ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Barton, Thomas; Peuker, Andreas

Article — Published Version

Extraktion und Analyse von Schlüsselwörtern für eine automatisierte Literaturauswertung zum Thema Empfehlungssysteme

HMD Praxis der Wirtschaftsinformatik

Provided in Cooperation with:
Springer Nature

Suggested Citation: Barton, Thomas; Peuker, Andreas (2022) : Extraktion und Analyse von Schlüsselwörtern für eine automatisierte Literaturauswertung zum Thema Empfehlungssysteme, HMD Praxis der Wirtschaftsinformatik, ISSN 2198-2775, Springer Fachmedien Wiesbaden GmbH, Wiesbaden, Vol. 60, Iss. 6, pp. 1312-1327, https://doi.org/10.1365/s40702-022-00909-1

This Version is available at:
https://hdl.handle.net/10419/307631

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

HMD (2023) 60:1312–1327
https://doi.org/10.1365/s40702-022-00909-1

FORSCHUNG FÜR DIE PRAXIS

Extraktion und Analyse von Schlüsselwörtern für eine automatisierte Literaturauswertung zum Thema Empfehlungssysteme

Thomas Barton · Andreas Peuker

Eingegangen: 4. März 2022 / Angenommen: 29. August 2022 / Online publiziert: 22. September 2022
© Der/die Autor(en) 2022

Zusammenfassung Mit der zunehmenden Anzahl an wissenschaftlichen Publikationen steigt die Komplexität zur Durchführung einer Literaturauswertung. Insbesondere die Analyse einer Vielzahl an wissenschaftlichen Publikationen ist mit manuellen Tätigkeiten verbunden, die in der Regel nur sehr zeitaufwendig umzusetzen sind. Um diesem Aufwand entgegenzuwirken, existieren unterschiedliche Methoden der deskriptiven Berechnung und des maschinellen Lernens, die zur Unterstützung einer wissenschaftlichen Literaturauswertung eingesetzt werden können. In diesem Zusammenhang kann Keyword Extraction genutzt werden, um Schlüsselwörter von Texten automatisiert zu erkennen. In diesem Beitrag wird vorgestellt, wie Keyword Extraction zur Unterstützung einer wissenschaftlichen Literaturauswertung zum Thema „Empfehlungssysteme“ eingesetzt werden kann.

Schlüsselwörter Automatisierte Literaturauswertung · Empfehlungssysteme · Keyword Extraction · YAKE! · TF-IDF

Thomas Barton (✉)
Hochschule Worms, Erenburgerstraße 19, 67549 Worms, Deutschland
E-Mail: barton@hs-worms.de

Andreas Peuker
HORNBACH Baumarkt AG, Hornbachstraße 11, 76879 Bornheim, Deutschland

Extraction and analysis of keywords for an automated literature review on the topic of recommender systems

Abstract With the increasing number of scientific publications, the complexity of conducting a literature review also increases. In particular, the evaluation of a large number of articles is associated with manual activities that are usually very time-consuming to realize. To counteract this effort, methods for keyword extraction can be used to support a scientific literature research. Keyword extraction can be used to automatically find relevant terms for a corpus. This paper proposes how keyword extraction can be used to support a scientific literature review on the topic of "recommender systems".

Keywords Automatic Literature Review · Recommender Systems · Keyword Extraction · YAKE! · TF-IDF

1 Einleitung

Die Auswertung von Literaturrecherchen ist ein sehr zeitaufwendiger Prozess, insbesondere zu Themen, die von hoher Relevanz und Aktualität sind (vom Brocke et al. 2009; Booth et al. 2016). Das Interesse eines Kunden an einem Objekt mit Hilfe von Empfehlungssystemen (engl. „recommender systems") zu prognostizieren, stellt ein solches Thema dar. Es begegnet uns täglich in Form von personalisierten Empfehlungen nicht nur bei der Konsumierung von Filmen über Streaming-Dienste oder bei der Durchführung von Online-Einkäufen (Cheng et al. 2016; Gupta et al. 2020). Für ein aktuelles und relevantes Forschungsgebiet lassen sich im Rahmen einer Literaturrecherche eine Vielzahl wissenschaftlicher Publikationen identifizieren, insbesondere wenn der Rechercheprozess auf einem konkreten Vorgehensmodell basiert. Für eine Literaturrecherche auf dem Gebiet der Wirtschaftsinformatik hat sich das Vorgehen nach vom Brocke et al. etabliert (vom Brocke et al. 2009) etabliert. Eine manuelle Analyse ist allerdings nur mit hohem Zeitaufwand durchführbar. Hier bietet sich der Einsatz von Methoden der einfachen deskriptiven Berechnung oder des Machine Learning an, um den Prozess der Literaturanalyse und -synthese zu unterstützen und ggfs. zu automatisieren (Tauchert et al. 2020). Ein wichtiger Schritt stellt hierbei die Beschreibung des Inhalts von wissenschaftlichen Publikationen mit Hilfe von Schlüsselwörtern dar. Dazu stehen verschiedene Verfahren aus dem überwachten und unüberwachten Lernen zur Verfügung (Rose et al. 2010; Siddiqi und Sharan 2015; Campos et al. 2018a, b; Sun et al. 2020). In wieweit Verfahren zur Extraktion von Schlüsselwörtern geeignet sind, um systematische Literaturauswertung zu unterstützen, erfolgt im Rahmen einer Literaturanalyse zu Empfehlungssystemen auf Basis der Forschungsplattform ACM (Association of Computing Machinery) Digital Library (ACM 2022). Dieser Beitrag soll aufzeigen, dass automatisch extrahierte Schlüsselwörter im Rahmen einer Literaturanalyse sehr hilfreich sein können, um die Inhalte von Publikationen zu beschreiben. Dazu wird eine Literaturrecherche zu Empfehlungssystemen herangezogen, die mit einer sehr umfangreichen Trefferliste einhergeht. Zur Beschreibung des Inhaltes werden alle Abstracts der selektierten

Publikationen analysiert. Der Aufbau des Beitrages ist wie folgt: Zunächst werden die Grundlagen dargelegt, die den Prozess der Literaturanalyse und der Extraktion von Schlüsselwörtern beschreiben. Im Anschluss daran wird in das Thema Literaturrecherche und die Methodik zur Extraktion von Schlüsselwörtern vorgestellt. Schließlich werden die erzielten Ergebnisse vorgestellt und diskutiert. Der Beitrag schließt mit einer Zusammenfassung und einem Fazit ab.

2 Grundlagen des Prozesses zur Literaturanalyse und Grundlagen der Key-Word-Extraktion

Auf dem Gebiet der Wirtschaftsinformatik existiert ein weit verbreitetes Prozessmodell, das die Durchführung einer wissenschaftlichen Literaturauswertung beschreibt (vom Brocke et al. 2009). Der Prozess umfasst fünf Phasen, die Abb. 1 zu entnehmen sind.

Die Durchführung ist dabei von manuellen Aktivitäten begleitet, die sich oftmals nur sehr zeitaufwendig umsetzen lassen. Insbesondere die Phase zur Literaturanalyse und -synthese ist ein sehr zeitaufwändiger Prozess, der in der Regel mit einer intensiven Auseinandersetzung von Forschenden mit den Inhalten von Publikationen einhergeht.

Um diese umfangreichen und arbeitsintensiven Tätigkeiten zu unterstützen, schlagen Tauchert et al. (2020) einen Prozess vor, um eine wissenschaftliche Literaturauswertung mit Hilfe von Machine Learning zu unterstützen. Dieser Prozess geht mit einer Abfolge von einzelnen Schritten einher. Er ist in Abb. 2 dargestellt.

Ein Prozessschritt dient dazu, den Inhalt von Publikationen zu beschreiben. Er ist mit Keyword Extraction bezeichnet. Mit seiner Hilfe kann die inhaltliche Analyse einer Vielzahl von Publikationen vereinfacht werden. Eine systematische Literaturrecherche zur Automatisierung von systematischen Literaturauswertungen ist der Gegenstand der Publikation von Van Dinter et al. (2021). Schlüsselwörter sollen den Inhalt von Texten in kurzer und präziser Form zusammenfassen (Feather und Sturges 2003). Die Schlüsselwörter können dabei entweder manuell oder automatisiert zugewiesen werden. Eine manuelle Durchführung erweist sich als sehr zeitaufwendig, sobald eine Vielzahl von Texten ausgewertet werden. Für eine automatisierte

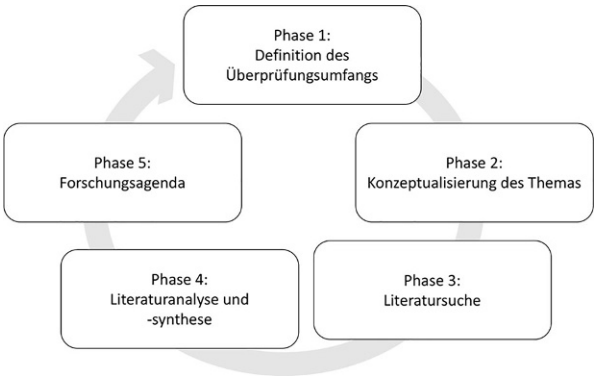

Abb. 1 Prozess für eine wissenschaftliche Literaturauswertung (eigene Abbildung in Anlehnung an vom Brocke et al. (2009))

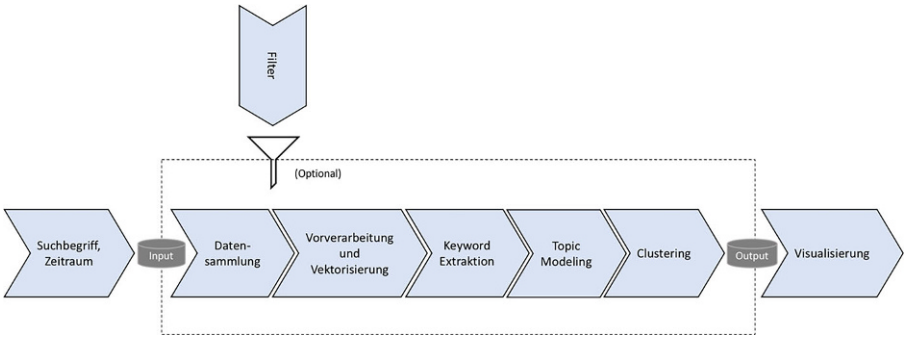

Abb. 2 Prozess für eine automatisierte Literaturauswertung (eigene Abbildung in Anlehnung an Tauchert et al. [2020])

Ermittlung ist hingegen eine Methode notwendig, die relevante Schlüsselwörter aus einem Text extrahiert. Solche Methoden lassen sich einem der folgenden Ansätze zuordnen (Siddiqi und Sharan 2015): Unter *statistische Ansätze* fallen diejenigen Methoden, die Schlüsselwörter basierend auf den statistischen Merkmalen eines Wortes extrahieren. Ein Merkmal ist beispielsweise durch die Wortfrequenz oder die inverse Dokumentenhäufigkeit gegeben. Methoden im Bereich des *überwachten Ansatzes* nutzen Modelle, die auf Basis von markierten Daten trainiert werden. Ein markierter Datenpunkt besteht dabei aus einem Text sowie den dazugehörigen Schlüsselwörtern. Nach dem Training versucht das Modell die Schlüsselwörter für einen neuen Text zu prognostizieren. Auch Methoden basierend auf dem *unüberwachten Ansatz* nutzen Modelle, um Schlüsselwörter aus Texten zu extrahieren. Der Unterschied zu überwachten Methoden besteht jedoch darin, dass das Training der Modelle keine markierten Daten benötigt. *Linguistische Methoden* nutzen die sprachlichen Merkmale (bspw. Semantik oder Syntaktik) eines Wortes, um Schlüsselwörter zu extrahieren. Den *hybriden Ansatz* verfolgen diejenigen Methoden, die zwei oder mehrere der genannten Methoden kombinieren. Es existieren zahlreiche Verfahren zur Extraktion von Schlüsselwörtern, wobei die „State-of-the-Art" Methoden entweder dem statistischen oder dem unüberwachten Ansatz zuzuordnen sind. In diesem Zusammenhang sind die folgenden Methoden zu nennen: Die Methode *TextRank* (Mihalcea und Tarau 2004) folgt dem unüberwachten Ansatz und extrahiert Schlüsselwörter basierend auf dem PageRank Algorithmus (Brin und Page 1998). Entworfen für die Erstellung einer Rangfolge von Webseiten in Online-Suchergebnissen erstellt der PageRank Algorithmus einen Graphen bestehend aus Knoten (Webseiten) und Kanten (Verlinkungen zwischen den Webseiten). Basierend auf der Verlinkungsstruktur der Webseiten wird ein Wert ermittelt, der die Wahrscheinlichkeit des Besuchs einer Webseite spiegelt. Dieses Vorgehen nutzt TextRank um Schlüsselwörter aus Texten zu extrahieren, wobei die Knoten ein jeweiliges Schlüsselwort und die Kanten die Verbindung der Schlüsselwörter in einem definierten Kontext kennzeichnen. Ein aktueller Überblick zur Extraktion von Schlüsselwörtern ist in dem Beitrag von Firoozeh et al. (2020) zu entnehmen.

TF-IDF (kurz für Term Frequency inverse Document Frequency) (Salton et al. 1975) fällt unter den statistischen Ansatz und nutzt neben der Frequenz (TF) auch die

inverse Häufigkeit (IDF), um Schlüsselwörter zu extrahieren. Formal ausgedrückt ergibt sich der Wert für das Wort *t* in dem Text *d* aus der Sammlung von Texten *D* aus Gl. 1.

$$TF - IDF\,(t, d, D) = TF\,(t, d) \cdot IDF\,(t, D)\,,$$

Gl. 1*: Berechnung von TF-IDF.*

wobei *TF(t,d)* die Häufigkeit von *t* in *d* kennzeichnet und *IDF(t,D)* die Anzahl an Texten darstellt, die *t* enthält.

YAKE! ist eine Methode basierend auf dem unüberwachten Ansatz und extrahiert Schlüsselwörter aus Texten unabhängig von Domain, Korpus oder Sprache. Der Extraktionsvorgang besteht aus vier Phasen: 1) In der ersten Phase wird ein Tokenisierungsverfahren angewendet, das den Text in einzelne Terme aufteilt, sobald ein Leerzeichen oder ein Sonderzeichen (bspw. Klammer, Komma oder Punkt) vorkommen. 2) Anschließend wird ein Wert für fünf Merkmale eines Terms ermittelt. Die Merkmale sind wie folgt beschrieben: Bei dem Merkmal *Großschreibung* wird davon ausgegangen, dass Terme, die mit einem großen Buchstaben beginnen, tendenziell eine größere Bedeutung für den Text haben. Gleiches gilt für Akronyme. Durch die *Wortposition* sollen Terme, die am Anfang eines Textes vorkommen, höher gewichtet werden. Die *Wortfrequenz* kennzeichnet die Häufigkeit eines Wortes, wobei Terme, die häufiger im Text erscheinen, höher gewichtet werden. Das Merkmal *Wortverwandtschaft zum Kontext* berechnet die Anzahl an unterschiedlichen Begriffen, die im Kontext eines Terms auftreten. Je mehr unterschiedliche Begriffe im Kontext eines Terms, desto bedeutungsloser ist der Term für den Text. Das letzte Merkmal *Wort in unterschiedlichen Sätzen* quantifiziert, wie oft ein Term in unterschiedlichen Sätzen auftritt. 3) Im dritten Schritt werden die Merkmalswerte kombiniert, sodass für jeden Term *w* das Gewicht *S(w)* resultiert. 4) Da ein Schlüsselwort aus mehreren Termen bestehen kann, wird schließlich eine sogenannte Kandidatenliste erstellt. Bei diesem Vorgang werden Terme basierend auf einem gleitenden Fenster (engl. „sliding window“) zusammengeführt, sodass Schlüsselwörter bestehend aus einer vordefinierten Anzahl an Termen resultierten. Die vordefinierte Anzahl wird durch den Parameter N-Gram angegeben, wobei N für die Anzahl an Termen steht, aus der sich ein Schlüsselwort zusammensetzen soll. Anschließend wird jedem Schlüsselwort *kw* einen finalen Wert *S(kw)* zugewiesen, wobei gilt: Je kleiner der Wert, desto aussagekräftiger ist das Schlüsselwort für den Text. Gl. 2 veranschaulicht die Berechnung des finalen Wertes.

$$S\,(kw) = \frac{\prod_{w \in kw} S\,(w)}{TF\,(kw) \cdot \left(1 + \sum_{w \in kw} S\,(w)\right)},$$

Gl. 2*: Finaler Wert nach YAKE!* wobei *S(kw)* den finalen Wert eines Schlüsselwortes und *S(w)* den kombinierten Merkmalswert für den Term *w* repräsentiert. *TF(kw)* kennzeichnet die Frequenz des Schlüsselwortes *kw*. Dabei gilt: Je kleiner *S(kw)*, desto relevanter ist *kw* für den Text (Campos et al. 2018b). In der Regel wird YAKE! zur Extraktion von Schlüsselwörtern aus einzelnen Texten eingesetzt.

Um die Relevanz von Schlüsselwörtern für mehrere Texte zu ermitteln, erfolgt die Berechnung des finalen Wertes nach Gl. 3.

$$\sum_{i=1}^{N} 1 - S\,(kw)_i,$$

Gl. 3: Finaler Wert nach YAKE! für mehrere Texte.

wobei N die Texte kennzeichnet, in denen das Schlüsselwort *kw* vorkommt. Somit kennzeichnet ein höherer Wert auch eine höhere Relevanz eines Schlüsselwortes für den gesamten Korpus.

3 Gegenstand der Literaturrecherche und Auswahl einer Methodik zur Extraktion von Schlüsselwörtern

Um den Nutzen einer Extraktion von Schlüsselwörtern bei einer Literaturrecherche zu untersuchen, wurde eine Literaturrecherche zu Empfehlungssystemen durchgeführt und analysiert. Empfehlungssysteme stellen einem Nutzer potenziell interessante Inhalte vor, die auf seinen persönlichen Interessen basieren. Diese Inhalte beziehen sich auf einen Entscheidungsprozess, zum Beispiel: Welche Nachrichten werden konsumiert? Welche Musik wird gehört? Welche Produkte werden gekauft? In diesem Zusammengang kennzeichnet ein Objekt den zu empfehlenden Inhalt (zum Beispiel eine Nachricht, ein Musiktitel oder ein Produkt). Das Ziel eines Empfehlungssystems besteht nun darin, das Interesse eines Objekts für einen Nutzer zu prognostizieren (Ricci et al. 2015). Um eine Empfehlung zu generieren existieren unterschiedlichen Konzepte: Beim kollaborativen Filtern (engl. „collaborative Filtering“) basiert eine Empfehlung auf dem Bewertungsverhalten von mehreren Nutzern. Demnach werden diejenigen Objekte vorgeschlagen, die ähnliche Nutzer auch als gut empfunden haben (Schafer et al. 2007). Konträr zum kollaborativen Filtern werden beim inhaltsbasierten Filtern (engl. „contentbased Filtering“) ausschließlich die Präferenzen eines Nutzers für die Erstellung einer Empfehlung berücksichtigt. Die Bewertungen von anderen Nutzern fließt demnach nicht in den Generierungsprozess (Meteren und Someren 2000). Hybrides Filtern kombiniert mehrere Konzepte, um die Nachteile eines Konzepts durch die Vorteile eines weiteren Konzepts zu beheben (Burke 2002).

Die Literaturauswertung wird auf der Forschungsplattform ACM Digital Library (ACM) durchgeführt (ACM 2022). ACM Digital Library ist eine der weltweit größten Datenbanken mit wissenschaftlichen Literaturbeiträgen zum Thema Informationstechnologie und erlaubt den Export von Teilen einer wissenschaftlichen Publikation, wie beispielsweise Abstract, Jahr oder definierte Schlüsselwörter der Autoren. Im Rahmen der durchgeführten Literaturanalyse muss der Suchbegriff „Recommender Systems“ in Titel, Abstract oder den vom Autor definierten Schlüsselwörtern enthalten sein. Die Suche erfolgt nur für wissenschaftliche Publikationen, die als „research papers“ klassifiziert sind. Diese Suche liefert 1843 wissenschaftliche Literaturbeiträge aus den Jahren 2007 bis 2021 zum Thema Empfehlungssysteme. Abb. 3 zeigt die Anzahl an wissenschaftlichen Literaturbeiträge in Abhängigkeit vom Publikationsjahr.

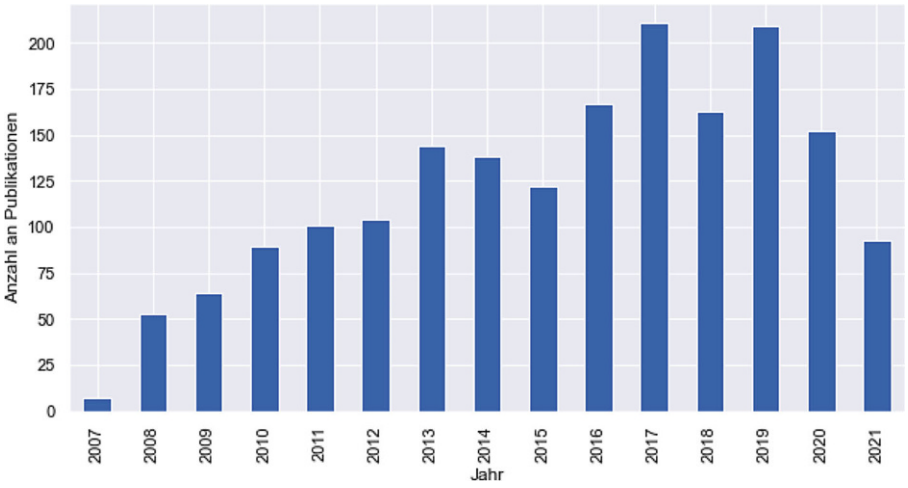

Abb. 3 ACM: Anzahl an wissenschaftlichen Publikationen zu „Recommender Systems“ von 2007 bis 2021

Dabei ist ein starkes Wachstum bezüglich der Anzahl an wissenschaftlichen Publikationen zu erkennen: Angefangen bei sieben Veröffentlichungen im Jahr 2007 steigt die Anzahl bis 2017 auf 211 wissenschaftliche Publikationen an. Die Abstracts sowie die definierten Schlüsselwörter der Autoren werden für die Analyse vorbereitet. Dazu wird neben dem Suchbegriff „recommender systems“ auch die Stoppwörter aus der Bibliothek „NLTK“ (Bird et al. 2009) aus dem Textkorpus entfernt. Stoppwörter beschreiben dabei diejenigen Begriffe, die keine Relevanz für die Beschreibung eines Textes aufweisen. Zudem erfolgt eine Normalisierung, indem alle Begriffe auf ihren Wortstamm („studies“ und „studying“ zu „study“) gekürzt werden (Anandarajan et al. 2019).

Um eine geeignete Methode zur Extraktion von Schlüsselwörtern für den Datensatz zu wählen, erfolgt ein Vergleich von unterschiedlichen Extraktionsalgorithmen. 2018 veröffentlichen Campos et al. eine Publikation, in der die „State-of-the-Art“ Methoden basierend auf einer Sammlung von wissenschaftlichen Literaturbeiträgen verglichen werden (Campos et al. 2018a). Als Metriken zur Bewertungsgrundlage werden *Precision, Recall* und *F1-Wert* eingesetzt. Diese sind in Gl. 4; Gl. 5 und Gl. 6 definiert (Sun et al. 2020).

$$Precision = \sum_{i=1}^{N} \frac{A_i \cap B_i}{A_i} / N$$

Gl. 4: Precision.

$$Recall = \sum_{i=1}^{N} \frac{A_i \cap B_i}{B_i} / N$$

Tab. 1 Vergleich von YAKE! und TF-IDF mittels Precision, Recall und F1-Wert

	YAKE!				TF-IDF			
Extrahierte Schlüsselwörter	5	10	15	20	5	10	15	20
Precision	0,4	0,4	0,5	0,6	0,3	0,4	0,4	0,5
Recall	0,8	0,4	0,33	0,3	0,6	0,4	0,267	0,25
F1-Wert	0,53	0,4	0,4	0,4	0,4	0,4	0,32	0,33

Gl. 5: Recall.

$$F1 - Wert = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Gl. 6: F1-Wert, wobei A_i die extrahierten Schlüsselwörter für Text i kennzeichnet. B_i repräsentieren die definierten Schlüsselwörter der Autoren, und N beschreibt die gesamte Anzahl an Texten im Korpus. In ihrer Untersuchung schneidet YAKE! am besten ab, gefolgt von TextRank und TF-IDF. Ähnliche Ergebnisse sind Peuker und Barton zu entnehmen: In der Publikation vergleichen die Autoren unterschiedliche Extraktionsalgorithmen für eine wissenschaftliche Literaturauswertung (Peuker und Barton 2021). Dabei werden die definierten Schlüsselwörter der Autoren mit den extrahierten Schlüsselwörtern eines Algorithmus verglichen und mittels F1-Wertes bewertet. Auch hier erzielt YAKE! die besten Resultate, gefolgt von TF-IDF. Um einen geeigneten Extraktionsalgorithmus für den in diesem Beitrag zugrunde liegenden Datensatz zu wählen, erfolgt ein weiterer Vergleich der Extraktionsalgorithmen. Da YAKE! und TF-IDF in den genannten Publikationen am besten abschneiden, werden die beiden Methoden auch für den Vergleich betrachtet. Hierfür werden die Top-10 definierten Schlüsselwörter der Autoren mit den extrahierten Schlüsselwörtern der Extraktionsalgorithmen mittels Precision, Recall und F1-Wert verglichen. Die Ergebnisse sind Tab. 1 zu entnehmen.

Dabei ist zu erkennen, dass YAKE! in allen Bereichen entweder gleich oder besser als TF-IDF abschneidet. Im Rahmen dieses Beitrags wird YAKE! als Extraktionsalgorithmus verwendet.

Tab. 2 Top-10 definierte und extrahierte Schlüsselwörter für die Jahre 2020 und 2021

Top-10 definierte Schlüsselwörter	Frequenz	Top-10 extrahierte Schlüsselwörter	Finaler Wert nach YAKE!
Collaborative Filtering	19	*Collaborative Filtering*	38,7
Deep Learning	14	*Neural Network*	15,9
Conversational	12	*Matrix Factorization*	14,9
Neural Network	9	Social Network	11,9
Machine Learning	8	*Machine Learning*	7,9
Fairness	8	*Deep Learning*	6,9
Matrix Factorization	8	Implicit Feedback	6,0
Reinforcement Learning	7	Social Media	5,9
Sequential	7	User Profile	5,0
Learn to Rank	7	Group Members	5,0

4 Ergebnisse

Nach Anwendung des Extraktionsalgorithmus auf die vorbereiteten Abstracts stehen neben den definierten Schlüsselwörtern der Autoren auch die extrahierten Schlüsselwörter von YAKE! zur Verfügung. Um die Relevanz von Begriffen für einen Textkorpus zu ermitteln, können beide Schlüsselwortkategorien entweder separat voneinander oder in Kombination genutzt werden.

Bei einer separaten Ermittlung wird für die definierten Schlüsselwörter der Autoren die Frequenz genutzt, um Begriffe nach Relevanz zu ordnen. Für die extrahierten Schlüsselwörter hingegen wird der finale Wert nach YAKE! (siehe Gl. 3) für eine Sortierung genutzt. Tab. 2 zeigt neben den Top-10 definierten Schlüsselwörtern der Autoren auch die Top-10 extrahierten Schlüsselwörter nach YAKE!. Um einen Überblick über aktuelle Themen im Bereich der Empfehlungssysteme darzustellen, basiert die Auswertung auf wissenschaftlichen Literaturbeiträgen aus den Jahren 2020 und 2021. Die Schlüsselwörter sind anhand der Frequenz bzw. dem finalen Wert nach YAKE! sortiert. Die übereinstimmenden Schlüsselwörter sind dabei hervorgehoben.

Zunächst ist zu bemerken, dass sowohl die definierten als auch die extrahierten Schlüsselwörter eine wesentliche Bedeutung im Bereich der Empfehlungssysteme einnehmen. Besondere Bedeutung erhält das Schlüsselwort „Collaborative Filtering“ (zu Deutsch „kollaboratives Filtern“), das sowohl bei den definierten als auch bei den extrahierten Schlüsselwörtern die Top-1 darstellt. Das Schlüsselwort „Matrix Factorization“ (zu Deutsch Matrixfaktorisierung) beschreibt die wohl bekannteste Methode zur Umsetzung von kollaborativen Filtern (Koren et al. 2009). Die definierten und die extrahierten Schlüsselwörter können auch in Kombination verwendet werden, um die Relevanz von Schlüsselwörtern für einen Textkorpus zu ermitteln. Um die beiden Schlüsselwortkategorien in Relation zu setzen, werden die Werte zunächst normalisiert. Eine Methode für die Normalisierung ist durch die Min-Max-Methode gegeben. Dabei nehmen die Schlüsselwörter in beiden Kategorien einen Wert zwischen 0 und 1 an (Borkin et al. 2019). Die Normalisierung von einem Schlüsselwort soll durch Gl. 7 verdeutlicht werden:

$$x\prime = \frac{x - min\,(x)}{max\,(x) - min\,(x)}$$

Abb. 4 Identifizierte Top-10 Schlüsselwörter für die Jahre 2020 und 2021

Gl. 7: Min-Max-Normalisierung.

Dabei kennzeichnet x den Wert eines Schlüsselwortes und $x\prime$den normalisierten Wert. Die Schlüsselwortkategorien können nun in Kombination zur Beschreibung von Inhalten genutzt werden, indem die normalisierten Werte der Schlüsselwörter addiert werden. Hierfür sind in Abb. 4 die Top-10 Schlüsselwörter zu entnehmen, die Kategorie übergreifend den höchsten Wert aufweisen.

Neben den Konzepten des kollaborativen Filterns („Collaborative Filtering“) und der Matrixfaktorisierung („Matrix Factorization“) tauchen mit „Machine Learning“, „Neural Network“, „Deep Learning“ auch moderne Methoden aus den Bereichen KI und Machine Learning als Schlüsselwörter auf. Um den Fokus auf aktuelle Themen im Bereich der Empfehlungssysteme darzustellen, bezieht sich die in Abb. 4 gezeigte Auswertung ausschließlich auf die Jahre 2020 und 2021. Daraus lässt sich ableiten, dass in den letzten Jahren auch Methoden des Machine Learning besondere Aufmerksamkeit im Bereich der Empfehlungssysteme erfahren. Neuronale Netze bestehen aus mehreren Schichten wie Eingangs-, Zwischen- und Ausgangsschicht. Modelle mit mehreren Zwischenschichten werden als tiefe neuronale Netze bezeichnet und sind dem Bereich Deep Learning zugeordnet. Dadurch ist es möglich, sehr komplexe Zusammenhänge in den Daten zu modellieren (Choi et al. 2020). Das Interesse an diesen Methoden ist besonders in den letzten Jahren gestiegen, als die hoch bewerteten Internetunternehmen wie Amazon, Google oder Facebook Empfehlungsalgorithmen basierend auf Deep Learning in ihre Anwendung integrieren. In diesem Sinne werden tiefe neuronale Netze (engl. „Deep Neural Networks“) von Unternehmen wie Google oder Facebook eingesetzt, um Empfehlungen zu erstellen (Cheng et al. 2016; Gupta et al. 2020). Im Jahr 2016 veröffentlichte Amazon sogar das Deep Learning Framework „Deep Scalable Sparse Tensor Neural Engine“ (kurz DSSTNE) unter Open-Source-Lizenz, mit dem das Unternehmen die personalisierten Empfehlungen für ihre Kunden generiert (Amazon 2021).

Über die Identifizierung von aktuellen Trends hinaus lassen sich die von Autoren definierten und die extrahierten Schlüsselwörter dazu einsetzen, die Relevanz von

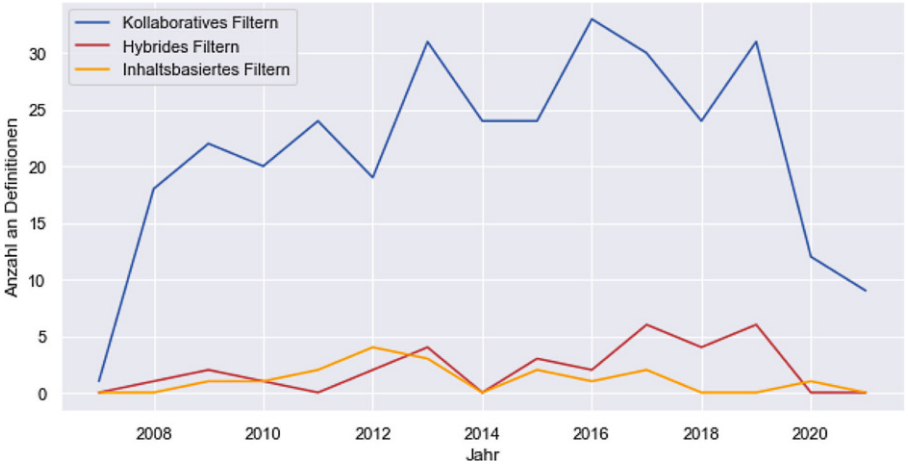

Abb. 5 Konzepte zur Empfehlungsgenerierung anhand der Anzahl an Definitionen der Autoren

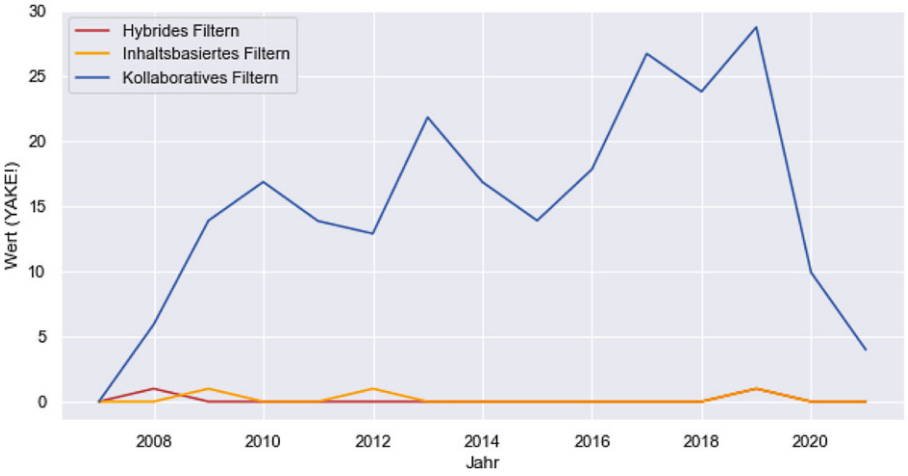

Abb. 6 Konzepte zur Empfehlungsgenerierung anhand dem Wert nach YAKE!

Themen in einem zeitlichen Verlauf dazustellen. Dadurch ist auch eine Gegenüberstellung von Themen in Abhängigkeit der Zeit möglich. Dies ist Abb. 5 basierend auf den definierten Schlüsselwörtern der Autoren zu entnehmen.

Dabei wird deutlich, dass in der wissenschaftlichen Literatur über den gesamten Zeitraum gesehen kollaboratives Filtern sowohl nach den definierten als auch nach den extrahierten Schlüsselwörtern das meistgenutzte Konzept zur Erstellung von Empfehlungen darstellt. Demnach werden insgesamt 320 Publikationen mit dem Schlüsselwort für kollaboratives Filtern („collaborative filtering“) gekennzeichnet. Dagegen ist die Relevanz von inhaltsbasierten und hybriden Filtern vergleichsweise niedrig. In Bezug auf den gesamten Zeitraum werden dabei 32 Publikationen durch hybrides Filtern und 17 Publikationen durch inhaltsbasiertes Filtern gekennzeichnet.

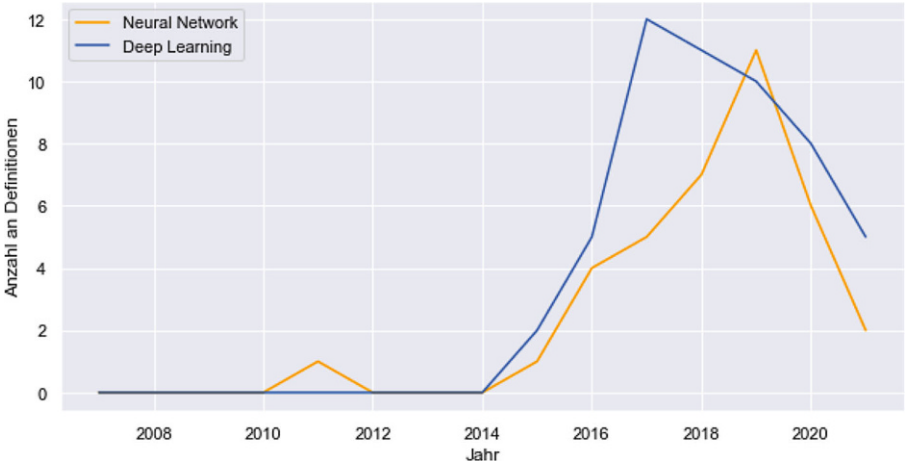

Abb. 7 Neuronale Netze und Deep Learning anhand der Anzahl an Definitionen der Autoren

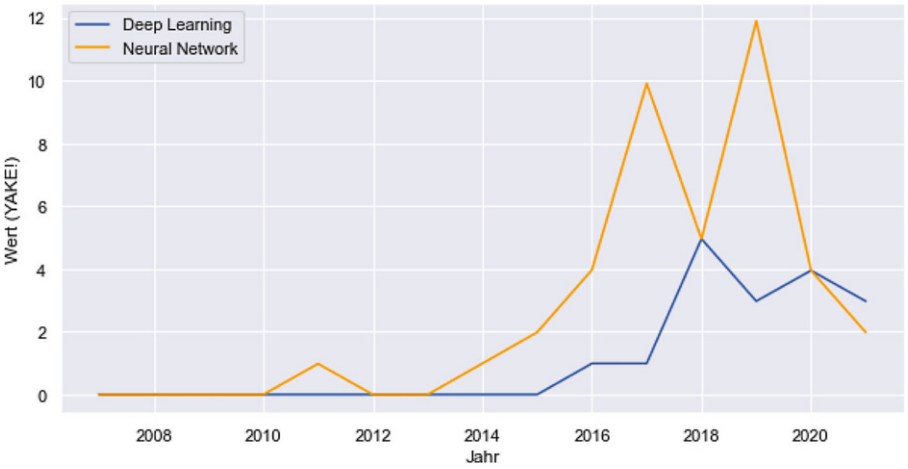

Abb. 8 Neuronale Netze und Deep Learning anhand dem Wert nach YAKE!

Ähnliche Ergebnisse sind auch durch die extrahierten Schlüsselwörter zu erkennen. Dabei wird in den Abstracts fast ausschließlich das kollaborative Filtern als Konzept zur Erstellung von Empfehlungen beschrieben. Dies ist Abb. 6 zu entnehmen.

In den letzten Jahren haben neuronale Netze und Deep Learning zunehmendes Interesse erfahren. Diese Entwicklung lässt sich auch anhand der Schlüsselwörter darstellen. Demnach zeigen Abb. 7 und 8 die Entwicklung von neuronalen Netzen und Deep Learning für die Jahre 2007 bis 2021 auf. Gegenübergestellt ist die Entwicklung basierend auf den definierten Schlüsselwörtern der Autoren (Abb. 7) sowie den extrahierten Schlüsselwörtern mittels YAKE! (Abb. 8).

Hierbei wird deutlich, dass die Autoren ab 2014 vermehrt das Schlüsselwort Deep Learning definieren, in den Abstracts jedoch die Bezeichnung neuronale Netze verwenden. Da Methoden des Deep Learning dem Term „Neural Network“ zuzuordnen sind, ist dies nicht als Abweichung anzusehen (LeCun et al. 2015). So beschreiben beispielsweise Convolutional Neural Networks (kurz CNN’s) neuronal Netze im Bereich des Deep Learning, die speziell für die Erkennung von Bildern geeignet ist (Albawi et al. 2017). Durch CNN’s können beispielsweise ähnliche Produkte anhand ihrer Artikelbilder in einem Webshop gefunden und als Grundlage für eine Empfehlung genutzt werden. Recurrent Neural Networks (kurz RNN’s) sind darauf ausgelegt, sequenzielle oder zeitliche Muster zu erkennen (Medsker und Jain 2001). Im Hinblick auf Empfehlungssysteme kann einem Kunden somit ein Produkt vorgeschlagen werden, basierend auf der Reihenfolge von bereits gekauften Artikeln.

Durch das beschriebene Vorgehen kann die Relevanz der Themen in den Abstracts bzw. in den definierten Schlüsselwörtern separiert betrachtet werden. Soll jedoch die allgemeine Entwicklung der Themen dargestellt werden, so bietet sich eine Auswertung mittels dem Kategorie übergreifenden Verfahren an. Somit werden die Schlüsselwörter, die in beiden Kategorien vorkommen, höher gewichtet. Der Abb. 9 ist die zeitliche Entwicklung der Themen basierend auf dem Kategorie übergreifenden Verfahren zu entnehmen.

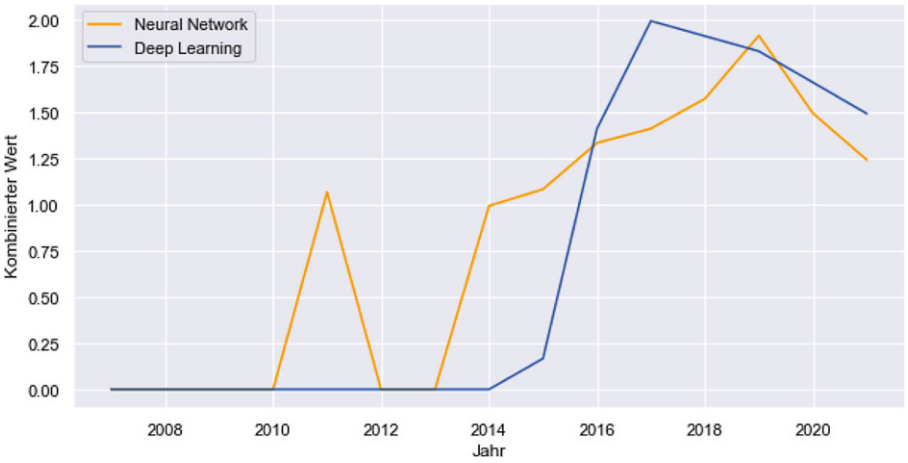

Abb. 9 Neuronale Netze und Deep Learning anhand des kombinierten Wertes

Im Gegensatz zu einer separaten Betrachtung der Schlüsselwörter ist zu erkennen, dass Deep Learning im Jahr 2017 die größte Relevanz einnimmt, da das Schlüsselwort von den Autoren vermehrt in den Abstracts beschrieben und als Schlüsselwort definiert wird.

Um den Einsatz von Methoden im Bereich Deep Learning näher zu untersuchen, erfolgt nun eine Betrachtung von CNN's und RNN's. Basierend auf den extrahierten Schlüsselwörtern stellt Abb. 10 die Relevanz der Methoden im zeitlichen Verlauf dar.

Hier ist zu erkennen, dass ab 2014 vermehrt RNN's zur Erstellung von Empfehlungen eingesetzt werden. Besonders im Jahr 2017, in dem auch die höchste Relevanz von Deep Learning zu erkennen ist, werden zunehmend RNN's in den Abstracts beschrieben. Hingegen ist ein besonderes Interesse an dem Einsatz von CNN's im Jahr 2019 zu erkennen.

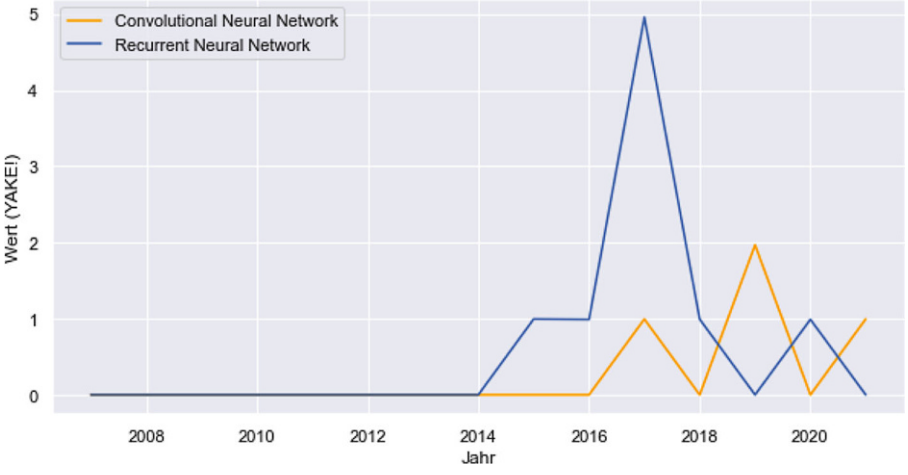

Abb. 10 CNN's und RNN's anhand dem Wert nach YAKE!

5 Zusammenfassung und Ausblick

Durch die Extraktion von Schlüsselwörtern aus den Abstracts von wissenschaftlichen Publikationen ist es möglich, eine Vielzahl von Publikationen für eine wissenschaftliche Literaturauswertung zusammenzufassen. Zudem kann die Entwicklung von themenrelevanten Begriffen in einem zeitlichen Verlauf dargestellt werden. Auf diese Weise lassen die Themen identifizieren, die zu einem gewissen Zeitpunkt oder in einem bestimmten Zeitraum von besonderer Relevanz für Forschende sind. Diese Themen können dann wieder Ausgangspunkt für weitere Recherchen sein. Im Rahmen dieses Beitrags wurde dies anhand einer Datenbasis von wissenschaftlichen Publikationen zum Thema Empfehlungssysteme aus den Jahren 2007 bis 2021 demonstriert.

Darauf aufbauend wäre es denkbar, die semantische Bedeutung von themenrelevanten Begriffen für eine wissenschaftliche Literaturauswertung zu nutzen: Durch sogenannte Worteinbettungen (engl. Word Embedding) ist es möglich, Begriffe durch Vektoren zu repräsentieren. Anhand der Vektoren kann die Ähnlichkeit von Begriffen durch ein Distanzmaß berechnet werden. Ein Begriff kann somit durch ähnliche Begriffe näher beschrieben werden (Mikolov et al. 2013).

Weiterführende Untersuchen im Bereich der wissenschaftlichen Literaturauswertung können daraus bestehen, die Extraktion von Schlüsselwörtern auch auf Volltexte auszudehnen, Methoden wie Topic Modeling oder Dynamic Topic Modeling einzusetzen oder das Potenzial von Word Embedding zu untersuchen.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf http://creativecommons.org/licenses/by/4.0/deed.de.

Literatur

Association of Computing Machinery (ACM) (2022) ACM Digital Library. https://dl.acm.org. Zugegriffen: 25. Febr. 2022

Albawi S, Mohammed TA, Al-Zawi S (2017) Understanding of a convolutional neural network. In: Bayat O (Hrsg) International Conference on Engineering and Technology Antalya, S 1–6

Amazon.com, Inc.: Deep Scalable Sparse Tensor Network Engine (DSSTNE). https://github.com/amazon-archives/amazon-dsstne. Zugegriffen: 13. Dez. 2021

Anandarajan M, Hill C, Nolan T (2019) Text preprocessing. In: Sharda R, Chen H (Hrsg) Practical text analytics: maximizing the value of text data. Advances in analytics and data science, 2. Aufl. Springer, Cham

Bird S, Kein E, Loper E (2009) Natural language processing with python. O'Reilly Media
Booth A, Sutton A and Papaioannou D (2016) Systematic approaches to a successful literature review. Sage Publications Ltd, London
Borkin D, Némethová A, Michalčonok G, Maiorov K (2019) Impact of data normalization on classification model accuracy. Res Pap Fac Mater Sci Technol Slovak 27(45):79–84
Brin S, Page L (1998) The anatomy of large-scale hypertextual web search engine. Comput Networks ISDN Syst 30(1–7):107–117
vom Brocke J, Simons A, Niehaves B, Reimer K, Plattfaut R, Cleven A (2009) Reconstructing the giant: on the importance of rigour in documenting the literature search process. In: Newell S, Whitley E, Pouloudi N, Wareham J, Mathiassen L (Hrsg) Proceedings of the 17. European conference on information systems Verona
Burke R (2002) Hybrid recommender systems: survey and experiments. User Model User-Adap Inter 12(2):331–370
Campos R, Mangaravite V, Pasquali A, Jorge AM, Nunes C, Jatowt A (2018a) A text feature based automatic keyword extraction method for single documents. In: Pasi G, Piwowarski B, Azzopardi L, Hanbury A (Hrsg) ECIR 2018: advances in information retrieval. Springer, Cham, S 684–691
Campos R, Mangaravite V, Pasquali A, Jorge AM, Nunes C, Jatowt A (2018b) YAKE! Collection-independent automatic keyword extractor. In: Pasi G, Piwowarski B, Azzopardi L, Hanbury A (Hrsg) ECIR 2018: advances in information retrieval. Springer, Cham, S 806–810
Cheng H-T, Koc L, Harmsen J, Shaked T, Chandra T, Aradhye H, Anderson G, Corrado G, Chai W, Ispir M, Anil R, Haque Z, Hong L, Jain V, Liu X, Shah H (2016) Wide & deep learning for recommender systems. In: Karatzoglou A, Hidasi B, Tikk D, Sar-Shalom O, Roitman H, Shapira B, Rokach L (Hrsg) Proceedings of the 1st Workshop on Deep Learning for Recommender Systems Boston, S 7–10
Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP (2020) Introduction to machine learning, neural networks, and deep learning. Trans Vis Sci Tech 9(2):14
van Dinter R, Tekinerdogan B, Catal C (2021) Automation of systematic literature reviews: a systematic literature review. Inf Softw Technol 136(4):106589. https://doi.org/10.1016/j.infsof.2021.106589
Feather J, Sturges P (2003) International encyclopedia of information and library science. Routledge
Firoozeh N, Nazarenko A, Alizon F, Daille B (2020) Keyword extraction: issues and methods. Nat Lang Eng 26(3):259–291
Gupta U, Wu C-J, Wang X, Naumov M, Reagen B, Brooks D, Cottel B, Hazelwood K, Hempstead M, Jia B, H-HS L, Malevich A, Mudigere D, Samelyanskiy M, Xiong L, Zhang X (2020) The architecture implications of Facebook's DNN-based personalized recommendation. In: Tullsen D, Esmaeilzadeh H (Hrsg) IEEE International Symposium on High Performance Computer Architecture (HPCA) San Diego, S 488–501
Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. IEEE Comput Soc 42(8):42–49
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
Medsker LR, Jain LC (2001) Recurrent neural networks: design and applications. CRC Press, Boca Raton
Meteren R, Someren M (2000) Using Content-based filtering for recommendation. In: Potamias G, Moustakis V, van Someren M (Hrsg) Proceedings of ECML 2000 workshop: machine learning in information age, S 47–56
Mihalcea R, Tarau P (2004) Textrank: bringing order into texts. In: Lin D, Wu D (Hrsg) Proceedings of the 2004 conference on empirical methods in natural language processing Barcelona, S 404–411
Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: Bengio Y, LeCun Y (Hrsg) International conference on learning representations Scottsdale, S 1–12
Peuker A, Barton T (2021) Comparison of different keyword extractors for an automated literature review. In: Böhm S, Suntrayuth S (Hrsg) Proceedings of the 5th international workshop on entrepreneurship, electronic and mobile business
Ricci F, Rokach L, Shapira B (2015) Recommender systems: introduction and challenges. In: Ricci F, Rokach L, Shapira B (Hrsg) Recommender systems handbook, S 1–34
Rose S, Engel D, Cramer N, Cowley W (2010) Automatic keyword extraction from individual documents. In: Berry MW, Kogan J (Hrsg) Text mining: applications and theory. John Wiley & Sons, Chichester, S 1–20
Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18(11):609–664
Schafer JB, Frankowski D, Herlocker J, Sen S (2007) Collaborative filtering recommender systems. In: Brusilovsky P, Kobsa A, Nejdl W (Hrsg) The adaptive web. Springer, Berlin Heidelberg, S 291–324

Siddiqi S, Sharan A (2015) Keyword and keyphrase extraction techniques: a literature review. Int J Comput Appl 109(2):18–23

Sun C, Hu L, Li S, Li T, Li H, Chi L (2020) A review of unsupervised keyphrase extraction methods using within-collection resources. Symmetry 12(11):1864. https://doi.org/10.3390/sym12111864

Tauchert C, Bender M, Mesbah N, Buxmann P (2020) Towards an integrative approach for automated literature reviews using machine learning. In: Bui TX (Hrsg) Proceedings of the 53rd Hawaii international conference on system sciences Maui, S 762–771