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

Khai phá dữ liệu (Data mining)

Cao Học Ngành Khoa Học Máy Tính

Giáo trình điện tử

Biên soạn bởi: TS. Võ Thị Ngọc Châu

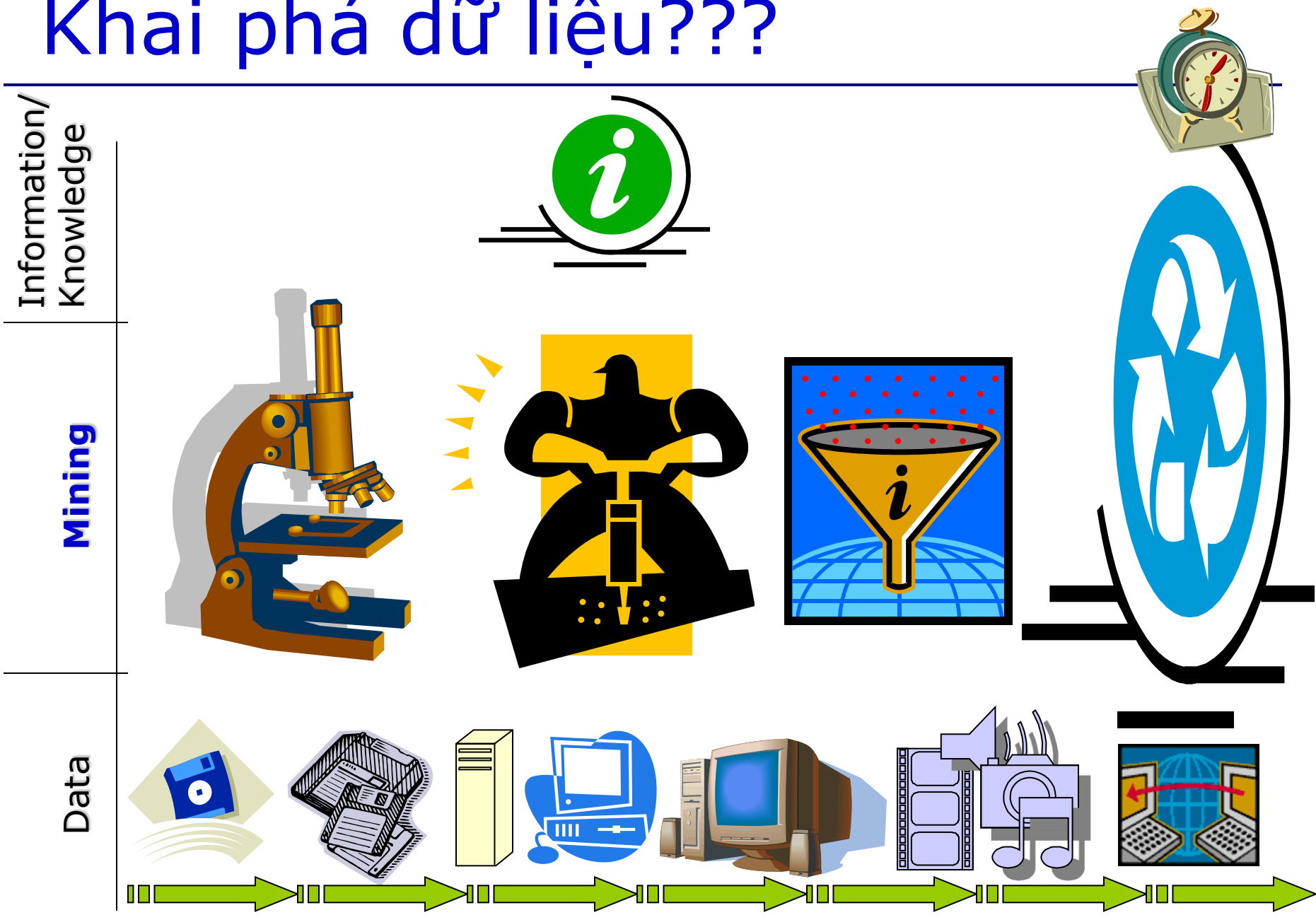
(chauvtn@cse.hcmut.edu.vn, chauvtn@hcmut.edu.vn)

Học kỳ 2 – 2014-2015

Khai phá dữ liệu ???

- Obama campaign's secret strategy – 2012
- Knowing your customers
- Predict final status of undergrad students
- Predict heart disease
- Car classification
- …

Khai phá dữ liệu???



KPDL – Lĩnh vực liên ngành

- Phân tích - thiết kế giải thuật (algorithm design and analysis)
- Quản lý dữ liệu (data management)
- Truy hồi thông tin (information retrieval)
- Máy học (machine learning)
- Thống kê (statistics)
- Trực quan hóa (visualization)
- ...

Mục tiêu của môn học

- Giới thiệu cho sinh viên tổng quan về các quá trình khám phá tri thức, khai phá dữ liệu, và quá trình tiền xử lý dữ liệu
- Giới thiệu cho sinh viên những hỗ trợ từ các lĩnh vực nghiên cứu khác trong khoa học máy tính dành cho lĩnh vực khai phá dữ liệu cũng như những giá trị lợi ích mà khai phá dữ liệu đóng góp trong các lĩnh vực ứng dụng khác nhau
- Trình bày các giải thuật và kỹ thuật chính trong giai đoạn tiền xử lý dữ liệu
- Trình bày các giải thuật và kỹ thuật khai phá dữ liệu chính gồm: hồi qui dữ liệu, phân loại dữ liệu, gom cụm dữ liệu, và phân tích kết hợp – tương quan
- Tạo khả năng cho sinh viên phát triển và tận dụng các giải thuật và kỹ thuật khai phá dữ liệu cho các ứng dụng và loại dữ liệu khác nhau

Tài liệu tham khảo

- [1] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques", Third 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] Hillol Kargupta, Jiawei Han, Philip S. Yu, Rajeev Motwani, and Vipin Kumar, "Next Generation of Data Mining", Taylor & Francis Group, LLC, 2009.
- [6] Daniel T. Larose, "Data mining methods and models", John Wiley & Sons, Inc, 2006.
- [7] Ian H.Witten, Frank Eibe, Mark A. Hall, "Data mining : practical machine learning tools and techniques", Third Edition, Elsevier Inc, 2011.
- [8] Florent Messeglia, Pascal Poncelet & Maguelonne Teisseire, "Successes and new directions in data mining", IGI Global, 2008.
- [9] Oded Maimon, Lior Rokach, "Data Mining and Knowledge Discovery Handbook", Second Edition, Springer Science + Business Media, LLC 2005, 2010.

Nội dung

- Chương 1: Tổng quan về khai phá dữ liệu
- Chương 2: Các vấn đề tiền xử lý dữ liệu
- Chương 3: Hồi qui dữ liệu
- Chương 4: Phân loại dữ liệu
- Chương 5: Gom cụm dữ liệu
- Chương 6: Luật kết hợp
- Chương 7: Khai phá dữ liệu và công nghệ cơ sở dữ liệu
- Chương 8: Ứng dụng khai phá dữ liệu
- Chương 9: Các đề tài nghiên cứu trong khai phá dữ liệu
- Chương 10: Ôn tập

Nội dung – Tài liệu tham khảo

- Chương 1: Tổng quan về khai phá dữ liệu [1, 2, 7, 9]
- Chương 2: Các vấn đề tiền xử lý dữ liệu [1]
- Chương 3: Hồi qui dữ liệu [1-7]
- Chương 4: Phân loại dữ liệu [1-7]
- Chương 5: Gom cụm dữ liệu [1-7]
- Chương 6: Luật kết hợp [1-7]
- Chương 7: Khai phá dữ liệu và công nghệ cơ sở dữ liệu [1, 2]
- Chương 8: Ứng dụng khai phá dữ liệu [3, 5, 9]
- Chương 9: Các đề tài nghiên cứu trong khai phá dữ liệu [5, 8]
- Chương 10: Ôn tập [1-9]

Nội dung - Lịch học

- Chương 1: Tổng quan về khai phá dữ liệu (T.1)
- Chương 2: Các vấn đề tiền xử lý dữ liệu (T.2-3)
- Chương 3: Hồi qui dữ liệu (T.4-5)
- Chương 4: Phân loại dữ liệu (T.6-7)
- Chương 5: Gom cụm dữ liệu (T.8-9)
- Chương 6: Luật kết hợp (T.10-11)
- Chương 7: Khai phá dữ liệu và công nghệ cơ sở dữ liệu (T.12)
- Chương 8: Ứng dụng khai phá dữ liệu (T.13)
- Chương 9: Các đề tài nghiên cứu trong khai phá dữ liệu (T.14)
- Chương 10: Ôn tập (T.15)

Hiểu biết - Kỹ năng đạt được

- Hiểu các bước trong quá trình khám phá tri thức
- Mô tả được các khái niệm, công nghệ, và ứng dụng của khai phá dữ liệu
- Giải thích được các tác vụ khai phá dữ liệu phổ biến như hồi qui, phân loại, gom cụm, và khai phá luật kết hợp
- Nhận dang được các vấn đề về dữ liệu trong giai đoạn tiền xử lý cho các tác vụ khai phá dữ liệu
- Hiểu cách sử dụng khai phá dữ liệu để có được các quyết định tốt hơn
- Sử dụng được các giải thuật và công cụ khai phá dữ liệu để phát triển ứng dụng khai phá dữ liệu
- Được chuẩn bị về kiến thức để có thể nghiên cứu trong lĩnh vực khai phá dữ liệu

Đánh giá kết quả học tập

- Tiểu luận: 30%
- Kiểm tra: 20%
- Thi cuối kỳ: 50%

→ Đạt: 30%*Tiểu luận + 20%*Kiểm tra + 50%*Thi cuối kỳ >= 5.0

Hình thức đánh giá kết quả học tập

- Tiểu luận: 30%
  - Nội dung báo cáo: 20%
  - Nội dung trình bày: 10%
- Kiểm tra: 20%
  - 2 bài kiểm tra vào tuần 7, 14
  - 10%/bài/20 phút
  - 10 câu trắc nghiệm (1đ/câu)
- Thi cuối kỳ: 50%
  - 40 câu trắc nghiệm (0.25đ/câu)+1 câu viết (2đ/câu)
  - Thời gian thi: 120 phút

Tiểu luận

- 1 đề tài/sinh viên CH, 2 đề tài/sinh viên NCS
- Sinh viên chọn đề tài và bắt đầu thực hiện tiểu luận từ tuần thứ **1**.
- Sinh viên nộp bài làm tiểu luận vào tuần thứ **15**.
  - Nộp bài trễ: -2 điểm
- Bài nộp cho tiểu luận gồm:
  - Báo cáo: .doc, .docx, .pdf
  - Trình bày: .ppt, .pptx, .pps
  - Sản phẩm (nếu có, để kiểm tra kết quả đạt được của tiểu luận)

Đề tài của Tiểu luận

- **01**. 2014 Pairwise dynamic time warping for event data
- **02**. 2014 Outlier detection for temporal data - a survey
- **03**. 2013 Stratified sampling for feature subspace selection in random forests for high dimensional data
- **04**. 2013 Ensemble learning for wind profile prediction with missing values
- **05**. 2013 An Optimized Cost-Sensitive SVM for Imbalanced Data Learning
- **06**. 2012 Transfer spectral clustering
- **07**. 2012 The Move-Split-Merge metric for time series
- **08**. 2012 Substructure clustering - a novel mining paradigm for arbitrary data types
- **09**. 2012 Secure Bayesian model averaging for horizontally partitioned data
- **10**. 2012 sDTW - computing DTW distances using locally relevant constraints
- **11**. 2012 Predicting student failure at school with high dimensional and imbalanced data
- **12**. 2012 Piecewise evolutionary segmentation for feature extraction in time series models
- **13**. 2012 Mining top-k frequent patterns without minimum support threshold
- **14**. 2012 Mining low support discriminative patterns from dense and high-dimensional data
- **15**. 2012 Hiding Sensitive Association Rules without Altering the support of sensitive items

Đề tài của Tiểu luận (tt)

- **16**. 2012 Finding association rules in semantic web data
- **17**. 2012 Analysis of preprocessing vs cost-sensitive learning for imbalance data sets
- **18**. 2012 An assessment of the effectiveness of a random forest classifier for land-cover classification
- **19**. 2011 Weighted dynamic time warping for time series classification
- **20**. 2011 Temporal data clustering via weighted clustering ensemble with different representations
- **21**. 2011 Scalable k-nn search on vertically stored time series
- **22**. 2011 Learning a tensor subspace for semi-supervised dimensionality reduction
- **23**. 2011 Incremental K-clique clustering in dynamic social networks
- **24**. 2011 Face recognition by generalized two-dimensional FLD method and multi-class SVM
- **25**. 2011 Clustering Very Large Multi-dimensional Datasets with MapReduce
- **26**. 2011 An optimal summetrical null space criterion of Fisher discriminant for feature extraction and recognition
- **27**. 2011 An evoluationary algorithm to discover quantitative association rules in multidimensional time series

Đề tài của Tiểu luận (tt)

- **28**. 2011 Activity knowledge transfer in smart environments
- **29**. 2011 A unique property of single-link distance and its application in data clustering
- **30**. 2011 A multi-objective artificial immune algorithm for parameter optimization in SVM
- **31**. 2010 Multi-objective genetic fuzzy classifiers for imbalanced and cost-sensitive datasets
- **32**. 2010 Linguistic data mining with fuzzy FP-trees
- **33**. 2010 Knowledge-Based Interactive Postmining of Association Rules Using Ontologies
- **34**. 2010 Incremental Clustering for Trajectories
- **35**. 2010 Fractal Mining - self similarity-based clustering and its applications
- **36**. 2010 Data clustering - 50 years beyond K-means
- **37**. 2010 Boosting support vector machines for imbalanced data sets
- **38**. 2010 An intelligent decision-support model using FSOM and rule extraction for crime prevention
- **39**. 2010 An efficient algorithm for incremental mining of temporal association rules
- **40**. 2010 A Survey on Transfer Learning

Đề tài của Tiểu luận (tt)

- **41**. 2010 A grid portal for solving geoscience problems using distributed knowledge discovery services
- **42**. 2009 Mining frequent trajectory patterns in spatial–temporal databases
- **43**. 2009 Graph Clustering Based on Structural-Attribute Similarities
- **44**. 2009 Applying web usage mining for personalizing hyperlinks
- **45**. 2009 A scalable framework for cluster ensembles
- **46**. 2008 On the k-NN performance in a challenging scenario on imbalance and overlapping
- **47**. 2008 Learning decision trees for unbalanced data
- **48**. 2008 Incrementally fast updated frequent pattern trees
- **49**. 2008 Efficient similarity search over future stream time series
- **50**. 2007 Mining nonambiguous temporal patterns for interval-based events
- **51**. 2007 ARMADA - an algorithm for discovering richer relative temporal association rules from interval-based data
- **52**. 2007 A kernel-based two-class classifier for imbalanced data sets
- **53**. 2006 Feature-based Similarity Search in Graph Structures

Đề tài của Tiểu luận (tt)

- **54**. 2006 Evaluating misclassifications in imbalanced data
- **55**. 2006 Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles
- **56**. 2005 Smooth function approximation using neural networks
- **57**. 2004 Towards parameter-free data mining
- **58**. 2004 Mining Sequential Patterns by Pattern-Growth - the prefix span approach
- **59**. 2003 Towards systematic design of distance functions for data mining applications
- **60**. 2001 On the Surprising Behavior of Distance Metrics in High Dimensional Space
- **61**. 2001 Fuzzy c-means clustering of incomplete data
- **62**. 2000 Privacy-preserving data mining
- **63**. 1997 Scaning reduce technich - Using a Hash-based Method with Transaction Trimming for Mining AR
- **64**. 1996 BIRCH - an efficient data clustering method for very large databases
- **65**. 1994 Fast algorithms for mining association rules
- **66**. …

Đề tài #2 của Tiểu luận (NCS)

- **1.** 2013 Reducing the size of databases for multirelational classification - a subgraph-based approach
- **2.** 2012 Hierarchical approaches
- **3**. 2012 From Combinatorial Optimization to Data Mining
- **4**. 2010 Fuzzy c-means and fuzzy swarm for fuzzy clustering problem
- **5.** 2008 The impact of overfitting and overgeneralization on the classification accuracy in data mining
- **6**. 2008 Higher order mining
- **7.** 2007 Cost-sensitive boosting for classification of imbalanced data
- **8.** 2006 Statistical Comparisons of Classifiers over Multiple Data Sets
- **9**. 2004 Privacy-preserving data mining - Why, how, and when
- **10**. …

Yêu cầu đối với sinh viên

- Sinh viên nên có mặt tại lớp hơn 75%.
- Sinh viên phải có mặt tại lớp vào tuần 7, 14.
- Sinh viên nên đọc trước tài liệu tham khảo cho mỗi chương.
- Sinh viên nên làm các bài tập của mỗi chương trong các tài liệu [1, 6].
- Sinh viên nên tham khảo thêm các tài liệu học tập khác, đặc biệt từ nguồn Internet.
- Sinh viên nên thực hành các công cụ liên quan.

Thực hành

- Oracle 10g/11g DBMS và Oracle 10g/11g Data Mining
  - www.oracle.com
- MS SQL Server 2005/2008 DBMS và Business Intelligence Development Studio
  - www.microsoft.com
- WEKA (the University of Waikato, New Zealand)
  - www.cs.waikato.ac.nz/ml/weka
- Other open source data mining/statistical systems such as R

A Brief History of Data Mining Society

- 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.

Where to Find References?

- 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.

Where to Find References?

- Publishers of Interest
  - ACM
  - IEEE
  - Springer
  - Elsevier

Thông tin liên lạc

- TS. Võ Thị Ngọc Châu

  (chauvtn@cse.hcmut.edu.vn, chauvtn@hcmut.edu.vn)

- Lịch tiếp sinh viên
  - Thứ 4 hàng tuần, 1:30-5:30 pm
- Tài khoản môn học
  - http://www.cse.hcmut.edu.vn/~chauvtn/data_mining/HK2%20-%202014%20-%202015/

Hỏi & Đáp …