# Information Retrieval

## Was ist Information Retrieval?

Wörtlich übersetzt bedeutet **Information Retrieval** die **Rückgewinnung von Informationen**. Laut Definition bezieht sich Information Retrieval auf einen Prozess, bei dem auf Basis eines Informationswunsches aus einem großen unstrukturierten Datenbestand die Informationen herausgesucht werden, die zum Informationswunsch passen.

Information Retrieval gehört damit zu den zentralen Aufgaben einer Suchmaschine: Suchmaschinen sind Informations- und Datensammler. Die gesammelten Daten werden ausgewertet, aufbereitet, gespeichert und wiedergewonnen.

## Bedeutung von Information Retrieval

Die immer größere Menge digital verfügbarer Dokumente bringt zugleich die Forderung nach einer **schnellen, zielgerichteten Suche** mit sich. Im klassischen Sinne bezieht sich das auf die Suche nach Textdokumenten. Die Rückgewinnung von Informationen muss aber grundsätzlich für alle multimedialen Dokumente möglich sein.

Neben der Hauptanwendung Suchmaschinen ist der Information Retrieval Prozess u.a. auch relevant für digitale Bibliotheken, Bilddatenbanken oder multimediale Archive.

Dabei hat die Ausprägung der Suche Einfluss auf die Erfordernisse bzw. Methoden des Information Retrieval. Dieser Einfluss äußert sich z.B. folgendermaßen:

- **Datenbestand, in dem gesucht wird:** große Unterschiede zwischen selbstverwaltetem Datenbestand und Datenbestand im Internet
- **Informationswunsch:** konkrete vs. eher vage Vorstellung bei der Suche
- **Dokumententyp:** Texte in verschiedenen Formaten (z.B. doc-, pdf-, html-Datei), Videos, Bilder, Audio-Dateien

Problematisch bei der Auswahl der passenden Informationen ist zudem das **unsichere Wissen des Information Retrieval Systems**, d.h. es hat keine Kenntnis von den Dokument-Inhalten. Das Retrieval System kann nur bestimmte Methoden nutzen, z.B. Textstatistik oder Termgewichtung, hat aber

Probleme mit bestimmten Wortverwendungen, z.B. mit Synonymen oder Homonymen.

Um den Informationswunsch besser erfüllen zu können, also ein besseres Ergebnis liefern zu können, gibt es beim Information Retrieval verschiedene Möglichkeiten, den Suchwunsch genauer einzuordnen, z.B. indem der Kontext der Suche berücksichtigt wird – genau das machen Suchmaschinen wie Google. Die Suchmaschine bezieht dafür bspw. frühere Anfragen mit ein.

## Modelle des Information Retrieval

Es gibt verschiedene Retrieval-Modelle, von denen einige aufeinander aufbauen. Zu den wichtigsten **Information Retrieval Modellen** gehören:

### Boolesches-Modell

- ältestes Information-Retrieval-Modell basierend auf der Booleschen Logik von 1854
- Inhalte werden nur anhand der Operatoren “und”, “oder”, “nicht” gefunden
- die Inhalte werden nicht sortiert – **es gibt kein Ranking der Ergebnisse**

### Linktopologisches Modell

- basiert nicht auf der Auswertung der Dokumentinhalte, sondern auf der **Auswertung der Verlinkungsstruktur zwischen Dokumenten** – dadurch ergibt sich ein Ranking der Dokumente
- die Struktur erlaubt eine Aussage zur Autorität von Dokumenten
- hierzu gehört bspw. der PageRank **von Google**, entwickelt von Larry Page und Sergey Brin

### Textstatistik

- Untersuchung der Terme innerhalb eines Dokuments
- Gewichtung geschieht über **WDF und IDF**
- **WDF:** Within Document Frequenzy – relative Häufigkeit eines Terms innerhalb eines Dokuments
- **IDF:** Inverse Document Frequency – Häufigkeit, mit der ein Dokument mit einem bestimmten Term in einer Datenbank vorkommt
- **Vektormodell** gehört ebenfalls zum Textstatistik-Modell: Jeder Text entspricht einem Punkt im Raum, die Winkel der Vektoren geben Hinweise zur Ähnlichkeit der Wörter zueinander

### Cluster Modell

- Zusammenfassung von Dokumenten nach Ähnlichkeit
- kann das Suchverfahren beschleunigen, da nur der Zugriff auf einen Dokumentenpool erforderlich ist
- Probleme können entstehen, wenn die Cluster nicht vollständig oder sehr umfangreich sind

## Wie nutzen Suchmaschinen Information Retrieval?

Jede Internetsuchmaschine nutzt Information Retrieval, um Suchanfragen zu bearbeiten. Bei Suchmaschinen ist es wichtig, die “ermittelten” Informationen zu bewerten und nach Wichtigkeit/Relevanz zu sortieren – daraus ergibt sich das Ranking. Sobald Sie einen Suchbegriff in das Suchfeld eingeben, gibt die Suchmaschine relevante Informationen zu Ihrem Suchbegriff aus den gespeicherten Datenbeständen wieder (SERP).

Mit SEO wird demzufolge versucht, die Rückgewinnung der Informationen der optimierten Seite zu verbessern – einer der Maßnahmen ist bspw. die WDF*IDF Optimierung von Webseiten

**Weiterführende Informationen:** Information Retrieval – Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web.

SEO-Küche Internet Marketing GmbH & Co. KG
Fraunhoferstr. 6, 83059 Kolbermoor
Telefon 08031 / 2575-100
Telefax 08031 / 2575-101
E-Mail: info@seo-kueche.de

SEO-Küche Internet Marketing GmbH & Co. KG,
HRA 11167 AG Traunstein
pers. Haftende Gesellschafterin:
SEO-Küche Verwaltungs GmbH, Kolbermoor, HRB 22414 AG Traunstein
Geschäftsführer: Christian Brunnenmayer, Patrick Keller, Oliver Lindner
Ust-IdNr.: DE 286 985 708, Steuer Nr. 156/174/08500

HypoVereinsbank
IBAN DE45 700202700015260147
BIC HYVEDEMMXXX