ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Lecture 1: Introduction to Web Mining

Content

1. What is WWW?
2. What is data mining?
3. What is web mining?

1. What is WWW?

- WWW (web) affects our daily life
  - Huge information, welknown, accessible and searchable
  - Billions linked web pages created by millions authors
- Web changes the way of information retrieval
  - Before, we ask friend/family for a book
  - With Internet, everything is simple with some clicks right at home or office
- Web is an important transaction channel
  - We can buy almost everything without going to the shop
  - We can easily connect to make friend, discuss, share with anyone in the world
  - Web is a virtual world reflecting the real world

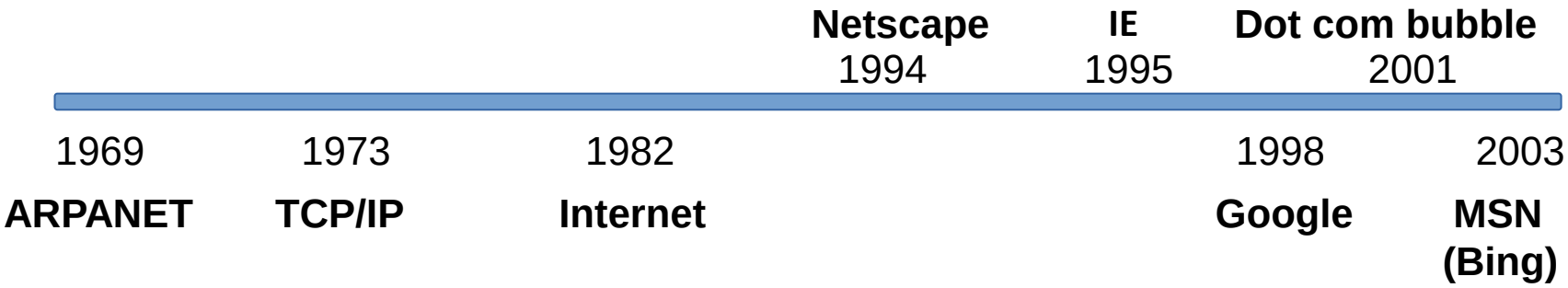
www definition

- "*Web is a computer network allowing a user from a computer to access information stored in a remote computer in the network*"
- Web is based on client-server architecture
  - User uses a client to access data in a server
  - Web browsing is done by a browser (IE, Firefox, Chrome):
    - Send request to server
    - Get response from server
    - Compile HTML
    - Display graphical content
- Web documents are hypertext allowing the author to link their document to any document on internet via hyperlink
  - To view linked document, user only needs to click on the hyperlink
  - Hypertext is invented by Ted Nelson in 1965
  - Hypertext allows to embed multimedia (image, video, voice)

Web history

- Web is invented by Tim Berners-Lee (CERN) in 1989 in a proposal on a protocol for managing distributed documents:
  - Hierarchy structure shows limitations
  - Propose a protocol to request information stored in a remote computer
  - Propose a common document format allowing a document to link to other documents
- Initial components of web:
  - Server
  - Browser
  - HTTP
  - HTML
  - URL

Web history (cont)



Web history (cont)

- Mosaic was created in 1993 at Illinois University
  - First browser with graphical interface and mouse interaction
  - Runs on UNIX, Macintosh and Windows
- In 1994, Mosaic was publicly released as Netscape
- In 1995, Internet Explorer of Microsoft was released

Web history (cont)

- ARPANET (1969) is developed by ARPA
- TCP/IP (1973) allows networks to connect
- Internet was born in 1982 based onTCP/IP

Web history (cont)

- Information shared on Web raises the need for efficient information retrieval
- Excite search engine was introduced by Stanford in 1993
- Yahoo! was found in 1994, providing information in hierarchical structure
- Google was found in 1998
- Microsoft introduces MSN in 2003 (Bing)
- W3C (The World Wide Web Consortium) was found in 1994 by MIT and CERN
  - Leading the development of Web
  - Building standards for Web
  - Setting up specifications and reference to support interaction between Web applications
- WWW was first organized in 1994
- 1995 – 2001, Web is developed and expanded
- 2001: dotcom bubble

2. What is data mining?

2.1 Definition of DM

2.2 History of DM

2.3 Data types

2.4 Discoverable patterns

2.5 Techniques in DM

2.6 Applications of DM

2.7 Challenges in DM

2.1 Definition of DM

- Known as Knowledge Discovery in Databases
- “the process of discovering useful patterns or knowledge from data sources”
- Patterns should be: correct, useful, and understandable
- Data sources: DB, text, image, Web v.v.
- DM is an interdiscipline domain including machine learning, statistics, DB, artificial intelligence, information retrieval, and visualization
- Main tasks in DM: supervised learning (classification), unsupervised learning (clustering), association rule mining, sequential mining

Definition of DM (cont)

- Data analyst selects data sources and targets based on domain knowledge
- Preprocessing:
  - Raw data is typical unsuitable for mining
  - Need cleansing to remove noise and abnormaly
  - When data is too large or contains irrelevant attribute, it requires sampling or feature/attribute selection
- DM: Applying techniques on clean data to discover knowledge
- Post processing: Choose useful pattern/knowledge using evaluation and/or visualization methods
- The process repeats until satisfied
- Traditional techniques are based on structured data. With the development of Web, semi-structured and non-structured data becomes more important

2.2 History of DM
DBMS (70'-80')

- Hierarchical DBMS
- Network DBMS
- Data modeling: entity – relation model
- Indexing and accessing
- Query language: SQL
- User interface, form, report
- Query processing and optimization
- Transaction, concurrency management, recovery
- OLTP

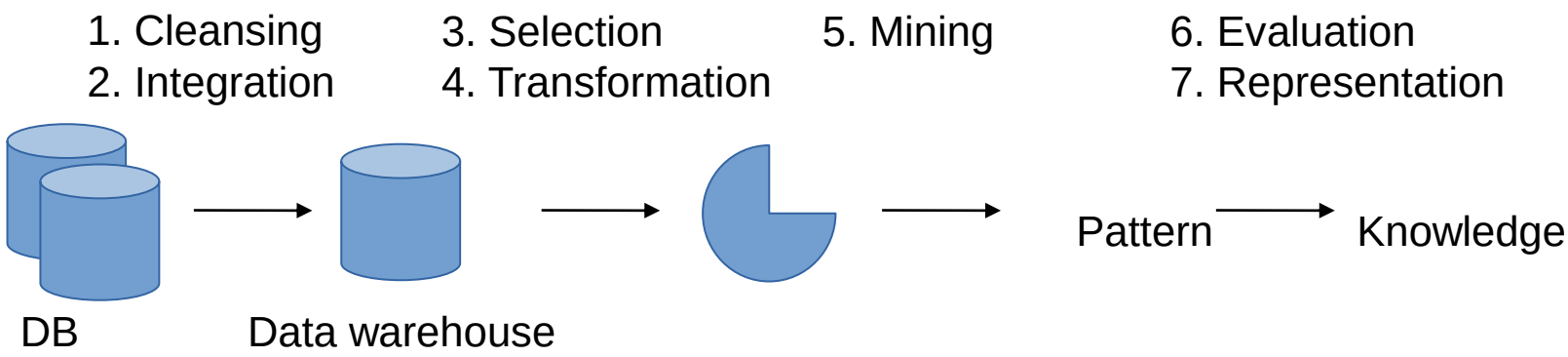
Advanced DBMS (80’-now)

- Advanced data models: Extended relation model, object relation model
- Complex data management: spatial, temporal, multimedia, sequence; structural objects, moving objects
- Data streams and meta-physic data systems
- Web DBs (XML, semantic web)
- Uncertainty data management and data cleansing
- Heterogenous source integration
- Text DBMS and integration with information retrieval
- Big data management
- Tuning DBMS and flexible systems
- Advanced query: ranking
- Cloud computing and parallel data processing
- Data policy and security

Advanced data analytics (80'-now)

- Data warehouse and OLAP
- DM and knowledge discovery: classification, clustering, abnomaly detection, association and correlation, summarization, comparision, pattern discovery, trend and variance analysis
- Complex data mining: stream, sequence, text, spatial, temporal, multimedia, web, network
- Applications of DM: business, social, commerce, bank, telecom, science and technology, social network

2.3 Data types



Phases in DM

Data from DB

- DBMS includes a set of relation data called DB and programs to manage and access data.
- The programs provide mechanisms to
  - Define DB structure and data storage
  - Describe and manage concurrency, share, distribute
  - Guarantee consistency and security
- A relational DB contains tables
  - Each table contains a set of attribute (column, field)
  - The records (row) in a table represents an object identified by a unique key and described by attributes
- DBs are accessed using queries
  - Queries are transformed into a set of relational operations like combination, selection and are then optimized
  - A query allow to retrieve a specific part of data
- In relational DB mining, main tasks are detecting trend, data patterns or variance analysis

Data warehouse

- Data warehouse is an information storage collected from multiple sources and is stored in a unique schema
- Data warehouse is constructed by a process including cleansing, integration, transformation, load, and regular data refresh.
- Data in data warehouse is organized in and object-oriented approach. Data is *summarized* and is stored to provide information in historical view for decision support (for organization)
- Data warehouse is modeled by a multidimensional data, called data cube
  - Each dimension is one or a set of attributes in the schema
  - Each cell stores a value like count or sum
  - A data cube provides a multidimensional view and allow pre-computation and quick access of *summarized* data

Data warehouse (cont)

- Data warehouse supports OLAP
- OLAP is based on domain knowledge to represent data in various abstract levels. Two fundamental operations in OLAP are drill-down and roll-up allowing users to observe data in different summarization level. E.g:
  - Drill-down observes monthly data from quarter data
  - Roll-up observes country data from province data
- General multi-dimensional observation techniqes can combine multiple dimensions at different detail levels. That leads to discovery important patterns

Transaction data

- A record in a transaction DB represents a transaction, including a unique identifier and components involving in the transaction
- Typical transaction types include transfer, purchase, buying, order, mouse click
- Transactional DB might miss additional tables like seller information or branch information
- Transactional data mining focuses on detection frequent sets. E.g, answering the question "*Which products are bought (buy customers) together?*"

Other data types

- Temporal data (stock), sequence (biology), space (map), industry design (construction design, system components, board), hypertext and multimedia, graph and networks
- Challenges on data structure (sequence, tree, graph, network) and semantic (order, connectivity)
- Some applications:
  - Temporal data: Detect transaction trend to plan customer care, money distribution; Detect stock trend for investment; Abnomaly detection based on clustering or occurrence frequency
  - Spatial data: Estimate poverty rate based on distance to main road; Detect community based on distance between items
  - Textual data: Estimate customer satisfaction based on review content
  - Multimedia: Object detection and classification, goal detection in videos

2.4 Discoverable patterns

- Functions in data mining: Describe and discriminate data, frequent pattern mining, association and correlation, classification and regression, clustering and abnomaly detection
- Two types of data mining tasks: descriptive and predictive

Describe and discriminate data

- Data description is summarizing common characteristics or features of a class of data
- Data statistics
- Roll-up in OLAP describe data in a specific dimension. E.g: Summarizing common characteristics of customer paying more than 100 million/year → 40-50 years old, job (could drill down in job dimension)
- Attribute oriented inference allows generalizing and describing data without user interaction step by step
- Data discrimination is comparing general features of a class with other reflective classes.
  - Reflective classes are provided by user. Their data could be retrieved from DB.
  - Comparative description could be represented by rules
  - E.g: Compare customer who frequently vs rarely buy hi-tech → 80% customers frequently buy hi-tech are at 20-40 and have college degree, 60% customers rarely buy hi-tech don't have college degree, drill-down in *education* or *income* could have other useful information

Frequent patterns, association and correlation

- *Set of frequent items* includes items frequently occurring together in a transactional DB (e.g, milk and bread are frequently bought together at food stores); *frequent sequential patterns* (e.g: customers usually buy computer, camera and memory card in that order); *frequent structural patterns*
- Association analysis

  mua(X, máy tính) → mua(X,phần mềm) [support = 1%, confidence = 50%)

  - Những sản phẩm nào thường được mua cùng nhau trong cùng một giao dịch
  - X: khách hàng
  - Độ tự tin (độ chắc chắn) (confidence/certainty) thể hiện khả năng khách hàng mua phần mềm nếu biết khách hàng mua máy tính
  - Độ hỗ trợ (support) thể hiện tỉ lệ giao dịch mà máy tính và phần mềm được mua cùng nhau trên tổng số giao dịch được phân tích
  - Luật kết hợp theo một chiều mua

  máy tính → phần mềm [1%, 50%]

  - Luật kết hợp đa chiều (liên quan đến nhiều thuộc tính)

  tuổi(X,20..29)^thu nhập(X,20..30tr)→mua(X,laptop) [2%,60%]

  - Có 2% có tuổi 20-29, thu nhập 20-30tr đã mua laptop trong tổng số khách hàng được phân tích (thu thập từ CSDL quan hệ)
  - Có 60% khả năng những người trong độ tuổi 20-29 và có thu nhập 20-30tr sẽ mua laptop
  - Luật kết hợp không thỏa mãn nếu ở dưới ngưỡng độ hỗ trợ tối thiểu (minimum support threshold) và đồ tự tin tối thiểu (minimum confidence threshold)

Classification and regression

- Classification is the process of searching a model to describe and discriminate data classes
  - Model is induced from training data (items with class labels)
  - Model is used to predict class labels of unknown items
  - Model could be represented by rules, decision trees, math/probabilistic equations, or neural networks
- Regression modelizes continuous functions

Clustering

- Items are clustered by maximizing intra-cluster similarity and minimizing inter-cluster similarity
- A cluster could be considered as a class to induce regularity
- Clustering is used in taxonomy construction

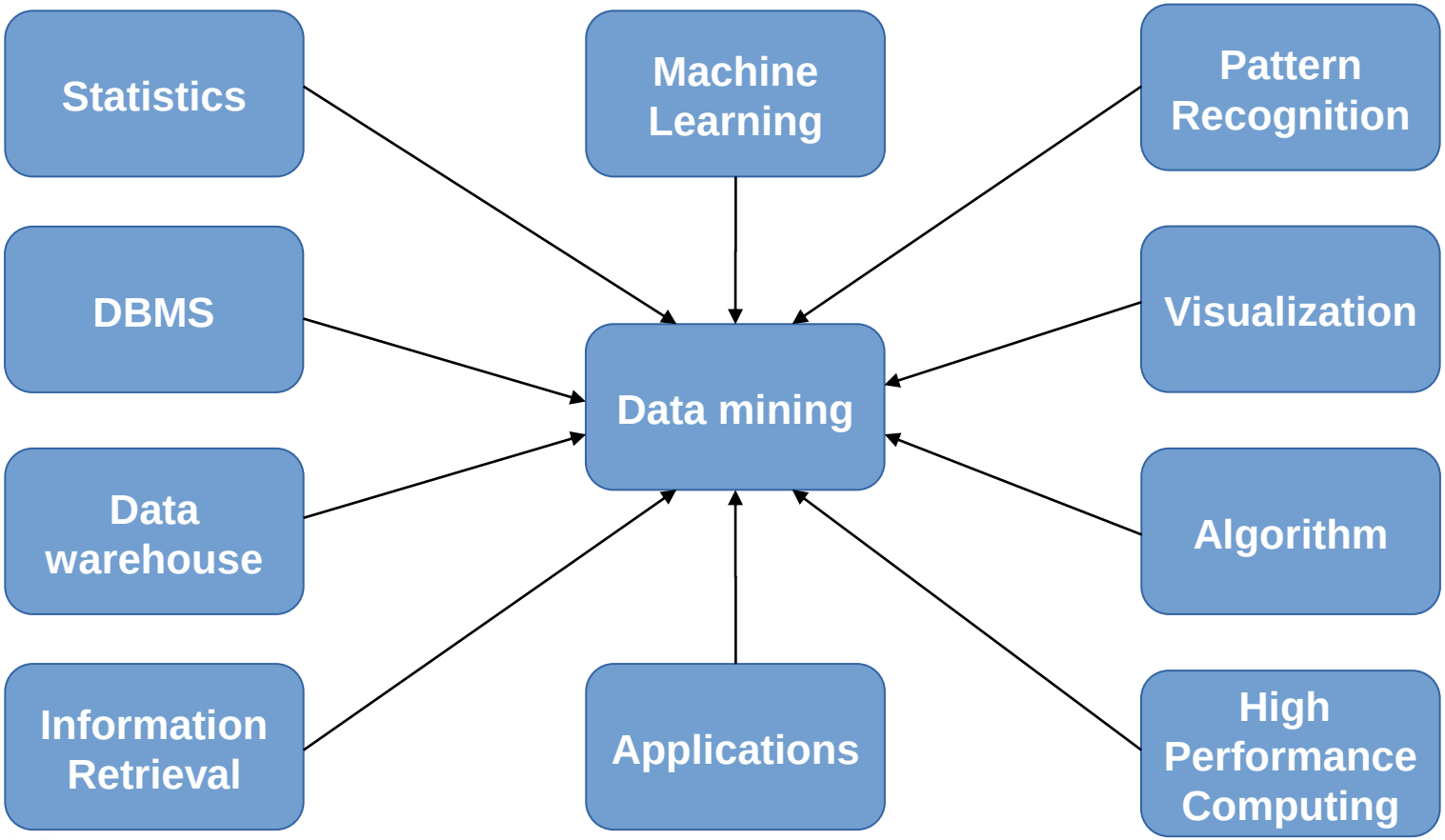
Abnomaly detection

- Abnomaly are items that don't fit to common behaviour or model of data set
- Most DM techniques consider abnomaly as noise or exception. However, in some applications like fraud detection, these items are importants
- Typical statistics tests could detect abnomaly based on data modeling following a distribution or a probabilistic model
- Distance-based methods detect abnormal items far from data clusters
- Density-based methods could detect abnormal items in a region where general statistic distribution couldn't detect
- E.g: Fraud card transaction could be detected based on transaction volume when comparing to normal transactions; or based on information of location, transaction type or transaction frequency

Potential patterns

- Potential patterns
  - *i)* understandable *ii)* correct in new data or pivot data with a certainty *iii)* potentially useful *iv)* novel
  - or confirm an assumption from user
- A potential pattern represents knowledge
- Objective metrics based on structure and statistics of patterns
  - In association analysis, support of an association rule $X => Y$ represents transaction fraction satisfying the rule. Confidence represents the probability $P(Y|X)$
  - In classification, accuracy of a classification rule represents data fraction correctly classified; coverage represents data fraction satisfied the rule
- Subjective metrics based on user belief on data
  - Not expected by users (vs belief)
  - Provide strategic information for decision support (e.g "a earthquake is typical followed by aftershocks")
  - As expectation to confirm an assumption from users
- A DM system could generate all potential patterns?
- A DM system could generate only potential patterns?

2.5 Techniques in DM



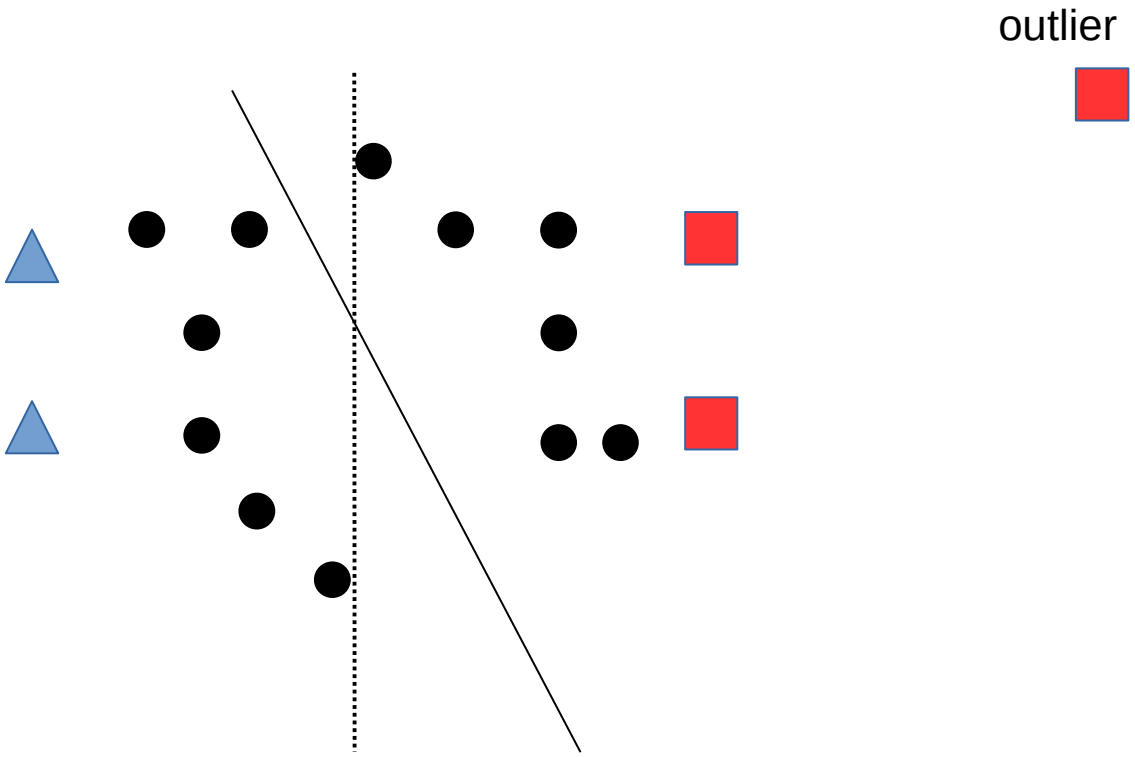
Statistics

- Study the collection, analysis, explanation and visualization of data
- Statistic model is a set of function describing behaviour of items in a target class in terms of random variables and relevant probability distribution. Fnferential statistics modelize data using stochastic process and uncertainty of observations
- The output of data description and discrimination coud be statistic models
- Pattern mining coud use statistic models to detect and process noise and missing data
- The output of DM could be verified by statistical hypothesis test. The output is statistically significant it has a small chance to be generate by random
- Most statistic methods have high computational complexity

Machine learning

- Study the learning ability of machine (or improve learning results) using data
- A main research direction is to make computer programs to automatically recognize complex patterns in data and make intelligent decisions based on data
- Supervised learning (classification) is based on labeled data
- Unsupervised learning (clustering) uses unlabeled data
- Semi-supervised learning is based on both labeled and unlabeled data. An approach is to use labeled data to build model and use unlabeled data to tune class boundary
- Active learning: Allow user participating in learning process. It aims at optimizing model quality under a constraint on the number of labeled data

Machine learning (cont)



..... Boundary based on labeled data

——— Boundary based on unlabeled data

DBMS and data warehouse

- DBMSs create, operate and use Dbs for organizations and end-users. DBMSs establish principles on data modeling, query language, query processing, optimization, data storage, indexing and acess. DBMSs could process data with complex structure
- DM require processing big data, in real-time and streaming data. DM could applied techniques in DBMS and integrate DM functions into DBMSs
- Data warehouse integrates data from multiple sources, supports OLAP operations and multi-dimensional DM

Information retrieval

- Science of searching for documents or information inside documents
- Documents are texts or multimedia (image, sound, video)
- Two assumptions
  i) data is non-structural
  ii) query includes kewords and doesn't have complex structure
- Topic models detect main topics in a document collection as long as in each document

2.6 Applications of DM
Business intelligence

- Organizations require knowledge about their business environment: customer, market, supply, resource and competitors
- Business intelligence provides historical and present view, and predicts organization operations
- Business intelligence includes report, OLAP, business performance management, competitive intelligence, standardization and predictive analysis
- DM helps organization effectively analyze market, compare customer feedbacks to similar products, analyze strength and weakness of competiors, keep close customers, and make wise business decision
- OLAP tools based on data warehouse and multi-dimensional DM; classification and predictive analysis play a center role in analyzing market, supply and sale; clustering is applied in customer relationship management to group similar customers; data description techniques could be use to understand customer groups to provide suitable services

Web search

- DM techniques are applied in search engine including crawling (which pages and frequency), indexing (which page and which part), search (ranking, advertisement, personalization)
- Search engines use cloud computing infrastructure with thousands nodes
- Search models and query classifiers are built offline; queries are processed online; models and classifiers are update based on change of queries and web data
- Personalizing search results give answers based on user profile and search history; rare queries are challenging

2.7 Challenges in DM
Mining methods

- Mining new and various knowlege: Combine clustering and ranking to create high-quality clusters and ranking in large networks
- Mining in multi-dimensional space by combining attribute dimensions
- Combining techniques from relevant fields: natural language processing in text mining, software engineering in bug mining
- Mining in semantically linked environment: knowledge from a set of items enforce mining knowledge from related items
- Noise, uncertainty, missing data: Could negatively effect DM and generate wrong patterns. Data cleansing, preprocessing, abnomaly detection, uncertainty inference could be applied.
- Pattern evaluation based on subjective criteria; guide the mining process with potential patterns or with user to increase pattern quality and limit search space

User interaction

- Flexible user interface, help users easily interact with system.
  - Data sampling, analyzing common characteristics of data, estimating mining results
  - Changing search objectives, tuning mining requirement
- Fundamental knowledge, constraints, regularity and other information of application domain need to be integrated into the system to evaluate patterns or to guide the mining process
- Develop DM languages to do ad hoc mining taks and to support data description
- Represent and visualize results lively and flexibly so that users understand the results and use them directly in decision making

Efficace and extendability

- DM algorithms need to be effective and extendable to extraction information from a large volume of data in a dynamic environment. Execution time needs to be predictable , short, and acceptable
- Parallel and distributed computing: Data is separated into pieces and is processed in parallel processes; The processes could communicate to share data and information; Their results are then combined
- Incremental mining allow updating new data without restart mining process from the beginning to enrich mined knowledge

Effects of DM on the society

- Limit bad effects, bring benefit to the society
- Evaluate data sensitivity, ensure data policy
- Integrate DM into current systems to increase service quality without requiring user knowledge on DM techniques

3. What is web data mining?

- Web is the largest public data source
- Main characteristics of web data:
  1. The volume of information/data in Web is increasingly huge. Information coverage is large and various. We could find almost everything in web
  2. Many data types: Structured tables, semi-structured web pages, non-structural texts, multimedia files (image, sound, video)
  3. Information in Web is heterogeneous. Different web pages could show the same information or similar information in different format. This makes information integration complicated.
  4. Information in Web is linked. Hyperlinks between web pages exist inside a website or between websites. In a website, hyperlinks form an (internal) information organization mechanism. Between websites, hyperlinks implicitly the role of web pages (densely linked web pages typically have high quality and/or high impact)

What is web data mining (cont)

5. Information on Web is misleading:
   - A typical web page contains pieces of information (main content, redirection, ad, copyright, user policy, v.v.)
   - Web basically doesn't have content proof, anyone could publish anything in Web. Most content in Web has low quality, errors, or incorectness
6. Web is a business and commercial channel: All commercial website provide users with buying, purchase, information collection. Websites need to automate tasks (e.g suggestion, customer care) to improve effectiveness.
7. Web is dynamic. Content in Web continuously changes. Follow and monitor change is an important requirement to web apps
8. Web is a virtual society. It is not only data, information, services, but also interaction between human, organization and automatic systems. User could easily and immediately connect to anyone from anywhere in the world. They could express their opinion on anything in forums, blogs, review pages, or social networks. That information create a new type of data for DM or social network analysis

What is web data mining (cont)

- Study information distribution in the web
- Study characteristics and classify web pages
- Monitor web evolution
- Study relationship between web agents like web pages, users, communities and activities in the web

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attentions!