.029	4,0	0,0
50	bildungsserver.de	1.022	8,7	12,5	197	376,1	765,7	9.262	43,4	929	1,2	0,4
51	lesekompetenz	1.013	3,3	4,0	45	197,9	441,4	3.905	59,4	1.011	5,8	0,4
52	zeitungsartikel	1.005	2,6	2,8	26	71,6	228,6	3.495	27,2	1.004	4,0	0,0

[1] Anzahl der Aktionen: Die Subspalten enthalten respektive das arithmetische Mittel ($\overline{a}$), die Standardabweichung ($s(a)$ und das Maximum.

[2] Einstieg auf Redaktionsseiten: Die erste Subspalte ($\sum$) gibt an, wie viele Einstiege mit der jeweiligen Suchphrase auf einer Redaktionsseite erfolgten, die zweite Subspalte ($\overline{h}$) enthält das arithmetische Mittel der Hierarchieebene und die dritte Subspalte ($s(h)$) gibt die Standardabweichung für die Hierarchieebene an.

D.3. Häufigste Verweisseiten

Tabelle 12.8.: Häufigste Verweisseiten

			Anz. Aktionen[3]			Besuchsdauer (Sek.)			Einst. Red.seiten[4]		
	Verweisseite	$\sum$	$\overline{x}$	s	max	$\overline{x}$	s	max	$\sum$	$\overline{x}$	s
-	*gesamt*	214.601	3,9	5,9	452	172,3	425,3	13.305	149.736	3,6	2,3
1	wiki.bildungsserver.de	17.431	4,6	5,1	116	237,6	470,5	7.253	15.853	4,5	1,5
2	de.wikipedia.org	13.006	4,0	5,0	73	200,4	457,1	9.648	11.089	4,1	2,0
3	www.wer-weiss-was.de	6.646	2,4	2,4	39	107,0	288,5	5.952	3.748	5,7	1,3
4	www.google.de	5.901	2,6	4,1	171	118,6	346,1	5.210	2.597	5,3	1,7
5	www.kimeta.de	5.233	2,1	4,4	111	89,6	283,4	8.013	0		
6	www.fachportal-paedagogik.de	4.775	5,0	6,9	100	265,5	543,0	7.963	356	3,6	1,9
7	www.schule-bw.de	4.189	4,5	5,8	98	186,3	423,1	5.942	2.839	1,6	1,6
8	www.feiertagskalender.ch	3.965	2,3	2,0	28	66,1	176,1	3.254	3.965	3,0	0,0
9	www.unicef.de	3.425	3,1	4,1	85	105,4	317,7	8.265	3.425	1,0	0,0
10	www.schulministerium.nrw.de	3.173	9,2	14,0	181	231,5	451,0	6.213	2.066	3,0	0,0
11	lernarchiv.bildung.hessen.de	3.135	2,8	4,2	99	132,1	362,0	6.921	1.568	3,3	2,4
12	www.hamburger-bildungsserver.de	2.645	3,5	4,8	68	162,0	391,0	4.560	2.225	5,2	1,9
13	www.zeit.de	2.631	5,1	4,1	61	223,9	386,0	4.451	2.629	1,1	0,5
14	iq.lycos.de	2.414	2,9	3,0	41	102,3	283,2	4.108	2.085	5,0	1,9
15	rzblx10.uni-regensburg.de	2.354	7,2	9,2	108	322,7	632,4	6.663	1.990	1,0	0,2
16	www.schulweb.de	1.918	3,6	4,5	56	172,9	419,0	4.557	434	1,8	1,9
17	forum.kindergarten-workshop.de	1.879	3,6	4,4	50	186,2	407,2	3.477	1.214	6,0	1,5
18	www.bmbf.de	1.838	2,6	2,8	28	112,8	279,9	2.630	1.807	6,0	0,1
19	www.nibis.de	1.588	3,7	5,4	90	197,4	491,3	6.129	397	1,8	1,4
20	bildungsserver.berlin-brandenburg.de	1.392	4,4	6,6	143	217,8	500,0	5.709	1.170	3,5	2,9
21	www.cambridge-exams.de	1.349	2,8	2,9	25	99,6	278,0	3.526	1.349	2,4	0,9
22	www.stellenboersen.de	1.307	12,2	14,6	133	306,2	473,8	3.894	0		
23	www.fragfinn.de	1.302	2,4	2,6	31	90,7	255,2	2.770	247	6,2	1,6
24	de.forestle.org	1.144	4,2	14,1	452	184,0	434,8	5.066	678	4,9	2,1
25	www.was-studiere-ich.de	1.117	2,7	2,8	36	139,6	344,2	3.074	968	6,0	0,0
26	www.autenrieths.de	1.067	6,1	7,9	96	219,9	446,3	4.388	867	4,2	0,8
27	www.gutefrage.net	1.058	2,9	2,7	34	140,5	326,9	3.549	1.022	4,6	1,4
28	www.arbeitsagentur.de	1.039	6,1	9,2	114	242,1	560,4	5.873	521	5,6	1,3
29	news.feed-reader.net	993	2,0	2,7	49	93,4	305,4	3.978	0		
30	www.uni-protokolle.de	969	3,8	5,4	83	175,8	551,0	13.305	320	4,5	2,1
31	www.bebis.de	967	3,9	5,3	82	172,1	454,7	7.958	635	4,2	3,0
32	www.cosmiq.de	936	2,9	2,9	29	86,0	215,8	2.476	816	5,1	1,9
33	www.wege-ins-studium.de	918	3,6	4,2	48	222,5	491,6	6.901	918	5,8	0,8
34	www.studienwahl.de	917	3,5	4,7	74	178,1	449,0	5.107	717	4,5	2,6
35	www.gbiu.de	911	2,3	3,3	54	65,8	207,2	2.513	140	4,3	2,5
36	www.bildungsbericht.de	870	3,8	5,0	54	244,5	575,6	7.577	93	6,2	0,8
37	www.alpenhotel-oberjoch.de	865	1,6	1,3	16	29,9	142,5	3.006	865	3,0	0,0
38	www.dipf.de	835	7,8	8,9	77	335,0	579,2	4.764	396	2,1	2,1
39	www.schulaufgaben4you.de	833	2,5	3,2	35	62,5	205,7	2.917	745	5,0	0,1
40	de.answers.yahoo.com	797	2,9	3,9	52	108,3	300,9	3.610	709	5,3	1,9

weiter auf der nächsten Seite

Fortsetzung

			Anz. Aktionen			Besuchsdauer (Sek.)			Einst. Red.seiten		
	Verweisseite	$\sum$	$\overline{x}$	s	max	$\overline{x}$	s	max	$\sum$	$\overline{x}$	s
41	wiki.zum.de	781	4,4	7,2	105	204,2	463,4	3.497	659	5,2	1,5
42	217.160.159.160	756	1,7	1,8	18	52,3	226,2	3.082	756	6,0	0,0
43	www.google.com	743	2,3	3,1	55	109,7	311,5	3.349	212	5,1	2,1
44	www.sachsen-anhalt.de	728	3,8	6,5	140	222,4	470,8	3.941	691	4,5	1,5
45	www.schule.de	710	4,7	6,7	96	142,9	313,9	2.821	688	1,1	0,8
46	berufenet.arbeitsagentur.de	701	8,5	10,9	138	276,5	543,3	7.657	215	1,0	0,0
47	www.cct-germany.de	690	4,7	5,7	54	249,7	478,6	3.276	685	4,6	1,5
48	www.jugendherberge.de	682	3,0	2,7	23	120,8	330,2	4.915	682	3,3	0,6
49	www.karriere-mit-zukunft.de	636	3,6	5,4	105	224,7	474,3	3.457	541	5,1	1,9
50	infobub.arbeitsagentur.de	634	5,2	6,0	45	247,6	506,9	6.212	520	3,3	1,2
51	blog.bildungsserver.de	631	3,2	3,9	39	183,8	440,3	3.565	591	5,2	2,0
52	navigationshilfe.t-online.de	631	4,4	5,3	49	201,3	436,7	3.309	395	4,2	2,5
53	www.deutschland.de	624	4,2	5,5	50	185,7	430,3	5.941	624	1,0	0,0
54	www.sachsen-macht-schule.de	558	3,1	5,4	88	86,5	259,4	3.142	469	3,1	0,5
55	www.schule.bayern.de	544	5,0	6,6	80	227,9	484,7	5.078	418	1,3	1,1
56	dms.bildung.hessen.de	541	5,2	7,5	73	269,2	668,0	5.991	470	1,1	0,7
57	www.lesen-in-deutschland.de	536	3,3	4,7	55	184,2	523,3	7.030	468	5,6	3,3
58	www.hueber.de	531	3,1	4,9	60	101,6	323,4	3.216	311	2,3	1,1
59	www.goethe.de	528	3,4	5,3	87	142,6	421,3	6.618	505	2,1	2,0
60	search.1und1.de	527	3,5	4,6	47	151,8	325,3	2.559	371	5,5	1,7
61	www.daad.de	513	2,8	3,8	37	184,1	433,4	4.129	513	5,2	1,8
62	www.mk.niedersachsen.de	508	4,6	6,1	67	182,9	439,5	4.340	486	1,4	1,2

[3] Anzahl der Aktionen: Die Subspalten enthalten respektive das arithmetische Mittel, die Standardabweichung und das Maximum.

[4] Einstieg auf Redaktionsseiten: Die erste Subspalte ($\sum$) gibt an, wie viele Einstiege mit der jeweiligen Suchphrase auf einer Redaktionsseite erfolgten, die zweite Subspalte ($\overline{x}$) enthält das arithmetische Mittel der Hierarchieebene und die dritte Subspalte (s) gibt die Standardabweichung für die Hierarchieebene an.

D.4. Zeitliche Verteilung der Zugriffe

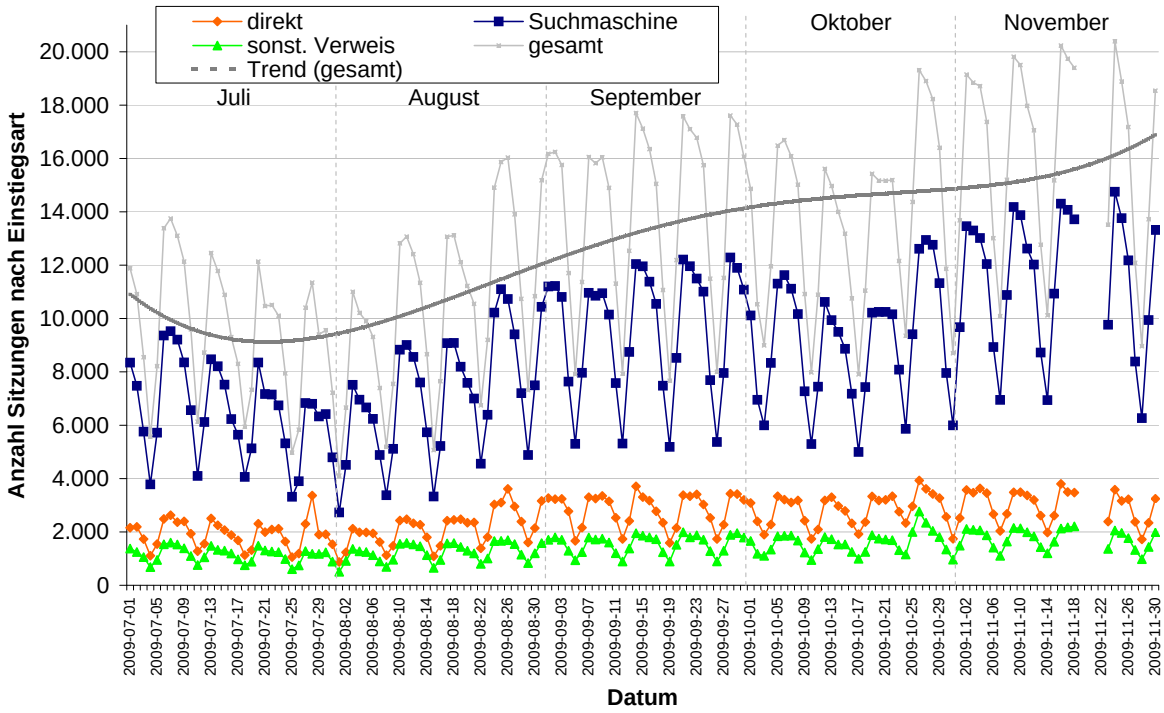


Abbildung 12.6.: Absolute Häufigkeiten der täglichen Sitzungen nach Einstiegsart

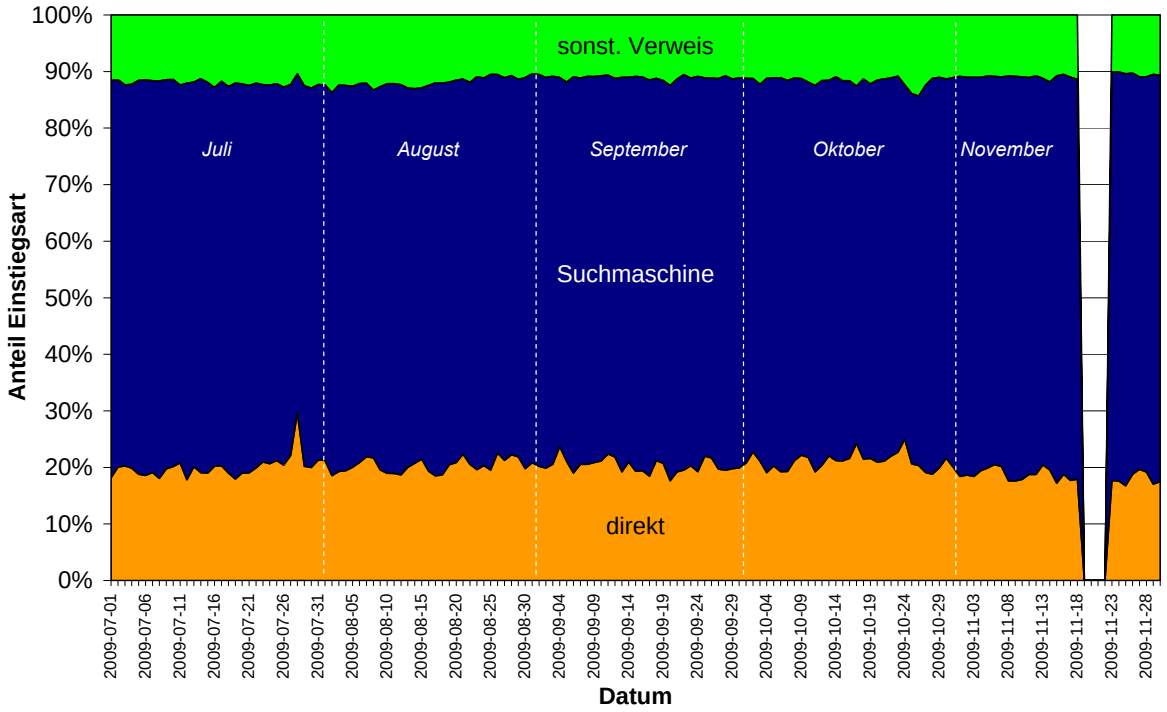


Abbildung 12.7.: Anteile der Einstiegsarten an den täglichen Sitzungen dia02-2

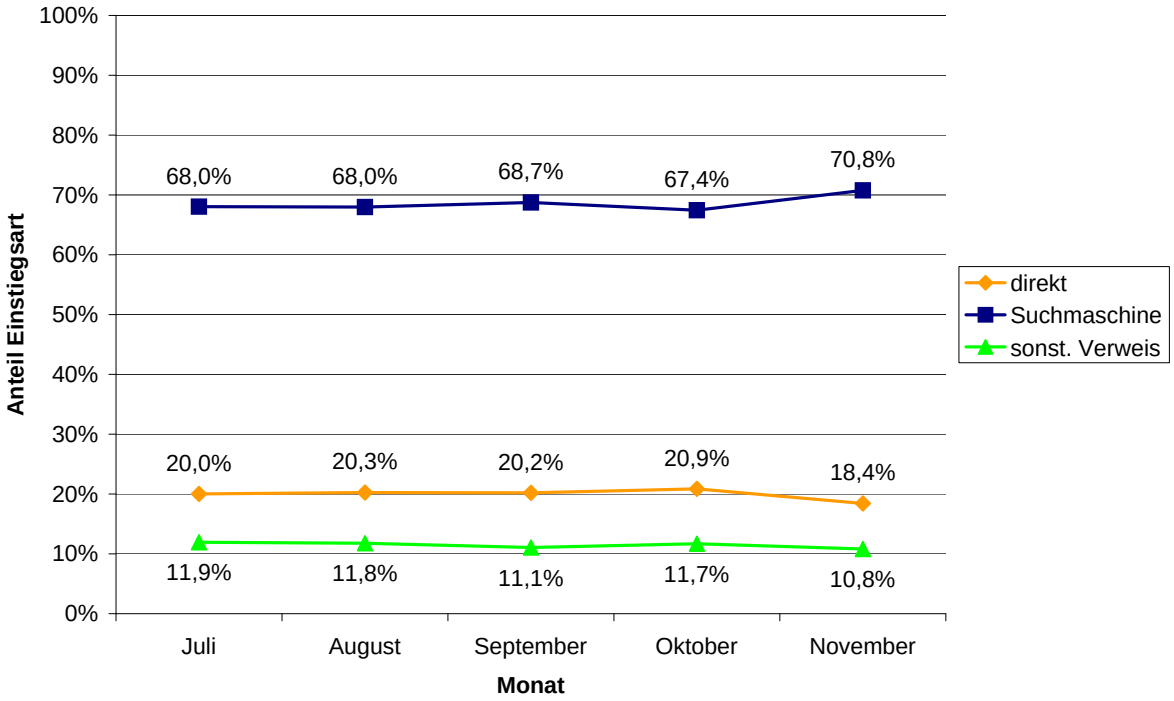


Abbildung 12.8.: Anteile der Einstiegsarten an den monatlichen Sitzungen

	direkt	**Suchmaschine**	**sonst. Verweis**	***Summe***
Montag	66.862	236.380	39.156	342.398
Dienstag	64.168	226.468	36.679	327.315
Mittwoch	63.680	225.567	36.928	326.175
Donnerstag	57.749	195.431	32.335	285.515
Freitag	46.979	146.770	24.876	218.625
Samstag	32.729	103.599	18.346	154.674
Sonntag	42.588	156.854	26.240	225.682
Summe	374.755	1.291.069	214.560	1.880.384

Tabelle 12.9a.: Sitzungen nach Wochentagen und Einstiegsarten (Kontingenztabelle)

	direkt	**Suchmaschine**	**sonst. Verweis**	***Summe***
Montag	68.238,9	235.090,0	39.069,1	342.398,0
Dienstag	65.232,9	224.734,0	37.348,1	327.315,0
Mittwoch	65.005,7	223.951,3	37.218,0	326.175,0
Donnerstag	56.902,3	196.034,2	32.578,5	285.515,0
Freitag	43.571,3	150.107,6	24.946,1	218.625,0
Samstag	30.826,1	106.199,0	17.649,0	154.674,0
Sonntag	44.977,8	154.952,9	25.751,3	225.682,0
Summe	374.755,0	1.291.069,0	214.560,0	1.880.384,0

Tabelle 12.9b.: Sitzungen nach Wochentagen und Einstiegsarten (Indifferenztabelle)

	direkt	**Suchmaschine**	**sonst. Verweis**	***Summe***
Montag	27,8	7,1	0,2	35,1
Dienstag	17,4	13,4	12,0	42,8
Mittwoch	27,0	11,7	2,3	41,0
Donnerstag	12,6	1,9	1,8	16,3
Freitag	266,5	74,2	0,2	340,9
Samstag	117,5	63,7	27,5	208,7
Sonntag	127,0	23,3	9,3	159,6
Summe	595,8	195,3	53,3	$\chi^2 = 844,4$

Tabelle 12.9c.: Sitzungen nach Wochentagen und Einstiegsarten (Berechnung Chi Quadrat)

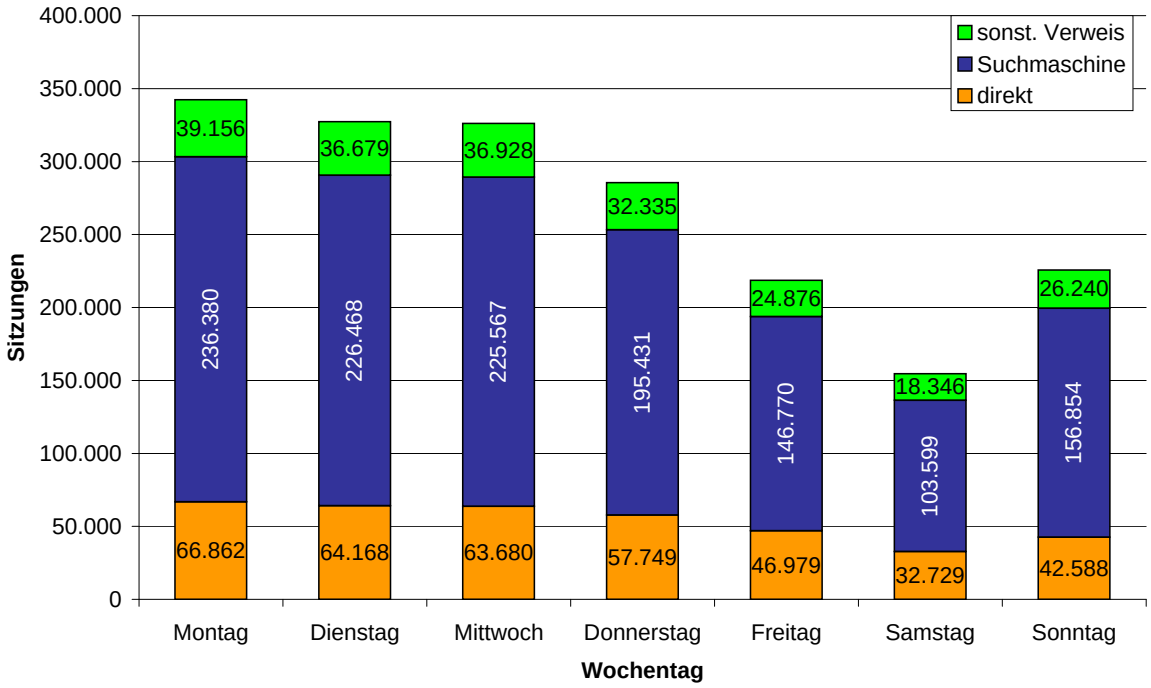


Abbildung 12.9.: Absolute Häufigkeiten der Sitzungen nach Wochentag und Einstiegsart

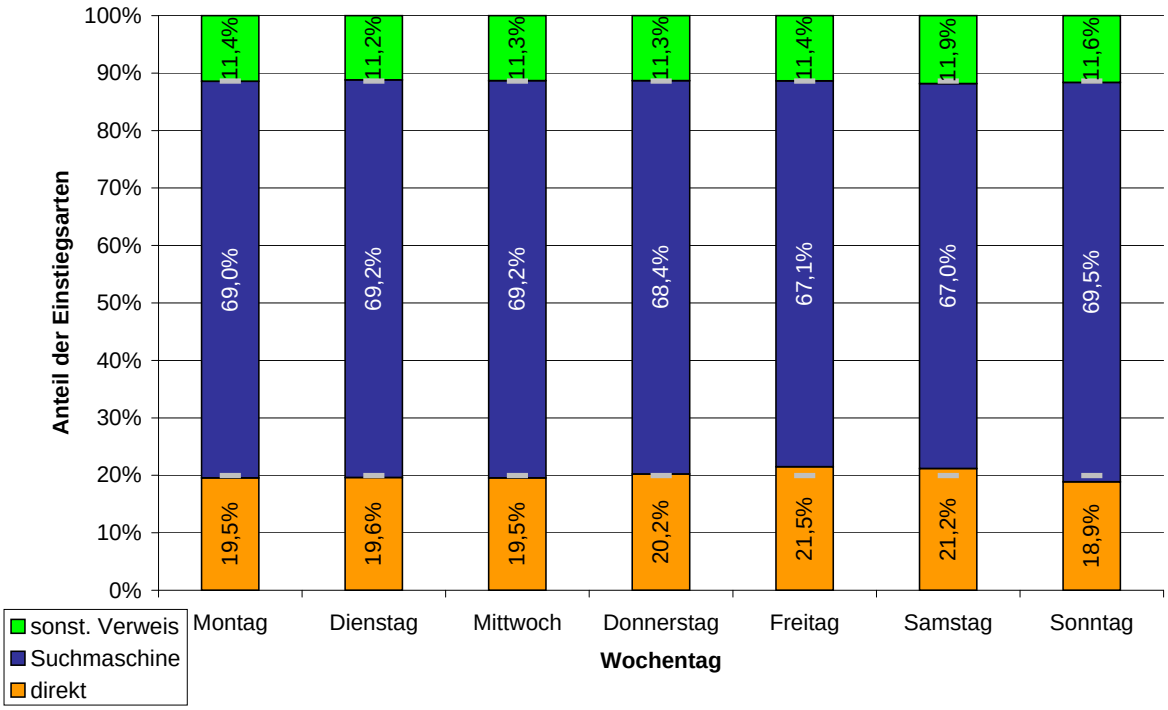


Abbildung 12.10.: Anteil der Einstiegsarten nach Wochentagen

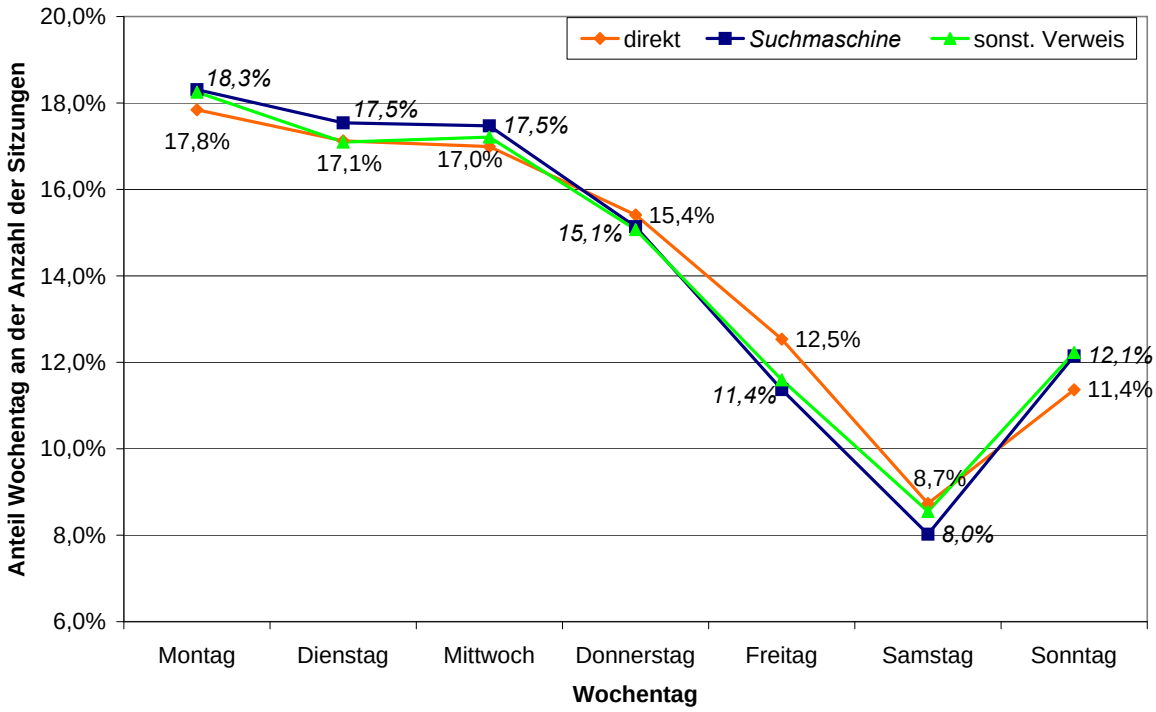


Abbildung 12.11.: Anteil der Wochentage an der Gesamtzahl der Sitzungen nach Einstiegsart

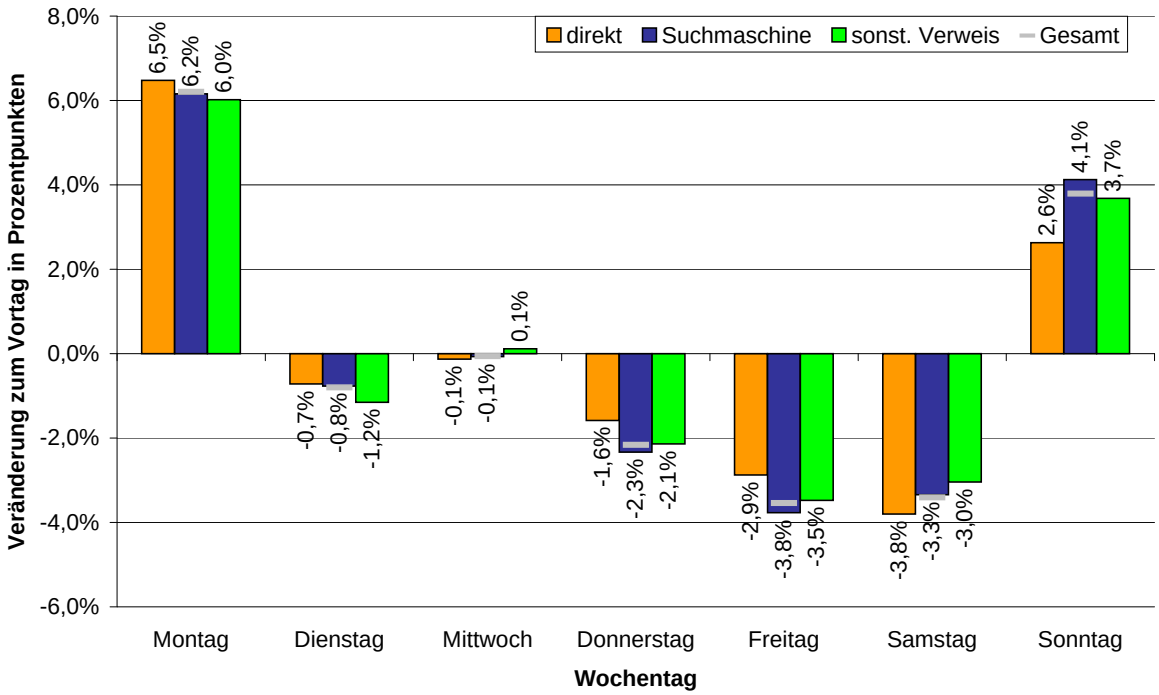


Abbildung 12.12.: Veränderung der Anteile aus Abb. 12.11 zum Vortag

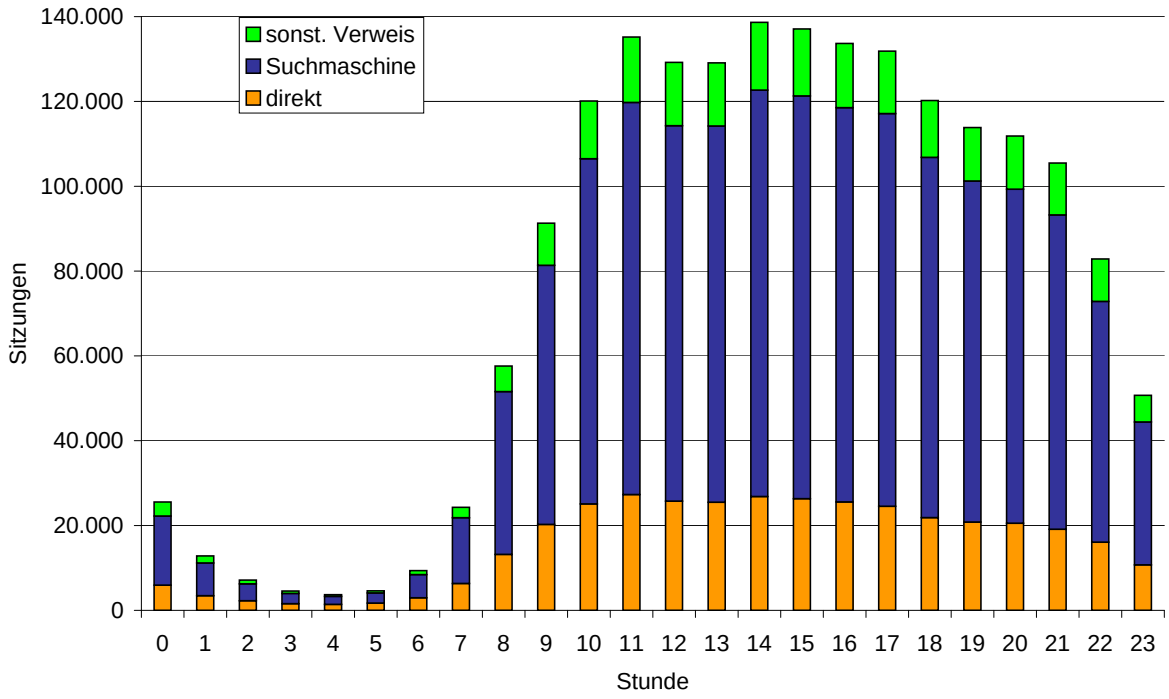


Abbildung 12.13.: Absolute Häufigkeiten der Sitzungen nach Tagesstunde und Einstiegsart

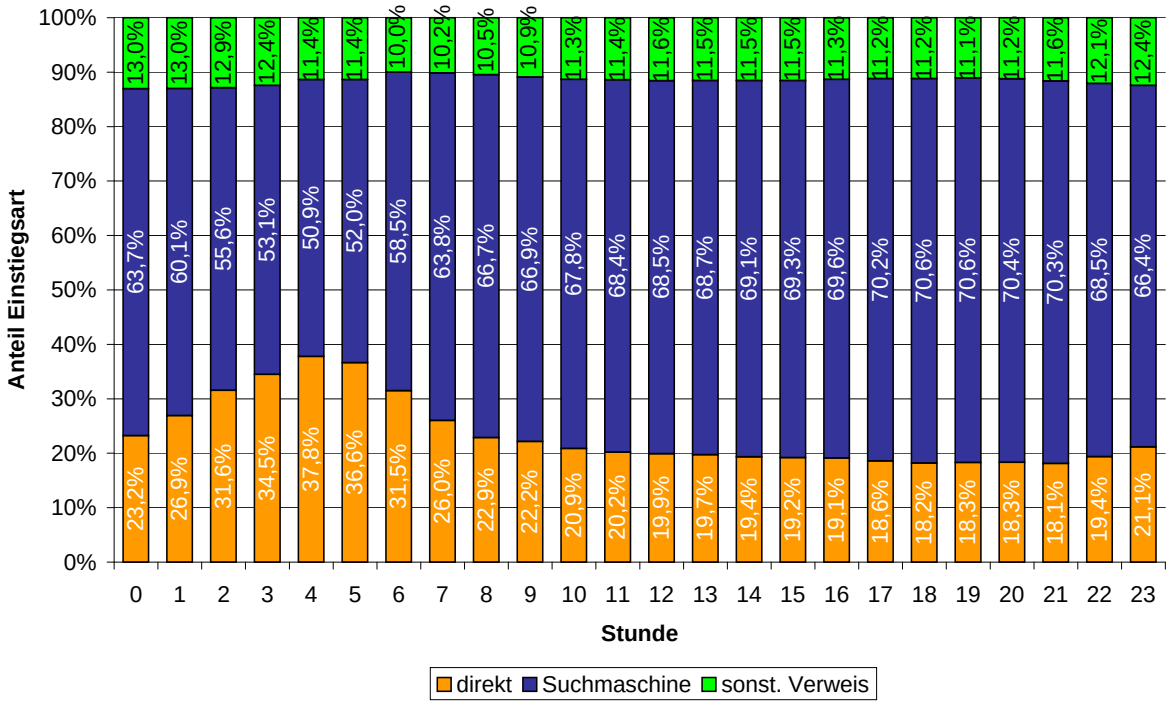


Abbildung 12.14.: Anteile der Einstiegsarten am Beginn der Sitzungen nach Tagesstunde

E. Auswertungsergebnisse zur DBS-Nutzung

E.1. Anzahl der Aktionen

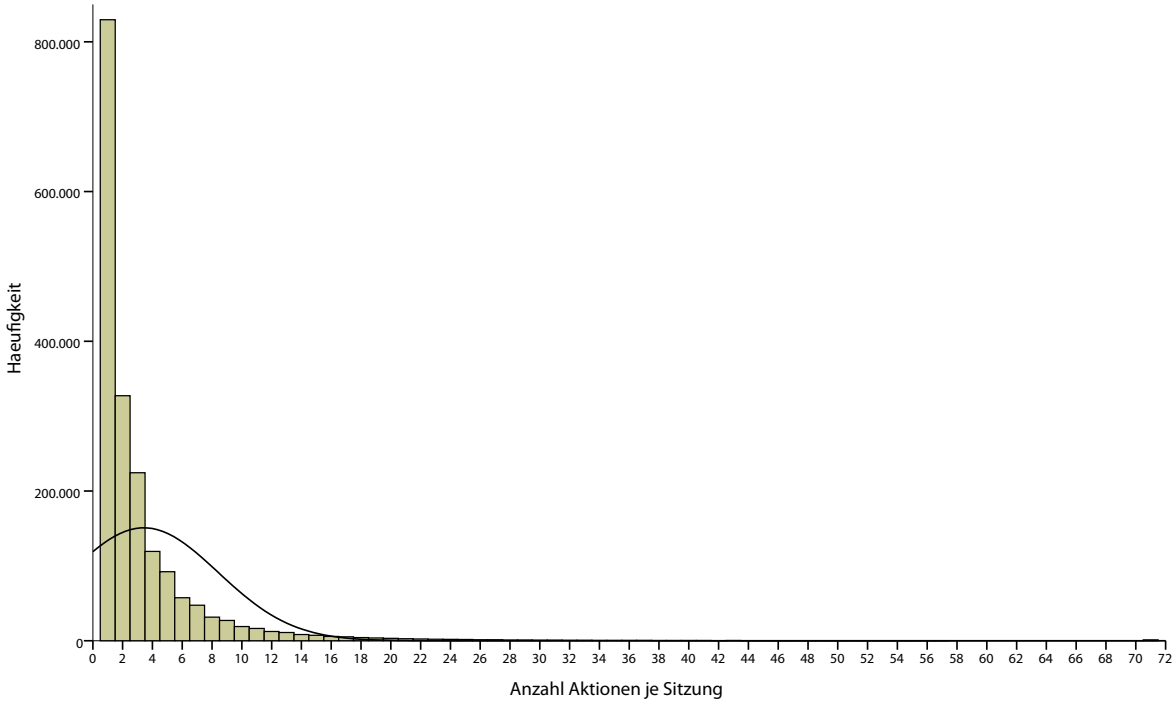
Tabelle 12.10.: Häufigkeitsverteilung der Aktionszahl je Sitzung

a[5]	h[6]	%	a	h	%	a	h	a	h	a	h
1	829.503	44,10%	47	196	0,010%	93	16	140	3	193	1
2	327.263	17,40%	48	152	0,008%	94	14	141	3	194	1
3	224.459	11,93%	49	191	0,010%	95	17	142	5	196	1
4	119.249	6,34%	50	168	0,009%	96	11	143	2	197	1
5	92.232	4,90%	51	167	0,009%	97	16	144	1	199	1
6	57.440	3,05%	52	171	0,009%	98	23	145	2	202	1
7	47.508	2,53%	53	127	0,007%	99	8	146	2	204	1
8	31.472	1,67%	54	139	0,007%	100	8	147	2	210	1
9	27.162	1,44%	55	136	0,007%	101	10	148	1	212	2
10	19.039	1,01%	56	121	0,006%	102	16	149	2	213	1
11	16.442	0,87%	57	122	0,006%	103	14	150	2	216	1
12	12.365	0,66%	58	96	0,005%	104	6	151	3	223	1
13	11.030	0,59%	59	93	0,005%	105	17	152	2	226	2
14	8.402	0,45%	60	84	0,004%	106	7	155	2	229	1
15	7.344	0,39%	61	82	0,004%	107	11	156	2	236	1
16	5.863	0,31%	62	84	0,004%	108	8	157	2	237	1
17	5.294	0,28%	63	84	0,004%	109	9	158	5	238	1
18	4.297	0,23%	64	69	0,004%	110	10	159	4	241	3
19	3.796	0,20%	65	59	0,003%	111	13	162	1	244	1
20	3.167	0,17%	66	69	0,004%	112	5	163	1	248	1
21	2.794	0,15%	67	58	0,003%	113	11	164	3	250	1
22	2.345	0,12%	68	49	0,003%	114	4	165	2	252	1
23	2.121	0,11%	69	54	0,003%	115	2	166	1	261	1
24	1.804	0,10%	70	44	0,002%	116	11	167	2	269	1
25	1.737	0,09%	71	43	0,002%	117	4	169	2	271	1
26	1.399	0,07%	72	32	0,002%	118	3	170	3	281	1
27	1.337	0,07%	73	41	0,002%	119	4	171	1	285	1
28	1.105	0,06%	74	47	0,002%	120	4	173	1	287	1
29	1.017	0,05%	75	40	0,002%	121	4	174	1	292	1
30	889	0,05%	76	30	0,002%	122	6	175	1	296	1
31	846	0,04%	77	26	0,001%	123	5	176	3	298	1
32	742	0,04%	78	35	0,002%	124	3	177	1	299	1
33	651	0,03%	79	29	0,002%	125	8	178	1	301	1
34	604	0,03%	80	37	0,002%	126	3	179	1	357	1
35	584	0,03%	81	22	0,001%	127	6	180	1	360	1
36	450	0,02%	82	26	0,001%	129	2	181	3	368	1
37	458	0,02%	83	26	0,001%	130	9	182	1	377	1
38	411	0,02%	84	24	0,001%	131	5	183	1	408	1
39	394	0,02%	85	30	0,002%	132	3	184	3	416	1
40	327	0,02%	86	24	0,001%	133	5	185	3	428	1
41	346	0,02%	87	20	0,001%	134	4	186	1	452	1
42	266	0,01%	88	22	0,001%	135	3	188	1	470	1
43	334	0,02%	89	24	0,001%	136	5	189	3	483	1

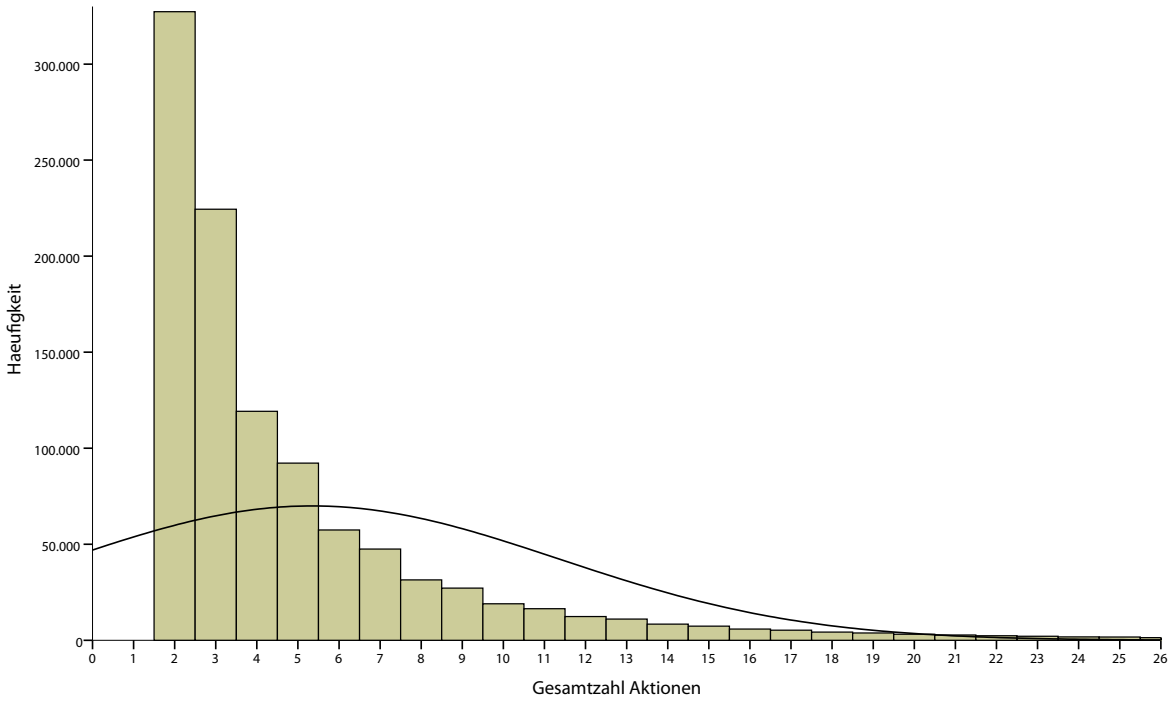
weiter auf der nächsten Seite

[5] a = Anzahl der Aktionen
[6] h = Häufigkeit der Aktionsanzahl

44	256	0,01%	90	21	0,001%	137	6	190	2	495	1
45	233	0,01%	91	21	0,001%	138	1	191	3	513	1
46	221	0,01%	92	19	0,001%	139	1	192	3		



(a) Histogramm der Aktionsanzahl je Sitzung mit Darstellung der Normalverteilung



(b) Histogramm der Aktionsanzahl je Sitzung unter Ausschluss des Wertes 1 mit Darstellung der Normalverteilung

Abbildung 12.15.: Histogramme der Aktionszahl

Tabelle 12.11.: Häufigkeitsverteilung der Aktionszahl je Sitzung und Einstiegsart

Anz. Aktionen	direkt		Suchmaschine		sonst. Verweis		Gesamt
$\sum$	374.843	19,93%	1.291.313	68,66%	214.601	11,41%	1.880.757
$\overline{x}$	3,41		3,39		3,88		3,45
s	6,81		4,90		5,93		5,46
1	206.327	24,87%	538.070	64,87%	85.106	10,26%	829.503
2	51.888	15,86%	238.639	72,92%	36.736	11,23%	327.263
3	28.589	12,74%	170.232	75,84%	25.638	11,42%	224.459
4	19.160	16,07%	85.123	71,38%	14.966	12,55%	119.249
5	13.816	14,98%	66.306	71,89%	12.110	13,13%	92.232
6	10.005	17,42%	39.706	69,13%	7.729	13,46%	57.440
7	7.481	15,75%	33.583	70,69%	6.444	13,56%	47.508
8	5.661	17,99%	21.470	68,22%	4.341	13,79%	31.472
9	4.751	17,49%	18.651	68,67%	3.760	13,84%	27.162
10	3.707	19,47%	12.616	66,26%	2.716	14,27%	19.039
11	3.053	18,57%	11.050	67,21%	2.339	14,23%	16.442
12	2.565	20,74%	8.080	65,35%	1.720	13,91%	12.365
13	2.169	19,66%	7.278	65,98%	1.583	14,35%	11.030
14	1.715	20,41%	5.472	65,13%	1.215	14,46%	8.402
15	1.527	20,79%	4.764	64,87%	1.053	14,34%	7.344
16	1.262	21,52%	3.758	64,10%	843	14,38%	5.863
17	1.164	21,99%	3.353	63,34%	777	14,68%	5.294
18	1.037	24,13%	2.677	62,30%	583	13,57%	4.297
19	825	21,73%	2.420	63,75%	551	14,52%	3.796
20	729	23,02%	1.994	62,96%	444	14,02%	3.167
21	630	22,55%	1.770	63,35%	394	14,10%	2.794
22	608	25,93%	1.411	60,17%	326	13,90%	2.345
23	511	24,09%	1.313	61,90%	297	14,00%	2.121
24	467	25,89%	1.064	58,98%	273	15,13%	1.804
25	416	23,95%	1.055	60,74%	266	15,31%	1.737
26	328	23,45%	873	62,40%	198	14,15%	1.399
27	334	24,98%	798	59,69%	205	15,33%	1.337
28	315	28,51%	644	58,28%	146	13,21%	1.105
29	244	23,99%	618	60,77%	155	15,24%	1.017
30	217	24,41%	544	61,19%	128	14,40%	889
31	237	28,01%	505	59,69%	104	12,29%	846
32	206	27,76%	433	58,36%	103	13,88%	742
33	169	25,96%	380	58,37%	102	15,67%	651
34	177	29,30%	332	54,97%	95	15,73%	604
35	158	27,05%	344	58,90%	82	14,04%	584
36	129	28,67%	250	55,56%	71	15,78%	450
37	137	29,91%	256	55,90%	65	14,19%	458
38	119	28,95%	230	55,96%	62	15,09%	411
39	112	28,43%	231	58,63%	51	12,94%	394
40	97	29,66%	186	56,88%	44	13,46%	327
41	105	30,35%	185	53,47%	56	16,18%	346
42	83	31,20%	147	55,26%	36	13,53%	266
43	83	24,85%	202	60,48%	49	14,67%	334
44	68	26,56%	151	58,98%	37	14,45%	256
45	80	34,33%	125	53,65%	28	12,02%	233
46	88	39,82%	106	47,96%	27	12,22%	221
47	59	30,10%	110	56,12%	27	13,78%	196
48	56	36,84%	74	48,68%	22	14,47%	152

weiter auf der nächsten Seite

Fortsetzung

Anz. Aktionen	direkt		Suchmaschine		sonst. Verweis		Gesamt
49	53	27,75%	118	61,78%	20	10,47%	191
50	52	30,95%	91	54,17%	25	14,88%	168
51	48	28,74%	96	57,49%	23	13,77%	167
52	46	26,90%	107	62,57%	18	10,53%	171
53	44	34,65%	69	54,33%	14	11,02%	127
54	60	43,17%	57	41,01%	22	15,83%	139
55	37	27,21%	67	49,26%	32	23,53%	136
56	42	34,71%	61	50,41%	18	14,88%	121
57	35	28,69%	68	55,74%	19	15,57%	122
58	42	43,75%	46	47,92%	8	8,33%	96
59	29	31,18%	54	58,06%	10	10,75%	93
60	30	35,71%	44	52,38%	10	11,90%	84
61	30	36,59%	34	41,46%	18	21,95%	82
62	21	25,00%	54	64,29%	9	10,71%	84
63	26	30,95%	48	57,14%	10	11,90%	84
64	34	49,28%	27	39,13%	8	11,59%	69
65	10	16,95%	40	67,80%	9	15,25%	59
66	24	34,78%	38	55,07%	7	10,14%	69
67	19	32,76%	27	46,55%	12	20,69%	58
68	17	34,69%	22	44,90%	10	20,41%	49
69	13	24,07%	34	62,96%	7	12,96%	54
70	16	36,36%	16	36,36%	12	27,27%	44
71+	451	39,42%	516	45,10%	177	15,47%	1.144
Gesamt	374.843	19,93%	1.291.313	68,66%	214.601	11,41%	1.880.757

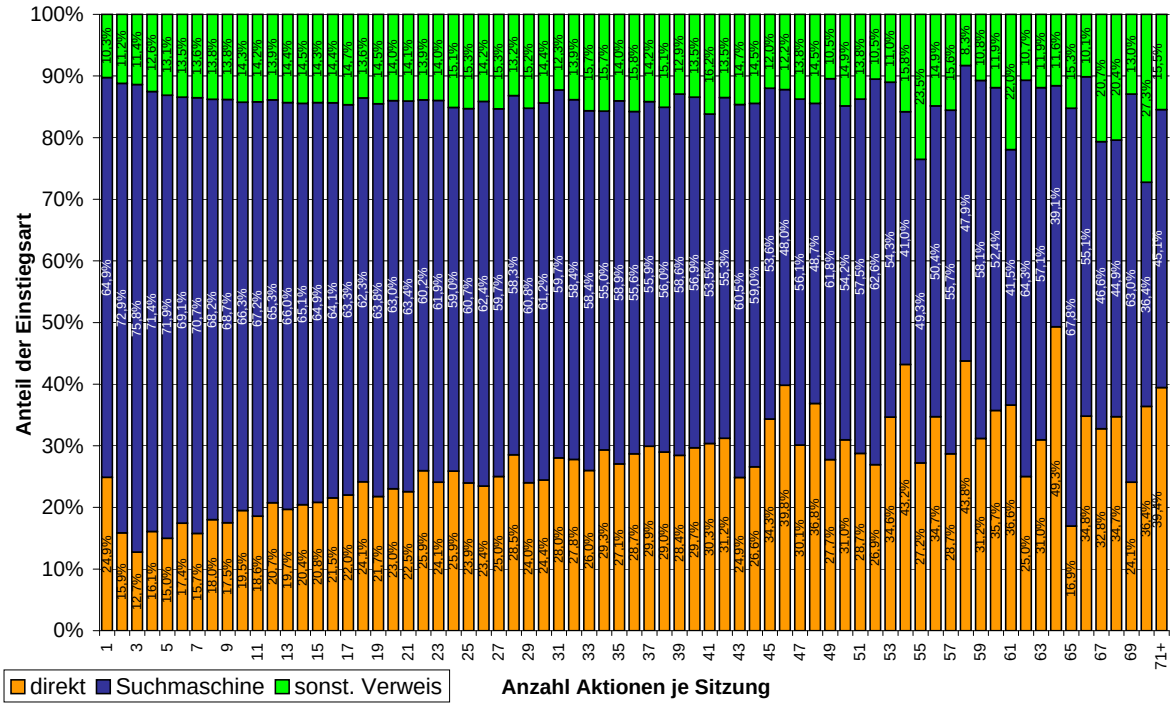


Abbildung 12.16.: Anteile der Einstiegsarten nach Anzahl der Aktionen je Sitzung

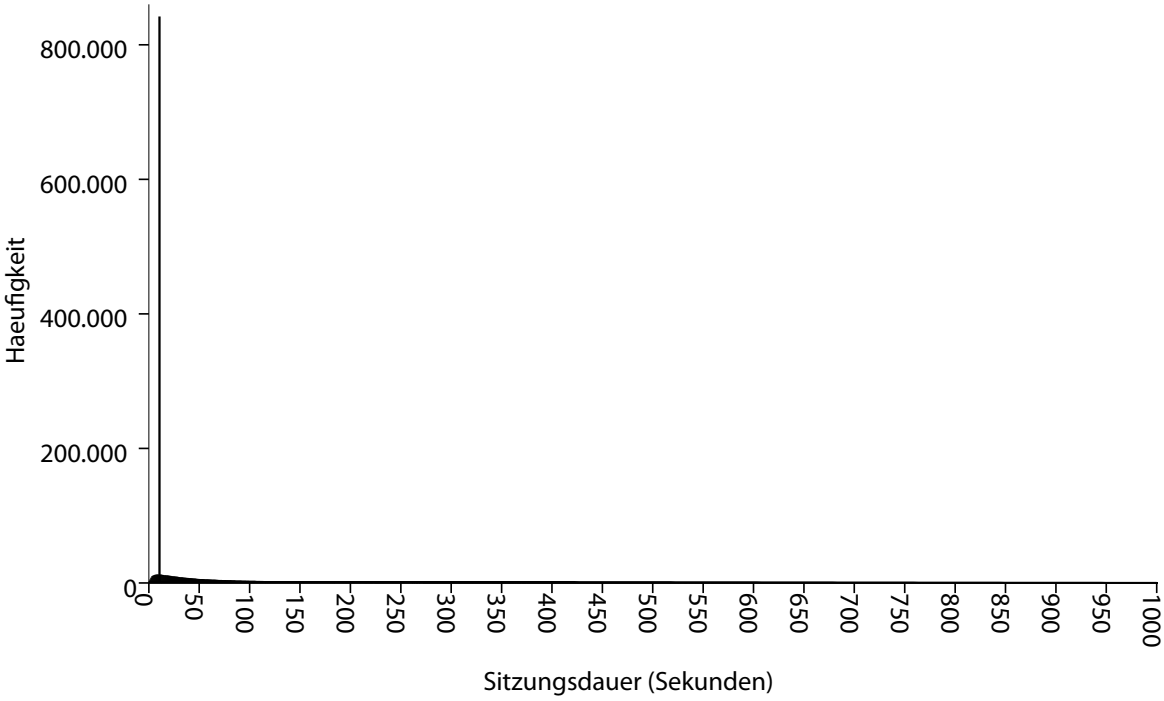
E.2. Sitzungsdauer

	mit 1-Akt.-Sitzungen		**ohne 1-Akt.-Sitzungen**	
Sitzungsdauer (Sek.)	**Häufigkeit**	**Prozent**	**Häufigkeit**	**Prozent**
$\leq$ 10	929.713	49,4%	100.210	9,5%
11 – 30	191.754	10,2%	191.754	18,2%
31 – 60	168.277	8,9%	168.277	16,0%
61 – 180	241.555	12,8%	241.555	23,0%
181 – 600	193.020	10,3%	193.020	18,4%
601 – 1800	127.670	6,8%	127.670	12,1%
$\geq$1801	28.768	1,5%	28.768	2,7%
Gesamt	1.880.757	100,0%	1.051.254	100,0%

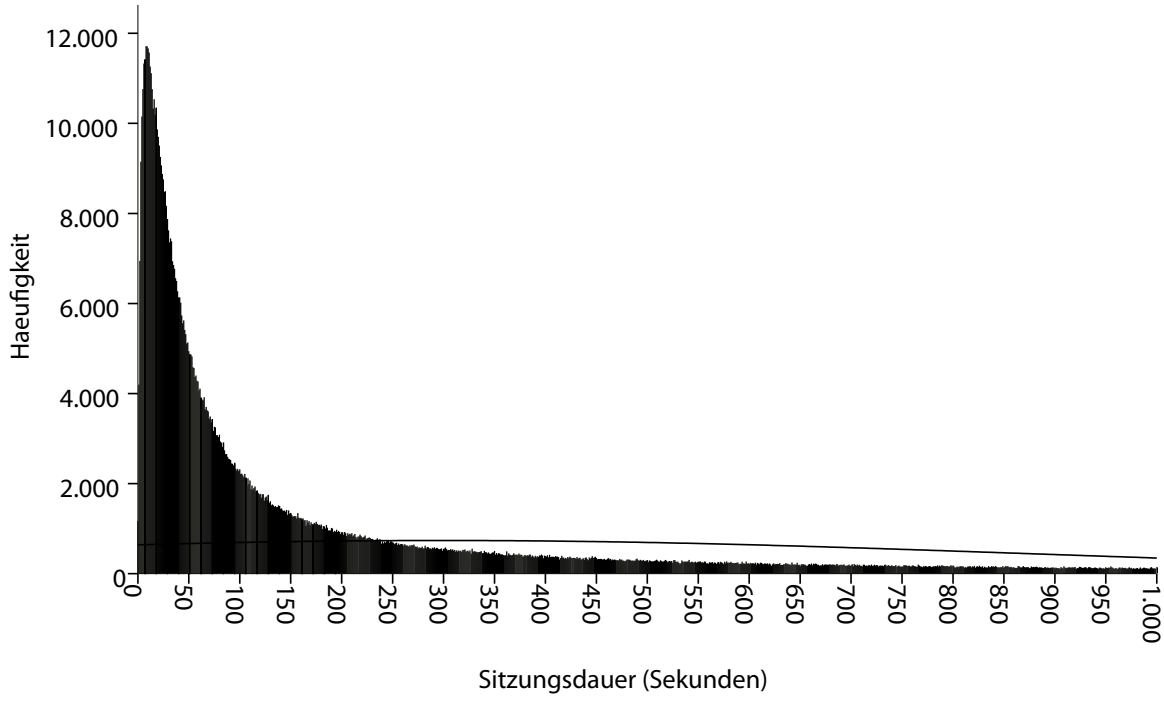
Tabelle 12.12.: Häufigkeit von Sitzungsdauern nach Klassen

Sitzungsdauer (Sek.)	**direkt**		**Suchmaschine**		**sonst. Verweis**		**Gesamt**	
$\leq$10	17.940	17,9%	71.433	71,3%	10.837	10,8%	100.210	100%
11–30	22.352	11,7%	146.177	76,2%	23.225	12,1%	191.754	100%
31–60	20.451	12,2%	126.300	75,1%	21.526	12,8%	168.277	100%
61–180	34.601	14,3%	174.594	72,3%	32.360	13,4%	241.555	100%
181–600	34.904	18,1%	133.889	69,4%	24.227	12,6%	193.020	100%
601–1800	29.286	22,9%	84.110	65,9%	14.274	11,2%	127.670	100%
$\geq$1800	8.982	31,2%	16.740	58,2%	3.046	10,6%	28.768	100%
Gesamt	168.516	16,0%	753.243	71,7%	129.495	12,3%	1.051.254	100%

Tabelle 12.13.: Häufigkeit von klassierten Sitzungsdauern nach Einstiegsart unter Ausschluss der Ein-Aktionen-Sitzungen

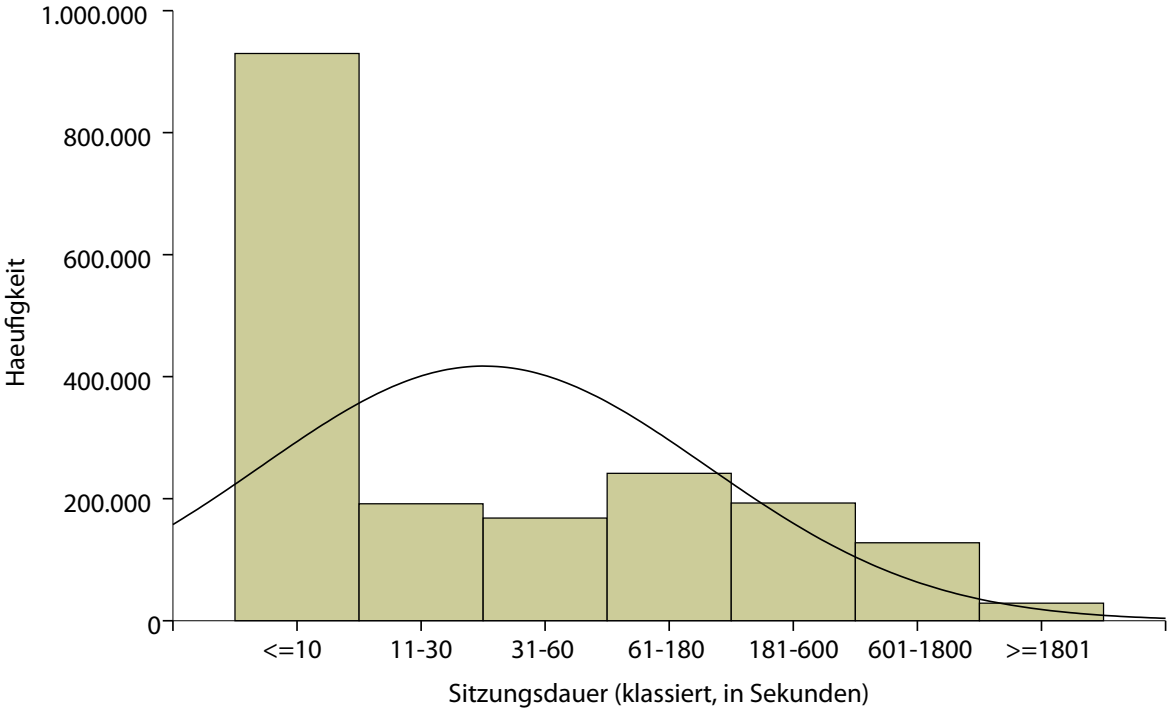


(a) Histogramm der Sitzungsdauer mit Darstellung der Normalverteilung

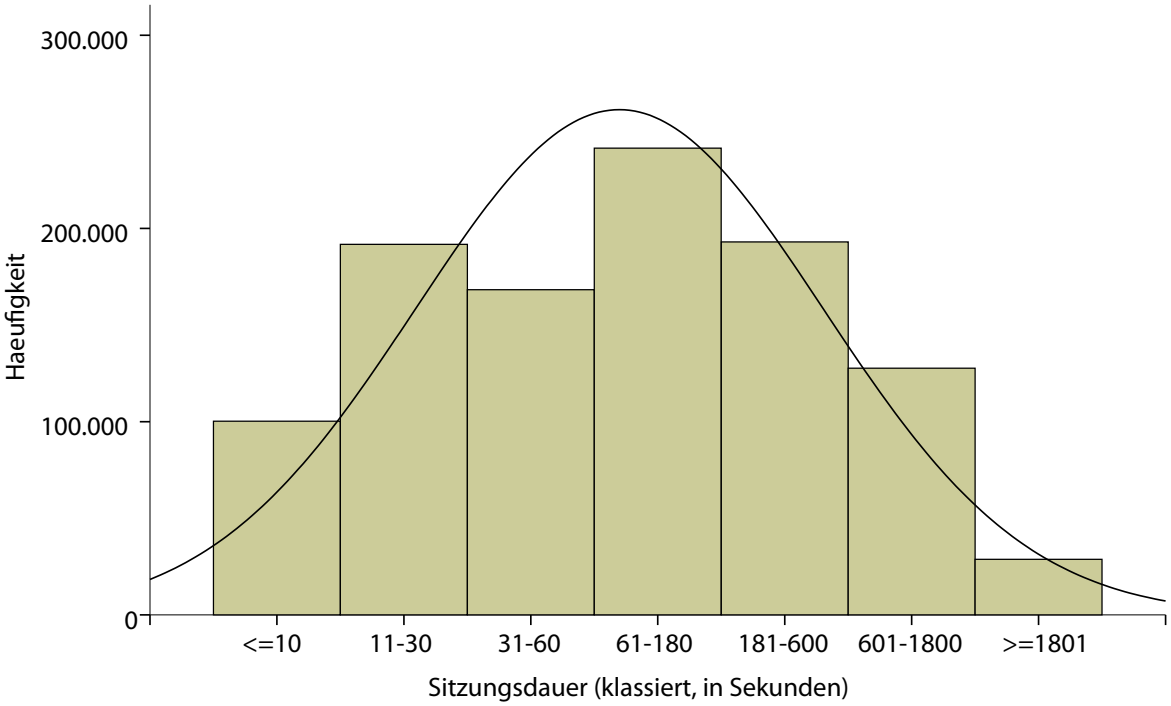


(b) Histogramm der Sitzungsdauer unter Ausschluss der Ein-Aktionen-Sitzungen mit Darstellung der Normalverteilung

Abbildung 12.17.: Histogramme der Sitzungsdauer bis 1.000 Sekunden



(a) Histogramm der klassierten Sitzungsdauern einschließlich der Ein-Aktionen-Sitzungen mit Normalverteilung



(b) Histogramm der klassierten Sitzungsdauern ausschließlich der Ein-Aktionen-Sitzungen mit Normalverteilung

Abbildung 12.18.: Häufigkeit der klassierten Sitzungsdauern

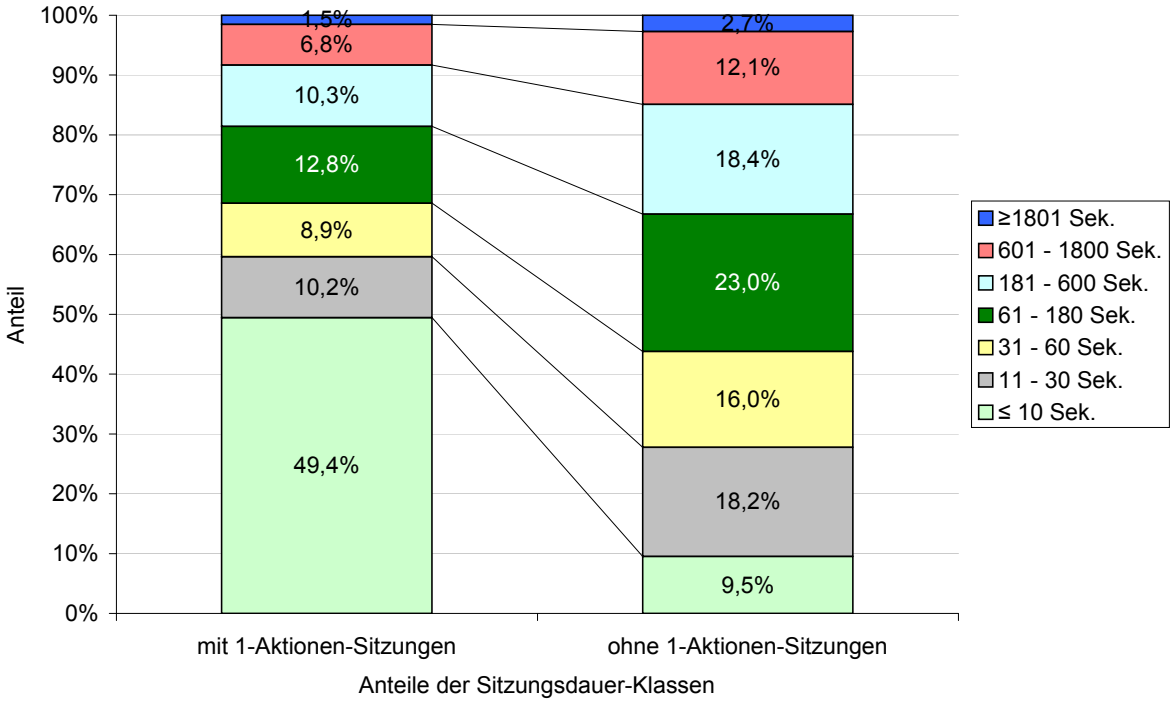
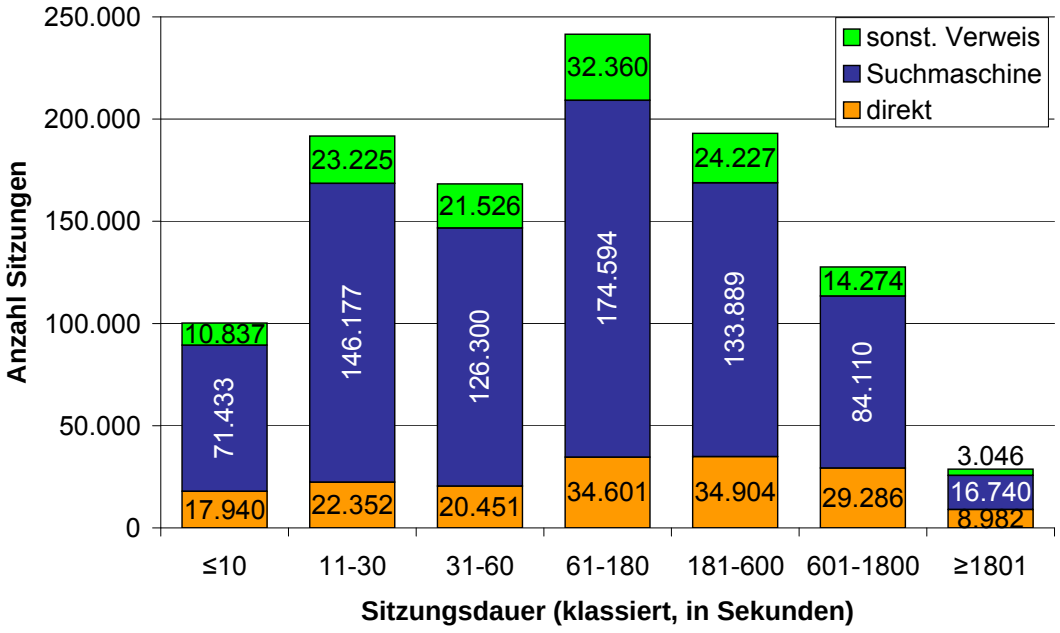
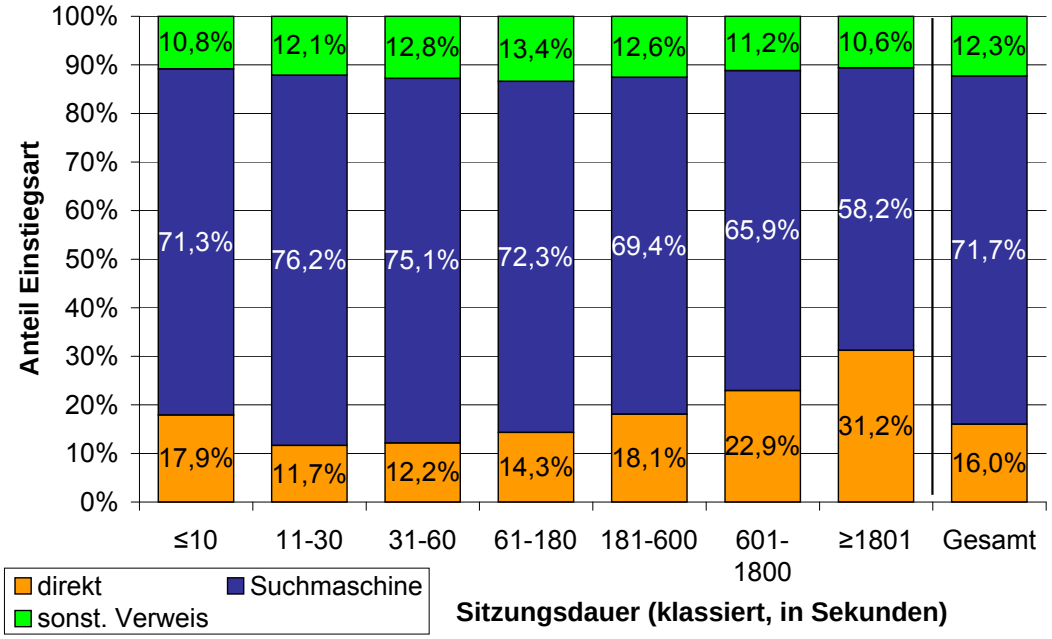


Abbildung 12.19.: Anteile der Sitzungsdauerklassen mit und ohne Berücksichtigung der Ein-Aktionen-Sitzungen



(a) Häufigkeiten



(b) Anteile

Abbildung 12.20.: Häufigkeiten und Anteile der Einstiegsarten an Sitzungsdauerklassen

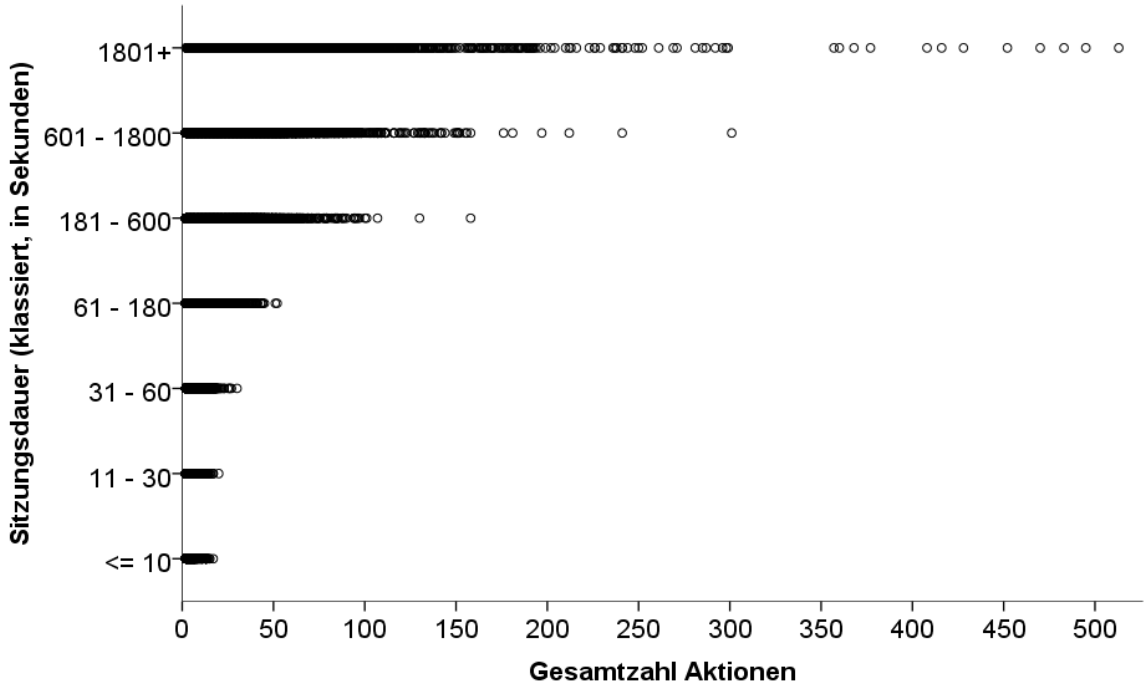


Abbildung 12.21.: Grafische Darstellung des Zusammenhangs zwischen Anzahl der Aktionen und klassierter Sitzungsdauer

E.3. Absprünge

	direkt	Suchmaschine	sonst. Verweis	Gesamt
Absprünge	206.327	538.070	85.106	829.503
Fortsetzungen	168.516	753.243	129.495	1.051.254
Gesamt	374.843	1.291.313	214.601	1.880.757
Absprungrate	55,0%	41,7%	39,7%	44,1%
Fortsetzungsrate	45,0%	58,3%	60,3%	55,9%
Gesamt	100,0%	100,0%	100,0%	100,0%

Tabelle 12.14.: Absprünge und Absprungrate nach Einstiegstyp

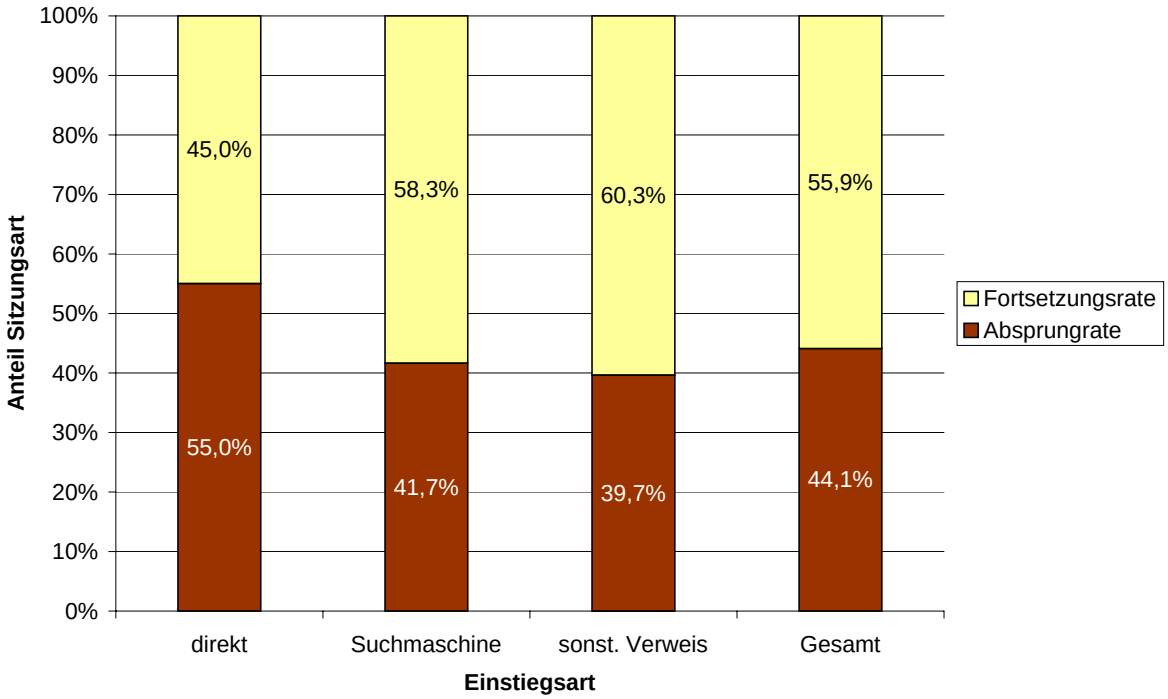


Abbildung 12.22.: Verteilung von Absprung- und Fortsetzungsrate nach Einstiegsart

E.4. Wiederkehrende Besucher

	direkt	Suchmaschine	sonst. Verweis	Gesamt
Neuer Besucher	196.890	1.084.122	182.710	1.463.722
Wiederkehrender Besucher	177.953	207.191	31.891	417.035
Gesamt	374.843	1.291.313	214.601	1.880.757
Neuer Besucher	52,5%	84,0%	85,1%	77,8%
Wiederkehrender Besucher	47,5%	16,0%	14,9%	22,2%
Gesamt	100,0%	100,0%	100,0%	100,0%

Tabelle 12.15.: Anteile neuer und wiederkehrender Besucher nach Einstiegstyp

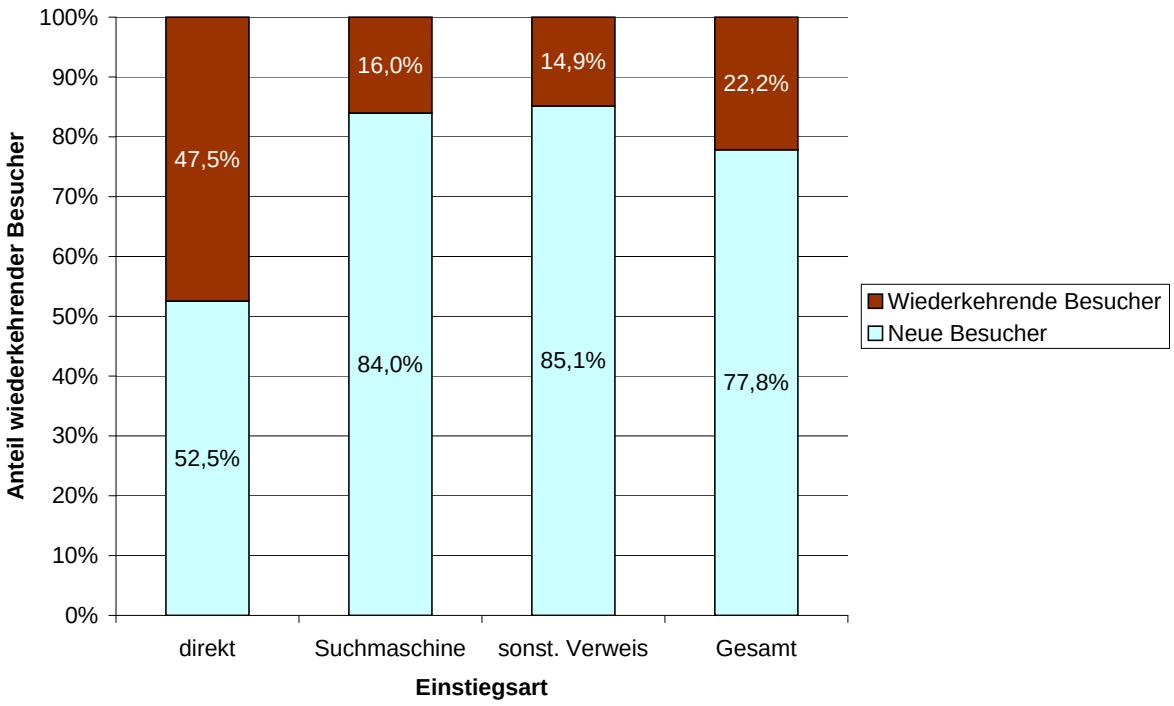


Abbildung 12.23.: Anteile wiederkehrender und neuer Besucher nach Einstiegsart

F. Auswertungsergebnisse zum DBS-Ausstieg

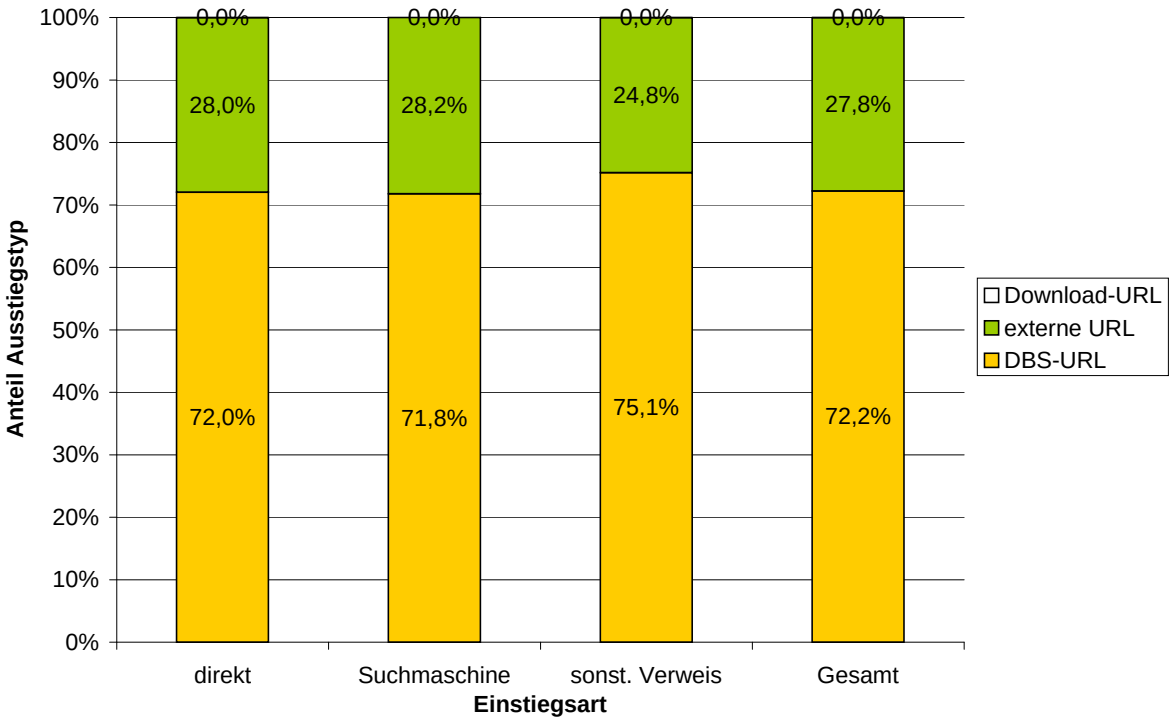


Abbildung 12.24.: Typ der letzten Sitzungsaktion nach Einstiegsart

G. Ergebnisse der Web Usage Mining-Methoden

G.1. Inhaltskategorien

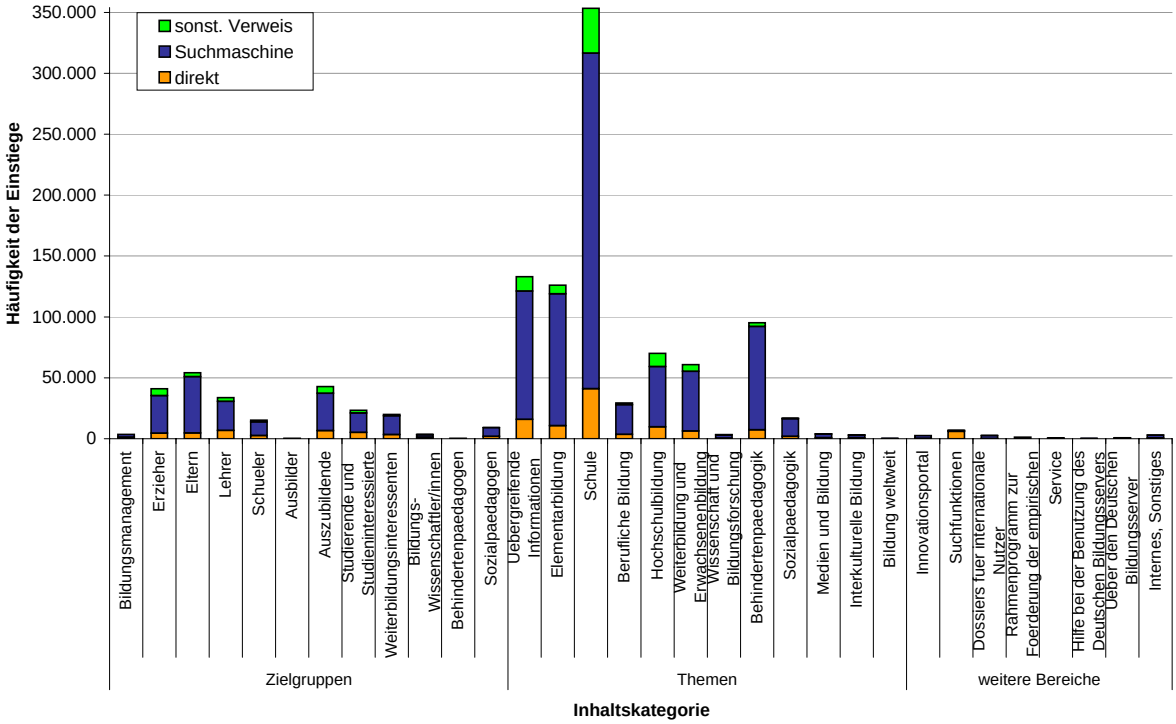


Abbildung 12.25.: Häufigkeiten der Einstiege in Inhaltskategorien nach Einstiegsart

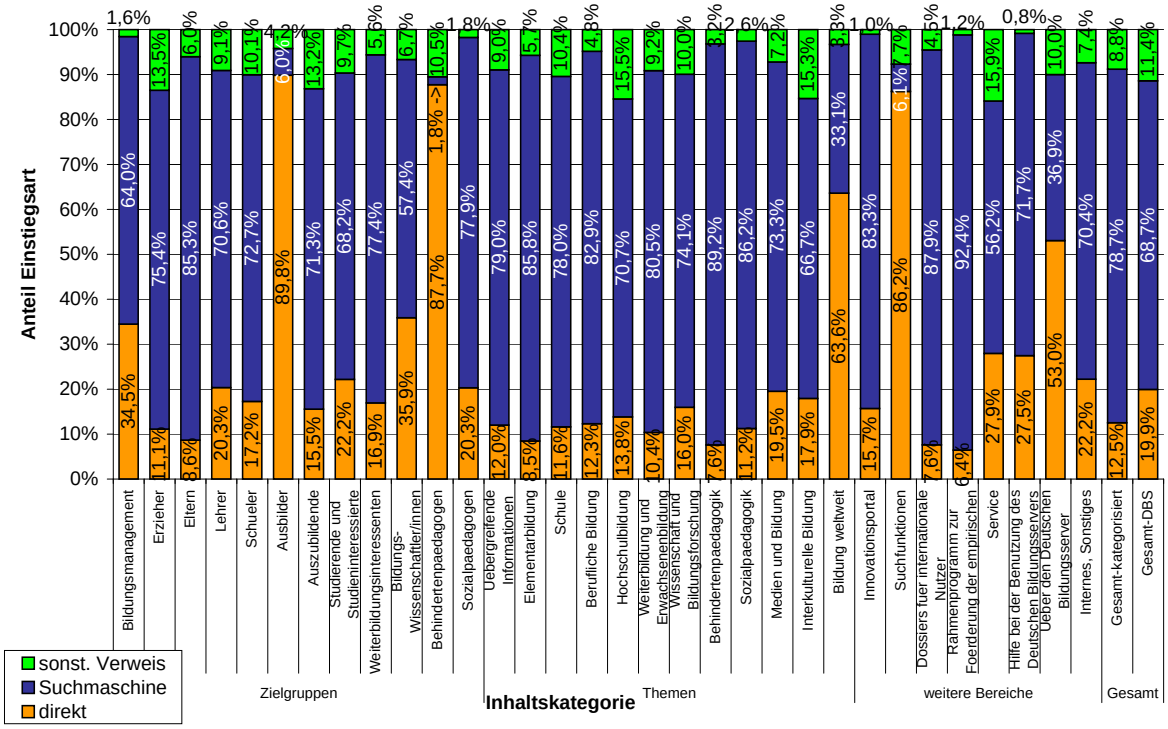


Abbildung 12.26.: Anteile der Einstiege in Inhaltskategorien nach Einstiegsart

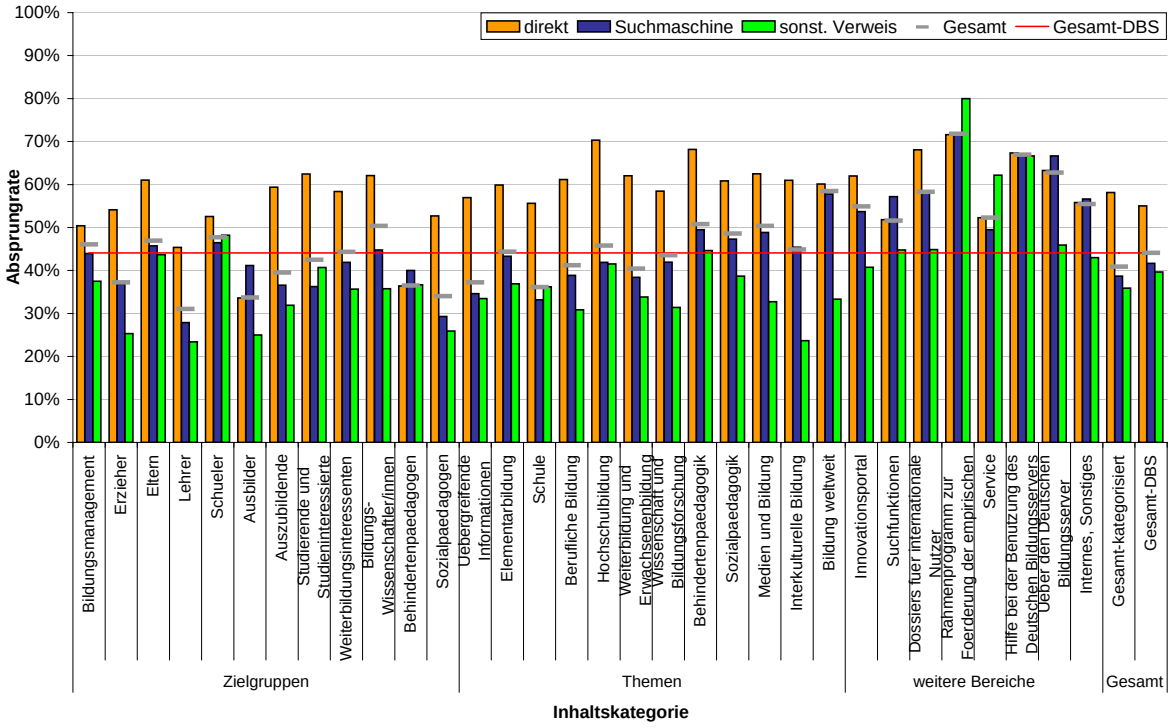


Abbildung 12.27.: Absprungrate nach Inhaltskategorien und Einstiegsart

	Verweisseite (Typ „sonstiger Verweis“)	**Einst.**	**Abspr.**	**Forts.**	**Abspr.rate**
1	http://eu.daad.de/eu/sokrates/infos-studierende/05393.html	402	202	200	50,3%
2	http://www.tektorum.de/andere-themen/3961-stipendien-fuer-architekten.html	113	55	58	48,7%
3	http://www.hs-nb.de/hochschule-nb/studium/studienberatung/	64	45	19	70,3%
4	http://wiki.bildungsserver.de/infoboerse/index.php/Stipendien_und_F%C3%B6rderm%C3%B6glichkeiten	207	42	165	20,3%
5	http://www.erasmus-berichte.de/erasmus/web.htm	93	32	61	34,4%
6	http://www.maschinenbau.tu-darmstadt.de/studieren/stipendien_foerderung/studienstiftungen/index.de.jsp	47	27	20	57,5%
7	http://ctech.e-technik.uni-ulm.de/ctech/bin/view/CTech/CTechFees	29	19	10	65,5%
8	http://www.hfm-weimar.de/v1/studium/stipendien/stiftungen.php	52	16	36	30,8%
9	http://www.gls-sprachenzentrum.de/242_auslandsstudium.html	30	15	15	50,0%
10	http://wiki.bildungsserver.de/infoboerse/index.php/Zweiter_Bildungsweg	79	14	65	17,7%
11	http://www.hmt-rostock.de/daad.html	26	14	12	53,9%
12	http://lotse.uni-muenster.de/psychologie/lernen_und_forschen/aus_und_weiterbildung/stipendien-de.php	39	13	26	33,3%
13	http://www.familien-wegweiser.de/bmfsfj/generator/wegweiser/stichwortverzeichnis,did=45808.html	24	12	12	50,0%
14	http://www.bmfsfj.de/bmfsfj/generator/wegweiser/stichwortverzeichnis,did=45808.html	20	11	9	55,0%
15	http://www.cct-germany.de/index.php?ID=297&action=2&texttyp=0&zielgruppe=0&lokalisierung=DE-GER	27	11	16	40,7%

Tabelle 12.16.: Verweis-URLs für Kategorie „Studierende und Studieninteressierte“ mit den häufigsten Absprüngen (Ausschnitt)

G.2. Funktionskategorien

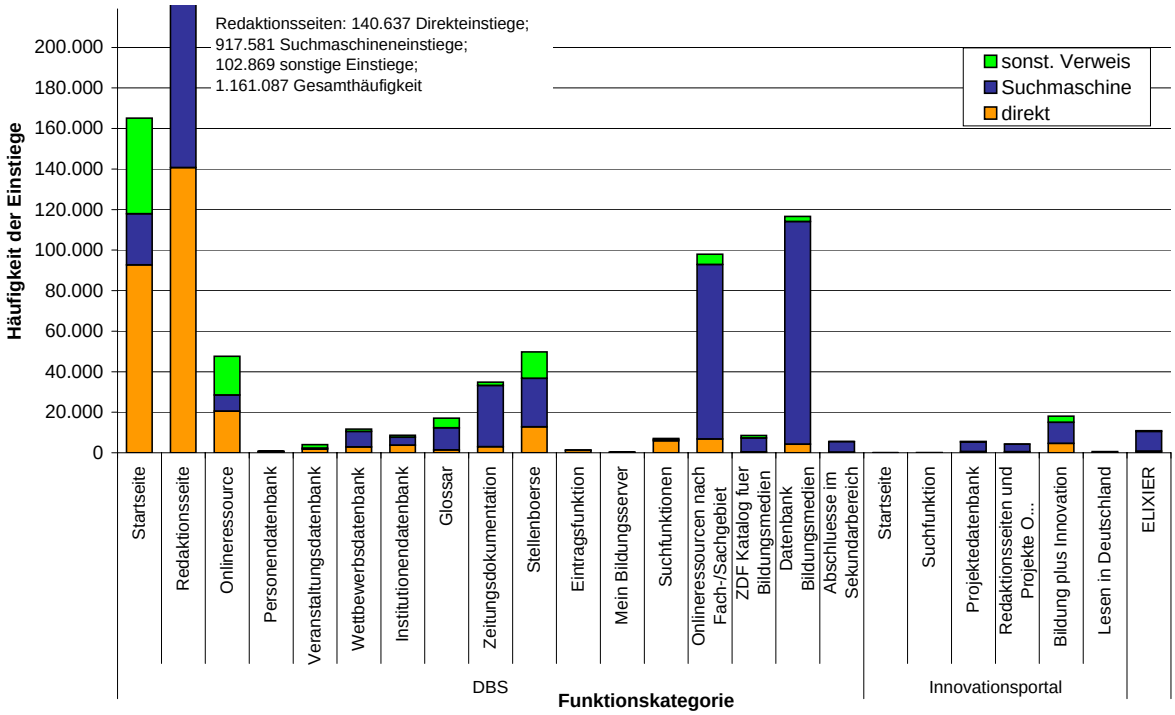


Abbildung 12.28.: Häufigkeiten der Einstiege in Funktionskategorien nach Einstiegsart

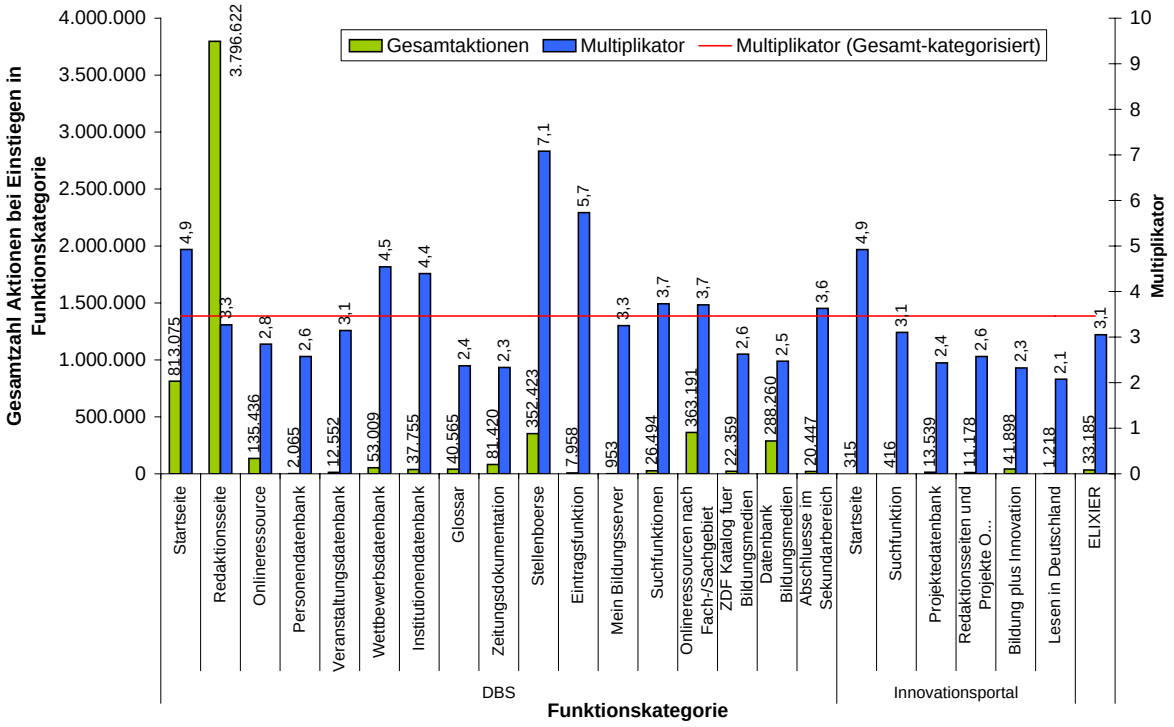


Abbildung 12.29.: Häufigkeit aller Aktionen, die auf Einstiege in den jeweiligen Kategorien folgten, und Multiplikator der Einstiege

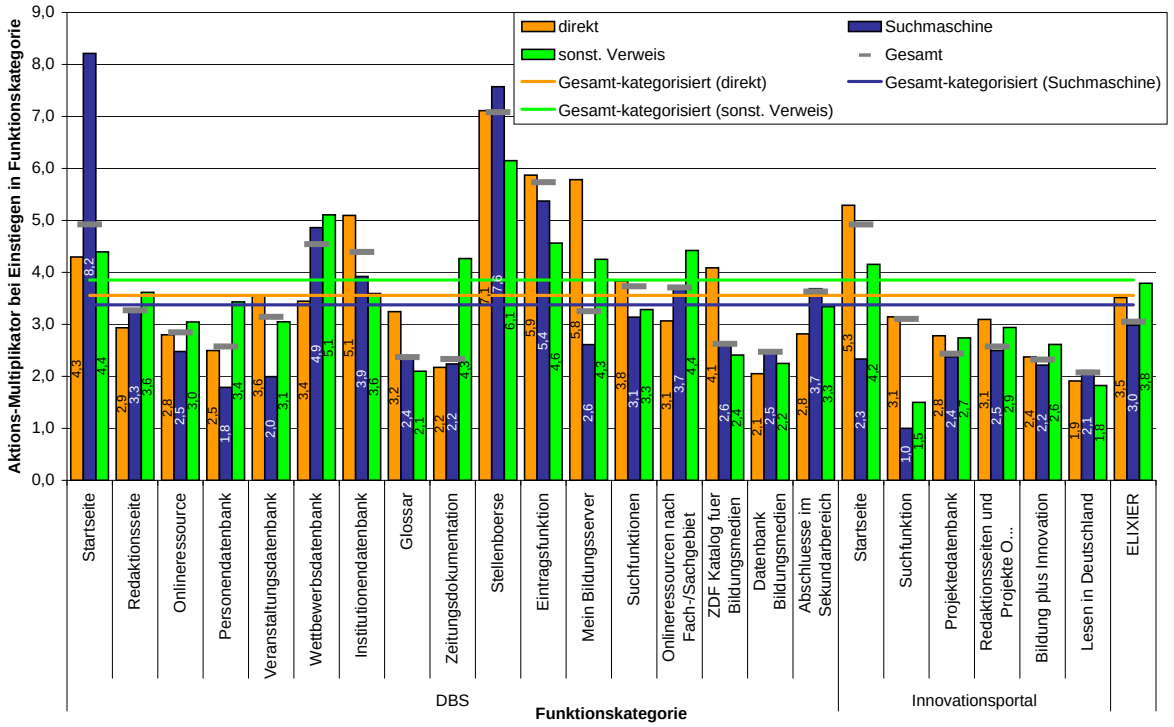


Abbildung 12.30.: Aktions-Multiplikatoren für Einstiege in Funktionskategorien nach Einstiegsart

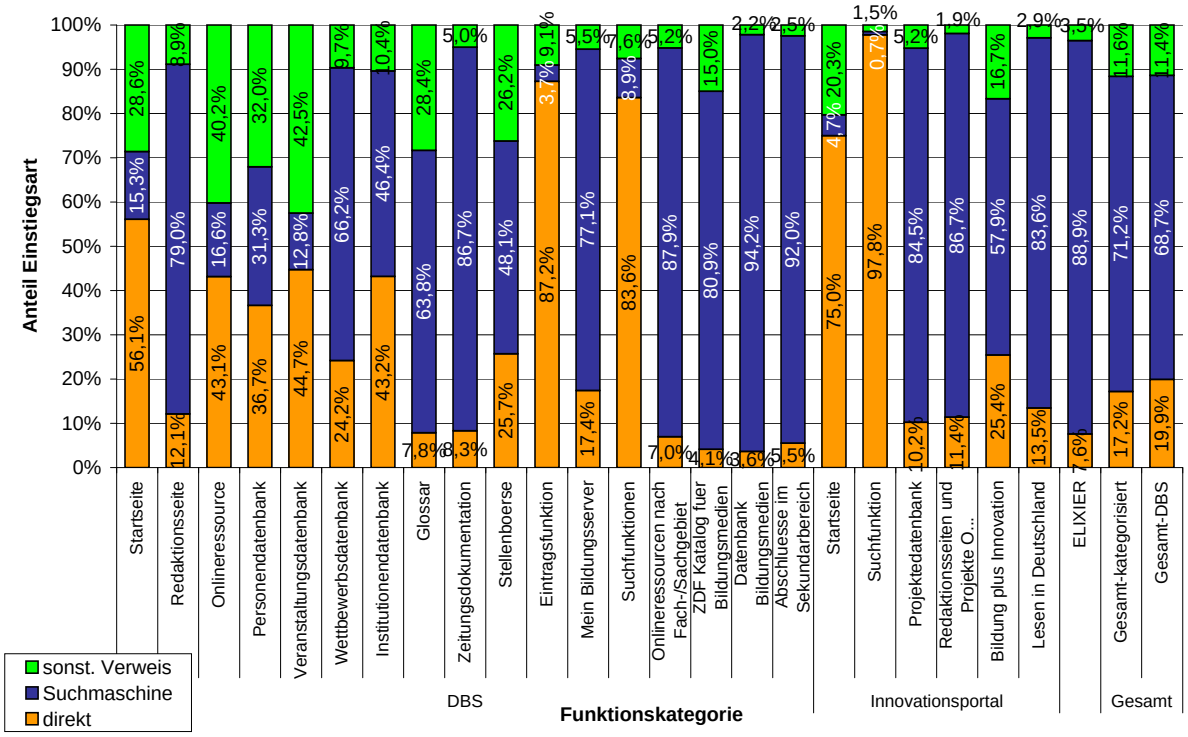


Abbildung 12.31.: Anteile der Einstiege in Funktionskategorien nach Einstiegsart

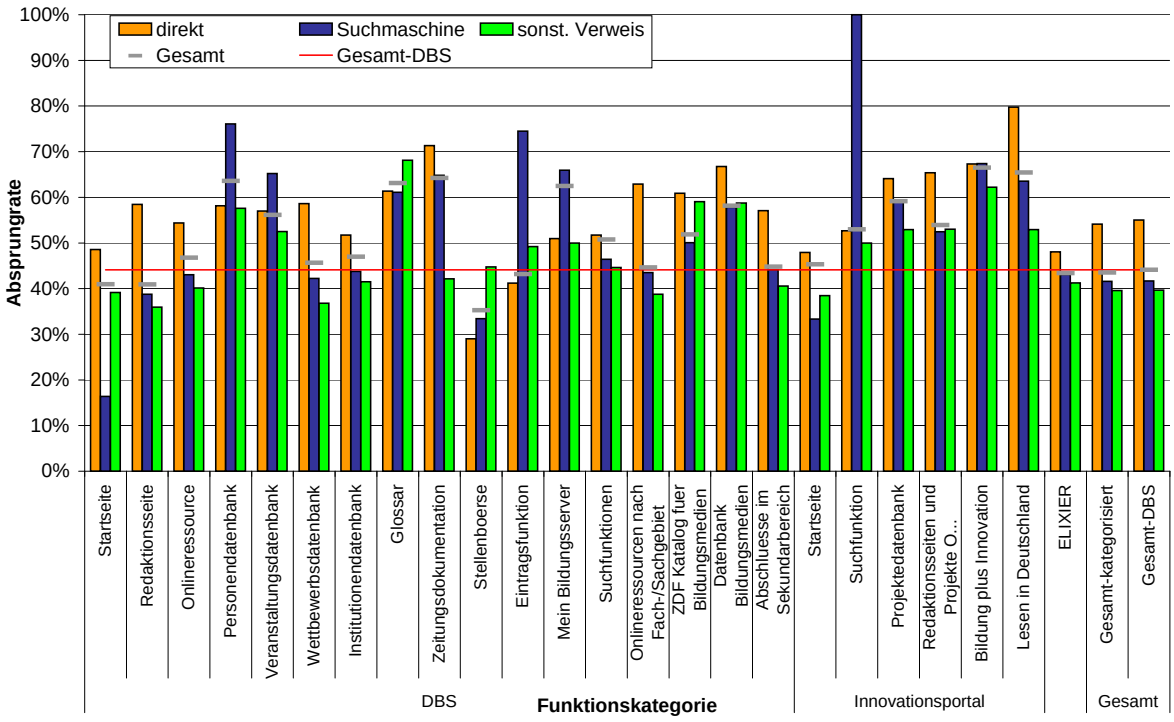


Abbildung 12.32.: Absprungrate nach Funktionskategorien und Einstiegsart

G.3. Einstiegstiefe

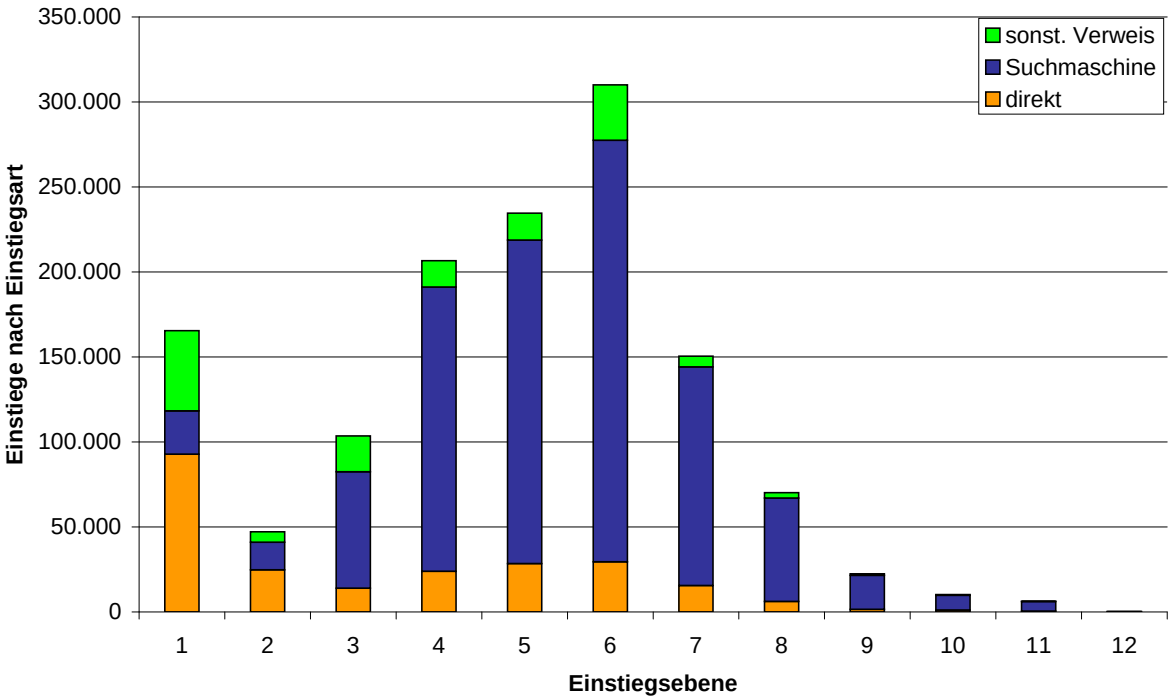


Abbildung 12.33.: Häufigkeiten der Einstiege nach Einstiegs-Hierarchieebene und Einstiegsart

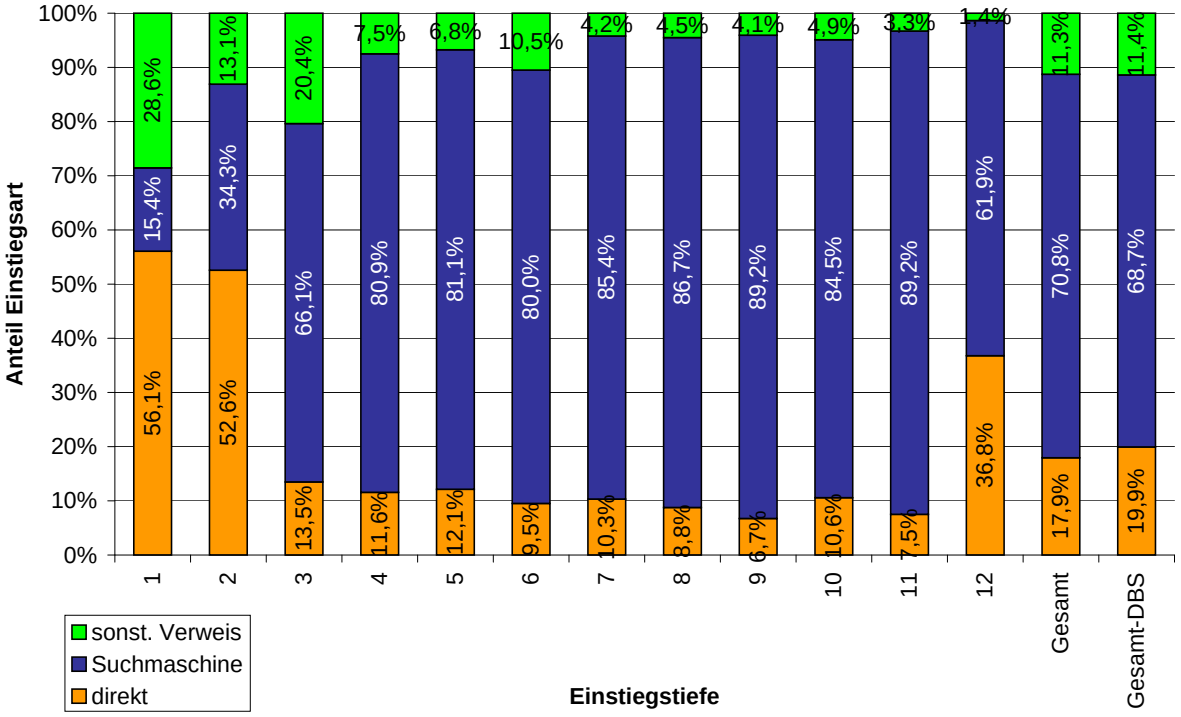


Abbildung 12.34.: Anteile der Einstiegsarten nach Einstiegs-Hierarchieebene

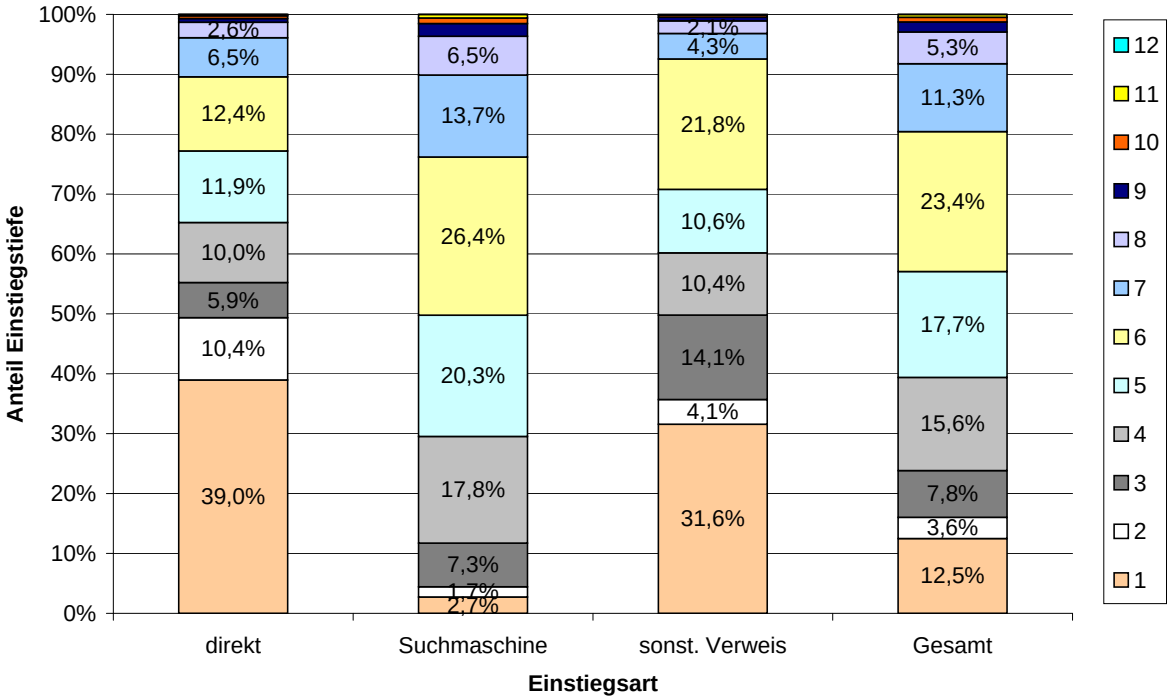


Abbildung 12.35.: Anteile der Einstiegs-Hierarchieebene nach Einstiegsart

G.4. Bewegung durch die Hierachieebenen

	direkt	Suchmaschine	sonst. Verweis	Gesamt
Aufwärts	44.187	213.810	38.475	296.472
Seitwärts	51.136	217.076	33.988	302.200
Abwärts	86.487	325.511	75.401	487.399
Gesamt	181.810	756.397	147.864	1.086.071
Aufwärts	24,3%	28,3%	26,0%	27,3%
Seitwärts	28,1%	28,7%	23,0%	27,8%
Abwärts	47,6%	43,0%	51,0%	44,9%
Gesamt	100,0%	100,0%	100,0%	100,0%

Tabelle 12.17.: Häufigkeiten und Anteile der Hierarchiebewegungen an den Einstiegsarten

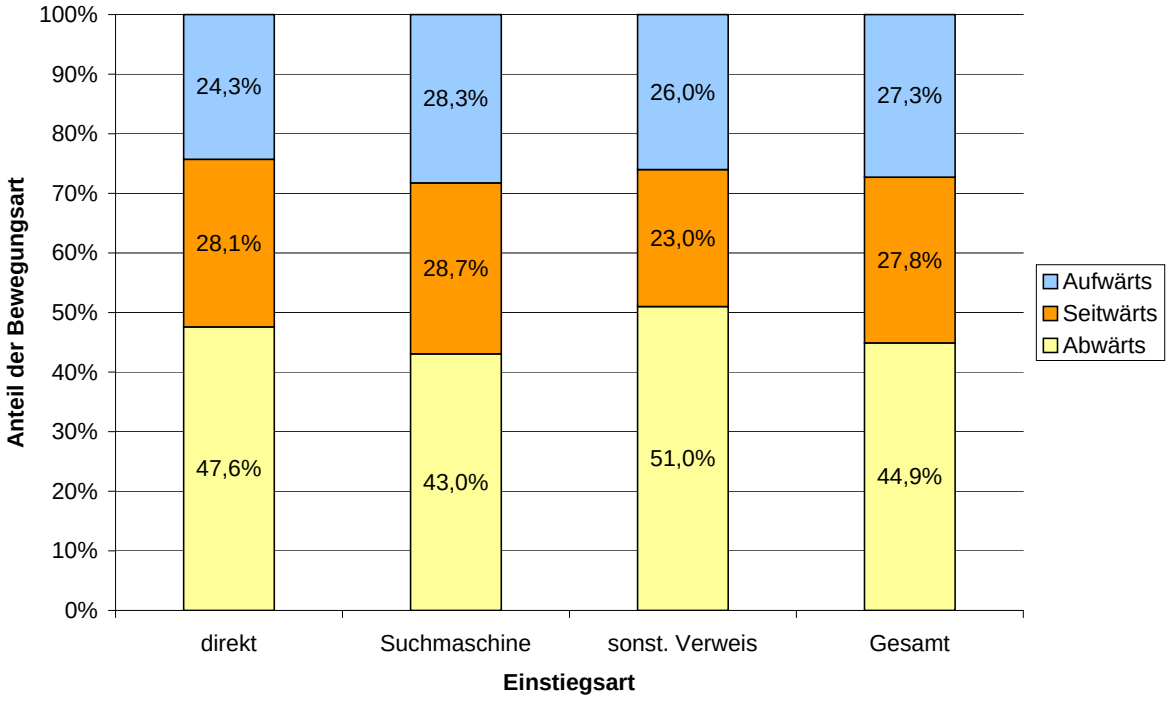


Abbildung 12.36.: Anteile der Hierarchiebewegungen an den Einstiegsarten

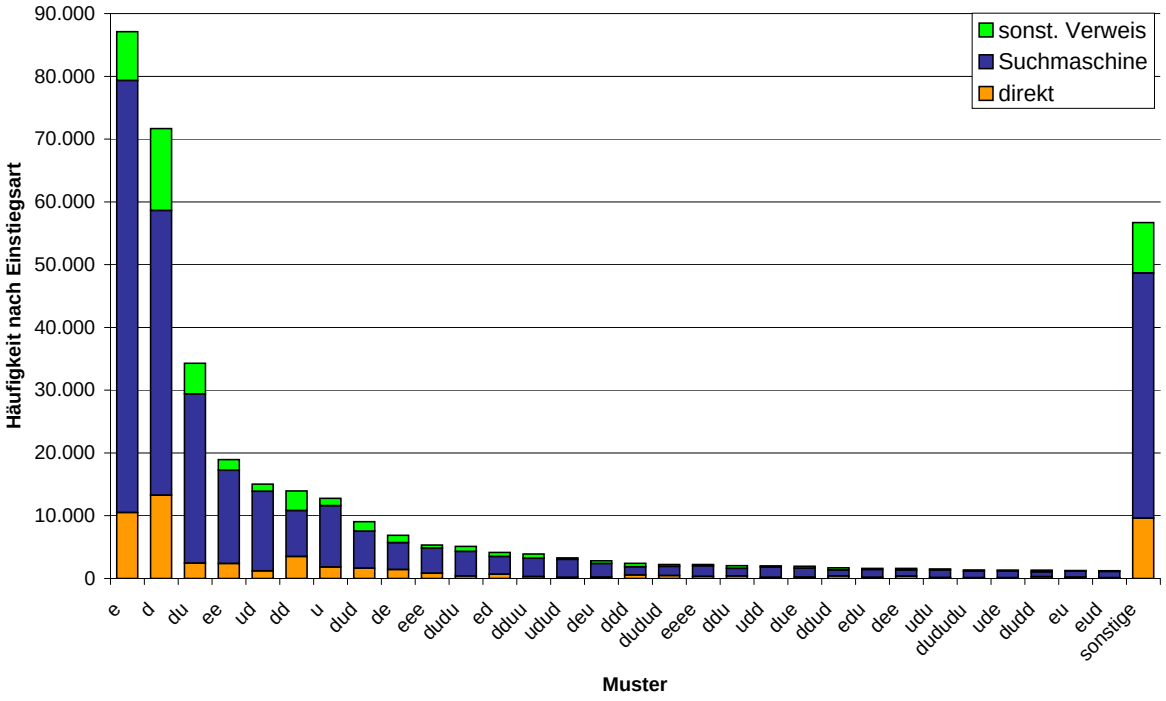


Abbildung 12.37.: Häufigkeiten der 30 häufigsten einfachen Muster nach Einstiegsarten

Abbildung 12.38.: DBS-Bildschirmfoto mit Hervorhebung der möglichen Hierarchiebewegungen

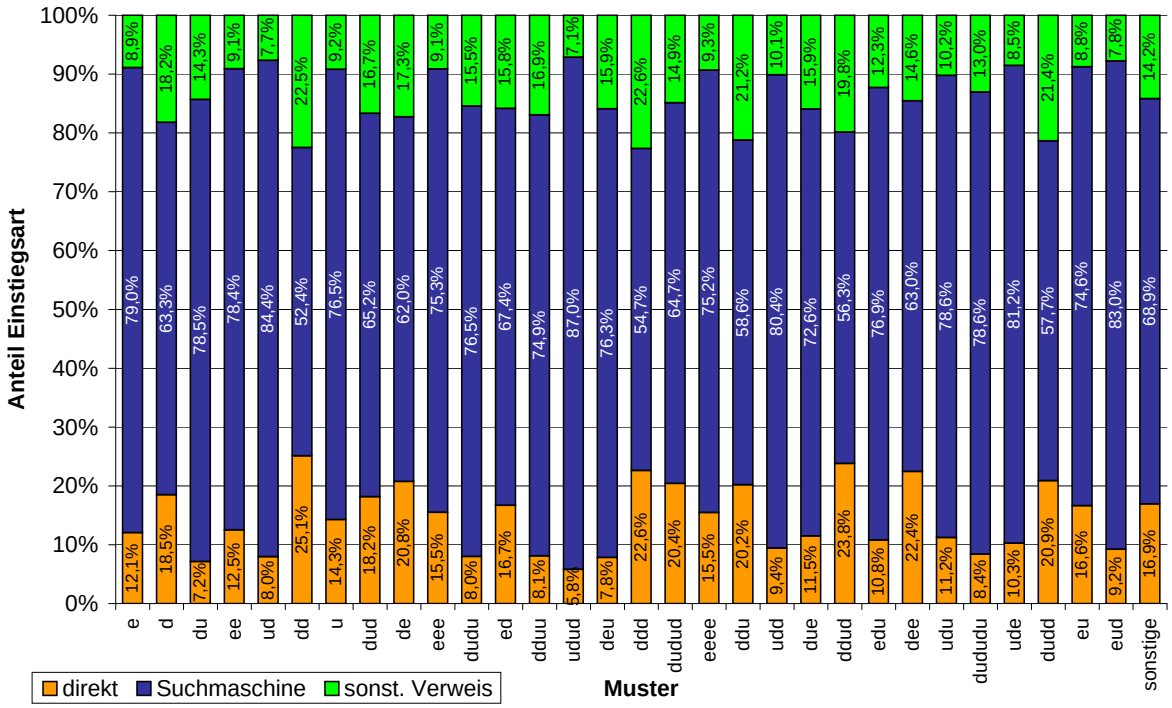


Abbildung 12.39.: Anteile der Einstiegsarten an den 30 häufigsten einfachen Mustern

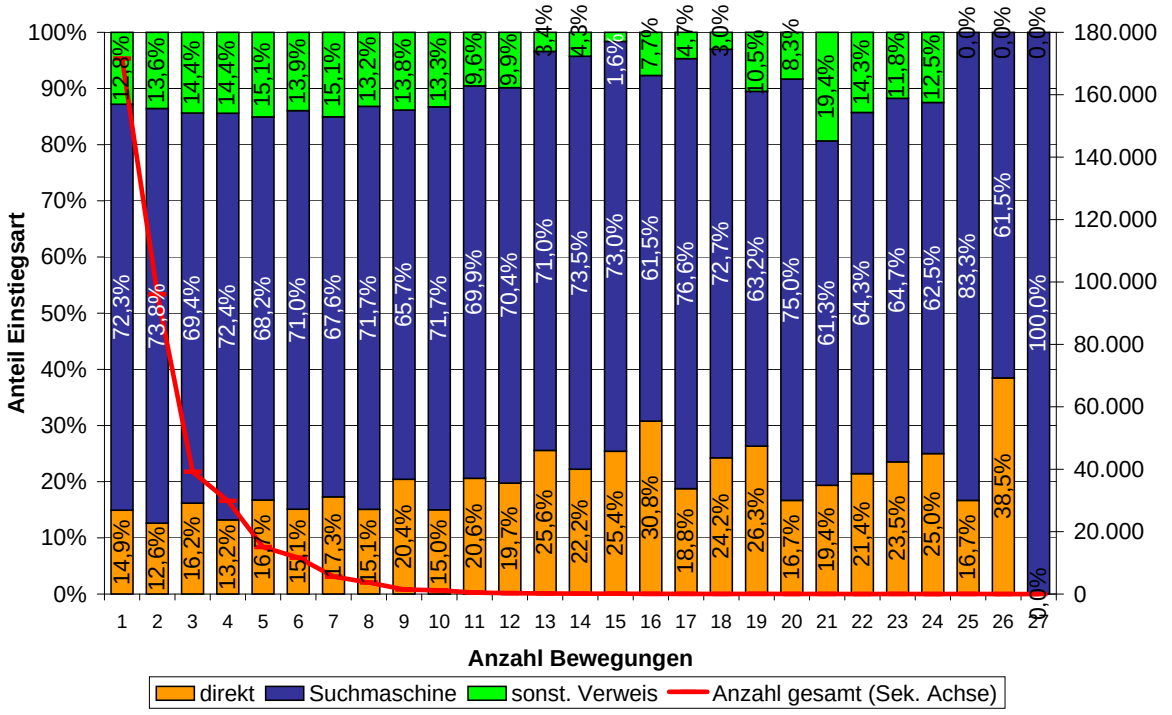


Abbildung 12.40.: Anteile der Einstiegsarten nach Komplexität der Muster

G.5. Bewegung zwischen Navigationsbereich und Suchfunktion

	direkt	Suchmaschine	sonst. Verweis	Gesamt
Thema zu Thema	135.344	578.771	102.630	816.745
Thema zu Zielgruppe	21.942	144.884	23.877	190.703
Zielgruppe zu Zielgruppe	68.658	300.925	70.373	439.956
Zielgruppe zu Thema	32.259	139.221	27.857	199.337
Suche zu Suche	20.633	35.610	11.180	67.423
Navigation zu Suche	13.463	50.617	9.048	73.128
Suche zu Navigation	14.673	39.467	9.465	63.605
Gesamt	306.972	1.289.495	254.430	1.850.897
DBS-Sitzungen gesamt	374.843	1.291.313	214.601	1.880.757

Tabelle 12.18.: Häufigkeiten der Bewegungen zwischen den Navigationsbereichen und der Suche

	direkt	Suchmaschine	sonst. Verweis	Gesamt
Kategoriewechsel innerhalb Zielgruppe oder Thema	36.690	158.302	32.786	227.778
Kategoriewechsel zwischen Zielgruppe und Themen	57.370	295.021	53.706	406.097
Bewegung innerhalb derselben Kategorie	173.408	749.079	144.330	1.066.817
Gesamt	267.468	1.202.402	230.822	1.700.692

Tabelle 12.19.: Häufigkeiten der Bewegungen innerhalb des Navigationsbereichs

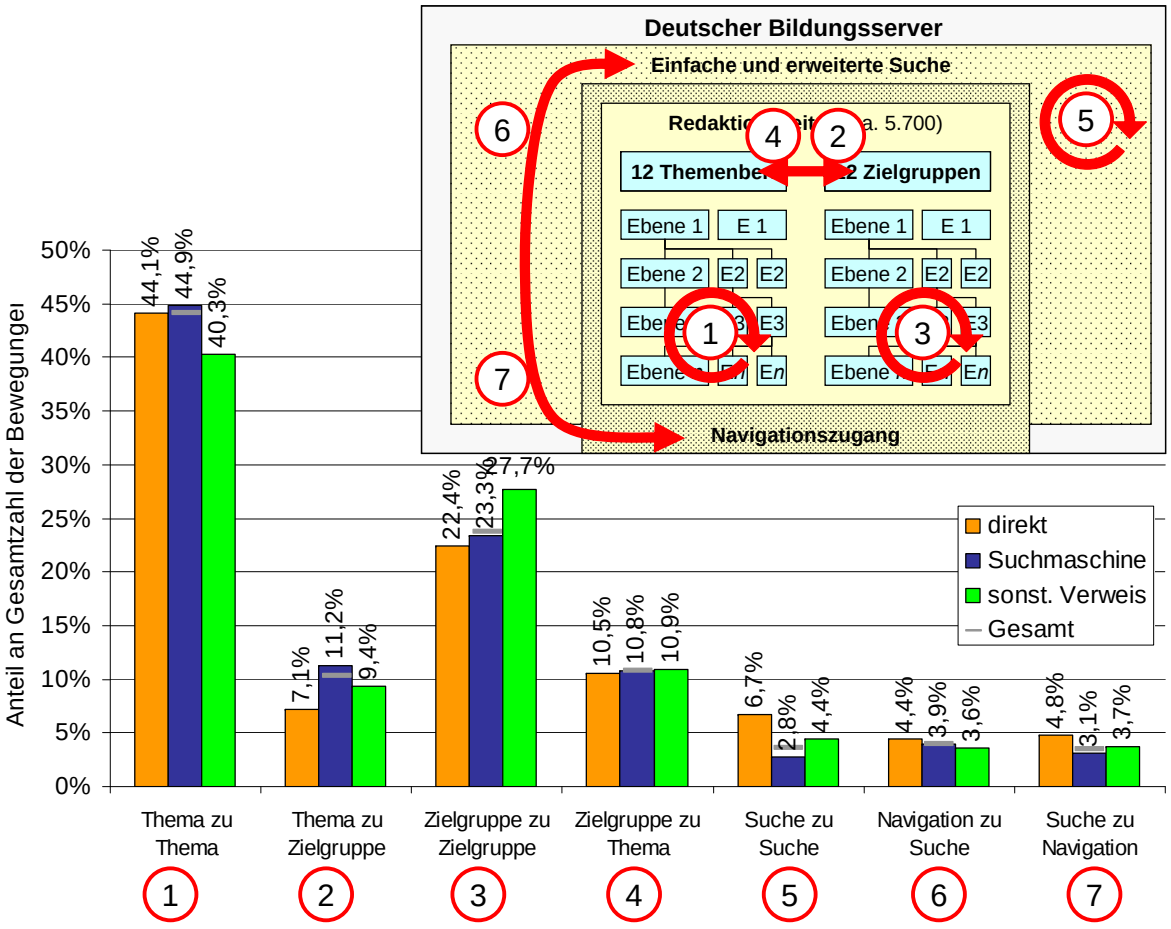


Abbildung 12.41.: Durchschnittlicher Anteil definierter Bewegungsarten zwischen Navigation und Suche an den Gesamtsitzungen

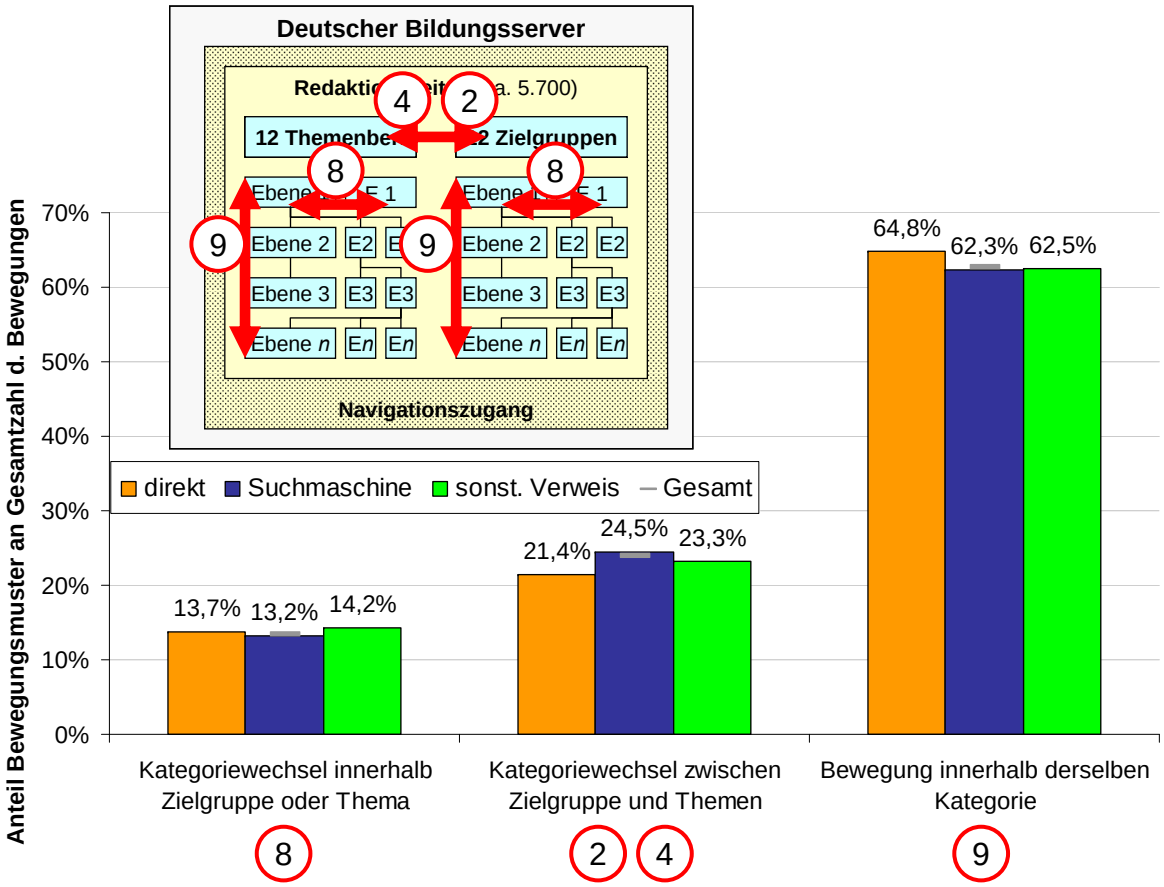


Abbildung 12.42.: Durchschnittlicher Anteil definierter Bewegungsarten innerhalb der Navigation an den Gesamtsitzungen