MS-Anchor: Linktexte als Ranking-Features im Zeitalter von Deep Learning

Maximilian Probst

Bachelorarbeit
Martin-Luther-Universität Halle-Wittenberg

05.10.2021

Einleitung

Was ist Anchor-Text?

Einleitung

Was ist Anchor-Text?
Beispiel:

Please visit this example page.

```
<a href="http://www.example.com">example page</a>
```

Einleitung

"Anchors often provide more accurate descriptions of web pages than the pages themselves."
– Sergey Brin und Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, 1998

Einleitung

- Anchor-Text bietet Vorteile für viele traditionelle Retrieval-Modelle
- Retrieval-Modelle entwickeln sich allerdings immer weiter
  $\rightarrow$ ist Anchor-Text noch immer von Nutzen?

Einleitung

- Anchor-Text bietet Vorteile für viele traditionelle Retrieval-Modelle
- Retrieval-Modelle entwickeln sich allerdings immer weiter
  $\rightarrow$ ist Anchor-Text noch immer von Nutzen?
- Grundlage für Anchor-Texte hier: MS Marco-Datensatz

Extraktion der Anchor-Texte

- MS Marco-Datensatz beinhaltet kaum Anchor-Text $\rightarrow$ Common Crawls nutzen
- genutzt wurden Common Crawls der Jahre 2016-2021
- ca. 2-4 Mrd. Webseiten und 50-80 TiB komprimierte Daten pro Crawl

Extraktion der Anchor-Texte

- parsen der Common Crawl Daten

Extraktion der Anchor-Texte

- parsen der Common Crawl Daten
- Filterschritte Anwenden
  - nur MS Marco-Seiten betrachten
  - seiteninterne Links entfernen
  - Stopwort-Anchor-Text entfernen
  - zu lange Anchor-Texte entfernen

Extraktion der Anchor-Texte

- parsen der Common Crawl Daten
- Filterschritte Anwenden
  - nur MS Marco-Seiten betrachten
  - seiteninterne Links entfernen
  - Stopwort-Anchor-Text entfernen
  - zu lange Anchor-Texte entfernen
- Anchor-Kontext bestimmen

Extraktion der Anchor-Texte

- parsen der Common Crawl Daten
- Filterschritte Anwenden
  - nur MS Marco-Seiten betrachten
  - seiteninterne Links entfernen
  - Stopwort-Anchor-Text entfernen
  - zu lange Anchor-Texte entfernen
- Anchor-Kontext bestimmen
- Anchor-Text-Daten im JSONL-Format speichern

Extraktion der Anchor-Texte

Struktur der JSONL Daten:

Extraktion der Anchor-Texte

Struktur der JSONL Daten:

`AnchorElement`

Extraktion der Anchor-Texte

Struktur der JSONL Daten:

`AnchorElement`

↳ `anchorText`

Extraktion der Anchor-Texte

Struktur der JSONL Daten:

`AnchorElement`

- ↳ `anchorText`
- ↳ `anchorContext`

Extraktion der Anchor-Texte

Struktur der JSONL Daten:

`AnchorElement`

- ↳ `anchorText`
- ↳ `anchorContext`
- ↳ `targetUrl`

Extraktion der Anchor-Texte

Struktur der JSONL Daten:

```
AnchorElement
    ↳ anchorText
    ↳ anchorContext
    ↳ targetUrl
    ↳ targetMsMarcoDocIds[]
```

Extraktion der Anchor-Texte

Struktur der JSONL Daten:

```
AnchorElement
    ↳ anchorText
    ↳ anchorContext
    ↳ targetUrl
    ↳ targetMsMarcoDocIds[]
    ↳ document
```

Extraktion der Anchor-Texte

Struktur der JSONL Daten:

```
AnchorElement
    ↳ anchorText
    ↳ anchorContext
    ↳ targetUrl
    ↳ targetMsMarcoDocIds[]
    ↳ document
        ↳ scrUrl
        ↳ recordID
        ↳ trecID
        ↳ infoID
```

Extraktion der Anchor-Texte

Struktur der JSONL Daten:

```
AnchorElement
    ↳ anchorText
    ↳ anchorContext
    ↳ targetUrl
    ↳ targetMsMarcoDocIds[]
    ↳ document
        ↳ scrUrl
        ↳ recordID
        ↳ trecID
        ↳ infoID
        ↳ naughtyWords[]
```

Extraktion der Anchor-Texte

Table: Übersicht der Extrahierten Anchor-Text-Daten.

Crawl	Anzahl Seiten	Anchor Texte	∅ Anchor Text pro	
			Tgt Seite	Src Seite
2016-07	1,73 Mrd.	1,05 Mrd.	331	3,09
2017-04	3,14 Mrd.	0,95 Mrd.	171	2,49
2018-13	3,20 Mrd.	0,83 Mrd.	136	2,10
2019-47	2,55 Mrd.	0,55 Mrd.	148	1,98
2020-05	3,10 Mrd.	0,67 Mrd.	159	1,99
2021-04	3,40 Mrd.	0,52 Mrd.	150	1,87

Extraktion der Anchor-Texte

Table: Übersicht der Extrahierten Anchor-Text-Daten.

Crawl	Anzahl Seiten	Anchor Texte	∅ Anchor Text pro	
			Tgt Seite	Src Seite
2016-07	1,73 Mrd.	1,05 Mrd.	331	3,09
2017-04	3,14 Mrd.	0,95 Mrd.	171	2,49
2018-13	3,20 Mrd.	0,83 Mrd.	136	2,10
2019-47	2,55 Mrd.	0,55 Mrd.	148	1,98
2020-05	3,10 Mrd.	0,67 Mrd.	159	1,99
2021-04	3,40 Mrd.	0,52 Mrd.	150	1,87

Insgesamt:

- 17 Mrd. Webseiten betrachtet
- 4,5 Mrd. Anchor-Texte extrahiert
- 52% der MS Marco-Dokumente abgedeckt

Anchor-Text als Retrieval-Feature

- Retrieval-Performance mittels Anserini prüfen

Anchor-Text als Retrieval-Feature

- Retrieval-Performance mittels Anserini prüfen
- Samples aus Anchor-Texten entnehmen (Schwellwert=2000 Anchor-Texte)

Anchor-Text als Retrieval-Feature

- Retrieval-Performance mittels Anserini prüfen
- Samples aus Anchor-Texten entnehmen (Schwellwert=2000 Anchor-Texte)

Erste Ergebnisse mittels BM25:

	Content-Baseline	Anchor-Text (CC 19-47)
MAP	0,23	0,08
R@100	0,71	0,24
R@1000	0,88	0,30

Anchor-Text als Retrieval-Feature

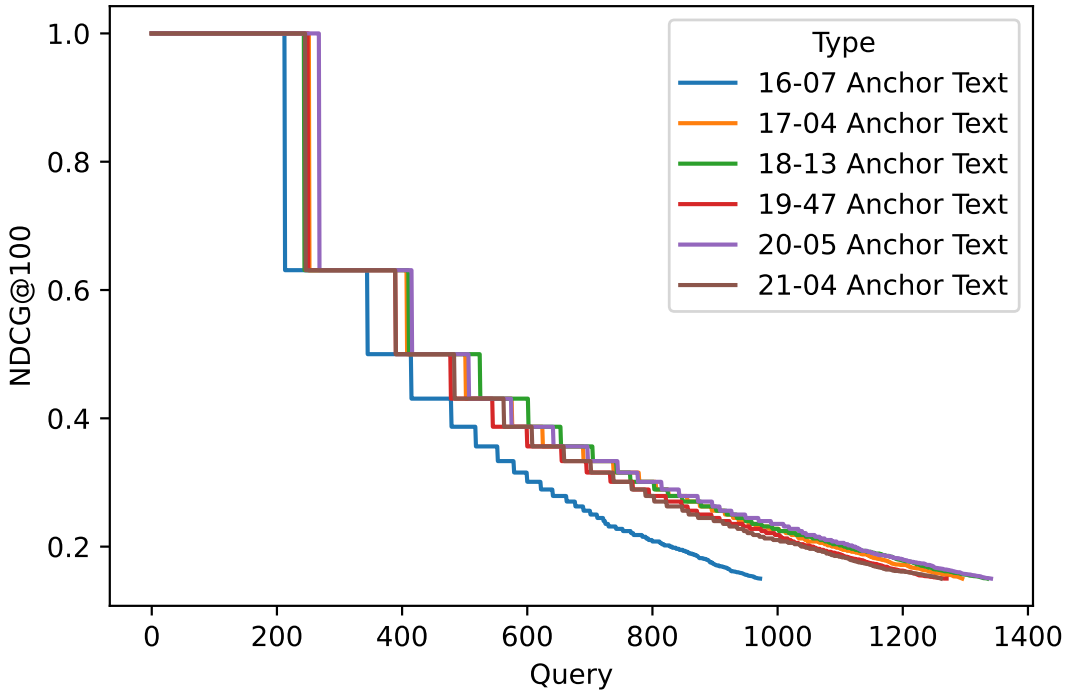

Figure: NDCG@100 je Query für Anchor-Texte der Jahre 2016 bis 2021.

Anchor-Text als Retrieval-Feature

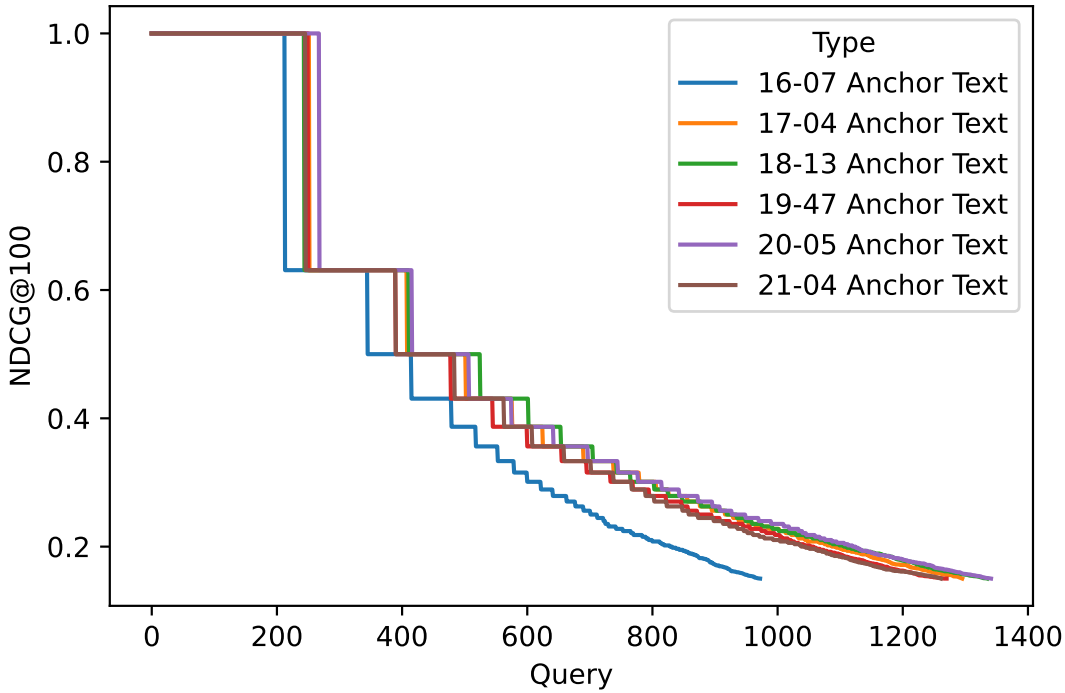

Figure: NDCG@100 je Query für Anchor-Texte der Jahre 2016 bis 2021.

- Retrieval-Performance Prinzipiell eher Zeitunabhängig
- aber: CC 16-07 schnitt unterdurchschnittlich schlecht ab

Anchor-Text als Retrieval-Feature

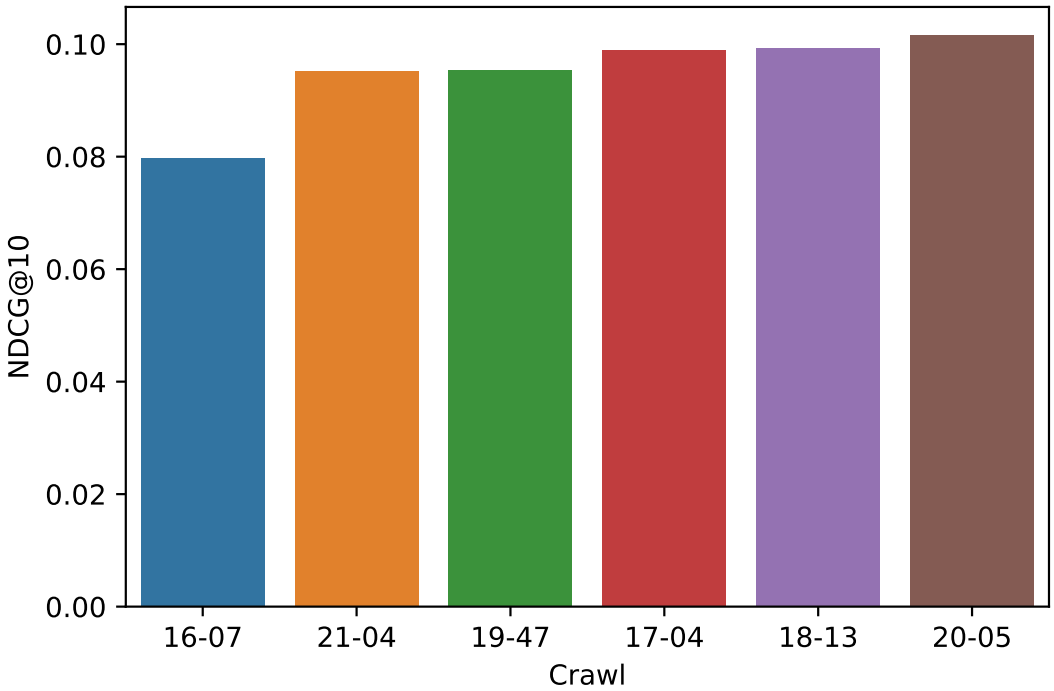

Figure: NDCG@10 je Crawl für Anchor-Texte der Jahre 2016 bis 2021.

Anchor-Text als Retrieval-Feature

Aggregation von Anchor-Texten

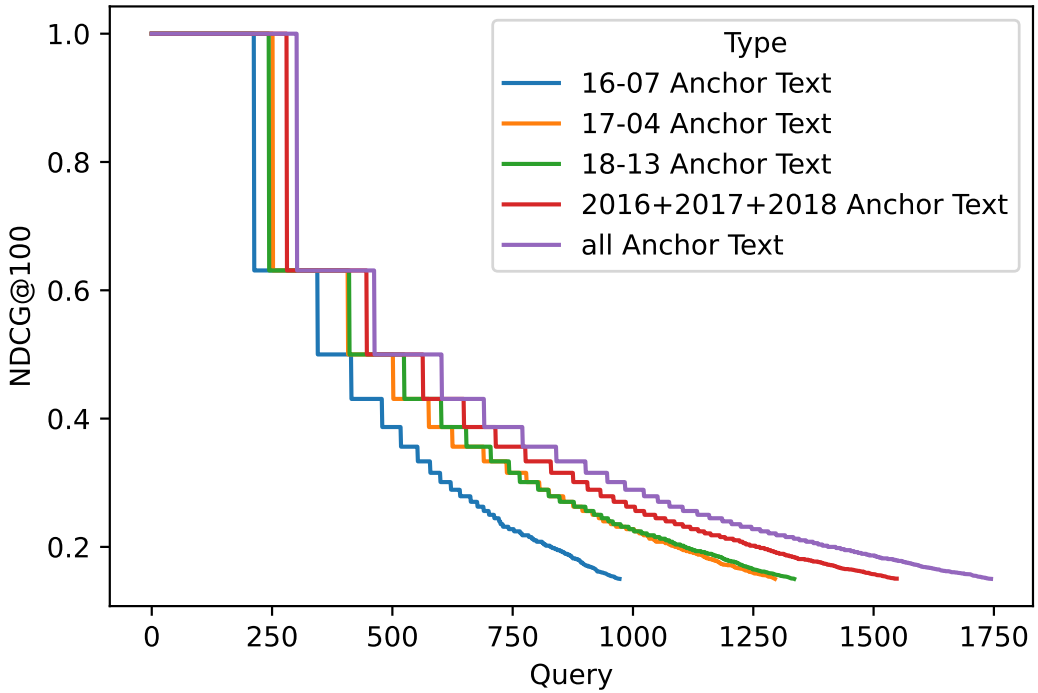

Figure: NDCG@100 je Query für Anchor-Texte der Jahre 2016 bis 2018, sowie deren Kombination und die Kombination aller Jahre (2016–2021.)

Anchor-Text als Retrieval-Feature

Aggregation von Anchor-Texten

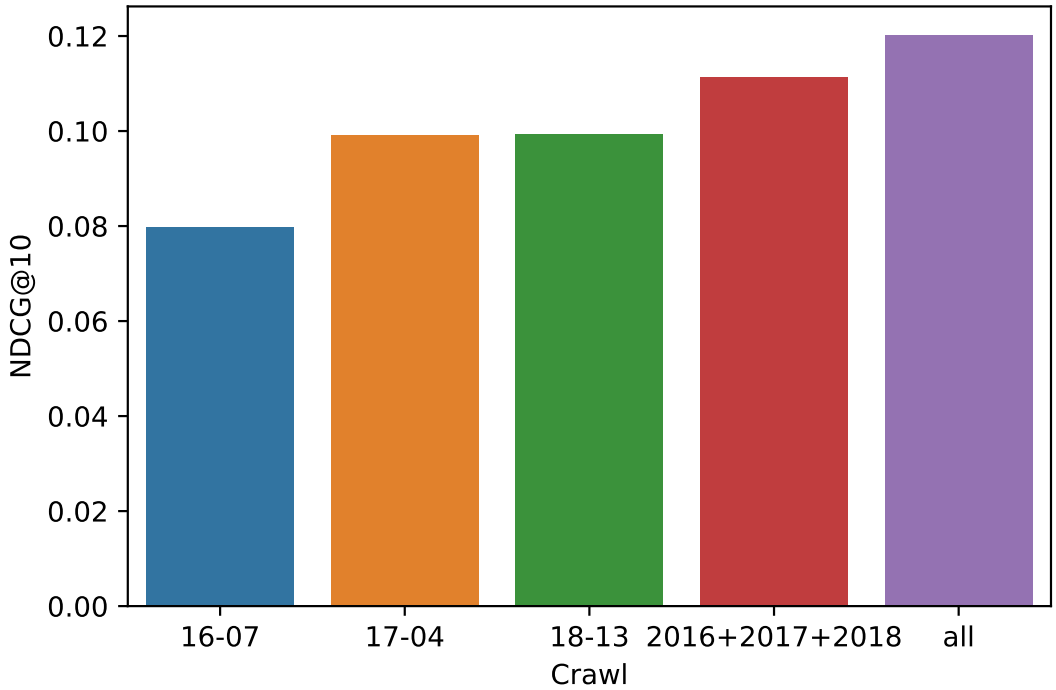

Figure: NDCG@10 je Query für Jahr der Jahre 2016 bis 2018, sowie deren Kombination und die Kombination aller Jahre (2016–2021).

Anchor-Text als Retrieval-Feature

ORCAS-Datensatz

- 18 Mio. Query-URL-Paare aus Bings Query-Logs
- für Forschungszwecke frei verfügbar

Anchor-Text als Retrieval-Feature

ORCAS-Datensatz

- 18 Mio. Query-URL-Paare aus Bings Query-Logs
- für Forschungszwecke frei verfügbar

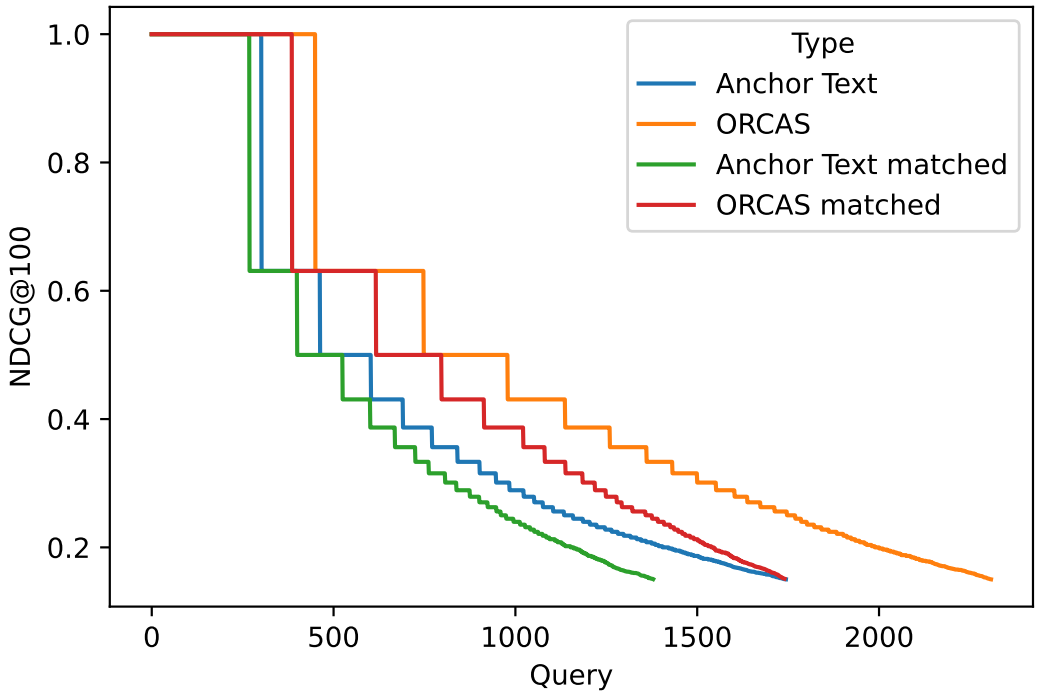

Figure: NDCG@100 pro Query für Anchor-Texte und ORCAS-Queries, sowie für auf deren Schnittmenge reduzierter Teildatensätze.

Anchor-Text für navigational Queries

Anchor-Text für navigational Queries

- Intention von Suchanfragen kann in 3 Arten unterteilt werden

Anchor-Text für navigational Queries

- Intention von Suchanfragen kann in 3 Arten unterteilt werden
  - informational Queries
  - navigational Queries
  - transactional Queries

Anchor-Text für navigational Queries

- Intention von Suchanfragen kann in 3 Arten unterteilt werden
  - informational Queries
  - navigational Queries
  - transactional Queries
- Bsp. informational Query: “wetter morgen” $\rightarrow$ breites Ergebnisspektrum möglich
- Bsp. navigational Query: “accuweather” $\rightarrow$ präzision nötig

Anchor-Text für navigational Queries

- Intention von Suchanfragen kann in 3 Arten unterteilt werden
  - informational Queries
  - navigational Queries
  - transactional Queries
- Bsp. informational Query: “wetter morgen” $\rightarrow$ breites Ergebnisspektrum möglich
- Bsp. navigational Query: “accuweather” $\rightarrow$ präzision nötig
- ca. 39% aller Anfragen sind informational
- ca. 25% aller Anfragen sind navigational

Anchor-Text für navigational Queries

Anchor-Text für navigational Queries

- Nick Craswell, David Hawking und Stephen Robertson mit "Effective site finding using link anchor information" erreichten gute Ergebnisse auf navigational Queries

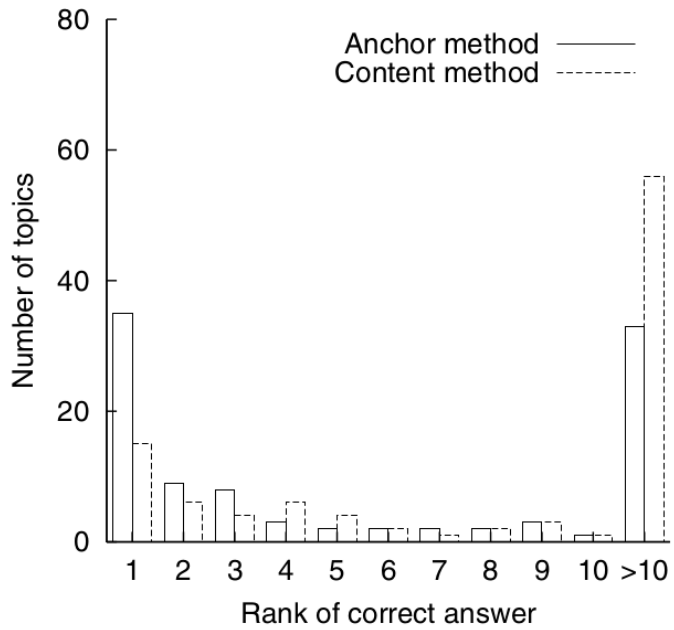

Figure: zufällige entry Pages (Craswell et al. 2001).

Anchor-Text für navigational Queries

Anchor-Text für navigational Queries

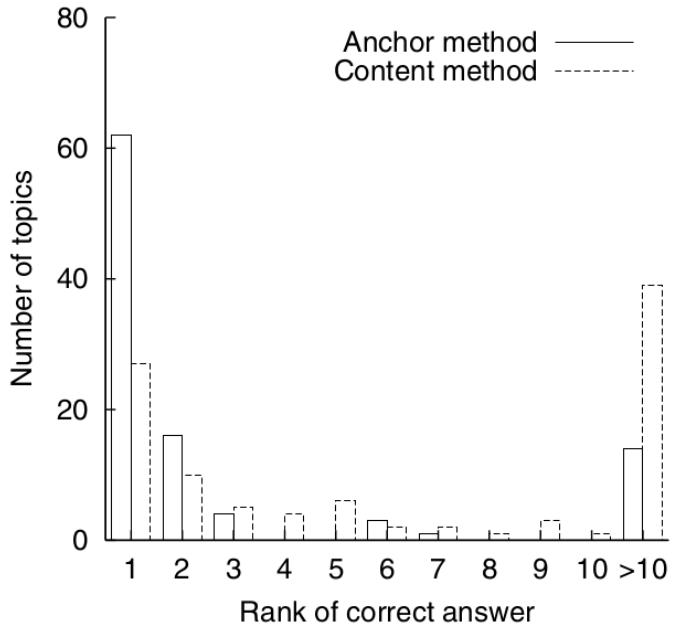

Figure: Yahoo entry Pages (Craswell et al. 2001).

Anchor-Text für navigational Queries

Experiment wird nachgestellt

- 100 zufällige Query-Paare erstellen
  (z.B. <sigir2001, http://www.sigir2001.org/>)
- 100 zufällige Query-Paare mit Alexa top 500 Seiten von MS Marco erstellen

Anchor-Text für navigational Queries

Anchor-Text für navigational Queries

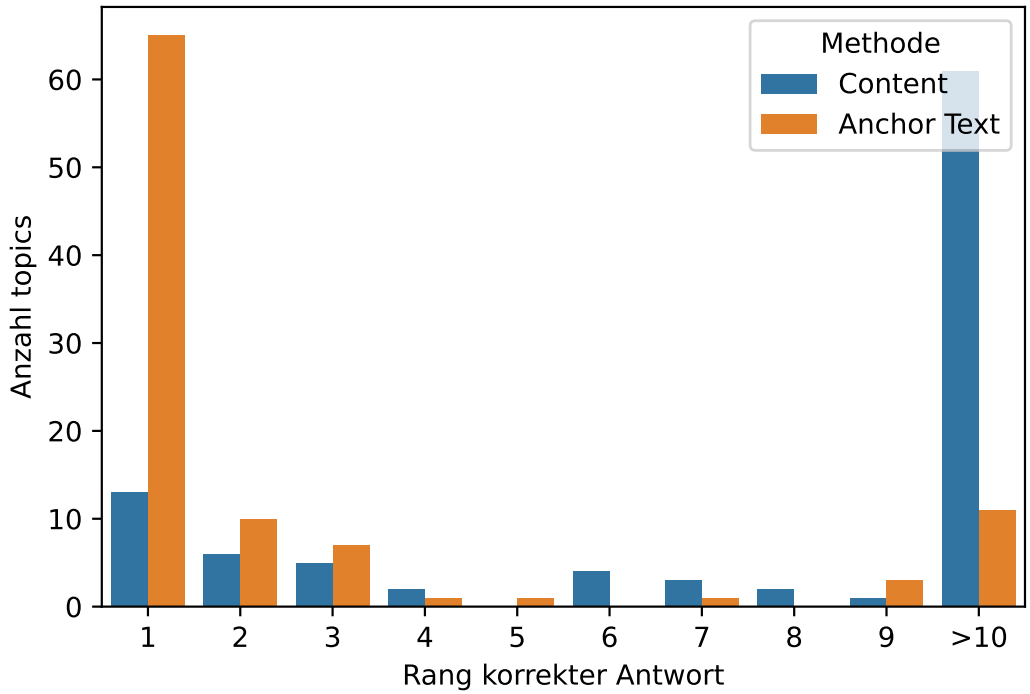

Figure: zufällige entry Pages auf MS Marco.

Anchor-Text für navigational Queries

Anchor-Text für navigational Queries

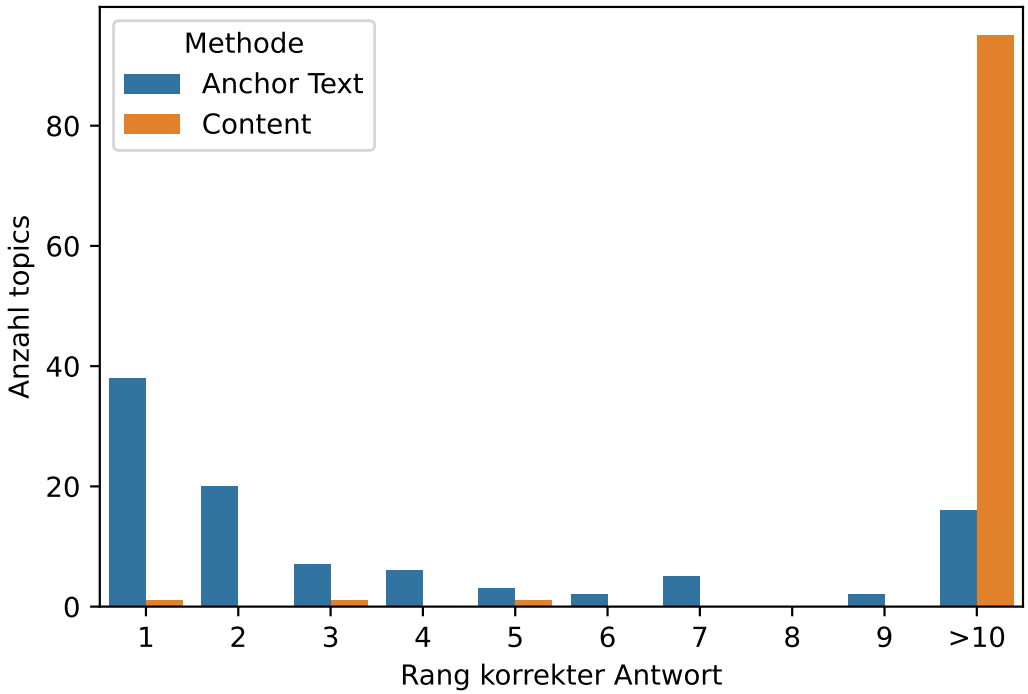

Figure: beliebte entry Pages auf MS Marco.

Anchor-Text für navigational Queries

Begründung:

Anchor-Text für navigational Queries

Begründung:

- Seite ist bekannter $\rightarrow$ über sie wird mehr berichtet

Anchor-Text für navigational Queries

Begründung:

- Seite ist bekannter $\rightarrow$ über sie wird mehr berichtet
- eigener Name wird oft nur selten erwähnt

Kombination von Anchor-Text mit anderen Features

- Anchor-Text mit anderen Features zu kombinieren Sinnvoll
- genutzt wird LambdaMART in Implementierung von LightGBM

Kombination von Anchor-Text mit anderen Features

- Anchor-Text mit anderen Features zu kombinieren Sinnvoll
- genutzt wird LambdaMART in Implementierung von LightGBM

Table: Ergebnis des Learning-to-Rank-Verfahrens gegenüber der Anserini Baseline und doc5Tquery (Deep Learning 2020).

		MRR	NDCG@10
Learning-to-Rank	LambdaMART	0,944	0,596
Anserini Baseline	BM25	0,852	0,527
doc5Tquery	BM25	0,937	0,589

Kombination von Anchor-Text mit anderen Features

- Anchor-Text mit anderen Features zu kombinieren Sinnvoll
- genutzt wird LambdaMART in Implementierung von LightGBM

Table: Ergebnis des Learning-to-Rank-Verfahrens gegenüber der Anserini Baseline und doc5Tquery (Deep Learning 2020).

		MRR	NDCG@10
Learning-to-Rank	LambdaMART	0,944	0,596
Anserini Baseline	BM25	0,852	0,527
doc5Tquery	BM25	0,937	0,589

- Anchortext als siebtwichtigstes Feature von 50

Fazit

Fazit

- Anchor Texte können bei richtiger Nutzung durchaus hilfreich sein

Fazit

- Anchor Texte können bei richtiger Nutzung durchaus hilfreich sein
- der Datensatz kann somit zu weiteren Forschungen beitragen