Khoa Khoa Học & Kỹ Thuật Máy Tính
Trường Đại Học Bách Khoa Tp. Hồ Chí Minh

Khai Phá Dữ Liệu - Data Mining

Mã MH: 055004

TRAN MINH QUANG
quangtran@hcmut.edu.vn
http://www.cse.hcmut.edu.vn/quangtran
http://researchmap.jp/quang

2017/2/25

COURSE OBJECTIVES

- To introduce students to knowledge discovery process, data mining process, and data preprocessing in general
- To introduce students to the support of other research areas in computer science for the data mining area as well as to the benefits of data mining to many various application domains
- To present main algorithms and techniques in data preprocessing
- To present main data mining algorithms and techniques for regression, classification, clustering, association – correlation analysis
- To enable students to develop and utilize data mining algorithms and techniques for many different applications and kinds of data

LEARNING OUTCOMES

- Understand the steps in the overall knowledge discovery process
- Describe basic concepts, technologies, and applications of data mining
- Explain popular data mining tasks including regression, classification, clustering, and association rules mining
- Identify data related issues in the data preprocessing phase for data mining tasks
- Understand how to use data mining to make better business decisions
- Develop and use data mining algorithms and tools for data mining application development
- Have sufficient knowledge to do research on the data mining area

KHAI PHÁ DỮ LIỆU (KPDL)

COURSE OUTLINE

Introduction to the course and the lecture methods (W1)

PART A: Background on DM (W1 to W3)

- Basic concepts and elements in DM (Lecture)
- Data pre-processing (Lecture)
- Overview on DM techniques: Prediction, Classification, Clustering, Association rule (Lecture)

PART B: Student presentations on DM topics/papers

- W4: G1: Classification techniques
  - Logistic regression
  - Decision tree
- W5: G2: Classification
  - Bayesian network
  - ANN
- W6: G3: Clustering techniques
  - K-means based method
  - Density based method

COURSE OUTLINE

PART B: Student presentations on DM topics/papers

- W7: G4:
  - EM method for clustering
  - Apriori: (Association rule)
- W8 : G5: Paper presentation
  - FP-growth
  - One related paper
- W9 + W10: Paper presentation
  - G6, G7, G8, G9

  Note: Selective new and novel papers
- W11 + W12: Invited lectures (Prof. Agnieszka Darzinska-Glebocka, Bialystok Unvi. Of Tech. Poland, date will be noted later)

PART C: Mini project

- W13 – W15
  - Presentation for the mini project: (8 groups): 3-3-3G/session

EVALUATIONS

- Midterm exam: 20% (presentation: topics/papers)
- Mini-project: 30% (report + presentation)
- Final exam: 50%

Note

- Pass: the average score >= 5
- Absent 3 lectures: -1 point on total score
- W1: create groups; each group selects the mini project
- W4-W12: Groups present DM topics/papers + Invited lectures
- W13-W15: Mini project presentation

MINI PROJECT

1. Network traffic classification (a part of data is available) – G1
2. SDN and Big data: Network traffic analysis at the controllers – G3
3. Crowd-sourcing models in social networks – G8
4. Context-aware data/service distribution model in IoT: A case study in Health-care systems – G7
5. Context-aware data/service distribution model in IoT: A case study in traffic control systems – G5
6. Collect data and analyze flood patterns in HCM City – G9

MINI PROJECT

7. Collect data and analyze stock price - G6
8. Collect data and analyze gold price – G4
9. Collect and analyze the traffic accident data
10. Pattern discovery from Vietlott lottery – G2
11. Investigation on Big Data
12. Investigate the Microsoft Azure Machine learning tools
13. Data mining in low computing power systems

SOURCE FOR RELEVANT PAPERS

- Publishers:
  - ACM
  - IEEE
  - Springer
  - Elsevier
- From the Internet
  - Google scholar
  - Labs/research groups who are strong on DM research

Source for relevant data/papers

- Data mining and KDD (SIGKDD member CDROM):
  - Conference proceedings: KDD, and others, such as PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery
- Database field (SIGMOD member CD ROM):
  - Conference proceedings: ACM-SIGMOD, ACM-PODS, VLDB, ICDE, EDBT, DASFAA
  - Journals: ACM-TODS, J. ACM, IEEE-TKDE, JIIS, etc.
- AI and Machine Learning:
  - Conference proceedings: Machine learning, AAAI, IJCAI, etc.
  - Journals: Machine Learning, Artificial Intelligence, etc.
- Statistics:
  - Conference proceedings: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization:
  - Conference proceedings: CHI, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

RELEVANT DM COMMUNITY

- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2001 conferences, and SIGKDD Explorations
- More conferences on data mining
  - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, etc.

REFERENCES

[1] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, 2012.

[2] David Hand, Heikki Mannila, Padhraic Smyth, "Principles of Data Mining", MIT Press, 2001.

[3] David L. Olson, Dursun Delen, "Advanced Data Mining Techniques", Springer-Verlag, 2008.

[4] Graham J. Williams, Simeon J. Simoff, "Data Mining: Theory, Methodology, Techniques, and Applications", Springer-Verlag, 2006.

[5] ZhaoHui Tang, Jamie MacLennan, "Data Mining with SQL Server 2005", Wiley Publishing, 2005.

[6] Oracle, "Data Mining Concepts", B28129-01, 2008.

[7] Oracle, "Data Mining Application Developer's Guide", B28131-01, 2008.

[8] Ian H.Witten, Eibe Frank, "Data mining : practical machine learning tools and techniques", 2nd Edition, Elsevier Inc, 2005.

[9] Florent Messeglia, Pascal Poncelet & Maguelonne Teisseire, "Successes and new directions in data mining", IGI Global, 2008.

[10] Oded Maimon, Lior Rokach, "Data Mining and Knowledge Discovery Handbook", 2nd Edition, Springer Science + Business Media, LLC 2005, 2010.

Q&A

quangtran@hcmut.edu.vn