Automatisierung der Sacherschließung (AutoSE)

Moritz Fürneisen, Anna Kasprzik
ZBW – Leibniz-Informationszentrum Wirtschaft
Frankfurt am Main, 11.10.2019

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Die ZBW ist Mitglied der Leibniz-Gemeinschaft.

Timeline: Automatisierung der Sacherschließung an der ZBW

- 2002 – 2004: DFG-Projekt AUTINDEX, mit der Universität des Saarlandes
  - ✓ Ergebnis: erster Prototyp für halbautomatisierte Sacherschließung
- 2009 – 2011: Projekt zur Evaluierung kommerzieller Software-Lösungen;
  - ✓ Wahl: *Decisiv Categorization* von *Recommind* (statistischer Ansatz, PLSA)
- 2012 – 2014: Reorientierungsphase
  - ✓ Formulierung von Anforderungen für den Gebrauch in der Praxis
- 2014 – 2018: Projekt AutoIndex – *do it yourself / Open Source…*
  - ✓ Ergebnis: Prototyp mit Fusion-Lösung, drei Datenreleases
- 2019: ~~Projekt~~ AutoSE – Neuanfang auf Basis bestehender Ergebnisse und Ziele

Wer sind wir, was wollen wir?

Wer?

1 Leitung, 1 wissenschaftl. Mitarbeiter, 1 Software-Entwickler (ab 2020), (1 Hiwi), 0.2 Thesaurusmanager [+ bibl. Bewertungsgruppe]

Warum?

Ziel: so viele Abläufe wie möglich im Erschließungsworkflow automatisieren

und dabei die Qualität der von der ZBW generierten Metadaten erhalten

Problemstellung:

Input – bibliographische Daten (Titel, Keywords, Abstract, Textkörper, ToC, …)

Output – eine Menge von Deskriptoren aus dem STW

Bis 2018 (Projekt AutoIndex)

forschungsbasierte Entwicklung eines Fusion-Ansatzes:

- Kombination mehrerer Machine-Learning-Methoden auf der Basis bestehender Open-Source-Lösungen
- wissenschaftliche Weiterentwicklung, z.B. zur Bewältigung von *concept drift* (dynamische Daten) oder für eine automatisierte Qualitätsabschätzung
- Grundlage: *short text* (Titel, Keywords, (Abstracts))
- Vokabularbasis: Standardthesaurus Wirtschaft (STW)

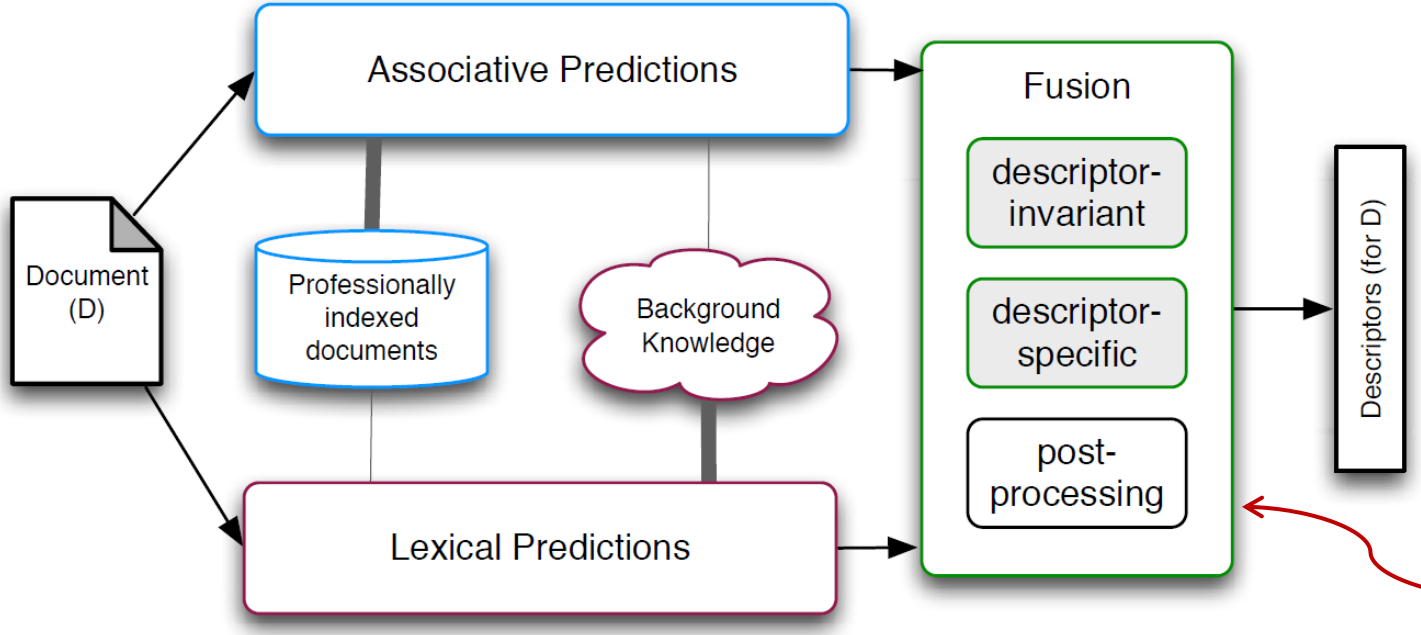
Lexikalische und assoziative Machine-Learning-Methoden

Im Fusion-Ansatz werden 2 Arten von Maschine-Learning-Methoden eingesetzt:

- **(statistische) assoziative Methoden:**
  verwenden statistische *features*, die aus den Wortvorkommen abgeleitet sind
- **lexikalische Methoden:**
  verwenden die Information aus dem Thesaurus und *features* wie Position des ersten Vorkommens, *tf-idf*, Länge

Toepfer, M. & Seifert, C. „Fusion architectures for automatic subject indexing under concept drift“ Int J Digit Libr (2018). https://doi.org/10.1007/s00799-018-0240-3

Fusion-Ansatz – schematische Darstellung



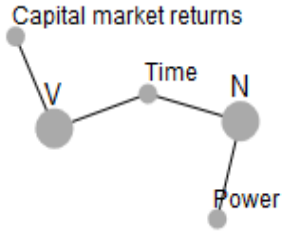
gesteuert durch zusätzliche Regeln wie z.B. „min2“ und „2ausVB“

Intellektuelle Evaluierung – *„releasetool“*

Title: **Improved calendar time approach for measuring long-run anomalies**

Keywords: long-run anomalies | standardized abnormal returns | test specification | power of test

Abstract: Although a large number of recent studies employ the buy-and-hold abnormal return (BHAR) methodology and the calendar time portfolio approach to investigate the long-run anomalies, each of the methods is a subject to criticisms. In this paper, we show that a recently introduced calendar time methodology, known as Standardized Calendar Time Approach (SCTA), controls well for heteroscedasticity problem which occurs in calendar time methodology due to varying portfolio compositions. In addition, we document that SCTA has higher power than the BHAR methodology and the Fama-French three-factor model while detecting the long-run abnormal stock returns. Moreover, when investigating the long-term performance of Canadian initial public offerings, we report that the market period (i.e. the hot and cold period markets) does not have any significant impact on calendar time abnormal returns based on SCTA.



Automatically Assigned Subjects

(explain)

Rating				Subject	Categories
--	0	+	++		
☐	☐	☐	☐	Power	N
☐	☐	☐	☐	Time	V N
☐	☐	☐	☐	Capital market returns	V

Document-level Quality

- ☐ good
- ☐ fair
- ☐ reject
- ☐ skip

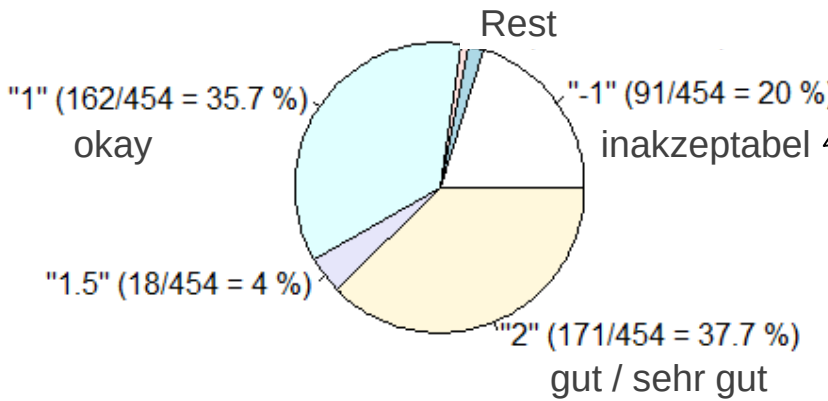
Missing Subjects

Add Missing Subject

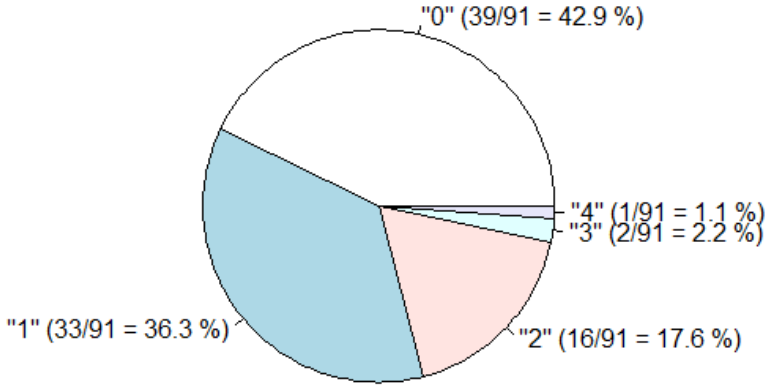
Submit

Intellektuelle Evaluierung des letzten Datenrelease

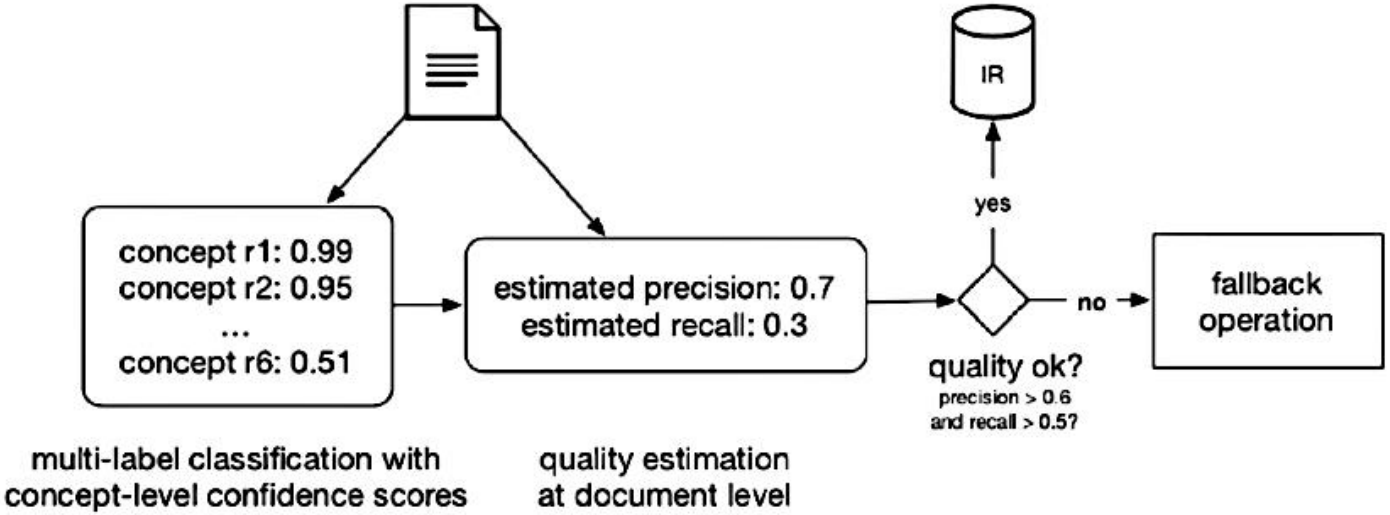
auf Deskriptorebene:



Anzahl schädlicher Deskriptoren
in abgelehnten Dokumenten:



Automatisierte Qualitätsabschätzung



Toepfer, M. & Seifert, C. „Content-Based Quality Estimation for Automatic Subject Indexing of Short Texts under Precision and Recall Constraints“ Proc. of TPDL (2018)

Automatisierte Qualitätsabschätzung

Category	Symbol	Description
Volume	#_Char	Number of characters (incl. white-space)
Volume	#_WS	Number of whitespace characters
Content	$TERM_i$	Variables for vocabulary terms (binary or numeric)
Content	#_W_OOV	Number of unknown terms
Content	#_SPECIAL	Number of special characters, e.g. "?"

Toepfer, M. & Seifert, C. „Content-Based Quality Estimation for Automatic Subject Indexing of Short Texts under Precision and Recall Constraints" Proc. of TPDL (2018)

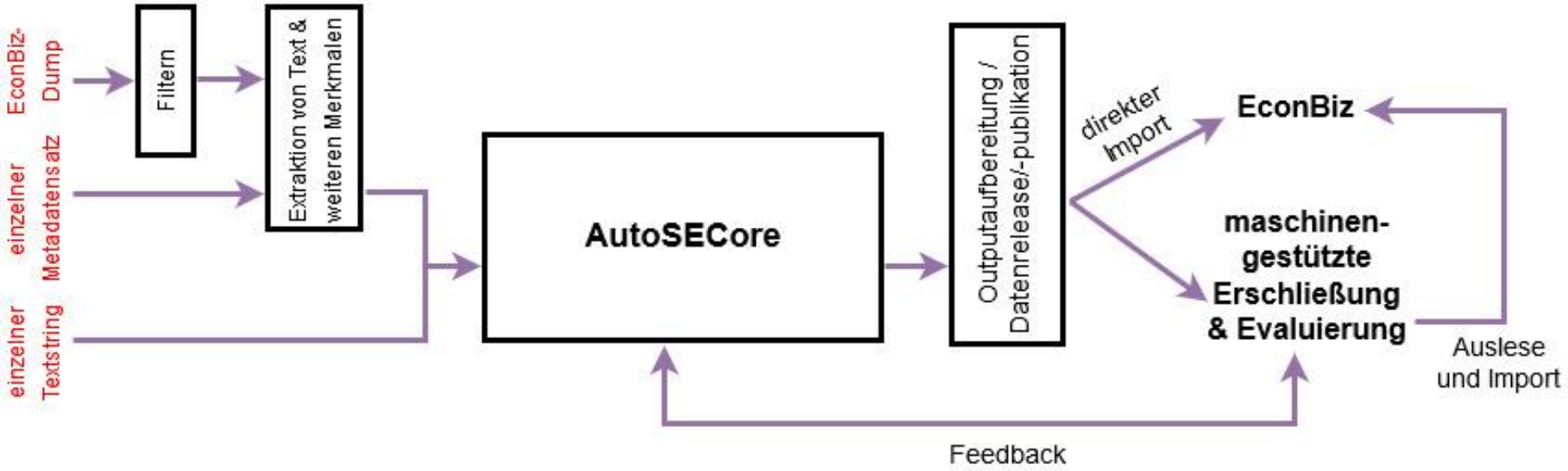
Was machen wir jetzt? – Transfer in die Praxis

Planung des Aufbaus einer Software-Architektur für die permanente Integration unserer Machine-Learning-Lösungen in den Erschließungsbetrieb

- Evaluierung des Open-Source-Tools Annif* als Plattform für unsere Backends
- Evaluierung unserer eigenen Prototypen auf ihre Eignung für die Bereitstellung als Dienst, Optimierung der Parameter
- Optionen: sowohl maschinengestützte als auch vollautomatisierte Erschließung soll möglich sein

* Annif: Finnische Nationalbibliothek, Osma Suominen; http://annif.org/

High-Level-Darstellung der geplanten Zielarchitektur



Was machen wir jetzt? – Wissenschaftliche Weiterentwicklung

- Evaluierung von Ansätzen mit neuronalen Netzen, Deep Learning
- Identifizieren und Weiterentwicklung von Möglichkeiten zur automatisierten Qualitätsabschätzung (inkl. Active Learning)

perspektivisch: automatisierte

- Integration externer Informationsquellen nach Semantic-Web-Prinzipien, z.B. aus der LOD-Cloud
- Extraktion von Termen als Thesauruskandidaten
- Extraktion und Auslese von Strukturelementen aus Textdokumenten (→ Formalerschließung)

Frühe Arbeiten zu Erschließung mit neuronalen Netzen @ ZBW:
Galke et al. 2017 https://arxiv.org/abs/1705.05311 Mai, Galke & Scherp 2018 https://arxiv.org/abs/1801.06717

Danke!

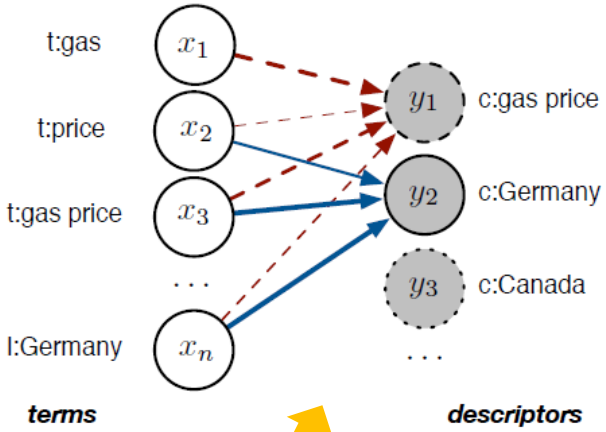
Kontakt:

- Moritz Fürneisen
  m.fuerneisen@zbw.eu
  Tel.: 040 42834-366
- Dr. Anna Kasprzik
  a.kasprzik@zbw.eu
  Tel.: 040 42834-425

Backup-Folien, zur Erläuterung

Lexikalische und assoziative Machine-Learning-Methoden

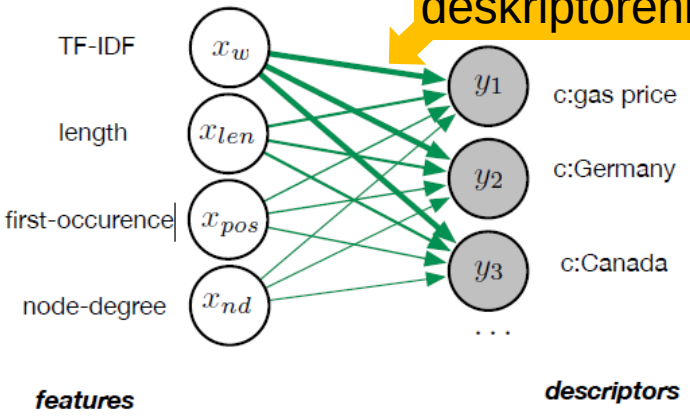
assoziativer Ansatz



deskriptorenspezifische Parameter!

lexikalischer Ansatz

deskriptoreninvariant!



Lexikalische und assoziative Machine-Learning-Methoden

… haben jeweils Stärken und Schwächen:

Aspekt	lexikalisch	assoziativ
benötigte Menge Trainingsdaten	++	–
Erkennen unbekannter Konzepte	++	– –
Erkennen von Synonymen	– –	++
Ambiguität	o	+
Ausnutzen von Thesaurusrelationen	+	o

➢ deshalb müssen wir sie kombinieren!

Toepfer, M. & Seifert, C. „Fusion architectures for automatic subject indexing under concept drift“ Int J Digit Libr (2018). https://doi.org/10.1007/s00799-018-0240-3

Deskriptorenspezifische vs deskriptoreninvariante Methoden

- eine **deskriptorenspezifische** Methode lernt die Bedingungen für ein spezifisches Konzept (z.B. „Deutschland“), kann dieses Wissen aber nicht auf ähnliche Konzepte (z.B. „Kanada“) übertragen
- eine **deskriptoreninvariante** Methode kann generalisieren und ihr Wissen auf Konzepte anwenden, die nicht in den Trainingsdaten vorgekommen sind

➢ wichtig, da wir generell ein Problem mit dem Mangel an Trainingsdaten haben!

Toepfer, M. & Seifert, C. „Fusion architectures for automatic subject indexing under concept drift“ Int J Digit Libr (2018). https://doi.org/10.1007/s00799-018-0240-3

Project AutoIndex – what has been done so far?

development/optimization of various classifiers

(based on existing lexical and associative methods)

lexical

associative

based on n-gram features

quadflor

dict

kNN

BRLR

maui

mausi

zaptain
stwlearn

BRSVM

monq

stwfsa

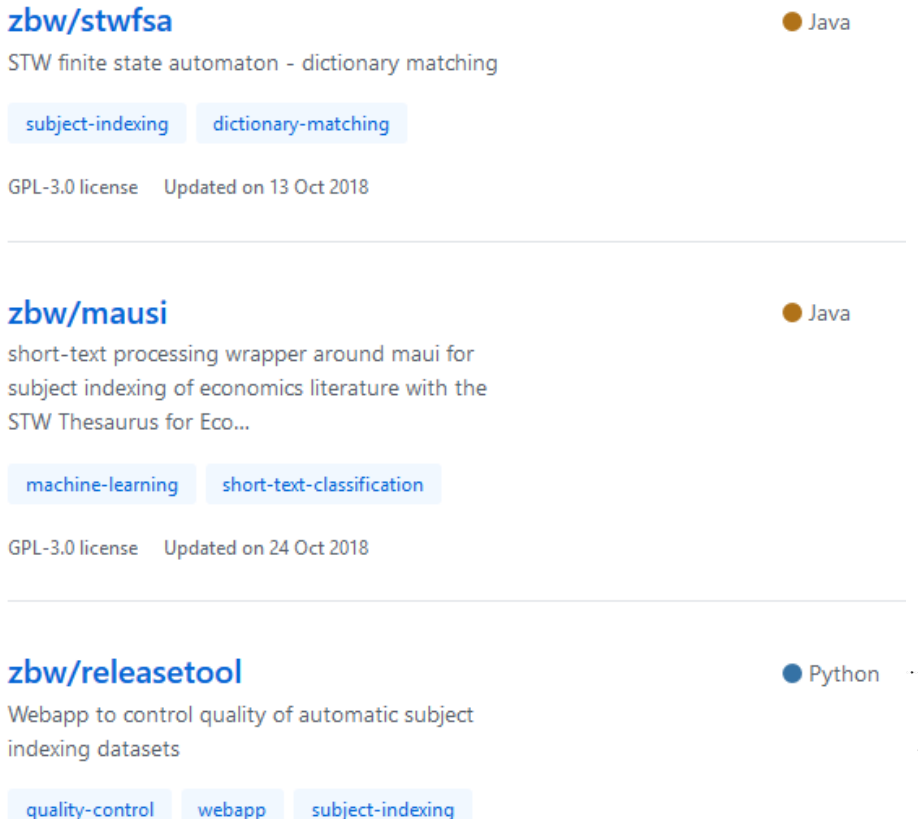
Rhack

...

https://github.com/zelandiya/maui https://github.com/HaraldKi/monqjfa

Project AutoIndex – what has been done so far?

- languages / frameworks: Java, Python, JavaScript; Django; *gitolite*; MariaDB
- agile development
- unit tests
- two servers:
  - *dwalin* – hardware (CentOS; *releasetool*)
  - *gloin* – virtual (CentOS; development)

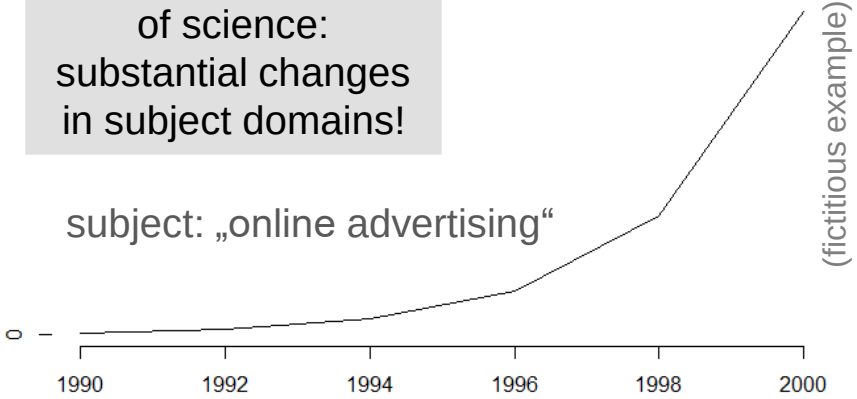


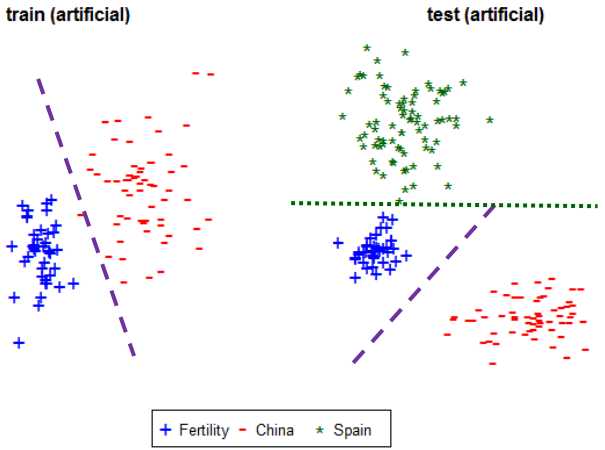
Challenge: Concept Drift

distributions of **terms and concepts differ between training and test data**

extreme case: concepts totally **miss training data**

inherent property of science: substantial changes in subject domains!





Experiments (Toepfer & Seifert 2017/18)

compare: **lexical** vs. **associative** vs. **fusion** [+ postprocessing]

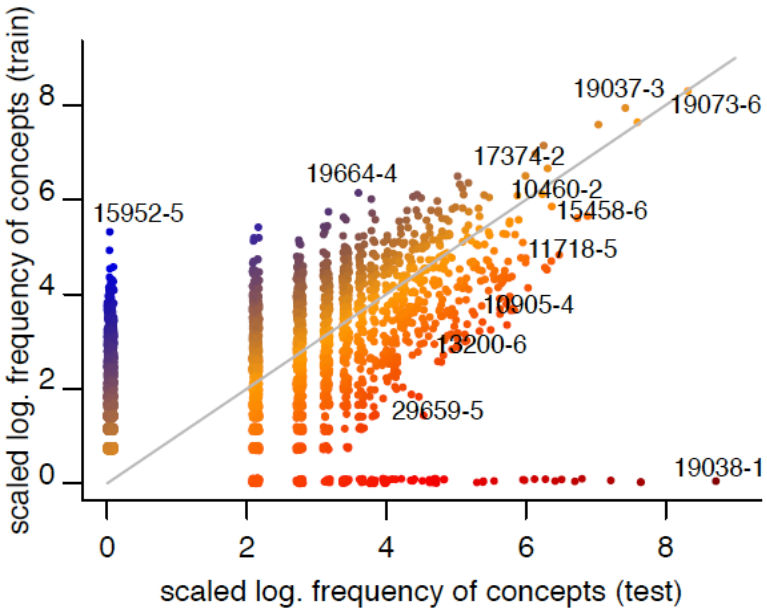
- precision, recall, F1?
- robust to concept-drift?

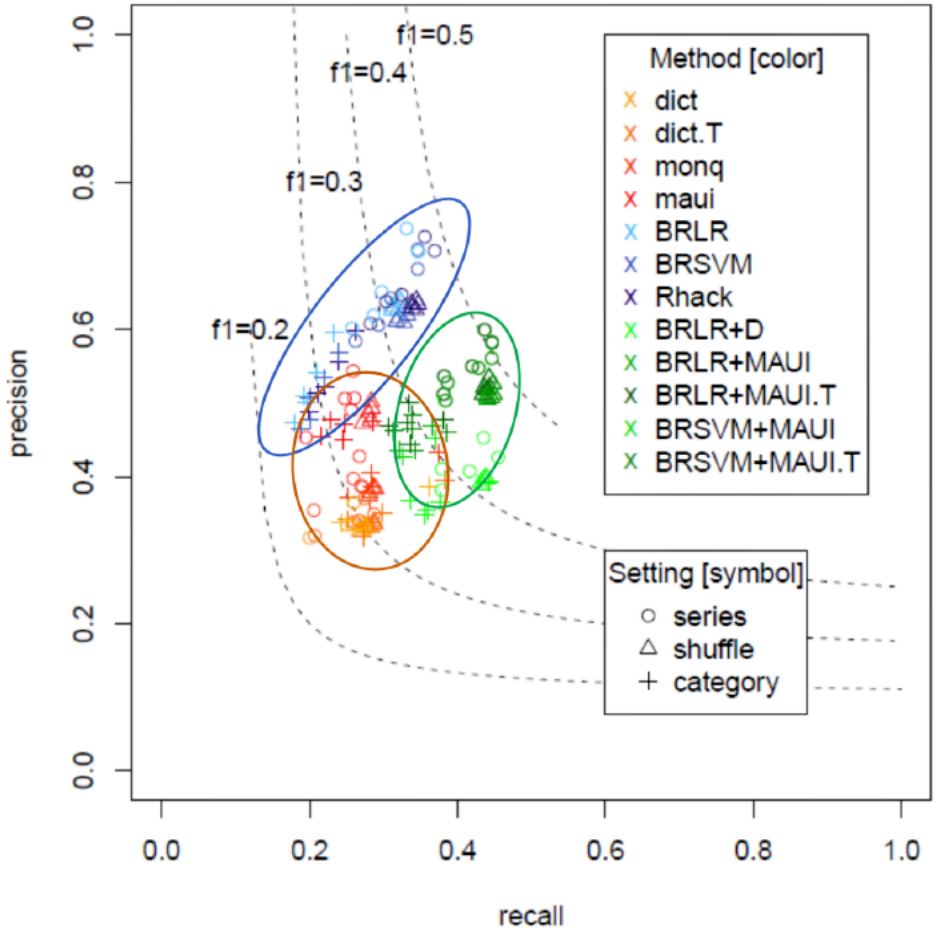
data:

- content: title & author keyword fields concatenated
- in total: 20,195 documents
- 3 settings, with 5 folds for each case:
  - shuffle: **random** subsets/folds of all documents
  - **explicit concept drift**: with/without certain concept categories
  - **implicit concept drift**: with/without certain working paper series

Results

- Fusion systems outperformed lexical and associative approaches in terms of F1
- Hard because of concept drift





Slide by: Christin Seifert

Martin Toepfer, Christin Seifert:
Content-Based Quality Estimation for Automatic Subject Indexing of Short Texts Under Precision and Recall Constraints. TPDL 2018: 3-15

Martin Toepfer, Christin Seifert:
Content-Based Quality Estimation for Automatic Subject Indexing of Short Texts under Precision and Recall Constraints. CoRR abs/1806.02743 (2018)

Martin Toepfer, Christin Seifert:
Towards Semantic Quality Control of Automatic Subject Indexing. TPDL 2017: 616-619

Martin Toepfer:
Machine Learning Architectures for Scalable and Reliable Subject Indexing - Fusion, Knowledge Transfer, and Confidence. TPDL 2017: 644-647

Martin Toepfer, Christin Seifert:
Descriptor-Invariant Fusion Architectures for Automatic Subject Indexing. JCDL 2017: 31-40