Information Retrieval

Norbert Fuhr

10. April 2006

Teil I

Einführung

Beispiele

- Internet-Suche
- Suche in Online-Dokumentationen
- Suche in Digitalen Bibliotheken
- Suche in Bildarchiven

Unterschiede zur Suche in klassischen Datenbanken:

- Schwierigkeit, passende Anfrage zu formulieren
- iterative Anfrageformulierung (abhängig von Antworten)
- viele Antworten, aber wenige davon relevant
- Rangordnung der Antworten (statt Antwortmenge)
- Repräsentation des Inhalts von Dokumenten inadäquat / unsicher

Was ist IR?

Salton (1968):
Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.

Definition der Fachgruppe IR in der GI:
Im Information Retrieval (IR) werden Informationssysteme in bezug auf ihre Rolle im Prozeß des Wissenstransfers vom menschlichen Wissensproduzenten zum Informations-Nachfragenden betrachtet. Die Fachgruppe „Information Retrieval“ in der Gesellschaft für Informatik beschäftigt sich dabei schwerpunktmäßig mit jenen Fragestellungen, die im Zusammenhang mit *vagen Anfragen* und *unsicherem Wissen* entstehen.

IR = Unsicherheit und Vagheit in IS

Vagheit: Benutzer kann seinen Informationswunsch nicht präzise spezifizieren

- vage Anfragebedingungen
- iterative Frageformulierung

Unsicherheit System besitzt unsicheres (unzureichendes) Wissen über den Inhalt der verwalteten Objekte

- unsichere Repräsentation
  ($\rightsquigarrow$ fehlerhafte Antworten)
- unvollständige Repräsentation
  ($\rightsquigarrow$ fehlende Antworten)

IR = inhaltsorientierte Suche

(engere Definition)
Suche auf verschiedenen Abstraktionsstufen:

Syntax Dokument als Folge von Symbolen
(z.B. Zeichenkettensuche in Texten, Farbe/Textur/Kontur in Bildern)

Semantik Bedeutung eines Dokumentes
(z.B. Textsemantik, in einem Bild vorkommende Objekte)

Pragmatik Nutzung eines Dokumentes (Zweck)
(z.B.: Löst das Dokument mein Problem? Was ist die Aussage des Textes / Bildes?)

IR beschäftigt sich mit der Semantik und Pragmatik von Dokumenten

IR-Anwendungen

- Internet-Suchmaschinen und -Kataloge
- Digitale Bibliotheken
- Suche in Nachrichten-, Mail-, News-Archiven
- Suche in technischen Dokumentationen
- Nachrichtenfilter

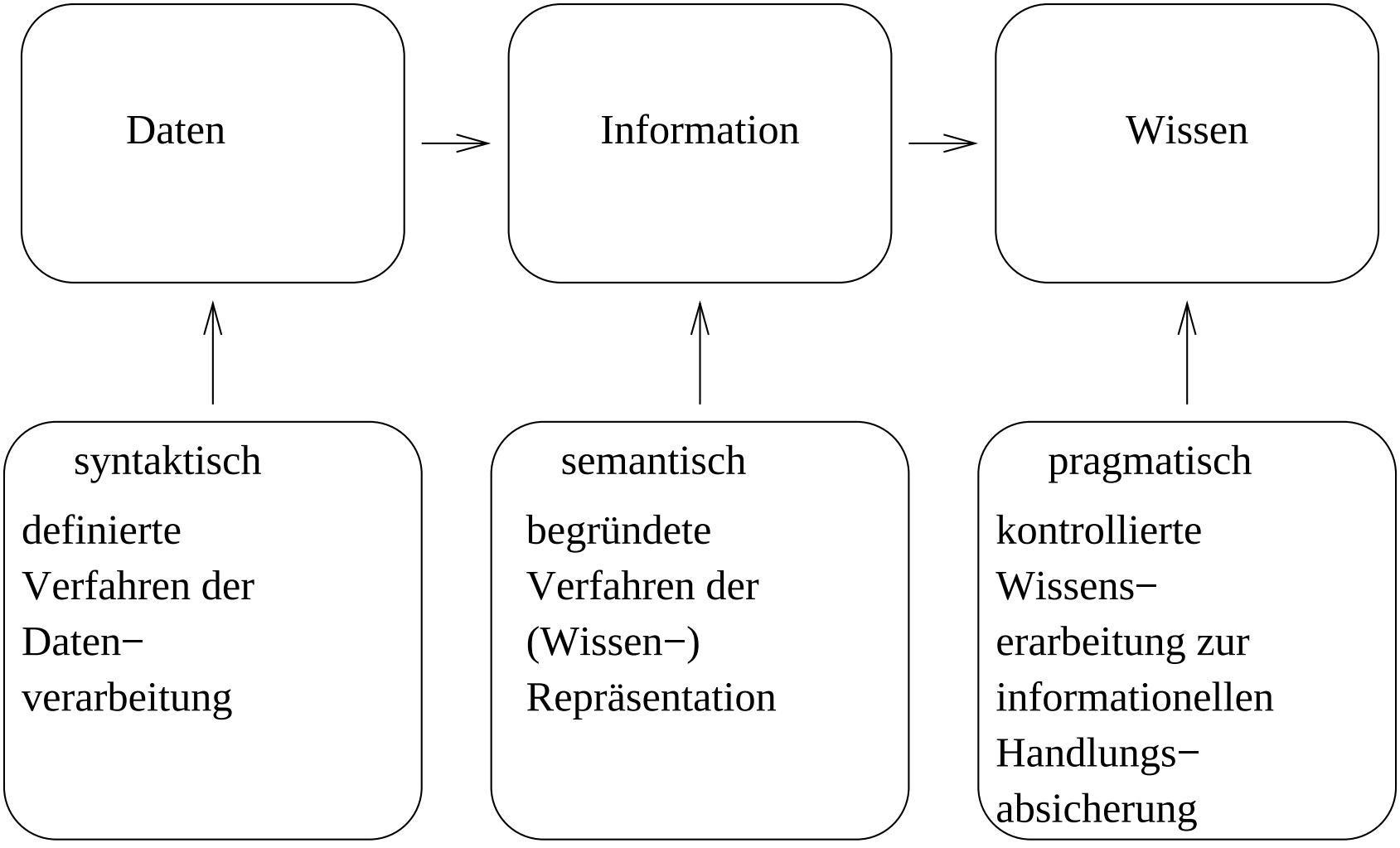
Teil II

IR-Konzepte

Dimensionen des IR

Matching	exakt	partiell, best match
Inferenz	Deduktion	Induktion
Modell	deterministisch	probabilistisch
Klassifikation	monothetisch	polithetisch
Anfragesprache	formal	natürlich
Fragespezifikation	vollständig	unvollständig
gesuchte Objekte	die Fragespezif. erfüllende	relevante
Reaktion auf Datenfehler	sensitiv	insensitiv

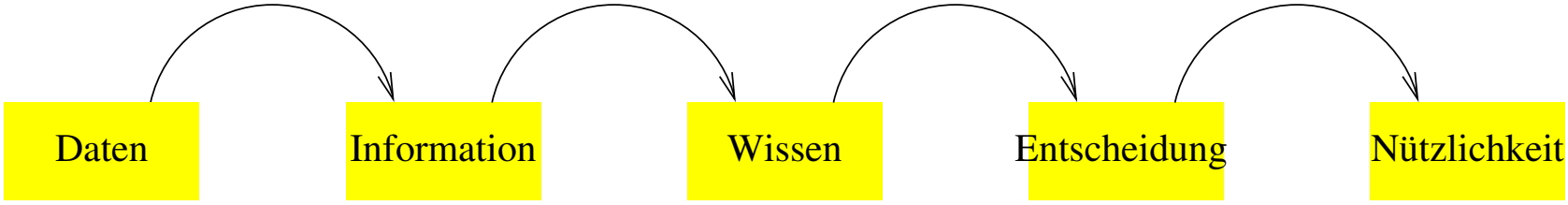
Daten — Information — Wissen



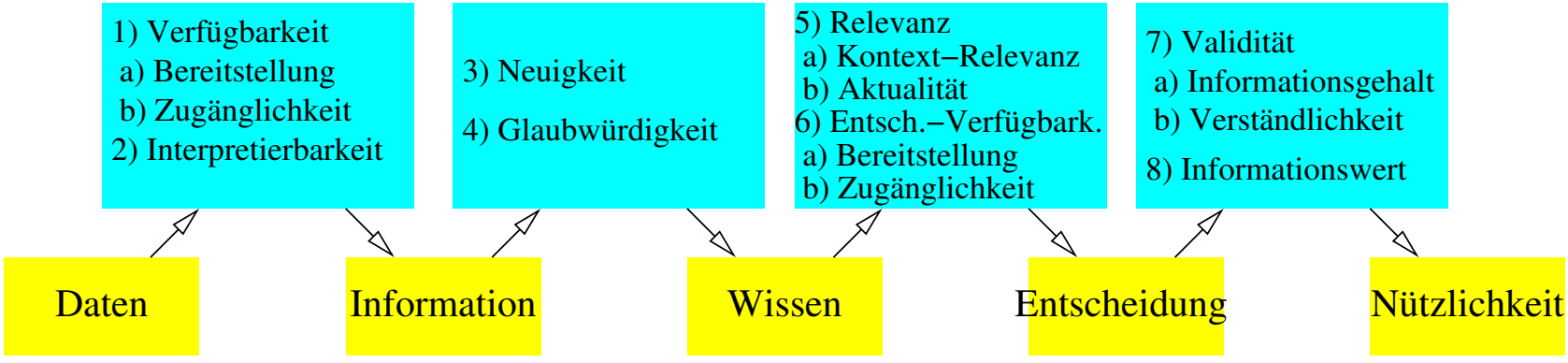
Information vs. Wissen

- Wissen ist die Teilmenge von Information, die von jemandem in einer konkreten Situation zur Lösung von Problemen benötigt wird
  (und häufig nicht vorhanden ist)
- Nach Wissen wird in externen Quellen gesucht.
- Die Transformation von Information in Wissen ist ein Mehrwert erzeugender Prozess

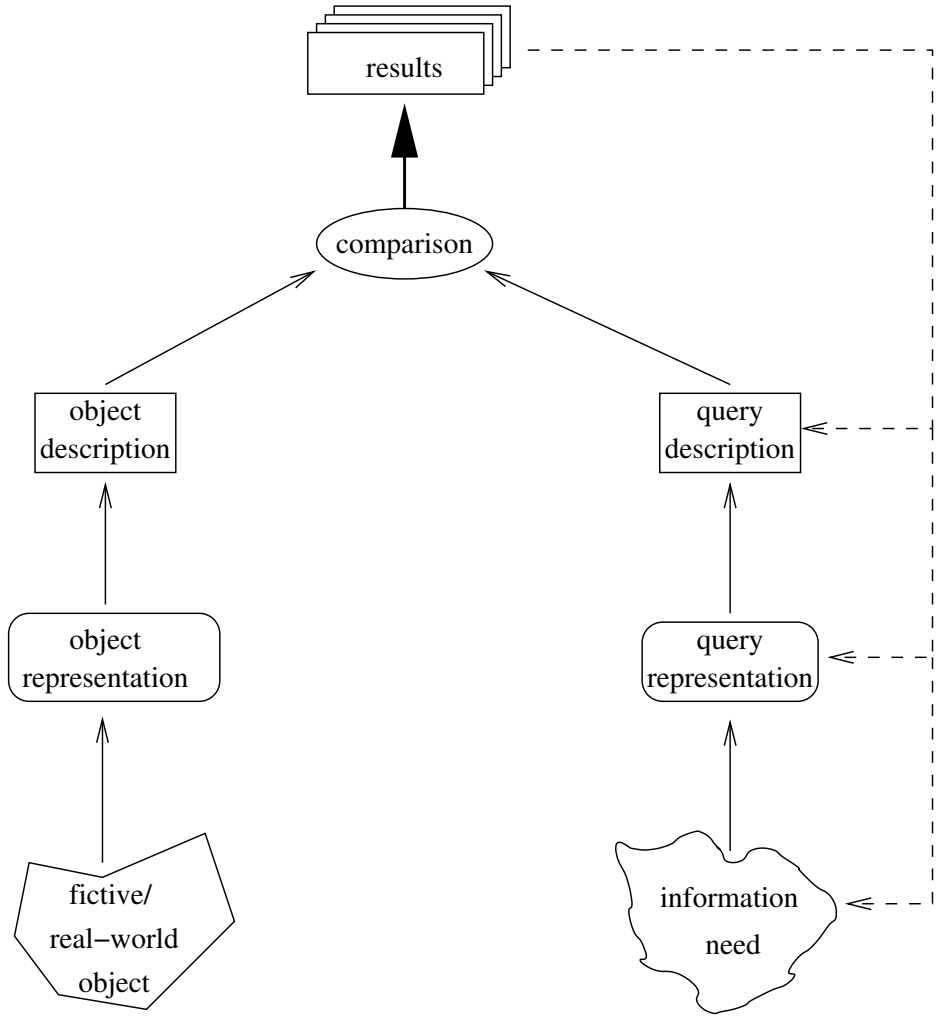
Wissen zur Entscheidungsunterstützung



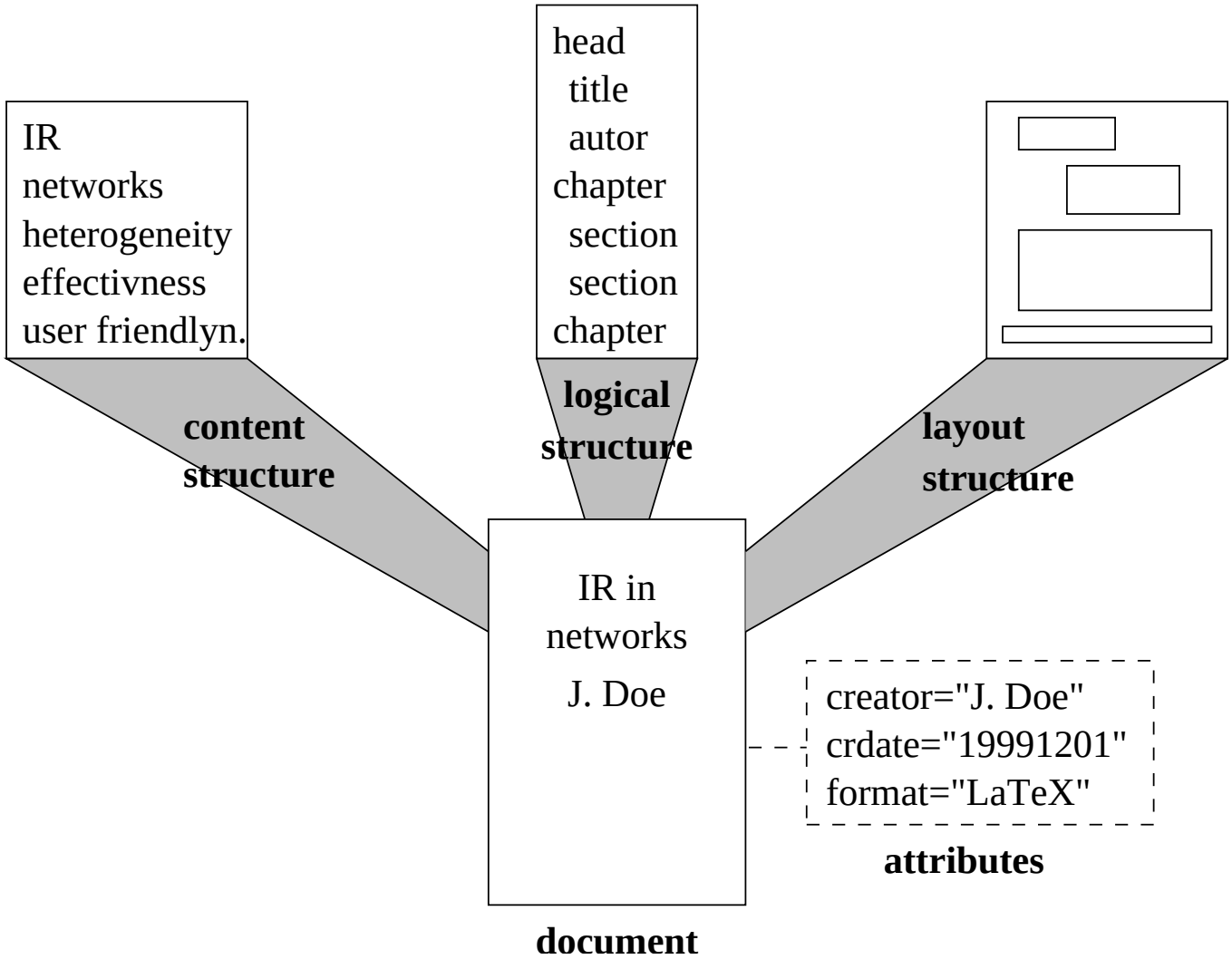
Qualitätskriterien:



Rahmenarchitektur für IR-Systeme



Sichten auf Dokumente



Anfragen und Sichten