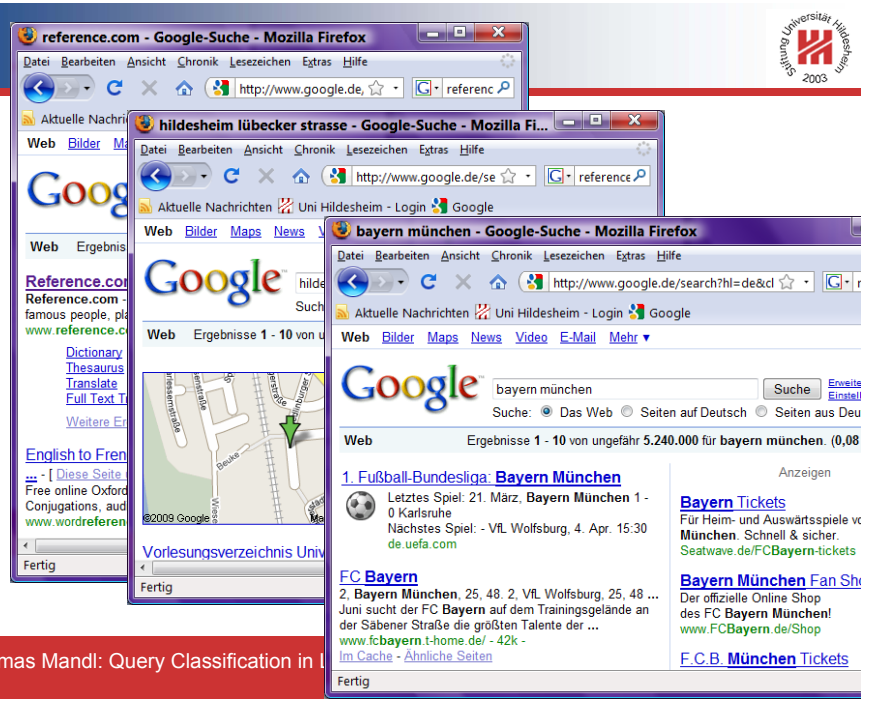
Thomas Mandl
Information Science
University of Hildesheim
mandl@uni-hildesheim.de

Query Classification in Logfile-Analysis: Evaluation Issues and User Satisfaction

Query Classification

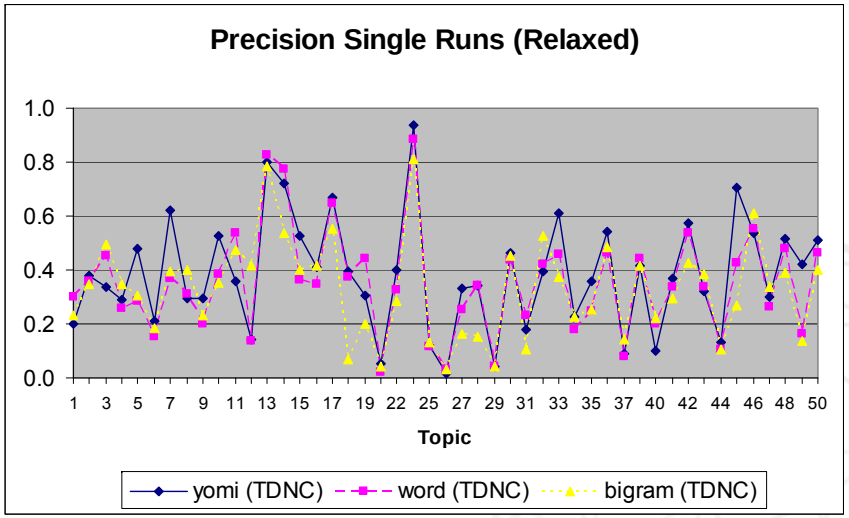
- To adapt the user interface (result display)
  - User behavior for navigational and informational queries differs
    - eye tracking study
    - URL or Snippet fixed for a longer time
    - *(Cutrell & Guan 2007)*
  - Geographical queries often require a map (Google)

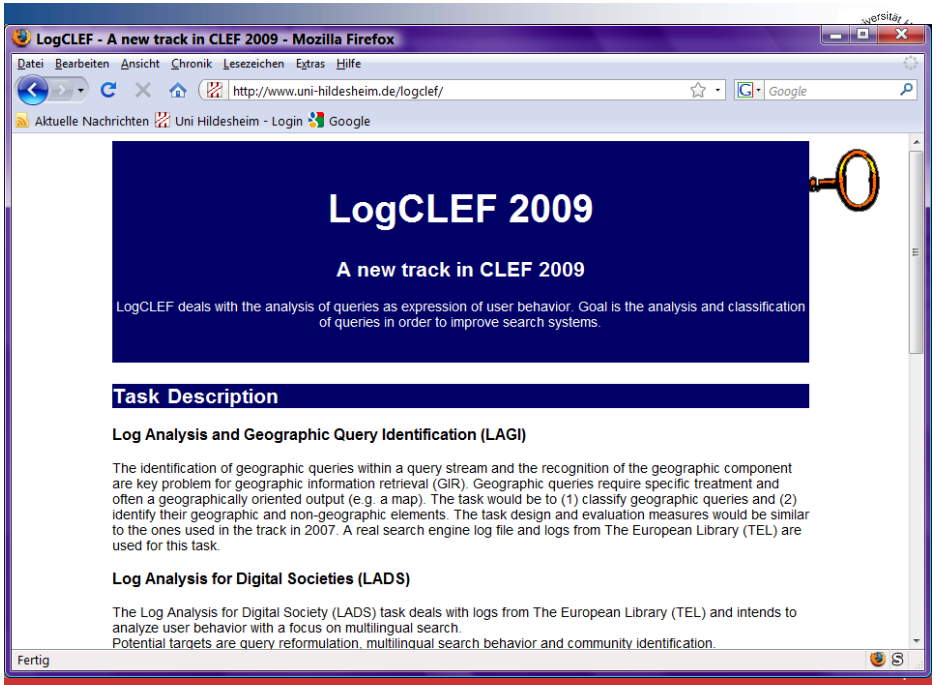


Query Classification

- Retrieval performance
  - „More work needs to be done on customizing methods for each topic“ (*Harman 2005*)
  - Geographic Queries may require geographic reasoning (GeoCLEF)
  - Difficult queries are solved well by several systems *(Hauff et al. 2009, Lease et al. 2009)*
  - Analysis of errors *(RIA Workshop, Savoy 2007)*

Example: NTCIR Experiments 2004





LogFile Analysis

- Geographic Query Classification and Parsing
  - Similar to GeoCLEF 2007 subtask
  - Find geo queries in a search engine log
  - Extract geo component
  - Analyzing options for log data
  - Support from MITRE Corp.
- Log Analysis in Digital Societies (LADS)
  - TEL query and activity logs
  - Analysis of user behavior
  - Permission for log release likely

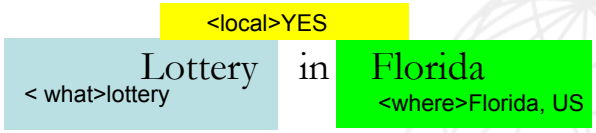
Data

- Query log from the MSN search engine
  - in English
  - 800.000 queries (collected August 2006)
  - 500 queries were labelled and used for evaluation
    - 100 queries for training
    - 400 for testing

Task 2009?

- Find queries with a geographic scope
  - Extract geo component

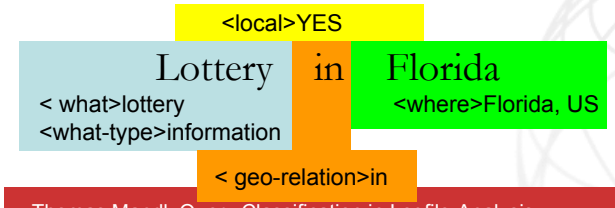
Example:



Task 2007

- Find queries with a geographic scope
  - Extract where component
  - Extract geo-relation-type
  - Extract what component
  - Classify what type {information, yellow page, map}

Example:



New Data

- LogFile from TUMBA: 350.652 lines of tumba! logs, anonymized and manually reviewed

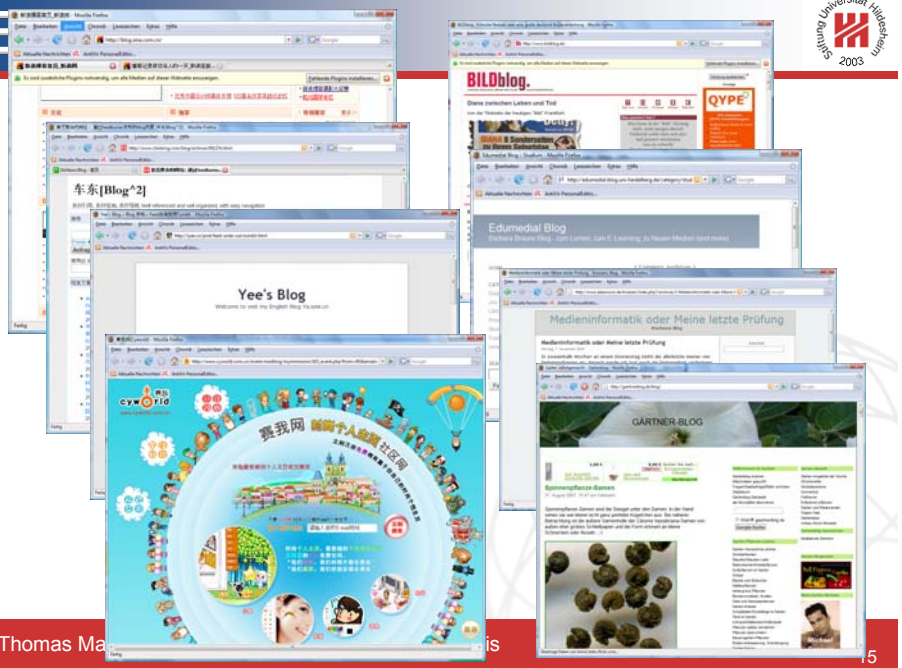
[Jan/2004] GET /pesquisa?terms=roswell+fanfiction HTTP/1.1
[Jan/2004] GET /pesquisa?site=ZZZ&terms=english HTTP/1.1 "ZZZ/index-down.html"
[Jan/2004] GET /pesquisa?terms=Tabela+Nacional+de+Incapacidades HTTP/1.1 "http://www.tumba.pt/"
[Jan/2004] GET /pesquisa?terms=sexo HTTP/1.1 "http://www.tumba.pt/"
[Jan/2004] GET /pesquisa?terms=desenvolvimento+motor HTTP/1.1 "ZZZ/"
[Jan/2004] GET /pesquisa?terms=termas+s.+gemil HTTP/1.1 "http://www.tumba.pt/"
[Jan/2004] GET /pesquisa?terms= HTTP/1.1
[Jan/2004] GET /pesquisa?terms=empreendimentos+tur%EDsticos HTTP/1.1
[Feb/2004] GET /pesquisa?terms=pesquisa+emprega HTTP/1.1 "http://www.tumba.pt/"
[Feb/2004] GET /pesquisa?site=ZZZ&terms=horarios+biologia HTTP/1.1 "ZZZ/„
[Mar/2004] GET /pesquisa?site=ZZZ&terms=ae HTTP/1.1 "ZZZ/index-down.html"
[Mar/2004] GET /pesquisa?site=ZZZ&terms= HTTP/1.1 "ZZZ"
[Mar/2004] GET /pesquisa?site=ZZZ&terms=ucdr HTTP/1.1 "ZZZ/49.htm"
[Mar/2004] GET /pesquisa?terms=educa%E7%E3o+pr%E9-escolar HTTP/1.1 "http://www.tumba.pt/"
[Mar/2004] GET /pesquisa?terms=anuncios+pessoais HTTP/1.1 "http://www.tumba.pt/"
[Mar/2004] GET /pesquisa?terms=sexo HTTP/1.1 "http://www.tumba.pt"
[Mar/2004] GET /pesquisa?terms=carvao+mineral HTTP/1.1
[Mar/2004] GET /pesquisa?site=ZZZ&terms=ZZZ@hotmail.com HTTP/1.1 "ZZZ/index-down.html"
[Mar/2004] GET /pesquisa?terms=wwwgov.pt HTTP/1.1

New Data

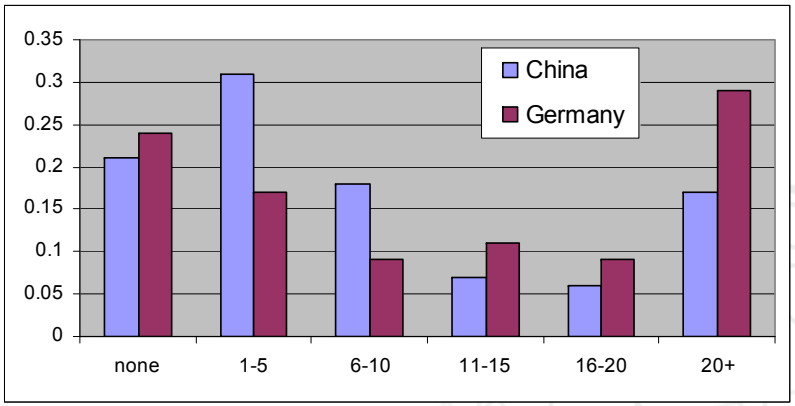
- LogFile from TUMBA: some classified and places annotated

[Nov/2003] GET /pesquisa?terms=viagem+nova+iorque HTTP/1.1 "http://www.tumba.pt/"
[Jan/2004] GET /pesquisa?terms=massey+ferguson+portugal HTTP/1.1
[Jan/2004] GET /pesquisa?terms=gastronomia+torres+vedras HTTP/1.1
[Feb/2004] GET /pesquisa?terms=origem+do+vidro+na+marinha+grande HTTP/1.1
[Feb/2004] GET /pesquisa?terms=%22serra+do+mar%E3o%22 HTTP/1.1 "http://www.tumba.pt/"
[Mar/2004] GET /pesquisa?site=ZZZ&terms=sistema+financeiro+portugu%EAs HTTP/1.1 "ZZZ"
[Mar/2004] GET /pesquisa?terms=peniche HTTP/1.1
[Mar/2004] GET /pesquisa?terms=serra+da+estrela HTTP/1.1 "http://www.tumba.pt/"
[Mar/2004] GET /pesquisa?terms=sites+brasileiros HTTP/1.1 "http://www.tumba.pt/"
[Oct/2003] GET /pesquisa?terms=jornais+de+leiria HTTP/1.1 "http://www.tumba.pt/" 1
[Oct/2003] GET /pesquisa?site=ZZZ&terms=mexico HTTP/1.1 "ZZZ/85.htm" 1

International Dimension

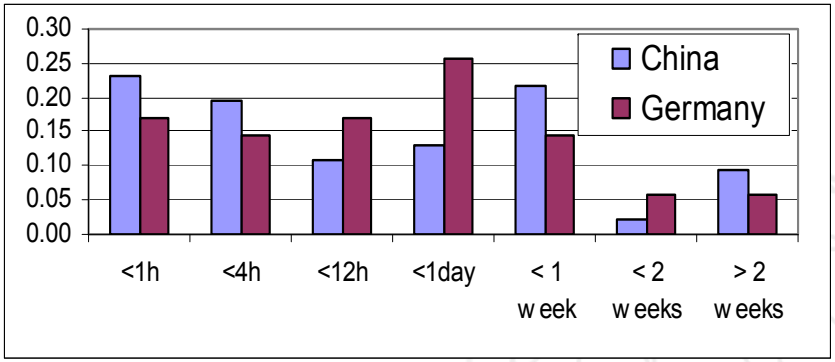


Results: High vs. low Context



- Germans use more external links

Results: Long-term Orientation



- More Chinese reactions after more than one week

- What about logs from social media?
- International / cultural aspects?

User Satisfaction

Satisfaction

- Selection of Queries from a Google Log
- Some test persons reconstructed the context and intended information need
- Other test persons performed searches
  - Judged relevance of top results
  - Reported their overall satisfaction
- A measure was found that correlated between satisfaction and relevance of top three hits
  - *(Huffmann & Hochster 2007)*

User Satisfaction

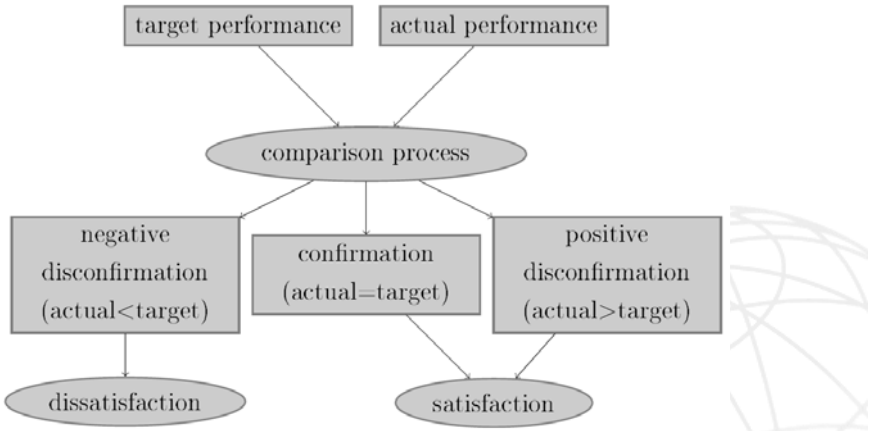
- Two tasks
  - In 5 minutes, as many relevant documents as possible should be found
- Analysis
  - Precision-oriented: time to find one single document
  - Number of documents found
- Differently performing systems are created artificially
  - 5 different MAP levels between 0.55 and 0.95
  - Results lists with 100 documents
  -

Turpin & Scholer 2006

Experiment Design

- Between groups
  - Two groups of users
  - Each test user sees only one system
  - Hard to achieve different satisfaction ratings
- Within groups
  - All users are presented all systems (usually both test conditions)
  - Order of presentation is balanced
  - Users can compare (nice for design studies)
  - Not realistic for IR (search twice?)

C/D paradigm



(Homburg et al.1999: 85)

User test

- Sample
  - 89 female subjects between ages 17 and 32 years
- Three CLEF-Topics
  - (10 min. per task)
  - Renewable energies
  - Castor transports in Germany
  - Child labour in Asia

Lamm & Mandl et al. 2009

Two factor design

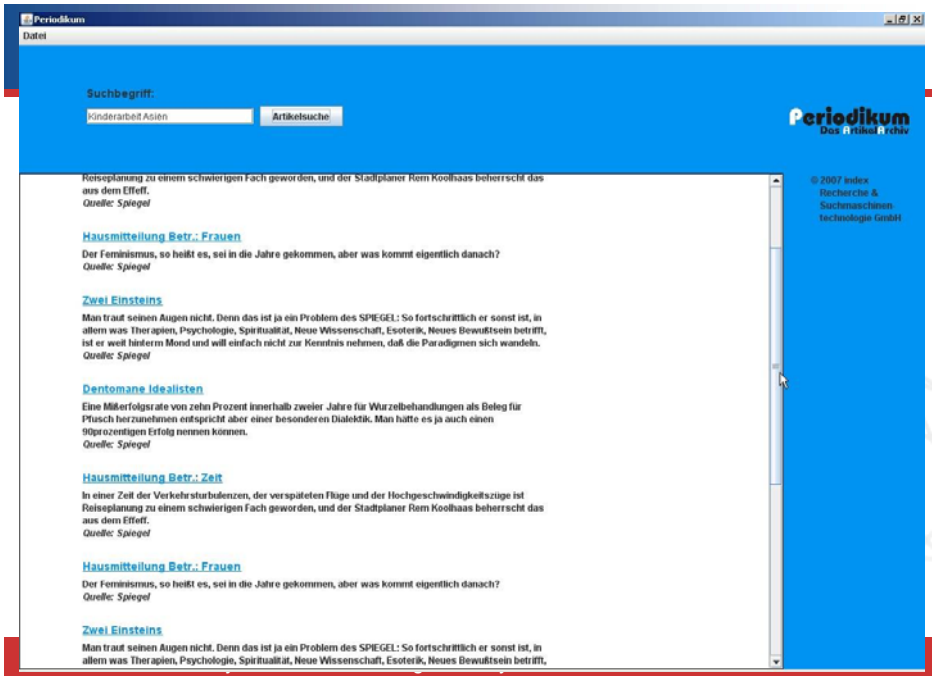
Expectations	System performance: low	System performance: high
high	20 subjects	20 subjects
low	20 subjects	20 subjects

Lamm & Mandl et al. 2009

Analysis

- User satisfaction
  - Questionnaire using a 7-point scale
- User performance
  - Effectiveness measures
    - Completeness (recall)
    - Accuracy (precision)
- Analysis of variance
  - Investigation of group differences

Lamm & Mandl et al. 2009



Effectiveness measures

- User precision

$$\frac{\text{Documents correctly identified as relevant}}{\text{Total number of relevant documents in list}}$$

- User recall

$$\frac{\text{Documents correctly identified as relevant}}{\text{Documents judged as relevant}}$$

Correctly means: Like the CLEF assessors

Results I

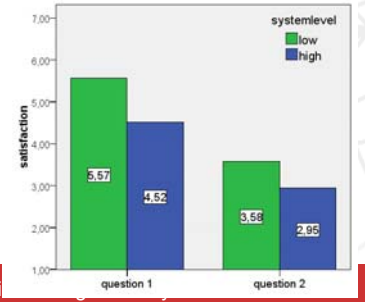
- Expectations
  - Expected influence, but not significant
  - Manipulation likely not sufficient
- System performance
  - Significant influence
  - User satisfaction
    - Precision oriented questions
  - User performance
    - Compensation for recall measures
    - Significant differences for precision measures

Results II

User satisfaction: indirect questions

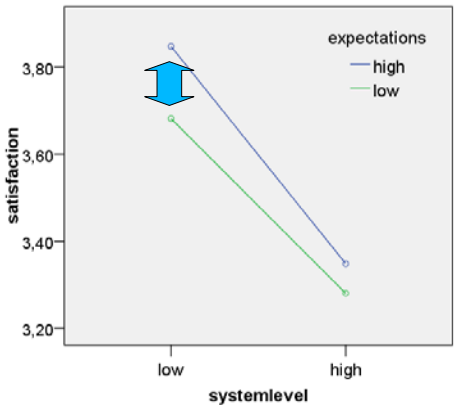
(1) „The filtering of articles could have been better.“

(2) „Most articles have been relevant with respect of the queries.“



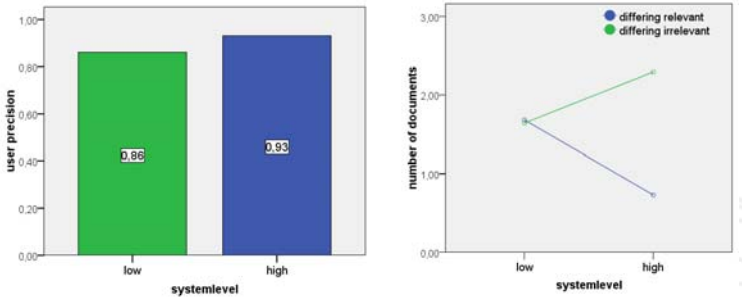
Results III

User satisfaction
C/D paradigm



Results IV

User performance: Assimilation effect



Percentage differences:

User precision: 8%

disagreement on relevant docs: 57%

disagreement on relevant docs irrelevant: 18%

- Results:
  - Requested satisfaction correlates significantly with system level
  - Expectation plays an estimated role, but not significantly
  - Precision-oriented measures higher for better system
  - Recall-oriented measures equal
  - -> User compensate
    - More documents are observed
    - Criteria for relevance are relaxed
    -

(Lamm, Mandl et al. 2009)

Compensation

- Relevance judgments are context dependent
  - Especially the number of relevant documents already encountered is important
  - Individual threshold might be lowered
  -

(Scholer & Turpin 2008)

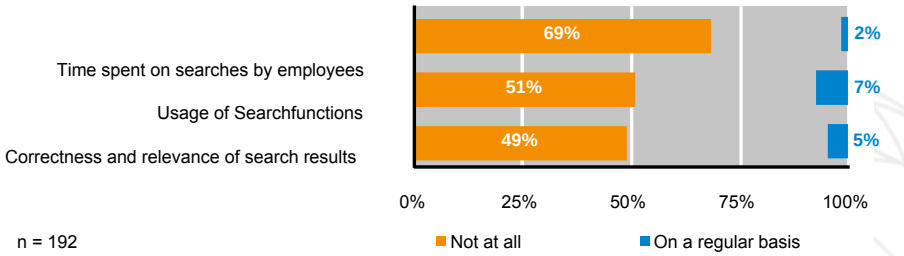
- Consistent with experience from relevance judgment for batch experiments

Scientific Challenge

- User Satisfaction and Log Analysis
  - How can we determine satisfaction from logs?
  - Logs as a basis for user experiments?

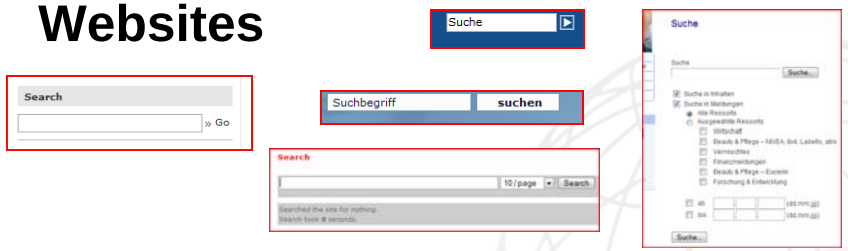
Practical Applications

- Do enterprises control search quality (by log analysis)?



Results from a survey among 192 decision makers from Swiss Enterprises

Evaluating the Search Applications of Industry and Public Websites



Credit to:
Martin Braschler, Ben Heuwing,Peter Schäuble, Jürg Stuker, Christa Womser-Hacker, Janko Zehe, Susanne Franke

Overview

- Consortium
- Goal: To Check Current State of Site Search
- Approach
  - Indicators
  - 35 websites were tested
- Results
  - Main Indicators
  - Google-Baseline
  - Interpretation

namics

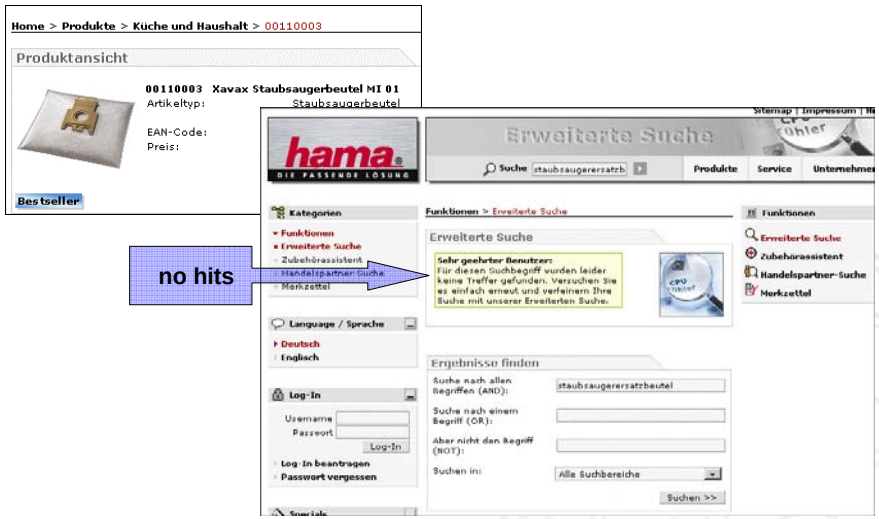
Zürcher Hochschule Winterthur
Institut für angewandte Informationstechnologie
InIT

eurospider
relevancy retrieval

HTW Chur
Hochschule für Technik und Wirtschaft

Fachhochschule Ostschweiz
University of Applied Sciences

Example



Search Index

- Completeness
  - All pages indexed?
- Freshness
  - New pages indexed?
  - Modified pages updated?
- Query / Document Analysis
  - Tokenization
  - Normalization
    - E.g. Jürg = JÜRG

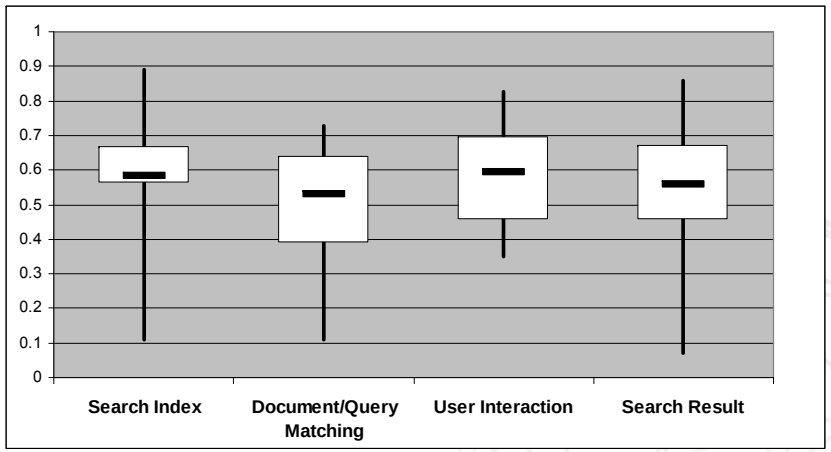
User Interaction

- Presentation of hit list
  - Snippet and other elements of a hit
  - Visual Coding
- User guidance
  - Interaction and functions
    - E.g. Enter starts search? Similar doc search?
  - Error handling, sub collections, zero hits, typos ...
- Performance

Search Result

- Navigation Tasks
  - E.g. Can the *Jobs* page be found quickly?
- Information Tasks
  - Depend on company
- Fact Queries (QA style)
  - E.g. Can the number of employees be found?

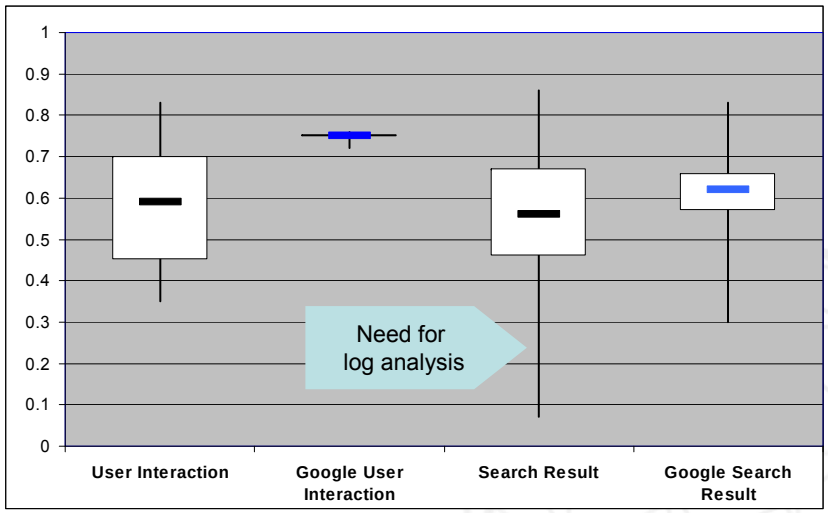
Results

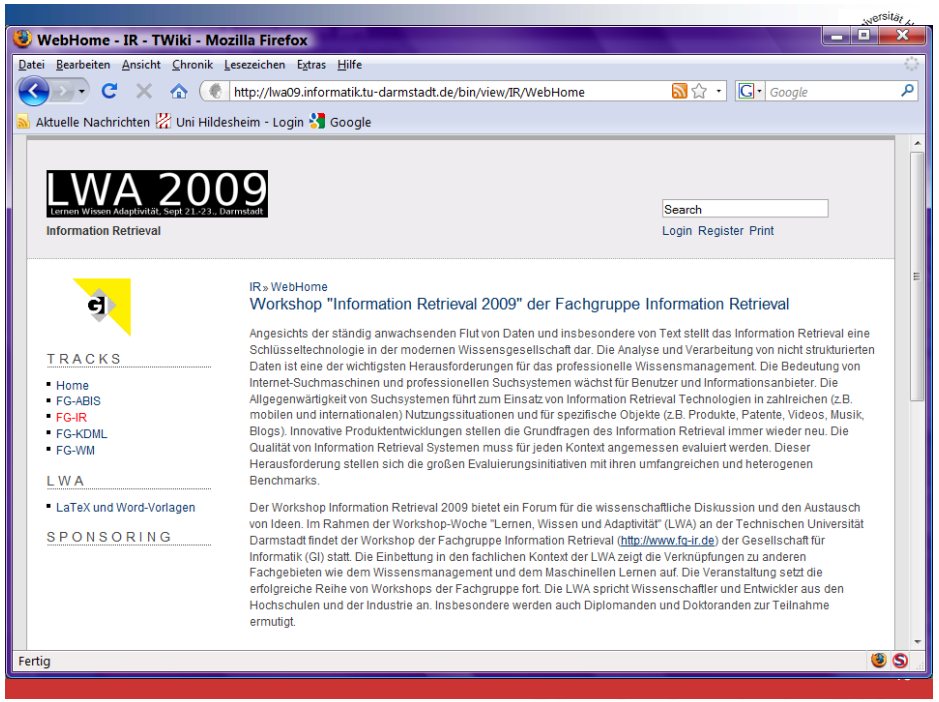


Google Baseline

- How well does Google with site search?
  - www.google.de -> site:www.xyz.de
- Approach
  - For half of the sites
  - search tasks were also done with Google
  - and evaluated

Google Baseline - Comparison





- Thank you for your
- attention!