Die Macht im Computer

Algorithmen: Nützliche Hilfsmittel oder Gefahr?

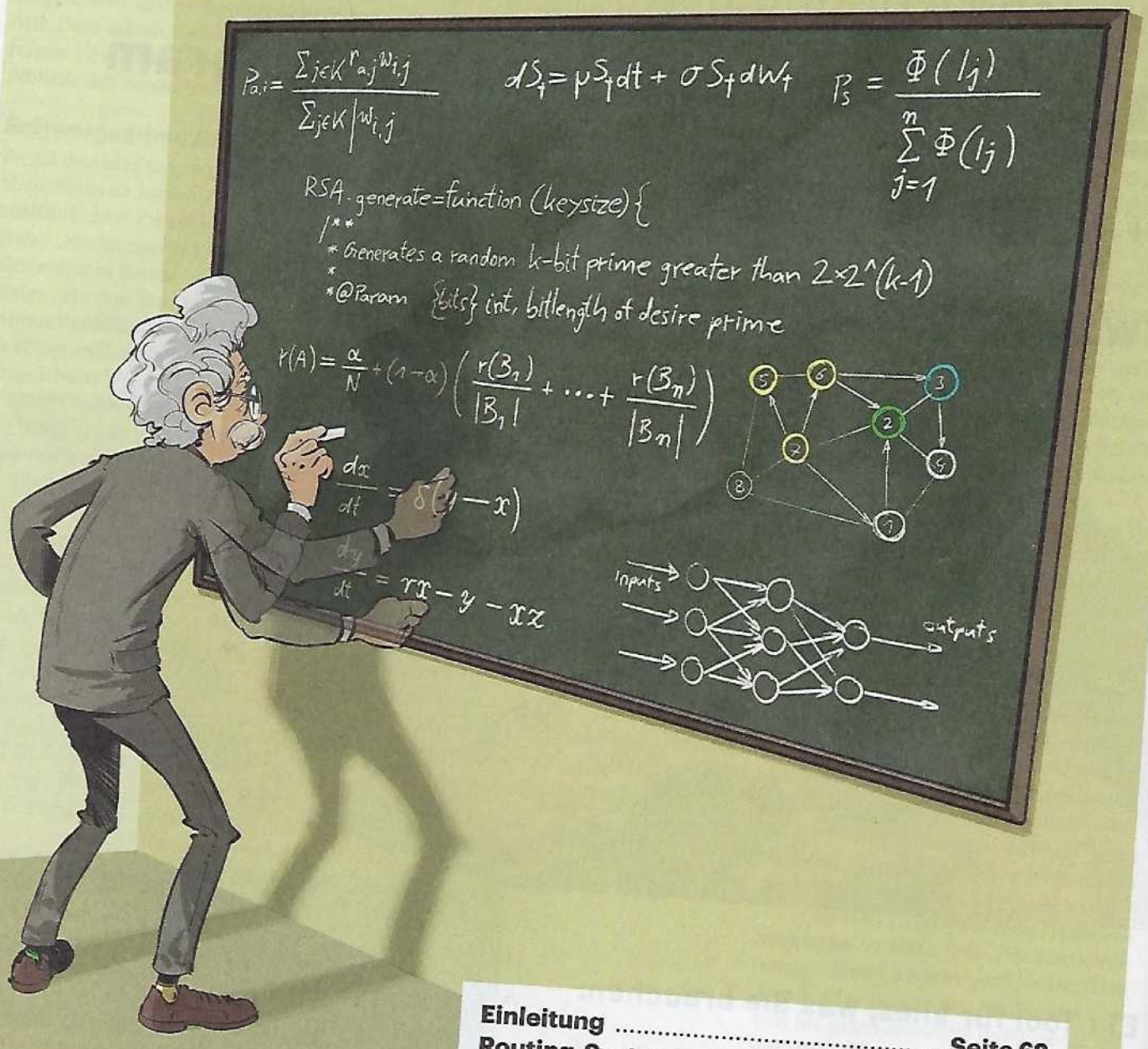


Bild: Albert Hulm

Wer hat Angst vorm Rechenwerk? Die Sorge vor einer Übermacht der Algorithmen beschäftigt mittlerweile Politik und Öffentlichkeit. Da kann etwas Grundwissen über ihre Funktion nicht schaden.

Von Wolfgang Stieler

Sie sind überall, und sie sind mächtig: Algorithmen entscheiden, ob wir einen Job bekommen oder nicht, wie kreditwürdig wir sind, welche Nachrichten wir sehen, was wir lesen, sehen und hören. Für viele Menschen klingt das mittlerweile mehr wie eine Drohung als eine Verheißung. So bedrohlich, dass inzwischen auch die Politik aufgewacht ist: Die Bundesregierung etwa berief eine Datenethikkommission ein.

Im Herbst 2019 legten die 16 von der Bundesregierung eingesetzten Experten ein erstes Gutachten vor. Mit drastischen Forderungen: Unter anderem empfahl die Kommission ein risikoadaptiertes Regulierungssystem für den Einsatz algorithmischer Systeme, eine Bewertung von Programmen nach „Kritikalität" und „Schädigungspotenzial" und sogar ein komplettes Verbot der allerschädlichsten Programme.

Ist das eine vernünftige Abwägung der Risiken technischer Entwicklung oder die Angst vor der Allmacht der Maschine? Eine Angst, die sich mehr aus popkulturellen Bildern und überzogenen Marketing-Versprechen speist als aus konkretem Wissen? Dieser Schwerpunkt soll diese Diskussion versachlichen: Er zeigt in den folgenden drei Artikeln anhand von Vorschlagsalgorithmen, Googles PageRank-Verfahren und Routing-Systemen, was Algorithmen so bedeutsam macht und wo ihre Risiken und Nebenwirkungen liegen.

Der Kontext macht den Unterschied

Im Kern sind Algorithmen in der IT zunächst nichts weiter als eine Folge von Rechenanweisungen. Was sie so mächtig und gelegentlich problematisch macht, ist nicht ihre Ausführung, sondern der Kontext, in dem sie entworfen und verwendet werden. Denn um abstrakte Probleme für Computer berechenbar zu machen, muss die Problemstellung auf ein mathematisches Modell abgebildet werden. Der betreffende Algorithmus löst dann dieses mathematisch abstrahierte Problem.

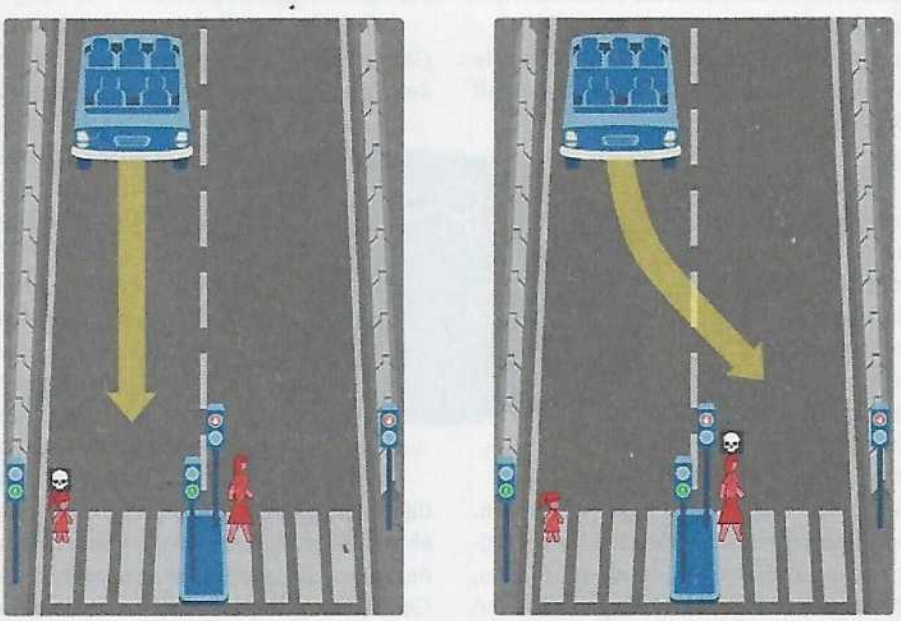
Die abstrakte Lösung kann dann auf alle möglichen konkreten Probleme angewandt werden. Einem Sortier-Algorithmus beispielsweise ist es egal, ob er Zahlenwerte sortiert, Produkte nach Beliebtheit oder Bilder nach ihrem Motiv. So gesehen sind Algorithmen wirklich so mächtig, präzise, unbestechlich und objektiv, wie viele Menschen glauben.

Ein kritischer Punkt dabei ist jedoch, wie genau eigentlich unscharfe, subjektive Größen wie Beliebtheit, Schönheit oder auch die Eignung für einen Job in Zahlenwerte übersetzt werden – die sogenannte Objektivierung. Denn das mathematische Modell kann immer nur einen kleinen Ausschnitt aus der Realität abbilden. Welche Größen wie genau mathematisch abgebildet werden, kann jedoch großen Einfluss auf das Ergebnis einer Berechnung haben. Wer unsinnige Daten eingibt, wird unsinnige Ergebnisse bekommen. „Garbage in, garbage out" wie der Informatiker sagt.

Dazu kommt, dass oftmals nicht alle Daten zur Verfügung stehen, die für eine exakte Berechnung notwendig wären. Ein autonomes Auto beispielsweise muss seine nächsten Aktionen anhand äußerst lückenhafter Sensordaten planen – und noch dazu damit rechnen, dass sich die Situation während der Planung weiter verändert.

Die Antwort der Informatik darauf sind Vereinfachungen, die oft, aber nicht zwingend immer zum Ziel führen: Sie vergleichen – ähnlich, wie es der Mensch tun würde – die Situation mit früheren Erfahrungen, versuchen die Konsequenz einer falschen Entscheidung abzuschätzen, treffen Annahmen über die zukünftige Entwicklung – und im Zweifelsfall lassen sie den Zufall per Münzwurf entscheiden.

Diese Erkenntnis ist alles andere als beruhigend. Denn sie bedeutet, dass die Bewertung und Regulierung von Algorithmen von ihrem Kontext abhängt. Die Annahmen, unter denen Algorithmen zum Einsatz kommen, sind genauso zu hinterfragen wie die Datenbasis und die heuristischen Methoden. An dieser Diskussion wird kein Weg vorbeiführen. Der Artikel ab Seite 74 gibt den aktuellen Stand dieser Diskussion um die „Ethik der Algorithmen" wieder. *(jo@ct.de)* **ct**



Das Auto kann nicht mehr bremsen, wen soll es überfahren? Derzeit sind solche Überlegungen noch Gedankenspiele. Eines Tages müssen selbstfahrende Autos aber vielleicht genau solche Entscheidungen treffen.

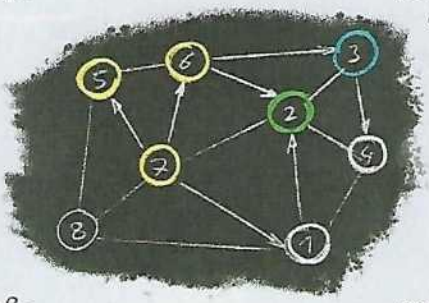
Zielfinder

Kürzeste Wege berechnen mit Dijkstras Algorithmus – und viele zusätzliche Daten

Ein Kartendienst wie Google Maps ist eine hochkomplexe Plattform, die weiß, wo es die nächste Tankstelle gibt und die ganz nebenbei Daten für die Routenfindung während der Nutzung erhebt. Für die Kernaufgabe, die Berechnung der optimalen Route von A nach B, kommt ein uraltes Verfahren zum Einsatz.

Von Jo Bager

Sie haben Ihr Ziel erreicht." Ein Router wie Google Maps ist längst mehr als ein einfaches Dienstprogramm, das den Nutzer von A nach B bringt. Google Maps ist vielmehr eine komplexe Reiseplattform, die die nächstgelegenen Bäcker inklusive Öffnungszeiten und Kundenbewertungen anzeigt. Und Google ist ein Meister darin, die Crowd dafür einzuspannen, Daten zu diesem Pool beizusteuern. So stammen die Bilder für viele Sehenswürdigkeiten von Freiwilligen – und die Stoßzeiten berechnet Google automatisch aus der Anzahl der Nutzer mit aktivierter Google-Maps-App.



Die Berechnung der Routen ist nur so gut wie die zugrundeliegenden Daten. Google Maps zum Beispiel kartiert längst nicht mehr nur Autostraßen. Es kann Routen auch für Fahrradfahrer, für Fußgänger, für öffentliche Verkehrsmittel und sogar für Flüge berechnen. Dafür bezieht es Kartenmaterial zum Beispiel von Ländern und Gemeinden mit ein, aber auch von Non-Profit-Organisationen oder Firmen.

Außerdem schickt es immer mal wieder Vermessungsautos los, die alte und neue Strecken abfahren, um sie zu kartieren. Dabei erfassen die Fahrerteams nicht nur die reine Strecke, sondern auch den Straßentyp, also etwa, ob es sich um eine Autobahn und Mautstraße handelt. Wie Google Maps und Co. aus den Wegdaten ihre Routen berechnen, zeigt der Kasten. Die aktuelle Herausforderung von Kartenanbietern wie Google Maps ist es, ihre Karten für das autonome Fahren wesentlich genauer zu machen [2].

Die Fahrzeuge, die die Wege kartieren, fertigen auch die StreetView-Aufnahmen der Straßen an und sind auch mit etlichen Sensoren ausgestattet. So entsteht nach und nach ein 3D-Scan der Welt.

Der Nutzer als Datenlieferant

Daten-Crowdsourcing wendet Google auch bei der Anzeige des Live-Verkehrs an. Dabei merken die Nutzer oft nicht einmal, dass sie Google Informationen zuliefern: Jeder Nutzer von Google Maps bezieht nämlich nicht nur seinen Weg aus Googles Cloud, sondern sendet auch permanent Signale dorthin.

Für viele Verbindungen ist nicht die Geschwindigkeit entscheidend, die auf den Straßen grundsätzlich zurückgelegt werden kann, sondern diejenige, die dort tatsächlich gefahren wird. Und die liegt etwa bei einem Unfall oder der Rush Hour oft deutlich darunter.

Damit der Router aber immer die genaue aktuelle Position und Verkehrslage berücksichtigen kann, muss Google regelmäßig die aktuelle Position an Googles Server senden. Daraus kann Google dann wieder die Geschwindigkeit des einzelnen Fahrzeugs ableiten. Und aus den Informationen der Masse der Google-Maps-Daten kann Google Informationen für das Streckennetz ableiten. An einer Stelle fahren alle Maps-Nutzer langsam: Dann muss dort eine Störung sein, die dann auch für die Routenberechnung zugrunde gelegt werden kann.

Dieses Prinzip dürfte in Zukunft noch ausgebaut werden. Volvo zum Beispiel experimentiert damit, dass sich die Autos des Herstellers über die Cloud zu Straßenbedingungen austauschen, sich also etwa über vereiste Straßenabschnitte warnen können. Im Zuge des autonomen Fahrens werden Autos sensor- und rechnertechnisch noch aufgerüstet.

Google wiederum arbeitet eng mit etlichen Autoherstellern zusammen, um seine Software in deren Autos zu integrieren. Man benötigt also nicht viel Fantasie, um sich auszumalen, wie sich Autos – und Google – in Zukunft noch viel detaillierter austauschen werden. *(jo@ct.de)* **ct**

Literatur

[1] Jo Bager: Weg-weisend, Routenplaner finden den optimalen Weg, c't 4/1995, S. 252
[2] Michael Link: Wegweiser, Wie sich Kartenhersteller aufs autonome Fahren vorbereiten, c't 7/2020, S. 126

Wo fließt der Verkehr flüssig, wo zäh – Kartendienste wie Google Maps sammeln solche Informationen von ihren Nutzern ein.

Der Algorithmus von Dijkstra

Grundlage für das Routing von Google Maps oder einem anderen Routenplaner ist ein Verfahren, das der niederländische Mathematiker Edsger W. Dijkstra im Jahr 1959 veröffentlicht hat. Kürzeste Wege zwischen zwei Punkten eines Wegenetzes zu berechnen ist ein klassisches Problem der mathematischen Graphentheorie. Graphen, das sind Gebilde aus Knoten (Punkten) und Kanten (Verbindungen zwischen jeweils zwei dieser Knoten).

Unter diesem Netzwerk kann man sich also zum Beispiel Ortspunkte und Streckenabschnitte zwischen diesen Punkten vorstellen, die jeweils ein (positives) „Gewicht" haben. Dieses Gewicht kann für die Länge der Wegstrecke stehen, aber auch für die Reisedauer oder die Kosten.

Dijkstras Kürzester-Weg-Algorithmus berechnet in diesem Netzwerk die kürzesten Wege von einem Knoten A zu allen anderen Knoten, die von A aus erreichbar sind. Das könnte man natürlich auch naiv versuchen, indem man alle Wegkombinationen durchprobiert und die jeweils kürzesten behält. Das ist aber mit einem riesigen Overhead verbunden.

Dijkstras Algorithmus kommt mit viel weniger Aufwand aus: Im ersten Schritt wählt er den Knoten, der direkt mit A über die Kante mit dem geringsten Gewicht verbunden ist – also im übertragenen Sinn den Punkt mit der geringsten Entfernung. Das sei Knoten B. Der Algorithmus merkt sich den Knoten, die Distanz dorthin und die Kante.

Im zweiten Schritt betrachtet der Algorithmus alle Knoten, die mit A oder B direkt über eine Kante verbunden sind, und wählt den Knoten aus, der die nächstkleinere Gesamtentfernung von A aus hat. In jedem weiteren Schritt geht der Algorithmus entsprechend vor: Ausgehend von den bisher erreichten Knoten sucht er denjenigen Knoten, der mit einem der Knoten der bisherigen Teillösung direkt verbunden ist und die kürzeste Gesamtentfernung zu A hat. Die Entfernung berechnet sich dabei aus der Summe der Gewichte der Kanten von A zum jeweiligen Knoten.

Auf diese Weise entsteht ein Graph, der von A aus zu allen anderen (erreichbaren) Knoten jeweils die kürzesten Wege enthält. Meist handelt es sich um genau einen Weg. Je nach Aufgabenstellung kann der Algorithmus den gesamten Graph der kürzesten Wege von A aus berechnen oder abbrechen, wenn er einen Zielknoten erreicht.

Diverse Heuristiken verbessern die Performance. So kann man zum Beispiel vom Start- und Zielpunkt aus losrechnen lassen und die Suche beenden, sobald sich beide Teilwege treffen. Manche Implementierung berücksichtigt auch die Richtung der Routen: Wer etwa von Hannover aus nach Berlin kommen will, wird kaum die A2 Richtung Dortmund berücksichtigen. Eine Verallgemeinerung und Erweiterung des Dijkstra-Algorithmus nennt sich A*.

Der A*-Algorithmus untersucht immer die Knoten zuerst, die wahrscheinlich auf dem kürzesten Weg zum Ziel führen. Um den viel versprechendsten Knoten zu ermitteln, wird allen bekannten Knoten jeweils ein Schätzwert zugeordnet, der angibt, wie lang der Pfad vom Start zum Ziel unter Verwendung des betrachteten Knotens im günstigsten Fall ist. Der Knoten mit dem niedrigsten Schätzwert wird als nächster untersucht.

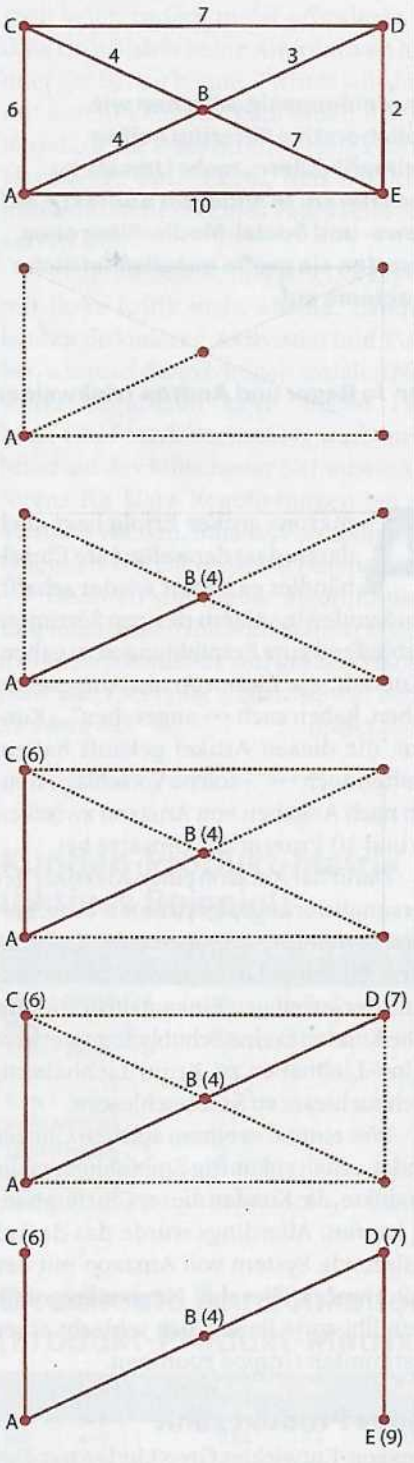
Die Zeichnung veranschaulicht die grundlegende Arbeitsweise des Dijkstra-Algorithmus. Um die kürzeste Verbindung von A nach E zu berechnen, benötigt das Verfahren vier Schritte – was bei dem Beispiel mit nur fünf Knoten der Gesamtlösung gleichkommt. Bei jedem Knoten der Teillösung hält das Verfahren die jeweilige Distanz zu A fest.

Der kleine Beispielgraph mit im Verhältnis zu den Knoten relativ vielen Kanten täuscht dabei. Bei einem großen Straßenwegenetz muss der Algorithmus in einem Arbeitsschritt nur jeweils einen kleinen Bruchteil der Kanten berücksichtigen. Daher ist der Algorithmus sehr effizient. Bereits 1995 konnte man sich mit einem handelsüblichen PC mit 386er Prozessor und 4 MByte RAM Routen innerhalb Deutschlands und Europas berechnen lassen [1].

Heute spielt die Rechenleistung auf dem Endgerät meist keine Rolle mehr – wenngleich ein aktuelles Smartphone viel mehr Rechenpower hat, als ein PC von 1997, und Apps wie OSMAnd Routen auch auf dem Mobilgerät berechnen können. Nichtsdestotrotz lassen viele weit verbreitete Apps ihre Routen in der Cloud berechnen und laden sie herunter.

Dijkstras Algorithmus

Der zugrundeliegende Graph, die vier Lösungsschritte und das Ergebnis. Die jeweils betrachteten Kanten sind gestrichelt, die (Teil-) Lösungen fett rot. Die hellblauen Kanten werden im jeweiligen Schritt nicht betrachtet.



Empfehlungen mit Nebenwirkungen

Subtile Manipulatoren: Vorschlagsalgorithmen steuern Menschenmassen

Empfehlungsalgorithmen wie Collaborative Filtering helfen Onlinehändlern, mehr Umsatz zu generieren. In anderem Kontext – auf News- und Social-Media-Sites etwa – werfen sie große gesellschaftliche Probleme auf.

Von Jo Bager und Andrea Trinkwalder

Amazons großer Erfolg liegt auch daran, dass der weltgrößte Einzelhändler es immer wieder schafft, den Kunden in seinem riesigen Sortiment verblüffend gute Empfehlungen zu geben. „Kunden, die diesen Artikel angesehen haben, haben auch *** angesehen", „Kunden, die diesen Artikel gekauft haben, kauften auch ***" – solche Vorschläge steuern nach Angaben von Amazon zwischen 10 und 30 Prozent der Umsätze bei.

Dafür hat Amazon einen Klassiker der Personalisierungsalgorithmen entscheidend verfeinert, das sogenannte Collaborative Filtering. Dabei geht es darum, anhand der jeweiligen Einkaufshistorie ähnliche Kunden in eine Schublade zu stecken: Krimi-Liebhaber zu Krimi-Liebhabern, Sachbuchleser zu Sachbuchlesern.

Wer einmal in einem solchen Cluster landet, erhält zukünftig Empfehlungen für Produkte, die Kunden dieses Clusters häufig kaufen. Allerdings wurde das darauf basierende System von Amazon mit der Zeit immer träger und Neukunden ohne Bestellhistorie ließen sich schlecht einer bestimmten Gruppe zuordnen.

Jedes Produkt zählt

Amazon-Entwickler Greg Linden hat dieses Verfahren einen entscheidenden Schritt weitergedacht. Sein „Item-to-Item Collaborative Filtering" schöpft zwar aus derselben Kunden-Produkt-Tabelle wie das einfache kollaborative Filtern, berechnet aber die Ähnlichkeit zweier Produkte direkt:

$$A(i,j) = \frac{\text{Anzahl Kunden mit } i \text{ und } j}{\sqrt{\text{Anz. Kunden mit } i \times \text{Anz. Kunden mit } j}}$$

Damit zwei Produkte als ähnlich gelten, müssen sie bei möglichst vielen Kunden gemeinsam im Warenkorb liegen, und zwar unabhängig davon, ob sie gleichzeitig oder nacheinander gekauft wurden. Damit Produkte, die insgesamt selten gekauft werden, nicht automatisch schlecht abschneiden, wird diese Zahl in Beziehung zu den absoluten Verkaufszahlen jedes einzelnen Produkts gesetzt.

So ähneln sich zwei Produkte, die bei eintausend Kunden gemeinsam im Warenkorb liegen, aber je 10.000-mal unabhängig voneinander verkauft wurden, nicht sonderlich. In diesem Fall wird 1000 als Zähler in die obige Gleichung eingesetzt. Da beide Produkte jeweils 11.000-mal verkauft wurden (10.000-mal unabhängig, 1000-mal gemeinsam), ergibt sich folgende Gleichung, die in einem nur geringen Ähnlichkeitswert von 0,09 resultiert:

$$A(i,j) = \frac{1000}{\sqrt{11.000 \times 11.000}}$$

Sehr nah beieinander liegen hingegen zwei Produkte, die nur 100 gemeinsame Kunden haben, aber lediglich an 10 Kunden alleine verkauft wurden. Sie hätten einen Ähnlichkeitswert von gerundet 0,91. Das Item-to-Item Collaborative Filtering ist zwar ähnlich rechenintensiv wie sein Vorgänger, hat aber den Vorteil, dass die Ähnlichkeiten vorab ermittelt und in einer separaten Item-Item-Tabelle gespeichert werden können. Während des Einkaufs reduziert sich der Verarbeitungsaufwand auf einige schnelle Abfragen in dieser reinen Produktmatrix, siehe die Beispieltabellen auf Seite 67 rechts unten.

In vielen Anwendungsfällen funktioniert das Item-to-Item Collaborative Filtering hervorragend. Wenn ein neues Produkt im Sortiment ist, kann das Verfahren allerdings zunächst keine verlässlichen Vorhersagen treffen. Auch Geschenkkäufe oder Freundschaftsdienste können die Vorhersagen verschlechtern.

Nicht nur Versandhändler setzen auf Collaborative-Filtering-Verfahren, sondern auch Online-Radiostationen und Video-Plattformen. Sie bieten so maßgeschneiderte Programme an oder schlagen Titel vor, die der Kunde noch nicht kennt, aber vermutlich mögen wird. Um Empfehlungen zu verbessern, versuchen sie zudem, auch die tatsächlichen Inhalte zu berücksichtigen, also beispielsweise den Musikgeschmack des Kunden anhand der Art und des Klangs seiner Lieblingsmusik zu klassifizieren.

Soziale Spalter

Soziale Netze stehen vor ganz ähnlichen Herausforderungen wie große Einzelhändler: Aus einem riesigen Angebot müssen sie für jeden einzelnen Nutzer individuelle Inhalte zusammenstellen. Der durchschnittliche Facebook-Nutzer etwa hat 155 „Freunde", die pro Tag hunderte Posts generieren. Viele Nutzer schaffen es nicht, sich durch diesen Berg komplett durchzuarbeiten. Deshalb trifft das soziale Netzwerk eine Vorauswahl.

$$P_{a,i} = \frac{\sum_{j \in K} r_{a,j} w_{i,j}}{\sum_{j \in K} |w_{i,j}|}$$

Dabei geht es nicht ums Verkaufen. In der Welt der sozialen Netze ist die Aufmerksamkeit die wichtigste Währung: Nur wenn ein Mitglied bei der Sache ist, nimmt es Werbung wahr und ist bereit, damit zu interagieren. Facebook etwa will die Interaktion und Aktivität des Nutzers maximieren: Posts, die eine angeregte Diskussion anstoßen oder gerne geteilt und gelikt werden, spült es weiter oben in die Newsfeeds der Mitglieder.

Dabei sammelt Facebook zunächst sämtliche Posts aus dem Freundeskreis,

Manchmal gibt Facebook Hinweise darauf, warum ein Posting im Newsfeed auftaucht – wie hier bei der Hannover-96-Meldung –, meist ist die Auswahl aber völlig intransparent.

Gruppen sowie Seiten, die man gelikt hat. Diesen Pool muss der Facebook-Algorithmus auf wenige hundert reduzieren und anschließend nach Relevanz sortieren. Am Ende bleiben nur etwa 20 Prozent des ursprünglichen Nachrichtenpools übrig.

Seit 2013 erledigt diesen Job nach Facebooks Angaben ein neuronales Netz, das dafür annähernd 100.000 Gewichtungen unterschiedlicher Faktoren vornimmt. Klingt beeindruckend – aber solange Facebook nicht verrät, wie hoch der Einfluss der jeweiligen Faktoren am Ende ausfällt, bleibt Skepsis angebracht, ob das komplexe Verfahren auch wirklich differenzierte Empfehlungen zustande bringt.

Reißerisch macht Klicks

Abgesehen von der Black-Box-Problematik ist das Grundproblem bei der schlichten Maximierung von Interaktion folgendes: Reißerische, skandalträchtige Themen – und damit auch potenzielle Falschnachrichten – bekommen mehr Aufmerksamkeit, weil sie hitzige Diskussionen entfachen und den Nutzer mitsamt seines Freundeskreises aktiv halten.

Um Fake News weniger Aufmerksamkeit zu verschaffen, modifiziert Facebook den Sortieralgorithmus immer mal wieder – zum Beispiel so, dass er Posts von Freunden sowie vertrauenswürdige und lokale Nachrichtenseiten bevorzugen soll. Meldungen aus Quellen, auf die man seit Jahren kaum reagiert hat, soll die Technik hingegen auf die hinteren Ränge verweisen. Bisher konnten solche Maßnahmen aber einseitige, polarisierende Berichterstattung nicht wirksam eindämmen.

All das deutet darauf hin, dass das Problem grundsätzlicherer Natur ist und in der Ausrichtung des Gesamtsystems auf möglichst hohe Interaktion liegt. Diese Metrik, auf die der Algorithmus nach wie vor optimiert, läuft einer ausgewogenen Berichterstattung im Newsfeed offenbar zuwider, allen Reparaturen zum Trotz.

Daran ändern auch die Faktenchecker nichts, mit denen Facebook weltweit zusammenarbeitet – hierzulande etwa das Recherchenetzwerk Correctiv. Bewertet ein Faktenprüfer einen Beitrag als Falschmeldung, will Facebook ihn „weiter unten im News Feed" anzeigen. Zudem können Faktenchecker solche Artikel mit „zusätzlichen Informationen" versehen. Inhalte von Seiten, die wiederholt Falschmeldungen teilen, sollen so weniger verbreitet werden.

Ein grundsätzliches Problem

Aber kann menschliche Intervention die Probleme lösen, die Facebooks Algorithmen aufwerfen? Roger McNamee, Ex-Facebook-Investor und jetzt Kritiker des Netzwerks, glaubt nicht mehr daran: „Der Schaden wird innerhalb von Sekunden angerichtet, Moderation dauert viel zu lange, selbst wenn man Millionen von Leuten dafür abstellt."

Twitter-CEO Jack Dorsey hält die Themenauswahl der Vorschlagsalgorithmen großer Plattformen ebenfalls für fragwürdig – und bezieht in die Kritik explizit seine eigene mit ein. Die Algorithmen seien zudem meist proprietär, sodass man bisher keine Alternativen habe oder gar bauen könne. Twitter will daher mit einem kleinen Team einen offenen Standard zur Dezentralisierung sozialer Netzwerke entwickeln, den der Kurznachrichtendienst eines Tages selbst einsetzen soll.

Roger McNamee und Jack Dorsey sind mit ihrer Kritik nicht alleine. Landauf, landab diskutieren Aktivisten und Politiker, wie man die mächtigen sozialen Netzwerke regulieren kann. Selbst Facebook-Chef Mark Zuckerberg warb im Februar auf der Münchener Sicherheitskonferenz für klare Regulierungen bei den Themen Wahlen, Inhalte, Portabilität und Offenheit von Daten sowie Datenschutz.

Denn so wunderbar algorithmische Empfehlungsverfahren im Kontext E-Commerce funktionieren: Auf großen Plattformen wie Facebook stellen sie ein echtes Problem dar. *(jo@ct.de)* c't

Kunden-Produkt-Matrix (fiktives Beispiel)

	DVD	Kartoffelchips	Buch	Rotwein
Kunde A (Vielfernseher)	1	1	0	0
Kunde B (Bücherwurm)	0	0	1	1
Kunde C (multimedial veranlagt)	1	1	1	1
Kunde D (medienabstinent)	0	1	0	1
Summe	2	3	2	3

Errechnete Ähnlichkeiten (Produkt-Produkt-Matrix)

	DVD	Kartoffelchips	Buch	Rotwein
DVD	1	0,82	0,5	0,41
Kartoffelchips	0,82	1	0,41	0,67
Buch	0,5	0,41	1	0,82
Rotwein	0,41	0,67	0,81	1

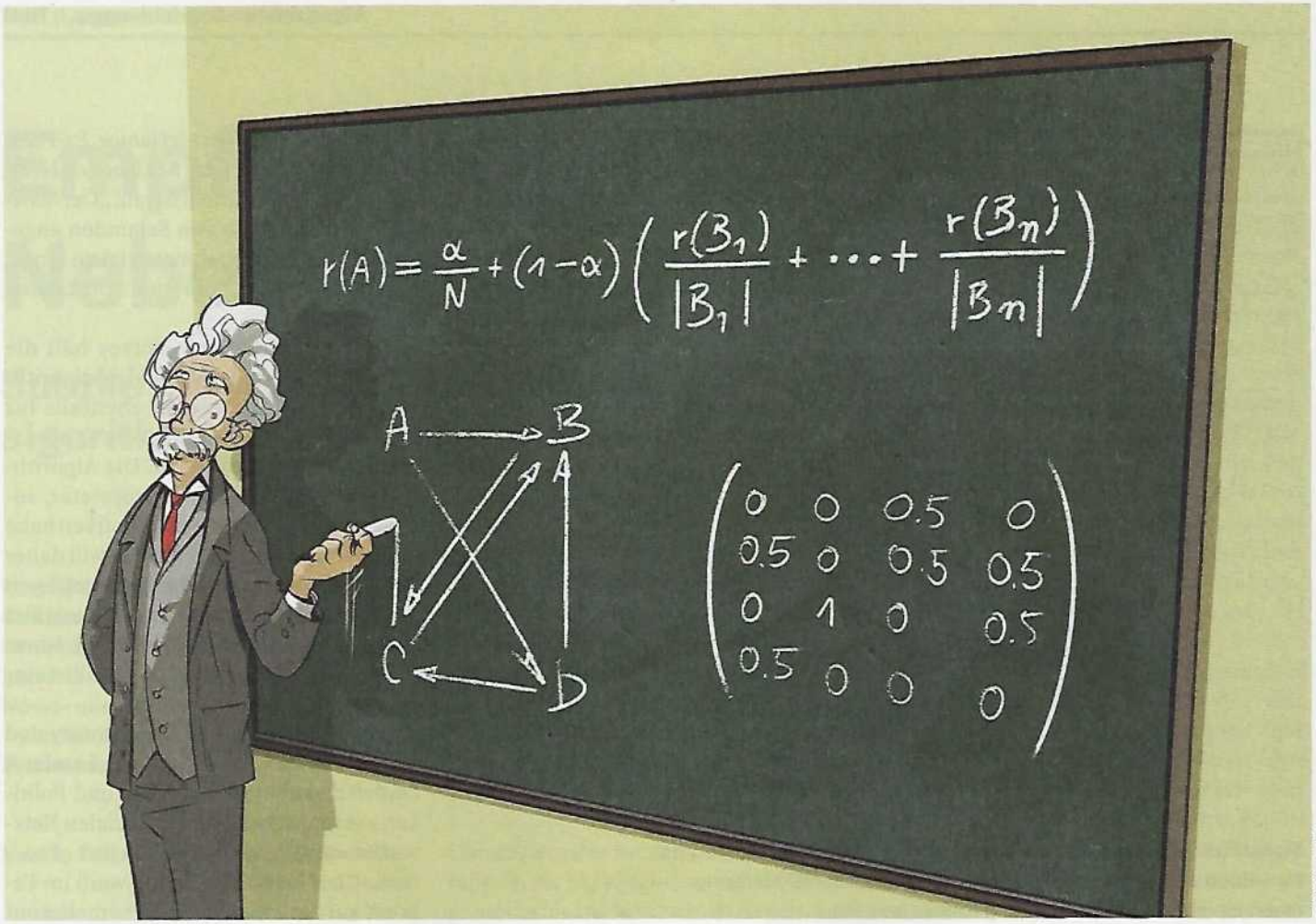


Bild: Albert Hulm

Das Orakel von Mountain View

Eine kurze Geschichte und ein Ausblick in die Zukunft des Google-Algorithmus

Zwei Studenten adaptieren ein jahrzehntealtes Verfahren aus der Soziologie – und legen so den Grundstein für die wichtigste Website der Welt: Google. Seither wird der Algorithmus der Suchmaschine beständig fortentwickelt.

Von Jo Bager

Google ist der einflussreichste Gatekeeper im Web. Wen Google nicht anzeigt, der existiert nicht. Weltweit hat der Suchdienst einen Marktanteil von 93 Prozent, hierzulande sogar von 95 Prozent.

Wer verstehen will, wie Google so erfolgreich werden konnte, muss sich die Anfänge der Suchmaschine vergegenwärtigen, die 1998 an den Start ging. Der von menschlichen Redakteuren editierte Katalog von Yahoo war seinerzeit die wichtigste Orientierungshilfe für Websurfer. Allerdings explodierte das Web in dieser Zeit geradezu und Yahoo kam da nicht hinterher. Es gab auch schon andere Suchmaschinen, AltaVista etwa oder Lycos.

Anders als von Menschen verfasste Kataloge schicken Suchmaschinen sogenannte Crawler durchs Web: Programme, die automatisiert Websites besuchen und deren Inhalte abrufen. Aus diesen Inhalten bauen Suchmaschinen einen Index auf. Gibt der Benutzer einen Begriff oder eine Phrase ins Suchfeld ein, durchsuchen sie ihren Index danach und bringen die gefundenen Seiten in eine Ausgabe-Rangfolge, das Ranking. Schließlich geben sie Verweise auf die wichtigsten gefundenen Webseiten auf der Ergebnisseite aus.

Für den Betrieb einer Suchmaschine sind also mehrere Programme notwendig: der Crawler, der Indexer, der Teil, der für das Ranking zuständig ist, und die Ausgabe. Wenn man vom Algorithmus einer Suchmaschine spricht, zum Beispiel vom „Google-Algorithmus", ist der Ranking-Algorithmus gemeint.

Schon in den Frühzeiten der Web Economy war es wichtig, dass eine Webseite ein gutes Ranking erhält, also unter den ersten Treffern einer Suche erscheint. AltaVista & Co. machten es Website-Betreibern allerdings nicht gerade schwer, ihre Seiten unter den ersten Treffern zu platzieren, denn die Suchmaschinen zogen nur die Inhalte der indexierten Webseiten selbst für die Berechnung des Ranking heran. Website-Betreiber mussten den potenziellen Suchbegriff nur häufig genug auf ihren Seiten oder in den Meta-Keywords unterbringen, um eine gute Platzierung zu erhalten.

Die Verlinkung zählt

Hier kommen Larry Page und Sergey Brin ins Spiel, zwei Studenten der Universität Stanford – und ein Verfahren aus der Soziometrie, das schon in den dreißiger Jahren des vergangenen Jahrhunderts entwickelt wurde. Jacob Levy Moreno hatte damals die sozialen Strukturen innerhalb einer Gruppe als Netzwerk von Sympathie- und Antipathie-Bekundungen analysiert. Dabei genießt dasjenige Mitglied der Gruppe den höchsten sozialen Status, das die Sympathien der meisten Gruppenmitglieder erhält.

Der Statistiker Leo Katz hat Morenos Modell 1953 verfeinert. Dabei gehen die Sympathiebekundungen nicht direkt in die Bewertung ein. Vielmehr sind Sympathien mehr wert, wenn sie von einem Gruppenmitglied stammen, das wiederum viele Sympathien genießt.

Pages und Brins Leistung war es, Katz' Modell auf das Web zu übertragen und daraus eine funktionierende Suchmaschine zu bauen: Das gesamte Web ist die Gruppe, jede Webseite ein Gruppenmitglied und jeder Link eine Sympathiebekundung. Page und Brin nannten ihr Verfahren PageRank. Aus Sicht der verlinkten Seite heißen die Links im Suchmaschinen-Jargon auch Backlinks; bei einer Seite mit vielen Backlinks sagt man auch, dass sie eine hohe Linkpopularität besitzt. Die Mathematik hinter Page Rank beschreibt der Kasten auf Seite 71 genauer.

Page erhielt ein Patent auf PageRank (siehe ct.de/yz11). Die beiden Studenten gründeten das Unternehmen Google, ansässig im kalifornischen Mountain View, das die gleichnamige Suchmaschine betreibt.

Spam-Wettkampf

Der PageRank war lange Zeit das wichtigste Kriterium des Google-Rankings. Google hat den PageRank einer Seite ab dem Jahr 2000 sogar per Browser-Werkzeugleiste veröffentlicht. Nutzer konnten sich so den PageRank der besuchten Seiten anzeigen lassen, um deren Qualität schnell einschätzen zu können.

Der sichtbare PageRank ließ sich vermarkten. Sogenannte Search Engine Optimizers (SEO) nutzten das Werkzeug daher auch gerne: Dienstleister, die versuchen, das Ranking von Websites zu verbessern.

SEOs probierten ein riesiges Arsenal an Tricks aus, um Google ein Schnippchen zu schlagen. So wurden sogenannte Link-Farmen gebaut – große Systeme von Websites mit zigtausenden einander massiv verlinkenden Seiten, die auf diese Weise einen hohen PageRank erhalten sollten. Kommentarbereiche in Blogs und anderen Sites wurden mit Beiträgen geflutet, die Links auf Webseiten erhielten, welche damit in Google-Rankings aufgewertet werden sollten.

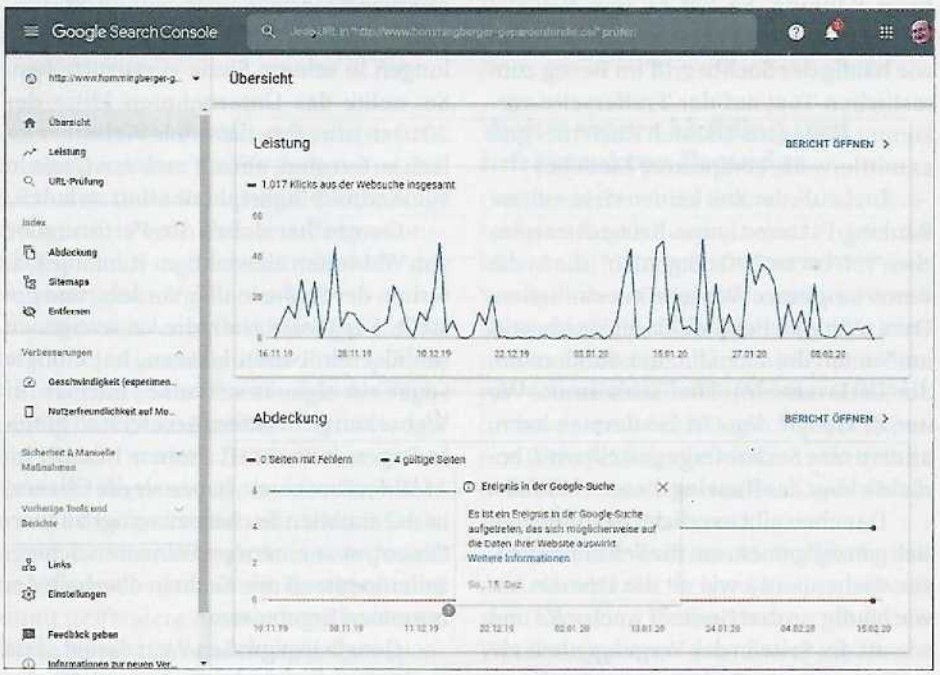
Lange Jahre lieferten sich SEOs mit Google ein Hase- und Igel-Rennen, bei dem erstere es immer mal wieder schafften, Treffer auf den ersten Google-Suchergebnisseiten zu platzieren. Google war und ist bis heute gezwungen, seine Suchmaschine immer wieder anzupassen – oder sogar das Web und die ihm zugrundeliegende Technik generell zu beeinflussen.

Für den Kommentar-Spam zum Beispiel entwickelte Google eine eigene Erweiterung des HTML-Link-Tag namens nofollow:

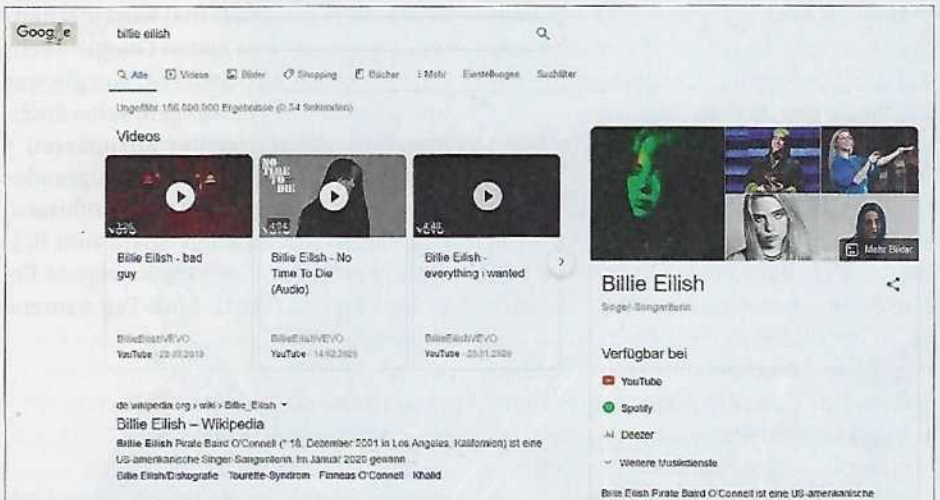
```
<a href="http://example.com/"
rel="nofollow">Beispiel</a>
```

Wenn ein Website-Betreiber einen Link damit auszeichnet, dann vererbte der Link nicht seine Linkpopularität, verhalf der verlinkten Seite also nicht zu einem besseren Ranking. Sobald Webmaster ihre Kommentarfunktionen entsprechend anpassten und Links auf diese Weise auszeichneten, gab es für Suchmaschinenoptimierer keinen Anreiz mehr, Spam-Kommentare abzusetzen.

Mittlerweile ist die Situation komplexer: Auch nofollow-Links können für das sogenannte Linkprofil einer Seite wichtig sein. Seiten, die keine nofollow-Links enthalten, können negativ auffallen. Und Links können helfen, Vertrauen zu erben. So sind nofollow-Links von der Wikipedia sehr begehrt, weil Google der Wikipedia vertraut.



Pflichtprogramm für Website-Betreiber: Google zeigt Webmastern in der Search Console, wie die Suchmaschine ihre Site sieht – und wie sie sie verändern müssen, wenn sie bessere Rankings erzielen wollen.



Patchwork: Die Suchergebnisseiten bedienen sich verschiedener Quellen, etwa des Video- und Newsindex. Der Kasten rechts speist sich aus dem Knowledge Graph – und indirekt unter anderem aus der Wikipedia.

Die Werkzeugleiste mit dem PageRank gibt es seit 2016 nicht mehr, was auch damit zu tun hat, dass der PageRank als Ranking-Faktor an Bedeutung verloren hat. Er war ohnehin nie der einzige Faktor beim Ranking. So hat es zum Beispiel immer schon eine Rolle gespielt, ob und wie häufig der Suchbegriff im Bezug zum restlichen Text auf der Trefferseite vorkommt (Keyword-Dichte). Auch hier gibt es mittlerweile komplexere Modelle.

Im Laufe der Zeit kamen viele weitere Ranking-Faktoren hinzu. Beobachter sprechen von bis zu 200 „Signalen", die in die Bewertung einer Webseite mit einfließen. Dazu zählt zum Beispiel, ob der Suchbegriff im Namen der Domain, der Subdomain, der URL oder im Titel vorkommt. Wo immer Google den Ort bestimmen kann, an dem eine Suchabfrage gestellt wird, berücksichtigt das Ranking diesen ebenfalls.

Daneben gibt es etliche Faktoren, die sich ganz allgemein um die Seite und Website drehen, etwa wie alt die Domain ist, wie häufig sie den Besitzer wechselte und wie oft die Seite in der Vergangenheit aktualisiert wurde. Sogenannter Duplicate Content, also Inhalte, die woanders in identischer Form vorkommen, können zur Abwertung führen. Der PageRank ist und bleibt aber nach wie vor ein wichtiger Faktor beim Ranking.

Das Web nach Googles Gusto

Seine große Bedeutung für Webmaster hat sich Google in der Vergangenheit öfter zunutze gemacht, um technische Entwicklungen in seinem Sinne voranzutreiben. So stellte das Unternehmen Mitte der 2010er-Jahre fest, dass viele Websites einfach zu fett sind, um auf mobilen Geräten wie Android-Smartphones flott zu laden.

Google hat daher die Performance von Webseiten als wichtiges Ranking-Kriterium der Suche in den Vordergrund gestellt. Für Webmaster, die keine eigenen Mobilseiten bauen können, hat Google sogar ein eigenes schlankes Format für Webseiten geschaffen: Accelerated Mobile Pages, kurz AMP. Stehen Inhalte im AMP-Format bereit, haben sie die Chance, in der mobilen Suche bevorzugt zu werden, etwa in einem prominenten Schlagzeilenkarussell mit Kacheln oberhalb der normalen Ergebnisse.

Google legt großen Wert darauf, dass nach wie vor neutrale Algorithmen für das Ranking verantwortlich sind und niemand von Hand an den Suchergebnissen herummanipuliert. Dennoch spielt der menschliche Faktor eine wichtige, wenn auch n indirekte Rolle.

So misst Google live die Interaktion der Besucher mit den Ergebnissen d Suchmaschine. Die Links in den Suc ergebnisseiten führen die Google-Nutz nicht direkt zu den Trefferseiten, sonde zuerst zu einem Google-Skript, das zum Ziel leitet. Google kann so messe auf welche Links einer Ergebnisseite d Benutzer klicken – Hinweise auf brauc bare Treffer – und von welchen Websi sie schnell wieder zurückkommen – H weise darauf, dass sie dort nicht gefund haben, was sie suchen.

Daneben lässt Google die Quali seiner Suchergebnisse laufend dur menschliche Tester bewerten. D 168-seitigen Katalog mit den Bewertun kriterien und reichlich Beispielen hat Go gle veröffentlicht (siehe ct.de/yz11). Side-by-Side-Tests zum Beispiel erhalt Tester zwei Versionen einer Suchergebn seite präsentiert, einmal mit einer zur D kussion stehenden Neuerung und einm ohne. Die Tester müssen dann angeb welche Version sie besser finden u warum. Nach eigenen Angaben hat Go gle allein im Jahr 2018 insgesamt 654.6 „Experimente" durchgeführt, die zu 32 Neuerungen führten.

Diese kleineren Updates finden a sehr häufig statt und Otto Normalanwe der und Website-Betreiber bekomm davon nicht viel mit. Anders verhält es s bei größeren Updates. Diese wesentlich Aktualisierungen des Suchalgorithmus schüttern häufig das gesamte Web, w durch größere Ranking-Umgewichtung oft ganze Branchen prominenter oder geschlagener in den Suchergebnissen a tauchen können. Die großen Updates t gen auch eigene Namen wie Panda (20 Penguin (2012), Hummingbird (20 oder RankBrain (2015).

Viele Töpfe

Die Ergebnisse der Google-Treffersei stammen schon lange nicht mehr nur einem Wortindex, auch wenn das nach vor der wichtigste Datentopf ist, aus d Google seine Ergebnisseiten zusamm baut. Daneben betreibt Google allerdi viele weitere spezialisierte Indizes, e für Ortsinformationen, Bilder, Vide Produkte und Bücher.

Google stellt für die Inhaltekategor eigene Suchfunktionen bereit. Jeder In stellt dabei andere Merkmale heraus. der Bildersuche etwa kann Google n

Die Mathematik hinter PageRank

Der PageRank-Algorithmus berechnet die relative Wichtigkeit einer Webseite innerhalb eines Netzwerks, und zwar in Form einer Wahrscheinlichkeitsverteilung: Er gibt für alle Seiten die Wahrscheinlichkeit an, dass ein Surfer, der zufällig auf Links von Seite zu Seite klickt, auf einer bestimmten Seite landet:

$$r(A) = \left(\frac{r(B_1)}{|B_1|} + \ldots + \frac{r(B_n)}{|B_n|}\right)$$

Bei B_1 bis B_n handelt es sich um die Seiten, die auf A verlinken, $|B_i|$ ist die Anzahl der jeweils ausgehenden Kanten. Die Formel ist rekursiv: Jeder PageRank einer Seite hängt vom PageRank aller anderen Seiten ab.

Um die PageRanks zu berechnen, baut man zunächst eine Matrix, die die Verlinkung innerhalb des Netzwerks von Webseiten abbildet (siehe Abbildung rechts). Dabei stellt jedes Matrixelement m(x,y) die Wahrscheinlichkeit dar, mit der ein Surfer, der sich auf der Seite x befindet und zufällig auf einen darin befindlichen Link klickt, zur Webseite y gelangt. Die Wahrscheinlichkeiten auf einer Seite sind gleich verteilt, weshalb alle von einer Webseite ausgehenden Links, also alle Links einer Spalte das gleiche Gewicht von 1/(Anzahl der Links) der Seite teilen. Links auf sich selbst zählen nicht, daher stehen dort jeweils Nullen.

Die Page-Rank-Berechnung startet mit einem Vektor, der für jede Webseite denselben Wert 1/(Anzahl der Webseiten) enthält. Der Vektor wird rechts mit der Matrix multipliziert. Diese Multiplikation wird mit dem Ergebnis so oft wiederholt, bis die Änderungen im Ergebnisvektor unter einen gewissen Schwellenwert fallen. In den ersten beiden Schritten ergibt sich:

$$\begin{pmatrix} 0 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} = \begin{pmatrix} 0.125 \\ 0.375 \\ 0.375 \\ 0.125 \end{pmatrix}$$

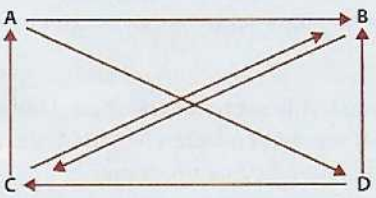
$$\begin{pmatrix} 0 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} 0.125 \\ 0.375 \\ 0.375 \\ 0.125 \end{pmatrix} = \begin{pmatrix} 0.1875 \\ 0.3125 \\ 0.4375 \\ 0.0625 \end{pmatrix}$$

Bei der eingangs genannten Formel handelt es sich um eine Vereinfachung – mit der sich das Prinzip aber besser verstehen lässt. Tatsächlich rechnet die Formel noch einen Dämpfungsfaktor mit ein, der berücksichtigt, dass Surfer nicht endlos weitersurfen, sondern mit einer gewissen Wahrscheinlichkeit aufhören – siehe die Formel auf der Tafel im Aufmacher.

Das Web umfasst heute zig Milliarden von Webseiten. Matrixmultiplikationen in dieser Größenordnung sind natürlich sehr aufwendig. Page hat in einem Paper bereits 1998 viele Optimierungen angedeutet, die in der Praxis zum Einsatz kommen, um den Rechenaufwand in Grenzen zu halten (siehe ct.de/yz11). Die Details der Implementierung sind aber Firmengeheimnis – wie viele andere Details des Rankings.

Links und Matrix

Als Basis für die Berechnung des PageRank wird die Verlinkungsstruktur in eine Matrix übertragen.



	A	B	C	D
A	0	0	0.5	0
B	0.5	0	0.5	0.5
C	0	1	0	0.5
D	0.5	0	0	0

der Größe, nach Farben oder Nutzungsrechten fahnden; die Produktsuche ermöglicht es unter anderem, einen Preisrahmen vorzugeben. Die News-Suche hat ebenfalls einen eigenen Index mit sich schnell aktualisierenden Nachrichtensites. Je nach Suchanfrage streut Google Inhalte dieser Quellen in die Ergebnisse des Hauptindex auf den Suchergebnisseiten ein.

Künstliche Intelligenz spielt schon seit einigen Jahren eine große Rolle bei der Google-Suche. So hat Google bereits 2012 begonnen, eine vernetzte Datenbank des Weltwissens aufzubauen, den sogenannten Knowledge Graph. Dieser Graph besteht aus hunderten von Millionen sogenannter Entitäten und ihren Beziehungen zueinander.

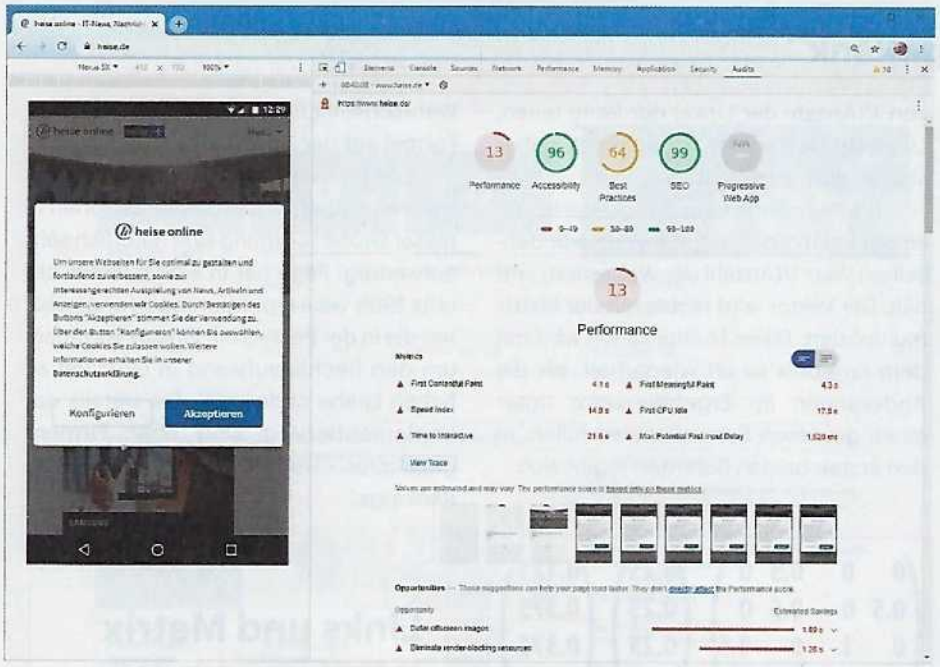
Seine Inhalte für den Knowledge Graph bezieht Google aus der Wikipedia, Wikidata, dem CIA World Factbook und aus vielen anderen Websites, die ihre Informationen extra dafür aufbereiten. Inhalte aus dem Knowledge Graph zeigt Google in Form kleiner Informationskästen rechts neben oder oberhalb der normalen Suchergebnisse an.

Der Knowledge Graph und generell KI werden eine immer größere Rolle spielen, insbesondere beim Verständnis der Suchanfragen. Erst im Dezember hatte Google ein Update namens BERT (Bidirectional Encoder Representations from Transformers) scharf geschaltet, mit dem die Suchmaschine komplette, als Satz an sie gestellte Abfragen besser als solche erkennt und somit treffendere Antworten geben kann. Diese Neuerung ist auch deshalb so wichtig, weil die Suche immer häufiger in Form des Assistenten unter Android und auf den Google-Home-Lautsprechern zum Einsatz kommt.

Webmasters Liebling und Infrastruktur-Provider

Google macht Webmastern mit vielen kostenlosen Angeboten das Leben leichter. So können sie nicht nur per Google AMP ihre Seiten hosten. Google bietet schicke Fonts an, die sie in ihre Websites einbinden können und hostet populäre JavaScript-Librarys – schnell und für den Webmaster Bandbreite sparend. Mit Google Analytics können Website-Betreiber zudem sehr detailliert die Besucherströme auf ihren Angeboten analysieren.

Last, not least steckt Google hinter dem mit Abstand meistverbreiteten Browser: Chrome. Einige darin enthaltene Werkzeuge dienen dem Zweck, Seiten so zu gestalten, dass die Suchmaschine sie gut finden und indexieren kann. All diese Angebote sind kein reiner Selbstzweck. Sie helfen Google dabei, die weltweiten Sur-



Google beherrscht die Browser-Szene mit seinem Browser Chrome. Die darin enthaltenen Werkzeuge sollen Entwickler dazu bringen, suchmaschinengerechte Seiten zu bauen.

fermassen besser zu verstehen. Und die damit erhobenen Daten haben wiederum Einfluss auf die Suchmaschine.

Lukrative Versteigerung

Spricht man über die Google-Suche, muss man die Werbung erwähnen. Anzeigen sind zwar auf den Suchergebnisseiten als solche gekennzeichnet und Google wird nicht müde zu betonen, dass man sich eine Platzierung in der Suche nicht kaufen kann. Werbung ist aber ein wesentlicher Bestandteil der Ergebnisseiten, die jeder Google-Nutzer zu Gesicht bekommt. Und nicht ganz nebensächlich kommt hier ebenfalls ein interessanter Algorithmus zum Einsatz.

Auf der Suche nach einem Geschäftsmodell für ihre Suchmaschine kopierten Page und Brin kurzerhand das Modell einer anderen Suchmaschine, Overture. Seither können Werbetreibende mit Google Adwords (heute Google Ads) Platzierungen auf den Suchergebnisseiten zu bestimmten Suchbegriffen ersteigern: Wer einen höheren Preis bezahlt, dessen Anzeige stellt Google potenziell prominenter dar. Außer dem Gebot gehen allerdings noch weitere Faktoren ins Ranking mit ein.

Diese Entscheidung für das Erlösmodell Werbung hat den Kurs des gesamten Unternehmens Google entscheidend mitgeprägt. Seither ist Google ein Werbekonzern, der ein riesiges Werbenetz mit tausenden Partner-Websites betreibt und den Großteil seiner Umsätze mit Online-Marketing macht.

Foul gespielt

Google tritt auf verschiedenen Gebieten sowohl als „neutraler" Gatekeeper für das restliche Netz, gleichzeitig aber auch als eigener Player in Erscheinung. Das gilt zum Beispiel bei der Produkt- und Preisrecherche. Google liefert bei Suchanfragen, die sich um Produkte drehen, Links zu Preisvergleichsdiensten – betreibt aber zudem selber einen.

Solche Doppelfunktionen führen oft zu Ärger mit den jeweiligen Konkurrenten. Andere Webdienste-Anbieter kritisieren immer wieder, dass Google seine Suchmaschine nutze, um eigene Angebote auf den Suchergebnisseiten in den Vordergrund zu stellen.

So hat die EU-Kommission 2017 eine Geldbuße in Höhe von 2,4 Milliarden Euro über Google verhängt: Der Suchmaschinen-Betreiber habe seine marktbeherrschende Stellung als Suchmaschine missbraucht, um den eigenen Preisvergleichsdienst zu bevorzugen. Als zusätzliche Auflage musste Google seine Suchergebnisseiten anpassen.

Der Streit flammt derzeit wieder auf. So haben 41 Google-Konkurrenten Beschwerde bei der EU-Kommission eingelegt, weil Google seinem eigenen Shopping-Service wettbewerbswidrige Vorteile verschaffe. Weitere Wettbewerbsverfahren in verschiedenen Ländern laufen ebenfalls noch.

Staatliche Eingriffe

Als zentraler Wegweiser im Netz finden sich auch immer wieder Links auf illegale Inhalte unter Googles Treffern. Regierungen und Strafverfolgungsbehörden ersuchen Google daher immer wieder, bestimmte Treffer zu löschen. Google ist nach eigenem Bekunden bestrebt, in jedem Land die Gesetze zu erfüllen und monierte Links aus den Suchergebnislisten zu streichen.

Auf den Trefferseiten weist Google auf solche Löschungen hin: „Als Reaktion auf ein rechtliches Ersuchen, das an Google gestellt wurde, haben wir <x> Ergebnis(se) von dieser Seite entfernt." Dazu gibt es einen Link auf die Seite der Lumen Database, wo Interessierte weiterführende Informationen zu Sperrungen finden. In seinem Transparenzbericht fasst Google zusammen, welche Länder wie viele Inhalte seit 2010 aus welchen Gründen entfernt haben wollten.

Ein paar Jahre lang betrieb Google eine Version seiner Suchmaschine für China, die den strengen Zensurvorgaben der chinesischen Regierung genügte. Später hat es diese aus Protest gegen die chinesische Zensur eingestellt. Seither sperrt China fast alle Google-Dienste.

Grundstein für ein Imperium

So breit Google heute aufgestellt sein mag: Die Suche ist und bleibt der Markenkern und der wichtigste Dienst des Konzerns. Im Laufe der Jahre hat Google viel Arbeit investiert, um die Suche immer wieder anzupassen und zu erweitern. Auch wenn PageRank heute nicht mehr so einen zentralen Einfluss auf die Suche hat: Google definiert sich bis dato durch die hohe Anzahl an technischen Mitarbeitern und den hohen Stellenwert von Algorithmen in allen seinen Produkten.

Website-Betreiber müssen immer darauf achten, was Google von ihnen will. Das Unternehmen aus Mountain View setzt viele der Standards in der Web- und Mobilwelt. Aber auch alle anderen Unternehmen der IT-Szene sollten im Auge behalten, was Google in seinen Forschungs- und Entwicklungszentren treibt. Vielleicht entsteht dort ja gerade so etwas wie ein PageRank der Zukunft. *(jo@ct.de)* **ct**

Weiterführende Informationen: *ct.de/yz11*

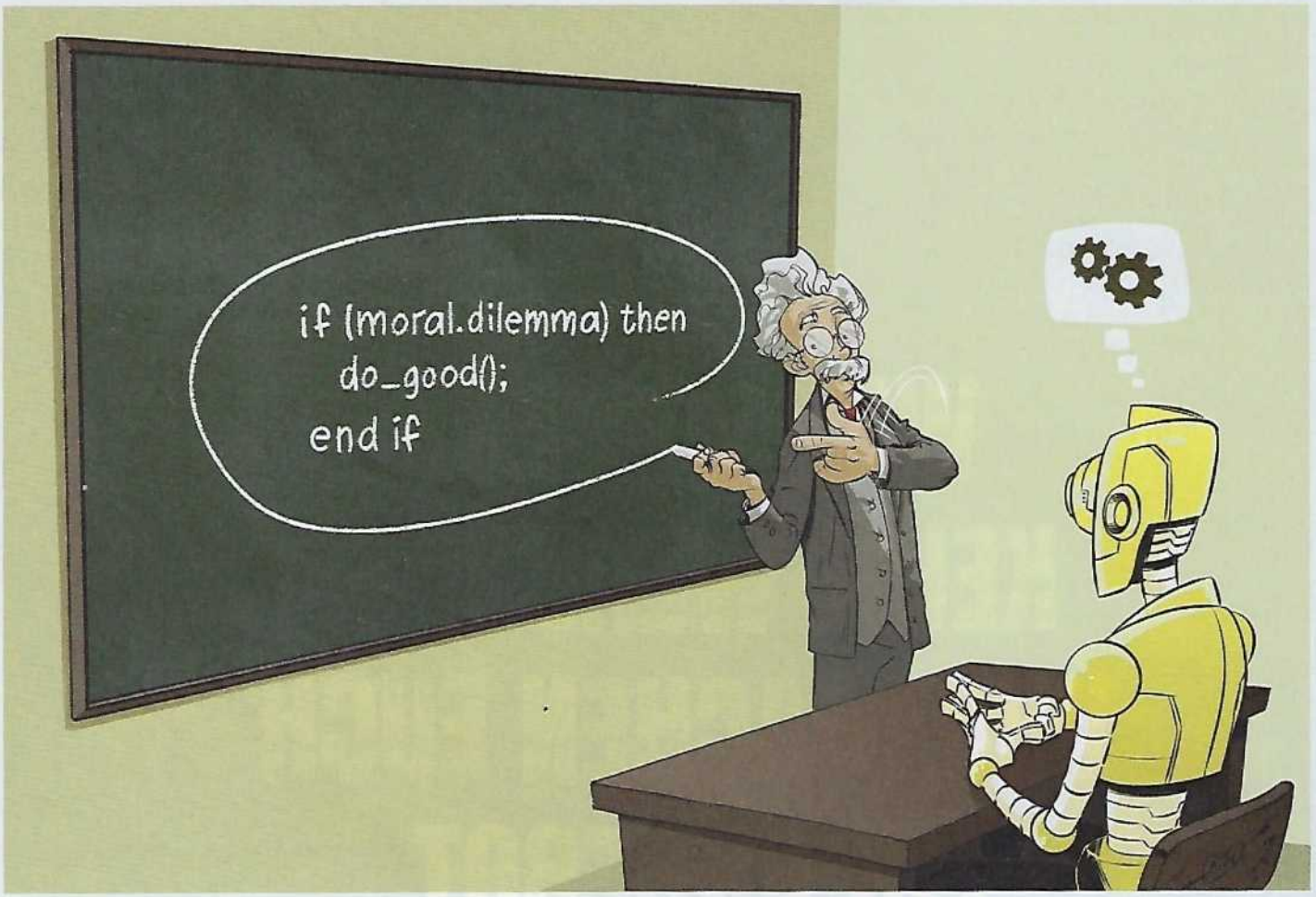


Bild: Albert Hulm

Ethik für Algorithmen

Wie man verhindern will, dass Software Böses tut

Algorithmen sollen fair sein, transparent und nachvollziehbar – solche oder ähnliche ethischen Leitlinien haben derzeit Konjunktur. Aber braucht man überhaupt neue Regeln und lassen sich diese immer umsetzen?

Von Sylvester Tremmel

Algorithmen sind fester Bestandteil unserer Gesellschaft. Sie erleichtern das Leben in unzähligen Situationen, vernetzen Menschen mit Gleichgesinnten und navigieren in unbekannten Städten. Dieselben Algorithmen können aber zum Problem werden, wenn man sie nicht kontrolliert und versteht. Ein System, das Gleichgesinnte zusammenführt, schließt schon fast zwingend Andersdenkende aus. Wenn es zu viel oder zu unkontrolliert Einfluss gewinnt, kann es Gesellschaften spalten und letztlich demokratische Mechanismen gefährden.

Solche Befürchtungen sind nicht aus der Luft gegriffen, immer wieder hört man von Systemen, die an gesellschaftlich relevanten Punkten zum Einsatz kommen und dann unbeabsichtigtes Verhalten zeigen. Bekanntes Beispiel ist ein von Amazon entwickeltes System zur Beurteilung von Bewerbern – bei dem sich zeigte, dass es systematisch Frauen benachteiligt [1]. Noch bedenklicher ist der ebenfalls recht berühmte Fall des Systems COMPAS, das in den USA die Rückfallwahrscheinlichkeit von Straftätern berechnet. Es hat damit direkt Einfluss auf die Länge der verhängten Haftstrafen. Studien werfen dem System vor weitgehend nutzlos zu sein und – noch schlimmer – das Rückfallrisiko von schwarzen Angeklagten grundsätzlich zu über- und das von weißen Angeklagten zu unterschätzen [2].

Ethische Leitlinien

Aufgrund solcher Fälle erfährt das Thema Ethik in der IT zurzeit große Aufmerksamkeit. Projekte wie „Algorithmenethik" der Bertelsmann Stiftung oder die „AG Ethik" der Initiative D21 schlagen Regelwerke zur Einhegung solcher Systeme vor, die mal als „algorithmisches System", mal als „künstliche Intelligenz" (KI) bezeichnet werden. Auch die Politik hat diverse Initiativen gestartet: auf Bundesebene unter

anderem die Datenethikkommission und die Enquete-Kommission „Künstliche Intelligenz". Und die EU-Kommission hat Mitte Februar ein Weißbuch zur Künstlichen Intelligenz vorgestellt.

Zusätzlich verpassen sich Firmen von Google bis hin zur Telekom ethische Leitlinien, sei es, weil sie die gesellschaftlichen Bedenken teilen oder einfach nur, um nicht amoralisch oder altbacken zu wirken. Die gemeinnützige Organisation AlgorithmWatch pflegt eine Liste solcher Ethikleitlinien, die bereits über 80 Einträge aus Politik, Wirtschaft und Gesellschaft enthält (alle Links unter ct.de/yxrv).

In den meisten Regelwerken finden sich inhaltlich recht ähnliche Forderungen. Teilweise sind das Selbstverständlichkeiten, die ohnehin gelten würden. „Wir unterstellen unsere KI-Systeme Recht und Gesetz", schreibt etwa die Telekom, ist mit solchen Banalitäten aber nicht alleine. Außerdem finden sich in vielen Leitlinien Forderungen nach Verantwortlichkeit, Sicherheit, Nichtdiskriminierung und der Möglichkeit, Entscheidungen des Systems zu revidieren. Solche Forderungen mögen zwar offensichtlich wirken, aber sie sind nicht inhaltsleer, sondern weisen auf tatsächliche Unzulänglichkeiten hin.

Handlungsbedarf besteht zum Beispiel für Forscher und Entwickler: Wie man Nichtdiskriminierung in einem „algorithmischen System" konkret realisiert, weiß nämlich niemand so genau. Wie implementiert man diese Eigenschaft im Programmcode? Wie testet man sie? Kontrolliert man Systeme regelmäßig? Wie sieht so eine Kontrolle aus?

Dass diese Fragen noch nicht zufriedenstellend beantwortet sind, ist den meisten Beteiligten durchaus bewusst. Algo.Rules, ein Gemeinschaftsprojekt der Bertelsmann Stiftung und des iRights.Lab, schreibt: „Die aktuelle Herausforderung der Algo.Rules besteht [...] in ihrer harten Umsetzung. Unsere Analyse hat gezeigt, dass viele der anderen, bestehenden Kriterienkataloge daran gescheitert sind, weil sie entweder nie einen Anspruch auf Implementierung hatten oder ihre Umsetzungsstrategien nicht gefruchtet haben."

Unklar ist auch, welche Systeme überhaupt von solchen Regeln betroffen sein sollen. Schließlich kann und will man nicht plötzlich auch Taschenrechner-Software und Tetris-Klone auf Sicherheit und Diskriminierungspotenzial hin prüfen. Viele Leitlinien reden relativ schwammig von gesellschaftlicher Relevanz und ähnlichen Kriterien, ohne sie genauer zu definieren. Etwas konkreter sind Vorschläge für abgestufte Vorgehensweisen, wie sie etwa von Datenethikkommission (DEK) kommen. Die DEK schlägt in ihrem Abschlussgutachten vor, Anwendungen nach ihrer „Systemkritikalität" einzuteilen. Zum Beispiel die Algorithmen in einem Getränkeautomaten wären laut DEK so unkritisch, dass sie nicht weiter geprüft werden müssten. Algorithmen, die Kreditwürdigkeiten beurteilen, wären dagegen sehr kritisch und sollten sogar unter kontinuierlicher Kontrolle stehen.

Allgemein definiert die DEK die Systemkritikalität als eine „Kombination aus der Wahrscheinlichkeit eines Schadenseintritts und der Schwere des zu befürchtenden Schadens." Solche Ansätze sind in der Technikphilosophie zwar verbreitet, aber die DEK lässt letztlich offen, wie die Berechnung dieser „Kombination" konkret aussehen könnte.

Schwierige Transparenz

Eine weitere Forderung, die sich in fast allen Leitlinien findet, lautet, algorithmische Systeme transparent und nachvollziehbar zu gestalten. Es soll also bekannt sein, wenn ein solches System eingesetzt wird und wie es allgemein arbeitet.

Man erhofft sich davon verschiedene Effekte. Zum einen wird fehlerhaftes Verhalten besser erkennbar: Operiert ein System mit einem Zusammenhang zwischen Geschlecht und Einstellungswahrscheinlichkeit, liegt etwas im Argen. Zum anderen soll im konkreten Einzelfall überprüft werden können, ob – und wenn ja warum – eine Entscheidung fehlerhaft war. Grundsätzlich soll Transparenz verhindern, dass Menschen entmündigt werden. Wenn man etwa gar nicht wüsste, dass ein algorithmisches System in Entscheidungen involviert ist, könnte man diese auch nicht hinterfragen.

Transparenz klingt gut, erweist sich aber bei selbstlernenden Systemen als handfeste technische Herausforderung. Solchen Systemen muss man nicht jeden

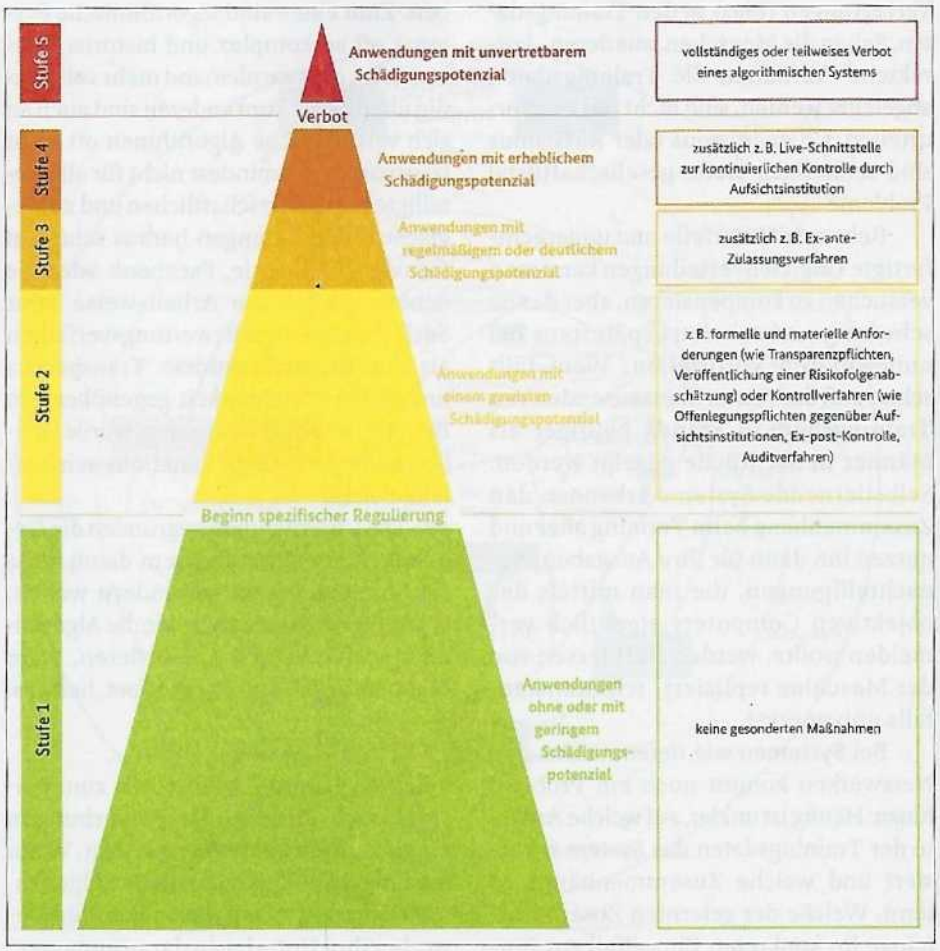


Bild: Gutachten der Datenethikkommission der Bundesregierung

Fünf Stufen „Systemkritikalität" schlägt die Datenethikkommission vor. Algorithmische Systeme in Stufe 1 will sie nicht regulieren, Stufe-5-Systeme verbieten.

Bild: Fraunhofer HHI

Eine KI erkennt ein Pferd. Erst die erklärende Heatmap offenbart: Erkannt wurde eigentlich nur ein Copyright-Zeichen.

Zusammenhang manuell in Form von Code eingegeben – darin liegt ihr großes Potenzial. Stattdessen leiten die Systeme Zusammenhänge selbstständig aus (großen Mengen von) Trainingsdaten ab.

Was genau so ein System aber „lernt", wenn es Trainingsdaten analysiert, ist auch für seine Entwickler und Anwender nur schwer zu sagen. Ein Problem sind Verzerrungen (Bias) in den Trainingsdaten. Schon die Menschen, aus deren „korrektem" Verhalten die Trainingsdaten abgeleitet werden, sind nicht frei von Vorurteilen. Chauvinismus oder Rassismus sind schließlich uralte gesellschaftliche Probleme.

Bekannte Vorurteile und ungerechtfertigte Ungleichverteilungen kann man versuchen zu kompensieren, aber das ist schwierig und scheitert spätestens bei unbewussten Vorurteilen. Wem fällt schon auf, dass unter abertausenden von Trainingsbildern Frauen häufiger als Männer in der Küche gezeigt werden. Selbstlernende Systeme erkennen den Zusammenhang beim Training aber und nutzen ihn dann für ihre Ausgaben. Benachteiligungen, die man mittels des objektiven Computers eigentlich vermeiden wollte, werden stattdessen von der Maschine repliziert, schlimmstenfalls unbemerkt.

Bei Systemen wie tiefen neuronalen Netzwerken kommt noch ein Problem hinzu: Häufig ist unklar, auf welche Aspekte der Trainingsdaten das System fokussiert und welche Zusammenhänge es lernt. Welche der gelernten Zusammenhänge im konkreten Einzelfall zur Entscheidung führten, lässt sich ebenfalls nur schwer feststellen – wenn überhaupt. Wissenschaftler forschen daher intensiv an erklärbarer KI (explainable AI) [3]. Interessante Fortschritte gibt es beispielsweise im Bereich der Bilderkennung, aber „Glassbox"-KIs, deren Verhalten allgemein verständlich, nachvollziehbar und erklärbar ist, gibt es noch nicht.

Auch bei klassisch programmierter Software ist Transparenz oft nicht gegeben. Zum einen sind algorithmische Systeme oft so komplex und historisch gewachsen, dass sie niemand mehr vollständig überblickt. Zum anderen sind auch an sich verständliche Algorithmen oft nicht transparent – zumindest nicht für alle Beteiligten. Aus wirtschaftlichen und strategischen Überlegungen heraus schützen Firmen wie Google, Facebook oder die Schufa die genaue Arbeitsweise ihrer Such-, Sortier- und Bewertungsverfahren als Geschäftsgeheimnisse. Transparenz und Nachvollziehbarkeit gegenüber den Betroffenen zu gewährleisten, würde letztlich bedeuten, deren Funktionsweise offenzulegen.

Die Unternehmen begründen die Geheimhaltung unter anderem damit, dass sie Manipulationen verhindern wollen. Wäre allgemein bekannt, wie die Algorithmen sortieren und aussortieren, wäre Missbrauch Tür und Tor geöffnet, heißt es.

Parteiische Computer

Solches „Gaming" befürchten zum Beispiel auch Firmen, die Bewerbungen algorithmisch analysieren lassen. Wenn die Entscheidungskriterien des Algorithmus bekannt wären, dann könnten Bewerber ihre Unterlagen daran anpassen. Sie würden dann ihre Anschreiben und Lebensläufe auf bekannte Schwachstellen oder Eigentümlichkeiten des Algorithmus hin optimieren. Für die Firmen ist das selbst dann ein Problem, wenn der Bewertungsalgorithmus menschliche Personaler nur unterstützen soll: Menschen neigen dazu, die Ergebnisse eines Computers als neutral, unvoreingenommen und letztlich korrekt hinzunehmen. Eine Bewerbung, die eigentlich zu schön ist, um wahr zu sein, wird dann vielleicht nicht kritisch hinterfragt. Das hauseigene Bewertungssystem hat ihr ja gute Noten gegeben.

Dieses ungerechtfertigte Vertrauen in Maschinen ist ein schon länger bekanntes Problem, man nennt es „Automation Bias". (Mit dem oben beschrieben Bias in Trainingsdaten hat das nichts zu tun.) Größere Transparenz – etwa durch Begründungen oder Grafiken – kann den Automation Bias paradoxerweise noch verschärfen: Anstatt Ergebnisse anhand der mitgelieferten Argumente zu hinterfragen, nehmen Menschen wahr, dass der ohnehin faire Computer auch noch gute Gründe für seine Entscheidung präsentiert [4].

Trotzdem sind Transparenz und Nachvollziehbarkeit sinnvolle Forderungen an algorithmische Systeme. Dem Automation Bias versucht man mit besseren „Erklärungen" beizukommen. Wichtig ist dabei etwa, Ergebnisse nicht einfach als Fakt zu präsentieren, sondern einzuordnen und auf Unsicherheiten hinzuweisen.

Moralische Maschinen

Der Automation Bias ist ein Effekt, der auch die gesamte Debatte über ethische Algorithmen beeinflusst. Menschen neigen dazu, die Fähigkeiten von Computern zu überschätzen. Das gilt besonders, wenn sie mit Namen wie „künstliche Intelligenz" betitelt werden. Bei solchen Begriffen stellt man sich leicht Computersysteme vor, die Bewusstsein und Willen haben und dadurch auch Verantwortung für ihr Tun übernehmen können. Darüber machen sich zwar nicht nur Philosophen Gedanken, zum Beispiel unter dem Schlagwort „E-Person", aber aktuelle KIs sind weit von solchen Fähigkeiten entfernt.

Deshalb ist auch die Debatte um „ethische Algorithmen" kurios: In ihr kommt die Idee zum Ausdruck, man könnte moralische Ansprüche direkt an Algorithmen richten. Das ist ungewöhnlich, weil Algorithmen und algorithmische Systeme technische Produkte sind. Moralische Ansprüche richten sich normalerwei-

se gegen die Urheber und Anwender solcher Produkte, aber nicht gegen die Dinge selbst. Man fordert ja auch nicht „ethische Messer". Stattdessen sind Messerproduzenten und -nutzer aufgefordert, sich moralisch zu verhalten.

Solange KIs noch keine Science-Fiction-gleichen Fähigkeiten haben, ist es deshalb auch nicht sinnvoll, moralische Ansprüche an sie zu stellen. Wer fordert, dass ein Algorithmus nicht diskriminieren darf, meint nach Stand der Technik eigentlich etwas anderes: Der Algorithmus darf nur so implementiert und eingesetzt werden, dass es nicht zu Diskriminierung kommt. Letztlich werden also doch Ansprüche an Entwickler und Nutzer gestellt.

Leitlinien für Entwickler

Das wirft die Frage auf, wie nötig viele der Leitlinien für KIs sind. Richtlinien für die Entwickler von KIs – und von allen anderen Softwareprodukten – gibt es nämlich bereits, in Form von Berufsethiken. Ähnlich dem Hippokratischen Eid der Mediziner, haben auch Ingenieure und Informatiker ethische Grundsätze. In Deutschland stellen etwa der Verein Deutscher Ingenieure und die Gesellschaft für Informatik (GI) solche Anforderungen an ihre Mitglieder. International gibt es unter anderem den „Code of Ethics and Professional Conduct" der amerikanischen Association for Computing Machinery (siehe ct.de/yxrv).

Ganz unabhängig von konkreten Systemen oder Algorithmen schreiben beispielsweise die Leitlinien der GI vor, dass GI-Mitglieder bereit sind, „das eigene und das gemeinschaftliche Handeln im gesellschaftlichen Diskurs kritisch zu hinterfragen und zu bewerten". GI-Mitglieder sollen auch „Verantwortung für die sozialen und gesellschaftlichen Auswirkungen [ihrer] Arbeit" tragen und ermöglichen, dass „von IT-Systemen Betroffene" selbstbestimmt handeln können.

Verschiedentlich wird gefordert, neben solchen bestehenden Leitlinien auch noch gezielte Berufsethiken für „Algorithmiker" oder Datenwissenschaftler einzuführen. Ob das nützlich ist, scheint fraglich: Damit Professionsethiken nicht bloße Papiertiger sind, muss die Öffentlichkeit sie nämlich kennen und einfordern. Außerdem müssen sich die Mitglieder der Profession den Leitlinien verpflichtet fühlen, wie beim jahrtausendealten Eid des Hippokrates. Schon die Informatik – eine relativ junge Wissenschaft – tut sich hier schwer. Die gesellschaftlichen Ansprüche an Informatiker und ihre moralischen Verpflichtungen sind noch relativ stark im Fluss. Nicht wenige Informatiker haben außerdem noch nie von den Leitlinien der GI gehört – oder sie bereits wieder vergessen.

Fazit

Bei der Vorstellung des Weißbuchs zur Künstlichen Intelligenz forderte EU-Kommissionspräsidentin Ursula von der Leyen, man müsse „Hochrisiko-KI – das heißt KI, die potenziell die Rechte von Menschen beeinträchtigt – testen und zertifizieren". Bis das aber wirklich so einfach läuft „wie bei Autos, chemischen Produkten, Kosmetika oder Spielzeug", wie Frau von der Leyen sich das vorstellt, bleiben noch viele Fragen zu klären.

Unterwegs wird sich wohl auch die Politik von einigen Vorstellungen trennen müssen: Es sei immens wichtig, Daten ohne Bias zu haben, sagt Frau von der Leyen. Algorithmische Systeme ohne Bias könne es nicht geben, steht dagegen in den Leitlinien der Initiative D21, weswegen „klare und verbindliche Richtlinien zum Umgang mit diesen formuliert werden" müssten.

Vielversprechende Forschungsansätze wie „erklärbare KI" können helfen, derartige Richtlinien zu erfüllen. Zuerst aber müssen Gesellschaft und Politik diese Vorgaben schaffen, idealerweise in einer sachlichen und übersichtlichen Diskussion. Momentan versuchen eher zu viele Köche den Brei zu verbessern. *(syt@ct.de)* c't

Literatur

[1] Martin Holland, Amazon: KI zur Bewerbungsprüfung benachteiligte Frauen, heise online: https://heise.de/-4189356

[2] Peter-Michael Ziegler, Im Namen des Algorithmus, Wenn Software Haftstrafen verhängt, c't 25/2017, S. 68

[3] Arne Grävemeyer, Intelligenztest für KI, Wie sich KI-Entscheidungen überprüfen lassen, c't 6/2020, S. 58

[4] Will Douglas Heaven, Why asking an AI to explain itself can make things worse, MIT Technology Review: www.technologyreview.com/s/615110

Ethische Leitlinien: *ct.de/yxrv*

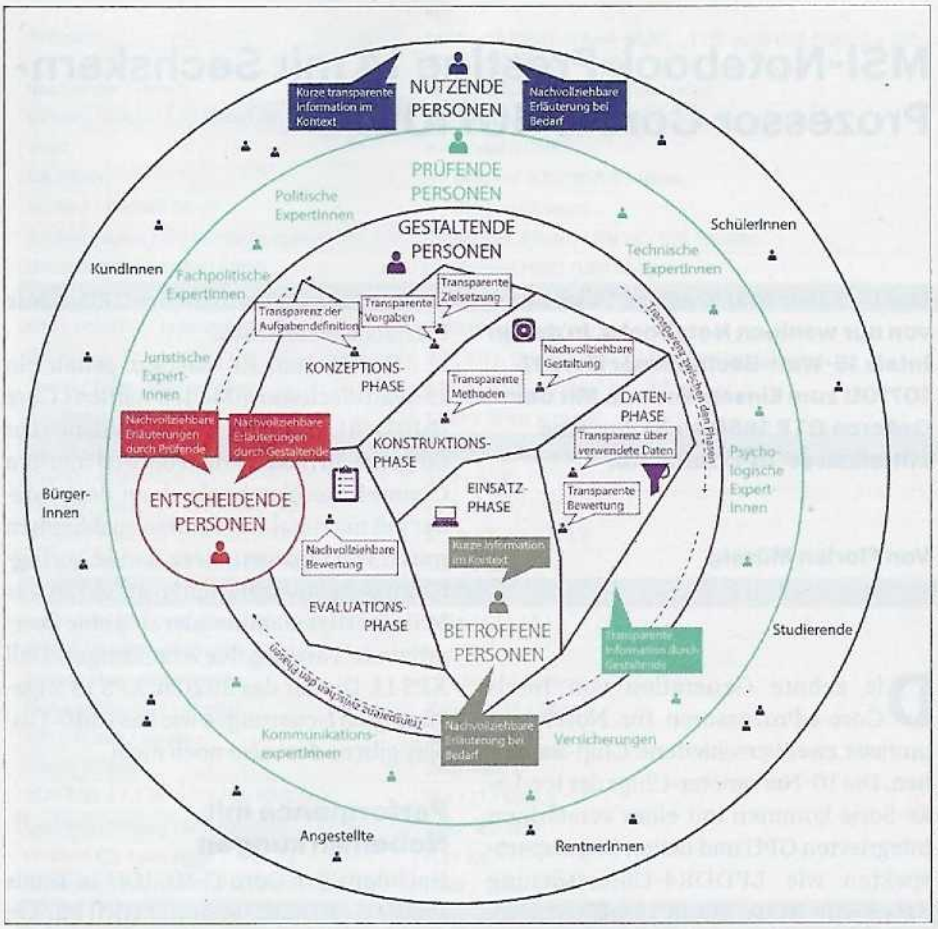


Bild: Initiative D21, #Algomon: 9 Leitlinien zum ethischen Umgang mit Algorithmen: Monitoring

Transparenz nach Personenkreis. Schon das Übersichtsdiagramm der „AG Ethik" der Initiative D21 zeigt, dass das Thema komplex ist.