### COURSE DESCRIPTION

<table>
<thead>
<tr>
<th>Dept., Number</th>
<th>CSC 177</th>
<th>Course Title</th>
<th>Data Warehousing and Data Mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semester hours</td>
<td>3</td>
<td>Course Coordinator</td>
<td>Meiliu Lu</td>
</tr>
<tr>
<td>URL (if any):</td>
<td></td>
<td><a href="http://gaia.ecs.csus.edu/~mei/">http://gaia.ecs.csus.edu/~mei/</a></td>
<td></td>
</tr>
</tbody>
</table>

#### Catalog Description

Data mining is the automated extraction of hidden predictive information from databases. Data mining has evolved from several areas including: databases, machine learning, algorithms, information retrieval, and statistics. Data warehousing involves data preprocessing, data integration, and providing on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data, which facilitates effective data mining. This course introduces data warehousing and data mining techniques and their software tools. Topics include: data warehousing, association analysis, classification, clustering, numeric prediction, and selected advanced data mining topics. Prerequisite: At least a C- grade in CSC 134 and STAT 50 and full CSC major status.

#### Textbook


#### Course Goals

Study various subjects in data warehousing and data mining that include:

1. Basic concepts on knowledge discovery in databases.
2. Concepts, model development, schema design for a data warehouse.
3. Data extraction, transformation, loading techniques for data warehousing.
4. Concept description: input characterization and output analysis for data mining.
5. Data preprocessing.
6. Core data mining algorithms design, implementation and applications.
7. Data mining tools and validation techniques.

#### Prerequisites by Topic

*Thorough understanding of:*

- Entity-relationship analysis.
- Physical design of a relational database.
- Probability and statistics – estimation, sampling distributions, hypothesis tests.
- Concepts of algorithm design and analysis.
Basic understanding of:
- Relational database normalization techniques.
- SQL.

Exposure to:
- Bayesian theory.
- Regression.

Major Topics Covered in the Course

1. Introduction to the process of knowledge discovery in databases.
2. Basic concepts of data warehousing and data mining.
3. Data preprocessing techniques: selection, extraction, transformation, loading.
4. Data warehouse design and implementation: multidimensional data model, case study using Oracle technology.
6. Information theory and statistics in data mining: from entropy to regression.
7. Data mining core algorithms: statistical modeling, classification, clustering, association analysis.
8. Credibility: evaluating what has been leaned from training data and predicting model performance on new data, evaluation methods, and evaluation metrics.
9. Weka: a set of commonly used machine learning algorithms implemented in Java for data mining.
10. C5 and Cubist: Decision tree and model tree based data mining tools.
11. Selected advance topics based on students’ interests such as: web mining, text mining, statistical learning.
12. Case studies of real data mining applications (paper survey and invited speaker).

Outcomes

Thorough understanding of:
- Process and tasks for Knowledge discovery in databases.
- Differences between a data warehouses OLAP and operational databases OLTP.
- Multidimensional data model design and development.
- Techniques for data extraction, transformation, and loading.
- Machine learning schemes in data mining.
- Mining association rules (Apriori).
- Classification and prediction (Statistical based: Naïve Bayes, regression trees and model trees; Distance based: KNN, Decision tree based: 1R, ID3, CART; Covering algorithm: Prism).
- Cluster analysis (Hierarchical algorithms: single link, average link, and complete link; Partitional algorithms: MST, K-means; Probability based algorithm: EM).
- Use of data mining tools: C5, Cubist, Weka.
Basic understanding of:
• Data warehouse architecture.
• Information theory and statistics in data mining.
• Credibility analysis and performance evaluation.

Exposure to:
• Mining complex types of data: multimedia, spatial, and temporal.
• Statistical learning theory.
• Support vector machine and ANN.

Laboratory Projects

1. Design and implement a data warehouse database (4 weeks).
2. Explore extraction, transformation, loading tasks in data warehousing (1 week).
3. Explore data mining tools and algorithms implementation (3 weeks).
4. Design and implement data mining application (3 weeks).

Estimated Curriculum Category Content (Semester hours)

<table>
<thead>
<tr>
<th>Area</th>
<th>Core</th>
<th>Advanced</th>
<th>Area</th>
<th>Core</th>
<th>Advanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithms</td>
<td>1</td>
<td></td>
<td>Data Structures</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Software Design</td>
<td>0.5</td>
<td></td>
<td>Prog. Languages</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Comp. Arch.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Oral and Written Communications

1. Three written reports (term project proposal, research paper review, and term project report).
2. Two oral presentations (10 minutes for paper review and 15-20 minutes for term project).

Social and Ethical Issues

No significant component.

Theoretical Content

1. Data warehouse schema and data cube computation.
2. Information theory and statistics in data mining.
3. Data mining algorithms and their output model performance prediction.
4. Evaluation metrics (confusion matrix, cost matrix, F measure, ROC curve).
Problem Analysis

1. Design of a data warehouse.
2. Design of a process of ETL (Extraction, Transformation, Loading).
3. Design of a data mining application.
5. Analysis and comparison of data mining schema.

Solution Design

1. Design of a data warehouse.
2. Design of a process of ETL (Extraction, Transformation, Loading).
3. Design of a data mining application.
5. Analysis and comparison of data mining schema.