BENNO STEIN im Gespräch mit **PETRA LÖFFLER**

«KORRELATIONEN SIND ÜBERALL DA, WO SIE GESUCHT WERDEN.»

Benno Stein ist Professor für Content Management und Webtechnologien an der Bauhaus-Universität Weimar und Sprecher des Digital Bauhaus Lab.[1] Er forscht auf den Gebieten des Maschinellen Lernens, des Data-Mining sowie des Information Retrieval und beschäftigt sich unter anderem mit der Analyse großer Datenmengen in sozialer Software. Beispielsweise entwickelt er mit seinem Forscherteam Technologien und Algorithmen, um Vandalismus und Datenmissbrauch in Wikipedia einzuschränken. Zum Gespräch bringt Benno Stein einen Ausschnitt aus der amerikanischen Quiz-Show *Jeopardy* mit, in der das von IBM entwickelte Computerprogramm «Watson» schwierige Fragen schneller als seine menschlichen Mitspieler beantwortet. Als Datenquelle nutzt «Watson» unter anderem die Wissensplattform Wikipedia.

Petra Löffler **Was macht aus Ihrer Sicht «Watson» so besonders?**
Benno Stein Die größte Schwierigkeit für Computerprogramme wie «Watson», Wissen zu verarbeiten, besteht darin, den Kontext zu bestimmen und daraus etwas zu machen – d. h. Wissen in Anwendung zu bringen. «Watson» gelingt das so gut, dass es schwierige Fragen in der Regel schneller beantwortet als die besten menschlichen Mitspieler. Es gibt für Computersysteme kaum eine schwierigere Situation als diese. «Watson» muss nicht nur die Frage korrekt erfassen, sondern auch Antwortmöglichkeiten vergleichen und die richtige Antwort, also die mit der größten Wahrscheinlichkeit aus der Menge an verfügbaren Daten herausfiltern.

P. L. Big Data ist derzeit in aller Munde, vor allem weil die Analyse großer Datenmengen durch leistungsstarke Computer verspricht, konkrete praktische Probleme wie die Finanzkrise, den Klimawandel oder Epidemien zu lösen, dadurch Entscheidungsprozesse zu vereinfachen und Zukunft vorhersehbarer zu machen.

1 Zur Homepage der Forschergruppe siehe: *www.webis.de*.

B.S. Mit großen Datenmengen geht man bereits seit ca. fünfzig Jahren um. Wettersimulationen gibt es schon lange, Crashtest- und Atombombensimulationen ebenfalls. Seit Jahrzehnten werden enorme Rechnerkapazitäten in die Lösung komplexer Fragestellungen investiert. Dabei ist die Komplexität solcher Problemstellungen zum Teil sehr gut beherrschbar und skalierbar. Die Situation von «Watson» ist im Vergleich dazu anders gelagert: Die Datenmengen sind kleiner, die Erwartungen an das Ergebnis zugleich höher. Die Datenmenge, die etwa im Zusammenhang mit dem Large Hadron Collider am Genfer CERN verarbeitet wird, ist mit 15 Petabytes pro Jahr immens. An deren Analyse sind über 170 Computercluster weltweit beteiligt.[2] Die Datenmenge ist also nicht das Problem. Neu bei Big Data ist dagegen die Art der Datenquellen. Ein Großteil der Daten kommt aus den sozialen Medien. Und es geht verstärkt darum, durch die Analyse dieser meist unstrukturierten Daten neue Geschäftsmodelle zu entwickeln. Die Daten heißen unstrukturiert, weil sie individuell und quasi ohne formale Vorgaben gemacht werden. Sie sind aus ökonomischer Sicht wertvoll, weil sie uns in ihrer Gesamtheit Dinge über gesellschaftliche Gruppierungen verraten, die sich bei der Analyse von Einzelnen nicht erkennen lassen. Big Data Analytics, also die Auswertung großer Datenmengen, kann jedoch keine gesellschaftlichen Probleme lösen. Eine gerechte Ressourcenverteilung zum Beispiel ist kein Computerproblem, sondern ein Problem von Machtverteilung.

P.L. An welchen konkreten Projekten arbeiten Sie am unlängst gegründeten Digital Bauhaus Lab?

B.S. Wir wollen Data Analytics vor allem bei der semantischen Erschließung von Texten voranbringen. Es geht dabei zum Beispiel um das Erkennen von Paraphrasierungen eines Textabschnitts, von Umschreibungen also, die die Textaussage im Kern nicht verändern. Das ist eine menschliche Domäne und erfordert eine Menge Hintergrundwissen, um z. B. Plagiate zu erkennen. Da kann man mit Big Data Analytics nachhelfen, indem man Rückschlüsse aus der Kontextanalyse und der Zeichenabfolge in Texten zieht. Wir testen verschiedene Dokumentmodelle an vier bis sechs Millionen Wikipedia-Artikeln. Mit Graph- und Clusteringanalysen können Ähnlichkeiten erkannt werden, die verlässlich sind. Das dauert auf unseren Rechnern nicht sehr lange. Mit dem Vergleich von Modellen sind sie dann ein paar Tage beschäftigt.

P.L. Es geht bei diesem Modellvergleich also um eine Optimierung der entsprechenden Modellierung von Korrelationen, eine Verbesserung der Gradierung von Ähnlichkeiten. Wie viele Datensätze braucht man, um diese Testdurchläufe durchzuführen und eine hinlänglich verlässliche Aussage über Paraphrasierungen treffen zu können?

B.S. Die Herausforderung liegt wie gesagt nicht in der Datenmenge, überhaupt nicht. Die Herausforderung ist, nach meiner Erfahrung mit Big Data, die

2 Vgl. www.lhc-facts.ch.

richtige Frage zu stellen. Eine entsprechende Datenanalyse zur Überprüfung der Frage durchzuführen, ist dagegen in der Regel nicht allzu schwierig.

P.L. Und was bedeutet es, die richtige Frage zu stellen?
B.S. Wir stellen ein paar solcher Fragen. Es geht bei unserer Forschung darum, an Sachverhalte heranzukommen, an die Computer vermeintlich nicht herankommen. Moralität ist ein weiteres Beispiel. Was wir seit Neuestem untersuchen, ist der so genannte Morning Morality Effect, den Forscher an der Harvard University und der University of Utah kürzlich beschrieben haben. Es geht dabei um die Annahme, dass Menschen morgens moralischen Verfehlungen stärker widerstehen können als nachmittags, einfach weil man ausgeschlafen prinzipientreuer und reflektierter ist. Der moralische Imperativ nimmt demnach im Laufe des Tages mit zunehmender Erschöpfung ab und damit auch die Fähigkeit, die eigenen Entscheidungen zu überprüfen. Dieser Effekt ist statistisch signifikant, und wir sind jetzt dabei, dafür im Internet Belege zu suchen, also Korrelationen zwischen Moralität, Ort und Zeit zu finden. Die Frage, gibt es moralische Suchanfragen oder nicht, kann man Moralität algorithmisch bestimmen, ist für Informatiker sehr spannend. Solche Hypothesen aufzustellen ist also die eigentliche Herausforderung bei der Analyse großer, unstrukturierter Datenmengen.

P.L. Es geht also um Entscheidungslogiken, die algorithmisch modelliert werden sollen. Aber die Wahrscheinlichkeit gerade moralischer oder unmoralischer Handlungen lässt sich auf diese Weise nicht vorhersagen. Welche Rolle spielen unwahrscheinliche oder marginale Ereignisse bei solchen Modellierungen?
B.S. Natürlich gibt es eine ganze Reihe von Faktoren, die bei einer solchen Suchanfrage zusammenwirken und die wir nur bedingt berücksichtigen können, z. B. der unerwartete morgendliche Anruf einer verärgerten Person oder sensible medizinische Daten wie der Hormonspiegel. Der Reiz, solche Fragen zu stellen und als Statistiker zu verarbeiten, liegt schlicht darin, dass sich viele Menschen für sie interessieren.

P.L. Die Suchanfragemodelle, die dabei entwickelt werden, werden immer genauer. Wie kann man Modelle bzw. Systeme, die sich durch Feedbackschleifen selbst optimieren, noch kontrollieren?
B.S. Im Grunde gar nicht. Die Kontrolle liegt bei den Menschen, die diese Modelle entwickeln. Das ist eine Grundannahme der Kybernetik.

P.L. Seit einigen Jahren werden die Programmierbarkeit menschlichen Verhaltens und damit kybernetische Denkansätze von Medientheoretikern und -aktivisten wie dem Autorenkollektiv Tiqqun verstärkt kritisiert. In aktuellen Debatten im Zusammenhang mit der NSA-Affäre taucht zudem

immer wieder die Frage auf, wem die Daten, die Internetnutzer willentlich und unwillentlich milliardenfach zur Verfügung stellen, eigentlich gehören? Wer hat die Verfügungsgewalt über diese Daten?

B.S. Der Kontext, in dem die Daten entstehen, ist entscheidend. Der Kontext, das sind u. a. soziale Softwareplattformen wie Facebook und Twitter, Suchmaschinen wie Google und Bing, eCommerce-Plattformen wie Amazon und Ebay oder Messaging-Dienstleister. Deren Benutzung führt zur Frage: Wem gehört das Netz – dem Infrastrukturbetreiber oder dem Serviceanbieter? Das ist letztlich eine Machtfrage. Nicht mangelnde Transparenz oder der Umfang der Datenmengen sind das Problem. Denn es gibt im Grunde nirgendwo mehr Transparenz als im Internet.

P.L. Die Machtfrage lässt sich auch auf eine andere Auseinandersetzung beziehen. Wollen wir ein zentrales oder ein dezentrales Netz, bevorzugen wir Konvergenz oder Divergenz? Wie sieht das Internet der Zukunft aus?

B.S. Diese Frage ist in der Geschichte des Internet schon oft und eindeutig behandelt worden. Viele relevante Netze sind dezentral und geheim. Ich rede von Netzen von Banken, zur Kontrolle von Kraftwerken und Satelliten, von GPS-Netzen. Diese Netze gehören denjenigen, die sie geschaffen haben, also den Betreibern, die ein ökonomisches Interesse daran haben, ihre Netze geschlossen zu halten.

P.L. Daneben gibt es Datenbanken und Netzwerke, die dezidiert auf die Zugänglichkeit für alle und auf User Generated Content setzen. Amateure werden motiviert, Daten zu generieren, um eine gewisse Datenbasis zu schaffen, auf deren Basis Suchanfragen überhaupt wirksam durchgeführt werden können. Daten werden hier als Gemeingut verstanden. In diesem Zusammenhang wird entweder das Wissen der Vielen gelobt oder deren Dummheit angeprangert. Wie stehen Sie zu dieser Debatte?

B.S. Es ist sehr schwierig, Wissen im Internet zu generieren. Vielleicht geht das auch gar nicht. 99,9 % der Daten im Internet sind irrelevant, wenn es um die Vermehrung von Wissen geht. Wikipedia ist, was die Qualität der Informationen angeht, eine Ausnahme. Dennoch ist Wikipedia vom Datenvolumen her gesehen vergleichsweise klein: Etwa sechs Millionen englische und deutsche Artikel und im Durchschnitt ca. 5 KB pro Artikel sind sehr wenig. Suchmaschinen wie Google oder Bing vermitteln uns ein extrem gefiltertes Bild des Internets. Sie filtern die 0,1 % der Daten heraus, die für eine Suchanfrage relevant sind. Um diese 0,1 % herauszufiltern, braucht man noch besseres algorithmisches Wissen, das den Unterschied erkennt. Eine wie auch immer umfangreiche Menge an Daten reicht also nicht aus, um uns schlau zu machen. Sie können nicht Informationen und damit Wissen durch algorithmische Hinzufügungen generieren.

P.L. Es braucht also schlaue Informatiker, die Programme wie «Watson» entwickeln, die dann bei der Suchanfrage den Wissensunterschied erkennen?
B.S. Die Aufgabe, aus einer Reihe von Differenzen hochwertige, also signifikante und vom Effekt her relevante Unterschiede herauszufiltern, ist sehr komplex und schwierig. Man muss am Anfang etwas Hochspekulatives hinzutun, wie zum Beispiel das Formulieren der Ausgangshypothese beim Morning Morality Effect, und dann diese Hypothese testen, also die algorithmische Leistung erbringen, dass der Effekt relevant ist. Dazu muss ich ein paar Millionen Suchanfragen statistisch analysieren. Da erscheint es einfacher, sich direkt zu fragen, ob es Moralität gibt, als sich auf komplexe Algorithmen zu verlassen.

P.L. Datenanalysen haben also Grenzen, Grenzen der Anwendbarkeit und der analytischen Leistungsfähigkeit von Computersystemen. Dennoch spielt die Optimierung bei der Suche nach Korrelationen eine wichtige Rolle.
B.S. Korrelationen sind überall dort, wo man sie sucht. Big Data Analytics wird derzeit hauptsächlich dafür eingesetzt, Hypothesen zu verifizieren. Man stellt eine Hypothese auf und entwickelt Modelle, die man algorithmisch miteinander vergleicht. Und dann überprüft man, ob die Korrelation statistische Aussagekraft hat. Man kann mit dieser Methode ja nichts beweisen, sondern nur argumentieren und statistische Relevanzen feststellen. Datenanalysen sind in dieser Hinsicht nicht kreativ, sondern leistungs- bzw. problemlösungsorientiert. Zum Beispiel kann man den Körperfettwert so einfach nicht direkt messen. Es besteht aber offenbar eine Korrelation zwischen dem Körperwiderstand und den verschiedenen Körpergewebetypen, und den kann man messen.

P.L. Bedeutet das, dass Korrelationen Kausalität ersetzen können?
B.S. Wir ersetzen unbewusst und leichtfertig Kausalität durch Korrelationen. Wenn zwei Ereignisse korreliert sind, also eines dem anderen in statistisch signifikanter Weise folgt, dann werten wir das automatisch als kausale Beziehung von Ursache und Wirkung. Objektiv feststellbar ist jedoch erst mal nur die zeitliche Abfolge der Ereignisse. Und auch mit jeder detaillierteren Modellierung der Ereignisse verlagert sich die Frage der Kausalität lediglich auf eine prinzipiellere Erkenntnisstufe.

P.L. Wie werden die Abläufe bei der Modellierungsoptimierung erfasst? Gibt es Aufzeichnungen darüber und Fehleranalysen?
B.S. Wichtig ist nicht das Ergebnis, sondern die Erklärung des Ergebnisses. Wir dokumentieren und erklären sehr genau, wie unsere Suchprozesse funktionieren und auch wie viel sie kosten. Wir machen zum Beispiel Qualitätsanalyse für Wikipedia und suchen automatisiert nach Fehlerquellen. Mit unseren Algorithmen finden wir Schwachstellen, an denen Vandalismus

geschieht, also zum Beispiel Menschen diffamiert oder falsche Angaben gemacht werden.

P.L. Der Anspruch ist also, Wikipedia besser zu machen?

B.S. Genau. Unsere Ausgangsfrage war: Kann man bestimmte Qualitätsfehler erkennen? Es gibt über 350 registrierte Qualitätsfehlerarten. Wir haben Algorithmen entwickelt, die verlässlich einige wichtige dieser Fehler und Schwachstellen aufspüren und jetzt mittels Bots, also automatisierten Suchabfragen, für Wikipedia verwendet werden können. Davon profitiert dann die gesamte Community.