INFSCI 2140
Information Storage and Retrieval
Lecture 1: Introduction

Peter Brusilovsky
http://www2.sis.pitt.edu/~peterb/2140-051/

INFSCI 2140 and your program

- Foundation course
- One of the key courses in any LS/IS program
- Information retrieval (IR) is one of the oldest and most developed areas of research in information science
- Hot research area with many crossroads
Business prospects for IR

- In the Web age IR became a part of any advanced computer application and started serving virtually any computer user
- Most of existing systems use simple old technologies. Advanced knowledge from IR field can improve the performance significantly

IR as a field of research

- American Society for Information Science (ASIS)
  - JASIS
  - Annual conference
- ACM SIGIR
  - IR conferences
- Information Processing and Management
- Information Retrieval
- Journal of Digital Information
Related fields

- Hypertext and Web
- Digital Libraries
- Multimedia (storage and retrieval)
- Data Mining
- Data Bases
- Knowledge Bases

The content of the course

- Storage and Retrieval
  - Two sides of the same coin
- Focus on *methods and technologies*
  - ...for computerized storage and retrieval of information in the form of documents
- Follows Korfhage book *plus*:
  - HCI (interaction and visualization)
  - Hypertext / hypermedia and WWW
  - User modeling / personalization
Course plan

- Introduction
- Classic Information Retrieval
- Improving Classic Information Retrieval
- What else in IR beyond Classic IR?
- Newest trends
  - User Modeling for IR
  - Web IR

Components of the grade

- Attendance (1 pt per lecture)
- Homework assignments
- Possible Quizzes
- Two Paper projects
- Midterm Exam
- Final group project
  - Will be announced at the next lecture
Paper project

- Read two papers
  - After 2000 and at least 8 pages long
  - On the Web or from recent conferences
- Provide an abstract while also trying to connect the paper to the course content
- Provide concept indexing:
  - What knowledge is required
  - What knowledge is communicated
- Present in public

Course Tools

- All information will be provided via course home pages (see first slide)
- Blackboard system will be used as a course tool for:
  - Posting course materials, assignments, and quizzes
  - Learning about and communication with each other
  - Asking questions and getting answers
  - Submitting assignments
  - Viewing grades and feedback
- Other tools will be used on later stages
Overview of Lecture 1

- Information systems - a design view
- Documents and queries
- Documents and surrogates
  - What’s in surrogates?
  - What’s in documents?

What is information

- Something that
  - is represented by a set of symbols
  - has some structure
  - can be read and to some extent understood by user of information

(Meadow)
What is Information Retrieval

- An information retrieval system is a device interposed between a potential user of information and the information collection itself. For a given information problem, the purpose of the system is to capture wanted items and to filter out unwanted items (Harter).

- Information retrieval deals with the representation, storage, and access to documents or representatives of documents (documents surrogates) (Salton).

Information System: Design View

- **Ectosystem** - factors that an IS designer can’t control
  - People involved
  - Available equipment and technology
  - The form in which information is available

- **Endosystem** - factors that an IS designer can specify, control, and manipulate
People in Ecosystem

- **User (community of users)**
  - The one who will be using the system (may need both to store or retrieve info)
- **Funder**
  - The one who bears the cost of operating
  - Has a global need in this system
- **Server**
  - The one who operates the IS and provides services

Example: S-T Info in Russia

- **Content**: unpublished project reports and Ph.D. thesis
- **Users**: Russian academics and researchers
- **Funder**: Russian Government
- **Server**: AUSTIC - a dedicated institution
Case Studies

- Pitt Library
- Amazon.com
- Google

More examples for homework
- Prepare analysis in a word file
- Submit via CourseWeb or e-mail
- Prepare to present in class

What we can control: Endosystem

- Media
  - Many forms from hardcopy to digital
- Representation
  - Storage formats, encoding, compression
- Devices
  - From file drawers to computers
- Algorithms and Data Structures
  - Maximize the usefulness of services
Design’s prospect

Information system

Ectosystem

Endosystem

Design

Performance and evaluation

- **User**
  - How effective the system is in helping me to satisfy my information needs

- **Server**
  - How efficient is the system
  - How well the system meets the needs of the user community

- **Funder**
  - Does the benefits justify costs
Life Cycle of an IS

- Design
- Work
- Evaluation
- Redesign
- Death

Information Objects

- The goal of information retrieval is to obtain information that is contained in one or more documents (information objects, information items)
- Examples
  - Good textbook for IS2140
  - Course for the Fall 2004
  - Fragment from Steelers last game
Information Need

- What kind of information you need to find - what you have in mind
- Example - Pension in Zurich
  - Relevant web sites will provide the user with necessary information and forms needed to actually make a reservation in a pension in Zurich.

Abstraction: Data

- Real world and its representation
  - Book -> library card -> database record
  - Course -> course syllabus -> Web page
- IS store information about real world as a collection of data abstracted from real world objects
- The amount and kind of stored information influence the search process
Real World and Abstraction

Reality

Data

Information Need

Information Access

Paradigms of Information Access

- Low interactive - query-based
  - Information Filtering and SID
  - Information Retrieval
- Highly interactive - behavior-based
  - Hypertext browsing
  - Information visualization
Abstraction: Query

- The user has an information need (IN) in her mind. It can be implicit or explicitly verbalized.
- The IS can’t understand the IN directly. IN has to be abstracted into a form that matches the information system.
- This abstraction is a query.

Real World and Abstraction

```
Reality ---------> Data
    |                |
    v
Information Need ---------> Query
```
Classic paradigm of information retrieval (ad-hoc retrieval)

- Set of documents
  - Kegels
- User comes with a query
  - A ball
- IR returns some documents in response to a query
  - Bowling model

What is document?

- Any object that can be stored
- Granularity
  - Book
  - Chapter
- Types
  - Programs, images, music, comp. programs
What is a query?

- Statement of an information need
- (Formal?) representation of an information need
  - Request to a librarian
  - IN described in NL
  - “Like that”
- Is a query a document?

Documents

- Document is stored and can be retrieved
- Variety of documents
  - Temporal
    - Ephemeric documents
    - Changing docs
  - Media
    - Real object: books, CDs, vine bottles
    - Digital objects: text, music, pictures, movies...
Formatting aspect

- Formatted and unformatted
- Mixture
- Metadata

Document surrogates

- Stored description of a document to be used for retrieval and presentation
- Surrogates are incomplete by its nature
  - Can't store all document: space, nature, design choice
- How to produce surrogates
  - By humans - rules, practice
  - By computers - programs
Documents and surrogates

- Metadata, Content data
  - Digitally stored, used for search, presentation, and selection

- Digital Document
  - Digitally stored, used for presentation and selection, not used for search

- Externally stored document / object
  - Externally stored, not used for search

Documents and surrogates (2)

Documents → Surrogates

Information Need ↔ Query
Examples of surrogates

- Document ID (system use!)
- Metadata: author, title, year
- Content representation
  - Keywords
  - Abstract / Extract

Case Study: Photo archive

- Photos are stored, but are not searchable
- Searchable are *descriptions*
- Description: what, when, where
  - Content (abstract vs. classifier)
  - Time (granularity!)
  - Location (coding scheme vs words)
Case Study: PhD Thesis

Case Study: Movie Rental Store
What’s in a surrogate?

- Metadata (about it)
  - Usually well-formalized, stored in a formatted database

- Content description
  - Rarely formalized, non-formatted storage
  - Keywords, terms…
  - Full-text abstract/extract
  - Restricted on unrestricted vocabulary

Unrestricted vs. restricted NL

- What could be use to describe the content (abstracts, keywords, terms, classifiers…)

- Controlled vocabulary
  - Words/terms to describe document content only can come from this vocabulary

- Unrestricted vocabulary
  - Any NL sentences /phrases can be used
Controlled vocabulary

- Effectiveness of the overall system - storage and search
- Reliability and precision of search
  ...but...
- Overhead
- Hard to force users and info providers
- Need thesaurus
- Loosing fine elements of meaning

Unrestricted NL

- More complicated logistics, slower search, limited search options
- Lower reliability and precision of search
  ...but...
- Almost no no overhead for humans
- No thesaurus
- Can express any meaning
What’s in a digital document?

- Digital documents
  - Text
  - Rich text and hypertext
  - Images
  - Multimedia
  - Compound documents
- From real to digital document
  - Digitize (code) and store

Digital Representation

- Computer store information digitally in binary format
- Ultimately everything is ones and zeroes
  - characters, numbers
  - e-mail messages, poems
  - pictures, video, music
- Binary coding:
  - $7 = 00000111$, ‘!’ $= 000100011$;
Representing text

- Coding problem
- Replace every symbol for a 1 byte code
  - Meaningful symbols and control symbols
- Problems of different coding - same symbol has multiple codings
- Standards: ASCII, ANSI, KOI-8…
- Too many coding standards, 1 byte??
- Universal standards, unicode

Rich text and hypertext

- Rich text - fonts, formatting, styles…
  - NROFF/TROFF, TeX/LaTeX
  - proprietary word processor, RTF
  - HTML
- Hypertext - links, anchors…
  - System-dependent way
  - No standard yet, open hypermedia
  - HTML’s HREF tag
Representing images

- Image as a matrix of dots (pixels)
- How many bytes per pixel?
  - 1/2 for 16 colors
  - 1 for 256
  - 2 for $2^{16}$
- An image consumes lots of space

Image formats

- Simple bitmap formats
  - BMP, Pict
- Publishing software formats
  - Photoshop, Canvas...
- Complex formats for BW/Color images
  - JBIG, TIFF, GIF, JPEG
Music and multimedia

- **Sound:**
  - Digitized music: aiff, wav, … MP3
  - Encoded music: MIDI

- **Multimedia = moving pictures + sound**
  - QuickTime, WPM…
  - MPEG standard
  - SMILE Web standard

- **Streaming vs. non-streaming**

Compression issues

- **Storage space vs. access time**
  - Uncompressed surrogates, compressed documents

- **Classic text compression**
  - Huffman codes; Ziv-Lempel codes

- **Image and multimedia compression**
  - .gif vs .jpeg

- **Loss of information in encoding and compression**
Assignment 1

- Due Thursday 9/9
  - Try various features
  - Home page with picture (4pt)
  - Submit an answer to Korfhage search problem to the discussion forum, get ready to present in class (up to 2pt)
  - Submit a Word (or ASCII) file with an analyzed example of an information system (via dropbox) (3pt)