

Der Nutzen von Webkommentaren für das keyword-basierte Multimedia-Retrieval

Masterverteidigung

Steffen Becker

Bauhaus-Universität Weimar

Übersicht

- ▶ Multimedia-Retrieval
- ▶ Text-Retrieval
- ▶ Kommentarkorpus
- ▶ Experimente
- ▶ Zusammenfassung und Ausblick

Multimedia-Retrieval

Internet Bilder Videos Shopping News Karten Mehr | MSN Hotmail

bing^{Beta}

sketch meat menu site:youtube.com

Videos Internet Videos Mehr ▾

SUCHVERLAUF

sketch meat menu vikings...
sketch meat menu...
sketch meat menu site:youtube.com
sketch meat site:youtube.com
sketch site:youtube.com

SafeSearch: Aus [Ändern](#)

Es wurden keine Ergebnisse für **sketch meat menu vikings site:youtube.com** gefunden.

Suchtipps:

- Vergewissern Sie sich, dass die Wörter richtig geschrieben sind.
- Versuchen Sie, Suchbegriffe umzuformulieren oder Synonyme zu verwenden.
- Versuchen Sie es mit weniger spezifischen Suchbegriffen.
- Formulieren Sie Ihre Anfragen so präzise wie möglich.

[Alle anzeigen](#)
[Alle löschen](#) · [Deaktivieren](#)

Multimedia-Retrieval – Annotationen



Search

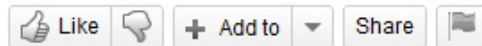
Browse

Monty Python - Spam

zumpzump

1 video

Subscribe



3,262,771

Uploaded by zumpzump on Feb 14, 2007

The origin of SPAM.

If you like this video, support Monthy Python buying their DVD.

Category:

[Entertainment](#)

Tags:

[spam](#) [montypython](#)

License:

Standard YouTube License

12,417 likes, 305 dislikes

Artist: [Monty Python](#)

Advertisement

As Seen On:

[Site Home](#)

Show less

Multimedia-Retrieval – Kommentare

Alle Kommentare (5.885)


[Alle anzeigen](#)

Auf dieses Video antworten...

this made it to encyclopedia dramatica cuz of its absurdly epic lulz

[DrMopZodiac](#) vor 6 Monaten

1:26 Click it over and over. Epic win.

[KentGamers](#) vor 6 Monaten 2 

Spam Spam Spam Spam Spam~

[RandomGeek12](#) vor 6 Monaten

I DONT LIKE SPAM!!!!

[TheNerdyCanuck](#) vor 6 Monaten

Smelly Processed Argentinian Meat.....S.P.A.M.

[FullMoonDrummer](#) vor 6 Monaten 2 

[@maplesyrupgl](#) I didn't. thanks for the info. i actually googled it and ur right :D :D

"According to the Internet Society and other sources, the term spam is derived from the 1970 Spam sketch of the BBC television comedy series "Monty Python's Flying Circus".[12][12] The sketch is set in a cafe where nearly every item on the menu includes Spam canned luncheon meat. As the waiter recites the Spam-filled menu..."

[malshania](#) vor 6 Monaten

244 people got spam sandwiches

[zaktt](#) vor 6 Monaten

Over 200 people don't like SPAM

[JuggaloFlob](#) vor 7 Monaten

[Zurück](#)

[97](#)

[98](#)

[99](#)

[100](#)

[101](#)

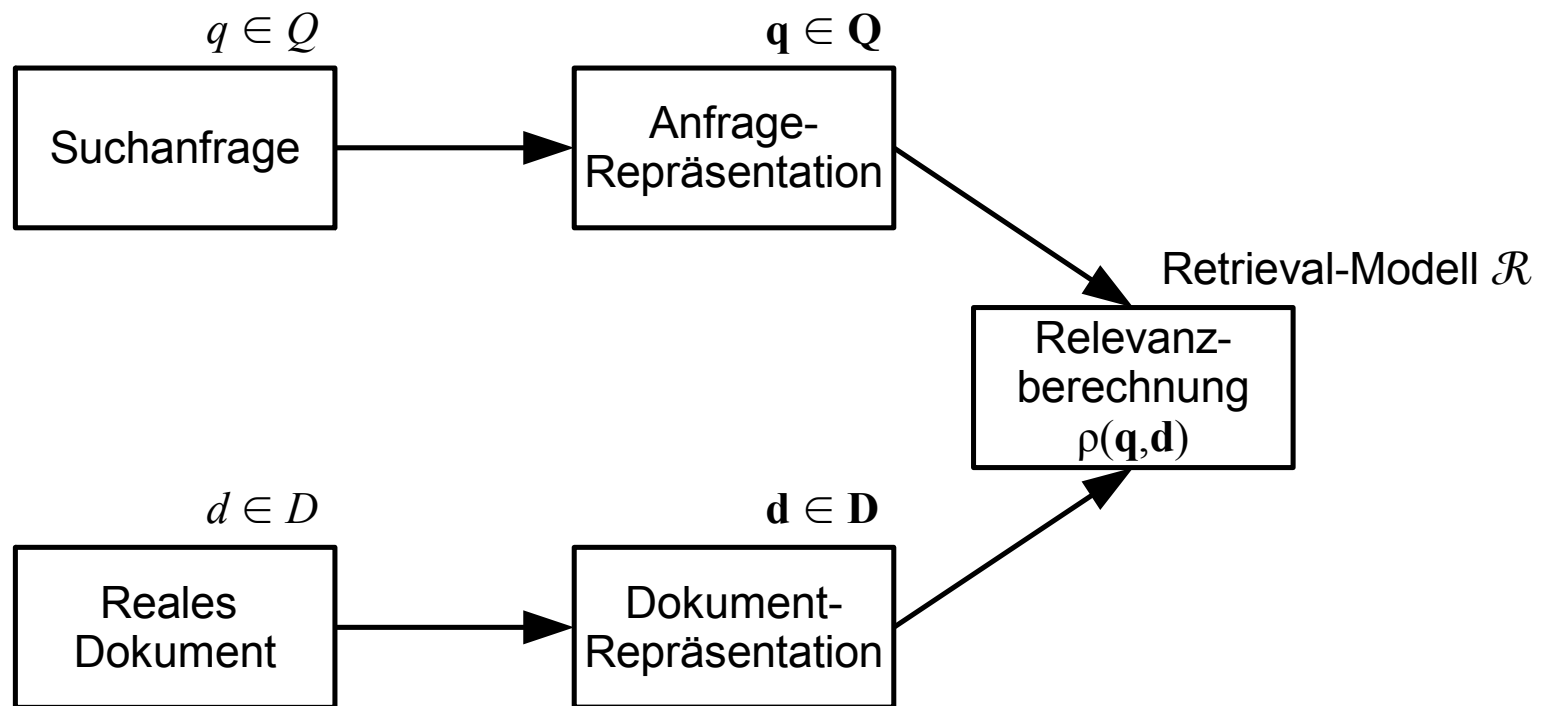
[102](#)

[103](#)

[Weiter](#)

[Alle Kommentare anzeigen »](#)

Text-Retrieval – Allgemeines Modell



Text-Retrieval – Vorgehen

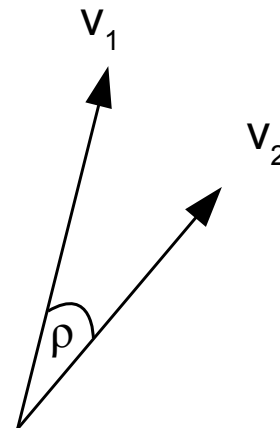
▶ Bag of Words:

- ▶ Wortzerlegung + Stammformreduktion + Stoppwortentfernung

▶ Vektorraum-Modell:

- ▶ Termhäufigkeit und Inverse Dokumenthäufigkeit als Vektor
- ▶ Ähnlichkeitsmaß: Skalarprodukt der Vektoren

	v_1	v_2
sketch	$\begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$
meat		
menu		



Text-Retrieval – Evaluierung

▶ Relevanzkorpus:

- ▶ Anfragen + Dokumente mit bekannter Relevanz

▶ Qualitätsmaße:

- ▶ Qualität der Ergebnisdokumente: Recall, Precision
- ▶ Qualität der Relevanzsortierung: NDCG, Rangkorrelation

Kommentarkorpus – Portale

► Übersicht über gecrawlte Portale:

Portal	Thema/Medium	Dokumente	Kommentare
Last.fm	Musik	7.200	300.000
YouTube	Video	300.000	80 Mio
IMDb	Film	25.000	1 Mio
Picasa	Bilder	4 Mio	7 Mio
HuffPost	Nachrichten/Text	66.000	3 Mio
Blogger	Blogs/Text	100.00	1 Mio

Kommentarkorpus – Analyse

► Textmengenverhältnisse pro Dokument:

Portal	Annotationslänge	Kommentare	Kommentarlänge	Textverhältnis
Last.fm	266	39	12	1:2
YouTube	117	238	17	1:107
IMDb	29	46	210	1:302
Picasa	6	2	9	1:5
HuffPost	695	43	26	1:4
Blogger	296	10	77	1:10

► Große Mengenunterschiede zwischen den Portalen

► Bei allen Portalen deutlich mehr Kommentar- als Annotationstext

Experimente – Übersicht

- ▶ Vergleich von Retrieval-Eigenschaften:
 - ▶ Wortschatzvergleich
 - ▶ Diskriminierungseigenschaften

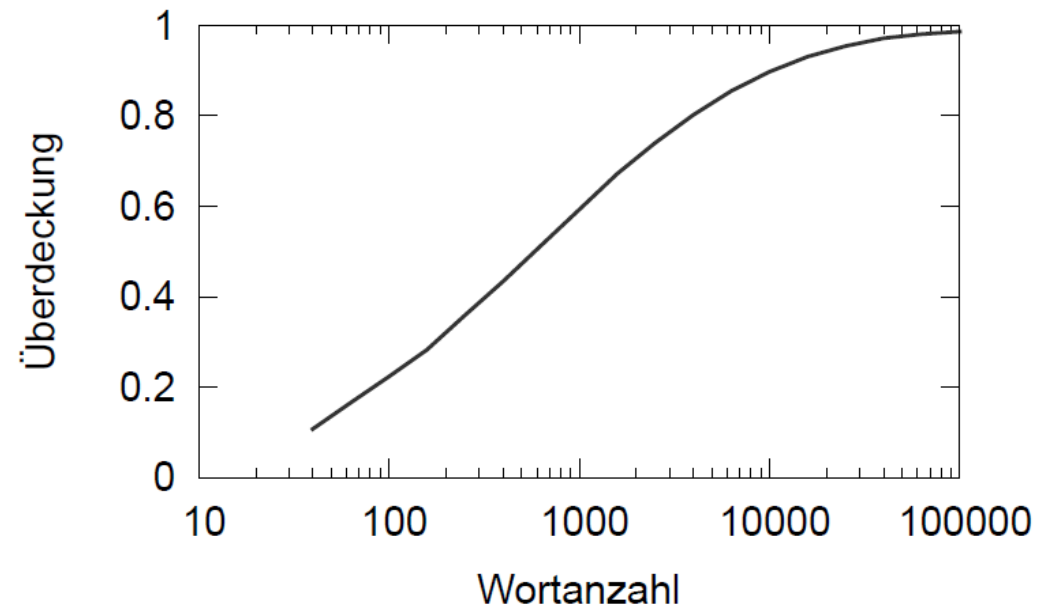
- ▶ Vergleich des Retrieval-Verhaltens:
 - ▶ Vergleich von Retrieval-Ergebnissen
 - ▶ Vergleich verschiedener Ergebnissortierungen
 - ▶ Manuelle Relevanzeinschätzung

Experimente – Wortschatzvergleich

► Durchschnittliche Meßwerte

Portal	Überdeckung
Last.fm	0,40
YouTube	0,43
IMDb	0,68
Picasa	0,03
HuffPost	0,18
Blogger	0,18

► Überdeckung bei IMDb



Experimente – Diskriminierungseigenschaften

► Durchschnittliche KL-Divergenz:

Portal	Annotationen	Kommentare	Mittlere TFIDF-Terme
Last.fm	6,45	6,25	9,40
YouTube	8,62	4,82	7,27
IMDb	8,07	3,31	7,71
Picasa	5,99	9,33	10,39
HuffPost	5,19	5,63	5,88
Blogger	7,08	7,56	8,14

► Ähnliche Diskriminierungseigenschaften

► Durch größere Wortmenge potentiell mehr diskriminierendere Wörter in den Kommentaren

Experimente – Retrieval-Ergebnisvergleich

- ▶ Suche auf Annotationen und Kommentaren mit AOL-Query-Log
- ▶ Durchschnittliche Messwerte mit Suchwortanzahl 1

Portal	Verbesserung	Überdeckung	Rangkorrelation
Last.fm	77	0,57	0,03
YouTube	18	0,28	0,28
IMDb	116	0,67	0,14
Picasa	3	0,1	0,26
HuffPost	2	0,09	0,24
Blogger	3	0,13	0,37

- ▶ Treffermengen durch Kommentare deutlich steigerbar
- ▶ Deutlich unterschiedliche Relevanzsortierungen

Experimente – Relevanzsortierungen

- ▶ Suche auf YouTube mit AOL-Query-Log, Top 100 Ergebnisse
- ▶ Durchschnittliche Messwerte mit Suchwortanzahl 1

	Bing-Sortierung	
	Überdeckung	Rangkorrelation
Kommentarsortierung	0,31	-0,02
Annotationssortierung	0,86	0,20
YouTube-Sortierung	0,27	0,21

- ▶ Deutlich unterschiedliche Ergebnisse der Suchmaschinen, kaum Übereinstimmung der Relevanzsortierungen
- ▶ Suchmaschinen als Vergleichsbasis ungeeignet

Experimente – Manuelle Relevanzsortierung

▶ Top 3 YouTube-Ergebnisse (nur mit Kommentaren gefunden)

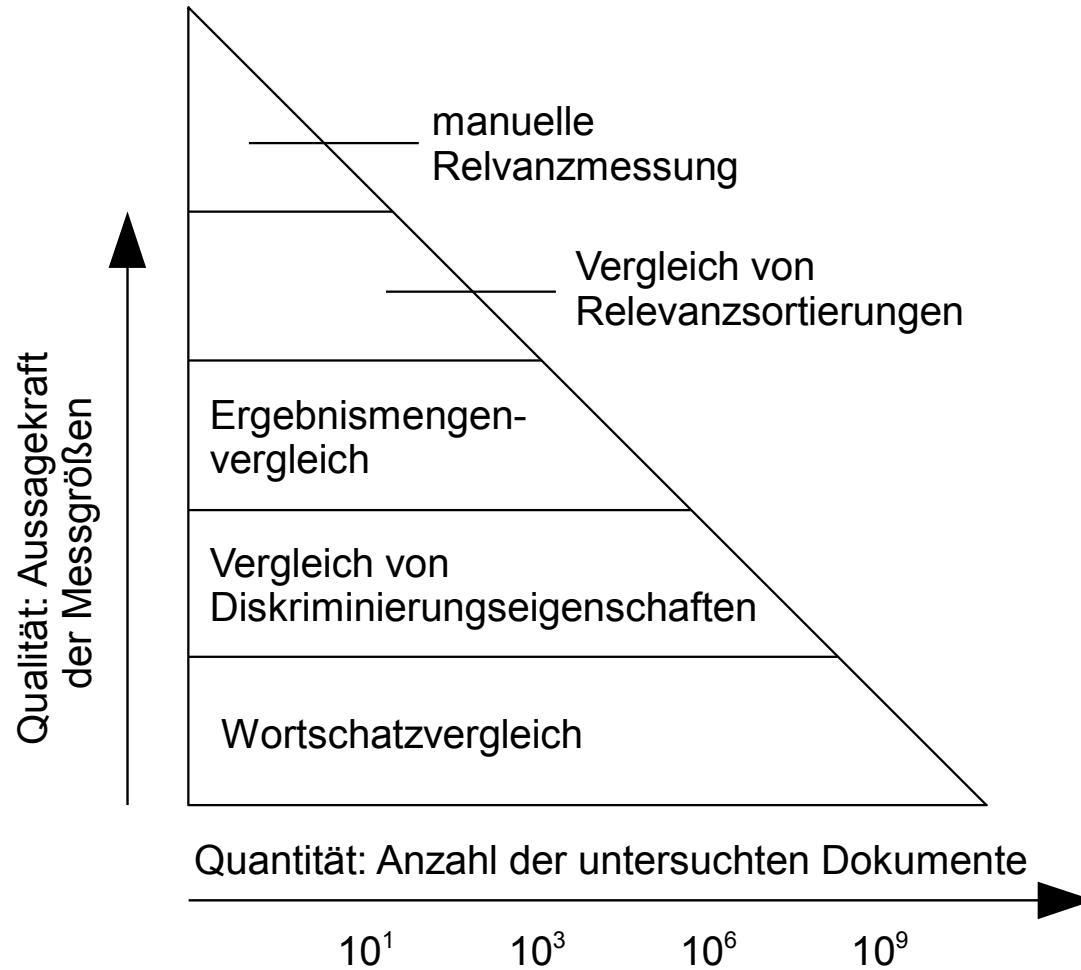
▶ Relevanz aller Treffer:

Relevanz	irrelevant	verwandt	relevant
Anteil	28 %	27 %	45 %

▶ Verteilung der relevanten und verwandten Treffer:

Anzahl	0	1	2	3
Anteil	7 %	21 %	26 %	46 %

Zusammenfassung und Ausblick



Zusammenfassung und Ausblick

- ▶ Kommentare als sinnvolle Ergänzung:
 - ▶ Wortschatz unterscheidet sich
 - ▶ Diskriminierungseigenschaften ähnlich bis höher
 - ▶ Ergebnismengen enorm steigerbar

- ▶ Qualität der Sortierung nicht eindeutig bewertbar:
 - ▶ Websuchmaschinen keine Vergleichsbasis
 - ▶ Manuelle Relevanzbestimmung zu aufwändig

- ▶ Schaffung eines Relevanzkorpus notwendig:
 - ▶ Genauere Messung
 - ▶ Entwicklung spezieller Relevanzfunktionen