Information Retrieval

Einführung und Überblick

von

Markus Schütze

Überblick

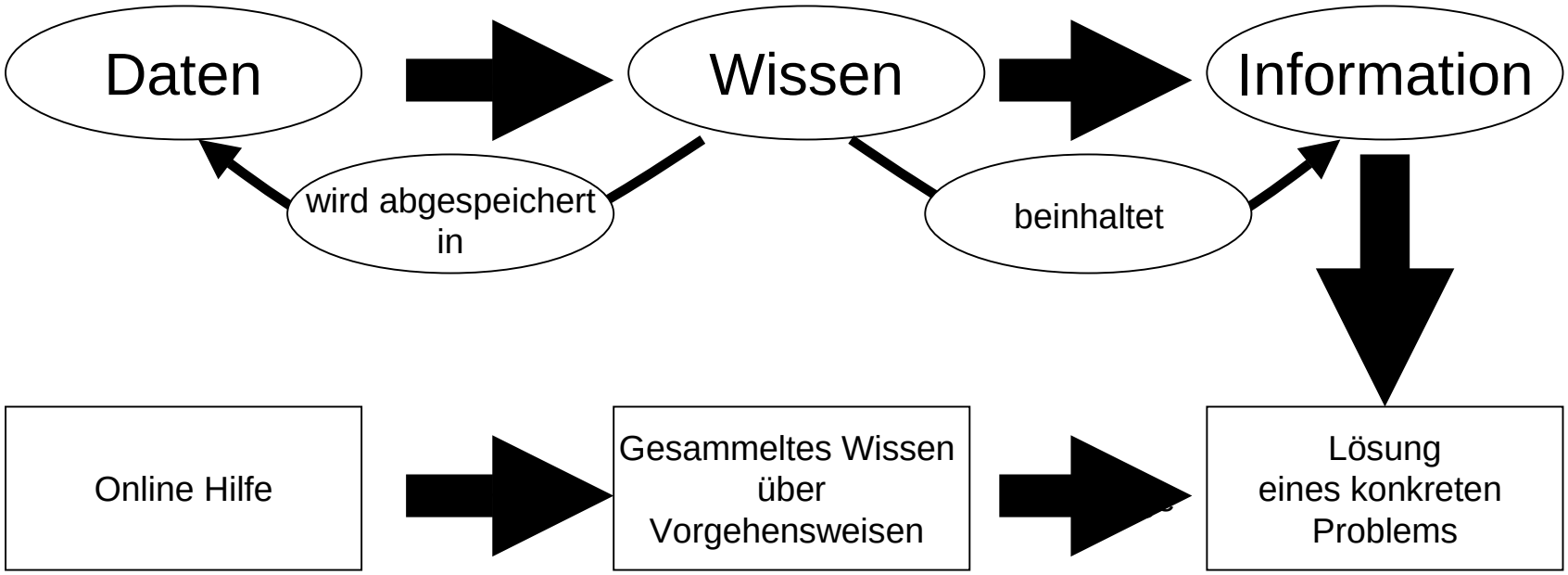
- Grundlegende Begriffe
  - Information
  - Information Retrieval
  - Indextermerstellung
- Information Retrievalmodelle (IR)
  - Überblick
  - formale Definition des IR-Prozesses
  - Gewichtung
  - Boolsches Retrival Modell
  - Vektorbasiertes IR
  - Probabilistisches IR
- Retrieval Evaluation
  - Wichtige Größen
  - Recall und Precision
  - TREC

Der Begriff ‚Information‘

Information ist : *Die Teilmenge von Wissen die von einer bestimmten Person oder Gruppe in einer konkreten Situation zur Lösung von Problemen benötigt wird.*

oder anschaulicher

Information ist Wissen in Aktion



Was ist ‚Information Retrieval' ?

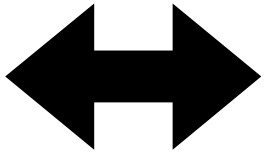
„Gegenstand des Information Reterieval ist die Repräsentation, Speicherung und Organisation von Information und der Zugriff zu Informationen. Dabei gibt es grundsätzlich keine Einschränkungen in der Art der Informationen."

Gerard Salton

Information Retrieval vs. Data Retrieval

Data Retrieval

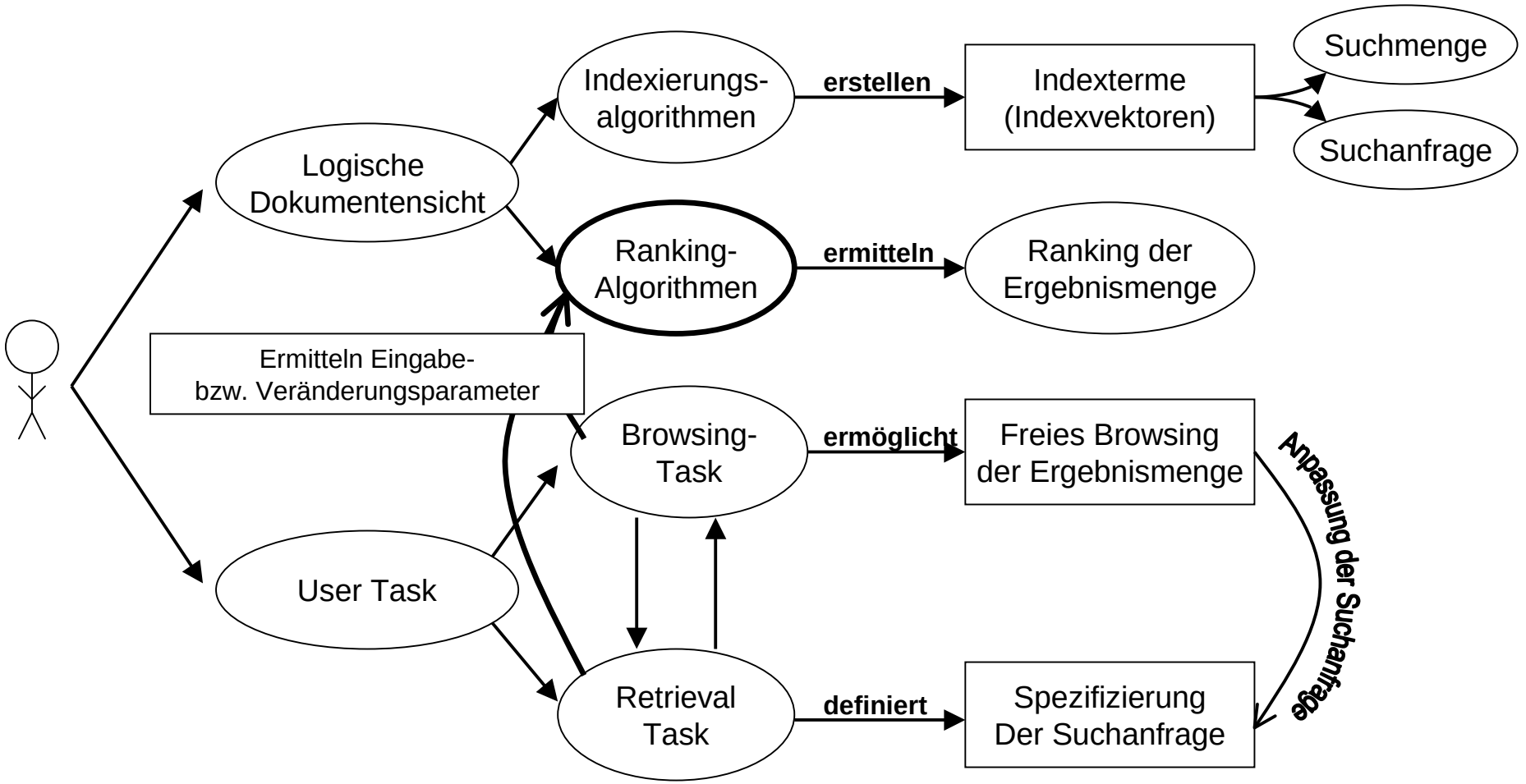
- exact matching
- deterministisches Modell
- Formale Anfragesprache
- Vollständige Fragespezifikation
- gesuchte Objekte: Fragespezifikation erfüllende
- Sensitive Reaktion auf Eingabefehler



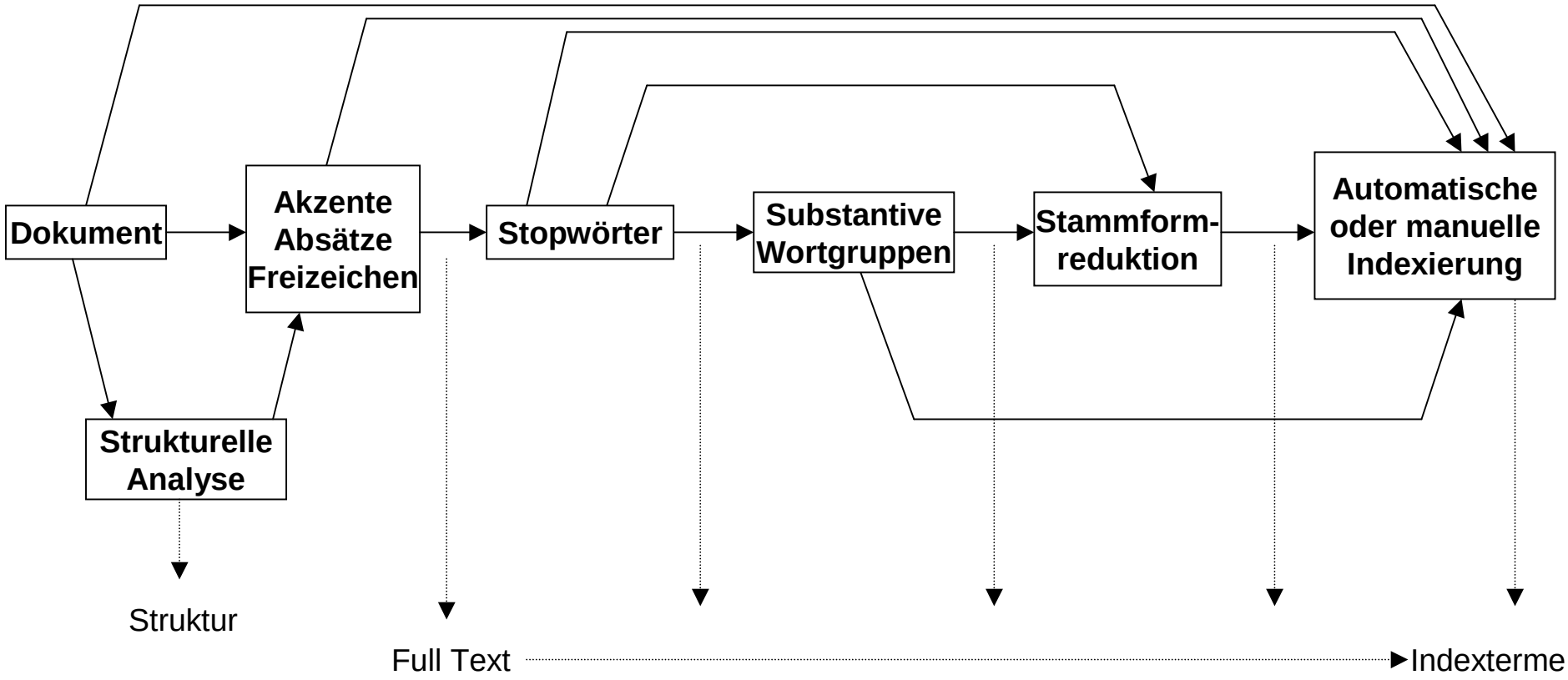
Information Retrieval

- Partiell, best match
- Probabilistisches Modell
- Natürliche Anfragesprache
- Unvollständige Fragespezifikation
- Gesuchte Objekte: relevante Suchobjekte
- Insensitive Reaktion auf Eingabefehler

Der Retrieval Prozess



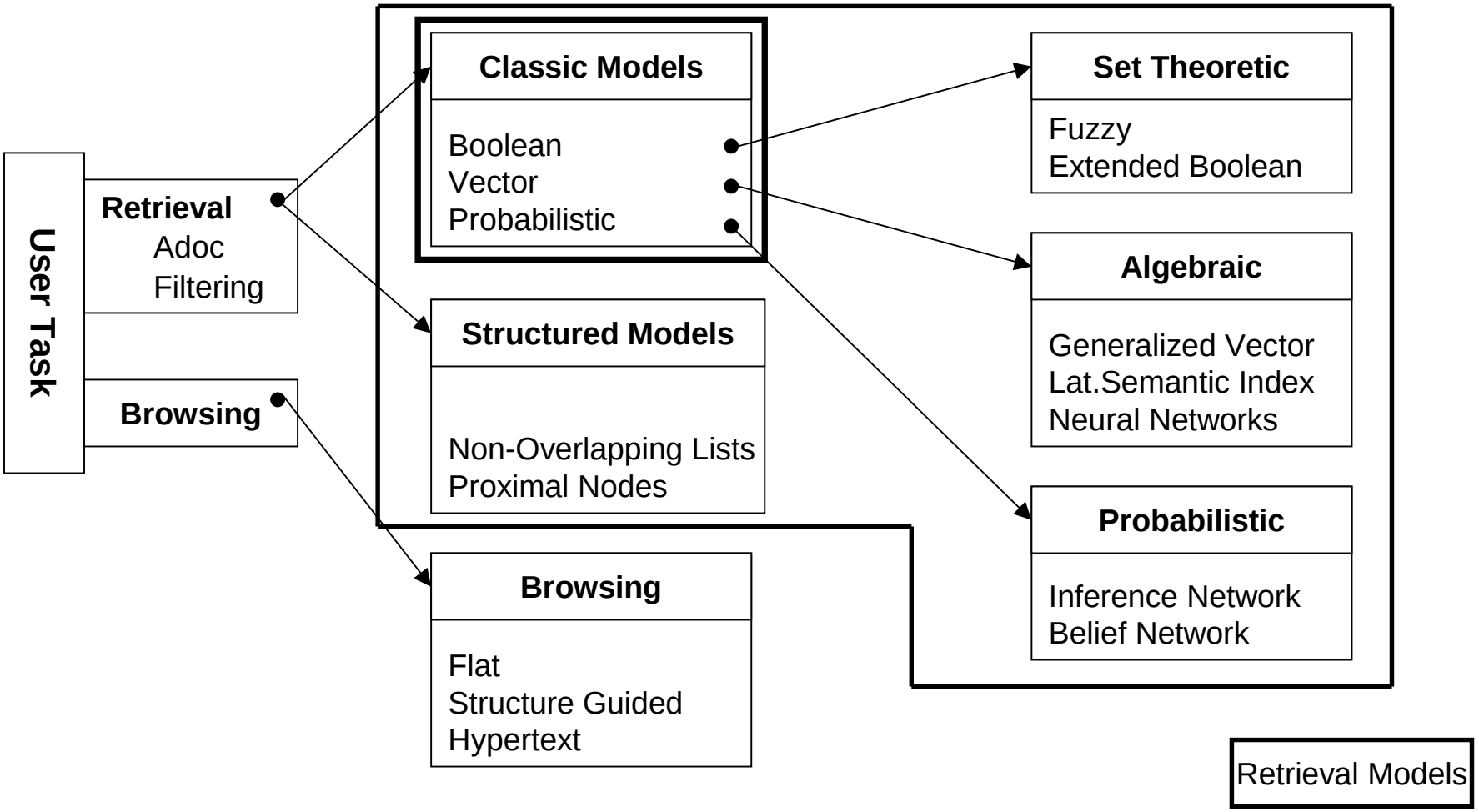
Die Indextermerstellung



Jedes Dokument bekommt einen Indextermvektor $d_j = (k_1, ..., k_i)$ zugeordnet.

mit i = max. aller ermittelten Indexterme

Retrieval Algorithmen



Formelle Definition des Information Retrieval

- *Ein Information Retrieval System ist ein 4er-Tupel [D,Q,F,R(q_i,d_j)] mit*
  - ***(1)*** ***D*** *ist ein Set aus zusammengesetzten logischen Sichten der Dokumente der Suchmenge*
  - ***(2)*** ***Q*** *ist die Suchanfragemenge*
  - ***(3)*** ***F*** *ist ein Schema, welches beschreibt wie Dokumente und Suchanfragen modelliert sind*
  - ***(4)*** ***R(q_i,d_j)]*** *ist eine Rankingfunktion, die eine reele Zahl zwischen einer Suchanfrage $q_i \varepsilon Q$ und einem Dokument $d_j \varepsilon D$ abbildet*

Ranking : definiert die Reihenfolge, die Relevanz zwischen Dokumenten d, die in Beziehung

Um Ranking durchzuführen, Müssen wir die Schlüsselwörter, Die Suchanfrage und Dokumente beschreiben bzw. **gewichten**

Gewichtung von Indextermen

Sei k_i ein Schlüsselwort (Indexterm), d_j ein Dokument und $w_{i,j} >= 0$
Dann ist eine Gewichtung azzoziiert mit dem Paar $(d_j, w_{i,j})$.

Seit t die Anzahl von Schlüsselwörtern innerhalb des IR_Systems und k_i ein im Dokument d_j. Sei $K = \{ k_1 , .., k_i \}$ der gesamte Satz vergebener Schlüsselwörter. Eine Gewichtung $\boldsymbol{w_{i,j}}$ wird für jedes Schlüsselwort k_i, das im Dokument erscheint vergeben.

Jedes Dokument d_j ist mit einen Gewichtungsvektor

$$d_{i,j} = (w_{1,j}, w_{2,j}, \ldots, w_{i,j})$$

Die Funktion g_i liefert uns das Gewicht zurück, das mit dem Schlüsselwort k_i t-dimensionalen Vektor $d_{i,j}$ vorkommt, folglich :

$$g_i(d_{i,j}) = w_{i,j}$$

Das Boolsche Modell

- Einfaches Modell
  - **binärer** Entscheidungsalgorithmus
- Basierend auf boolscher Algebra
  - **große Beliebtheit**
- einfach und formal
  - Queries haben **präzise** Semantik
- aber :
  - für Benutzer ist es sehr schwierig , seine natürlichsprachliche Anfrage in einen boolschen Ausdruck umzuformen.
  - **kein partielles Matching**

Die Rankingfunktion

- Gewichtung der Schlüsselwörter k_i von Suchanfrage und Dokumenten:

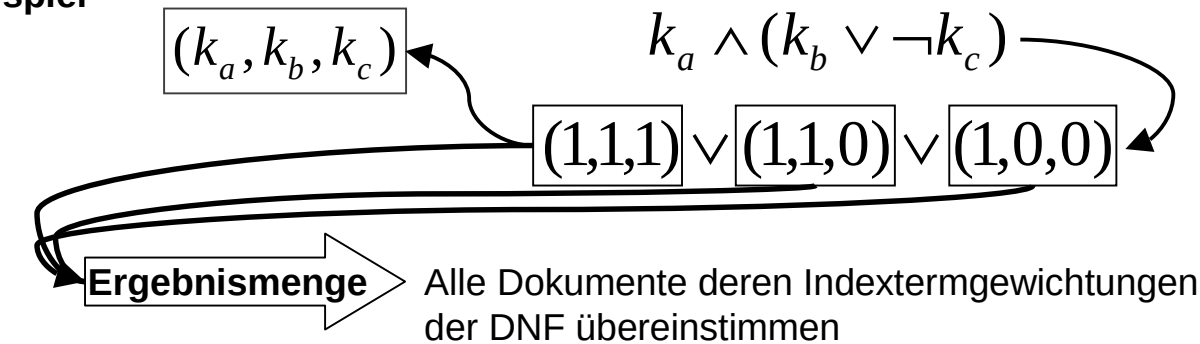
$$w_{i,j} = \{0,1\}$$

- Spezifizierung der Suchanfrage :

  beliebige Anzahl von Schlüsselwörtern, die mit and, or, not zusammengesetzt werden

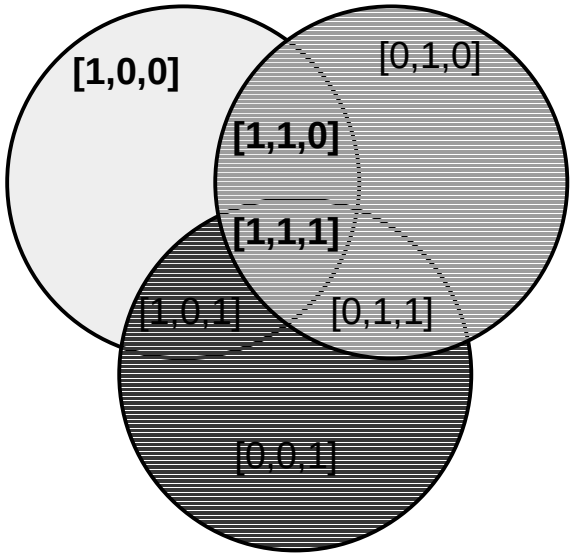
  » interne Repräsentation als Disjunktion von Konjunktionen (**Disjunktive Normalform** (DNF))

Beispiel



Alle Dokumente deren Indextermgewichtungen mit einem Element der DNF übereinstimmen

Graphische Veranschaulichung



Als relevant eingestufte Dokumentteilmengen : [1,0, 0] ; [1,1,0] ; [1,1,1]

Was ist aber mit [0,1,0] ???

Formal

$$sim(d_j, q) = \begin{cases} 1 & falls \quad \exists q_{cc} \mid (q_{cc} \in q_{DNF}) \wedge (\forall k_i, g_i(d_j) = g_i(q_{cc})) \\ 0 & anderfalls \end{cases}$$

mit

$w_{i,j} \in \{0,1\}$

q_{DNF}	Disjunktive Form der Query
q_{cc}	Komponente von q_{DNF}

Vektormodell

- Berücksichtigt partielles Matching im Gegensatz zum boolschen Retrieval Modell
- Berechnet den Grad der Abweichung zwischen Suchanfrage und jedem Element der Suchmenge
- Beruht auf algebraischer Vektorrechnung
- Präsentiert als Ergebnis des Retrievalprozesses dem Benutzer eine aufsteigendsortierte Ergebnisliste

Gewichtung

- Im vektorriellen Ansatz werden Suchanfrage und jedes Element der Suchmenge mit einen Gewichtungsvektor ‚versehen‘
- Die einzelnen Komponenten der Vektoren werden gebildet durch die Gewichtung der vergebenen Indexterme $w_{i,j}$. $w_{i,j}$ ist hier aber eine positive nicht binäre Zahl.

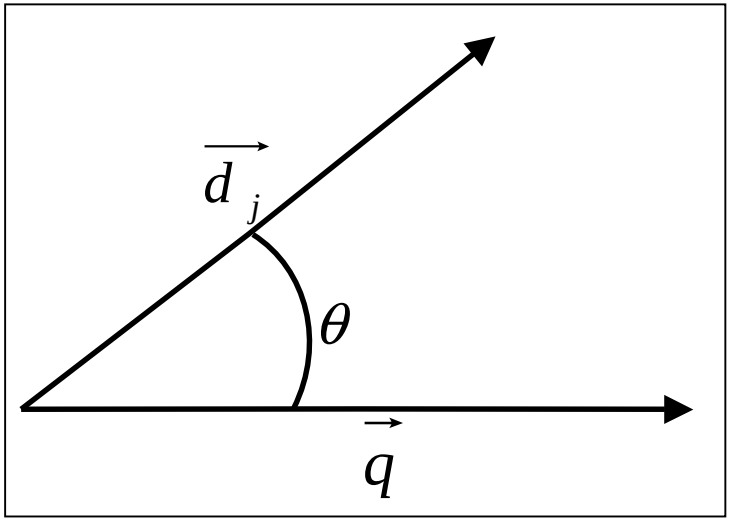
Suchanfrage-vektor

$$\vec{q} = (w_{1,q}, w_{2,q}, ..., w_{t,q})$$

Dokument-vektor

$$\overrightarrow{d_j} = (w_{1,j}, w_{2,j}, ..., w_{t,j})$$

Die Rankingfunktion



- Berechnet den Winkel zwischen Suchanfragevektor und Dokumentenvekor durch Cosinus. (mit Hilfe des Skalarproduktes)
- Je kleiner der Winkel zwischen Suchanfragevektor und Dokumentenvektor ist, desto relevanter wird das Dokument für den Benutzer eingestuft

Die Rankingfunktion formal

$$\theta = sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|}$$

$$= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w^2_{i,j} \times \sum_{j=1}^{t} w^2_{j,q}}}$$

Normierung

- Um im Vektormodell aussagekräftige Vergleiche zwischen Dokumenten ermitteln zu können, müssen die einzelnen Gewichtungen der Indexterme normiert werden.
- Wichtiges Maß hierfür :
  - Die Häufigkeit des Auftretens eines Indextermes in der Menge der Suchdokumente
  - Wird verfeinert in weitere Maße :
    - Dokument frequency
    - Inverse Dokument frequency

Dokument frequency

- Beschreibt die Häufigkeit des Auftrehtes eines Indextermes in einem Dokument

$$f_{i,j} = \frac{\text{Häufigkeit eines Schlüsselwortes im Dokument}}{\text{max aller Schlüsselwörter im Dokument}}$$

Inverse Dokument frequency

- Beschreibt das Auftrehten eines Indexterms in in allen Dokumenten der Suchmenge
- Beschreibt den Effekt, dass Schlüsselworte, die in vielen Dokumenten vorkommen, schlechtere Kandidaten sind.

$$idf_i = \frac{\#\,\text{aller Dokumente}}{\#\,\text{der Dokumente in welchen der Indexterm } k_i \text{ erscheint}}$$

Normalisierte Gewichtungen

- Für Dokumentvektoren

$$w_{i,j} = f_{i,j} \times \log \frac{\#\,\text{aller Dokumente}}{\#\,\text{aller Dokumente mit Indexterm } k_i}$$

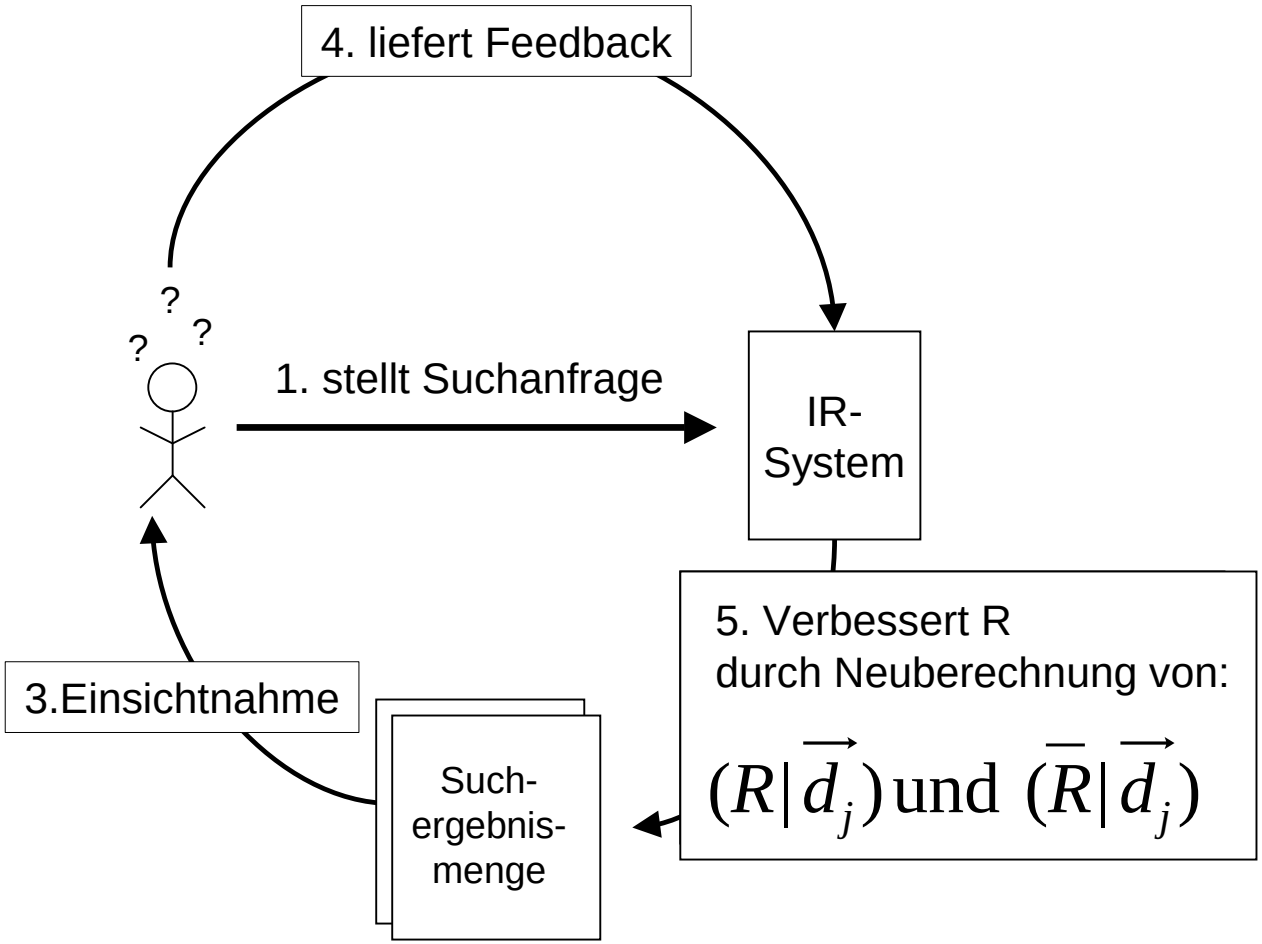
- Für Suchanfragevektoren

$$w_{i,q} = (0.5 + \frac{0.5 \times \#\,\text{vom Indexterm } k_i}{\max \text{aller Schlüsselwörter in Query}} \times \frac{\#\,\textit{Dokumente}}{\#\,\text{der Dokumente mit Indexterm } k_i})$$

Probabilistisches Modell

- Nimmt an , dass eine Menge existiert, die genau die Wünsche des Benutzers enthält.
  = ideal answer set R
- Ermittelt eine Grundergebnismenge, die iterativ im Laufe des IR-Prozesses verbessert wird.
- basiert auf Statistik- und Stochastikalgorithmen
- Gewichtungen von Dokumenten und Suchanfrage sind hier wieder binär $w_{i,j}$= { 0, 1 }

Grundidee


4. liefert Feedback
? ? ?
1. stellt Suchanfrage
IR-System
5. Verbessert R
durch Neuberechnung von:
$(R|\vec{d}_j)$ und $(\bar{R}|\vec{d}_j)$
3.Einsichtnahme
Such-ergebnis-menge

Kritik

- Das Modell berücksichtigt, dass der Benutzer oft nicht weiß, welche Informationen er genau sucht
- Am Anfang des Retrievalprozesses ist nicht klar was R ist
- Die Häufigkeit der Indexterme in Dokumenten wird nicht berücksichtigt

Retrieval Evaluation

- Beschäftigt sich mit der Ermittlung von Retrievaleffizienz und –effektivität
  - Retrievaleffizienz ist die Größe die Kosten und Zeit des Retrievalvorganges beschreibt
  - Retrievaleffektivität misst die Fähigkeit des IR-Systems, Informationen nachzuweisen, die der Benutzer auch benötigt

⇨ Retriavaleffiziens und –effektivität bestimmen somit die Leistung eines Informationretrievalsystems

Wichtige Kenngrößen

- Recall
- Precision
- Aufwand zur Formulierung der Suchanfragen
- Zeit
- Form der Ergebnisrepräsentation
- Abdeckung mit der Datenbank

Recall und Precision

- Recall: Fähigkeit des Systems, alle relevanten Daten nachzuweisen

$$\text{Recall} = \frac{\#\,\text{der nachgewiesenen relevanten Dokumente}}{\#\,\text{aller relevanten Dokumente}}$$

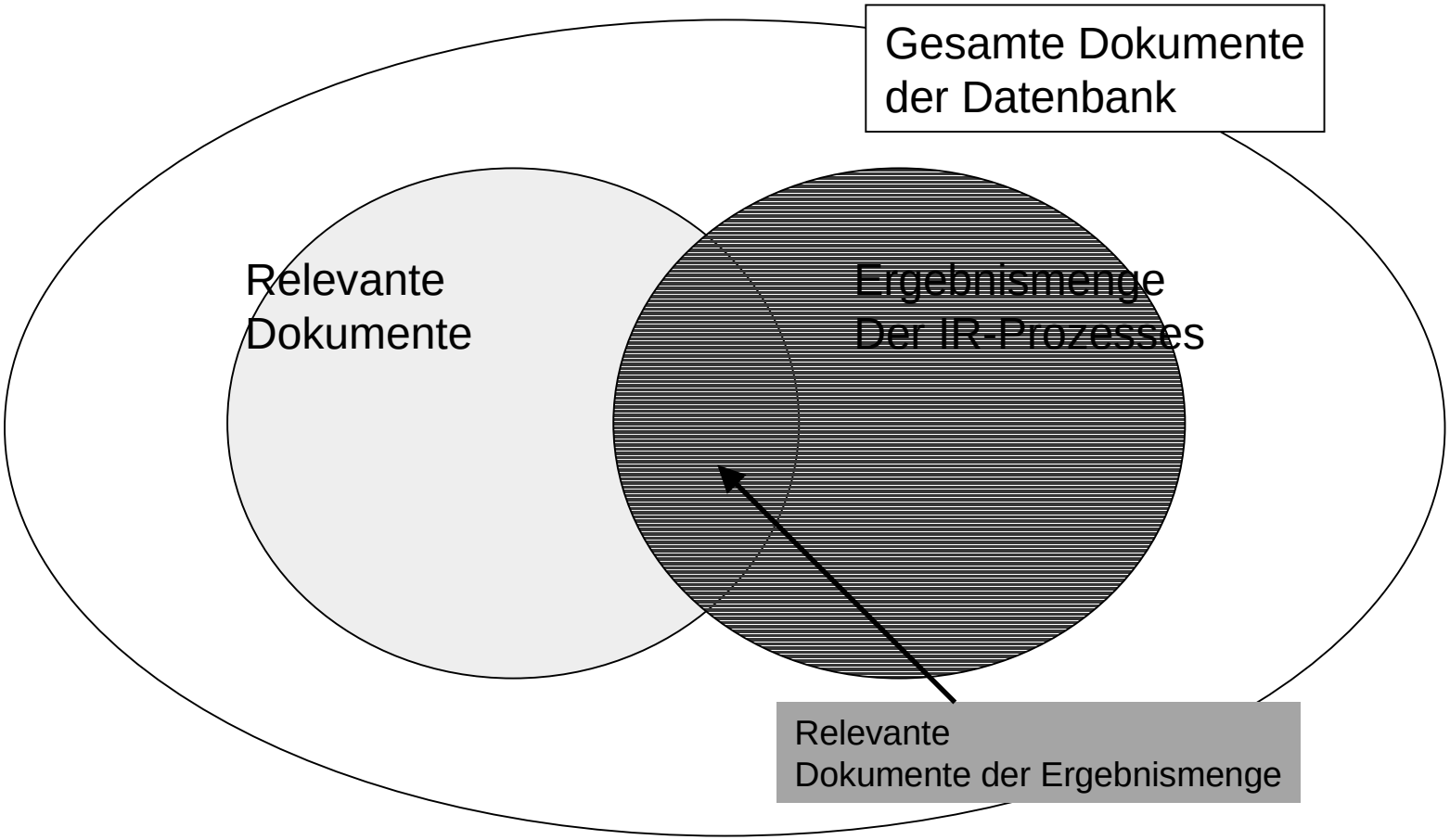
Maß für die Quantitätseinschätzung

- Precision: Fähigkeit des Systems, nur relevanten Dokumente nachzuweisen

$$\text{Precision} = \frac{\#\,\text{der nachgewiesenen relevanten Dokumente}}{\#\,\text{aller nachgewiesenen Dokumente}}$$

Maß für die Qualitätseinschätzung

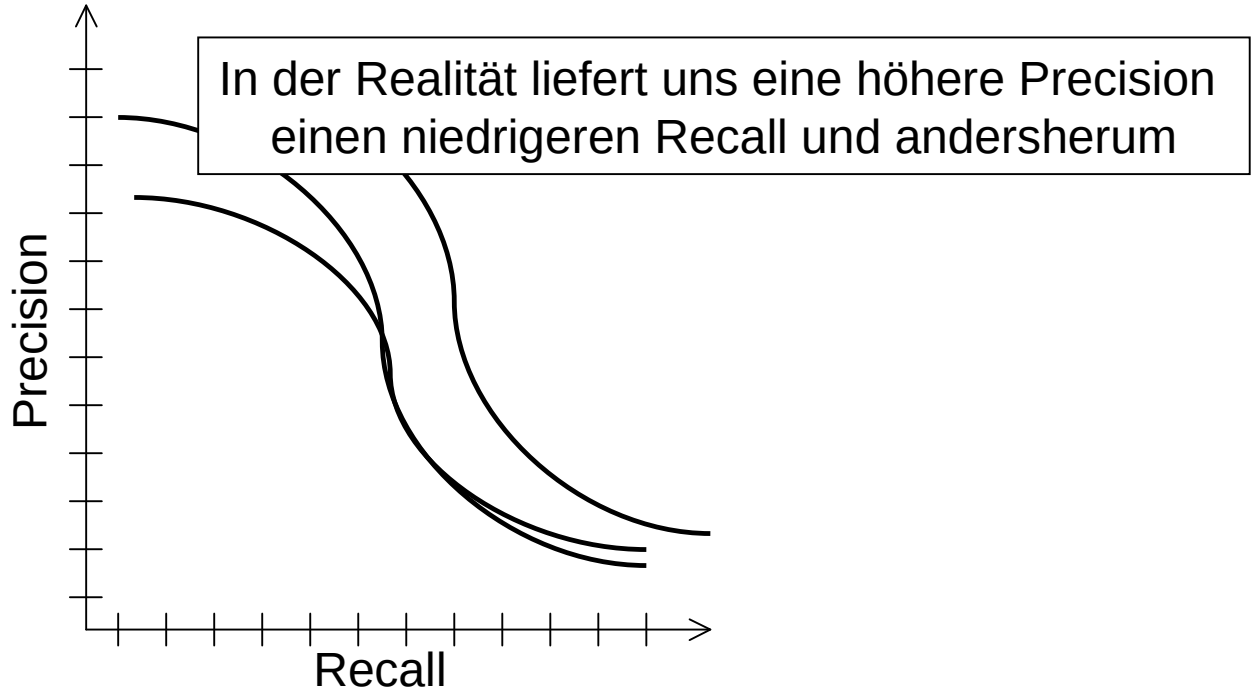
Recal und Precision veranschaulicht



Recall und Precision Kurven

- Recall und Precision eingetragen in eine gemeinsames Diagramm, liefern uns eine gute Möglichkeit die Leistung abzuschätzen, die ein IR-System bietet.

Skizze



TREC

- TREC = **T**ext **Re**trieval **C**onference
- Neben europäischen CLEF eine der maßgeblichen Instanzen, die für die Definition von normierten Test-Retrieval-Collektionen zuständig ist

- **Ziel :** Vergleichbare Benchmarks zu schaffen
- **Mittel:** Zusammenstellung einer wohl definierten Dokumentenmenge (Collektionen) und darauf spezifizierten Suchanfragen

Die TREC-Collektion

WSJ	Wall Street Journal
AP	Assoziated Press
ZIFF	Computer Selects
FR	Fedarel Register
DOE	US DOE Publications
SKMN	San Jose Mercury News
PAT	US Patens
FT	Financiel Times
CR	Congressional Records
FBIS	Foreign Broadcast Information Service
LAT	LA Times

- Besteht aus mehren Submengen von Dokumenten unterschiedlichster Richtungen

Beispiel eines Dokumentes

Spezifiziert mit Metadatenstruktur (SGML)
Ermöglicht automatisches Parsen der Dokumente

```
<doc>
<docno> WSJ880406-0090 </docno>
<h1> AT&T Unveils Services to Upgrade Phone Networks Under Global Plan <h1>
<author> Janet Guyon (WSJ Staff) </author>
<dateline> New York </dateline>

<text>
Amerikan Telephone & Telegraph Co. Introduced the first of new
generation of phone services with broad ....
</text>

</doc>
```

Beispiel einer Suchanfrage

Werden in der TREC-Collection TOPIC‘S genannt

<top>

<num> Number: 168 </num>
<title> Topic: Financing AMTRAK </title>

<desc> Description:
A document will adress the role of the Federal Goverment in
Financing the operation of National railroad Transportation Coporation (AMTRAK)

<narr> Narrative: A relevant document must provide information on
The goverment‘s responsibility to make AMTRAK an economically viabla entity.
It could also discuss the privatization of AMTRAK as an alternative to continuing
Goverment subsides given to air and bus transportation with
Those provided to AMTRAK would also be relevant.
</narr>

</top>

Wird vom gleichen Algorithmus geparst,
der auch das zur TOPIC passende Dokument parst
Und in Indexterme zerlegt

TREC-Benchmark

Besteht aus mehreren Subbenchmarks

Beispiele:

Filtering Task, Routing Task

Ermittelt folgende Kenngrößen:

Recall-precision

average Precision Histogram

Dokument Level averages

Summary table statistics

ENDE