



Fachhochschule Köln,
Campus Gummersbach
Institut für Informatik und Ingenieurwissenschaften

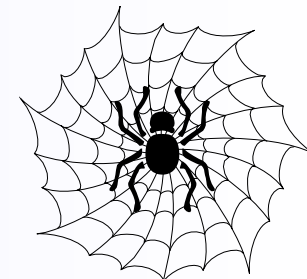
Information Retrieval

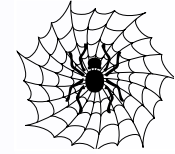
WPF 45

Überblick

Prof. Dr. Heide Faeskorn - Woyke

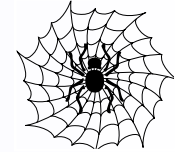
Fachhochschule Köln
Campus Gummersbach
Institut für Informatik
faeskorn@gm.fh-koeln.de





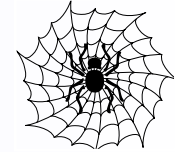
Ziele

- **Inhaltliche Ziele**
 - Grundlagen und wesentliche Komponenten von IR und Text Mining kennen und anwenden können
 - Standards und Architekturen kennenlernen
 - Ein spezielles Anwendungsgebiet von IR/TextMining selbständig bearbeiten und präsentieren
- **Nebenziele (Soft Skills)**
 - Selbstständige Recherche (insbesondere Internet, aber nicht nur) auch in **englischen** Texten
 - Anwendung von **Präsentationsprogrammen**
 - **Erstellung** von aussagefähigen Präsentationen
 - **Halten** von Referaten / Präsentationen
 - **Technische** Sachverhalte **verständlich** vermitteln



Vorläufiger Zeitplan WPF 45 Information Retrieval

Termin	Thema
26.03.09	Vorstellung des WPF's
02.04.09	Festlegung der Referatsthemen und Grundlagen IR
23.04.09-14.05.09	Grundlagen von IR/Text Mining (FW, LE), evt. ein Termin mehr...
21.05.09- 02.06.09	Beratungstermin für Referate: Donnerstags, ab 15 Uhr in 2.230 Teilnehmer bereiten Referate vor und verschicken diese per Email an alle Teilnehmer des Fachs
02.06.09-11.06.09	Beratungstermin für Referate: Donnerstags, ab 15 Uhr in 2.230 Teilnehmer bewerten und geben Anregungen zu Referaten von vier Kommilitonen
11.06.06-18.06.09.	Beratungstermin für Referate: Donnerstags, ab 15 Uhr in 2.230 Anregungen werden in die Referate eingearbeitet
Ab 18.06.09	Teilnehmer halten Referate



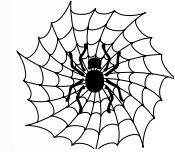
Was ist von den Teilnehmern zu leisten? (1)

- **Teilnahme** an den Veranstaltungen 😊
- Übernahme eines **Referatsthemas**
 - In dieser Präsentation finden Sie Vorschläge
 - Bitte davon einen Vorschlag auswählen oder ein eigenes Thema vorschlagen
 - Benotet werden inhaltliche und formale Aspekte
 - geht zu **40 %** in die Endnote ein
- **Schriftliche Ausarbeitung** des Referats:
 - geht zu **40 %** in die Endnote ein
- **Bewertung der Referate** von 3-4 Kommilitonen
 - geht zu **20 %** in die Endnote ein



Was ist von den Teilnehmern zu leisten? (2)

- **25 Teilnehmer** werden zugelassen
 - **Voraussetzung:** Abgeschlossenes Grundstudium
 - Bei mehr als 25 Anmeldungen entscheidet die DBS-Klausurnote
- **die Übernahme eines Referatsthemas**
 - Powerpoint-Präsentation
 - Incl. Übungsaufgabe für die Teilnehmer, falls möglich
 - Jeder Teilnehmer hat insgesamt 45 Minuten Zeit
 - Geht zu 40 % in die Endnote ein
- **Schriftliche Ausarbeitung**
 - Umfang: 10-15 Seiten
 - Auf wissenschaftlichen Stil achten
 - Geht zu 40 % in die Endnote ein



Was ist von den Teilnehmern zu leisten? (3)

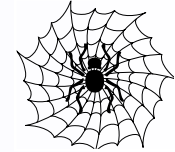
- **Benutzerzentrierte Bewertung:**
 - Vor dem Referatstermin (siehe Zeitplan) wird ein Entwurf per Email an alle Teilnehmer verschickt
 - Teilnehmer werden in 4-erGruppen eingeteilt
 - Teilnehmer geben fundierten Kommentare zu den Referaten einer anderen Gruppe ab
 - Diese Kommentare gehen mit in die Endnote ein, und zwar nicht in das Referat, welches bewertet wird, sondern für denjenigen, der bewertet.
 - Referenten haben noch mindestens eine Woche Zeit, die Anmerkungen zu verwenden und die eigenen Referate zu verbessern
 - geht zu **20 %** in die Endnote ein



Was ist von den Teilnehmern zu leisten? (3)

• **Achtung**

- Verspätete Abgaben von Referatsentwürfen sind bei diesem Benotungssystem nicht möglich, da sonst keine Bewertungen durchgeführt werden können
 - Müssen daher mit 5 bewertet werden
- Falls jemand seine Bewertung für einen anderen Kommilitonen nicht durchführt, gilt das Gleiche:
 - Wird mit 5,0 bewertet
- Dafür haben Sie ja im Semester eine Pause, um die Referate vorzubereiten 😊



Inhaltsübersicht

1. Einführung

1.1 Motivation

1.2 Suchmaschinen

1.3. Unterschiede IR –Suche zur Datenbanksuche

2 IR Grundprinzip

2.1 Text Retrieval

2.2 Aufgaben für ein IRS

2.3 Index, invertierte Liste und Signatur

2.4. Recall, Precision, Fallout

2.5 Rangfolge und Relevanz

3 Vagheit in der Sprache

3.1 Stoppworte

3.2 Stemming

3.3. Mehrwörtergruppen

3.4. Thesauri

4 Traditionelle IR Modelle

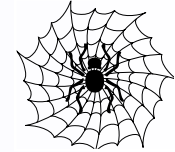
4.1 Boolesche Modell

4.2 Fuzzy-Modell

4.3 Vektorraummodell

4.4 Probabilistisches Modell

4.5 Latent Semantic Index



Inhaltsübersicht

- 5.1 IR im WWW**
- 5.2 Aufbau einer Suchmaschine**
- 5.3 Crawling**
- 5.4 Funktionen einer Suchmaschine**
- 5.5 Der PageRank-Algorithmus**
- 5.6 Probleme bei Suchmaschinen**

- 6.1. Metadatenansatz und Dublin Score**
- 6.2. Semantic Web**
TXT → XML → RDF(S)



Diese Inhaltangabe wird noch um Text-Mining-Inhalte verstärkt



Referatsthemen (1)

Referatsthemen:

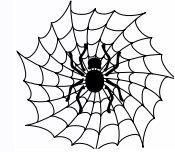
1. Probabilistische Retrieval Modelle
 - Ferber, Kapitel 10, Stock, Kapitel , Kap. 21
2. Latent Semantic Indexing und verwandte Modelle
 - Henrich
3. Natural Language Processing : Automatische Texterkennung
 - Stock und GI-Herbsttreffens (Heyer)+ Semantic Web –Buch, Granitzer
4. Bewertung und Vergleich von IR-Systemen (TREC)
 - Ferber, Mandl, Henrich
5. Alternativen zur globalen Suche: Kataloge, Cluster und Browsing (Henrich)
6. Interaktives Information Retrieval (Fuhr)



Referatsthemen (2)

Referatsthemen:

7. Multi Media Retrieval (Schmitt)
8. Peer-to-Peer Information Retrieval (Henrich)
9. Klassifizierung von WEB-Suchmaschinen (Stock, Web & Datenbanken)
10. Digitale Bibliotheken
11. WEB-Datenextraktion (Semantik Web, Baumgartner) , Springer und Datenbankspektrum)
12. Lucene von Apache



Literatur (Auswahl)

- Fachgruppe IR der GI: <http://www.uni-hildesheim.de/~fgir/>
- Ferber, Reginald: Information Retrieval, Suchmodelle und Data Mining Verfahren für Textsammlungen und das WEB, dpunkt, 2003. <http://information-retrieval.de/>
- Fuhr, Norbert: Skriptum Information Retrieval: http://www.is.informatik.uni-duisburg.de/courses/ir_ss06/folien/irskall.pdf
- Henrich: Information Retrieval, Information Retrieval, http://www.unibamberg.de/fileadmin/uni/fakultaeten/wiai_lehrstuehle/medieninformatik/Dateien/Publikationen/2008/henrich-ir1-1.2.pdf
- Pellegrini, T, Blumauer, A: Semantic Web, Wege zur vernetzten Wissensgesellschaft, Springer, 2005
- Rahm, E., Vossen, G. : Web und Datenbanken. Konzepte, Architekturen, Anwendungen, Springer, 2000.
- Rijsbergen: Information Retrieval: <http://www.dcs.gla.ac.uk/Keith/Preface.html>
- Schmitt, Ingo: Ähnlichkeitssuche in Multimediatatenbanken, Oldenbourg, 2006.
- Stock, W.: Information Retrieval, Oldenbourg, 2007.