Internet-Suchmaschinen

Prof. Dr. Gerd Stumme
Dr. Andreas Hotho
Dipl.-Inform. Beate Krause

Wintersemester 2007/08

ENDOWED CHAIR OF THE HERTIE FOUNDATION
Knowledge and Data Engineering
DEPARTMENT OF MATHEMATICS & COMPUTER SCIENCE

Organisatorisches

Präsenzübung bedeutet

- **selbständiges Bearbeiten** des Übungsblattes in Kleingruppen à 3-4 Personen unter Betreuung des Assistenten
- **kein prinzipielles Wiederholen** des Vorlesungsstoffs
- **kein Vorrechnen** der Musterlösung etc. (Diese wird später zur Verfügung gestellt.)
- **Nötig dafür:**
  - selbständige Vorlesungsnachbereitung **vor** der Übung
  - Mitbringen des Skriptes
  - eigene Aktivität entfalten

Organisatorisches

Vorlesung

- Beginn: 23. Oktober 2007
- Dienstag 14:15-15:45 Uhr, Raum 0443

Übungen

- Beginn: 1. November 2007
- Donnerstag 8:30-10:00 Uhr, Raum 0443
- wird als Präsenz- und Praxisübung abgehalten (s. nächste Folie)
- Programmierhausaufgaben

Unterlagen

- siehe Literatur

Prüfung

- Die Prüfung wird je nach Teilnehmerzahl mündlich oder schriftlich abgehalten.

Organisatorisches

Warum ein neues Übungskonzept?

- aktives Erarbeiten des Vorlesungsstoffes bringt mehr
- Zusammenhänge im Stoff erkennen
- strukturiertes Denken und selbständiges Arbeiten lernen
- Teamarbeit lernen
- Erklären lernen (als Tutor und als Teilnehmer)
- Klausurtraining ;-)
- *Ihr Studium der ... haben Sie abgeschlossen. Zu Ihren persönlichen Stärken zählen Sie Eigeninitiave, Kommunikations- und Kooperationsbereitschaft, Teamarbeit.* (Typischer Anzeigentext)

Organisatorisches

Praxisübung – Implementieren einer Suchmaschine

- Ausgabe der ersten Praxisaufgabe zur ersten Übung am 1.11.07
- Am 8.11.07 Fragestunde zur Praxisaufgabe
- Abgabe der ersten Praxisaufgabe am 15.11.07
- Praxisaufgaben im 14 Tagerhythmus
- 4 von 6 Aufgaben müssen für die Teilnahme an der Prüfung abgegeben werden

Organisatorisches

Sprechstunden nach Absprache:

Gerd Stumme:	stumme@cs.uni-kassel.de	0561/804-6251
Andreas Hotho:	hotho@cs.uni-kassel.de	0561/804-6252
Beate Krause:	krause@cs.uni-kassel.de	0561/804/6254

FG Wissensverarbeitung, FB Mathematik/Informatik

Raum 0440, Wilhelmshöher Allee 73

Informationen im Internet: http://www.kde.cs.uni-kassel.de

Hier ist u.a. folgendes zu finden:

- aktuelle Ankündigungen
- Folienkopien
- Übungsblätter
- Literaturempfehlungen
- Termine

ENDOWED CHAIR OF THE HERTIE FOUNDATION
Knowledge and Data Engineering
DEPARTMENT OF MATHEMATICS & COMPUTER SCIENCE

Literatur

Wesentliche Quellen

- Ricardo Baeza-Yates & Berthier Ribeiro-Neto. Modern Information Retrieval, New York, NY: ACM Press; 1999; 513 pp. (ISBN: 0-201-39829-X.)
- Ian H. Witten, Alistair Moffat, and Timothy C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images, Morgan Kaufmann Publishing, San Francisco, ISBN 1-55860-570-3.
- Reginald Ferber. Information Retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. dpunkt-Verl.: Heidelberg 2003.
- Rijsbergen. C.J van, Information retrieval, http://www.dcs.gla.ac.uk/Keith/Preface.html
- Konzepte des Information Retrieval, http://irgroup.cs.uni-magdeburg.de/dt/vorlesungen/WS03-04_KIR.htm
- Intelligent Information Retrieval and Web Search, http://www.cs.utexas.edu/users/mooney/ir-course/

Literatur

Weiteres Material

- R.R. Korfhage. Information storage and retrieval. Wiley: New York, 1997
- G. Salton / M.J. McGill. Information Retrieval - Grundlegendes für Informationswissenschaftler. McGraw-Hill: Hamburg etc., 1987
- Machine Learning, Tom Mitchell, McGraw Hill, 1997.

Wir wollen wissen …

- Wie funktionieren Google und MSN Search?
  - Wie sammeln sie Informationen?
  - Welche Tricks benutzen sie?
  - Mögliche Nutzung außerhalb des Webs?
- Wie kann man diese Ansätze verbessern?
  - Verstehen von natürlicher Sprache?
  - Benutzerinteraktion?
- Was kann man tun, um diese Ansätze zu beschleunigen?
  - Schnellere Computer? Caching? Kompression?
- Wie entscheiden wir, ob die Ansätze funktionieren oder nicht?
  - Im allgemeinen für alle Anfragen, oder für spezielle Anfragen?
  - Für spezielle Dokumentensammlungen oder das Web?
  - Maße?
- Was kann man noch mit diesen Ansätzen machen?
  - Andere Medien?
  - Andere Aufgaben?

Übersicht

- Einführung
- Boolesches und Vektorraum-Retrieval-Modelle
- Elementares Tokenizing, Indexing, und die Implementierung von vektorraumbasiertem Retrieval
- Performanz-Bewertung von Retrieval-Systemen
- Anfrage-Operationen (Relevance Feedback, Anfrageerweiterung)
- Anfragesprachen und –paradigmen
- Strukturelle Anfragen
- Texteigenschaften
- Web-Suche: Einführung, Crawling, Interfaces, Link-Analyse
- Empfehlungssysteme
- Text-Clustering & -Klassifikation
- Informations-Extraktion
- Aktuelle Suchmaschinen, Trends, Suche im Web 2.0

Einführung

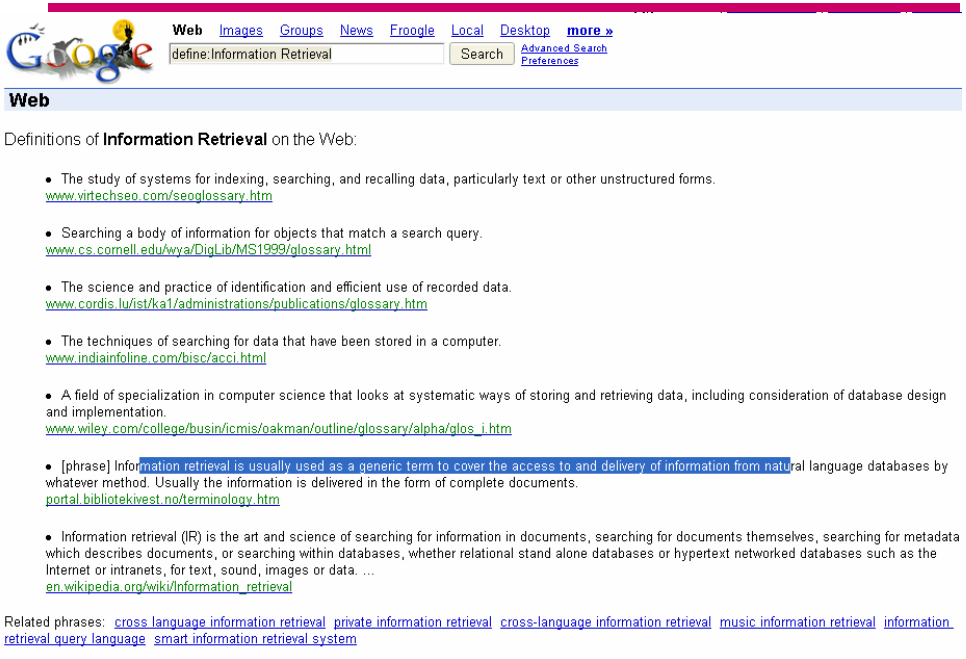
Was ist Information Retrieval (IR)?

Information-Retrieval (IR) (Informationswiedergewinnung, gelegentlich Informationsbeschaffung)

ist eine Forschungsrichtung, die sich mit computergestützter, inhaltsorientierter und unscharfer Suche in unstrukturierten Datenmengen beschäftigt.

http://de.wikipedia.org/wiki/Information-Retrieval
http://www.ib.hu-berlin.de/~is/web-lehrsammlung/Begriffe/Retrieval.htm

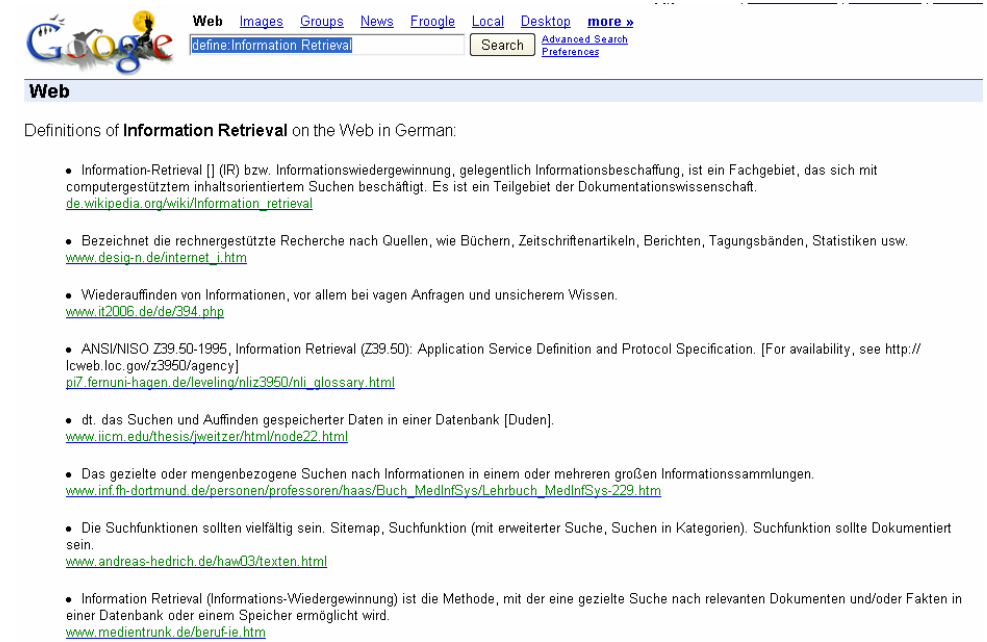
Einführung



Was ist Information Retrieval (IR)?

- Indexierung und Retrieval (Finden, Wiederfinden) von Texten
- Suchen nach Seiten im World Wide Web ist die aktuelle “killer app”
- Beschäftigt sich in erster Linie mit dem Finden der *relevanten* Dokumente gemäß einer gegebenen Frage (Query)
- Beschäftigt sich außerdem mit dem *effizienten* Finden von Dokumenten in *großen* Dokumentensammlungen

Einführung



Information Retrieval - Data Retrieval

	Data Retrieval	Information Retrieval
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Classification	Monothetic	Polythetic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive

C.J. van Rijsbergen, 1979 S.1

Eine typische IR-Aufgabe:

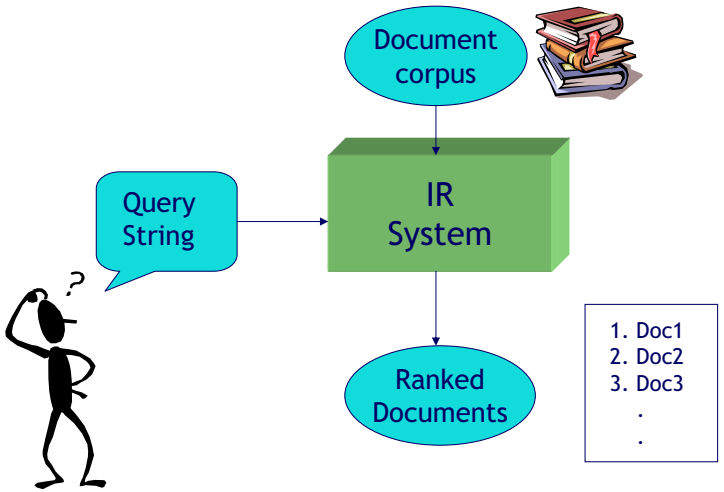
Gegeben:

- Textkorpus mit natürlichsprachlichen Textdokumenten.
- Eine Benutzeranfrage in Form eines Textstrings.

Finde:

- Eine geordnete Menge an Dokumenten, die relevant zur Anfrage sind.

IR System



Drei Phasen des IR

- Fragestellung (Information Need)
- Bestimmung einer Antwort (Response)
- Bewertung der Antwort (Evaluation)

→ Interaktiver Prozess

Frage stellen

- Fragesteller = “user”
  - Befindet sich in einem bestimmten Umfeld / Bewusstsein - ein kognitiver Zustand
  - Ist sich seiner Wissenslücken bewusst
  - Kann diese Lücken evtl. nicht genau bestimmen
- Paradox des FOA (*Finding Out About*):
  - Wenn der Nutzer in der Lage ist, die richtige Frage zu stellen, besteht häufig keine Notwendigkeit mehr für diese Frage.
    - “The need to describe that which you do not know in order to find it.” Roland Hjerppe
- Anfrage
  - Ausdruck dieses schlecht definierten Zustandes

Frage beantworten

- Wenn der Antwortende ein Mensch ist:
  - Ist er in der Lage, die schlecht gestellte Frage in eine bessere umzuformulieren?
  - Kennt der Antwortende die Antwort?
  - Kann er diese Antwort in Worten ausdrücken?
  - Wird der Anfrager diese Antwort verstehen?
  - Haben beide das notwendige Hintergrundwissen?
- Wenn der Antwortende ein Computersystem ist...

Bewertung der Antwort

- Wie gut wird die Frage beantwortet?
  - Wurde die Antwort vollständig beantwortet oder nur teilweise?
  - Wurden Hintergrundinformationen zur Verfügung gestellt?
  - Wurden Hinweise für weitergehende Untersuchungen gegeben?
- Wie relevant ist die Antwort für den Frager?

Relevanz

Relevanz ist eine subjektive Beurteilung und kann folgendes einschließen:

- Richtiges Thema
- Aus der richtigen Zeit (zeitgemäß)
- Aus vertrauenswürdiger Quelle (verlässlich)
- Antwort berücksichtig die Ziele des Nutzers und die beabsichtigte Nutzung der Information (*information need*)

Relevanz (Forts.)

In welcher Art kann ein Dokument relevant sein für eine Anfrage?

- Präzise Antwort auf eine präzise Frage.
  - Wer ist in Meiers Grab begraben? Meier.
- Frage wird teilweise beantwortet.
  - Wo ist Söhrewald? In der Nähe von Kassel.
- Weitere Informationsquellen vorschlagen.
  - Was ist Lymphedema? Schau in diesem medizinischen Lexikon nach.
- Hintergrundinformationen geben.
- Den Fragesteller an relevante, ihm bekannte Informationen erinnern.

Relevanz bei der Stichwort-Suche [Keyword Search]

- Die einfachste Form der Relevanz ist das wortwörtliche Vorkommen des Anfragestrings im Text.
- Eine weniger restriktive Idee ist, dass **die einzelnen** Wörter* aus der Anfrage häufig im Textdokument vorkommen müssen (*bag of words*).

* Siehe http://www.spiegel.de/kultur/zwiebelfisch/0,1518,307445,00.html zum Unterschied zwischen Worten und Wörtern!

Probleme mit Stichwörtern

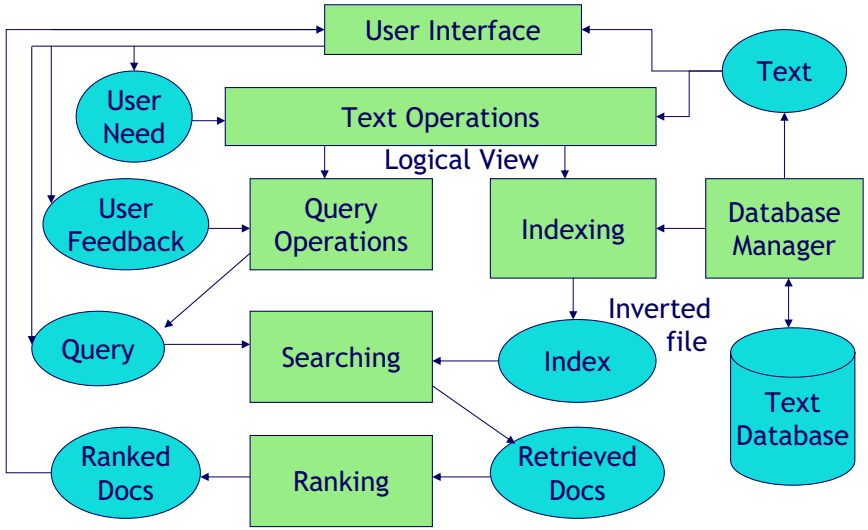
Man findet relevante Dokumente nicht bei synonymen Termen.
- "restaurant" vs. "café"
- "Auto" vs. "PKW"

Man erhält irrelevante Dokumente durch mehrdeutige Terme.
- "Bank" (Finanzinstitut vs. Sitzgelegenheit)
- "Apple" (company vs. fruit)
- "bit" (unit of data vs. act of eating)

Intelligentes IR

- Bedeutung des Wortes wird mit in Erwägung gezogen.
- Reihenfolge der Wörter in der Anfrage wird beachtet.
- Anpassung an den Anwender durch direktes oder indirektes Feedback.
- Zuverlässigkeit der Quelle wird beachtet.

IR-System-Architektur



IR-Systemkomponenten

Text Operations berechnet die Wörter des Indexes (tokens).
- Stopword removal
- Stemming

Indexing konstruiert einen *invertierten Index* aus Wörtern mit Zeigern zu den Dokumenten.

Searching findet mit Hilfe des invertierten Index Dokumente, die Tokens aus der Anfrage enthalten.

Ranking gewichtet alle gefundenen Dokumente gemäß einer Relevanzmetrik.

IR-Systemkomponenten (Forts.)

User Interface ist für die Interaktion mit dem Anwender verantwortlich:
- Anfrage entgegennehmen und Dokumente präsentieren.
- Relevance feedback.
- Visualisierung der Ergebnisse.

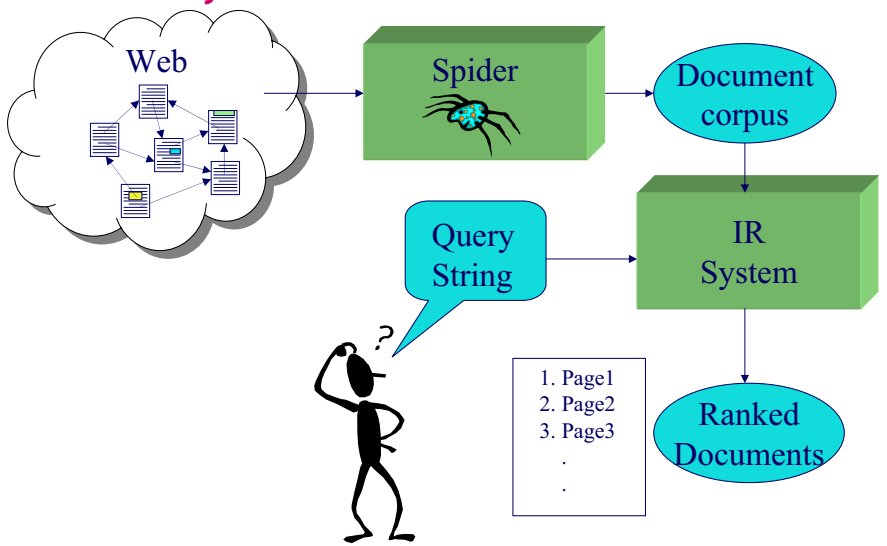
Query Operations verändert die Anfrage zur Verbesserung der Ergebnisse:
- Anfrageerweiterung (Query expansion) mittels Thesaurus.
- Anfrageanpassung mittels Relevance Feedback.

Anwendung: Web-Suche

Web-Suche ist die Anwendung des IR auf HTML-Dokumente des World Wide Web.

Unterschiede:
- Man muss die Dokumente für den Korpus im Web einsammeln (Crawling)
- Ausnutzung der strukturierten Layout-Information in HTML (XML).
- Unkontrollierbare Veränderung der Dokumente
- Ausnutzung der Linkstruktur

Web Search System



Weitere IR-nahe Aufgaben

- Automated document categorization (Kategorisieren)
- Automated document clustering (Gruppieren)
- Automated Text Summarization (Zusammenfassen)
- Question answering (Frage/Antwort)
- Information filtering (spam filtering) (Filtern)
- Information extraction (Extrahieren)
- Information integration (Integrieren)
- Recommending information or products (Empfehlen)
- Ranking in Web 2.0

Geschichte des WWW und des IR

bis 1960:

- "Informationsexplosion" nach dem Ende des zweiten Weltkrieges führt zur Notwendigkeit, diese besser zu organisieren.
- 1945 Vannevar Bush verfolgt mit der Memex-Maschine ähnliche Ideen wie sie im heutigen Web zu finden sind (Assoziation von Informationen mit Links).

Geschichte des WWW und des IR

1960-70er:

- Initiale Untersuchung von Text-Retrieval-Systemen für "kleine" Korpora bestehend aus Zusammenfassungen wissenschaftlicher Publikationen sowie Gesetzes- und Geschäftsdokumenten.
- Die Entwicklung einfacher Boolean- and Vector-Space-Modelle.
- Prof. Salton und seine Studenten an der Cornell Universität waren die führenden Forscher auf diesem Gebiet.
- 1965 Ted Nelson prägt den Begriff „Hypertext".

Geschichte des WWW und des IR (Forts.)

1980er:

- Systeme mit großen Dokumentensammlungen entstehen, viele laufen in Unternehmen:
  - Lexis-Nexis
  - Dialog
  - MEDLINE

Geschichte des WWW und des IR (Forts.)

1990er:

- Suche nach “FTPbaren” Dokumenten im Internet
  - Archie
  - WAIS
- Suche im World Wide Web
  - Lycos
  - Yahoo
  - Altavista

Geschichte des WWW und des IR (Forts.)

auch 1990er:

- Organisierte Wettkämpfe
  - NIST TREC (Text REtrieval Conference)
- Recommender-Systeme
  - Amazon
- Automatisiertes Text-Kategorisieren & -Clustern

Geschichte des WWW und des IR (fort.)

2000er

- Analyse der Links für die Web-Suche
  - Google
- Automatisierte Informationsextraktion
  - Whizbang
  - Burning Glass
- Frage/Antwort (Question Answering)
  - TREC Q/A track

Geschichte des WWW und des IR (fort.)

auch 2000er:

- Multimedia-IR
  - Image
  - Video
  - Audio und Musik
- Mehrsprachiges IR (Cross-Language IR)
  - DARPA Tides (Translingual Information Detection, Extraction and Summarization)
- Zusammenfassen von Dokumenten

Verwandte Forschungsgebiete

- Datenbanken (Database Management)
- Bibliothekswesen (Library and Information Science)
- Künstliche Intelligenz (Artificial Intelligence)
- Sprachverarbeitung (Natural Language Processing)
- Maschinelles Lernen (Machine Learning)
- Data Mining

Datenbanken (Database Management)

- Fokussiert auf strukturierte Daten, die in relationalen Tabellen gespeichert sind und nicht auf freien Text.
- Beschäftigt sich mit der effizienten Abarbeitung von wohldefinierten Anfragen in einer formalen Sprache (SQL).
- Klare Semantik für Daten und Anfragen.
- Aktuell beschäftigt man sich auch mit semi-strukturierten Daten wie XML (bringt DB näher zu IR)

→ Datenbanken-Vorlesung im Sommersemester

Bibliothekswesen (Library and Information Science)

- Fokussiert auf die Mensch-Maschine-Schnittstelle des IR (human-computer interaction, user interface, visualization).
- Beschäftigt sich mit der effektiven Kategorisierung menschlichen Wissens.
- Beschäftigt sich mit der Analyse des Verhältnisses zwischen Personen und Publikationen.
- Aktuelle Arbeiten im Bereich der Digitalen Bibliotheken *bringen das BW näher an* IR.

Künstliche Intelligenz (Artificial Intelligence)

- Fokussiert auf Methoden zur Akquisition, Repräsentation und zum Ableiten von (neuem) Wissen.
- Formalismen zur Repräsentation von Wissen und Anfragen sind:
  - Prädikatenlogik
  - Beschreibungslogiken
  - Bayesian Networks
- Aktuelle Arbeiten im Bereich Semantic Web und Ontologien schaffen einen engeren Bezug zu IR.

→ Vorlesung Künstliche Intelligenz - Di 16-18 Uhr

Sprachverarbeitung (Natural Language Processing)

- Fokussiert auf die syntaktische, semantische und pragmatische Analyse von natürlichsprachlichem Text
- Die sytaktische und semantische Analyse könnte eine bedeutungsbezogene anstatt einer stichwortbasierten Suche ermöglichen.

Sprachverarbeitung in Richtung IR:

- Methoden zur Wortsinnerkennung von mehrdeutigen Wörtern im Kontext (*word sense disambiguation*).
- Methoden zur Identifikation von spezifischen Informationen in Texten (*information extraction*).
- Methoden zur Beantwortung von natürlichsprachlichen Anfragen auf Dokumentkorpora.

Maschinelles Lernen (Machine Learning, ML) KDD, Data Mining

- Fokussiert auf die Entwicklung von Systemen, die in der Lage sind, ihre Leistung anhand ihrer Erfahrung zu steigern.
- Automatische Klassifikation von Beispielen basierend auf Lernmethoden die auf beschrifteten Trainingsbeispielen basieren (*supervised learning*).
- Automatisierte Methoden zum Gruppieren von unbeschrifteten Beispielen in bedeutungsvolle Gruppen (*unsupervised learning*).

→ **Knowledge-Discovery-Vorlesung**

ML in Richtung IR:

Text-Kategorisierung
- Automatisches Klassifizieren in Hierarchien (Yahoo).
- Adaptive Filter/Recommender.
- Automatische Spamfilter.

Text-Clustern
- Clustern von IR Anfrageergebnissen.
- Automatisches Ableiten von Hierarchien (Yahoo).

Lernen für Informationsextraktion

Text Mining

Overview

- Introduction
  - What is IR, task, systems (in detail), history
  - Web search
  - IR related tasks
  - IR related research areas
- Boolean and Vector-Space Retrieval Models
  - Retrieval Models (Boolean, Statistical, Vector Space Model)
  - Weighting, Similarity Measure
- Basic Tokenizing, Indexing, and Implementation of Vector-Space Retrieval
  - Tokenizing, Stopwords, Stemming
  - Implementation of Sparse Vectors, Inverted Files, IDF computing
  - Retrieval with an Inverted Index
  - Analysis of time complexity
- Performance Evaluation of Information Retrieval Systems
  - Gold standard - Precision, Recall, F-Measure, Rank measures
  - Subjective relevance measures
  - Trec, Cystic Fibrosis Collection
- Query Operations (Relevance Feedback / Query Expansion)
  - Query Reformulation – Rochio Model, Pseudo Feedback, Thesaurus, Wordnet, statistical Thesaurus
  - Local vs. global Analysis of the query
- Query Languages
  - Boolean, Natural Language, Phrasal, Proximity and Structural Queries
  - Pattern Matching, Levenstein Distance, Regular Expressions
- Text Properties and Languages
  - Zipf's Law
  - Meta Data
- Web Search: Introduction
  - WWW history, Challenges for IR, Statistics about the web, web search principle
- Web Search: Spidering
  - Spiders, spider programming in java, link extraction, multi threaded spider, topic directed spider
- Web Search: Interfaces
  - Interface, Clustering
  - Apache TomCat, Servlet, Session Tracking, Simple Search Servlet
- Web Search: Link Analysis
  - Meta Search Engines, Bibliometrics, Hits, PageRank, Google Ranking,
- Recommender Systems
  - Book recommender, collaborative filtering, content based recommender, combination
  - experiments movie domain
  - Active learning
- Text Clustering & Classification
  - Introduction of Clustering and Classification, specific text properties for clustering and classification
- Information Extraction
  - MUC, Simple pattern, Template based Extraction, Filler Extraction,
  - Learning for IE
  - Web Extraction (shop bot)
- Aktuelle Suchmaschinen, Trends, Suche im Web 2.0