Information Retrieval WS 2008/09

Benno Stein

Raum:	110b, B11
Tel:	3795
URL:	http://www.uni-weimar.de/medien/webis
E-Mail:	benno.stein@medien.uni-weimar.de
Sprechstunde:	nach Vereinbarung

Inhalt

Übungen

Organisation:

- Übungsaufgaben werden im Netz zur Verfügung gestellt
- bei praktischen Aufgaben: Teams (2 bis 3 Studenten/innen) möglich

Aufgaben:

- theoretisch
- praktisch

Beginn der Übungen: Ende Oktober

Ziele

- Problemfelder und Aufgaben des Information Retrieval entwicklen können
- formale Beschreibungen von Verfahren verstehen und in Algorithmen umsetzen können
- einschlägige Algorithmen sinnvoll anwenden
- sich selbst weiterbilden können

Angrenzende Gebiete

1. Statistik [Paradigmen, Modelle]
2. Algebra

3. Maschinelles Lernen [Techniken, Algorithmen]
4. Wissensverarbeitung
5. Heuristische Suche

6. Betriebliche Informationssysteme [Anwendungen]
7. Business Intelligence
8. Web-Technologie

Literatur

- Baeza-Yates, Berthier Ribeiro-Neto.
  Modern Information Retrieval
  Addison Wesley 1999, ISBN 0-201-39829-X.
- Soumen Chakrabarti.
  Mining the Web: Discovering Knowledge from Hypertext Data
  Morgan Kaufmann 2002, ISBN 1-55860-754-4.
- Reginald Ferber.
  Information Retrieval: Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web
  Dpunkt Verlag 2003, ISBN 3-89864-213-5
  information-retrieval.de/irb/ir.html
- William B. Frakes, Ricardo Baeza-Yates.
  Information Retrieval – Data Structures and Algorithms
  Prentice Hall, 1992, ISBN 0-13-463837-9
- Grossman, Frieder.
  Information Retrieval: Algorithms and Heuristics
  Springer 2004, 2nd Edition, ISBN 1402030045.

Literatur (Fortsetzung)

- Norbert Fuhr.
  Scriptum zur Vorlesung Information Retrieval
  www.is.informatik.uni-duisburg.de/teaching/lectures/ir_ss06/
- Witten, Moffat, Bell.
  Managing Gigabytes: Compressing and Indexing Documents and Images
  Morgan Kaufmann 1999, 2nd Edition, ISBN 1-55860-570-3.

Weitere Literatur, die direkt aus dem World Wide Web geladen werden kann, ist in den entsprechenden Kapiteln angegeben.

Kapitel IR: I

I. Einführung

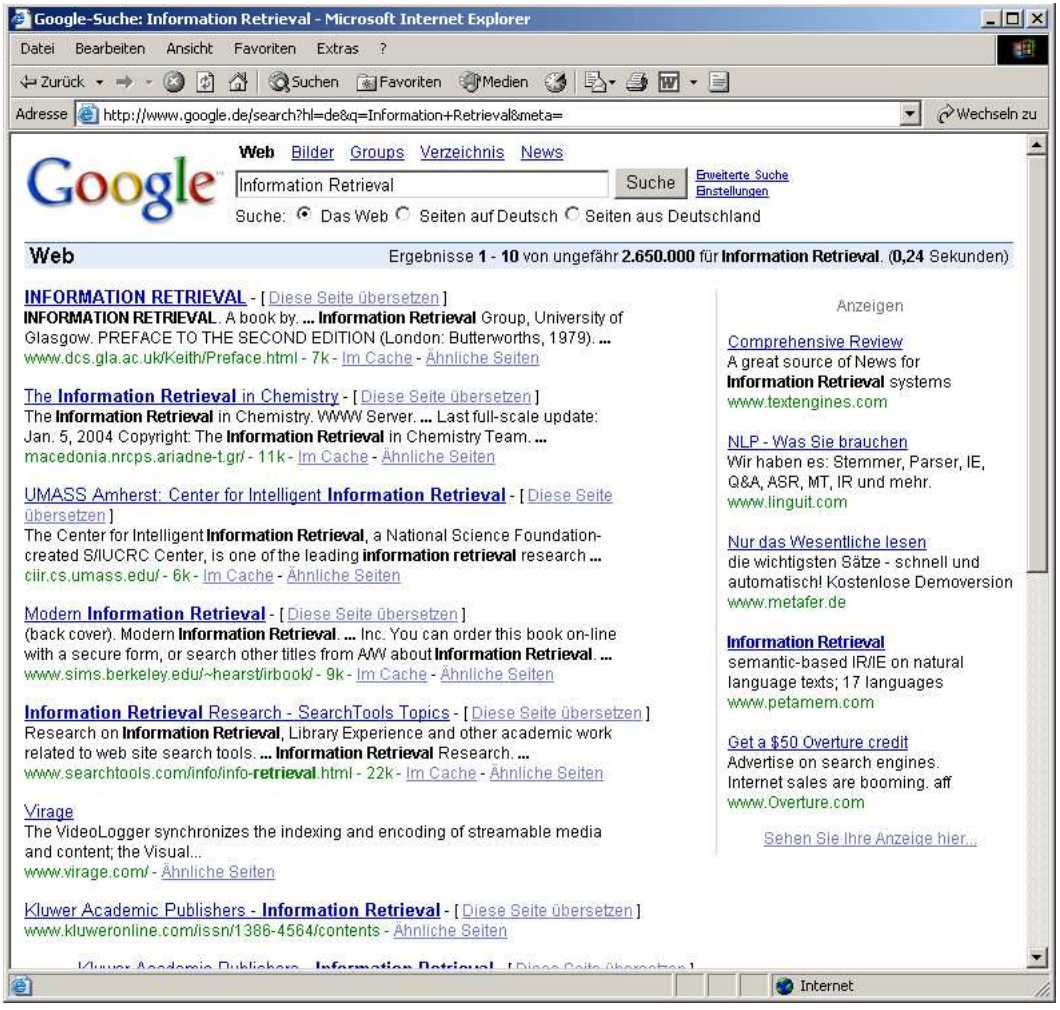
- Retrieval-Szenarien
- Begriffsbildung
- Einordnung Information Retrieval

Retrieval-Szenarien

„Liefere Dokumente, die die Terme «Information» und «Retrieval» enthalten.“

Retrieval-Szenarien

„Liefere Dokumente, die die Terme «Information» und «Retrieval» enthalten.“

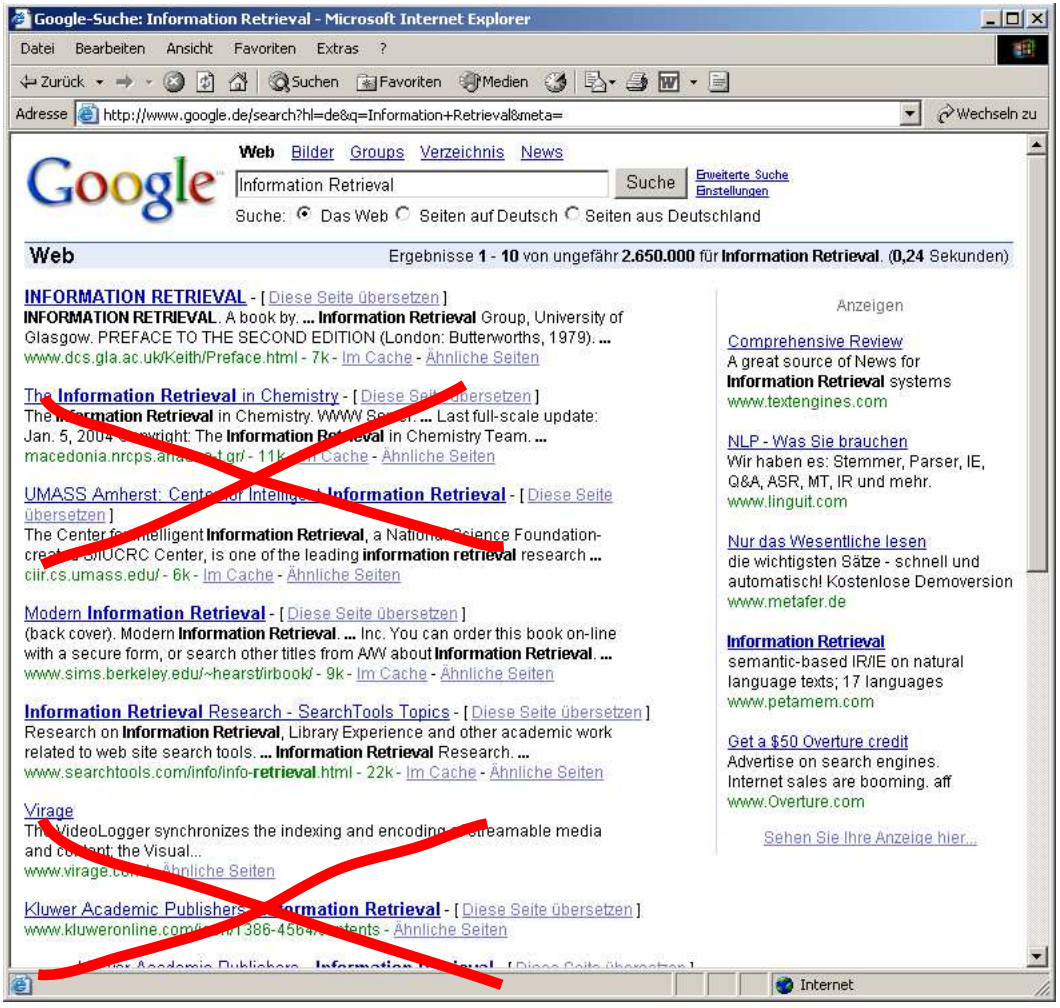


Retrieval-Szenarien

„Liefere Dokumente, die die Terme «Information» und «Retrieval» wissenschaftlich beschreiben.“

Retrieval-Szenarien

„Liefere Dokumente, die die Terme «Information» und «Retrieval» wissenschaftlich beschreiben.“

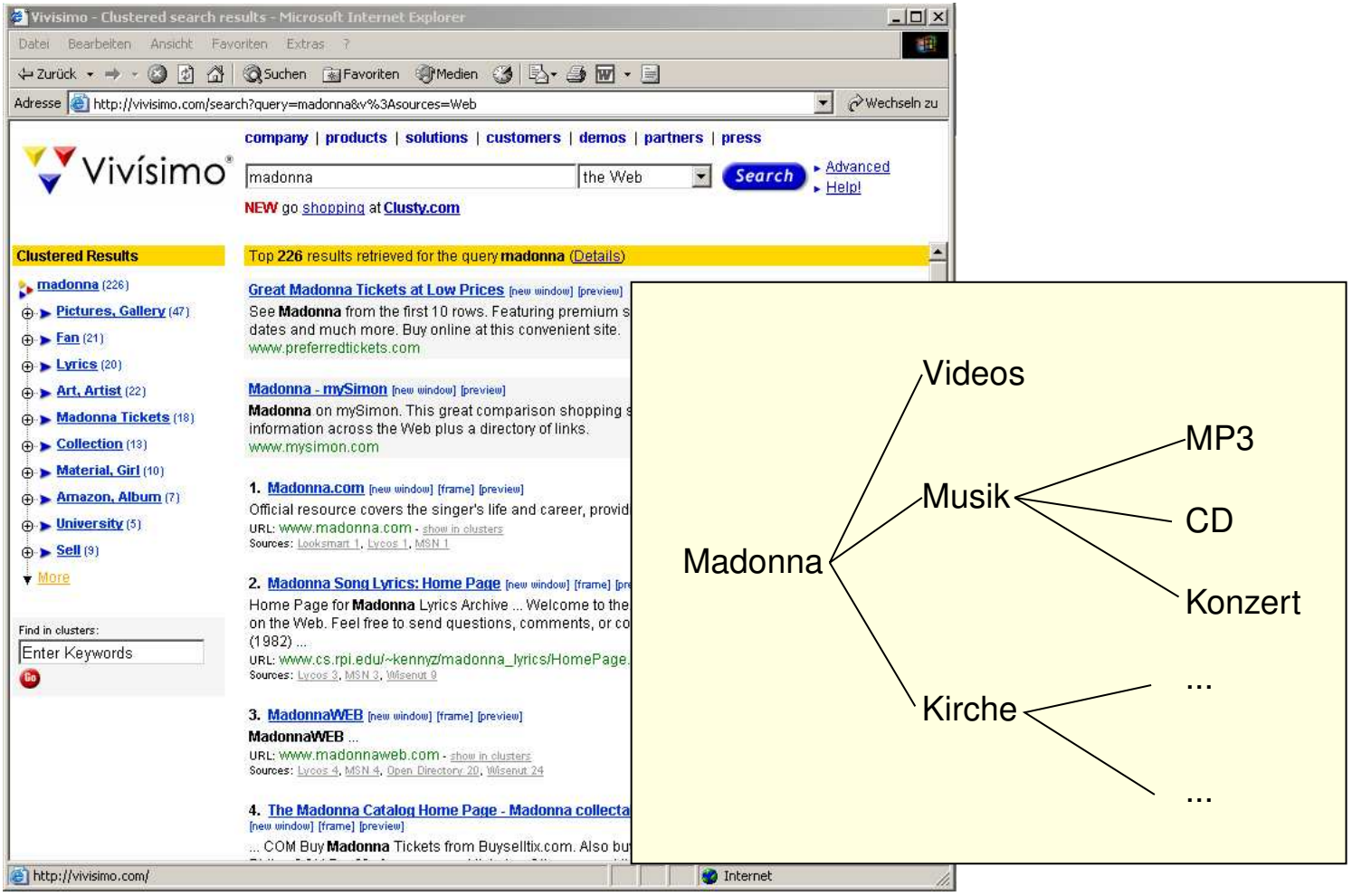


Retrieval-Szenarien

„Erstelle von denjenigen Dokumenten, die den Term «Madonna» enthalten, eine Taxonomie.“

Retrieval-Szenarien

„Erstelle von denjenigen Dokumenten, die den Term «Madonna» enthalten, eine Taxonomie.“

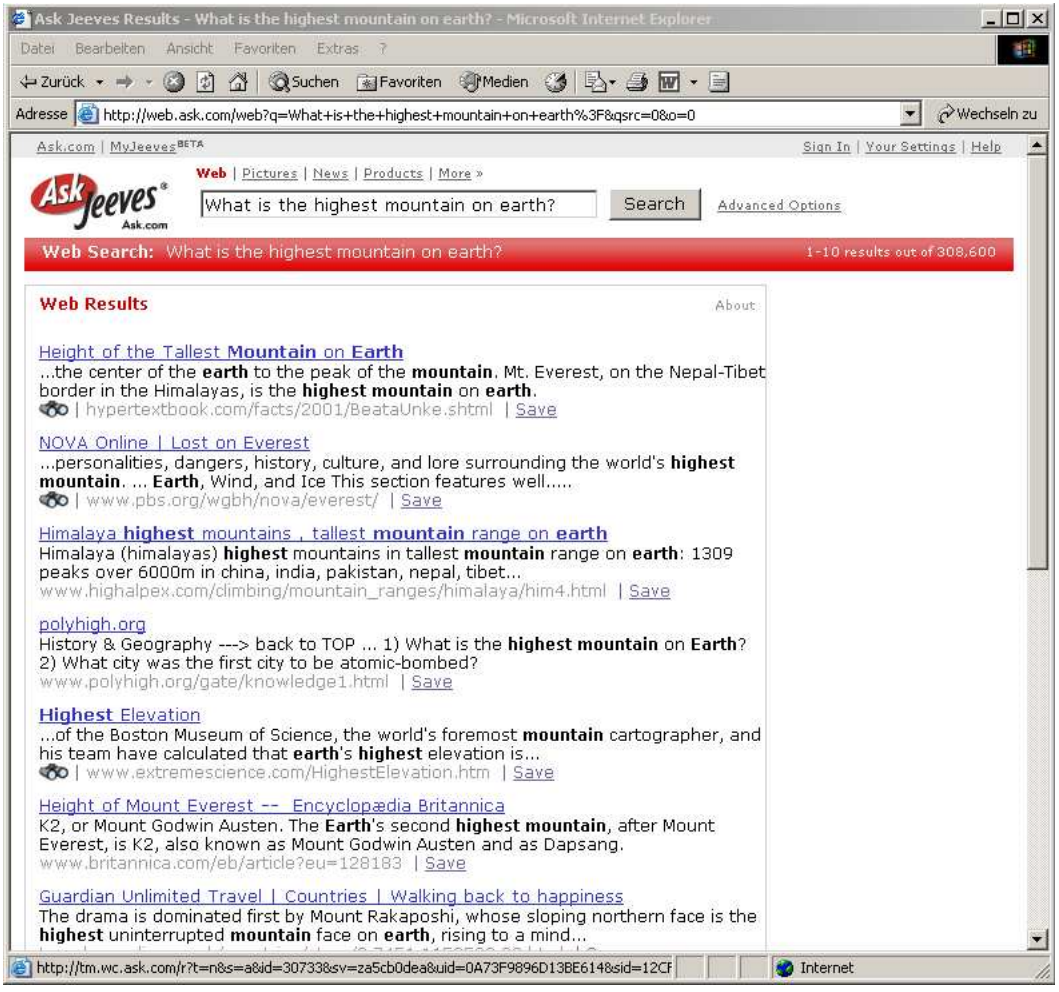


Retrieval-Szenarien

„Beantworte die gestellte Frage.“

Retrieval-Szenarien

„Beantworte die gestellte Frage.“



Retrieval-Szenarien

„Erstelle eine Ausarbeitung zu dem genannten Thema.“

Thema: Die Rolle von MacDonald's in der europäischen Esskultur.

Bemerkungen:

- Die illustrierten Szenarien sind auch auf anderen Datenquellen als dem World Wide Web denkbar:
  - Archive von Zeitungsredaktionen
  - Patientendaten von Krankenversicherungen
  - FBI
  - Bibliotheken
  - Datenbestände großer Firmen – Stichwort: OLAP
- Die illustrierten Szenarien enthalten vielfältige und unterschiedlich komplexe Herausforderungen:
  - Speicherung und effizienter Zugriff auf riesige Datenmengen
  - effiziente Suche
  - komplexe Suchanfragen *(Queries)*
  - Bewertung und Vergleich von Anfragen und Suchergebnissen
  - (visuelle) Aufbereitung von Suchergebnissen, Navigation, Benutzerführung
  - einfaches Textverstehen
  - automatische Textsynthese
- Unsere Suchmaschine `AIsearch` leistet eine thematische Sortierung.

Begriffsbildung

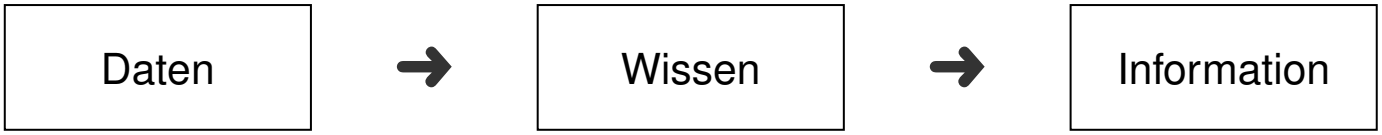
Suche in Dokumentkollektionen kann auf verschiedenen Abstraktionsstufen stattfinden. Vergleiche hierzu die Ebenen der Semiotik:

- Syntax
  Ein Dokument wird als Folge von Symbolen betrachtet. Beispiele:
  Zeichenkette in Texten, Histogramm oder Kontur in Bildern

- Semantik
  Ein Dokument wird auf der Ebene seiner Bedeutung betrachtet. Semantik hat immer etwas mit Interpretation zu tun.

- Pragmatik
  Ein Dokument wird hinsichtlich seines Verwendungszusammenhangs betrachtet. Beispiele:
  Enthält ein Dokument eine Lösung meines Problems?
  Was ist die Absicht des Autors des Textes?

Begriffsbildung



syntaktische Ebene, sigmatische Ebene	semantische Ebene	pragmatische Ebene
Beispiel: Datenbank als Sammlung von Werten	Beispiel: Interpretation der Werte in einer Datenbank	Transformation von Wissen, um die benötigte Information zu erhalten.

Definition 1 (Information [Kuhlen 90]**)**
Information ist die Teilmenge von Wissen, die von jemandem in einer konkreten Situation zur Lösung seines Problems benötigt wird.

Bemerkungen:

- Semiotik (griechisch: Zeichentheorie) ist die Lehre von den sprachlichen und nichtsprachlichen Zeichen und ihrer Verwendung. Die moderne Semiotik wurde insbesondere durch C. S. Peirce und C. W. Morris begündet; Gliederung in drei Bereiche: die Beziehung zwischen zwischen den Zeichen selbst (Syntaktik), zwischen dem Zeichen und dem Bezeichneten (Semantik), sowie zwischen dem Zeichen und seinem Verwender (Pragmatik).
- In der Semiotik kann weiterhin noch eine sigmatische Ebene unterschieden werden:

Ebene	Element
Syntax	Zeichen
Sigmatik	Daten
Semantik	Nachricht, Wissen
Pragmatik	Information

Begriffsbildung

Definition 2 (Information Retrieval [GI-Fachgruppe]**)**
Im Information Retrieval (IR) werden Informationssysteme in Bezug auf ihre Rolle im Prozess des Wissenstransfers vom menschlichen Wissensproduzenten zum Informationsnachfragenden betrachtet.

Das heißt, Information Retrieval ist eine inhaltsorientierte Suche und beschäftigt sich insbesondere mit der Semantik und Pragmatik von Dokumenten.

Besondere Retrieval-Herausforderungen:

1. vage Anfragen
2. unsicheres Wissen
3. Genauigkeit der Antwort
4. Effizienz

Bemerkungen [vgl. Wanner 2003]:

zu 1) Vage Anfragen sind dadurch gekennzeichnet, dass die Antwort a-Priori nicht eindeutig definiert ist. Hierzu zählen neben Fragen mit unscharfen Kriterien insbesondere solche, die nur im Dialog, interaktiv durch Reformulierung und in Abhängigkeit von den bisherigen Antworten beantwortet werden können. Häufig müssen mehrere Datenbasen zur Beantwortung einer einzelnen Anfrage durchsucht werden.

zu 2) Die Darstellungsformen des in einem IR-System gespeicherten Wissens kann vielfältig sein: Texte, multimediale Dokumente, Falldatenbanken, Regeln, semantische Netze, etc. Die Unsicherheit – aber auch die Unvollständigkeit – dieses Wissens resultiert oft aus der begrenzten Repräsentation von dessen Semantik. Weiterhin werden auch solche Anwendungen betrachtet, bei denen die gespeicherten Daten selbst (von der Natur der Sache her) unsicher oder unvollständig sind.

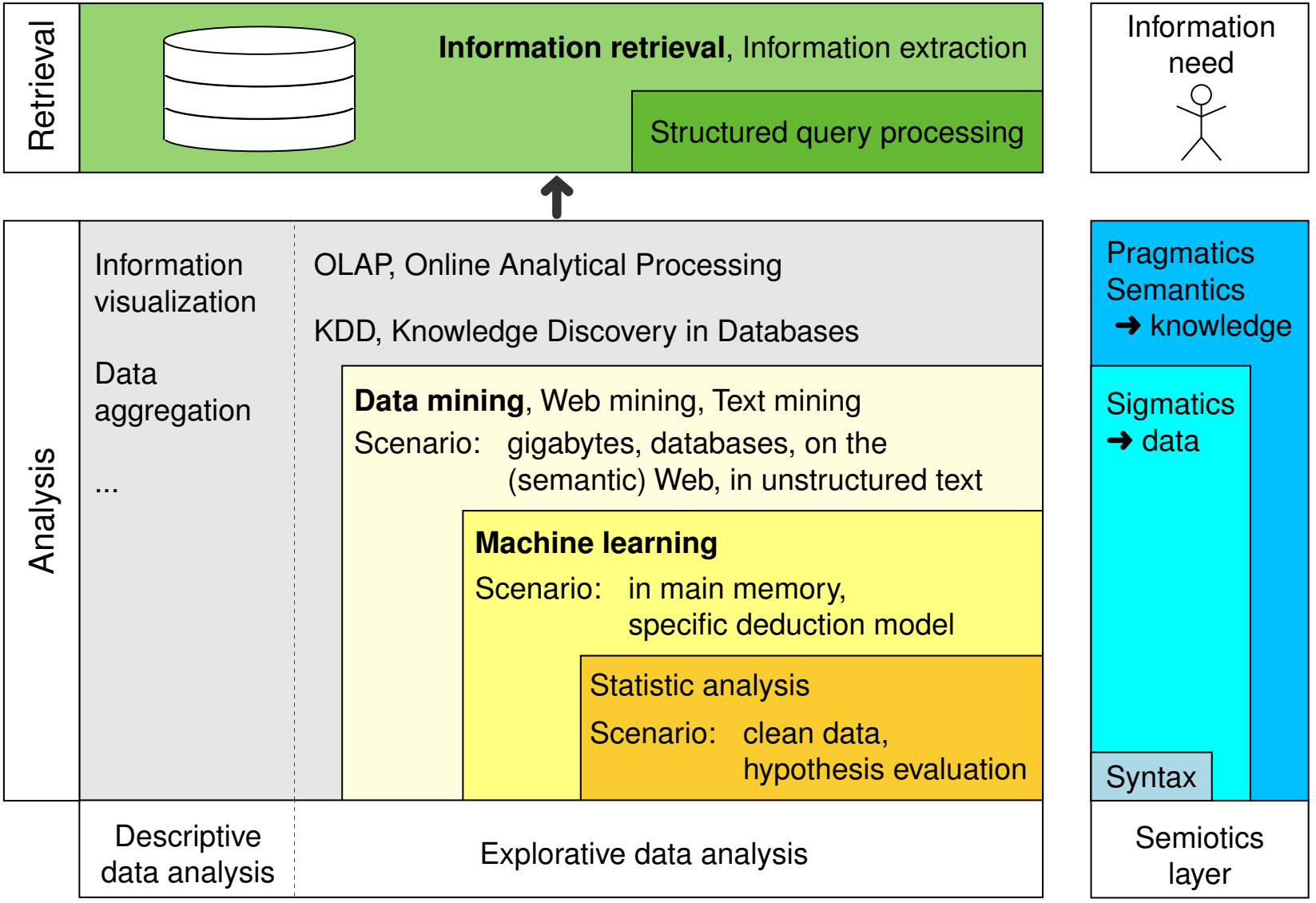
Einordnung Information Retrieval

Daten-Retrieval versus Text-IR

	Daten-Retrieval	Text-IR
Matching	exakt	partieller Match, bester Match
Inferenz	Deduktion	Induktion
Modell	deterministisch	probabilistisch
Klassifikation	monothetisch	polithetisch
Anfragesprache	formal	natürlich
Fragespezifikation	vollständig	unvollständig
gesuchte Objekte	Fragespezifikation erfüllend	relevante
Reaktion auf Datenfehler	empfindlich	robust

[vgl. Rijsbergen 1979, Fuhr 2004]

Einordnung Information Retrieval

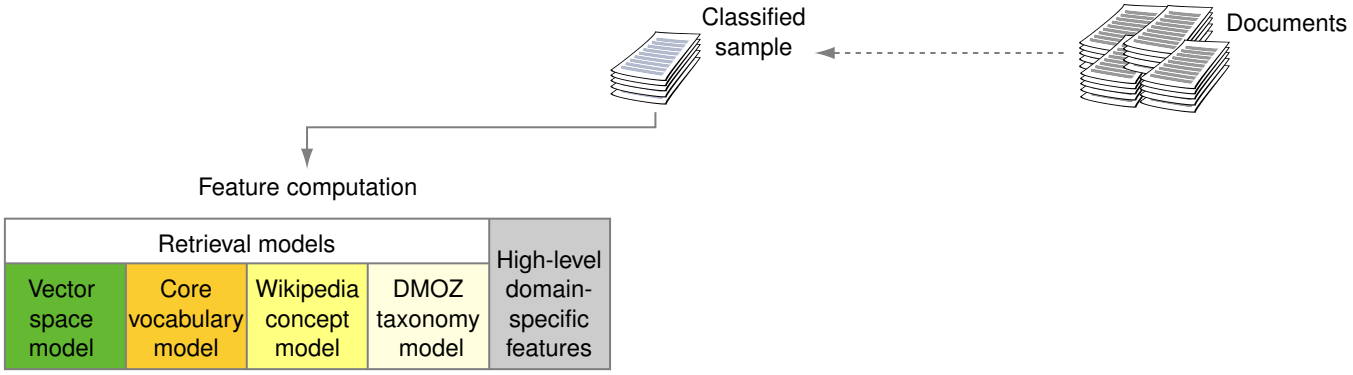


Einordnung Information Retrieval

Anwendung: Personenkategorisierung im Web (Spock-Challenge)

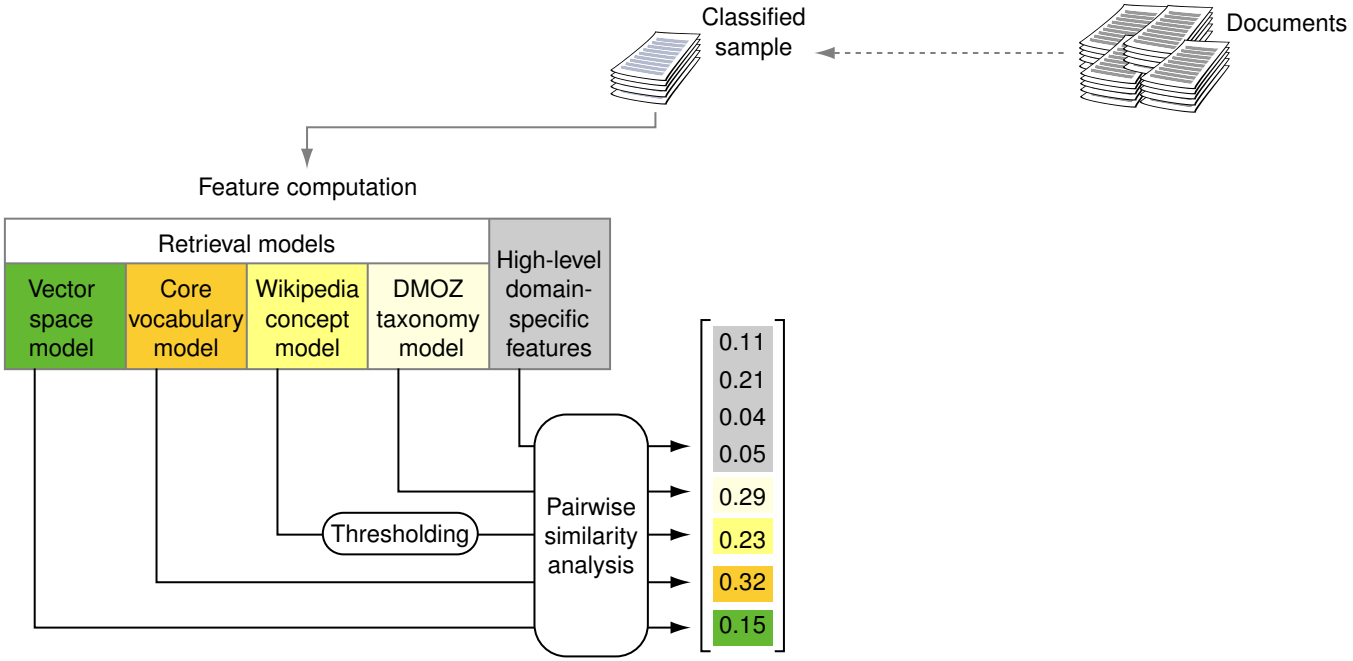
Einordnung Information Retrieval

Anwendung: Personenkategorisierung im Web (Spock-Challenge)



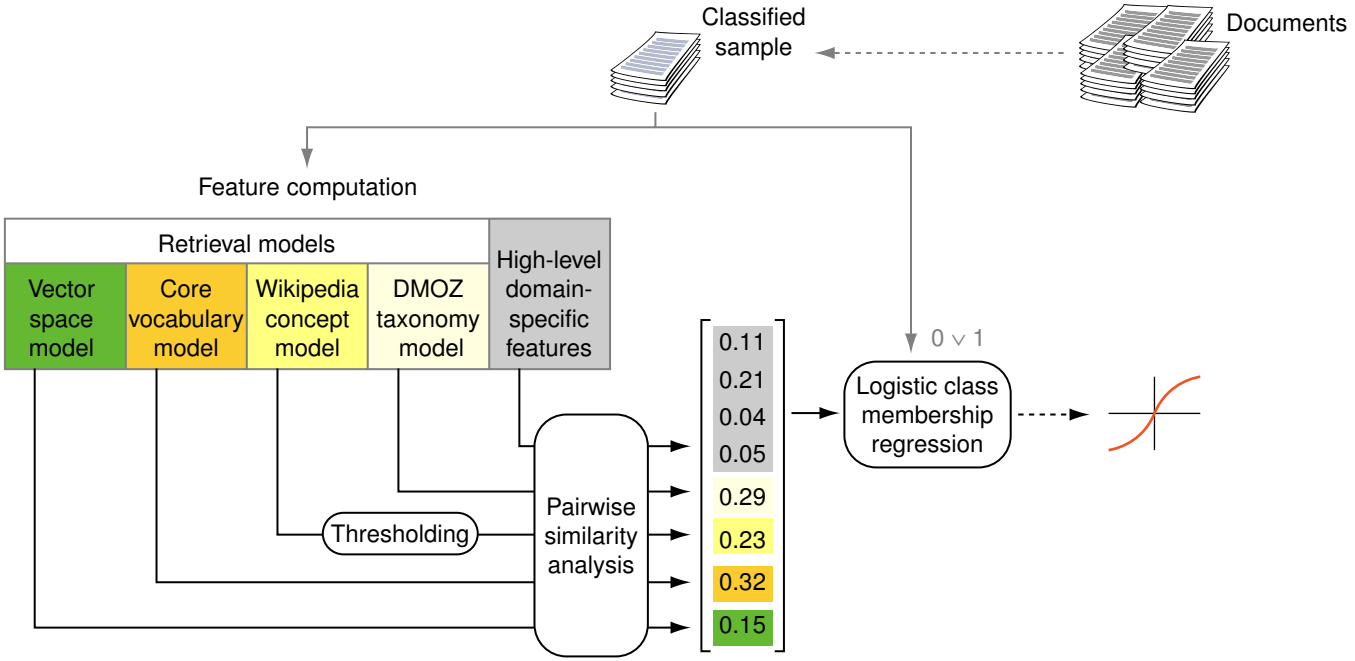
Einordnung Information Retrieval

Anwendung: Personenkategorisierung im Web (Spock-Challenge)



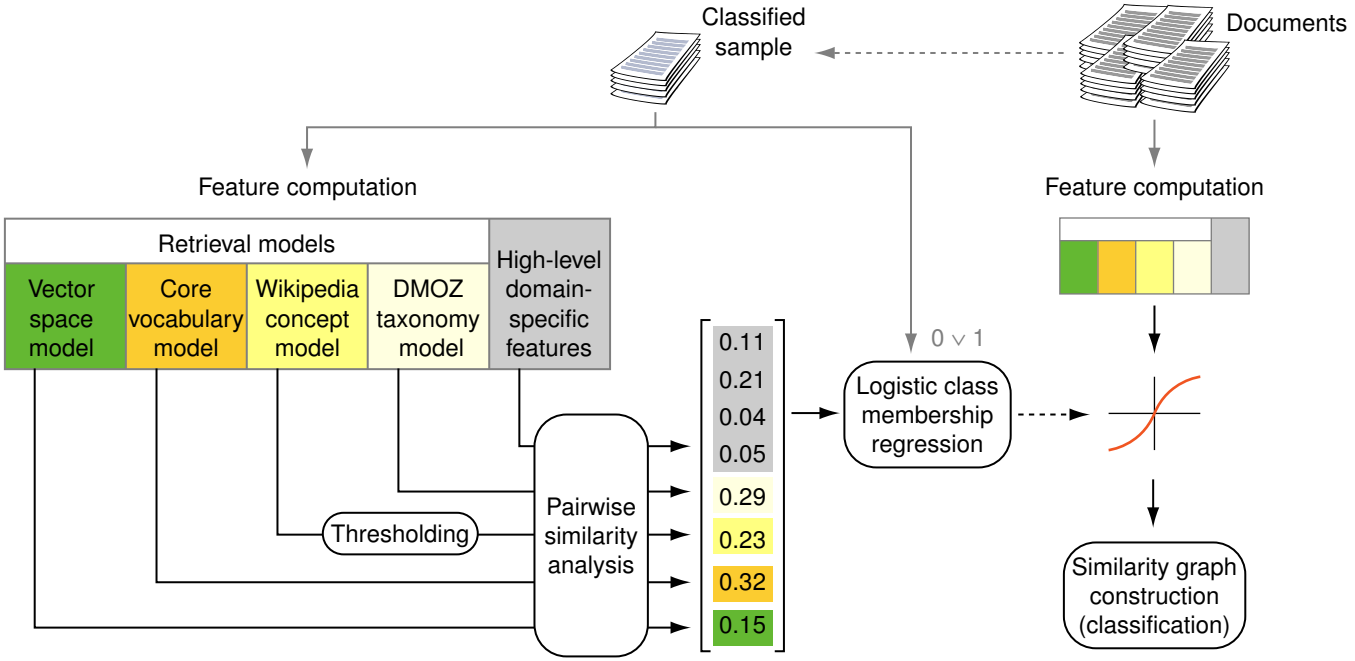
Einordnung Information Retrieval

Anwendung: Personenkategorisierung im Web (Spock-Challenge)



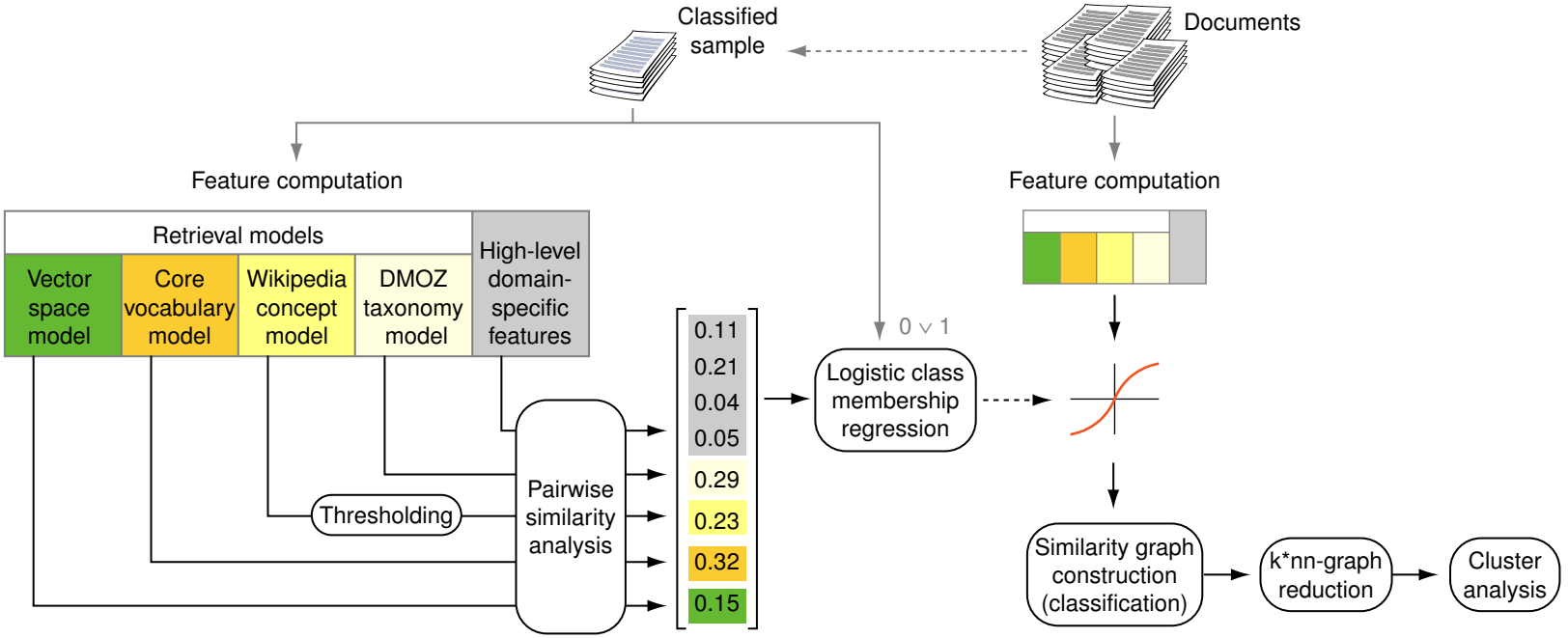
Einordnung Information Retrieval

Anwendung: Personenkategorisierung im Web (Spock-Challenge)



Einordnung Information Retrieval

Anwendung: Personenkategorisierung im Web (Spock-Challenge)



Einordnung Information Retrieval

Methoden und Techniken

- Modellierung von Dokumenten und Text
- (approximatives) String-Matching
- Textvorverarbeitung und Indexing
- Benutzerinteraktion und Visualisierung
- Benutzermodellierung und Personalisierung
- Relevanzanalyse
- verteilte und Peer-to-Peer Softwaretechnik
- Kategorisierung, Klassifikation
- Natural Lange Processing (NLP)
- Web-Technologie
- Datenstrukturen, effiziente Symbolverarbeitung