Information Retrieval – Introduction and Survey

Norbert Fuhr

University of Duisburg-Essen

Germany

fuhr@uni-duisburg.de

What is Information Retrieval?

"Information Retrieval deals with *uncertainty* and *vagueness* in information systems"
(IR Specialist Group of German Informatics Society, 1991)

- *Uncertain* representations of the semantics of objects (text, images,…)
- *Vague* specifications of information needs (iterative querying)

What to Retrieve?

"Retrieve that amount of knowledge which a user needs in a specific situation for solving his/her current problem" (Kuhlen 1991)

- Consider specific user, situation and problem
  → *contextual retrieval*
- How to get this information
  → *global information access*

Workshop "Challenges in Information Retrieval and Language Modeling", 2002
http://ciir.cs.umass.edu/irchallenges/

Global information access

"Satisfy human information needs through natural, efficient interaction with an automated system that leverages world-wide structured and unstructured data in any language."

Information access

- Information properties
  - media
  - structure
  - heterogeneity
- Access methods

Information Media

- Text
- Facts
- 2D: graphics, images
- Speech
- Video
- 3D

Information structure

- Unstructured
- Semi-structured (XML)
- Fully structured
- Hyperlinked (Web)

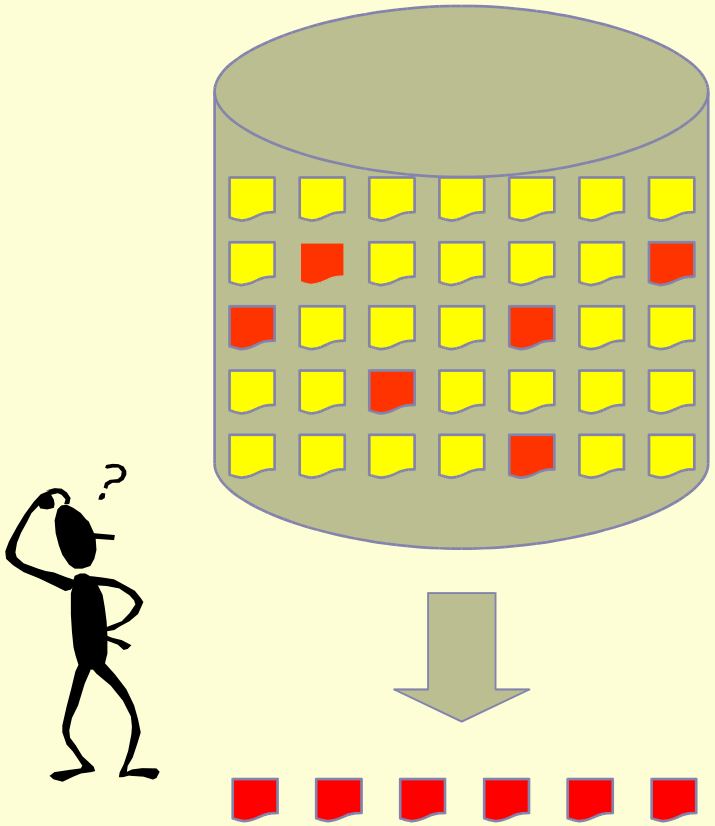
Heterogeneity

- Language: multilingual
- Media: multimedia
- Heterogeneous structures
- Heterogeneous services

Information Access Methods

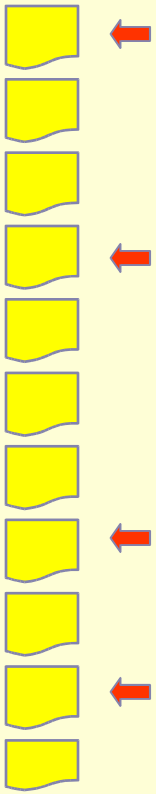
Ad-hoc retrieval

One time queries (e.g. Web search)



Filtering/Routing

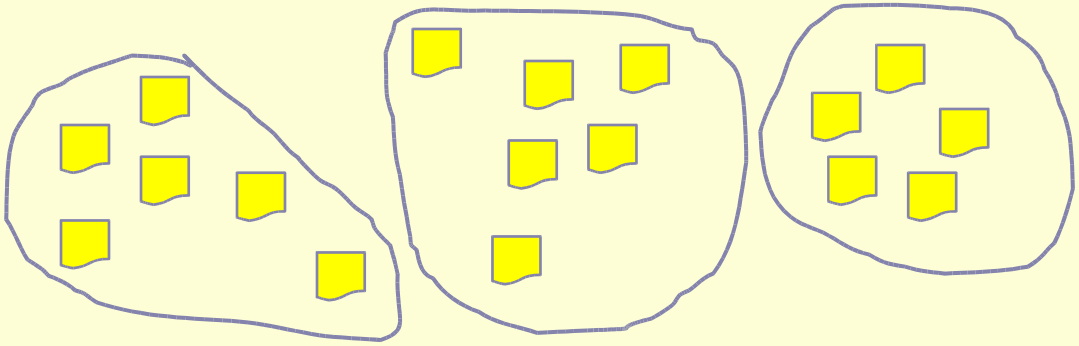
Constant search profile (e.g. Spam filtering)



Information Access (2):

- Categorization/Clustering:

  Group documents into predefined classes/ adaptive clusters



- Topic Detection and Tracking:

  Cluster news in stream

Information Access(3): Summarization

for browsing / survey on retrieval results

Sheffield Speech and Hearing Research Group - Speech ...
... We believe that **statistical methods** are well suited to this situation ... Content/style
models for non-extractive **summarization**;; Multi-**document summarization**;; ...
www.dcs.shef.ac.uk/spandh/projects/s3l/ - 4k - Cached - Similar pages

Text Interpretation: Extracting Information
... state approximation and other shallow but effective sentence processing **methods**, and (2) the emergence of weak heuristic and **statistical methods** that help to ...
cslu.cse.ogi.edu/HLTsurvey/ch7node5.html - 11k - Cached - Similar pages

Citations: Nonparametrics: **statistical methods** based on ranks - ...
... EL Lehmann. Nonparametrics: **Statistical Methods** based on Ranks. Holden-Day, S.Francisco, 1975. Columbia Multi-**Document Summarization**: Approach and.. ...
citeseer.nj.nec.com/context/224403/0 - 33k - Cached - Similar pages

Info. Access(4): Inform. Extraction

A.C Nielsen Co. said George Garrick, 40 years old, president of Information Resources Inc.'s London-based European Information Services operation, will become president and chief operating officer of Nielsen Marketing Research USA, a unit of Dun & Bradstreet Corp. He succeeds John Costello, who resigned in March

Inform. Access(5): Question answering

Find text passage answering fact query

Hoppe report 2000
... Shooting rate: 1-2/sec; recording **length**: 700 msec; surveying speed: 0.8 m/sec on **River Rhine**, 1.4 m/sec on Main and Neckar; daily survey **length**: 20-25 km; ...
comp1.geol.unibas.ch/report2000/report_2_1.htm - 25k - Cached - Similar pages

Bed Breakfast **Rhine** Hotel Im Malerwinkel Bacharach Middle **River** ...
Bed Breakfast **Rhine** Hotel Im Malerwinkel Bacharach Middle **River** Valley of ... Koblenz
Rhineland Bicycle Palatinate Rooms way **Length** Schlafzimmer rail ...
www.loreleytal.com/bacharach/im-malerwinkel/e/ - 8k - Cached - Similar pages

[PDF]202.1017 Fld Riwa **river** without
File Format: PDF/Adobe Acrobat - View as HTML
... m 3 /s **River** ID card name **Rhine** (Rhein, Rijn) origin Switzerland destination North
Sea character glacier **river** contents melt- and rainwater **length** 1,320 km ...
www.riwa.org/pdf.php?pdf=generalinfo.pdf - Similar pages

Current IR Research

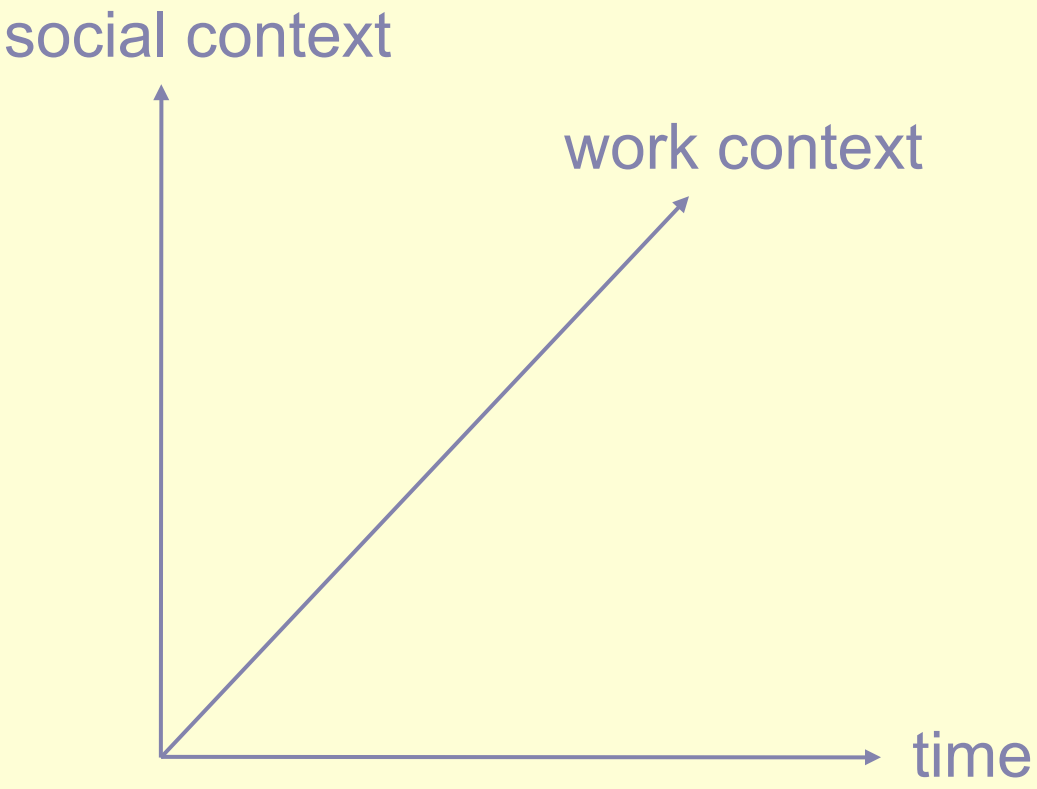
… focuses on models, methods and systems for information properties and access methods:

$$\{\{\text{Media}\} \times \{\text{Structure}\} \times \{\text{Heterogeneity}\}\} \times \{\text{Access methods}\}$$

Contextual retrieval

"Combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs."

Considering Context



Time-dependence

- Batch retrieval
- Constant information needs (Filtering $\rightarrow$ adaptation)
- Interactive retrieval
- Personalization:
  - Preferences
  - Seen items
  - Evolving interests

Interactive retrieval: Levels of search activities

1. **Move**: Low-level search function
   (e.g. type in search term, view retrieved document)
2. **Tactic**: several moves to further a search
   (e.g. broaden/narrow a query)
3. **Stratagem**: set of actions on a single domain
   (e.g. citation database, tables of contents of journals)
4. **Strategy**: complete plan for satisfying an information need
   (e.g. subject search, browse relevant journals, find referenced articles)

Interactive Retrieval: Current Research

- Evaluation results: quality differences between methods in batch retrieval vanish in interactive retrieval
- Empirical studies: information seeking as a sequence of interconnected but diverse searches
- Specific methods for interactive retrieval required:
  - information seeking: ‘berrypicking’
  - tactics & stratagems

Work context

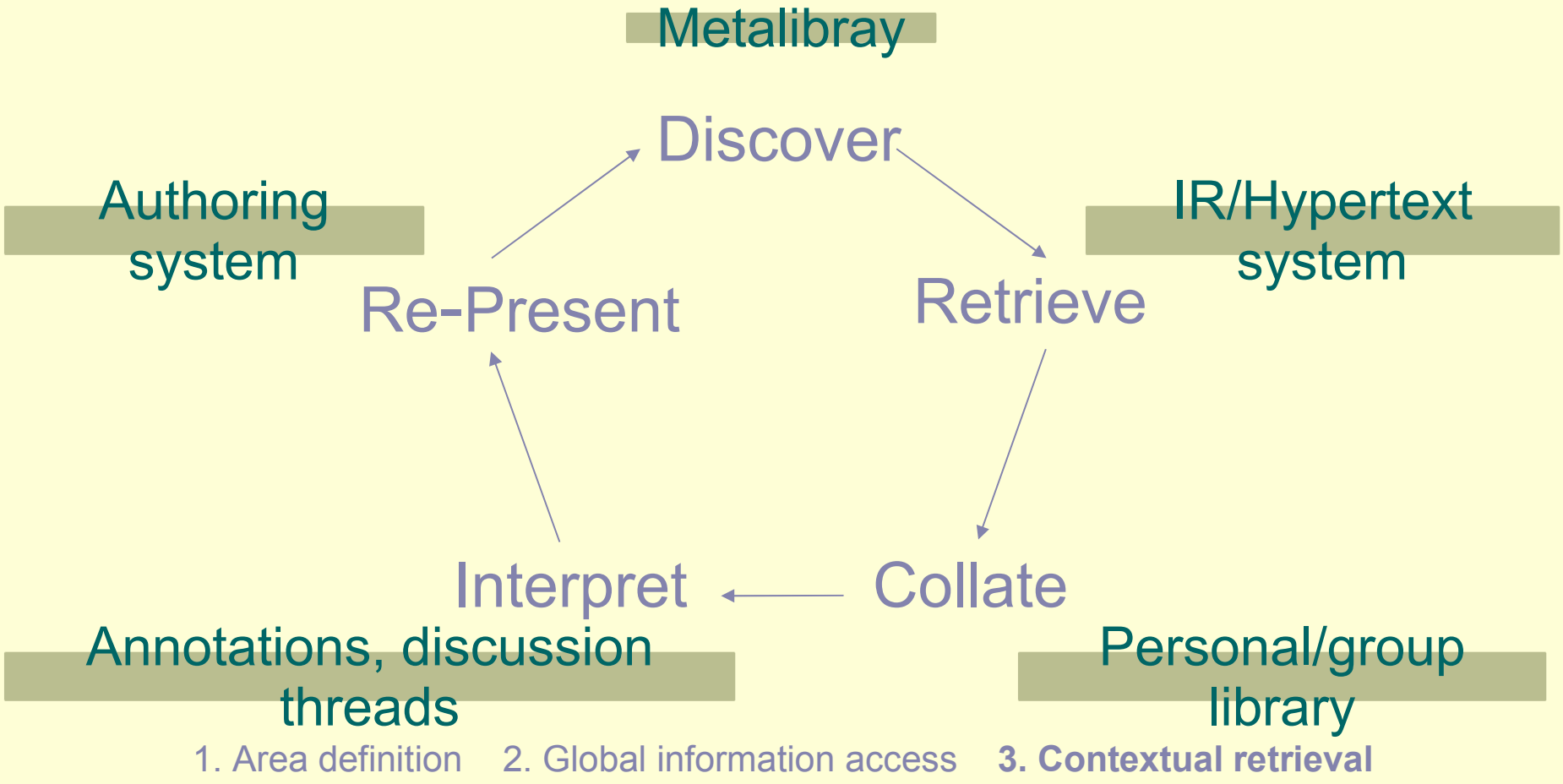
- Context-free
- Task-specific searches
- Workflow (application-specific)

Workflow: Generic problem solving scheme

1. Problem understanding (Hypermedia system with introductory/survey articles)
2. Identification of possible solutions (Hierarchical hypermedia system)
3. Selection of optimum solution (Information retrieval system)

→ integrated systems required

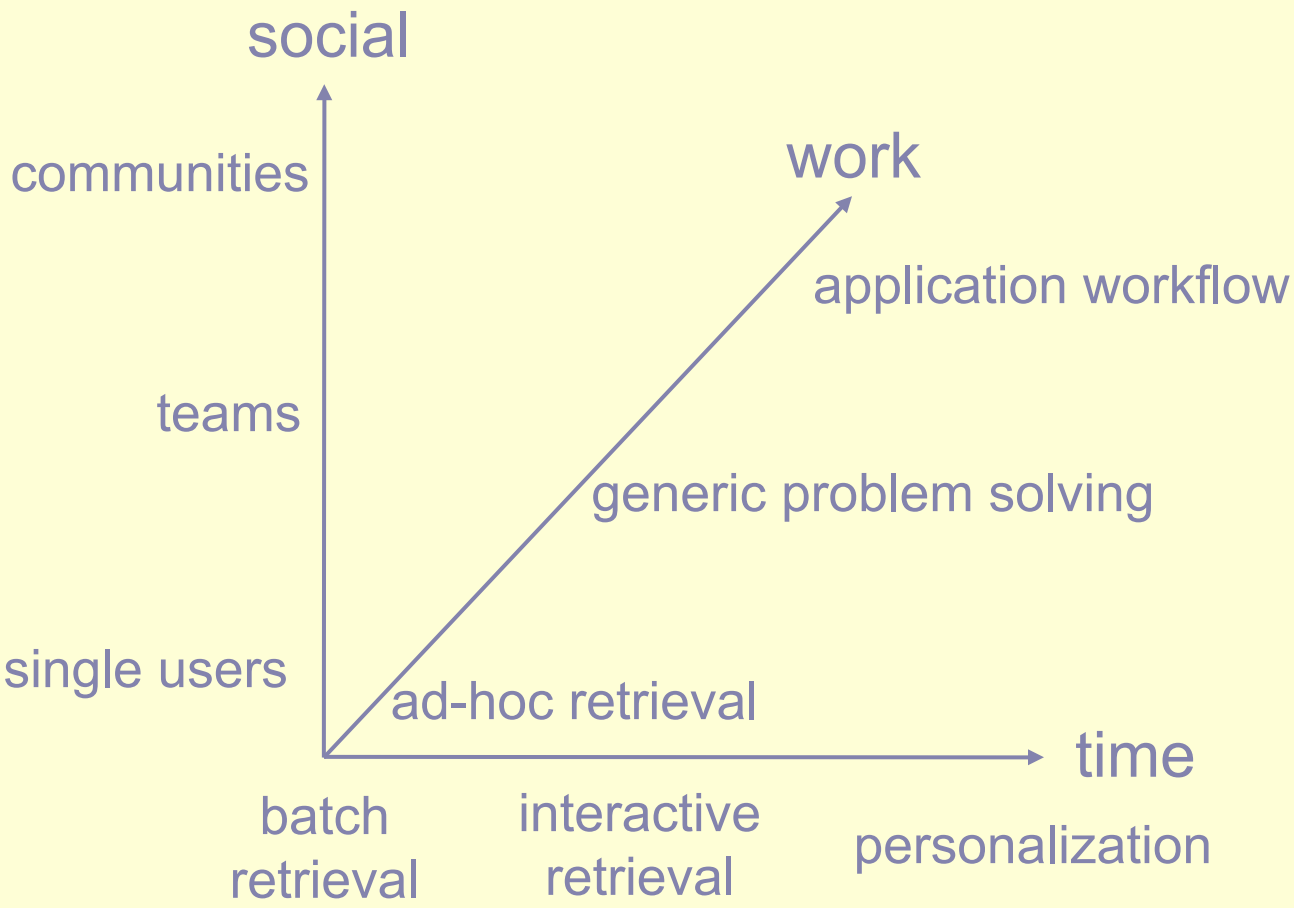
Workflow example: Digital Library Life Cycle



Social context

- Single user
- (Fixed) user groups
  - Collaborative information access
- (Open) communities

Context dimensions



Conclusion

- Global information access
  - Focus of current research
- Contextual retrieval
  - Promises significant quality improvements
  - More research necessary

Organisation

- Vorlesung: für Kommedia-Studenten (alte PO) nur bis Mitte des Semesters
- Übungen
  - freiwillig für Kommedia
  - verpflichtend für DAI

    „Sage es mir, und ich vergesse es;
    zeige es mir, und ich erinnere mich;
    lass' es mich tun, und ich behalte es.“ (Konfuzius)

Organisation(2)

- Prüfung/Leistungsnachweis:
  - Kommedia: zusammen mit 2. Informatik-Fach
  - DAI: Leistungskontrolle: mündlich, im September