

WEB Mining

I. Course code: 0700020

Class hours: 32

Credits: 2

II. Suitable specialty: Computer Science and Technology

III. Prerequisites: Foundation of Computer Science

IV. Course goals:

Web mining is the application of data mining techniques to discover patterns from the World Wide Web. Web mining can be divided into three different types – Web usage mining, Web content mining and Web structure mining. The course propose: Introduce the fundamentals of Web data mining. Introduce various techniques for Web data mining. Develop the ability to design Web data mining applications

V. Teaching method:

Classroom lectures, Classroom discussion, Group project practice.

VI. Contents of teaching:

Classroom teaching: 32 class hours

1. Background 3 class hours
 - 1) A Brief Introduction of the Web and the Internet
 - 2) What is web mining?
 - 3) Foundations: data mining

2. Information Retrieval and Web Search 2 class hours
 - 1) Basic Concepts of Information Retrieval
 - 2) Information Retrieval Models
 - 3) Relevance Feedback
 - 4) Evaluation Measures
 - 5) Text and Web Page Pre-Processing
 - 6) Inverted Index and Its Compression
 - 7) Latent Semantic Indexing
 - 8) Web Search
 - 9) Meta-Search and Combining Multiple Rankings
 - 10) Web Spamming

3. Social Network Analysis 3 class hours
 - 1) Social Network Analysis
 - 2) Co-Citation and Bibliographic Coupling
 - 3) PageRank and HITS

4) Community Discovery

4. Web Crawling 3 class hours
 - 1) A Basic Crawler Algorithm
 - 2) Universal Crawlers
 - 3) Focused Crawlers
 - 4) Topical Crawlers

5. Structured Data Extraction: Wrapper Generation 2 class hours
 - 1) Wrapper Induction
 - 2) Instance-Based Wrapper Learning
 - 3) Automatic Wrapper Generation: Problems
 - 4) String Matching and Tree Matching

6. Information Integration 2 class hours
 - 1) Introduction to Schema Matching
 - 2) Pre-Processing for Schema Matching
 - 3) Schema-Level Matching
 - 4) Domain and Instance-Level Matching
 - 5) Combining Similarities
 - 6) Integration of Web Query Interfaces

7. Opinion Mining and Sentiment Analysis 3 class hours
 - 1) The Problem of Opinion Mining
 - 2) Document Sentiment Classification
 - 3) Sentence Subjectivity and Sentiment Classification
 - 4) Opinion Lexicon Expansion
 - 5) Aspect-Based Opinion Mining
 - 6) Mining Comparative Opinions
 - 7) Opinion Search and Retrieval
 - 8) Opinion Spam Detection

8. Web Usage Mining 2 class hours
 - 1) Data Collection and Pre-Processing
 - 2) Data Modeling for Web Usage Mining
 - 3) Discovery and Analysis of Web Usage Patterns
 - 4) Recommender Systems and Collaborative Filtering
 - 5) Query Log Mining

9. Group Practice teaching : 10 class hours
 - 1) Group project design and discussion 6 class hours
 - 2) Presentation and evaluation of group research or project 4 class hours

VII.Examination and grading:

The score uses a hundred-mark system.Total Score100%:

- 1) Classroom Performance 10%
- 2) Weekly Summaries 10%
- 3) Homework 30%
- 4) Final Exam /Team Project 50%.

VIII.Textbook and reference:

- [1] Web Data Mining, Bing Liu, Second Edition, July 2011
- [2] Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More Oct 20, 2013 by Matthew A. Russell
- [3] Data Mining: The Textbook Apr 14, 2015,by Charu C. Aggarwal
- [4] Mining the Web: Discovering Knowledge from Hypertext Data,Oct 23, 2002,by Soumen Chakrabarti
- [5] Advanced Topics in Information Retrieval (The Information Retrieval Series) 6.2011

Syllabus writer: Niuzhendong