Information Retrieval And Storage

Jianping Zhang

Spring 2002
Simple View of IR Systems

- **Given:**
  - a collection of documents
  - a user information need (query)

- **Find:**
  - a subset of documents relevant to the query
Information

• Text

• Images

• Audio

• Video
IR Issues

• Storage
  How information should be represented, organized, and stored.

• Retrieval
  How user information needs are satisfied.
Information Retrieval vs Data Retrieval

- **Information Retrieval**
  - unstructured data
  - inexact

- **Data Retrieval**
  - structured data
  - exact
Motivations

• **Library**
  access to books, journals, and other documents

• **Web**
  access to distributed online multimedia documents in an open environment

• **Digital Library**
  access to distributed online multimedia documents in a controlled environment
Web Impacts

- Distributed Documents
- Document Quality
- Scalability
- Links
- Users with different backgrounds
User Tasks

• User Pull
  – Retrieval
  – Browsing

• System Push
  – Filtering
  – Routing
IR History

- Binary Retrieval
- Ranked Retrieval
- Concept-Based Retrieval
- Information Extraction And Summarization
- Distributed Retrieval
- Text Mining
- Multimedia Retrieval
Language Technologies

- Information Retrieval
- Question Answering
- Information Extraction And Summarization
- Information Filtering
- Document Clustering And Categorization
- Text Mining
- Topic Detection And Tracking
- Machine Translation
- OCR And Speech Recognition
Language Technologies

- Statistical
- Natural Language Understanding
IR Processes

- **Indexing**
  represent and organize documents for efficient and effective retrieval

- **Query Formulation**
  transform user queries to some internal representations

- **Retrieval**
  match user query with documents

- **Retrieval Evaluation**
  displayed retrieval documents and modify query based on user feedback
Logic View of Documents

- Index Terms
- Full text
- Representative Keywords
- Concepts And Topics
- Stop words
- Stemming
- Content
- Segments
- Noun Groups
IR System Structure
Topics Covered

- Modeling
- Evaluation
- Retrieval Algorithm
- Indexing Algorithm
- Web Search Engine
- Text Categorization And Clustering
- Topic Detection And Tracking
- Text Mining
- Information Agent and filtering and routing
IR Models

General Approach to IR and determine indexing and retrieval algorithms

- Binary Model
- Vector Model
- Probabilistic Model
- Other Models
Evaluation

• **Efficiency**
  – storage
  – response time

• **Effectiveness**
  – recall
  – precision
Indexing Algorithm

- Text Operations
- Text Representations
- Indexing
Retrieval Algorithms

- Query Operations
- Searching
Web Search Engine

• Gatherer
• Indexer
• Searcher
• Retriever
Text Categorization And Clustering

- **Categorization**
  - automatically assign documents to different categorization
  - learn classification rules

- **Clustering**
  automatically generate groups (clusters) of documents
Topic Detection And Tracking

- detection of topics of documents
- new event detection
- new information detection
Text Mining

• Data Mining
  Discover useful pattern from data for example credit fraud use detection

• Text Mining
  Discover useful pattern from data for example detection of aviation safety problems from aviation safety documents
Information Filtering And Routing

- information retrieval agents
- documents and email routing
- user profiling and machine learning
Intelligent Agents

- **Autonomous**
  Agents have the ability to pursue a goal, and operate largely without direct control from humans, or from other agents.

- **Social (Communicating)**
  Agents often work collaboratively.

- **Adaptable (Reactive)**
  Agents know about and respond to things going on in their environment.

- **Proactive**
  Agents take initiative in order to get closer to their defined goals.