15. Buckower Mediengespräche 07. und 08. Oktober 2011
„Medientechnologien versus Handlungsstrategien: Der Spielraum des Rezipienten“

Günther Schatter

Maschinelle Wissensgenerierung in Netzen.

Kollaborative und klandestine Methoden der Konstruktion epistemischer Systeme.

Information is not Knowledge.
Knowledge is not Wisdom.
Wisdom is not Truth.
Truth is not Beauty.
Beauty is not Music.
Music is the best.
(Frank Zappa 1940–1993)

Im Einklang mit dem Titel der Tagung sollen hier neuere technische Entwicklungen im Bereich der netzbasierten Kommunikation danach befragt werden, in welchem Maße Subjekte durch ihr Handeln im Netz an der Generierung von Wissen beteiligt werden. Der mögliche Spielraum, der den Akteuren – also nicht allein den Rezipienten – gegeben ist, wird hier in einem durch zwei Koordinaten beschriebenen Handlungsfeld verortet und beispielhaft dargestellt. Die gegebenen Freiräume und Schranken ziehen Möglichkeiten, aber auch Zumutungen und Bedrohungen nach sich. Verschiedene Anwendungen zeigen, dass gemeinschaftlich erzeugtes Wissen ein hohes Potenzial besitzt, um gesellschaftliche, wissenschaftliche und ökonomische Fragen auf oft überraschende Weise einer Lösung zuzuführen. Andererseits ist die maschinengestützte Wissensgenerierung unter Nutzung kollektiver Ressourcen hinsichtlich ihrer sozio-ökonomischen Konsequenzen weitgehend unerforscht und erfordert vielfältige ethische, juristische, kulturelle und bildungspolitische Antworten und schließlich auch Auskünfte epistemologischer Art zum Verhältnis von Wissen und Wahrheit im Zeitalter aufkommender maschineller Kognition.

Hintergrund

Seit gut zwei Jahrzehnten ist der vernetzte Hypertextdienst World Wide Web als Informations- und Kommunikationssystem fest im Arsenal der Medien verankert. Von Beginn an waren Suchmaschinen erforderlich, um mit Hilfe von Schlüsselwortsuchen bzw. Mustervergleich gewünschte Inhalte in einem verteilten Dokumentensystem näherungsweise ausfindig zu machen. Wegen der Unschärfe der Verfahren geschieht dies bislang üblicherweise durch Bereitstellung von Fundstellennachweisen in Form von Hypertextadressenlisten (URL) und nicht der gewünschten Informationen in Form einer Antwort selbst. Hauptgründe dafür sind das fehlende Verständnis für Bedeutungs- und Sinnzusammenhänge der Sprache durch Maschinen bzw. Algorithmen als auch die Nichtstrukturiertheit auszuwertender Dokumente. Technische und

wissenschaftliche Entwicklungen verschieben auch hier die Grenzen des Machbaren hin zu Bereichen mit mehr Komfort aber auch zu wachsenden Gefährdungen. Davon soll hier berichtet werden.

Beim Begriff des *Data Mining* handelt es sich um statistische Methoden, um in Datenbanken systematisch Wissensmuster u. a. durch Kategorienbildung zu finden und zu extrahieren. Seit dem Jahr 1996 wird vom *Web Mining* gesprochen, wenn Verfahren des Data Mining auf Aktionen, Inhalte und Strukturen des WWW angewendet werden.[1] Vielfältige Forschungsanstrengungen zielen seit Anfang des Jahrtausends darauf ab, das ursprünglich dokumentenorientierte WWW zu einem wissensorientierten System weiter zu entwickeln, um Unzulänglichkeiten zu vermeiden und eine bessere Qualität von Systemreaktionen zu ermöglichen. Eines der nutzbringenden wesentlichen Merkmale – von zahlreichen denkbaren dieser Entwicklung – ist die Ablösung von Such- durch Antwortmaschinen. Dem entspricht der Übergang von verweisenden zu bedeutungsorientierten Strukturen bzw. das Problem der Überführung von unstrukturierter Information in verwertbares Wissen: Aus vorhandenen Daten werden abgeleitete Daten mit Wissensqualität algorithmisch produziert. Begründung finden solche Entwicklungen insbesondere in zwei Tatsachen:

Zum einen ist ein ungebrochenes Wachstum des Datenaufkommens (*Big Data*) zu beobachten, das sich lt. Studien alle zwei Jahre verdoppelt.[2] Neben dem Zuwachs, der durch generierte Inhalte wie dem zunehmenden Bewegtbildtransfer entsteht, ist auf die Durchmischung von Kommunikation und Transaktionen hinzuweisen, die tendenziell stark zunehmen wird. Geschäftliche Transaktionsdaten erzeugen neue Daten, die als Datenschatten oder -spur bezeichnet werden und den doppelten Umfang der von den Nutzer/-innen bewusst erzeugten Daten (Texte, Fotos, Video- oder Musikdateien) aufweisen sollen.[2] Probleme der Verarbeitungskapazität werden sich durch die Integration von verteilten Sensorsystemen evtl. künftig stärker zeigen, wenn Informationen von physischen Objekten in das Internet eingespeist werden. Jene Entwicklungen sind durch die Begriffe *Internet der Dinge* bzw. *Ubiquitous/Pervasive Computing* beschrieben und werden u. a. durch die Internet Protocol Version 6 (IPv6) vorangetrieben. Der steil anwachsende Energiekonsum wird in diesem Zusammenhang oft übersehen.

Weiterhin ist die maschinelle Bedeutungserschließung von Texten und anderen sinntragenden Strukturen (Bilder, Ton, Bewegtbild) aufwändig, langsam und fehleranfällig. Daher liegt es nahe, semantische Informationen nicht erst während der Suche extrahieren zu lassen, sondern Dateien aller Art möglichst frühzeitig mit maschinenlesbaren semantischen Attributen konsequent und einheitlich strukturiert auszustatten.

Seit der Verkündung eines Manifests für ein semantisches Netz[3] im Jahr 2001 sind vielfältige Teilergebnisse zur Durchsetzung wissensorientierter Netze erbracht worden. In erster Linie betrifft dies die Definition von Hilfsmitteln zur formalen Wissensdarstellung und für automatisches Schließen mit Hilfe von Ontologiesprachen. Weiterhin wurden beispielsweise Abfragesprachen zur Wissensbereitstellung und Methoden der Informationsextraktion entwickelt. Mit leistungsfähigen statistischen Algorithmen gelingt es zunehmend, verstreute Daten im World Wide Web automatisch zu erfassen, zu klassifizieren und zu neuartigen Wissenszusammenhängen zu bündeln. Klassische Verfahren der Informatik wie Mustererkennung und maschinelles Lernen werden um neue Methoden ergänzt, die spezifisch für dezentral vernetzte Systeme mit Nutzerinteraktionen sind. All dies ist notwendig, da die in menschlicher Sprache beschriebenen Sachverhalte im WWW in einer von Maschinen verwertbaren Form bereitgestellt werden müssen. Die wissenschaftlich-technischen Grundlagen und Voraussetzungen des *Semantic Web* werden an dieser Stelle nicht weiter verfolgt.[4, 5, 6]

Eine weitere Entwicklung setzte ebenfalls seit Beginn des Jahrtausends ein – die der Herausbildung von Komponenten des *Social Web* (bisher oft Web 2.0).[7] Typisch für die Wissensrepräsentationen in sozialen Netzen ist die Methode einer freien Verschlagwortung (*folksonomy*, *social tagging*). Die Kategorisierung geschieht hier regellos ohne kontrolliertes Vokabular und Indexierungsvorgaben, sondern eher pragmatisch, hierarchiefrei und intuitiv. Daher werden diese Verfahren der Wissensrepräsentation von traditionell arbeitenden Institutionen oft kritisch bis herablassend betrachtet. Dennoch werden hier auf selbstlose Weise wichtige Wissensbestände insbesondere der Populärkultur im Ansatz einer Systematisierung zugeführt, die ansonsten nicht stattfinden würde. Diese Entwicklungen stehen im Zusammenhang mit der häufig zitierten *kollektiven Intelligenz*.[8, 9] Deren analytische Begründung steht zwar noch im großen Maße aus[10], jedoch sind Methoden der automatischen Textanalyse, der Abbildung auf Wortnetze[11, 12] und des Data Mining[13] hilfreich, um solch unscharfe Kategorisierungen zu analysieren und zu konsolidieren.

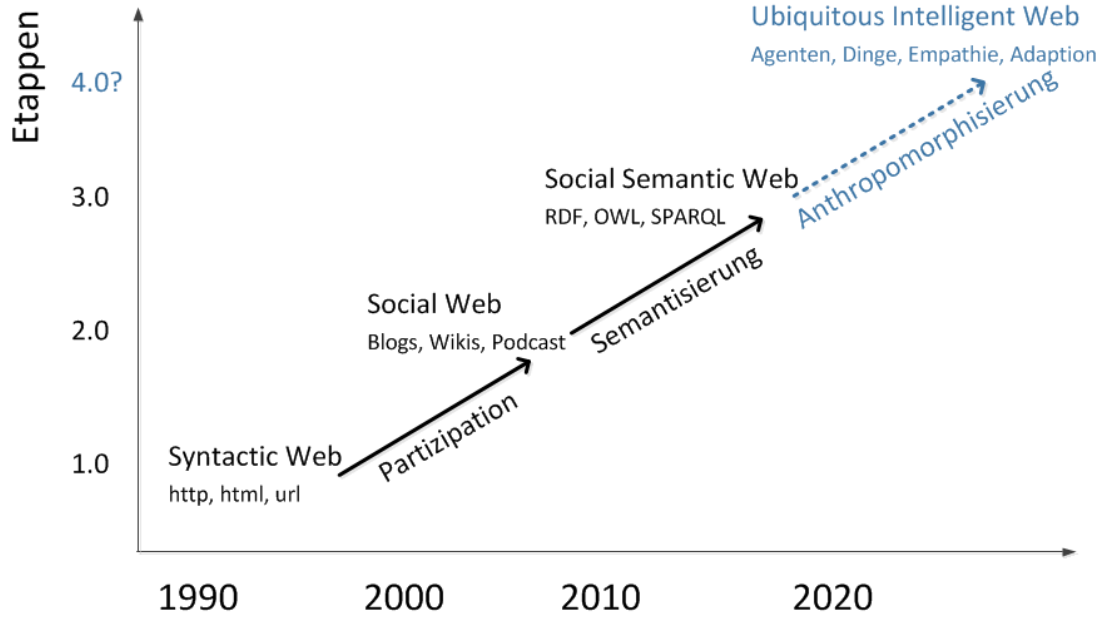

Abb. 1: Etappen der WWW-Entwicklung und -Nutzung

Beide Entwicklungen – also semantische als auch soziale Netze – beginnen sich zu verbinden; Strukturen eines Social Semantic Web werden sichtbar, die gelegentlich als Web 3.0 bezeichnet werden.[14] Die Informatik entwickelt Methoden und Werkzeuge, um die jeweiligen Vorzüge von Ontologien als auch Folksonomien miteinander zu verschmelzen. Die bisherigen Etappen der WWW-Entwicklung werden durch die vorsichtige Annahme einer hypothetischen Phase 4.0 ergänzt und sind schematisch in Abb. 1 dargestellt. Künftige Systeme werden vermutlich durch das Internet der Dinge, einfühlsame Agentensysteme, Assoziationsfähigkeit, multimodale Schnittstellen, einer Lern- und Vergessensfähigkeit u. a. m. gekennzeichnet sein.

Unstrukturierte Daten

Seit Aufkommen des WWW und insbesondere seit seiner Hinwendung zu kommerziellen Anwendungen existieren mannigfaltige Verfahren, um jegliche Nutzerinteraktionen automatisch zu analysieren und daraus Wissen zu extrahieren. Gemeinsam ist diesen Methoden, dass sie meist ohne ausdrückliche Ankündigung

durchgeführt werden und ein stilles Einverständnis der Akteure im WWW voraussetzen. Im Web Mining werden alle denkbaren Handlungen der Nutzer/-innen als auch Inhalte und Strukturen benutzt, um Aussagen über verschiedenste Sachverhalte zu erlangen, Trends abzuschätzen, Modelle zu konstruieren, Kausalitäten zu begründen u. a. m.[15]

Im einfachsten Fall der Wissensgewinnung wird von Daten ausgegangen, die bei Nutzeraktionen von Webbrowsern erfasst werden wie: Zahl und Abfolge der Seitenabrufe, Verweildauer, ggf. Herkunftslink, Version von Browser und Betriebssystem, IP-Adresse mit ggf. Hostname, Sprache, Cookies, mit JavaScript zusätzlich Bildschirmauflösung und installierte Plugins.[16, 18] Provider haben den kompletten Zugang zu den Verhaltensdaten und zusätzlich die zugehörigen Kundendaten, womit sich Geschäftsmodelle für einen Datenhandel begründen lassen. Dieses *Web Usage Mining* nutzt Logdateien von *Profile-Engines* und dient u. a. zur Berechnung von Werbepreisen. Gegenmaßnahmen wie Werbeblocker oder Spurenverwischer können absichtsvoll diese Datensammlung unterlaufen.[17] Die Verknüpfungsstruktur von WWW-Angeboten ist nicht unmittelbar mit Handlungen von Nutzern verbunden, erlaubt aber ebenso eine automatische Auswertung von Informationsbeständen, um Schwerpunkte der Navigation und Nutzung als auch den Wertgehalt von Informationsangeboten einzuschätzen, um ein Ranking oder eine Klassifizierung zu ermöglichen (Webometrie).[15, 16]

Die Häufigkeitsanalyse von Stichworten bei Suchanfragen ist ebenfalls eine frühe Form des Web Mining, um den Bedeutungswandel von nachgefragten Themen zu analysieren. Ein bekanntes Beispiel hierfür ist die Vorhersage von Grippeerkrankungen durch Häufigkeitsanalyse von Begriffen in Suchanfragen, die auf ein Interesse an Erkältungskrankheiten, Medikamenten und Therapien bedeuten. Die Ergebnisse können mit den realen Krankheitszahlen gemeinsam nach Regionen und Zeiträumen aufgerufen und in grafischen Darstellungen veranschaulicht werden.[19] Mit dem Vorlauf einiger Tage sind die Ergebnisse überraschend genau und damit gut als Frühwarnsystem für Epidemien einsetzbar. In ähnlicher Weise kann mit diesen Instrumenten die Nachfrage nach Gütern, Medienprodukten und Ereignissen abgebildet werden.[15, 16]

Das wohl wesentlichste Verfahren automatischer Wissensgenerierung stellt z. Zt. die automatische inhaltliche Erschließung von Webressourcen dar. *Web Content Mining* versucht, auch unstrukturierte Ressourcen zu analysieren. Vor allem kommerzielle Anbieter sind hochgradig interessiert, ihr Marketing zielgerichtet zu entwickeln, um verstohlen aus vielfältigen Quellen Wissen zu extrahieren. Provider können auch E-Mail-Korrespondenzen in Analysen einbeziehen.[20] Unter dem Begriff *Web Analytics* werden Werbemaßnahmen, Kundenprofilerstellung, Warenkorbanalysen, Empfehlungsdienste, Preissuchmaschinen usw. gezielt erstellt.[21] Die scheinbare Objektivität und Selbstlosigkeit von Empfehlungsdiensten und Preissuchmaschinen ist skeptisch zu bewerten, da häufig aus wirtschaftlichen Interessen heraus Angebote selektiv ausgeblendet bzw. propagiert werden.

Ein webbasiertes Bibliometriesystem[22] für die Wissenschaft erlaubt auf statistischer Grundlage die Erstellung von Rankings und Vergleichen für Autoren, Universitäten, Staaten usw. nach der formalen Auswertung von wissenschaftlichen Publikationen und Zitierhäufigkeiten als angenommene Maßzahl für wissenschaftliche Qualität und Produktivität.[23] Die verwendeten Daten sind semistrukturiert, die Ergebnisse erlauben zudem Schlüsse auf Entwicklungstrends der Wissenschaft wie das Anwachsen und Absinken der Konjunktur von wissenschaftlichen Themen, die in grafischen Darstellungen verdeutlicht werden. Quasi im Gegenzug sichern kollaborativ arbeitende Gruppen wissenschaftliche Standards durch die Aufdeckung von

Plagiaten.[24] Amorphe Gruppen, deren Mitglieder sich meist nicht kennen, arbeiten mit temporären klaren Zielstellungen in einer starken Form der Kooperation zusammen – ähnlich wie eine dynamischen Projektgruppe. Mit vergleichbarem Enthusiasmus arbeiten Self-Tracker in losen Netzwerken, um Daten über den eigenen Körper zu gewinnen, die anschließend ausgetauscht werden, um Befindlichkeitsanalysen, Ursachenforschung und Therapieversuche in kollaborativer Weise mit Gleichgesinnten zu unternehmen.[25]

In Prognosebörsen[26] wird versucht, die Verfahren kollektiver Intelligenz[8] für Wirtschaftsvorhersagen (u. a. Arbeitslosenzahlen) zu nutzen. Mit virtuellen Aktien werden Indikatoren handelbar. Im Zentrum steht die Erforschung der möglichen Prognosegüte, jedoch sorgen ausgesetzte Belohnungen für Verzerrungen der Schätzungen. Wichtig ist bei diesen Verfahren die taktische Unverbundenheit der Akteure.[9] Abstimmungen zur Erfassung von Kollektivmeinungen sind als *E-Voting* u. a. für kulturelle, sportliche und politische Fragen zur Publikums- bzw. Meinungserforschung bis hin zur regulären Wahlentscheidung üblich geworden. Simple Formen erschöpfen sich im binären Gefällt mir-Knopf (*Like-Button*), der jedoch trotz oder wegen seiner harmlosen Anmutung in der Lage ist, Wissen über Nutzer auf externen Webseiten und personenbezogene Nutzerprofile herzustellen.[27]

Die Auswertung des Kommunikationsgeschehens sozialer Netze stellt ein völlig neuartiges Hochleistungslabor für empirische Studien der Sozialforschung dar. Hier kann auf die herkömmliche Erhebung von Daten über soziale Tatsachen oft verzichtet werden, lediglich eine statistische Analyse des semantischen Rauschens von halböffentlicher oder publizistischer Kommunikation muss zielgerichtet betrieben werden, um regionales bis globales Bewusstsein widerzuspiegeln zu können. Neben Blogs sind soziale Netze als auch Foren und vor allem Kurznachrichtendienste wie Twitter (seit 13.07.2006) mit seiner Verkürzung und Beschleunigung von Interesse. So können spontane Äußerungen von Menschengruppen quasi in Echtzeit und im bislang unbekanntem Umfang erfasst werden. Daraufhin lassen sich Arbeitshypothesen und Theorien entwickeln bzw. überprüfen oder Steuerungsinstrumente für Entscheidungsprozesse konzipieren.

Am Beispiel der Hedonometrie, d. h. der Glücksvermessung, zeigt sich die Leistungsfähigkeit der Methoden, um den Gemütszustand von Gruppen bis zur Größe von Nationen nach einer Extraktion von Schlüsselwörtern aus unstrukturierten Texten quantitativ darzustellen. In einem solchen Großversuch wurden beispielsweise 4,6 Mrd. Twittertexte von 63 Mio. Nutzer im Zeitraum von 33 Monaten auf ca. 10000 Schlüsselwörter untersucht.[28] Damit sind Tages- bis Jahrestrends, Sprachgebrauch, nationale und kulturelle Besonderheiten im Zusammenhang mit Emotionen im großen Maßstab darstellbar. Das Projekt We feel fine stellt eine Echtzeitvariante dar, welche die aktuelle Befindlichkeit von Menschengruppen auswertet und grafisch veranschaulicht.[29] Ähnlich gelagerte Beispiele sozialwissenschaftlicher Forschungen stellen der Sprachgebrauch in der Literatur[30], die emotionsbasierte Analyse von Liedtexten und Blogs[31] als auch die Vorhersage von Filmerfolgen[32] dar.

Das publizierte Vorgehen und der Umfang der verwendeten Daten der Sozialforschung lassen ahnen, in welchen Maßstäben in Anwendungsfällen jenseits der wissenschaftlichen Erkenntnisgewinnung gearbeitet werden kann. In diesem Zusammenhang ist an die Personenüberwachung als auch an die Marktforschung, die Kundenprofilbestimmung u. a. wirtschaftsgeleitete Interessen zu denken. Im Jahr 2011 wurde der Grimme-Online-Award einem Projekt zuerkannt, das aus den Vorratsdaten eines Telekommunikationsnetzes, die nach sechs Monaten bei einem Mobilfunkbetreiber vorlagen, plastische Bewegungsprofile und Lebensgewohnheiten

erzeugt hat. Eine Datei aus 35831 Zeilen Zeit- und Positionsdaten wurde in eine grafische Animation plastisch rückübersetzt.[33]

Strukturierte Daten

Der Übergang zur Analyse speziell aufbereiteter – d. h. mit Zusatzinformationen angereicherter – Texte und anderer Datentypen ist ein Versprechen höherer Geschwindigkeit und Präzision für den maschinellen Aufbau von strukturierten Faktenbeständen und der damit möglichen Beantwortung von Wissensfragen. Die Repräsentation expliziten Wissens ist mit Kategorien wie Korrektheit, Darstellungskomplexität, Angemessenheit und der Effizienz von möglichen Schlussfolgerungen verbunden.[34] Für die maschinelle Generierung faktenorientierten Wissens sind komplexe Verfahren vorzusehen, die sich mit den Schritten Extraktion von Bedeutung, Disambiguierung, Relationierung von Konzepten und Indexierung in Ontologie-Datenbanken nur grob beschreiben lassen.[6] Das globale Projekt der Herausbildung semantisch aufbereiteter Wissensbestände stellt eine extrem ressourcenaufwändige Anstrengung dar, die zunächst nur für Teilbereiche des WWW umgesetzt werden kann.[35] In vielen Forschungsvorhaben werden automatische Verfahren entwickelt, die durch Lernvorgänge die Bedeutung von Inhalten (Texte, Bilder, Audio- und Videoaufzeichnungen usw.) erfassen und in eine standardisierte Beschreibung transferieren können. In Trainingsmengen müssen jedoch vorab qualifizierte manuelle Annotationen erfolgen, die über den Erfolg automatischer Verfahren später entscheiden.

Die Transformation der Enzyklopädie Wikipedia in eine semantische Version als Semantic Media Wiki demonstriert sowohl Möglichkeiten als auch Probleme der Vorstellung von einem Social Semantic Web. Durch die Integration von Erweiterungen wie Attributes und Typed Links soll es gelingen, Fakten aus unterschiedlichen Lexikoneinträgen unter gemeinsamen Gesichtspunkten der Suche zu betrachten.[36, 37] Informationen können so in verschieden Zusammenhängen mehrfach genutzt werden, indem typisierte Verweise, Seitenattribute und Daten mit vielfältigen Kategorien in Beziehung gesetzt werden oder auch für Sprachmodifikationen bereitgestellt werden. Auf diese Weise könnten künftig auch komplexe Anfragen mit mehreren gekoppelten Fragestellungen erfolgen und ein automatischer Abgleich der Wissensbestände möglich werden. Ein etwas andersartiges Vorhaben stellt DBpedia dar.[38] Hier werden Inhalte aus Wikipedia in strukturierter Form entnommen und mit weiteren Datensätzen aus Datenbanken kombiniert, um dieses Nachschlagewerk integrativ zu erweitern. Dies ist ein typisches Beispiel für den Ansatz *Linked Open Data* (LOD).[39] Um Gruppenstrukturen zu beschreiben, existieren Ontologiebeschreibungen, die geeignet sind, soziale Beziehungen zu modellieren ohne auf kommerziell betriebene Plattformen zurückgreifen zu müssen.[40]

Ein erstes experimentelles Beispiel von Systemen, welches semantische Anfragen erlaubt, ist das im Jahr 2005 aktivierte System Wolfram Alpha.[41] Ein zellulärer Automat erzeugt Datenzusammenstellungen und Grafiken und ist in Teilbereichen recht aussagefähig. Google Squared (03.06.2009–05.09.2011) stellte auf Anfrage eine Starttabelle mit Fakten aus dem WWW zur Verfügung, die vom Nutzer verändert werden konnte[42]. Das System funktionierte mit der Kombination aus zwei Schlüsselworten brauchbar. Das Projekt ReVerb identifiziert und extrahiert logische Beziehungen aus englischen Sätzen als Grundlage automatischer Schlussfolgerungen. Es ist Bestandteil des Projekts KnowItAll, nutzt u. a. die Wissensbestände von Wikipedia und ist in der Lage, auf einfache Fragen zu antworten.[43] Diese Prototypen stellen die Vorboten künftiger in

natürlicher Sprache auskunftsfähiger Systeme dar, die – im Gegensatz zum populären IBM-System Watson[44] – mit speziellen Wissensdatenbanken im WWW verbunden sein werden.

Fazit

In Abb. 2 wird der Versuch einer Systematisierung unternommen. Der Grad der Beteiligung der Nutzer wird in der Horizontale in qualitativen Schritten durch einen Kategorienbezug aufgetragen, der einem zunehmenden Grad an absichtsvollem Handeln bzw. Bewusstheit entspricht. Die Attribute bewegen sich von ungewollter Beobachtung des Handelns, über Gleichgültigkeit bis zu zielgerichtetem Tun. Auf der Ordinate ist die Qualität der Nutzerbeteiligung wiederum in diskreten Qualitätsschritten mit zunehmendem Anspruchsniveau aufgetragen. Diese bewegt sich zwischen einfachen Navigations- und Suchvorgängen über Abstimmungen und textbasierten Meinungsäußerungen bis hin zur Bereitstellung systematischer Beiträge und wissenschaftlicher Texte.

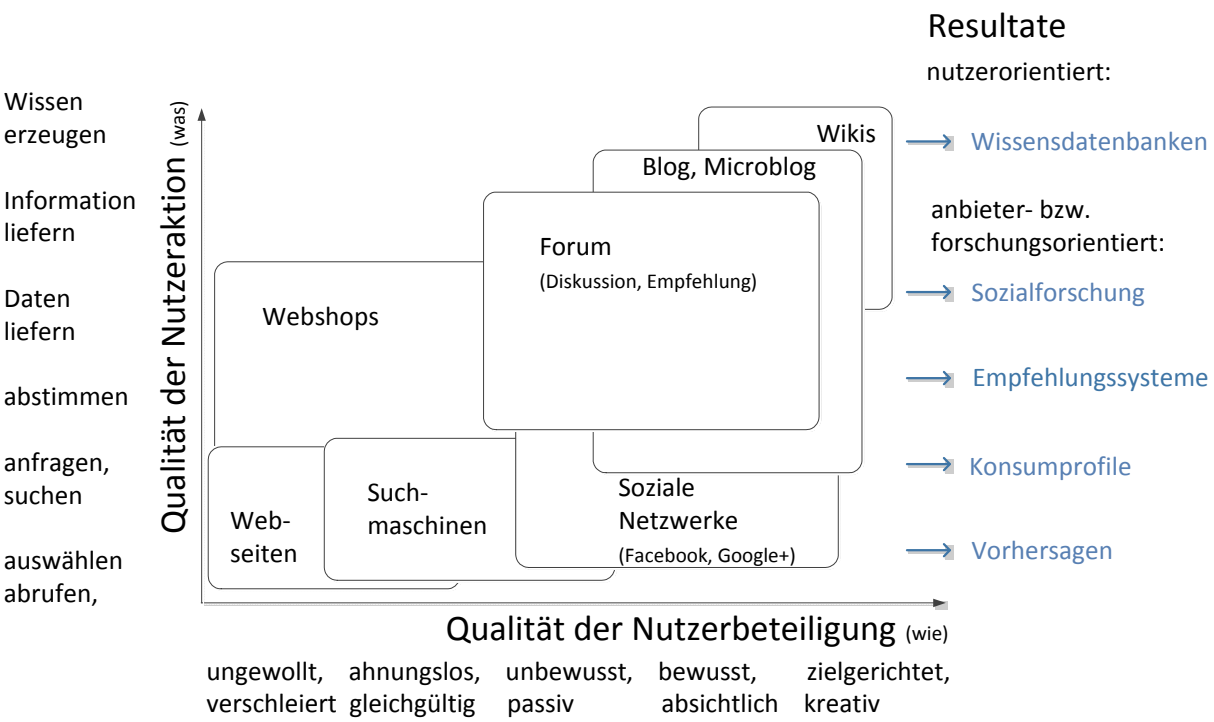

Abb. 2: Wissensgenerierung in Netzen in Abhängigkeit von Nutzerbeteiligung und Nutzerinteraktion

Im zweidimensionalen Lösungsraum finden sich Hauptgruppen von Anwendungen, die den Variablenkombinationen schematisch durch eine Flächenzuweisung entsprechen. Aus den sich überlappenden Anwendungen lassen sich verschiedene Wissensrepräsentationen ableiten, die zwei Kategorien von Resultaten zugeordnet werden. Ursprünglich wurden einfache unstrukturierte Daten dazu herangezogen, um für Dienstanbieter Akzeptanzwerte bestimmen zu können. Die quantitative Forschung nutzt zunehmend unstrukturierte Informationen z. B. für Trendberechnungen; aus strukturierten Wissensbeständen werden künftig nutzerorientierte Antworten gewonnen werden, wie in den vorangegangenen Abschnitten erläutert wurde.

Ausblick

Forschungsaufgaben sind in großer Zahl vorhanden, um Verfahren der Wissensgenerierung zu vervollkommnen. Neben der Erhöhung von Genauigkeit und Relevanz bei der Faktengewinnung, sind viele Detailfragen zu lösen. Unscharfe und unvollständige zeitliche Faktenbezüge müssen geklärt und aktualisiert werden. Das Altern von Fakten und damit verbundene Aktualitätsbezüge wurden bislang nur wenig berücksichtigt, die Mehrdeutigkeit von Aussagen ist zu lösen. Neben der Textverarbeitung warten weitere Mediengattungen auf automatische Erschließung und Annotation wie Audioarchive, Fernsehbeiträge, Filme usw. Damit steht auch eine Diskussion an, welche die Zukunft, Dauerhaftigkeit und Funktionsbestimmung von traditionellen Archiven, Sammlungen, Bibliotheken, Medienzentren usw. betrifft. Weiterhin stellen sich Fragen zur Zuverlässigkeit, zur Überprüfung und zu Korrekturverfahren künftiger maschineller Wissensdatenbanken.

Um eine nahtlose Kommunikation mit Menschen zu ermöglichen, müssen die Schnittstellen zur Technik verbessert werden. Hierzu zählen die Verwendung natürlicher Sprache einschließlich Fremdsprachen, multimodale Schnittstellen, das inhaltliches Verstehen von Fragen, die Nutzung einer Dialoggeschichte, die vorhergehende Anfragen berücksichtigt, eine Gedächtnisfunktion und das Zulassen mehrstelliger Fragerelationen. Die Informatik hat in Kooperation mit den epistemologischen Disziplinen auch bedeutende Probleme erkenntnistheoretischer Art praktisch zu bearbeiten. Wie kann Wissen, das nicht expliziter faktenorientierter Natur ist, kognitiv modelliert werden? Wie kann es gelingen, angeborenes (natives) Wissen in Bestände zu integrieren?[45] Weiter stellen sich Fragen nach der Einbeziehung von implizitem, episodischem oder auch von handlungsorientiertem prozeduralen Wissen.

Mit Hilfe des Web Minings lassen sich viele nützliche Aussagen sowohl für Anbieter, Nutzer und Forscher gewinnen. Mit der Verfeinerung der Methoden ist jedoch anzunehmen, dass durch Aggregatoren auch Ungewolltes aufgedeckt werden kann, das in die Persönlichkeitsrechte von Subjekten eingreifen kann und über die bislang üblichen Fragestellungen des Datenschutzes weit hinausgeht. Darauf weisen beispielsweise aktuelle Personensuchmaschinen hin, die Kontaktdaten und Ansätze für Persönlichkeitsprofile erzeugen. Diese Algorithmen sind prinzipiell zu weit höherer Leistung als gegenwärtig in der Lage, wenn Techniken des maschinellen Lernens, Geodaten, Wissensdatenbanken usw. mit vorgefunden Personeninformationen kombiniert werden. Damit ließen sich Spekulationen und hypothetische Aussagen gewinnen. Maschinen werden bald auch öffentlich Schlüsse ziehen können. So ist nicht undenkbar, dass Risikoberechnungen zur Finanz- und Gesundheitssituation, zu Kriminalitätsgefahren, zum partnerschaftlichen Verhalten (Treueprognose) und zu Lebensgewohnheiten von Menschen ggf. als Dienstleistung angeboten werden. Mediale Fingerabdrücke, Gesichtserkennung, die Identifikation von Sprechern und Autoren gelangen in allgemeine Reichweite, um einfache Datensammlungen durch Individualitätsmerkmale zu bereichern, womit vielfältige Negativutopien denkbar sind. Gefahren der Verwechslung aber auch der Denunziation sind sehr wahrscheinlich, aber auch versehentliche Veröffentlichung, gegenseitige Überwachung und Verdächtigung sind im Bereich des Möglichen.

Der Datenschutz wurde bislang meist aus einem Ansatz betrieben, Bürger/-innen vor den Zumutungen des Staates, von Organisationen oder von Unternehmen zu bewahren. Im Umfeld des Datenschutzes werden nun aber auch Perspektiven einer anderen Form einer Dystopie diskutiert, die mit der Verfügbarkeit von Techniken und Daten für viele zusammenhängt. In welchem Maße werden Bürger Daten über andere Bürger sammeln, sie mit technischen Mitteln beobachten? Kommt die Rasterfahndung für alle, wie kann die Kontrolle der Kontrolleure erfolgen?[46, 47] Vermutlich werden sich neuartige ethische Fragestellungen im Anschluss weiterer

technischer Entwicklungen ergeben und Korrekturen der Technik erfordern, damit deren Vorteile im gesellschaftlichen Konsens genutzt werden können.

Bei all dem stellen sich aber auch Fragen nach der Gläubigkeit, die eines Tages dem maschinell dargebotenem Wissen entgegengebracht werden wird und in welcher Weise es für welche Zwecke eingesetzt werden soll? Wo gibt es unabweisbare Notwendigkeiten für automatisierte Wissenssysteme, wo liegen die Schranken? Aber auch Probleme des Vertrauens und der Skepsis schließen sich an – Fragen, die nicht sonderlich weit entfernt liegen und die von der Forschung, von der Politik und auch künftig von der Medienbildung zu beantworten sind.

Literatur

1 Etzioni, Oren: The World Wide Web: quagmire or gold mine? Communications of the ACM 39(1996) H. 11, S. 65-68. citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.2897&rep=rep1&type=pdf (Zugriff 15.11.2011) Die Begriffe Web bzw. Data Mining werden üblicherweise nicht ins Deutsche übersetzt.

2 Gantz, John; Reinsel, David: The 2011 Digital Universe Study: Extracting Value from Chaos. IDC View. 28.06.2011. www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf (Zugriff 15.11.2011)

3 Berners-Lee, Tim; Hendler, James; Lassila, Ora: The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American. 284 (2001), H. 5, S. 34–43.

4 Hitzler, Pascal; Krötzsch, Markus; Rudolph, Sebastian; Sure, York: Semantic Web. Springer Berlin, 2008.

5 Stuckenschmidt, Heiner: Ontologien. Konzepte, Technologien und Anwendungen. Springer Berlin, 2009.

6 Kendal, Simon; Creen, Malcom: An Introduction to Knowledge Engineering. Springer London, 2007.

7 O'Reilly, Tim: Was ist Web 2.0? Entwurfsmuster und Geschäftsmodelle für die nächste Software Generation. 2005. (Deutsche Übersetzung) www.oreilly.de/artikel/web20_trans.html (Zugriff 15.11.2011)

8 Surowiecki, James: The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nation. Doubleday New York, 2004. (Deutsch: Die Weisheit der Vielen: Warum Gruppen klüger sind als Einzelne. Goldmann, 2007)

9 Aulinger, Andreas, Pfeiffer, Max (Hrsg.): Kollektive Intelligenz. Methoden, Erfahrungen und Perspektiven. Steinbeis Stuttgart, 2009.

10 Lorenz, Jan; Rauhut, Heiko; Schweitzer, Frank, Helbing, Dirk: How social influence can undermine the wisdom of crowd effect. Proc.National Academy of Sciences of the United States of America. 16.05.2011. www.pnas.org/content/early/2011/05/10/1008636108.full.pdf+html (Zugriff 15.11.2011)

11 Wordnet. A lexical database for English. Princeton University. wordnet.princeton.edu (Zugriff 15.11.2011)

12 Melo, Gerard de; Weikum, Gerhard: Towards a Universal Wordnet by Learning from Combined. Proc. 18th ACM Conference on Information and Knowledge Management (CIKM 2009). Hong Kong, 2009. www.mpi-inf.mpg.de/~gdemelo/papers/demelo-uwn-cikm2009.pdf (Zugriff 15.11.2011)

13 Witten, Ian H.; Frank, Eibe; Hall, Mark A.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Amsterdam, 2009.

14 Gruber, Tom: Where the Social Web Meets the Semantic Web. 5th International Semantic Web Conference ISWC, 7. November 2006. tomgruber.org/writing/social-web-meets-semantic-web.pdf (Zugriff 15.11.2011)

15 Kosala, Raymond; Blockeel, Hendrik: Web Mining Research: A Survey, SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining ACM 2(2000) H. 1 www.acm.org/sigs/sigkdd/explorations/issues/2-1-2000-06/kosala.pdf (Zugriff 15.11.2011)

16 Aden, Timo: Google Analytics: Implementieren. Interpretieren. Profitieren. Hanser München, 2010.

17 Howe, Danile C.; Nissenbaum, Helen; Toubiana, Vincent: TrackMeNot. New York University. cs.nyu.edu/trackmenot (Zugriff 15.11.2011)

18 Shahabi, Cyrus; Zarkesh, Amir M.; Adibi, Jafar l.; Shah, Visha: Knowledge Discovery from Users Web-Page Navigation. Proc. Workshop on Research Issues in Data Engineering IEEE RIDE97, April 1997. dblab.usc.edu/Users/shkim/papers/ride97.pdf (Zugriff 15.11.2011)

19 Google Flu Trends. www.google.org/flutrends (Zugriff 15.11.2011)

20 Kaushik, Avinash: Email Marketing: Campaign Analysis, Metrics, Best Practices. 18.07.2011. www.kaushik.net/avinash/email-marketing-campaign-analysis-metrics-practices (Zugriff 15.11.2011)

21 Kaushik, Avinash: Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity. John Wiley & Sons, 2009.

22 Microsoft Academic Search. academic.research.microsoft.com/About/Help.htm (Zugriff 15.11.2011)

23 Hirsch, Jorge E.: An index to quantify an individual's scientific research output. 2005. arxiv.org/PS_cache/physics/pdf/0508/0508025v5.pdf (Zugriff 15.11.2011)

24 GuttenPlag – kollaborative Plagiatsdokumentation: Eine kritische Auseinandersetzung mit der Dissertation von Karl-Theodor Freiherr zu Guttenberg. 2011. de.guttenplag.wikia.com/wiki/GuttenPlag_Wiki (Zugriff 15.11.2011)

25 Quantified Self: Self knowledge through numbers. quantifiedself.com (Zugriff 15.11.2011)

26 Handelsblatt: Handel mit Inflation und Arbeitslosenzahl. 8.11.2010. www.handelsblatt.com/politik/konjunktur/nachrichten/handel-mit-inflation-und-arbeitslosenzahl/3631408.html, www.eix-market.de (Zugriff 15.11.2011)

27 Kraska, Sebstian: Datenschutz und der Facebook Like-Button: Was Webseiten-Betreiber beachten müssen. www.datenschutzbeauftragter-online.de/datenschutz-facebook-like-button-was-webseiten-betreiber-beachten-muessen (Zugriff 15.11.2011)

28 Dodds, Peter S.; Harris, Kameron D.; Kloumann, Isabel M. ; Bliss, Catherine A.; Danforth, Christopher M.: Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. Physics and Society 29.07.2011. arxiv.org/PS_cache/arxiv/pdf/1101/1101.5120v4.pdf (Zugriff 15.11.2011)

29 Kamvar, Sepandar D.; Harris, Jonathan: We Feel Fine and Searching the Emotional Web. WSDM 11 Hong Kong. www.wefeelfine.org/wefeelfine.pdf (Zugriff 15.11.2011)

30 Google Books Ngram Viewer. books.google.com/ngrams (Zugriff 15.11.2011)

31 Dodds, Peter S.; Danforth, Christopher M.: Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. Journal of Happiness Studies 11(2009) H. 4, S. 441-456. www.springerlink.com/content/757723154j4w726k/fulltext.pdf (Zugriff 15.11.2011)

32 Asur, Sitaram; Huberman, Bernardo A.: Predicting the Future with Social Media. Social Science Electronic Publishing, 26.03.2010. www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf (Zugriff 15.11.2011)

33 Verräterisches Handy. www.zeit.de/datenschutz/malte-spitz-vorratsdaten www.opendatacity.de/vorratsdatenspeicherung/ (Zugriff 15.11.2011)

34 Stock, Wolfgang; Stock, Mechtild: Wissensrepräsentation. Oldenbourg München, 2008.

35 Weikum, Gerhard; Theobald, Martin: From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. ACM Symposium on Principles of Database Systems (PODS), Indianapolis, 2010. www.mpi-inf.mpg.de/~weikum/pods2010-weikum&theobald.pdf (Zugriff 15.11.2011)

36 Krötzsch, Markus; Vrandecic, Denny: Semantic MediaWiki. In: Fensel, Dieter (Hrsg.): Foundations for the Web of Information and Services. Springer Berlin, 2011. S. 311-326.

37 Semantic MediaWiki (SMW). semantic-mediawiki.org/ (Zugriff 15.11.2011)

38 DBpedia. dbpedia.org/About (Zugriff 15.11.2011)

39 Heath, Tom; Bizer, Christian: Linked Data. Evolving the Web into a Global Data Space. Morgan & Claypool, 2011.

40 The Friend of a Friend (FOAF) project. www.foaf-project.org/ (Zugriff 15.11.2011)

41 Wolfram Alpha computational knowledge machine. www.wolframalpha.com/ (Zugriff 15.11.2011)

42 Wikipedia Google Squared: en.wikipedia.org/wiki/Google_Squared (Das Testsystem ist seit Oktober 2011 nicht mehr erreichbar.)

43 Fader, Anthony; Soderland, Stephen; Etzioni, Oren: Identifying Relations for Open Information Extraction. Conference on Empirical Methods in Natural Language Processing, 2011. www.cs.washington.edu/homes/afader/bib_pdf/emnlp11.pdf, reverb.cs.washington.edu, www.cs.washington.edu/research/knowitall (Zugriff 15.11.2011)

44 IBM: The DeepQA Project. http://www.research.ibm.com/deepqa/deepqa.shtml (Zugriff 15.11.2011)

45 Fodor, Jerry A.: The Modularity of Mind. 12. Aufl. der Erstausgabe 1983. MIT Press Cambridge, 2001. Das Buch entstand aus gemeinsamer Arbeit mit Noam Chomsky.

46 Schneider, Norbert: Die digitalen Menschenleser. FAZ 10.08.2010, S. 33.

47 Internet & Gesellschaft Co:llaboratory: Abschlussbericht - Gleichgewicht und Spannung zwischen digitaler Privatheit und Öffentlichkeit. Berlin, November 2011. collaboratory.de/downloads/Abschlussbericht4dina5.pdf (Zugriff 22.11.2011)