Course Overview

闫宏飞
北京大学，信息科学技术学院，计算机系
http://net.pku.edu.cn/~yhf, yhf AT net.pku.edu.cn

Thanks to Chengxiang Zhai for sharing many of these slides.
Course Objectives

• Introduce information retrieval (IR) and Search Engine (SE)
  – Foundation: Basic concepts, principles, methods, etc
  – Trends: Frontier topics

• Prepare students to do research in IR and/or related fields
Course Management

- Course website: http://net.pku.edu.cn/~course/2010-ccf-adl
- Course group discussion: http://groups.google.com.hk/group/cs410pku
- Questions: Post the questions on the group discussion forum
What is Text Info. Management?

• TIM is concerned with technologies for managing and exploiting text information effectively and efficiently

• Importance of managing text information
  – The most natural way of encoding knowledge
    • Think about scientific literature
  – The most common type of information
    • How much textual information do you produce and consume every day?
  – The most basic form of information
    • It can be used to describe other media of information
  – The most useful form of information!
Text Management Applications

Access

Select information

Organization

Add Structure/Annotations

Mining

Create Knowledge
Examples of Text Management Applications

• Search
  – Web search engines (Google, Yahoo, ...)
  – Library systems

• Recommendation
  – News filter
  – Literature/movie recommender

• Categorization
  – Automatically sorting emails

• Mining/Extraction
  – Discovering major complaints from email in customer service
  – Business intelligence
  – Bioinformatics

• Many others...
Elements of Text Info Management Technologies

Focus of the course

Retrieval Applications

Information Access

Mining Applications

Knowledge Acquisition

Search

Filtering

Summarization

Categorization

Visualization

Mining

Extraction

Clustering

Natural Language Content Analysis

Text
Text Management and Other Areas

Human-computer interaction
Software engineering
Web
Probabilistic inference
Machine learning
Natural language processing
Storage
Compression

TM Algorithms

Text

User
Related Areas

- Information Retrieval
- Databases
- Info Science
- Software engineering
- Computer systems
- Applications (Web, Bioinformatics...)
- Natural Language Processing
- Pattern Recognition
- Data Mining
- Machine Learning
- Models
- Statistics
- Optimization
- Algorithms
- Systems
Publications/Societies (Incomplete)

- **Learning/Mining**
  - ICML
  - ICML, NIPS, UAI
  - ACM SIGKDD
  - AAAI
  - ACL
  - COLING, EMNLP, ANLP

- **Info Retrieval**
  - ACM SIGIR

- **Applications**
  - WWW
  - RECOMB, PSB
  - TREC

- **Info. Science**
  - JCDL
  - ASIS

- **Databases**
  - ACM SIGMOD
  - VLDB, PODS, ICDE

- **Statistics**

- **NLP**
  - HLT
  - COLING, EMNLP, ANLP

- **Software/systems**
  - SOSP
  - OSDI

- **ACM CIKM, Databases**

- **ACM CIKM, Databases**

- **WWW, RECOMB, PSB**

- **TREC, SOSP, OSDI**

- **JCDL, ASIS**

- **ACM SIGMOD, VLDB, PODS, ICDE**

- **ICML, NIPS, UAI, ACM SIGKDD, AAAI, ACL, COLING, EMNLP, ANLP, HLT, SOSP, OSDI, JCDL, ASIS, ACM SIGMOD, VLDB, PODS, ICDE**
Where to Publish IR Papers

• Core IR conferences:
  – ACM SIGIR, ACM CIKM
  – ECIR, AIRS

• Core IR journals
  – ACM TOIS, IRJ
  – IPM, JASIS

• Web Applications
  – WWW, WSDM

• Other related conferences
  – Natural Language Processing: HLT, ACL, NAACL, COLING, EMNLP
  – Machine Learning: ICML, NIPS
  – Data Mining: KDD, ICDM
  – Databases: SIGMOD, VLDB, ICDE

• …
Next lectures

• Web Search
• Crawling the Web
• Retrieval Models

• Appendix
  – Essential Background