Sozial Extra 5 2020: 285–290 https://doi.org/10.1007/s12054-020-00317-z Online publiziert: 1. September 2020

Durchblick: Data Mining

Data-Mining als Werkzeug empirischer Sozialforschung

Computergestützte Auswertung und Visualisierung großer Datenmengen aus dem Internet am Beispiel von Rezensionen sozialwissenschaftlicher Publikationen

Das Internet enthält eine stetig wachsende Menge an Daten, die auch von sozialarbeiterisch engagierten Menschen produziert und konsumiert, jedoch eher selten ausgewertet werden, was auch damit zu tun hat, dass deren Umfang die Möglichkeiten traditioneller Untersuchungsmethoden übersteigt. Die computergestützte Auswertung großer Datenmengen, sogenanntes „Data Mining", die in anderen Disziplinen bereits deutlich höhere Aufmerksamkeit erfährt, ist jedoch auch für Sozialwissenschaften relevant.

Im Jahr 2016, auf dem 36. Chaos Communication Congress – dem jährlich stattfindenden Kongress des Chaos Computer Clubs – hielt David Kriesel einen Vortrag mit dem Titel „SpiegelMining" (Kriesel 2016), in dem er von seinem Versuch berichtet, mittels „Data Mining" Erkenntnisse über den Aufbau und die inneren Abläufe von Spiegel-Online zu gewinnen. „Data Mining" bezeichnet den Prozess der (teil-) automatischen Extraktion von Wissen aus großen Datenbeständen, unter anderem durch systematische Anwendung statistischer Methoden. Ziel dieses Ansatzes ist es, vorher unbekannte Zusammenhänge, Muster und Trends ausfindig zu machen.

Konstantin Kirchheim

Otto-von-Guericke-Universität Magdeburg, Magdeburg, Deutschland

*1994, Master of Science (Informatik), seit 2020 wissenschaftlicher Mitarbeiter an der Otto-von-Guericke-Universität, Lehrstuhl für Software-Engineering.

konstantin.kirchheim@ovgu.de

Zusammenfassung Der Beitrag zeigt an einem Beispiel, wie Methoden der Informatik genutzt werden können, um die riesigen, teils ungenutzten Datenmengen, die im Internet vorhanden sind und sich aufgrund ihres Volumens einer Untersuchung durch traditionelle Methoden entziehen, ausgewertet werden können. Anhand eines Beispieldatensatzes von ca. 18.000 Rezensionen zu wissenschaftlichen Publikationen aus dem Bereich der Sozialen Arbeit wird gezeigt, wie entsprechende Daten gesammelt, verarbeitet, aggregiert und visualisiert werden können, um sie einer weiterführenden Analyse zugänglich zu machen.

Schlüsselwörter Data-Mining, Soziale Arbeit, Szientometrie, Digital Humanities

Inspiriert durch den Vortrag versuchen wir (der Informatiker Konstantin Kirchheim und der Student der Sozialen Arbeit, Tilman Kloss) die Methodik in einem interdisziplinären Projekt für Fragestellungen der Sozialarbeitsforschung nutzbar zu machen, indem wir eine bekannte Webseite, auf der Rezensionen zu wissenschaftlichen Publikationen aus dem Umfeld der Sozialen Arbeit veröffentlicht werden, untersuchen. Dazu haben wir ca. 18.000 solcher Rezensionen aus dem Zeitraum von 2001 bis 2019 abgerufen und ausgewertet. Das Ziel sollte gleichsam die Identifikation von Kernbereichen und blinden Flecken des wissenschaftlichen Diskurses der Sozialen Arbeit sowie eine Analyse von deren Verschiebung im Laufe der Zeit sein.

Der Umfang dieses Beitrags erlaubt es lediglich, einen kleinen Ausschnitt der gewonnenen Daten und Visualisierungen zu erläutern. Auf der diesen Artikel begleitenden Webseite www.extra-mining.de stehen jedoch zusätzliche Statistiken und hochauflösende Abbildungen zur Verfügung.

Erstellen einer Datenbasis

Zunächst müssen die (online) verfügbaren Daten zu einer Datenbasis zusammengetragen werden. Eine umfangreiche Einführung in dieses „Scraping" genannte Sammeln (und Auswerten) von Daten von Webseiten kann (Mitchell 2015) entnommen werden, im folgenden können lediglich die Grundlagen beleuchten werden. Stark vereinfach gesagt besteht eine Webseite aus einer oder mehreren Textdateien. Wenn ein Web-Browser eine Seite aufruft, fordert er den Server, der die Webseite bereitstellt, auf, ihm eine Kopie dieser Textdatei zuzusenden. Das Format dieses Textes nennt man HTML (Hypertext Markup Language), und es beschreibt dem Web-Browser, wie die Seite dargestellt werden soll. Die

Datenbasis besteht also aus den heruntergeladenen HTML-Dateien.

Bereinigung der Daten

Die Grundlage der Datenbasis bilden 20.281 abgerufene Rezensionen. Bei deren Auswertung zeigte sich, dass 714 nicht richtig verarbeitet werden konnten, und somit keine verwertbaren Ergebnisse liefern. Von den verbleibenden Rezensionen wurden nur die berücksichtigt, die im Zeitraum vom 01.01.2001 bis 31.12.2019 erschienen sind. Die Anzahl der Veröffentlichungen verschiedener Rezensenten variiert zum Teil stark. Es gibt sowohl Publizierende, die in einzelnen Themengebieten besonders aktiv sind, und dort einen substanziellen Teil der Rezensionen verfasst haben, als auch Autor_innen, die einen überdurchschnittlich hohen Anteil der insgesamt abgerufenen Rezensionen verfasst haben. Dies birgt die Gefahr, dass Meinungen, Präferenzen und Vorurteile einzelner die Ergebnisse der Auswertung verzerren. Da das Ziel eine möglichst ganzheitliche, von individuellen Personen unabhängige Betrachtung ist, werden die Veröffentlichungen des Rezensenten mit der höchsten Publikationsanzahl als statistische Ausreißer behandelt, und in der folgenden Auswertung nicht berücksichtigt. Trotz dieser Maßnahmen ist darauf hinzuweisen, dass für eine repräsentativere Untersuchungen Daten aus verschiedenen Quellen zusammengetragen werden sollten.

Extraktion von Merkmalen

Aus den verbleibenden 18.010 Rezensionen werden bestimmte Merkmale – genannt „Features“ – extrahiert. Zu den Features zählen zum einen Informationen über die Rezension selbst, wie z. B. deren Text, das Erscheinungsdatum, oder der Name des/der Rezensent_in. Zum anderen aber auch Informationen, die sich auf die rezensierte Publikation beziehen, wie z. B. deren Seitenanzahl, Erscheinungsdatum, Verlag, Preis oder ISBN. Um mit den extrahierten Features arbeiten zu können, werden diese üblicherweise in eine tabellarische Struktur einsortiert, bei der jede Zeile eine Rezension, und jede Spalte ein zugehöriges Merkmal enthält. Beispielsweise nennt jede Rezension den Verlag, bei dem die betrachtete Publikation erschienen ist. Nach der Feature-Extraktion enthält die Tabelle also eine Spalte, die den Verlag angibt. Dies erlaubt es z. B. zu ermitteln, wie häufig bestimmte Verlage vertreten sind, indem deren Auftreten in der entsprechenden Spalte gezählt wird.

Implizite Features

Jede Rezension enthält einen kurzen Abschnitt, in dem der/die Autor_in kurz vorgestellt wird, und der entweder mit „Rezensent“ oder „Rezensentin“ (und seit einiger Zeit mit „Rezension von“) betitelt ist. Implizit wird hier also das Geschlecht des/der Verfasser_in angegeben. Die untersuchte Webseite hat diese Bezeichnung innerhalb des Zeitraums, in dem die Daten gesammelt wurden (auch rückwirkend), zugunsten der generischen Formulierung angepasst, die keine Rückschlüsse mehr auf das Geschlecht der Verfasser_innen zulässt. Da die überwiegende Mehrheit der Rezensionen jedoch vor dieser Umstellung abgerufen wurde, steht dieses Feature für den Großteil des Untersuchungszeitraumes zur Verfügung.

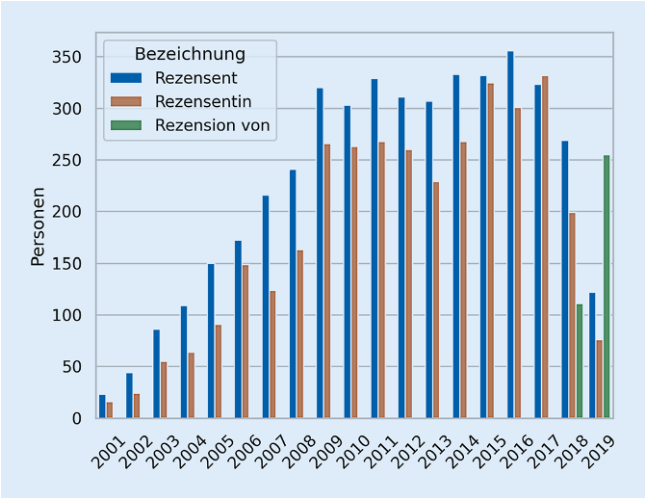

Abb. 1 Anzahl aktiver Rezensent_innen, kategorisiert nach Geschlecht

Durch Auszählen der entsprechenden Tabellenspalte ergibt sich, dass 62,9 % der Rezensionen von „Rezensenten“, 33,8 % von „Rezensentinnen“ und 3,1 % von Personen ohne spezifische Angaben verfasst wurden.

Kombination von Features

Durch die Kombination mehrerer Features können komplexere Zusammenhänge erschlossen werden. Die Verbindung des Erscheinungsdatums jeder Rezension und der Bezeichnung des/der Verfasser_in erlaubt es, die Anzahl der aktiven „Rezensenten“ und „Rezensentinnen“ über den Untersuchungszeitraum auszuwerten. Die entsprechende Einteilung ist in Abb. 1 dargestellt. Hierbei ist zu beachten, dass die Anzahl der Personen – nicht die der Rezensionen – für ein Jahr betrachtet wird. Eine Person, die in einem Jahr mehrere Rezensionen verfasst, hat wird also lediglich einmal aufgeführt.

Unscharfe Features

Einige Artikelmerkmale lassen sich weniger einfach auslesen als das Geschlecht der Autor_innen. Zum Beispiel geben auf der untersuchten Seite viele Personen zusätzlich zu ihrem Namen einen akademischen Grad,

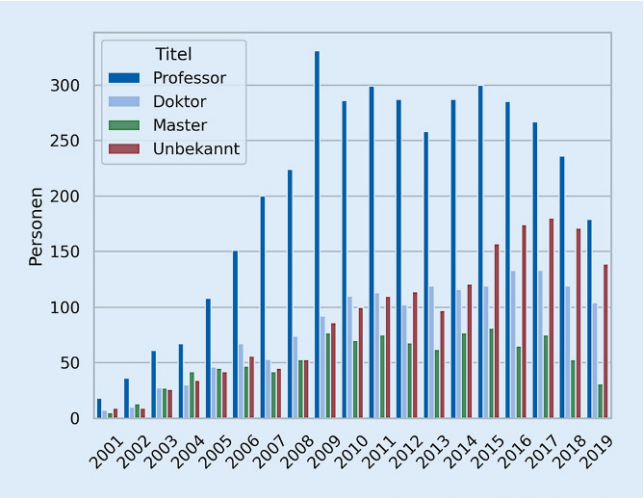

Abb. 2 Anzahl aktiver Rezensent_innen, kategorisieret nach akademischem Abschluss

z. B. „Diplom“, „Doktor“ oder „Professor“ an. Analog zu der geschlechtsbasierten Auswertung können die Verfasser_innen anhand ihres höchsten angegebenen akademischen Abschlusses in Kategorien eingeteilt werden. Bei diesem Feature ist eine eindeutige Kategorisierung jedoch schwieriger, da es einerseits verschiedene Möglichkeiten gibt, einen solchen Titel zu schreiben (z. B. „Prof.“ oder „Professor“) und es andererseits verschiedenartige, aber prinzipiell gleichwertige Titel gibt – beispielsweise verschiedene Varianten des Diploms. In solchen Situationen ist es gegebenenfalls notwendig, die Daten manuell zu bereinigen, oder sich damit zufrieden geben, dass die automatisierte Methode nur „hinreichend gut“ funktionieren. Man könnte diese Unschärfe als den Preis betrachten, den man dafür zahlt, die Daten überhaupt auswerten zu können. Allgemein muss zudem beachtet werden, dass sich online verfügbare Informationen – anders als in Printmedien – prinzipiell jederzeit ändern können. So ist im Falle der hinterlegte akademischen Titel (und prinzipiell auch der Geschlechter) darauf hinzuweisen, dass diese bei einer Aktualisierung auch rückwirkend in allen vorherigen Publikationen des/der Autor_in angepasst werden, so dass sich die Angabe nicht auf die zum Zeitpunkt der Veröffentlichung erworbenen, sondern die aktuellsten Titel bezieht.

Auch für den akademischen Grad lässt sich – durch Kombination mit dem Erscheinungsdatum der Rezension – die Anzahl der jährlich aktiven Rezensent_innen in den unterschiedlichen Kategorien ermitteln. Die entsprechende Aufschlüsselung finden sich in Abb. 2. Verfasser mit den Abschlüssen „Master“, „Magister“ oder verschiedenen Varianten des „Diploms“ wurden in der Kategorie „Master“ zusammengefasst.

„Versteckte“ Features

Neben den „offensichtlichen“ Merkmalen können die Rohdaten auch weitere, zunächst verborgene Informationen enthalten. So sind zu jeder Rezension Stichwörter der Deutschen Nationalbibliothek hinterlegt, die die rezensierten Publikationen bestimmten Kategorien zuordnen. Diese Stichwörter sind – da sie nicht durch den Web-Browser dargestellt werden – für einen menschlichen Betrachter nicht unmittelbar ersichtlich. Suchmaschinen können diese jedoch nutzen, um den Inhalt einer Internetseite einzuordnen, und so Nutzern passende Suchergebnisse zu liefern.

Statistiken zu den meistgenannten Stichwörtern – der Anteil der Nennungen durch „Rezensentinnen“ so wie der Anteil der Nennungen durch Verfasser_innen mit mindestens einem Professorentitel – sind in Tab. 1 aufgeführt. Man kann feststellen, dass die Verhältnisse der Geschlechter oder der Bildungstitel zwischen den Stichwörtern zum Teil erheblich voneinander abweichen.

Identifizieren von Themenschwerpunkten

Die hinterlegten Stichwörter erlauben es, die rezensierten Publikationen hinsichtlich der behandelten Themen zu untersuchen. Während ein einzelnes Stichwort eher keinen eigenständigen Themenkomplex bildet, ergeben sich thematisch zusammengehörige Bereiche aus den Beziehungen der Stichwörter untereinander.

Themen-Landkarten

Eine Methode, das Beziehungsgeflecht der Stichwörter zu visualisieren, sind die „Themen-Landkarten“, die Kriesel in seinem eingangs erwähnten Vortrag ebenfalls benutzt. Die grundlegende Idee ist, die strukturelle Nähe in den Daten in visuelle Nähe auf einer digitalen Landkarte zu überführen. Der Vorteil dieser Methode besteht darin, dass sie in der Lage ist, komplexe, multidimensionale Daten in einer auch für Laien intuitiv zugänglichen Form abzubilden und „erkundbar“ zu machen.

Stichwort-Netzwerke

Zur Konstruktion solcher Karten kann wie folgt vorgegangen werden: abstrakt lässt sich jedes Stichwort als Knoten in einem Netzwerk betrachten. Die Kanten des Netzwerks – die die Knoten verbinden – besitzen einen „Gewicht“ genannten Zahlenwert, der eine Beziehung zwischen den jeweiligen Knoten, zum Beispiel deren Ähnlichkeit, beschreibt.

Im Folgenden wird die Anzahl der gemeinsamen Nennungen jeweils zweier Stichwörter als Kantengewicht verwendet.

Die Verbindung zwischen den Knoten „Kind“ und „Sozialarbeit“ repräsentiert also die Anzahl des gleich-

Tab. 1 Meistgenannte Stichwörter

Stichwort	Nennungen	Rezensentinnen (%)	Professor (%)
Kind	878	47,11	43,16
Sozialarbeit	739	36,45	66,98
Jugend	574	35,57	46,86
Kindertagesstätte	436	58,55	35,32
Alter	340	44,84	62,64
Prävention	291	45,61	50,51
Psychotherapie	273	44,18	36,26
Eltern	231	50,00	35,93
Krankenhaus	229	21,83	71,11
Altenpflege	217	47,90	38,71

Tab. 2 Beispiel einer Stichwort-Matrix

	Kind	Sozialarbeit	Jugend	Alter
Kind	0	17	216	0
Sozialarbeit	17	0	11	3
Jugend	216	11	0	0
Alter	0	3	0	0

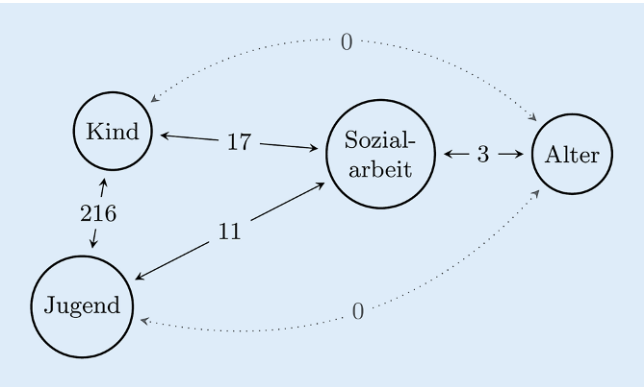

Abb. 3 Graph aus Tab. 2 nach der Anwendung von Forceatlas 2

zeitigen Auftretens beider Stichwörter in allen Rezensionen.

In Tab. 2 sind beispielhaft die gemeinsamen Nennungen der Stichwörter „Kind“, „Sozialarbeit“, „Jugend“ und „Alter“ aufgeführt. Da z. B. das Wort „Kind“ genauso häufig mit dem Wort „Jugend“ genannt wird wie das Wort „Jugend“ mit dem Wort „Kind“ ist diese Tabelle (oder Adjazenzmatrix) symmetrisch.

Netzwerkanalyse

Solche Netzwerke lassen sich mit Netzwerkanalyse-Tools verarbeiten. Ein Beispiel für ein solches Tool ist „Gephi“ (Bastian et al. 2009). Eine Funktion von Gephi erlaubt es, durch einen Algorithmus namens „Forceatlas 2“ (vgl. Jacomy et al. 2014) die Knoten eines Graphen unter Berücksichtigung der Kanten anzuordnen oder zu „layouten“. Dazu simuliert Gephi ein physikalisches System, bei dem sich Knoten, wie Partikel mit gleicher Ladung, gegenseitig abstoßen, während Kanten diese, wie Federn, zusammenziehen. Dabei ist die Stärke der Anziehung proportional zu dem Gewicht der Kante, in diesem Fall also der Anzahl der gemeinsamen Nennungen. Durch die anziehenden und abstoßenden Kräfte entsteht Bewegung, die die Knoten so verschiebt, dass sich häufig gemeinsam genannte Stichwörter aufeinander zu, selten gemeinsam genannte Stichwörter voneinander weg bewegen. Nach einer gewissen Zeit konvergiert das so simulierte physikalische System in einen stabilen Zustand. Wenn wir aus den Beispieldaten in Tab. 2 ein Netzwerk konstruieren, und auf dieses Forceatlas 2 anwenden, könnte sich ähnliche Struktur ergeben wie in Abb. 3.

Einer der Vorteile dieses iterativen Layouting-Ansatzes ist, dass man diesen in Echtzeit verfolgen kann. Es entsteht der Eindruck, dass sich das Netzwerk in einem kontinuierlichen Prozess aufspannt und ausrichtet. Somit bietet das Verfahren einen intuitive(re)n Einblick in den zugrundeliegenden Algorithmus, da der Prozess zu jedem Zeitpunkt angehalten, verlangsamt, oder das Netzwerk manipuliert werden kann. Einer der Nachteile von Forceatlas 2 ist dessen inhärenter Indeterminismus: Da der stabile Endzustand von dem Ausgangszustand abhängt, erhält man bei mehrfacher Ausführung mit hoher Wahrscheinlichkeit unterschiedliche Ergebnisse (vgl. Jacomy et al. 2014).

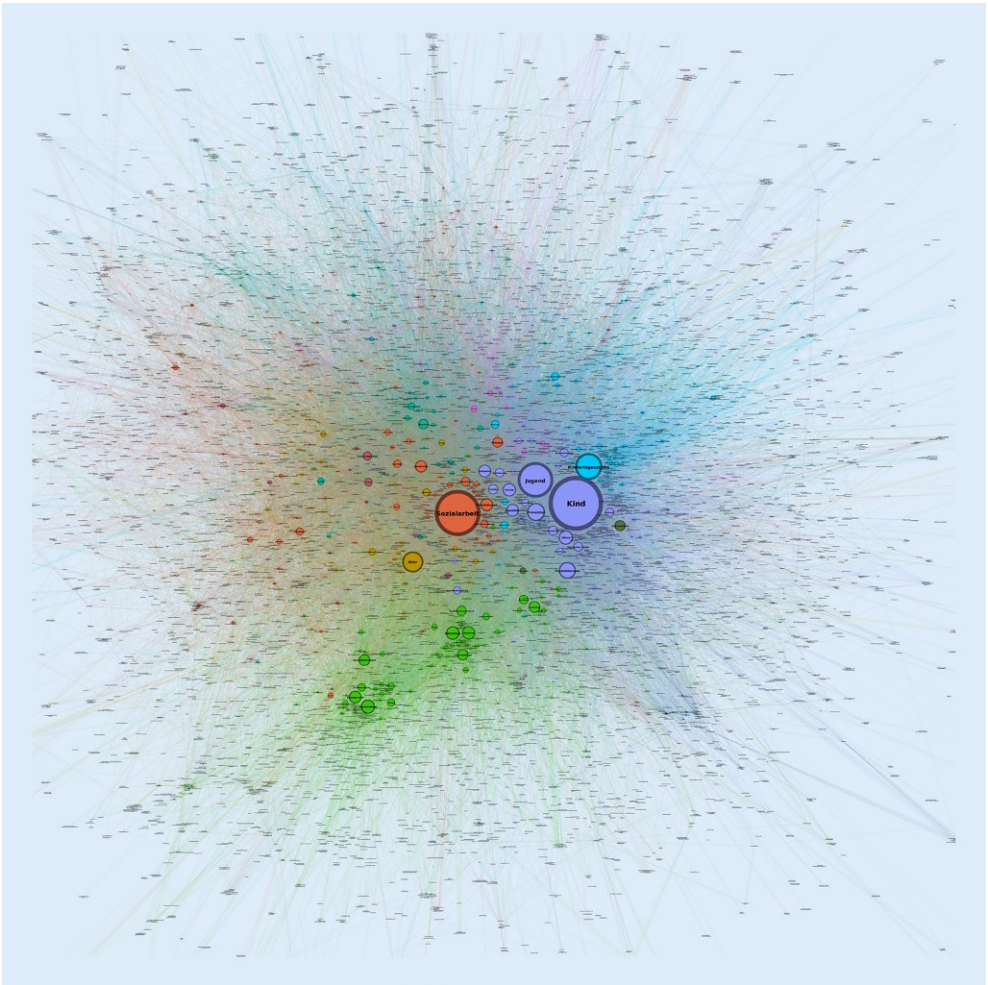

Abb. 4 Gesamter Stichwort-Graph

Themengebiete

Aus den Stichwörtern entsteht ein Netzwerk mit 7463 Knoten und 58.440 Kanten. An dieser Stelle sollte deutlich werden, dass traditionelle Methoden nicht in der Lage sind, entsprechende Datenmengen mit vertretbarem Aufwand zu bewältigen. In dem Netzwerk lassen sich nun größere Themengebiete identifizieren. Zum einen entstehen durch Forceatlas 2 in der Anordnung der Stichwörter zusammenhängende Gruppen, die sich bereits visuell als Themengebiete interpretieren lassen. Es existieren jedoch zahlreiche weitere Verfahren, die z. B. nach Gruppen von stark miteinander verbundenen Knoten suchen. Durch die „Louvain-Methode“ (Blondel et al. 2008) lassen sich ca. 300 verschiedene Themengebiete identifizieren und farbig hervorheben, wie in Abb. 4 dargestellt.

Zusammenfassung und Ausblick

Bisher wurde die ausgewertete Webseite in Bezug auf Geschlecht und Bildungsgrad ihrer Autorenschaft untersucht sowie bestimmte Themenbereiche identifiziert. Weder der Gehalt der Daten, noch die Möglichkeiten der vorgestellten Methode sind an dieser Stelle ausgeschöpft. Indem z. B. der Graph mit zusätzlichen Features anreichert wird, kann dessen Informationsgehalt weiter erhöht werden. So kann die Größe und Farbe der Knoten Eigenschaften des zugehörigen Stichwortes widerspiegeln, beispielsweise die Summe seiner Nennungen oder den Anteil der Nennungen durch männliche oder weibliche Rezensent_innen. Hochauflösende Beispielgrafiken können der genannten Webseite entnommen werden.

Neben den hier erwähnten Verfahren existieren zahlreiche weitere, die teilweise mächtiger, aber auch komplexer und somit für Laien schwieriger zu interpretieren sind, wie beispielsweise von Eckl et al. (2019, 2020) verwendet. Zudem wurden in diesem Artikel lediglich Meta-Daten der untersuchten Rezensionen betrachtet – es sei darauf verwiesen, dass auch der Inhalt der Rezensionen statistisch untersucht werden kann.

Die aggregierten Daten und Darstellungen können als Ausgangspunkt und zur Unterstützung weiterführender Untersuchungen dienen. Wenn beispielsweise die Geschlechterverhältnisse zwischen den identifizierten

Themengebieten signifikant schwanken, wirft dies die Frage nach der Ursache dieses Phänomens auf, deren Beantwortung auch für die Soziale Arbeit relevant ist. Eine exemplarische Analyse und Interpretation der hier vorgestellten Daten finden sich im Beitrag von Tilman Kloss in diesem Schwerpunkt.

Eingegangen. 2. April 2020

Angenommen. 23. Juli 2020

Funding. Open Access funding provided by Projekt DEAL.

Literatur

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: an open source software for exploring and manipulating networks*. Third international AAAI conference on weblogs and social media, San Jose. https://doi.org/10.13140/2.1.1341.1520.

Blondel, V., Guillaume, J.-L., Lambiotte, R., & Etienne, L. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Eckl, M., Prigge, J., Schildknecht, L., & Ghanem, C. (2020). *Zehn Jahre Soziale Passagen: Eine empirische Analyse ihrer Themen*. Berlin: Springer.

Eckl, M., Ghanem, C., Löwenstein, H., & Spensberger, F. (2019). Die Entwicklung der Sozialen Arbeit von separaten Gruppen hin zu einer Scientific Community: Eine Soziale Netzwerkanalyse. *Neue Praxis. Zeitschrift für Sozialarbeit, Sozialpädagogik und Sozialpolitik 5(19)*, 467–484.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. Public Library of Science. *Plos One*. https://doi.org/10.1371/journal.pone.0098679.

Kriesel, D. (2016). Spiegelmining—reverse engineering von Spiegel-Online. Chaos communication congress. Youtube. https://www.youtube.com/watch?v=-YpwsdRKt8Q. Zugegriffen: 02. April 2020.

Mitchell, R. (2015). *Web scraping with Python. Collecting more data from the modern web*. : O'Reilly Media, Inc.