Vorlesung

Indexierung, Suche, Relevanz

Stefania Racioppa & Brigitte Jörg

Hans Uszkoreit Vorlesung
Informationswissenschaft
und Informationssysteme

Indexierung

- Informationserschließung
  - Notationen
    - Formale Textbeschreibung
    - Kennzeichnungen in einer künstlichen Sprache
    - z. B. Bibliothekenverzeichnisse, Kataloge, Klassifikationen
  - Stich- vs. Schlagwörter
    - Inhaltliche Textbeschreibung
    - Wörter kommen im Text vor oder stehen in Relation zum Textinhalt
  - Deskriptoren vs. Nichtdeskriptoren
    - Kontrollierte inhaltliche Textbeschreibung
    - Wörter werden anhand eines Thesaurus selektiert

Ermittlung der Deskriptoren

- Extraktion
  - Deskriptoren werden dem Text entnommen
    - Höhere Indexierungskonsistenz

- Addition
  - Deskriptoren stehen in Relation zum Inhalt
    - Reichere Beschreibung
    - Indexierung von Bildern

Koordinierung der Deskriptoren

- Koordinative Indexierung
  - Gefundene Deskriptoren stehen gleichrangig nebeneinander
  - Retrieval durch einzelne Deskriptoren oder logische Verküpfungen

- Strukturierte Indexierung
  - Syntaktische Beziehung der gefundenen Deskriptoren bleiben erhalten
  - Genauere Inhaltswiedergabe
  - Retrieval durch feste Fügungen

Indexierungsmethoden

- Freie Indexierung
  - Alle gefundenen Deskriptoren sind zugelassen
  - Stoppwörter?

- Kontrollierte Indexierung
  - Thesaurusbasierte Extraktion
  - Schlagwortkatalog

Indexierungsmethoden

- Intellektuelle Indexierung
  - Intellektuelle Inhaltsanalyse
  - Manuell vergebene, repräsentative Schlagwörter
  - Kontrolliertes Vokabular
- Computergestützte Indexierung
  - Indizierung
  - Intellektuelle Vor- oder Nachbereitung
- Automatische Indexierung
  - Maschinelle Indexierung ohne Vor-/Nachbereitung

Automatische Indexierung

- Freitextverfahren
  - Alle Textwörter in Textform
  - Stoppwörter werden ausgeschlossen
  - Suchmaschinen im WWW (Trunkierung)

- Statistisches Verfahren
  - Häufigkeitsanalyse
  - Termgewichtung, inverse Dokumenthäufigkeit

- Informationslinguistische Verfahren

Informationslinguistische Verfahren

- Morphologisch-lexikalisches Verfahren
  - Morphologische Analyse
  - Wortformen- und Stammlexika
    - Arbeits- und kostenintensiv
- Morphosyntaktisches Verfahren
  - Textanalyse auf Wort- und Satzebene
  - Dependenzanalyse
    - Verhältnismäßig komplexe Lösungen
- Semantisches Verfahren
  - Tiefensemantische Beschreibung der Dokumentinhalte
  - Rollenindikatoren, Thesaurusrelationen

Zusammenfassung

- Textbeschreibung (Index)
  - Notationen
  - Stich- bzw. Schlagwörter
  - Deskriptoren
- Erschließung
  - Extraktion
  - Addition
- Koordinierung
  - Gleichordnend
  - Syntaktisch
- Indexierungsmethoden
  - Frei
  - Kontrolliert
- Indexierungstechniken
  - Intellektuell
  - Computergestützt
  - Automatisch
    - Freitextverfahren
    - Statistisches Verfahren
    - Informationslinguistisches Verfahren
      - Morphologisch-lexikalisches Verfahren
      - Morphosyntaktisches Verfahren
      - Semantisches Verfahren

IDX

- Automatische Indexierung
  - Freie Indexierung
  - Koordinative Indexierung
- Freitextverfahren
  - Morphologisch-lexikalisches Verfahren
  - Semantisches Verfahren (Thesaurus)
  - Statistisches Verfahren (Ansätze)

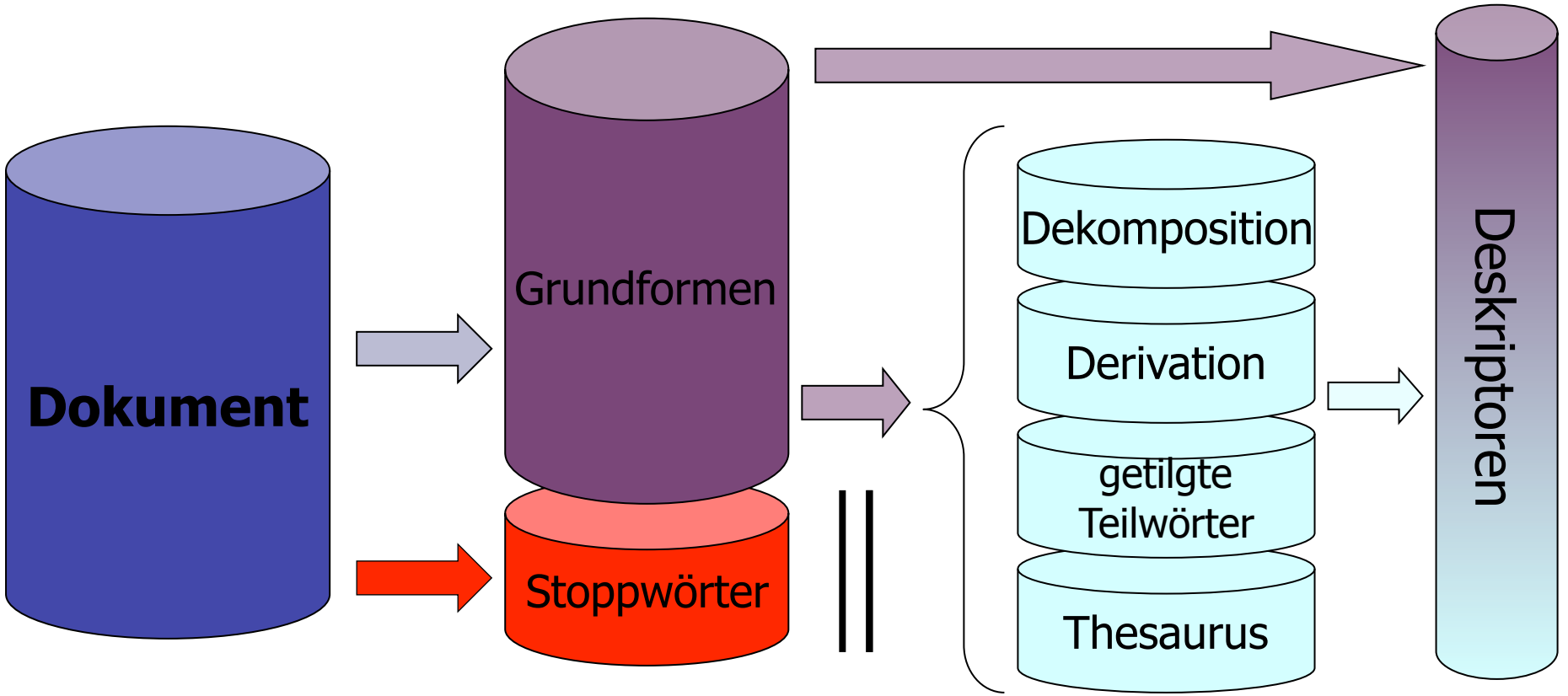
IDX: Komponenten

- **Identifikationswörterbuch**
  - Stammlexikon
  - Textwortformen → Grundformen
  - Dekomposition
  - Rechtschreibkontrolle, Normalisierung
    - Primus Korrekturmanager (**Brockhaus Duden Neue Medien)**

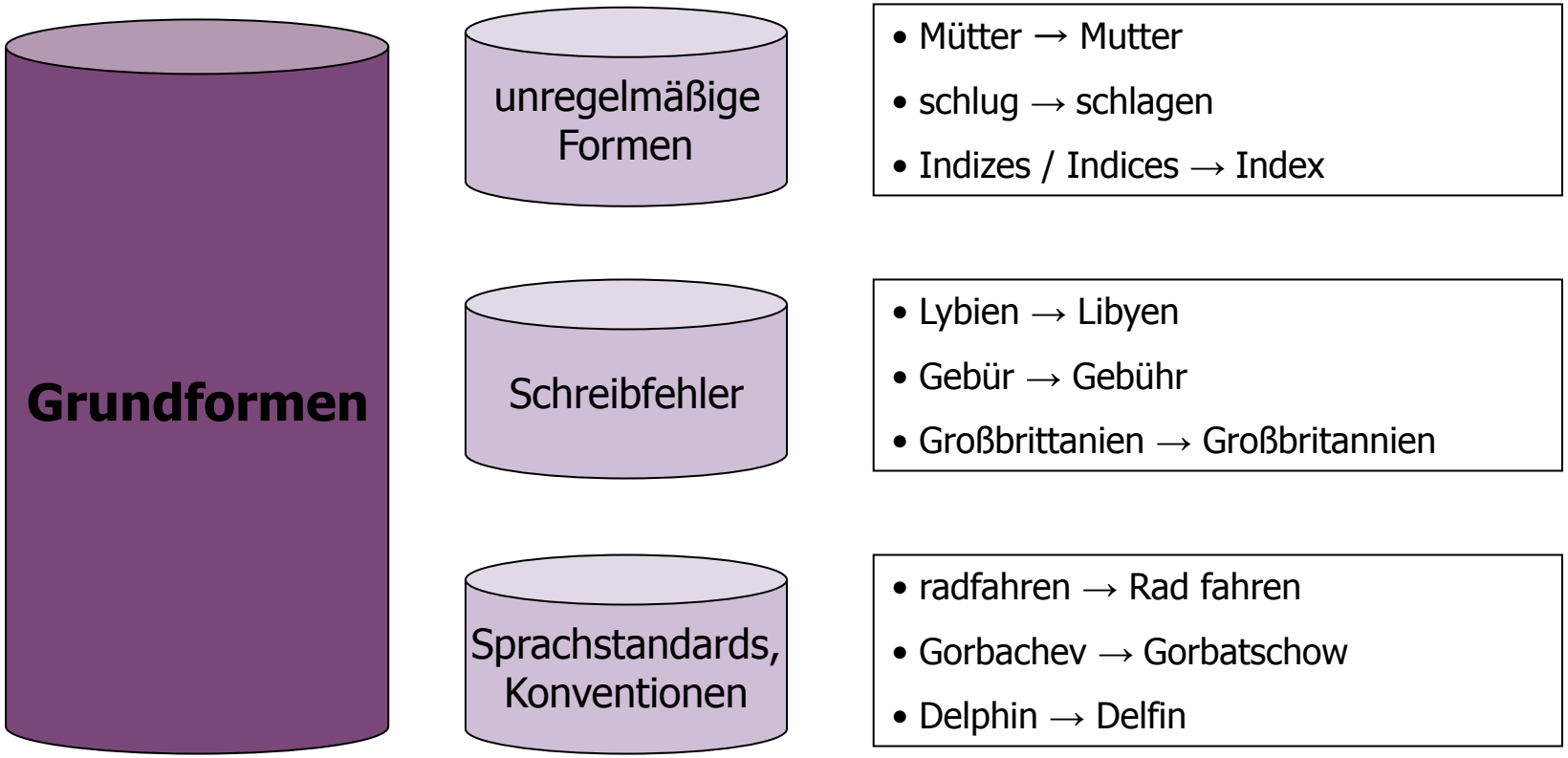
- **Relationenwörterbuch**
  - Sinnvolle Kompositazerlegung
  - Stoppwörter
  - Thesaurus
  - Disambiguierung

- **(opt.) Übersetzungswörterbuch**
  - Übersetzung auf Wortbasis

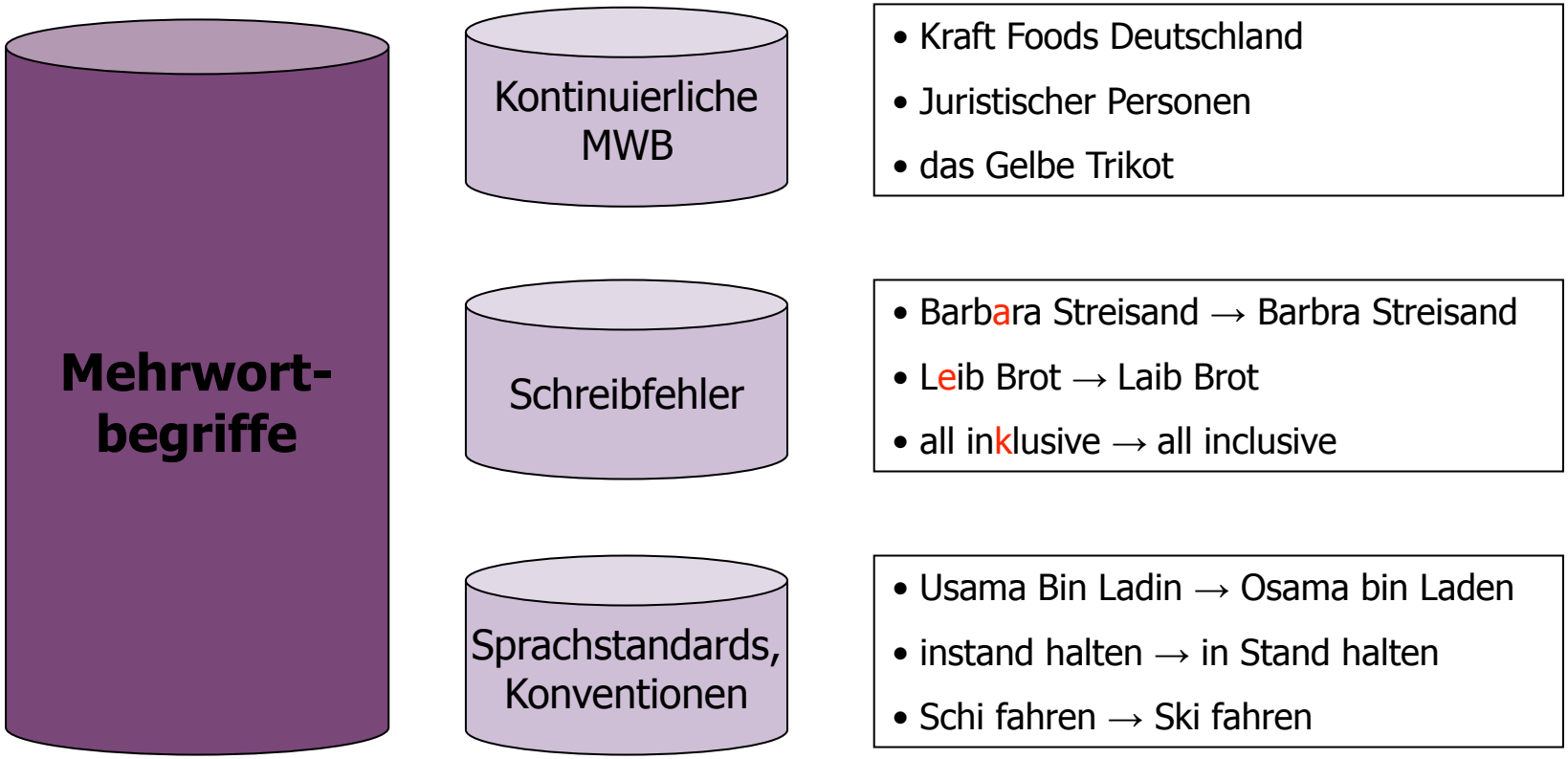
IDX: Informationsextraktion



Phase 0: Grundformermittlung



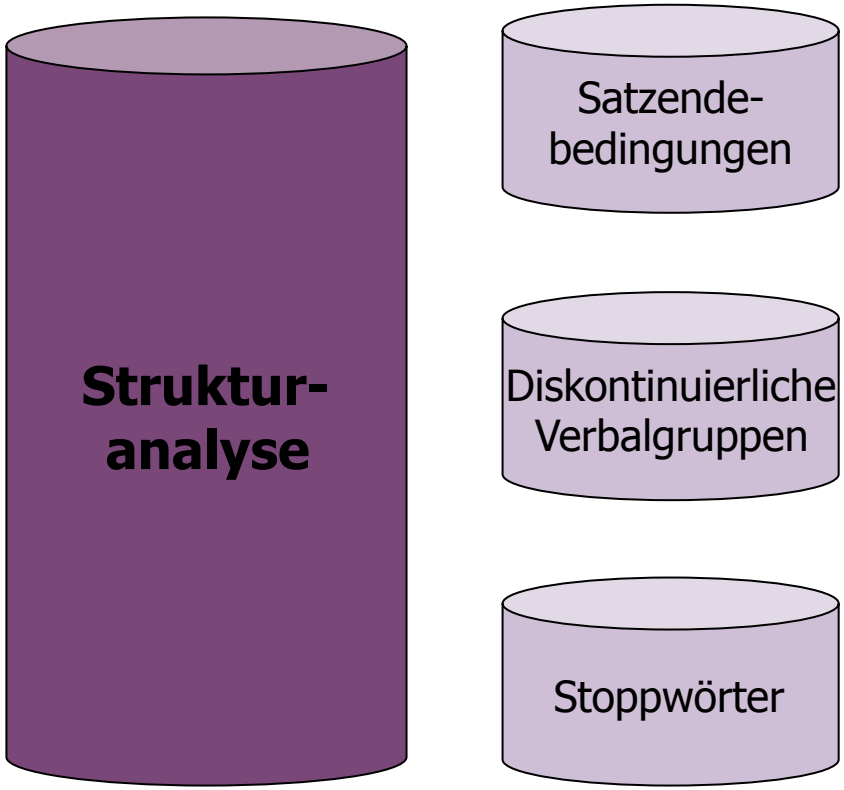
Phase M: Mehrwortbegriffe



Phase B: getilgte Teilwörter

- Wiederaufbau der Tilgung
  - Haus- und Hofwirtschaft
    - → Hauswirtschaft und Hofwirtschaft
- Keine semantische Verifikation

Phase 1: Strukturanalyse



- ... zu finden. Essen ist eine ...
- ... beim Grillen. Im Garten ist. ...

- ging von ihr weg → weggehen
- setzten den Kanzler in Kenntnis → in Kenntnis setzen

- wieder
- auf keinen Fall

Phase 2: Dekomposition

sinnvolle Bestandteile

- Bundeswehretat → Bundeswehr + Etat
- Computerbildschirm → Computer + Bildschirm
- Bundestagsdebatte → Bundestag + Debatte

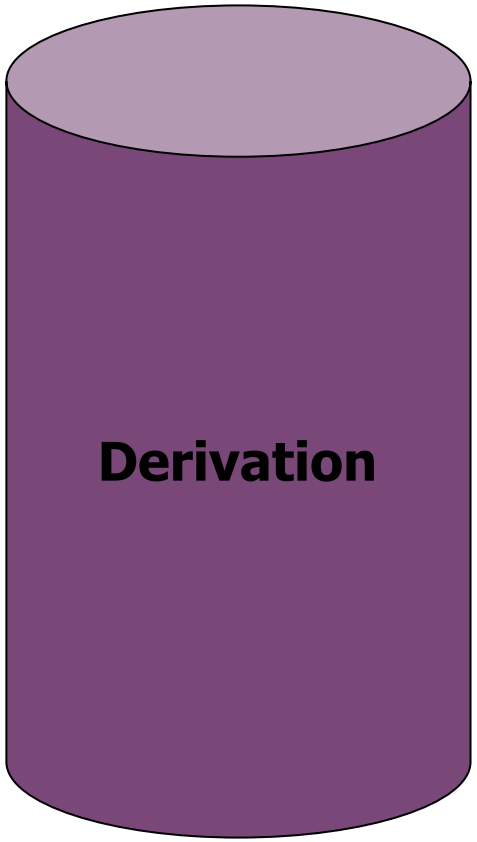
wenig sinnvoll

- Bundeswehretat → Bund + Wehretat ?
- Computerbildschirm → Computerbild + Schirm ?
- Bundestagsdebatte → Bund + Tagsdebatte ?

sinnlos

- Verbrechen → Verb + Rechen ???
- neunmalklug → neunmal + klug ???
- Opposition → OP + Position ???

Phase 2: Derivation



sinnvolle Bestandteile

- buchen → Buchung
- besteuern → Besteuerung
- beschließen → Beschluss

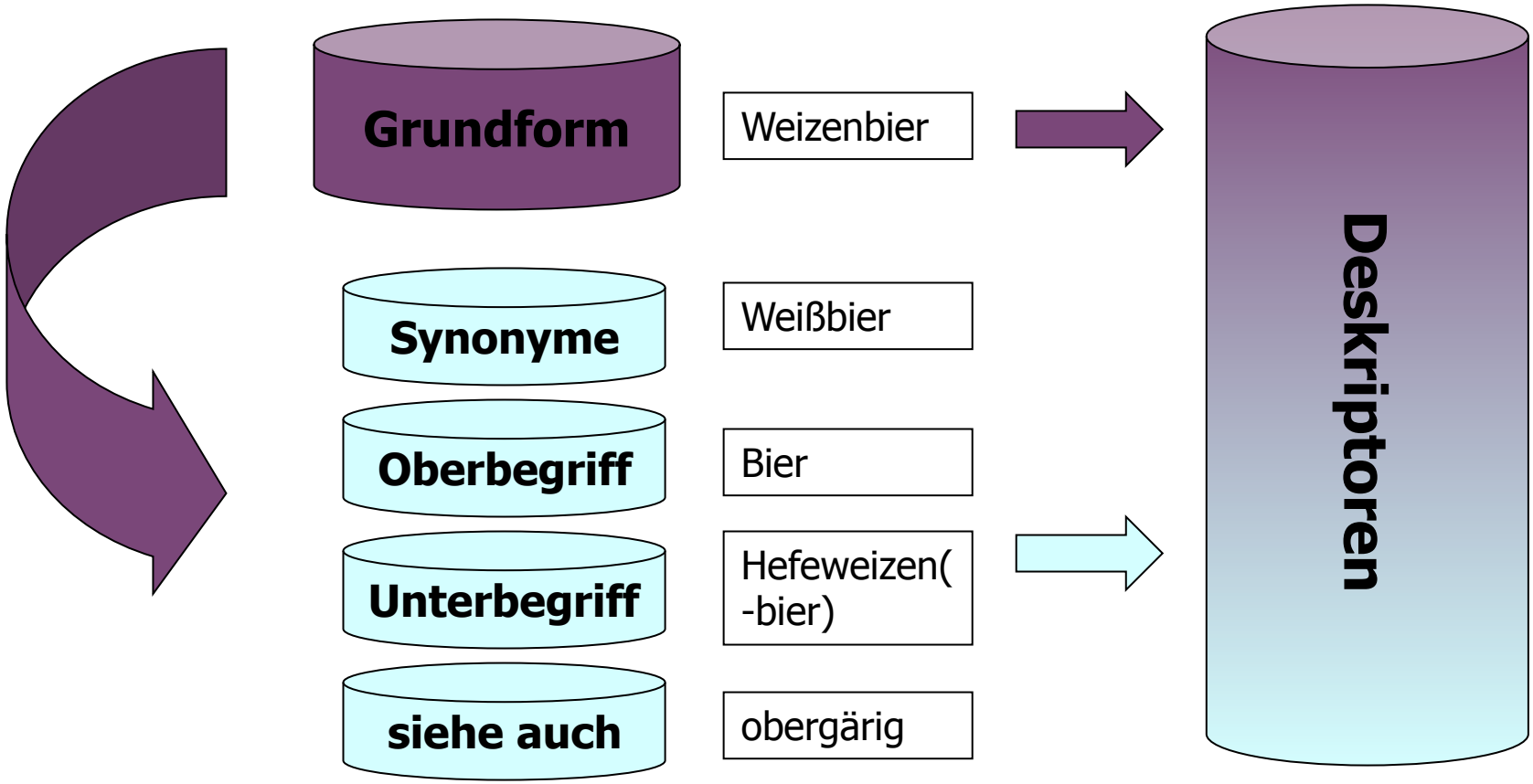
wenig sinnvoll

- stundet → Stunde?
- saßen → Sitzung ?
- raten (welche Bedeutung) → Rat ?

sinnlos

- gleichen → Gleichung ???
- malen → Mal ???
- buchen → Buch ???

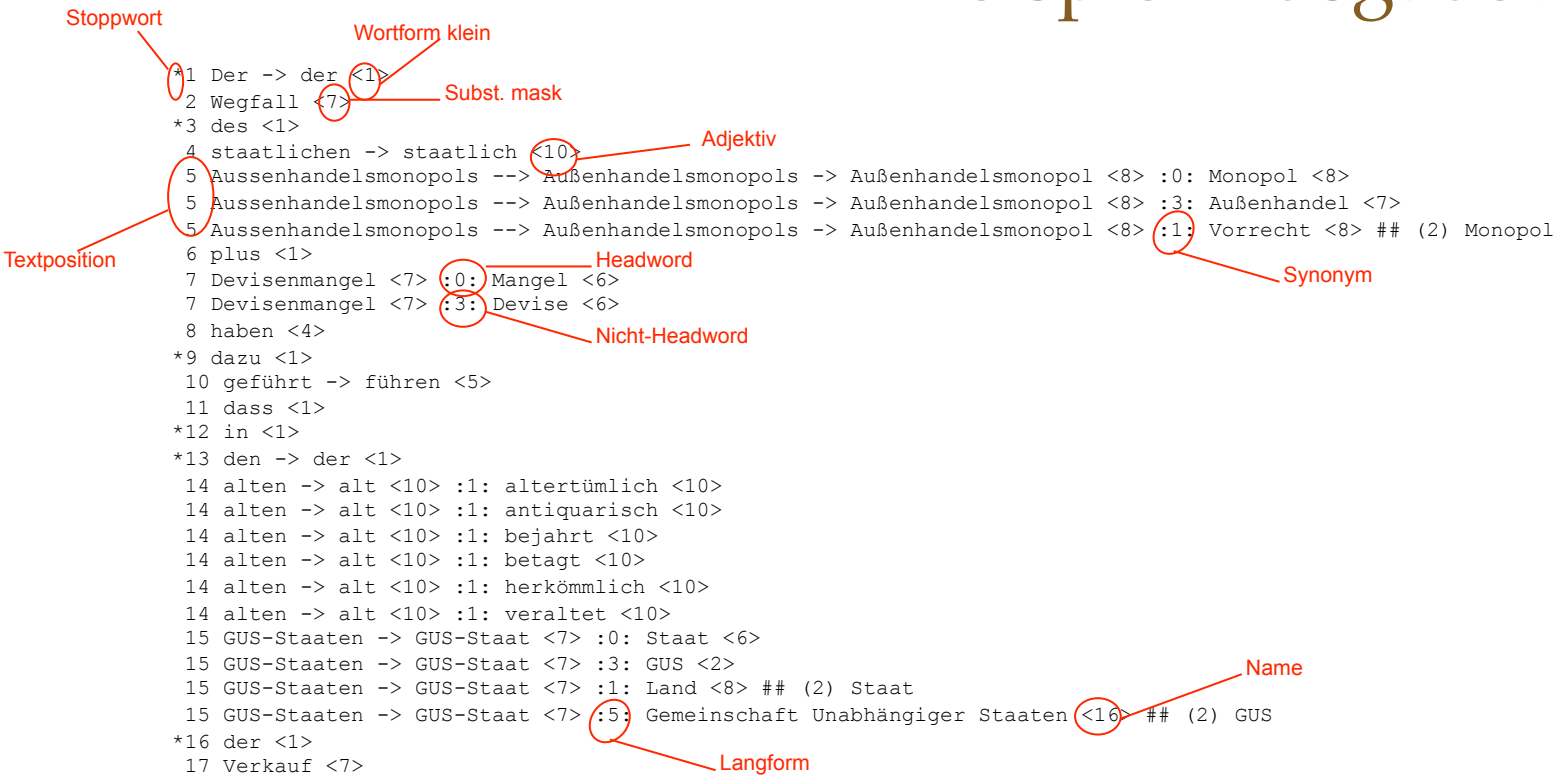
Phase G: Thesaurus



Phase T: Übersetzung

- Übersetzung der gefundenen Deskriptoren
  - Rang → position
- Ausgabe der semantischen Informationen
  - Rang → ≤44≥ dimension
- Ausgabe der Lesarten
  - Rang → <= Dienstgrad> rank

Beispiel Ausgabedatei




```
*1 Der -> der <1>
 2 Wegfall <7>
*3 des <1>
 4 staatlichen -> staatlich <10>
 5 Aussenhandelsmonopols --> Außenhandelsmonopols -> Außenhandelsmonopol <8> :0: Monopol <8>
 5 Aussenhandelsmonopols --> Außenhandelsmonopols -> Außenhandelsmonopol <8> :3: Außenhandel <7>
 5 Aussenhandelsmonopols --> Außenhandelsmonopols -> Außenhandelsmonopol <8> :1: Vorrecht <8> ## (2) Monopol
 6 plus <1>
 7 Devisenmangel <7> :0: Mangel <6>
 7 Devisenmangel <7> :3: Devise <6>
 8 haben <4>
*9 dazu <1>
 10 geführt -> führen <5>
 11 dass <1>
*12 in <1>
*13 den -> der <1>
 14 alten -> alt <10> :1: altertümlich <10>
 14 alten -> alt <10> :1: antiquarisch <10>
 14 alten -> alt <10> :1: bejahrt <10>
 14 alten -> alt <10> :1: betagt <10>
 14 alten -> alt <10> :1: herkömmlich <10>
 14 alten -> alt <10> :1: veraltet <10>
 15 GUS-Staaten -> GUS-Staat <7> :0: Staat <6>
 15 GUS-Staaten -> GUS-Staat <7> :3: GUS <2>
 15 GUS-Staaten -> GUS-Staat <7> :1: Land <8> ## (2) Staat
 15 GUS-Staaten -> GUS-Staat <7> :5: Gemeinschaft Unabhängiger Staaten <16> ## (2) GUS
*16 der <1>
 17 Verkauf <7>
*18 von <1>
 19 Westwaren -> Westware <6> :0: Ware <6>
 19 Westwaren -> Westware <6> :3: West <7>
*20 gegen <1>
 21 Geld <8>
*22 nur <1>
*23 noch <1>
*24 in <1>
 25 Einzelfällen => Einzelfälle -> Einzelfall <7> :0: Fall <7>
 25 Einzelfällen => Einzelfälle -> Einzelfall <7> :3: Einzel <8>
 26 funktioniert -> funktionieren <5>
```

Beispiel Ausgabedatei

```
*1 Der -> der <1>
 2 Wegfall <7>
*3 des <1>
 4 staatlichen -> staatlich <10>
 5 Aussenhandelsmonopols --> Außenhandelsmonopols -> Außenhandelsmonopol <8> :500: Monopol <8>
 5 Aussenhandelsmonopols --> Außenhandelsmonopols -> Außenhandelsmonopol <8> :503: Außenhandel <7>
 5 Aussenhandelsmonopols --> Außenhandelsmonopols -> Außenhandelsmonopol <8> :1: Vorrecht <8> ## (2) Monopol
 6 plus <1>
 7 Devisenmangel <7> :0: Mangel <6>
 7 Devisenmangel <7> :3: Devise <6>
 8 haben <4>
*9 dazu <1>
 10 geführt -> führen <5>
 11 dass <1>
*12 in <1>
*13 den -> der <1>
 14 alten -> alt <10> :1: altertümlich <10>
 14 alten -> alt <10> :1: antiquarisch <10>
 14 alten -> alt <10> :1: bejahrt <10>
 14 alten -> alt <10> :1: betagt <10>
 14 alten -> alt <10> :1: herkömmlich <10>
 14 alten -> alt <10> :1: veraltet <10>
 15 GUS-Staaten -> GUS-Staat <7> :0: Staat <6>
 15 GUS-Staaten -> GUS-Staat <7> :3: GUS <2>
 15 GUS-Staaten -> GUS-Staat <7> :1: Land <8> ## (2) Staat
 15 GUS-Staaten -> GUS-Staat <7> :5: Gemeinschaft Unabhängiger Staaten <16> ## (2) GUS
*16 der <1>
 17 Verkauf <7>
*18 von <1>
 19 Westwaren -> Westware <6> :0: Ware <6>
 19 Westwaren -> Westware <6> :3: West <7>
*20 gegen <1>
 21 Geld <8>
*22 nur <1>
*23 noch <1>
*24 in <1>
 25 Einzelfällen => Einzelfälle -> Einzelfall <7> :0: Fall <7>
 25 Einzelfällen => Einzelfälle -> Einzelfall <7> :3: Einzel <8>
 26 funktioniert -> funktionieren <5>
```


Suche

Vorraussetzung: Annahme

- über eine Struktur von Dokumenten
- über eine Struktur von Anfragen

Dokument:

- Menge / Multimenge von Termen (z.B.: Schlagwörter, Stichwörter)
- Menge von Notationen (künstliche Bezeichnung)

=> Deskriptoren (Index)

Term / Suchbegriff:

- Wort; mehrgliedriger Begriff; komplexes Freitextmuster
- Beschreibung (auch gewichtet - je nach Suchmodell bzw. Indexierungsverfahren)

Suchmodelle – Grundlagen

T: $\{t_1, ..., t_n\}$: Menge aller Terme in einer Dokumentenkollektion (Indexierungsvokabular)

Q: Menge aller erlaubten Anfragen des jeweiligen Suchmodells

q: Frageformulierung

d: Dokument

$\vec{\mathbf{d}}$: $\{d_1, ..., d_n\}$: Beschreibung des Dokumentes als Vektor von Indexierungsgewichten, wobei d_i das Gewicht von d für den Term T_i angibt.

Boolesches Suchmodell

Frageterme sind ungewichtet (Gewicht: 0 oder 1) durch **Boolsche Operatoren** miteinander verknüpft. Die Menge Q der erlaubten Anfragen kann man wie folgt definieren:

- **jeder Term $t_i \in T$ ist eine Anfrage**
- **q**
- **NOT q**
- **q_1 AND q_2**
- **q_1 OR q_2**

Fuzzy-Suchmodell

Verwendet die gleiche Struktur der Anfragen wie die Boolsche Suche, allerdings in Kombination mit gewichteter Indexierung (beschränkt auf das Intervall [0,1]. Gewichtete Anfragen sehen dann wie folgt aus:

- **jeder gewichtete δ Term $t_i \in T$ ist eine Anfrage**
- **$\delta(q, \vec{d})$**
- **$\delta(\text{NOT } q, \vec{d}) := 1 - \delta(q, \vec{d})$**
- **$\delta(q_1 \text{ AND } q_2, \vec{d}) := \delta(q_1, \vec{d}) \cdot \delta(q_2, \vec{d})$**
- **$\delta(q_1 \text{ OR } q_2, \vec{d}) := \delta(q_1, \vec{d}) + \delta(q_2, \vec{d}) - \delta(q_1, \vec{d}) \cdot \delta(q_2, \vec{d})$**

Fuzzy-Suchmodell - Beispiel

Ein **Dokument** d hat die folgenden Indexierungsgewichte:

- 0.9 Alpen
- 0.5 Rodeln
- 0.8 Abfahrtsski
- 0.3 Skilanglauf

Für die Anfrage **q=„Alpen AND (Rodeln OR Skilanglauf)“** ergibt sich dann folgende gewichtete Anfrage:

$$\delta(q_1,d) \cdot (\delta(q_2, d) + \delta(q_3, d) - \delta(q_2, d) \cdot \delta(q_3, d))$$

$$0.9\ (0.5 + 0{,}3 - 0.5 \cdot 0.3) = \mathbf{0.584}$$

Suchmodell Vekorraum

Dem Vektorraummodell liegt eine geometrische Interpretation zugrunde, bei der Dokumente und Anfragen als Punkte in einem Vektorraum aufgefasst werden, der durch die Terme der Kollektion aufgespannt wird.

t_i	q_i	d_{1i}	d_{2i}	d_{3i}	d_{4i}
Rodeln	**2**	**1**	**0.5**	**1**	**1**
Skilanlauf	**2**	**1**	**1**	**1**	**1**
Wintersportort	**1**	**1**		**1**	
Alpen	**1**		**1**	**1**	**0.5**
Heli-Ski	**-2**		**1**		
$\delta(q,d_m)$		**5**	**2**	**6**	**4.5**

Tabelle: *Möglicher Fragevektor q_i und vier Beispiel-dokumente mit Indexierungsgewichten*

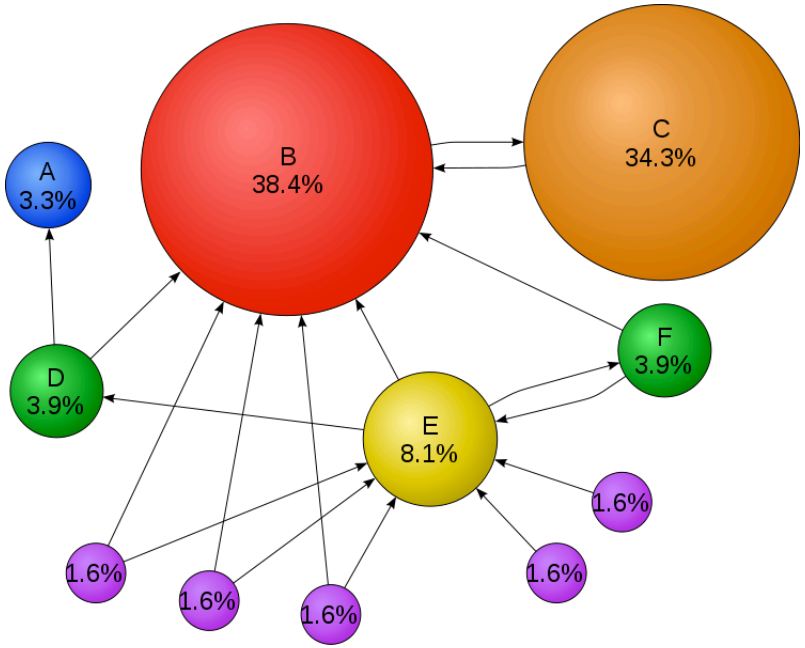
Anfragen werden somit als Vektor $\vec{q}=\{q_1, \ldots q_n\}$ dargestellt, wobei q_i das Fragetermgewicht von q_i für den Term t_i angibt.

Suchanfrage: Wintersportort in den Alpen, der Rodeln und Skilanglauf aber keinen Heli-Ski bietet.

- Boolsche Operatoren möglich
- Suchergebnisse optimiert für die breite Masse
- Gewichtung (Relevanz) gemäß **PageRank**
  beinflusst durch Jon Kleinberg (Authorities,Hubs)
  Eugene Garfield (Citation Analysis)

Google Pagerank

Measuring the relative importance of a document within the set of documents by assigning numerical weighting to each element of a hyperlinked set of documents.



PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important".

Source: http://en.wikipedia.org/wiki/PageRank

YAHOO!

- Boolsche Operatoren möglich
- Suchergebnisse optimiert für die breite Masse
- Gewichtung (Relevanz): „**...** ranks results according to their relevance to a particular query by analyzing the **web page text**, **title** and **description accuracy** as well as its **source**, **associated links**, and **other unique document characteristics**"

- Boolsche Operatoren möglich
  (ggf. NLP Erweiterung in der Zukunft)
- Suchergebnisse optimiert für die breite Masse
- Gewichtung (Relevanz) noch unklar

WolframAlpha

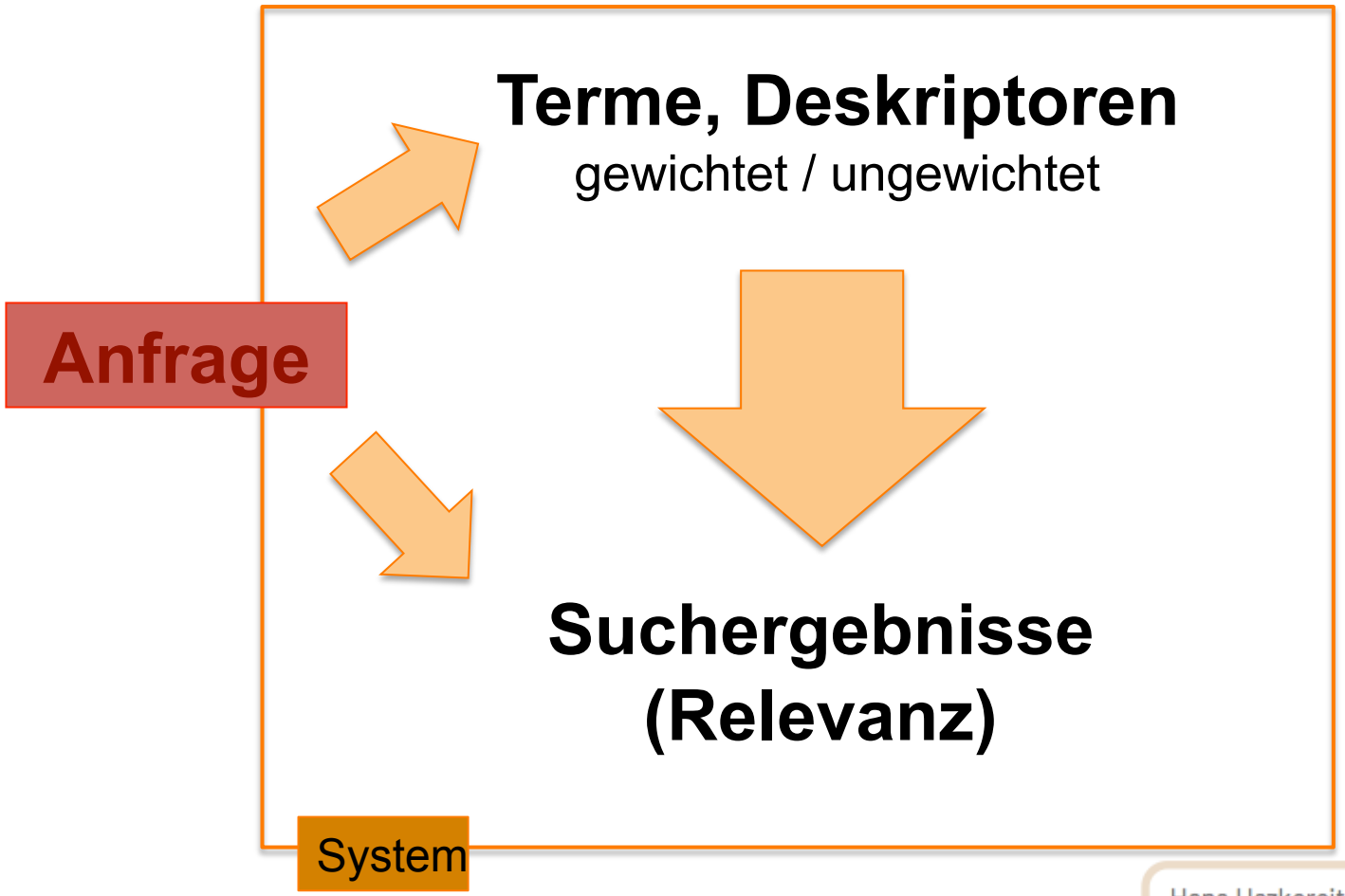
... keine **Suchmaschine –** sondern

... eine **Wissensmaschine ...**

– Unterschiede in der Suchformulierung

– Ergebnisse sind eher konkrete Antworten
als eine Liste von gefundenen Dokumenten

Suche -> Relevanz



Relevanz

- **Wichtige Frage bei der Informationswiedergewinnung**
  - Subjektive Relevanz
  - Objektive Relevanz

- **Betrachtung: Immer in Bezug auf eine Suchanfrage**
  - Genauigkeit der Such-Formulierung
  - Möglichkeiten der Such-Formulierung

Relevance Ranking

“Queries given to search engines or other retrieval systems are often not very specific, and lead to a large number of matching documents. In these cases the retrieval system should have a good estimate of the relevance of the documents to the user's needs, so that "good" documents show up early in the enumeration. A large number of factors should enter into a good ranking method, including the positions of the query terms in the document, linguistic context of the matches, link popularity, classification of the documents, user models etc. "Classical" methods compute a measure of "distance" between the query and the retrieved document, such as TF/IDF or cosine similarity. For hyperlinked documents, methods which make use of the hyperlink structure have proved very effective for relevance ranking. Google was the first large-scale search engine to make use of hyperlink structure for relevance ranking.”

Source: http://www.lt-world.org/

Relevanz

Cranfield-II Experimente (Begründung von Evaluierungsprinzipien)

-> Bewertungsmethoden, -verfahren, -ansätze, zur Messung: wie gut die Systeme in de Lage sind, die an sie gestellten Anforderungen zu erfüllen, relevante Dokumente zu liefern und nicht-relevante zurückzuhalten.

-> Effektivität / Effizienz

-> Ein **effektives** IR-System verfügt über die Fähigkeit, relevante Dokumente wiederaufzufinden und gleichzeitig nicht-relevante zurückzuhalten.

Bei Ranking-Systemen wie den gängigen Suchmaschinen spielt die Positionierung der Ergebnis-objekte zusätzlich eine wichtige Rolle. Es geht darum, die relevantesten Dokumente in den vordersten Rängen der Ergebnislisten zu präsentieren. **Dahinter steht die Annahme, dass ein derartiges System den Benutzer am besten zufriedenstellen wird.**

Relevanz

- Häufig ist es die Relevanzbestimmung, welche Kritik an der Retrievalmessung hervorruft.
- Es wird ein Widerspruch zwischen der statistisch-quantitativen Anwendung von Maßen un der relativ unscharfen , nur schwer in quantitativen Kategorien fassbaren Basis der Relevanzbewertung gesehen.

Das traditionelle Verständnis des Relevanzbegriffes geht von einer Relation zwischen einer bestimmten Anfrage und den Ergebnisdokumenten aus. Die Forderung nach objektiver Relevanzbestimmung durch einen unabhängigen Juror ist jedoch **schwer einlösbar**.

- In neueren Studien hat man sich intensiv mit der Subjektivität von Relevanzurteilen und deren Konsequenzen auseinandergesetzt.

  -> Effektifitätsbewertung: -> siehe Recall / Precision (rein quantitativ)

  -> weitere wichtige Größen z.B.: Zeitverhalten, Benutzungsaufwand, Input- und Outputgestaltung, (Impact) …