

Diplomarbeit

Interessenprofile in virtuellen Identitäten



Sebastian Kurt, <kurt@inf.fu-berlin.de>,
Freie Universität Berlin
Fachbereich Informatik
Arbeitsgruppe Netzbasierte Informationssystem

Zusammenfassung

Mit dieser Arbeit wird das Auffinden, Extrahieren und Auswerten von Interessen in virtuellen Profilen betrachtet. Die Nutzer von Internetdiensten übergeben durch unterschiedliche Aktionen Daten an Dienste. Diese bilden ein virtuelles Abbild der Vorlieben einer Person. Der Vergleich zu den per Fragebogen ermittelten realen Interessenschwerpunkten zeigt inwieweit Übereinstimmung zwischen den Welten herrscht. Mit dem virtuellen und realen Interessenprofilabdruck besteht die Möglichkeit grob zu charakterisieren.

Die Arbeit wurde mit Blick auf die von Tobias Bielohlawek (DFKI) geschriebene Diplomarbeit „Mediated Identity“ und der zugehörigen Software *WhoAmI* verfasst. Die Software *ThatsMe* welche im folgenden beschrieben und implementiert wird, bietet eine Antwortmöglichkeit auf die Frage „Wer bin Ich virtuell?“.

Inhaltsverzeichnis

1. Einführung	8
1.1. Anwendungsszenarien	9
2. Problem	10
3. Aktuelle Situation	12
3.1. Öffentliche Wahrnehmung rund um virtuelle Identität	12
3.1.1. Jahressteuergesetz mit Bürgeridentifikationsnummer	13
3.1.2. Terrorbekämpfung vs. Überwachungsgesellschaft	14
3.1.3. Vorratsdatenspeicherung	15
3.1.4. Gläserner Bürger	15
3.1.5. Identitätsdiebstahl	16
3.1.6. Online-Reputation	16
3.2. Dienste rund um virtuelle Identitäten	17
3.2.1. Identitätsmanagement	17
3.2.2. Dienste die „Identität“ in den Vordergrund stellen	17
3.2.3. Suchmaschinen	18
3.2.4. Personensuchmaschinen	19
3.2.5. Datenschutz	20
3.2.6. Visualisierung von Profildaten	20
3.2.7. Werbeindustrie	21
3.3. Zahlenbasis	21
3.3.1. Marktsegmentierung	23
3.3.2. Nutzerzahlen	23
4. Modelltheoretischer Ansatz	25
4.1. Definitionen und Zahlen	25
4.1.1. Definitionen	25
4.1.2. Zahlen	31
4.2. Online-Identitäten	33
4.2.1. Identität	33
4.2.2. Identitätsmanagement	33
4.2.3. Online-Identitäten und Beweggründe dafür	34
4.2.4. Daten aus Online-Identitäten	34
4.3. Modell eines Interessenprofil	35
4.3.1. Entstehung des allgemeinen Interessenprofils	35
4.3.2. Die Klassifikationen der Lexika	35
4.3.3. Begründung der Verwendung von Lexika	36
4.4. Segmentierungsmodelle	36

4.4.1.	Sinus-Milieus	37
4.4.2.	Semiometrie-Ansatz	38
4.4.3.	@facts Online-Nutzertypen 2007	38
4.4.4.	Modell des Allgemeinen Interessen-Struktur-Test	39
4.4.5.	Ansätze zur Einteilung aus der Werbeindustrie	39
4.4.6.	Daten die Nutzer unterteilen	40
4.5.	Zusammenhang zwischen Interessen und Nutzergruppen	41
4.5.1.	Interessen von Nutzern mit zugehörigen Nutzergruppen	41
4.5.2.	Nutzergruppen mit zugehörigen Interessen von Nutzern	42
5.	Erstellung eines Interessenprofils	43
5.1.	Allgemeine Quellen im Internet	43
5.1.1.	Auswahlkriterien für Quellen	43
5.2.	Verwendete Quellen	44
5.2.1.	Zugang und Identifizierung	45
5.2.2.	Enthaltene Inhalte	45
5.3.	Einordnung und Bewertung der Quellen	48
5.3.1.	Eigenschaften und Gewicht von Quellen	48
5.4.	Tags	49
5.4.1.	Beschränkung auf Tags	51
5.4.2.	Dimensionen des Tagging	51
5.4.3.	Freiheiten bei der Verwendung von Tags	54
5.5.	Das eigentliche Verfahren	55
5.5.1.	Gewicht eines Dienstes	57
5.5.2.	Gewicht eines Tags	57
5.5.3.	Gewicht einer Kategorie	57
5.6.	Lexika	57
5.6.1.	Wikipedia	58
5.6.2.	Freebase	58
5.7.	Weitere Verfahren zur Erkennung von Bedeutungen	58
5.7.1.	Collaborative filtering	59
5.7.2.	Crowdsourcing	59
6.	Die Implementierung von <i>ThatsMe</i>	60
6.1.	Softwaredesign	60
6.1.1.	Zielbestimmung	60
6.1.2.	Produkteinsatz	61
6.1.3.	Qualitätsanforderungen	61
6.1.4.	Benutzungsoberfläche	61
6.1.5.	Technische Produktumgebung	62
6.1.6.	Spezielle Anforderungen an die Entwicklungs-Umgebung	62
6.2.	Umsetzung der Kriterien des Verfahrens in der Implementierung	62
6.2.1.	Datenbankmodell	62
7.	Gefundene Ergebnisse	65
7.1.	Testreihen	65

7.1.1.	Daten der Anfrage	65
7.1.2.	Anfrageziele	67
7.1.3.	Antwortverwertungen	67
7.2.	Beispielhafter Ablauf und erhaltenes Profil	68
7.2.1.	Pfade zwischen Tags und Kategorien	69
7.2.2.	Beispiele bemerkenswerter Zuordnungen	70
8.	Vergleich der Ergebnisse mit realen Interessenprofilen	72
8.1.	Auswertung der Fragebogen	72
8.1.1.	Allgemeine Fragen zum Thema	72
8.1.2.	Fragen mit direktem Bezug zur Analyse	74
8.1.3.	Auswahl von Testpersonen	75
8.2.	Vergleich der realen mit virtuellen Interessenschwerpunkten	77
8.2.1.	Ergebnisse des Vergleiches der sieben Hauptkategorien	77
8.2.2.	Ergebnisse des Vergleiches der 15 ausgewählten Begriffe	81
8.2.3.	Ergebnisse des Vergleiches der frei genannten Interessen	82
8.2.4.	Bemerkungen zu ausgefilterten Kategorien	84
8.3.	Probleme mit Lösungsvorschlägen	84
8.3.1.	Vorteile von vielen Nutzern	84
8.3.2.	Sprache, Schreibweise und Bedeutung	84
8.3.3.	Antwortkategorien	86
8.3.4.	Gewicht von Diensten	86
9.	Zusammenfassung und Ausblick	88
9.1.	Zusammenfassung	88
9.1.1.	Details in „LongTail“ und der Einfluss des „Power law“	88
9.2.	Ausblick	88
9.2.1.	Grundsätzliche Weiterentwicklungen	88
9.2.2.	Dynamische Bewertung und Rangfolge der Koeffizienten	89
9.2.3.	Natural Language Processing (NLP)	89
9.2.4.	Konzept des Semantic Web	92
9.2.5.	Weiterverwendung der Software	92
A.	Anhang mit Bildern und Tabellen	93
	Literaturverzeichnis	103
	Abbildungsverzeichnis	104

Verwendete Abkürzungen

Beziehungsweise (bzw.)
Application Programming Interface (API)
Uniform Resource Locator (URL)
Really Simple Syndication in RSS 2.0 (RSS)
Chaos Computer Club (CCC)
Social Network (SN)
Oder ähnlich (o. ä.)
Deutsch (dt.)
Englisch (engl.)
Und andere (u. a.)
Et cetera (etc.)
Vergleiche (vgl.)
Abschnitt (Abs.)
Abbildung (Abb.)
Tabelle (Tab.)
Zum Beispiel (z. B.)
Und andere (et al.)
Datenbank (DB)
Freebase (FB)
Wikipedia (WiP)
Identifikationsbezeichnung oder Identifikationsnummer (ID)
Radio Frequency Identification (RFID)
Versus (vs.)
Circa (ca.)
Virtueller Interessenprofilabdruck (VIPA)
Realer Interessenprofilabdruck (RIPA)
Keine Angabe (k. a.)
Beispiel (Bsp.)
Umgangssprachlich (ugs.)
Fragebogenauswahlmöglichkeit (FAM)
Virtuell (virt.)
Taghäufigkeit, wie oft werden Tags verwendet (TH)
Objekt-Tag-Dienst-Tupel (OTD)
Virtuelle Realität (VR)
Name der Software, welche für die praktische Umsetzung des Verfahrens implementiert wurde (ThatsMe)
Virtuelle Realität (UGC)
Interessenschwerpunkt (ISP)
Beispielsweise (bspw.)

1. Einführung

Mit der Verlagerung vieler Bereiche des öffentlichen und privaten Lebens in das Internet bringt jeder zunehmend persönliche Informationen in die virtuelle Welt ein. Zu Beginn kann dies die Gestaltung des Avatars¹ in SecondLife² oder einem Online-Computerspiel sein³.

Trends, wie das Betreiben eines Weblogs, sind ein größerer Schritt, sich und seine Meinung im Netz zu präsentieren. Zu lesen ist von den überwiegend positiven Auswirkungen des Bürgerjournalismus und der freien Meinungsäußerung. In wissenschaftlichen, unternehmerischen und organisatorischen Bereichen tragen Weblogs zur besseren Kommunikation bei.[Sch06]

Die sozialen Gemeinschaften (Myspace⁴, Facebook⁵, StudiVZ⁶ und unzählige weitere) fordern ihre Teilnehmer dazu auf, die Profile für ihr Netzwerk zu füllen. Nur so sei gewährleistet, von allen Freunden gefunden zu werden und viele neue Bekanntschaften zu ergründen.

Zahlreiche „Web2.0“-Dienste überzeugen den Nutzer zur Registrierung. Die „Nachteile“ der Preisgabe an persönlichen Daten werden überwogen von unbegrenzten Vorteilen, welche sich durch die Anmeldung und Nutzung des Dienstes ergeben. Dass ein Kundenstamm von entscheidender Wichtigkeit ist, haben aber auch die etablierten Internetdienste erkannt. Mit Meinungsportalen zur Produktbewertung bindet bspw. Amazon⁷ die Nutzer enger an den Dienst. Der Vorteil, den registrierte Nutzer für einen Dienst bieten, geht über die Sicherheitsfrage hinaus, dabei spielt Kundenbindung eine große Rolle.⁸

Die Nutzung von Suchmaschinen, Statistiken⁹ nach überwiegend Google¹⁰, war bislang nur Kennern ein datenschutzrechtliches Problem. Doch auch hier fallen Profildaten an. Jede Suchanfrage setzt einen Teil des Puzzles dazu¹¹. Dies lässt Bürgerrechtler gegen die Betreiber aufbegehren¹².

All dies führt dazu, dass Menschen im World Wide Web immer gläserner werden. Berichte von Identitätsdiebstahl und digitalen Doppelgängern¹³ erscheinen neben Veröffentlichungen um dem entgegenzuwirken. Der FU Berlin stiftete die Bundesdruckerei eine Professur für sichere Identität.

¹Grafik, die den Teilnehmer eines Chats darstellt[Dud06, S. 221]

²<http://www.secondlife.com/> vom 11.10.2007.

³http://www.n24.de/wissen_technik/multimedia/article.php?articleId=131372 vom 11.10.2007.

⁴<http://www.myspace.com> vom 11.10.2007.

⁵<http://www.facebook.com> vom 11.10.2007.

⁶<http://www.studivz.de> vom 11.10.2007.

⁷<http://www.amazon.de> vom 11.10.2007.

⁸<http://www.alistapart.com/articles/identitymatters> vom 11.10.2007.

⁹<http://www.heise.de/newsticker/meldung/74552> vom 11.10.2007.

¹⁰<http://www.google.com> vom 11.10.2007.

¹¹<http://www.heise.de/newsticker/meldung/90069> vom 11.10.2007.

¹²<http://www.heise.de/newsticker/meldung/93261> vom 11.10.2007.

¹³<http://www.spiegel.de/netzwelt/web/0,1518,495618,00.html> vom 11.10.2007.

Andere Quellen rufen „Das Ende der Geheimnisse“ aus und fragen „wer in einer Welt leben will, in der alles öffentlich ist“. Geheimnisse seien in der Vergangenheit ein Garant für hohes Ansehen einer Person gewesen.¹⁴

Damit dies auch so bleibt, bieten unterschiedliche Firmen ihre Dienstleistungen an. Zum einen diejenigen, die Identitäten schützen wollen indem sie die Identität versichern und hohe Geldsummen nach dem Mißbrauch bieten¹⁵. Dann die, welche versuchen, alle Fundstellen von Identitätsdaten aufzuzeigen und sich um den Ruf der Person dahinter kümmern¹⁶. Zu guter Letzt die, die bewußt falsche künstlich generierte Identitäten anbieten¹⁷ beziehungsweise die Werber, die unter falscher Identität Produkte und Dienstleistungen anpreisen¹⁸.

Auf dem Gebiet der Werbetreibenden besteht zu Profildaten die Motivation, zielgerichtet Botschaften an Mann und Frau zu bringen. Teilweise kann der Nutzer mit seinem Profil und dem Verhalten im Internet Geld verdienen. Inwieweit Offenheit im virtuellen Raum Bezug zur Wirklichkeit hat, soll hier geklärt werden.

1.1. Anwendungsszenarien

An dieser Stelle sollen einige Szenarien beschrieben werden, wie das Verfahren und die entwickelte Software genutzt werden könnten.¹⁹ In Abbildung A.5 sind die Beispiele grafisch dargestellt.

Persönlicher Vergleich von Interessen Ein Person möchte erfahren, inwieweit ihr virtuelles Abbild mit den realen Vorlieben übereinstimmt (Bsp. vor Bewerbungsanfragen für Berufe)

Dienstweiter Vergleich von Interessen Der Anbieter eines Dienstes oder Dritte mit Zugriff auf die Daten von Nutzern möchten erkunden, welche Interessengebiete die Nutzer bevorzugen

Gruppeninteressen Die gemeinsamen Interessen einer Gruppe (Bsp. Freunde, Arbeitsgruppe, Gruppen in einem Dienst) von Personen sollen ermittelt werden

Benutzer mit speziellen Interessen erhalten Zu gegebenen Interessen sollen Nutzer (Bsp. in einem Dienst) gefunden werden (Bsp. zielgerichtet Werbung)

¹⁴Das Ende der Geheimnisse von Adam Soboczynski, S. 64, DIE ZEIT, Nr.11 am 8.März 2007.

¹⁵<http://www.lifelock.com/> vom 11.10.2007.

¹⁶<http://naymz.com/> vom 11.10.2007.

¹⁷<http://www.zulugrid.com/2006/06/16/false-identity-generator/> vom 11.10.2007.

¹⁸Tarnen und täuschen von Maximilian Geyer et al., STERN, S. 188, 16/2007.

¹⁹Einzig der erste Fall wird in der Arbeit verfolgt.

2. Problem

In der Einführung wurden Ansätze erwähnt, die zahlreiche Fragen und Problem aufwerfen. Dabei muss zum einen die Unterscheidung zwischen Gut und Böse oder auch negativ und positiv für den Nutzer betrachtet werden. Andererseits ist zu trennen zwischen Bekämpfung des Terrorismus als auch der Kriminalität durch die Organe des Staates und der zivilen Nutzung des Internet und dem damit einhergehenden Datenschutz.

Zahlreiche Begriffe stehen im Raum wenn in den Medien zum Thema berichtet wird. Einige sollen näher beleuchtet, andere bewußt ausgeklammert werden. Zu umfangreich ist die Problematik um sie komplett zu durchleuchten. Aktuell stehen vor allem die Politik und die Vorschläge von Innenminister Wolfgang Schäuble in der Diskussion. Eine Reihe von negativ belegten Begriffen stehen mit ihm im Zusammenhang. So sind Vorratsdatenspeicherung, Onlinedurchsuchung, Überwachungsgesellschaft, Flugdatentransfer, Kontoabfrage oder der Schäuble-Katalog zu nennen. Um die potenzielle Bedrohung der Privatsphäre dieser Maßnahmen aufzuzeigen, werden diese Ausdrücke in dieser Arbeit erneut auftauchen und näher betrachtet.

In diesem Zusammenhang liest man auch von der Identifikationsnummer¹. Diese soll, vergeben vom Bundeszentralamt für Steuern, einen Menschen von der Geburt bis 20 Jahre nach dem Tod begleiten. In ihr werden unter anderem Name, Künstlername, Geschlecht, Geburtstagsdatum, Adresse oder Doktorgrad gespeichert. Datenschützer beklagen allerdings die Möglichkeit, viele bisher getrennt lagernde Register zu kombinieren. Auf die Funktion dieser eindeutigen Kennung wird in Abschnitt 4.2.4 bei den Ausführungen zum Modell spezieller eingegangen.

Auch hier steht dem positiven Ansatz mit kritischen Auswirkungen ein kriminelles Problem gegenüber. Der Mißbrauch von Identitäten aus dem Internet nimmt zu. Schlagworte die hier fallen, sind Phishing, Pharming und Spoofing². Dabei dreht es sich im Allgemeinen darum, dass unberechtigte Personen die gesammelten Daten für ihre Aktionen nutzen und dadurch unbescholtene Bürger schädigen. Sei es durch Auktionen unter falschem Namen bei Ebay oder Kreditkartenbetrug mit ausgespähten Zugangsdaten³.

Doch die Problematik ist vielschichtiger. Meldungen, nach denen sich auf dem Suchmaschinenmarkt eine Abgrenzung anhand der Datenschutzmaßnahmen durchsetzt,⁴⁵ folgen neuen Sicherheitsbedenken welche ausführlichere Datensammlungen aufwerfen. Microsoft, Ask.com und Yahoo geben bekannt, die personenbezogene Speicherdauer der Suchanfrage zu verkürzen und danach nur noch anonymisiert Daten vorzuhalten. Die Ankündigung von Google es zu ermöglichen, Inhalte aus dem Such-Index zu entfernen, zielt ebenfalls in diese Richtung⁶.

¹<http://www.heise.de/newsticker/meldung/90890> vom 11.11.2007.

²<http://www.spiegel.de/netzwelt/web/0,1518,495618,00.html> vom 11.11.2007.

³<http://www.heise.de/newsticker/meldung/85160> vom 11.11.2007.

⁴<http://www.heise.de/newsticker/meldung/93116> vom 11.11.2007.

⁵<http://www.heise.de/newsticker/meldung/93193> vom 11.11.2007.

⁶<http://www.golem.de/0704/51760.html> vom 11.11.2007.

Gefahr für den Datenschutz zieht aus anderen Kreisen auf. Im Bereich der „Web2.0“-Dienste tritt nach außen, was das Geschäftsmodell bislang verbarg. So kündigt Youtube an, mehr Nutzerdaten sammeln zu wollen, um damit besser werben zu können⁷, StudiVZ ändert die AGBs⁸ und möchte am liebsten per Handy werben und Facebook ermöglicht mit Beacon⁹ Werbenden Zugriff auf die Mitglieder. All dies nicht ohne Kritik¹⁰.

Allerdings unterscheiden sich die Konzepte der oben genannten grundlegend von den in der Arbeit untersuchten Methoden. Bei diesen liegt der Schwerpunkt auf der eigenhändigen Eingabe des Nutzers und der freien Verfügbarkeit für Dritte. Wobei diese differenziert zu betrachten ist und separat in Abschnitt 5.1.1 erläutert wird.

Dabei muss auf die Fragestellung, inwiefern die Aktivitäten der Internetnutzer mit ihren realen Aktionen übereinstimmen, genauso eingegangen werden wie auf die Inhalte, welche Nutzer online hinterlegen¹¹ (vgl. 3.1). Auch die Nutzung zu beruflichen Zwecken oder im Bildungsbereich fällt in den Betrachtungsraum. So lassen sich aus privatem Material (Bsp. Fotos vom Urlaub) wohl leichter persönliche Interessen entdecken als aus den Daten, welche für berufliche Zwecke (Bsp. Links zum Thema des Berufs) hinterlegt sind.

Ein Hauptproblem der Arbeit liegt in der Verwertung der unzähligen Informationen zu aussagekräftigeren Argumenten. So kann bspw. das Konzept der „Tags“ einerseits ein guter Lieferant für Fakten sein, doch problematisch dabei ist die Vergleichbarkeit zwischen Nutzern und ihren Tags sowie die Bedeutung eines benutzerspezifisch vergebenen Schlagwortes. Dies ist für die Einordnung in Interessenprofile wichtig.

Um die Nutzer in Gruppen einteilen zu können und ihre Interessen zu kategorisieren, werden verschiedene Ansätze diskutiert. Neben Klassifikationen aus der Werbebranche oder den Medienunternehmen, gelangen auch klassische Interessenmodelle in die Überlegung. So wird neben AdSense von Google (vgl. 4.4.5), die Studie „@facts extra - Online-Nutzertypen 2007“ von SevenOne¹² (vgl. 4.4.3) und der Allgemeine Interessen-Struktur-Test zur Erfassung schulisch-beruflicher Interessen (vgl. 4.4.4) betrachtet.

⁷<http://www.heise.de/newsticker/meldung/88611> vom 11.11.2007.

⁸http://www.n24.de/wirtschaft_boerse/unternehmen/article.php?articleId=175742 vom 20.12.2007

⁹<http://www.facebook.com/business/?beacon> vom 20.12.2007

¹⁰<http://www.heise.de/newsticker/meldung/100642> vom 20.12.2007

¹¹Tagesaktuelle Ereignisse oder Geschehen mit privatem Bezug.

¹²Werbevermarkter von ProSiebenSat.1.

3. Aktuelle Situation

Um einen Überblick zu schaffen, was sich aktuell auf dem Gebiet der virtuellen Identität tut, behandelt dieses Kapitel die derzeitige Situation. In den Medien ist die Domäne der Identität in vielen Facetten im Gespräch. Neben Datenschutz schürt man vor allem mit der negativen Seite die Aufmerksamkeit. Vieles davon steht nur indirekt im Zusammenhang mit der Arbeit. Abschnitt 3.1 geht deshalb auch auf die Abgrenzung von bestimmten Bereichen, die in den Medien vermischt werden, ein.

Im Zeitalter der Social Networks und den darin enthaltenen enormen Datenmengen entdecken findige Geschäftsleute neue Chancen, den Nutzer an sich zu binden. Bei diesen Diensten etablieren sich unterschiedliche Formen. Einige kümmern sich um den Ruf des Nutzers im Internet. Andere monetarisieren das Surfverhalten des Anwenders. Die Werbeindustrie mit personenbezogener Ansprache an Kunden ist dabei eine der treibenden Kräfte. Um bei der schiereren Menge der Dienste nicht den Überblick zu verlieren, treten auch Aggregatoren auf. Diese bieten dem Nutzer die Möglichkeit, seine Profile und Spuren zu sammeln. Welche Aspekte dabei noch vorzufinden sind, bespricht Abschnitt 3.2.

Am Ende des Kapitels steht ein zahlenmäßiger Überblick. Darin wird der aktuelle Stand über demografische Charakteristika und die Verbreitung innerhalb der Gesellschaft aufgezeigt. Anhand von Studien und Umfragen sollen Zahlen und Statistiken das Marktsegment erläutern. Bei der Größenordnung, welche die Nutzerzahlen im Internet erreicht haben, müssen auch geographische Eigenschaften und weitere Betrachtungen zu den Zielgruppen analysiert werden.

3.1. Öffentliche Wahrnehmung rund um virtuelle Identität

Wie bereits in der Problemstellung erwähnt, wird der Begriff der Identität in vielerlei Hinsicht kommuniziert. Dass dabei Vorteilhaftes durch negative Berichterstattung an der Entwicklung gehindert wird, bleibt genau so wenig aus wie eine zunehmende Abstumpfung gegenüber Gefahren im Zusammenhang mit der Privatsphäre.

Kaum eine Tageszeitung kommt ohne einen Artikel zu Bundesinnenminister Schäubles Plänen aus. Dabei spielt der Politiker Terrorismusbekämpfung und die freie Informationsgesellschaft gegeneinander aus. Mit der „Vorratsdatenspeicherung“ und dem „Bundestrojaner“ soll die Überwachung der Kommunikationswege und die von Computern verdächtiger Personen ermöglicht werden. Doch der Widerstand wächst. Weil das „Recht auf informationelle Selbstbestimmung“¹ verletzt werde, planen Datenschützer zu klagen².

Zu den Kritiker zählen auch die Mitglieder des CCC. In ihrer monatlichen Radiosendung und dem zugehörigen Podcast³ gehen sie auf das Thema „Anonymität, Pseudonymität und

¹Geprägt vom Bundesverfassungsgericht 1983.

²http://www.tagesschau.de/aktuell/meldungen/0,OID6763748_REF1,00.html vom 01.12.2007.

³<http://chaosradio.ccc.de/cr121.html> vom 28.10.2007

der neue Trend zur uneingeschränkten Offenheit“ ein. Auch sie zeigen Vor- und Nachteile auf, sehen die Ansammlung von Identitäten an vielen Stellen aber mit Vorbehalt.

Zu den Schlagworten in diesem Zusammenhang ein Überblick.

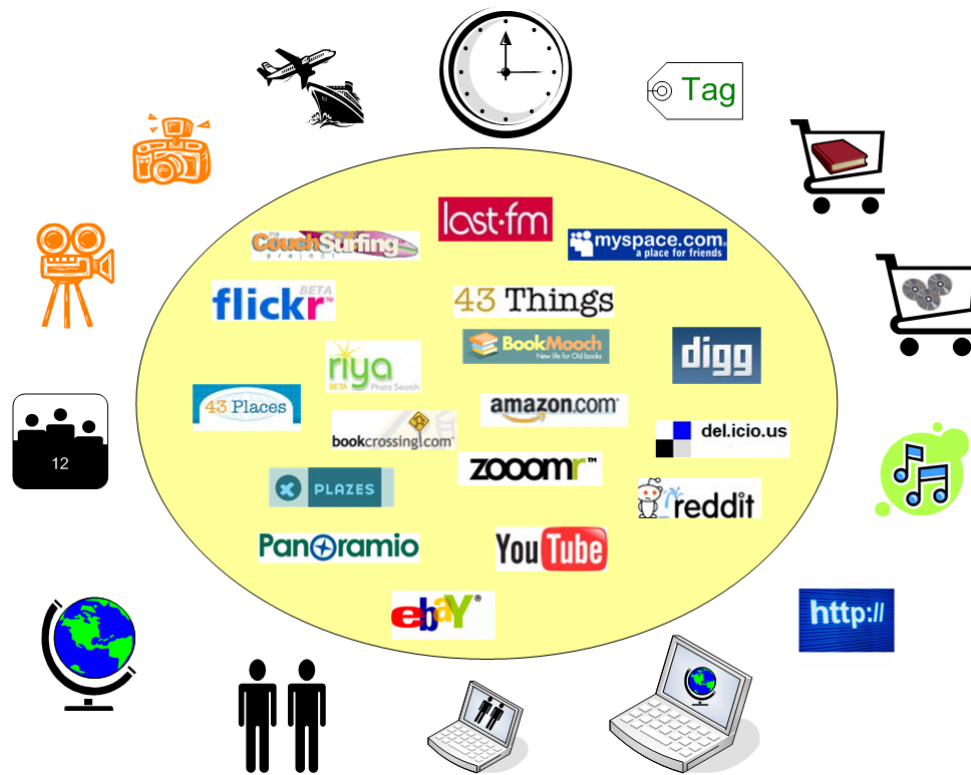


Abbildung 3.1.: Beispiele für Inhalte von Internetdiensten (eigene Darstellung)

3.1.1. Jahressteuergesetz mit Bürgeridentifikationsnummer

In die Kritik geraten ist der Entwurf des Bundeszentralamts für Steuern zum neuen Jahressteuergesetz. Auch dieses stört das oben angesprochene Recht nach Meinung der Gegner. Beanstandet werden unterschiedliche Punkte, zum einen die übermäßige Speicherung von nicht zwangsläufig benötigten Daten, zum anderen die Vergabe einer Identifikationsnummer. Das entstehende Überwachungspotenzial sei enorm. “Wegen der steuerrechtlichen Relevanz vieler Alltagsvorgänge von geschäftlichen Transaktionen bis zum einfachen Bezahlen einer Rechnung wird die neue Steuer-ID allgegenwärtig sein“⁴. Beachtet werden muss dabei, dass die Nummer von Geburt an bis 20 Jahre nach dem Tod jeden Deutschen eindeutig erkennbar macht. Den Vorteil, dadurch leichter gegen Steuerbetrüger vorgehen zu können, erkaufte man sich mit risikoreichen Nachteilen. Durch die “praktisch nicht möglich[e]“⁵ Hinderung der Erstellung von Persönlichkeitsprofilen entsteht ein großes Potenzial für Mißbrauch. Ein natürliches Problem (siehe Abs. 5.2.1) wird für Institutionen mit Zugriff

⁴<http://www.heise.de/newsticker/meldung/91963> vom 01.12.2007.

⁵Thilo Weichert vom Unabhängige Landeszentrum für Datenschutz Schleswig-Holstein
<http://www.heise.de/newsticker/meldung/91963> vom 01.12.2007.

auf die Daten aus der Welt geschafft. Die datenbankübergreifende, eindeutige Zuordnung wäre dadurch leicht möglich.⁶

Was bei der Volkszählung und den vorab verhandelten Klagen 1987 noch als einer der Hauptanklagepunkte diente, die Anonymität der Bürger, könnte hierdurch per se abgeschafft werden. Ein Bevölkerungsregister wäre eine mögliche Option. Der Bundesdatenschutzbeauftragte Peter Schaar fordert eine Zweckbindung, welche es verhindert, die erhobenen Daten für andere als die eigentlichen Zwecke zu nutzen.⁷

3.1.2. Terrorbekämpfung vs. Überwachungsgesellschaft

Nach den Anschlägen von New York, London und Madrid wurden verstärkt Maßnahmen ergriffen, die den organisierten Terrorismus bekämpfen sollen. Dabei pendelt die Wahrnehmung zwischen zwei Begriffen.

Einerseits beklagen kritische Beobachter „Trends zur Überwachungsgesellschaft“⁸, wobei vor allem die Vorratsdatenspeicherung (Abs. 3.1.3) Missbehagen erzeugt. Auf der anderen Seite tritt erneut Innenminister Schäuble bei der Terrorbekämpfung mit der Datensammlung unter dem Begriff „Antiterrordatei“ auf den Plan. Spiegel vergleicht die Datenbank mit der „Software eines Internetversands“⁹ allerdings sind die Inhalte weitaus gefährlicher. So sollen erstmals deutschlandweit alle Sicherheitsbehörden auf dem gleichen Datenbestand Abfragen zu Terrorverdächtigen tätigen können.

Weltweiter Datenaustausch beim Flugdatentransfer

Ebenfalls im Zusammenhang mit der Terrorismusbekämpfung aber noch weltumspannender geht der Austausch von auf Flughäfen gesammelten Personendaten. Die USA mit ihrem „Inselstatus“ und der durch die amerikanische Regierung erkannten Terrorbedrohung „hätten gute Gründe, wissen zu wollen, wer in die USA fliegt“ so Innenminister Schäuble.¹⁰ Doch es gibt dazu auch andere Ansichten. Der EU-Datenschutzbeauftragte Peter Hustinx bezweifelt „ob das Ergebnis dieser Verhandlungen voll mit den europäischen Grundrechten übereinstimmt“¹¹.

Vom EU-Parlament kommt ebenfalls Kritik zu dem Abkommen. Die Landesregierungen sollen „den vorliegenden Entwurf sorgfältig [...] überprüfen“¹². Die Punkte, welche bemängelt werden, stehen im Kontext des Datenschutzes. Es sei vertraglich nicht geregelt, wofür die Daten verwendet werden (Zweckbestimmung) und die Dauer der Speicherung sei von drei auf 15 Jahre erhöht worden. Zusätzlich können über Umwege Informationen zu europäischen Behörden kommen, die sonst klar geregelte Verfahren anwenden müssen, um an die Daten zu gelangen. Es ist ebenfalls nicht ausgeschlossen, dass Daten an Drittländer in aller Welt weitergegeben werden.¹³ Die Frage, welche Daten (derzeit 34 Datenfelder)

⁶<http://www.spiegel.de/politik/deutschland/0,1518,498687,00.html> und <http://www.heise.de/newsticker/meldung/91963> vom 01.12.2007.

⁷<http://www.heise.de/newsticker/meldung/92101> vom 01.12.2007.

⁸<http://www.heise.de/newsticker/meldung/91459> vom 07.12.2007.

⁹<http://www.spiegel.de/netzwelt/web/0,1518,474924,00.html> vom 07.12.2007.

¹⁰<http://www.heise.de/newsticker/meldung/91958> vom 07.12.2007.

¹¹<http://www.heise.de/newsticker/meldung/91958> vom 07.12.2007.

¹²<http://www.heise.de/newsticker/meldung/92608> vom 20.12.2007

¹³<http://www.heise.de/newsticker/meldung/92450> vom 07.12.2007.

übermittelt werden, stellt einen weiteren prekären Punkt dar.¹⁴

3.1.3. Vorratsdatenspeicherung

Ein anderer Baustein, welcher im Zusammenhang mit der Diskussion zur Informationsfreiheit nicht unerwähnt bleiben darf, ist die „kontroverse Debatte“ um die Vorratsdatenspeicherung.¹⁵

3.1.4. Gläserner Bürger

In Verbindung mit Identität und Datenschutz wird der Begriff des „Gläsernen Bürgers“. Dabei zeigen die Medien, an welchen Stellen Menschen Datenspuren hinterlassen, welche ihr Verhalten dokumentieren. Dabei kann man unterscheiden zwischen den real existenten Gefahren und denen die durch zukünftig denkbare Techniken entstehen können.

Der Alltag birgt einige Beispiele für die erste Gruppe. Es beginnt (am Morgen) mit der Nutzung des Mobiltelefons. Der Telekommunikationsanbieter speichert Verbindungsdaten und SMS-Gebrauch. Wer Kundenkarten (Bsp. Payback) zum Einkaufen verwendet, hinterlässt auch hier ausführliche Daten. Auch das Suchen im World Wide Web führt dazu Profildaten zu erzeugen. Abschnitt 3.2.3 nennt Details.¹⁶

Die Liste der vorstellbaren Fallen für Nutzerdaten ist ungleich länger. Diskussionen um RFID, digitales Fernsehen und TV-Grundverschlüsselung¹⁷ oder Mautsysteme auf den Straßen bergen genügend Stoff, um aufmerksame Zuhörer zu faszinieren.¹⁸ Was die unzähligen Internetdienste, welche eine Registrierung des Nutzers erfordern, mit den Datenbergen machen, ist zwar gesetzlich geregelt¹⁹ aber Risiken bestehen dennoch. Ein bei StudiVZ für den Datenschutz zuständiger Mitarbeiter führt dabei die Crawler an. Alle Daten, die im WWW öffentlich stehen und auch durch Anmeldeformulare geschützte Informationen, können durch die Datensammler und deren „automatisierte Abfragen gelesen werden“.²⁰ Nur eine Anonymisierung oder das bewusste Geheimhalten verhindert, dass Angaben in die falschen Hände geraten.

Allein der ausreichend informierte Bürger kann abschätzen, was zu seinem Vorteil geschieht und was die Bürgerrechte einschränkt. Der Satz, „Wer nichts zu verbergen hat, hat auch nichts zu befürchten“ wird wiederholt verwendet. Dieser Vorwand wird von Michael Lohmann im Bezug auf die drohende Einschränkung von Bürgerrechten analysiert.²¹ Bundesinnenminister Wolfgang Schäuble sieht aber vor allem im privaten Bereich erhebliche Gefahren und vermutet dort eine Bedrohung.²²

¹⁴<http://www.heise.de/newsticker/meldung/92608> vom 07.12.2007.

¹⁵<http://www.heise.de/newsticker/meldung/92349/> vom 20.12.2007

¹⁶http://www.n24.de/wissen_technik/multimedia/article.php?articleId=140241 vom 07.12.2007.

¹⁷<http://www.heise.de/newsticker/meldung/95415> vom 07.12.2007.

¹⁸<http://www.spiegel.de/netzwelt/web/0,1518,487773,00.html> vom 07.12.2007 Flash-Animation bei Panopti.com via SPIEGEL online.

¹⁹<http://www.datenschutz-berlin.de/doc/de/sonst/umiukdg1.htm> vom 20.12.2007.

²⁰<http://www.heise.de/newsticker/meldung/88793> vom 07.12.2007.

²¹<http://www.heise.de/tp/r4/artikel/23/23625/1.html> vom 07.12.2007.

²²<http://www.heise.de/newsticker/meldung/85023> vom 07.12.2007.

3.1.5. Identitätsdiebstahl

Virtuelle Identitäten sind anfällig für Identitätsdiebstahl. Nicht nur wohlhabende Personen werden davon bedroht²³. Auch der normale Nutzer rückt zunehmend in das Ziel der Angreifer. Unter den Schlagworten „Phishing, Pharming und Spoofing“ berichten die Medien von den kriminellen Machenschaften. Das Vorgehen ist einfach zu beschreiben. Die im Internet gefundenen Namen, Adressen, Bank- und Kreditkartendaten werden dazu mißbraucht, mit falscher Identität Geschäfte abzuwickeln. Geschädigt werden dabei nicht nur die Identitätsbestohlenen. Auch die mit getürkten Informationen hinter das Licht geführten „Online-Marktplätzen wie eBay (Versteigerung gestohlener Waren über einen fremden Account), Online-Versandhäuser, Bezahlssysteme wie PayPal oder Online-Beratungen und Kontaktbörsen“²⁴ erleiden Verluste.

3.1.6. Online-Reputation

In Wissenschaft, Politik und Medien ist der Ruf einer Person jeher ein wichtiges Gut. Auch in anderen Bereichen des Lebens erhält die Reputation immer mehr Aufmerksamkeit. In einer Zeit, wo sich alles und jeder im Internet bewerten lässt, stehen damit auch Personen unter dem Druck der Öffentlichkeit, die bisher nur im kleinen Kreis begutachtet wurden. Am Beispiel von Professoren und Lehrern kann man dies zeigen. Bislang fragte man Kommilitonen oder Mitschüler zur Meinung über eine Lehrkraft. Die Zahl der Einschätzungen war im persönlichen Gespräch gering. Mit dem Erscheinen von Bewertungsdiensten im WWW wurde diese ungleich größer. Die Transparenz brachte noch mehr Probleme mit sich. Berichte über Klagen mit der Überschrift „Rufschädigung oder Meinungsfreiheit?“²⁵ machen es deutlich. Positive Bewertungen werden gern gesehen. Sobald aber die negative Einschätzung überhand nimmt oder sogar beleidigend wird, entstehen Probleme. Nicht nur in Deutschland auch an schweizer Universitäten wird über das Thema diskutiert.²⁶ Neben Bewertungen zu Berufsständen wie Mediziner²⁷, Lehrern²⁸ und Hochschulprofessoren²⁹ ist auch der Ruf der Privatperson gefragt. Unter der Überschrift „Karrierekiller Google“ berichtet die Wirtschaftswoche in der Ausgabe 47/2006. Dabei wird vor allem die gefährliche Seite hervorgehoben, welche es notwendig macht, dass man sich „mindestens einmal pro Monat [der Imagepflege] widmen sollte“. In dem zu dem Artikel veröffentlichten Interview³⁰ nennt Reputationsforscherin Susanna Wieseneder einen Grund dafür, „die Leute lästern lieber, als dass sie loben“.

²³<http://www.heise.de/newsticker/meldung/94548> vom 07.12.2007.

²⁴<http://www.e-recht24.de/artikel/strafrecht/188.html> vom 07.12.2007.

²⁵http://www.n24.de/wissen_technik/multimedia/article.php?articleId=134142 vom 07.12.2007.

²⁶<http://www.spiegel.de/unispiegel/wunderbar/0,1518,494873,00.html> vom 07.12.2007.

²⁷<http://checkthedoc.de/> oder <http://ximt.de/> vom 07.12.2007.

²⁸<http://spickmich.de/> vom 07.12.2007.

²⁹<http://www.meinprof.de/> vom 07.12.2007.

³⁰<http://www.wiwo.de/pswiwo/fn/ww2/sfn/slink/bid/182996/index.html> vom 07.12.2007.

3.2. Dienste rund um virtuelle Identitäten

Die Liste der Anbieter, welche Dienste rund um die Online-Identität anbieten, ist lang.³¹ Unterschieden wird dabei zwischen verschiedenen Bereichen, zum einen dem Identitätsmanagement, bei welchem die Verwaltung der Erkennungsmerkmale im Mittelpunkt steht. Online-Reputation ist ein weiteres Segment. Hierbei steht der Ruf und das „Image“ einer Person im WWW im Vordergrund. Ein dritter großer Bereich sind die Aggregatoren. Diese verknüpfen die Profile unterschiedlicher Dienste zu einem großen Profil. Relativ neu treten Suchmaschinen für Personen auf. Im Kontext mit Interessenprofilen sind auch die Anbieter personalisierter Startseiten zu betrachten. Zu den genannten Kategorien sind in den folgenden Abschnitten jeweils Beispiele zu finden.

3.2.1. Identitätsmanagement

Im Bereich der Verwaltung der eigenen Identität wird von Datenschützern immer wieder zur Sparsamkeit bei der Preisgabe der persönlichen Informationen aufgerufen. Die Registrierung bei Anbietern erfordert wiederholt gleiche Informationen. Neben einer eindeutigen Kennung (E-Mail oder Benutzername) werden Passwort, Name und weitere Personendaten abgefragt.

Diesem Problem widmen sich Systeme beim Identitätsmanagement. Neben der freien Lösung openID³² bietet auch Microsoft mit CardSpace³³ Lösungen zum sogenannten „single-sign-on“-Verfahren an. Mit nur einer Kennung kann man sich bei unterschiedlichen Diensten anmelden. Doch auch diese Systeme haben Nachteile.

3.2.2. Dienste die „Identität“ in den Vordergrund stellen

Eine Vielzahl von Anbietern bieten Dienstleistungen rund um die Daten der Nutzer an, einerseits für Produzenten andererseits für Konsumenten. Anhand von RSS, Plugins in die Blog-Software und ähnlichem lassen sich die erzeugten Daten chronologisch in einen „Lifestream“ verpacken.³⁴ Eine Überblick über weitere Aggregatoren in diesem Sektor bietet TechCrunch³⁵.

Personalisierte Startseiten

Vor einigen Jahren prägten große Portal-Seiten die Eingangspunkte in das World Wide Web. Diese waren sehr allgemein und sprachen die breite Masse an. Ein gewichtiger Nachteil war die mangelnde Personalisierung für den Einzelnen. Viele kleinere Nischen waren nicht repräsentiert. Mit dem Aufkommen von RSS und der zunehmenden Individualisierung von Angeboten auf spezielle Gruppen kamen auch „Personalisierte Startseiten“ auf. Hierbei stellt sich der Nutzer sein eigenes Portal zusammen. Neben Angeboten direkt vom Diensteanbieter können auch fremde Informationen eingebunden werden. Dies geschieht häufig über RSS oder URL.

³¹<http://mashable.com/2007/09/10/online-identity/> vom 07.12.2007.

³²<http://openid.net/> vom 07.12.2007.

³³<http://www.microsoft.com/net/cardspace.aspx> vom 07.12.2007.

³⁴<http://lifestreamblog.com/define/> vom 12.12.2007.

³⁵<http://www.techcrunch.com/2007/03/15/streakr-search-makes-social-networks-bare-all/> vom 12.12.2007.

Allen Anbieter sind die abgedeckten Interessengebiete gleich. Neben Wetterberichten und dem Horoskop sind es die Kategorien Beliebte Gadgets, Nachrichten, Tools, Kommunikation, Spaß, Finanzen, Sport, Lifestyle, Technologie, die bspw. iGoogle anbietet³⁶. Neben Fotos, Videos und Musik aus anderen Internetseiten ist es ebenfalls möglich, seine E-Mail-Postfächer einzublenden. Die vier derzeit verbreitesten Anbieter sind iGoogle³⁷, Netvibes³⁸, Pageflakes³⁹ und Protopage^{40, 41}.

3.2.3. Suchmaschinen

Das Wort „googeln“ steht im Duden[Dud06, S. 436] und viele Begriffe lassen sich durch eine Anfrage an die Internetsuchmaschine Google schnell erklären. Bislang war dies für Menschen nicht so einfach. Zwar zeigten die Ergebnisse zahlreiche Hinweise auf eine gesuchte Person. Doch wenig Internetaffine oder Individuen mit weitverbreiteten Namen waren kaum oder nicht eindeutig zu finden. Mit den Personensuchmaschinen gibt es auf Menschen spezialisierte Recherchedienste. Doch auch die allgemeinen Vertreter bieten zahlreiche Ergebnisse, wenn mit ihnen nach Personen geforscht wird. Neben den Resultaten, die innerhalb von Profilen von sozialen Netzwerken, der Freunde-Suchmaschine Stayfriends⁴², privaten Internetauftritten und den Namensnennungen in Firmeninternetseiten entstehen, wird noch mehr Material vorgehalten.

Um diese nicht öffentlich zugänglichen Datensammlungen drehen sich Kritiken vieler Datenschützer. Suchmaschinenbetreiber begreifen vermehrt, dass Datenschutz ein Punkt ist, der Kunden beschäftigt. Bedenklich sind die Profile, die mit gespeicherten Suchanfragen erstellt werden können und bei den Suchmaschinen vorliegen. „Ask.com lässt Nutzer [ihre] Suchgeschichte löschen“.⁴³ Google anonymisiert Suchanfragen nun bereits nach 18 Monaten und nicht wie bisher nach 31 Jahren⁴⁴, startet aber gleichzeitig den Dienst „Web-History“. Dieser ermöglicht es, seine Suchergebnisse unter Berücksichtigung der im Vorfeld getätigten Anfragen und besuchten Websites „stärker personalisiert[e] und zielgerichtet“⁴⁵ zu verbessern.⁴⁶ Mit dem Webprotokoll kann der Nutzer seine Geschichte bei Google einsehen.⁴⁷ Der Ansatz dieser Arbeit stellt die reinen Personendaten aber gegenüber den Interessen in den Hintergrund.

Microsoft und Ask.com verlangen ein nachhaltigeres Vorgehen der gesamten Branche⁴⁸. Auch die „Datenkrake“ Google fordert globalen Datenschutz-Standard⁴⁹. Auf einer Konfe-

³⁶<http://www.google.com/ig/directory?root=/ig&igtb=Home&dpos=top> vom 11.12.2007.

³⁷<http://www.google.com/ig> vom 11.12.2007.

³⁸<http://www.netvibes.com/> vom 11.12.2007.

³⁹<http://www.pageflakes.com/> vom 11.12.2007.

⁴⁰<http://www.protopage.com/> vom 11.12.2007.

⁴¹<http://mashable.com/2007/06/29/personalized-homepages/> vom 12.12.2007 zeigt einen Vergleich der Funktionen von 14 Seiten.

⁴²<http://www.stayfriends.de/> vom 11.12.2007.

⁴³<http://www.golem.de/0707/53632.html> vom 11.12.2007.

⁴⁴<http://www.heise.de/newsticker/meldung/92807> vom 11.12.2007.

⁴⁵<http://www.heise.de/newsticker/meldung/93722> vom 11.12.2007.

⁴⁶<http://www.golem.de/0708/53874.html> vom 11.12.2007.

⁴⁷<http://www.google.com/history?hl=de> vom 12.12.2007.

⁴⁸<http://www.heise.de/newsticker/meldung/93116> vom 11.12.2007.

⁴⁹<http://www.spiegel.de/netzwelt/web/0,1518,505698,00.html> vom 11.12.2007.

renz der UNESCO⁵⁰ mahnt die Firma die Schaffung weltweit geltender Datenschutzregelungen an.⁵¹ Das Thema „Datenschutz ist [...] zu einem Wettbewerbsargument geworden.“⁵²

Die Bestrebungen, die bereits gespeicherten Daten möglichst gut zu sichern, sind die eine Seite. Ein anderer Ansatz versucht bereits beim Entstehen der Daten anzusetzen. Die Suche wird anonymisiert, der Suchende ist nicht mehr durch Cookies oder ähnliche Mechanismen wieder erkennbar. Für Google existiert ein solcher Dienst bereits⁵³.

Noch drastischer ist der Schritt ohne Google das Internet zu benutzen. „Ein Tag ohne Google“ titelt Spiegel-Online⁵⁴. Der Autor kommt zu dem Schluss, dass bei den Diensten die Google bietet, gute Alternative existieren. Für Suchanfragen ist aber ein Ausweichen nicht ohne Einschränkungen möglich.

3.2.4. Personensuchmaschinen

Allgemeine Suchmaschinen gibt es verschiedene. Vor allem Google ist weit verbreitet. Doch Spezialisten auf dem Gebiet der Personensuche sind vermehrt anzutreffen.

Wer hier nicht gefunden wird, kann neu angelegt und mit Informationen versehen werden. Personen lassen sich durch Tags näher beschreiben und so in Kategorien einordnen. Auf den Suchbegriff „Berlin“ antwortet Spock⁵⁵ als erstes Ergebnis mit Berlins Regierendem Bürgermeister Klaus Wowereit⁵⁶. Auch sonst sind Menschen, die in den Medien präsent sind, zu finden. Da es ein amerikanischer Dienst ist, tendiert die Gewichtung bei den Persönlichkeiten in diese Richtung. Die Bedenken, die dabei auftreten, stammen aus dem Bereich des Datenschutzes und der Privatsphäre.

Zum einen können Menschen Schlagworte ohne Überprüfung zugewiesen werden.⁵⁷ Zum anderen ist nicht klar, wie man sich gegen einen Eintrag zu seiner Person wehren kann. „Da der Dienst in den USA sitzt, greifen deutsche Datenschutzgesetze nicht.“⁵⁸ Umfangreiche Sicherungen und die gegenseitige Bewertung sollen Mißbrauch verhindern.⁵⁹

Spock ist nicht allein. Mit Wink liefert ein weiterer Wettbewerber auf Anfrage Informationen zu Personen. Vom Aufbau her ähnlich, kann man auch hier ein eigenes Profil anlegen. In dieser Suchmaschine werden Profile von Myspace, LinkedIn and Bebo durchsucht. Alle zwei Wochen sollen neue Quellen dazukommen.⁶⁰ Neuere Angaben nennen auch Xing, Flickr oder Facebook als Datenursprung.⁶¹

⁵⁰http://portal.unesco.org/ci/en/ev.php-URL_ID=24772&URL_DO=DO_TOPIC&URL_SECTION=201.html vom 11.12.2007.

⁵¹<http://www.heise.de/newsticker/meldung/95973/> vom 11.12.2007.

⁵²<http://www.spiegel.de/netzwelt/web/0,1518,496020,00.html> vom 11.12.2007.

⁵³<http://lifelife.com/software/privacy/search-google-anonymously-at-googlonymous-287857.php> vom 11.12.2007.

⁵⁴<http://www.spiegel.de/netzwelt/web/0,1518,488347,00.html> vom 11.11.2007.

⁵⁵<http://www.spock.com/> vom 11.12.2007.

⁵⁶<http://www.spock.com/q/Berlin> vom 14.09.2007.

⁵⁷<http://nerds.computernotizen.de/2007/08/14/spock-und-der-mob-ist-da/> vom 11.11.2007.

⁵⁸<http://netzpolitik.org/2007/spockcom-alptraum-der-datenschuetzer/> vom 11.11.2007.

⁵⁹http://www.n24.de/wissen_technik/multimedia/article.php?articleId=141202 vom 11.11.2007.

⁶⁰<http://www.techcrunch.com/2006/11/10/wink-now-searches-myspace-linkedin-and-beebo/> vom 11.11.2007.

⁶¹<http://www.golem.de/0708/53993.html> vom 11.11.2007.

Weitere Personensuchmaschinen sind Squidoo⁶², WikiYou⁶³ oder ZoomInfo⁶⁴. Neue sind für den deutschen Markt angekündigt.⁶⁵ Alle haben gemein vor allem zu Stars aus Film und Fernsehen und anderen Medien Daten zu beherbergen. Auch mit Srchr⁶⁶ und Social-Grapes⁶⁷ lassen sich Soziale Netzwerke gezielt durchsuchen.

Da bislang noch viele Dienste ihre Nutzerprofile hinter einer Registrierung bzw. Anmeldung verstecken, haben Suchmaschinen keinen Zugriff darauf. Doch das kann sich ändern. Facebook macht einen Schritt hin zur Durchsuchbarkeit seiner Daten. Vorerst haben auch nicht registrierte Nutzer Zugriff auf Namen und Bild und können Kontakt über Facebook aufnehmen. Später sollen diese Informationen auch in Suchmaschinen erscheinen. Bisher kam in den Index nur was explizit vom Besitzer öffentlich gemacht wurde.⁶⁸ Geplant ist auch die Änderungen in Profilen per RSS-Feed zur Verfügung zu stellen⁶⁹.

3.2.5. Datenschutz

Der Bereich des Datenschutzes wurde schon in einigen der vorhergehenden Abschnitte gestreift. In dieser Passage soll ihm nochmals besonderes Augenmerk gelten. Zahlreiche technische Neuerungen aber auch altbekannte Ansätze werfen Fragen auf. So gerät die GEZ in Verdacht, gegen den Datenschutz zu verstossen, indem Schwarzseher ausfindig gemacht werden⁷⁰. Die bei Wirtschaftsauskunfteien gespeicherten Daten erleichtern einerseites das Auffinden verschollen geglaubter Personen zum Klassentreffen. Andererseits ist Mißbrauch ebenfalls vorstellbar.⁷¹ Die Diskussionen um die Gefahren, welche von den über Googles Streetview⁷² zugängigen Daten ausgehen, beschränken sich derzeit auf den amerikanischen Markt⁷³. Aber bereits mit der Ankündigung, den Dienst auch in Kanada einzuführen, erheben sich neue Proteste⁷⁴. Die Privatsphäre der kanadischen Bürger müsse respektiert werden.

3.2.6. Visualisierung von Profildaten

Die Sammlung von Daten zu Personen ist ausgiebiges Thema gewesen. Damit all die Informationen aber auch anschaulich verwertet und begutachtet werden können, bedarf es der Visualisierung. Vor allem die sozialen Netzwerke bieten mit API-Zugriff und verknüpften Daten gute Ausgangsbasen.

Nicht nur die Dienste, welche Inhaber der Daten sind, bieten Darstellungen an. Sie zeigen, wer mit wem in Beziehung steht und wen man über weitere Kontakte kennt. Im

⁶²<http://www.squidoo.com/squidwho/hq> vom 11.11.2007.

⁶³<http://www.wikiyou.com/> vom 11.11.2007

⁶⁴<http://www.zoominfo.com/> vom 11.11.2007.

⁶⁵<http://www.zweinux.cc/datenschutzer-wetzt-eure-messer-yasni-kommt/> vom 11.11.2007.

⁶⁶<http://www.srchr.com> vom 12.12.2007.

⁶⁷<http://socialgrapes.com> vom 12.12.2007.

⁶⁸<http://www.golem.de/0709/54555.html> vom 11.11.2007.

⁶⁹Bei Myspace und Xing <http://www.zweinux.cc/nach-myspace-nun-xing-newsfeed-ab-montag/> vom 12.12.2007.

⁷⁰<http://www.heise.de/newsticker/meldung/96126> vom 11.11.2007.

⁷¹<http://www.spiegel.de/netzwelt/web/0,1518,452447,00.html> vom 11.11.2007.

⁷²<http://maps.google.com/help/maps/streetview/> vom 11.11.2007.

⁷³<http://www.heise.de/newsticker/meldung/91937> vom 11.11.2007

⁷⁴<http://www.heise.de/newsticker/meldung/95921> vom 11.11.2007.

kleinen wird dies auf den Profildaten bei OpenBC und StudiVZ verwendet. Weitaus größer illustrieren Drittanbieter die Netzwerke. Fldgt⁷⁵ visualisiert neben den Beziehungen der Flickr-Benutzer auch deren Tags und stellt übersichtlich dar, wo Schwerpunkte liegen. Cloudalicious⁷⁶ stellt die Entwicklung der Tags zu Links bei Delicious in Diagrammform dar. So bietet sich dem Betrachter eine neue Sicht auf Daten, die sonst nur in Textform zu sehen sind. Auch die Kontaktnetzwerke von Delicious werden visualisiert⁷⁷.

3.2.7. Werbeindustrie

Die Daten und Vorlieben der Nutzer sind für die Werbebranche von Interesse. Dienstleister wie navtracks⁷⁸ bieten ihren Service mit dem Satz „Erkennen Sie das Profil Ihrer Web-Besucher“. Ein Großteil der bei Übernahmen gezahlten Beträge beruht auf der Annahme mit den zahlreichen Kundendaten Gewinne zu erzielen.⁷⁹

3.3. Zahlenbasis

Um die Relevanz einzelner Dienste einschätzen zu können, ist es notwendig die zugrunde liegenden Statistiken zu untersuchen. Auch die Aussagekraft der Interessenprofile in virtuellen Identitäten hängt stark damit zusammen wie die Nutzer soziale Netzwerke und Dienste darum annehmen. Geringe Verbreitung schmälert die Bedeutung. Nutzen nur homogene Gruppen die Dienste sind auch die Inhalte der Profile eher eingegrenzt auf einen Themenbereich. Zu Beginn sind die fortschrittlichen Anwender in der Überzahl. „Early Adopters“ probieren aus und beteiligen sich an den Testphasen. Erst mit zunehmendem Bekanntheit steigen die Nutzerzahlen. Inwieweit dies für die zu Interessenprofilen beitragenden Anbieter zutrifft, zeigen Studien und Umfragen. Digg, eine Nachrichten-Börse, bei der Nutzer Neuigkeiten sammeln und bewerten, möchte bspw. weg vom Image der „Geeks“. Eine Ausrichtung hin zum Thema Politik sei bereits im Gange, so der Gründer Kevin Rose im Interview mit Technology Review[Pon07].

Da die Dienste ihre genauen Kundendaten geheimhalten und damit auch verbergen, wie viele aktive Produzenten den passiven Konsumenten oder gar „Karteileichen“ gegenüberstehen, helfen Schätzungen etwas Licht ins Dunkel der Nutzerzahlen zu bringen (vgl. 3.3.2).

Ein Schlagwort der Web2.0-Anbieter muss dabei gesondert betrachtet werden. Was man im deutschen mit Nischenmarkt beschreibt, ist auch als „Long Tail“ bekannt. Dieses Konzept, gezielt nur bestimmte, eingegrenzte Personenkreise anzusprechen, schränkt die Vielfalt weiter ein. So suchen sich Anbieter ihre Zielgruppen am Rande der großen Massen aus.

Bei Nutzern bestehen zum Teil zusätzlich Barrieren, die eine Durchdringung eines Dienstes in einem Markt erschweren. Dem entgegen erleichtern manche Gegebenheiten die Verwendung von Diensten. Seien es sprachliche Anforderungen, wegen denen der Anwender

⁷⁵<http://www.fldgt.com/visualize> vom 11.11.2007.

⁷⁶<http://cloudalicio.us> vom 11.11.2007.

⁷⁷<http://de.skurt.de> vom 11.11.2007.

⁷⁸<http://www.navtracks.de> vom 12.12.2007.

⁷⁹http://www.tagesschau.de/aktuell/meldungen/0,,OID7177888_REF1,00.html vom 12.12.2007.

einen Dienst favorisiert oder auch der Personenkreis, welcher die Inhalte einer Internetseite prägt (siehe dazu auch Abs. 9.2.3). Bei sozialen Netzwerken spielt auch die Akzeptanz unter Gleichgesinnten eine Rolle. Die Arbeit widmet sich diesen Details im Abschnitt 4.1.1.

Mit der zunehmenden Verbreitung des Internets als Massenmedium neben Zeitung, Radio und Fernsehen steigt auch die alltägliche Nutzung. Ob beruflich oder privat E-Mail und World Wide Web erreichen immer mehr Menschen. In Deutschland surfen bereits mehr als 60 % der Bevölkerung⁸⁰ im Internet. Dabei spielt auch die Demographie eine Rolle. Der Umstand, dass die Studien meist auf regionale Märkte wie Deutschland oder USA beschränkt sind, macht die Situation unübersichtlich und einen Vergleich schwer.

In Deutschland besteht vor allem in der Breite Nachholbedarf. Es kann keinesfalls von einer durchdringenden Verbreitung gesprochen werden. Dies zeigt eine Umfrage des Marktforschungsinstitut Forrester Research.⁸¹

Zu diesen Fakten passen auch Daten die „Dialago/Market Research Online in Deutschland und Frankreich mit je 1000 Interviews sowie mit jeweils 500 Fragebögen in Italien, Schweden und Russland im Auftrag von Fujitsu Siemens“ gesammelt und analysiert wurden. Demnach kennen zwar 71 % Second Life doch 62 % der Deutschen haben noch nie Web2.0 ausprobiert, nur 3 % nutzen Flickr regelmäßig und sind Blog-Muffel. Im europäischen Vergleich hinken sie hinter Russland (Flickr beliebter) und Italien (bevorzugen Blogs) her. Im Bereich des Onlineshopping zeichnet sich das Bild positiver, 82 % der Deutschen haben dies bereits einmal getan⁸².

Auch Stern.de wollte wissen, welche Begrifflichkeiten allgemein bekannt sind und was nur von Experten verwendet wird. Mit der Untersuchung die das Hamburger Meinungsforschungsinstitut SirValUse durchführte, kam man zu folgendem Ergebnis. 90 % kennen Web2.0-Angebote (75 % nutzen diese auch), dazu zählen für Deutsche „Enzyklopädien (Wikipedia) und Online-Diskussions-/Informations-Foren“. Amerika identifiziert damit Video- und Fototauschbörsen. Dass Blogs bei den Deutschen relativ unpopulär sind, zeigt der Vergleich DE: 13 % USA: 32 %, noch dahinter reihen sich Linksammlungen (Social Bookmarking) ein. Aktiv (in dem Inhalte eingestellt werden) werden in Deutschland hauptsächlich Blogs genutzt. In Amerika ist das Spektrum breiter, es reicht über Foren, soziale Netzwerke zu Online-Linksammlungen. Laut Umfrage hat die Selbstdarstellung einen überwiegenden Adressaten, die Gruppe der Freunde und Bekannten. Dabei sticht hervor, dass Amerikaner es als wichtiger ansehen, „persönliche Daten ins Netz zu stellen“ als die Deutschen.⁸³

Die Ergebnisse, welche Stern.de vermittelt stehen im Widerspruch zu der Studie von PricewaterhouseCoopers (PWC). Dort kennen nur 15 % den Begriff „Web2.0“. Selbst unter Technikern sind es hierbei nur 30 %. Dies liegt wohl an der Auswahl der Befragten und des Umfrageumfeldes. SirValUse fragt online je 500 Nutzer in USA und Deutschland während bei PWC repräsentativ 500 deutsche Haushalte ausgewählt wurden.⁸⁴

⁸⁰Zahlen aus <http://www.nonliner-atlas.de/> vom 11.11.2007 zum Zeitraum Frühjahr 2007.

⁸¹<http://www.zweinull.cc/deutschland-auf-dem-weg-zum-web-20-entwicklungsland/> vom 12.12.2007.

⁸²<http://www.heise.de/newsticker/meldung/95280> vom 11.11.2007.

⁸³<http://stern.de/computer-technik/internet/591483.html> vom 12.12.2007.

⁸⁴http://www.pwc.de/portal/pub/!ut/p/kcxml/04_Sj9SPykssy0xPLMnMz0vM0Y_QjzKLd4p3djUBSZnFG8Q76kfCRIL0vfV vom 12.12.2007.

3.3.1. Marktsegmentierung

Aufgrund der Differenzierung für spezielle Zielgruppen trifft man auf für spezielle Märkte und Umfeldler adaptierte Anbieter. Diese stammen selbst bei ausschließlicher Sprachanpassung nicht zwangsläufig von der gleichen Firma. Kopien sind in diesem Sektor eine verbreitete Methode, um am Erfolg zu partizipieren.

Diese Aufteilung des Marktes ist auch eine Schwierigkeit für die Erstellung von Interessenprofilen. Damit befasst sich Abschnitt 4.1.1. Da sich keine standardisierten Schnittstellen etablieren, weil jeder Dienst seine Daten für sich behalten will und Wettbewerbern nur in fremden Marktsegmenten ermöglicht, Verbindungen herzustellen, ist ein unkomplizierter, einheitlicher Zugriff nicht möglich. So muss für jeden weiteren Dienst, der zu einem Profil beitragen soll, festgelegt werden, wie die Daten extrahiert werden.

3.3.2. Nutzerzahlen

Die Bedeutung eines Dienstes hängt stark von der Masse seiner Nutzer ab. Für Anwender spielt der Vorteil, viele Gleichgesinnte anzutreffen dabei eine Rolle. Für diese Arbeit liegt der Gewinn darin, die Chancen zu erhöhen Profildaten einer Person in einem verwendeten Dienst zu finden.

Dass hierbei der Unterschied zwischen Zahlenangaben der Anbieter und bereinigten, relevanten Werten erheblich sein kann, zeigen verschiedene Beispiele. Dabei kommt es ganz darauf an, was als Nutzer gesehen wird. Die Annahme, Nutzer, die seit mehr als drei Monaten nicht mehr in ihren Profilen tätig waren, als inaktiv zu bezeichnen, spielt dabei eine Rolle. Im Fall von Myspace sind demnach ein Drittel der Nutzer nicht mehr aktiv. Wenn man allerdings von einer Nutzerzahl von weltweit 180 Millionen ausgeht, liegt noch immer ein großer Berg an Daten vor. Allein in Deutschland sind nach Angaben des Betreibers vier Millionen Kunden beheimatet.⁸⁵

Auch für Ebay sind Recherchen zum Unterschied zwischen Gesamtbenutzerzahl und der Anzahl aktiver Nutzer bekannt. Zum 30. Juni 2007 sollen bis zu 66 % der Nutzer und damit knapp 157 Millionen für den Zeitraum von 12 Monaten weder ver- noch gekauft haben. Allerdings stieg die Zahl der Nutzer von 202 Millionen weltweit auf 241 Millionen. Dies bedeutet Ebay wuchs im Quartal um 10 Millionen Angemeldete, verliert aber gleichzeitig ebenso viele Nutzer in die Inaktivität.⁸⁶

Im Bereich der sozialen Netzwerke herrscht ein Konkurrenzkampf. Große Gegner sind unter anderem Friendster, Facebook, StudiVZ und Myspace. 84 % der Nutzer unterhalten laut einer Studie des Marktforschungsinstitut Parks Associates aus dem Juni 2007⁸⁷ bei maximal zwei Anbietern ein Profil. Die Marktführer machen es für kleine Anbieter immer schwieriger, Nutzer zu gewinnen.

Eine weitere Studie vergleicht das Wachstum von sieben Gemeinschaften anhand der Besucherzahlen auf dem amerikanischen Markt. Dabei liegt Facebook für den Zeitraum Juni 2006 bis Juni 2007 bei 270 % Zuwachs auf Platz zwei, Myspace mit 72 % im Mittelfeld. Ein Netzwerk namens „Tagged“ führt mit 774 % die Liste der Steigerungen an. Allerdings

⁸⁵<http://www.zweinull.cc/myspace-ein-drittel-der-nutzer-inaktiv/> vom 11.11.2007.

⁸⁶Zahlen <http://wortfilter.de/News/news2248.html> und Bewertung <http://www.kriegs-recht.de/zweidrittel-der-ebay-nutzer-karteileichen-teil-ii/> vom 11.11.2007.

⁸⁷http://newsroom.parksassociates.com/article_display.cfm?article_id=4418 vom 11.11.2007 Befragung von 402 US-Nutzern.

zählt man hier nur 13 Millionen Nutzer, was im Vergleich zu Facebook mit 52 Millionen und Myspace mit 114 Millionen recht gering ist.

Dabei wird auch auf die regionale Aufteilung des Publikums eingegangen. In Nordamerika haben Myspace mit 62 % und Facebook mit 68 %, in Lateinamerika Orkut mit 49 %, in Europa Bebo mit 62 %, im pazifischen Raum Friendster mit 88 % hohe Anteile ihrer Nutzer. Für Afrika (hier Mittel- und Ostafrika) liegen Hi5 mit 9 % und Facebook mit 6 % relativ dicht beieinander.⁸⁸

Zu den Statistiken, die sich um die Zahlen der Nutzer im Internet drehen, gehört auch eine Betrachtung außerhalb der sozialen Netzwerke. Private Internetpräsenzen zeigen eine weitere Ausprägung von Informationen, die zum Interessenprofil beitragen können. Nach einer Umfrage des Meinungsforschungsinstituts Forsa von 1000 Personen im Alter ab 14 Jahren besitzen 20 % eine private Internet-Präsenz. Es kam aber auch heraus, dass Community-Profile beliebter als eigene Homepages sind.⁸⁹

Einige Nutzerzahlen treten ohne die Möglichkeit zum Vergleich auf, sollen aber nicht vorenthalten werden.

- „Mehr als 826.000 Einwohner Londons haben ein Profil bei Facebook. Damit ist die britische Metropole die Stadt mit der weltweit höchsten Zahl an Facebook-Nutzern“⁹⁰
- Xing (OpenBC): „Kundenzahl wuchs nach 2,13 Millionen im ersten Quartal des Jahres (2007) auf 3,52 Millionen“ Grund Zukäufe von fremdsprachigen Communities⁹¹

⁸⁸<http://www.comscore.com/press/release.asp?press=1555> vom 11.11.2007.

⁸⁹<http://www.heise.de/newsticker/meldung/94313> vom 11.11.2007.

⁹⁰<http://www.zweinull.cc/privates-surfen-am-arbeitsplatz-ein-alter-hut/> vom 11.11.2007.

⁹¹<http://www.heise.de/newsticker/meldung/94676> vom 11.11.2007.

4. Modelltheoretischer Ansatz

Bevor das folgende Kapitel das Verfahren zur Erstellung eines Profils beleuchtet, konzentriert sich dieser Abschnitt auf das Modell des Interessenprofils. Im Unterschied zu den Personenprofilen geht es bei Interessenprofilen nicht um demographische Faktoren. Daher ist eine Abgrenzung der Daten notwendig, die für Interessen trotzdem von Belang sind und bei denen die Angaben zur Person vernachlässigt werden. Im Umgang mit den Vorlieben einer Person sind Begriffe wichtig die immer wieder auftauchen. Diese sollen vorab zum besseren Verständnis und um hervorzuheben, warum sie wichtig sind, erläutert werden.[Hee04]

4.1. Definitionen und Zahlen

Bevor anhand von unterschiedlichen Vorlagen das Interessenprofil modelliert wird, sind Begriffe aus dem Umfeld zu klären. Des Weiteren soll in diesem Bereich auf die Statistiken im Zusammenhang mit den betrachteten Bereichen eingegangen werden. Neben der allgemeinen Internetnutzung stehen dabei Zahlen zum Web2.0-Gedanken im Vordergrund.

4.1.1. Definitionen

Bereits der Titel der Arbeit „Interessenprofile in virtuellen Identitäten“ birgt Begriffe, die näher beleuchtet werden müssen. Im Umfeld stehen aber noch weitere Fachausdrücke. Einige davon sind denjenigen, die sich mit dem Internet beschäftigen aus dem Zusammenhang geläufig. Doch verbergen sich bereits in den Definitionen wertvolle Informationen im Umgang mit Problemen und Fragen der Arbeit.

Virtuell

Alltagssprachlich übersetzt bedeutet es „scheinbar“ oder „augenscheinlich“. Oft verwendet wird es im Zusammenhang mit der Realität, als virtuelle Realität VR. Dazu schreibt[Sum95, S. 1597] „an image produced by a computer that surrounds the person looking at it and seems almost real“. Klarer im Kontext dieser Arbeit wird es in Verbindung mit dem Begriff der Identität (Abs. 4.2).

In der Philosophie behandelt unter anderem Baudrillard Virtualität im Zusammenhang mit Massenmedien. Die Aufsätze des ersteren „beschäftigen sich durchgehend mit der Auflösung der Wirklichkeit [...] in den binären Code“[Vai00, S. 170]. Ein „Ersatzbegriff für die Wirklichkeit“ ist dabei neben Virtualität, Simulation und Hyperrealität auch das Simulakrum. Dieses wird erklärt als „reproduction[s] of objects or events“ oder auch enger gefasst als „ein abstraktes System von Zeichen, dass in einer spezifischen Beziehung zur materiellen Welt steht und ein Konstruktionsmodell von Wirklichkeit bildet“.[Vai00, S. 178][ed00] Mit

dem Buch *System der Dinge* [Bau07] aus der Medientheoretik beschreibt er seine Sicht weiter. Die Ansicht Mikos' über Baudrillard's Theorien führt zurück zum Thema der Arbeit. „Visuelle Medien [...] verdoppeln Realität nicht, sie bilden Realität ab“.[Mik94, S. 190]

Eine weitere Definition bezogen auf „computergenerierte künstliche[n] Paralleltwelten“ bietet Hülsmann, wobei auch hier die Ansicht „das Virtuelle ist nicht der Gegensatz zum Realen“ vertreten wird. Für den Autor wird die Gesellschaft heutzutage dadurch geprägt, dass der „computererzeugte Raum ein Teil des Sozialraumes ist und gleichberechtigt neben dem physischen Raum steht“.[Hül00a, S. 47]

Interessenprofile

Bevor der Begriff in seiner zusammengesetzten Bedeutung erläutert werden kann, stehen die beiden Teile für sich allein. Interesse, auch bezeichnet als Vorlieben oder in manchen Bereichen auch als Geschmack (siehe Musik), „ist Teilnahme, Aufmerksamkeit, Liebe zu einer Sache“.[AMH73, S. 323] Unter dem Profil versteht der Duden ein „charakteristisches Erscheinungsbild“.[Dud06, S. 810] Fasst man die Begriffe zusammen, erhält man die für eine Person spezifische Darstellung der Neigungen. In den Wirtschaftswissenschaften hat das Interessenprofil Einfluss auf die Zugehörigkeit zur Zielgruppe. Dabei werden die Interessen unterteilt. Einerseits ist von inhaltlichen Vorlieben (Beispiele: Sport, Reise, Technik) die Rede, andererseits von dienstorientierten Interessenschwerpunkten (Beispiele: Spiele, Shopping, Kommunikation).[OFG07, S. 10f] Die Ansichten weiterer Bereiche, wie die der Marktforschung und Werbung, betrachtet Abschnitt 4.4.

Klassifikationen

Ordnen und Gruppieren liegt in der Natur des Menschen. Doch auch „...wenn wir [...] noch so methodisch verfahren, jeder Wunsch, die Erscheinungen dieses großen Gebietes nach Gruppen zu ordnen, kann nur einen provisorischen Charakter haben...“[Her88, S. 82]. Einen Grund, aus wirtschaftlicher Sicht für das Klassifizieren birgt die Bemerkung „Classification makes shopping [...] very much easier“[Hun05, S. 3]. In der Definition zu Klassifikationen bei Hunter heißt es „As humans we are able to recognize a member of a particular class because it displays certain characteristics common to that class but not to others“.[Hun05, S. 1]

Zwei der drei Beispiele für lang erprobte und angewandte Klassifikationen zeigen, dass bereits lange vor dem Auftreten des Computers und elektronischer Datenverarbeitung auf diesem Gebiet geforscht wurde. Mit der Dewey Decimal Classification (DDC) entwickelte Melvil Dewey bereits um 1900 eine Ordnung, die ihre Anwendung in Bibliotheken fand und auch heute noch, in mittlerweile mehr als 20 Auflagen angepasst, eingesetzt wird.[Hun05, S. 49] Eine weitere Kategorisierung, die aber verstärkt in den wissenschaftlichen Bibliotheken im Einsatz ist und ebenfalls gegen Ende des 19. Jahrhunderts entwickelt wurde, ist die Library of Congress Classification.[Hun05, S. 53] Moderner und noch spezieller ordnet das ACM Computing Classification System die Veröffentlichungen.[Hun05, S. 48] Die Liste mit allgemeinen sowie speziellen Klassifikationen ließe sich weiter fortsetzen. Dabei stellen die Ordnungen für Bibliotheken die Klasse mit der längsten „Lebenserfahrung“.

Alle Vertreter der Categoriesysteme lassen sich wiederum ordnen. Unterscheidungen bestehen anhand der betrachteten Subjekte, abstrakte oder konkrete (Bsp. Mut vs. Haus).

Weiterhin wird anhand der Vorgehensweise beim Erstellen differenziert. Ausprägungen dabei sind der „top-down“-Ansatz der hierarchischen Klassifikationen oder „bottom-up“ bei der „Faceted“-Klassifikation. Für Erstere stellt die enumerative Variante einen Spezialfall dar. Beispiele sind botanische Ordnungssysteme aber auch die DDC gehört in diese Klasse.[Hun05, S. 5] Für Facetten-orientierte Gliederungen sollte klar sein, was eine Facette ist. Das Oxford Dictionary beschreibt es als „one side of a many sided body“¹. Ein Vorteil der letzteren besteht in leichter Erweiterbarkeit, dagegen sprechen lange und komplexe Systeme sowie im entsprechenden Anwendungsfall die unpassende Verwendung für Regale in Bibliotheken. Vorteilhaft für die enumerativen, hierarchischen Klassifikationen ist die lange Erfahrung und weit verbreitete Benutzung. Kehrseite ist die Erweiterbarkeit mit neuen Konzepten, welche teilweise vollständige Überarbeitungen zur Folge haben.[Hun05, S. 68] In [Bro04, S. 37] wird noch das „analytic-synthetic scheme“ als Klassifikationsart genannt.

Strukturell wird weiterhin unterschieden zwischen disjunkten Klassifikationen, ohne Überschneidung der Klassen, und nicht disjunkten Klassifikationen, mit (begrenztem) Überschneiden. Zusätzliches Kennzeichen, relevant für die Arbeit vor allem zweite Ausprägung, sind exhaustive und nicht exhaustive Klassifikationen. Dabei werden alle bzw. nur die 'wichtigsten' Objekte in die Gruppierung einbezogen. Die Prinzipien der Datenreduktion und der Abstraktion, „Objektmenge zerfällt in kleine, prinzipiell gut unterscheidbare Gruppen“ werden eingesetzt, um Ordnung zu erhalten.[Boc74]

Weitere Merkmale beschreibt Broughton im Kapitel „Need for Classification“ in [Bro04, S. 4ff]. Mit der Reihenfolge Gruppieren, Ordnen, Reihenfolge der Attribute festlegen, wird hierbei auf den Ablauf eingegangen. Der erste Schritt entscheidet nach „characteristics of division“. Dies können Eigenschaften und Attribute sein, die „all members of a group have in common“. Darauf folgt das Urteil über „relationship[s] between groups“. Speziell betrachtet werden dabei zusammengesetzte Subjekte, welche mehr als einen Aspekt besitzen. Relevant für Ordnungen sind weiterhin die semantischen Beziehungen zwischen den Begriffen. Einerseits die hierarchischen, „thing-kind“ oder taxonomisch (Beispiel: Käse -> Camembert), „whole-part“ oder partitiv (Europa -> Deutschland) und instantive Beziehungen (Beispiel: Künstler -> Leonardo da Vinci), andererseits nicht hierarchische Bindungen wie die überlappende Zusammensetzung (Beispiel: Papagei und Haustier).[Bro04, S. 25] Auch in der Herangehensweise bestehen Unterschiede. Sucht man etwas, wozu Details bekannt sind, wird dies als „known item retrieval“ bezeichnet. Während dessen bei der Recherche anhand des Inhalts die Rede von „subject retrieval“ ist.[Bro04]

Im Speziellen wird auf die Details der verwendeten Klassifikationen im Abschnitt 4.3.2 eingegangen. Ein Aspekt, der zum nächsten Terminus, den Thesauri überleitet, ist die Verwendung begrenzten Vokabulars in Klassifikationen.

Thesaurus

Die Arbeit mit dem Titel „Vom Nutzen unscharfen Begriffswissens“ [Kra91] drückt in der Überschrift einen Teil der Erklärung des Ansatzes aus. Thesauri oder auch Fachwortschätze stellen Beziehungen zwischen Begriffen her. In Verbindung mit den beschränkten Vokabularen von Klassifikationen können dadurch Probleme im Umgang mit eben diesen vermindert

¹<http://www.askoxford.com/results/?view=dict&freesearch=facet&branch=13842570&textsearchtype=exact> vom 12.12.2007.

werden. Dabei bauen Thesauri ein semantisches Netz auf. „Die Bedeutung eines Begriffes ergibt sich ausschließlich durch seine Beziehung[en] zu anderen Begriffen“.[Kra91, S. 259] Bei den Beziehungen existieren vier Arten. Zum einen Generalisierung und Spezialisierung, andererseits positive sowie negative Assoziation². Diese sind dabei nicht im Sinne von „entweder-oder“ zu verstehen. Vielmehr bilden sie auf Grundlage der „Fuzzy-Set-Theory“³ eine Ähnlichkeitsrelation ab.[Kra91]

Synonyme sind auch im WWW von Bedeutung. Neben der eigentlichen Anfragesprache ist es für den Benutzer wichtig, „richtige Benennung in den Anfragen zu verwenden“.[Kra91, S. 258] Am Beispiel Suchmaschine wird dieser Zustand deutlich. Die Anfrage „Anleitung“ bringt nicht zwingend befriedigende Ergebnisse. So sind neben den Begriffen, die das passende Resultat in der gleichen Sprache (Beispiel: Hilfe, Erklärung, Schema) beschreiben, auch Worte, die in anderen Sprachen mehr oder weniger gut abstecken (Beispiel: engl. tutorial, engl. instruction) was gesucht wird. Dieses Problem wird für diese Arbeit im Bereich der Tags (Abs. 5.4) deutlich. Während sich Autoren und Leser bspw. ein breites Spektrum an Synonymen in der Literatur wünschen, erschwert diese Vielfalt die Suche und den Vergleich.

Web2.0

Der durch Tim O’Reilly geprägte Begriff des Web2.0 kam 2004 auf. Damals wurde die erste Konferenz unter dem Titel „Web 2.0 Conference“ in San Francisco⁴ veranstaltet. Mittlerweile hat sich der Begriff für vielerlei Entwicklungen im Internet durchgesetzt. In dem Artikel „What Is Web 2.0?“[O’R05] vergleicht O’Reilly ältere Konzepte mit ihren Entsprechungen in der nächsten Version. In Anlehnung an „A Pattern Language“ von Christopher Alexander[Ale77] stellt er Entwurfsmuster für das World Wide Web⁵ (siehe Übersicht 4.1.1) auf. Einige Autoren versuchen bereits den Begriff „Web3.0“ zu etablieren. Diese Version soll das von Tim Berners-Lee erdachte „Semantic Web“ darstellen.

1. The Long Tail
2. Data is the Next Intel Inside
3. Users Add Value
4. Network Effects by Default
5. Some Rights Reserved
6. The Perpetual Beta
7. Cooperate, Don’t Control
8. Software Above the Level of a Single Device

²Bedeutungsähnlich vs. gegensätzlich.

³Erlaubt mehr als nur 0 und 1, also auch 0,7 als logischen Wert.

⁴<http://www.web2con.com/web2con/coverage.csp> vom 24.11.2007.

⁵<http://www.oreilly.de/artikel/web20.html?page=3#designpatterns> vom 24.11.2007.

Das zweite Pattern („Data is the Next Intel Inside“) hängt stark mit dem Begriff „User generated Content“ UGC zusammen, welcher ein wesentliches Prinzip der Internetdienste im Web2.0 ist. Ein Diskussionspunkt ist, warum der Nutzer seine Daten und vor allem seine Arbeit kostenlos und in Mengen den Internetseiten anvertraut. Anreize, welche dem Nutzer geboten werden, gibt es demnach unterschiedliche. Im Vortrag von Felix Peters auf der re:publica 2007⁶ werden drei Belohnungen („Geld“, „Reputation“, „Sex“) genannt.[Pet07]

- ... weil der Nutzer es eh tut (Beispiel: Lesezeichen speichern, Bilder für Freunde online stellen)
- ... weil der Nutzer berühmt werden kann (Beispiel: Video hochladen welches sich viele Menschen anschauen und darauf reagieren)
- ... weil sie neue Menschen kennenlernen können und mit Bekannten in Kontakt bleiben können (Beispiel: Soziale Netzwerke)
- ... weil sie dafür bezahlt werden

Welche Daten die Testpersonen dieser Arbeit online zeigen, listet Abschnitt 8.1.

Soziales Netzwerk

Der Begriff „sozial“ („gemeinnützig, wohltätig“ [Dud06, S. 947]) als Attribut eines Netzwerkes deutet auf die Ausrichtung der Online-Gemeinschaften hin. Garton definiert ein soziales Netzwerk SN als „a set of people (or organizations or other social entities) connected by a set of social relationships, such as friendship, co-working or information exchange.“

„Vitamin B“ (mit „B“ als Abkürzung für Beziehung⁷) als ein Vorteil von Netzwerken kommt auch bei Singh zur Sprache. Dabei geht er auf die sozialen Bindungen in Netzen ein und nennt Belege, dass die Menge der einfachen Verbindungen mehr Einfluss als ihre Verbindungsstärke hat. Freemans drei Theorien („point, control and independence“) zur Zentralität von Knoten in Netzwerken betrachten wie Ressourcen in SN verknüpft sein können. Damit lassen sich Aussagen über die Wichtigkeit von Personen (Einzelgänger oder „Netzwerker“) in Online-Gemeinschaften treffen.[Fre79][Sin07]

Ein weiterer Aspekt der SN unterscheidet und auch in Abschnitt 5.3 zur Sprache kommt, ist die Spezialisierung bzw. Generalisierung eben dieser. Iskold identifiziert dabei Beispiele für Netzwerke mit allgemeinem Ansatz (Bsp. Myspace, Facebook) und solche die ein eindeutiges Spezialthema haben (Beispiel: Flickr, LastFM, Delicious). Die Vorteile (auf lange Sicht) der breiter aufgestellten Gemeinschaften benennt er mit dem Bestreben der Nutzer, einen Dienst für alle Aktionen zu verwenden und nicht gleichzeitig mehrere Zugänge aktuell zu halten. Der Nutzen bei den Spezialisten liegt darin, dass diese sich auf ihre Fachgebiete konzentrieren können (Beispiel: mehr Funktionen). Interessant ist auch die wirtschaftliche Betrachtung der möglichen Zusammenarbeit von Spezialisten und Generalisten.[Isk07]

Der Ansatz SN „ähnlich [...] der freiwillige[n] Feuerwehr“ in Krisenfällen einzusetzen, zielt in die Richtung, dass der Anwender im Web2.0 Daten online stellt. So sollen zusammengebrochene Informationskanäle überbrückt werden.[Röt07]

⁶<http://re-publica.de/> vom 25.11.2007.

⁷<http://synonyme.woxikon.de/synonyme/beziehung.php> vom 25.11.2007.

Vom Standpunkt des Datenschutzes und der Privatheit bzw. Offenheit sind SN differenziert zu betrachten. Einerseits hat ein Unternehmen Zugriff auf die Daten aller Nutzer. Dies ist nicht überall so im Fokus der Öffentlichkeit wie in Deutschland. Andererseits haben Nutzer Zugang zu Personenkreisen, mit denen sie bisher nur wenig im realen Leben zu tun hatten. Die Möglichkeiten, sich gegenseitig kennen zu lernen oder auch auszuspionieren, sind um ein Vielfaches zahlreicher als zu der Zeit, als die Sozialen Netzwerke noch nicht bestanden. Andererseits besteht zwischen den Vertretern der Gemeinschaften der Unterschied, ihre Daten auch nach außen frei zugänglich zu machen. Die sogenannten „walled gardens“⁸ stehen offenen Diensten (wie Myspace) gegenüber.[Chi07] Differenziert werden müssen dabei allerdings noch weitere Merkmale (vgl. 5.2.1).

Ein letzter Punkt, der Interessierte mit spektakulären Bildern lockt, ist die Visualisierung von Sozialen Netzwerken. Dabei stehen vor allem Graphabbildungen im Vordergrund. Heer zeigt verschiedene Netzwerkkabbildungen anhand von 1,5 Millionen Nutzer-Profilen aus Friendster⁹. [Hee04]

Taxonomy - Folksonomy - Ethnolocation

Der Aufbau der Categoriesysteme durch Benutzer ist eine Eigenschaft, welche die verwendeten Lexika (Wikipedia und Freebase) eint. Deren Entstehung spiegelt einen Gegensatz zum Ansatz der klassischen Taxonomien (Abs. 4.1.1) wie DDC wieder. Noch deutlicher wird dieser Aspekt allerdings in den Diensten. Hier wird das Konzept des „Tagging“ (Abs. 5.4) verwendet. Die dadurch entstehenden Taxonomien wurden anfangs auch als „Folksonomies“ [Van07] bezeichnet. Merholz spricht von „ethnolocation“, ein Begriff geprägt von Susan Leigh Star, bei dem die Ordnung durch die Nutzer und deren Anbringen von Metainformation an die Objekte entsteht.[Mer04]

So lassen sich zwischen Begriffen synonyme Beziehungen herstellen. Eine kritische Betrachtung liefert Shirky. Die Untersuchung verschiedener Tagging-Ansätzen bietet Boyd et. al. [MNBD06]. Dabei wird als ein weiterer Vorteil das zusätzliche Anfügen von Metadaten erwähnt. Dies bestätigt auch Trant an einer Studie zum Tagging für Museumsgegenstände¹⁰. Neue Begriffe, die zuvor nicht mit dem Objekt verknüpft waren und es leichter auffindbar machen, konnten assoziiert werden. Dabei unterschieden sich die Schlagworte aus „Professional and Public Vocabularies“. Auch die Sichtweise, mit welcher Besucher die Werke sehen, konnte dadurch analysiert werden. Dies war vorher nur schwer möglich.[Tra06]

Ausgehend von der Definition für Ontologien, „an explicit specification of the conceptualization of a domain“ und der Anmerkung, dass diese nicht nur *von* sondern auch meist *für* Mitglieder einer Domäne erstellt wird, beschreibt Mika das Entstehen von Ontologien in Online-Communities. Diese geschieht dabei nicht explizit, indem sich Nutzer auf einen Konsens einigen und Konzepte festlegen. Der Weg den das Tagging hierbei ermöglicht verläuft implizit. Die entstehende Folksonomy lässt sich mit Hilfe der Graphentheorie auswerten.

⁸Nach außen abgeschlossene Dienste wie Facebook oder StudiVZ.

⁹<http://www.friendster.com/> vom 25.11.2007.

¹⁰<http://www.steve.museum/> vom 30.11.2007.

Mashup

Der Überblick bei ProgrammableWeb¹¹ macht die Vielzahl der Möglichkeiten deutlich. Der Dienst listet 2533 Mashups und 555 APIS auf (Stand 24.11.2007). Doch was dahinter steckt, wird deutlich mit der Erklärung aus dem Wörterbuch. Einerseits „crush something [...] until it is soft and smooth“ [Sum95] lenkt bereits in Richtung zerkleinern und vermischen. Dict.cc antwortet mit „etw. vermischen“¹². Am Beispiel GoogleMaps¹³ und der Kombination mit verschiedensten Diensten, die Daten mit Geoinformationen auf den Landkarten darstellen, wird es deutlich. Zu Beginn der technischen Entwicklung waren dabei meist nur zwei Dienste über ihre API verknüpft. Heutzutage sind häufig größere Vernetzungen anzutreffen. Mit OpenSocial¹⁴ von Google oder auch der API zu Facebook¹⁵ sind unzählige neue Möglichkeiten durch Schnittstellen zu nutzerstarken Diensten hinzugekommen.

4.1.2. Zahlen

Wie weit verbreitet die Nutzungsmöglichkeiten und durchdrungen der Markt rund um das Thema Internet sind, klärt dieser Abschnitt. Dabei wird auf lokale und demographische Unterschiede eingegangen. Ob Bloggen, Soziale Netzwerke und Accounts bei „Location Based Services“¹⁶ Phänome der technikaffinen Nerds, „Early Adopters“¹⁷ und Informatikstudenten sind, wird anhand von Statistiken präsentiert.

Unter dem Begriff *Web2.0* können sich in Deutschland sowie den USA, einer Stern.de-Studie zufolge (499 Pers. in D und 502 Pers. in USA, Befragung über das Online-Panel Toluna¹⁸), nur wenige etwas vorstellen, „lediglich zehn Prozent in Deutschland und 20 Prozent in den USA“. Doch kennen „90 % in Deutschland und den USA [...] Web-2.0-Angebote“. Dabei sind in absteigender Reihenfolge Enzyklopädien (Wikipedia), Online-Diskussions-/Informations-Foren und Internet-Video- sowie Internet-Fotoalben vertraut. Die Motivationen hinter der Nutzung ähneln sich zwischen den Märkten ebenfalls. Auf vorderen Plätzen liegen „Austausch mit anderen Nutzern (D: 67 %; USA: 60 %) und gegenseitiges Geben und Nehmen (D: 55 %; USA: 47 %)“. Doch gibt es auch merkbare Unterschiede. „Für Amerikaner ist es wesentlich wichtiger als für Deutsche, 'persönliche' Daten ins Netz zu stellen. 42 Prozent der US-Befragten bekannten sich dazu, im Gegensatz zu 28 Prozent der Deutschen.“ Bei der Nutzung der Dienste liegt Wikipedia in beiden Ländern gleich auf, während Youtube (D: 21 %, USA 40 %), Flickr (D: 6%, USA 17%) und Myspace (D: 11 %, USA 29 %) im amerikanischen Raum weitaus häufiger verwendet werden. Die persönliche Bedeutung der Web2.0-Angebote „für die Befragten ist gering“. [San07]

PricewaterhouseCoopers¹⁹ (PWC) legt ebenfalls Zahlen („basieren[d] auf Interviews mit 501 repräsentativ ausgewählten Haushalten in Deutschland“) vor. Sie stellen überrascht fest, „dass mit dem häufig genutzten Schlagwort Web 2.0 nur 15 Prozent der Befragten

¹¹<http://www.programmableweb.com/> vom 24.11.2007.

¹²<http://www.dict.cc/englisch-deutsch/mash.html> vom 24.11.2007.

¹³<http://maps.google.com/maps> vom 24.11.2007.

¹⁴<http://code.google.com/apis/opensocial/> vom 24.11.2007.

¹⁵<http://developers.facebook.com/> vom 24.11.2007.

¹⁶<http://www.ibm.com/developerworks/ibm/library/i-lbs/> vom 26.11.2007.

¹⁷http://www.zeit.de/2000/36/200036_early_adapters.xml vom 26.11.2007.

¹⁸<http://www.toluna-group.com/de/> vom 26.11.2007.

¹⁹Wirtschaftsprüfungsgesellschaft, <http://www.pwc.de> vom 25.11.2007.

etwas anfangen können“. Die Einflüsse von Demographie (Geschlecht, „Bildungsniveau und Haushaltseinkommen“) werden angemerkt (siehe auch 4.5).[AG07]

Statistiken aus dem (N)ONLINER Atlas 2007 von TNS Emnid („etwa 50.000 Telefoninterviews mit der deutschsprachigen Wohnbevölkerung“) geben Einblicke in die Verbreitung von Breitband-Internetzugängen, wie DSL und dem mobilen Internet(1 %) in Deutschland. Nahezu zwei Drittel aller Deutschen oder „60,2 Prozent der Deutschen surfen im Internet“. Auch auf die Altersstruktur sowie geschlechtsspezifische Unterschiede wird eingegangen.[Göp07]

Die öffentlich-rechtlichen Fernsehanstalten liefern mit der ARD/ZDF-Onlinestudie 2007 (AZOS, „1142 Interviews mit Onlinern und 680 Interviews mit Offlinern“) gleichfalls aktuelles Zahlenmaterial. Den „40,8 Millionen Deutsche[n] ab 14 Jahren [mit] Zugang zur Internet-Welt“ dient dies „vor allem der Informationsbeschaffung“ (72 %), „Unterhaltungsmedium“ ist es nur für relativ wenige (14 %). Im Zusammenhang mit Web2.0 „steht weiterhin der passive Abruf und nicht das aktive Erstellen von Inhalten im Vordergrund“. Dabei steht Wikipedia vor den Videoportalen in der Rangliste. Die für die Arbeit von vordergründiger Wichtigkeit stehende aktive Mitarbeit haben nur wenige Nutzer bereits geleistet („erst 6 Prozent [...] Beitrag für Wikipedia verfasst, 7 Prozent einen Film [...] eingestellt und 2 Prozent [...] Spielfigur in einer virtuellen Welt [...] geschaffen.“). AZOS untersucht sechs Web2.0-Angebotsformen näher. Dies sind „Online-Enzyklopädie Wikipedia“, „Bilder- und Videocommunitys“, „Lesezeichensammlungen (Social Bookmarking[, bleibt eine Randscheinung mit 3 %])“, „Soziale Netzwerke“ (berufliche und private), Weblogs (Anstieg der Besucher von 7 auf 11 %, „zahlreiche Formen, die von einem persönlichen Tagebuch bis zum journalistisch professionell gemachten Watchblog reichen“) und „Virtuelle Spielwelten“. Vor allem Jugendliche nutzen die „Chance zur Eigendarstellung: Bereits jeder zweite Teenager, der private Netzwerke nutzt, hat auch ein eigenes Profil angelegt.“ Speziell zu Weblogs kommen die Autoren zu der Aussage, „Im Vergleich zu anderen Web-2.0-Angeboten sind die Nutzer von Weblogs insgesamt aber durch die ständige Interaktion [, Aktive Nutzung bestehend aus Schreiben und Kommentieren,] in der Blogosphäre engagierter und vergleichsweise aktiv.“²⁰[AZ07]

Teilweise auf Blogs bezogen ist die von FOCUS veröffentlichte Studie „Communications Networks (CN) 11.0.“(24.765 Befragte im Alter von 14 bis 69 Jahren). Danach nutzen „gute zehn Prozent der 33 Millionen Onliner in Deutschland“ Weblogs. Davon „zählen 71 Prozent zu den rein passiven Blognutzern“.[Net07]

Die Blogstudie der Universität Leipzig²¹ identifiziert fünf Typen von Blognutzern („Social Networker“, „Selbstdarsteller“, „Informationssucher“, „Wissensdurstigen“ und „Aktiven Konsumenten“) und bewertet die Relevanz von Weblogs. Die Aussage „Sag mir, wie du bloggst, und ich sage dir, wer du bist“ verdeutlicht den Zusammenhang zum Ansatz dieser Arbeit, die Blogs einer Person mit in die Interessenerkennung einzubeziehen. Die fünf Blognutzertypen treten ungefähr gleich verteilt auf. Die Studie bezeichnet sie als „mehrheitlich investigative Multiplikatoren“, da sie Wissen sammeln, es aktiv weitergeben und viele Kontakte in Netzwerken besitzen. Die Relevanz der Inhalte sehen Blognutzer in der Einzigartigkeit (77 %) und in dem „Insiderwissen [das] an die Öffentlichkeit kommt“ (84 %).[ZB07]

Von einem Vergleich der Statistiken sieht die Arbeit ab. Da sich die Berechnungen der

²⁰<http://www.ard-zdf-onlinestudie.de/> vom 26.11.2007.

²¹<http://www.blogstudie2007.de/> vom 26.11.2007.

Anteile auf verschiedene Ausgangsbasen beziehen, sind Differenzen möglich. Ein Beispiel dafür ist die Menge der „Onliner“. FOCUS (Abschnitt 4.1.2) gibt diese mit 33 Millionen an, während die ARD/ZDF-Onlinestudie die Zahl bei 40 Millionen sieht.

4.2. Online-Identitäten

Die Daten, die Personen im Laufe der Zeit an unterschiedlichen Plätzen im Internet hinterlassen, bilden ein weiteres Ich. Hülsmann beschreibt dies mit der Bemerkung, „doch mit zunehmender Dauer der Teilnahme an virtuellen Interaktionen können Menschen Online-Identitäten entwickeln“.[Hül00a, S. 97] Um von einer virtuellen Identität zu sprechen, muss eine Abgrenzung zum realen Ich geschehen. In der Philosophie haben sich mit diesem Thema bekannte Namen beschäftigt.

Gegenüber den reinen Personendaten[Kur07] muss zunächst eine Abgrenzung erfolgen. In dieser Arbeit geht es nicht vordergründig um Alter, Geschlecht oder andere demographische Teile der Daten. Diese haben zwar Einfluss auf die Interessen, werden aber für die Analyse nicht herangezogen.

4.2.1. Identität

Bereits Locke (1632-1704) eröffnete die Diskussion „mit seinem kurzen Text *Über Identität und Differenz*“. Meuter diskutiert dies und weist auf den Ansatz Lockes zur Identität hin. „Um festzustellen, worin die Identität der Person besteht, [müssen wir] zunächst untersuchen, was Person bedeutet“. Die beiden Eigenschaften Selbstbetrachtung und Vernunft dienen als Unterscheidung zu anderen Wesen. Personale oder auch zeitübergreifende Identität definiert sich über „ein individuelles menschliches Leben“ geprägt durch „die verschiedenen Lebensphasen und [...] unterschiedliche[n] Situationen, die wir erleben“.[Meu95, S. 9f] Konkret auf die Fragen der virtuellen Identitäten und dem Bezug zum realen Abbild einer Person wird Luhmann zitiert. Die Zerlegung „in mehrere Selbsts [...] um der Mehrheit sozialer Umwelten und der Unterschiedlichkeiten der Anforderungen gerecht werden zu können“, stellt ein Problem dar. Als Beispiele für Selbsts dienen „musikalische[s] [...] für die Oper [oder] strebsame[s] [...] für den Beruf“.[Meu95, S. 223]

MacLuhan sagte schon 1966 Medien mit der Interaktivität des Internets voraus[McL03, S. 98] und erklärte das Verschmelzen mit den Worten „the world [...] becoming a global village“[McL03, S. 265]. Er fasst Identität genereller, indem er sie darauf bezieht, wohin sich Menschen zugehörig fühlen, womit sie sich identifizieren. Er sieht sogar einen Grund für Gewalt darin begründet und beschreibt dies als den „quest for identity“.[McL03, S. 266]

4.2.2. Identitätsmanagement

Ein spezielles Thema im Bereich der Identität umschreibt die Verwaltung eben dieser. Unter dem Titel Identitätsmanagement fassen Unternehmen Konzepte auf, welche „der Identifizierung, der Authentifizierung und der Transaktionsabwicklung in der Online-Welt“[DSH07] dienen. Im Bereich des Datenschutzes assoziiert es „die personenbezogenen Daten unter der Kontrolle des Nutzers zu belassen“. Neben den kommerziellen Produkten von Microsoft

(CardSpace)²² oder der Liberty Alliance²³ existieren auch in der Open Source-Gemeinde verbreitet Ansätze wie OpenID.

Ziel aller Systeme ist es, die vertraulichen Daten beim Nutzer zu belassen und nur die für Transaktionen notwendigen Details zu den Diensten zu übertragen. Das Bild der Visitenkarte (und ihrer Daten) dient dabei als Vergleich. Nur Daten zur Identifizierung werden weiter gegeben. Fakten zur Autorisierung verbleiben beim Nutzer. Vorstellbar ist auch, „dass eine nicht mehr benötigte Lieferadresse oder Bankverbindung automatisch im System eines Händlers gelöscht wird.“ Dienste (Anonymisierungsdienst), die die Kapselung der Daten gewährleisten sollen, könnten auch in Verbindung mit Dritten „als Schiedsgericht bei Streitigkeiten dienen“.[Kre07]

Den derzeit ausführlichsten Bericht zum Identitätsmanagement („Verkettung von Identitäten“) bespricht Abschnitt 4.2.4.

4.2.3. Online-Identitäten und Beweggründe dafür

Die Motivation für Personen, Daten im Internet zu veröffentlichen, hat neben der vordergründigen sozialen Komponente (Freunde kennenlernen, Freundschaft pflegen), verschiedene Hintergründe (Vgl. 4.1.1).

Ein Teil der Online-Identität dreht sich um die Reputation („Ruf, Ansehen“[Dud06, S. 826]) einer Person. Dabei ist das Image im Internet von zwei Richtungen her interessant. Einerseits möchte niemand negativ auffallen und mit Alkohol, Drogen und anderen nachteiligen Fakten in Zusammenhang gebracht werden. Andererseits besteht das Bestreben, sich möglichst positiv zu präsentieren. Dabei kommen vorteilhafte Interessen, Schulabschlüsse oder Aktivitäten in Betracht. Beide Bereiche decken die Dienste ab, welche sich um die Online-Reputation kümmern (Abs. 3.1.6). Vor allem im Bereich der Bewerbungen und Jobsuche sehen diese Aktivitäten ihren Ansatz. Da ein Teil der Personalabteilungen in Umfragen zugibt, Suchmaschinen zu nutzen, um Bewerber zu überprüfen (77 %), 10 % stöbern auch direkt in Sozialen Netzwerken. Suchmaschinenstatistiken sagen aus, dass es 25 bis 50 Millionen Namenssuchen pro Tag gibt.[Zel07]. Dabei trägt die positive Beeinflussung des virtuellen Ansehens zu Abbildung von Interessen bei. Wie schon oben erwähnt, sind dies hier bspw. Aktivitäten, Hobbys oder interessierende Themen einer Person, welche sich in den Profilen wiederfinden.

4.2.4. Daten aus Online-Identitäten

Einen sehr ausführlichen Überblick, was bei der Verkettung von Identitäten möglich ist, veröffentlichen das unabhängige Landeszentrum für Datenschutz in Schleswig-Holstein und der Forschungsbereich Datenschutz und Datensicherheit der TU Dresden. Auf mehr als 230 Seiten werden die Fragen „Wo werden welche Daten über mich erhoben? Wer kann sie miteinander verknüpfen, wo werden Profile über mich erstellt? Wie lassen sich diese verketteten Daten auswerten?“ beantwortet. In vier Anwendungsszenarien ²⁴ zeigen die Autoren auf, dass durch die „Verkettung von anonymen Profilen“ enorme Gefahrenpotentiale entstehen.[Bor07][TUD07]

²²<http://cardspace.netfx3.com/> vom 21.11.2007.

²³<http://www.projectliberty.org/liberty/about> vom 21.11.2007.

²⁴ „Überwachung mit Hilfe von Alltagsgegenständen“, „Internet-Suchmaschinen“, „Arbeitnehmer und ortsbezogene Dienste“, „Ambient Assisted Living“

4.3. Modell eines Interessenprofil

Die Interessen einer Person bilden deren Interessenprofil. Dabei kann nicht von einer eindeutigen Beschreibung eines Menschen ausgegangen werden. Einerseits existiert kein Standard, der die Interessen einheitlich regelt. Andererseits sind die Vorlieben einer Person abhängig vom aktuellen Geschehen und anderen Einflüssen (Zeitpunkt, Ort, Freunde, Umfeld). Dieser Abschnitt beschreibt das dieser Arbeit zu Grunde liegende allgemeine Interessenprofil und schildert Details zu den Quellen, welche zur Erstellung beigetragen haben. Diese sind zum einen zwei Internet-Lexika (Wikipedia englisch und deutsch²⁵), Freebase²⁶ sowie die Kategorisierung der Nachrichten bei dem Dienst Digg.

4.3.1. Entstehung des allgemeinen Interessenprofils

Da das Profil nicht nur für eine Person starken Schwankungen ausgesetzt ist, sondern auch Kontext abhängig ist, kann nur oberflächlich vom „Interessenprofil“ als etwas klar Abgegrenztem gesprochen werden. Dahinter stehen viel mehr Schwerpunkte in Interessengebieten. Zusätzlich zu diesem Hindernis kommt die Hürde der vielfältigen Benennungen und Zuweisung zu Oberkategorien. Der Versuch, disjunkte Interessenschwerpunkte zu klassifizieren, ergibt die sieben Kernbereiche Kultur, Entertainment, Wirtschaft, Gesellschaft, Sport, Technik, Wissenschaften. Die Haupt- und Unterkategorien der Lexika sowie von Digg dienen als Quelle.

Vergleicht man die Rubriken mit den in Abschnitt 4.4 besprochenen Segmenten, lässt sich teilweise eine Zuweisung herstellen. Am Beispiel der SevenOne-Studie (Abs. 4.4.3) sieht man, dass diese nur fünf Hauptkategorien aufweist und bei diesen auch eine leicht abweichende Einordnung vornimmt (Bsp. Technik und Wissen zusammengefasst, Unterkategorien anders verteilt).

4.3.2. Die Klassifikationen der Lexika

Aus den oben beschriebenen Datenbanken lassen sich vier Klassifikationen entnehmen. Die kleinste und am wenigsten differenzierte kommt von Digg. Sie ist auf den Anwendungsfall „Einordnung einer Internetadresse“ angepasst. Die acht Hauptkategorien besitzen vier bis zwölf Unterkategorien. Der Schwerpunkt liegt hier auf Themen aus dem amerikanischen Raum. Dementsprechend ist die Klassifikation in Englisch ausgeführt.

Die existierenden Wikipedia-Ableger besitzen alle ein Categoriesystem, welches aufgrund der Datenmenge der jeweiligen Sprache entsprechend mit Unterordnungen gefüllt ist.

Sowohl Wikipedia als auch Freebase verwenden als größte Struktur nicht die Hauptkategorien. WiP setzt darüber die *Portale*²⁷ (Geographie, Geschichte, Gesellschaft, Kunst, Religion, Sport, Technik, Wissenschaft) und gruppiert darunter die *Rubriken*. FB bezeichnet die unschärfste Stufe oberhalb der *Domänen* als Kategorien²⁸.

Die Abhängigkeit der Ergebnisse der Einordnung vom Datenstamm des jeweiligen Lexikons wird in Abschnitt 9.1 weiter ausgeführt. Theoretisch besteht die Möglichkeit weitere

²⁵<http://wikipedia.org/> vom 27.11.2007.

²⁶<http://www.freebase.com/view/> vom 27.11.2007.

²⁷http://de.wikipedia.org/wiki/Portal:Wikipedia_nach_Themen vom 28.11.2007.

²⁸<http://www.freebase.com/view/allDomains> vom 28.11.2007.

Lexika in anderen Sprachen oder mit speziellen Themengebieten in die Analyse einzubeziehen. Auch andere große und allgemeine Enzyklopädien existieren (Beispiel: DMOZ²⁹ oder Cyc/OpenCyc³⁰), welche die Ergebnisse verfeinern könnten. Dies wird im Rahmen dieser Arbeit nicht weiter verfolgt. Einen Überblick über zusätzliche Möglichkeiten dabei bietet Abschnitt 9.2.3. Die Verfahren der Zuordnung von Tags zu den Kategorien bespricht detailliert Abschnitt 7.2.1.

4.3.3. Begründung der Verwendung von Lexika

Den breitesten Ansatz einer Ordnung für Interessen bieten allgemeine Lexika. Im Zeitalter des Buches waren sie bestrebt, ein vollständiges Abbild der Welt wiederzugeben und damit dem Leser die Möglichkeit zu bieten, aus nahezu allen Bereichen Wissen zu erlangen. Den Versuch, ein vollständiges Bild der Welt zu zeichnen, führen die Enzyklopädien im Internet fort.

Die in dieser Arbeit als Grundlage der Einordnung von Begriffen verwendeten Lexika *Wikipedia* und *Freebase* bieten einen reichhaltigen Bestand an Wissen. Der Ansatz beider Dienste, den Inhalt vom Benutzer erstellen zu lassen, sorgt dafür, dass Interessen der Nutzer breitgefächert vorhanden sind. Während Wikipedia die Qualitätssicherung nahezu vollständig den Nutzern überlässt, steht bei Freebase die Firma Metaweb im Hintergrund.

Neben dem eigentlichen Softwaresystem, welches die Dienste anbieten, sind zwei Bereiche von Bedeutung. Zum einen die reinen Daten wie Texte, Zahlen oder Bilder. In Wikipedia sind diese in unzähligen Überarbeitungen im Laufe der Lebenszeit des Dienstes entstanden, diskutiert und überarbeitet. Hierbei haben Nutzer in unterschiedlichen Rollen differenzierte Rechte. In Freebase ist es ebenfalls möglich, allerdings nur nach Registrierung, Daten zu erstellen und zu verändern. Die Inhalte stammen aber hierbei auch aus der Wikipedia und anderen lizenzfrei zugängigen Quellen. Der zweite Bereich kann mit dem Begriff Struktur betitelt werden. Hierunter fällt die Ordnung, die die Daten in Kategorien einteilt. Bei Wikipedia hat der Nutzer auch hier freie Hand und in Diskussionen wird versucht, einen Konsens zu finden. Bei Freebase wird stärker Wert auf die Struktur gelegt. Dies zeigt sich daran, dass nicht die Daten im Vordergrund stehen sondern die Ordnung. Ein Grund dafür ist die semantische Ausrichtung der Datenbank. Auch Operationen (Transitivität) die der Struktur beitragen sind in Planung.

Ein weiterer Vorteil, derzeit nur für Wikipedia praktisch nutzbar, ist die Vielsprachigkeit. 16 Wikipedias mit über 100.000 Artikel³¹ bieten reichhaltigen Inhalt für die Einordnung von Begriffen. Im Rahmen dieser Arbeit wird nur die deutsch- sowie die englischsprachige Version³² betrachtet (vgl. Thesauri 4.1.1 und Sprache von Tags 4.4.6 als weitere Verwendung der Lexika).

4.4. Segmentierungsmodelle

Eine Einteilung der Menschen in Gruppen ist neben den soziodemographischen Ansätzen auch auf Grundlage anderer Aspekte üblich. Dies geschieht dabei nicht erst seit es Untersu-

²⁹<http://www.dmoz.org/> vom 28.11.2007.

³⁰<http://www.cyc.com/> vom 28.11.2007.

³¹<http://de.wikipedia.org/wiki/Wikipedia:Sprachen> vom 28.11.2007.

³²Dt. Wikipedia 661.000 Artikel, engl. Wikipedia 2.076.000 Artikel (Stand 16.11.2007).

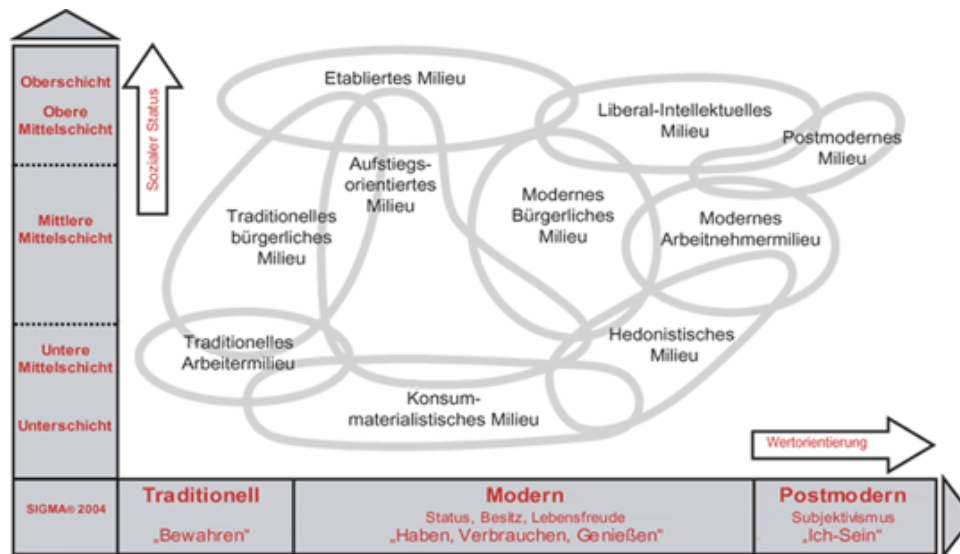


Abbildung 4.1.: SIGMA Milieus in Deutschland³⁹

chungen zum Nutzungsverhalten im Internet gibt. Exemplarisch stehen die „Sinus-Milieus, bei denen Menschen, die sich hinsichtlich ihrer Lebensauffassung und Lebensweise ähneln [...] oder der Semiometrie-Ansatz, bei dem Zielgruppen hinsichtlich ihrer soziokulturellen Werte beschrieben werden.“ [OFG07, S. 7] Nicht ausführlich betrachtet wird hier die Segmentierung im Rahmen der ARD/ZDF-Onlinestudie 2007 aus Abschnitt 4.1.2.

4.4.1. Sinus-Milieus

Mit der für den deutschen Markt auf zehn Milieus (siehe Abbildung 4.1) beschränkten Einteilung nach den Dimensionen Wertorientierung und sozialem Status, bietet die Gesellschaft für internationale Marktforschung und Beratung³³ einen Ansatz zur Segmentierung. Beispiele darin sind das „Postmoderne Milieu“³⁴, „Sie sind selbstbewußte Lifestyle-Architekten, die sich ohne Bauanleitung aus ihrem individuellen ‚construction kit‘ einen Lebensstil nach ihrem persönlichen Maß schneiden“, Toleranz von Widersprüchen, multiple Identitäten³⁵ oder auch das Konsum-materialistisches Milieu³⁶, umschrieben mit „Milieu der wirtschaftlich und sozial Randständigen mit geringen Chancen am Arbeitsmarkt nachindustrieller Gesellschaften: alte wie auch neue Armut“³⁷. Mit dem Fragebogen „Target It!“³⁸ kann man sich persönlich nach seinen Aussagen zu bestimmten Themen einordnen lassen.

³³<http://www.sigma-online.com> vom 23.11.2007.

³⁴6,9 % der Wohnbevölkerung 16 und älter

³⁵http://www.sigma-online.com/de/SIGMA_Milieus/SIGMA_Milieus_in_Germany/Postmodernes_Milieu_/# vom 23.11.2007.

³⁶10,5 % der Wohnbevölkerung 16 und älter

³⁷http://www.sigma-online.com/de/SIGMA_Milieus/SIGMA_Milieus_in_Germany/Konsummaterialistisches_Milieu/# vom 23.11.2007.

³⁸http://target-it.sigma-online.com/Target_It/ vom 23.11.2007.

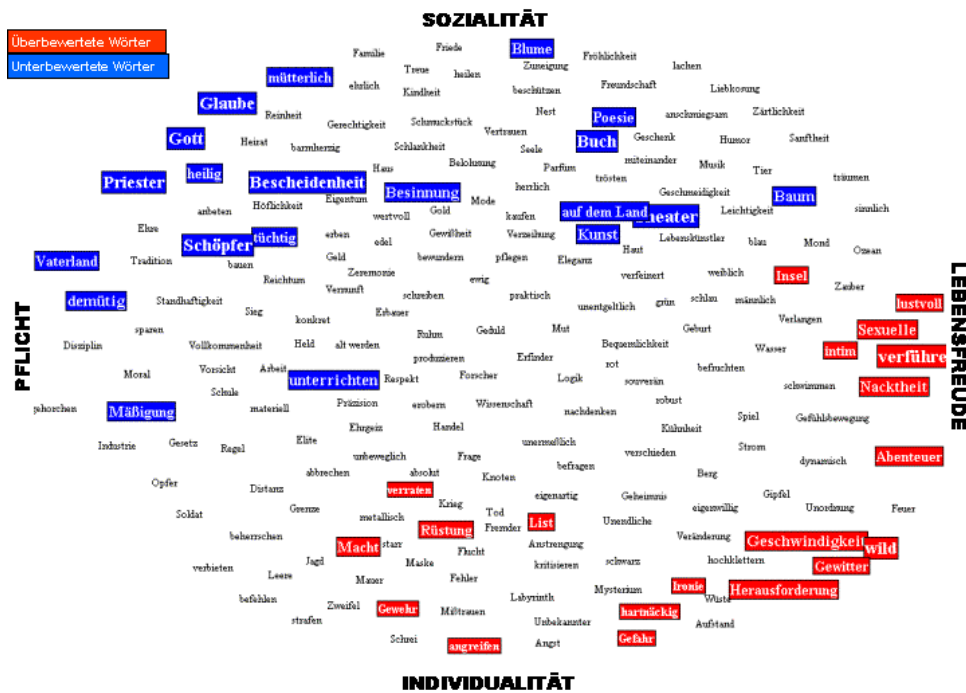


Abbildung 4.2.: Beispiel einer Markenpositionierung⁴³ mit Infratest

4.4.2. Semiometrie-Ansatz

Dieser Ansatz von TNS Infratest⁴⁰ nutzt 210 Begriffe die der Befragte bewerten muss und anhand seiner Einschätzung (sehr unangenehm bis sehr angenehm) in die vier Dimensionen Pflicht, Sozialität, Lebensfreude, Individualität eingeordnet wird (Beispiel siehe Abbildung 4.2). Der Fragebogen⁴¹ für ausführliche Bewertung sowie die Möglichkeit eines persönlichen Semiogramm⁴² stehen auch hier zur Verfügung.

4.4.3. @facts Online-Nutzertypen 2007

Speziell für den Onlinemarkt wurde von Forsa im Auftrag von SevenOne Interactive die Studie „@facts Online-Nutzertypen 2007“ erstellt. Diese Firma ist Mitglied in AGOF⁴⁴. Sie werben mit Phrasen wie „Wo surfen Sportfans, die ein Auto kaufen wollen?“⁴⁵ (siehe dazu auch Abschnitt 1.1).

Die Studie unterteilt die Anwender im Internet in sieben „verschiedene möglichst homogene“ Gruppen (siehe Übersicht 4.4.3), welche durch ein „clusteranalytisches Verfahren“ ermittelt wurden. Innerhalb eines Types „ähneln sich [die Nutzer] also hinsichtlich ihrer Interessenschwerpunkte bei der Internetnutzung“. Die Fokuse der Vorlieben wurden mit

⁴⁰ http://www.tns-infratest.com/02_business_solutions/02017_Semiometrie.asp vom 23.11.2007

⁴¹ http://www.tns-infratest.com/pdf/business_solutions/Semiometrie-FB-Basis2007.pdf vom 23.11.2007.

⁴² http://www.tns-infratest.com/02_business_solutions/02017c_Semiogramm.asp vom 23.11.2007.

⁴⁴ Arbeitsgemeinschaft Online-Forschung e.V. <http://www.agof.de/> vom 23.11.2007.

⁴⁵ <http://www.agof.de/agof.3.html> vom 23.11.2007.

sechs Kategorien bei den Inhalten (News, Lifestyle, Technik & Wissen, Entertainment, Sport & Auto, Freizeit) und fünf Klassen bei den Angeboten (Mediaanwendungen, Kommunikation, Service & Shopping, Spiele, User generated Content) erkannt. Einige diese Kategorien finden sich in dem Modell des Interessenprofils dieser Arbeit wieder.[OFG07]

- Multi-Interest & User generated Content (Typ 1)
- Entertainment & Communication (Typ 2)
- Fun & Game (Typ 3)
- Music & Video (Typ 4)
- Free Time Planning (Typ 5)
- Service, Shopping & Lifestyle (Typ 6)
- Low Interest (Typ 7)

4.4.4. Modell des Allgemeinen Interessen-Struktur-Test

Stiftung Warentest⁴⁶ listet 23 „Onlineverfahren zur Selbsteinschätzung“ (14 für Erwachsene und neun für Jugendliche)⁴⁷. Diese zielen dabei aber mehr auf die berufliche Ausrichtung.

Im Allgemeinen Interessen-Struktur-Test (AIST-R) als Beispiel wird mit Hilfe eines Interessenfragebogens zur Erfassung schulisch-beruflicher Interessen eine Person in sechs Dimensionen⁴⁸ eingeordnet. „Diese sechs Typen stehen zueinander in einer kreisförmigen Beziehung, was sich auch in den Bevorzungen äußert. Meist werden im Profil zwei oder auch drei Bereiche dominieren, die nebeneinander liegen“. Auch für den Freizeitinteressenbereich findet sich ein angepasster Test⁴⁹.

Zum AIST-R kann anhand der „Latent Class Analyse“ (LCA) aufgezeigt werden „inwieweit typische Interessenprofile der Gesamtstichprobe [...] identifizierbar sind“.[TD96, S. 2] Dabei wird der Test durch Normierung und Klassifikation geeicht. Um die Aussagekraft der Ergebnisse zu bewerten, wird jede Einordnung einer Person mit der erwarteten Wahrscheinlichkeit verglichen („Männer eher in Bereich R, Frauen eher in A und S, weniger Unterschied in I und C, in E keine Geschlechterunterschiede“[TD96, S. 4]) (Abkürzungen siehe A.3).

4.4.5. Ansätze zur Einteilung aus der Werbeindustrie

Hohe Relevanz am Interessenprofil von Personen und der Einteilung in Interessengruppen hat die Werbeindustrie. Hauptauschlaggebend dabei ist zielgerichtete Werbung. Die folgenden Passagen nennen Beispiele für Anwendungen des Zusammenspiels von Werbung und Interessenanalysen. Google und Facebook sind die gewichtigsten Vertreter.

⁴⁶<http://www.test.de/unternehmen/> vom 21.11.2007.

⁴⁷<http://www.test.de/themen/bildung-soziales/weiterbildung/test/-Onlinetests-zur-Selbsteinschaetzung/1493119/1493119/1496244/> vom 21.11.2007.

⁴⁸Konventionell, unternehmerisch, intellektuell, realistisch, sozial, künstlerisch.

⁴⁹<http://www.stangl-taller.at/STANGL/WERNER/BERUF/TESTS/FIT/> vom 21.11.2007.

Google AdWords und AdSense

Um Internetnutzern mit speziell auf die Person abgestimmten Einblendungen Botschaften zu vermitteln, wird der Umweg über Segmentierungsmodell bei Google vermieden. Die Firma beschränkt sich mit ihren Produkten AdWords⁵⁰ und AdSense⁵¹ darauf, aus Schlagworten zu erkennen, welche Werbung eingeblendet werden soll. Der Werbetreibende gibt bei AdWords Stichworte vor, für die seine Anzeige geschaltet werden soll. Auf der Gegenseite bietet AdSense die Möglichkeit, Werbeplatz anhand des Inhalts (relevante Stichworte) zu vermarkten.

4.4.6. Daten die Nutzer unterteilen

Neben den reinen Segmentierungsmodellen lassen sich Personen im Internet auch anhand weiterer Merkmale gruppieren. In Sozialen Netzwerken (oder allg. Web2.0-Quellen) sind dies neben Eigenschaften der Person wie Sprache, Alter, Bildung oder Beruf auch ihre Interessen und Kontakte. Ein weiterer differenzierender Bereich sind aus den Eigenschaften resultierende und darüber hinausgehende Gruppenzugehörigkeiten. Einige Beispiele dafür sind in der Übersicht 4.4.6 aufgelistet. Die Gruppen haben aufgrund ihres Fokus unterschiedliche Größen, so ist verständlicherweise die Anzahl der „Schüler vom Lessing-Gymnasium“ kleiner als die der spanisch sprechenden Personen.

- Auf ein Land oder eine Sprache lokalisierter Dienst (Youtube)
- Einteilung der Nutzer anhand ihres Wohnortes (Qype⁵²)
- Neben StudiVZ (Studenten) existiert auch SchülerVZ
- „Networks“ bei Facebook (Beispiel: FU Berlin, Germany)
- Gruppen wie „Nudeln machen ist auch kochen!“ bei StudiVZ
- Personen, die gleiche Applikationen in Diensten nutzen (Beispiel: „Cities I've Visited“ bei Facebook)
- Foto-Themen-Gruppen bei Flickr (Bsp. Panoramas⁵³)

Sprache

Englisch als Sprache des Internets und der fortschrittliche amerikanische Markt bieten im Bereich des World Wide Web ein starkes Zugpferd für alle englischsprachigen Dienste. Zahlreiche Start-Up-Unternehmen zielen auf dieses Absatzgebiet. Auch die schiere Größe des bereits bestehenden Nutzervolumens sorgt für hohe Beteiligungen.

Vor allem Menschen mit guter Bildung und wenigen Ressentiments gegenüber fremden Sprachen nutzen Dienste, die nicht in ihrer Muttersprache verfügbar sind. All jene die sich hierbei schwerer tun, warten eher ab bis, lokalisierte Dienste zugänglich sind.

⁵⁰<http://adwords.google.de/> vom 24.11.2007.

⁵¹<https://www.google.com/adsense> vom 24.11.2007.

⁵²<http://www.qype.com/> vom 21.11.2007.

⁵³<http://flickr.com/groups/52241735229@N01/> vom 28.11.2007.

Dies wirkt sich auch auf die nationale Verbreitung aus. In Ländern, in denen traditionell die englische Sprache bereits von klein auf präsent ist, siehe skandinavische Nationen (keine separate Übersetzung von Filmen höchstens Untertitel), fällt es den Menschen leicht, die englischsprachigen Dienste zu nutzen. Allerdings liegt dies meist auch daran, dass es sich nicht lohnt, für den überschaubaren Kundenkreis eine Übersetzung anzubieten⁵⁴. Auch ein Grund, warum Menschen in größeren Nationen und damit größeren Märkten, siehe Beispiel Deutschland oder Frankreich, nicht so leicht fremdsprachige Angebote annehmen. Die Erwartungshaltung für eine lokalisierte Version ist größer, wiederum aber auch die Selbstverständlichkeit alles Englische zu verstehen geringer.

4.5. Zusammenhang zwischen Interessen und Nutzergruppen

Nachdem die möglichen Ergebnisse der Klassifikation von Interessen (Abs. 4.3.2) anhand der Lexika und Beispiele für geordnete Nutzergruppen aufgezeigt wurden, sollen diese beiden Konzepte verknüpft werden. Der folgende Abschnitt soll die Frage beleuchten, ob einerseits die Interessen eines Nutzer Aussagen über seine Zugehörigkeit zu einer Nutzergruppe zulassen. Andererseits inwiefern Anwender mit gleichen Eigenschaften auch gleiche Vorlieben haben. Es muss dabei darauf geachtet werden, dass neben soziodemographische Faktoren auch soziokulturelle Eigenschaften einen Einfluss auf die Gruppierung haben (siehe auch 4.1.2).[OFG07, S. 7][Hül00a, S. 105]

Die Schlussfolgerung aus den Annahmen würde es Betreibern leichter machen. Einerseits könnten sie von unkompliziert erfassbaren, demographischen Fakten auf die Interessen ihrer Nutzer schließen. Andererseits besteht die Kontrollmöglichkeit, ob die in einem Dienst vorhandenen Interessen die statistischen Daten bestätigen und auf bestimmte Nutzergruppen hinweisen. Ebenfalls zu ergründen wäre, ob eine Nutzergruppe sich für neue Themen interessiert. So könnten Zeitschriftenverlage die Magazine für Jugendliche und Kinder veröffentlichen, über die Interessen eben dieser Auskunft erhalten und ihre Angebote dementsprechend anpassen.[Hom06] Auch für andere beschreibbare Gruppierungen ist dies denkbar, solange sich ihre Interessen in Zusammenhang mit demographischen oder kulturellen Eckdaten untersuchen lassen (vgl. 1.1).

4.5.1. Interessen von Nutzern mit zugehörigen Nutzergruppen

Die Entwicklung des Interessenspektrums soll als Beispiel für den ersten Fall herangezogen werden. Das Institut für Demoskopie Allensbach untersucht dieses in [DA93]. Dabei steigt die Anteilnahme am Politikgeschehen mit zunehmendem Alter[DA93, S. 3f]. Zwischen Frauen und Männern ist allerdings ein Unterschied erkennbar. Ebenfalls Einfluss hat die Herkunft (neue bzw. alte Bundesländer). Mit Statistiken zum „Interesse an Wohnen und Einrichten“, „gesunder Ernährung, gesunder Lebensweise“, Literatur, Sport und der „berufliche[n] Weiterbildung“[DA93, S. 16ff] werden Zahlen aufgezeigt, welche den Zusammenhang zwischen demographischen Faktoren und den Interessenschwerpunkten belegen.

⁵⁴Siehe Untertitel bzw. Kinofilmübersetzungen

4.5.2. Nutzergruppen mit zugehörigen Interessen von Nutzern

Alter, Geschlecht, Einkommen und Herkunft als Kriterien der Demographie können Auswirkungen auf die Zugehörigkeit einer Person zu einer Gruppe haben, wie es Fall zwei annimmt. Beispiel dafür ist Typ 2 (Entertainment & Communication). Mit fast zwei Dritteln (63,2 %) liegt hier der höchste Frauenanteil. Typ 3 (Fun & Game) und 4 (Music & Video) sind besonders ausgeprägt bei den Jugendlichen (bis 19 Jahre). Der Schulabschluss Abitur/Studium ist in der Gruppe Free Time Planning (Typ 5) ausschlaggebend mit 46,7 %. Einzelne Interessen sind dabei nicht anhand von demographischen Faktoren zu erkennen.

5. Erstellung eines Interessenprofils

Im vorhergehenden Kapitel wurde das Modell und Daten in den Profilen besprochen. Nun folgt die Auswahl der Quellen und deren Bewertung, darauf die Priorisierung und Einordnung nutzerbezogener Daten. Neben den in der Implementierung enthaltenen praktisch umgesetzten Methoden kommen hierbei auch die theoretischen Ansätze zur Sprache. Prozeduren, die mit Mehraufwand weitere Fortschritte bringen können, werden betrachtet. Der Bewertung der Ergebnisse im Vergleich mit den realen Interessen von Testpersonen widmet sich Kapitel 8.

5.1. Allgemeine Quellen im Internet

Die „Web Trand Map 2007/V2“ (siehe Abb. 5.1) liefert einen Ausschnitt der relevanten Dienste im Internet. Die Auswahl der Linien gibt einen Anhaltspunkt, welche Rubriken vertreten sind. Einen zweiten Abriss über Dienst mit Nutzerprofilen bietet ShowYourself¹ (Abb. A.6 mit 25 Quellen). Die für die Arbeit ausgewählten Quellen (siehe Überblick 5.2) sind dabei so gewählt, dass eine für Interessen vielfältige Mischung die Analyse beeinflusst. Da im Internet nicht alle Themen gleichmächtig repräsentiert werden und viele Spezialitäten nur in Nischen ihren Platz haben, wurden auch allgemeine Dienste gewählt. Bei denen hat der Nutzer entsprechende Freiheiten sich individuell zu informieren bzw. Daten zu speichern. Eine Bewertung der Eigenschaften bespricht Abschnitt 5.3.1.

5.1.1. Auswahlkriterien für Quellen

Aus der Vielzahl aller virtuellen Quellen mit den gespeicherten Interessendaten wurde eine Auswahl getroffen, welche folgende Kriterien erfüllen musste. Dabei ist ein Problem für die Erkennung der Interessen, dass es zu den Themen jeweils mehr als einen Wettbewerber auf dem Markt gibt. Dadurch sind auch die Nutzer verstreut. Diesem Problem widmet sich Abschnitt 9.2.3.

API-Zugriff

Zwar bieten viele Dienste ihre Daten zu Nutzern frei im Netz an, doch ist der Zugriff darauf nur über den Browser möglich. Einige Quellen verfügen zusätzlich über eine Programmierschnittstelle (API), welche den Zugriff über JSON oder REST ermöglicht (siehe 5.2.1).

Freie Verfügbarkeit von Daten

Desweiteren sind die Daten nicht in allen Diensten für jedermann verfügbar. Restriktive Dienste geben Informationen nur nach Anmeldung am System preis. Dazu kommt die

¹<http://www.dbachrach.com/showyourself/> vom 3.12.2007.

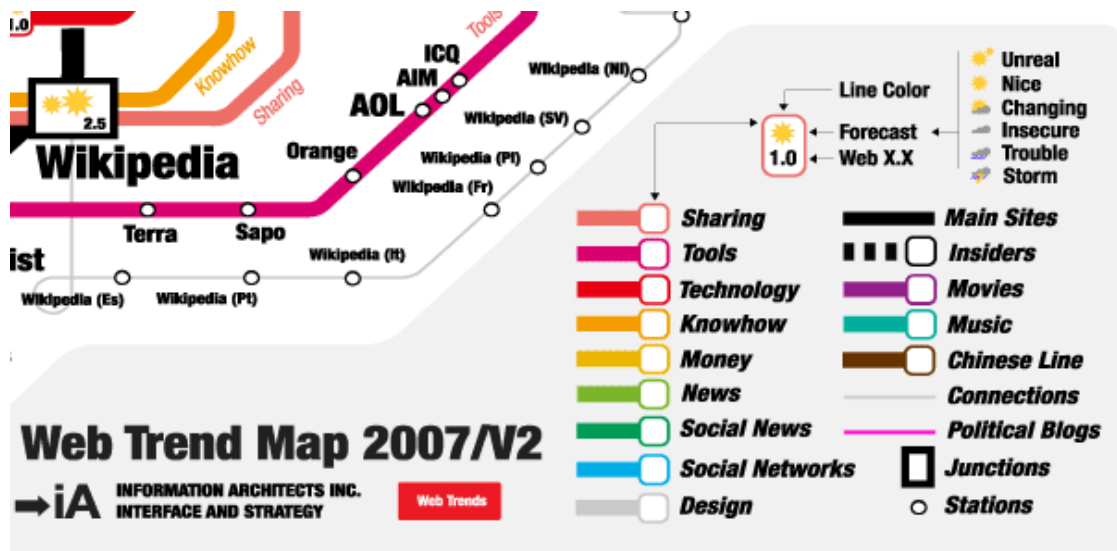


Abbildung 5.1.: Rubriken im Internet, Ausschnitt aus WebTrendsMap

http://www.informationarchitects.jp/slash/iA_WebTrends_2007_2_1024_768.gif vom 01.12.2007

Möglichkeit für Anwender nur Freunden (authorisierten Personen, Gruppen) Zugang zu gewähren (Beispiel: Facebook). Denen gegenüber stehen Dienste, die direkte Adressen zu den Nutzerprofilen anbieten (in der Form `www.dienst.com/nutzername` o. ä., siehe [Kur07, S. 34]). Eine Alternative dabei ist, die Nutzernamen zu kodieren. Anstelle des Loginnamens wird die ID, häufig eine Buchstabenanzahlkombination, verwendet.

Auch bei den Diensten welche Zugriff via API erlauben, bestehen Unterschiede welche Informationen nach außen gegeben werden. Einerseits sind dies nur Daten zum identifizierten Nutzer, andererseits auch Daten anderer Anwender im Zusammenhang mit den eigenen Daten. Seltener bietet die API Zugriff auf Objekte anderer Nutzer.

5.2. Verwendete Quellen

Interessen verteilen sich aufgrund ihrer Vielseitigkeit auf eine große Anzahl an Diensten. Daher ist ein breites Spektrum an Lieferanten notwendig, um alle Facetten der Vorlieben einer Person im WWW aufzuzeigen. Mit den Diensten Flickr, 43Things, LastFM, Technorati, Upcoming, Digg und Delicious sind unterschiedliche Bereiche abgedeckt.

In der Tabelle 5.2 sind die Dienste und ihre Eigenschaften aufgelistet. Während bei Delicious, Digg, Technorati und 43Things ein breites Spektrum an Interessen abgedeckt werden kann, schränkt sich dies bei Flickr, Upcoming und erst recht LastFM ein. Im folgenden Abschnitt wird auf diese Einschränkung aufgrund der Inhalte näher eingegangen. Allen Diensten gemein ist dabei die Beschränkung auf virtuell repräsentierbare Objekte.

Tabelle 5.1.: Programmierschnittstellen der Quellen, (Stand 01.12.2007)

Dienst	API-Adresse
Flickr	http://flickr.com/services/api/
43things	http://www.43things.com/about/view/web_service_api
Technorati	http://technorati.com/developers/api/
Upcoming	http://upcoming.yahoo.com/services/api/
Delicious	http://del.icio.us/help/api/ oder http://del.icio.us/help/json/
LastFM	http://www.audioscrobbler.net/data/webservices/
Digg	http://apidoc.digg.com/

5.2.1. Zugang und Identifizierung

Programmierschnittstellen

Bei der Definition zu *Mashup* (Abs. 4.1.1) kommen APIS bereits zur Sprache. Diese abstrakten Schnittstellen ermöglichen den definierten Zugriff auf die Daten der Dienste. In Tabelle 5.1 sind die API der Quellen aufgelistet. Die Schnittstellen regeln auch die Autorisierung des Benutzers gegenüber dem Dienst. Die drei Zugangsstufen erläutert Abschnitt 7.2 im Kapitel zur Implementierung.

Eindeutige Kennung

Im Unterschied zu der Suche nach Personeninformationen im Internet[Kur07] ist bei der Recherche nach Interessen eines Menschen in dieser Arbeit die Identifikation nicht von Belang. Die Personenmerkmale sind ohne eindeutiges Merkmal nicht zweifelsfrei einer Person zuzuordnen. Es muss beachtet werden, dass zwei Personen den gleichen Benutzernamen bei unterschiedlichen Diensten verwenden könnten. Dieses Problem umgeht das Verfahren im hier beschriebenen Ansatz. Vom Nutzer werden die Kennungen bei den jeweiligen Diensten für die Ermittlung seines Interessenprofils vorausgesetzt.

5.2.2. Enthaltene Inhalte

Die erste Gruppe, mit Ausnahme von 43Things, benutzt als Gegenstand Internetadressen. Die Themen dieser Links kommen aus allen Bereichen des Lebens. 43Things und Upcoming bilden eine zweite Klasse. Die Objekte *Veranstaltungen* bzw. *Lebensziele* lassen ebenfalls einen weiten Spielraum bei Interessen zu. Enger wird die Spanne bei Flickr mit dem Objekt *Foto*. Am Geringsten ist die Entscheidungstoleranz bei LastFM mit dem großen, für sich stehenden Interessengebiet *Musik*. Welche Einfluss die Inhalte auf die Analyse haben, beschreibt Abschnitt 5.3. Nicht vertreten in den Quellen sind Videoportale, Bücher-Communities aber auch Einkaufs- und Auktionsplattformen. Die Objekte in Diensten haben unterschiedliche Beziehungen zu den Nutzern (Abs. 5.2.2). Dies wirkt auf die Nähe und damit auf die Gewichtung der assoziierten Interessen ein. Verwendete Details zu den Objekten eines Dienstes zeigt Abschnitt 5.4.1, was ferner zu finden ist und nicht direkt in die Analyse einfließt, beschreibt Abschnitt 9.2.3.

Tabelle 5.2.: Verwendete Quellen und deren Inhalte (Stand 01.12.2007)

Dienst	Gegenstand	Adresse	Sprache
Flickr	Foto	http://flickr.com/	mehrsprachig
43things	Lebensziel	http://www.43things.com/	englisch
Technorati	Weblog (Blogsuchmaschine)	http://technorati.com/	englisch
Upcoming	Termin (Terminatenbank)	http://upcoming.yahoo.com/	englisch
Delicious	Lesezeichen/URLs	http://del.icio.us	englisch
LastFM	Musik (Titel, Alben, Interpreten)	http://last.fm/	mehrsprachig
Digg	Nachrichten(URLs)	http://digg.com/	englisch

Objekt Internetadresse

Die größte Freiheit unter den betrachteten Objekten bieten URLs. Da sich hinter der Adresse unterschiedlichste Themen verbergen können, lagern hier viele Daten für Interessenprofile. Die drei Dienste Delicious (Lesezeichen), Digg (Nachrichten) und Technorati (Blogs) bieten daher abweichende Tendenzen bei den URLs.

Die größte Streuung an Inhalten liegt in den reinen Lesezeichendiensten (Vertreter Delicious) vor. Alles was sich im Internet adressieren lässt, kann von Nutzern darin als Favorit abgespeichert werden.

Ebenfalls URLs verwenden News-Dienste (Vertreter Digg), um Objekte zu referenzieren. In ihnen werden die Adressen eher mit dem Anreiz veröffentlicht, andere darüber zu informieren. Laut Angaben der Betreiber lag der Fokus 2004 eher auf „Geek-Themen“. Mittlerweile sind die Inhalte breiter gestreut, „Politik ist bereits jetzt einer der populärsten Bereiche auf der Site und wird Technologie bald überholen“. [Pon07]

Bei Technorati sind die Objekte *Weblogs*. Nutzer können sich ihre Favoriten sowie die selbstbetreuten Blogs in ihrem Account hinterlegen. Zu den Themen, welche von Bloggern wiedergegeben werden, existieren ähnliche Zahlen wie bei Digg. Im Jahr 1997 begann es mit „Neuigkeiten für Nerds“² bei Slashdot. [Sch06, S. 116]

Fittkaumass zählt bei den Themen von Blogs auf den ersten fünf Plätzen die Themen „Computer, Internet“ (Aktiv und Passiv[AP]:16 %, Passiv[P]: 51 %), „Nachrichten“ (AP:7 %, P:60 %), „Politik“ (AP:9 %, P:49 %), „Unterhaltung, Freizeit“ (AP:8 %, P:46 %), „Reisen, Urlaub“ (AP:8 %, P:44 %). [FM05]

Die Statistiken zu Weblogs (Inhalte, Demographie) und anderen Bereichen gibt Abschnitt 4.1.2 wieder.

Objekt Foto

In den Foto-Communities haben sich in den letzten Jahren enorme Mengen an Bildmaterial angesammelt. Die Verbreitung der digitalen Fotografie („Mehr als die Hälfte der Deutschen fotografieren digital“, „Im Schnitt [...] 55 Bilder pro Monat“³) ist ein Motor dieser Entwicklung. In Flickr wurde am 14.11.2007 das zweimilliardste Foto eingestellt, „täglich [kommen] etwa 4 Millionen neue“ hinzu.⁴ Neben den Amateuren und professionellen Fotografen, die Fotodienste nutzen, um ihre Bilder zu präsentieren, ist der örtliche Bezug von Fotos in

²Bsp. Releases des Linux-Kernels, Anime-Zeichentrickfilme.

³<http://www.golem.de/0708/54343.html> vom 01.12.2007.

⁴<http://www.heise.de/newsticker/meldung/98988> vom 01.12.2007.

letzter Zeit ein Fokus.⁵ „Geotagging“ ermöglicht das einfache Verknüpfen von Orten mit dort fotografierten Aufnahmen.⁶

Objekt Lebensziel

Ein weitläufiges Thema bieten die Ziele, welche sich Menschen im Leben stellen. Neben Reisen zu fernen Orten finden sich Vorhaben, körperlich besser in Form zu kommen oder auch bestimmte Menschen kennenzulernen.

Objekt Musik

Während bei den Fotografien überwiegend eigene Werke als Objekte dienen, ist der Ansatz bei Musik anders. Das virtuelle CD-Regal mit Lieblingsinterpreten kann hier gespeichert werden. Anzumerken ist für den verwendeten LastFM, dass hier die Software Audioscrobler⁷ automatisch beim Musikhören auf dem Rechner die Vorlieben (Titel, Album, Interpret) an den Dienst schickt.

Objekt Termin

Jede Art von Veranstaltung, virtuell und real, kann in Terminbörsen eingetragen werden. Es lassen sich Teilnehmer einladen und Fotos damit verknüpfen.

Arten von Objektbeziehungen zum Nutzer

Mehrfach wurde bereits in anderen Abschnitten darauf verwiesen, dass Nutzer und Objekte in verschiedenen Beziehungen stehen. Diese Relation hat unterschiedlichen Einfluss auf die Interessen. Je näher jemand mit einem Gegenstand verbunden ist, umso aussagekräftiger sind die daraus erhaltenen Interessen. Je weniger Aktivität aufgebracht werden muss, um die Beziehung zum Objekt zu erzeugen, desto lockerer ist das Verhältnis.[Kur07, S. 19]

Die Beziehungen lassen sich je nach Objekt unterschiedlich benennen. Übersicht 5.2.2 ordnet einige Beispiele ja nach Dienst und versucht diese absteigend nach Nähe zu sortieren.

Die engste Beziehung haben Besitzer, Autoren oder Fotografen eines Objektes. Als „Väter“ erschaffen sie den Gegenstand und versehen ihn mit der Grundlage an Daten. Weiter entfernt stehen Kommentatoren. Danach folgen Erweiterer. Noch lockerer ist die Beziehung von Fans, sie markieren Objekte als ihre Favoriten. Nur passiv ist das Verhältnis von Lesern.

Flickr Fotograf, Kommentator, Erweiterer(Tag oder Gruppe), Fan

LastFM Hörer, Fan

43Things Autor(„Zielhaber“), Folgender, Erfüller, Erweiterer

Upcoming Veranstalter/Eintragender, Teilnehmer, Berichterstatter, Beobachter

Digg Entdecker („Submission“), Bewerter („Dugg“), Kommentator, Kommentarbewerter

⁵<http://www.golem.de/0711/56109.html> vom 01.12.2007.

⁶<http://flickr.com/places> vom 01.12.2007.

⁷<http://www.audioscrobler.net/> vom 01.12.2007.

Technorati Autor („Blogger“), Kommentator(direkt, via Trackback oder als Antwort im eigenen Blog), Abonnent(RSS-Leser)

Delicious Entdecker (privat), Für-Sich-Entdecker, Link aus dem Netzwerk erhalten

Die Anbieter speichern zu den Objekten statistische Daten. Anhand dieser können sie erkennen, und für andere Nutzer sichtbar machen, was interessante und wichtige Gegenstände sind. Beispiele dafür sind die Zahlen wie „Views“ zu einem Video bei YouTube⁸, die Anzahl der Personen, die ein Lesezeichen bei Delicious ebenfalls gespeichert haben oder die Beobachter einer Veranstaltung bei Upcoming (vgl. A.1 und A.2).

In den Statistiken (Abs. 4.1.2) finden sich insbesondere zu Blogs auch Zahlen, welche Aktivitäten im Internet führend sind. In der Implementierung (Abs. 6) sind nur ausgewählte Rollen beachtet worden.

5.3. Einordnung und Bewertung der Quellen

Daten, die die verwendeten Dienste zum Interessenprofil beisteuern, sind nicht gleichwertig. Dies liegt an verschiedenen Faktoren. Drei (Persönlichkeit, Spezialisierung, Tag-Häufigkeit) davon werden verwendet, um eine Priorisierung der Aussagekraft zu gewährleisten. Da auch die Faktoren untereinander nicht die gleiche Relevanz haben, gehen sie wiederum gewichtet in die Berechnung des Gewichts eines Dienstes ein.

5.3.1. Eigenschaften und Gewicht von Quellen

Generalisierung

Die Spezialisierung auf ein Thema am Beispiel LastFM (Musik) versus Delicious (Lesezeichen) verdeutlicht, welchen Einfluss diese Eigenschaft auf die Interessen aus dem entsprechenden Dienst auf die Interessen hat. Die Vielfalt innerhalb des Themas *Musik* ist zwar hoch, doch nach außen hin werden die Musikrichtungen nicht separat als Interesse gewertet. Im Gegensatz dazu können sich hinter *Lesezeichen* eine Vielzahl an möglichen Vorlieben verstecken. Da sich Lesezeichen allerdings auch nur auf virtuelle Ressourcen beziehen, besteht auch hier eine Schieflage gegenüber der Realität. Der höchste Grad an Generalisierung wird Objekten in 43Things beigemessen. Sie haben einen starken Bezug zur wirklichen Welt.

Quellen mit hoher Spezialisierung bedienen in der Regel nur wenige Interessengebiete. Umso stärker die Generalisierung der Inhalte eines Dienstes ist, umso mehr Interessenschwerpunkte können daraus hervorgehen (Abb. 5.2).

Persönlichkeit

Auch der Grad der Individualität eines Dienstes für einen Benutzer wirkt auf das Gewicht eines Dienstes ein. Am wenigsten Einfluss auf die Inhalte eines Dienstes hat der Benutzer, wenn er nur passiv daran Anteil hat. Besteht die Möglichkeit, Kommentare zu geben oder Objekte im Dienst als Favoriten zu kennzeichnen, erhöht dies den persönlichen Einfluss. Die größte Bedeutung haben Quellen, wo der Nutzer nahezu allein den Inhalt bestimmt

⁸<http://www.youtube.com> vom 01.12.2007.

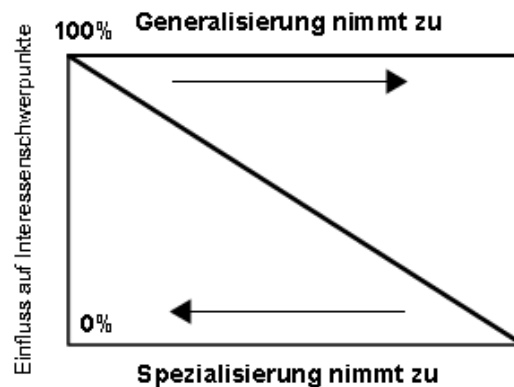


Abbildung 5.2.: Einfluss der Generalisierung der Objekte eines Dienstes auf die ISP (eigene Darstellung)

(Beispiel: Technorati mit eigenen Blogs oder Delicious). Die Persönlichkeit ließe sich auch als variabler Wert verwenden (siehe Abs. 9.2.2).

Tag-Häufigkeit

Die Verwendung von Tags bei Diensten ist unterschiedlich verbreitet. Ebenso sind Nutzer variabel beim Einsatz des Taggings. Ein Grund dafür ist, dass sich manche Objekte leicht mit Schlagworten versehen lassen, andere weniger gut. Dies ist nicht nur für einen Nutzer eine Maßzahl, sondern auch für einen Dienst. Bei der Quelle Digg ist das Tagging nicht im Funktionsumfang enthalten. Dort dient eine Vorauswahl an Kategorien („Container“, „Topic“) dazu Objekte mit Informationen zu versehen. Da aber trotzdem jeweils zwei Begriffe zu einem Objekt zugewiesen werden, ist TH gesetzt.

Gewicht eines Dienstes

Die drei obigen Faktoren wurden für die Analyse festgelegt (Tabelle 5.3). Die Formel $p \cdot g \cdot 1 - th$ verbindet die Bausteine zu einem Wert. Mit diesem haben die Daten aus den Diensten verschiedenen Einfluss auf die Interessenschwerpunkte (Abs. 5.5.1). Mit einer dynamischen Gewichtung ließe sich eine weitere Optimierung der Analyse konstruieren (Abs. 9.2.2). Die Faktoren des Gewichts spiegeln nicht unbedingt die gefundenen Zahlen in den Resultaten wieder, da hier nur sehr wenige Testkandidaten einfließen. Die Korrelation zwischen den drei ausgewählten Faktoren wird nicht betrachtet (Verwendung der Formel im Verfahren 5.1).

5.4. Tags

Ein zentrales Konzept des Web2.0 ist das „Tagging“ für die Inhalte. Dabei ordnet der Nutzer dem Objekt Schlagworte zu. Nicht alle Dienste bieten dabei freies Tagging für die Ordnung im System an (siehe vorgeschriebene Kategorien in Digg). Da Anwendern ansonsten kaum Beschränkungen bei Wortwahl und Schreibweise unterliegen, verwendet jeder sein eigenes

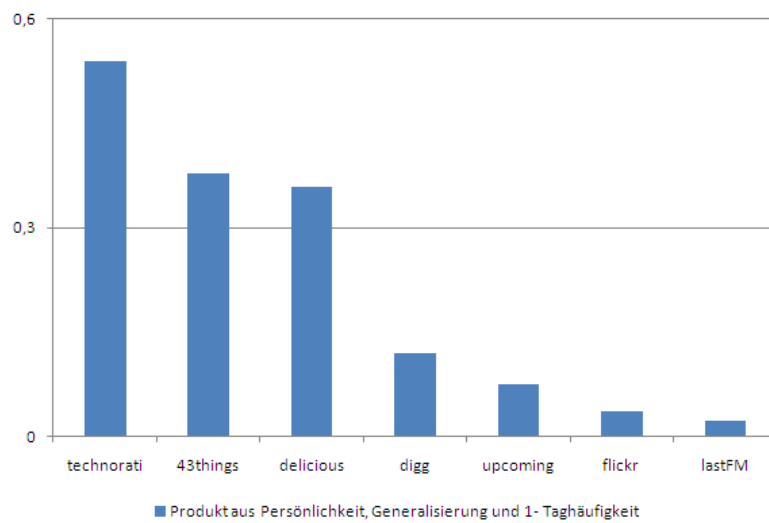


Abbildung 5.3.: Gewicht eines Dienst als Produkt dreier ausgewählter Faktoren (eigene Darstellung)

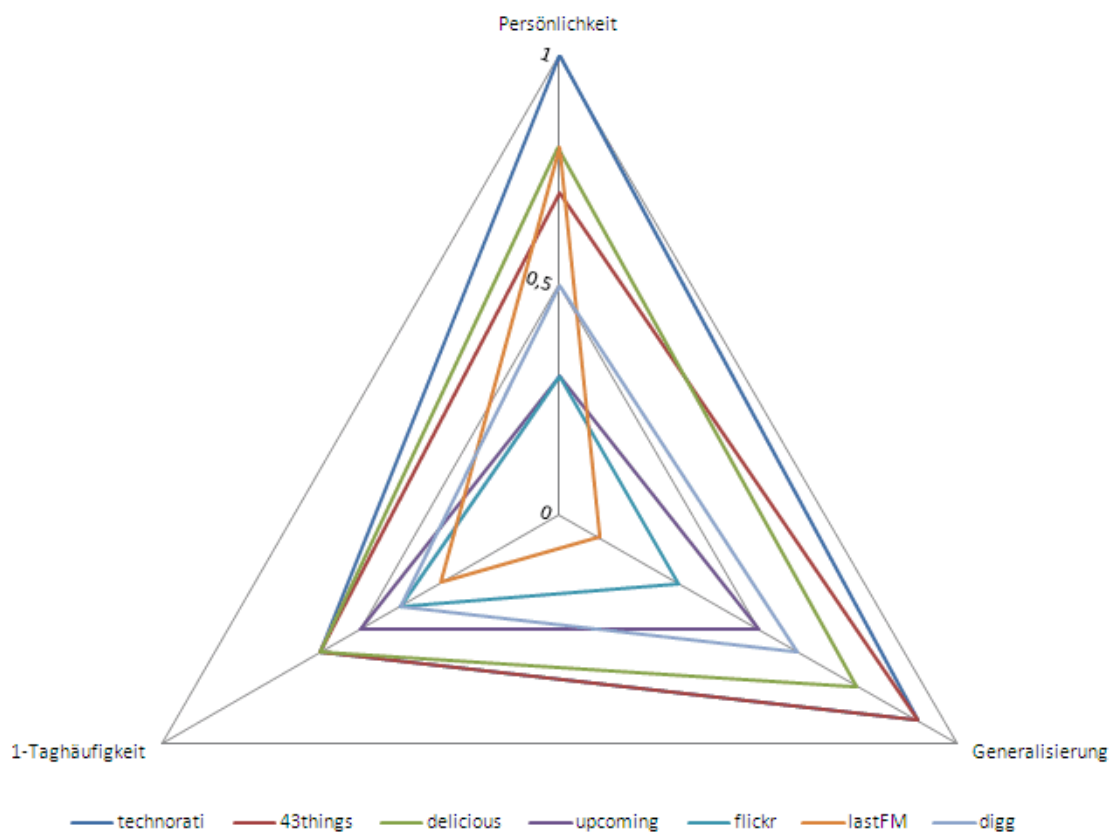


Abbildung 5.4.: Persönlichkeit (P), Generalisierung (G) und Taghäufigkeit TH als die drei Dimensionen der Gewichtung eines Dienstes (eigene Darstellung)

Tabelle 5.3.: Gewichtung der Dienste

Dienst	Produkt	P	G	I-TH
Technorati	0,54	1	0,9	0,6
43things	0,38	0,7	0,9	0,6
Delicious	0,36	0,8	0,75	0,6
Upcoming	0,08	0,3	0,5	0,5
Flickr	0,04	0,3	0,3	0,4
LastFM	0,02	0,8	0,1	0,3
Digg	0,12	0,5	0,6	0,4

Vokabular. Die Funktionen und Arten von Tags können dabei unterschiedlich sein, von Bewertungen über den Gebrauch als Merkzettel bis zu maschinell verwertbaren Schlagworten (siehe Abs. 5.4.2). Die Konsequenzen des Tagging für die Ordnung in Diensten bespricht Abschnitt 4.1.1. Eine Restriktion im Zusammenhang mit Tags, die entscheidenden Einfluss auf die Arbeit hat, bespricht der folgende Abschnitt.

5.4.1. Beschränkung auf Tags

In den Diensten findet man neben den eigentlichen Objekten (siehe 5.2.2) und direkten Personendaten (Alter, Geschlecht, Name, Herkunft) auch indirekte Informationen zum Nutzer. Doch für die Analyse der Interessen werden in dieser Arbeit ausschließlich die Tags an den Objekten genutzt. Eine Ausnahme bilden dabei die Kategorien bei Digg. Diese werden vom System wie Schlagworte behandelt. Dabei muss später bei der Bewertung darauf geachtet werden, dass für diese Begriffe teilweise abgeänderte Regeln gelten. Wie die Schlagworte verwertet werden, um aus ihnen auf Interessenschwerpunkte zu schließen, behandelt Abschnitt 7.1.3.

Nicht in die Betrachtung einbezogen werden Gruppen und Kategorien, denen Objekte zugeordnet sind. Im System wird eine ID (URL, Name oder die ID des Objektes in Bezug zum jeweiligen Dienst) gespeichert, um damit die Tags zu referenzieren (DB-Schema siehe 6.2). Jedoch werden Name, Beschreibung, andere Nutzer und weitere Daten zum Objekt, welche außerhalb der Tags stehen, nicht für die Auswertung eingesetzt (vgl. 9.2.3).

5.4.2. Dimensionen des Tagging

Die Betrachtung der Dimensionen des Taggings von Boyd ergibt sieben Faktoren (siehe Abbildung A.3 im Anhang). Der tripartite Ansatz (Abs. 5.4.2) mit den Mengen Objekt, Nutzer und Tag bildet dabei eine Grundlage. In der Übersicht 5.4 sind die Dimensionen für die in der Arbeit verwendeten Quellen dargestellt. Dabei ist Digg ausgelassen, da dort kein Tagging angeboten wird. Auffällig dabei ist Flickr, es verwendet im Gegensatz zu allen anderen Diensten bei der Gruppierung das „Set-model“. LastFM und Delicious bieten dem Anwender Vorschläge bereits angefügter Tags an. LastFM bietet dem Nutzer zusätzlich Objekte die gefallen könnten an. Wenn die Tags außen vorgelassen werden, ähneln sich alle Dienste bei der Verbindung der Objekte. Gesondert zu erwähnen bei den Verbindungen der Objekte, sind die Ortsinformationen der Objekte (Beispiel: Upcoming) oder die Gruppen bei Flickr. In der Quelle werden auch die möglichen Potentiale zur Weiterentwicklung der einzelnen Ansätze sowie weitere Details im Umgang mit Tags besprochen.[MNBD06]

Tabelle 5.4.: Dimensionen des Tagging (1) für verwendete Dienste (Stand: 02.12.2007, - = keine Angaben)

Dienst	Tagging Rights	Tagging Support	Type of Object
Flickr	Permission-based	Viewable tagging	any (Foto)
43things	Free-for-all	Viewable tagging	any (Lebensziel)
Technorati	Self-tagging	- (basierend auf Beiträgen im Blog)	any (Blog)
Upcoming	Permission-based („Who can see it?“)	Viewable tagging	any (Event)
Delicious	Free-for-all	Suggestive tagging	any (URL)
LastFM	Free-for-all	Suggestive tagging	any (Album, Interpret, Titel)

Tabelle 5.5.: Dimensionen des Tagging (2) für verwendete Dienste (Stand: 02.12.2007, - = keine Angaben, Connectivity = Conn.)

Dienst	Source of Material	Resource Conn.	Social Connectivity	Aggregation
Flickr	Supplied by participant	Grouped	Linked	Set-model
43things	Supplied by participant	none	Linked („People doing this“)	Bag-model
Technorati	Supplied by participant	Linked (Tags)	None	-
Upcoming	Supplied by participant	Grouped (nach Orten)	Linked	Bag-model
Delicious	Supplied by participant	Linked (Tags)	Linked („Network“)	Bag-model
LastFM	Supplied by participant and system	Linked (Band,Tags)	Linked („Freunde“)	Bag-model

Neben den Eigenschaften von Taggingssystemen ist das Verhältnis zum Nutzer in verschiedenen Arbeiten von Interesse. Nicht nur das „Wie“ [Sin05] sondern auch das „Warum“ [Sin06] werden analysiert. Der erste der beiden Gründe steht in engen Zusammenhang mit der Entstehung der „Folksonomies“. Der entscheidende Unterschied zum Kategorisieren ist der fehlende Schritt der Auswahl aus einem abgegrenzten Vokabular beim Tagging.[Sin05]

Ansätze, wie sich die durch Tagging entstehenden Ordnungen verbessern lassen [Pin05], geben Hinweise auf Verbesserungsmöglichkeiten (Abs. 8.3). Den drei „Tricks“ der Profis („knowing the complete vocabulary“, „Synonyms“, „Hierarchy“) stellt Pind fünf Vorschläge, welche Laien das Taggen vereinfachen können anbei. Dies sind Vorschlags-Funktionen mit vorhandenen Tags (aller Nutzer), „showing related tags“ (Synonyme), Hervorhebung von bevorzugten Tags, „Infer hierarchy from the tags“ (Zusammenhänge wie zwischen „Volvo“ und „Auto“ u.a.) und „adjust tags on old content“.[Pin05] Vor allem der letzte Punkt ist wichtig, um das hinzu gewonnene Wissen zu verwenden. Aber die obige Frage *Warum?*, ist hierbei für viele normale Nutzer nicht beantwortet.

Die Analyse zu den Regelmäßigkeiten eines Nutzers beim Taggen (am Beispiel Delicious) von Golder und Huberman [GH05] bietet Auswertungen zur Häufigkeit der Benutzung (sehr unterschiedlich über alle Nutzer) und dem Verhältnis der Anzahl der Lesezeichen zu verwendeten Tags (Abwechslung lässt mit der Zeit nach). Für die Arbeit von Bedeutung ist die Aussage „users’ interests develop and change over time“ [GH05, S. 4]. Die Autoren zeigen, dass sich anhand der Schlagwortverwendung erkennen lässt, welche Bereiche zu welchem Zeitpunkt für einen Nutzer von besonderem Interesse waren. Ein Dienst, der dies für URLs visualisiert, ist Clouldalicious⁹.

Arten von Tags

Golder und Huberman [GH05] identifizieren sieben Arten (Aufgaben) von Tags. In Übersicht 5.4.2 sind diese mit Beispielen aufgelistet [GH05, S. 5]. Den Einfluss der Unterteilung auf die Analyse bespricht 5.4.3.

1. „Identifying What (or Who) it is About“, Beispiel: „uno“, „w3c“, „soccer“

⁹<http://cloudalicio.us> vom 03.12.2007.

Eigene versus fremde Tags

Zweierlei Typen von Tags sind mit Objekten einer Person, in den Diensten verbunden, wo es möglich ist, dass derselbe Gegenstand von mehreren Personen getaggt wurde (siehe 5.4.2).

In unmittelbarem Zusammenhang mit Interessenprofilen stehen Tags, die die untersuchte Person selbst vergeben hat. Beispiele dafür sind die Tags, die gespeicherte Lesezeichen einer Person oder den Inhalt von Fotos beschreiben. Dazu gehören auch die Schlagworte, welche Beiträge im persönlichen Weblog beschreiben.

Tags, welche nicht durch die Person selbst an die Objekte des Nutzers gelangt sind, sondern durch Fremde, bilden die zweite Gruppe. Diese indirekte Beziehung findet sich da wieder, wo der untersuchte Mensch Objekte in seinem Profil gespeichert hat, welche er sich mit anderen teilt.

Mit der Verwendung von indirekt vergebenen Tags steht mehr Information zu einem Objekt bereit. Die Assoziation zu einem Interessenschwerpunkt fällt leichter.[Mik05] Allerdings verfälschen die allgemeinen, fremden Schlagworte die persönlich von einer Person verwendeten Begriffe. Daher werden im Verfahren dieser Arbeit keine indirekten Beschreibungen für die Erstellung des Interessenprofils verwendet.

5.4.3. Freiheiten bei der Verwendung von Tags

Der Vorteil der Freiheit von Tags für den Nutzer ist ein Nachteil für die Weiterverwendung. Da beim Tagging nicht mit einem abgeschlossenen Vokabular (wie in klassischen Taxonomien) gearbeitet wird, tauchen unterschiedliche Probleme auf. Einen Filter, der die Tags eines Nutzers vor der Weiterverwendung bearbeitet, schlagen Al-Khalifa und Davis vor. Ihr Ansatz versucht Delicious als Unterstützung für „semantic annotation of web resources from an educational perspective“[AKD06a, S. 1] zu nutzen. Anders als im Rahmen dieser Arbeit beschränken sie das Vergleichsgebiet für Folksonomy und Ontologie auf eine Domäne („teaching 'CSS' in a 'web design' course“)[AKD06a, S. 3]. Ihr Experiment [AKD06b] zeigt „folksonomies hold more semantic value than keywords extracted using machines“ und hebt daher den Vorteil des in dieser Arbeit beschriebenen Verfahrens heraus. Schritt 1 ihrer Methode stellt dabei „tags extraction and normalization“[AKD06a, S. 4] dar. Schritt 2 betrifft das eigentliche Verfahren (Abs. 5.5).

Die Normalisierung der Tags betrifft in [AKD06a] die folgenden Schritte. Zuerst werden alle Worte in Kleinbuchstaben umgewandelt, dann nicht-englische Zeichen gelöscht. Da sich die Autoren auf eine Sprache beschränken, verhindert dies Probleme. Darauf folgt ein Stemming und eine Gruppierung von ähnlichen Begriffen. Der letzte Schritt dieser Phase, das Eliminieren von generellen Tags, wird anhand des Vergleiches mit der Domänen-Ontologie bewerkstelligt (siehe Abs. 5.5).

Zu den obigen Schritten, die dem „clean up the noise in people's tags“[AKD06a, S. 4] dienen, einige Anmerkungen im Rahmen dieser Arbeit. Leichtere Probleme wie unterschiedliche Schreibweisen (Groß/Klein, Getrennt/Zusammen, falsche Rechtschreibung) lassen sich im Allgemeinen durch Vergleiche mit Wörterbüchern¹¹ lösen. Schwieriger ist die Sprache eines einzelnen Wortes zu ergründen. Da Nutzer in der globalisierten Welt nicht ausschließlich ihre Muttersprache zum Taggen verwenden, mischen sich fremdsprachige (mehrheitlich

¹¹Vgl. Google mit automatischer Vorschlagsfunktion für alternative Schreibweisen.

englisch bei deutschen Benutzern) Begriffe an den Objekten (siehe auch Abs. 8.3.2).

Ein Hindernis, welches ausschlaggebend für den Erfolg der Analyse von virtuellen Interessen ist, stellt die Semantik von Begriffen dar. Schon an einfachen Beispielen, wie Bank (Geldinstitut oder Gartenmöbel) zeigt sich, dass hier nur schwer aus einzeln stehenden Worten eine Bedeutung und damit Assoziation zu Bereichen erkennbar ist. Eine Lösungsmöglichkeit bespricht Abschnitt 9.2.3.

Variabilität der Tags

Unter der Annahme, dass ein Nutzer beim wiederholten Taggen (auch in verschiedenen Diensten) immer gleiche Schreibweisen verwendet, wird in dieser Arbeit das Normalisieren der Tags nur mit geringem Aufwand verfolgt. Alle Tags werden in Kleinbuchstaben gespeichert und bei der Verarbeitung gefiltert. Um die Veränderungen der genutzten Worte zu prüfen, kann anhand der Levenshtein-Distanz das Taggen des Benutzers bewertet werden. Um genaue Aussagen zu machen sind weitere Verfahren notwendig, welche in der Arbeit nicht vertieft werden.

Ordnen von Tags

Um Tags zu gruppieren, benutzen Dienste „bundles“ (Delicious) oder bieten Möglichkeiten verwandte Tags zu ergründen (vgl. 9.2.3). Auch die Mengen aus Google Sets dienen hierbei (siehe Abs. 9.2.3).

5.5. Das eigentliche Verfahren

Nachdem die Herkunft der Daten, deren Bewertung und die Beschränkung auf Tags erläutert wurde, folgt die Beschreibung, was mit den Daten geschieht, um daraus ein Interessenprofil zu erstellen.

1. Extrahieren
2. Normalisieren und Filtern
3. Anfrage mit den verbliebenen Tags an FUTEF
4. Eintragen der neu erhaltenen Kategorien
5. Einordnen der Unterkategorien in Hauptkategorien mit CatGraph
6. Interessenschwerpunkte berechnen

Mit der Angabe des Benutzernamen (und eventuell weiteren Identifizierungen zu einer Person) werden aus den Quellen Objekte eingesammelt. Details der Implementierung bespricht Abschnitt 6. In Verbindung mit dem Objekt wird der Benutzer und die Besitzart (Abs. 5.2.2) abgespeichert. Die Tags eines Objektes behalten Verknüpfungen zum Dienst und dem Objekt.

Wie auch in [AKD06a] beginnt das Verfahren nach dem Extrahieren der Tags aus den Quellen, mit der Normalisierung. Das Aussieben der generellen Tags (Menge A) geschieht

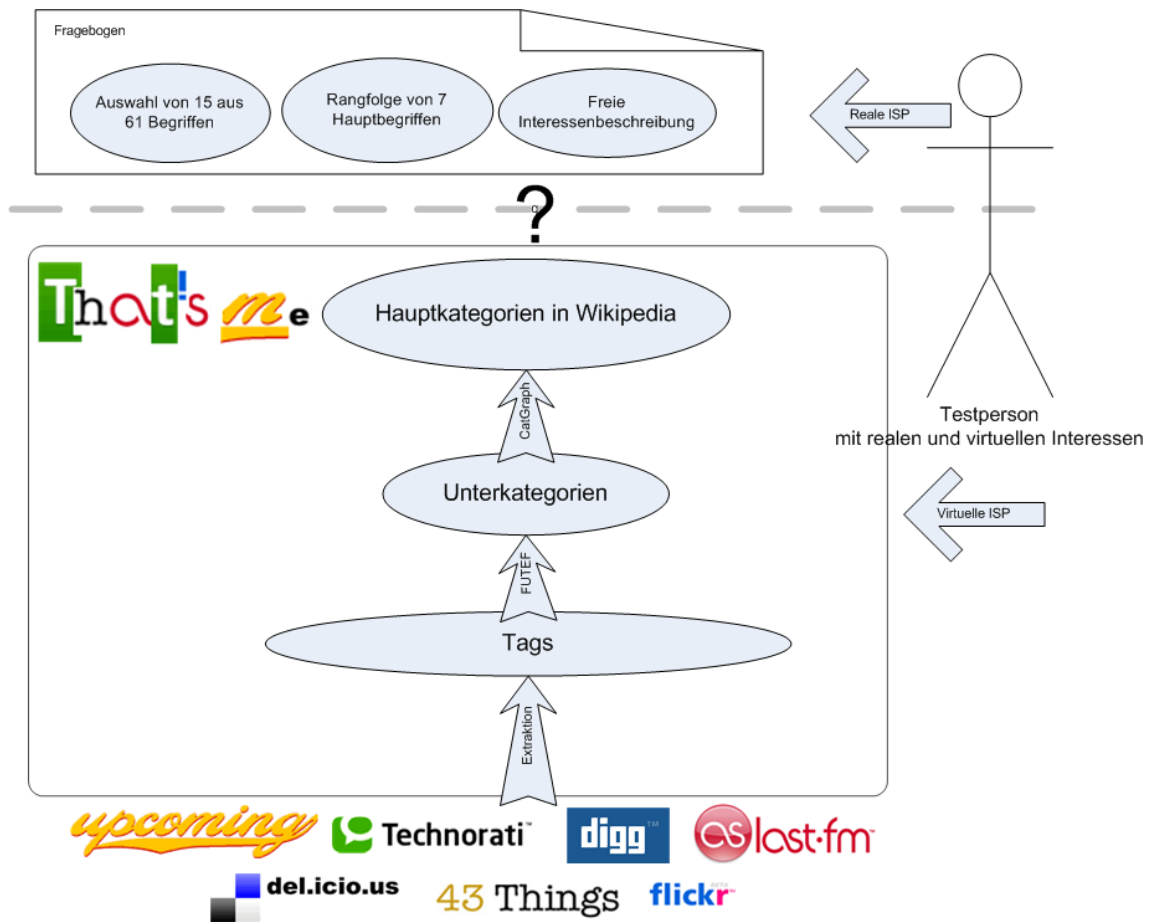


Abbildung 5.6.: Schema des Verfahrens (eigene Darstellung)

anhand der bereits vorhanden Kategorien in der Datenbank. Für den Start wurden dazu 154 Rubriken aus den Lexika sowie Digg (Hauptkategorien und erste Ebene der Unterategorien) eingegeben. Für diese wird eine Anfrage an die Lexika gestellt.¹²

Zu den verbliebenen Tags werden die Kategorien aus Wikipedia und Freebase geholt. Die Tags, für die keine Antwort zurückkommt (Menge C), sind den Lexika unbekannt. Die Antworten aller einordenbaren Schlagworte (Menge B) werden in der Datenbank hinterlegt. Nun wird erneut ein Abgleich der generellen Tags mit den Rubriken in der DB vorgenommen (weitere Einteilung in Menge A und B).

Für die Rangfolge der ISP wird anhand der Formeln 5.1 stufenweise ausgerechnet, mit welchem Gewicht die Elemente Einfluss auf das Ergebnis haben. Tags, die ohne Antwort aus den Lexika zurückkommen, können auf diesem Weg nicht mit Semantik versehen werden (Abs. 9.2.3).

¹²Das Aussieben wurde in der Umsetzung des Verfahrens nicht verwirklicht, da es wichtige Informationen verschluckt.

$$G_{Dienst} = \text{Persönlichkeit} \cdot \text{Generalisierung} \cdot (1 - \text{Taghäufigkeit}) \quad (5.1)$$

$$G_{Tag} = \sum G_{Dienst} \text{ in denen ein Tag auftritt} \quad (5.2)$$

$$G_{Kat} = \sum G_{Tag} \text{ aller Kategorien die zugeordnet wurden} \quad (5.3)$$

$$G_{ISP} = \sum G_{Kat} \text{ aller Unterkategorien die zugeordnet wurden} \quad (5.4)$$

5.5.1. Gewicht eines Dienstes

Wie bereits in Abschnitt 5.3.1 erläutert wurde, haben Dienste aufgrund ihrer Eigenschaften unterschiedlich starke Einwirkung auf die Vielfalt der Interessen. Die Formel 5.1 zur Berechnung des Dienstgewichts bildet den ersten Faktor. Jeder Tag, der aus einer Quelle stammt, wird später mit diesen Wert gewichtet.

5.5.2. Gewicht eines Tags

Das Tag-Gewicht ergibt sich aus der Summe des Auftretens bei den Diensten (Anzahl der Verwendung) in Verbindung mit dem Dienst-Gewicht. Nicht mit bewertet wird die Anzahl fremder Tags. Dieser Einfluss lenkt von den Interessen des betrachteten Nutzers ab.

5.5.3. Gewicht einer Kategorie

Mit den Gewichten von Diensten und Tags wird der Einfluss der Kategorie auf die Interessenschwerpunkte ermittelt. Dieser setzt sich zusammen aus der Summe der Gewichte der Tags, die die jeweilige Kategorie als Resultat ergeben.

5.6. Lexika

Die Einordnung der Tags in Kategorien geschieht mit Hilfe der Lexika. Als Teil des Verfahrens soll in diesem Absatz auf die Antworten der Enzyklopädien eingegangen werden. Da Digg nur theoretisch zum Vergleichen von Kategorien mit Tags (siehe 5.5) dient, werden in diesem Segment keine Einzelheiten geschildert. Details der Implementierung beschreibt Abschnitt 6.

Gemein ist den Lexika neben den Eigenschaften aus 4.3.2, dass neben den Kategorien die für die Ordnung der Daten sorgen auch „Systemkategorien“ in der Klassifikation enthalten sind. Diese werden für die Analyse der Interessen ignoriert. Ebenso existieren Rubriken, die aufgrund der häufigen Beziehungen zu anderen Bereichen vermehrt in Antworten enthalten sind. Ein Beispiel dafür sind Ortsangaben und damit verbundene „Geographie“ als Kategorie. Auf der Seite der Interessenschwerpunkte kann das einerseits zu einem Überhang eben dieser führen. Andererseits sind Interessen wie „Reisen/Urlaub“ dadurch übermäßig beeinflusst. Dieses Hindernis lässt sich mit Hilfe des Kontext zu einem Objekt überprüfen (Abs. 9.2.3).



Abbildung 5.7.: Systemkategorien in Freebase¹³

5.6.1. Wikipedia

Für die Visualisierung des Netzwerk an Kategorien in Wikipedia dient die Software Cat-Graph¹⁴. Beispiele für Systemkategorien in WiP (Englisch) sind „Categories need diffusion“, „Main topic classification“ und viele mehr. Die Wurzel der Klassifikation markiert die eine Systemkategorie. Für die deutsche Ausgabe ist dies „!Hauptkategorie“ in der englischen Version „!Main topic classifications“. Abbildung A.8 zeigt ein Beispiel.

Die Kategorien welche WiP einem Tag zuweist, lassen sich in zwei Gruppen teilen. Zum einen die Unterkategorien, welche nicht weiter betrachtet werden, zum anderen die Oberkategorien, die sehr umfangreich ausfallen. In der Analyse wird deshalb im Kategoriegraph nur zwei Schritte (Vater und Großvater) nach oben gelaufen. Zusätzlich werden die Kinder der „!Hauptkategorie“ (und ihrer englischen Entsprechung) sowie deren Kinder zur Analyse herangezogen.

5.6.2. Freebase

Die Struktur hinter den Daten von Freebase ist derzeit¹⁵ in einigen Bereichen dichter als in anderen (siehe 9.2.3). Auch FB hat Systemrubriken, welche unter der Kategorie „System“ (Abb. 5.7) zusammengefasst sind. Den Aufbau des System von Freebase zeigt Abbildung A.7.

5.7. Weitere Verfahren zur Erkennung von Bedeutungen

Auch andere Ansätze, die Semantik in digitalen Objekten zu erkennen, existieren. Für die Domäne der Kunst in Museen wurde kurz in Abschnitt 4.1.1 ein Beispiel erwähnt. Ein weiteres aus einem ähnlichen Bereich stellt [Bro02] dar. Mit Hilfe der „Facet analytical theory“ wird dabei untersucht, wie „classification schemes can be applied to the organization of digital resources“. [Bro02] Die folgenden Abschnitte erläutern andere Ansätze die automatische Erkennung behandeln oder vereinfachen.

¹⁴<http://tools.wikimedia.de/~dapete/catgraph/> vom 03.12.2007.

¹⁵Freebase bezeichnet seinen derzeitige Status mit „alpha“, Stand 03.12.2007.

5.7.1. Collaborative filtering

Ein Verfahren, welches nur kurz erwähnt wird, aber für die Erkennung von Vorlieben in großen Datenmengen eine Rolle spielt (Vgl. Empfehlungen bei Amazon¹⁶, ist „Collaborative filtering“.

5.7.2. Crowdsourcing

Dem Problem der mangelnden künstlichen Intelligenz von Computern tritt ein anderer Ansatz entgegen. Mit „Crowdsourcing“ werden die Lösungen komplexer Probleme auf die Massen an Internetnutzern verteilt. Die menschlichen Vorteile ermöglichen so bspw. die Erkennung von Bildern in einer Qualität mit ausschließlich automatischer Verarbeitung noch nicht effizient möglich sind (Vgl. GalaxyZoo¹⁷ oder MechTurk¹⁸ von Amazon.).

¹⁶<http://www.cs.helsinki.fi/u/gionis/linden03amazon.pdf> vom 20.12.2007

¹⁷<http://www.galaxyzoo.org/> vom 20.12.2007.

¹⁸<http://www.mturk.com/> vom 20.12.2007

6. Die Implementierung von *ThatsMe*

Neben der Theorie der Gewinnung von Interessenprofilen aus virtuellen Identitäten, die in den vorhergehenden Abschnitten besprochen wurde, erläutert der folgende Teil die Umsetzung in die Praxis. Mit *ThatsMe* kann der Anwender Daten aus den Quellen extrahieren. Diese können analysiert werden. Als Ergebnis erhält der Benutzer die Rangfolge seiner virtuellen Interessenschwerpunkte.

Für die Beschreibung des Softwaredesign dient ein abgespecktes Pflichtenheft. Die Erweiterbarkeit auf weitere Quellen erläutert 9.2.5.

6.1. Softwaredesign

Eng an dem theoretischen Modell der Interessenanalyse orientiert sich der Entwurf der Software. In der Spezifikation ist festgelegt, welche Teile notwendig, welche zumindest angedacht sind und was nicht realisiert wurde.

6.1.1. Zielbestimmung

Musskriterien

Die Kriterien welche unabdingbar für die Software sind, beschreibt der folgende Abschnitt. Ohne sie wäre eine Analyse nicht durchführbar.

1. Identifizierung des Nutzers gegenüber dem System, damit die Daten aus den Quellen einer Identität zugeordnet werden können
2. Eingabemöglichkeit der Identifikation (Abs. 7.2) des Anwenders gegenüber den Diensten
3. Extraktion der Daten aus den Diensten und Speicherung der Daten
4. Anfragen mit den Daten an Lexika und Speichern der Antworten (Kategorien) der Lexika
5. Anzeige der Ergebnisse als virtuelles Interessenprofil

Sollkriterien

Die Erfüllung dieser Kriterien wird angestrebt. Das System sollte sich die Identifikations eines Nutzers (gegenüber *ThatsMe*) merken. So ist gewährleistet, dass ein wiederkehrender Anwender nicht alle bereits getätigten Schritte erneut abarbeiten muss. Aus den Tags sollten, vor der Anfrage an Lexika und dem Vergleich mit den Kategorien, diejenigen ausgefiltert werden, welche für die Analyse ungeeignet sind (Abs. 5.4.3). Aus den Kategorien sollten solche ausgefiltert werden die keinen Mehrwert für das Verfahren haben oder durch ihre Menge die Ergebnisse verfälschen.

Kannkriterien

Alle Kriterien in diesem Bereich bilden Verbesserungen des Verfahrens (Abs. 9.2) ab. Dazu zählen vorgelagerte Aktivitäten, Veränderungen im Verlauf des Prozesses und nachträgliche Modifikationen.

Für die Erweiterung der anzugebenden Profile (bei Quellen) kann eine einfache Schnittstelle vorgesehen werden. Damit können Tags auch aus anderen Diensten die Interessenschwerpunkte beeinflussen. Um weitere Lexika zur Einordnung der Tags in Kategorien heranzuziehen, kann eine Schnittstelle die Verbindung zu den Enzyklopädien herstellen, welche erweiterbar ist. Für den Überblick wäre es interessant, auch andere Daten eines Nutzers außer den Endergebnissen sehen zu können (Tags, Objekte). Die Darstellung der Interessen könnte (grafisch) die vielfältigen Verbindungen zwischen Tags, Unterkategorien und Hauptkategorien abbilden.

Abgrenzungskriterien

Die folgenden Punkte sind nicht Bestandteil der Software. Fremde Benutzer können die Daten anderer Anwender nicht sehen. Ein Vergleich von Personen innerhalb von *ThatsMe* ist nicht vorgesehen. Ebenso wird dem Nutzer keine Möglichkeit geboten seine Tags zu optimieren (Bsp. Schreibweisen, andere Worte) und damit außerhalb der Quellen zu verändern. Die Software wird nicht mit Blick auf Performance oder Sicherheitsaspekte entworfen. Da *ThatsMe* nur lokal im Einsatz ist, wird auf Neuregistrierung und Benutzerverwaltung verzichtet.

6.1.2. Produkteinsatz

Anwendungsbereiche

Die Daten eines Benutzer im Internet beschreiben eine Person. Diese Beschreibung kann der Benutzer dazu anwenden um für sich interessante Inhalte anzuzeigen. Mit der Angabe eines realen Interessenprofils lässt sich ein Vergleich zwischen virtueller Darstellung und der Wirklichkeit machen.

Betriebsbedingungen

Das System läuft für die Analyse auf einem lokalen Testsystem. Es wird während der Laufzeit durch den Autor der Arbeit betreut.

6.1.3. Qualitätsanforderungen

Das System wird als Proof-of-Concept entwickelt. Die Qualitätssicherung wird nur im Bezug auf das Ergebnis des Verfahrens hin angewandt.

6.1.4. Benutzungsoberfläche

Die Schnittstelle für den Benutzer stellt ein auf HTML/CSS basierendes Konstrukt dar. Für jede Testperson existiert ein Zugang in dem die Daten der jeweiligen Person extrahiert, gespeichert, verwertet und angezeigt werden.

6.1.5. Technische Produktumgebung

Software: für Server und Client

Die Software ist in PHP unter Zuhilfenahme des ZEND Frameworks¹ implementiert. Für die Anbindung an die Dienste wurde zusätzlich JavaScript verwendet. Auf dem Server läuft eine MySQL-Datenbank².

Produkt-Schnittstellen

Es sind keine Schnittstellen implementiert. Die Erweiterbarkeit wird in Abschnitt 9.2.5 besprochen.

6.1.6. Spezielle Anforderungen an die Entwicklungs-Umgebung

Entwicklungs-Schnittstellen

Um Daten zu extrahieren, wurden der Schnittstellen der Dienste Flickr, Delicious, LastFM, 43Things, Upcoming, Technorati und Digg sowie der Lexika Wikipedia³ und Freebase⁴ verwendet. Zusätzlich wurde die FUTEF-API⁵ verwendet, welche Suchen in den Inhalte der englischen Wikipedia ermöglicht. Die Adressen zu den API der Dienste zeigt 5.2.1.

6.2. Umsetzung der Kriterien des Verfahrens in der Implementierung

Neben der Theorie, die das Verfahren stützt, war die praktische Umsetzung ein Teil der Arbeit. Einerseits um die Ansätze an Beispielen erläutern zu können, andererseits um Hindernisse, die das Modell außen vor lässt, zu ergründen. Der folgenden Abschnitt beschreibt die Einzelheiten auf diesem Gebiet.

6.2.1. Datenbankmodell

Als Grundlage der Speicherung der Daten wurde eine relationale Datenbank verwendet. Es wurden sechs Entitäten und acht Beziehungen identifiziert (Abb. 6.1). Diese Relationen zwischen Tag, Objekt und Benutzer beschreibt das Konzept des tripartiten Graph (Abs. 5.4.2). Dienste bilden einen weiteren Rahmen der drei. Die Zuordnung der Tags in die Kategorien welche die Lexika ordnen, bilden den VIPA eines Benutzers.

Die Umsetzung in Tabellen mit Attributen zeigt das Datenbankmodelldiagramm in Abb. 6.2. Darin sind auch die Fremdschlüsselbeziehungen, welche die Relationen darstellen, eingezeichnet. Hervorzuheben dabei die Oberkategorie-Beziehung, welche es ermöglicht, dass Kategorien mehr als nur einer Rubrik als Kind zugeordnet ist (Bsp. „Programming languages“ ist Unterkategorie von „Computing“, „Education“, „Science“).

¹<http://framework.zend.com/> vom 10.10.2007.

²<http://www.mysql.com/> vom 10.10.2007.

³<http://tools.wikimedia.de/dapete/catgraph/> CatGraph-Tool vom 10.12.2007.

⁴<http://www.freebase.com/view/freebase/api> vom 10.12.2007.

⁵<http://api.futef.com/apidocs.html> vom 10.12.2007.

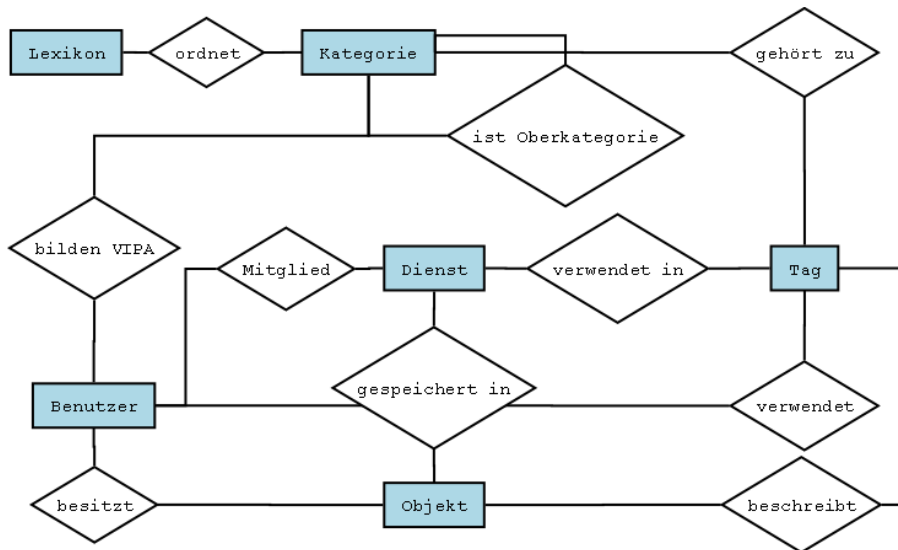


Abbildung 6.1.: Entity-Relationship-Modell zu *ThatsMe* (eigene Darstellung)

Das Speichern der Zugangsdaten von Benutzern zu Diensten wurde nicht implementiert. Für eine stetige Beobachtung von Benutzern wäre dies notwendig. Ebenso wurde auch darauf verzichtet, die taggende Person mit einem Benutzer zu verbinden. Faktisch wird aber sobald gegeben der Benutzername zu einem Tag als Tagger gespeichert. Über die Beziehung „Objekte eines Benutzers“ sind alle Tags, die an den Ressourcen eines Anwender stehen, zuweisbar.

Feinheiten des Verfahren in der Datenbank

Einige Detaillösungen die im Abschnitt 7.1 zur Sprache kommen, werden in den kommenden Zeilen aus der Sicht der Umsetzung wiedergegeben.

Mit dem Feld Tagger werden die Informationen gespeichert um unterscheiden zu können, welche Tags direkt von *ThatsMe*-Nutzern vergeben wurden und welche von fremden Personen an die Objekte geschrieben wurden. Da die zeitliche Komponente der Nutzerdaten nicht im Vordergrund der Analyse stand, wurden eingeordnete Tags markiert (DB-Schema, Tabelle *Tag*, Spalte *aktuell*) und nicht erneut verwendet. Dies steigert die Effizienz der Software, ist aber falls sich die Kategorien in den Enzyklopädien oder die Lexika (als einordnende Komponente) selbst ändern, unpraktikabel. Das anfangs erwähnte Aussortieren von bekannten Begriffen (Abs. 5.4.3), welches im Verfahren letztlich nicht eingesetzt wurde, spiegelt sich in *Omega* in der Tabelle *Tag* wieder.

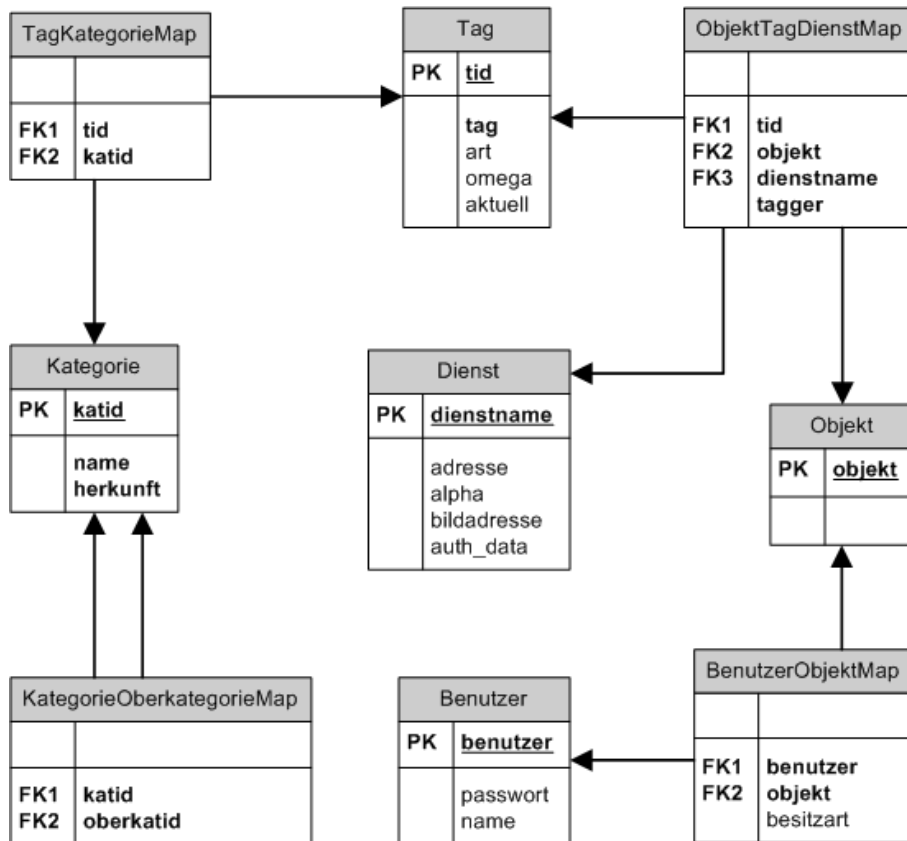


Abbildung 6.2.: Tabellenstruktur in der *ThatsMe*-Datenbank (eigene Darstellung)

7. Gefundene Ergebnisse

Mit dem in Kapitel 5 beschriebenen Verfahren als Grundlage der Implementierung wurden Testläufe mit den Daten der Versuchspersonen durchgeführt. Den Ablauf und die dabei erhaltenen Erkenntnisse bespricht das folgende Kapitel. Im darauffolgenden Kapitel findet ein Vergleich der realen Einschätzungen der Testpersonen anhand von Fragebögen mit den Resultaten der Analyse statt.

7.1. Testreihen

Nicht jede Kombination von Daten zur Einordnung und Lexika bringt erfolgversprechende Ergebnisse. Innerhalb der Antworten der Lexika sind weiterhin nicht alle Teile aussagekräftig und zur Weiterverwendung geeignet. Mit einigem zeitlichen Aufwand wurden daher verschiedene Folgen von Daten der Anfrage, Anfrageziele und Antwortverwertungen getestet. Die einzelnen Faktoren besprechen die folgenden Abschnitte.

7.1.1. Daten der Anfrage

Als Startpunkt der Analyse stehen die Daten der Nutzer in virtuellen Profilen. Sie bilden die Quelle für den virtuellen Interessenprofilabdruck VIPA.

Die Beschränkung auf Tags (Abs. 5.4.1) schließt dabei Einflüsse anderer Daten in den Profilen aus. Nach dem Extrahieren aus den Quellen speichert die Datenbank 2.771 unterschiedliche Tags. Inwiefern diese wirklich verschieden sind, bespricht Abschnitt 5.4.3. Im nächsten Abschnitt werden Kennzahlen zu den Tags an den Objekten erläutert und verglichen.

Kennzahlen der Tags in den Ergebnissen

Wenn man die einzelnen Taganzahlen der Dienste summiert, kommt man auf 3.021, es werden also Tags auch dienstübergreifend verwendet. Die Menge teilt sich unterschiedlich auf. Zum einen 35 Kategorien aus Digg¹, welche als Tags interpretiert werden. Sie haben den Vorteil bereits englischsprachig zu sein und aus einem festen Vokabular zu stammen. Andererseits 1005 Tags aus Delicious, schon durch die Vielzahl an zugewiesenen Objekten (Tab. 7.1) haben sie den größten Einfluss auf den VIPA. Verglichen mit Flickr, bei ungefähr doppelter Anzahl der Objekte und weniger als dem doppelten der Tags, ist die Mehrfachverwendung und damit das Nutzen ähnlicher Tags bei Delicious weitaus stärker ausgeprägt (Zahl der eindeutigen Paare von Objekt, Tag und Dienst: Delicious mit 87.747 vs. Flickr mit 5.629). Diese Daten zeigen, dass die Testpersonen Tags bei Delicious zum Organisieren verwenden, während dies bei Flickr nicht ausgeprägt ist. Die anderen Dienste liefern kaum Material. Interessant wäre eine Betrachtung der Verteilung von verwendeten

¹In der Datenbank bei Art eines Tags markiert mit „diggcontainer“, „diggtopic“.

Tabelle 7.1.: Zahlen in der Datenbank nach dem Extrahieren, Objekt-Tag-Dienst (OTD)

Dienst	OTD	Tags	Objekte	t/o	o/t
Technorati	116	114	9	12,67	0,08
43things	135	98	21	4,67	0,21
Delicious	87747	1005	344	2,92	0,34
Upcoming	63	23	4	5,75	0,17
Flickr	5269	1781	686	2,6	0,39
LastFM	0	0	0	0	0
Digg	372	35	176	0,2	5,03

Tags an Objekten um deren Bedeutung im Dienst zu verstehen. In [GH05] werden Statistiken zu Delicious genauer analysiert. Hierbei spielen die Potenzgesetze eine starke Rolle. Wenige Tags werden häufig verwendet, viel eher selten. Wie sich dies auf die Ergebnisse auswirkt, erläutert Abschnitt 9.1.1.

Verfeinerung und Auswahl der Daten

Da Benutzer, neben dem eigentlich Zweck (Organisation), ihre Tags in weiteren Funktionen benutzen, müssen diese nicht beachteten Schlagworte ausgefiltert werden (Übersicht 7.1.1). Wie dies in der Software realisiert wurde, beschreibt Abschnitt 6.2.1.

- Tags mit denen Lesezeichen anderen Benutzer übermittelt werden (Präfix „for:“)
- Tags mit geographischen Informationen (Präfix: „geo:“, „ge:“, auch in Flickr verwendet)
- Tags die Erinnerungsfunktionen haben (Präfix: „toDo“, „toRead“, „toListen“ auch in anderen Schreibweisen)
- Tags die nur aus Ziffern bestehen (Bsp. Jahreszahlen)
- Tags aus weniger als zwei Zeichen bestehend (Bsp. Abkürzungen)
- Tags mit mehr als 23 Zeichen (Bsp. zusammengesetzt Worte und Sätze die mangels Leerzeichen zusammengerückt wurden)
- Tags aus dem Wortschatz von Dublin Core² (Präfix: „dc:“)

Alle restlichen Tags wurden nach ihrer Verwendungshäufigkeit sortiert. Die meistverwendeten nutzte die Anfrage an die Lexika um sie einzuordnen (Abs. 7.1.2). Dabei wurden zuerst alle Tags die von Nutzer selbst an Objekten angebracht wurden („Eigentags“) eingeordnet, danach die welche über Favoriten und andere Besitzarten von Objekten (Abs. 5.2.2) gesammelt wurden („Fremdtags“). Die angefragten Tags ergaben über 17.000 Kategorien.

Der direkte Vergleich der Tags mit den bereits vorhandenen Kategorien erfolgte nicht. Dieser Schritt filtert Begriffe anhand bekannter Konzepte (Abs. 5.4.3) heraus und mindert dadurch den Aufwand. Da allerdings auch bekannte Begriffe in anderen Kategorien auftreten, würde deren Fehlen in der Einordnung ein Lücke hinterlassen.

²<http://www.dublincore.org/> vom 17.12.2007.

7.1.2. Anfrageziele

Neben den Daten für die Inhalte der Anfrage aus dem vorangegangenen Abschnitt ist das Ziel der Anfrage von entscheidendem Einfluss. Im theoretischen Teil wurden Wikipedia (dt., engl.), Freebase und Digg als mögliche Quellen beschrieben. Nach Ablauf der Tests kristallisierte sich ein Vertreter als passend heraus. So wurde der direkten Anfrage an die Lexika und deren Kategorien ein Schritt vorgeschaltet. Mit der FUTEF-API³ ist es möglich die Inhalte der Wikipedia zu durchsuchen. Im Gegensatz zum Vergleich mit den Kategorien ist dies weitaus ergiebiger. Es werden so auch Themen korrekt eingeordnet, bei denen ein Tag die gleich Form wie ein Kategorienname besitzt. In der Antwort kommen neben den Kategorien jedes einzelnen Artikels auch die zehn meistgenannten Rubriken der gesamten Antwortmenge zurück⁴. Da hierbei allerdings der interessante „LongTail“ (Abs. 9.1.1) verloren geht, wird dieser Teil ignoriert.

7.1.3. Antwortverwertungen

FUTEF ist derzeit nur für die englische Version der Wikipedia verfügbar. Dieser Fakt zeigt eines der Haupthindernisse, die Sprachunterschiede, auf. Alle deutschen Tags, die nicht in einem englischen Wikipedia-Artikel erscheinen, bleiben ohne Antwort in Form von Kategorien.

Die Kategorien mit denen FUTEF auf die Tags antwortet, fallen mit 17.489 sehr zahlreich aus. Ein Blick auf die Ergebnisse zeigt, dass das Categoriesystem der Wikipedia eher einem festen Schlagwortkatalog (siehe Abs. 4.1.1) entspricht als einer Klassifikation.[Vos06] Einige Beispiele (Übersicht 7.1.3) machen deutlich welcher Art die Kategorien⁵ sind. Auch hierbei wird der erhöhte Anteil an künstlerischen Einflüssen (Musik, Film, Literatur) sichtbar. Wie in Freebase ist diese Domäne besser strukturiert als andere. Einzig die zeitliche Einordnung von Objekten in die Geschichte ist annähernd gleichmächtig vertreten.

- Kategorien mit Jahreszahlen der Formen
 - Geboren im Jahr „<Jahreszahl> births“, Gestorben im Jahr „<Jahreszahl> deaths“, Gründungen „established“
 - Lieder aus einem Jahr „<Jahreszahl> songs“, Musikalben aus einem Jahr „<Jahreszahl> albums“
- Berufsgruppen (Bsp. vor allem Künstler „actors“, „artists“, „composers“, „writers“, „novelists“, „songwriters“, „singers“, „musicians“)
- Ortsbezogene Kategorien (Bsp. „Companies based in North Carolina“, „Geo Coordinates“)

Ein Filter sortiert auch hier die großen Gruppen mit wenig erfolgversprechendem Inhalt aus. Die restlichen knapp 13.000 werden nach ihrer Verwendungshäufigkeit sortiert. Nach ungefähr 500 Einordnungen kommen nur noch sehr vereinzelt neue Kategorien hinzu. Dies ist die Anzahl der Kategorien in der Wikipedia. Die Zuweisung zu den 43 Hauptkategorien

³<http://futef.com/> vom 18.12.2007.

⁴<http://api.futef.com/apidocs.html> vom 17.12.2007.

⁵<http://de.wikipedia.org/wiki/Wikipedia:Kategorien> vom 20.12.2007.

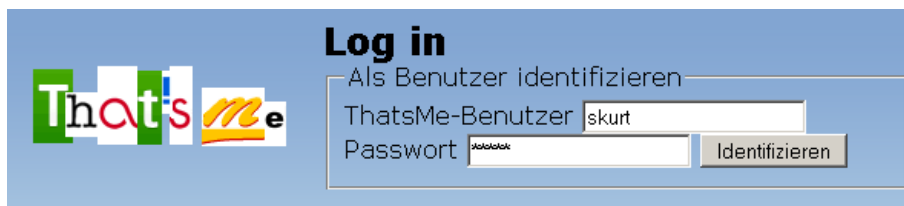


Abbildung 7.1.: Login-Formular *ThatsMe*-Software (eigene Darstellung)

(siehe Übersicht A.2). Schaut man sich die Verteilung, wie oft eine Kategorie einem Tag zugeordnet wurde an, erkennt man das wenigen Kategorien viele Tags (nur die ersten 100 haben zweistellige Anzahlen) und umgekehrt sehr viele Tags (ab Platz 5000) nur einer Kategorie zugeordnet sind.

Die Antwort von Catgraph enthält alle Kategorien die über der angefragten Kategorie stehen, verwertet werden nur die Kinder erster Ordnung der „Main topic classification“. Da der Algorithmus bei nicht eingeschränkter Anzahl den vollständigen Kategoriebaum zurückgibt, wird das Ergebnis auf 100 Rubriken begrenzt. Für die Analyse werden demzufolge drei Ebenen der Kategorien gespeichert. Als Wurzel die „Main topic classification“, darunter die 43 Hauptkategorien, denen zugeordnet die 13.000 Kategorien aus der FUTEF-Anfrage. Im folgenden Abschnitt finden sich Beispiele für Zuordnungen. Bemerkenswert ist die Kategorie „German language“ welche teilweise an deutschen Begriffen vermerkt ist.

7.2. Beispielhafter Ablauf und erhaltenes Profil

Mit der Identifizierung eines Anwenders gegenüber der Software beginnt der Prozess (Abb. 7.1). Anhand der hier angegebenen eindeutigen Kennung werden alle Objekte, die aus den Diensten extrahiert wurden, in der Datenbank einem Nutzer zugeordnet. Um Daten zu einer Person aus einer Quelle zu holen, übergibt der Benutzer (Schritt eins) den dort verwendeten Log-in an das System (Abb. 7.2). Nur bei einer Quelle (Upcoming) ist es zusätzlich notwendig, *ThatsMe* zu autorisieren Daten abzufragen. Je nachdem wie viele Objekte der jeweilige Dienst im Profil des Benutzers vorhält und wie ausgeprägt das Tagging dabei war, erhält *ThatsMe* unterschiedlich zahlreich Tags zur Analyse. In Schritt zwei werden diese genutzt, um die Begriffe mit Hilfe von FUTEF und damit der englischsprachigen Wikipedia Kategorien zuzuordnen (Abs. 5.5). Dabei werden alle bereits eingeordneten Tags, die durch andere Nutzer oder vorausgehende Durchläufe in die Datenbank gelangt sind, nicht erneut behandelt.

Sind alle Tags eines Nutzers den Klassifikationen zugeordnet, zeigt das Interessenprofil die berechnete Zusammensetzung der Interessenschwerpunkte anhand der virtuellen Identität an. Dabei werden die Treffer in Unterkategorien zu denen ihrer jeweiligen Oberkategorien addiert.

Die anfangs angedachte Einordnung mit Hilfe verschiedener Enzyklopädien wurde nach Sichtung der Ergebnisse nicht realisiert. Die Software bietet dem Nutzer daher nicht an, sein ermitteltes Profil mit Kategorien aller Lexika sowie einzelner Zuordnungen aus Freebase bzw. Wikipedia anzuzeigen. Die Möglichkeit zu unterscheiden, ob nur selbstvergebene Tags einbezogen werden oder auch fremde Schlagworte an den eigenen Objekten in die Analyse



Abbildung 7.2.: Eingabeformular für Benutzernamen bei den Diensten (eigene Darstellung)



Abbildung 7.3.: Anzeige der Ergebnisse für Testperson 4 (eigene Darstellung)

einfließen, bleibt bestehen. In der statistischen Auswertung bestimmen nur die eigenen Tags den VIPA.

7.2.1. Pfade zwischen Tags und Kategorien

Die Pfade von Objekten mit deren Tags, über Kategorien aus FUTEF hin zu den Hauptkategorien sind zahlreich (Abb. 7.6. Einige Beispiele für beide Richtungen zeigt der folgende Abschnitt. Dabei muss angemerkt werden, dass auch unpassende oder gar falsche Zuordnungen geschehen und einige Teilgebiete aufgrund ihrer Vielzahl an Einträgen überbewertet sind. Dies korrigiert der Glättungsfaktor (Abb. 7.4), welcher Kategorien abwertet welche eine hohe Anzahl an Unterkategorien haben.

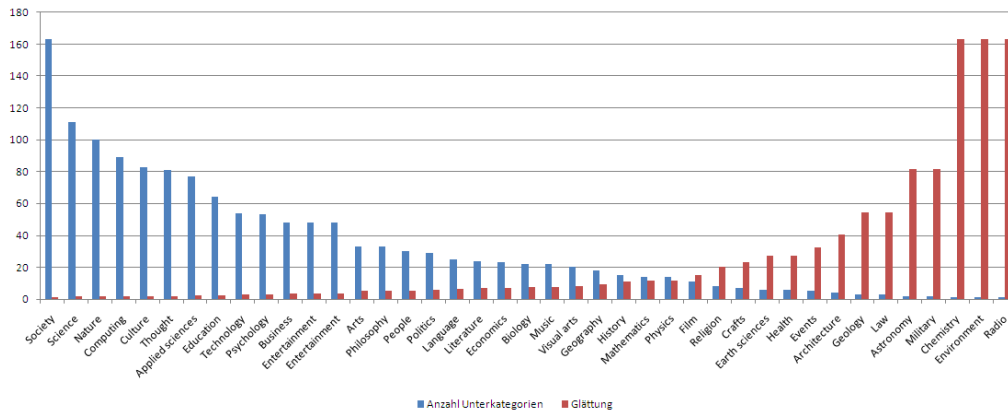


Abbildung 7.4.: Glättung für (vielgenannte) Kategorien (eigene Darstellung)

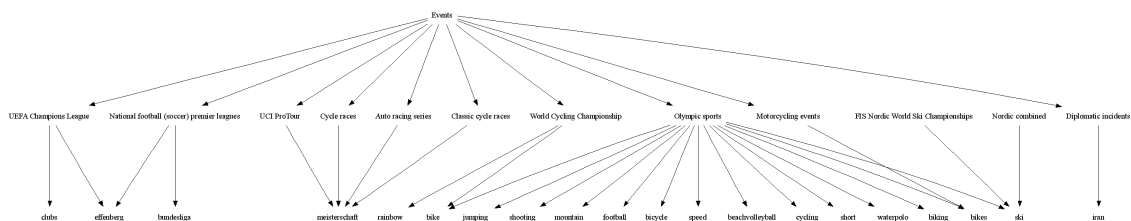


Abbildung 7.5.: Ausschnitte von Unterkategorien der Kategorie *Events* und deren Tags (eigene Darstellung)

Top Down - Hauptkategorien und deren Tags

Der erste Weg führt vom Allgemeinen zum Speziellen. Oberhalb der zahlreichen Unterkategorien abstrahieren die Hauptkategorien noch stärker. Dabei sind viele Details erst sichtbar, wenn man in die Tiefe des Verfahrens eintaucht. Wie die Ergebnisse zeigen, ist bereits ein kleines Beispiel wie die Rubrik „Events“ (ausschnittsweise in Abb. 7.5) mit 199 Paaren von Unterkategorien und Tags sehr differenziert.

Bottom Up - Vom Tag an Objekten zu Hauptkategorien

Die andere Richtung, von unten nach oben beschreibt die Abbildung der Tags und ihrer Resultate in den Kategorien. Hierbei fächert die Struktur nicht so weit auf, da am Ende anstelle von 2.770 Tags nur die 43 Hauptkategorien stehen. In Abb. 7.6 ist ein Beispiel für das Tag „digital“ zu sehen.

7.2.2. Beispiele bemerkenswerter Zuordnungen

Obwohl die Masse der Zuordnungen kaum überschaubar ist, sollen einige einzelne Beispiel genannt werden. Dabei ist zu erkennen, dass auch einzelne deutschsprachige Begriffe mit den englischen Inhalten zugewiesen werden können. So ist bspw. dem Tag „spiegel“ die Unterkategorie „Political scandals“ zugeordnet. Bei den Objekten, die mit diesem Tag ver-

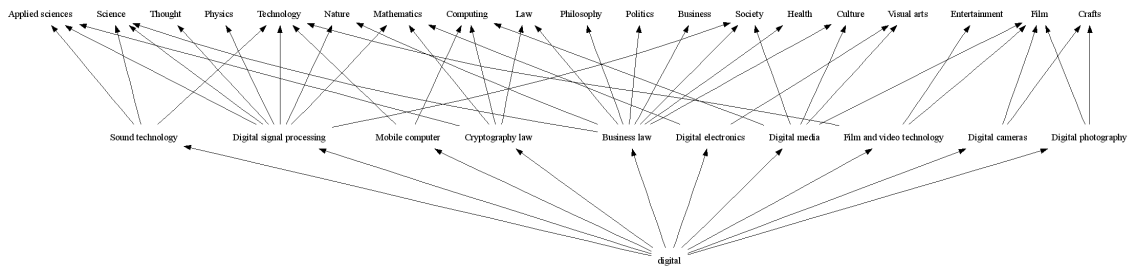


Abbildung 7.6.: Zuordnung von Hauptkategorien zum Tag *digital* (eigene Darstellung)

sehen wurden, handelt es sich um Artikel aus SPIEGEL-Online⁶. Der mehrdeutige Begriff wurde hierbei zufällig mit dem richtigen Spiegel verknüpft.

Namen von Objekten, die nicht als Kategorien auftauchen und somit wegen dem Vergleich von Inhalten und deren Titeln (Methode der FUTEF-API) in den Ergebnissen sind, erhalten ihre passenden Kategorien (Bsp. Tags „surveyor“ bzw. „voyager“ [Marssonden] zu „Current spaceflights“ oder „effenberg“ [Fußballspieler] zu „UEFA Champions League“).

⁶<http://www.spiegel.de/> vom 18.12.2007.

8. Vergleich der Ergebnisse mit realen Interessenprofilen

In diesem Kapitel werden die erhaltenen Ergebnisse der Interessenschwerpunkte in der virtuellen Welt mit denen durch einen Fragebogen eingeschätzten realen Gegenstücken verglichen. Davor steht eine Beurteilung der erhaltenen Resultate auf den Fragebögen. Am Ende werden die Probleme des Verfahrens zur Gewinnung von Interessenprofilen aus virtuellen Identitäten und daraus resultierende Lösungsvorschläge diskutiert.

8.1. Auswertung der Fragebogen

Um das Verfahren neben den theoretischen Betrachtungen auch praktisch zu testen, wurden exemplarisch Testpersonen und ihre Online-Profile verwendet. Bevor die Bewertung der Aussagekraft des virtuellen Interessenprofils geschehen kann, muss eine Referenz zu den realen Interessen hergestellt werden. Den dazu verwendeten Fragebogen bespricht der folgende Abschnitt.

Demographie und andere Faktoren haben Einfluss auf die Interessen einer Person. Ob ein Mensch sich virtuell genauso darstellt, wie er im wirklichen Leben auftritt, ist nicht eindeutig geklärt[Boy07b][Boy07c]. Inwiefern diese Einflüsse bei den verwendeten Testpersonen auftreten, soll ein Fragebogen prüfen. Im Anhang (Abb. A.4) befindet sich eine Kopie des Formulars, welches per E-Mail verschickt wurde.

Mit Fragestellungen in den zwei Bereichen *Allgemein* und *ThatsMe* werden Daten zum Nutzer erhoben. Teil 1 bezieht sich dabei pauschal auf die Nutzung des Internets und dem Verhältnis der Person zu den Daten im World Wide Web. Im zweiten Teil werden Daten abgefragt, die speziell zur Analyse im Zusammenhang mit der Software gebraucht werden. Im folgenden werden die Fragen aufgelistet, erläutert und die Antworten der Tester besprochen. Der genaue Wortlaut der Antworten wurde gespeichert, ist aber aus Platzgründen nicht in der Arbeit enthalten.

Die Fragen wurden vom 24.11.2007 bis zum 27.11.2007 beantwortet. Das Extrahieren der Daten aus den angegebenen Diensten fand zwischen dem 10.12.2007 und dem 13.12.2007 statt. Es ist eine Verschiebung der Interessenschwerpunkt über den Zeitraum möglich. Aufgrund der geringen Anzahl an Daten und dem Fehlen von Eintragsdaten an den gespeicherten Objekten wird diese Einschränkung nicht weiter verfolgt.

8.1.1. Allgemeine Fragen zum Thema

Welche Dienste(Internetseiten) besuchst Du im Internet?

Die genannten Internetseiten zeichnen das breite Spektrum der Dienste ab. Bei den Nachrichtenseiten antworten die Befragten mit „Spiegel, Google News“ (Person 1[P1]), „FAZ.net“ (Person 2[P2]) oder „Spiegel Online, NY Times Online“ (Person 3[P3]). Person 4 (P4) gibt

keine Nachrichten bezogene Seite an. P1, P2 und P3 nennen Blogs als Quellen (P1: "zu Rails, SEO, MacOS X, Startups, Kaiserslautern, Musik", P2: „Bloglines“). Im Bereich Onlineshopping werden „die typischen Onlineshops Ebay, Amazon etc.“ (P1) genannt. Bei den Video-Communities sind „MyVideo, YouTube“ (P3) vertreten. Person 3 nennt auch noch den Dienst Jimdo, mit welchem man eine eigene Homepage ins WWW stellen kann. Die Sozialen Netzwerke nennt nur P3 explizit zu dieser Frage („StudiVZ, Facebook, LinkedIn, Xing,“).

Wie nutzt Du die oben genannten Dienste?

Die zweite Frage zeigt die Verhaltensweisen im World Wide Web. Person 1 ist „bei privaten Blogs [...] recht aktiv (sozialer Kontakt)“. Sie bezeichnet aber „große Blogs [...] als „zu anonym“. Person 2 kommentiert Blogs mit Hilfe des Dienstes „Bloglines“ sowie bei Youtube. Aktiv ist sie unter anderem bei Flickr („Fotos einstellen“), bei Kicktipp.de („tippen“), im Groupwaresystem BSCW und bei LastFM. Person 3 nutzt ebenfalls LastFM („vergebe auch Tags“). Den Social Bookmarking Dienst Delicious nutzt sie zum Speichern, Taggen und Kommunizieren. In den Sozialen Netzwerken hat sie Profile und ist „teilweise auch Mitglied von Gruppen“. „Änderung im Berufsleben“ aktualisiert sie hier auch („außer StudiVZ“).

Welche Daten stellst Du ins WWW? Machst Du Dir Gedanken wer Deine Daten liest?

Alle vier Personen machen sich Gedanken. P1 ist der Meinung „leider wohl viel zu viele: Name, Adresse, Alter, Lebenslauf, Hobbies, Fotos, Videos etc. sind online“. Tarnt diese aber durch „verschiede[ne] Pseudonyme“ und verteilt die Daten. Person 3 stellt „private Daten (persönliche Kontaktdaten, private Fotos, etc.) nur in Social Networks“ und beschränkt die Sichtbarkeit auf ihre Kontakte. Bei beruflichen Daten ist sie weniger vorsichtig („gerne detaillierter“). Person 4 „achte[t] darauf, dass diese Information nicht mit meiner Identität in Verbindung gebracht werden“. Nimmt aber an, dass das „vermutlich gar nicht geht“.

Bist Du dabei ehrlich oder verdrehst Du die Wirklichkeit auch mal ein wenig?

Person 4 hat ihrer Meinung nach nur wenig „preisgegeben“, war dabei dabei aber ehrlich. Person 3 bleibt „bei der Wirklichkeit“, lässt „höchstens etwas weg“. P1 und P2 „verdreh[e]n die Wirklichkeit auch mal ein wenig“(P2). Dabei existiert P1 virtuell unter „falsche[m] Namen“ oder „unterschiedlich[er] Schreibweisen von Ortsnamen“.

Wenn Du wüsstest, dass Deine Interessen im Internet für Jedermann zugänglich sind, würdest du etwas an Deinem Online-Verhalten ändern? Was wäre anders für Dich?

Während P2, P3 und P4 keine Änderungen angeben, denkt Person 1 daran „mehr Hintertüren z.B. anonyme Proxyserver, mehr SSL [zu] benutzen“. Person 3 merkt an nie „private Gedanken und Emotionen bzw. meine subjektive Meinung über andere Menschen“ online zu verbreiten. Person 4 kann sich Missbrauch in einem Maß, dass sie „davon belästigt oder gestört würde“ „noch nicht“ vorstellen.

Tabelle 8.1.: Nutzungshäufigkeiten der Testpersonen für die Dienste (Legende [nach Angaben der Nutzer]: ++ = „oft“, „immer“, „täglich“; + = wöchentlich; o = monatlich, „manchmal“; – = „selten“; -- = keine Angabe, „nie“)

	43Things	Delicious	Digg	Flickr	LastFM	Technorati	Upcoming
P1	--	++	--	++	++	o	--
P2	--	++	--	++	+	o	--
P3	--	+	+	--	+	--	--
P4	--	o	--	-	-	--	--

8.1.2. Fragen mit direktem Bezug zur Analyse

Die folgenden Fragen sollen dazu dienen die realen Interessenschwerpunkte der Testpersonen zu identifizieren. Diese Erkenntnisse erlauben einen Vergleich mit dem virtuellen Gegenstück.

Wenn ja, wie oft benutzt Du die bei *ThatsMe* verwendeten Dienste?

Neben der mit den Antworten beschriebenen Nutzung (Abb. 8.1), gaben die Testpersonen ihre Benutzernamen bei den Diensten für die Analyse bekannt. Die Kombination beider Informationen bietet Daten zur Untersuchung.

Die Nutzung der Quellen 43Things und Technorati mit den generellsten Inhalten ist bei den Testpersonen kaum verbreitet. Dem gegenüber sind Spezialisten (Flickr und LastFM) gut vertreten. Person 4 nutzt die Dienste wenn dann eher passiv („lesend“) und ist daher für eine Analyse nur mäßig geeignet. Person 1 und 2 geben die zeitweise Nutzung von Technorati (P1: „manchmal“, P2: „2x/Monat“) an und sind auch sonst nach eigener Einschätzung für die verwendeten Dienste aktiver als die anderen beiden. Die Angaben von Person 3 zeigen, wenn sie einen Dienst nutzt, dann „mehrmals wöchentlich“. Inwieweit das eigene Urteil zur Nutzung mit den gefundenen Daten übereinstimmt, bespricht Abschnitt 7.

Nenne Bereiche des Lebens für die Du Dich interessierst!

Diese Einschätzung bietet kaum objektiv auswertbare Merkmale, vermittelt aber einen Eindruck über die Person. So sagt Person 4, „Ich bin ein typischer Geek“. Proband 1 beschreibt seine Interessen mit „technisch, Kultur, Reisen, musikalisch“.

Da die hier angegebenen Begriffe aus dem eigenen Vokabular der untersuchten Personen kommen, besteht die Möglichkeit, dass verwendete Tags ähnlich lauten oder auch Kategorien damit übereinstimmen. Dies prüft Abschnitt 8.2.3.

Ordne die folgenden Interessenschwerpunkte Deinen Vorlieben gemäß absteigend!

Die sieben Hauptkategorien wurden nach Analyse der Ober- und Unterkategorien aus Freebase, Wikipedia und Digg aufgestellt. Dabei war eine Aufteilung von „Arts & Entertainment“ in die Bereiche *Kultur* und *Entertainment*, sowie „Technik und Wissen“ in *Technik* und *Wissenschaften* aufgrund der zahlenmäßigen Dominanz notwendig. Die anderen drei Bereiche sind namentlich *Wirtschaft*, *Gesellschaft* und *Sport*.

Tabelle 8.2.: Kategorierangfolge der Interessen der Testpersonen

	P1	P2	P3	P4	Mittelwert
Entertainment	5	2	1	2	2,5
Technik	1	1	5	5	3
Wissenschaften	2	3	6	1	3
Gesellschaft	6	4	2	3	3,75
Kultur	3	5	4	6	4,5
Sport	4	6	7	4	5,25
Wirtschaft	7	7	3	7	6

Mit der Rangfolge der sieben Hauptbereiche (siehe Tabelle 8.2) ordnet jede Testperson die Hauptkategorien nach ihren Präferenzen. Neben den 15 Rubriken aus Frage 8.1.2 sind dies Faktoren mit welchen die Übereinstimmung zwischen realen und virtuellen Vorlieben gemessen wird. Der Bereich *Entertainment* liegt gefolgt von *Technik* und *Wissenschaften* in den Antworten an der Spitze. Danach kommen *Gesellschaft* und *Kultur*. Auf den hinteren Plätzen der Interessen der Testpersonen liegt *Sport* gefolgt von *Wirtschaft*.

Wähle aus den folgenden Bereichen die 15, welche Dich am meisten interessieren!

Die 61 Bereiche stellen eine Auswahl der Kategorien aus den Lexika und Digg dar. Da sich englische und deutsche Begriffe überschneiden, wurden diese Doppelungen entfernt und zusammengefasst (Auflistung siehe Fragebogen 8.1).

Aus 61 Bereichen (Abb. A.4) mussten die Testpersonen 15 (ohne Einordnung nach Präferenz) auswählen. Die Begriffe waren dabei zufällig angeordnet und bezeichneten unterschiedlich große Themenbereiche (Bsp. *Internet* als großer Bereich versus *Kampfkunst*). Teilweise waren dabei auch Kategorien, die hierarchisch auf verschiedenen Stufen stehen (*Spiele* mit der Untermenge *Computerspiele*) oder überlappende Themen als Inhalt haben (Bsp. *Gesundheit und Fitness* versus *Medizin und Gesundheit*), enthalten. Die Antworten zeigt Tabelle 8.3. Wie aus ihnen Erkenntnisse für die Analyse gezogen wurden, beschreibt Abschnitt 8.2.2.

8.1.3. Auswahl von Testpersonen

Die Auswahl der Testpersonen eignet sich nicht dazu allgemeine Aussagen über die Verwendung der Analyse für die Erkennung von Interessenprofilen zu machen. Mit vier Personen, die noch dazu aus einer dem Internet zugewandten Domäne kommen (u.a. Informatiker), ist die Gruppe nicht repräsentativ besetzt.

Mittels einer Themengewichtung der Dienste kann bestimmt werden, welcher Dienst stärkeren Einfluss auf welchen Interessenschwerpunkt hat. Kombiniert mit der Rangfolge der Interessen der Testpersonen zeigt es, dass Personen mit *Entertainment* auf den vorderen Plätzen auch in den Diensten mit diesen Themen aktiv sind (Bsp. P2 bei Flickr und LastFM, P3 bei LastFM). Auch wenn in Delicious (oder anderen Lesezeichen-Diensten) nicht mehr nur Wissenschaftler und Technikbegeisterte zu finden sind, ist ihr Einfluss doch spürbar. Da alle vier Tester diesen Dienst benutzen und *Wissenschaften* und *Technik* gemeinsam auf dem zweiten Platz liegen, wird auch hier die Gewichtung deutlich. Zu den anderen Schwerpunkten und Themen sind Verbindungen nicht sichtbar. Dies liegt zum einen daran, dass bspw. kein Dienst mit dem Fokus auf *Sport* in der Auswahl ist. Das zu

Tabelle 8.3.: Auswahl der Testpersonen von je 15 Bereichen (Frage *ThatsMe.4*)

	P1	P2	P3	P4
Spiele		x	x	x
Filme und Kino	x		x	x
Computer	x	x		x
Internet	x	x	x	
Software	x	x		x
Fotografie und Zeichnungen	x	x		x
Programmieren	x	x		x
Musik	x		x	x
Science Fiction		x		x
Technik/Hardware	x	x		
Essen und Rezepte	x		x	
Gesundheit und Fitness	x		x	
Kommunikation und Sprache		x	x	
Reisen	x		x	
Comics/Comedy		x		x
TV/Fernsehen			x	x
Mathematik		x		x
Computerspiele		x		x
Kultur	x		x	
Geographie und Sehenswürdigkeiten	x		x	
Lehre			x	
Philosophie und Denken		x		
Spielekonsolen				x
Raumfahrt und Luftfahrt	x			
Menschen/Prominente			x	
Religion und Glauben		x		
Umwelt und Natur	x			
Astronomie				x
Wirtschaft			x	
Fußball				x
Gesellschaft		x		
Geschichte			x	

Digg angesprochene Thema Politik, welches sich nach Wikipedia in der Hauptkategorie Gesellschaft einordnet, zeigt ebenfalls kein signifikantes Auftreten bei den Interessen der Testpersonen (4.Platz).

Daten von drei anderen Personen wurden herangezogen um die Ergebnisse besser einordnen zu können und weitere Facetten von *ThatsMe* zu nutzen (keiner der vier Tester nutzt 43Things oder Upcoming). Zum einen der Autor der Studie (markiert mit SK), zum anderen zwei englischsprachige Nutzer deren Benutzernamen zur Verfügung standen. Die beiden letztgenannten Tester füllten keinen Fragebogen aus. Sie sollen zur Überprüfung, inwieweit die Sprache der Tags Einwirkungen auf die Ergebnisse hat, beitragen. Der Autor der Studie füllte nur den zweiten Teil des Fragebogens aus und taucht nicht in alle Statistiken auf, um den Einfluss aufgrund des Hintergrundwissen zu minimieren.

Zahlen zu den Daten der Tester bei den Diensten

In der Tabelle 8.4 wird ersichtlich, in welchem Maß ein Nutzer bei einem Dienst (wenn überhaupt) zum Zeitpunkt der Analyse aktiv war. Dabei wird nur die Quantität der Daten betrachtet, die zeitliche Komponente bleibt außen vor. Bei den Tags wird unterschieden zwischen eigenen und fremden Schlagworten. Die Daten aus LastFM wurden wegen ihrem verstärkten Einfluss im Bereich Musik nicht extrahiert und flossen so nicht in die Auswertung ein.

Tabelle 8.4.: Anzahl von Tags (e = eigene, f = fremde Tags) und Objekten eines Nutzers bei einem Dienst

		P1	P2	P3	P4	SK	Σ
Objekt	43Things	0	0	0	0	21	21
Tag	eigen	0	0	0	0	9	9
	fremd	0	0	0	0	89	89
Objekt	Delicious	50	50	50	50	50	250
Tag	eigen	46	116	126	144	194	626
	fremd	123	106	126	113	104	572
Objekt	Digg	0	30	0	25	108	163
Tag	eigen	0	0	0	0	5	5
	fremd	0	14	0	18	26	58
Objekt	Flickr	100	136	0	100	154	490
Tag	eigen	24	156	0	12	225	417
	fremd	0	372	0	0	564	936
Objekt	LastFM	0	0	0	0	0	0
Tag	eigen	0	0	0	0	0	0
	fremd	0	0	0	0	0	0
Objekt	Technorati	2	2	0	0	3	7
Tag	eigen	9	40	0	0	41	90
	fremd	-	-	-	-	-	-
Objekt	Upcoming	0	0	0	0	4	4
Tag	eigen	0	0	0	0	23	23
	fremd	-	-	-	-	-	-

8.2. Vergleich der realen mit virtuellen Interessenschwerpunkten

Der Vergleich der Interessenschwerpunkte soll prüfen, inwieweit die eigene Einschätzung der Vorlieben mit den virtuellen Aussagen übereinstimmt. Der eigentliche Vergleich von realen und virtuellen Interessen basiert auf drei Ansatzpunkten. Alle geschehen auf der Ebene der 40 Hauptkategorien.

1. Vergleich der sieben Themenschwerpunkte
2. Vergleich der 15 Einzelnennungen
3. Vergleich der frei genannten Interessengebiete

Die Ergebnisse sind mit vier Personen lediglich ein exemplarischer Vergleich, welcher statistisch nicht valide sein kann. Die Korrelationen der Rubriken untereinander werden nicht näher in Betracht gezogen.

8.2.1. Ergebnisse des Vergleiches der sieben Hauptkategorien

Mit der Platzierung der sieben Hauptkategorien, stellte jede Testperson im Fragebogen eine Rangfolge auf. Um diese zu vergleichen, wird das Auftreten einer Kategorie in den Ergebnissen ermittelt. Die Übereinstimmung mit den realen Einschätzungen ist grafisch in den Abb. 8.1(a), 8.1(b), 8.1(c), 8.2(a), 8.2(b) dargestellt.

Bei der Auswertung werden in diesem Bereich verschiedene Merkmale verglichen, erstens die Rangfolge wie sie im Fragebogen notiert wurde, im Vergleich zu den Werten in virtuellen Profilen. Die Durchschnittsdifferenz (diff) gibt dabei an wie reale und virtuelle Einordnung insgesamt zueinander passt. Als zweites welche Bereiche sehr große Unterschiede zwischen VIPA und RIPA aufweisen. Drittens der Trend der Rubriken zueinander, stehen z. B. die gleichen drei auf den ersten Plätze nur untereinander vertauscht.

Tabelle 8.5.: Differenz zwischen virtuell und real (gering ist passender), Mittelwert der virt. Plätze

	p1	p2	p3	p4	sk
Gesamtdifferenz real/virtuell	1,1	2,4	1,6	2,9	2,0
Mittelwert Platzierung virtuell	3,1	2,1	2,9	2,5	2,2

Da *Sport* bei den Kategorien aus Wikipedia keine Verwendung findet und wie in Abschnitt 7.5 beschrieben viele Sportbegriffe den *Events* zugeordnet wurden, stellt diese Kategorie den Ersatz.

Testperson 1

Person 1 hat unter allen Testpersonen die geringsten Differenzen (Gesamtdiff. 1,1) zwischen RIPA und VIPA. Die Bereiche Technik, Wissenschaften, Kultur liegen exakt auf den gleichen Plätzen. Entertainment ist nur um einen Platz unterschiedlich (Abb. 8.1(a)). Vergleicht man die Platzierungen untereinander rangiert Sport treffend an letzter Stelle. Technik ist auch zahlenmäßig gegenüber Wissenschaften im Vorteil. Die anderen vier Kategorien bilden das ausgeglichene Mittelfeld.

Die maximale Anzahl an Kategorien liegt bei P1 weit unter der Hälfte im Vergleich zu gleichen Werten bei den anderen Testern. Daher spielt der Glättungsfaktor (vgl. 7.4) eine übermäßige Rolle.

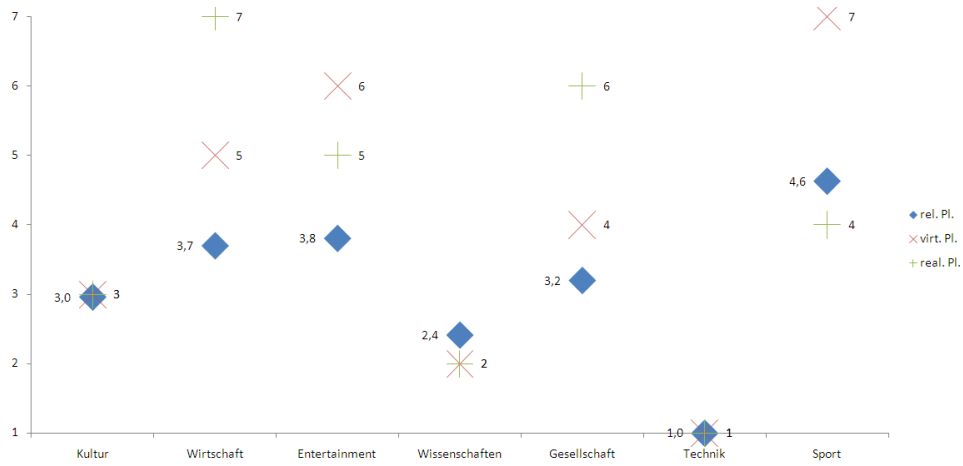
Ein Grund für die Übereinstimmungen bei den sieben Hauptkategorien liegt sicher auch in den Antworten zu Frage 8.1.2 welche die Hauptinteressen hervorheben. Der Einfluss von Wissenschaft hat wohl mit dem derzeitigen Betätigungsfeld zu tun (Tags wie „diplomarbeit“).

Testperson 2

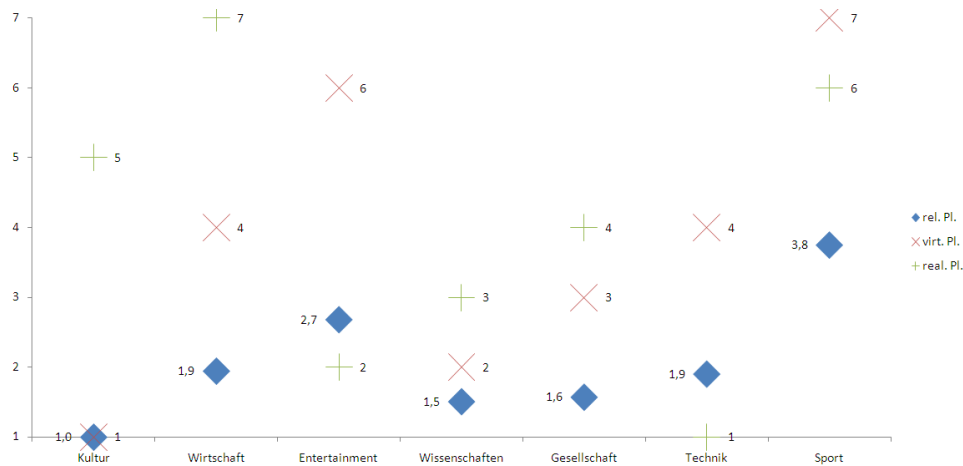
Bei Person 2 sind die Übereinstimmungen weniger deutlich. Auf den vorderen Plätzen liegen Wissenschaft und Gesellschaft nah beieinander. Zur realen Einordnung passend, befindet sich Sport auf den hinteren Plätzen. Auch bei zahlenmäßiger Einordnung liegt es sichtbar zurück. Kultur geht hier knapp als meist eingeordnete Kategorie hervor. Im Mittelfeld besteht keine klare Abgrenzung der anderen Rubriken. Bei Tester 2 tritt eine vergleichsweise hohe Anzahl an Zuordnungen im Bereich Sport auf, was den realen Interessen nicht entspricht.

Testperson 3

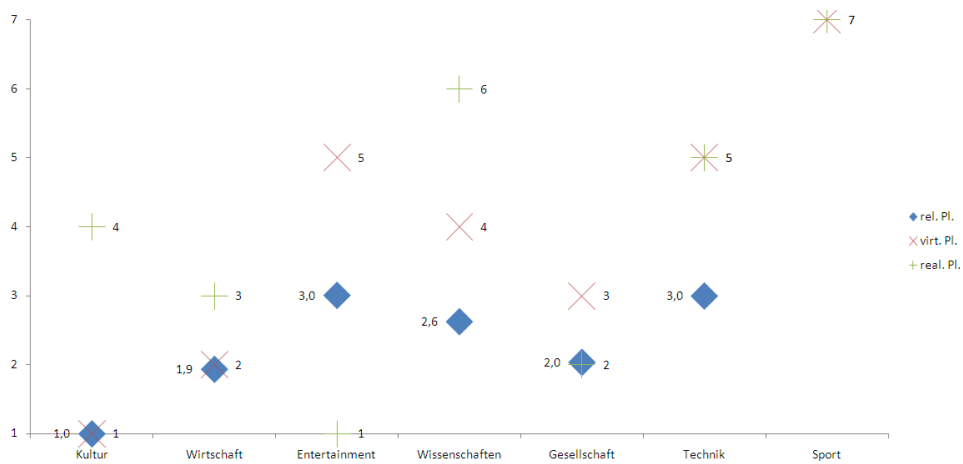
Virtuelle und reale Interessen von Testperson 3 liegen am zweit nächsten (Gesamtdiff. 1,6) beieinander. Deutlich wird das Desinteresse an Sport (jeweils Platz 7), Gesellschaft und Wirtschaft tauschen jeweils die Platzierungen. Die größte Abweichung zeigt Entertainment. Bei Betrachtung der Zahlen hebt sich Sport erneut auf dem letzten Platz hervor. Auch hier findet man ein Mittelfeld ohne große Abweichungen. Kultur hebt sich daraus ein wenig nach vorn ab.



(a) P1

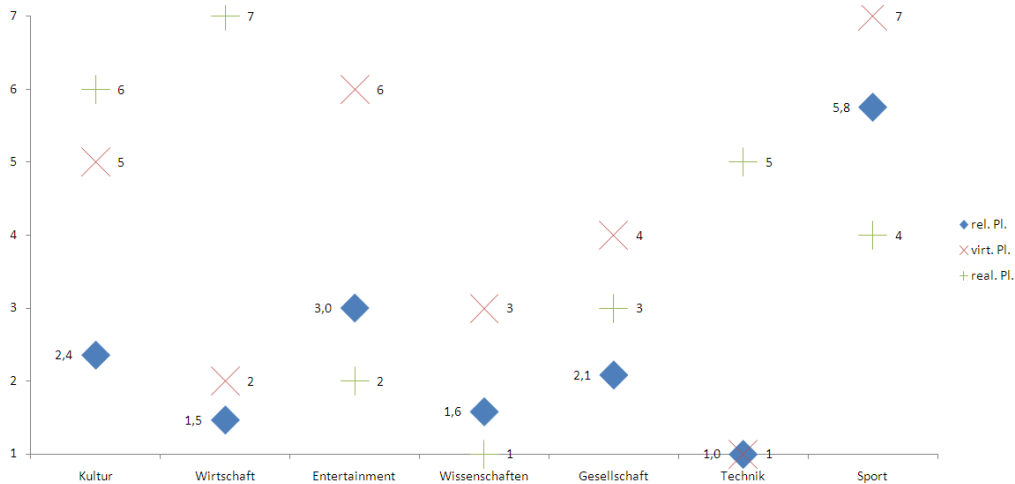


(b) P2

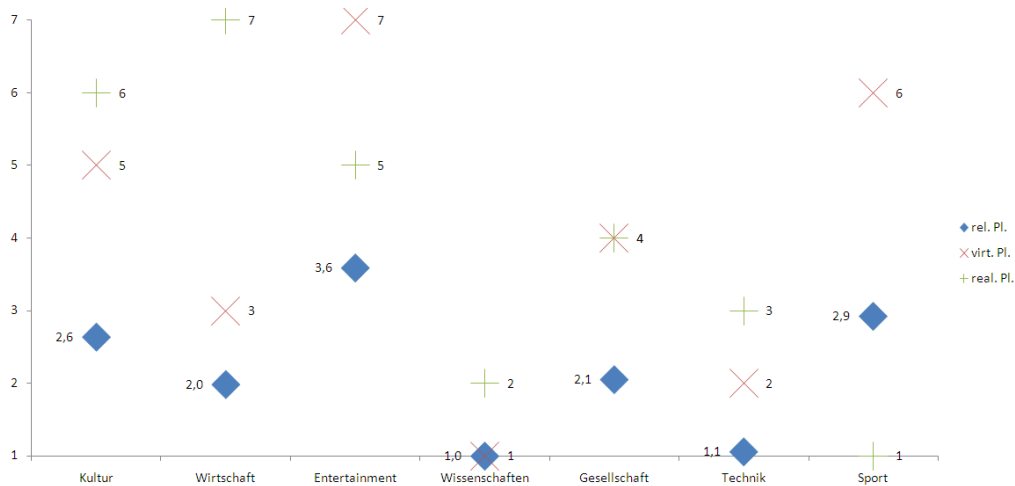


(c) P3

Abbildung 8.1.: Virtuelle vs. reale Platzierungen der sieben Hauptkategorien (eigene Darstellung, rel. Pl. = geglättete relative Platzierung im Vergleich zur Maximumanzahl der ISP dieser Person)



(a) P4



(b) SK

Abbildung 8.2.: Virtuelle vs. reale Platzierungen der sieben Hauptkategorien (eigene Darstellung, rel. Pl. = geglättete relative Platzierung im Vergleich zur Maximumanzahl der ISP dieser Person)

Testperson 4

Bei Tester 4 bestehen die deutlichsten Unterschiede (Gesamtdiff. 2, 9). Nur in den Bereichen Gesellschaft und Kultur liegen die Plätze nah beieinander. Die anderen Einordnungen weisen teils deutliche Abweichungen auf (Differenz Wirtschaft 5, Differenz Technik 4). Die Zahlen sagen aus, dass in der Rubrik Sport Zuordnungen erfolgen. Da allerdings in den anderen Rubriken vergleichsweise viele Kategorien gefunden wurden, wird die absolute Zahl abgewertet. Der Einfluss der Glättung ist hierbei spürbar.

Testperson SK

Nah zusammen liegen bei Tester SK die Bereiche Wissenschaften, Technik und Kultur. Deutliche Abweichungen belegen Sport und Wirtschaft. Die Zahlen in der Kategorie belegen den realen vorderen Platz von Sport (Im Vergleich zu den anderen Testern die meisten Sport-Zuordnungen), aufgrund der Vielzahl in die anderen Bereichen schlägt sich dies aber nicht auf die Platzierung nieder. Die Plätze 1 und 2 sind zu recht an Wissenschaften und Technik vergeben, auch zahlenmäßig heben sie sich hervor. Mit den unterschiedlichsten Einflüssen (alle sieben Dienste) aller betrachteten Personen liegen hier auch die Anzahlen der Zuordnungen der Kategorien absolut gesehen hoch.

8.2.2. Ergebnisse des Vergleiches der 15 ausgewählten Begriffe

Die 61 Bereiche wurden im Fragebogen FAM zufällig angeordnet zur Auswahl gestellt. Die Testpersonen sollten daraus 15 wählen und sie ohne Reihenfolge notieren. Um diese mit den aus der Analyse erhaltenen Ergebnissen zu vergleichen, wurde jeweils das Maximum der Kategorieanzahl (geglättet mit der Anzahl der Kategorien, siehe 7.4) einer Person ermittelt und als 100 % definiert. Mit der virtuellen Entsprechung jeder realen Auswahlmöglichkeit (siehe Abschnitt 8.2.2) wird verglichen. Dabei stehen reale ISP als „1“ und ihre Pendanten mit dem Verhältnis zum Maximum aller Kategorien dieser Person. In den folgenden Abschnitten sind die Profilabdrücke beider Welten jeweils übereinander gelegt.

Virtuelle Entsprechungen realer Angaben

Da zum Zeitpunkt der Erstellung des Fragebogens noch nicht deutlich war, welches die idealen Interessenbeschreibungen zur Auswahl für die Testpersonen sind, wurden 61 breit gestreute Rubriken und Themen aus den Lexika und Digg verwendet. Um die beiden Datensätze zu vergleichen müssen sie miteinander verbunden werden. Dazu wurden den 38 von den Testern verwendeten Rubriken je eine der 43 Hauptkategorien der Wikipedia zugewiesen. Diese Zuordnung war nicht überall eindeutig und selbstverständlich.

So existieren keine Kategorien „Comics“ oder „Science Fiction“ und auch „Spielekonsolen“ stehen als Instanzen verschiedener Gebiete. Bei Wirtschaft wurden erneut „Business“ und „Economics“ zusammengelegt. Ebenso gliedert sich Sport wieder der Kategorie „Events“ an.

Übereinstimmungen und Abweichungen

Das Diagramm 8.3 verdeutlicht es, je nachdem ab welchem relativen Wert (prozentualer Anteil an maximaler Anzahl der Kategorien) ein Gebiet als interessant in einem virtuellen Profil gezählt wird, stimmen umso mehr ISP überein. Den deutlichsten Abfall haben die Kurven im Intervall zwischen 40 % und 50 %. Um möglichst viele Treffer zu erhalten, wurde für die Aussagen über VIPA zu RIPA als untere Grenze die 40 % Marke verwendet. Damit gelangt die Analyse in diesem Bereich zu 65 % (oder 37 von 57 möglichen) Übereinstimmungen.

Diese setzen sich wie folgt zusammen. Am meisten Einfluss (5/5) haben die der Kategorie *Computer* zugeordneten Rubriken. Dies verwundert nicht da in diesem Gebiet zum

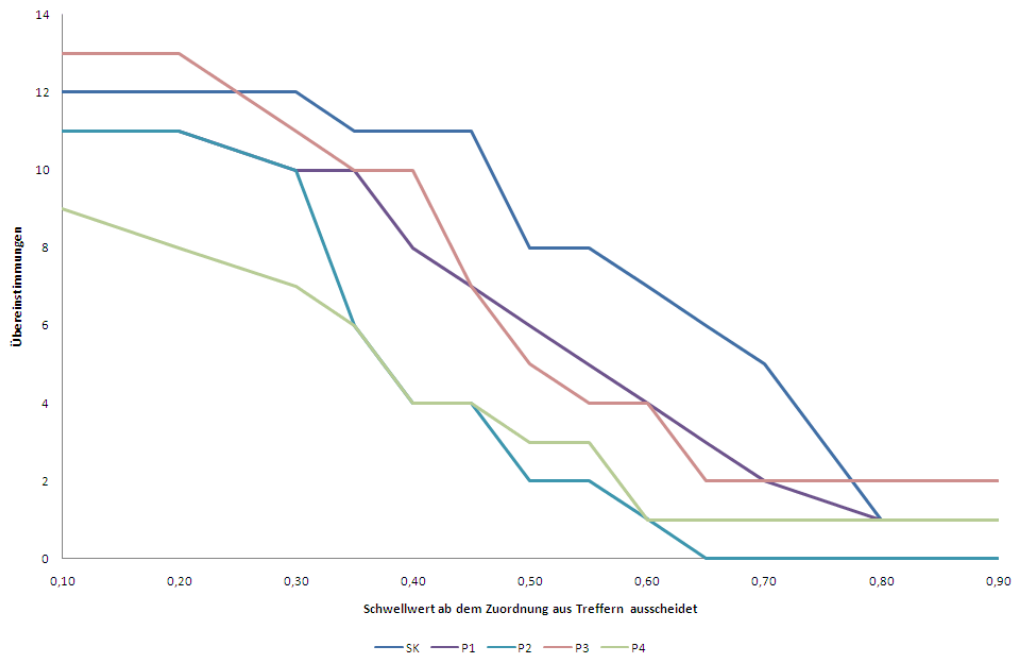


Abbildung 8.3.: Einfluss einer unteren Schwelle auf Anzahl der Übereinstimmungen (eigene Darstellung, Zahlen siehe Tabelle A.1, Darstellung interpoliert)

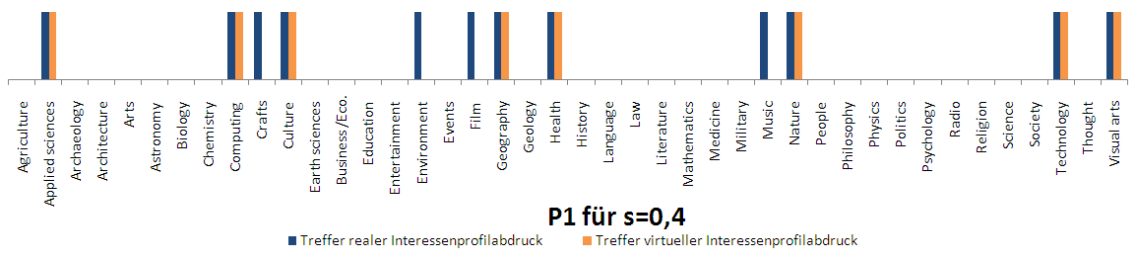
einen viele Objekt aus den Diensten angesiedelt sind und zum anderen gleich sechs Fragebogenauswahlen liegen. Darauf folgen mit 3 von 5 Treffern *Applied Sciences*, *Visual Arts* und *Language*. Dahinter mit je zwei Treffern bei fünf Testern *Culture*, *Business*, *Education*, *Film*, *Geography*, *History*, *Music*, *Nature* und *Technology*.

Die Diagramme 8.4(a), 8.4(b), 8.4(c), 8.4(d) und 8.4(e) stellen bildlich die jeweiligen virtuellen und realen Interessenprofilabdrücke (VIPA und RIPA) dar. Dabei wurde aus Übersichtsgründen auf die Anzeige der FAM verzichtet.

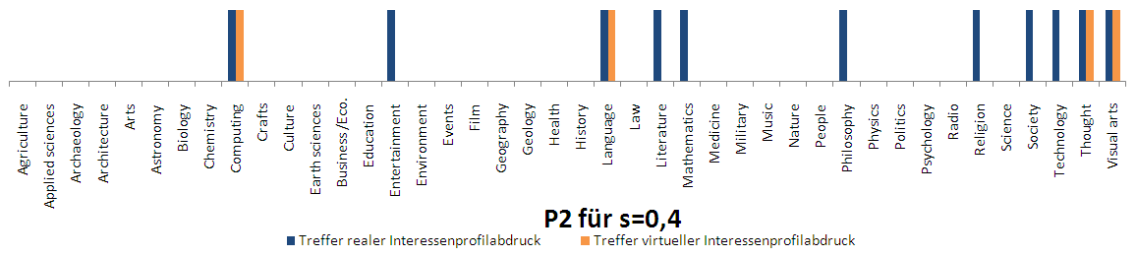
Aufgrund der Einschränkung bei der Abbildung von realen zu virtuellen Rubriken in Abschnitt 8.2.2 minimiert sich die Anzahl der 15 Angaben vom Fragebogen (P1:12, P2:11, P3:13, P4:9 und SK:12). Prozentual die meisten Treffer erhält Person SK (11 von 12 möglichen, 92 %). Darauf folgt P3 mit 77 % (10/13) und P1 mit 67 % (8/12). Virtuell weniger ähnlich wie real denken P4 mit 44 % (4/9) und P2 36 % (4/11).

8.2.3. Ergebnisse des Vergleiches der frei genannten Interessen

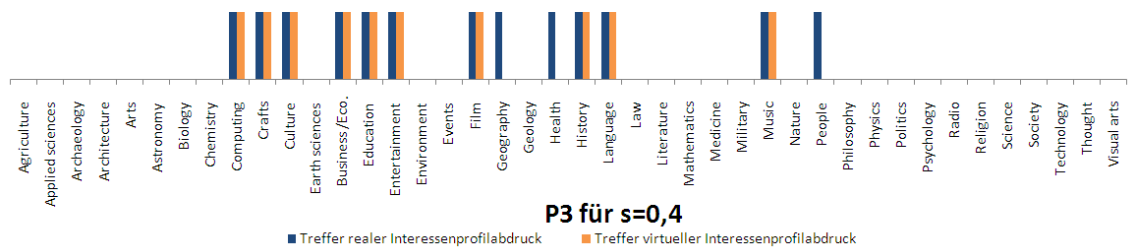
Als dritte Beurteilung der Übereinstimmung werden die Bereiche des Lebens aus Frage 8.1.2 mit den Ergebnissen verglichen. Bei Person 1 finden sich außer „musikalisch“ alle Begriffe wieder. P2 hat nur in den Bereichen Fotografie und Computer Treffer. Die Resultate zu Tester 3 bestätigen die Angaben auf dem Fragebogen. Person 4 und SK machen in diesem Bereich keine verwertbaren Angaben.



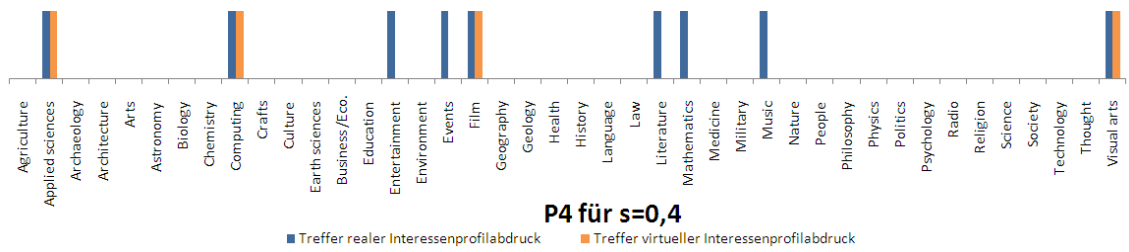
(a) P1



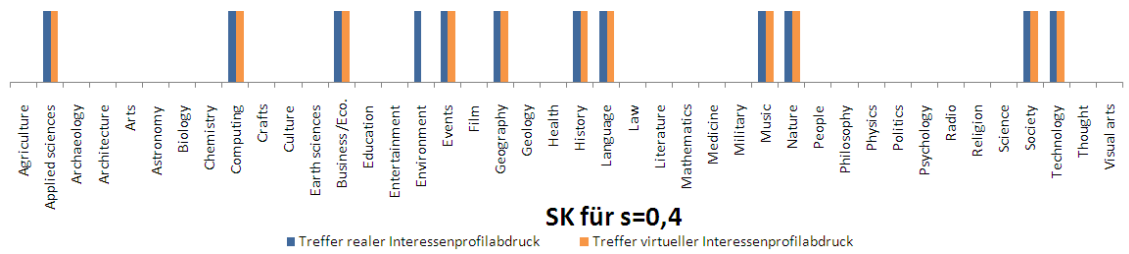
(b) P2



(c) P3



(d) P4



(e) SK

Abbildung 8.4.: VIPA vs. RIPA der Tester mit Schwellenwert s (eigene Darstellung)

8.2.4. Bemerkungen zu ausgefilterten Kategorien

Aus den gefundenen Ergebnissen wurden Kategorien ausgefiltert (Abs. 7.1.3). Zwei Rubriken benötigen davon zusätzliche Betrachtung. Zum einen sind dies Daten aus dem Bereich der Kunst (Film, Musik, Malerei). Ihre zahlenmäßige Stärke rückt die Resultate der Profilerkennung mehr in ihre Richtung. Ein Ausfiltern ist hierbei von Vorteil und verunreinigt nicht ausschlaggebend. So bleibt bspw. „Music“ auch mit den Filterungen im vorderen Drittel der Nennung der Kategorien.

Anders lagert der Fall bei allen geographischen Informationen. Neben Orten von Fotos oder Veranstaltungen besitzen Objekte (Bsp. Reisedaten) Tags, die deren räumliche Zuordnung vermerken sollen. Werden allzu viele geographische Kategorien aussortiert, sinkt deren Einfluss auf den VIPA. Wie die Anzahl der Kategorien in diesem Bereich zeigt („Geographie“ liegt auf dem viertletzten Platz), entspricht dies nicht den erwartenden zahlreichen Nennungen.

8.3. Probleme mit Lösungsvorschlägen

Der folgende Abschnitt beschreibt die Probleme der Analyse und versucht Lösungsmöglichkeiten vorzuschlagen. Dabei sollen hier vorerst nur naheliegende Verbesserungen angesprochen werden. Grundsätzliche Optimierungen bespricht das nächste Kapitel.

8.3.1. Vorteile von vielen Nutzern

Umso mehr Benutzer ihre Profile mit *ThatsMe* analysieren und dazu Tags aus virtuellen Identitäten in das System einfließen, umso schneller und besser kann der Prozess ablaufen. Dies liegt einerseits an der Option, dass weitere Benutzer neue Tags (die beim letzten Abfragen nicht am Objekt standen) zu den bestehenden Gegenständen beisteuern. Andererseits erhöht sich durch die Menge an bestehenden Zuordnungen von Kategorien zu Tags die Chance, das nicht erneut für ein Schlagwort die Lexika angefragt werden müssen. Allerdings ist zu erwähnen, dass Änderungen an den Kategorien in den Enzyklopädien im Laufe der Lebenszeit des Systems derzeit keinen veränderten Einfluss haben, da bereits bestehende Zuweisungen nicht erneut geprüft werden.

Mit fortschreitender Dauer und Einordnung von Tags ist davon auszugehen, dass die begrenzte Anzahl an Kategorien die in den Lexika enthalten sind, zunehmend vollständig in *ThatsMe* gespeichert sind. Auch beim Tag-Vokabular der Nutzer ist nach anfänglich großen jeweiligen Änderungen für neue Nutzer und ihre Diensten eine Sättigung absehbar. Danach kommen nur noch wenige bislang unbekannte Begriffe hinzu. Beide Entwicklungen sind allerdings nicht unbedingt stabil. Die Inhalte und damit die Kategorien in den Lexika sowie die Objekte, welche getaggt werden, sind „User generated content“ (Abs. 4.1.1). Daraus folgt eine Anpassung an aktuelle Themen und Weiterentwicklung der Begrifflichkeiten. Wie in Abschnitt 5.6.1 beschrieben wird bei der Verarbeitung der Anfrage allerdings nur ein Teil der Kategorien in der Antwort verwendet.

8.3.2. Sprache, Schreibweise und Bedeutung

Ein Haupthindernis der Untersuchung stellen die Sprachen der Tags dar. Da nicht ohne weiteres klar ist, ob ein Nutzer Objekte in Englisch oder Deutsch (oder anderen Landes-

sprachen) mit Schlagworten versieht, wird jedes Tag allen Lexika vorgelegt. Dies führt zu Fehlzuordnungen. Eng damit verbunden ist die Problematik der unterschiedlichen Schreibweisen. Neben Einzahl und Mehrzahl, können Abkürzungen oder alternative Rechtschreibungen die Möglichkeiten, einen Begriff in unterschiedlicher Weise zu notieren, steigern.

Ein Versuch, die sprachliche Herkunft eines Begriffs heraus zu finden, indem man ihn bei Google in die Suchmaske eingibt und die daraus resultierenden Top-Level-Domain analysiert, schlägt fehl.[MT05] Algorithmen, die die Sprache von Texten erkennen, untersuchen dies anhand von mindestens fünf Worten. Ein einzelner Tag kann mangels weiterer Informationen nicht zugeordnet werden. Der Kontext am Objekt (mit den anderen Tags) lässt eine Schlussfolgerung auch nur mit bestimmter Wahrscheinlichkeit zu, da für alle Tags am Objekt die gleiche Frage der sprachlichen Herkunft besteht.

Ein Ansatz der Lösung könnte sein, das Vokabular der Tags zu beschränken. Ein Vorteil von Klassifikationen besteht in deren beschränktem Wortschatz, welcher zur Beschreibung von Objekten nur bekannte Begriffe erlaubt. Wären den Nutzern nur bestimmte Tags erlaubt, könnten diese einmalig zugeordnet und damit klar bestimmt werden.

Dieses Konzept erweitert das Metatagging. Die Beschränkung der Auswahl an Worten in Kategorisierungen wird dabei mit dem Konzept der Folksonomy (Abs. 4.1.1) verbunden. Das Vokabular kann erweitert werden, indem Sprache, Schreibweisen und Bedeutungen zu Schlagworten mit jedem Tag verbunden werden (vgl. 9.2.3). Einen Teil dieser Aufgabe erledigen Thesauri (vgl. 4.1.1).

Für die Kategorien sind die Hindernisse mit Sprache, Schreibweise und Semantik geringer, da sie direkt aus den Lexika stammen und dort weitgehend festgeschrieben sind. Dass auch in den Enzyklopädiën Probleme in diesem Bereich bestehen, zeigt das vielfache Auftreten von Weiterleitungen, Begriffserklärungen und die Verwendung der Suchfunktion.

Erstere sind selbst in den Antworten, welche die Kategorie-Einordnung (Abs. 7.1.3) gibt, enthalten (Bsp. Science-Fiction zu Science Fiction). Sei es durch Unterscheidung von Begriffen mit bzw. ohne Bindestrich oder wie bspw. beim Englischen (Bsp. Sport zu Sports), indem für Kategorien meist der Plural verwandt wird¹.

Die Begriffserklärung² wird vor allem für mehrdeutige Worte eingesetzt und ermöglicht es dem menschlichen Benutzer, sich das gewünschte Konzept auszuwählen. Für die automatisierten Anfragen im Rahmen dieser Arbeit stellten Rückfragen der Lexika Hindernisse dar.

Eine Weiterentwicklung für die Lexika wäre die Verknüpfung der Kategorien mehrerer Sprachen zu großen Systematiken. Wikipedia existiert in rund 200 Sprachen, dabei zehn Ausgaben mit mehr als 50.000 Artikeln³. Vergleicht man die Anzahl der 130.000 im Duden[Dud06, S. 5] erfassten Wörter und geht davon aus, dass nicht alle enthaltenen Begriffe auch als Tags verwendet werden, ist eine solche Größe wohl das Mindestmaß für aussagekräftige Zuordnungen. Doch jede Fassung hat ein eigenes Categoriesystem. Der Verknüpfung findet auf der Ebene der Konzepte statt.

¹In der englischen Wikipedia werden Weiterleitungen mit der Systemkategorie „Wikipedia category redirects“ markiert.

²<http://de.wikipedia.org/wiki/API>, Beispiel für den Begriff API vom 14.12.2007.

³<http://www.sap.info/public/DE/de/index/Category-12603c61b2d8e182c-de/-1/articleContainer-11326431f1cbb6b9c7> vom 14.12.2007.

8.3.3. Antwortkategorien

Ein weiteres Problem bereitet die Tatsache, dass Begriffe einer Vielzahl an Kategorien zugeordnet sind. Dabei ist für den Betrachter nicht sofort klar, warum bspw. die englische Ausgabe der Wikipedia zu Volleyball mit 41 Hauptkategorien⁴ antwortet. Dies erschwert eine genaue Zuordnung. In der Anfrage müsste unterschieden werden können, ob ein Begriff in der Kategorie nur erwähnt wird oder eine Teilmenge davon darstellt.

8.3.4. Gewicht von Diensten

Die Gewichte der Dienste lassen sich weitaus detailreicher ermitteln. Dies würde eine gezieltere Einordnung der Resultate ermöglichen. Allerdings beeinträchtigt ein Mangel oder ein Überfluss an einordenbaren Daten deutlich mehr die Analyse.

Ein Möglichkeit an dieser Stelle des Verfahrens zu optimieren, wäre ein standardisierter Testbenutzer. Dieser müsste vordefinierte Interessen (in nur einer Sprache) in den Diensten hinterlassen und damit eine Eichung (auf ein Lexika) ermöglichen. Eine solche Kalibrierung ist aufgrund der Vielzahl an Faktoren nicht realistisch. Allerdings wäre für bestimmte Anwendungsfälle und die damit verbundene Beschränkung auf bestimmte Domänen etwas Derartiges denkbar. Abschnitt 9.2.3 beschreibt was mit der Beschränkung auf Themen möglich wird.

Weitere Faktoren haben Einfluss auf die Aussagekraft (Vgl. 9.2.2). Dies könnte bspw. die letzte Nutzung des Dienstes sein. Umso weiter Aktivitäten zurückliegen, umso weniger Bedeutung muss den daraus resultierenden Interessen beigemessen werden. Schreibt eine Person z. B. an der Diplomarbeit sind die Themen in Delicious darauf orientiert. Nachdem die Arbeit fertiggestellt wurde, nimmt die Aktivität auf diesem Gebiet wieder ab.

Neben den für Interessen von Nutzern nicht interessanten Daten lassen sich aus unterschiedlichen Stücken der Personenprofile verwertbare Informationen gewinnen. So sind neben den Tags auch Kategorien in denen eine Person aktiv ist von Bedeutung. Am Beispiel von Ebay oder anderen Plattformen, zum Thema Handeln können das bestellte Warengruppen sein.

Auch Gruppen stellen relevante Angaben dar. Bei Sozialen Netzwerken können die Nutzer diesen beitreten und damit ausdrücken sich mit dem Thema der Gruppe zu identifizieren. Da Personen auch selbst Gruppen gründen können ist die Anzahl groß und die Unterschiede gering. Neben einigen sehr großen Gruppen existieren unzählige kleine Gruppen in denen sich Menschen mit gleichen Hobbys, Freunden, Angewohnheiten oder anderen Gemeinsamkeiten treffen.

Drei weitere Klassen von Daten sind interessant für Interessenprofile.

Orte

Zum einen besuchte Orte einer Person, dies können Urlaube oder Geschäftsreisen sein. Je nachdem tragen sie bei entsprechender Anzahl zum Interesse „Reisen“ bei oder schränken Vorlieben für bestimmte Länder oder Gegenden ein. Sind sie aussagekräftig genug lassen sich daraus noch weitere Details ableiten.

⁴<http://tools.wikimedia.de/~dapete/catgraph/graph.php?wiki=wikipedia&lang=en&cat=Volleyball&d=0&n=0&format=png&>
vom 11.12.2007.

Zeitangaben

In Quellen sind auch Zeitangaben vorhanden. Diese bringen einen weiteren Faktor in die Betrachtung der Interessen. Der Geschmack einer Person ändert sich im Laufe der Zeit. Es gelangen ständig neue Eindrücke hinzu. Ob die gefundenen Daten wirklich die aktuellen Interessen darstellen ist fraglich. Nicht jedes Profil kann aufgrund der Menge an Daten auf dem laufenden gehalten werden. Zu vielen Angaben und deren Änderungen werden keine Datumsangaben gespeichert oder angezeigt.

Daher werden die Zeitangaben bewusst aus der Betrachtung ausgeklammert.

Kontakte

Die dritte Klasse von Informationen welche Einfluss auf Interessenprofile hat sind Kontakte. Dieses Konzept ist wohl das wichtigste innerhalb von Sozialen Netzwerken. Auf die Vorlieben eines Menschen wirkt es, wenn auch nur indirekt, doch merklich ein. Neben den realen Freunden mit denen man bspw. Schule oder Universität teilt, setzt sich das Personennetzwerk aus virtuellen Freunden zusammen. Aus beiden Gruppen hat man über längere Zeit eher mit denen noch Kontakt mit denen man gleiche Interessen teilt. So kann durchaus geschlussfolgert werden, wenn auch nicht explizit in einem Profil angegeben, dass eine Person aus einem Bekanntenkreis der sich stark für Musik interessiert ebenfalls eine hohe Wahrscheinlichkeit für Interesse an Musik hat.

Wie schon in vorhergehenden Kapitel erwähnt existieren zahlreiche Dienste die sich auf gleiche oder ähnliche Inhalte beziehen. Auch ist nicht immer eindeutig ausgewiesen, dass ein Anbieter nur ein Segment bedient. Dies führt zu einer großen Anzahl an Quellen welche Daten zu Nutzern und Interessen vorhalten. Nicht alle können im Rahmen dieser Arbeit Einfluss in das Interessenprofil haben. Aus der allgemeinen Auflistung stechen einige hervor. Dieser Auswahl ist gemein, ausschlaggebende Vorteile zu besitzen, um Anteil an dem Ergebnis zu nehmen.

In der Arbeit [Kur07, S. 36] befindet sich eine Liste von Beispielprofilen innerhalb unterschiedlicher Dienste. Die dort genannten 16 Kategorien tragen direkt oder indirekt zum Interessenprofil bei. Der Fokus lag bei der Betrachtung aber pauschal auf Personendaten, also eher demografische Faktoren. Fakten zu Interessen

9. Zusammenfassung und Ausblick

In diesem abschließenden Kapitel werden die Ergebnisse der Arbeit zusammengefasst und auf weitere Entwicklungsmöglichkeiten hingewiesen. Dabei sollen nicht nur Optimierungen angesprochen werden, sondern auch abweichende Ansätze, die aus den Erkenntnissen sichtbar werden.

9.1. Zusammenfassung

Die Gewinnung von Interessenprofilen aus virtuellen Identitäten ist machbar. Sind zu einer Person genügend zuordenbare Daten in virtuellen Interessenprofilen vorhanden, bringen Zuordnungen mit Hilfe von Klassifikationen verwertbare Ergebnisse. Soll ein Vergleich von realen und virtuellen Interessenschwerpunkten erfolgen, muss vorab geklärt werden, wie die Zuordnungen der Begrifflichkeiten erfolgen. Passt die Einschätzung der Interessen des wirklichen Lebens nicht zu den Aussagen, die eine Klassifikation zu den virtuellen Gegebenheiten macht, entstehen Schwierigkeiten bei der Auswertung.

9.1.1. Details in „LongTail“ und der Einfluss des „Power law“

Eine automatische Analyse hat gegenüber dem manuellen Verfahren den Vorteil große Datenmengen zu verarbeiten. Dabei fallen Details nicht weiter auf. Der menschliche Betrachter kann aber Feinheiten erkennen, die sich verstecken. Dabei spielen die Potenzgesetze und der sog. „LongTail“ eine Rolle. In diesem Schwanz des Funktionsgraphen befinden sich alle Daten, die nur wenige Nennungen besitzen. Diese Nuancen verschwinden im groben Raster welches die großen Kategorien abzeichnet, sind aber für eine charakteristische Beschreibung der Interessen einer Person wichtig.

9.2. Ausblick

Mit diesem Ausblick sollen die theoretischen Betrachtungen, mit welchen Methoden die Ergebnisse des Verfahrens verbessert werden können, aufgezeigt werden. Ebenso beschrieben werden alternative Anwendungen, welche abweichende Voraussetzungen und Annahmen der Software eröffnen können.

9.2.1. Grundsätzliche Weiterentwicklungen

Neben den in Abschnitt 8.3 beschriebenen Hindernissen des Vergleichs soll der folgende Teil prinzipielle Änderungen am Verfahren besprechen.

9.2.2. Dynamische Bewertung und Rangfolge der Koeffizienten

In dieser Arbeit wurde die Bewertung der Dienste statisch ausgelegt. Mit der Verwendung von dynamisch berechneten Kriterien ließe sich der subjektive Eindruck durch fest definierte Zahlen einschränken. Am Beispiel des Verhältnisses von selbsterstellten zu favorisierten Objekten ist für die Persönlichkeit eine adaptivere Lösung denkbar. Abgestimmt auf den einzelnen Nutzer und seine Verwendung eines Dienstes ließen sich die Ergebnisse optimieren.

Die Rangfolge der Koeffizienten und damit der Multiplikatoren jedes einzelnen können in die Berechnungen einfließen. Weitere Dimensionen eines Dienstes sind die „Vielfalt“ oder der Benutzungsgrad von Tags

9.2.3. Natural Language Processing (NLP)

Die Abstraktion auf Tags vermeidet die Verwendung von weiteren informationsreichen Daten wie Text. Dieser könnte mit Hilfe von NLP analysiert werden und die Daten daraus in die Interessen einfließen.

Standardmodell für Interessen

Der Vergleich als Grundlage der Evaluation des Verfahrens muss einige Hindernisse überwinden. Wie schon festgestellt wurde, existiert kein Standardmodell um Interessen einzuordnen. Daher kann ein Abgleich nur oberflächlich vollzogen werden. Ein Weg, der das Problem löst, besteht darin die realen Interessen des Nutzers nur anhand der Kategorien der Lexika beschreiben zu lassen. Da die Antworten der Tag-Zuordnung passend zu der Selbsteinschätzung der Anwender vorliegen, ist ein Abgleich einfacher. Diese Arbeit macht nur teilweise von dieser Verbesserung Gebrauch. Im Fragebogen werden verschiedene Bereiche unterschiedlicher Klassifikationen vermischt.

Freebase Suggested-Metatagging

Eine Erweiterung des in Abschnitt 8.3.2 besprochenen Metatagging ist das „(Freebase) Suggested-Metatagging“. Dabei wird als Tag nicht ein mehrdeutiger Begriff verwendet, sondern eine eindeutige Identifizierung (ID aus einem Namensraum) um ein Konzept mit einem Objekt zu verbinden. Hinter der ID stecken weitere Informationen zur Semantik und der Beziehung zu anderen Konzepten. Mit Hilfe des Eingabefelds kann aus einer Datenbank ein Begriff vorgeschlagen werden. Für Freebase existiert ein Beispiel¹.

Über Dienste verstreute Nutzer

Die zahlreichen „CopyCats“ (Umsetzungen erfolgreicher Geschäftsideen anderer Märkte für den deutschen Markt) tragen dazu bei, dass die Nutzer selbst in einem Themengebiet mehr oder weniger verstreut über die einzelnen Anbieter sind. Dies erschwert es, möglichst vielen Nutzern die Software auf ihre verwendeten Dienste hin anzubieten. Auch ist eine API bei weitem nicht Standard bei den Anbietern. So ist der einfache Zugang zu den Daten

¹http://mqlx.com/~willmoffat/libs/freebase-suggest/examples/suggest_demo_tagging.html vom 21.12.2007.

versperert. Große Anbieter wie Google mit „OpenSocial“² (in Verbindung mit Myspace und anderen) oder Facebook³ bieten Nutzern die Möglichkeit auf Profildaten zuzugreifen. Dies erleichtert den Zugang zu großen Mengen von Anwendern und ihren Daten.

Beschränkte Domänen

Die Spezialdienste (Bsp. Musik) bieten in ihren Themengebieten einen weit höheren Grad an Detaillierung als dies bei generellen Profilen der Fall sein kann. So zeigt die Einordnung der Tags aus LastFM in Freebase eine recht umfassende Klassifizierung des virtuellen Musikgeschmacks. Anzumerken ist dabei, dass Freebase in diesem Bereich umfangreich strukturiert ist. Natürlich ist es für den einzelnen Dienst mit seinen Nutzern aufgrund der vorhandenen (auch nicht öffentlich zugängigen) Daten um ein Vielfaches leichter ebensolche Erkenntnisse zu gewinnen. Doch zeigt dieser Fall, dass mit den frei verfügbaren Daten von Personen in einigen Bereichen ohne selbstständiges Datensammeln einfach Ergebnisse kombiniert werden können. In Domänen wie Reisen, Sport oder Kunst ist ein ähnliches Vorgehen möglich, wenn dazu entsprechende Dienste sowie gute Klassifikationen vorhanden sind.

Manueller Eingriff

Die automatisierte Verwertung der Tags stößt teilweise an ihre Grenzen. Greift ein erfahrener Benutzer ein, kann die Analyse verbessert werden. Einerseits können Schreibweisen von Tags so optimiert werden, dass sie erfolgreiche Antworten in den Lexika hervorrufen (Bsp. Rechtschreibfehler, ugs. Redewendungen, zusammengesetzte Worte [2006fifaworldcup]). Andererseits schaffen manuelle Einordnungen für Tags denen bislang keine Kategorien zugewiesen werden konnten, Abhilfe. In diesem Bezug lassen sich auch fehlende Einträge von Kategorien oder Konzepten in den Lexika anlegen und mit den jeweiligen Klassifikationen verknüpfen.

Geographische Daten

Die Betrachtung der geographischen Daten bringt weitere Erkenntnisgewinne. So ließen sich unterschiedliche Einordnungen für Angaben zum Heimatort bzw. zu davon abweichenden Orten einrichten. Auf Interessen hat dies nur geringen Einfluss, daher wird es nicht ausführlicher betrachtet.

Weitere Enzyklopädien

Die in Abschnitt 9.2.3 mit der Einschränkung auf bestimmte Domänen möglichen Verbesserungen sind auch denkbar, wenn weitere Lexika einbezogen werden. Dies müssen dabei keine Vertreter universeller Inhalte sein, auch Spezial-Klassifikationen bergen mit dem in ihrem Umfeld verwendeten Vokabular Potential für Optimierungen der Einordnung. Von Vorteil für die einfache Verwendung weiterer Klassifikationen⁴ ist ein einfacher Zugriff (via

²<http://code.google.com/apis/opensocial/> vom 21.12.2007.

³<http://developers.facebook.com/> vom 21.12.2007.

⁴Überblick weiterer Enzyklopädien

<http://de.wikipedia.org/wiki/Wikipedia:Gr%C3%B6%C3%9Fenvergleich> vom 21.12.2007.

API). Zusätzlich sollte auf die Sprache des einzuschätzenden Benutzers mit der Sprache des Lexika eingegangen werden.

Kontext eines Tags

Dieser Abschnitt beschreibt den Kontext, in dem sich ein Tag in Verbindung mit seinen Objekten befindet. Dieser wird im Verfahren weitgehend vernachlässigt, birgt aber einige Informationen die genutzt werden könnten. Denkbar wären Anfragen an Lexika ähnlich denen an Suchmaschinen. Während letztere Antworten die Internetseiten beinhalten, welche die gesuchten Worte aufweisen, sollten ähnliche Anfragen an die Lexika mit mehreren Tags die Kategorien liefern, die diese enthalten. Dabei spielen die Beziehungen der Kategorien eine Rolle.

Problematisch ist dabei die Erkennung des Kontext eines Tags. Nicht zwangsläufig ist dieser durch die anderen Daten am Objekt bestimmt. Mit Datensammlungen lassen sich auch hier Verfeinerungen betreiben. So können mit den „Related Tags“ von Flickr⁵ (Related Tag Browser⁶, Synomiser⁷) oder dem Tool „Google Sets“⁸ Begriffe auf ihre Zusammengehörigkeit hin überprüft werden.

Der Versuch jeweils Paare von Tags in der Anfrage zu benutzen, wurde aufgrund des hohen Aufwandes und der vergleichsweise geringen Erfolgsquote nicht weiterverfolgt.

Vorteile großer Datenbestände

Was bereits in Abschnitt 2 angesprochen wurde, soll hier erweitert werden. Ein Vorteil von Firmen mit großen Datenbeständen (Bsp. Google, Facebook, u.ä.) gegenüber dem verwendeten Verfahren liegt im Zusammensuchen verstreuter Details. Während die Analyse auf unterschiedlichste Quellen zugreift und dabei viele Informationen aufgrund des mangelnden API-Zugriffs, aus Aufwandssicht oder fehlenden Zusatzinformationen übergeht, nutzen große Dateninhaber ihre Vorzüge dabei. Einerseits haben sie die Daten leicht zugänglich in Datenbanken gespeichert, andererseits wissen die Anbieter vollständig über die Struktur der Daten bescheid. Zusätzlich verfügen Firmen wie Google oder Facebook über einen immensen Informationsvorsprung. Sie haben Zugriff auf Daten wie Zugriffszahlen bestimmter Ressourcen. Daraus lässt sich erkennen welches Element von Interesse und Wichtigkeit ist.

Mutmaßung über Geheimdienste oder Behörden und deren Datenbestände werden dabei außer Betracht gelassen. Die Diskussionen zur lebenslang eindeutigen Steuernummer oder der Vorratsdatenspeicherung sind weitere Nuancen der Ansammlung von Daten und der daraus resultierenden Möglichkeiten der Verwertung.

Zeitliche Faktoren

Eine Komponente die nur am Rande betrachtet wurde, ist die Zeit. Während in der Arbeit Interessenschwerpunkte einer Person nur zu genau einem Zeitpunkt erforscht werden konnten, besteht die Möglichkeit dies über längere Zeiträume zu tun. Im Abschnitt 9.2.3 wurden

⁵<http://flickr.com/photos/tags/triathlon/clusters/> vom 21.12.2007.

⁶http://www.airtightinteractive.com/projects/related_tag_browser/app/ vom 21.12.2007.

⁷<http://www.powerhousemuseum.com/dmsblog/index.php/2006/10/23/synonymiser-beta-proof-of-concept/> vom 21.12.2007.

⁸<http://labs.google.com/sets> vom 21.12.2007.

die großen Datensammlungen verschiedener Institutionen erwähnt. Bei ihnen sind Daten von der ersten Registrierung eines Nutzers bis zum Austreten oder Löschen (falls dies möglich ist) vorhanden. Ein plakatives Beispiel für Soziale Netzwerke ist der Beziehungsstatus (Bsp. solo, vergeben). Der Betreiber könnte leicht verfolgen, ob sich die Interessen eines Nutzers ändern, nachdem er seinen Status gewechselt hat.

Die Dauer der Speicherung von Daten zu Benutzer ist in der Diskussion. Einige Firmen (Suchmaschinen⁹) wollen diese einschränken. Damit werden auch deren Möglichkeiten Interessen von Personen über Zeiträume zu analysieren, begrenzt.

Mehr als nur Tags

Werden außer Schlagworten noch mehr Informationen in das Verfahren einbezogen, kann dies zu Verbesserungen führen. So sind Gruppen, denen Personen angehören, Quellen für Aussagen zu Interessen. Auch die Kontakte einer Person bieten zusätzliches Wissen, da teilweise Freunde nur aufgrund gleicher Interessen in den Kontakten anzutreffen sind.

9.2.4. Konzept des Semantic Web

Ein Punkt der bspw. Überlegungen zum Kontext von Tags teilweise inhaltlos macht, ist der Ansatz des Semantischen Web. Da hierbei Bedeutungen von Daten an eben diesen spezifiziert sind, ergeben sich neue Möglichkeiten. Eine Zuordnung wird erleichtert.

9.2.5. Weiterverwendung der Software

Die Software *ThatsMe* ist für das Verfahren nur auf einem Level verwirklicht worden, der einfaches Datensammeln und Auswertungen zu wenigen Nutzern ermöglicht. Dabei war die Bedienung beschränkt, Fehlerprüfung und andere Sicherheitsmaßnahmen wurden nicht implementiert.

Für den weiteren Einsatz bestehen Chancen in verschiedenen Bereichen. So könnten Nutzer ihre Daten beim Dienst speichern. Implementierungen für einzelne Dienste oder abgestimmt auf spezielle Domänen sind möglich.

Für den Anwendungsfall, der die Interessenprofile der Nutzer und deren Verwertung im Ganzen als Ergebnis betrachtet, ist das Verfahren nur beschränkt anwendbar. Allenfalls Besitzer von Personendaten können mit ihrem Wissen und dem Zugriff auf alle Daten Einblicke in die Interessen der Nutzer erhalten. Rechtlich reichen diese Betrachtungen in das Thema des Datenschutzes und sind Schwerpunkt aktueller Diskussionen (StudiVZ¹⁰ oder Facebook¹¹).

⁹<http://www.heise.de/newsticker/meldung/96310/> vom 21.12.2007.

¹⁰<http://www.zweinull.cc/eine-kleine-studivz-presseschau/> vom 21.12.2007.

¹¹<http://www.heise.de/newsticker/meldung/100023> vom 21.12.2007.

A. Anhang mit Bildern und Tabellen

Abbildung auf der Titelseite: Zusammengesetztes Bild aus den Namen der Dienste (eigene Darstellung)



Abbildung A.1.: Statistische Daten in Delicious

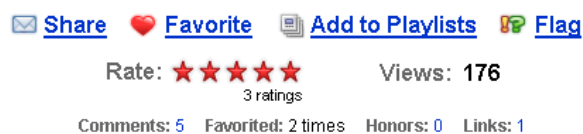


Abbildung A.2.: Statistische Daten in Youtube

Tabelle A.1.: Einfluss einer unteren Schwelle auf Anzahl der Übereinstimmungen (Diagramm dazu 8.3)

x	SK	P1	P2	P3	P4
0,10	12	11	11	13	9
0,20	12	11	11	13	8
0,30	12	10	10	11	7
0,35	11	10	6	10	6
0,40	11	8	4	10	4
0,45	11	7	4	7	4
0,50	8	6	2	5	3
0,55	8	5	2	4	3
0,60	7	4	1	4	1
0,65	6	3	0	2	1
0,70	5	2	0	2	1
0,80	1	1	0	2	1
0,90	1	1	0	2	1

1. **Tagging Rights** - who can tag what?
 - a. Self-tagging - users can only tag their own contributions (e.g. Technorati)
 - b. Permission-based - users decide who can tag their resources (e.g. Flickr), or some other design on the continuum between self-tagging and free-for-all
 - c. Free-for-all - any user can tag any resource
2. **Tagging Support** - how does the interface support tag entry?
 - a. Blind tagging - user cannot see the other tags assigned to the resource they're tagging
 - b. Viewable tagging - users can see the other tags assigned to the resource they're tagging
 - c. Suggestive tagging - user sees suggested tags for the resource they're tagging
3. **Aggregation** - how are tags aggregated for a given resource?
 - a. Bag-model - the same tag can be assigned to a resource multiple times (allowing statistics to be generated and users to see if there is agreement among taggers about the content of the resource). This is very del.icio.us.
 - b. Set-model - a tag can be applied only once to a resource, like in Flickr.
4. **Type of Object** - what kind of resource is being tagged?
 - a. "any object that can be virtually represented can be tagged or used in a tagging system"
5. **Source of Material** - how does the resource get there?
 - a. Supplied by participant - user contributes the resource
 - b. Supplied by system - like in the [ESP Game](#)
6. **Resource Connectivity** - how are resources connected to each other in the system?
 - a. Linked - resources are linked (with hyperlinks or other links)
 - b. Grouped - resources can be assigned to groups (like Flickr groups)
 - c. None
7. **Social Connectivity** - how are users connected to each other in the system?
 - a. Linked - users are connected as contacts, friends or other social links
 - b. Grouped - users can join groups
 - c. None

Abbildung A.3.: Faktoren die Tagging beeinflussen aus http://atomiq.org/archives/2006/12/taxonomy_of_tagging_systems.html vom 01.12.2007, gekürzte Fassung zu [MNBD06]

Fragebogen zur Diplomarbeit „Interessenprofile in virtuellen Identitäten“

Deine Daten werden anonymisiert gespeichert. Bitte antworte in Wortgruppen oder Sätzen.

Allgemein

1. Welche Dienste(Internetseiten) besuchst Du im Internet?
2. Welche davon liest Du nicht nur und benutzt sie auch aktiv, indem du Daten einträgst/kommentierst/änderst? Was machst Du dabei?
3. Welche Daten stellst Du ins World Wide Web? Machst Du Dir Gedanken wer Deine Daten liest?
4. Bist du dabei ehrlich oder verdrehst du die Wirklichkeit auch mal ein wenig?
5. Wenn Du wüsstest, dass Deine Interessen im Internet für Jedermann zugänglich sind, würdest du etwas an Deinem Online-Verhalten ändern? Was wäre anders für Dich?
6. Was ist Dein Beruf? Wie alt bist Du? Mann oder Frau?

ThatsMe¹

1. Wenn ja, wie oft benutzt Du die bei ThatsMe verwendeten Dienste (Flickr, 43Things, LastFM, Technorati, Upcoming, Digg, Delicious)?
2. Nenne Bereiche des Lebens für die Du Dich interessierst!
3. Ordne die folgenden Interessenschwerpunkte Deinen Vorlieben gemäß absteigend (Kultur, Entertainment, Wirtschaft, Gesellschaft, Sport, Technik, Wissenschaften)!
4. Wähle aus den folgenden Bereichen die 15 welche Dich am meisten interessieren:

Religion und Glauben,	Kommunikation und Sprache,	Theater & Oper,	
Planen/Bauen/Architektur,	Fotografie und Zeichnungen,	Computer,	
Gesundheit und Fitness,	Klatsch und Tratsch,	Mathematik,	
Rekorde/Auszeichnungen,	Raumfahrt und Luftfahrt,	Computerspiele,	
Umwelt und Natur,	Essen und Rezepte,	Verkehrsmittel,	
Medizin und Gesundheit,	Wirtschaftsnachrichten,	Kampfkunst,	
Anderer Sport,	Geographie / Sehenswürdigkeiten,	Astronomie,	
Menschen/ Prominente,	TV/Fernsehen,	Chemie,	Spielkonsolen,
Philosophie und Denken,	Fußball,	Hockey,	Filme und Kino,
Finanzen und Geld,	Militärwesen,	Kultur,	Meteorologie,
Technik/Hardware,	Organisationen,	Tennis,	Venture Capital,
Politik,	Science Fiction,	Mythologie,	Programmieren,
Lehre,	Software,	Reisen,	Extremsport,
Recht,	Spiele,	Basketball,	Weltnachrichten,
Internet,	Biologie,	Football,	Geschichte,
Golf,	Musik,	Gesellschaft,	Wirtschaft
	Literatur,		

Vielen Dank für die Beantwortung der Fragen. Die Ergebnisse findest Du auf www.videntity.de.

Sebastian Kurt, kurt@inf.fu-berlin.de

¹ ThatsMe, die Software welche Deine virtuellen Interessen erkennt

Abbildung A.4.: Fragebogen mit dem die Einschätzungen der Testpersonen vorgenommen wurden (eigene Darstellung)

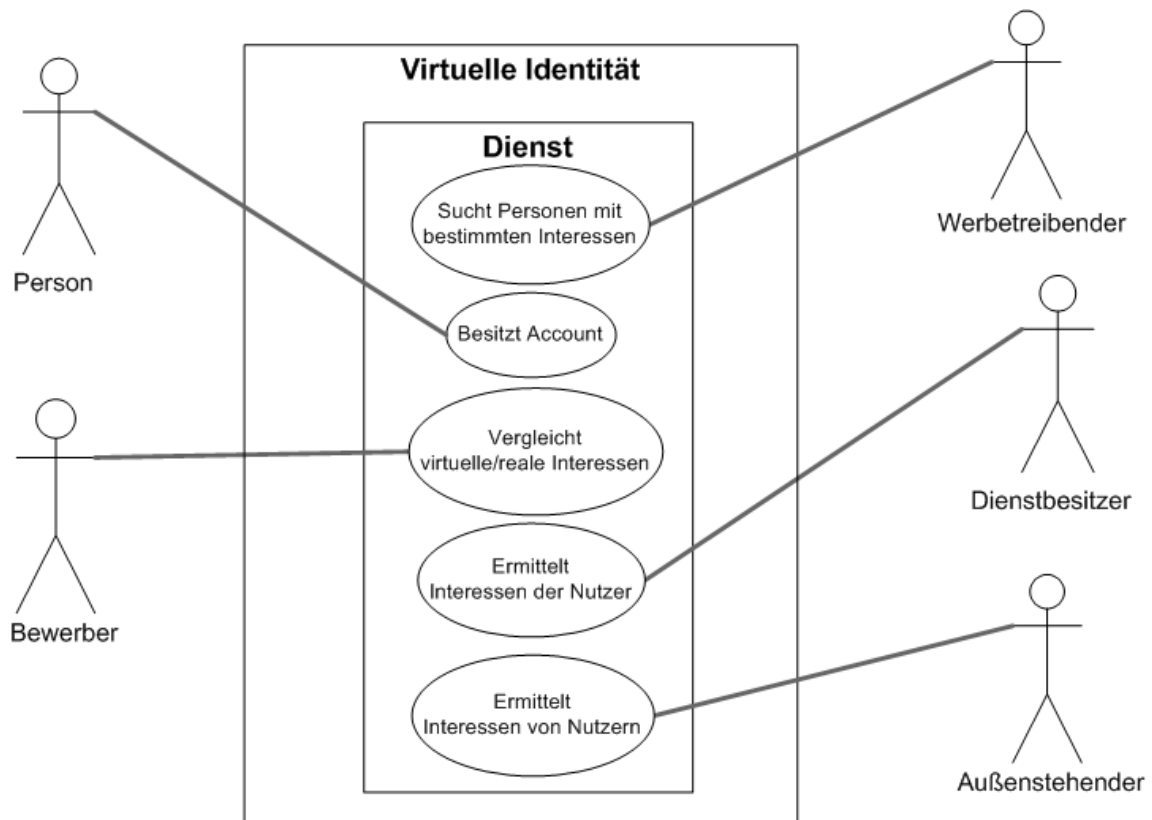


Abbildung A.5.: Anwendungsszenarien (eigene Darstellung)

Tabelle A.2.: 43 Hauptkategorien der englischen Wikipedia (Stand 20.12.2007)

Agriculture	Culture	History	Physics
Applied sciences	Earth sciences	Language	Politics
Archaeology	Economics	Law	Psychology
Architecture	Education	Literature	Radio
Arts	Entertainment	Mathematics	Religion
Astronomy	Environment	Medicine	Science
Biology	Events	Military	Society
Business	Film	Music	Technology
Chemistry	Geography	Nature	Thought
Computing	Geology	People	Visual arts
Crafts	Health	Philosophy	


























Your Name	
	Digg/username
	Flickr/username
	Myspace/username
	Facebook/Your Name
	Friendster/Your Name
	Virb/username
	LinkedIn/Public Profile Name
	Twitter/username
	YouTube/username
	Last.fm/username
	Del.icio.us/username
	Wikipedia/username
	Wishlist/Your Name
	Skype/username
	AIM/screen name
	GMail/Your Name
	coComment/username
	iJigg/username
	PureVolume/username
	Upcoming/Your Name
	Kongregate/username
	Zaadz/username
	Technorati/username
	MyBlogLog/username
	Blog/Your Name

Abbildung A.6.: Beispiele für Dienste (Abb. nach ShowYourself <http://www.dbachrach.com/showyourself/> vom 03.12.2007)

Tabelle A.3.: Gruppierungen im Allgemeiner-Interessen-Struktur-Test

R: Praktisch-technische Interessen (Realistic)
I: Intellektuell-forschende Interessen (Investigative)
A: Künstlerisch-sprachliche Interessen (Artistic)
S: Soziale Interessen (Social)
E: Unternehmerische Interessen (Enterprising)
C: Konventionelle Interessen (Conventional)

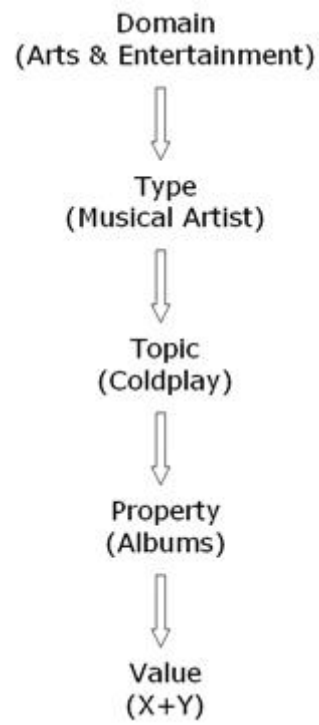


Abbildung A.7.: Aufbau von Freebase mit Beispielen, Abb. nach Carsten Pötter¹

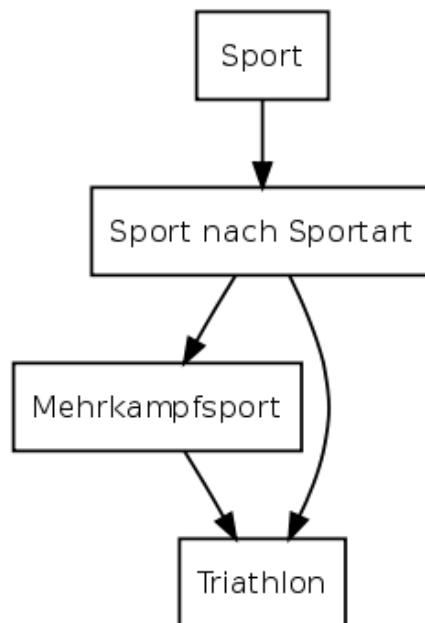


Abbildung A.8.: Kategorien zum Tag *Triathlon* nach CatGraph²

Literaturverzeichnis

- [AG07] AG, PricewaterhouseCoopers: *IFA soll HDTV den Durchbruch bringen*. http://www.pwc.de/portal/pub/!ut/p/kcxml/04_Sj9SPykssy0xPLMnMz0vM0Y_QjzKLd4p3djUBSZnFG8Q76kfCRIL0vfV9PfJzU_UD9AtyI8odHRUVASpBSEg!/delta/base64xml/L3dJdyEvd0ZNQUFzQUMvNElVRS82X0JfQ0VS?siteArea=49c234c4f2195056&content=e5e33b4a6dbce6f&topNavNode=49c4e4a420942bcb. Version: 2007
- [AKD06a] AL-KHALIFA, Hend S. ; DAVIS, Hugh C.: *FolksAnnotation: A Semantic Metadata Tool for Annotating Learning Resources Using Folksonomies and Domain Ontologies*. In: *the Second International IEEE Conference on Innovations in Information Technology*, IEEE Computer Society, 2006
- [AKD06b] AL-KHALIFA, Hend S. ; DAVIS, Hugh C.: *Folksonomies versus automatic keyword extraction: An empirical study*. In: *IADIS Web Applications and Research 2006 (WAR2006)*, 2006
- [AKD07] AL-KHALIFA, Hend S. ; DAVIS, Hugh C.: *Exploring The Value Of Folksonomies For Creating Semantic Metadata*. In: *International Journal on Semantic Web and Information Systems (IJSWIS)* 3 (2007), March, Nr. 1, 13–39. <http://eprints.ecs.soton.ac.uk/13555/>
- [Ale77] ALEXANDER, Christopher: *A Pattern Language: Towns, Buildings, Construction (Center for Environmental Structure Series)*. Oxford University Press, USA, 1977. – ISBN 0195019199
- [AMH73] ANNELIESE MÜLLER-HEGEMANN, Jonny G. Gerhard Butzmann B. Gerhard Butzmann: *Meyers Jugendlexikon*. VEB Bibliographisches Institut, 1973
- [AZ07] ARD ; ZDF: *ARD/ZDF-Onlinestudie 2007*. media perspektiven 8/2007. <http://www.ard-zdf-onlinestudie.de/>. Version: September 2007
- [Bau07] BAUDRILLARD, Jean: *Das System der Dinge*. In: Campus Verlag; Auflage: 3., Aufl., 2007. – ISBN 3593384701
- [BL00] BOWKER, Geoffrey C. ; LEIGH STAR, Susan: *Sorting Things Out: Classification and Its Consequences (Inside Technology)*. The MIT Press, 2000. – ISBN 0262522950
- [Boc74] BOCK, Hans H.: *Automatisch Klassifikation*. Göttingen : Vandenhoeck & Ruprecht in Göttingen, 1974
- [Bor07] BORCHERS, Detlef: *Datenschützer: „Verkettung digitaler Identitäten“ gefährdet Privatsphäre*. <http://www.heise.de/newsticker/meldung/98370>. Version: 2007
- [Boy07a] BOYD, Danah: *None of this is Real: Identity and Participation in Friendster*. in press, 2007 <http://www.danah.org/papers/NoneOfThisIsReal.pdf>
- [Boy07b] BOYD, Danah: *Social Network Sites: Public, Private, or What?* In: *Knowledge Tree* 13 (2007). http://kt.flexiblelearning.net.au/tkt2007/?page_id=28
- [Boy07c] BOYD, Danah: *Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life*. MacArthur Foundation on Digital Learning, Identity Volume (ed. David Buckingham). MIT Press., 2007 <http://www.danah.org/papers/WhyYouthHeart.pdf>

- [Bri07] BRIEGLEB, Volker: Sony baut Videoportal zu Talentschmiede um. In: *Heise Online* (2007). <http://www.heise.de/newsticker/meldung/92749>
- [Bro02] BROUGHTON, Vanda: Facet analytical theory as a basis for a knowledge organization tool in a subject portal. In: LÓPEZ-HUERTAS, María J. (Hrsg.) ; MUÑOZ-FERNÁNDEZ, Francisco J. (Hrsg.): *Challenges in knowledge representation and organization for the 21st century: integration of knowledge across boundaries: Proceedings of the the Seventh International ISKO Conference 10-13 July 2002, Granada, Spain*, Ergon Verlag, July 2002. – ISBN 3899132475, 135-142
- [Bro04] BROUGHTON, Vanda: *Essential classification*. Facet, 2004
- [Chi07] CHITU, Ionut A.: Improving Google's Social Network. In: *Google Operating System* (2007). <http://googlesystem.blogspot.com/2007/07/googles-social-networking-projects.html>
- [DA93] DEMOSKOPIE ALLENSBACH, Institut für: *Politisches Interesse und Entwicklung des Interessenspektrums zwischen dem 20. und 30. Lebensjahr*. Bundesministerium für Frauen und Jugend, 1993
- [DSH07] DATENSCHUTZ SCHLESWIG-HOLSTEIN, PRIME-Projekt und Unabhängiges Landeszentrum f.: „White Paper“ zu *Identitätsmanagement*. <http://www.datenschutzzentrum.de/presse/20070627-prime-whitepaper.htm>. Version: 2007
- [Dud06] DUDENREDAKTION: *Duden*. Bibliographisches Institut & F.A. Brockhaus AG, 2006
- [(ed00] (EDS.), Wojciech H. Kalaga; Tadeusz R.: *Signs of culture: simulacra and the real*. Lang, 2000
- [FM05] FITTKAU, Susanne ; MAASS, Holger: *Weblogs: Ein überschätztes Phänomen?* http://www.fittkaumaass.com/download/W3B21_Studie_Weblog.pdf. Version: 2005
- [Fre79] FREEMAN, Linton C.: Centrality in social networks: Conceptual clarification. In: *Social Networks* 1 (1979), Nr. 3, 215–239. <http://moreno.ss.uci.edu/27.pdf>
- [Gac98] GACKENBACH, Jayne: *Psychology and the Internet : Intrapersonal, Interpersonal, and Transpersonal Implications*. Academic Press, 1998. – ISBN 0122719506
- [GH05] GOLDBERGER, Scott ; HUBERMAN, Bernardo A.: *The Structure of Collaborative Tagging Systems*. <http://arxiv.org/abs/cs.DL/0508082>. Version: Aug 2005
- [Göp07] GÖPFERT, Yvonne: *Online-Nutzer mögen Breitband-Internetzugang*. <http://www.golem.de/0706/53084.html>. Version: 2007
- [Hee04] HEER, Jeffrey: Vizster. Visualizing Online Social Networks. In: *InfoSys 247 Information Visualization* (2004). http://www.cs.berkeley.edu/~jheer/vizster/early_design/
- [Her88] HERMES, Hans-Joachim: *Wissensorganisation im Wandel : Dezimalklassifikation, Thesaurusfragen, Warenklassifikation ; proceedings, Aachen, 29. Juni - 1. Juli 1987*. Indeks Verlag, 1988. – ISBN 3886720187
- [Hül00a] HÜLSMANN, Thorsten: *Geographien des Cyberspace*. Bibliotheks- und Informationssystem, 2000
- [Hül00b] HÜLSMANN, Thorsten: *Geographien des Cyberspace*. Oldenburg : Bibliotheks- und Informationssystem der Carl von Ossietzky Universität Oldenburg, 2000. – 118 Seiten S. – ISBN 3-8142-0756-4

- [Hom06] HOMANN, Meike: *Zielgruppe Jugend im Fokus der Werbung : verbale und visuelle Kodierungsstrategien jugendgerichteter Anzeigenwerbung in England, Deutschland und Spanien*. Kovac, 2006
- [Hun05] HUNTER, Eric J.: *Classification made simple*. Ashgate, 2005
- [IDe07] INITIATIVE D21 E.V., TNS I.: *(N)ONLINER Atlas Deutschlands größte Studie zur Nutzung und Nicht-Nutzung des Internets*. <http://www.nonliner-atlas.de/>. Version: 2007
- [Isk07] ISKOLD, Alex: The Expansion of Social Networks. In: *Read/WriteWeb* (2007). http://www.readwriteweb.com/archives/the_expansion_of_social_networks.php
- [KA06] KALKA, Jochen ; ALLGAYER, Florian: *Zielgruppen*. Moderne Industrie, 2006. – ISBN 3636030663
- [KE04] KEMPER, Alfons ; EICKLER, André: *Datenbanksysteme - Eine Einführung, 5. Auflage*. Oldenbourg, 2004. – ISBN 3-486-27392-2
- [Kra91] KRACKER, Martin: *Vom Nutzen unscharfen Begriffswissens*. In: *Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation*. International Society for Knowledge Organization / Deutsche Sektion, 1991
- [Kre07] KREMPL, Stefan: *Datenschutz beim Identitätsmanagement*. <http://www.heise.de/newsticker/meldung/91826>. Version: 2007
- [Kur07] KURT, Sebastian: Grenzen und Potentiale semantischer Aggregation von personenbezogenen Internet-Daten / XML-Clearinghouse. Version: 2007. http://www.xml-clearinghouse.de/reports/Grenzen_und_Potentiale_semantischer_Aggregation_von_personenbezogenen_Internet-Daten.pdf. 2007. – Forschungsbericht
- [LA05] LAMBIOTTE, R. ; AUSLOOS, M.: *Collaborative tagging as a tripartite network*. <http://arxiv.org/abs/cs.DS/0512090>. Version: December 2005
- [LMD06] LIU, Hugo ; MAES, Pattie ; DAVENPORT, Glorianna: Unraveling the Taste Fabric of Social Networks. In: *Int. J. Semantic Web Inf. Syst.* 2 (2006), Nr. 1, 42-71. <http://web.media.mit.edu/~hugo/publications/drafts/IJSWIS2006-tastefabrics.pdf>
- [LS05] LEHMANN, Kai ; SCHETSCHKE, Michael: *Die Google-Gesellschaft*. Bielefeld : transcript-Verl., 2005
- [McL03] MCLUHAN, Marshall: *Understanding me : lectures and interviews*. McClelland & Stewart, 2003. – ISBN 026213442X
- [Mer04] MERHOLZ, Peter: *adaptive path » metadata for the masses*. <http://www.adaptivepath.com/publications/essays/archives/000361.php>. Version: 2004
- [Meu95] MEUTER, Norbert: *Narrative Identität : das Problem der personalen Identität im Anschluß an Ernst Tugendhat, Niklas Luhmann und Paul Ricoeur*. M und P, Verl. für Wiss. und Forschung, 1995. – ISBN 347645133X
- [Mik94] MIKOS, Lothar: *Zur Popularität von Cyberspace*. In: *Gemeinschaften, Virtuelle Kolonien, Öffentlichkeiten*. Hg. v. Manfred Faßler u. Wulf R. Halbach, 1994
- [Mik05] MIKA, Peter: Ontologies Are Us: A Unified Model of Social Networks and Semantics. In: *International Semantic Web Conference*, Springer, 2005 (LNCS), 522-536
- [MNBD06] MARLOW, Cameron ; NAAMAN, Mor ; BOYD, Danah ; DAVIS, Marc: Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. In: *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, 2006

- [MT05] MAYR, Philipp ; TOSQUES, Fabio: *Google Web APIs – an Instrument for Webometric Analyses?* <http://www.citebase.org/abstract?id=oai:eprints.rclis.org:3704>. Version: 2005
- [Net07] NETWORKS, FOCUS C.: *Communication Networks 11.0 - Deutschlands große Markt-Media-Studie*. <http://wissensforum.medialine.de/2007/09/12/cn-110-mehr-als-314-mio-deutsche-nutzen-blogs/>. Version: 2007
- [OFG07] ONLINE-FORSCHUNG, Arbeitsgemeinschaft ; GMBH, SevenOne I.: *@facts extra Online-Nutzertypen 2007*. http://www.sevenoneinteractive.net/downloads/pods/pID41b5a52ac47d89.66302066/070723_Final_at-facts-extra_NEU-ges.pdf. Version: 2007
- [O’R05] O’REILLY, Tim: *What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software*. <http://www.oreilly.de>. <http://www.oreilly.de/artikel/web20.html>. Version: 2005
- [PBMW98] PAGE, Lawrence ; BRIN, Sergey ; MOTWANI, Rajeev ; WINOGRAD, Terry: The Page-Rank Citation Ranking: Bringing Order to the Web / Stanford Digital Library Technologies Project. Version: 1998. <http://citeseer.ist.psu.edu/page98pagerank.html>. 1998. – Forschungsbericht
- [Pet07] PETERS, Felix: *Partizipatorisches Internet. Von wegen: Dein User ist Stinkfaul!* <http://programm.re-publica.de/programm/events/32.de.html>. Version: 2007
- [Pin05] PIND, Lars: *Folksonomies: How we can improve the tags*. <http://pinds.com/2005/01/23/folksonomies-how-we-can-improve-the-tags>. Version: Januar 2005
- [Pon07] PONTIN, Jason: *Unsere User bestimmen*. Internet. <http://www.heise.de/tr/artikel/93632>. Version: 2007
- [RSM05] RENNER, Karl-Heinz ; SCHÜTZ, Astrid ; MACHILEK, Franz: *Internet und Persönlichkeit*. Hogrefe Verlag GmbH + Co., 2005. – ISBN 3801718522
- [Rüt02] RÜTTGER, Michael: *MySQL für Dummies*. mitp-Verlag, 2002. – ISBN 382663022X
- [Röt07] RÖTZER, Florian: Virtuelle Nachbarschaftshilfe durch „Social Networking“-Websites bei Katastrophen. In: *Heise Online* (2007). <http://www.heise.de/newsticker/meldung/85407>
- [San07] SANDER, Ralf: *Web 2.0 - gern genutzt, aber was ist das?* <http://www.stern.de/computer-technik/internet/591483.html>. Version: 2007
- [Sch06] SCHMIDT, Jan: *Weblogs Eine kommunikationssoziologische Studie*. Uvk, 2006. – ISBN 3896695800
- [Shi05a] SHIRKY, Clay: Folksonomies & Tags: The rise of user-developed classification. In: *IMCExpo*, 2005
- [Shi05b] SHIRKY, Clay: Ontology is Overrated: Categories, Links, and Tags. In: *O’Reilly ETech*, 2005
- [Sin05] SINHA, Rashmi: *A cognitive analysis of tagging (or how the lower cognitive cost of tagging makes it popular)*. http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html. Version: September 2005
- [Sin06] SINHA, Rashmi: *A social analysis of tagging (or how tagging transforms the solitary browsing experience into a social one)*. http://www.rashmisinha.com/archives/06_01/social-tagging.html. Version: Januar 2006

- [Sin07] SINGH, Shiv: Social Networks And Group Formation. Theoretical Concepts to Leverage. In: *Boxes and Arrows* (2007). <http://www.boxesandarrows.com/view/social-networks>
- [SP07] SHNEIDERMAN, Ben ; PREECE, Jennifer: 911.gov. In: *POLICYFORUM* (2007). <http://www.cs.umd.edu/hcil/911gov.pdf>
- [Sum95] SUMMERS, Della: *Longman Dictionary of Contemporary English*. Langenscheidt-Longman GmbH, 1995
- [Tay03] TAYLOR, Arlene G.: *The Organization of Information : Second Edition (Library and Information Science Text Series)*. Libraries Unlimited, 2003. – ISBN 1563089696
- [TD96] TARNAI, Christian ; DOTTERWEICH, Florian: *Anwendung der Latent Class Analyse zur Identifikation typischer Interessenprofile*. Münster : Institut für sozialwissenschaftlich Forschung e.V., 1996
- [Tra06] TRANT, Jennifer: *Social Classification and Folksonomy in Art Museums: early data from the steve.museum tagger prototype*. <http://www.archimuse.com/papers/asist-CR-steve-0611.pdf>. Version: 2006
- [TUD07] TECHNISCHE UNIVERSITÄT DRESDEN, Unabhängige L. u.: *Verkettung digitaler Identitäten*. <https://www.datenschutzzentrum.de/projekte/verkettung/2007-uld-tud-verkettung-digitaler-identitaeten-bmbf.pdf>. Version: 2007
- [Vai00] VAHINGER, Dirk: *Auszug aus der Wirklichkeit : eine Geschichte der Derealisierung vom positivistischen Idealismus bis zur virtuellen Realität*. Fink, 2000. – ISBN 3770534778
- [Van05] VANDER WAL, Thomas: *Folksonomy Definition and Wikipedia*. <http://www.vanderwal.net/essays/051130/folksonomy.pdf>. Version: November 2005
- [Van07] VANDER WAL, Thomas: *Folksonomy*. <http://vanderwal.net/folksonomy.html>. Version: 2007
- [Vos06] VOSS, Jakob: *Collaborative thesaurus tagging the Wikipedia way*. <http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0604036>. Version: 2006
- [WM02] WIDHALM, Richard ; MÜCK, Thomas: *Topic Maps : Semantische Suche im Internet (Xpert.press)*. Springer, 2002. – ISBN 3540417192
- [WW07] WU, Fei ; WELD, Daniel S.: Autonomously semantifying wikipedia. In: *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. New York, NY, USA : ACM, 2007. – ISBN 9781595938039, 41–50
- [ZB07] ZERFASS, Ansgar ; BOGOSYAN, Janine: *Blogstudie 2007 : Informationssuche im Internet - Blogs als neues Recherchetool*. <http://www.blogstudie2007.de/>. Version: February 2007
- [Zel07] ZELENKA, Anne: *Why You May Need an Online Persona*. <http://webworkerdaily.com/2007/03/28/why-you-may-need-an-online-persona/>. Version: 2007

Abbildungsverzeichnis

3.1.	Beispiele für Inhalte von Internetdiensten	13
4.1.	SIGMA Milieus in Deutschland	37
4.2.	Beispiel einer Markenpositionierung mit Infratest	38
5.1.	Rubriken im Internet	44
5.2.	Einfluss der Generalisierung der Objekte eines Dienstes auf die ISP	49
5.3.	Gewicht eines Dienstes als Produkt dreier ausgewählter Faktoren	50
5.4.	Persönlichkeit, Generalisierung und Taghäufigkeit zur Gewichtung eines Dienstes	50
5.5.	Tripartiter Graph	53
5.6.	Schema des <i>ThatsMe</i> -Verfahrens	56
5.7.	Systemkategorien in Freebas	58
6.1.	Entity-Relationship-Modell zu <i>ThatsMe</i>	63
6.2.	Tabellenstruktur in der <i>ThatsMe</i> -Datenbank	64
7.1.	Login-Formular <i>ThatsMe</i> -Software	68
7.2.	Eingabeformular für Benutzernamen bei den Diensten	69
7.3.	Anzeige der Ergebnisse für Testperson 4	69
7.4.	Glättung für (vielgenannte) Kategorien	70
7.5.	Ausschnitte von Unterkategorien der Kategorie <i>Events</i> und deren Tags	70
7.6.	Zuordnung von Hauptkategorien zum Tag <i>digital</i>	71
8.1.	Virtuelle vs. reale Rangfolgen (1)	79
8.2.	Virtuelle vs. reale Rangfolgen (2)	80
8.3.	Einfluss einer unteren Schwelle auf Anzahl der Übereinstimmungen	82
8.4.	VIPA vs. RIPA der Testpersonen	83
A.1.	Statistische Daten in Delicious	93
A.2.	Statistische Daten in Youtube	93
A.3.	Faktoren die Tagging beeinflussen	94
A.4.	Fragebogen mit dem die Einschätzungen der Testpersonen vorgenommen wurden	95
A.5.	Anwendungsszenarien	96
A.6.	Beispiele für Dienste	97
A.7.	Aufbau von Freebase mit Beispielen	98
A.8.	Kategorien zum Tag <i>Triathlon</i>	98

Eidesstattliche Erklärung

Ich versichere hiermit an Eides Statt, dass diese Arbeit von niemand anderem als meiner Person verfasst worden ist. Alle verwendeten Hilfsmittel wie Berichte, Bücher, Internetseiten oder ähnliches sind im Literaturverzeichnis angegeben. Zitate aus fremden Arbeiten sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

Sebastian Kurt Berlin, 20. Dezember 2007

Persönliches Nachwort

Mit den letzten Worten möchte ich mich bei allen bedanken die mich vor und während der Arbeit auf meinem Weg unterstützt haben. Dazu zählen unter anderem meine Eltern, Lisa, Tobias, Daniel, Sebastian, Ronald, Diana, Nils und Olaf. Neben den Testpersonen die zum Erfolg beitrugen, danke ich auch den wissenschaftlichen Mitarbeitern Magnus (Seine Idee gab den Anstoss) und Elena³.

Einen speziellen Gruß an die Organisatoren der *re:publica 2007*. Ohne sie hätte ich den Vortrag von Stephan⁴ nicht gehört und damit auch nie Kontakt mit Tobias (Autor der Diplomarbeit „WhoAmI“) vom DFKI erhalten. Das Bloggen⁵ während der Arbeit hat einige neue Denkanstösse gegeben und mich ermuntert auch weiterhin meine Erkenntnisse mit Interessierten zu teilen.

³<http://www.sti-innsbruck.at/about/team/details/elena-simperl/>.

⁴<http://www.dfki.uni-kl.de/baumann/>.

⁵<http://vIdentity.de>.