Leo Wendt

tf-idf

Information Retrieval

Gliederung

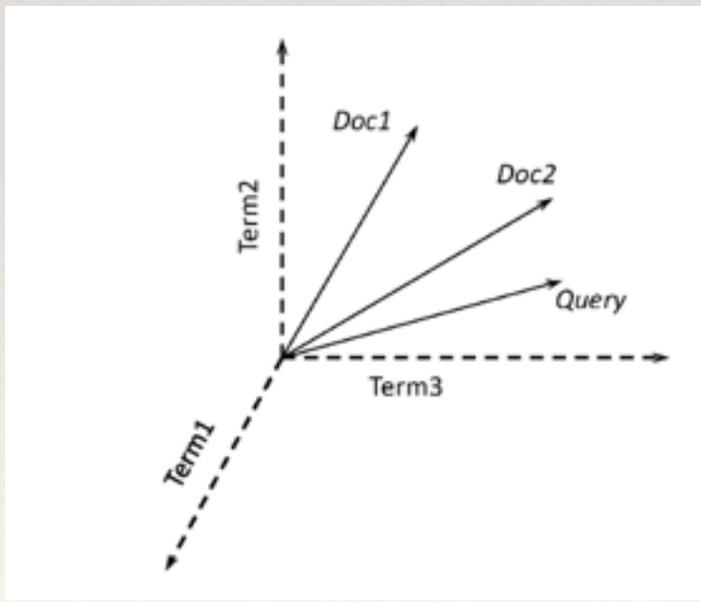
- Einordnung
- Ranking mit tf-idf
- Beispiel
- Fragen

Einordnung

- statistisches Verfahren zur Angabe der Relevanz eines Wortes von einem Dokument innerhalb einer Dokumentensammlung
- Anwendung als Gewichtungsfaktor (Scoring/Ranking) beim

Vector Space Model

	D_1	D_2	D_3	D_4
aquarium	1	1	1	1
bowl	0	0	1	0
care	0	1	0	0
fish	1	1	2	1
freshwater	1	0	0	0
goldfish	0	0	1	0
homepage	0	0	0	1
keep	0	0	1	0
setup	0	1	0	0
tank	0	1	0	1
tropical	1	1	1	2



Darstellungen: Ralf Krestel

Ranking mit tf-idf | tf

- Ausgangsposition: „bag of words“ -> Gewicht ist Anzahl der Vorkommnisse des Terms
  - Beobachtung: Wiederholung ist ein Zeichen für Betonung (Luhn)
- Idee: Einführung der „term frequency“ (tf)
  - höhere term frequency -> höheres Gewicht
- unterschiedliche Berechnungen

Ranking mit tf-idf | tf

- Beobachtung: einfache Termhäufigkeit ungeeignet, da so lange Dokumente bevorteilt werden
- Relevanz eines Terms ist nicht direkt proportional zur Anzahl (500 / 1000)

- Lösung 1: Benutzung einer logarithmischen Skala
  - $tf(t,d) = \log(f_{t,d} + 1)$
- Lösung 2 (häufiger verwendet): Normieren mit Anzahl des am häufigsten vorkommenden Terms
  - $tf(t,d) = f_{t,d} \ / \ \max\{f_{t,d} : t' \in d\}$

Ranking mit tf-idf | idf

- Ausgangsposition: Betrachtung der Relevanz des Terms isoliert von anderen Dokumenten
- Problem: häufig in einer Dokumentensammlung auftretende Terme sind weniger aussagekräftig (Bsp.: Psychologie in Informatikliteratur vs in Psychologieliteratur)
- Idee: Einführung der „inverse document frequency“ (idf)
  - höhere idf -> höheres Gewicht
- Bsp. zur Berechnung: $idf_t = \log(N/n_t)$
  - je größer idf, desto geringer ist n_t

Ranking mit tf-idf

- tfidf(t,d,D) = tf(t,d) * idf(t,D)

Beispiel

term	tf(doc 1)	tf(doc 2)	tf(doc 3)	n_t	idf
Methode	4250 / 50000 = 0.085	3400 / 43000		850	
der	50000/50000 = 1	43000/43000		1000	log(1000/1000) = 0.00
Wasser	7600 / 50000 = 0. 152	4000 / 43000		400	0.40
Bioreaktor	600 / 50000 = 0.012	0 /43000		25	log(1000/25) =1.60

N = 1000

Beispiel

doc 3

term	abs. Häufigkeit	tf	idf	tf-idf
Methode	3		log(1000/800)=0.09	
der	10		log(1000/1000) = 0.00	0
Wasser	5	5 / 10 = 0.5	0.40	0.5 * 5 = 2.25
Bioreaktor	8	8 / 10 = 0.8	log(1000/25)=1.60	0.8 * 8 = 6.40

Fragen

- Wofür stehen jeweils tf und idf und was bedeuten sie?
- Was sind mögliche Nachteile?

Literatur

- Ruud Koot et. al.: tf-idf. Abrufbar unter: https://en.wikipedia.org/wiki/Tf–idf.
- Ralf Krestel: Retrieval Models I. Abrufbar unter: https://www.dropbox.com/s/pualu5zg6sx842q/04-retrievalModelsI.pdf?dl=0.
- Boolesche- und Vektorraum-Modelle. Abrufbar unter: https://www.kde.cs.uni-kassel.de/lehre/ws2006-07/IR/folien/4Folie_02_IRmodels_d.pdf.
- Jens Wolff (2004): Vorlesung Information Retrieval. Abrufbar unter: http://www.iai.uni-bonn.de/III/lehre/vorlesungen/InformationRetrieval/WS04/Vorlesung-041104neu.pdf.
- Michael Dittenbach (2010): Storing and Ranking Techniques - tf-idf term weighting and cosine similarity. Abrufbar unter: http://www.ir-facility.org/scoring-and-ranking-techniques-tf-idf-term-weighting-and-cosine-similarity.