

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

- I. Einführung
- II. Grundlagen des Information Retrieval
- III. Retrieval-Modelle
- IV. Indexkonstruktion
- V. Anfragekonstruktion
- VI. Indexing und Suche
- VII. Web-spezifische Aspekte
- VIII. Schlüsselworte, Labeling, Summarization
- IX. Multilinguales Retrieval
- X. Multimedia-Retrieval

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 entwickeln 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.“

The screenshot shows a Microsoft Internet Explorer browser window displaying a Google search for "Information Retrieval". The address bar shows the search URL: <http://www.google.de/search?hl=de&q=Information+Retrieval&meta=>. The search results are displayed in German, showing the first 10 results out of approximately 2,650,000. The results include:

- INFORMATION RETRIEVAL** - [[Diese Seite übersetzen](#)]
INFORMATION RETRIEVAL. A book by ... **Information Retrieval** Group, University of Glasgow. PREFACE TO THE SECOND EDITION (London: Butterworths, 1979). ...
www.dcs.gla.ac.uk/Keith/Preface.html - 7k - [Im Cache](#) - [Ähnliche Seiten](#)
- The Information Retrieval in Chemistry** - [[Diese Seite übersetzen](#)]
The **Information Retrieval** in Chemistry. WWW Server. ... Last full-scale update: Jan. 5, 2004 Copyright: The **Information Retrieval** in Chemistry Team. ...
macedonia.nrcps.ariadne.tgr/ - 11k - [Im Cache](#) - [Ähnliche Seiten](#)
- UMASS Amherst: Center for Intelligent Information Retrieval** - [[Diese Seite übersetzen](#)]
The Center for Intelligent **Information Retrieval**, a National Science Foundation-created S/UCRC Center, is one of the leading **information retrieval** research ...
ciir.cs.umass.edu/ - 6k - [Im Cache](#) - [Ähnliche Seiten](#)
- Modern Information Retrieval** - [[Diese Seite übersetzen](#)]
(back cover). Modern **Information Retrieval**. ... Inc. You can order this book on-line with a secure form, or search other titles from **AWW** about **Information Retrieval**. ...
www.sims.berkeley.edu/~hears/tirbook/ - 9k - [Im Cache](#) - [Ähnliche Seiten](#)
- Information Retrieval Research - SearchTools Topics** - [[Diese Seite übersetzen](#)]
Research on **Information Retrieval**, Library Experience and other academic work related to web site search tools. ... **Information Retrieval** Research. ...
www.searchtools.com/info/info-retrieval.html - 22k - [Im Cache](#) - [Ähnliche Seiten](#)
- Virage**
The VideoLogger synchronizes the indexing and encoding of streamable media and content; the Visual...
www.virage.com/ - [Ähnliche Seiten](#)
- Kluwer Academic Publishers - Information Retrieval** - [[Diese Seite übersetzen](#)]
www.kluweronline.com/issn/1386-4564/contents - [Ähnliche Seiten](#)

On the right side of the page, there is an "Anzeigen" (Ads) section with several advertisements:

- Comprehensive Review**
A great source of News for **Information Retrieval** systems
www.textengines.com
- NLP - Was Sie brauchen**
Wir haben es: Stemmer, Parser, IE, Q&A, ASR, MT, IR und mehr.
www.linguit.com
- Nur das Wesentliche lesen**
die wichtigsten Sätze - schnell und automatisch! Kostenlose Demoverision
www.metafer.de
- Information Retrieval**
semantic-based IR/IE on natural language texts; 17 languages
www.petamem.com
- Get a \$50 Overture credit**
Advertise on search engines. Internet sales are booming. aff
www.Overture.com

At the bottom of the page, there is a link: [Sehen Sie Ihre Anzeige hier...](#)

Retrieval-Szenarien

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

Retrieval-Szenarien

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

Google-Suche: Information Retrieval - Microsoft Internet Explorer

Adresse <http://www.google.de/search?hl=de&q=Information+Retrieval&meta=>

Web Bilder Groups Verzeichnis News

Information Retrieval Suche Erweiterte Suche Einstellungen

Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland

Web Ergebnisse 1 - 10 von ungefähr 2.650.000 für Information Retrieval. (0,24 Sekunden)

INFORMATION RETRIEVAL - [[Diese Seite übersetzen](#)]
INFORMATION RETRIEVAL. A book by ... **Information Retrieval** Group, University of Glasgow. PREFACE TO THE SECOND EDITION (London: Butterworths, 1979). ...
www.dcs.gla.ac.uk/Keith/Preface.html - 7k - [Im Cache](#) - [Ähnliche Seiten](#)

~~The **Information Retrieval** in Chemistry~~ - [[Diese Seite übersetzen](#)]
~~The **Information Retrieval** in Chemistry. WWW Search ... Last full-scale update: Jan. 5, 2004 ... Copyright: The **Information Retrieval** in Chemistry Team. ...~~
~~macedonia.nrps.ana.edu/~tgr/ - 11k - [Im Cache](#) - [Ähnliche Seiten](#)~~

~~UMASS Amherst: Center for Intelligent **Information Retrieval**~~ - [[Diese Seite übersetzen](#)]
~~The Center for Intelligent **Information Retrieval**, a National Science Foundation-created JUCRC Center, is one of the leading **information retrieval** research ...~~
~~citr.cs.umass.edu/ - 6k - [Im Cache](#) - [Ähnliche Seiten](#)~~

~~Modern **Information Retrieval**~~ - [[Diese Seite übersetzen](#)]
~~(back cover). Modern **Information Retrieval**. ... Inc. You can order this book on-line with a secure form, or search other titles from AWW about **Information Retrieval**. ...~~
~~www.sims.berkeley.edu/~hears/irbook/ - 9k - [Im Cache](#) - [Ähnliche Seiten](#)~~

~~**Information Retrieval** Research - SearchTools Topics~~ - [[Diese Seite übersetzen](#)]
~~Research on **Information Retrieval**, Library Experience and other academic work related to web site search tools. ... **Information Retrieval** Research. ...~~
~~www.searchtools.com/info/info-retrieval.html - 22k - [Im Cache](#) - [Ähnliche Seiten](#)~~

~~Virage~~
~~The VideoLogger synchronizes the indexing and encoding of streamable media and content; the Visual...~~
~~www.virage.com/ - [Ähnliche Seiten](#)~~

~~Kluwer Academic Publishers **Information Retrieval**~~ - [[Diese Seite übersetzen](#)]
~~www.kluweronline.com/01386-4564/ - [Ähnliche Seiten](#)~~

Anzeigen

[Comprehensive Review](#)
A great source of News for **Information Retrieval** systems
www.textengines.com

[NLP - Was Sie brauchen](#)
Wir haben es: Stemmer, Parser, IE, Q&A, ASR, MT, IR und mehr.
www.linguit.com

[Nur das Wesentliche lesen](#)
die wichtigsten Sätze - schnell und automatisch! Kostenlose Demoversion
www.metafer.de

[Information Retrieval](#)
semantic-based IR/IE on natural language texts; 17 languages
www.petamem.com

[Get a \\$50 Overture credit](#)
Advertise on search engines. Internet sales are booming. aff
www.Overture.com

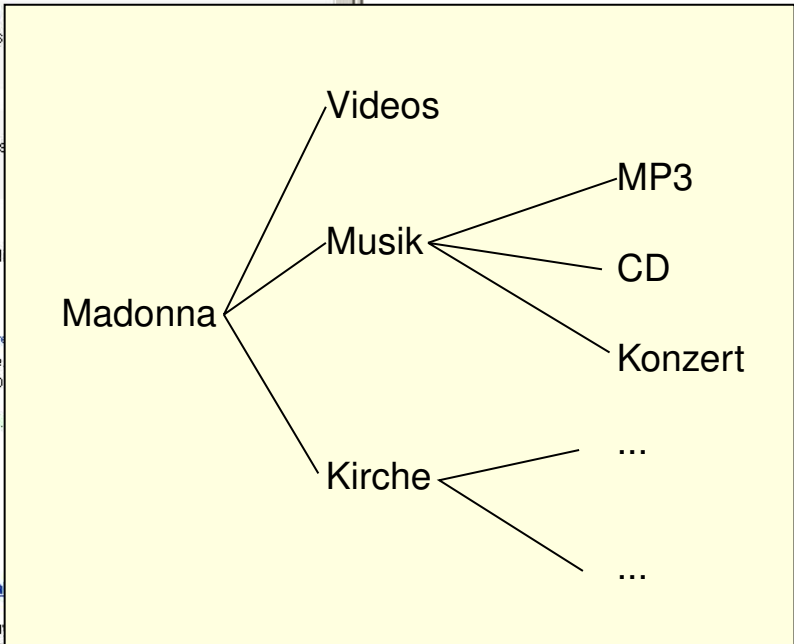
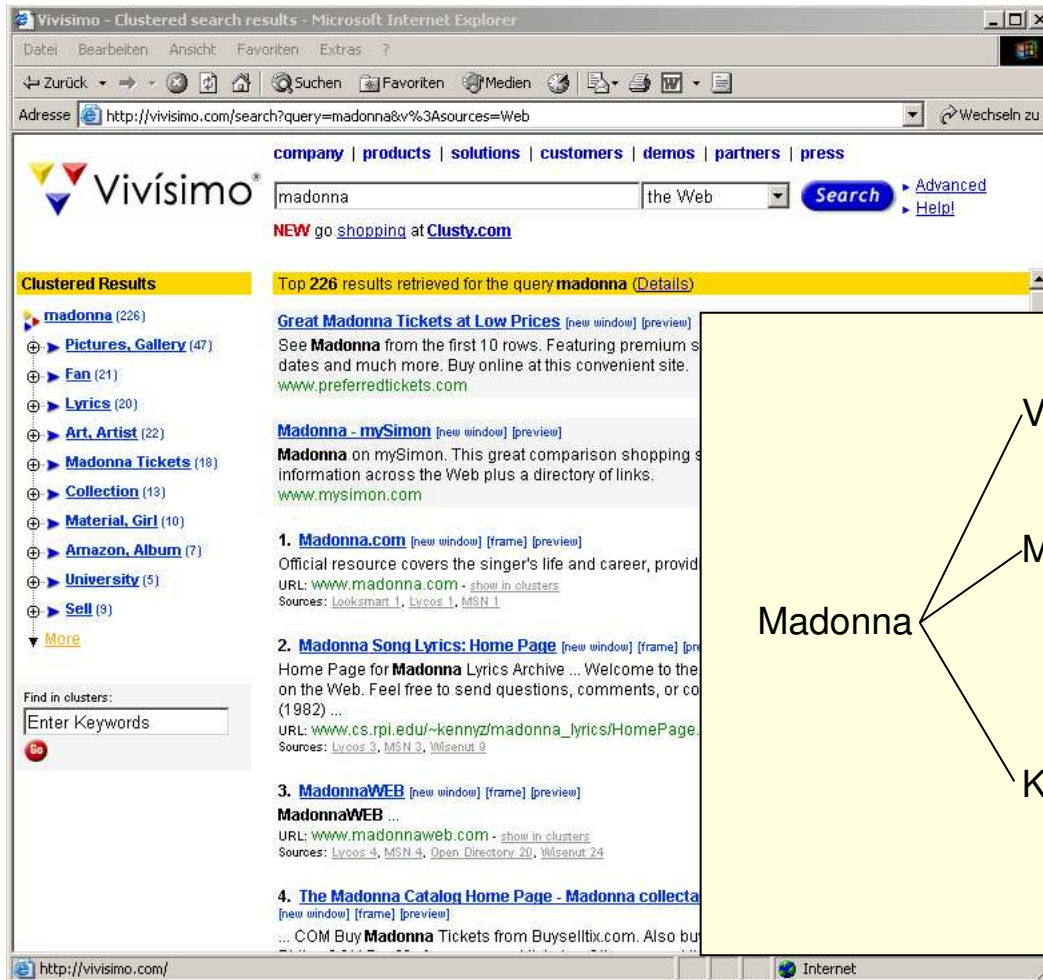
[Sehen Sie Ihre Anzeige hier...](#)

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.“



[Web](#) | [Pictures](#) | [News](#) | [Products](#) | [More »](#)

What is the highest mountain on earth?

Search

[Advanced Options](#)

Retrieval-Szenarien

„Beantworte die gestellte Frage.“

The screenshot shows a Microsoft Internet Explorer browser window with the title "Ask Jeeves Results - What is the highest mountain on earth? - Microsoft Internet Explorer". The address bar contains the URL "http://web.ask.com/web?q=What+is+the+highest+mountain+on+earth%3F&qsrc=0&o=0". The search results page features the Ask Jeeves logo and a search bar with the query "What is the highest mountain on earth?". Below the search bar, a red banner indicates "Web Search: What is the highest mountain on earth? 1-10 results out of 308,600". The "Web Results" section lists several search results:

- Height of the Tallest Mountain on Earth**: ...the center of the **earth** to the peak of the **mountain**. Mt. Everest, on the Nepal-Tibet border in the Himalayas, is the **highest mountain on earth**.
hypertextbook.com/facts/2001/BeataUnke.shtml | [Save](#)
- NOVA Online | Lost on Everest**: ...personalities, dangers, history, culture, and lore surrounding the world's **highest mountain**. ... **Earth**, Wind, and Ice This section features well....
www.pbs.org/wgbh/nova/everest/ | [Save](#)
- Himalaya highest mountains , tallest mountain range on earth**: Himalaya (himalayas) **highest** mountains in tallest **mountain** range on **earth**: 1309 peaks over 6000m in china, india, pakistan, nepal, tibet...
www.highalpe.com/climbing/mountain_ranges/himalaya/him4.html | [Save](#)
- polyhigh.org**: History & Geography ---> back to TOP ... 1) What is the **highest mountain** on **Earth**? 2) What city was the first city to be atomic-bombed?
www.polyhigh.org/gate/knowledge1.html | [Save](#)
- Highest Elevation**: ...of the Boston Museum of Science, the world's foremost **mountain** cartographer, and his team have calculated that **earth's highest** elevation is...
www.extremescience.com/HighestElevation.htm | [Save](#)
- Height of Mount Everest -- Encyclopædia Britannica**: K2, or Mount Godwin Austen. The **Earth's** second **highest mountain**, after Mount Everest, is K2, also known as Mount Godwin Austen and as Dapsang.
www.britannica.com/eb/article?eu=128183 | [Save](#)
- Guardian Unlimited Travel | Countries | Walking back to happiness**: The drama is dominated first by Mount Rakaposhi, whose sloping northern face is the **highest** uninterrupted **mountain** face on **earth**, rising to a mind...

The browser's status bar at the bottom shows the URL "http://tm.wc.ask.com/r?t=n&s=a&id=307338&sv=za5cb0dea&uid=0A73F9896D13BE614&sid=12CF" and the "Internet" icon.

Retrieval-Szenarien

„Erstelle eine Ausarbeitung zu dem genannten Thema.“



Thema: Die Rolle von MacDonal'd'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 [AIssearch](#) 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

Daten



Wissen



Information

syntaktische Ebene,
signifikative 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 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.

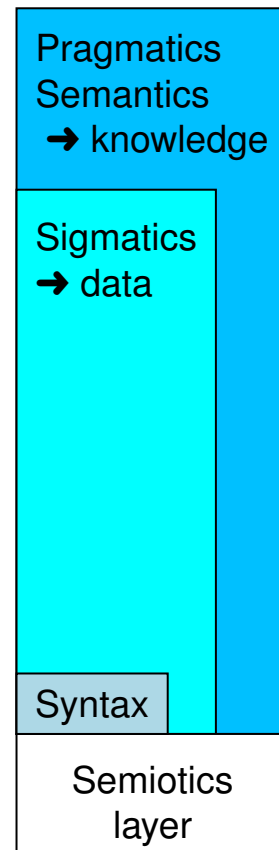
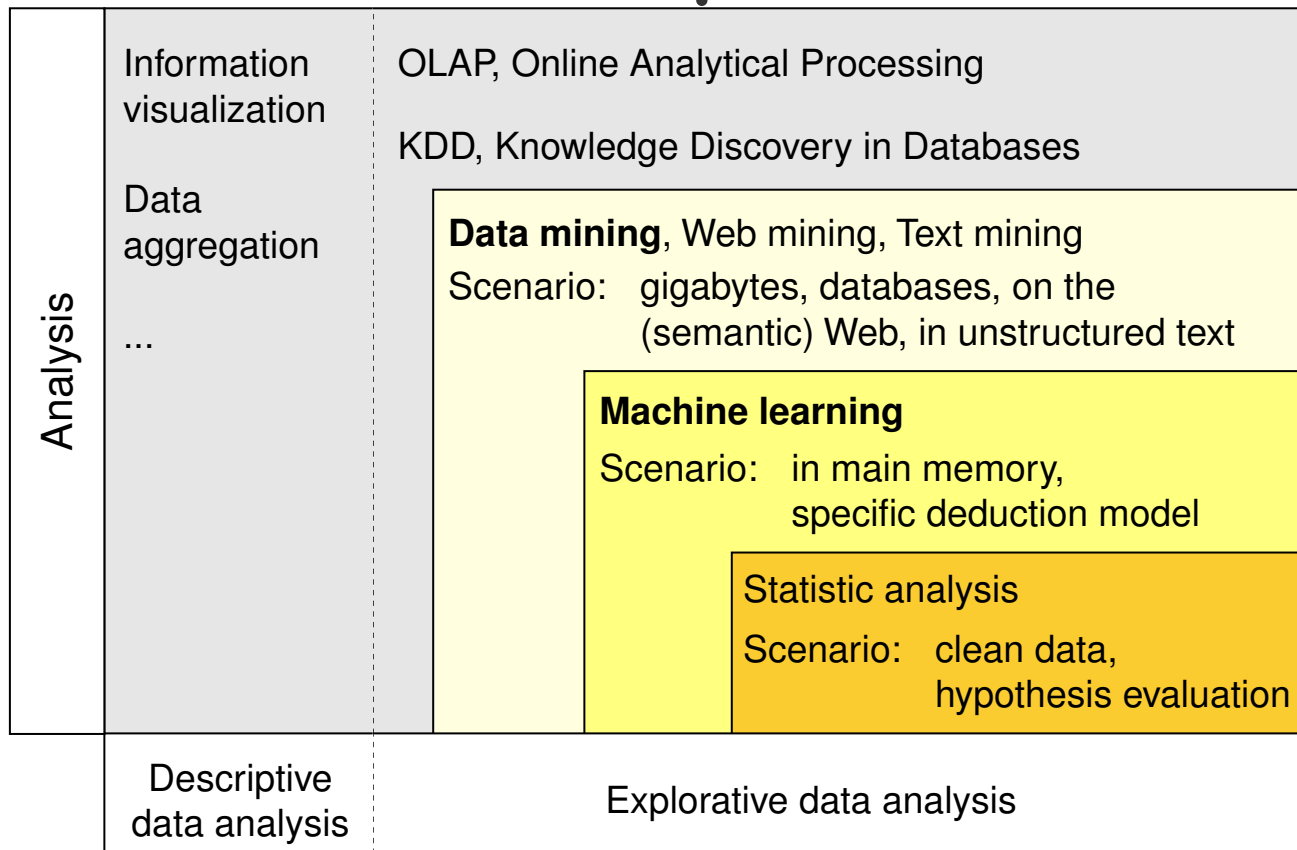
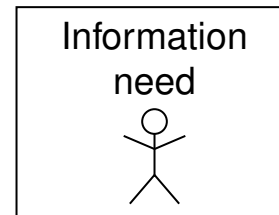
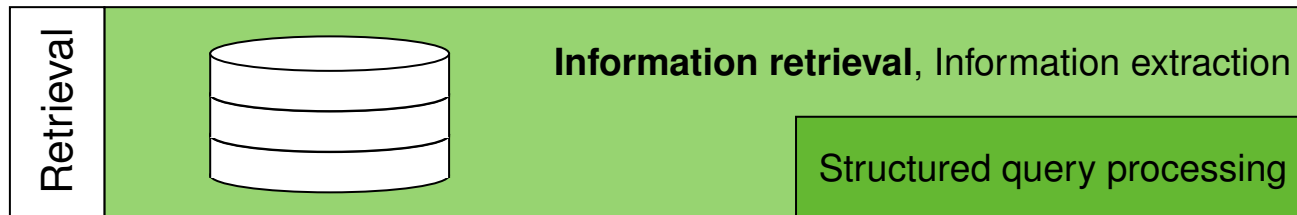
Einordnung Information Retrieval

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



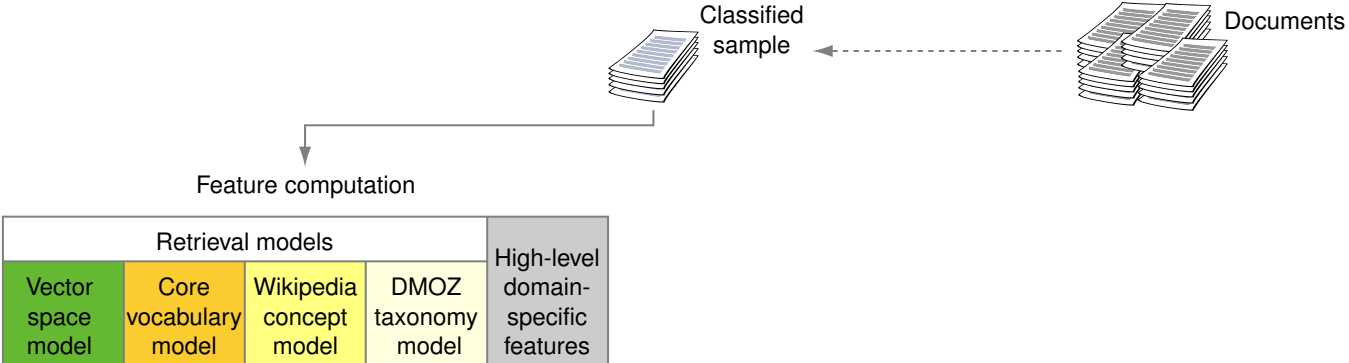
Einordnung Information Retrieval

Anwendung: Personenkategorisierung im Web (Spock-Challenge)



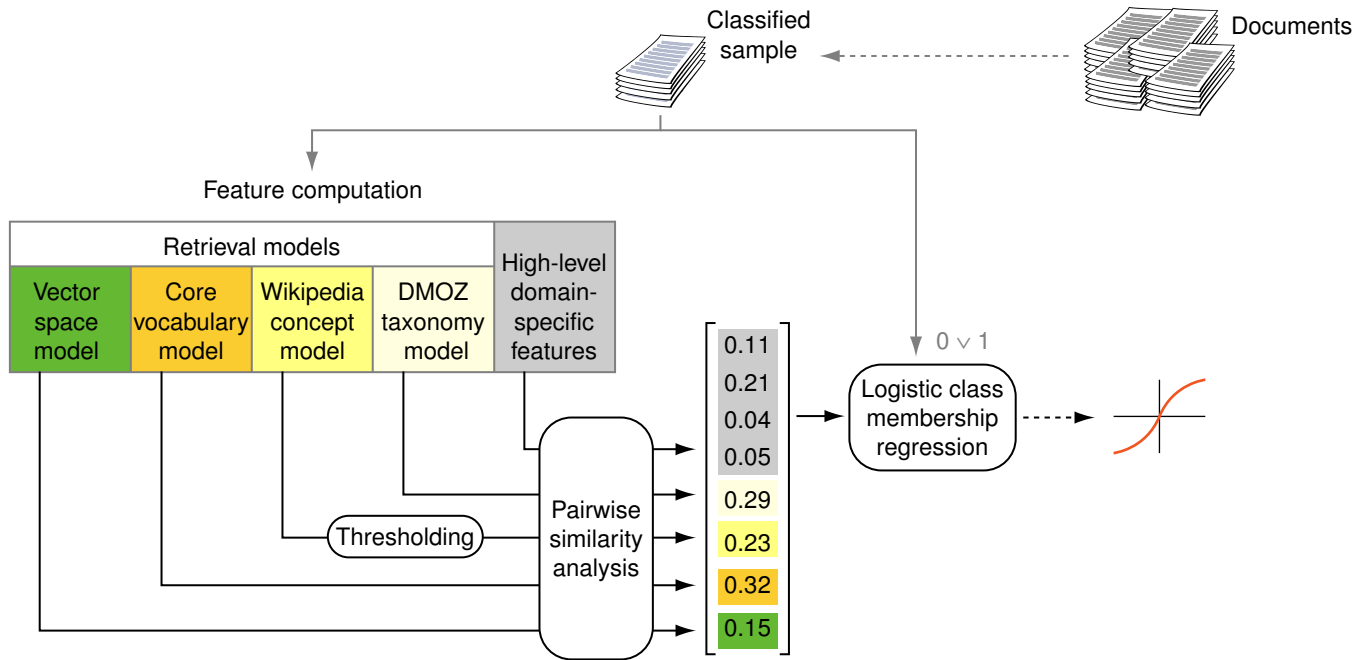
Einordnung Information Retrieval

Anwendung: Personenkategorisierung im Web (Spock-Challenge)



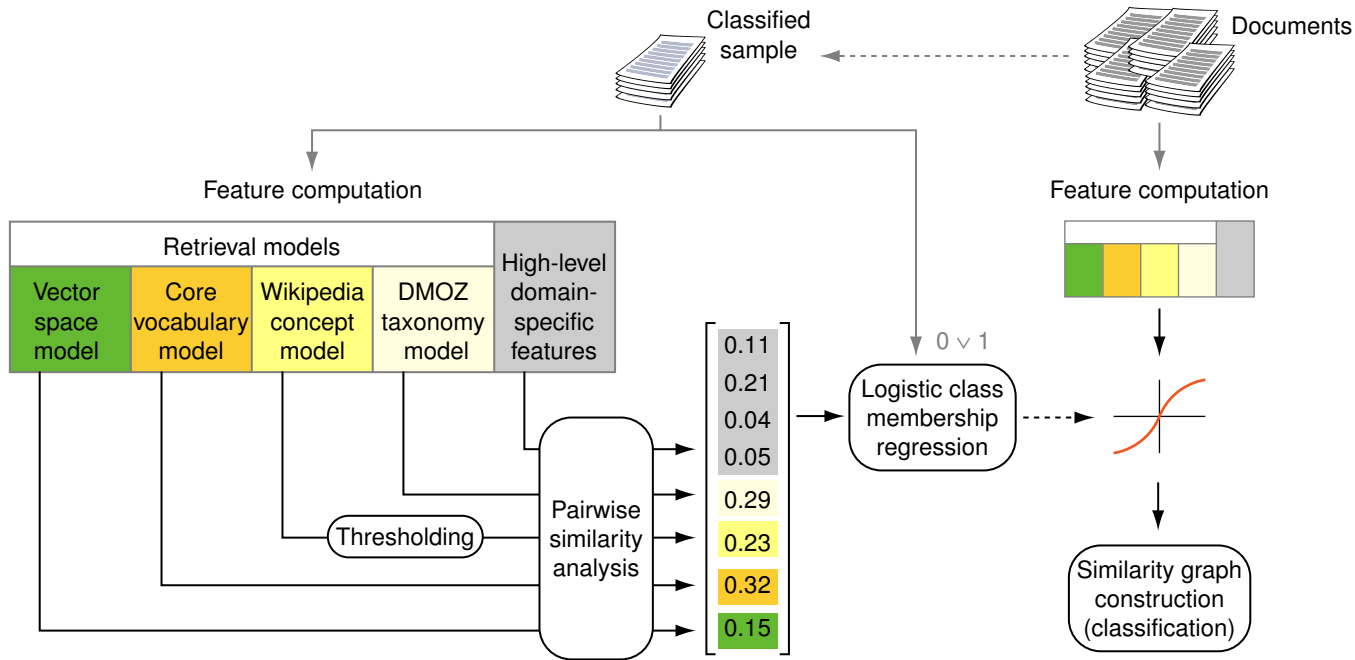
Einordnung Information Retrieval

Anwendung: Personenkategorisierung im Web (Spock-Challenge)



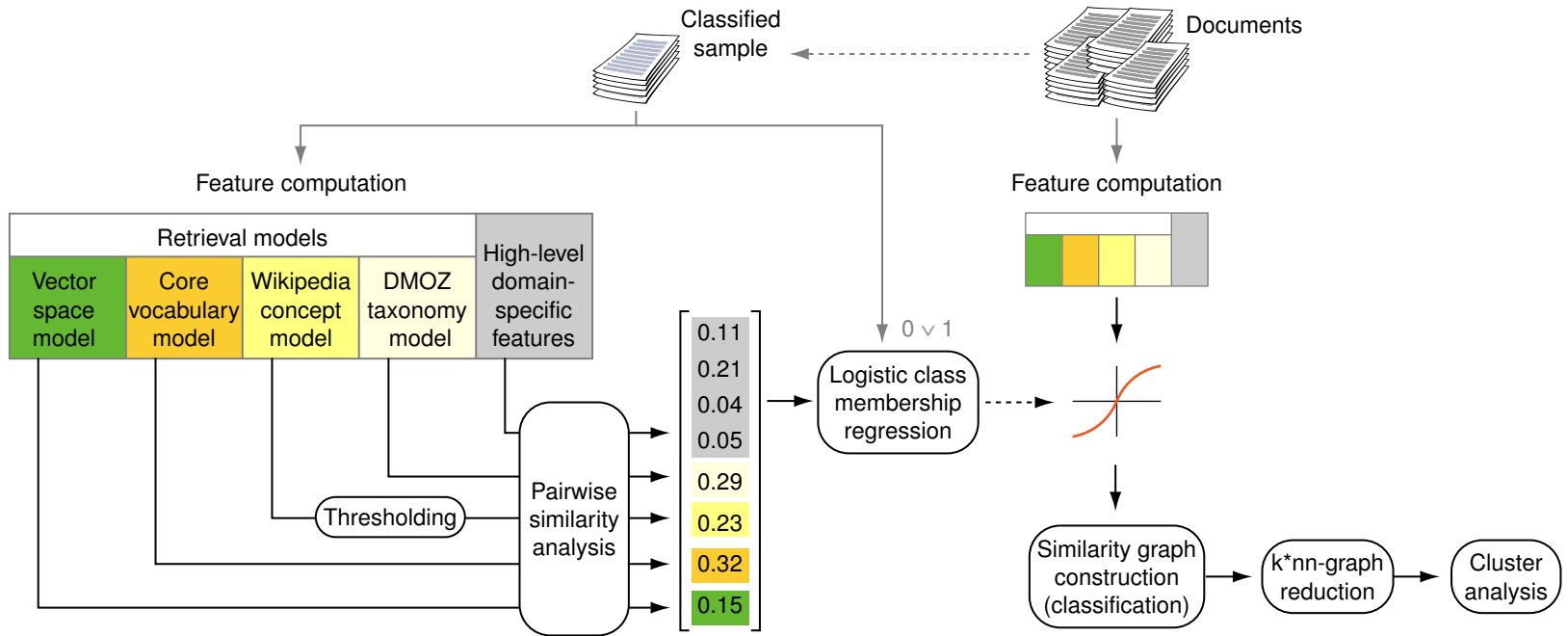
Einordnung Information Retrieval

Anwendung: Personenkategorisierung im Web (Spock-Challenge)



Einordnung Information Retrieval

Anwendung: Personenkategorisierung im Web (Spock-Challenge)



Einordnung Information Retrieval

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 Language Processing (NLP)
- ❑ Web-Technologie
- ❑ Datenstrukturen, effiziente Symbolverarbeitung