

Datenkonzentration bei Suchmaschinen – die Macht von Google & Co.

Zusammengefasster Beitrag für die Sommerakademie 2007 des Unabhängigen Landeszentrums für Datenschutz Schleswig-Holstein (ULD)

Jan Schallaböck, ULD, Referat 10

Einleitung

Für die Nutzung des Internets dienen Suchmaschinen wie Yahoo, Google oder AOLSearch¹ dem Nutzer zum Auffinden von Inhalten über Suchanfragen. Eine Suchanfrage ist dabei ein einzelner oder eine Aneinanderreihung von Suchbegriffen, auf die die jeweilige Suchmaschine eine Liste von Internetseiten mit ihrer jeweiligen Internetadresse zurückliefert. Je nach Nutzungsverhalten geben die Suchanfragen in vielfältiger Weise Aufschluss über die Gewohnheiten, Interessen und Tätigkeiten eines Nutzers.



User ID	Search Keywords	Date	Website
4417749	pine straw lilburn delivery	2006-03-06 18:35:02	
4417749	panasonic vacuum cleaners parts	2006-03-22 13:03:52	http://www.vacuum-depot.com
4417749	osteopenia	2006-03-20 19:47:03	http://www.medterms.com
4417749	osteopenia	2006-03-20 19:47:03	http://health.ivillage.com
4417749	oogle	2006-03-16 22:34:11	
4417749	nicotine effects on the body	2006-03-26 10:31:15	http://www.geocities.com
4417749	new zealand real estate	2006-03-09 21:41:48	http://www.realenz.co.nz
4417749	new york state history rapp road	2006-05-18 13:14:43	http://www.woodrow.org
4417749	new york state history rapp road	2006-05-18 13:14:43	http://www.nysm.nysed.gov

Abbildung 1: Screenshot AOLSearchDatabase

AOL veröffentlichte im August 2006 einen Auszug von Suchanfragen von über 500.000 Nutzern ihrer Suchmaschine zu Forschungszwecken². Dabei waren die Nutzernamen durch Zahlen ersetzt, die einen Rückschluss auf die dahinter stehenden Personen verhindern sollten. Einige der Suchanfragen wiesen allerdings eindeutige Hinweise auf ihren Verfasser auf, so

¹ Daneben existieren eine Vielzahl anderer Suchdienste im Internet, deren Behandlung aber den Rahmen der vorliegenden Studie sprengen würde.

² <http://www.heise.de/newsticker/meldung/76474/>.

dass in diesen Fällen die Rückführung möglich wurde. Diese Veröffentlichung verdeutlichte ein weiteres Mal die Sensibilität derartiger Datenbestände. AOL löschte nach massiver Kritik von Nutzern und Bürgerrechtsorganisationen den Datenbestand zwar wieder aus dem Netz³, er ist allerdings nach wie vor im Internet verfügbar.⁴

Beim Suchmaschinenbetreiber anfallende Daten

Bei einem Suchvorgang fallen beim Betreiber der Suchmaschine zumeist⁵ folgende Daten an:

<ul style="list-style-type: none">• Suchanfrage (ein oder mehrere Suchbegriff(e))• IP-Adresse des anfragenden Computers• Benutzter Browser und Betriebssystem• Die vom Nutzer ausgewählte Internetseite	<ul style="list-style-type: none">• Datum und Uhrzeit der Anfrage• Ggf. der Inhalt eines Cookies⁶• Weitere Angaben eines http-requests⁷• Ggf. weitere Daten
--	--

Tabelle 1: Anlässlich eines Suchvorgangs beim Suchmaschinenbetreiber anfallende Daten

Technische Realisierung einer Verkettung von Suchanfragen

Über die IP-Nummer

Die Verkettung der Suchanfragen untereinander, aber auch mit anderen Daten kann bei Suchmaschinen insbesondere über die IP-Nummer erfolgen, die den abfragenden Computer ausweist. Durch den Einsatz von Anonymisierungsproxies⁸ und dynamischen IP-Adressen⁹ weisen diese Ergebnisse jedoch für den Betreiber Verfälschungen und Ungenauigkeiten auf.

Über Cookies

Daher wird oftmals der Einsatz von Cookies⁵ vorgezogen, die einen Computer, bzw. einen Browser auch dann noch eindeutig identifizieren, wenn die IP-Adresse zwischenzeitlich geändert wurde. Cookies werden oft auch im Interesse des Nutzers verwendet, um ihn beispielsweise für die Nutzung personalisierter Dienste⁶ wieder zu erkennen und so den Aufwand einer erneuten Anmeldung zu ersparen.

Umgang mit dem Datenbestand

Der Umgang der Suchmaschinenbetreiber mit den von ihnen gesammelten ist uneinheitlich. So speichert AOL den Datenbestand nur für einen Zeitraum dreißig Tagen⁷, während Google derzeit unbefristet speichert.⁸

³ <http://www.eff.org/Privacy/AOL/spreadtheword.php/>.

⁴ <http://www.aolsearchdatabase.com/>.

⁵ Zur Funktionsweise vgl. <http://de.wikipedia.org/wiki/HTTP-Cookie/>.

⁶ Schon die Suchanfrage selber kann einen solchen personalisierten Dienst darstellen.

⁷ So zumindest <http://www.eff.org/deeplinks/archives/005162.php>. In Deutschland werden nach <http://www-w.aol.de/Portalkontakt-Datenschutz/lange-speichert-AOL-meine-Daten-1255476409-7.html/> Nutzungsdaten gelöscht oder anonymisiert, sofern sie nicht für Abrechnungszwecke benötigt werden.

⁸ Google hat kürzlich in einer Presseerklärung mitgeteilt, dass sie nach einer Intervention des Norwegischen Datenschutzbeauftragten von dieser Praxis abzuweisen beabsichtigen und nach einem Zeitraum von 18 - 24 Monaten die Daten nur noch in anonymisierter Form speichern wollen (<http://www.eff.org/deeplinks/archives/005162.php/>).

Allen Anbietern dienen die gespeicherten Anfragen zunächst zur Optimierung der Leistung ihrer Suchmaschine, gleichzeitig werden sie aber oftmals auch eingesetzt um zielgerichtete Werbung zu platzieren. Auch eine Weitergabe der Daten an Dritte zu Marktforschungszwecken ist nicht unüblich.⁹

Yahoo verwendet die entstandenen Daten ebenfalls zu Werbe- und Marktforschungszwecken, sowie zur Verbesserung ihres Dienstleistungsangebotes. In der Datenschutzerklärung wird jedoch darauf hingewiesen, dass die Verwendung dadurch ausgeschlossen werden kann, dass im Browser die Nutzung von Cookies deaktiviert wird.¹⁰

Verkettbarkeit bei der Nutzung von Diensten der Google Inc.

Dienste der Google Inc.

Die Firma Google Inc. ist der größte Suchmaschinenbetreiber der Welt. Die Zugriffe werden wie oben dargestellt zu Marketingzwecken verkettet. Zu Google Inc. gehören neben der Suchmaschine eine Vielzahl weiterer Dienste, so z.B.:

• Google-Suche	jedes Suchwort, das ein Nutzer bei Google eingibt;
• Goolge-Desktop	ein Index der Computerdateien eines Nutzers, der E-Mails, der Musik, Fotos und der Chat- und Webbrowser-Historie;
• Google-Talk	Instantmessage-Chats zwischen Nutzern;
• Google-Maps	erfragte Adressanfragen, oft inklusive der Wohnadresse des Nutzers, um Richtungen anzuzeigen;
• Google-Mail (Gmail)	E-Mailhistorie eines Nutzers mit der Voreinstellung, E-Mails "für immer" zu speichern;
• Google-Calendar	Terminplan eines Nutzers, wie er vom Nutzer eingegeben wird;
• Google-Orkut	Tool für Social-networking, das persönliche Informationen speichert wie Name, Wohnort, Status der Beziehung usw.;
• Google-Reader	ATOM/RSS-Feeds, die ein Nutzer liest;
• Google Video/YouTube	Videos, die sich ein Nutzer anschaut;
• Google-Checkout	Kreditkarten- oder Zahlungsinformationen zur Benutzung auf anderen Webseiten.

Tabelle 2: Google Dienste und Datenbestände (Auszug)

Im März 2007 hat Google Inc. das Unternehmen DoubleClick Inc. erworben. DoubleClick Inc. vermarktet Werbung durch zielgerichtete Schaltung von Anzeigen im Internet (targeted advertising).

⁹ Vgl. etwa http://about.aol.com/aolnetwork/mem_commitments/, in der deutschsprachigen Datenschutzerklärung wird jedoch darauf hingewiesen, dass die Daten nur in anonymisierter Form oder mit ausdrücklicher Einwilligung weitergegeben wird - <http://www.aol.de/Portalkontakt-Datenschutz/Welche-Daten-erhebt-AOL-AOL-Kunden-Zweck-1255476409-3.html/>.

¹⁰ Siehe <http://privacy.yahoo.com/privacy/de/index.html/>. Ob diese Lösung rechtlichen Anforderungen gerecht wird, sofern nur auf diesem Wege dem Kopplungsverbot gem § 12 Abs. 3 TMG entgegen getreten werden soll, ist allerdings zweifelhaft.

Anhand des Beispiels der Möglichkeiten von Google Inc. soll nachfolgend dargestellt werden, welche Möglichkeiten zur Verknüpfung von digitalen Identitäten existieren.

Voranzustellen ist dabei zunächst, dass zwischen der technisch möglichen Verknüpfung und der rechtlich zulässigen Verknüpfung zu unterscheiden ist. In dem nachfolgenden Szenario wird auf die Verkettungsmöglichkeiten im Bereich der Datenbestände und Systeme der Google Inc. eingegangen. Die Frage der rechtlichen Zulässigkeit oder der tatsächlichen Praxis wurde dabei bewusst außer Acht gelassen, um technische Potentiale in den Vordergrund der Betrachtung zu rücken.

Verkettung von Datenbeständen bei der Fa. DoubleClick

DoubleClick ist Hersteller einer Software, die zur Vermarktung von Werbung Internet eingesetzt wird (DART-Produktreihe). Sie unterhält Kooperationsverträge mit Werbetreibenden und Werbevertriebsfirmen. Auf den Seiten, die Werbung unter Nutzung der Software der Firma Doubleclick einblenden, wird ein Cookie der DoubleClick bezogen und auf dem aufrufenden Computer gespeichert. Hierdurch erhält DoubleClick die Information unter welcher IP-Adresse eine bestimmte Internetseite zu welcher Zeit aufgerufen wird. Durch Verkettung dieser Adressen kann so ein umfangreiches Surf- und somit Interessensprofil erstellt werden. Durch den Cookie ist es möglich den Nutzer verschiedener Internet wieder zu erkennen.

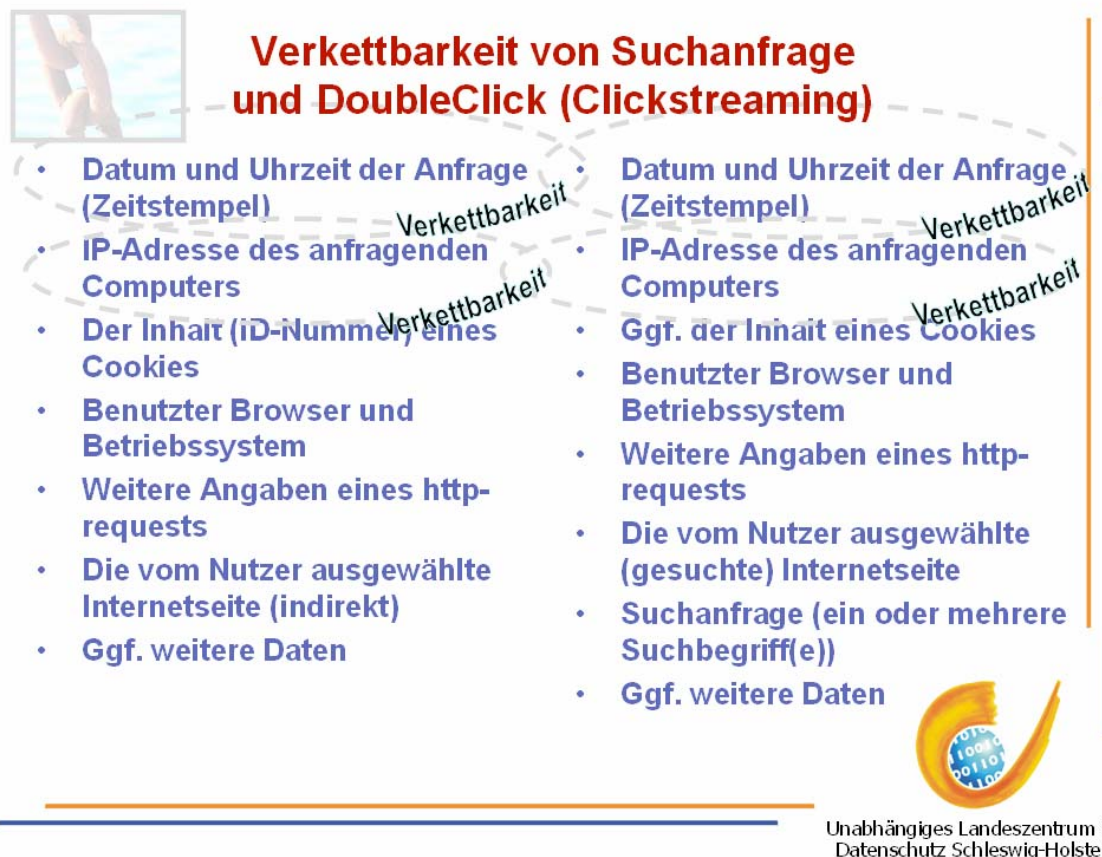


Abbildung 2: Verkettung über Zeitstempel und IP-Adresse

Dieses Interessensprofil lässt sich weiter verdichten indem es mit den bei Google verfügbaren Suchanfragen über die Vergleichsoperatoren IP-Adresse und Zeitraum verknüpft wird.

Der Umfang dieses Profils hängt dabei maßgeblich davon ab, wie stark der Nutzer Seiten frequentiert, die mit DoubleClick kooperieren, und in wie weit er google.com als Suchmaschine nutzt.

Derzeit unterhält DoubleClick Kooperationsverträge zu einer Reihe von Werbevermarktern in Deutschland und weltweit, so z.B. die TOMORROW FOCUS AG¹¹, die eine Reichweite von 32,6% der Internetnutzer hat¹². Zu denen von Ihnen mit Werbung beschalteten Seiten zählen unter anderem die Seiten: Amica Online, Focus Online, Playboy.de und Map24.com. Daneben weisen die über den Vermarkter AdLINK Internet Media GmbH mit Werbung beschalteten Seiten einen Double-Click-Cookie auf¹³. Über die Adlink Internet Media GmbH werden eine Vielzahl der Internetauftritte von Tageszeitungen (so z.B. Berliner Zeitung, WAZ, Hamburger Morgenpost, u.a.¹⁴ mit Werbung beschickt.

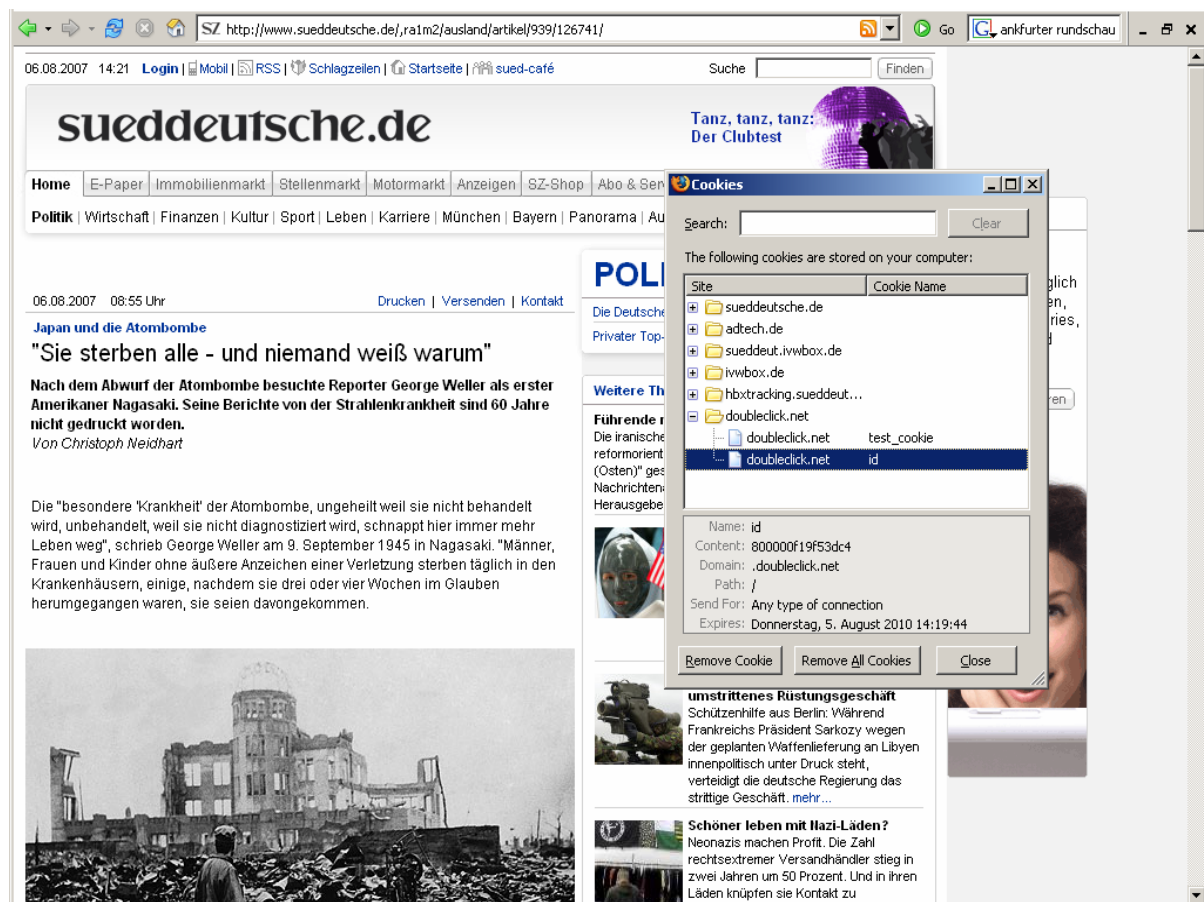


Abbildung 3: DoubleClick-Cookie bei Sueddeutsche.de

Beispielszenario

Möglichkeiten und Grenzen der Verkettung der Datenbestände bei den von Google Inc. angebotenen Diensten sollen nachfolgend anhand zweier fiktiver Szenarien dargestellt werden. Ausdrücklich wird erneut darauf hingewiesen, dass nicht behauptet werden soll, dass

¹¹ Vgl. http://www.slazenger.de/pb/detail/DoubleClick_und_die_TOMORROW_FOCUS_AG_weiten_Zusammenarbeit_aus.html.

¹² AGOF e.V., Internet Facts 2006-IV.

¹³ Stichprobentest des Verfassers.

¹⁴ AGOF e.V., aaO.

die Google Inc. ein derartiges Profil erstellt; das Beispiel soll lediglich die Aktionsmöglichkeiten erläutern.

Nutzung von Google Inc.-Diensten durch Bob

Bob, 31, Informatiker. Bob ist kritischer Computernutzer. Er nutzt Anonymisierungsdienste beim Surfen und lässt sich, um Profilbildung zu vermeiden, regelmäßig eine neue IP-Adresse zuweisen. Er schließt damit aus, dass eine Verkettung seines Nutzungsverhaltens über längere Zeiträume möglich ist. Cookies hat er deaktiviert, die Speicherung gestattet er nur in solchen Ausnahmefällen, in denen er sich davon überzeugt hat, dass dies in seinem Interesse liegt. In Einzelfällen sieht Bob davon ab, das Internet zu benutzen, weil er nicht überblicken kann, welche Daten dabei entstehen.

Bei Bob ergeben sich ohne weiteres keine Möglichkeiten zur Bildung umfangreicher Profile und Verkettungen. Abfolgen von Suchanfragen, bei denen er keine zwischenzeitliche Erneuerung der IP-Adresse durch den Anonymisierungsdienst vornehmen lässt, lassen sich allerdings dennoch zusammen führen. Sein Surfverhalten kann auch nicht über Cookies verkettet werden. Allerdings kann er eine Verkettung über sog. WebBeacons mit seinem Verhalten nicht ausschließen.

Bei WebBeacons werden pixelgroße – und damit für den Nutzer nicht auffällige - Grafiken in Webseiten von deren Betreiber integriert. Diese Grafiken stammen jedoch nicht vom Betreiber selbst, sondern von Dritter Seite und werden auch von dort herunter geladen, so dass dort jedes Mal, wenn ein Internetangebot eines Partners aufgerufen wird, auch eine Datenspur in Form der IP-Adresse des Nutzers entsteht. Über diese IP-Adresse lässt sich wiederum das Verhalten des Nutzers über die verschiedenen Angebote verfolgen.

Nutzung von Google Inc.-Diensten durch Alice

Alice, 26, ist Studentin und arbeitet an einer Masterarbeit zur Entwicklungszusammenarbeit. Abends liest sie oft ihre private E-Mail und chattet über gmail, einem Dienst der Google Inc. Dass dabei Werbung korrespondierend zum Inhalt ihrer E-Mail geschaltet wird empfindet sie als praktisch. Auch die Suchmaschine nutzt sie intensiv. Webseiten, die sie aufsucht, findet sie oft leichter über Suchbegriffe bei google.com als sich die Domainnamen zu merken. Hierdurch entstehen auch Spuren über die von ihr aufgesuchten Internetangebote. Nachfolgend dargestellt ist eine Sammlung von Einträgen, die sich in den Logdateien der jeweiligen Dienste befinden könnte:

Uhrzeit	Eintrag
18.30.10	gmail.com: Anmeldung unter dem Benutzernamen Alice1981
18.31.23	gmail.com: Aufruf diverser mails
19.15.11	google.de, Eingabe der Suchanfrage: entwicklungshilfeorganisationen jobs
19.16.33	google.de, Eingabe der Suchanfrage: auswärtiges amt jobs
19.16.38	google.de: Anwahl des ersten Suchtreffers in der Liste: Auswärtiges Amt - Tätigkeit bei Internationalen Organisationen
anschließend	Surfen auf den Seiten des Auswärtigen Amtes
19.22.38	Erneute Anmeldung bei gmail.com
19.23.12	gmail.com: Mail verfassen angeklickt
19.27.49	Besuch der Internetseite travelchannel.de

19.29.22	travelchannel.de: Anwahl des Angebotes für Pauschalreisen unter 300 Euro
19.31.12	travelchannel.de: Anwahl des Angebotes für Luxusreisen
19.32.30	travelchannel.de: Anwahl eines Angebotes für die französische Mittelmeerküste
19.33.21	travelchannel.de: Erneutes Laden des Angebotes für Luxusreisen
19.34.12	travelchannel.de: Anwahl des Angebotes für Reisen nach Afrika
19.35.43	travelchannel.de: Anwahl des Angebotes für Tansania/Sansibar
19.36.12	travelchannel.de: Anwahl des Angebotes für den Breezes Beach Club von Meiers Weltreisen
19.41.12	google.de: Suchanfrage: Sansibar virtualtourist
anschließend	Surfen auf den Seiten von virtualtourist.com
20.12.14	gmail.com: mail verschicken an kip@auswaertiges-amt.de mit folgendem Betreff: Anmeldung Karrieremesse und dem Text: "Sehr geehrter Herr..."
20.13.01	amica.de: Anwahl der Seite: Seite Kuss-Check
20.41.02	Erneute Anmeldung bei gmail
20.43.12	gmail.de: chat mit Nutzerin Alicia1980 mit folgendem Inhalt: "Hey Alicia..."
21.33.12	Abmeldung bei gmail

Tabelle3: Vereinfachtes Logfile Alice

Oben dargestellt ist eine Sammlung von Einträgen, die sich – in technisch standardisierter Form (vorliegend wurden sie zum Zwecke der Lesbarkeit in Stichpunkte übersetzt und vereinfacht) - in den Logdateien der jeweiligen Dienste befinden könnten. Sämtliche der oben bezeichneten Suchanfragen, Seitenaufrufe oder anderen Aktionen lassen sich untereinander und zu einem einzigen Profil verketten. Die dazu erforderlichen Daten liegen alle zumindest während der Erbringung des Dienstes in einem System vor, über das Google Inc. aufgrund seiner Beteiligungen die Kontrolle hat. Die Webseiten travelchannel.de und Amica.de nutzen DoubleClick und hinterlassen ein entsprechendes Cookie. Eine Verkettung wäre jeweils über Zeitstempel und IP-Adresse oder über die Verknüpfung von Cookies möglich. Lediglich das Surfverhalten auf den Seiten des Auswärtigen Amtes und bei virtualtourist.com ist nicht offensichtlich für Google Inc. nachverfolgbar.

Die Auswertungsmöglichkeiten der oben dargestellten Profile sind vielfältig. So stellen Hu et al. in einem kürzlich veröffentlichten Papier dar, wie das Geschlecht und Altersgruppe eines Nutzers aus dem Surfverhalten abgeleitet werden kann¹⁵.

Typische Auswertungsverfahren umfassen vor allen Dingen die Aggregation der Datenbestände und zielen nicht nur auf Profilbildung ab. In der Kombination mit den im Vorangegangenen dargestellten Auswertungsmöglichkeiten der individuellen Daten z. B. zum demographischen Hintergrund eines Nutzers lässt sich Werbung zielgruppenorientiert präsentieren. Während zielgruppengerechte Schaltung von Werbung kein Novum darstellt, sind die Möglichkeiten der Erfassung der Verhaltensweisen, die dann - unter Umständen - letztlich bis zu einer Kauf eines Produktes führen, in der nunmehr möglichen Dichte bemerkenswert.

¹⁵ Hu, Jian/Zeng, Hua-Jun/Li, Hua/Niu, Cheng/Chen, Zheng: Demographic Prediction Based on User's BrowsingBehavior; abrufbar unter: <http://www.cs.bell-labs.com/cm/cs/who/pfips/temp/web/www2007.org/papers/paper686.pdf> (Juli 2007).

So können Verfahren entwickelt werden, die Aussagen treffen, wie hoch die statistische Wahrscheinlichkeit, dass der Nutzer einem Werbelink folgt, und die abgeleitet von den vorhandenen individuellen Nutzercharakteristika sind. Die Genauigkeit statistischer Verfahren bestimmt sich maßgeblich aus dem sog. Hebesatz, also der Anzahl der erfassten Personen und Entscheidungen. Es kann mittels der vorgestellten Möglichkeiten sehr viel genauer prognostiziert werden, wie spezifische Verhaltensweisen bei bestimmten Einflüssen gesteuert werden, da der Prognose nahezu eine Vollerhebung des Nutzerverhaltens zugrunde liegt.

Rechtliche Betrachtungen

Eine detaillierte rechtliche Prüfung der oben dargestellten Szenarien würde den Rahmen dieser Abhandlung sprengen. Eine solche an dieser Stelle vorzunehmen ist auch nicht sinnvoll, da mit Beispielen gearbeitet wurde, die grade nicht beanspruchen, die tatsächliche Praxis darzustellen, sondern lediglich Möglichkeiten erläutern sollten.

Auffällig ist jedoch aus rechtlicher Sicht, dass auch ohne Vorliegen eines Personenbezuges im Sinne des BDSG, das Schutzziel des Datenschutzrechtes betroffen ist. Selbst wenn Alice im Einzelfall nicht individualisierbar ist, d.h. Google Inc. (oder auch Dritte) aus dem Profil nicht die natürliche Person ableiten können, droht doch die Gefahr eines erheblichen Informationsungleichgewichtes zwischen den Parteien. Alice kann nicht überblicken, was die Gegenseite über sie weiß und dieser Informationsvorsprung kann auch zur Verhaltensbeeinflussung genutzt werden.

Eine Verhaltensweise wie sie Bob an den Tag legt, der die Nutzung bestimmter Dienste vermeidet, ist daher nachvollziehbar. Ob sich derartige Einflussnahmen in relevanter Zahl tatsächlich darstellen oder darzustellen drohen, bedürfte allerdings einer empirischen Untersuchung. Festgestellt werden kann jedoch, dass eine solche Verhaltensanpassung zu verhindern erklärtes Ziel des Datenschutzes ist. Das Bundesverfassungsgericht legte in seinem Volkszählungsurteil dar, dass "wer unsicher ist, ob abweichende Verhaltensweisen jederzeit notiert und als Information dauerhaft gespeichert, verwendet oder weitergegeben werden, wird versuchen, nicht durch solche Verhaltensweisen aufzufallen. (...) Dies würde nicht nur die individuellen Entfaltungschancen des Einzelnen beeinträchtigen, sondern auch das Gemeinwohl, weil Selbstbestimmung eine elementare Funktionsbedingung eines auf Handlungsfähigkeit und Mitwirkungsfähigkeit seiner Bürger begründeten freiheitlichen demokratischen Gemeinwesens ist."¹⁶

¹⁶ BVerfGE 65,1, 42.