

Angela Vorndran, Stefan Grund, Claudia Grote, Jan Eberhardt

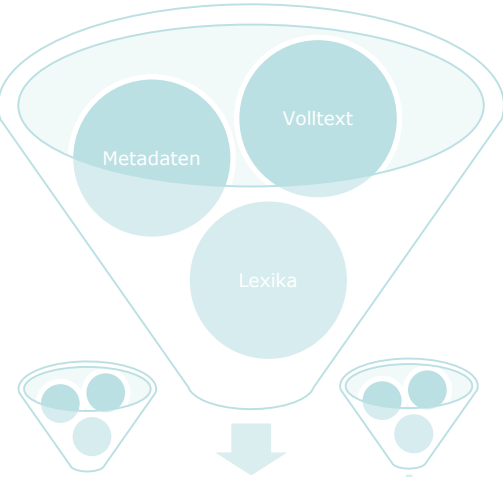
Die Maschine überwacht die Maschine

- Wie maschinelle Verfahren automatisierte Prozesse verbessern können

Gliederung

1. Motivation
2. Maschinelle Verfahren zur Beurteilung der Eignung für die maschinelle Sachgruppenvergabe
 1. Datenlage
 2. Maschinelle Verfahren
 3. Fazit
3. Metrik zur Beurteilung von Werkbündeln

Motivation



- Automatisch erzeugte Inhalte spielen immer wichtigere Rolle in bibliographischen Metadaten
- Möglichkeiten der intellektuellen Evaluierung sind begrenzt wegen hoher Anzahlen

130;943;355;420;090;420;90;530;029;355;
100;430;100;540;K;960;363;435;150;437;1
1;970;370;437;290;460;290;600;200;439;2
10;980;380;439;200;439;200;570;301;480;
17;990;390;440;300;470;300;610;020;350;
20;350;400;450;327;500;327;624;33;363;3
29;355;410;460;010;980;380;439;030;360;
30;360;30;470;090;420;90;530;100;430;10
33;363;33;480;K;960;363;435;060;390;60;
15;370;45;490;060;390;60;500;020;350;40

Maschinelle Unterstützung zur Beurteilung der Inhalte

Maschinelle Beurteilung der Werkbündel in Culturegraph

Explorative Untersuchung: maschinelle Klassifikationsverfahren zur Beurteilung der Eignung für die maschinelle Sachgruppenvergabe

Einsatz von maschinellen Klassifikationsverfahren zur Vorhersage der Eignung für maschinelle Prozesse

- Hintergrund: Aktuelles Verfahren für automatische Vergabe von Sachgruppen eingesetzt seit 2012
- Klassifikator: Support Vector Machine (SVM)
- Averbis Extraction Platform (AEP)
- Angewendet in:
 - Reihe O, Reihe B, Reihe H
 - Alle Netzpublikationen außer Belletristik
 - Formate: PDF seit 2012, Epub seit 2015
 - Sprachen: Deutsch, Englisch
- Umfang: 2.021.064 Publikationen (8.3.2019)

Idee

Bestand mit intellektuell und
maschinell vergebenen Sachgruppen

```

019@ faXA-DE-BE
021A faFeedback aus der Sicht von Kindern und L
028A f91029598215f8Hoya, Fabian [Tn3]fBVerfasse:
029F fSmf91043386068f8Springer Fachmedien Wiesb
032@ fg11fa1. Auflage 2019
033A fpWiesbadenfnSpringer Fachmedien W
034D faOnline-Ressource
037A faLizenzpflichtig
039D faErscheint auch alsfnDruck-Aus
041A f9040533697f8Schüler [Ts1]
041A/01 f9041353315f8Leseverstehen [Ts1]
041A/02 f9040334482f8Leseunterricht [Ts1]
045E fe370fEpfDie-sgfD2018-09-26
045E fe370fEifD2018-07-25
045E fe370fEmfAep-sgfK0,99998fD2018-07-25
045F feDLG22gerfa372.47
045F/01 fa372.4
045G feDLG22gerfa372.47
045G/01

```

\$E i: intellektuell
 p: aus Parallelpublikation
 m:maschinell erzeugt

Einsatz
maschineller
Lernverfahren

relevante Merkmale der Datensätze,
die richtige maschinelle Sachgruppe
erhalten, ermitteln

Vorhersage einer
richtigen/falschen
Klassifikation

Ermitteln von priorisiert
nachzubearbeitenden
Fällen

Sind maschinelle Lernverfahren
für diese Fragestellung und
bibliographische Metadaten
geeignet?

Verwendete Daten

- Set aus 65.825 Netzpublikationen mit intellektuell und maschinell vergebenen Sachgruppen
- Vergleich der intellektuellen und ersten maschinell vergebenen Sachgruppe

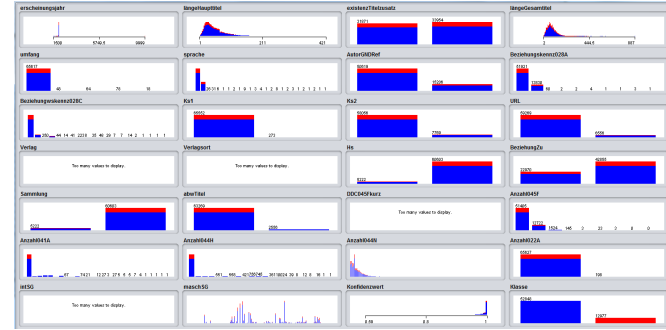
```
045E fe370fEpfHie-sgfd2018-09-26
045E fe370fEifD2018-07-25
045E fe370fEmfHaep-sgfk0,999
045F feDDC22gerfa372.47
045F/01 fa372.47
045G feDDC23gerfa372.47
045G/01 fa372.47
```

Klasse 1:
gleich

Klasse 0:
ungleich

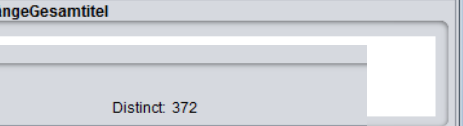
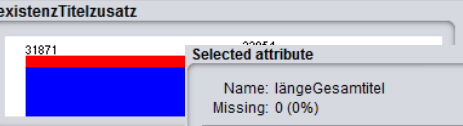
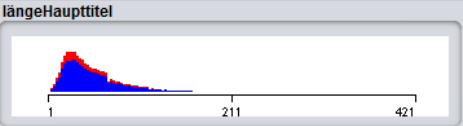
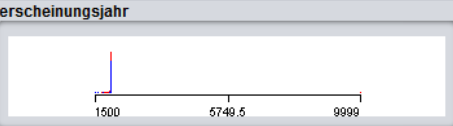
```
044N fbnoSchemefaProjektevaluierung
044N fbnoSchemefaEvaluierungskurzbericht
044N fbnoSchemefaPEV
044N fbCBNRMfaBewirtschaftung der natur"lich
045C ff333.7fF0,970fg320fG0,898fd2017-03-16
045D fKK A8_03_20160930_de
045E fe333.7fEiFD2017-03-20
045E fe330fEmfHainbfK0,990fd2017-03-16
046X f0a
101@ fa2
208@/01 fa15-03-17fbf
201B/01 f016-03-17ft04:03:51.000
201U/01 f0utf8
002@/01 f0300774505
```

Auswahl von Attributen



- Nutzung von Metadaten für die maschinellen Verfahren (kein Volltext!)
- Aussagekräftige Metadateninhalte
 - Welche Daten können einen Hinweis auf die Eignung zur maschinellen Vergabe von Sachgruppen geben?
 - Oft sehr heterogen, z.B. Verlagsname, Schlagwort
- Verändern von Variablentypen kann helfen
 - Einteilen in Bereiche z.B. Seitenumfang (0-50, 50-100, 100-200)
 - Vereinfachen zu binären Variablen (Feld belegt – nicht belegt)
 - Häufigkeiten ermitteln (Anzahl von Schlagwörtern)
 - Metrisierung (Länge des Titels)





umfang

sprache

AutorGNDRef

längeGesamtittel

Selected attribute

Name: Hs
Missing: 0 (0%)
Distinct: 2

Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	JE	5222	5222.0
2	NH	60603	60603.0

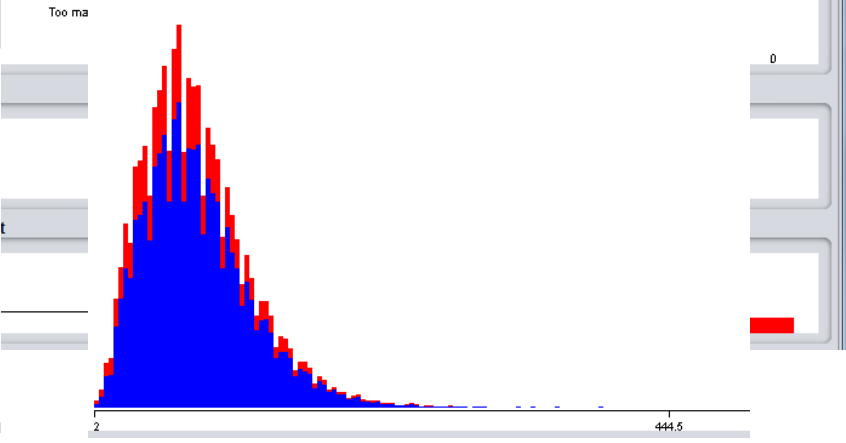
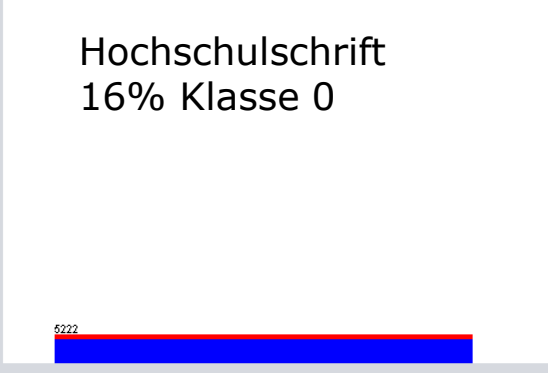
Keine
Hochschulschrift
20% Klasse 0

Selected attribute

Name: längeGesamtittel
Missing: 0 (0%)
Distinct: 372

Statistic	Value
Minimum	2
Maximum	887
Mean	81.309
StdDev	43.153

Class: Klasse (Nom) Visualize All



Getestete Verfahren

- Wichtige Attribute zur Vorhersage ermitteln
 - Information Gain
- Maschinelle Klassifikation (0,1)
 - probabilistisch: Naive Bayes
 - Regelbasiert: Decision Table
 - Distanzbasiert: IBk
- Entscheidungsbaum
 - J48

Bedingte
Wahrscheinlichkeiten für
bestimmte Klasse bei
auftreten der jeweiligen
Attribute

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Klassenzuordnung
basierend auf
erlernten Regeln

Ähnlichkeit von
neuem Dokument zu
bereits klassifizierten

Attribute mit größter Aussagekraft

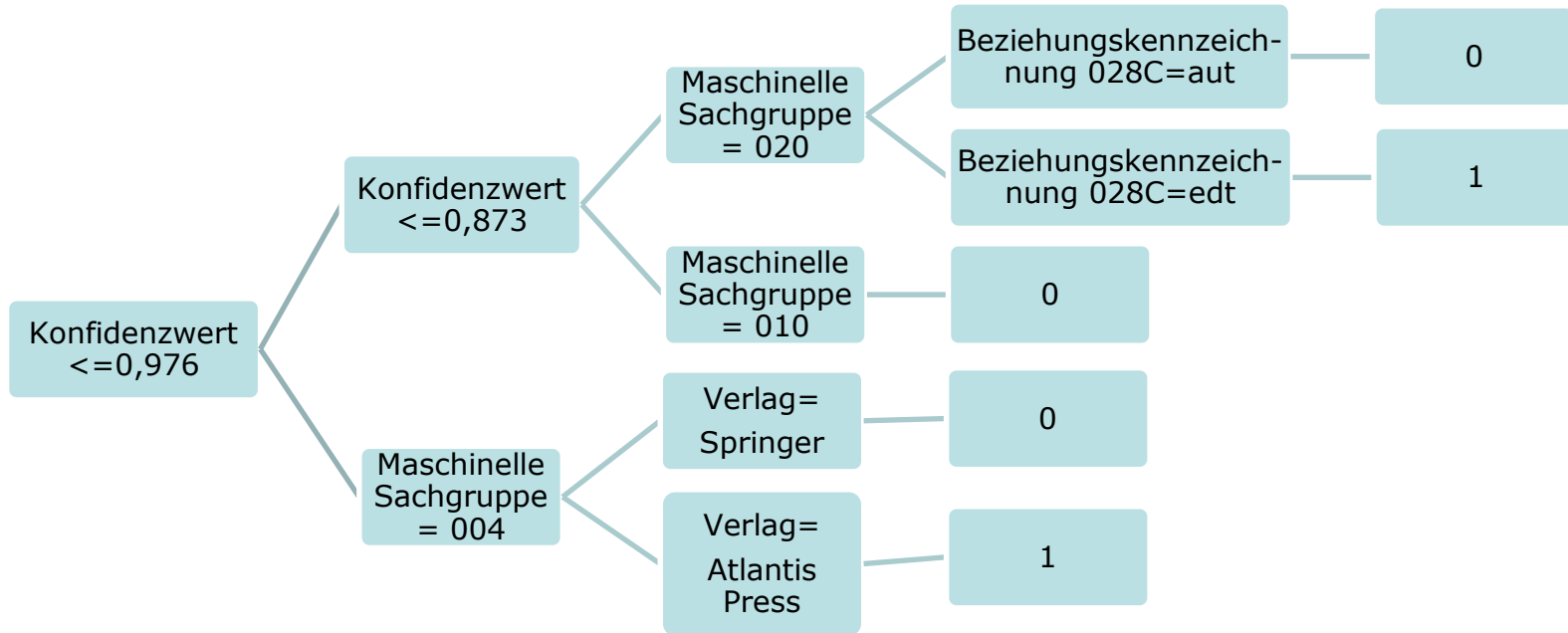
Information Gain	Attribut
0,09937176	Konfidenzwert
0,0610188	Intellektuell vergebene Sachgruppe
0,05740721	Verlag
0,0365195	Maschinell vergebene Sachgruppe
0,0279311	Verlagsort
0,02274862	Vollständige DDC-Notation (Pica-Feld 045F, verkürzt)
0,00578433	Anzahl von Deskriptoren aus Fremddaten (Pica-Feld 044N)
0,00393881	Erscheinungsjahr
0,00278322	Länge des Gesamttitels

... und unwichtigste Attribute	
0,00002062	Ist eine erste Körperschaft eingetragen?
0,0000161	Existiert ein Titelzusatz?
0,00000134	Anzahl von in der Manifestation verkörperter Werke (022A)

Ergebnisse der maschinellen Klassifikation (0,1)

Verfahren	Anteil richtig klassifizierter Datensätze in % - alle Attribute	Anteil richtig klassifizierter Datensätze in % - nach maschineller Sachgruppenvergabe*	Anteil richtig klassifizierter Datensätze in % - Lieferzustand*
Naive Bayes	78,6	79,71	75,14
Entscheidungsbaum (J48)	96,81	82	80,29
Decision Table	90	82,05	80,32
IBk	88,86	77	72,35
		*ohne intellektuelle SG und DDC	*ohne intellektuelle SG und DDC, maschinelle SG, Konfidenzwert

Beispiel Ergebnisse Entscheidungsbaum



Fazit

- Die getesteten Verfahren bieten für die verschiedenen Fragestellungen hilfreiche Informationen
- Ermittlung des Informationsgehaltes der einzelnen Attribute
 - Konfidenzwert ist zuverlässiger Anhaltspunkt
 - Zugehörigkeit zu einer Sachgruppe ist relevant
 - Großer Einfluss des Verlags und des Erscheinungsjahres bieten Ansatzpunkte für weitere Untersuchungen,
 - welche Ausprägungen geben Ausschlag?
 - liegen Gründe im Lieferformat oder Workflow?

Fazit

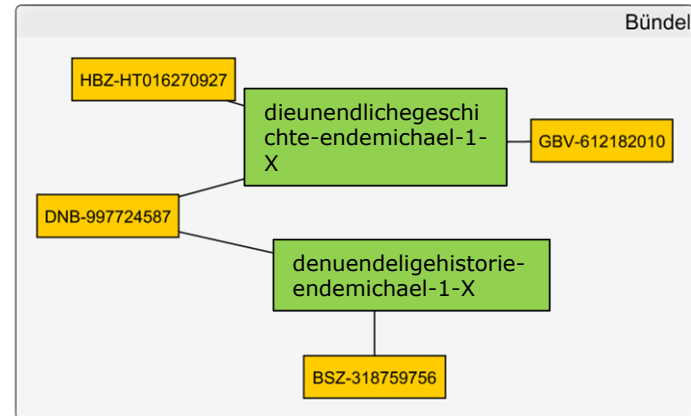
- Einsatzzweck: Auswahl zu überprüfender Titeldatensätze nach automatischer Sachgruppenvergabe
 - Die Datensätze, die Klasse 0 zugeordnet werden, werden priorisiert überprüft
- Einsatzzweck: Vorhersage der Korrektheit der automatisch vergebenen Sachgruppe vor dem Prozess
 - gelieferte Metadaten haben geringere Aussagekraft
 - Bieten Anhaltspunkt für weitere Analyse der Ausprägungen der relevanten Merkmale

Metrik zur Evaluierung von Werkbündeln

- Hintergrund:
 - In Culturegraph werden Publikationen über einen Datenabgleich zu Werken gebündelt
 - Werke enthalten verschiedene Auflagen, Ausgaben und Übersetzungen
 - Grundlage: Culturegraph-Bestand (Metadaten deutscher Bibliotheksverbände, des Österreichischen Bibliothekenverbands und der DNB)
- Im Datenbestand von >171 Mio. Titeldaten ist eine intellektuelle Evaluierung immer nur punktuell möglich
- Automatisiert ermittelte Kennzahlen sollen Aufschluss über die Richtigkeit eines Werkbündels geben

Clustering der Schlüssel nach dem Verfahren der Breitensuche

- Für jeden Datensatz liegen ein oder mehrere Schlüssel vor
- Bündel werden anhand von Schlüsseln zusammengestellt
- Dabei werden in einem Bündel alle Datensätze gesammelt, die einen gemeinsamen Schlüssel mit einem anderen Mitglied besitzen



Homogenität der Bündel ermitteln

- Entropie für relevante Angaben berechnet: AutorIn, Titel, Sprache → misst die Heterogenität der Daten
- Schritt 1: Entropie = 0 -> alle Angaben gleich, Cluster korrekt

"Clusternummer" "Feld" "Clustergröße" "Entropie" "Anteil des häufigsten Inhaltes" "häufigster Feldinhalt"

"6" "title" "17" "0.0" "1.0" "kantismoralreligion"

"6" "creator" "17" "0.0" "1.0" "Wood, Allen W."

In allen 17
Mitgliedern des
Clusters ist der Titel
und der Autor gleich

Heterogene Bündel beurteilen

- Schritt 2 (in Planung): Regeln für die Beurteilung bei Entropie $\neq 0$ z.B.
 - Titel abweichend, AutorIn gleich, Sprache abweichend -> möglicherweise korrektes Cluster, da verschiedene Übersetzungen
 - Schwellenwerte ermitteln für als wahrscheinlich korrekt anzusehende Cluster –geringer Entropiewert = Schreibfehler, andere Auflage, Namensvariante

Abkürzungen im
Titel sorgen für
Ungleichheit

```
"114898" "title" "8" "0.974" "0.5"  
"solidaritätderkirchemitisaeldiethelogis  
esverhältnissesderkirchezumjudentumna  
ziellenverlautbarungen"  
"114898" "creator" "8" "0.0" "1.0" "Wirth, Wolfgang"  
"114898" "language" "8" "0.0" "1.0" "ger"
```

Hauptsachtitel	Solidarität Der Kirche Mit Israel
Zusatz	Die Theologische Neubestimmung Des Verhältnisses Der Kirche Zum Judentum Nach 1945 Anhand Der Offiziellen Verlautbarungen
Person	aut Wirth, Wolfgang

Hauptsachtitel	Solidarität Der Kirche Mit Israel
Zusatz	D. Theol. Neubestimmung D. Verhältnisses D. Kirche Zum Judentum Nach 1945 Anhand D. Offiziellen Verlautbarungen
Person	aut Wirth, Wolfgang 130544299

Fazit

- Durch die Anwendung einer Metrik können wir das Verfahren verbessern, indem wir Problemfälle identifizieren und Algorithmen zur Schlüsselbildung anpassen können
- Es wird aber vermutlich auch „Problembündel“ geben, die nicht zu korrigieren sind und die daher von vorneherein aus den maschinellen Verfahren zur Datenübernahme herausgenommen werden sollten.

Vielen Dank für Ihre
Aufmerksamkeit.
Fragen?

Kontakt: Angela Vorndran

a.vorndran@dnb.de

Stefan Grund

s.grund@dnb.de