MINING MASSIVE DATA SETS

1.Thông tin về học phần (General Information):

Tên học phần (Course name): Mining Massive Data Sets

Mã học phần (Course code): INT_E14126

Số tín chỉ (Number of credits): 3

Loại học phần (Course type): Elective

Học phần tiên quyết (Prerequisites):

Học phần trước (Previous courses):

Học phần song hành (Parallel courses):

Các yêu cầu đối với học phần (Course requirements):

- Phòng học lý thuyết (Lecture room): Projector, microphone and speaker, black board or white board.
- Phòng thực hành (Laboratory): Computers, microphone, speaker.

Giờ tín chỉ đối với các hoạt động (Teaching and Learning hours):

- Lý thuyết (Lectures): 32h
- Bài tập (Exercises): 4h
- Bài tập lớn (Projects): 4h
- Thực hành (Labs): 4h
- Tự học (Individual reading): 1h

Địa chỉ Khoa/Bộ môn phụ trách học phần (Address of the Faculty/Department in charge of the course):

- Địa chỉ (Address): Khoa Công nghệ Thông tin 1 - Học viện Công nghệ Bưu chính Viễn thông, Km10, Nguyễn Trãi, Hà Đông, Hà Nội
  Faculty of Information Technology 1 - Posts and Telecommunications Institute of Technology, Km10, Nguyen Trai Street, Ha Dong District, Hanoi.
- Điện thoại (Phone number): (024) 33510432

2. Mục tiêu học phần (Objectives)

Về kiến thức (Knowledge):

The aim of this course is to provide learners with the fundamental concepts of mining techniques and big data. The learners are required to understand:

- basic concepts and fundamental tasks in data mining.
- application of data mining and big data.
- various kind of data mining system.

Kỹ năng (Skills):

Learners will learn a set of skills to develop various kinds of data mining system. These skills consist of :

- analyzing a data mining problem.
- designing a simple data mining system.
- Evaluating the data mining system.

Thái độ, Chuyên cần (Attitude):

- Learners are required to attend all classes, do exercises and assignments.

3. Tóm tắt nội dung học phần (Description)

Understanding the components of data mining system and big data tool is the basic knowledge and skill set for every data scientist. On completion of this course, learners will be able to understand the fundamentals of data mining technique as well as to apply these techniques to solve the problem. The data mining system development consists of analyzing the problem, designing the system architecture, implementing the system, and evaluating the system.

4. Nội dung chi tiết học phần (Outlines)

Chapter 1: Introduction to Data mining

1.1. Introduction to data mining
- 1.1.1. Statistical modeling
- 1.1.2. Machine learning
- 1.1.3. Computational approaches to modeling
- 1.1.4. Summarization
- 1.1.5. Feature extraction

1.2. Statistical limits on data mining
- 1.2.1. Total information awareness
- 1.2.2. Bonferroni's principle
- 1.2.3. Example of Bonferroni's principle

1.3. Feature of data mining
- 1.3.1. Importance of word in documents
- 1.3.2. Hash functions
- 1.3.3. Indexes
- 1.3.4. Second storage
- 1.3.5. The base of natural logarithms
- 1.3.6. Power Laws

Chapter 2: Large-Scale file systems and Map-reduce

2.1. Distributed file systems
- 2.1.1. Physical organization of compute nodes
- 2.1.2. Large-scale file system organization

2.2. Map-Reduce
- 2.2.1. The map tasks
- 2.2.2. Grouping and Aggregation
- 2.2.3. The reduce tasks
- 2.2.4. Combiners
- 2.2.5. Details of Map-reduce execution
- 2.2.6. Coping with node failures

2.3. Algorithms using Map-reduce
- 2.3.1. Matrix-Vector multiplication by Map-reduce
- 2.3.2. If the vector v cannot fit in main memory
- 2.3.3. Relational-algebra operations
- 2.3.4. Computing Selections by Map-Reduce
- 2.3.5. Computing Projections by Map-Reduce
- 2.3.6. Union, Intersection, and Difference by Map-Reduce
- 2.3.7. Computing Natural Join by Map-Reduce
- 2.3.8. Generalizing the Join Algorithm
- 2.3.9. Grouping and Aggregation by Map-Reduce
- 2.3.10. Matrix Multiplication
- 2.3.11. Matrix Multiplication with One Map-Reduce Step

2.4. Algorithms using Map-reduce
- 2.4.1. Workflow systems
- 2.4.2. Recursive extension to map-reduce

6.2. Market baskets and the A-priori algorithm
    6.2.1 Representation of market-basket data
    6.2.2. Use of main memory for itemset counting
    6.2.3. Monotonicity of itemsets
    6.2.4. Tyranny of counting pairs
    6.2.5. The A-priori algorithm
    6.2.6. A-priori for all frequent itemsets
6.3. Handling larger datasets in main memory
    6.3.1. The algorithm of Park, Chen, and Yu
    6.3.2. The multistage algorithm
    6.3.3. The multihash algorithm
6.4. Limited-pass algorithm
    6.4.1. The simple, randomized algorithm
    6.4.2. Avoiding errors in sampling algorithm
    6.4.3. The algorithm of Savasere, Omiecinski and Navathe
    6.4.4. The SON algorithm and Map-reduce

Chapter 7: Clustering
7.1. Introduction to Clustering Techniques
7.2. Hierarchical clustering
    7.2.1. Hierarchical clustering in a Euclidean space
    7.2.2. Efficiency of Hierarchical clustering
    7.2.3. Alternative rules for controlling hierarchical clustering
    7.2.4. Hierarchical clustering in non-euclidean spaces
7.3. K-means algorithms
    7.3.1. K-means basics
    7.3.2. Initializing clusters for K-means
    7.3.3. Picking the right value of K
    7.3.4. The algorithm of Bradley, Fayyad, and Reina
    7.3.5. Processing data in the BFR algorithm
7.4. The CURE algorithm
    7.4.1. Initialization of CURE
    7.4.2. Completion of the CURE algorithm
7.5. Clustering in Non-Euclidean spaces
    7.5.1. Representing clusters in the GRGPF algorithm
    7.5.2. Initializing the cluster tree
    7.5.3. Adding points in the GRGPF algorithm
    7.5.4. Splitting and merging clusters
7.6. Clustering for streams and parallelism
    7.6.1. The stream-computing model
    7.6.2. A stream-clustering algorithm
    7.6.3. Initializing buckets
    7.6.4. Merging buckets
    7.6.5. Answering queries
    7.6.6. Clustering in a parallel environment

5. Học liệu (Textbooks)

5.1. Học liệu bắt buộc (Required Textbooks)

[1] Mining massive datasets – Anand Rajaraman and Jeffrey U. Ullman, Cambridge University Press, ISBN: 1107015359, 2011.

5.2. Học liệu tham khảo (Reference Textbooks)

[2] Analytics in a Big Data World: The Essential Guide to Data Science and its Applications - Bart Baesens, Wiley, ISBN: 978-1-118-89270-1, 2014.

6. Phương pháp, hình thức kiểm tra – đánh giá kết quả học tập học phần (Grading Policy)

Grading method	Percentage	Group/Individual
- Attendance	10%	Individual
- Exercises	10%	Individual
- Mid-term projects/exam	20%	Group or individual
- Final examination	60%	Individual

Trưởng Bộ môn **(Head of Department)**	**Giảng viên biên soạn** **(Lecturer)**
Ngô Xuân Bách	**Vũ Hoài Nam**