Information Retrieval

Nikolai Henning, Jule Cremer
Computerlinguistische Grundlagen I WS 21/22

Inhalt

- Definition
- Geschichte
- Grundsätzliches Vorgehen
  - Boolsche Retrieval
  - Vektorraummodell
- Term Document Matrix
- Inverted Index
- Precision und Recall
- Problematiken

“Gegenstand des IR ist die Repräsentation, Speicherung und Organisation von Informationen und der Zugriff zu Informationen. Dabei gibt es grundsätzlich keine Einschränkungen in der Art der Informationen.” - Gerard Salton

Definition

-dt. (Wieder-)Gewinnung von Informationen, Informationsrückgewinnung

-aus einer großen Menge von unsortierten Daten spezielle Informationen bereitstellen

-Grundvoraussetzung: große Datenmenge, Wissensbasis

-Aufgaben eines IR-Systems: Repräsentation von Daten in Wissen, Datenmenge durchsuchen, die Informationen, die einem Anfragenden möglichst gut bei einer Problemlösung helfen können, darin bewerten und gewichten

-Bereiche: Internetrecherche, E-Mail/Computer Suche…

Geschichte

-1945 Vannevar Bush: “As We May Think” eine Zukunftsvision der Informationsbeschaffung und -organisation

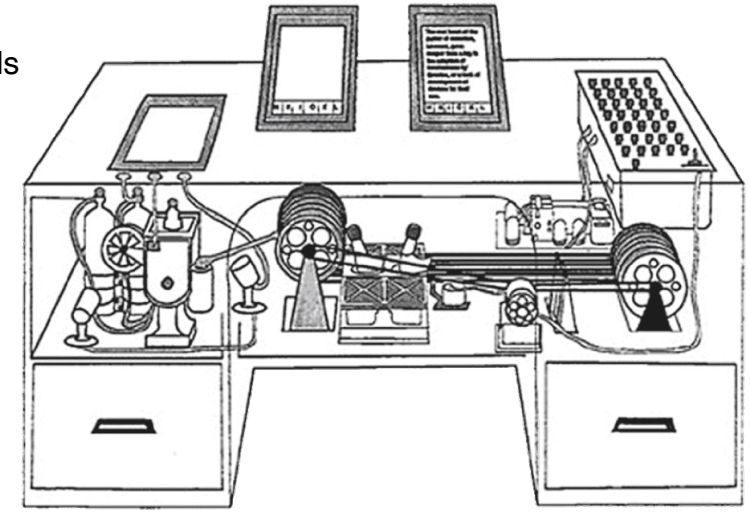
-Memex, eine Maschine so groß wie ein Schreibtisch, die als Wissensspeicher und Rechercheapparat dienen sollte

- 1950er Hans Peter Luhn: textstatistische Verfahren

-1960er Salton: Vektorraummodell

- 1970er: verschiedene Retrieval-Systeme werden auf kleine Textkorpora angewandt

-1991: durch das Internet wird Information Retrieval zum Massenphänomen (WAIS)



Vorgehen

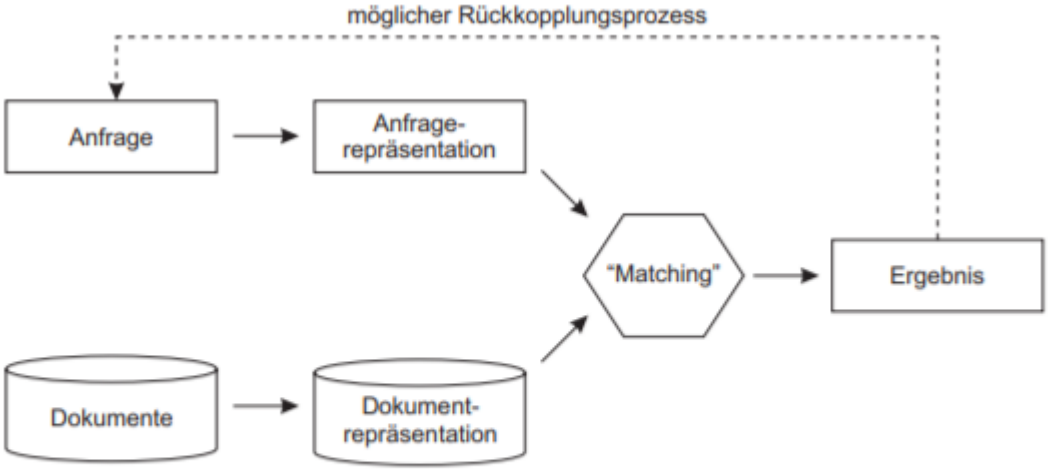


Abbildung 1.8 — Verallgemeinertes Modell der Anfragebearbeitung beim Information Retrieval

Boolsches Retrieval

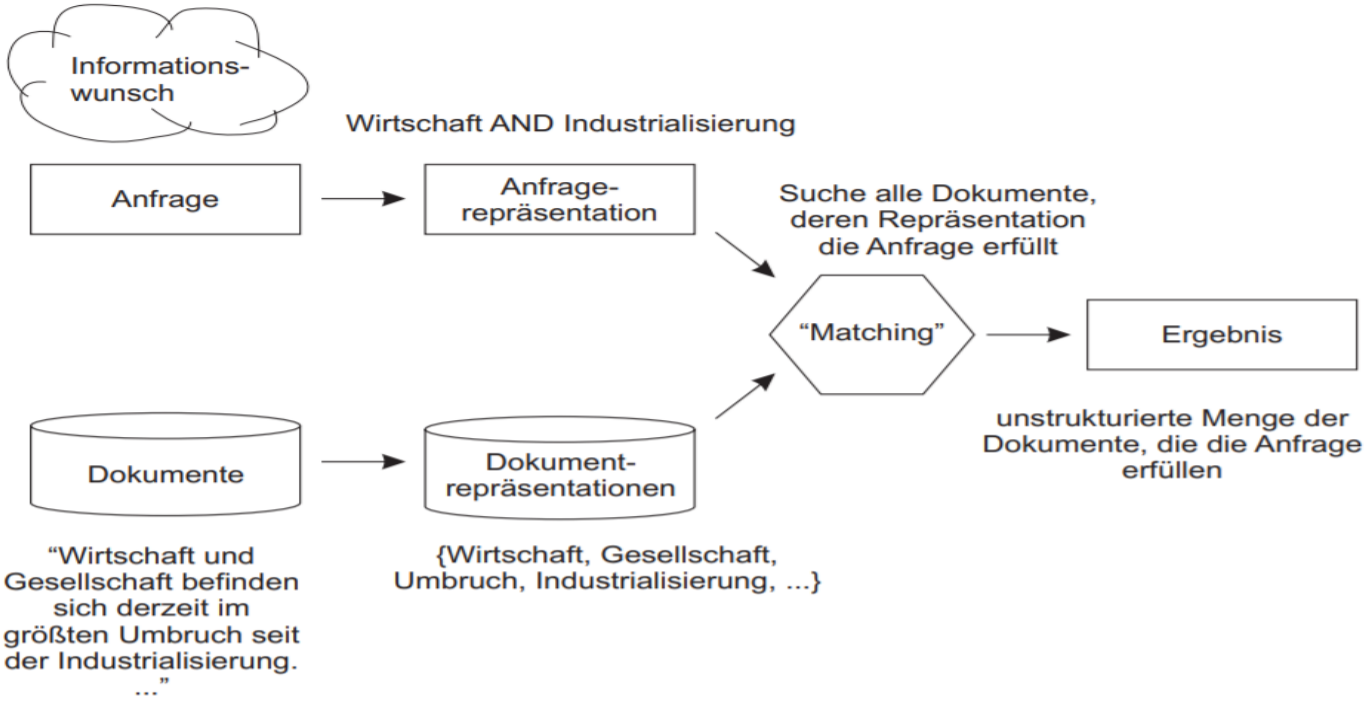


Abbildung 1.9 — Modell der Anfragebearbeitung beim Booleschen Retrieval

Vektorraummodell

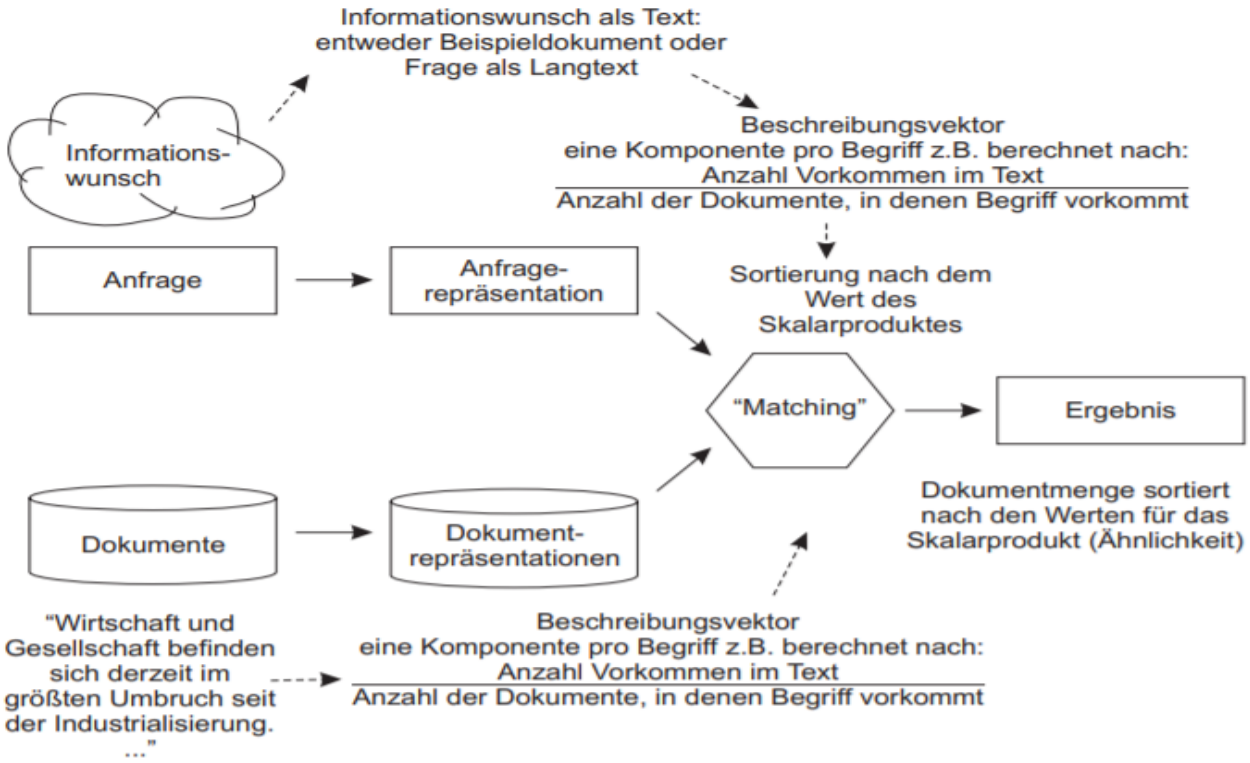


Abbildung 1.10 — Modell der Anfragebearbeitung beim Vektorraummodell

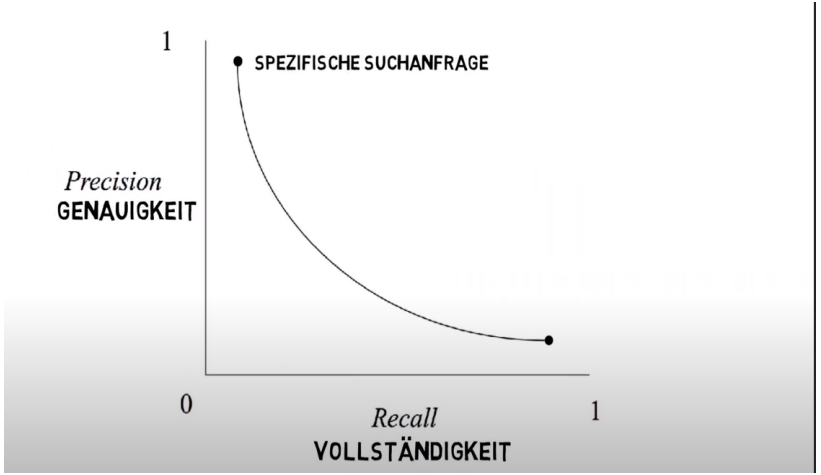
Term Document Matrix

- Term Document Matrix
- Indexieren
- Tokenization
- Rechenoperation
- Nachteile

Inverted Index

- Indexieren
- Tokenization
- TF und IDF
- Rechenoperation

Precision und Recall



Problematiken

- Relevanz und Pertinenz
- Natürlichsprachige Suchanfragen
- Unwissenheit des Nutzers
- Zugang zu Daten über den Nutzer

Quellen

https://www.uni-bamberg.de/fileadmin/uni/fakultaeten/wiai_lehrstuehle/medieninformatik/Dateien/Publikationen/2008/henrich-ir1-1.2.pdf

https://www.xovi.de/was-ist-information-retrieval/

https://www.ionos.de/digitalguide/online-marketing/suchmaschinenmarketing/information-retrieval-wie-suchmaschinen-daten-abrufen/

https://www.youtube.com/playlist?list=PLaZQkZp6WhWwoDuD6pQCmgVyDbUWl_ZUi