# Automatisierte Erstellung von Pressedossiers durch Textmining

## Kontextualierung im journalistischen Umfeld

Marcel Schöneberg

marcel.schoeneberg@haw-hamburg.de

Hochschule für Angewandte Wissenschaften Hamburg (HAW)
Fakultät für Technik und Informatik
Department Informatik

Masterseminar Präsentation

16.12.2014

# Zielsetzung

- Ursprüngliche Idee: Trenderkennung in Social Media
- Ziele von Projekt 1:
  - Datenbasis schaffen
  - Vorverarbeitungen durchführen und verstehen
  - (Hierarchisches) Clustering als Experiment
    - Daten kennenlernen
    - These: 'Ausgewachsene' Trends/Themen müssten einen Cluster bilden
    - Untercluster zeigen ggf. kleinere Trends auf
    - Ggf. als Basis für spätere Zeitreihenanalysen nutzen

[Sch14]

# Erfolge

- Vorhandene Datenbasis mit über 27000 Tweets zum Hashtag-Piraten
- Zeitraum: 11.2011 - 05.2014
- Erfahrungen mit Vorverarbeitung (z.B. Normalisierung, Stemming, NLP)
- Clustering mit verschiedenen Clustermengen und unterschiedlichen Untermengen des Korpus
- Diverse Stolpersteine kennengelernt

# Misserfolge - Begründung für den Fokuswechsel

- Weitgehend nutzlose Resultate
- Cluster zumeist eher Sammlung aus unzusammenhängenden Begriffen (ggf. 'Zufallstreffer')
- Auch Nutzung eines auf Nomen basierenden Korpus nicht erfolgversprechend
- Texte grundlegend zu kurz
- 'schlechte' Sprache (Abkürzungen usw.)
- Clustering zu undurchsichtig
- Trenderkennung ist ein schwammiges Ziel

→ Durchaus möglich [1] aber u.A: komplexe **Vorverarbeitung** etc. notwendig

⟶ Keine gute Ausgangsbasis

[1] Siehe Fazit von [Sch14]

# Einführung in die Thematik I

- Ziel: Erstellung von Dossiers $\rightarrow$ Kontextualisierung eines Artikels
- Fokus auf textbasierten Dossiers: Keine multimedialen Inhalte usw.
- Dossiers: **Verschiedene** (Experten-) **Definitionen** vorhanden [Hä14]
  - Arbeitsthese: Zusammenstellung von 'ähnlichen' Artikeln

  $\rightarrow$ Inhaltliche / semantische Nähe

  $\longrightarrow$ **Erster** Schritt: Nicht für alles geeignet $\rightarrow$ Abbildung 1
- Ähnlichkeit als Basis $\rightarrow$ **Distanzfunktionen**

$\longrightarrow$ (Zunächst) möglichst verständlich halten

# Einführung in die Thematik II

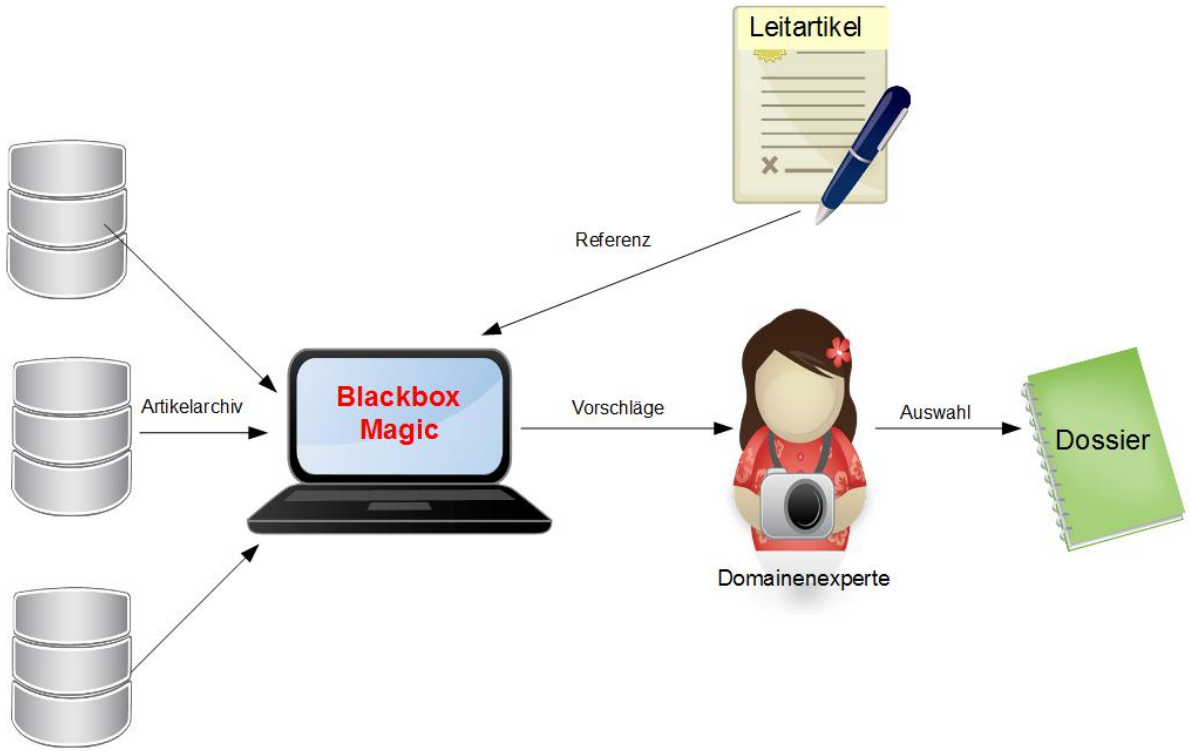



Abbildung: Realisierbarer Workflow

# Einführung in die Thematik III

- Artikelarchiv: nach **Sortierung** ca. 2700 Artikel (Englisch)
  - von Journalisten
  - Basierend auf dem Eurozine Netzwerk `www.eurozine.com`
  - Größere Länge als Tweets
  - Meta-Informationen vorhanden (Verlinkungen auf Inhaltsverzeichnisse, redaktionelle Focalpoints etc.) $\rightarrow$ Sinnvoll nutzbar?
  - semi-strukturiert in XML (Autor, Abstract, Überschriften usw.)

# Einführung in die Thematik IV

- *Beispiel*
- Artikelarchiv hat ebenfalls Probleme:
  - Nicht alle englischsprachig
  - Enthält auch Zusammenfassungen etc.
  - z.T. Gedichte (recht kurz, sprachlich besonders)
  - Invalides XML

# Fragestellungen und Ziele

- Automatisierbare Dossiererstellung möglich?
- Was sind Pressedossiers bzw. wie werden diese benutzt?
- Lassen sich fachliche Anforderungen in einen Algorithmus überführen?
- Sind journalistische Definitionen eines Dossiers mit Informatik umsetzbar?
- Wie kann man ein generiertes Ergebnisses bewerten?

$\longrightarrow$ Versuch den Algorithmus übersichtlich zu halten

# Erste Ideen I

- Viele Diskussionen und Ideen
- Simple Distanzfunktionen nutzen
- Facetten / abweichende Sichtweisen finden
- Vorhandene Clustering Techniken benutzen
- Vorschlagssysteme aufgrund von Lernalgorithmen / Crowd
- Metainformationen (Autoren, Herkunft) nutzen $\rightarrow$ besserer Überblick
- Ggf. Distanzfunktionen verbessern

# Erste Ideen II

→ Zu einfach gedacht

- Geht nicht auf **fachliche Anforderungen** ein
- Clusteringverfahren oft sehr **undurchsichtig** (vgl. Projekt 1)
- Vorschlagssysteme brauchen (oft) eine Basis - Nicht vorhanden → **Coldstart problem**
- Viele Systeme setzen auf Personalisierung oder die Crowd → Ziel: **automatisierte und neutrale** Ergebnisse (z.B. [PL10], [KFD12] , [LCLS10])
- Ideen → 'Den zweiten Schritt vor dem ersten tun'
- Erlernen von Ähnlichkeit nicht im Fokus (Neuronale Netze usw.)
  - Nicht ohne Weiteres anpassbar
  - Nicht als erster Schritt

# Related Work

- Forschungsprojekt der DPA
- Overview Project http://overview.ap.org/
- XML Ähnlichkeiten [TCY09]
- Distanzfunktionen [HRJM13]
- Content-based Recommender Systems [LGS]
- Überblick über Recommender Systems in der Breite [BOHG13]
- Article Recommender Systems [BDD+12]

# Konkrete Schritte I

- Ziel ist höchst experimentell $\rightarrow$ Schrittweise **annähern und verbessern**
- Verbesserung der Distanzfunktion zwischen Dokumenten
- Bag of Words-Ansatz
- Ausgangsbasis: **der** Leitartikel – später ggf. gewichtete Wortmenge

# Konkrete Schritte II

- **Gewichtete** Wertungen von 'wichtigeren' Teilen → Gute Werte noch zu ermitteln
  1. Abstract
  2. Title
  3. Subheadings
  4. Text
- Anreicherung mit themenspezifischen **Kategorien**
  - Wortmengen pro Thema (z.B. Wirtschaft, Politik usw.) einbringen
  - Gewichtung des 'Bag of words' zugunsten der (hierarchischen) Kategorien
- **Erweiterung** durch Erkenntnisse einer Fachexpertin

# Distanzfunktionen

- Vieles hat wurde im Rahmen des Seminars schon angesprochen [Sip14]
- Zunächst Nutzung einer einfachen euklidischen Distanz (auf Basis der Gewichtungen)
- Verschiedene 'Relevanzmaße' für Termhäufigkeiten (Termvorkommen, TF-IDF)
- Diverse Kriterien von Distanz möglich (Relevanz, Themenbreite, Neuigkeit etc.)

$\rightarrow$ Diverse Funktionen und Möglichkeiten vorhanden

# Weitere mögliche Schritte

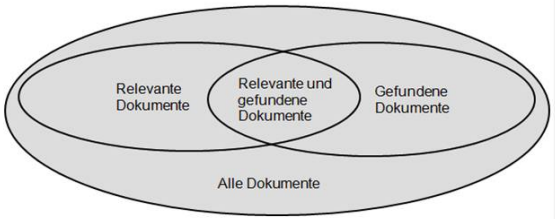
- Autoren nutzen um Herkunft des Artikels zu bestimmen
- Ggf. Einbeziehung von Inhaltsverzeichnissen der (Themen-)Zeitschriften
- Linguistische Verbesserungen (z.B. Nutzung von Ontologien etc.)
- Erkennung von Facetten eines Themas (europäische Sicht auf Ursprungsartikel)

# Bewertung von Ergebnissen

- Verifikation der Ergebnisse über Focalpoints-Zusammenfassungen
- Testkorpus als Verifikationsbasis:
  - Focalpoint-Artikel (redaktionell eingeordnet)
  - (fast) zufällige Artikel (nicht im Focalpoint)
  - → Berechnete Dossierartikel sollten möglichst den Focalpointartikeln entsprechen

→ Nutzung von Eigenschaften wie:

- Recall (Trefferquote): $\frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$
- Precision (Genauigkeit): $\frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$



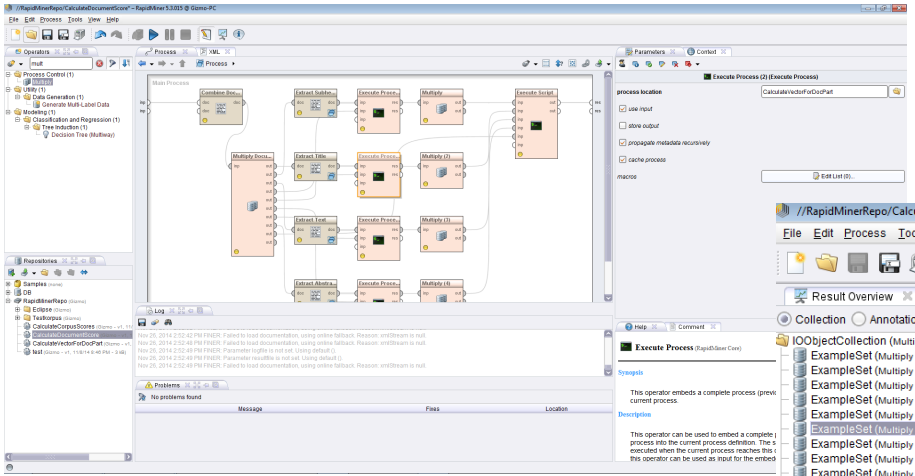

[PR-]

- Wissen einer Domänenexpertin

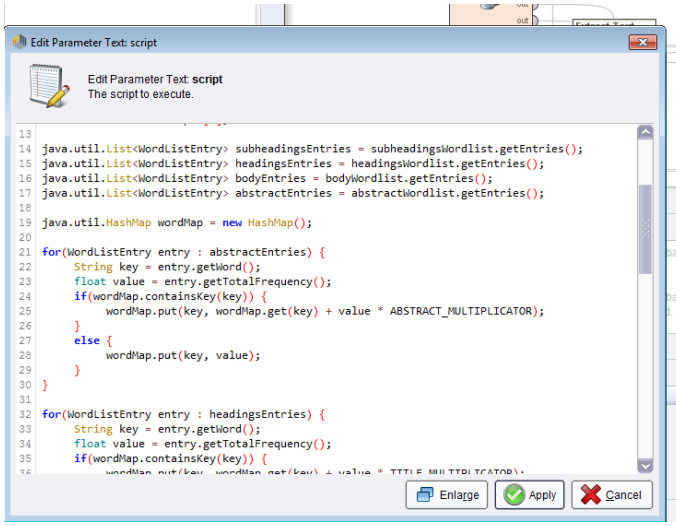
# Aktueller Fortschritt: Hauptprojekt I

- Aufgesetzter Testkorpus (Focalpoint Artikel + beliebige Artikel als Vergleichsbasis)
- Rapidminer [2] Verständnis und Hürden
- KDD Preprocessing
- Erste Versuche lauffähig
- Domänenübergreifendes Projekt $\rightarrow$ Grundlegendes gemeinsames Verständnis

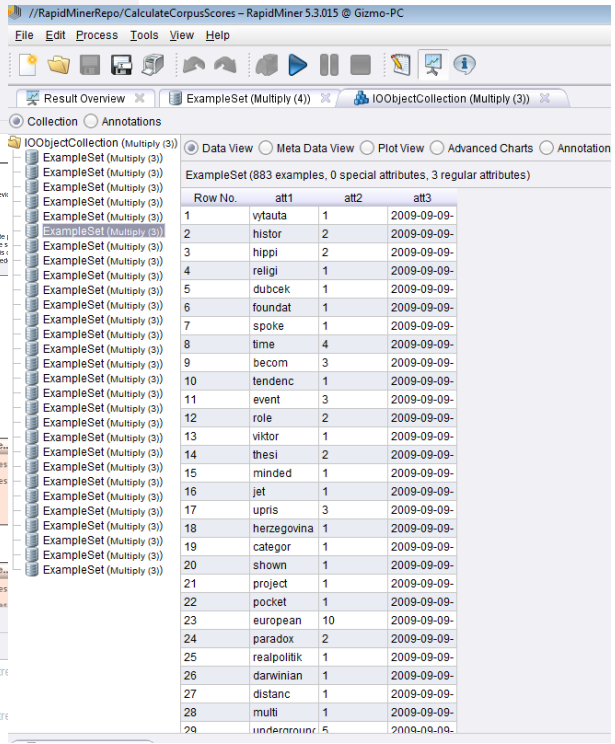
# Aktueller Fortschritt: Hauptprojekt II



(a) Process




(b) Script




(c) Resultwindow

[2] https://rapidminer.com/

# Chancen und Risiken I

- **Chancen**
  - Viel Interesse - 'Editors don't scale'
    - HAW NextMedia [3] gefördert durch BMBF
  - DPA Forschungsprojekt
- **Risiken**
  - Zu abweichende Meinungen von 'Was ist ein Dossier'
  - Gefahr das Experten ihr Wissen nicht weitergeben können - 'Habe ich im Gefühl' [McD83]

  $\rightarrow$ Definitionen zu komplex oder zu unspezifisch um sie automatisiert umzusetzen
  - Aufeinander aufbauender Prozess - wenn eins schief geht $\rightarrow$ Problem

[3] http://nextmedia-haw.de/

# Fragen

Vielen Dank für die Aufmerksamkeit!

Fragen?

# Literature I


Bancu, Cristian ; Dagadita, Monica ; Dascalu, Mihai ; Dobre, Ciprian ; Trausan-Matu, Stefan ; Florea, Adina M.:
ARSYS – Article Recommender System.
In: **Proceedings of the 2012 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing**.
Washington, DC, USA : IEEE Computer Society, 2012 (SYNASC '12). –
ISBN 978–0–7695–4934–7, 349–355

Bobadilla, J. ; Ortega, F. ; Hernando, A. ; GutiéRrez, A.:
Recommender Systems Survey.
In: **Know.-Based Syst.** 46 (2013), Juli, 109–132.
http://dx.doi.org/10.1016/j.knosys.2013.03.012. –
DOI 10.1016/j.knosys.2013.03.012. –
ISSN 0950–7051

Hälker, Nina:
**Dienen textminingbasierte Dossiers dem Wachsen einer gemeinsamen europäischen Erzählung?**
2014. –
Arbeitspapier

Harispe, Sébastien ; Ranwez, Sylvie ; Janaqi, Stefan ; Montmain, Jacky:
Semantic Measures for the Comparison of Units of Language, Concepts or Entities from Text and Knowledge Base Analysis.
In: **CoRR** abs/1310.1285 (2013).
http://arxiv.org/abs/1310.1285

# Literature II


Kirshenbaum, Evan ; Forman, George ; Dugan, Michael:
A Live Comparison of Methods for Personalized Article Recommendation at Forbes.Com.
In: **Proceedings of the 2012 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II**.
Berlin, Heidelberg : Springer-Verlag, 2012 (ECML PKDD'12). –
ISBN 978–3–642–33485–6, 51–66

Li, Lihong ; Chu, Wei ; Langford, John ; Schapire, Robert E.:
A Contextual-bandit Approach to Personalized News Article Recommendation.
In: **Proceedings of the 19th International Conference on World Wide Web**.
New York, NY, USA : ACM, 2010 (WWW '10). –
ISBN 978–1–60558–799–8, 661–670

Lops, Pasquale ; Gemmis, Marco de ; Semeraro, Giovanni:
In: **Recommender Systems Handbook**

McDermott, John P.:
Extracting Knowledge From Expert Systems.
In: Bundy, Alan (Hrsg.): **IJCAI**, William Kaufmann, 1983, 100-107

Peng, Chi-Chieh ; Liu, Duen-Ren:
Combining Reputation and Content-based Filtering for Blog Article Recommendation in Social Bookmarking Websites.
In: **Proceedings of the 12th International Conference on Electronic Commerce: Roadmap for the Future of Electronic Business**.
New York, NY, USA : ACM, 2010 (ICEC '10). –
ISBN 978–1–4503–1427–5, 8–14

**Precision-Recall diagram**.
http://wikis.gm.fh-koeln.de/wiki_ir/uploads/InformationRetrieval/Recall/ir_bild.jpg

# Literature III


Schöneberg, Marcel:
Erkennung von Trends in sozialen Netzwerken.
Version: 10 2014.
http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2014-proj/schoeneberg.pdf.
2014. –
Projektbericht

Sippel, Sigurd:
Recommendations for cocktail recipes.
(2014), 11.
http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2014-sem/sippel/folien.pdf

Tekli, Joe ; Chbeir, Richard ; Yetongnon, Kokou:
Survey: An Overview on XML Similarity: Background, Current Trends and Future Directions.
In: **Comput. Sci. Rev.** 3 (2009), August, Nr. 3, 151–173.
http://dx.doi.org/10.1016/j.cosrev.2009.03.001. –
DOI 10.1016/j.cosrev.2009.03.001. –
ISSN 1574–0137