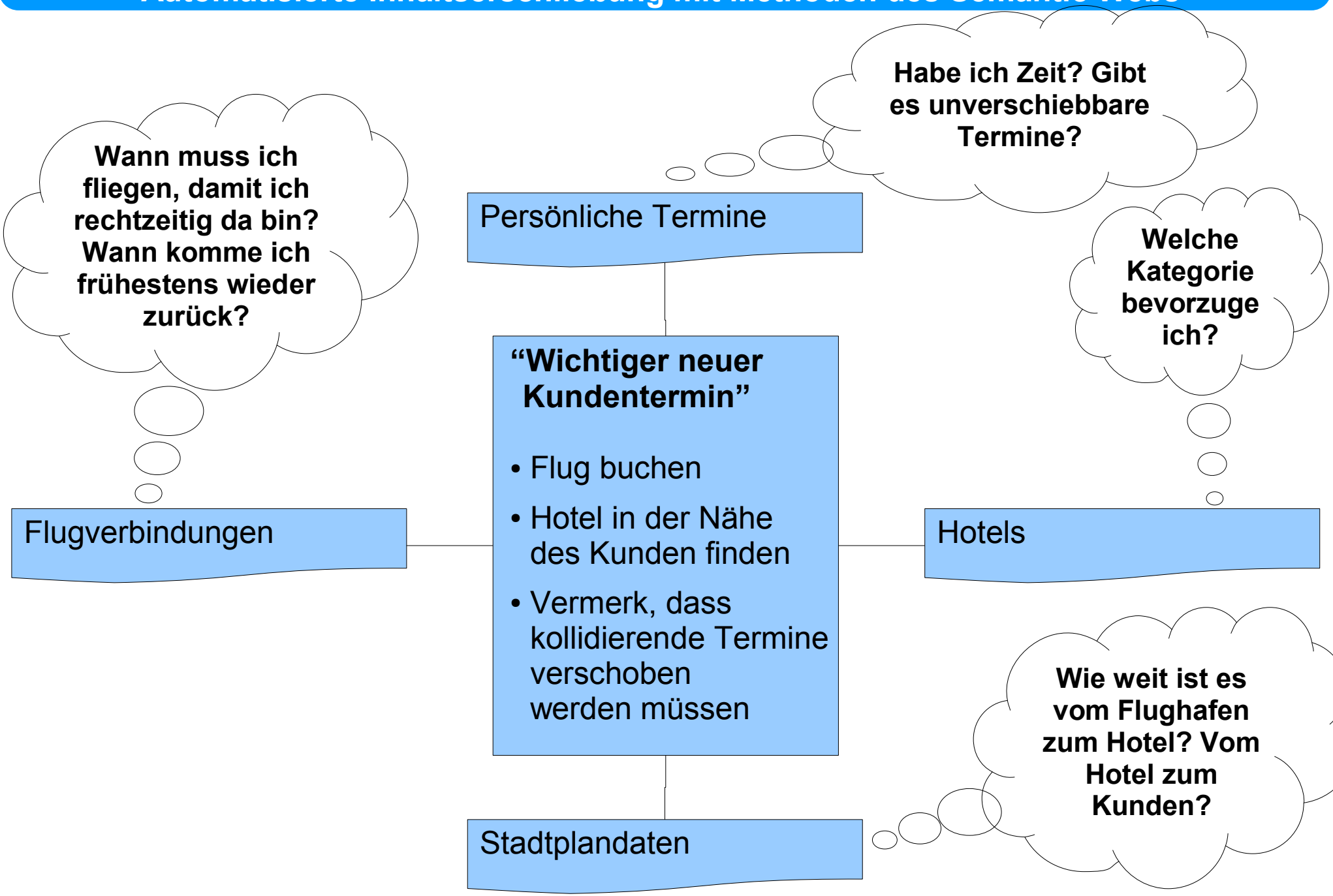
Automatisierte Inhaltserschließung mit Methoden des Semantic Webs

Kai Eckert
Institut für Informatik
Universität Mannheim

97. Deutscher Bibliotekartag
03. - 06. Juni 2008
Mannheim

Semantic Web

- Formale Beschreibung der **Bedeutung von maschinenlesbaren Daten**.
- Ziel:
  - Interpretation der Daten durch Maschinen und
  - Verknüpfung der interpretierten Daten zur Ableitung von Ergebnissen.
- Voraussetzung: Einordnung der Daten in eine Wissensbasis und Ableitung nach vordefinierten Regeln.


Habe ich Zeit? Gibt es unverschiebbare Termine?
Persönliche Termine
Wann muss ich fliegen, damit ich rechtzeitig da bin? Wann komme ich frühestens wieder zurück?
Welche Kategorie bevorzuge ich?
“Wichtiger neuer Kundentermin”
• Flug buchen
• Hotel in der Nähe des Kunden finden
• Vermerk, dass kollidierende Termine verschoben werden müssen
Flugverbindungen
Hotels
Wie weit ist es vom Flughafen zum Hotel? Vom Hotel zum Kunden?
Stadtplandaten

Wissensrepräsentation

- Ontologien
  - Formale Beschreibung relevanter Konzepte in einem Wissensgebiet: *Flug, Hotel, Termin, Zeitraum, Zeitpunkt.*
  - Zusätzlich Inferenzregeln: *Ein Termin kollidiert mit einem anderen, wenn sich die Zeiträume der Termine überschneiden.*
- Durch die gemeinsame Nutzung von Ontologien werden Daten vergleichbar und austauschbar.

Begrifflichkeiten

- **Ontologie** als Überbegriff für repräsentiertes Wissen
  - **Thesaurus** als hierarchische Begriffsstruktur
  - **Klassifikation** zur eindeutigen Einordnung von Inhalten
  - **Schlagwortkatalog** zur vereinheitlichten Inhaltsrepräsentation

Frühere Indexierungsprojekte

- Milos I und II
  - Automatische Erschließung unter Verwendung von Wörterbüchern und Thesauri. Ergänzung zur manuellen Verschlagwortung.
- CARMEN
  - Schwerpunkt heterogene Metadaten, Umgang mit Metadaten unterschiedlicher Herkunft und Qualität.

DFG Projekt

- Automatische Indexierung von Volltexten und Abstracts
- Kooperation mit Collexis, Einsatz der Collexis Suchmaschine
- Verbesserung der Fachrecherche in betriebswirtschaftlicher Literatur
- Standard Thesaurus Wirtschaft
- Einsatz weiterer Thesauri zur Abdeckung von Fachgebieten

Thesaurusbasierte Inhaltserschließung

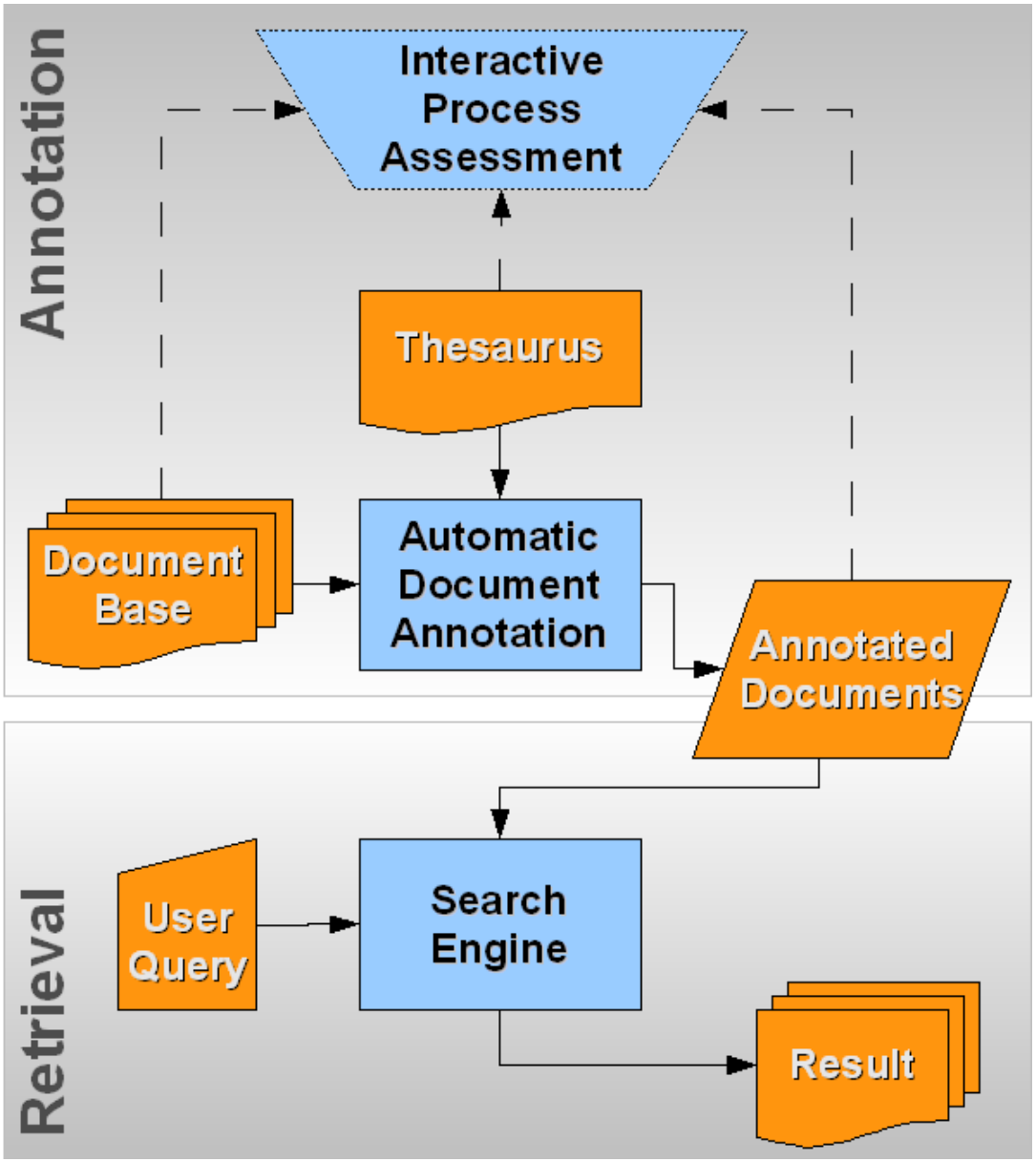
- Manuelle Erschließung
  - Traditionelle bibliographische Erschließung
  - Erschließung nach einheitlichen Kriterien, hohe Qualität
- Automatische Erschließung
  - Vorteilhaft bei großen, schnell wachsenden Dokumentenmengen:
    - News, Websites, Wissenschaftliche Veröffentlichungen (Zeitschriftenbeiträge, Konferenzbeiträge)

Probleme der automatischen Erschließung

- Thesaurus:
  - Qualität und Eignung für das Fachgebiet
- Indexierungssystem:
  - Fallstricke der automatisierten Sprachverarbeitung
  - Qualität der Vorverarbeitungsschritte (Normalisierung, Stemming)
  - Leistungsfähigkeit der Disambiguierung

Überwachung durch menschliche Experten notwendig.

Semtinel Architektur



Erforderliche Thesaurus-Revisionen

- Traditionell
  - **Erweitern und Anpassen,** um Veränderungen im Fachvokabular abzubilden.
  - **Löschen** oder **Zusammenführen** von seltenen Begriffen.
  - **Aufteilen**, **Erweitern** oder **Einschränken** von häufig genutzten Begriffen.
  - **Überprüfen der Struktur,** zur Vermeidung von Unausgewogenheiten in der Bildung von Unterklassen.
- Neu
  - Identifikation von **problematischen Begriffen** für die automatische Erschließung.

Intuitive Identifikation von problematischen Konzepten

- Sehr **hohe** Anzahl Zuordnungen:
  - Zu allgemein – sollte aufgeteilt werden
  - Nicht signifikant
  - *Fehlerhafte Zuweisungen*
- Sehr **geringe** Anzahl Zuordnungen:
  - Zu spezialisiert – sollte mit anderen Begriffen zusammengeführt werden
  - Fehlende Synonyme
  - Nicht signifikant
  - *Fehlende Zuweisungen*

Berücksichtigung der Thesaurus-Hierarchie

- **Hohe** Anzahl → **Höher** in der Hierarchie
  - Allgemeinere Begriffe
- **Niedrige** Anzahl → **Niedriger** in der Hierarchie
  - Speziellere Konzepte

IC Diff Analyse

Informationsgehalt:

- Vorgestellt von Resnik
- Basiert auf der Auftrittswahrscheinlichkeit in der Dokumentenbasis

$$IC(c) = -\log P(c)$$

Intrinsischer Informationsgehalt:

- Vorgestellt von Seco, Veale und Hayes
- Basiert auf der Anzahl der Unterbegriffe

$$IIC(c) = -\log\left(\frac{hypo(c)+1}{max}\right)$$

$$D_{IC}(c) = IC(c) - IIC(c)$$

Intuitiv: Ein Wert zwischen *-1* und *1*, der angibt, ob ein Begriff eine auffällige Häufigkeit hat bezüglich seiner Position im Thesaurus.

Semtinel Demo

DEMO

Weitere Schritte

- Bewertung duch Experten aus dem Bibliotheksbereich.
- Auswirkungen auf die Qualität der Retrieval-Ergebnisse.
- Entwicklung weiterer Analysemethoden.

Vielen Dank für Ihre Aufmerksamkeit.

Fragen oder Anregungen?

kai@informatik.uni-mannheim.de