11M1WCI431: Advanced Web Mining

Course Credit: 3
Semester: M.Tech, IV

Introduction

This course teaches students basic techniques to mine the Web and information networks (including social networks and social media). Detailed Topics include web Crawling, indexing, ranking and search algorithms using content and link analysis, Web clustering, classification, and mining algorithms, and social network analysis and on-line social media mining. The goal of this course is for students to gain knowledge on information search and web mining. Course covers techniques used to collect, analyze, and understand the data from Internet and the web (including social networks). At the end of the class, students should be able to understand the whole process of collecting information from the web, and carrying out system design for search and mining the web.
Advanced web mining

Course Objectives (Post-conditions)

Knowledge objectives:

To

1. Strengthen real network concept while crawling web.
2. Strengthen real large scale data structure understanding.
3. Strengthen the knowledge of database while storing semi structure data.
4. Develop the knowledge of web search engines, and related technologies.
5. Develop the skill to apply the learned knowledge in real problems.

Application objectives:

Web mining, refers to the automatic discovery of interesting and useful patterns from the data associated with the usage, content, and the linkage structure of Web resources, quickly become one of the most popular areas in computing and information systems because of its direct applications in e-commerce, e-CRM, Web analytics, information retrieval/filtering, Web personalization, and recommender systems.

Expected Student Background (Preconditions)

Elementary Knowledge about the Topics Introduction to Programming is required.

Topics Outline:

S NO	Topics	Hrs
1	Basic Search Engines and information retrieval, Architecture of a Search Engine, Basic Building Blocks(Text Acquisition, Text Transformation, Index Creation, User Interaction, Ranking Evaluation).	5
2	Crawls and Feeds, Deciding what to search, Crawling the Web, Directory Crawling, Document Feeds, Storing the Documents, Detecting Duplicates,	5

	Removing Noise.	
3	Processing Text, From Words to Terms, Text Statistics, Vocabulary Growth, Estimating Database and Result Set Sizes, Document Parsing, Overview, Tokenizing, Stopping, Stemming, Phrases and N-grams, Document Structure and Markup, Link Analysis, Anchor Text, PageRank, Link Quality, Information Extraction, Internationalization	5
4	Ranking with Indexes, Abstract Model of Ranking, Inverted indexes, Documents, Counts, Positions, Fields and Extents, Scores, Ordering, Compression, Entropy and Ambiguity, Delta Encoding, Bit-aligned codes, Byte-aligned codes, Looking ahead	5
5	Evaluating Search Engines, Why Evaluate? The Evaluation Corpus, Logging, Effectiveness Metrics, Recall and Precision, Averaging and Interpolation, Focusing On the Top Documents, Using Preferences, Efficiency Metrics, Training, Testing, and Statistics, Significance Tests, Setting Parameter Values, Bottom Line	5
6	Classification and Clustering, Classification and Categorization, Naïve Bayes, Support Vector Machines, Evaluation, Classifier and Feature Selection, Spam, Sentiment, and Online Advertising, Clustering, Hierarchical and*K*-Means Clustering, *K* Nearest Neighbor Clustering, Evaluation	5
7	Social Search, What is Social Search?, User Tags and Manual Indexing, Searching With Communities, Filtering and, Recommending, Document Filtering, Collaborative Filtering, Personalization, Peer-to-Peer and Metasearch, Distributed search, P2P Networks	5
8	Beyond Bag of Words, Feature-Based Retrieval Models, Term Dependence Models, Structure Revisited, XML Retrieval, Longer Questions, Better Answers, Words, Pictures, and Music, One Search Fits All?	5
	Total	40

References

1. Mining The Web, Soumen Chakarborty.
2. Modern Information Retrival, Cristofer Manning, Pravakar Raghaban.

Evaluation Scheme:

S.No	Examination	Marks
1	T-1	15
2	T-2	25
3	T-3	35
4	*Internal Marks	25

*Internal Marks Breakdown:

Assignments 9 marks (3x3)

Quizzes 12 marks (3x4)

Regularity 4 Marks