

Carina Fiedler

# INDEXING

# Inhalt

- ⦿ Motivation – Wozu braucht man Indizes?
- ⦿ Invertierter Index
- ⦿ Programmierung eines invertierten Index
- ⦿ Suchanfragen mit Hilfe eines Index

# Motivation – Wozu braucht man Indizes?

- Grundlage: kurze Suchanfragen, Elemente nicht in allen Dokumenten vorhanden
- Ohne Index: sequentielles Durchsuchen aller Dokumente => sehr hoher Zeit- und Rechenleistungsbedarf
- Index optimiert Geschwindigkeit und Durchführung, da nur der Index durchsucht werden muss

# Invertierter Index

- Bekannteste Datenstruktur für Textsuche
- Aufbau:

and	1	only	2
aquarium	3	pigmented	4
are	3 4	popular	3
around	1	refer	2
as	2	referred	2
both	1	requiring	2
bright	3	salt	1 4
coloration	3 4	saltwater	2
derives	4	species	1
due	3	term	2
environments	1	the	1 2
fish	1 2 3 4	their	3
fishkeepers	2	this	4
found	1	those	2
fresh	2	to	2 3
freshwater	1 4	tropical	1 2 3
from	4	typically	4
generally	4	use	2
in	1 4	water	1 2 4
include	1	while	4
including	1	with	2
iridescence	4	world	1
marine	2		
often	2 3		

Lexikon/Wörterbuch/Vokabular	Autreten/Postings
<ul style="list-style-type: none"> <li>- Liste aller vorkommenden Terme</li> <li>- Zeiger auf die Postings-Struktur</li> <li>- Datenstruktur: Hashtabelle, Baumstrukturen</li> </ul>	<ul style="list-style-type: none"> <li>- Liste aller Positionen, an denen das Wort in den Dokumenten auftritt</li> <li>- Dokumenten ID, Anzahl der Dokumente, in denen der Term auftritt, Position im Dokument</li> </ul>

# Programmierung eines Index

- ⦿ Sammlung der Indexterme
- ⦿ Alphabetische Sortierung der Terme und docID
- ⦿ Gruppierung von gleichen Termen
- ⦿ Aufspaltung in Wörterbuch und Postings
- ⦿ Sortierung der Postings nach docID

# Suchanfragen mit Hilfe eines Index

Invertierter Suchalgorithmus:

- ⦿ Finde gesuchte Elemente (Wörter, Muster etc.) im Lexikon
- ⦿ Frage die Auftrittspositionen für jeden Lexikoneintrag ab
- ⦿ Bewerte die Relevanz der Auftrittspositionen bzw. der Dokumente entsprechend Retrieval Modell

# Quellen

- ◉ [en.wikipedia.org/wiki/Search\\_engine\\_indexing](https://en.wikipedia.org/wiki/Search_engine_indexing)
- ◉ [en.wikipedia.org/wiki/Inverted\\_index](https://en.wikipedia.org/wiki/Inverted_index)
- ◉ [en.wikipedia.org/wiki/Database\\_index](https://en.wikipedia.org/wiki/Database_index)
- ◉ [nlp.stanford.edu/IR-book/html/htmledition/a-first-take-at-building-an-inverted-index-1.html](http://nlp.stanford.edu/IR-book/html/htmledition/a-first-take-at-building-an-inverted-index-1.html)
- ◉ [www.csee.umbc.edu/~ian/irF02/lectures/04Search-Inverted-Files.pdf](http://www.csee.umbc.edu/~ian/irF02/lectures/04Search-Inverted-Files.pdf)