An Overview of CS512 @Spring 2020

JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

JANUARY 21, 2020

Data and Information Systems (DAIS) Course Structures at CS/UIUC

- Three main streams: Database, data mining and text information systems
- Database Systems:
  - Database management systems (CS411: Fall + Spring)
  - Advanced database systems (CS511: Fall)
- Data mining
  - Intro. to data mining (CS412: Fall + Spring)
  - Data mining: Principles and algorithms (CS512: Spring (Han))
  - Network of Networks (Hanghang Tong)
- Text information systems
  - Introduction to Text Information Systems (CS410: Spring (Zhai))
  - Advance Topics on Information Retrieval (CS 598 or CS510: Fall (Zhai))
- Social & Economic Networks (CS 598: Hari Sundaram)

CS512 Coverage@2019: Mining Massive Text Corpora and Information Networks

- ❑ Class introduction + course technical overview (.5 week)
- ❑ Text mining 1: Text embedding (1.5 week)
- ❑ Text mining 2: Phrase mining (1.5 week)
- ❑ Text mining 3: Named entity/relation extraction and typing (1.5 week)
- ❑ Text mining 4: Mining patterns, relations and claims (1.5 week)
- ❑ 1st midterm exam (0.5week) — 2nd Lect. of 7th week
- ❑ Text mining 5: Mining sets and taxonomies (1 week)
- ❑ Text mining 6: Text cube: Construction and Exploration (1 week)
- ❑ Network mining 1: Heterogeneous information networks and network clustering (1 week)
- ❑ Network mining 2: Classification and link prediction in hetero. info. networks (1 week)
- ❑ Network mining 3: Other issues at mining heterogeneous information networks (1 week)
- ❑ Truth finding (1 week)
- ❑ 2nd midterm exams (0.5 week)—2nd Lect. of 15th week
- ❑ Class research project presentation (final week + exam week)

Class Information

- **Instructor:** Jiawei Han (www.cs.uiuc.edu/~hanj)
  - Lectures: Tues/Thurs 3:30-4:45pm (0216 SC)
  - Office hours: Tues/Thurs 4:45-5:30pm (2132 SC)
- **Teach Assistants** (using Piazza to seek for help when needed)
  - Xiaotao Gu (50%), Lucas (Liyuan) Liu (50%, online TA), Jiaming Shen
  - TA office hours: TBD
- **Prerequisites** (course preparation: Consent with instructor if not sure)
  - CS412 (offered every semester) plus
  - General knowledge on statistics, machine learning, natural language processing and text information systems
- **Course website** (bookmark it since it will be used frequently!)
  - https://wiki.cites.illinois.edu/wiki/display/cs512/Lectures
- **Major textbook: Recent research papers**

Textbooks & Recommended References

- **Textbooks**
  - Charu C. Aggarwal, Machine Learning for Text, Springer 2017
  - Chao Zhang and Jiawei Han, Multidimensional Mining of Massive Text Data, Morgan & Claypool Publishers, 2019
  - Xiang Ren and Jiawei Han, Mining Structures of Factual Knowledge from Text: An Effort-Light Approach, Morgan & Claypool Publishers, 2018
  - Jialu Liu, Jingbo Shang and Jiawei Han, Phrase Mining from Massive Text and Its Applications, Morgan & Claypool, 2017
  - Yizhou Sun and Jiawei Han, Mining Heterogeneous Information Networks: Principles and Methodologies, Morgan & Claypool, 2012
  - Recent published research papers (see course syllabus)
- **Other general reference books**
  - Jiawei Han, Micheline Kamber, *Jian Pei, Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011
  - K. P. Murphy, "Machine Learning: a Probabilistic Perspective", MIT Press, 2012

Course Work: Assignments, Exams and Course Project

- **Assignments:** (Two assignments, equal weight) **25%** total
  - One programming assignment (10%)
  - One mini-research assignment (15%)
- **Two midterm exams** (equal weight): **40%** in total
- **Research project proposal (3-5 pages): 2%** (due **at the end of 5th week**)
- **Class attendance (3%)**: Max misses w/o penalty: 3, then −0.3% for each miss
  - For online students, 3% will be folded into research/survey report
- **Final course project: 30%** (due at the end of semester)
  - Evaluated by class (50%) and TA + instructor (50%) collectively!
- **Class presentation on new papers and surveys** (Optional: max credit: 0.5%)
  - Topics and time slot (~15 minutes): Consent with instructor; maximal using TA-guided classical paper presentation slots

Research Projects Evaluation

- **Final course project:** 30% (due at the end of semester)
  - The final project will be evaluated based on (1) **technical innovation**, (2) **thoroughness of the work**, and (3) **clarity of presentation**
  - The final project will need to hand in: (1) **project report** (length will be similar to a typical 8- to 12-page double-column conference paper), and (2) **project presentation slides** (required for both online and on-campus students)
  - Each course project for every on-campus student will be evaluated collectively by instructor (plus TA) and other on-campus students in the same class
  - Online student projects will be evaluated by instructors and TA only
  - Single-person project is OK; encouraged to have 2-3 as a group, and/or team up with some senior graduate students (clearly specify the % of contributions)

Where to Find Reference Papers?

- Course research papers: Check reading list and references at each chapter
- Major conference proceedings on data mining and related disciplines
  - DM conferences: ACM SIGKDD (KDD), ICDM (IEEE, Int. Conf. Data Mining), SDM (SIAM Data Mining), ECMLPKDD (Principles KDD), PAKDD (Pacific-Asia)
  - Web and IR conferences: SIGIR, CIKM, WWW, WSDM
  - NLP conferences: ACL, EMNLP, NAACL
  - ML conferences: NIPS, ICML
  - DB conferences: ACM SIGMOD, VLDB, ICDE
  - Social network conferences: ASONAM
- Other related conferences and journals
  - IEEE TKDE, ACM TKDD, DMKD, ML, …
- Use course Web page, DBLP, Google Scholar, Citeseer

Questions for Short Discussion

- Two disciplines: Data mining vs. machine learning
  - What are the links and differences?
- Two courses: CS412 (Introduction to Data Mining) vs. CS512 (Advance Data Mining)
  - What are the links and differences?
- Two research projects: Mini-research assignment vs. your selected research projects
  - What are the links and differences?
- Discussion on course grading policy

Our Journey: From I ctures & Knowledge

MORGAN & CLAYPOOL PUBLISHERS

Mining Structures of Factual Knowledge from Text

An Effort-Light Approach

Xiang Ren
Jiawei Han

Xiang Ren: SIGKDD'18 Dissertation

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, Robert Grossman, *Series Editors*

MORGAN & CLAYPOOL PUBLISHERS

Multidimensional Mining of Massive Text Data

Chao Zhang
Jiawei Han

C. Zhang: SIGKDD'19 Dissertation Award Runner-Up

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, Robert Grossman, *Series Editors*

MORGAN & CLAYPOOL PUBLISHERS

Phrase Mining from Massive Text and Its Applications

ialu Liu
ngbo Shang
wei Han

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, Robert Grossman, *Series Editors*

Thi
MK
HESIS LECTURES ON
ATA MINING AND KNOWLEDGE D
Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, Robert Grossman, Serie
CLAYPOOL PUBLISHERS
ISCOVERY
Grossman, Series Editors

Han, Kamber
Data Mining, 3rd

nd Faloutsos (e
Link Mining, 2010

Sun and Han, Mining Heterogeneous Information Networks, 2012
Y. Sun: SIGKDD'13 Dissertation Award

Latent Entity
015
C. Wang: SIG ssertation Award