Nutzungsstatistiken elektronischer Publikationen

Erschienen in: Zeitschrift für Bibliothekswesen und Bibliographie. 2007, Heft 4-5, S. 234-237

[Abstract]

Einer der Vorteile des Open-Access-Publizierens wird in der erhöhten Sichtbarkeit der Dokumente gesehen. Damit verbunden ist die Annahme, die erhöhte Sichtbarkeit führe im Vergleich zu kostenpflichtig zugänglichen Dokumenten auch zu einer verstärkten Nutzung der Open-Access-Dokumente. Zugleich wird angenommen diese verstärkte Nutzung sei die Ursache der bei Open-Access-Dokumenten erhöhten Zitationshäufigkeiten. Generell fallen beim Zugriff auf elektronische Dokumente Nutzungsdaten an. Diese Nutzungsdaten können über Webserver- oder Linkresolver-Logs erhoben werden. Dieser Beitrag konzeptionalisiert Nutzungsstatistiken nicht allein als Prädikator für spätere Zitationshäufigkeiten oder –raten, sondern auch als Indikator, der die von Zitationen abweichenden und alternativen Auswirkungen einer Publikation ausdrückt. Werden Nutzungsdaten für statistische Auswertungen herangezogen, müssen typische Verzerrungen (wie z.B. automatisierte Zugriffe) beseitigt werden, um Interoperabilität zu erreichen. Im Idealfall können aus einem bereinigten Rohformat Nutzungsstatistiken nach verschiedenen Standards bzw. Konventionen (wie z.B. COUNTER, IFABC) erstellt werden. Die Aggregation und der Austausch der Daten verschiedener Server können über offene Schnittstellen - mit besonderem Augemerk auf die Deduplizierung von Dokumenten und Nutzern - erfolgen. Neben der Modellierung alternativer Impact-Maße sind mit erhobenen Nutzungsdaten auch Anwendungen wie etwa Recommender-Systeme, Austausch mit anderen Diensten oder als Unterstützung bei Portfolio-Entscheidungen möglich.

Wer publiziert (von lat. *publicus*: öffentlich), will gelesen und wahrgenommen werden. Dies ist im wissenschaftlichen Bereich ebenso der Fall wie bei Belletristik und Sachbuch. Open Access postuliert per se eine bessere Rezeption durch einfachere Zugänglichkeit. Quantitativ ist dies nur durch Statistiken belegbar. Bei konventionellen gedruckten Publikationen (Büchern, Zeitungen und Zeitschriften) sind keine genauen und direkten Nutzungswerte messbar, sondern es werden Hilfsmittel wie Verkaufs-, Ausleihzahlen oder Auflagenhöhen herangezogen, um daraus Nutzungszahlen abzuleiten. Im Bereich Hörfunk und Fernsehen wird die Nutzung in ähnlicher Weise indirekt durch Befragung repräsentativer Gruppen oder aus den Reichweiten der Sender und den potenziellen erreichbaren Hörer- und Seherzahlen im intendierten Sendegebiet interpoliert. Mit dem Online-Publizieren stehen nun neue, direktere und genauere Mess- und Auswertungsverfahren zur Verfügung. Hier können Ansätze der Web Analytics (auch Web Controlling, Web Analyse, Traffic Analyse, Clickstream Analyse) übernommen werden. Damit bezeichnet man die Sammlung und Auswertung des Verhaltens von Besuchern auf Websites. Was dabei als Nutzung zu bezeichnen ist, muss definiert werden[1]. Automatisch gemessen werden kann im elektronischen Umfeld immer nur der Zugriff. Zudem sind Publikationen nicht mit Webseiten gleichzusetzen. Wissenschaftliche Online-Publikationen stellen ein verteiltes, komplexes

Netzwerk dar, bei dem eine Publikation von verschiedenen Anbietern, in unterschiedlichen technischen Umgebungen, in unterschiedlichen Stufen des Publikationsprozesses und zu unterschiedlichen Bedingungen zur Verfügung gestellt werden kann. Daher ist neben der Definition, was als Nutzung zu betrachten ist, entscheidend, welche Teilbereiche des Publikationsnetzwerkes man in eine Datenerhebung einbezieht. Hier geht es unter anderem um die Frage, welche Publikationen (z.B. nur Zeitschriftenartikel, die im ISI Web of Science ausgewertet werden, nur begutachtete Publikationen etc.), welche Anbieter der Publikationen (Verlage, Aggregatoren, institutionelle oder fachspezifische Open-Access-Repositories) oder welche Publikationsstufen (vor, während oder nach der Begutachtung) berücksichtigt werden. In Zukunft sind mit der Ausweitung des Publikationsbegriffes sicherlich auch wissenschaftliche Primärdaten bzw. Lernobjekte in ihrer Nutzung zu berücksichtigen.

Webserver

Open-Access-Publikationen sind im Internet frei zugänglich und können nicht nur von beliebigen Nutzern gelesen, sondern in der Regel auch von Spidern indiziert werden. Die Indexierung durch sog. Webcrawler (auch als Spider bzw. Robots bezeichnet) ist Vorrausetzung für den Nachweis in (wissenschaftlichen) Suchmaschinen und daher im Sinne der erhöhten Sichtbarkeit der Open-Access-Publikationen durchaus erwünscht. Da maschinelle Zugriffe durch Spider keine Benutzung des Dokuments durch einen Leser ausdrücken, stellen sie eine für Webserver-Logs typische Verzerrung der Nutzungsdaten dar. Jeder Abruf einer Publikation – sei es durch einen Leser oder einen Spider – wird in den Logfiles des Webservers festgehalten. Dabei werden Informationen wie z.B. die IP-Adresse des aufrufenden Rechners, falls verwendet die Authentifizierung mit Benutzername und Passwort, der Zeitpunkt des Zugriffs, die aufgerufene Datei (die Publikation), der Statuscode des Webservers über den Aufruf, die Adresse der aufrufenden Seite (gegebenenfalls inklusive Suchstring, der zur aufgerufenen Publikation geführt hat) und Angaben über den eingesetzten Browser gesammelt. Wird der Zugriff auf Publikationen nicht über die Webserver-Logs erfasst, sondern z.B. über Vermittlerdienste wie Linkresolver, können einige der Verzerrungen entfallen.

Linkresolver

Ein Linkresolver ist ein im Bereich digitaler Bibliotheken genutztes System zur kontextabhängigen Anzeige von Diensten oder Publikationen. Es stellt einen zentralen Verknüpfungspunkt innerhalb der digitalen Bibliothek dar, indem es dem Nutzer die in einer bestimmten Situation sinnvollen und möglichen Dienste anbietet. Wenn sichergestellt ist, dass Dienste in ausreichendem Umfang verknüpft sind, so dass ein großer Teil ihrer

Nutzung über das Linkresolver-System erfolgt, so bietet sich hier ein alternativer Ansatz der Erhebung von Nutzungsdaten.[2] Da Linkresolver endnutzerorientierte (Meta-)Dienste sind, werden keine maschinellen Zugriffe erfasst und die Ausgangsqualität der Daten ist somit höher. Im Bereich wissenschaftlicher Publikationen sind jedoch z.B. fachliche oder institutionelle Repositories noch nicht ausreichend in Linkresolver-Systeme eingebunden, der erfasste Anteil des eingangs angesprochenen Publikationsnetzwerks ist mithin (noch) zu gering. Um derzeit größere Mengen an Nutzungsdaten von Open-Access-Publikationen zu erhalten, ist man daher auf die Auswertung von Webserver-Logs angewiesen.

Datenschutz

Unabhängig davon, wie Nutzungsdaten elektronischer Publikationen erhoben werden, setzen Gesetze der Speicherung und Sammlung dieser Daten Grenzen: § 13 Abs. 4 Nr. 2 des Telemediengesetzes (TMG) schreibt vor, dass so genannte Diensteanbieter (etwa Betreiber von Webservern) technisch und organisatorisch sicherstellen, dass anfallende personenbezogene Daten über den Ablauf des Zugriffs oder die sonstige Nutzung unmittelbar nach deren Beendigung gelöscht oder gesperrt werden. § 3a des Bundesdatenschutzgesetzes (BDSG) hält weiterhin fest: „Gestaltung und Auswahl von Datenverarbeitungssystemen haben sich an dem Ziel auszurichten, keine oder so wenig personenbezogene Daten wie möglich zu erheben, zu verarbeiten oder zu nutzen. Insbesondere ist von den Möglichkeiten der Anonymisierung und Pseudonymisierung Gebrauch zu machen, soweit dies möglich ist“. Demnach sind Betreiber von Open-Access-Angeboten oder anderer Services, die Nutzungsstatistiken etwa zum Zweck der Evaluierung oder Bestimmung alternativer Qualitätskriterien für wissenschaftliche Dokumente nutzen wollen, verpflichtet, anfallende Nutzungsdaten zu anonymisieren. Sollte die Aggregation der Nutzungsdaten verschiedener Server geplant sein, könnten die Nutzerinformationen auch pseudonymisiert und an zentraler Stelle gesammelt werden, solange sichergestellt ist, dass Personenbezug auch mit Hilfe von zusätzlichem Wissen nicht mehr hergestellt werden kann.[3]

Verzerrungen / Manipulation

Neben der Bereinigung der Nutzungsinformationen von datenschutzrelevanten Elementen ist es allerdings auch nötig, irrelevante Einträge aus den Logfiles zu entfernen. Vor allem Webserver-Logs unterliegen zahlreichen Verzerrungen, die z.B. durch Doppelklicks, maschinelle Zugriffe durch Webcrawler oder Überwachungstools, mehrfache Einträge etwa durch in HTML-Dokumente eingebundene Elemente wie Grafiken oder durch für die Webdarstellung optimierte PDF-Dateien zustande kommen. Andere Probleme ergeben sich aus der Summierung der Zugriffe auf Mehrdateiendokumente und der Gefahr, zu geringe

Nutzungszahlen zu ermitteln, wenn von verschiedenen Nutzern über einen Proxy unter *einer IP-Adresse* auf eine Publikation zugegriffen wird. Zudem ist der Umgang mit ungewöhnlichen Nutzungsphänomenen wie etwa softwaregesteuerten Massendownloads von Publikationen zum späteren Offline-Lesen oder dem künstlichen Erhöhen von Zugriffszahlen durch häufiges Anklicken oder Einsetzen spezieller Robots zu klären.

Es existieren zahlreiche Ansätze, solche Verzerrungen zu beseitigen: So kann das Proxy-Problem näherungsweise durch den Einsatz von Cookies gelöst werden. Zur Eliminierung von Webcrawlern kann eine Kombination verschiedener Strategien gewählt werden: Neben Abfragen von einschlägigen Verzeichnissen und Listen (z.B. der International Federation of Audit Bureaux of Circulations IFABC[4]) ist die Identifikation über den User Agent, über Zugriffe auf die robots.txt[5] oder automatisierte Abfragen über uwhois[6] möglich. Zusätzlich kann versucht werden, Spider und Massendownloads durch das Festlegen bestimmter Obergrenzen für Zugriffshäufigkeiten zu bestimmen: Wird von einer IP (oder von IPs aus einem bestimmten Bereich) innerhalb eines bestimmten Zeitraumes eine vordefinierte absolute oder prozentuale Zugriffshäufigkeit auf Dokumente überschritten, könnten diese aus den Logfiles eliminiert werden. Beide Grenzen (Zugriffshäufigkeit und Zeitraum) müssen allerdings transparent begründet und eindeutig definiert sein. Gleiches gilt für die Behandlung der Spiderzugriffe und Doppelklicks. Zumindest für den Umgang mit einigen dieser Phänomene existieren Modelle und mit COUNTER[7] sogar ein De-Facto-Standard für die Ermittlung von Zugriffen auf wissenschaftliche elektronische Publikationen.

Referenzen und Modelle

COUNTER findet im Umfeld kommerzieller Verlage, deren Dokumente vor freiem Zugriff geschützt sind, Anwendung und kennt daher manche Phänomene (z.B. Probleme durch Webcrawler) nicht, die beim Zugriff auf Open-Access-Publikationen auftreten. Für andere Verzerrungen definiert COUNTER Vorgehensweisen: So wird ein wiederholtes Laden einer PDF-Datei von derselben IP-Adresse aus innerhalb einer Zeitspanne von 30 Sekunden als Doppelklick gewertet und aus der Zugriffsstatistik entfernt, die Zeitspanne für HTML-Dateien liegt bei 10 Sekunden. Auch wenn COUNTER der gängige Standard zur Bestimmung der Zugriffe auf elektronische Dokumente ist, stellt sich die Frage, ob etwa ein wiederholtes Herunterladen eines PDF-Dokuments innerhalb von 30 Sekunden wirklich adäquat die Nutzung eines Dokuments beschreibt. LogEC, das Statistikmodul des Repository-Netzwerk RepEC, definiert Nutzung in anderen zeitlichen Dimensionen: Innerhalb eines Monats wird pro IP-Adresse nur ein Zugriff auf eine Datei berücksichtigt.[8] Einen wiederum anderen Rahmen wählt die VG Wort, die über das System METIS eine valide Grundlage zur Ausschüttung von Vergütungen an Autoren von Onlinedokumenten schaffen will, in dem sie - orientiert an den Vorgaben der deutschen Werbeindustrie und der Informationsgemeinschaft

zur Feststellung der Verbreitung von Werbeträgern e.V. (IVW) - die Zeitspanne auf 30 Minuten festlegt. Die Vielzahl der Vorgehensweise zeigt, dass Standardisierung und Offenheit der Formate nötig sind.

Normierung

Da derzeit nicht abzusehen ist, welche Standards sich in diesem Bereich mittelfristig durchsetzen werden, ist ein erster wichtiger Schritt, Nutzungsdaten aus unterschiedlichen Quellen über ein einheitliches Format syntaktisch austauschbar zu machen. Hier sind zum einen OpenURL ContextObjects[9] und zum anderen das SUSHI-Schema[10] zu nennen. Daten in dieser Form können von noch zu schaffenden Diensten über das OAI-Protokoll aggregiert und um Mehrfacheinträge bereinigt werden (Deduplizierung). Deduplizierung bezieht sich dabei zum einen auf die Publikation selbst (Referent). Hier ist sicherzustellen, dass nur Nutzungsdaten aggregiert werden, die sich mit hinreichender Sicherheit auf diese Ressource beziehen. Da noch keines der derzeit eingesetzten Systeme (DOI, URN) zur eindeutigen Identifizierung von Ressourcen umfassend genug ist, ist dies nur über metadatenbasierte Heuristiken zu erreichen. Zum anderen bezieht sich die Deduplizierung auf den Zugreifenden (Agent). Hier ist sicherzustellen, dass z.B. einheitliche Pseudonymisierungsalgorithmen verwendet werden, die es ermöglichen, später Strukturen in den erhobenen Daten erkennen und auswerten zu können. Basierend auf den so aggregierten und deduplizierten Rohdaten können dann unterschiedliche Zählweisen und Standards (wie COUNTER, LogEC oder IVW) bedient bzw. gefiltert werden. Erst diese Definition und der so entstandene Filter konstituieren dann Nutzung. Um die so gewonnen Nutzungszahlen für den Endnutzer transparent zu machen, ist es wichtig, die verwendete Konvention und ggf. Irrtumswahrscheinlichkeiten, die aufgrund der Datenerhebung oder -aufbereitung auftreten können, anzugeben.

Anwendungen und Dienste

Neben den standardisierten und transparent angebotenen Nutzungszahlen für einzelne Publikationen, die je nach Datenbasis bereits einen hohen Aussagewert haben, sind eine Reihe weiterer Anwendungen und Dienste denkbar, bzw. bereits prototypisch vorgestellt worden, die auf Nutzungsstatistiken elektronischer Publikationen basieren.

So können z.B. Rankings aufgrund von Nutzungszahlen berechnet werden. Hier sind verschiedene Ansätze basierend auf der Häufigkeit oder basierend auf strukturellen Zusammenhängen denkbar.[11] Einer der bekanntesten Algorithmen zur Berechnung von strukturbasierten Rankings ist der bei Google verwendete Page Rank. Neben der Bewertung von Publikationsleistungen im Umfeld von Forschungsevaluation[12] können auch Konsortien bzw. Bibliotheken diese Rankings als Entscheidungshilfe für den Aufbau ihres Portfolios an

elektronischen Ressourcen benutzen. Diese Rankings können eine wertvolle Ergänzung zu dem rein aus Häufigkeiten von Zitationen berechneten Impact Factor (IF) sein und sogar mit ihm kombiniert werden.[13]

Wertet man die Strukturmerkmale von Nutzungsdaten aus, kann darüber hinaus auch ein Recommender System erstellt werden, das aufgrund der aufeinander folgenden Nutzung zweier Ressourcen eine Verbindung herstellt und bei einer Suche zu einer gefundenen Ressource weitere Ressourcen vorschlägt.[14] Schließlich sind Modelle denkbar, in denen erhobene und aggregierte Nutzungsdaten von Dokumenten anderen Diensten oder Angeboten, etwa zur Anreicherung von Datenbanken oder zur Ergänzung von Verlagsstatistiken, bereitgestellt werden. In diesen Fällen wäre es sinnvoll, eine Policy zum Austausch der Daten zu definieren und deren Verwendung durch eine Creative Commons Lizenz zu regeln.

---

[1] Ansätze sind COUNTER (Counting Online Usage of Networked Electronic Resources), http://www.projectcounter.org und die nicht offen gelegten Verfahren der IVW (Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern), die im Rahmen von METIS (MEldesystem für Texte auf InternetSeiten) verwendet werden, http://www.vgwort.de/metis.php.

[2] Bollen, Johan, Van de Sompel, Herbert: An architecture for the aggregation and analysis of scholarly usage data. In: Joint Conference on Digital Libraries (JCDL2006), June 2006, S. 298-307. http://doi.acm.org/10.1145/1141753.1141821

[3] Auf die Verwendung der Nutzungsdaten zu diesem Zweck und die Pseudonymisierung ist in der Datenschutzerklärung des Webangebotes hinzuweisen. Ebenso auf die Möglichkeit des Nutzers, dieser Verwendung seiner Daten zu widersprechen (§ 15 Abs.3 TMG). Jede weitergehende Speicherung und Verarbeitung der Nutzungsdaten bedarf der Einwilligung (d.h. Erklärung zu Beginn der Online-Nutzung) des jeweiligen Nutzers unter Einhaltung der Vorschriften des § 13 Abs. 2 und 3 TMG.

[4] Homepage: http://www.ifabc.org/index.asp

[5] Robots exclusion standard vgl. http://www.robotstxt.org/wc/exclusion.html

[6] http://www.uwhois.com/, über uwhois-Abfragen kann festgestellt werden, auf wen eine IP-Adresse registriert ist.

[7] Vgl. COUNTER code of Practice for Journals and Databases, http://www.projectcounter.org/r2/COUNTER_COP_Release_2.pdf und COUNTER Code of Practice for Books and Reference Works, http://www.projectcounter.org/cop/books/cop_books_ref.pdf

[8] Dokumentiert unter: http://logec.repec.org/about.htm

[9] Vgl. Bollen, Van de Sompel, a.a.O. Tabelle 2

[10] SUSHI draft standard vom 20.9.2006. http://www.niso.org/standards/resources/Z39-93_DSFTU.pdf

[11] Bollen, Johan, Van de Sompel, Herbert, Smith, Joan, Luce, Rick: Toward alternative metrics of journal impact: a comparison of download and citation data. In: Information Processing and Management, Vol. 41, Issue 6, 2005, S. 1419–1440. http://dx.doi.org/10.1016/j.ipm.2005.03.024

[12] Vgl. die Aktivitäten des Instituts für Forschungsinformation und Qualitätssicherung (iFQ). http://www.forschungsinfo.de

[13] Ball, Philip: Prestige is factored into journal ratings. In: Nature online, 15 February 2006, http://www.nature.com/news/2006/060213/full/439770a.html

[14] Bollen, Johan, Nelson, Michael L., Geisler, Gary, Araujo, Raquel: Usage derived recommendations for a video digital library. In: Journal of Network and Computer Applications. Vol. 30, Issue 3, August 2007, S. 1059-1083. http://dx.doi.org/10.1016/j.jnca.2005.12.009