

Retrieval-Modelle zum Filtern, Ranken und Zusammenfassen von Web-Kommentaren

Vortrag zur Bachelorarbeit von
Fabian Loose
Steffen Becker

Bauhaus-Universität Weimar

Überblick

1. Einleitung

1.1 Kommentar-Retrieval

1.2 Retrieval-Aufgaben

2. Retrieval-Modelle

2.1 Kommentarqualität

2.2 Thematische Relevanz

2.3 Meinungsanalyse

3. Zusammenfassung/Diskussion

1.1 Kommentar-Retrieval

- ▶ Kommentare = Benutzergenerierte Inhalte zu einem geg. Thema

- ▶ Probleme bei hunderten bis tausenden Kommentaren:
 - ▶ Alle zu lesen sehr zeitaufwendig bzw. unmöglich
 - ▶ Nicht alle Kommentare lesenswert
 - ▶ Inhaltliche Wiederholungen

- ▶ Information-Retrieval:
 - ▶ Große Zahl von Kommentaren vs. Informationsbedürfnis
 - ▶ Retrieval-Modell:
 - ▶ formale Repräsentation der Kommentare
 - ▶ Bewertungsfunktion bezüglich Relevanz

1.2 Retrieval-Aufgaben

▶ Filtern:

- ▶ Entfernen unerwünschter Kommentare
- ▶ Relevanzkriterium: Kommentarqualität

▶ Ranken:

- ▶ Sortierung der Kommentare nach Relevanz zum geg. Thema
- ▶ Relevanzkriterium: Thematische Nähe

▶ Zusammenfassen:

- ▶ Kommentare mit gleicher Aussage zusammenfassen
- ▶ Übersicht über Gesamtheit schaffen
- ▶ Relevanzkriterium: Meinungsausdruck

2.1 Kommentarqualität

- ▶ Retrieval-Aufgabe: Filtern unerwünschter Kommentare
- ▶ Retrieval-Modell:
 - ▶ Darstellung als Vektor von Merkmalen
 - ▶ Bewertungsfunktion: binärer Klassifikator (erwünscht/unerwünscht)
- ▶ Merkmale:
 - ▶ Linguistische Stilmerkmale
 - ▶ Vandalismusmerkmale

2.1 Kommentarqualität – Merkmale

- ▶ Beispiel für Stilmerkmale:

- ▶ Dale-Chall Reading Grade Score

- ▶ Anzahl der Terme

- ▶ Beispiel für Vandalismusmerkmale:

- ▶ Anteil vulgärer Wörter an allen Wörtern

- ▶ Komprimierbarkeit eines Textes

2.1 Kommentarqualität – Slashdot Korpus

- ▶ Slashdot: Nachrichtenplattform im Web
- ▶ Viel genutzte Kommentarfunktion
- ▶ Moderation durch Benutzer
- ▶ Zuordnung von Kommentaren zu 8 möglichen Kategorien
- ▶ Punktesystem: -1 bis 5 Punkte (Scores)
- ▶ Korpus: alle Artikel und Kommentare der letzten 2,5 Jahren

2.1 Kommentarqualität – Ergebnisse

▶ Experiment:

- ▶ Trainierter Naive Bayes Klassifikator
- ▶ Zwei Klassen: (Positiv) und (Negativ + „Funny“)
- ▶ Stil- und Vandalismusmerkmale
- ▶ Art der Evaluierung: 10-fold cross-validation

	Precision	Recall	F-Measure
Positiv	0.823	0.810	0.817
Negativ + „Funny“	0.614	0.633	0.623

2.2 Thematische Nähe – Ähnlichkeitsmodelle

- ▶ Retrieval-Aufgabe: Ranken
- ▶ Retrieval-Modell: Klassische Modelle des Text-Retrieval
 - ▶ Darstellung als Vektor von Termen/Konzepten
 - ▶ Bewertungsfunktion: Ähnlichkeitsberechnung
- ▶ Getestete Modelle:
 - ▶ Vektorraummodell (VSM)
 - ▶ Latent Semantic Indexing (LSI)
 - ▶ Explicit Semantic Analysis (ESA)

2.2 Thematische Nähe – Vektorraummodell

- ▶ Repräsentation eines Dokumentes:

- ▶ Vektor von Termgewichten

- ▶ Jede Dimension steht für einen Term

- ▶ Termgewicht typischerweise Häufigkeit (allg. Wichtigkeit) eines Terms

- ▶ Ähnlichkeitsberechnung :

- ▶ Kosinus des Winkels zweier Dokumentvektoren

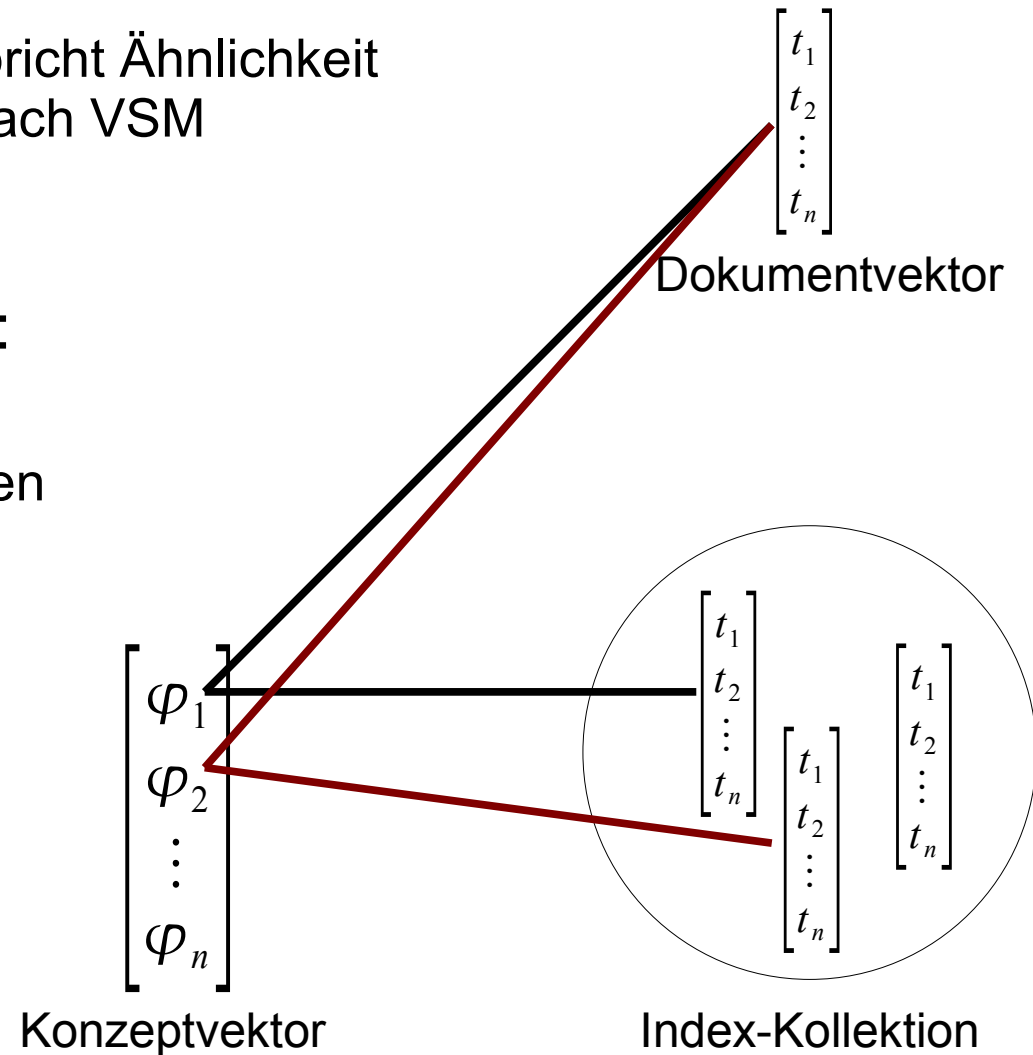
2.2 Thematische Nähe – ESA-Modell

▶ Repräsentation eines Dokumentes:

- ▶ Vektor von Konzepten
- ▶ Jede Dimension entspricht Ähnlichkeit zu Wikipedia-Artikel nach VSM

▶ Ähnlichkeitsberechnung :

- ▶ Kosinus des Winkels zweier Konzeptvektoren



2.2 Thematische Nähe – Kontinuitätsmodell

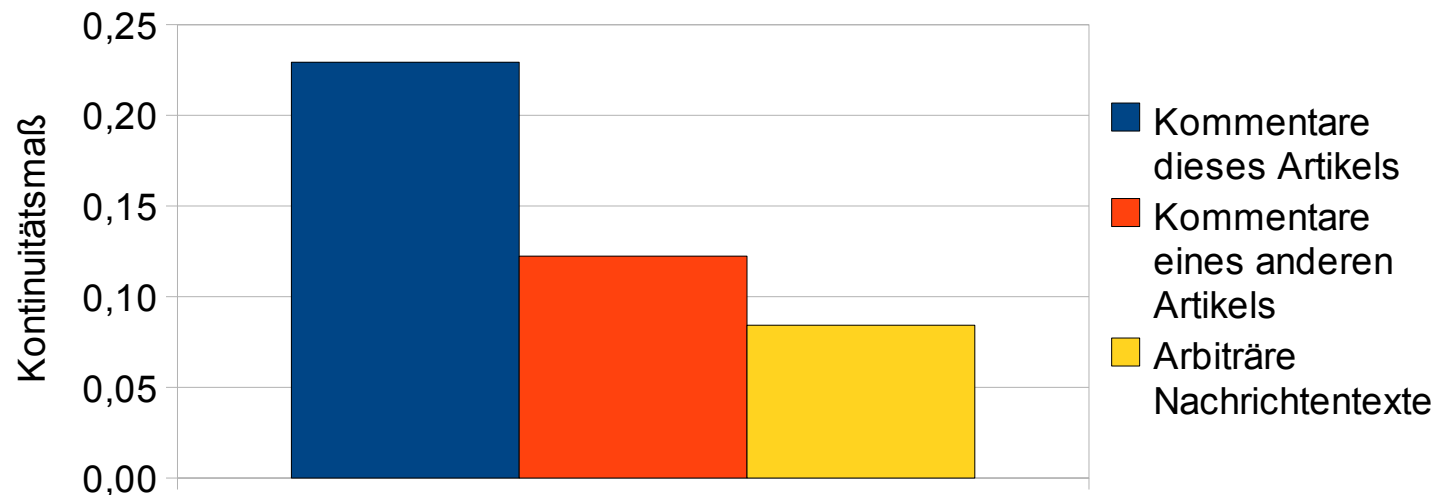
- ▶ Problem der reinen Ähnlichkeitsberechnung:
 - ▶ Artikel selbst wäre der beste Kommentar

- ▶ Idee:
 - ▶ Entfernen aller Terme des Artikels aus den Kommentaren
 - ▶ Ähnlichkeitsberechnung auf Konzeptebene mit ESA

- ▶ Interpretation des Ergebnisses:
 - ▶ Wie stark wird das Thema des Artikels durch den Kommentar ergänzt

2.2 Thematische Nähe – Experiment

- ▶ 10 Slashdot-Artikel
- ▶ Zu jedem Artikel 3 Gruppen von je 4 Kommentaren



2.3 Meinungsanalyse - Einführung

- ▶ Information-Retrieval & Computerlinguistik
- ▶ Zusammensetzung einer Meinungsäußerung in Text:
 - ▶ Meinungsinhaber
 - ▶ Meinungsausdruck
 - ▶ Meinungsgegenstand
- ▶ Aufgabenstellungen:
 - ▶ Erkennung der Subjektivität bzw. Objektivität eines Wortes/Textes
 - ▶ Erkennung der semantischen Orientierung (positiv/negativ)
 - ▶ Erkennung der Stärke der Subjektivität oder semant. Orientierung

2.3 Meinungsanalyse - Ansätze

Lexikalisch

Charakteristik:

- ▶ Wörterbuch
- ▶ Grammatik. Regeln

Vorteil:

- ▶ Universell einsetzbar

Nachteil:

- ▶ Nur bekannte Wörter identifizierbar

Korpus-basiert

- ▶ Extraktion von Merkmalen
- ▶ Trainierter Klassifikator

- ▶ Höhere Erkennungsleistung

- ▶ Vorklass. Trainingsmenge
- ▶ Featureauswahl schwierig

2.3 Meinungsanalyse - Wörterbucherstellung

▶ Ansätze:

▶ Thesaurus-basiert

▶ Korpus-basiert

▶ Wörterbuch:

▶ *The General Inquirer* als Grundlage

▶ Pointwise Mutual Information (PMI) zum Lernen der Orientierung neuer Wörter:

$$PMI(w_1, w_2) = \log_2 \left(\frac{p(w_1 \cap w_2)}{p(w_1)p(w_2)} \right) = \log_2 \left(\frac{\frac{1}{N} \text{hits}(w_1 \text{ NEAR } w_2)}{\frac{1}{N} \text{hits}(w_1) \frac{1}{N} \text{hits}(w_2)} \right)$$

$$O(w) = \sum_{w_{pos} \in Pos} PMI(w, w_{pos}) - \sum_{w_{neg} \in Neg} PMI(w, w_{neg})$$

2.3 Meinungsanalyse - Retrieval-Modell


- ▶ Retrieval-Aufgabe: Zusammenfassung gleicher Aussagen
- ▶ Retrieval-Modell:
 - ▶ Darstellung als Vektorpaar: positiv/negativ
 - ▶ Zusammenfassung aller pos./neg. Vektoren
- ▶ Besonders geeignet für kurze Kommentare
 - ▶ Wenig Fakten/Argumente, häufig pure Meinungsäußerungen
 - ▶ Einzelne Meinung wenig informativ
 - ▶ Als Zusammenfassung: allgem. Meinungsbild

2.3 Meinungsanalyse – Anwendung auf YouTube

- ▶ Annahme bei kurzen Kommentaren:
 - ▶ Kommentator = Meinungsinhaber
 - ▶ Thema (Video) = Meinungsgegenstand

- ▶ Vorgehen:
 - ▶ Laden aller Kommentare
 - ▶ Preprocessing auf Zeichenebene
 - ▶ Preprocessing auf Wortebene
 - ▶ Erzeugung der Vektorpaare für alle Kommentare
 - ▶ Darstellung als Opinion Cloud


2.3 Meinungsanalyse – Demonstration


Weltweit (alle) | Deutsch
Neues Konto | Quicklist (0) | Hilfe | Anmelden

Startseite Videos Kanäle Community

erweitert
Video hochladen

The Simpsons - Monster Mash



Bewerten: ★★★★★
Aufrufe: 205.592

Weiterleiten
Favorit
Playlists
Melden

MySpace
Facebook
Digg
mehr

Opinion Cloud
Kommentare
Statistiken & Daten

Display Mode: Pos/Neg
Scaling: Log
Threshold:


75%
25%

... awesome :-D :-) seasoned nice ... sorry tuck suck killed ...

thank lovely

great lol good

cool
well best funny XD better ...



Von: **emmilyn11**

Hinzugefügt: 3. März 2008

[\(Weitere Informationen\)](#)

Abonnieren

Scenes from the Simpsons Halloween specials to the s...


URL:

Einbetten:

<object width="425" height="344"><param name="movie" value="

► Mehr von: **emmilyn11**


▼ Ähnliche Videos



Simpsons send in the clone, Halloween special (fast forward)

04:39 Von: sk49zt


Aufrufe: 216.838



Simpsons Secret Episode

02:12 Von: TheSecondL


Aufrufe: 1.425.517



The Simpsons - Homer Forgets

05:28 Von: SpedUpSimpsons


Aufrufe: 68.468



The Simpsons Treehouse of Horror XVII

05:39 Von: XSinEvAnX

Aufrufe: 352.110







Simpsons halloween

03:20 Von: Daz988

Aufrufe: 59.500

Promotete Videos

 VIPs machen Politik (AL... 02:18 allesBUINTE	 Curse mit Silbermond 03:47 Curse	 Qualifiers 2010 Sloveni... 02:31 Videosport	 CEREMONY : MADAGASCAR 00:28 beijing2008
--	--	---	---

3. Zusammenfassung

- ▶ 3 Retrieval-Aufgaben

- ▶ 3 Relevanzkriterien

- ▶ 3 Retrieval-Modelle

- ▶ Möglicher Retrieval-Prozess:

 - ▶ 1. Filtern über Kommentarqualität

 - ▶ 2. Ranken über Kontinuitätsmodell

 - ▶ 3. Zusammenfassen kurzer Kommentare als Opinion Cloud

3. Zusammenfassung - Ausblick

- ▶ Weitere Relevanzkriterien finden
- ▶ Vorgestellte Modelle in Anwendung testen
- ▶ Umgang mit Diskussionssträngen/Threads
- ▶ Untersuchung der Manipulierbarkeit der Modelle
- ▶ Umgang mit Rechtschreibfehlern/Tippfehlern