General background

Driving forces – rapid evolution of:

- Computing power
- Memory
- Networking (internet)

- DB systems
- IR systems
- Hypertext (WWW)
- GUI/presentation tools (html)

Digital Libraries

- From Information retrieval to search engines
- e-books, e-libraries & related topics

Consequences:

Transformation of existing applications
Generation of new applications related to
data collection, organization, classification, access

Some examples:

(Classical) Libraries:

- Automation of catalogs (old stuff)
- On-line e-journals
- Collections of born-digital materials

New:

- Digitized collections (images, maps,..)
- on-line archives
- On-line, virtual museums

Portals, directories, search engines

- Yahoo – a directory (manual labor)
- Google – a search engine (fully automatic)
  IR technology & hypertext structure

- Amazon (& similar on-line sales companies)
  (book/dvd/.. Portal/directory)

Digital libraries & bibliographic services:

- ACM Digital Library acm-diglib
collection of all (full) papers from ACM journals
- SIGMOD digital anthology anthology
- DBLP dblp
collection of bibliographic information
- Citeseer citeseer
citation and impact factor data
Scientific data directories/repositories/portals:
- Bioinformatics: -- over 500 db's/data sources
- Astronomy – the world-wide telescope project
- Classical humanities – the Perseus digital library

Data and information services:
- Medline & US national library of medicine: nlmed
- Lexis-nexis (and many like it)

New kinds of services:
- Query subscription on XML/data streams
  - niagara, niagaracq
  - news
  - Stocks
  - satellite data

Issues:
- Heterogeneity → data transformation/integration
- Dependence on experiments → scientific experiment meta-data
- Huge volumes → fast, approximate, on-line stream processing

Library/IR basic concepts
Classification:
- Unique call number
- hierarchical & universal classification system
  - Dewey (1876)     LC (1900)

Axiom: unique position for a book
(+ secondary subjects)
- Unique main catalog
- Claim to universality

Issues:
- Fast streams
- Millions and more queries & subscribers
- Needs ultra-fast
  - stream processing
  - Query evaluation/data routing

A few more relevant buzzwords:
- Customer profiling
- Knowledge management
- E-learning
- E-publication
Operations in collection creation/maintenance:

Cataloging:/catalog
create metadata record for an item
(many style/spelling ... conventions used here)

Indexing: 
identify key terms (in all text/some fields)
• Controlled -- uses a fixed vocabulary
• Uncontrolled terms chosen by indexer

Abstraction: abstraction
create short description (of key ideas)

Metadata

Data describing a collection/item

Example: library bibliographic record for a book

See:

- Dublin Core & its use in stanford

Products:
- Bibliographic record db
- Author/title catalog
- Subject (header) catalog
  (provides entry points in on-line catalogs)

Currently, operations performed manually
• Expensive, slow, time-consuming
• Require experts
• Results are non-uniform (even with experts)

Automatic indexing & abstracting --- research areas

Technology/theory background

Structured data -- dbms
- Schema (structure) describes data precisely
- Queries & query language (based on structure)
- Indices, query optimization
  covered in DB course

Big challenge:
integration of multiple heterogeneous autonomous sources

Unstructured data -- (free) text:

Data = collection of texts
Query = set of terms (words)
Indices = inverted lists (words to locations)
Results = ranked answers

Issues/challenges:
• Answers are imprecise, approximate
• Difficult to evaluate answer goodness

Thesaurus:

A vocabulary of standard terms/concepts
• Hierarchical organization -- every term has
  - BT -- broader term, NT -- narrower terms
• Additional relationships
  - Synonyms, RT -- related terms

Used for:
• Indexing → standardization of index terms
• Querying → standardization of query terms
  (supposedly solves problem of non-uniform indexing)

Creation: complex, lengthy, community process
See also: ontology, KWIC (google them!)
Semi-structured data --- XML:
A standard for data exchange, also stored
covered in bsdi course
  • Self-describing data
  • Optional meta-data --- DTD/schemas
  • Validation tools
  • query language
  • Stream processing tools (under development)

Current move: extend to semantic web

Hypertext/www:
html, http, soap, ....
A browsing model
covered in bsdi course
Issues/ disadvantages:
  • No notion of query, just browsing
  • No structure on data
  • no data quality guarantees

Semi-structured data --- XML:
A standard for data exchange, also stored
covered in bsdi course
  • Self-describing data
  • Optional meta-data --- DTD/schemas
  • Validation tools
  • query language
  • Stream processing tools (under development)

Current move: extend to semantic web

The course
  • IR -- classical to Google
    - System architectures
    - Kinds of queries
    - Auxiliary data structures (indices) &
      efficient query processing
    - Compression, a bit of theory, uses in IR
    - Extensions to hypertext
      (using link structure)
  • "Conceptual" topics
    - E-books
    - E-publishing

Machine learning:
Used for automatic
  • Classification
  • Indexing
  • Abstracting
  • Clustering
Of free text/semi-structured data
covered in machine learning courses

End of technology survey

End of Introduction