DGI-Praxistage 2018

Manfred Hauer*

Ist Künstliche Intelligenz der Traktor des Dienstleistungssektors?

Eigene Erfahrungen mit maschinellen Verfahren in Bibliothek und Dokumentation

<https://doi.org/10.1515/iwp-2019-2003>

Statistische Ämter, Kaufleute, Ökonomen, Gesellschafts-, Natur- und Ingenieurwissenschaften versuchen Ähnliches in Gruppen zusammenzufassen, zu vermessen und Trends zu berechnen und nutzen dafür Statistik, mathematische Methoden. Dabei entstanden Modelle, die sich in der amtlichen Statistik oder der Gesellschaftswissenschaften lange am Rande bewegten. Sie sind der Ursprung dessen, was heute als „Artificial Intelligence“ (AI oder KI) und „Neuronale Netze“ begrifflich werbewirksam von Informatikern aufgegriffen und in technische Implementierungen umgesetzt wird. Es geht um Häufigkeiten, Korrelationen, Ähnlichkeiten, Gruppen/Cluster, Nähe von Clustern, Regeln und dynamische Veränderungen und auch um immer öfter daran angekoppelte technische Aktionen, z.B. die Steuerung von Fahrzeugen oder Antworten auf gesprochene oder geschriebene Aussagen und dahinter liegende Suchprozesse.

Bibliotheken dienten immer dazu, Aussagen zwischenzuspeichern und bei Bedarf zu liefern. Information Retrieval war zunächst die technische Repräsentation klassischer Papierspeicher: Kataloge/Datenbanken mit schneller Sortierung und Ähnlichkeitsfindung über Mengenlehre – AND/OR/NOT. Erst in Form der Randlochkarte und seit den 1950/60er Jahren als Datenbanken und seit den 70er Jahren als internationale Online-Datenbanken. Die Dokumentare, ein Teilbereich des Bibliothekswesens, erweiterten vor ca. 160 Jahren den Informationsraum durch mehr verbale Beschreibung mittels Deskriptoren, zunehmend sortiert in einem Thesaurus und durch Abstracts und bezogen auf schnell wachsende neue Medientypen, Aufsätze aus Zeitschriften und Patente.

Die Einbeziehung von Wechselwirkungen (Verlinkung und Gewichtung) zwischen den repräsentierten Objekten und Profilen/Profilgruppen waren für die heutigen

Internet-Giganten der Durchbruch. Also „neuronale“ Strukturen. Garfield war seit den 50er Jahren mit dem Social Science Citation Index der dokumentarische Vorläufer, aber es war noch zu früh für einen technischen Merge mit den restlichen Dokumentaren. Doch Linking fing schon viel früher an, mit Verweisen in Enzyklopädien, Fußnoten und Literaturzitaten – und es kommt aus dem Alltag: „Mein Papa hat gesagt, dass ...“ und schon steigt das Gewicht.

Heutige Information Retrieval-Systeme wie die Google Search oder Apache Lucene und seine Derivate (SOLR, ElasticSearch) und daraus abgeleitete Applikation (VU-Find, Primo u.v.a) nutzen Methoden, um eine möglichst hohe statistische Ähnlichkeit zwischen der Anfrage und Antworten zu liefern und soweit verfügbar Gewichtungen über Linking-Strukturen einzubeziehen. Hinter dem Ranking stehen maschinelle Indexierungs- und Klassifizierungsverfahren – Ähnlichkeitsoperationen. Sie stehen im Wettbewerb mit intellektueller Bearbeitung. Dies hat starke Auswirkungen auf bezahlte menschliche Arbeit und deren Akzeptanz. Als Nutzer will jeder in einer Sekunde die richtige Antwort, als Bibliothekar und Dokumentar will niemand seine vertraute Art der Erschließung verlieren.

1 Eigene Erfahrung mit maschinellen Verfahren

Schon im ersten Semester des sozialwissenschaftlichen Grundstudiums 1975 nutzte ich SPSS (Statistical Package für Social Science), das heute IBM gehört und deren KI-Wunder Watson möglich machte. Computerlinguistik gehörte zum Studium der Informationswissenschaft von 1981 an und die Nutzung von international zugänglichen Information Retrieval-Systemen, den „Online-Datenbanken“, wovon ab 1983 mein Beruf wurde, zunächst als Informationsvermittler. Die Migration von Hostsystemen in Richtung In-house-Hosts, dann kollaborativen Groupware-Lösungen und schließlich bis auf den eigenen PC und synchrone

*Kontaktperson: Dipl.-Inf.wiss. Manfred Hauer, M.A., AGI – Information Management Consultants, Mandelring 238 b, 67433 Neustadt an der Weinstraße, E-Mail: manfred.hauer@agi-imc.de

Web-Server prägten die folgenden Jahre. Zusätzliche maschinelle Verfahren, zunächst das linguistisch-regelbasierte Verfahren PrimusIDX, fanden fast keinen Markt. 1999 gab es einen zweiten Anlauf mit Autindex, das seit 2001 als Bestandteil von intelligentCAPTURE bis heute trägt. Autindex ist ein regelbasiertes Verfahren, das mit linguistischen Wörterbüchern, Grammatikregeln, Heuristiken, semantischen Klassen und einem großen, aber flachen Thesaurus und einfacher Statistik Texte analysiert. intelligentCAPTURE zielt auf Bibliothek und Dokumentation und seine Workflows werden über den Lookup in Bibliothekssysteme gesteuert. Der zu analysierende Text muss additiv zu den Titeldaten der Kataloge erst noch hereingeholt werden. Es geht darum, Texte auf Papier zu digitalisieren oder aus PDF-Dateien der Verlage zu übernehmen. Schon beim Digitalisieren/Scannen greifen Algorithmen zur Bildoptimierung. Während die PDF-Extraktion ein komplexes regelbasiertes Verfahren nutzt, ist die Übertragung eines Images in brauchbaren Text, Optical Character Recognition (OCR), ein recht anspruchsvolles Verfahren, in dem wir mit Abbyy FineReader 12 seit Herbst 2019 erstmals neuronale Netz-Technologie als ein Modul im Rahmen der Zeichen- und Layout-Erkennung nutzen. Darin spielen „Wörterbücher“ eine nicht ersetzbare Rolle, um die 200 Sprachen – vieles jenseits des lateinischen Alphabets, das selbst schon über 250 Sonderzeichen kennt – von Bild, Grafik, Linien, Schmutz, abgeschnittenen Rändern, Stempeln, Kommentar oder Gebrauchsspuren zu unterscheiden. Da die 200 OCR-Sprachen nicht zur einen Indexierungssprache Deutsch passen, bauen wir mittels Google Translate für 100 Sprachen die Brücke. Translate in der aktuellen Version ist selbstlernende KI, die intern Englisch als Knoten verwendet: Chinesisch nach Deutsch geht über Englisch und umgekehrt. Der Transfer von Fachbegriffen über die Sprachen hinweg gelingt zuverlässig.

Das erste selbstlernende maschinelle Klassifikationsverfahren haben wir 2019 für die Identifikation von Autoren und Titeln trainiert, um die bibliografischen Angaben von Aufsätzen aus Inhaltsverzeichnissen von Zeitschriftenheften zu erkennen. Der Classifier war für 1,2 Millionen Trainingsdaten (nur Aufsätze) aus unserem dandelon.com in nur einer Minute trainiert. Bis wir diese Trainingsdaten in dieser Menge hatten, waren 17 Jahre mit viel menschlicher Arbeit bei der Erfassung, Software-Implementierung, Marketing, Vertrieb, Hosting und Support nötig.

Durch das Sharing-Konzept über unser dandelon.com und parallel in vier weiteren Projekten wurden Inhaltsverzeichnisse von Monografien (2,6 Mio.) und Aufsatzdaten (1,2 Mio.) generiert mit einer Zugriffsmenge von 200.000 pro Tag, verteilt über viele Bibliotheken und Bibliotheksverbünde.

Von den Bibliotheken wurden 2018 pro Monat 4500 neue Buchtitel kollaborativ in der offenen dandelon.COM-munity digitalisiert und über die Verbünde bis hin zu WorldCat verteilt. Hinzu kommen 1000 neue eBooks pro Monat, sowie kleinere Projekte – insgesamt ein Zuwachs von 346.000 Monografien und 34.000 Aufsätzen in nur einem Jahr. Mengen, welche die früheren Fachinformationszentren selbst in zehn Jahren nicht erreichen konnten. Die fremdsprachigen Titel aus etwa 30 Sprachen werden ganz ohne menschlichen Aufwand in deutscher und teils zusätzlich englischer Sprache indexiert – eine Testkollektion gleich in sieben europäischen bzw. zusätzlich drei asiatischen Sprachen (Hauer, 2017).

All dies ist für ein sehr kleines, sich stets selbst finanzierendes Unternehmen wie AGI-Information Management Consultants mit 2, 5 Mitarbeitern eine Menge, die ohne KI und „Collaboration“ nicht möglich wäre.

2 Arbeitsmarkt im Wandel

Zurück zur Statistik – nicht als Mathematik, sondern zum darin dokumentierten gesellschaftlichen Wandel. Der Agrarsektor (primärer Sektor), dem um 1800 noch ca. 80 Prozent der Bevölkerung angehörten, ist heute auf 0,5 Prozent in Deutschland geschrumpft und ernährt mehr Menschen als jemals zuvor. Dazu waren technische Methoden notwendig, primär der Ersatz tierischer und menschlicher Arbeitskraft durch Motoren und Getriebe auf Rädern – Traktoren. „Ungetüme“ am Anfang, heute zunehmend digital optimierte Alleskönner oder Spezialisten für Bodenbearbeitung, Pflanzenschnitt, Aussaat, Ernte und Transporte – teils mit GPS, Bordcomputer und Sensoren, vernetzt. Das Wissen und Können der Bauern, der Agrarwissenschaftler und die Anforderungen der Agrarmärkte wandern in die Traktoren. Züchtung und Chemie komplementieren den Wandel. Niemand hätte es vor 100 oder 200 Jahren für möglich gehalten, mit so wenigen Menschen so viel zu produzieren.

Der Sekundäre Sektor hat im Verlauf der letzten 50 Jahre rund 50 Prozent an Beschäftigten verloren, eine Auswirkung von Rationalisierung und Verlagerung. Er wird durch Industrie 4.0 mit dann wohl zehn Prozent der heutigen Beschäftigten das Produktionsvolumen mindestens halten – nicht erst in 50 Jahren. Ob diese restlichen zehn Prozent dann wieder in den alten Industrienationen arbeiten, wird sich zeigen – China hat mehr Roboter im Einsatz als jedes andere Land (Jaeger 2019).

Der Tertiäre Sektor, Dienstleistungen, nahm lange zu, er verzeichnete grob eine Verdoppelung über 70 Jahre. Vieles davon geht auf Auslagerungen aus Produktionsbetrie-

Abbildung 1: Entwicklung der Erwerbstätigen in Deutschland nach Wirtschaftssektoren in % aller Erwerbstätigen (Inlandskonzept).
Quelle: Statistisches Bundesamt 2018.

ben zurück, wie Logistik, Gebäude-Management, Personal-Management, Kantinen, Finanzen, Steuern, Recht, IT und Informationsdienstleistungen. Das Statistische Bundesamt nennt einen Schlüssel für zukünftige Entwicklung: „Es ist noch nicht abzusehen, in welchem Ausmaß und in welcher Weise sich der Strukturwandel in Deutschland künftig auf die Erwerbstätigkeit in den Wirtschaftsbereichen auswirken wird. Bereits jetzt ist anhand der Zeitreihen bei der Zahl der Erwerbstätigen jedoch erkennbar, dass sich das Tempo der Tertiarisierung in der jüngsten Vergangenheit verlangsamt hat. Zukünftige Szenarien von Tertiarisierung hängen dabei auch von Gestaltungsoptionen im Handlungsfeld Digitalisierung ab.“ (S 2).

Die Marginalitätsquote zeigt die Wahrscheinlichkeiten für den Verlust der Arbeit und ist im Tertiären Sektor am höchsten.

Abbildung 2: Entwicklung der Marginalitätsquote in %.
Quelle: Statistisches Bundesamt.

Die Arbeitsplatzverluste können durch demographische Entwicklungen teilweise abgemildert werden. In einer mittleren Projektionsvariante würde das Erwerbspersonen-

potenzial in Deutschland altersstrukturbedingt zwischen 2015 und 2030 um über eine Million Menschen und bis 2060 sogar um fast sieben Millionen Menschen sinken (Fuchs, 2017).

Gelingt die weitere Taylorisierung, die Aufspaltung in einzelne Arbeitsschritte einerseits und die Integration von KI-Techniken zu umfassenderen Lösungen andererseits, was wahrscheinlich ist, so wird es bei den fast 80 Prozent Beschäftigten im Dienstleistungsbereich nicht bleiben. Noch prognostizieren Wissenschaftler mit klassischen statistischen Methoden die Auswirkung der Digitalisierung auf den Arbeitsmarkt:

Drei Thesen:

1. Frey/Osborne – Oxford 2013: 47 Prozent der Tätigkeiten in den USA sind automatisierbar.
2. ZEW Mannheim 2015: 42 Prozent der Tätigkeiten haben hohe Automatisierungswahrscheinlichkeit, aber durch Synthese mit interaktiven Tätigkeiten vermutlich nur 12 Prozent hohe Automatisierungswahrscheinlichkeit.
3. IAB Nürnberg 2015 (Wolter u. a.): Bis 2025 Verschiebung von 920.000 Arbeitsplätzen: 490.000 Verlust und 450.000 neue Arbeitsplätze.

Das Statistische Bundesamt beschreibt Auswirkungen auf den Einzelnen:

- „> Beschleunigung von Innovationszyklen und höhere Relevanz von persönlicher Weiterbildung und Beschäftigungsfähigkeit (Implikationen für Politik, Unternehmen, Erwerbstätige)
- > Entgrenzung von Erwerbsarbeit und Privatsphäre durch mobiles Arbeiten und mobile Erreichbarkeit
- > höhere Relevanz von Solo-Selbstständigkeit im Zusammenhang mit der Direktvermarktung von Dienstleistungen

- > möglicherweise höhere Relevanz von Nebenjobs und geringfügiger Beschäftigung im Gigwork-Dienstleistungssegment“ (S 35)

Die sozialen und personenbezogenen Tätigkeiten, vom Kindergarten über Burn-Out-Therapie bis zur Demenz-Station, gelten bislang noch weniger geeignet für Substitution als Tätigkeiten in Finanzwirtschaft, Handel und Logistik, sowie Öffentlichem Dienst. (Enquete-Kommission 2002, S 59). Stimmt das für immer? Und wird nicht hoch interaktive Technik auch im sozialen Bereich Einzug halten?

3 Die Wandlungen eines Information Professionals

Der lange Leidensweg der DGD/DGI korreliert mit der Abschaffung, Reduktion oder Auslagerung zunächst der Informationsdienste in Konzernen ab etwa 2000 und zeitlich verzögert den staatlichen Fachinformationseinrichtungen. Er zeigt sich auch in unserer 35-jährigen wechselvollen Unternehmensgeschichte: Zuerst Recherchedienste für interne Abteilungen, daraus entstehen interne Informationsabteilungen, die wir mit Software für die Online-Abfrage ausstatten, weiter mit ersten eigenen Retrieval-Applikationen ab 1992, schließlich umfangreiche, komplexe Software-Entwicklung mit dem Label „Information Center“, einer Art „SAP“ für internationale Informationsabteilungen für viele Aufgaben. Seit 2002 ein markanter Rückgang, der Markt zerbröckelt. Wir lösen Technologien aus Information Center heraus und wechseln zu staatlichen Bibliotheken. Der Marktwechsel gelingt, doch nicht ohne weitere Tätigkeiten im bibliothekarischen Umfeld. Und dennoch: Unser intelligentCAPTURE wird meist von studentischen Hilfskräften bedient und liefert Retrieval-Leistungen, die es vorher nicht gab.

KI wirkt in einem Bündel von Technologien, Mathematik & Informatik, Datentransfer, Rechenleistung, Datenverfügbarkeit, Sensorik, Robotik plus andere Grundlagen- und Anwendungswissenschaften, katalytisch, so unsere Einschätzung. Heute zerfallen KI-Leistungen noch in viele Puzzle-Steine, die integriert werden müssen – und

übermorgen? Ob aus 80 Prozent Dienstleistungssektor heute 0,5 Prozent in 200 Jahren werden, wird sich zeigen.

Deskriptoren: Beruf, Berufsbild, Künstliche Intelligenz, Entwicklungstendenz, Indexierungsverfahren, Dienstleistung

Literatur

- Bericht der Enquete-Kommission S 59, Deutscher Bundestag – 14. Wahlperiode 2002 – Drucksache 14/ 8800. <http://dipbt.bundestag.de/dip21/btd/14/088/1408800.pdf>
- Arbeitsmarkt im Wandel – Wirtschaftsstrukturen, Erwerbsformen und Digitalisierung. Statistisches Bundesamt | WISTA | 3 | 2018. https://www.destatis.de/DE/Publikationen/WirtschaftStatistik/2018/03/ArbeitsmarktWandel_032018.pdf?__blob=publicationFile
- Hauer, Manfred: Indexing of ebooks with intelligentCAPTURE 9.0, 2017, Handout. <https://www.agi-imc.de/internetXR.nsf#/search/query=id:DEAGIB2D67B77DC696142C12581D400512C5A>
- Fuchs, Johann u. a.: Projektion des Erwerbspersonenpotenzials bis 2060 – Arbeitskräfteangebot sinkt auch bei hoher Zuwanderung 2017 <http://doku.iab.de/kurzber/2017/kb0617.pdf>
- Jaeger, Lars: Mit Künstlicher Intelligent an die Weltspitze. Die Beispiele China und Deutschland. Open Password Pushdienst, 18. Jan. 2019. https://www.password-online.de/?wysija-page=1&controller=email&action=view&email_id=606&wysijap=subscriptions

Dipl.-Inf.wiss. Manfred Hauer, M. A.

AGI – Information Management Consultants
Mandelring 238 b
67433 Neustadt an der Weinstraße
<https://orcid.org/0000-0003-2549-1478>
GND ID: 1025266161
manfred.hauer@agi-imc.de

Manfred Hauer ist Inhaber der AGI-Information Management Consultants. Er kam über die Soziologie zur Informationswissenschaft und gründete 1983 sein Unternehmen. Dieses hat sich von der Informationsvermittlung über den Software-Vertrieb zum Anbieter umfassender Knowledge-Management-Lösungen (Information Center) entwickelt, oft mit mehreren tausend Nutzern in globalen Unternehmen.