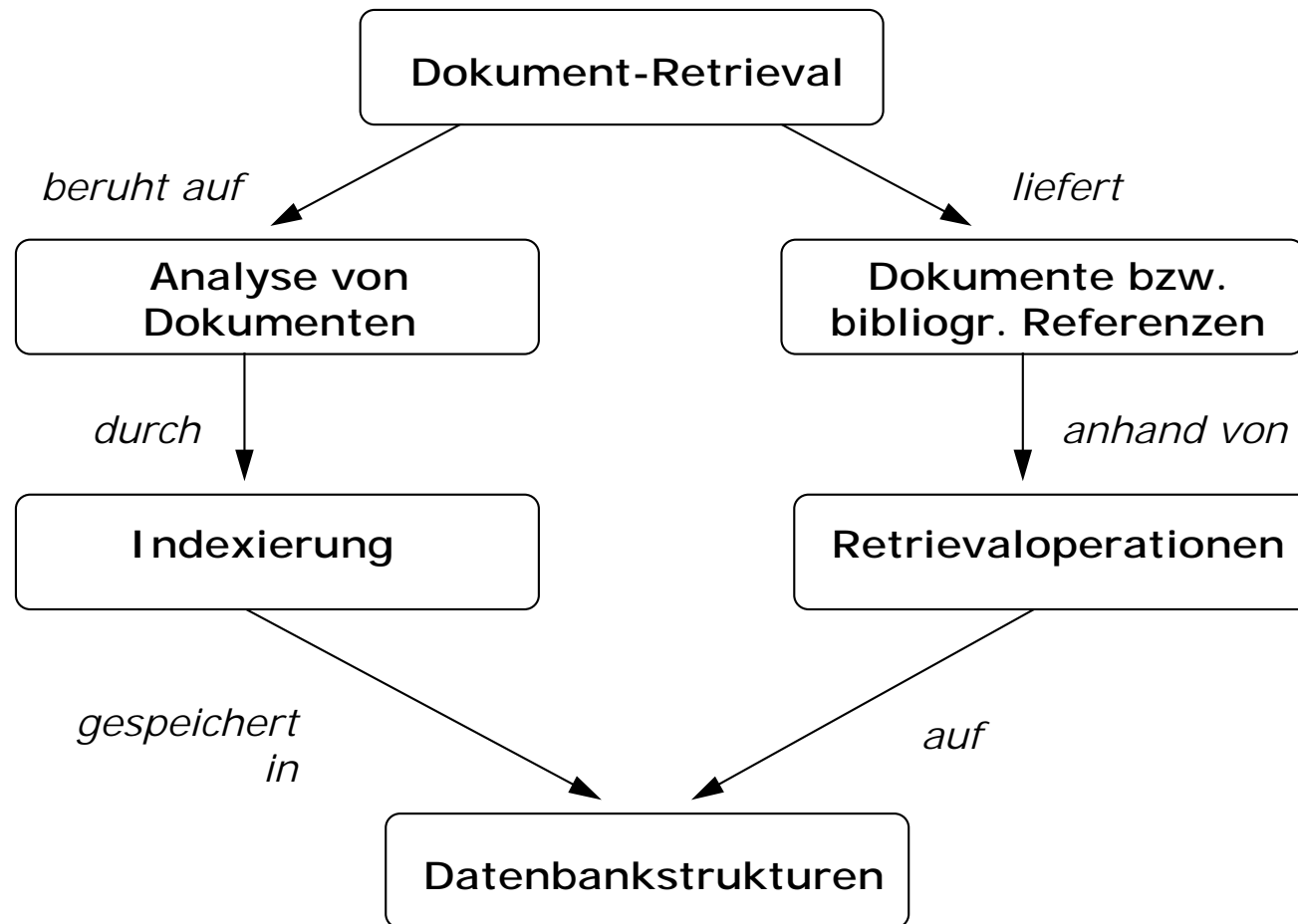


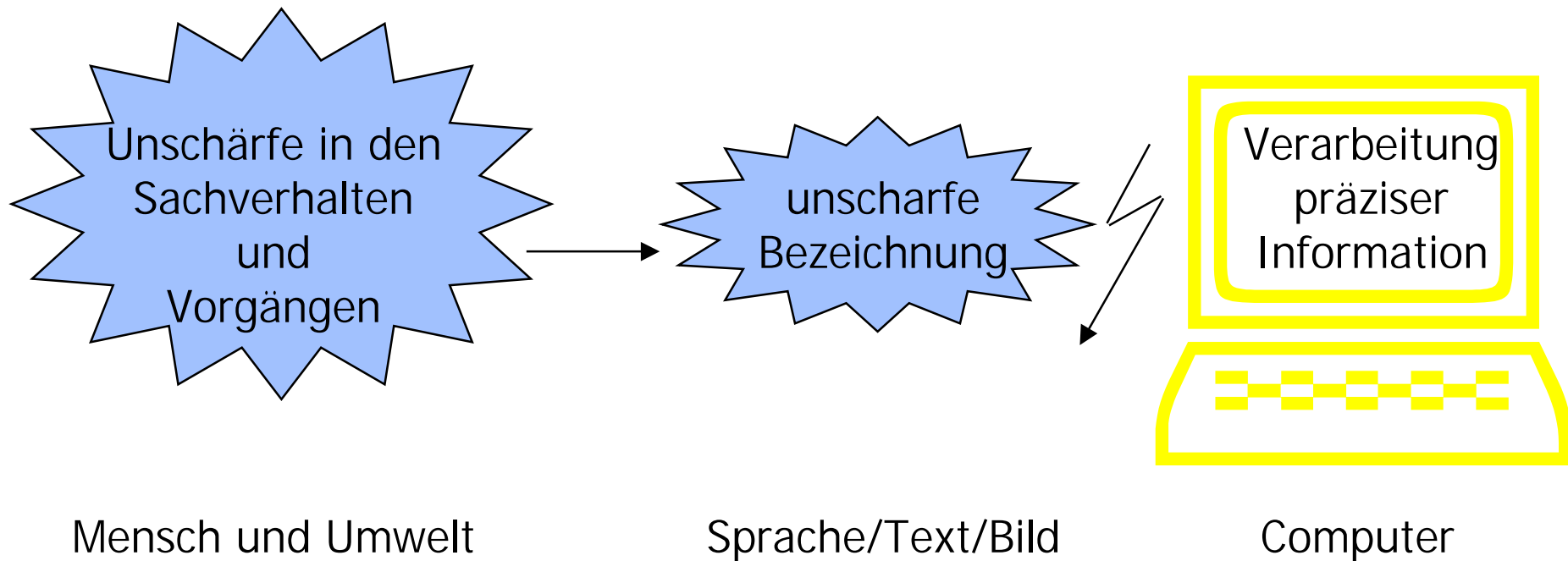
Automatische Indexierung

Prof. Holger Nohr
Fachhochschule Stuttgart

Grundmodell des Information Retrieval



Unschärfe im Information Retrieval

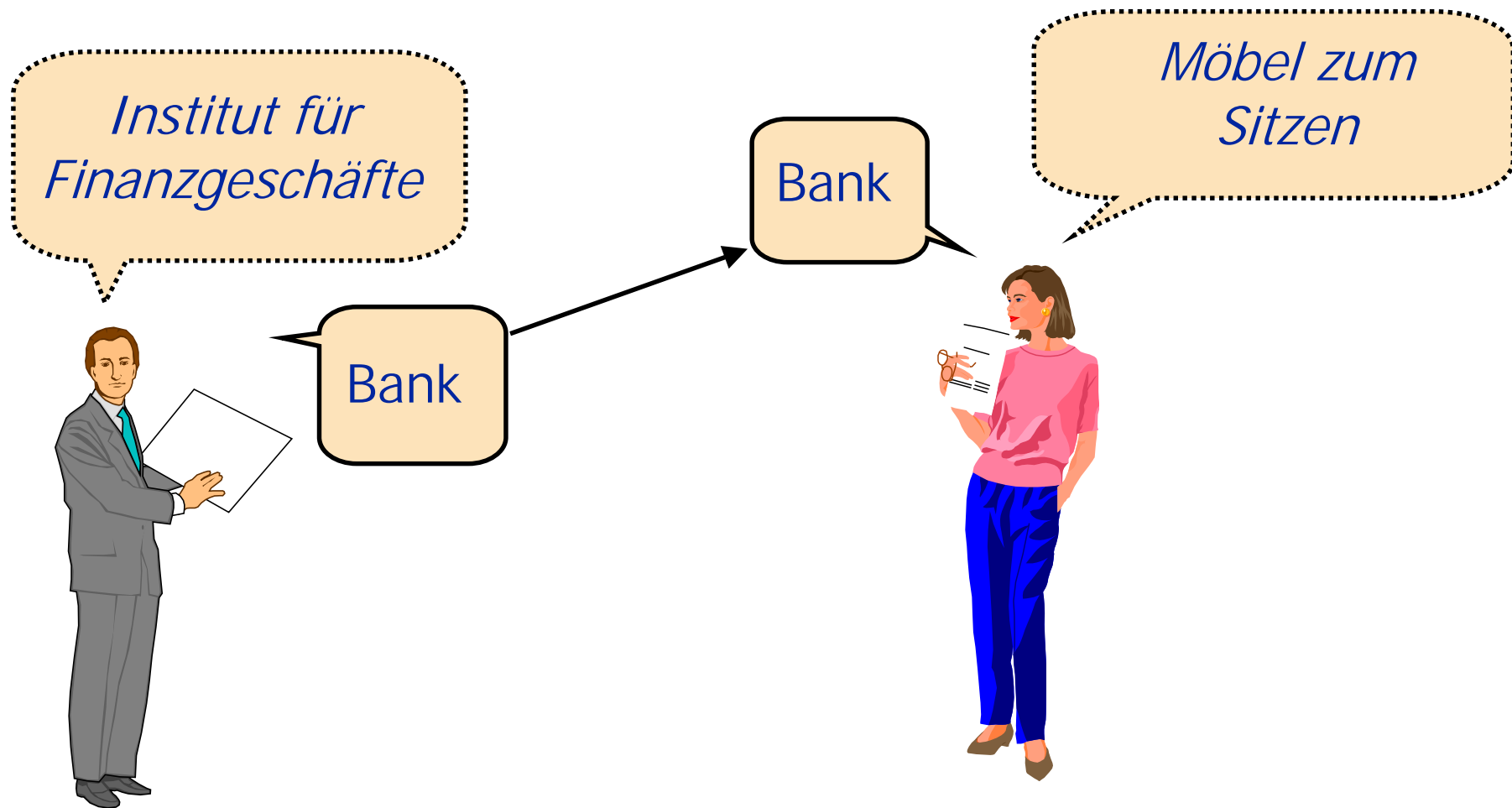


Benötigte Information:
Maschinenbaumarkt in Estland

Relevanz ?

Gefundene Information:
Markt für Werkzeug- und Verarbeitungsmaschinen im Baltikum

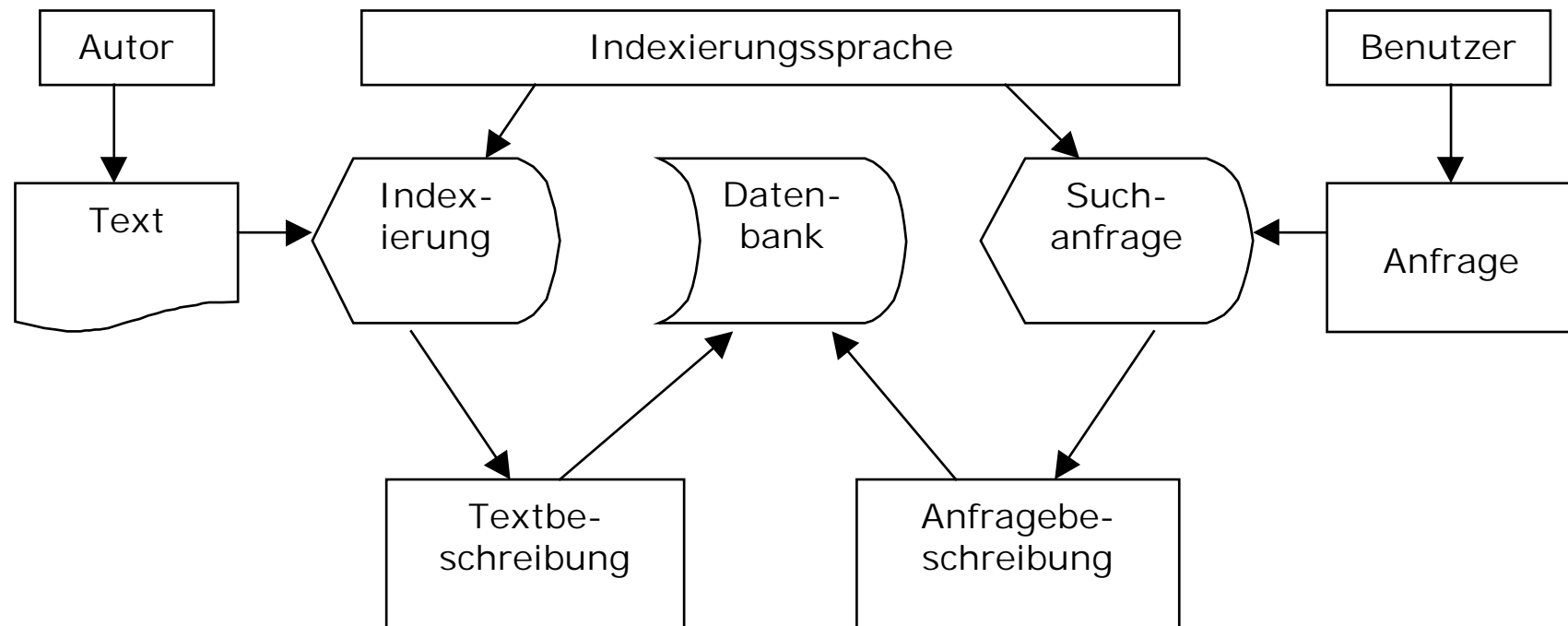
Begriff - Benennung - Kommunikation

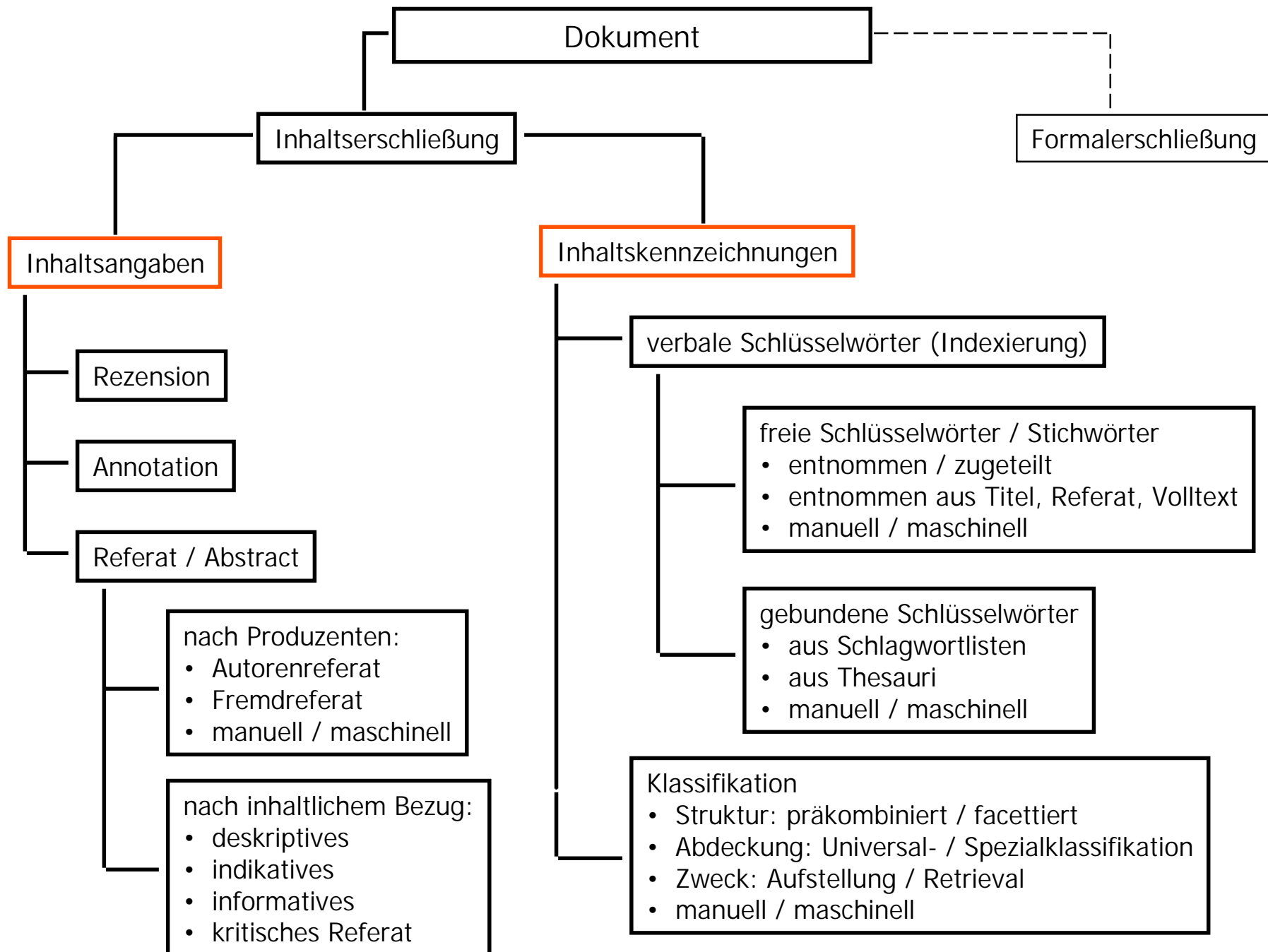


Niederschlag ?

- In den nächsten Tagen muß mit Niederschlag gerechnet werden.
- Die Ergebnisse fanden ihren Niederschlag in der Gesetzesvorlage.
 - Der erste Niederschlag erfolgte in der achten Runde.
 - Dies schlug sich negativ im Nettoverdienst nieder.
 - Sie war sehr niedergeschlagen.
 - Radioaktiver Niederschlag konnte nicht gemessen werden.
- Soldaten konnten den Aufstand nach zwei Tagen niederschlagen.
 - Der Räuber schlug den Geldboten nieder.
 - Das Landgericht schlug die Klage nieder.

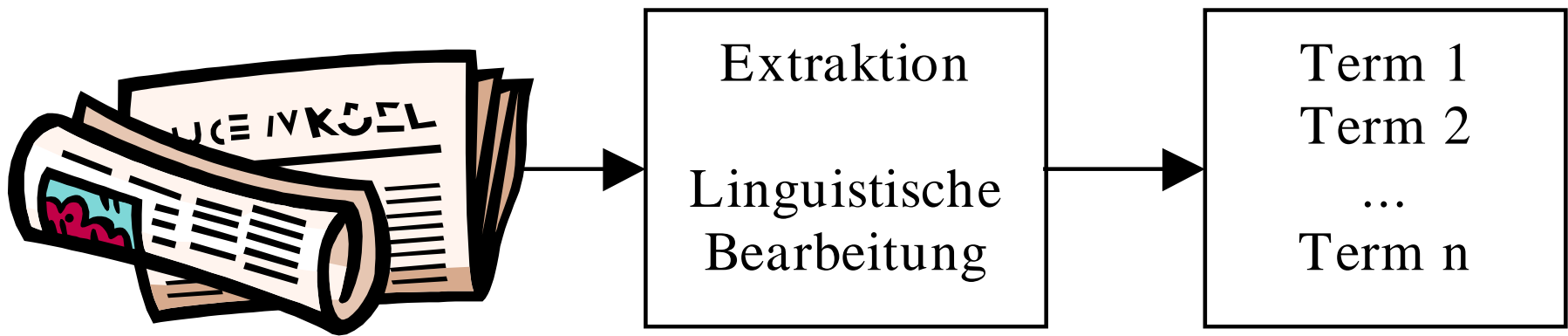
Modell der Informationerschließung

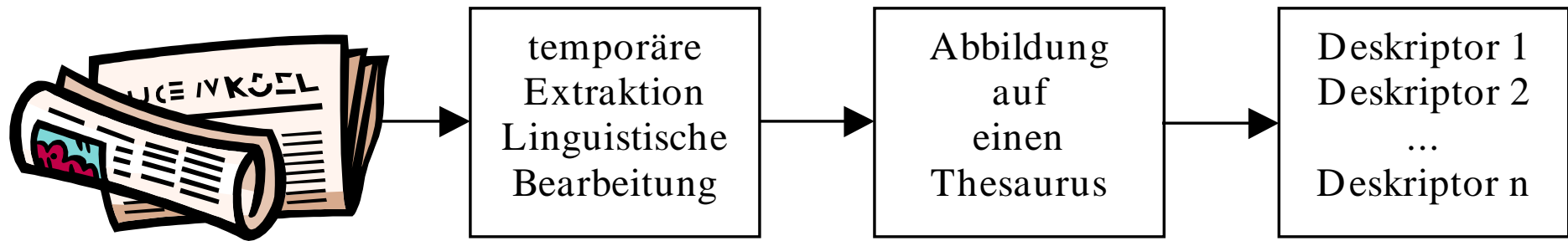




Verfahren der Automatischen Indexierung

- Freitextverstichwortung (ohne weitere Auswahl/Bearbeitung)
- Statistische Indexierungsverfahren
- Informationslinguistische Indexierungsverfahren
 - algorithmische Verfahren
 - wörterbuchgestützte Verfahren
- Pattern-Matching-Verfahren
- Begriffsortorientierte Indexierungsverfahren
- Automatic Text Summarization





Termfrequenz im Dokument

$$TF_{td} =$$

$$\frac{FREQ_{td}}{GESAMT_{td}}$$

$FREQ_{td}$ = Frequenz eines Terms im Dokument

$GESAMT_{td}$ = Gesamtzahl der Terme im Dokument

Termfrequenz im Dokument

Beispielrechnung

$$\text{TF}_{\text{td}} = \frac{4}{87} = 0,05$$

Termfrequenz in der Dokumentkollektion

$$TF_{tk} = \frac{FREQ_{tk}}{GESAMT_{tk}}$$

$FREQ_{tk}$ = Frequenz eines Terms i.d. Kollektion

$GESAMT_{tk}$ = Gesamtzahl der Terme i.d. Kollektion

Termfrequenz in der Dokumentkollektion

Beispielrechnung

$$\text{TF}_{\text{tk}} = \frac{350}{100.000} = 0,0035$$

Signifikanz eines Terms

$$S = \text{TF}_{\text{td}} - \text{TF}_{\text{tk}}$$

$$S = 0,05 - 0,0035 = 0,0465$$

Termhäufigkeitsansatz

Diesem Ansatz liegen folgende Annahmen zugrunde:

1. Häufig auftretende Wörter haben für die Bedeutung eines Dokuments eine höhere Signifikanz als Wörter mit einem geringem Vorkommen, sind also bessere Deskriptoren.

2. Seltener auftretende Wörter haben innerhalb einer Dokumentsammlung einen höheren Diskriminanzeffekt als häufig vorkommende Wörter, sind also bessere Deskriptoren.

Inverse Dokumenthäufigkeit (IDF)

$$\text{IDF}(t) = \frac{\text{FREQ}_{td}}{\text{DOKFREQ}_t}$$

Inverse Dokumenthäufigkeit (IDF)

Beispielrechnung

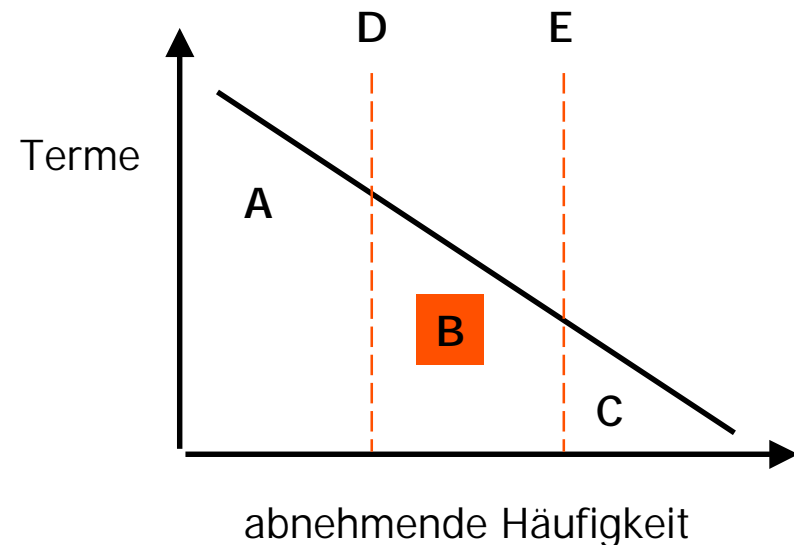
$$\text{IDF}(t) = \frac{4}{50} = 0,08$$

Statistische Verfahren

Statistischen Verfahren liegt die Annahme zugrunde, die Häufigkeit des Vorkommens von Termen in Dokumenten korreliert mit ihrer Bedeutung für den Inhalt dieser Dokumente.

Entscheidungsstärke ist die Fähigkeit eines Deskriptors, relevante Dokumente nach- und irrelevante Dokumente zurückzuweisen.

Entscheidungsstärkste Deskriptoren sind die Terme im mittleren Frequenzbereich (B). Hoch- (A) und niedrigfrequente (C) Terme erfüllen das Kriterium nicht. (nach H.P. Luhn)



Statistische Indexierung

Beispiel (1)

Text 1:

Computer werden im Information Retrieval eingesetzt. Es existieren Verfahren auf Computern für ein automatisches Retrieval. Moderne Computer ermöglichen ein effizientes Retrieval nach spezifischer Information.

Text 2:

Nutzer von Systemen zum Information Retrieval wurden befragt. Viele Nutzer waren mit der Funktionalität des Retrieval zufrieden. Die vorhandenen Systeme zum Information Retrieval genügen den Anforderungen der Nutzer. Es existieren eine Reihe von Systemen auf Computern.

Text 3:

Die Entwicklung neuer Systeme für das Information Retrieval wird von vielen Nutzern begrüßt. Die Entwicklung zielt auf neue Methoden des Retrievals mit Computern ab. Systeme zum effizienten Retrieval nach Information befinden sich derzeit in der Entwicklung.

Text 4:

Das Information Retrieval wird in Datenbanken durchgeführt. Verschiedene Datenbanken haben eine Oberfläche für den Nutzer, die ein zielgerichtetes Retrieval in Informationsräumen ermöglicht. Verschiedene Systeme für ein Retrieval in Datenbanken stehen derzeit dem Nutzer zur Verfügung.

Text 5:

Die Entwicklung von Systemen zum Retrieval in Informationsräumen ist für viele Nutzer von Datenbanken interessant. In Informationsräumen kann man navigieren und somit das Information Retrieval unterstützen. Der Informationsraum wird dreidimensional auf Computern visualisiert.

Statistische Indexierung

Beispiel (2)

Stoppwortliste:

Verfahren, Anforderung, Reihe, Methode,
Verfügung, Funktionalität, Oberfläche

Computerlinguistische Auswahl- und Bearbeitungsregeln:

nur Substantive, reduziert auf Nominativ und
Singular

Statistische Indexierung

Beispiel (3)

Indexterm	FREQ1	FREQ2	FREQ3	FREQ4	FREQ5	DOKFREQ
Computer	3	1	1	-	1	4
Information	2	2	2	1	1	5
Retrieval	3	3	3	3	2	5
Nutzer	-	3	1	2	1	4
System	-	3	2	1	1	4
Entwicklung	-	-	3	-	1	2
Datenbank	-	-	-	3	1	2
Informationsraum	-	-	-	1	3	2

Statistische Indexierung

Beispiel (4)

Termhäufigkeitsansatz:

$$\text{GEWICHTUNG} \sim \text{FREQ}_{ik} / \text{DOKFREQ}_k$$

Dabei wird eine Gewichtung ermittelt, die Relation herstellt aus der Häufigkeit eines Terms k im Dokument i (FREQ_{ik}) und umgekehrt proportional der Gesamtzahl der Dokumente (DOKFREQ_k), in denen der Term auftritt.

Statistische Indexierung

Beispiel (5)

Termgewichtung:

Indexterm	Text1	Text2	Text3	Text4	Text5
Computer	0.75	0.25	0.25	0	0.25
Information	0.4	0.4	0.4	0.2	0.2
Retrieval	0.6	0.6	0.6	0.6	0.4
Nutzer	0	0.75	0.25	0.5	0.25
System	0	0.75	0.5	0.25	0.25
Entwicklung	0	0	1.5	0	0.5
Datenbank	0	0	0	1.5	0.5
Informationsraum	0	0	0	0.5	1.5

Schwellenwert: Festlegung auf 0.5

Statistische Indexierung

Beispiel (6)

Ergebnis der Indexierung: Invertierter Index:

Indexterm	Texte				
Computer	Text 1				
Retrieval	Text 1	Text 2	Text 3	Text 4	
Nutzer		Text 2		Text 4	
System		Text 2	Text 3		
Entwicklung			Text 3		Text 5
Datenbank				Text 4	Text 5
Informationsraum					Text 5

Vektorraummodell

Im Vektorraummodell werden Fragen und Dokumente als Vektoren eines vieldimensionalen Vektorraumes aufgefaßt, der vom Vokabular aufgespannt wird. Die Retrievalfunktion versucht die räumliche Ähnlichkeit von Frage und Dokumenten zu bewerten.

3-dimensionaler Vektorraum:

t1 = automatisch

t2 = manuell

t3 = Indexierung

Dokumente:

Alles über automatische Indexierung

(1,0,1)

Manuelle und automatische Indexierung

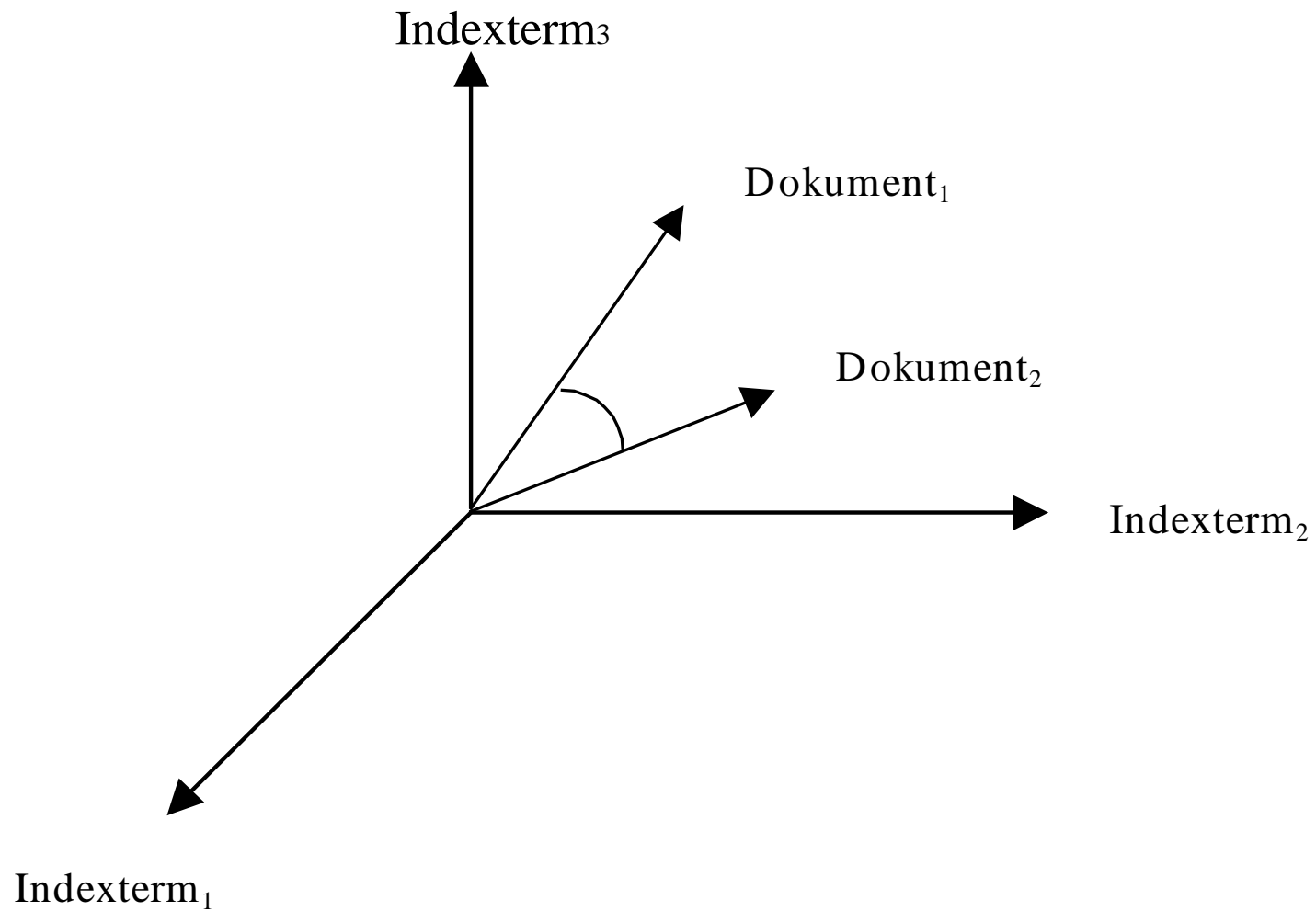
(1,1,1)

Frage:

Manuelle Indexierung

(0,1,1)

Vektorraummodell



Dokument - Dokument - Ähnlichkeit

$$\text{Ähn}(D_i, D_j) = \frac{1}{n} \sum_{k=1}^n g_{ik} g_{jk}$$

n = Anzahl der Indexterme

g_{ik} = Gewicht des Indexterms k im Dokument D_i

Informationslinguistik



Es geht allgemein darum:

- nicht sinntragende Wörter zu eliminieren
- grammatische Flexionsformen auf eine Grundform zurückzuführen (Wortstammanalysen)
 - Komposita zu zerlegen
 - Phrasen zu erkennen
- Pronomina korrekt zuzuordnen

Verfahren der Informationslinguistik

Regelbasiert:

Die Regeln einer Sprache werden (soweit benötigt) in einen Algorithmus gefaßt.

Der Algorithmus erkennt über eine Suffixliste Endungen (z.B. -*ing*) und wird das Wort *ringing* auf den Stamm *ring* reduzieren.

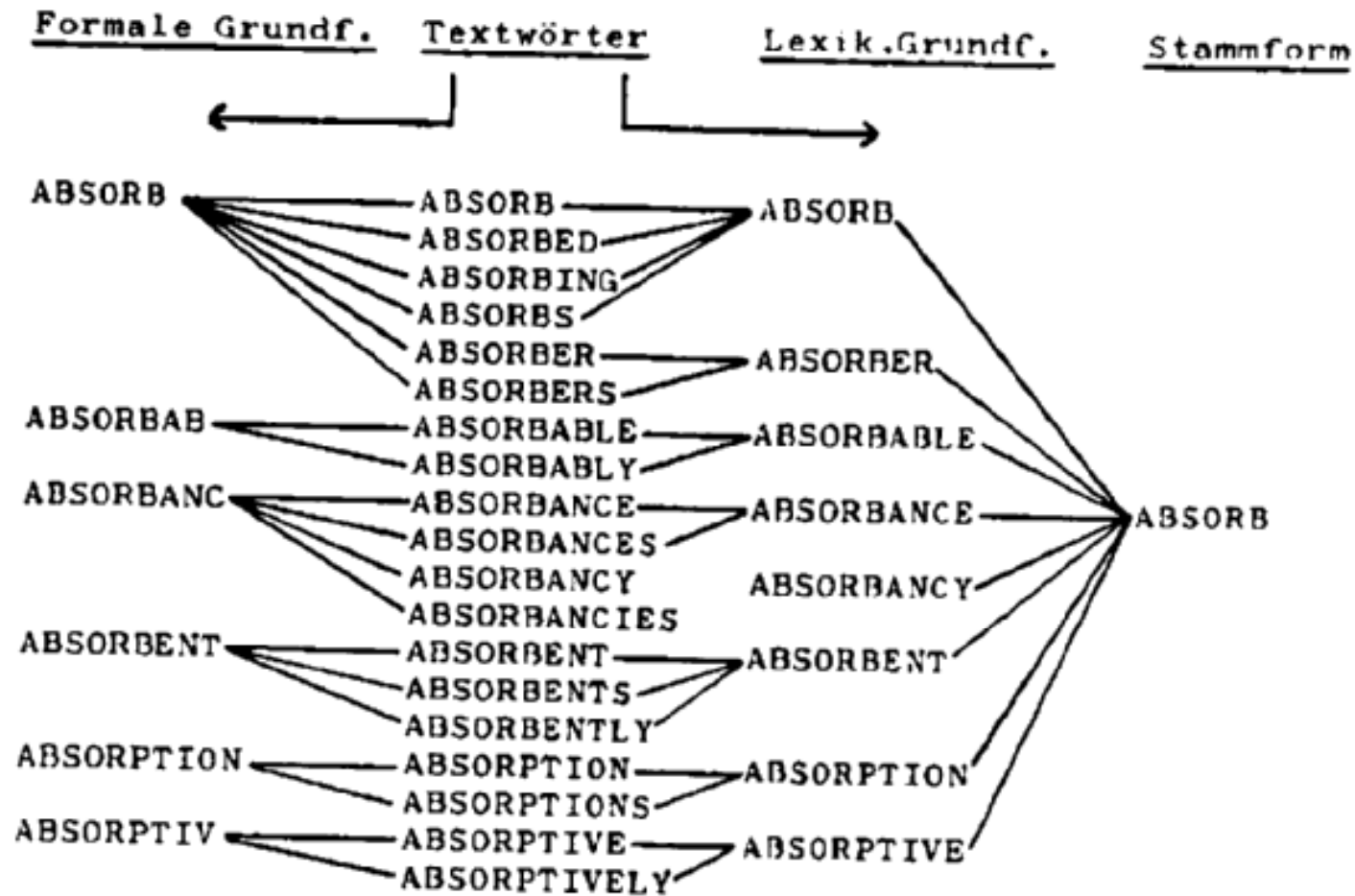
Wörterbuchbasiert:

Dem Verfahren liegen Wörterbücher zugrunde. Die Behandlung eines jeden Wortes muß in Wörterbüchern festgehalten werden.

Beispiele:

IDX, PASSAT

Morphologische Reduktion



R. Kuhlen:
Morphologische
Relationen durch
Reduktions-
algorithmen.
In: NfD 25, 1974,
S. 168-172

Beispiel für einen Reduktionsalgorithmus für die englische Sprache

Notation

%	alle Vokale, einschl. Y
*	alle Konsonanten
!	Länge des Wortes
/	„oder“
§	Leerzeichen
→	„zu“
←	„aus“
\	„nicht“

Regeln des Algorithmus

- a) $IES \rightarrow Y$
- b) $ES \rightarrow \S$ nach $* O/CH/SH/SS/ZZ/X$
- c) $S \rightarrow \S$ nach $* /E/%Y/%O/OA/EA$
- d) $S' \rightarrow \S$
 $IES' \rightarrow Y$
 $ES' \rightarrow \S$
- e) $'S \rightarrow \S$
 $' \rightarrow \S$
- f) $ING \rightarrow \S$ nach $**/%/X$
 $ING \rightarrow E$ nach $%*$
- g) $IED \rightarrow Y$
- h) $ED \rightarrow \S$ nach $**/%/X$
 $ED \rightarrow E$ nach $%*$

Fehlerklassen automatischer Reduktionsalgorithmen: *Understemming*

Regeln:

er	➡	leicht- <i>er</i>
en	➡	den Gift- <i>en</i>
es	➡	des Papier- <i>es</i>
e	➡	bei Licht- <i>e</i>
s	➡	des Wasser- <i>s</i>

Verschiedene Wortformen mit
gleicher Grund- bzw.
Stammform werden nicht
zusammengeführt:



des schlecht(est~en)
den schlecht(~en)
der schlecht(er~e)

die Them~en
des Thema~s

Fehlerklassen automatischer Reduktionsalgorithmen: *Overstemming*

Regeln:

er	➡	leicht- <i>er</i>
en	➡	den Gift- <i>en</i>
es	➡	des Papier- <i>es</i>
e	➡	bei Licht- <i>e</i>
s	➡	des Wasser- <i>s</i>

Verschiedene Wortformen mit
gleicher Grund- bzw.
Stammform werden falsch
zusammengeführt:



den Buch~en
des Buch~es

das Eis~en
des Eis~es

die Rind~en
die Rind~er

Funktionen von IDX

- Stoppworteliminierung
- Wortweise Übersetzung
- Grundformenermittlung
 - Bibliotheken = Bibliothek
- Dekomposition
 - Bibliotheksgebäude = Bibliothek, Gebäude
- Derivation
 - bibliothekarisch = Bibliothek
- Mehrworterkennung
 - wissenschaftliche Bibliothek
- Wortbindestrichergänzung
 - Kinder- und Jugendbibliothek = Kinderbibliothek, Jugendbibliothek
- Wortrelationierung
 - semantische Relation: Äquivalenz, Hierarchie

IDX-Indexierung im MILOS-Projekt

Gaus, Wilhelm:

Dokumentations- und Ordnungslehre:
Theorie und Praxis des Information
Retrieval

RSWK-Verschlagwortung:

Information Retrieval / Lehrbuch

Stichwortextraktion:

Dokumentations
Information
Ordnungslehre
Praxis
Retrieval
Theorie

IDX-Indexierung:

Dokumentation
Dokumentationslehre
Information
Information Retrieval
Lehre
ordnen
Ordnung
Ordnungslehre
praktisch
Praxis
Retrieval
theoretisch
Theorie

Recall und Precision

a: gefundene relevante Datensätze

b: gefundene nicht-relevante Datensätze (Ballast)

c: relevante Datensätze, die nicht gefunden wurden (Verlust)

$$\text{Recall} = a / a+c$$

(in %)

Anzahl der sowohl relevanten als auch
selektierten Datensätze

Anzahl der gespeicherten relevanten
Datensätze

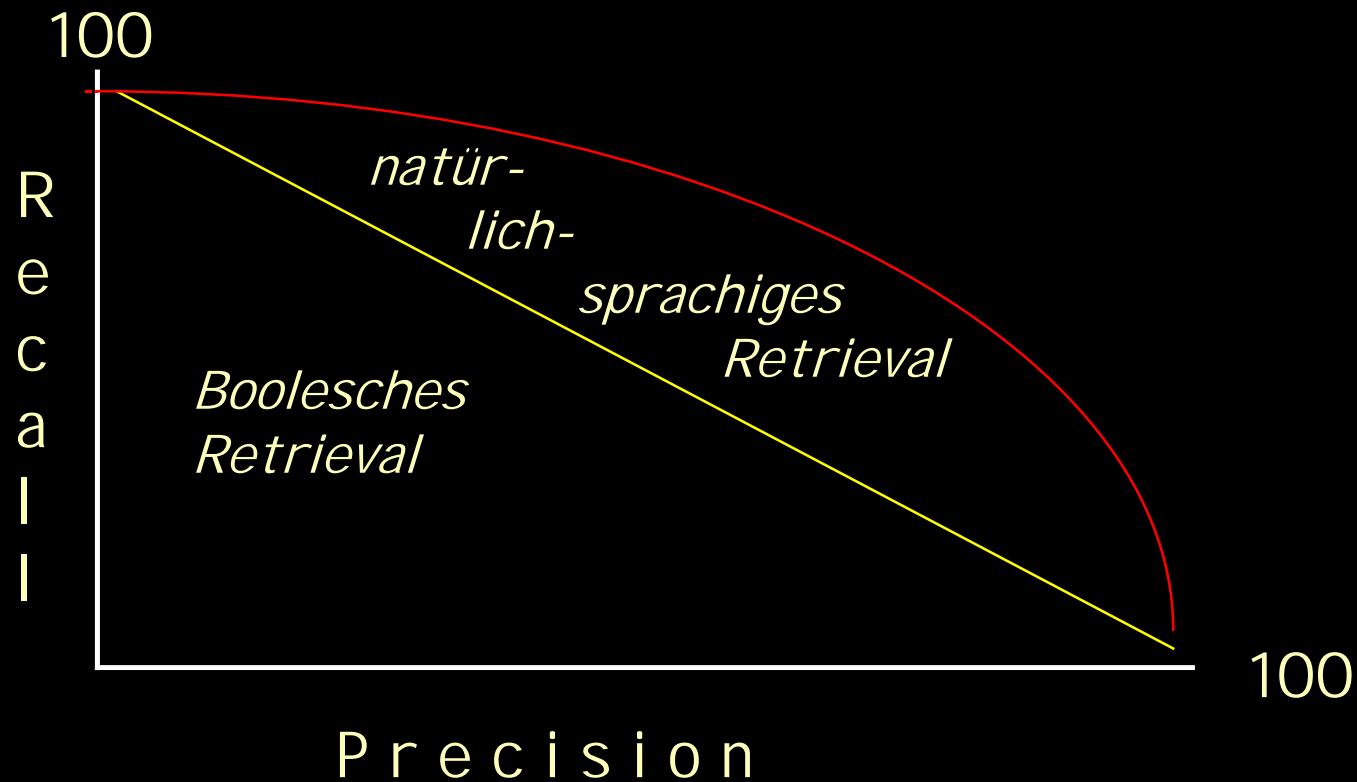
$$\text{Precision} = a / a+b$$

(in %)

Anzahl der sowohl relevanten als auch
selektierten Datensätze

Anzahl der selektierten Datensätze

Recall und Precision im Retrievalmodell



Retrievaltest zu MILOS I

Retrievaltest mit 40.000 Datensätzen und 50 Suchfragen

Methode	Recall	Precision	Einheitswert
Stichwort	14 %	59 %	0.84
Stichwort + Maschinelle Indexierung (IDX)	51 %	83 %	0.46
Stichwort + RSWK-Schlagwörter (Verstichwortet)	39 %	83 %	0.58

Retrievaltest zu MILOS II

Retrievaltest mit rund 190.000 Datensätzen
und 100 Suchfragen

Methode	"0-Treffer-Erg."	Precision
Stichwort	15	0,82
Maschinelle Indexierung (IDX)	3	0,75
RSWK (Verstichwortet)	30	0,95
Basic Index	0	0,803

Retrievaltest zu MILOS II

Ausgewählte Suchfragen:

Suchfrage	Suchformulierung	Stichwortregister		IDX-Register		RSWK-Register		Basic-Index	
		gefunden	relevant	gefunden	relevant	gefunden	relevant	gefunden	relevant
Hyperaktivität bei Kindern	<i>Hyperaktivität + Kindern</i>	0	0	0	0	0	0	4	4
	<i>Hyperaktivität + Kinder</i>	2	1	4	4	0	0	18	18
	<i>Hyperaktivität + Kind</i>	2	2	32	32	0	0	32	32
Homöopathische Mittel	<i>Homöopathische + Mittel</i>	5	3	0	0	0	0	43	43
	<i>Homöopathisch + Mittel</i>	0	0	145	145	0	0	145	145
	<i>Homöopathie + Mittel</i>	1	1	64	64	0	0	65	65
Interkontinentalrakete	<i>Interkontinentalrakete</i>	0	0	649	0	0	0	649	0

Das Verfahren AIR/X

- AIR/X ist ein an der TH Darmstadt (1978 bis 1985) entwickeltes Indexierungskonzept (*Leitung: G. Lustig*)
- AIR/X teilt Deskriptoren aus einem kontrollierten Vokabular zu
 - AIR/X ist ein **probabilistisches Indexierungsverfahren**
- AIR/X simuliert manuelles Indexieren; die Implementierung von AIR benötigt manuelle Indexierungsergebnisse als Vorgabe
- AIR/X ist in einer Pilotanwendung als AIR/PHYS beim FIZ Karlsruhe im Einsatz. (PHYS ist heute Teil von INSPEC.)

Indexierungsablauf von AIR

Textanalyse

Zerlegung in Wörter, Stoppworteliminierung,
Grundformenreduktion

Formelidentifizierung und -transformation

Formel = Deskriptor

Relationen und Relevanzbeschreibungen

Identifizierung von Textterm-Deskriptor-Relationen; Beschreibung
der Relevanz über ein Vektorraummodell

Berechnung des Gewichts der Deskriptoren

Zuteilung der Deskriptoren, wenn ihr Gewicht einen definierten
Schwellenwert überschreitet

Philosophie von AIR/X

Im Kern beruht die Philosophie von AIR/X auf einer
Einschätzung der Wahrscheinlichkeit:

Würde ein (menschlicher) Indexierer in Kenntnis bestimmter
Merkmale eines Dokuments (Terme im Text) einen
Deskriptor s zuteilen?

Probabilistischer Ansatz

Implementierung von AIR

Für die Vorbereitung auf ein Anwendungsgebiet wird eine große Menge manuell indexierter Dokumente benötigt.

Aus dieser Quelle wird die Wahrscheinlichkeit berechnet, daß beim Auftreten von Term t der Deskriptor s zuzuteilen ist.

Daraus leiten sich Wörterbucheinträge der Form

$\text{term} \longrightarrow \text{Gewicht} \longrightarrow \text{Deskriptor}$

ab.

AIR: Z-Relation

$$z(t,s) = \frac{h(t,s)}{f(t)}$$

$f(t)$ Anzahl der Dokumente aus einer Menge manuell indexierter Referate, in denen ein Term t auftritt

$h(t,s)$ Anzahl derjenigen unter diesen Dokumenten, denen der Deskriptor s manuell zugeteilt wurde

Wörterbuch AIR/PHYS in Zahlen

Anzahl der in die Berechnung eingegangenen Dokumente	392.000
durchschn. Dokumentlänge in Wörtern	103
Indexierungstiefe	8,8
<u>Deskriptoren aus den Thesauri (IDENTITÄT)</u>	→ 22.683
zugeteilte Deskriptoren	17.108
davon mind. dreimal zugeteilt	14.134
davon durch Relation Z abgedeckt	→ 10.002
<u>Einzelwörter</u>	702.337
davon mind. dreimal vorkommend	117.243
an Relationen beteiligt	85.017
Mehrwortgruppen	763.417
davon mind. dreimal vorkommend	546.198
an Relationen beteiligt	94.658
z-Werte zw. Einzelwörtern und Deskriptoren	159.930
z-Werte zw. Mehrwortgruppen und Deskriptoren	620.617
davon in PHYS/PILOT übernommen	170.697
z-Werte zw. Formeln und Deskriptoren	25.306
Term-Deskriptor-Relationen (USE; ohne Relation Z)	50.138
Deskriptor-Deskriptor-Relationen (BROADER-TERM, ENTHALTEN-IN, ABGRENZUNG)	192.907
insgesamt in PHYS/PILOT enthaltene Relationen	621.661

Quelle: M. Schwantner:
Entwicklung und Pflege des
Indexierungswörterbuches
PHYS/PILOT. In: Deutscher
Dokumentartag 1987.
Weinheim 1988

Die Relation Z generiert
57 % aller Einträge im
Wörterbuch.

Wörterbuch AIR/PHYS

Term	Descriptor	Regel	z(t,s)
stellar wind	stellar winds	Z	0,74
stellar wind	stellar winds	Identität	-
molecular outflow	stellar winds	Z	0,57
hot stellar wind	stellar winds	Z	0,76
molecular cloud	molecular clouds	Z	0,34
molecular cloud	molecular clouds	Identität	-
dense molecular cloud	molecular clouds	Z	0,35
incline joint	molecular clouds	Z	0,50
small crack extension	molecular clouds	Z	0,50

Erfahrungen mit AIR/PHYS

- Indexiert werden monatl. ca. 10.000 Dokumente aus der Physik
- Die Indexierung bezieht sich auf englischsprachige Titel und Abstracts
- früher durchschnittl. 9 manuell zugeteilte Deskriptoren heute durchschnittl. 12 maschinell zugeteilte Deskriptoren
- Manuelle Nacharbeitung:
 - ca. 4 der maschinellen Indexate werden gestrichen
 - ca. 4 Indexate werden manuell hinzugefügt

Retrievaltest

Das Pilotprojekt wurde durch einen Retrievaltest begleitet. 15.000 Dokumente aus PHYS wurden 300 Originalfragen unterzogen.

Ermittelt wurden die Werte Recall (r) und Precision (p) der maschinellen und der intellektuellen Indexierung:

Durchschnittswerte der maschinellen Indexierung:

$$p = 0,46; r = 0,57$$

Durchschnittswerte der intellektuellen Indexierung:

$$p = 0,53; r = 0,51$$

Indexierung und Retrieval

Die angewendeten Verfahren bleiben nicht auf die Indexierung beschränkt, sondern bearbeiten analog auch die Retrievalfragen:

(vgl. Folie zum Vektorraummodell)

