Kapitel IR: III

III. Retrieval-Modelle

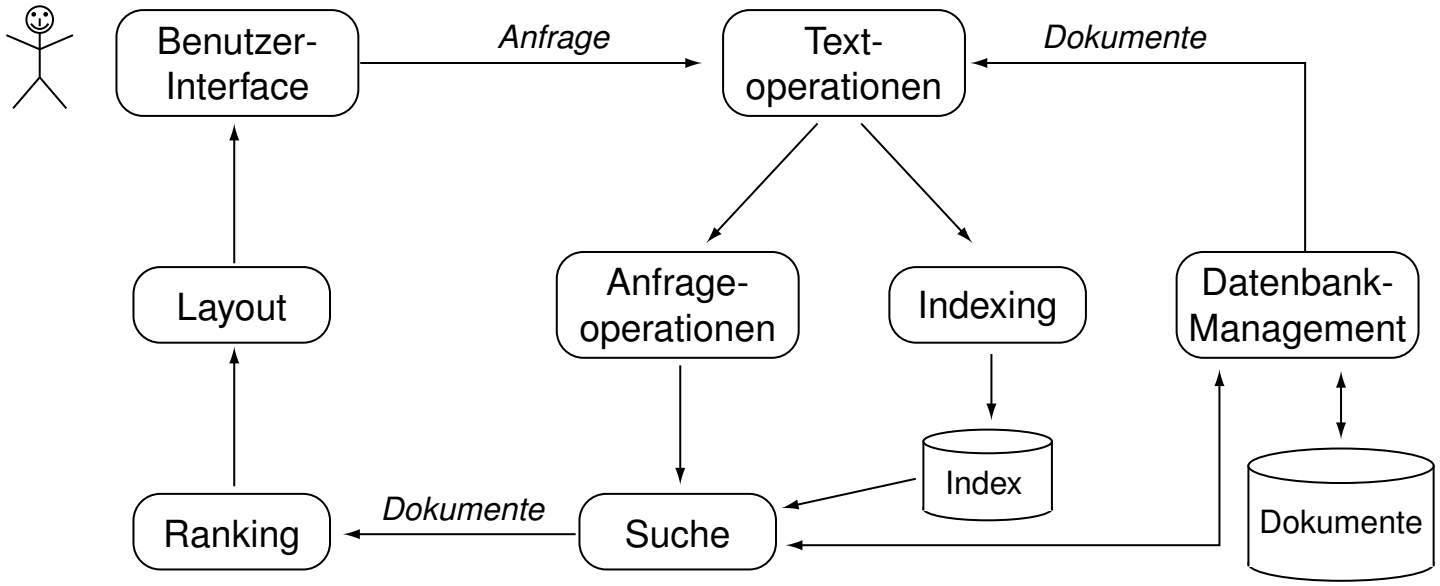
- Modelle und Prozesse im IR
- Klassische Retrieval-Modelle
- Bool'sches Modell
- Vektorraummodell
- Retrieval-Modelle mit verborgenen Variablen
- Algebraisches Modell

Modelle und Prozesse im IR

Prozesse im Text-IR

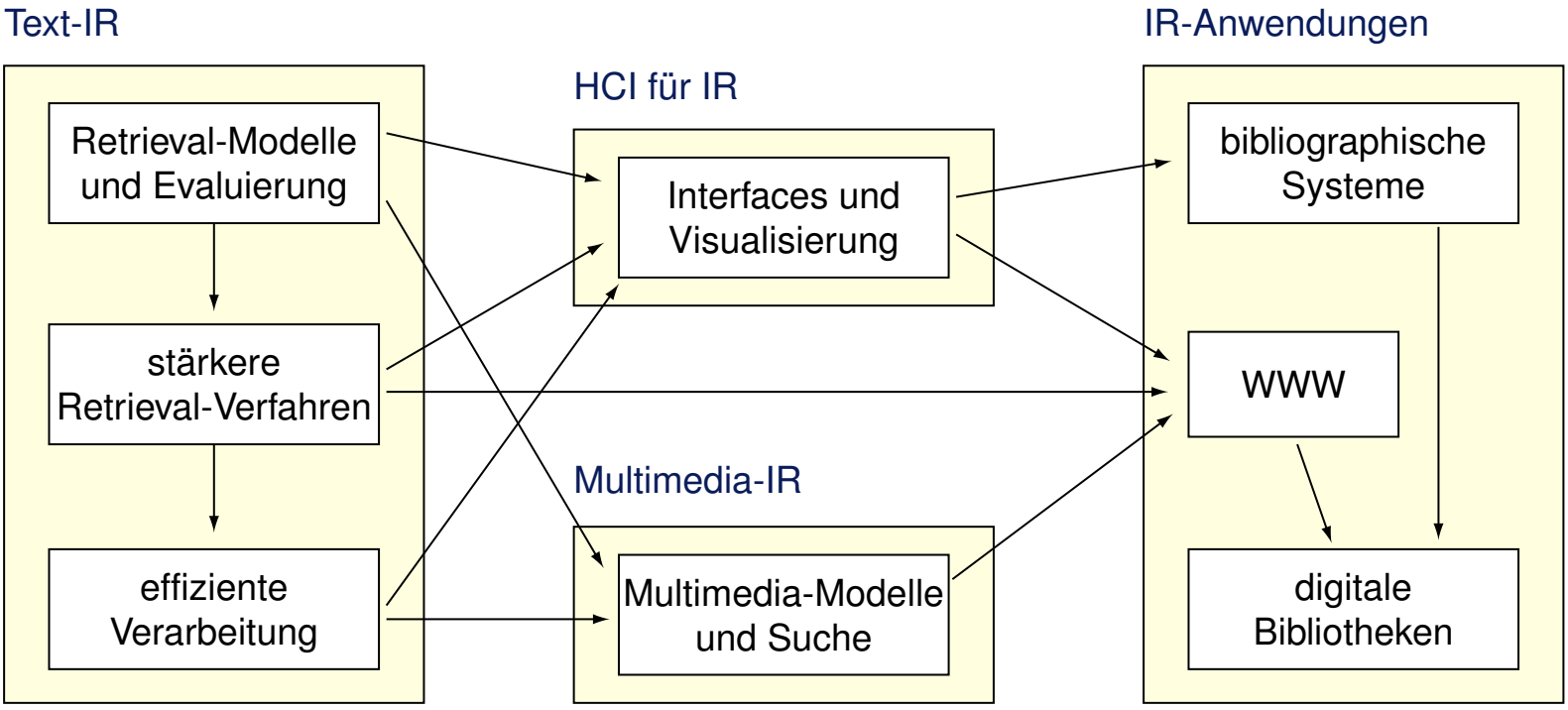
Prozesse im Text-IR können als die Operationalisierung des Wissenstransfers von einer Dokumentkollektion zum Informationsnachfragenden verstanden werden.

Beispiel für einen Retrieval-Prozess:



Modelle und Prozesse im IR

IR-Gebiete und Anwendungen



[vgl. Baeza-Yates/Ribeiro-Neto 1999]

Modelle und Prozesse im IR

Sichten auf ein Dokument

Die Automatisierung von Retrieval-Aufgaben erfordert die Modellierung und Repräsentation von Dokumenten auf einem Rechner. Dabei lassen sich drei *orthogonale* Sichten auf den Inhalt unterscheiden:

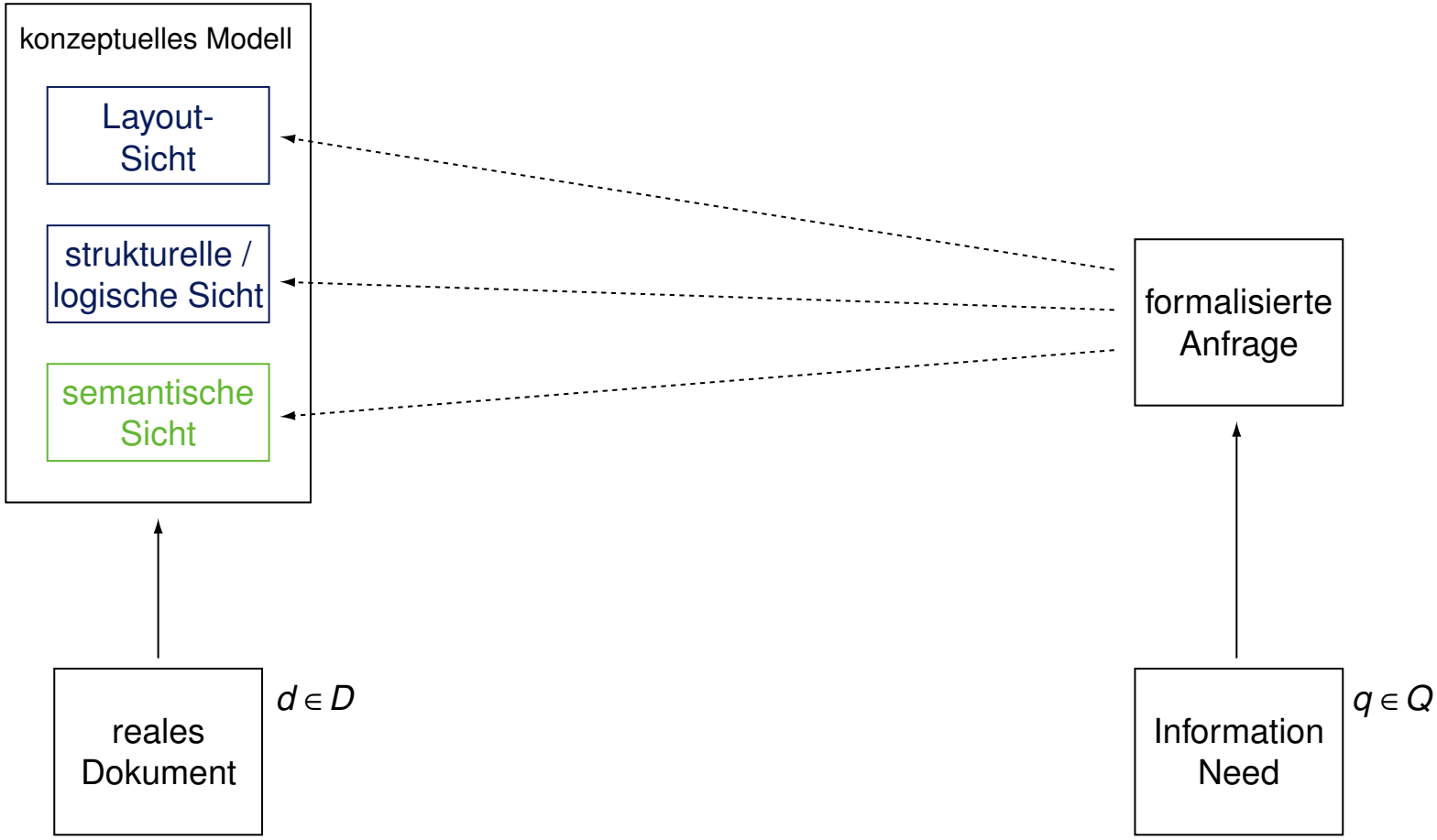
1. Layout-Sicht
   Darstellung eines Dokuments auf einem zweidimensionalen Medium.

2. Strukturelle bzw. logische Sicht
   Definiert den Aufbau bzw. die logische Struktur eines Dokuments.
   Beispiel:
   ```
   \documentclass[twocolumn,german]{article}
   \title{...}
   \author{...}
   \section{...}
   ```

3. Semantische Sicht
   Betrifft die Aussage eines Dokuments und ermöglicht dessen Interpretation.

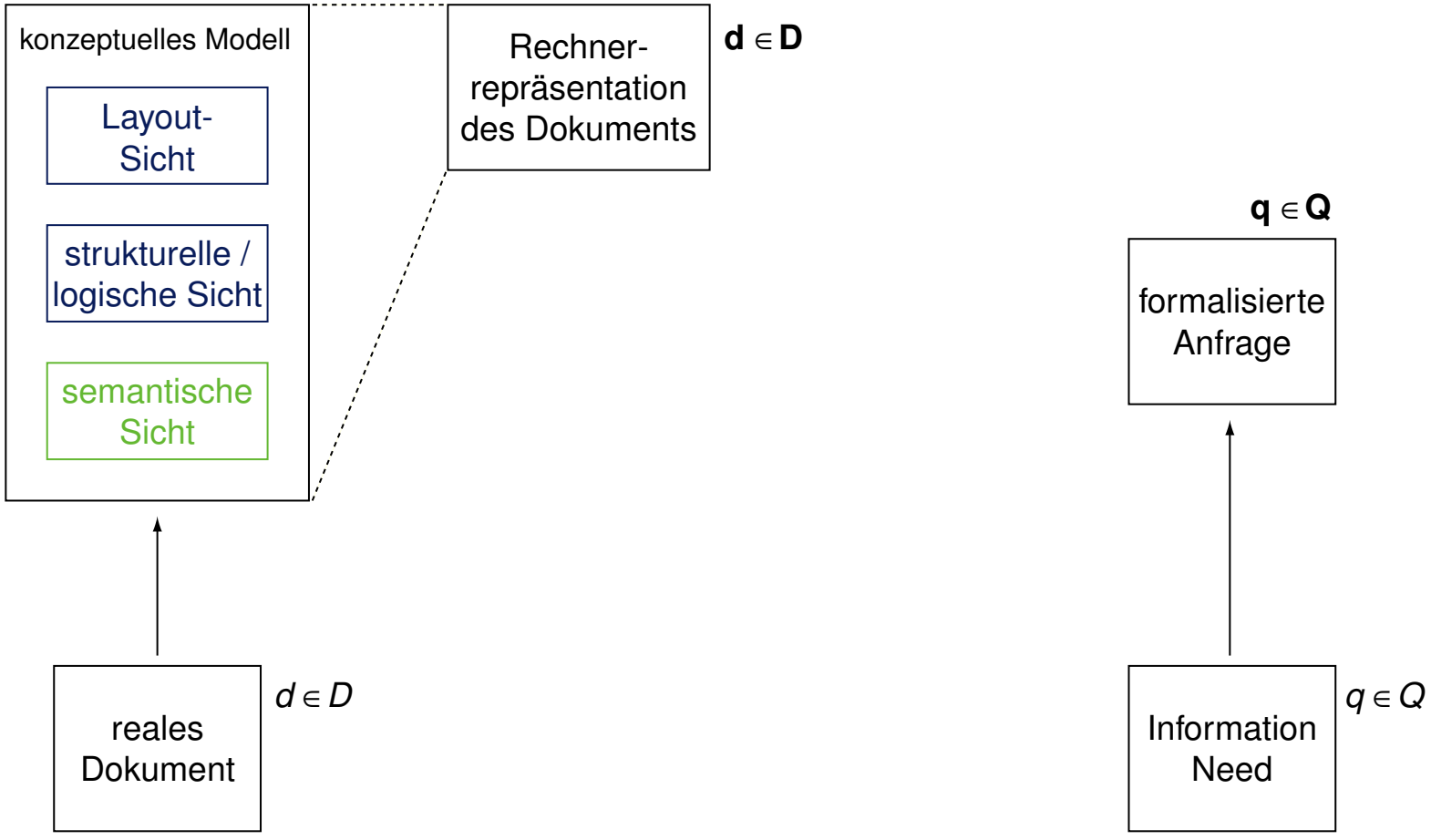
Modelle und Prozesse im IR

Vom Dokument zum Dokumentmodell



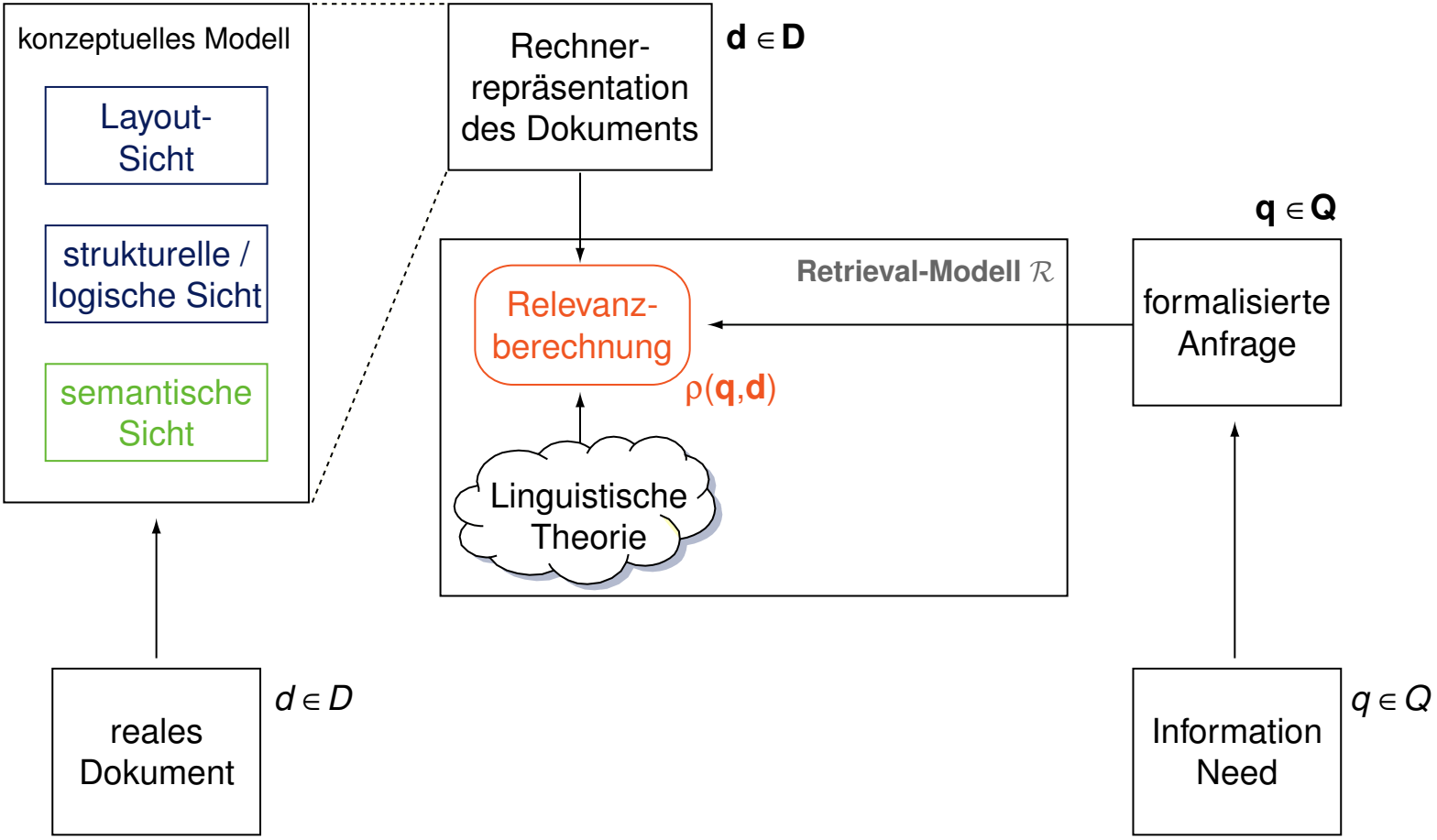
Modelle und Prozesse im IR

Vom Dokument zum Dokumentmodell



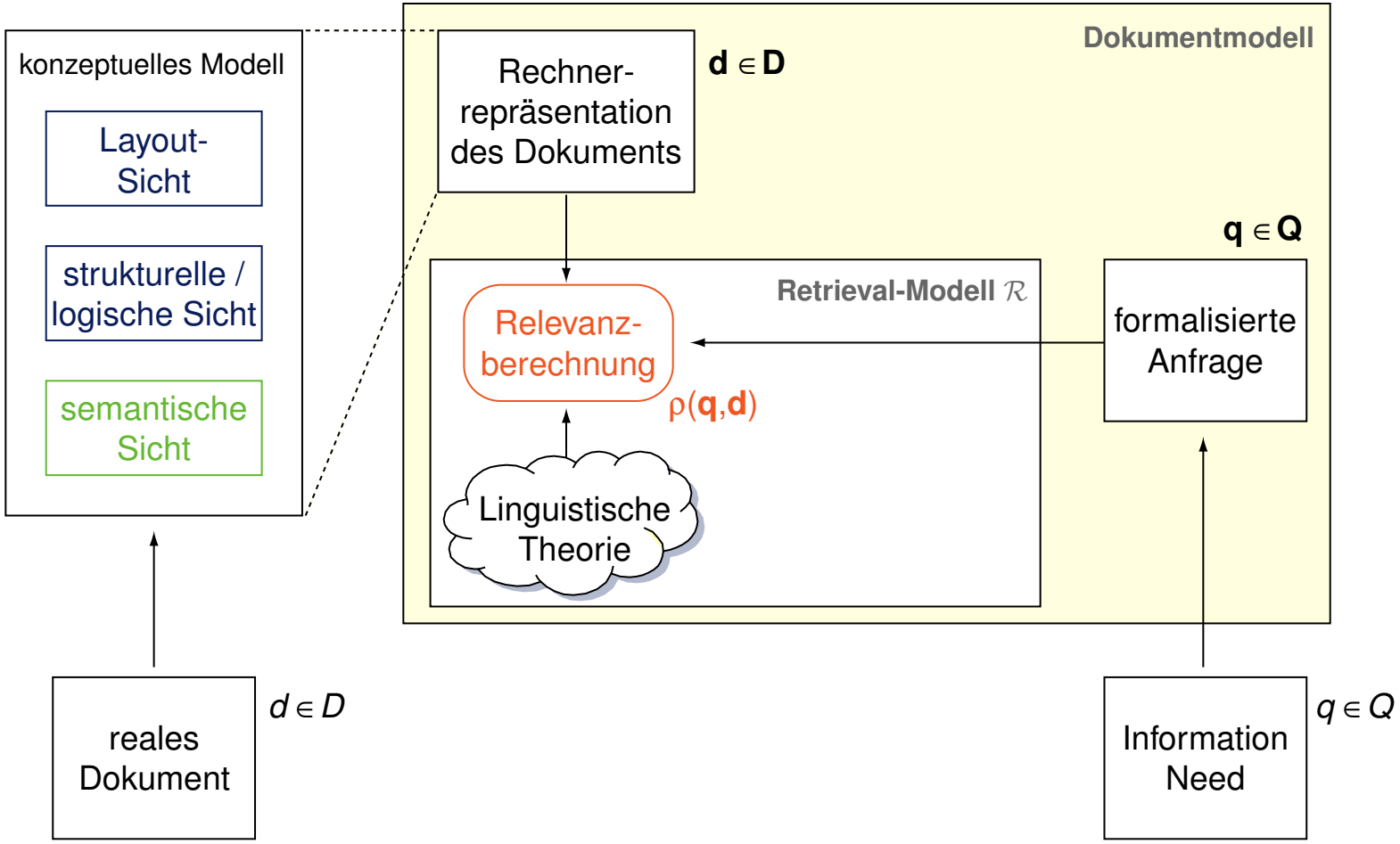
Modelle und Prozesse im IR

Vom Dokument zum Dokumentmodell



Modelle und Prozesse im IR

Vom Dokument zum Dokumentmodell



Modelle und Prozesse im IR

Definition 1 (Dokumentmodell, Retrieval-Modell, Retrieval-Funktion)
Sei D eine Menge von Dokumenten und Q eine Menge von Anfragen. Ein Dokument-Modell für D, Q ist ein Tupel $\langle \mathbf{D}, \mathbf{Q}, \rho_{\mathcal{R}} \rangle$, dessen Elemente wie folgt definiert sind:

1. $\mathbf{D}$ ist die Menge der Repräsentationen der Dokumente $d \in D$. In $\mathbf{d} \in \mathbf{D}$ können Layout-, logische und semantische Sicht codiert sein.

2. $\mathbf{Q}$ ist die Menge der formalisierten Anfragen.

Modelle und Prozesse im IR

Definition 1 (Dokumentmodell, Retrieval-Modell, Retrieval-Funktion)
Sei D eine Menge von Dokumenten und Q eine Menge von Anfragen. Ein Dokument-Modell für D, Q ist ein Tupel $\langle \mathbf{D}, \mathbf{Q}, \rho_{\mathcal{R}} \rangle$, dessen Elemente wie folgt definiert sind:

1. $\mathbf{D}$ ist die Menge der Repräsentationen der Dokumente $d \in D$. In $\mathbf{d} \in \mathbf{D}$ können Layout-, logische und semantische Sicht codiert sein.

2. $\mathbf{Q}$ ist die Menge der formalisierten Anfragen.

3. $\mathcal{R}$ ist ein Retrieval-Modell und formalisiert ein Prinzip, ein Paradigma oder eine linguistische Theorie.

   Auf der Grundlage von $\mathcal{R}$ ist die Retrieval-Funktion $\rho_{\mathcal{R}}(\mathbf{q}, \mathbf{d})$ definiert. Sie quantifiziert die Systemrelevanz zwischen einer formalisierten Anfrage $\mathbf{q} \in \mathbf{Q}$ und einer Dokumentrepräsentation $\mathbf{d} \in \mathbf{D}$:

   $$\rho_{\mathcal{R}} : \mathbf{Q} \times \mathbf{D} \rightarrow \mathbf{R}$$

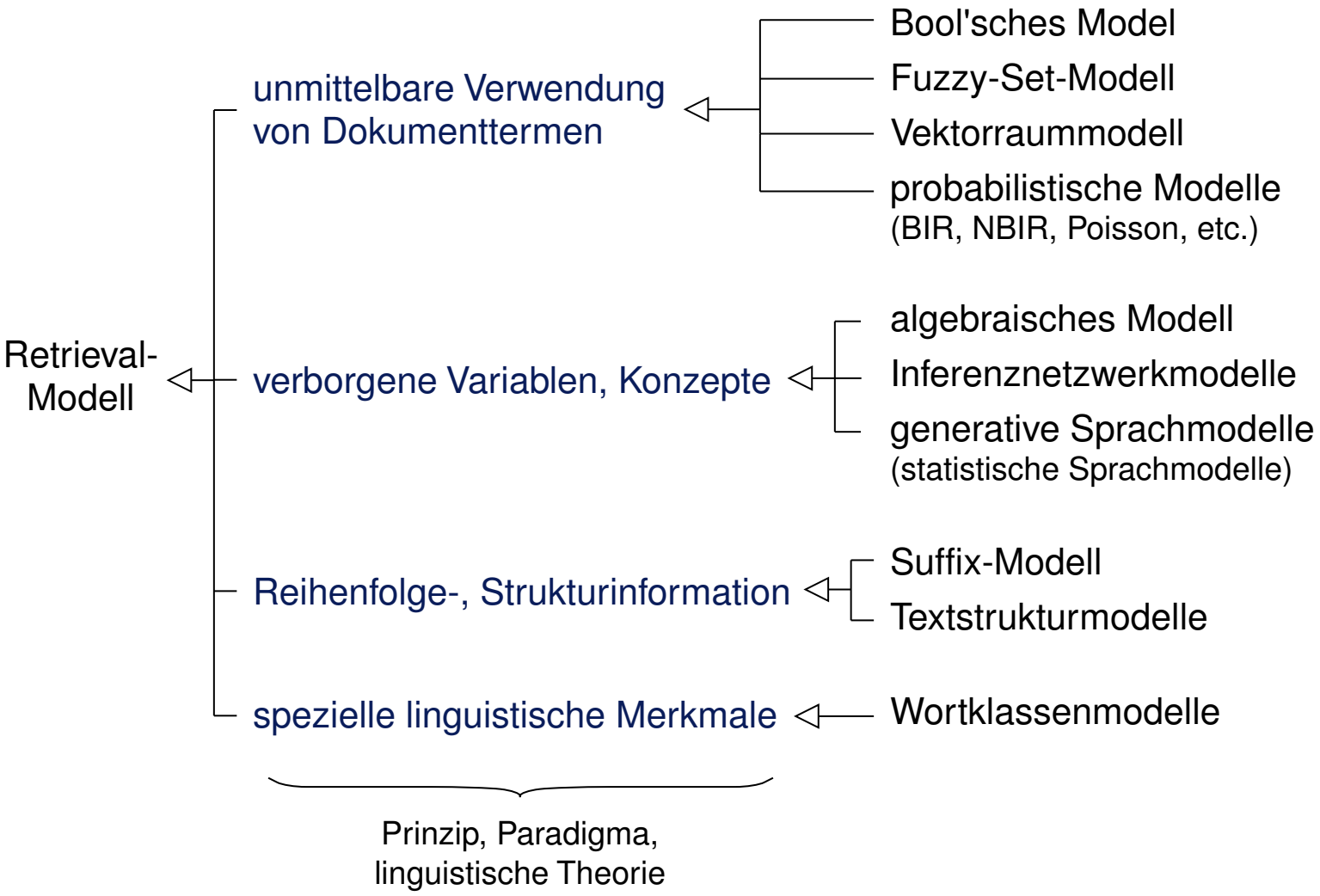
   Die von $\rho_{\mathcal{R}}$ berechneten Werte heißen Retrieval-Werte *(Retrieval Status Value, RSV)*.

Bemerkungen:

- Eine Dokumentrepräsentation enthält bestimmte Elemente und/oder quantifiziert bestimmte Aspekte des Originaldokumentes. Beispiele für Dokumentrepräsentationen sind Feature-Vektoren, Feature-Bäume oder Fingerprints.
- Ein Retrieval-Modell liefert die theoretische Basis dafür, wie man aus den drei Sichten eines Dokumentes Rückschlüsse ziehen kann, den Information-Need eines Anwenders zu erfüllen. Beispiele für Retrieval-Modelle sind das Vektorraummodell, das Binary-Independence-Modell oder das Latent-Semantic-Indexing-Modell.
- Retrieval-Modelle werden auch als Retrieval-Strategien bezeichnet.
- Die meisten Retrieval-Modelle basieren auf einer semantischen Sicht der Dokumente.
- Die Mengen $\mathbf{Q}$ bzw. $\mathbf{D}$ können auch intensional, durch Abbildungen $\alpha_Q : Q \rightarrow \mathbf{Q}$ und $\alpha_D : D \rightarrow \mathbf{D}$ definiert werden. [vgl. Fuhr 2004]
- Dokumentrepräsentationen und Retrieval-Modelle sind orthogonale Konzepte. Beachte jedoch, dass Retrieval-Modelle teilweise auch direkt mit bestimmten Repräsentationen assoziiert werden. So findet man in der Literatur den Begriff „Vektorraummodell“ nicht nur im Zusammenhang mit Retrieval-Modellen, sondern auch als Bezeichnung für die Dokumentrepräsentation in der Form eines Term-Vektors.

Modelle und Prozesse im IR

Taxonomie von Retrieval-Modellen



Modelle und Prozesse im IR

Taxonomie von Retrieval-Aufgaben

Collection-Time	Collection-Style: supervised	Collection-Style: semi-supervised	Collection-Style: unsupervised
short term			Categorization
long term			

Modelle und Prozesse im IR

Taxonomie von Retrieval-Aufgaben

Collection-Time	Collection-Style: supervised	Collection-Style: semi-supervised	Collection-Style: unsupervised
short term	Query formulation	Relevance feedback, Dialog	Categorization
long term	Profiling	Filtering, Classification	Monitoring

Modelle und Prozesse im IR

Taxonomie von Retrieval-Aufgaben

Collection-Time	Collection-Style: supervised	Collection-Style: semi-supervised	Collection-Style: unsupervised
short term	Google Deutschland	Ask Jeeves Ask.co.uk	 **aisearch** document categorization
long term	AvantGo	 Mozilla junk mail	amazon

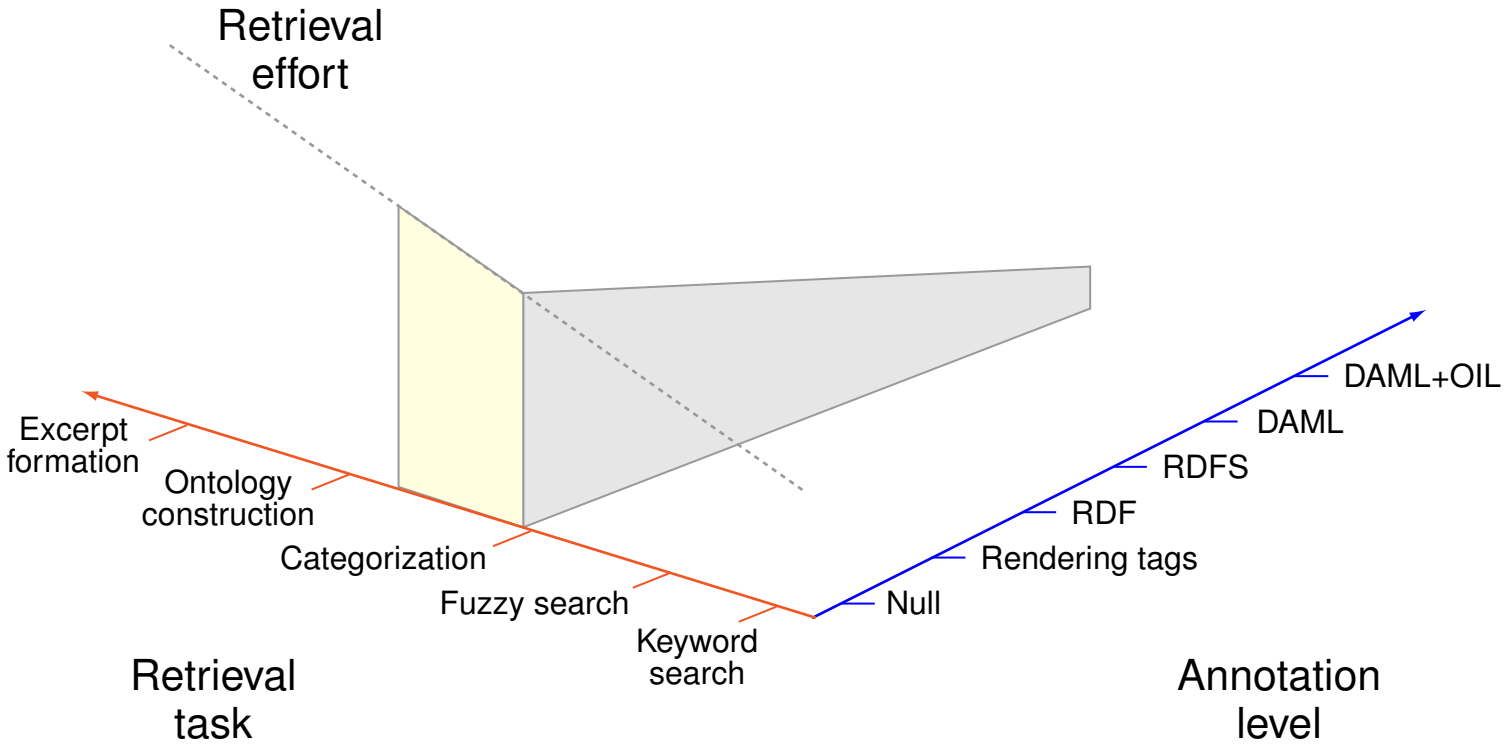
Bemerkungen:

- „Collection-Time = short term“ entspricht Ad-hoc-Retrieval: Die Dokumente in der Kollektion bleiben eher statisch, die Anfragen ändern sich.
- „Collection-Time = long term“ entspricht Filtering: Die Anfragen sind eher statisch (bekannt), und neue Dokumente kommen in die Kollektion bzw. existierende Dokumente ändern sich. Filtering-Aufgaben werden oft auf Basis von Benutzerprofilen definiert.

  Beispiel: Beobachtung des Aktienmarktes durch einen Computer und Benachrichtigung eines Anwenders beim Eintreffen bestimmter Ereignisse.

Modelle und Prozesse im IR

Taxonomie von Retrieval-Aufgaben



Bemerkungen:

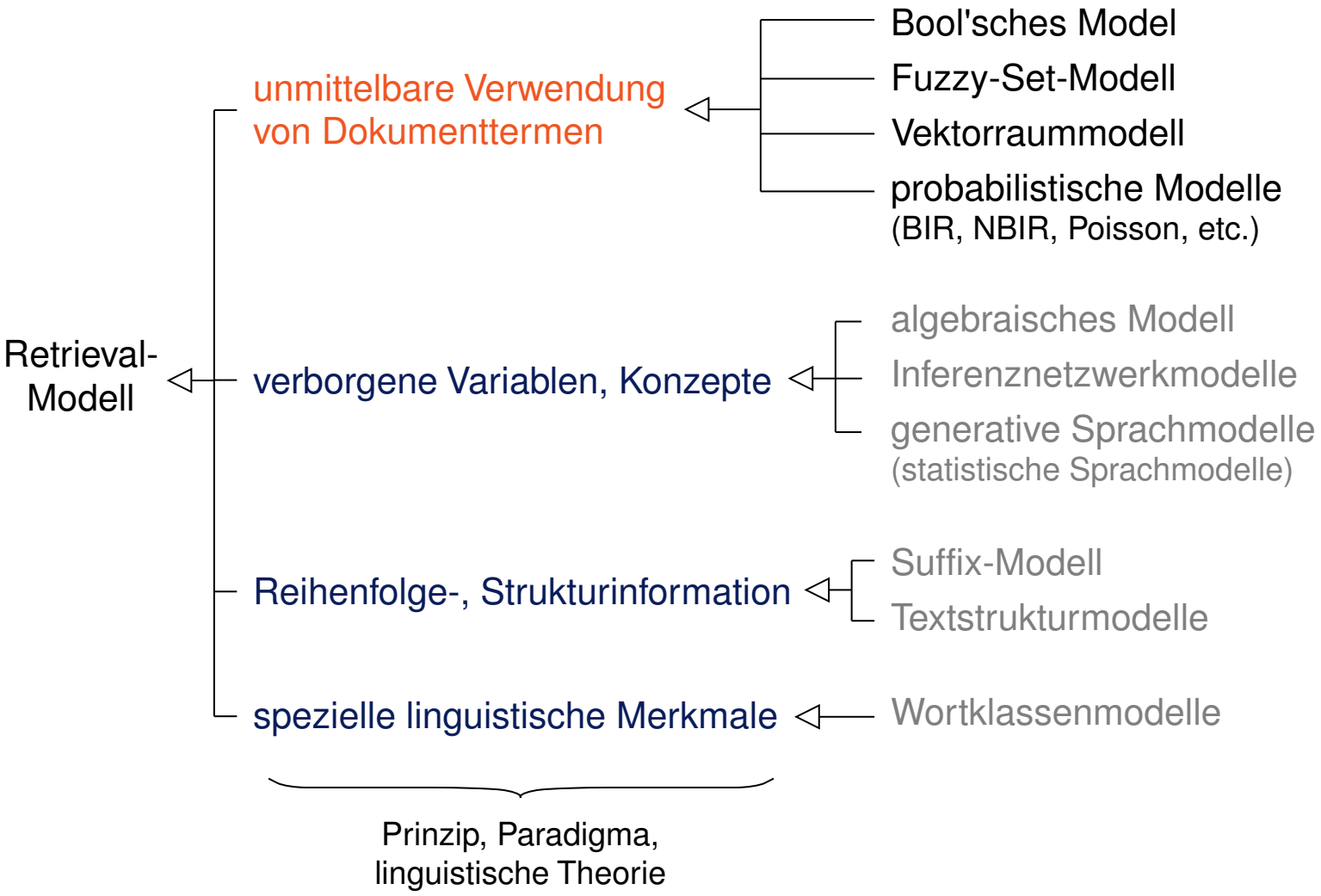
- Der Grad bzw. die Qualität der Textauszeichnung („tagging“) ist negativ korreliert mit dem Retrieval-Aufwand.

Kapitel IR: III (Fortsetzung)

III. Retrieval-Modelle

- Modelle und Prozesse im IR
- Klassische Retrieval-Modelle
- Bool'sches Modell
- Vektorraummodell
- Retrieval-Modelle mit verborgenen Variablen
- Algebraisches Modell

Klassische Retrieval-Modelle



Klassische Retrieval-Modelle

Die klassischen Retrieval-Modelle abstrahieren ein Dokument $d \in D$ zu einer unstrukturierten Menge von Indextermen, die sich quasi unmittelbar und *automatisch* aus d gewinnen lassen.

Die Dokumentrepräsentation $\mathbf{d}$ eines Dokumentes d besteht aus gewichteten Indextermen, die aus d stammen.

Unterscheidung der klassischen Retrieval-Modelle:

1. Art und Weise, wie sich Gewichte w_i für die Indexterme t_i berechnen.
2. Art und Weise, wie formalisierte Anfragen $\mathbf{q}$ konstruierbar sind.
3. Art und Weise, wie sich die Retrieval-Funktion $\rho_{\mathcal{R}}(\mathbf{q}, \mathbf{d})$ berechnet.
4. Art und Weise, wie die Menge relevanter Dokumente R konstruiert wird.

Bool'sches Modell

Dokumentmodell $\langle \mathbf{D}, \mathbf{Q}, \rho_{\mathcal{R}} \rangle$ [vgl. allgemeines Dokumentmodell]

Dokumentrepräsentationen $\mathbf{D}$.

Typischerweise bilden die Nomen eines Korpus in ihrer Grundform die Menge der Indexterme $T = \{t_1, \ldots, t_m\}$. Die Repräsentation $\mathbf{d}$ eines Dokumentes d ist eine Abbildung von T nach $\{0, 1\}$, wobei $\mathbf{d}(w) = 1$ bzw. $\mathbf{d}(w) = 0$ als „Term in d vorhanden" bzw. „nicht vorhanden" interpretiert wird.

Formalisierte Anfragenmenge $\mathbf{Q}$.

Eine formalisierte Anfrage $\mathbf{q} \in \mathbf{Q}$ entspricht einer logischen Formel über dem Alphabet $\Sigma = T$, in der die Junktoren $\wedge$, $\vee$, $\neg$ und Klammern verwendet werden können.

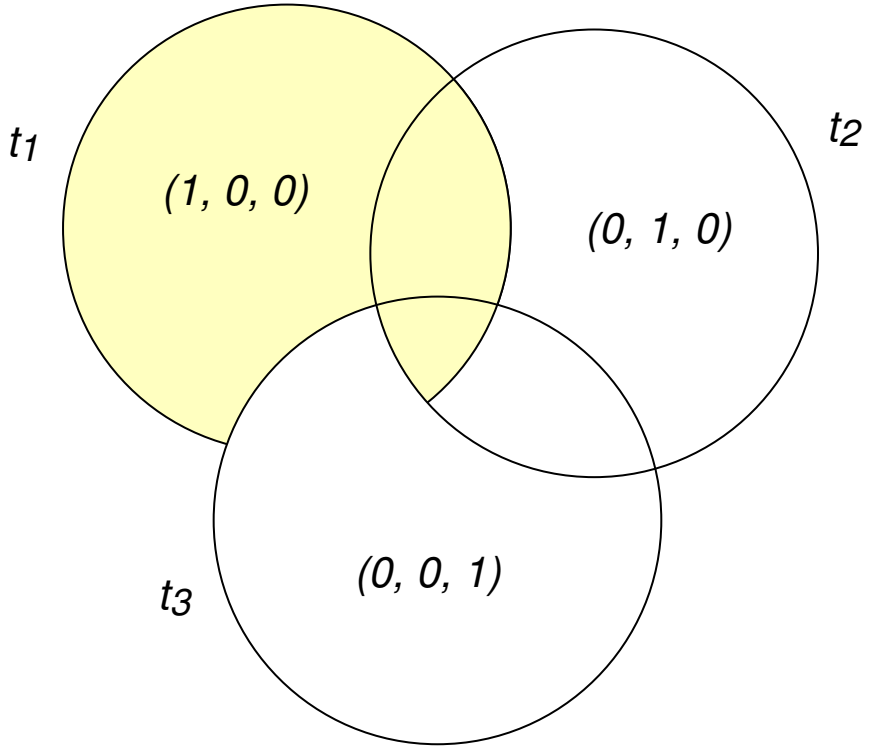
Retrieval-Funktion $\rho_{\mathcal{R}}$.

Die Dokumentrepräsentation $\mathbf{d}$ eines Dokumentes d induziert eine Interpretation $\mathcal{I}_{\mathbf{d}}$ für $\mathbf{q}$; man setzt $\rho_{\mathcal{R}}(\mathbf{q}, \mathbf{d}) = \mathcal{I}_{\mathbf{d}}(\mathbf{q})$.

Gilt $\rho_{\mathcal{R}}(\mathbf{q}, \mathbf{d}) = 1$, wird das Dokument d Element der Antwortmenge R.

Bool'sches Modell

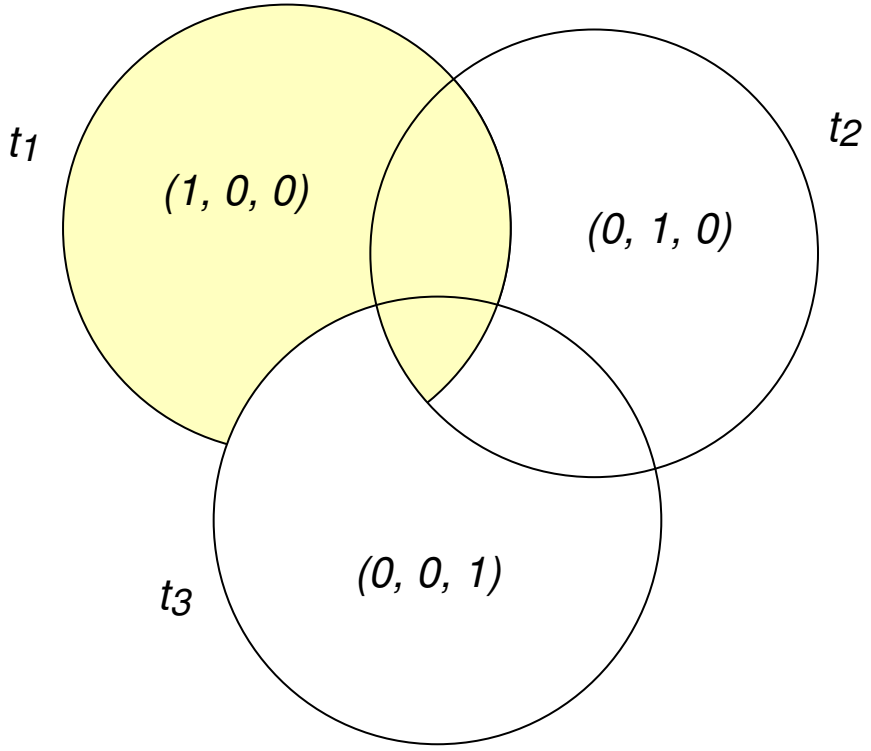
Retrieval-Funktion $\rho_{\mathcal{R}}$



Welche Anfrage ist illustriert?

Bool'sches Modell

Retrieval-Funktion $\rho_{\mathcal{R}}$



Welche Anfrage ist illustriert?

$$\mathbf{q} = (t_1 \wedge \neg t_2 \wedge \neg t_3) \vee (t_1 \wedge t_2 \wedge \neg t_3) \vee (t_1 \wedge t_2 \wedge t_3) \approx t_1 \wedge (t_2 \vee \neg t_3)$$

Bool'sches Modell

Beispiel

Dokumentrepräsentation:

$$\mathbf{d} = \{\ (\texttt{chrysler}, 1),\ (\texttt{deal}, 1),\ (\texttt{usa}, 1),\ (\texttt{china}, 0),\ (\texttt{cat}, 0),\ (\texttt{sales}, 1),\ (\texttt{dog}, 0),\ \ldots\ \}$$

Formalisierte Anfrage:

$$\begin{aligned}\mathbf{q} &= \texttt{usa} \wedge (\texttt{dog} \vee \neg\texttt{cat}) \\ &\approx (\texttt{usa} \wedge \texttt{dog}) \vee (\texttt{usa} \wedge \neg\texttt{cat}) \\ &\approx (\texttt{usa} \wedge \neg\texttt{dog} \wedge \neg\texttt{cat}) \vee (\texttt{usa} \wedge \texttt{dog} \wedge \neg\texttt{cat}) \vee (\texttt{usa} \wedge \texttt{dog} \wedge \texttt{cat})\end{aligned}$$

Induzierte Interpretation:

$\mathcal{I}_{\mathbf{d}}(\mathbf{q}) = 1$, wegen $\mathcal{I}_{\mathbf{d}}(\texttt{usa}) = 1, \mathcal{I}_{\mathbf{d}}(\texttt{dog}) = 0$ und $\mathcal{I}_{\mathbf{d}}(\texttt{cat}) = 0$.

Bemerkungen:

- Das Zeichen „$\approx$“ steht für „ist logisch äquivalent mit“.
- Was bedeutet die logische Äquivalenz?

Bool'sches Modell

Diskussion

Vorteile:

- Mächtigkeit: Prinzipiell kann mit einer Bool'schen Anfrage jede beliebige Teilmenge von Dokumenten aus einer Kollektion selektiert werden.
- einfache und genaue Implementierbarkeit

Nachteile:

- die Schwarz-Weiß-Aufteilung in die Menge R (bzw. $\bar{R}$) der relevanten (bzw. nicht-relevanten) Dokumente ist zu streng
- keine Ordung auf der Antwortmenge R hinsichtlich der Relevanz
- die Größe der Antwortmenge ist schwierig zu kontrollieren
- keine Möglichkeit zur Gewichtung von Fragetermen
- umständliche Formulierung von Anfragen
- schlechte Retrieval-Qualität

Vektorraummodell

Dokumentmodell $\langle \mathbf{D}, \mathbf{Q}, \rho_{\mathcal{R}} \rangle$ [vgl. allgemeines Dokumentmodell]

Dokumentrepräsentationen $\mathbf{D}$.

Typischerweise bilden die Wortsstämme aller Nicht-Stopworte eines Korpus die Menge der Indexterme $T = \{t_1, \ldots, t_m\}$. Der Wertebereich der Termgewichte ist $\mathbf{R}$; für die Gewichtsberechnung existieren verschiedene Konzepte.

Formalisierte Anfragenmenge $\mathbf{Q}$.

Eine formale Anfrage $\mathbf{q} \in \mathbf{Q}$ hat den gleichen Aufbau wie eine Dokumentrepräsentation $\mathbf{d} \in \mathbf{D}$.

Retrieval-Funktion $\rho_{\mathcal{R}}$.

Dokumentrepräsentationen und formalisierte Fragen werden als Punkte eines orthonormalen Vektorraums interpretiert, der durch die Terme aufgespannt wird.

Wichtige Ansätze zur Berechnung von $\rho_{\mathcal{R}}$ sind das cos-Ähnlichkeitsmaß und die euklidische Distanz.

Bemerkungen:

- Das Vektorraummodel wurde 1983 in dem Retrieval-System SMART in der Arbeitsgruppe von Gerhard Salton an der Cornell University eingesetzt. Die Entwicklungen und Überlegungen von Saltons Arbeitsgruppe reichen viele Jahre zurück.

Vektorraummodell

Retrieval-Funktion $\rho_{\mathcal{R}}$

Definition des Skalarproduktes $\mathbf{a}^T\mathbf{b}$ zwischen zwei Vektoren $\mathbf{a}$ und $\mathbf{b}$, mit φ als Winkel zwischen $\mathbf{a}$ und $\mathbf{b}$:

$$\begin{aligned} \mathbf{a}^T\mathbf{b} &= ||\mathbf{a}|| \cdot ||\mathbf{b}|| \cdot \cos(\varphi) \\ \Leftrightarrow \quad \cos(\varphi) &= \frac{\mathbf{a}^T\mathbf{b}}{||\mathbf{a}|| \cdot ||\mathbf{b}||} \end{aligned}$$

Vektorraummodell

Retrieval-Funktion $\rho_{\mathcal{R}}$

Definition des Skalarproduktes $\mathbf{a}^T\mathbf{b}$ zwischen zwei Vektoren $\mathbf{a}$ und $\mathbf{b}$, mit φ als Winkel zwischen $\mathbf{a}$ und $\mathbf{b}$:

$$\begin{aligned} \mathbf{a}^T\mathbf{b} &= ||\mathbf{a}|| \cdot ||\mathbf{b}|| \cdot \cos(\varphi) \\ \Leftrightarrow \quad \cos(\varphi) &= \frac{\mathbf{a}^T\mathbf{b}}{||\mathbf{a}|| \cdot ||\mathbf{b}||} \end{aligned}$$

Normalisiert man $\mathbf{a}$ und $\mathbf{b}$ – hier bezeichnet als $\mathbf{a}'$ und $\mathbf{b}'$ – gilt:

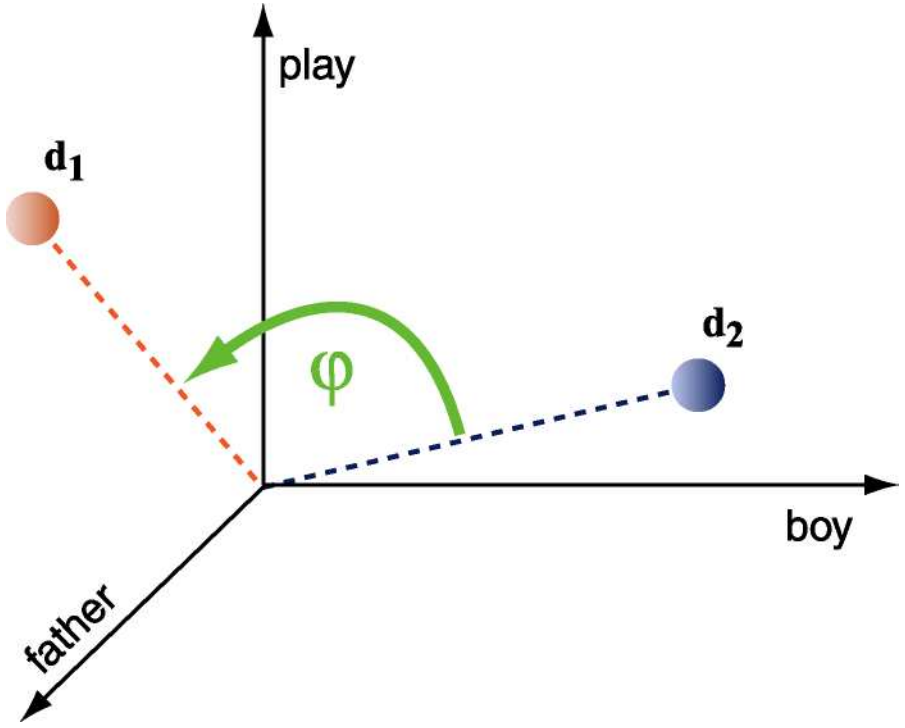
$$\cos(\varphi) = \frac{\mathbf{a}^T\mathbf{b}}{||\mathbf{a}|| \cdot ||\mathbf{b}||} = \frac{(\mathbf{a}')^T\mathbf{b}'}{||\mathbf{a}'|| \cdot ||\mathbf{b}'||} = (\mathbf{a}')^T\mathbf{b}' = \sum_{i=1}^{n} a'_i \cdot b'_i$$

$\langle \mathbf{D}, \mathbf{Q}, \rho_{\mathcal{R}} \rangle$ mit cos-Ähnlichkeitsmaß:

Definition von $\rho_{\mathcal{R}}(\mathbf{q}, \mathbf{d})$ als $\cos(\varphi)$, mit φ als Winkel zwischen $\mathbf{q}$ und $\mathbf{d}$.

Vektorraummodell

Retrieval-Funktion $\rho_{\mathcal{R}}$



Vektorraummodell

Beispiel

$$\mathbf{d_1} = \begin{pmatrix} \texttt{chrysler} & w_1 \\ \texttt{usa} & w_2 \\ \texttt{cat} & w_3 \\ \texttt{dog} & w_4 \\ \texttt{mouse} & w_5 \end{pmatrix} \quad = \quad \begin{pmatrix} \texttt{chrysler} & 1 \\ \texttt{usa} & 4 \\ \texttt{cat} & 3 \\ \texttt{dog} & 7 \\ \texttt{mouse} & 5 \end{pmatrix}$$

$$\mathbf{d'_1} = \begin{pmatrix} \texttt{chrysler} & 0.1 \\ \texttt{usa} & 0.4 \\ \texttt{cat} & 0.3 \\ \texttt{dog} & 0.7 \\ \texttt{mouse} & 0.5 \end{pmatrix}, \qquad \mathbf{d'_2} = \begin{pmatrix} \texttt{chrysler} & 0.4 \\ \texttt{usa} & 0.1 \\ \texttt{cat} & 0.7 \\ \texttt{dog} & 0.5 \\ \texttt{mouse} & 0.3 \end{pmatrix}$$

Der Winkel φ zwischen $\mathbf{d'_1}$ und $\mathbf{d'_2}$ ist etwa $38°$, $\cos(\varphi) \approx 0.79$.

Vektorraummodell

Retrieval-Modell $\mathcal{R}$

Zur Berechnung der Termgewichte w im Vektorraummodell hat sich der tf-idf-Ansatz bewährt [Sparck-Jones] :

1. In Dokument d_j ist die Bedeutung eines Terms t_i proportional zu seiner Häufigkeit.
   ➜ *term frequency* $tf(t_i, d_j)$; sie bezeichnet die Häufigkeit des Vorkommens von Term t_i in Dokument d_j.

2. Auf einen Korpus bezogen ist die Bedeutung eines Terms t_i umgekehrt proportional zur Anzahl $df(t_i)$ *(document frequency)* derjenigen Dokumente, die den Term t_i beinhalten.
   ➜*inverse document frequency* $idf(t_i)$.

$$idf(t_i) = \log_2(\frac{n+1}{df(t_i)+1})$$

   n bezeichne die Anzahl der Dokumente in dem betrachteten Korpus.

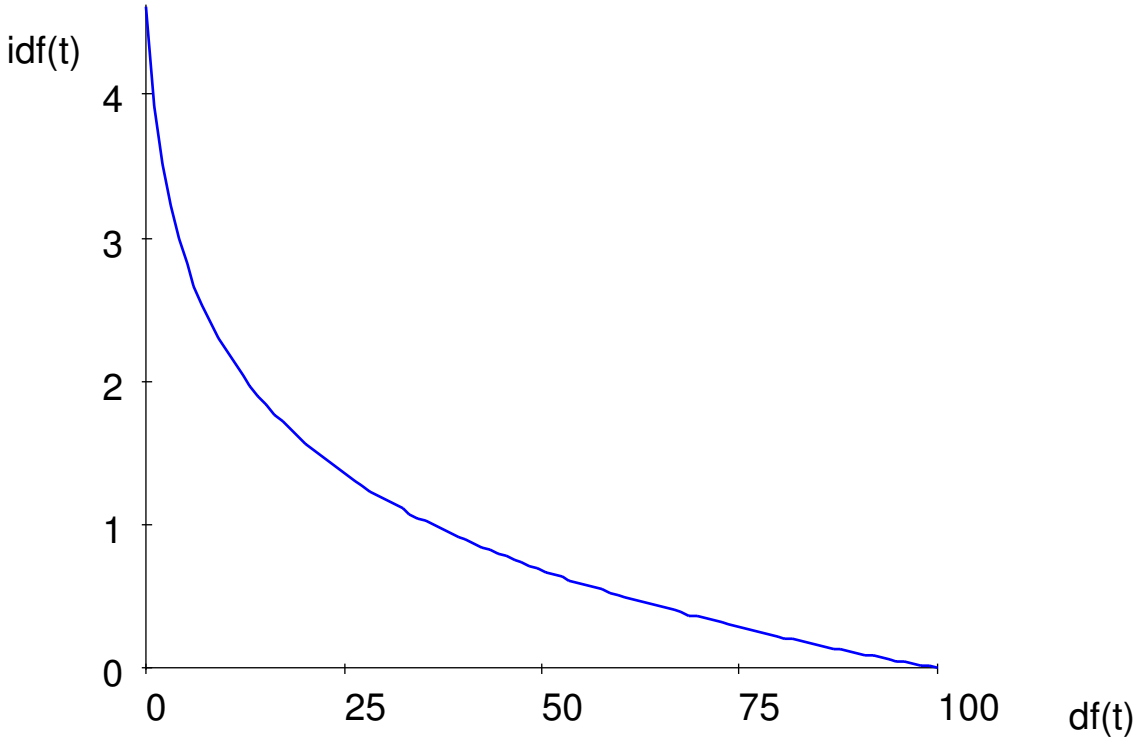
Bemerkungen:

- Idee hinter $idf(t)$: Ein Term t, der nur in wenigen Dokumenten vorkommt, besitzt hohes Diskriminierungspotential. Vergleiche das Konzept der Stopwort-Elimination hiermit.
- Zur Berechnung des Gewichts eines Terms t wird das Produkt aus $tf(t)$ und $idf(t)$ verwendet.

Vektorraummodell

Retrieval-Modell $\mathcal{R}$

Verlauf der Funktion $idf(t) = \ln(\frac{n+1}{df(t)+1})$ für Korpusgröße $n = 100$.



Vektorraummodell

Diskussion

Vorteile:

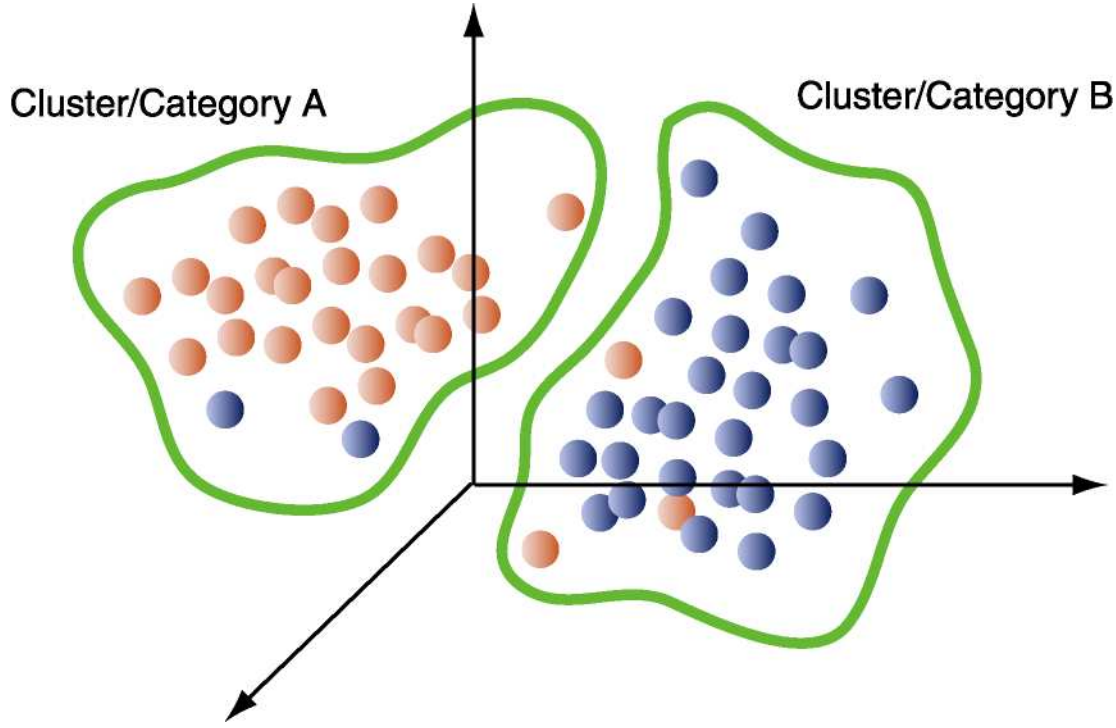
- Termwichtung verbessert die Retrieval-Performanz
- Im Gegensatz zur Schwarz-Weiß-Situation beim Bool'schen Modell erlaubt das „partielle Matching“ ein Retrieval von Dokumenten, die die Bedingungen in den Anfragen approximieren.
- Die Retrieval-Funktion $\rho_{\mathcal{R}}$ definiert eine Rangordnung unter den gefundenen Dokumente bzgl. ihrer Ähnlichkeit zur der Anfrage.

Nachteile:

- Indexterme werden als voneinander unabhängig angesehen

Vektorraummodell

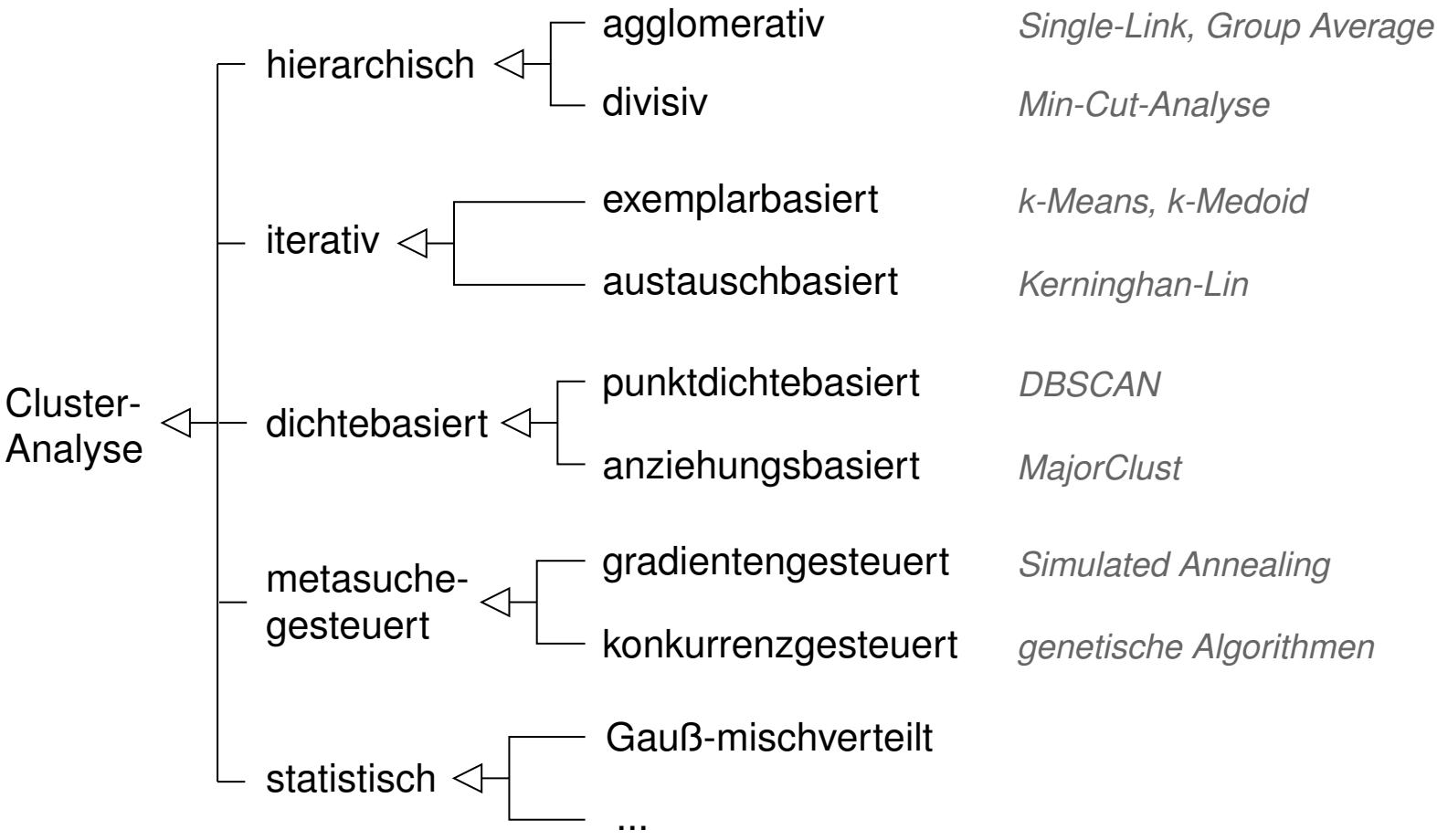
Cluster-Analyse im Vektorraum



1. Auswahl einer Retrieval-Funktion $\rho_{\mathcal{R}}$
2. Berechnung der Distanzmatrix mit Cluster-Abstandsmaß $d_{\mathcal{C}}$ z. B. als $1 - \rho_{\mathcal{R}}$
3. Berechnung eines Clusterings $\mathcal{C}$

Vektorraummodell

Cluster-Analyse im Vektorraum

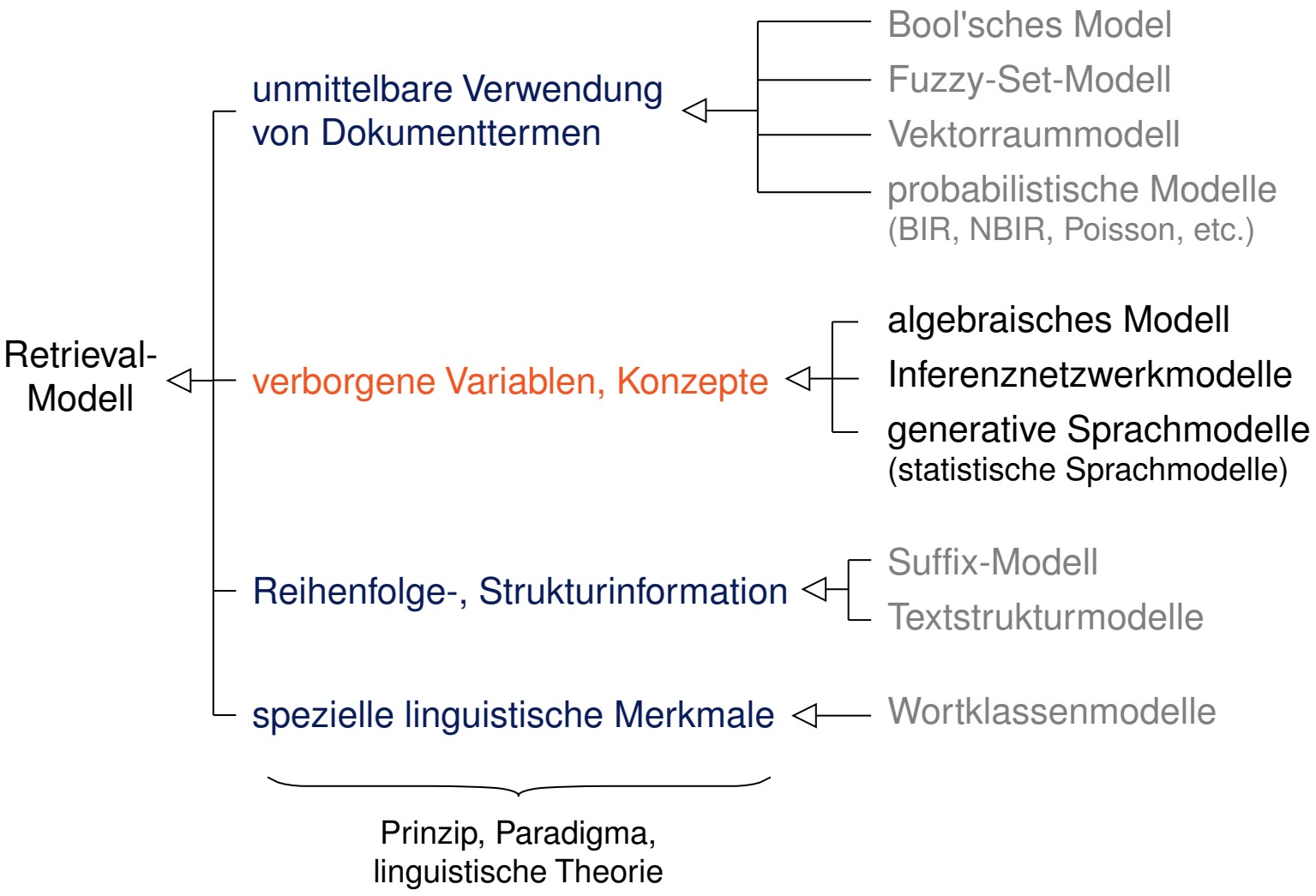


Kapitel IR: III (Fortsetzung)

III. Retrieval-Modelle

- Modelle und Prozesse im IR
- Klassische Retrieval-Modelle
- Bool'sches Modell
- Vektorraummodell
- Retrieval-Modelle mit verborgenen Variablen
- Algebraisches Modell

Retrieval-Modelle mit verborgenen Variablen



Retrieval-Modelle mit verborgenen Variablen

Offensichtlich haben die Terme eines Dokumentes $d \in D$ etwas mit der Semantik von d zu tun. Modelle mit verborgenen Variablen verlangen aber nicht, dass dieser Zusammenhang unmittelbar quantifizierbar ist.

Semantik *äußert sich* in den Termen von d, ist aber verknüpft mit dahinterliegenden Konzepten, Ideen oder einem intrinsischen Kanal, bzw. resultiert aus einem gemeinsamen kulturellen Hintergrund.

Unterscheidung von Retrieval-Modellen mit verborgenen Variablen:

1. Wofür steht eine verborgene Variable?
2. Art und Weise, wie verborgene Variablen mit d zusammenhängen.
3. Art und Umfang von Unabhängigkeitsannahmen.
4. Art und Weise, wie verborgene Variablen berechnet werden.
5. Art und Weise, wie sich die Retrieval-Funktion $\rho_{\mathcal{R}}(\mathbf{q}, \mathbf{d})$ berechnet.

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$ [vgl. ETH-Zürich 2001]

Überlegung:

In der $m \times n$ Term-Dokument-Matrix bestehen Korrelationen durch Synonyme, Kookkurrenzen, sich wiederholende Phrasen und n-Gramme.

Es besteht berechtigte Hoffnung, dass durch eine Koordinatentransformation die Dokumente des m-dimensionalen Vektorraums in einen Teilraum niedriger Dimension abgebildet und ausreichend genau approximiert werden können.

Idee:

Transformation der hochdimensionalen Dokumentvektoren in einen niedrigdimensionalen Raum bei möglichst genauer Erhaltung der Information.

Die hierbei entstehenden Linearkombinationen der Terme lassen sich als verborgene Konzepte interpretieren.

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$

Term-Dokument-Matrix:

	d_1	d_2	$\ldots$	d_n
t_1	w_{1_1}	w_{1_2}	$\ldots$	w_{1_n}
t_2	w_{2_1}	w_{2_2}	$\ldots$	w_{2_n}
$\vdots$				
t_m	w_{m_1}	w_{m_2}	$\ldots$	w_{m_n}

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$

Term-Dokument-Matrix:

	d_1	d_2	$\ldots$	d_n
t_1	w_{1_1}	w_{1_2}	$\ldots$	w_{1_n}
t_2	w_{2_1}	w_{2_2}	$\ldots$	w_{2_n}
$\vdots$				
t_m	w_{m_1}	w_{m_2}	$\ldots$	w_{m_n}

Kookkurrenz

	d_1	d_2	d_3	d_4
t_1	**2**	**7**	**4**	0
t_2	w_{2_1}	w_{2_2}	w_{2_3}	w_{2_4}
t_3	**2**	**6**	**3**	0
t_4	w_{4_1}	w_{4_2}	w_{4_4}	w_{4_4}

$t_1 \sim t_3$

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$

Term-Dokument-Matrix:

	d_1	d_2	$\ldots$	d_n
t_1	w_{1_1}	w_{1_2}	$\ldots$	w_{1_n}
t_2	w_{2_1}	w_{2_2}	$\ldots$	w_{2_n}
$\vdots$				
t_m	w_{m_1}	w_{m_2}	$\ldots$	w_{m_n}

Kookkurrenz

	d_1	d_2	d_3	d_4
t_1	**2**	**7**	**4**	0
t_2	w_{2_1}	w_{2_2}	w_{2_3}	w_{2_4}
t_3	**2**	**6**	**3**	0
t_4	w_{4_1}	w_{4_2}	w_{4_4}	w_{4_4}

$t_1 \sim t_3$

wiederholte Phrase

	d_1	d_2	d_3	d_4
t_1	**1**	**2**	**4**	0
t_2	w_{2_1}	w_{2_2}	w_{2_3}	w_{2_4}
t_3	**2**	**4**	**7**	0
t_4	**1**	**2**	**3**	0

$t_1 \sim 2 \cdot t_3 \wedge 1 \cdot t_4$

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$

Term-Dokument-Matrix:

	d_1	d_2	$\ldots$	d_n
t_1	w_{1_1}	w_{1_2}	$\ldots$	w_{1_n}
t_2	w_{2_1}	w_{2_2}	$\ldots$	w_{2_n}
$\vdots$				
t_m	w_{m_1}	w_{m_2}	$\ldots$	w_{m_n}

Kookkurrenz

	d_1	d_2	d_3	d_4
t_1	**2**	**7**	**4**	0
t_2	w_{2_1}	w_{2_2}	w_{2_3}	w_{2_4}
t_3	**2**	**6**	**3**	0
t_4	w_{4_1}	w_{4_2}	w_{4_4}	w_{4_4}

$t_1 \sim t_3$

wiederholte Phrase

	d_1	d_2	d_3	d_4
t_1	**1**	**2**	**4**	0
t_2	w_{2_1}	w_{2_2}	w_{2_3}	w_{2_4}
t_3	**2**	**4**	**7**	0
t_4	**1**	**2**	**3**	0

$t_1 \sim 2 \cdot t_3 \wedge 1 \cdot t_4$

Synonym

	d_1	d_2	d_3	d_4
t_1	**2**	**4**	**3**	0
t_2	w_{2_1}	w_{2_2}	w_{2_3}	w_{2_4}
t_3	**2**	**0**	**1**	0
t_4	**0**	**4**	**2**	0

$(t_1) \sim t_3 + t_4$

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$

Aus der linearen Algebra:

(1) Sei A eine $n \times n$-Matrix, λ ein Eigenwert von A mit Eigenvektor $\mathbf{x}$. Dann gilt:

$$A\mathbf{x} = \lambda\mathbf{x}$$

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$

Aus der linearen Algebra:

(1) Sei A eine $n \times n$-Matrix, λ ein Eigenwert von A mit Eigenvektor $\mathbf{x}$. Dann gilt:

$$A\mathbf{x} = \lambda\mathbf{x}$$

(2) Sei A eine symmetrische $n \times n$-Matrix vom Rang r. Dann ist A wie folgt darstellbar:

$$A = U \Lambda U^T$$

Λ ist eine mit den Eigenwerten von A besetzte $r \times r$-Diagonalmatrix
U ist eine $n \times r$-spaltenorthonormale Matrix: $U^T U = I$

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$

Aus der linearen Algebra:

(1) Sei A eine $n \times n$-Matrix, λ ein Eigenwert von A mit Eigenvektor $\mathbf{x}$. Dann gilt:

$$A\mathbf{x} = \lambda\mathbf{x}$$

(2) Sei A eine symmetrische $n \times n$-Matrix vom Rang r. Dann ist A wie folgt darstellbar:

$$A = U \Lambda U^T$$

Λ ist eine mit den Eigenwerten von A besetzte $r \times r$-Diagonalmatrix
U ist eine $n \times r$-spaltenorthonormale Matrix: $U^T U = I$

(3) Sei A eine $m \times n$-Matrix vom Rang r. Dann ist A wie folgt darstellbar:

$$A = U S V^T$$

S ist eine $r \times r$-Diagonalmatrix S
U ist eine spaltenorthonormale $m \times r$-Matrix
V ist eine spaltenorthonormale $n \times r$-Matrix

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$

Aus der linearen Algebra (Fortsetzung):

(4) Mit $A = USV^T$ gilt:

$$A^T A = (US\,V^T)^T(US\,V^T) = VSU^T US\,V^T = VS^2V^T$$

Die Spalten von V sind Eigenvektoren von $A^T A$.

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$

Aus der linearen Algebra (Fortsetzung):

(4) Mit $A = USV^T$ gilt:

$$A^T A = (US\,V^T)^T(US\,V^T) = VSU^T US\,V^T = VS^2V^T$$

Die Spalten von V sind Eigenvektoren von $A^T A$.

(5) und weiterhin:

$$AA^T = (US\,V^T)(US\,V^T)^T = US\,V^T VSU^T = US^2U^T$$

Die Spalten von U sind Eigenvektoren von AA^T.

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$

Aus der linearen Algebra (Fortsetzung):

(4) Mit $A = USV^T$ gilt:

$$A^T A = (US\,V^T)^T(US\,V^T) = VSU^T US\,V^T = VS^2V^T$$

Die Spalten von V sind Eigenvektoren von $A^T A$.

(5) und weiterhin:

$$AA^T = (US\,V^T)(US\,V^T)^T = US\,V^T VSU^T = US^2U^T$$

Die Spalten von U sind Eigenvektoren von AA^T.

(6) $A = US\,V^T$ lässt sich als Summe (dyadischer) Vektorprodukte schreiben:

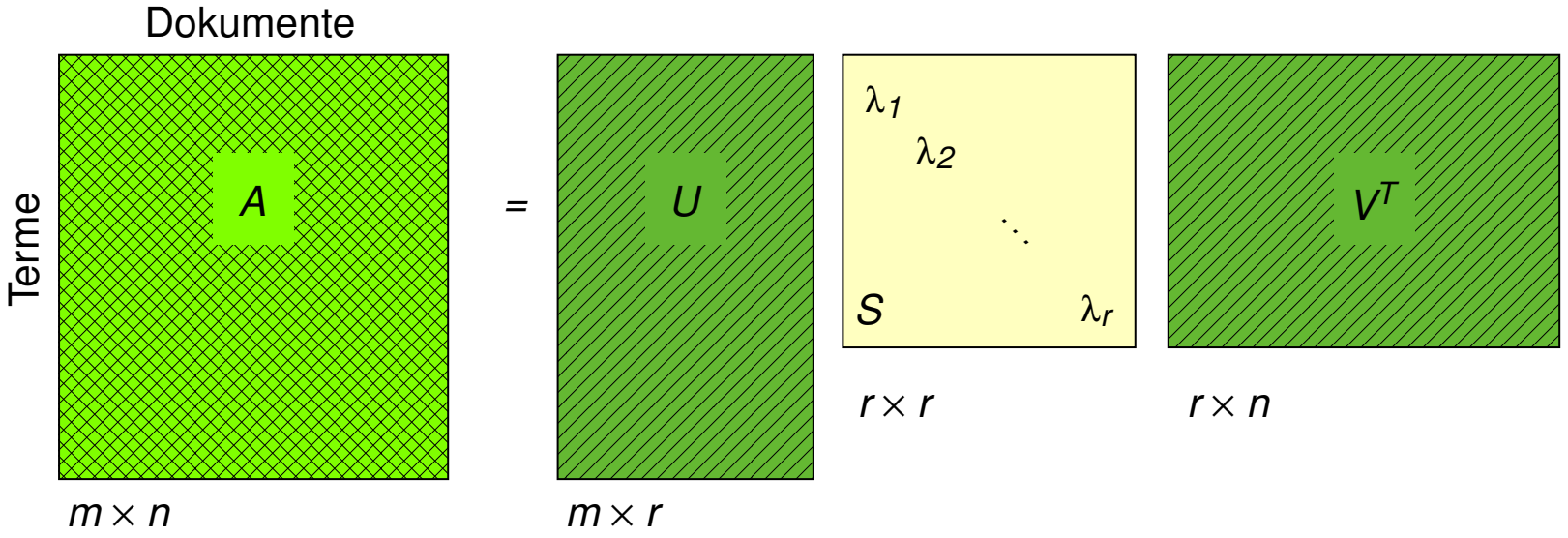
$$A = s_1(\mathbf{u}_1\mathbf{v}_1^T) + s_2(\mathbf{u}_2\mathbf{v}_2^T) + \ldots + s_r(\mathbf{u}_r\mathbf{v}_r^T)$$

Approximation von A durch Weglassen der Summanden mit den kleinsten Singulärwerten.

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$

Singulärwertzerlegung $A = U S V^T$:



U ist Spalten-orthonormal

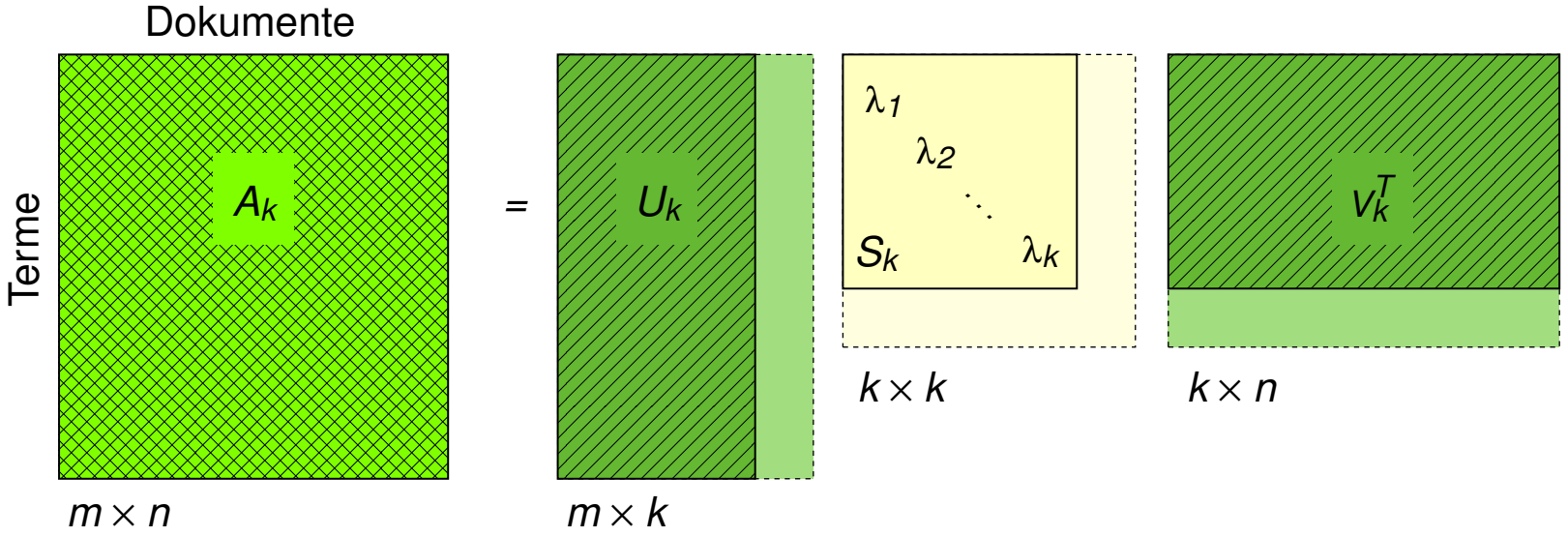
S ist diagonal, $r \leq \min\{m,n\}$

V^T *ist Zeilen-orthonormal*

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$

Dimensionsreduktion $A_k = U_k S_k V_k^T$:



U_k ist Spalten-orthonormal

S_k ist diagonal, $k<r$

V_k^T ist Zeilen-orthonormal

Bemerkungen:

- Die Eigenwerte für A ergeben sich aus der Gleichung $\det(A - \lambda I) = 0$. Die Gleichung definiert ein Polynom n-ten Grades und hat folglich n Wurzeln, die real oder komplex und mehrfach sein können. Die zugehörigen Eigenvektoren sind orthogonal.
- Eine symmetrische Matrix hat reale Eigenwerte. Eine positiv-definite Matrix hat nur positive Eigenwerte.
- Die Singulärwertzerlegung verallgemeinert die Eigenwertzerlegung auf allgemeine Rechtecksmatrizen.

- Matrix-Multiplikation und -Transposition: $(AB)^T = B^T A^T$
- Diagonalisierung bzw. Eigenzerlegung einer quadratischen Matrix A: $A = PDP^{-1}$, D ist Diagonalmatrix mit den Eigenwerten von A, P enthält die Eigenvektoren von A. A ist genau dann diagonalisierbar, falls sie n linear unabhängige Eigenvektoren besitzt.
- $U^T = U^{-1}$, falls U eine orthogonale Matrix ist.
- $U^T U = I$, falls U eine spaltenorthonormale Matrix ist.
- $U^T = U$, falls U eine symmetrische Matrix ist.

- Die Reduktion der $r \times r$-Diagonalmatrix S auf die kleinere $k \times k$-Diagonalmatrix S_k geschieht durch Weglassen der kleinsten Diagonalelemente bei entsprechender Sortierung der Spaltenvektoren in U_k und V_k.

Algebraisches Modell

Dokumentmodell $\langle \mathbf{D}, \mathbf{Q}, \rho_{\mathcal{R}} \rangle$ [vgl. allgemeines Dokumentmodell]

Dokumentrepräsentationen $\mathbf{D}$.

1. Die Dokumentrepräsentationen des Vektorraummodells werden zu einer $m \times n$ Term-Dokument-Matrix A zusammengefasst.
2. A wird durch Dimensionsreduktion zur Konzept-Dokument-Matrix $\mathbf{D} = V_k^T$. $\mathbf{D}$ repräsentiert die Dokumente im Konzeptraum (latent semantic space).

Formalisierte Anfragenmenge $\mathbf{Q}$.

Ausgangspunkt einer formalen Anfrage ist ihre Vektorraumrepräsentation $\mathbf{q}$. Durch folgende Operation wird $\mathbf{q}$ in den Konzeptraum transformiert:

$$\mathbf{q}' = \mathbf{q}^T U_k S_k^{-1}$$

Retrieval-Funktion $\rho_{\mathcal{R}}$.

$\rho_{\mathcal{R}}$ wird unmittelbar auf die Darstellungen der Dokumente und Anfragen im Konzeptraum angewandt. Dabei kommen Retrieval-Funktionen wie beim Vektorraummodell zum Einsatz.

Bemerkungen:

- Dieses auf Singulärwertzerlegung beruhende Retrieval-Modell wurde 1988 von Derweester et. al. entwickelt und unter dem Namen „Latent Semantic Indexing“, LSI, vorgestellt.

Algebraisches Modell

Beispiel [vgl. ETH-Zürich 2001]

Dokumentkollektion:

d_1	Human machine interface for Lab ABC computer applications.
d_2	A survey of user opinion of computer system response time.
d_3	The EPS user interface management system.
d_4	System and human system engineering testing of EPS.
d_5	Relation of user-perceived response time to error measurement.
d_6	The generation of random, binary, unordered trees.
d_7	The intersection graph of paths in trees.
d_8	Graph minors IV: Widths of trees and well-quasi-ordering.
d_9	Graph minors: A survey

Algebraisches Modell

Beispiel [vgl. ETH-Zürich 2001]

Dokumentkollektion:

d_1	Human machine interface for Lab ABC computer applications.
d_2	A survey of user opinion of computer system response time.
d_3	The EPS user interface management system.
d_4	System and human system engineering testing of EPS.
d_5	Relation of user-perceived response time to error measurement.
d_6	The generation of random, binary, unordered trees.
d_7	The intersection graph of paths in trees.
d_8	Graph minors IV: Widths of trees and well-quasi-ordering.
d_9	Graph minors: A survey

Anfrage $\mathbf{q}$ = { human, computer, interaction }

Bemerkungen:

- Anfrageauswertung im Originalraum unter dem Bool'schen Retrieval-Modell bei einer $\wedge$-Verknüpfung der Terme in $\mathbf{q}$: Result-Set $R = \emptyset$.
- Anfrageauswertung im Originalraum unter dem Bool'schen Modell bei einer $\vee$-Verknüpfung der Terme in $\mathbf{q}$: Result-Set $R = \{d_1, d_2, d_4\}$.
- Anfrageauswertung im Originalraum unter dem Vektorraummodell: Result-Set $R = \{d_1, d_2, d_4\}$.

Algebraisches Modell

Beispiel: Term-Dokument-Matrix A

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
human	1			1					
interface	1		1						
computer	1	1							
user		1	1		1				
system		1	1	2					
response		1			1				
time		1			1				
EPS			1	1					
survey		1							1
trees						1	1	1	
graph							1	1	1
minors								1	1

Terme, die nur in einem Dokument vorkommen, und Stopworte wurden nicht berücksichtigt.

Algebraisches Modell

Beispiel: Singulärwertzerlegung $A = USV^T$

$U =$

0.2214	-0.1132	0.2890	-0.4148	-0.1063	-0.3410	0.5227	-0.0605	-0.4067
0.1976	-0.0721	0.1350	-0.5522	0.2818	0.4959	-0.0704	-0.0099	-0.1089
0.2405	0.0432	-0.1644	-0.5950	-0.1068	-0.2550	-0.3022	0.0623	0.4924
0.4036	0.0571	-0.3378	0.0991	0.3317	0.3848	0.0029	-0.0004	0.0123
0.6445	-0.1673	0.3611	0.3335	-0.1590	-0.2065	-0.1658	0.0343	0.2707
0.2650	0.1072	-0.4260	0.0738	0.0803	-0.1697	0.2829	-0.0161	-0.0539
0.2650	0.1072	-0.4260	0.0738	0.0803	-0.1697	0.2829	-0.0161	-0.0539
0.3008	-0.1413	0.3303	0.1881	0.1148	0.2722	0.0330	-0.0190	-0.1653
0.2059	0.2736	-0.1776	-0.0324	-0.5372	0.0809	-0.4669	-0.0363	-0.5794
0.0127	0.4902	0.2311	0.0248	0.5942	-0.3921	-0.2883	0.2546	-0.2254
0.0361	0.6228	0.2231	0.0007	-0.0683	0.1149	0.1596	-0.6811	0.2320
0.0318	0.4505	0.1411	-0.0087	-0.3005	0.2773	0.3395	0.6784	0.1825

Algebraisches Modell

Beispiel: Singulärwertzerlegung $A = USV^T$

$U =$

0.2214	-0.1132	0.2890	-0.4148	-0.1063	-0.3410	0.5227	-0.0605	-0.4067
0.1976	-0.0721	0.1350	-0.5522	0.2818	0.4959	-0.0704	-0.0099	-0.1089
0.2405	0.0432	-0.1644	-0.5950	-0.1068	-0.2550	-0.3022	0.0623	0.4924
0.4036	0.0571	-0.3378	0.0991	0.3317	0.3848	0.0029	-0.0004	0.0123
0.6445	-0.1673	0.3611	0.3335	-0.1590	-0.2065	-0.1658	0.0343	0.2707
0.2650	0.1072	-0.4260	0.0738	0.0803	-0.1697	0.2829	-0.0161	-0.0539
0.2650	0.1072	-0.4260	0.0738	0.0803	-0.1697	0.2829	-0.0161	-0.0539
0.3008	-0.1413	0.3303	0.1881	0.1148	0.2722	0.0330	-0.0190	-0.1653
0.2059	0.2736	-0.1776	-0.0324	-0.5372	0.0809	-0.4669	-0.0363	-0.5794
0.0127	0.4902	0.2311	0.0248	0.5942	-0.3921	-0.2883	0.2546	-0.2254
0.0361	0.6228	0.2231	0.0007	-0.0683	0.1149	0.1596	-0.6811	0.2320
0.0318	0.4505	0.1411	-0.0087	-0.3005	0.2773	0.3395	0.6784	0.1825

$S =$

3.3409								
	2.5417							
		2.3539						
			1.6445					
				1.5048				
					1.3064			
						0.8459		
							0.5601	
								0.3637

Algebraisches Modell

Beispiel: Singulärwertzerlegung $A = USV^T$

$U =$

0.2214	-0.1132	0.2890	-0.4148	-0.1063	-0.3410	0.5227	-0.0605	-0.4067
0.1976	-0.0721	0.1350	-0.5522	0.2818	0.4959	-0.0704	-0.0099	-0.1089
0.2405	0.0432	-0.1644	-0.5950	-0.1068	-0.2550	-0.3022	0.0623	0.4924
0.4036	0.0571	-0.3378	0.0991	0.3317	0.3848	0.0029	-0.0004	0.0123
0.6445	-0.1673	0.3611	0.3335	-0.1590	-0.2065	-0.1658	0.0343	0.2707
0.2650	0.1072	-0.4260	0.0738	0.0803	-0.1697	0.2829	-0.0161	-0.0539
0.2650	0.1072	-0.4260	0.0738	0.0803	-0.1697	0.2829	-0.0161	-0.0539
0.3008	-0.1413	0.3303	0.1881	0.1148	0.2722	0.0330	-0.0190	-0.1653
0.2059	0.2736	-0.1776	-0.0324	-0.5372	0.0809	-0.4669	-0.0363	-0.5794
0.0127	0.4902	0.2311	0.0248	0.5942	-0.3921	-0.2883	0.2546	-0.2254
0.0361	0.6228	0.2231	0.0007	-0.0683	0.1149	0.1596	-0.6811	0.2320
0.0318	0.4505	0.1411	-0.0087	-0.3005	0.2773	0.3395	0.6784	0.1825

$S =$

3.3409								
	2.5417							
		2.3539						
			1.6445					
				1.5048				
					1.3064			
						0.8459		
							0.5601	
								0.3637

$V^T =$

0.1974	0.6060	0.4629	0.5421	0.2795	0.0038	0.0146	0.0241	0.0820
-0.0559	0.1656	-0.1273	-0.2318	0.1068	0.1928	0.4379	0.6151	0.5299
0.1103	-0.4973	0.2076	0.5699	-0.5054	0.0982	0.1930	0.2529	0.0793
-0.9498	-0.0286	0.0416	0.2677	0.1500	0.0151	0.0155	0.0102	-0.0246
0.0457	-0.2063	0.3783	-0.2056	0.3272	0.3948	0.3495	0.1498	-0.6020
-0.0766	-0.2565	0.7244	-0.3689	0.0348	-0.3002	-0.2122	0.0001	0.3622
0.1773	-0.4330	-0.2369	0.2648	0.6723	-0.3408	-0.1522	0.2491	0.0380
-0.0144	0.0493	0.0088	-0.0195	-0.0583	0.4545	-0.7615	0.4496	-0.0696
-0.0637	0.2428	0.0241	-0.0842	-0.2624	-0.6198	0.0180	0.5199	-0.4535

Algebraisches Modell

Beispiel: Dimensionsreduktion $A_k = U_k S_k V_k^T$

U_k

0.2214	-0.1132
0.1976	-0.0721
0.2405	0.0432
0.4036	0.0571
0.6445	-0.1673
0.2650	0.1072
0.2650	0.1072
0.3008	-0.1413
0.2059	0.2736
0.0127	0.4902
0.0361	0.6228
0.0318	0.4505

S_k

3.3409	
	2.5417

V_k^T

0.1974	0.6060	0.4629	0.5421	0.2795	0.0038	0.0146	0.0241	0.0820
-0.0559	0.1656	-0.1273	-0.2318	0.1068	0.1928	0.4379	0.6151	0.5299

Algebraisches Modell

Beispiel: Dimensionsreduktion $A_k = U_k S_k V_k^T$

U_k

0.2214	-0.1132
0.1976	-0.0721
0.2405	0.0432
0.4036	0.0571
0.6445	-0.1673
0.2650	0.1072
0.2650	0.1072
0.3008	-0.1413
0.2059	0.2736
0.0127	0.4902
0.0361	0.6228
0.0318	0.4505

S_k

3.3409	
	2.5417

V_k^T

0.1974	0.6060	0.4629	0.5421	0.2795	0.0038	0.0146	0.0241	0.0820
-0.0559	0.1656	-0.1273	-0.2318	0.1068	0.1928	0.4379	0.6151	0.5299

A_k

0.1621	0.4005	0.3790	0.4676	0.1760	-0.0527	-0.1151	-0.1591	-0.0918
0.1406	0.3698	0.3290	0.4004	0.1650	-0.0328	-0.0706	-0.0968	-0.0430
0.1524	0.5050	0.3579	0.4101	0.2362	0.0242	0.0598	0.0869	0.1240
0.2580	0.8411	0.6057	0.6974	0.3923	0.0331	0.0832	0.1218	0.1874
0.4488	1.2344	1.0509	1.2658	0.5563	-0.0738	-0.1547	-0.2096	-0.0489
0.1596	0.5817	0.3752	0.4169	0.2765	0.0559	0.1322	0.1889	0.2169
0.1596	0.5817	0.3752	0.4169	0.2765	0.0559	0.1322	0.1889	0.2169
0.2185	0.5496	0.5110	0.6281	0.2425	-0.0654	-0.1425	-0.1966	-0.1079
0.0969	0.5321	0.2299	0.2118	0.2665	0.1368	0.3146	0.4444	0.4250
-0.0613	0.2321	-0.1389	-0.2656	0.1449	0.2404	0.5461	0.7674	0.6637
-0.0647	0.3353	-0.1456	-0.3014	0.2028	0.3057	0.6949	0.9766	0.8487
-0.0431	0.2539	-0.0967	-0.2079	0.1519	0.2212	0.5029	0.7069	0.6155

Algebraisches Modell

Beispiel: Dimensionsreduktion $A_k = U_k S_k V_k^T$

U_k

0.2214	-0.1132
0.1976	-0.0721
0.2405	0.0432
0.4036	0.0571
0.6445	-0.1673
0.2650	0.1072
0.2650	0.1072
0.3008	-0.1413
0.2059	0.2736
0.0127	0.4902
0.0361	0.6228
0.0318	0.4505

S_k

3.3409
2.5417

V_k^T

0.1974	0.6060	0.4629	0.5421	0.2795	0.0038	0.0146	0.0241	0.0820
-0.0559	0.1656	-0.1273	-0.2318	0.1068	0.1928	0.4379	0.6151	0.5299

A_k

0.1621	0.4005	0.3790	0.4676	0.1760	-0.0527	-0.1151	-0.1591	-0.0918
0.1406	0.3698	0.3290	0.4004	0.1650	-0.0328	-0.0706	-0.0968	-0.0430
0.1524	0.5050	0.3579	0.4101	0.2362	0.0242	0.0598	0.0869	0.1240
0.2580	0.8411	0.6057	0.6974	0.3923	0.0331	0.0832	0.1218	0.1874
0.4488	1.2344	1.0509	1.2658	0.5563	-0.0738	-0.1547	-0.2096	-0.0489
0.1596	0.5817	0.3752	0.4169	0.2765	0.0559	0.1322	0.1889	0.2169
0.1596	0.5817	0.3752	0.4169	0.2765	0.0559	0.1322	0.1889	0.2169
0.2185	0.5496	0.5110	0.6281	0.2425	-0.0654	-0.1425	-0.1966	-0.1079
0.0969	0.5321	0.2299	0.2118	0.2665	0.1368	0.3146	0.4444	0.4250
-0.0613	0.2321	-0.1389	-0.2656	0.1449	0.2404	0.5461	0.7674	0.6637
-0.0647	0.3353	-0.1456	-0.3014	0.2028	0.3057	0.6949	0.9766	0.8487
-0.0431	0.2539	-0.0967	-0.2079	0.1519	0.2212	0.5029	0.7069	0.6155

$\mathbf{q}$

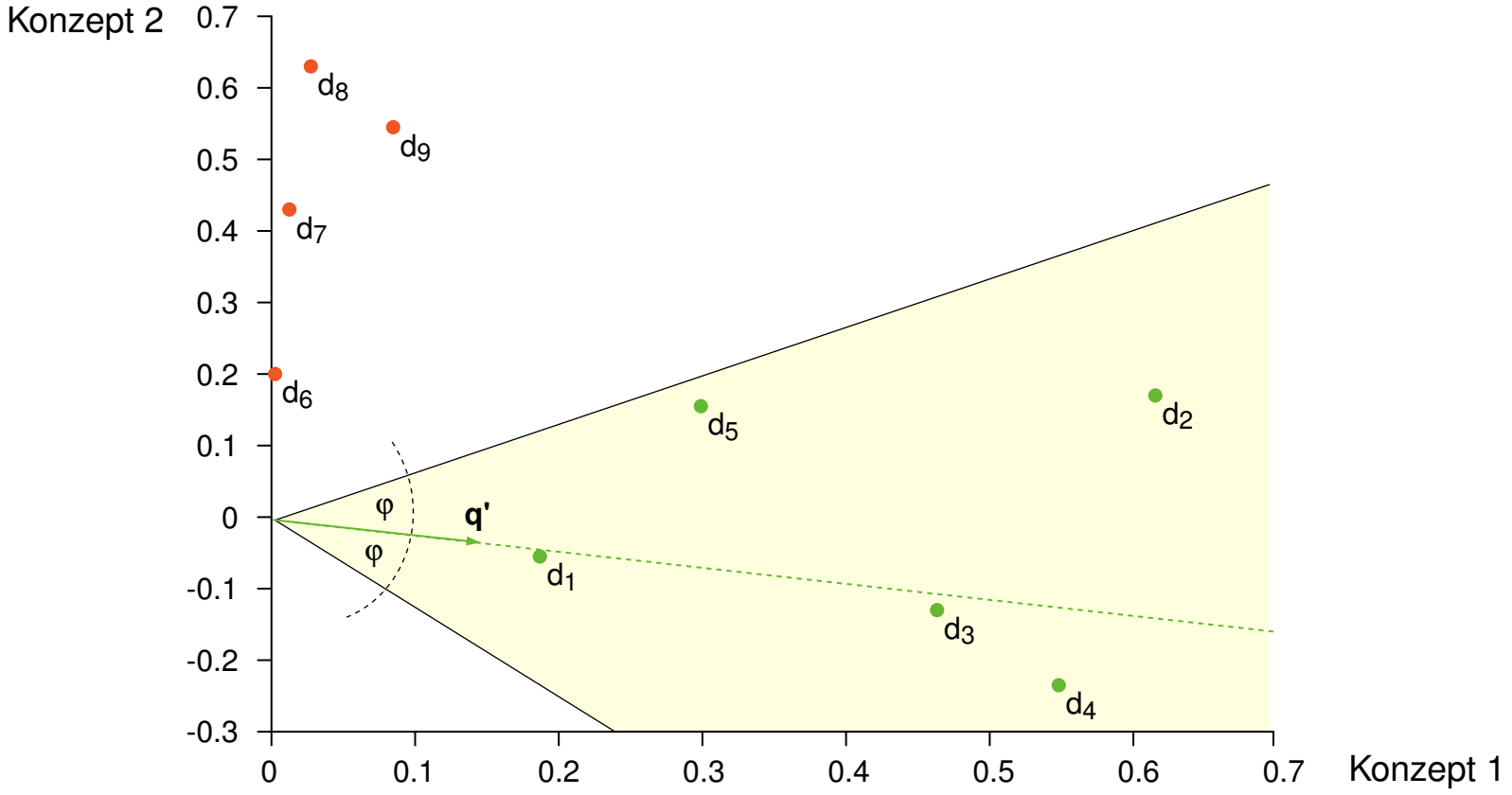
1
0
1
0
0
0
0
0
0
0
0
0

$\mathbf{q}' = \mathbf{q}^T U_k S_k^{-1}$

0.1382
-0.0276

Algebraisches Modell

Beispiel: Anfrageauswertung im Konzeptraum

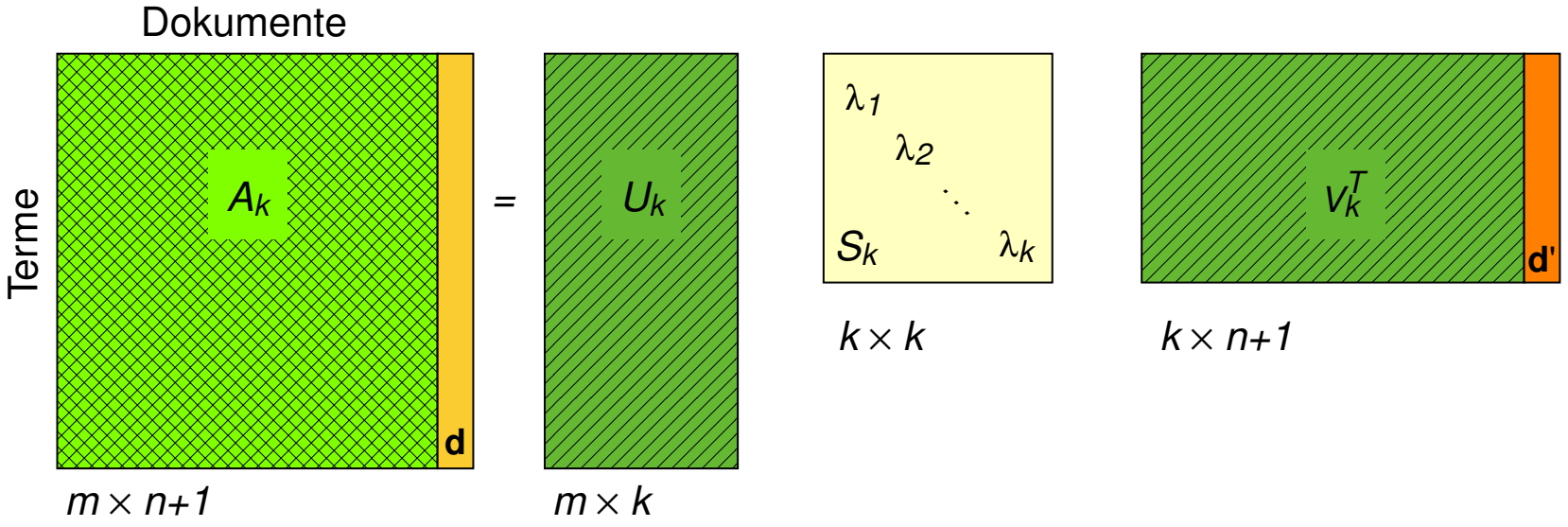


$\varphi = 30°$ ➜ Dokumente müssen zu dem Anfragevektor $\mathbf{q}'$ eine Ähnlichkeit >87% aufweisen.

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$ (Fortsetzung)

Einfügen neuer Dokumente:

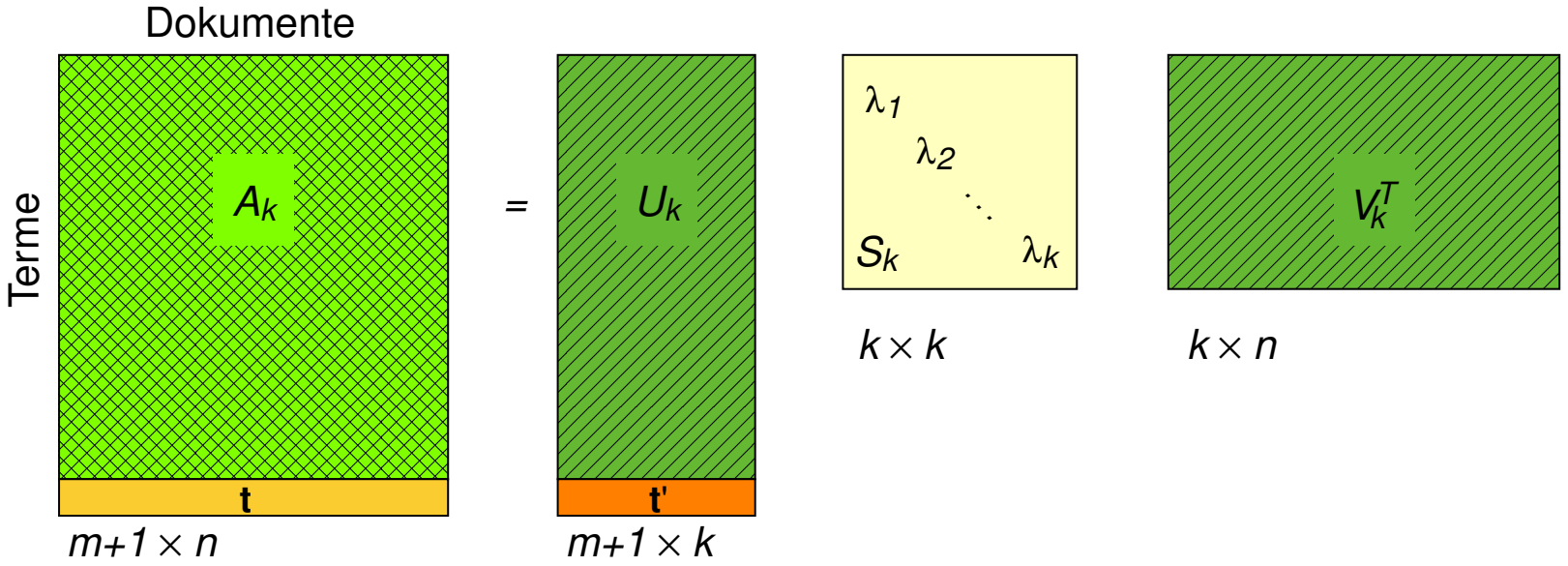


1. originalen Dokumentvektor $\mathbf{d}$ an A_k hängen
2. reduzierten Dokumentvektor $\mathbf{d}' = \mathbf{d}^T U_k S_k^{-1}$ berechnen (vgl. formalisierte Anfrage)
3. reduzierten Dokumentvektor $\mathbf{d}'$ an V_k^T hängen

Algebraisches Modell

Retrieval-Modell $\mathcal{R}$ (Fortsetzung)

Einfügen neuer Terme:



1. originalen Termvektor $\mathbf{t}$ unten bei A_k hinzufügen
2. reduzierten Termvektor $\mathbf{t}' = \mathbf{t}^T V_k S_k^{-1}$ berechnen
3. reduzierten Termvektor $\mathbf{t}'$ unten bei U_k hinzufügen

Algebraisches Modell

Beispiel 2 [vgl. ETH-Zürich 2001]

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
data	1	2	1	5	0	0	0
information	1	2	1	5	0	0	0
retrieval	1	2	1	5	0	0	0
brain	0	0	0	0	2	3	1
lung	0	0	0	0	2	3	1

Algebraisches Modell

Beispiel 2 [vgl. ETH-Zürich 2001]

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
data	1	2	1	5	0	0	0
information	1	2	1	5	0	0	0
retrieval	1	2	1	5	0	0	0
brain	0	0	0	0	2	3	1
lung	0	0	0	0	2	3	1

$A = U S V^T$, approximiert: $A_k = U_k S_k V_k^T$

Rang$(A) = 2$, und so folgt mit $k = 2$, dass $A_2 = A,\ U_2 = U,\ S_2 = S,\ V_2^T = V^T$:

Algebraisches Modell

Beispiel 2 [vgl. ETH-Zürich 2001]

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
data	1	2	1	5	0	0	0
information	1	2	1	5	0	0	0
retrieval	1	2	1	5	0	0	0
brain	0	0	0	0	2	3	1
lung	0	0	0	0	2	3	1

$A = U S V^T$, approximiert: $A_k = U_k S_k V_k^T$

Rang$(A) = 2$, und so folgt mit $k = 2$, dass $A_2 = A,\ U_2 = U,\ S_2 = S,\ V_2^T = V^T$:

$$A = \begin{pmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \\ 0 & 0.71 \end{pmatrix} \times \begin{pmatrix} 9.64 & 0 \\ 0 & 5.29 \end{pmatrix} \times \begin{pmatrix} 0.18 & 0.36 & 0.18 & 0.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.53 & 0.8 & 0.27 \end{pmatrix}$$

Bemerkungen:

- Es gibt zwei Konzepte, das Informatikkonzept {data, information, retrieval} und das Medizinkonzept {brain, lung}.

Algebraisches Modell

Beispiel 2: Dokumentähnlichkeitsmatrix $A^T A$

$$A^T A = \begin{pmatrix} 3 & 6 & 6 & 15 & 0 & 0 & 0 \\ 6 & 12 & 6 & 30 & 0 & 0 & 0 \\ 3 & 6 & 6 & 15 & 0 & 0 & 0 \\ 15 & 37 & 15 & 75 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 12 & 4 \\ 0 & 0 & 0 & 0 & 12 & 18 & 6 \\ 0 & 0 & 0 & 0 & 4 & 6 & 2 \end{pmatrix}$$

Algebraisches Modell

Beispiel 2: Dokumentähnlichkeitsmatrix $A^T A$

$$A^T A = \begin{pmatrix} \mathbf{3} & \mathbf{6} & \mathbf{6} & \mathbf{15} & 0 & 0 & 0 \\ \mathbf{6} & \mathbf{12} & \mathbf{6} & \mathbf{30} & 0 & 0 & 0 \\ \mathbf{3} & \mathbf{6} & \mathbf{6} & \mathbf{15} & 0 & 0 & 0 \\ \mathbf{15} & \mathbf{37} & \mathbf{15} & \mathbf{75} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{8} & \mathbf{12} & \mathbf{4} \\ 0 & 0 & 0 & 0 & \mathbf{12} & \mathbf{18} & \mathbf{6} \\ 0 & 0 & 0 & 0 & \mathbf{4} & \mathbf{6} & \mathbf{2} \end{pmatrix}$$

Interpretation. $A^T A$ zeigt die Dokument-Cluster.

Algebraisches Modell

Beispiel 2: Dokumentähnlichkeitsmatrix $A^T A$

$$A^T A = \begin{pmatrix} \mathbf{3} & \mathbf{6} & \mathbf{6} & \mathbf{15} & 0 & 0 & 0 \\ \mathbf{6} & \mathbf{12} & \mathbf{6} & \mathbf{30} & 0 & 0 & 0 \\ \mathbf{3} & \mathbf{6} & \mathbf{6} & \mathbf{15} & 0 & 0 & 0 \\ \mathbf{15} & \mathbf{37} & \mathbf{15} & \mathbf{75} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{8} & \mathbf{12} & \mathbf{4} \\ 0 & 0 & 0 & 0 & \mathbf{12} & \mathbf{18} & \mathbf{6} \\ 0 & 0 & 0 & 0 & \mathbf{4} & \mathbf{6} & \mathbf{2} \end{pmatrix}$$

Interpretation. $A^T A$ zeigt die Dokument-Cluster.

Erklärung. Wegen $A^T A = V S^2 V^T$ sind die Zeilen von V_k^T die Eigenvektoren von $A^T A$, sie beschreiben die unkorrelierten Hauptrichtungen für Dokument-Cluster:

$$V_2^T = \begin{pmatrix} 0.18 & 0.36 & 0.18 & 0.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.53 & 0.8 & 0.27 \end{pmatrix}$$

➜ Falls d_1 relevant ist, so sind es auch d_2, d_3, d_4, nicht jedoch d_5, d_6, d_7.

Algebraisches Modell

Beispiel 2: Termähnlichkeitsmatrix AA^T

$$AA^T = \begin{pmatrix} 31 & 31 & 31 & 0 & 0 \\ 31 & 31 & 31 & 0 & 0 \\ 31 & 31 & 31 & 0 & 0 \\ 0 & 0 & 0 & 14 & 14 \\ 0 & 0 & 0 & 14 & 14 \end{pmatrix}$$

Algebraisches Modell

Beispiel 2: Termähnlichkeitsmatrix AA^T

$$AA^T = \begin{pmatrix} \mathbf{31} & \mathbf{31} & \mathbf{31} & 0 & 0 \\ \mathbf{31} & \mathbf{31} & \mathbf{31} & 0 & 0 \\ \mathbf{31} & \mathbf{31} & \mathbf{31} & 0 & 0 \\ 0 & 0 & 0 & \mathbf{14} & \mathbf{14} \\ 0 & 0 & 0 & \mathbf{14} & \mathbf{14} \end{pmatrix}$$

Interpretation. AA^T zeigt die Term-Cluster bzw. Konzepte, evtl. Synonyme.

Algebraisches Modell

Beispiel 2: Termähnlichkeitsmatrix AA^T

$$AA^T = \begin{pmatrix} \mathbf{31} & \mathbf{31} & \mathbf{31} & 0 & 0 \\ \mathbf{31} & \mathbf{31} & \mathbf{31} & 0 & 0 \\ \mathbf{31} & \mathbf{31} & \mathbf{31} & 0 & 0 \\ 0 & 0 & 0 & \mathbf{14} & \mathbf{14} \\ 0 & 0 & 0 & \mathbf{14} & \mathbf{14} \end{pmatrix}$$

Interpretation. AA^T zeigt die Term-Cluster bzw. Konzepte, evtl. Synonyme.

Erklärung. Wegen $AA^T = US^2U^T$ sind die Spalten von U_k die Eigenvektoren von AA^T, sie beschreiben die unkorrelierten Hauptrichtungen für Konzepte:

$$U_2 = \begin{pmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \\ 0 & 0.71 \end{pmatrix}$$

Algebraisches Modell

Diskussion

Vorteile:

- automatische Entdeckung verborgener Konzepte
- syntaktische Erkennung von Synonymen
- semantische Erweiterung von Anfragen aufgrund syntaktischer Analyse – und nicht durch Relevanz-Feedback oder die Bemühung von Thesauri

Nachteile:

- die Wirkungsweise von LSI ist nicht vollständig verstanden; eine theoretisch fundierte Brücke zur Linguistik ist nur ansatzweise vorhanden
- LSI entfaltet die volle Wirkung nur in einer *geschlossenen Retrieval-Situation*: die Kollektion ist bekannt, gegeben und ändert sich nur wenig
- die Singulärwertzerlegung ist rechenaufwendig, $O(n^3)$