Angela Vorndran, Stefan Grund, Claudia Grote, Jan Eberhardt

Die Maschine überwacht die Maschine

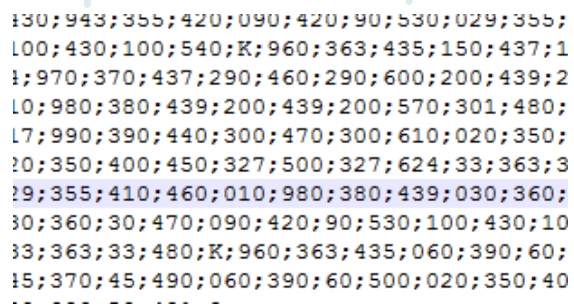
- Wie maschinelle Verfahren automatisierte Prozesse verbessern können

DEUTSCHE NATIONAL BIBLIOTHEK

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

Maschinelle Unterstützung zur Beurteilung der Inhalte

Maschinelle Beurteilung der Werkbündel in Culturegraph

Explorative Untersuchung: maschinelle Klassifikations-verfahren zur Beurteilung der Eignung für die maschinelle Sachgruppenvergabe

DEUTSCHE NATIONAL BIBLIOTHEK

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 fe370fEiff02018-07-25
045E fe370fEmfdaep-sgfK0,99998fD2018-07-25
045F feDDC22ger:fa372.47
045F/01 fa372.4
045G feDD
045G/01 f
```

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

Einsatz maschineller Lernverfahren

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

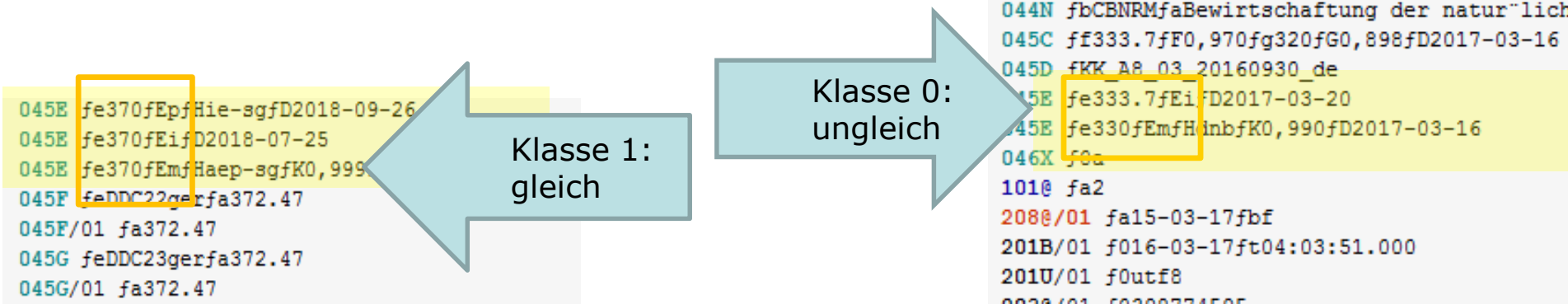
Vorhersage einer richtigen/falschen Klassifikation

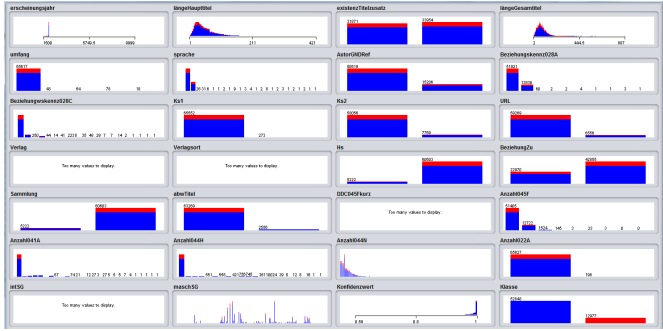
Ermitteln von priorisiert nachzubearbeitenden Fällen

Sind maschinelle Lernverfahren für diese Fragestellung und bibliographische Metadaten geeignet?

Verwendete Daten

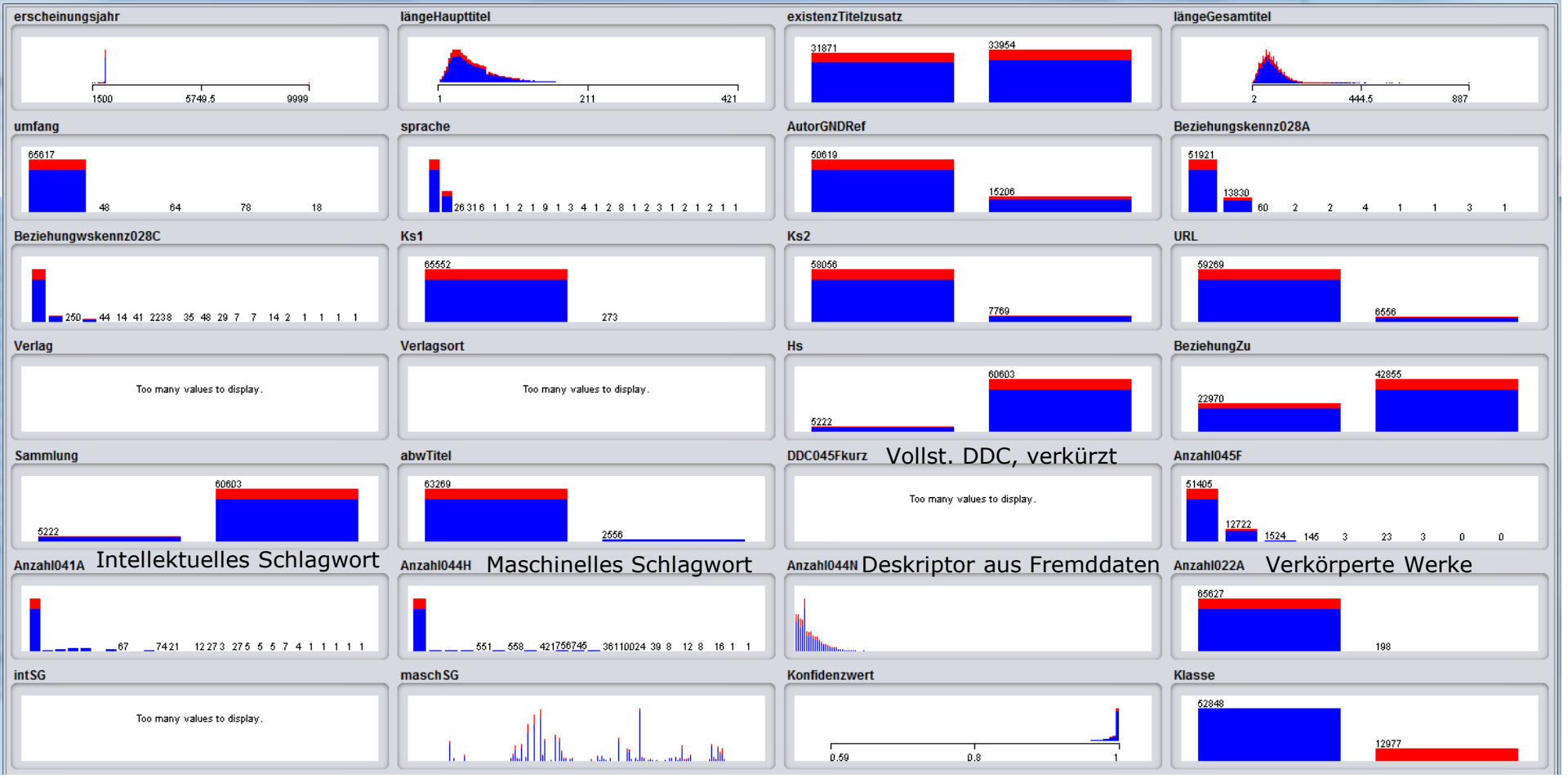
- Set aus 65.825 Netzpublikationen mit intellektuell und maschinell vergebenen Sachgruppen
- Vergleich der intellektuellen und ersten maschinell vergebenen Sachgruppe

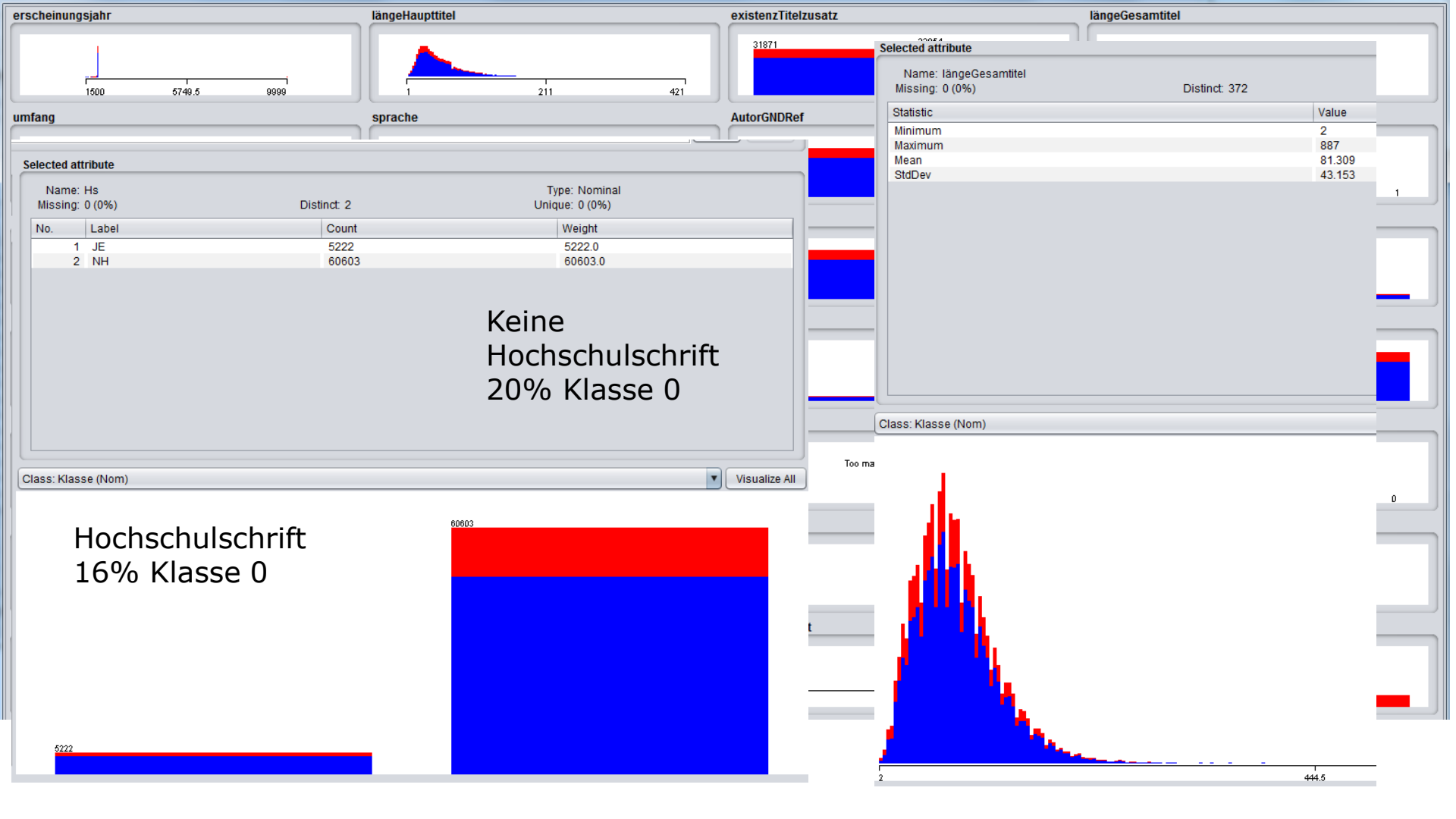




Auswahl von Attributen

- 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)


erscheinungsjahr
1500
5749.5
9999
längeHaupttitel
1
211
421
existenzTitelzusatz
31871
33954
längeGesamtitel
2
444.5
887
umfang
65617
48
64
78
18
sprache
26 31 6 1 1 2 1 9 1 3 4 1 2 8 1 2 3 1 2 1 2 1 1
AutorGNDRef
50619
15206
Beziehungskennz028A
51921
13830
60 2 2 4 1 1 3 1
Beziehungwskennz028C
250 44 14 41 2238 35 48 29 7 7 14 2 1 1 1 1
Ks1
65552
273
Ks2
58056
7769
URL
59269
6556
Verlag
Too many values to display.
Verlagsort
Too many values to display.
Hs
60603
5222
BeziehungZu
42855
22970
Sammlung
60603
5222
abwTitel
63269
2556
DDC045Fkurz
Vollst. DDC, verkürzt
Too many values to display.
Anzahl045F
51405
12722
1524 145 3 23 3 0 0
Anzahl041A
Intellektuelles Schlagwort
67 7421 12273 275 5 5 7 4 1 1 1 1 1
Anzahl044H
Maschinelles Schlagwort
551 558 421756745 36110024 39 8 12 8 16 1 1
Anzahl044N
Deskriptor aus Fremddaten
Anzahl022A
Verkörperte Werke
65627
198
intSG
Too many values to display.
maschSG
Konfidenzwert
0.59
0.8
1
Klasse
52848
12977


erscheinungsjahr
1500
5749.5
9999
längeHaupttitel
1
211
421
existenzTitelzusatz
31871
längeGesamttitel
umfang
sprache
AutorGNDRef
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
Class: Klasse (Nom)
Visualize All
Hochschulschrift 16% Klasse 0
60603
5222
Selected attribute
Name: längeGesamttitel
Missing: 0 (0%)
Distinct: 372
Statistic Value
Minimum 2
Maximum 887
Mean 81.309
StdDev 43.153
Class: Klasse (Nom)
Too ma
2
444.5

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

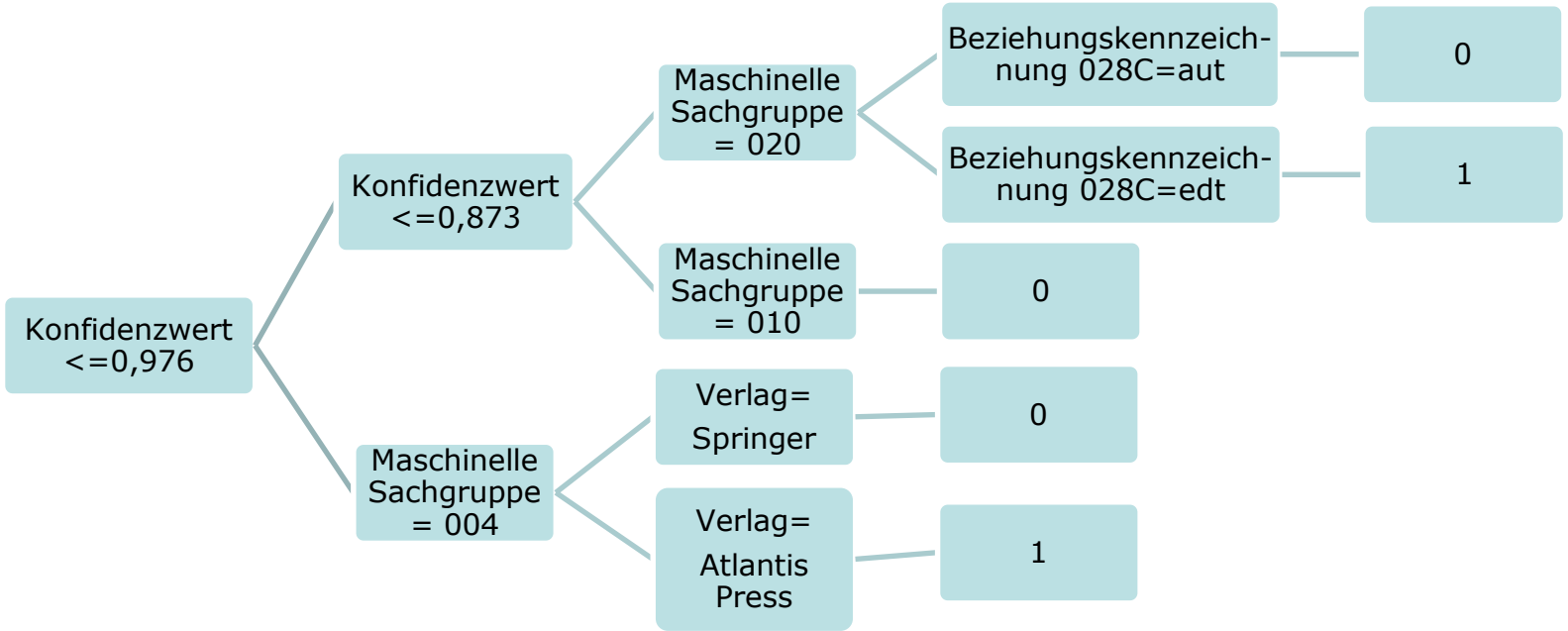
… und unwichtigste Attribute	
0,00002062	Ist eine erste Körperschaft eingetragen?
0,0000161	Existiert ein Titelzusatz?
0,00000134	Anzahl von in der Manifestation verkörperter Werke (022A)

DEUTSCHE NATIONAL BIBLIOTHEK

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



J48 Decision Tree, Daten nach Maschineller Sachgruppenvergabe

DEUTSCHE NATIONAL BIBLIOTHEK

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?

DEUTSCHE NATIONAL BIBLIOTHEK

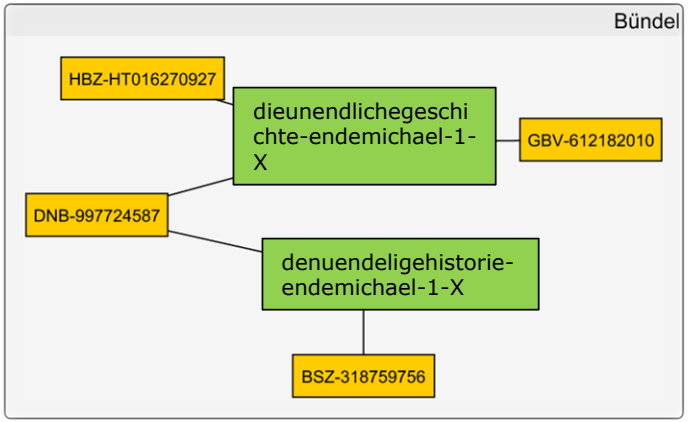
Fazit

- 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 Bibliotheksverbünde, des Österreichischen Bibliothekenverbunds 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

Clusterung 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" "kantsmoralreligion"
"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 ≠ 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

"114898" "title" "8" "0.974" "0.5"
"solidaritätderkirchemitisraeldietheologi[illegible]ungd
esverhältnissesderkirchezumjudentumna[illegible]anhandderoff
iziellenverlautbarungen"
"114898" "creator" "8" "0.0" "1.0" "Wirth, Wolfgang"
"114898" "language" "8" "0.0" "1.0" "ger"

Abkürzungen im Titel sorgen für Ungleichheit

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

DEUTSCHE NATIONAL BIBLIOTHEK

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.

DEUTSCHE NATIONAL BIBLIOTHEK

Vielen Dank für Ihre Aufmerksamkeit.

Fragen?

Kontakt: Angela Vorndran
a.vorndran@dnb.de

Stefan Grund
s.grund@dnb.de