

922 U3640

Web Retrieval and Mining

(Fall 2008)

Instructor: Pu-Jen Cheng

TA: Ruei-Cheng Chen

*Department of Computer Science & Information
Engineering*

National Taiwan University

Goal & Design

- Introduce “**Web Retrieval/Search**” and “**Web Mining**”
- Prepare students for doing research/development in related fields
- Targeted at computer science (senior) undergraduate students and graduate students
- Related to Information Retrieval, Data Mining, Machine Learning, and Natural Language Processing

Format

- **~ 4 Assignments**
 - **Construction of a simple information retrieval and data mining system**
 - **Coding (C/C++ & Java) & experimenting with the IR benchmark**
 - **Demonstration & report**
 - **Individual work**
- **Midterm**
- **No Final**
- **Course Project**
 - **Presentation & report (including idea, method & experiment)**
 - **Group work encouraged**
 - **Literature review is required for graduate students**

Grading

- **Assignments: 60%**
- **Midterm: 20%**
- **Project: 20%**

Schedule

- **Part I: Web Information Retrieval**
 - Text Processing (Assignment 1)
 - Retrieval Model (Assignment 2)
 - Link Analysis (Assignment 3)
 - Multimedia/Multilingual Information Retrieval
 - Learning to Rank
 - Crawling
- **Part II: Web Mining**
 - Classification (Assignment 4)
 - Clustering
 - Information Extraction
 - Information Filtering
 - Query Log Mining

Assignment

- **Extension of CMU's Statistical Language Modeling Toolkit**
 - <http://www.speech.cs.cmu.edu/SLM/toolkit.html>
- **Test on Benchmark**

Readings

- *Mining the Web: Discovering Knowledge from Hypertext Data*, by Soumen Chakrabarti, Morgan Kaufmann, 2002. (Selected Chapters)
- *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, by Bing Liu, Springer, 2006. (Selected Chapters) **Available online!**
- *Modern Information Retrieval*, by Ricardo Baeza-Yates, Berthier Ribeiro-Neto. (Selected Chapters)
- *Introduction to Information Retrieval*, by Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, (Selected Chapters) **Available online!**
- **Additional readings will be available online**

Questions?

<http://www.csie.ntu.edu.tw/~pjcheng/course/WM2008>

Office hours:

Pu-Jen Cheng: Wed. 9:00-12:00am, R323

Ruei-Cheng Chen: cobain@iis.sinica.edu.tw