Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology

Data Mining
Course ID: CO3029

Assoc. Prof. TRAN MINH QUANG
quangtran@hcmut.edu.vn
http://researchmap.jp/quang

2022/9/7

Course objectives

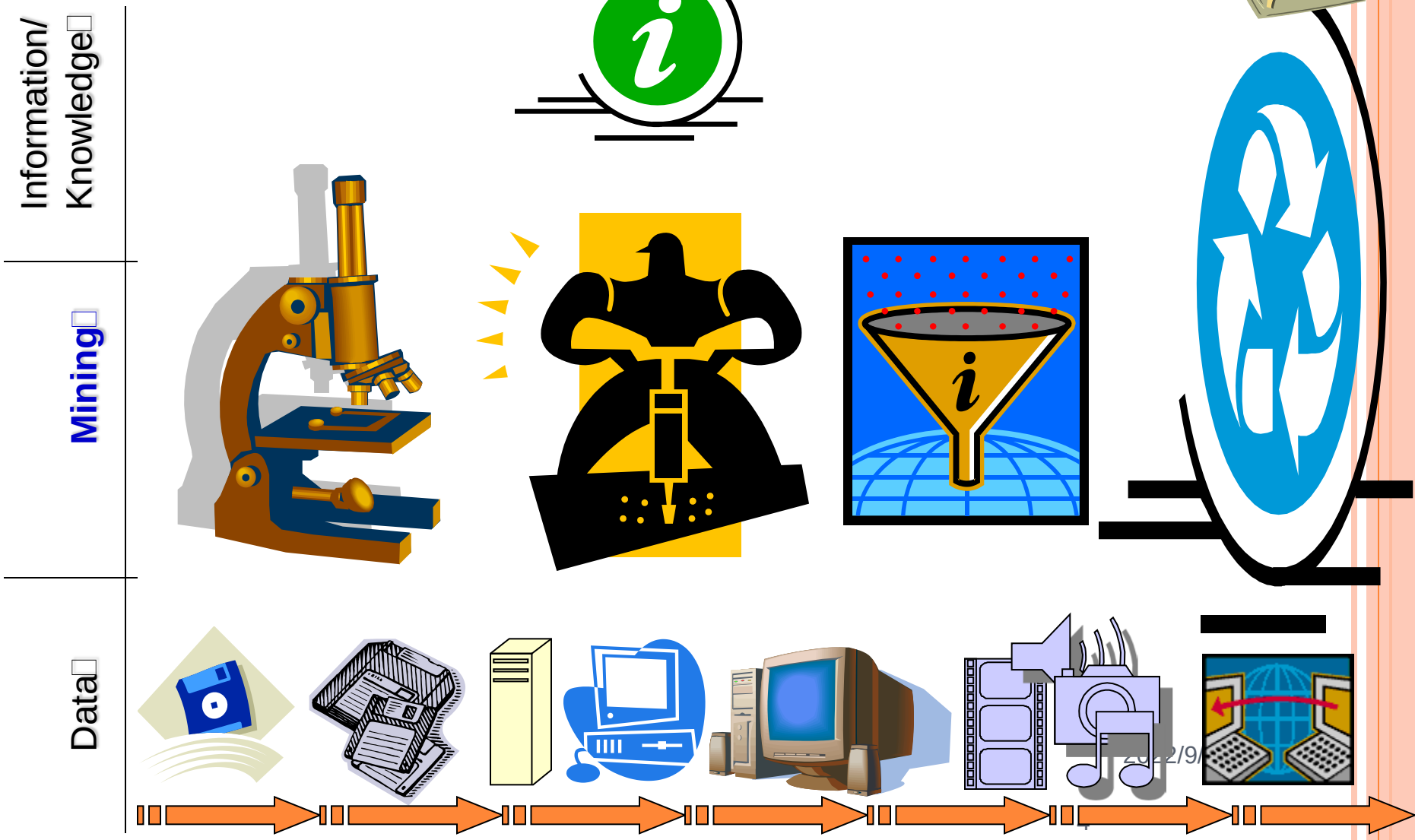
- This course aims to introduce the knowledge discovery concepts, process, technologies, and applications of data mining.

- It is also to discuss data preprocessing issues, data mining tasks, algorithms and tools that can be used to support data analysts and data mining application developers.

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
- Use data mining algorithms and tools for data mining application development
- Have sufficient knowledge to do research on the data mining area

DATA MINING



Content

- Chapter 1: Introduction to Data Mining
- Chapter 2: Data Preprocessing
- Chapter 3: Regressions
- Chapter 4: Data Classification
- Chapter 5: Data Clustering
- Chapter 6: Association rule mining
- Chapter 7: Mini project presentation
- Chapter 8: Research direction + Summary

COURSE OUTLINE

Introduction to the course and the lecture methods (W1)

PART A: Background on DM (W1 to W2)

- **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

<12 groups will presents and review for each other>

- **W3 : Classification**
  - **G1: Introduction to Classification, Logistic regression and applications**
  - **G2: Decision tree and applications**
- **W4: Classification**
  - **G3: Bayesian methods and applications**
  - **G4: ANN and applications**

COURSE OUTLINE

- W5: Clustering
  - G5: A selected paper on classification
  - G6: K-means based method
- W6: Clustering
  - G7: Density based method
  - G8: A selected paper on clustering
- W7: Association rule mining
  - G9: Apriori
  - G10: FP-growth
- W8: Association rule mining
  - G11: Introduction to deep learning
  - G12: Selected paper on association rule

PART C: Mini project

- W9– W12
  - Presentation for the mini project: (12 groups): 3 Groups/session (12 minutes for presentation + 12 minutes for Q&A)
  - Summary + Research direction

REFERENCES

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

[2] Trần Minh Quang, "Khai Phá Dữ Liệu và Kỹ Thuật Phân Lớp", NXB Đại Học Quốc Gia TP. HCM, 2020.

[3] Charu C. Aggarwal, "Data Classification: Algorithms and Applications" (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series) 1st Edition, 2014.

[4] Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, "Mining of Massive Datasets", 2nd Edition, Cambridge University Press 2014.

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, “Deep Learning”, The MIT Press, 2016.

FURTHER READ

"[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.

EVALUATION

- Exercise/Quiz: 5% + Presentation: 15% => **20%**
- Mini project: 30%
- Final exam: 50%: Multi choice and written questions

Note: No column is allowed NULL -> set to Zero for the total score

- Absence 2 lectures –1 point from total score
- -1 more point for each further absence
- Refer well text book and the Internet
- Practice Data mining tools: Weka, Python, R, Oracle and SQL Server,…

Mini Project

- 3- 4 Students/Group
- W2-8: Conduct projects
- W9-12: Present in the class (12 minutes + 12 minutes Q&A)

Mini Project_Evaluation

1. (2 points) Describe problems clearly (motivation, problem definition, main contributions of the project,...)
2. (3 points maximum based on levels) Proposed solution
   - Applying the existing techniques (1.5 points))
   - Improve the existing technique s (2.0 points)
   - Propose new one (3.0 points)
3. (2 points) Implementation the proposed method
4. (1 points) Evaluation and Discussion
5. (2 points) Presentation: Slide + presentation skills

MINI PROJECT

1. Estimating traffic condition in Ho Chi Minh City using classification techniques
2. Estimating traffic condition in Ho Chi Minh City using prediction techniques
3. Estimating traffic condition in Ho Chi Minh City using association rule mining techniques
4. Estimating traffic condition in Ho Chi Minh City using clustering methods
5. Discovery trends in common e-commercial websites
6. Develop a recommendation systems for e-commercial websites

MINI PROJECT

7. Analyze stock market using prediction techniques
8. Analyze stock market using classification method techniques
9. Analyze stock market using Association rule methods
10. Analyze stock market using Clustering methods
11. Analyze retail data using data mining techniques
12. Analyze user behavior social network using data mining techniques
13. Predict the real-estate price

MINI PROJECT

14. Investigate Oracle Data Mining and develop an application
15. Investigate Microsoft data mining tools and develop an application
16. Investigate Intelligent Miner (IBM) and develop an application
17. Data mining from Big Data: techniques, tools and applications
18. Investigate Mapreduce, Hadoop for big data analysis
19. Real-time data mining/processing
20. Investigate data mining tools on the cloud and develop an application
21. **Propose data a mining system which is potential for real-world application**

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.

Q&A

quangtran@hcmut.edu.vn