Informationswissenschaft
Universität Hildesheim
womser@uni-hildesheim.de

Evaluierung von Information Retrieval Systemen

Die IR-Studie: Einführung

www.ir-studie.de

Informationswissenschaft
Universität Hildesheim
womser@uni-hildesheim.de

Agenda

14.00 - 14.30 Uhr Begrüßung/Einleitung, Prof. Wormser Hacker

14.30 - 15.00 Uhr Methodik, Dr. Martin Braschler

15.00 - 15.45 Uhr Resultate Teil 1, Beispiele, Dr. Peter Schäuble

Pause

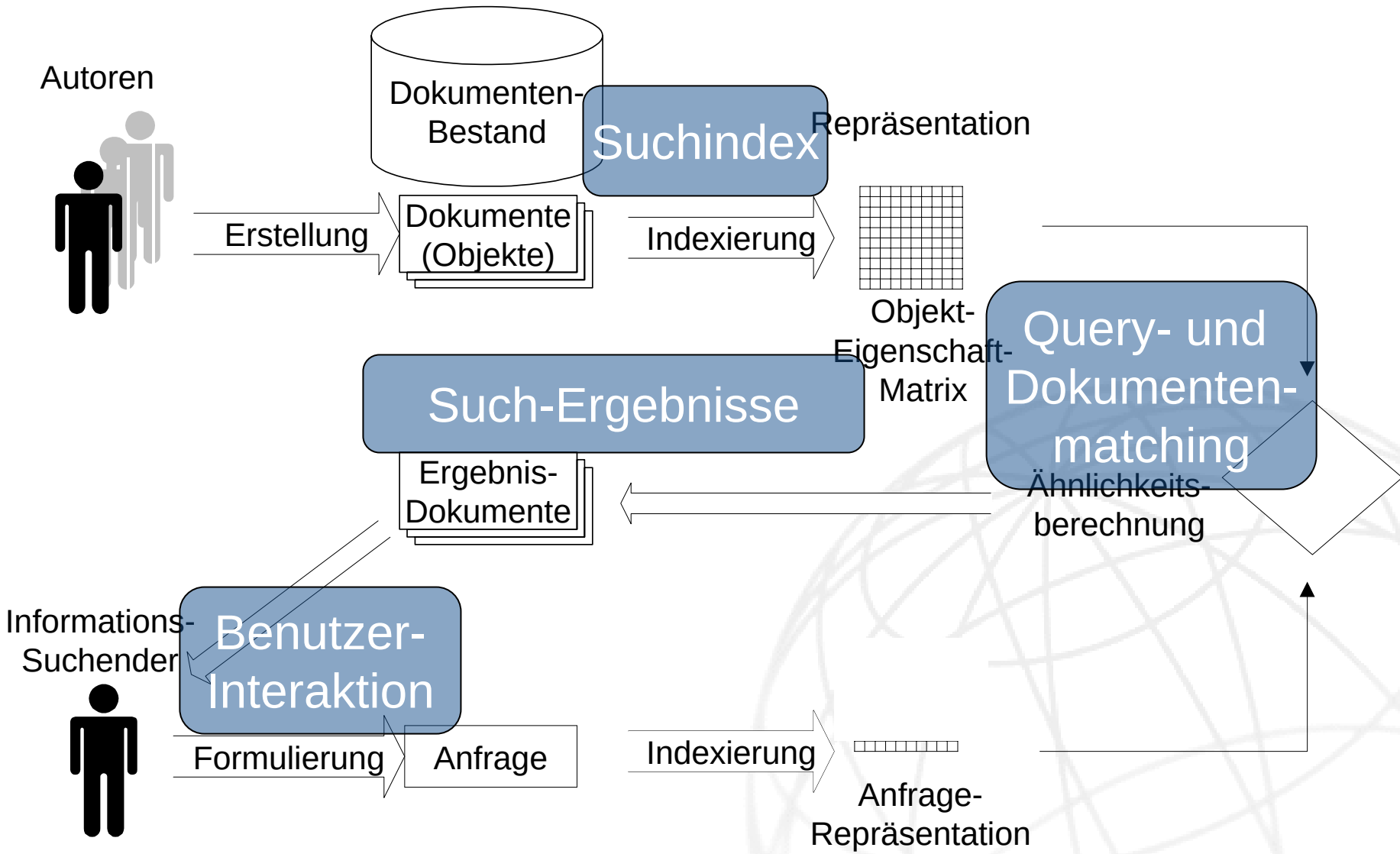
16.00 - 16.45 Uhr Resultate Teil 2, Beispiele, Jürg Stuker

16.45 - 17.30 Uhr Zusammenfassung, Dr. Thomas Mandl

17.30 - 18.00 Uhr Diskussion, Prof. Josef Herget

18.00 - … Get-together

Information Retrieval



„There must be some fundamental understanding of what it means to be good and what it means to be better“ (Bollmann/Cherniavsky 1983,3)

Am häufigsten evaluiert:

Qualität der Suchergebnisse

Recall und Precision

- „The ability of the retrieval system to uncover relevant documents is known as the recall power of the system“ (Lancaster 1968,55)

$$\text{Recall} = \frac{\textit{Anzahl gefundender relevanter Dokumente}}{\textit{Anzahl relevanter Dokumente}}$$

$$\text{Precision} = \frac{\textit{Anzahl gefundender relevanter Dokumente}}{\textit{Anzahl gefundener Dokumente}}$$

Cranfield-Paradigma der Evaluierung im Information Retrieval

- Objektive Relevanz wird von neutralem Beobachter beurteilt
- Beziehung zwischen dem Informations-wunsch und dem Dokument
- Keine individuelle und subjektive Relevanzbewertung
- Bis heute Testaufbau aller wichtigen Evaluierungsinitiativen im Information Retrieval (TREC, CLEF, NTCIR, INEX, ...)

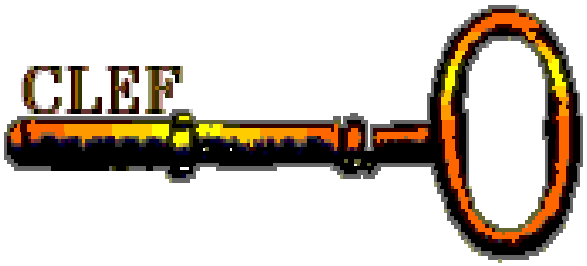
TREC: Text Retrieval Conference

- „TREC is a new ballgame for IR research and development“ (Sparck Jones 1994)
- Evaluierungsinitiative des National Institute of Standards and Technology (NIST) in den USA
- 1992: TREC-1 (Proceedings 1993)

NIST
National Institute of Standards and Technology

Cross-Language Evaluation Forum

Forschung zu cross- und multi-lingualen Information Retrieval Systemen



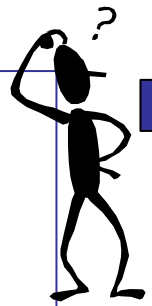
EU Förderung: DELOS NoE for Digital Libraries

- Testumgebung
- Systementwicklung
- Benchmarks
- Evaluierungsforschung

Ziele Evaluierungsinitiativen

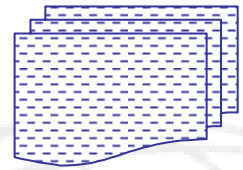
- Einheitliche Bewertungsmaßstäbe für Retrieval-Systeme finden (Standardisierung)
- Vergleich zwischen Systemen ermöglichen
- Weiterentwicklung von IR Systemen vorantreiben
- Anforderungen aus der Community aufnehmen und Methodik weiterentwickeln

Pooling Methode



Juroren erstellen Topics.

Systeme liefern die Top 1000 Dokumente für jedes Topic



Pooling aller einmal gefundenen Dokumente

Relevanz-Bewertung durch Juroren

Ellen Voorhees – CLEF 2001 Workshop

Quelle: TREC presentation slide No. 15

Text REtrieval Conference (TREC)

Wie zuverlässig ist die Evaluierung nach dem Cranfield-Paradigma?

Bedenken und Antworten

Einwand Subjektivität

- Hängt das Ergebnis von den Juroren ab, welche die Relevanz der Dokumente bewerten?
  - Juroren bewerten tatsächlich unterschiedlich
  - Dies wirkt sich aber nicht auf die Reihenfolge der Systeme aus
  - Der Vergleich fällt unabhängig von Juroren gleich aus

Buckley & Voorhees 2005

Einwand Anzahl Topics

- Reichen 50 Topics aus, um die Systeme zu vergleichen?
  - Zwischen zwei Systemen muss ein gewissen Unterschied bestehen, um statistisch sicher zu sein, dass eines besser ist als das andere
  - Ab 50 Topics liegt der Unterschied unter 5%
  - Teilweise auch deutlich unter 5% (absolut)

Sanderson & Zobel 2005

-> die Unterschiede in unserer Studie sind größer

Einwand Variablität

- Der Unterschied in der Performanz zwischen den Systemen ist wesentlich geringer als zwischen den Topics
  - Stabile Performanz über alle Anfragen wichtiger als hohe durchschnittliche Performanz
  - “Schwierige” Anfragen sollten höheres Gewicht bei der Evaluierung gewinnen

Buckley & Voorhees 2005, Mandl 2006

Einwand Aufgabentyp

- Können Evaluierungsergebnisse auf andere Aufgaben, Korpora und Benutzerintentionen übertragen werden?
  - Nein

Buckley & Voorhees 2005

-> Evaluierung muss für jeden Anwendungsfall erfolgen

Vorgehen IR-Studie

- Ganzheitliche Bewertung

Suchindex

Query- und Dokumenten-Matching

Such-Ergebnisse

Benutzer-Interaktion

Benutzer?

- Der Benutzer ist nie in der Rolle der Evaluierenden
- Führt eine Verbesserung der Systeme hinsichtlich der traditionellen Kennzahlen auch zu einer höheren Benutzerzufriedenheit?
- Führt ein gutes Ranking auch zu einer höheren Benutzerzufriedenheit?

Einwand Benutzerzufriedenheit

- Momentan ungeklärt
  - Evaluation: Bei mehreren Bild-Retrieval Systemen konnte kein Zusammenhang zwischen Präferenz und Qualität erkannt werden

    Maskari et al. 2006
  - Für einfache Suchaufgaben und kontrollierte Retrieval-Qualität ergab sich ebenfalls kein Effekt auf die Präferenz

    Turpin & Scholer 2006

Vorgehen IR-Studie

- Site-interne Suchsysteme
  - Website als wichtiges Kommunikationsinstrument
  - Entscheidet bei vielen Kontakten mit Benutzern bereits über Erfolg oder Misserfolg
  - Was der Benutzer nicht schnell findet, existiert für ich nicht

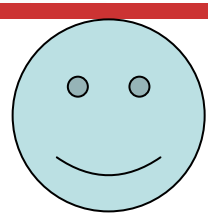
IR Psychologie

«The **unhappy customer**, on average, **will tell 27** other people ...»

«**Dissatisfied customers tell** an average of **ten** other people about their bad experience. Twelve percent tell up to twenty people.»

→ ***Bad news travels fast.***

IR Psychologie



On the other hand, satisfied customers will tell an average of *five* people about their positive experience.

→ ***Good news travels somewhat slower***

Eine Site-Suche sollte robust funktionieren

Ausreisser nach unten müssen vermieden werden!

- für jede Anfrage
- für jedes Bewertungskriterium

IR Psychologie

- Sehr gute Ergebnisse, schlechte Interaktion
  - -> Unzufriedener Benutzer
- Sehr gute Interaktion, schlechte Ergebnisse
  - -> Unzufriedener Benutzer

Vielen Dank für Ihre Aufmerksamkeit