

CSE4392/5334 DATA MINING

Lecture 3: Course
Project

CSE4392/5334 Data Mining, Fall 2009
Department of Computer Science and Engineering, University of Texas at Arlington
©Chengkai Li, 2009

TA

- Xiaonan (Michael) Li
- Office: [GeoScience](#) 237
- Phone: (817) 272-0896
- E-mail: xiaonan.li [AT] mavs [DOT] uta [DOT] edu

Lecture 3: Course Project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 2

WebCT

- WebCT now accessible at
<http://www.uta.edu/webct/>
- UTA NetID and password

Lecture 3: Course Project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 3

Project Overview

- **Must be done individually**
- **Defined projects**
 - 3 programming tasks (classification, association rule, clustering).
 - You can implement existing method or your own method.
 - The implementation will be evaluated over a given dataset.
 - Two tracks:
 - Structured data
 - Text data
- **Open-ended research project**
Talk to me if you want to pursue this.

Lecture 3: Course project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 4

0-tolerance to plagiarism

- We will use detection software to identify similar source codes and documents.

Lecture 3: Course Project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 5

Defined Projects: Schedule

- **Due time is always 11:55pm.**
- **10/07 Due: Classification (P1)**
- **11/04 Due: Association Rule Mining (P2)**
- **12/02 Due: Clustering (P3)**

Lecture 3: Course project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 6

Open-Ended Projects: Topics

Lecture 3: Course Project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 7

1. Apache Mahout: distributed machine learning algorithms on Hadoop

- Explore Mahout and Hadoop
- Experiment with certain distributed machine learning algorithms implemented in Mahout
- Compare with single-node machine learning algorithm

Lecture 3: Course Project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 8

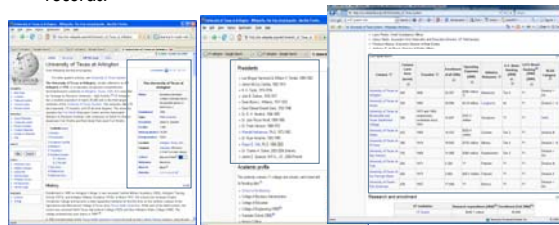
2. Learning Preference/Ranking Function

- Users may have complicated (implicit) functions in ranking objects, or rules in defining preferences.
- How to learn such functions or rules, based on training data (user's answers on probing questions.)
- The learned function can be used to rank future data.
- Ranking function => objects
- objects => ranking function

Lecture 3: Course Project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 9

3. Extracting Structured Records from Wikipedia

- Extract tables and lists from Wikipedia pages. Integrate them in relational database, to enable querying over such structured records.



Lecture 3: Course Project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 10

4. Classifying Wiki articles

- Each wiki page has a set of categories.
- Some categories are inherently more important than other categories.
- How to decide the degree of membership of this page in each category?

Lecture 3: Course Project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 11

5. Clustering Wiki articles

- Cluster the keyword query result of Wikipedia
- <http://wiki.clusty.com/>

Can you do even better?

We can use these information:

- keywords
- associated entities
- category
- infobox

Lecture 3: Course Project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 12

6. Quantifying the category information in Wikipedia

- Each wiki page has a set of categories.
- Some categories are inherently more important than other categories.
- Quantify the degree of membership of a page in a category?
 - "TF-IDF"
 - hyperlinks in the page
 - Pages in the same category?

Lecture 3: Course Project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 13

7. Schema Summary

- **interactive query formulation**
 - Suggest keywords to use
 - Suggest entities to join
 - Can further rank suggestions
 - Indicate "well-formedness" of users' queries
- **index, statistics**
 - materialize certain join paths
 - estimate result cardinality
- **query optimization**

Lecture 3: Course Project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 14

8. Significant Fact Finding

Lecture 3: Course Project CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 15