Structure of IR Systems

INST 734
Lecture 1, January 29, 2014
Jiaul Paik
Agenda

• Introductions and Motivation

• The Structure of Interactive IR systems

• Course Overview
Some Holistic Definitions of IR

• A *problem-oriented* discipline, concerned with the problem of the effective and efficient transfer of desired information between human generator and human user.

  Anomalous States of Knowledge as a Basis for Information Retrieval. (1980)

• A process for establishing a view on an information space from a perspective defined by the user.

  Douglas W. Oard, UMD iSchool Professor
Information Retrieval Systems

• Information
  – What is “information”?

• Retrieval
  – What do we mean by “retrieval”?
  – What are different types of information needs?

• Systems
  – How do computer systems fit into the human information seeking process?
Information
What do We Mean by “Information?”

• How is it different from “data”?
  – Information is **data in context**
    • Databases contain data and produce information
    • IR systems contain and provide information

• How is it different from “knowledge”?
  – Knowledge is a **basis for making decisions**
    • Many “knowledge bases” contain decision rules
An Example

• Data
  – 98.6°F, 99.5°F, 100.3°F, 101°F, ...

• Information
  – Hourly body temperature: 98.6°F, 99.5°F, 100.3°F, 101°F, ...

• Knowledge
  – If you have a temperature above 100°F, you most likely have a fever

• Wisdom
  – If you don’t feel well, go see a doctor
Information Hierarchy

• Data
  – The raw material of information

• Information
  – Data organized and presented in a particular manner

• Knowledge
  – “Justified true belief”
  – Information that can be acted upon

• Wisdom
  – Distilled and integrated knowledge
  – Demonstrative of high-level “understanding”
Information Hierarchy

Data

Information

Knowledge

Wisdom

More refined and abstract
## Databases vs. IR

<table>
<thead>
<tr>
<th></th>
<th>Databases</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What we’re retrieving</strong></td>
<td>Structured data. Clear semantics based on a formal model.</td>
<td>Mostly unstructured. Free text with some metadata.</td>
</tr>
<tr>
<td><strong>Queries we’re posing</strong></td>
<td>Formally (mathematically) defined queries. Unambiguous.</td>
<td>Vague, imprecise information needs (often expressed in natural language).</td>
</tr>
<tr>
<td><strong>Results we get</strong></td>
<td>Exact. Always correct in a formal sense.</td>
<td>Sometimes relevant, often not.</td>
</tr>
<tr>
<td><strong>Interaction with system</strong></td>
<td>One-shot queries.</td>
<td>Interaction is important.</td>
</tr>
<tr>
<td><strong>Other issues</strong></td>
<td>Concurrency, recovery, atomicity are all critical.</td>
<td>Issues downplayed.</td>
</tr>
</tbody>
</table>
Types of information

• Unstructured text (e.g., web pages)
• Structured documents (e.g., XML)
• Images
• Audio (sound effects, songs, etc.)
• Video
• Services
• Codes
Systems
Design Strategies

• Foster human-machine synergy
  – Exploit complementary strengths
  – Accommodate shared weaknesses

• Divide-and-conquer
  – Divide task into stages with well-defined interfaces
  – Continue dividing until problems are easily solved

• Co-design related components
  – Iterative process of joint optimization
Human-Machine Synergy

• Machines are good at:
  – Doing simple things accurately and quickly
  – Scaling to larger collections in sublinear time

• People are better at:
  – Accurately recognizing what they are looking for
  – Evaluating intangibles such as “quality”

• Both are pretty bad at:
  – Mapping consistently between words and concepts
Retrieval
Types of Information Needs

• Retrospective ("Retrieval")
  – “Searching the past”
  – Different queries posed against a static collection
  – Time invariant

• Prospective ("Recommendation")
  – “Searching the future”
  – Static query posed against a dynamic collection
  – Time dependent
What Do We Mean by “Retrieval?”

• Find something that you want
  – The information need may or may not be explicit

• Known item search
  – Find the class home page

• Answer seeking
  – Is Lexington or Louisville the capital of Kentucky?
Relevance

• **Relevance** relates a *topic* and a document
  – Duplicates are equally relevant, by definition
  – Constant over time and across users

• **Pertinence** relates a *task* and a document
  – Accounts for quality, complexity, language, …

• **Utility** relates a *user* and a document
  – Accounts for prior knowledge
Iterative Search

• Searchers often don’t clearly understand
  – The problem they are trying to solve
  – What information is needed to solve the problem
  – How to ask for that information

• The query results from a clarification process
Structure
Divide and Conquer

• Strategy: use *encapsulation* to limit complexity

• Approach:
  – Define *interfaces* (input and output) for each component
  – Define the *functions* performed by each component
  – Build each component (in isolation)
  – See how well each component works
    • Then redefine interfaces to exploit strengths / cover weakness
  – See how well it all works together
    • Then refine the design to account for unanticipated interactions

• Result: a hierarchical decomposition
Supporting the Search Process

1. Source Selection
2. Query Formulation
3. Search
4. Selection
5. Examination
6. Delivery
Supporting the Search Process

1. **Source Selection**
   - IR System

2. **Query Formulation**
   - Query

3. **Search**
   - Index
   - Ranked List

4. **Selection**
   - Document

5. **Examination**
   - Document

6. **Acquisition**
   - Collection

7. **Delivery**
The IR Black Box
Inside The IR Black Box

Query Representation Function

Query Representation

Ranking Function

Index

Document Representation Function

Document Representation

Hits
Representation
Counting Terms

• Terms tell us about documents
  – If “dog” appears a lot, it may be about dogs

• Documents tell us about terms
  – “the” is in every document -- not discriminating

• Documents are most likely described well by rare terms that occur in them frequently
  – Higher “term frequency” is stronger evidence
  – Low “document frequency” makes it stronger still
“Bag of Terms” Representation

- Bag = a “set” that can contain duplicates
  ➢ “The quick brown fox jumped over the lazy dog’s back” → \{back, brown, dog, fox, jump, lazy, over, quick, the, the\}

- Vector = values recorded in any consistent order
  ➢ \{back, brown, dog, fox, jump, lazy, over, quick, the, the\} → [1 1 1 1 1 1 1 2]
Bag of Terms Example

Document 1
The quick brown fox jumped over the lazy dog’s back.

Document 2
Now is the time for all good men to come to the aid of their party.

<table>
<thead>
<tr>
<th>Term</th>
<th>Document 1</th>
<th>Document 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>aid</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>all</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>back</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>brown</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>come</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>dog</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>fox</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>good</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>jump</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>lazy</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>men</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>now</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>over</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>party</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>quick</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>their</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Stopword List
- for
- is
- of
- the
- to
Learning From Linking Behavior

Diagram:
- Hub
- Authority

Nodes and edges represent the relationships between hubs and authorities.
Indexing

• Organizing documents
  – To facilitate faster retrieval

• Some relevant data structures
  – Inverted index
  – Hashing
  – Search trees

• Index construction algorithms
  – Static
  – Dynamic
  – Distributed
Retrieving
Types of Retrieval

• Boolean
  – Find documents that have “Maryland” and “University”

• Ranked
  – Order documents based on some score
Ranking

• Term weighting
  – How important a term for a document/object of interest?

• Weight combination
  – How to combine individual weights to produce composite score (the basis of ranking)
Types of Evidence
Some common evidences

• Content
  – Words in a document

• Metadata
  – Image tags

• User behavior data
  – Number of clicks on a document
Structure of the Course
Course Goals

• Appreciate IR system capabilities and limitations

• Understand IR system design & implementation
  – For a broad range of applications and media

• Evaluate IR system performance

• Identify current IR research problems
Course Design

• Readings provide background and detail

• Class provides organization and direction
  – We will not cover every detail

• Homework and project provide experience

• Final exam helps focus your effort
Assumed Background

• Knowledge of programming
• Basic scripting language
• Basic math (e.g.,) (don't worry!)
  1. probability
  2. matrix algebra
  3. concept of vector
Grading

• Homework (20%)
  – Mastery of concepts and experience using tools

• Term project (50%)
  – Details will be given on course Web page

• Final exam (30%)
  – In-class exam (open book)
Important Things to Know

• Office hours: 5 PM Thursdays
  – Or schedule by email, or ask after class

• Everything is on the Web
  – http://www.umiacs.umd.edu/~jiaul/teaching/ir

• Easiest way to reach me by email
  – jia.paik@gmail.com
Some Things to Do This Week

Homeworks
  – Will be given on Mondays based on the previous lecture
  – Due at 6 PM next Monday!!

• Do the reading **before** class
  – Read for ideas, not detail

• Explore the Web site
  – Start thinking about the term project