BÀI GiẢNG NHẬP MÔN KHAI PHÁ DỮ LIỆU

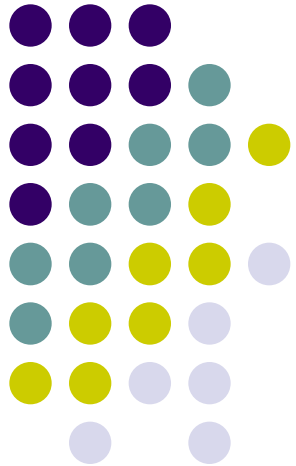
CHƯƠNG 1. GIỚI THIỆU CHUNG VỀ KHAI PHÁ DỮ LIỆU

PGS. TS. HÀ QUANG THỤY

HÀ NỘI 02-2011

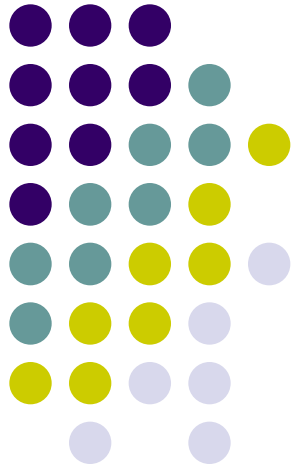
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐẠI HỌC QUỐC GIA HÀ NỘI

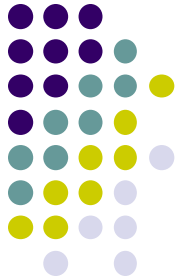


Nội dung

1. Nhu cầu của khai phá dữ liệu (KPDL)
2. Khái niệm KPDL và phát hiện tri thức trong CSDL
3. KPDL và xử lý CSDL truyền thống
4. Một số ứng dụng điển hình của KPDL
5. Kiểu dữ liệu trong KPDL
6. Các bài toán KPDL điển hình
7. Tính liên ngành của KPDL

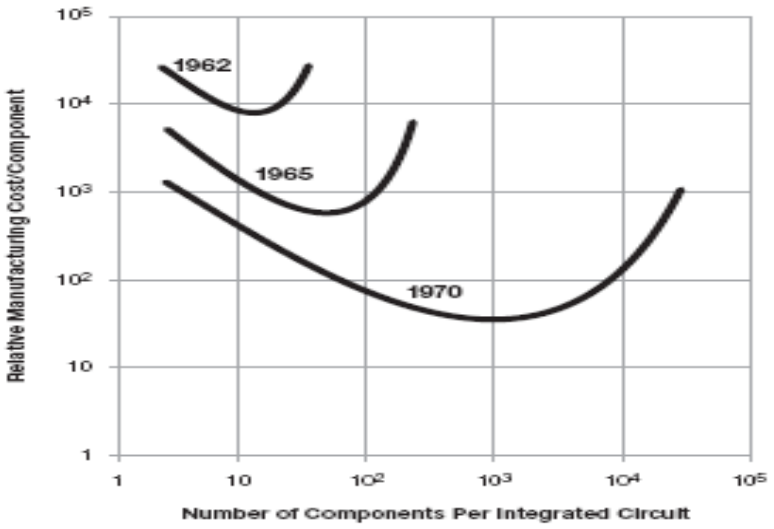


1. Nhu cầu về khai phá dữ liệu



- Sự bùng nổ dữ liệu
  - Lý do công nghệ
  - Lý do xã hội
  - Thể hiện
- Ngành kinh tế định hướng dữ liệu
  - Kinh tế tri thức
  - Phát hiện tri thức từ dữ liệu

Bùng nổ dữ liệu: Luật Moore



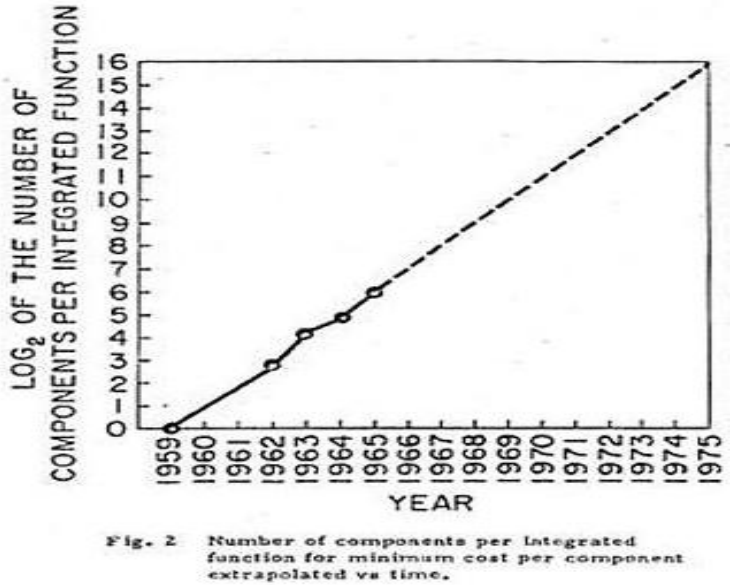
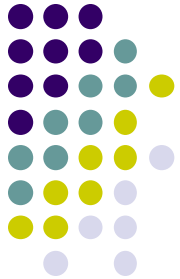


Fig. 2 Number of components per integrated function for minimum cost per component extrapolated vs time.

- Xuất xứ
  - Gordon E. Moore (1965). Cramming more components onto integrated circuits, *Electronics*, **38** (8), April 19, 1965. *Một quan sát và dự báo*
- "Phương ngôn 2x
  - Số lượng bán dẫn tích hợp trong một chíp sẽ tăng gấp đôi sau khoảng hai năm
  - Chi phí sản xuất mạch bán dẫn với cùng tính năng giảm một nửa sau hai năm
  - Phiên bản 18 tháng: rút ngắn chu kỳ thời gian

Luật Moore & công nghiệp điện tử



- Dẫn dắt ngành công nghệ bán dẫn
  - Mô hình cơ bản cho ngành công nghiệp mạch bán dẫn
  - *"Định luật Moore vẫn tạo khả năng cơ bản cho sự phát triển của chúng tôi, và nó vẫn còn hiệu lực tốt tại Intel… Định luật Moore không chỉ là mạch bán dẫn. Nó cũng là cách sử dụng sáng tạo mạch bán dẫn".* Paul S. Otellini, Chủ tịch và Giám đốc điều hành Tập đoàn Intel
  - *"toàn bộ chu trình thiết kế, phát triển, sản xuất, phân phối và bán hàng được coi là có tính bền vững khi tuân theo định luật Moore… Nếu đánh bại định luật Moore, thị trường không thể hấp thụ hết các sản phẩm mới, và kỹ sư bị mất việc làm. Nếu bị tụt sau định luật Moore, không có gì để mua, và gánh nặng đè lên đôi vai của chuỗi các nhà phân phối sản phẩm"*. Daniel Grupp, Giám đốc PT công nghệ tiên tiến, Acorn Technologies, Inc. (http://acorntech.com/)
- Thúc đẩy công nghệ xử lý, lưu giữ và truyền dẫn dữ liệu
  - Công nghệ bán dẫn là nền tảng của công nghiệp điện tử.
  - Định luật Moore với công nghiệp phần cứng máy tính: bộ xử lý Intel trong 40 năm qua (trang tiếp theo).
  - Bùng nổ về năng lực xử lý tính toán và lưu trữ dữ liệu.
  - Tác động tới sự phát triển công nghệ cơ sở dữ liệu (tổ chức và quản lý dữ liệu) và công nghệ mạng (truyền dẫn dữ liệu)

Luật Moore: Bộ xử lý Intel

Microprocessor	Year of Introduction	Transistors
4004	1971	2,300
8008	1972	2,500
8080	1974	4,500
8086	1978	29,000
Intel286	1982	134,000
Intel386™ processor	1985	275,000
Intel486™ processor	1989	1,200,000
Intel® Pentium® processor	1993	3,100,000
Intel® Pentium® II processor	1997	7,500,000
Intel® Pentium® III processor	1999	9,500,000
Intel® Pentium® 4 processor	2000	42,000,000
Intel® Itanium® processor	2001	25,000,000
Intel® Itanium® 2 processor	2002	220,000,000
Intel® Itanium® 2 processor (9MB cache)	2004	592,000,000

"Another decade is probably straightforward...There is certainly no end to creativity". Gordon Moore, Intel Chairman Emeritus of the Board Speaking of extending Moore's Law at *the International Solid-State Circuits Conference (ISSCC)*, February 2003.

Hệ thống ước và bội đơn vị đo

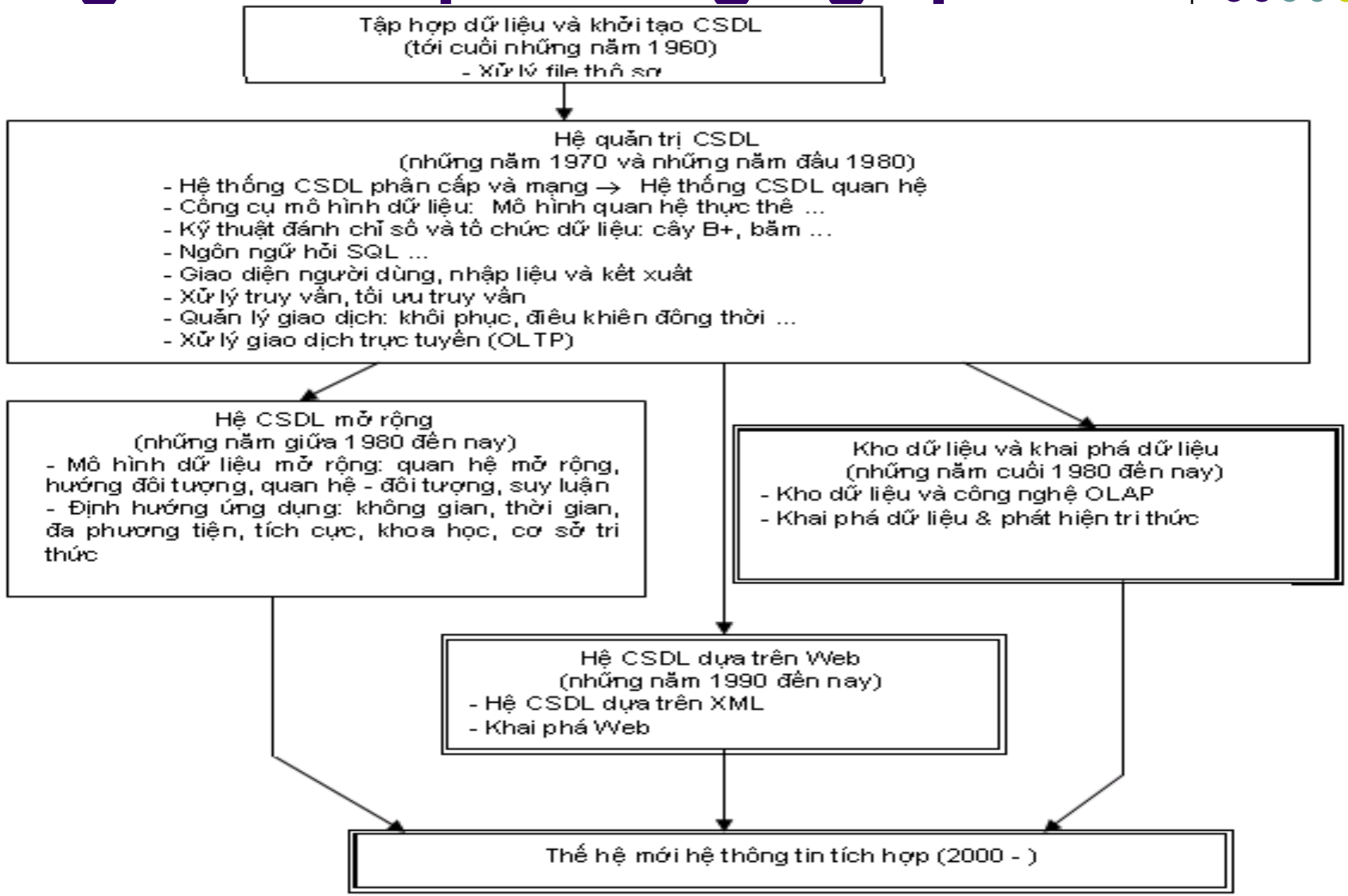
Biểu thức	Giá trị chính xác	Tiền tố	Biểu thức	Giá trị chính xác	Tiền tố
10^{-3}	0.001	mili	10^{3}	1.000	Kilo
10^{-6}	0.000001	micro	10^{6}	1.000.000	Mega
10^{-9}	0.000000001	nano	10^{9}	1.000.000.000	Giga
10^{-12}	0.000000000001	pico	10^{12}	1.000,000.000.000	Tera
10^{-15}	0.000000000000001	femto	10^{15}	1.000.000.000.000.000	Peta
10^{-18}	0.000000000000000001	atto	10^{18}	1.000.000.000.000.000.000	Exa
10^{-21}	0.000000000000000000001	zepto	10^{21}	1.000.000.000.000.000.000.000	Zetta
10^{-24}	0.000000000000000000000001	yocto	10^{24}	1.000.000.000.000.000.000.000.000	Yotta

Giá trị, cách đọc các bội và ước điển hình

Thiết bị thu thập – lưu trữ dữ liệu

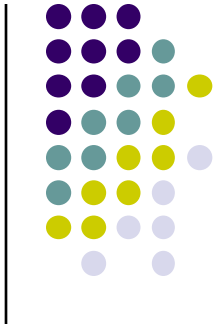
- Năng lực số hóa
  - Thiết bị số hóa đa dạng
  - Mọi lĩnh vực Quản lý, Thương mại, Khoa học…
  - Một ví dụ điển hình: SDSS
- Sloan Digital Sky Survey
  - http://www.sdss.org/
  - Đã tạo bản đồ 3-chiều có chứa hơn 930.000 thiên hà và hơn 120.000 quasar
  - Kính viễn vọng đầu tiên
    - Làm việc từ 2000
    - Vài tuần đầu tiên: thu thập dữ liệu thiên văn học = toàn bộ trong quá khứ. Sau 10 năm: 140 TB
  - Kính viễn vọng kế tiếp
    - Large Synoptic Survey Telescope
    - Bắt đầu hoạt động 2016. Sau 5 ngày sẽ có 140 TB

Bùng nổ dữ liệu: Công nghệ CSDL



- Tiến hóa công nghệ CSDL [HK0106]

Công nghệ CSDL: Một số CSDL lớn



- Tốp 10 CSDL lớn nhất
  - http://top-10-list.org/2010/02/16/top-10-largest-databases-list/
  - Library of Congress: 125 triệu mục; Central Intelligence Agency (CIA): 100 “hồ sơ: thống kê dân số, bản đồ…” hàng tháng; Amazon: 250 triệu sách, 55 triệu người dùng, 40TB; ChoicePoint: 75 lần Trái đất – Mặt trăng; Sprint: 70.000 bản ghi viễn thông; Google: 90 triệu tìm kiếm/ngày; AT&T: 310TB; World Data Centre for Climate
- Trung tâm tính toán khoa học nghiên cứu năng lượng quốc gia Mỹ
  - National Energy Research Scientific Computing Center: NERSC
  - tháng 3/2010: khoảng 460 TB
  - http://www.nersc.gov/news/annual_reports/annrep0809/annrep0809.pdf
- YouTube
  - Sau hai năm: hàng trăm triệu video
  - dung lượng CSDL YouTube tăng gấp đôi sau mỗi chu kỳ 5 tháng

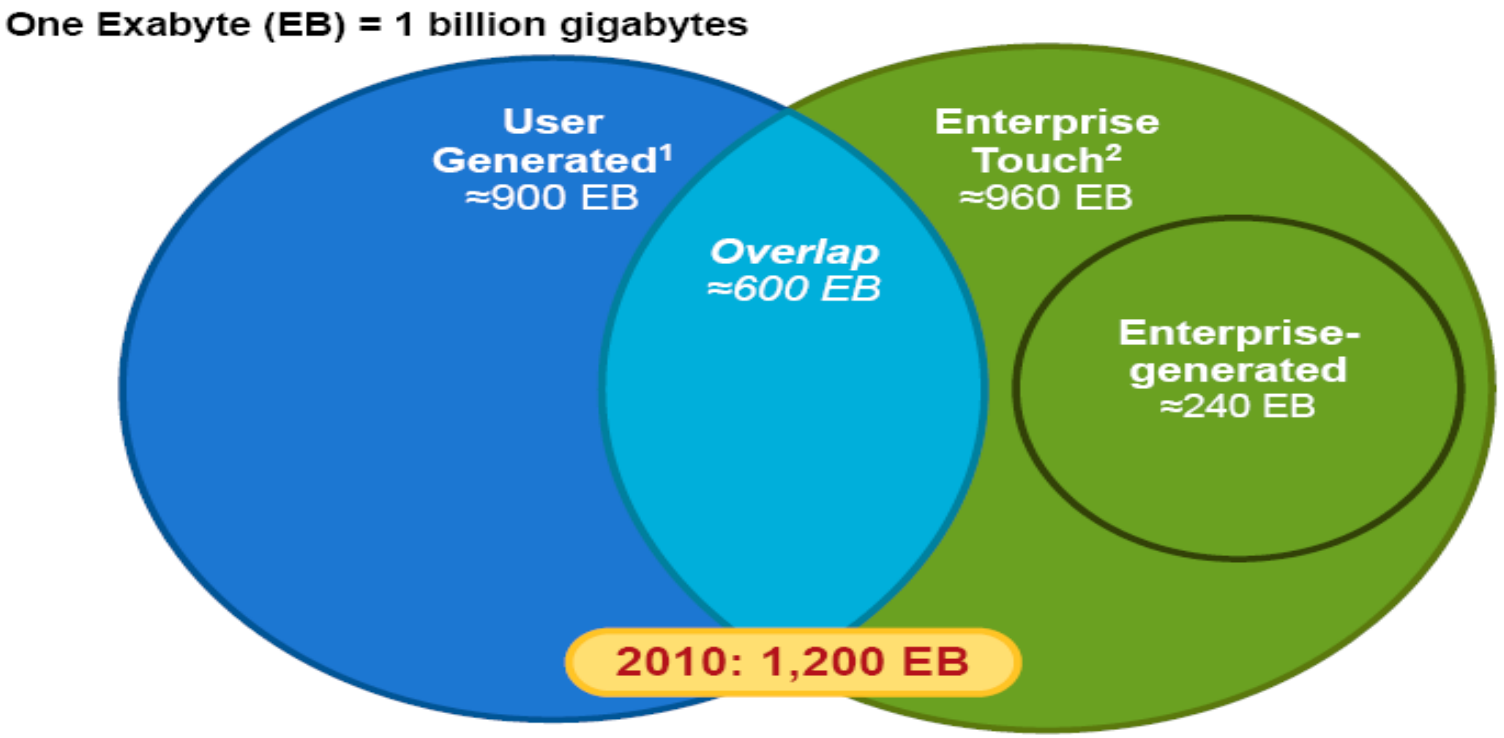
Bùng nổ dữ liệu: Công nghệ mạng

IP Traffic, 2009–2014							
	2009	2010	2011	2012	2013	2014	CAGR 2009–2014
By Type (PB per Month)							
Internet	10,942	15,205	21,181	28,232	36,709	47,176	34%
Managed IP	3,652	4,963	6,771	8,851	11,078	13,199	29%
Mobile Data	91	228	538	1,158	2,132	3,528	108%
By Segment (PB per Month)							
Consumer	11,602	16,534	23,750	32,545	43,117	55,801	37%
Business	3,083	3,862	4,740	5,697	6,801	8,103	21%
By Geography (PB per Month)							
North America	5,115	7,091	10,051	12,988	16,136	19,019	30%
Western Europe	3,495	4,818	6,712	9,261	12,417	16,158	36%
Asia Pacific	3,920	5,367	7,295	9,815	12,985	17,421	35%
Japan	1,068	1,539	2,149	2,855	3,591	4,300	32%
Latin America	438	680	1,026	1,527	2,274	3,479	51%
Central Eastern Europe	493	678	938	1,306	1,815	2,510	38%
Middle East and Africa	157	223	319	490	700	1,018	45%
Total (PB per Month)							
Total IP Traffic	**14,686**	**20,396**	**28,491**	**38,242**	**49,919**	**63,904**	**34%**

Source: Cisco VNI, 2010

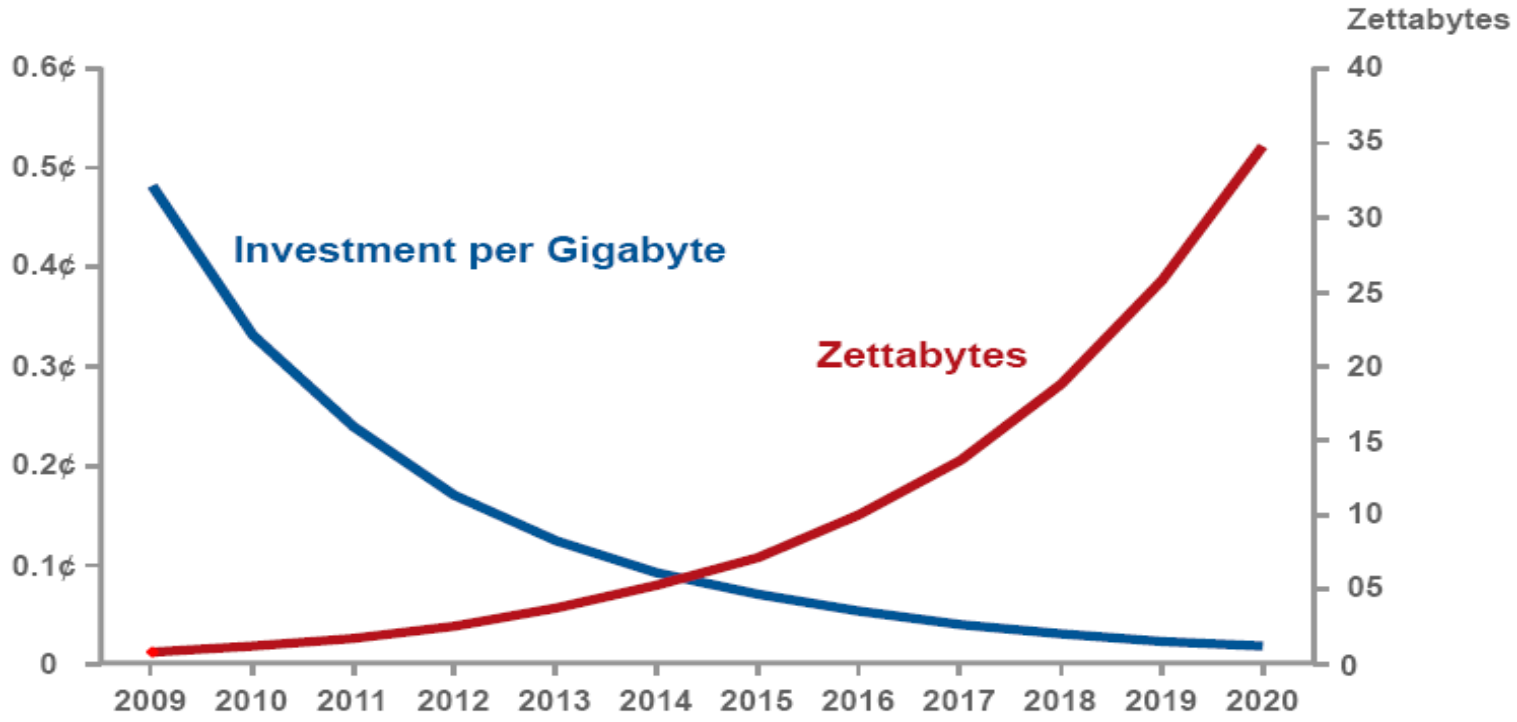
- Tổng lượng giao vận IP trên mạng
  - Nguồn: Sách trắng CISCO 2010
  - 2010: 20.396 PB/tháng, 2009-2014: tăng trung bình hàng năm 34%
- Web
  - 13 tỷ rưỡi trang web được đánh chỉ số (ngày 23/01/2011)
  - Nguồn: http://www.worldwidewebsize.com/

Bùng nổ dữ liệu: Tác nhân tạo mới



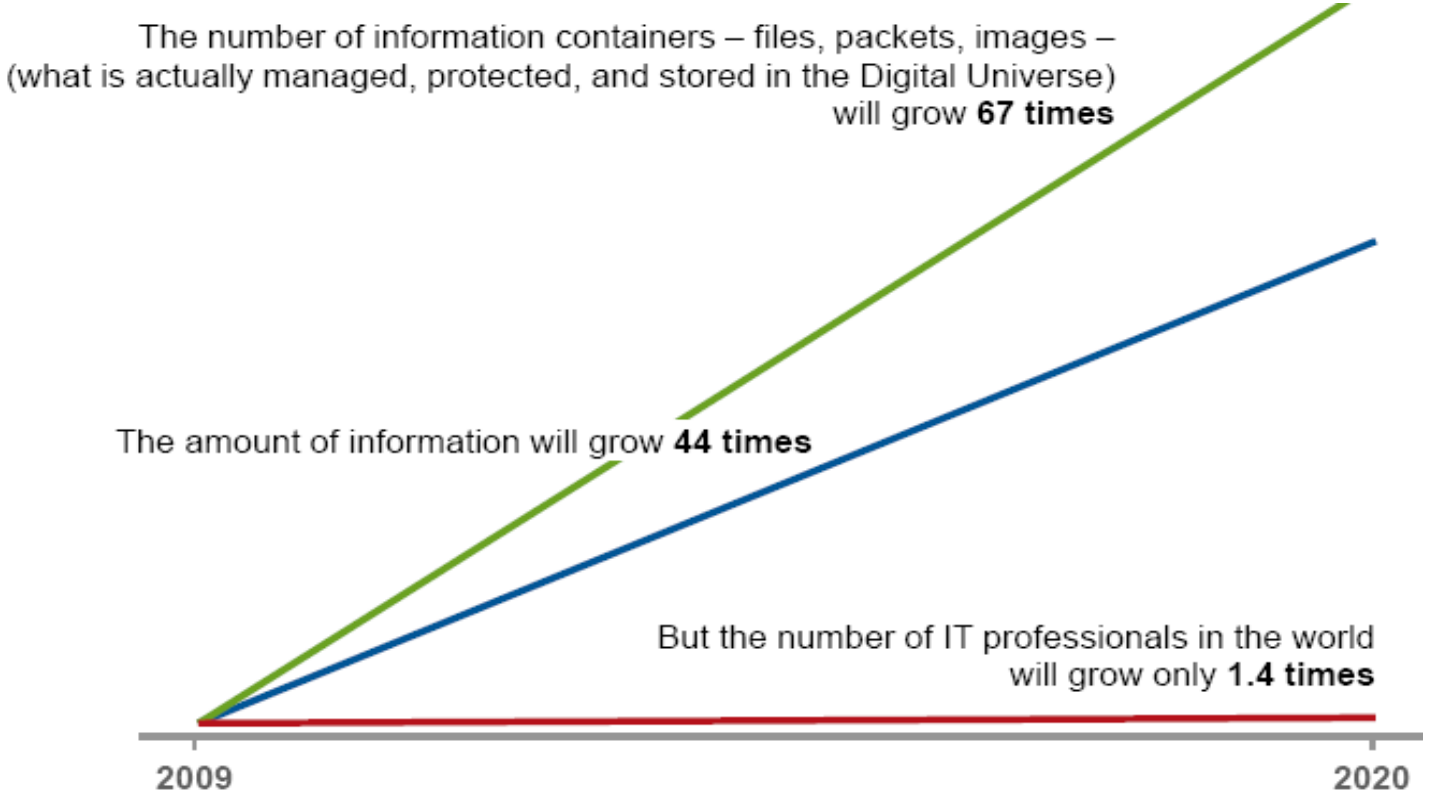
- Mở rộng tác nhân tạo dữ liệu
  - Phần tạo mới dữ liệu của người dùng ngày càng tăng
  - Hệ thống trực tuyến người dùng, Mạng xã hội…
  - Mạng xã hội Facebook chứa tới 40 tỷ ảnh
  - 2010: 900 EB do người dùng tạo (trong 1260 EB tổng thể). *Nguồn*: IDC Digital Universe Study, sponsored by EMC, May 2010

Bùng nổ dữ liệu: Giá thành và thể hiện



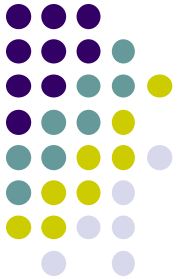
- *Nguồn:* IDC Digital Universe Study, sponsored by EMC, May 2010
- Giá tạo dữ liệu ngày càng rẻ hơn
  - Chiều hướng giá tạo mới dữ liệu giảm dần
  - 0,5 xu Mỹ/1 GB vào năm 2009 giảm tới 0,02 xu Mỹ /1 GB vào năm 2020
- Dung lượng tổng thể tăng
  - Độ dốc tăng càng cao
  - Đạt 35 ZB vào năm 2020

Nhu cầu nắm bắt dữ liệu



- Bùng nổ dữ liệu với tăng trưởng nhận lực CNTT
  - Dung lượng thông tin tăng 67 lần, đối tượng dữ liệu tăng 67 lần
  - Lực lượng nhân lực CNTT tăng 1,4 lần
  - *Nguồn*: IDC Digital Universe Study, sponsored by EMC, May 2010.

Nhu cầu thu nhận tri thức từ dữ liệu



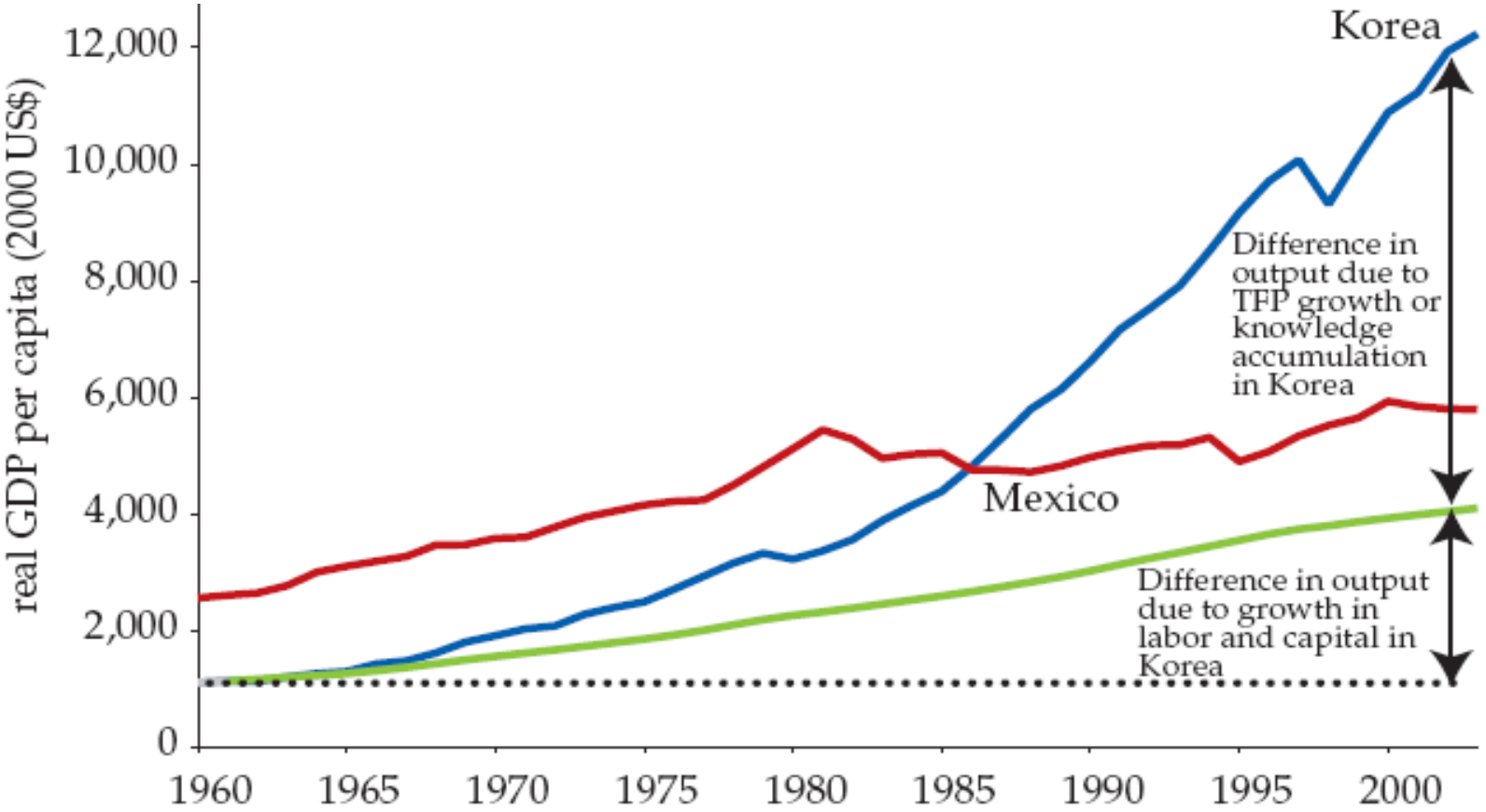
- Jim Gray, chuyên gia của Microsoft, giải thưởng Turing 1998
  - *"Chúng ta đang ngập trong dữ liệu khoa học, dữ liệu y tế, dữ liệu nhân khẩu học, dữ liệu tài chính, và các dữ liệu tiếp thị. Con người không có đủ thời gian để xem xét dữ liệu như vậy. Sự chú ý của con người đã trở thành nguồn tài nguyên quý giá.*

    Vì vậy, chúng ta phải tìm cách tự động phân tích dữ liệu, tự động phân loại nó, tự động tóm tắt nó, tự động phát hiện và mô tả các xu hướng trong nó, và tự động chỉ dẫn các dị thường.

    Đây là một trong những lĩnh vực năng động và thú vị nhất của cộng đồng nghiên cứu cơ sở dữ liệu. Các nhà nghiên cứu trong lĩnh vực bao gồm thống kê, trực quan hóa, trí tuệ nhân tạo, và học máy đang đóng góp cho lĩnh vực này. Bề rộng của lĩnh vực làm cho nó trở nên khó khăn để nắm bắt những tiến bộ phi thường trong vài thập kỷ gần đây" [HK0106].
- Kenneth Cukier,
  - *"Thông tin từ khan hiếm tới dư dật. Điều đó mang lại lợi ích mới to lớn… tạo nên khả năng làm được nhiều việc mà trước đây không thể thực hiện được: nhận ra các xu hướng kinh doanh, ngăn ngừa bệnh tật, chống tội phạm …*

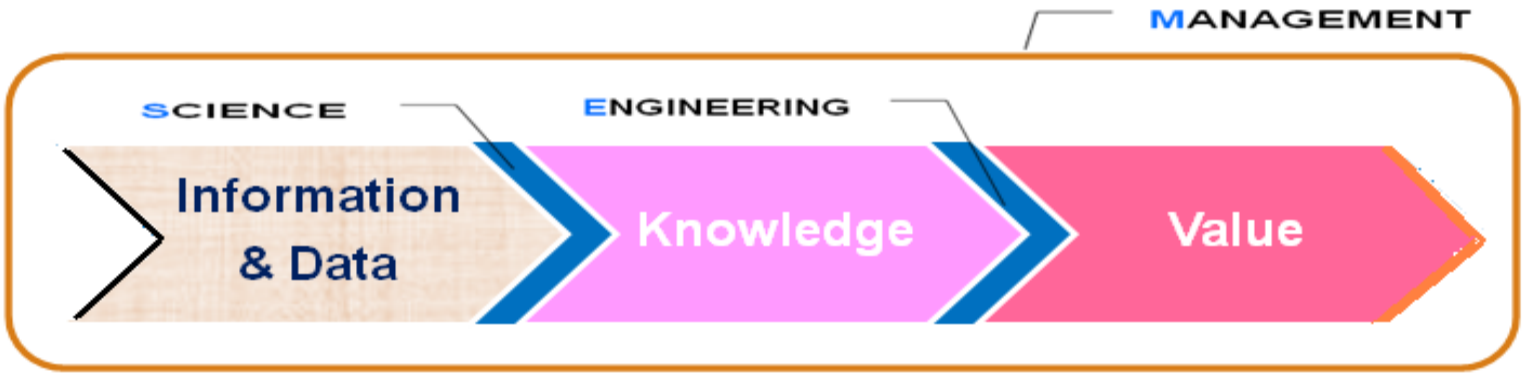
    Được quản lý tốt, dữ liệu như vậy có thể được sử dụng để mở khóa các nguồn mới có giá trị kinh tế, cung cấp những hiểu biết mới vào khoa học và tạo ra lợi ích từ quản lý". http://www.economist.com/node/15557443?story_id=15557443

Kinh tế tri thức



- Kinh tế tri thức
  - Tri thức là tài nguyên cơ bản
  - Sử dụng tri thức là động lực chủ chốt cho tăng trưởng kinh tế
- *Hình vẽ*: Năm 2003, đóng góp của tri thức cho tăng GDP/đầu người của Hàn Quốc gấp đôi so với đóng góp của lao động và vốn. TFP: Total Factor Productivity (The World Bank. *Korea as a Knowledge Economy*, 2006)

Kinh tế dịch vụ: Từ dữ liệu tới giá trị



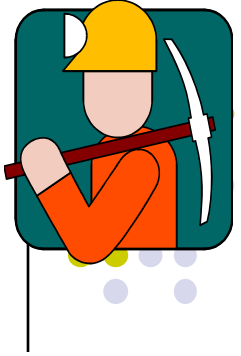
- Kinh tế dịch vụ
  - Xã hội loài người chuyển dịch từ kinh tế hàng hóa sang kinh tế dịch vụ. Lao động dịch vụ vượt lao động nông nghiệp (2006).
  - Mọi nền kinh tế là kinh tế dịch vụ.
  - Đơn vị trao đổi trong kinh tế và xã hội là dịch vụ
- Dịch vụ: dữ liệu & thông tin [redacted] tri thức [redacted] giá trị mới
  - Khoa học: dữ liệu & thông tin [redacted] thức
  - Kỹ nghệ: tri thức [redacted] ch vụ
  - Quản lý: tác động tới toàn bộ quy trình thi hành dịch vụ

Jim Spohrer (2006). A Next Frontier in Education, Employment, Innovation, and Economic Growth, *IBM Corporation*, 2006

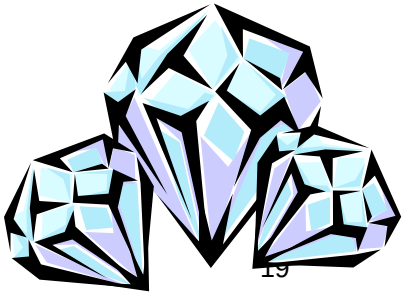
Ngành kinh tế định hướng dữ liệu

- Ngành công nghiệp quản lý và phân tích dữ liệu
  - "Chúng ta nhập trong dữ liệu mà đói khát tri thức"
  - Đáng giá hơn 100 tỷ US$ vào năm 2010
  - Tăng 10% hàng năm, gần gấp đôi kinh doanh phần mềm nói chung
  - vài năm gần đây các tập đoàn lớn chi khoảng 15 tỷ US$ mua công ty phân tích dữ liệu
  - *Tổng hợp của Kenneth Cukier*
- Nhân lực khoa học dữ liệu
  - CIO và chuyên gia phân tích dữ liệu có vai trò ngày càng cao
  - Người phân tích dữ liệu: người lập trình + nhà thống kê + "nghệ nhân" dữ liệu. Mỹ có chuẩn quy định chức năng
  - Tham khảo bài trao đổi "Tản mạn về cơ hội trong ngành Thống kê (và KHMT) của Nguyễn Xuân Long ngày 03/7/2009. http://www.procul.org/blog/2009/07/03/t%e1%ba%a3n-m%e1%ba%a1n-v%e1%bb%81-c%c6%a1-h%e1%bb%99i-trong-nganh-th%e1%bb%91ng-ke-va-khmt/

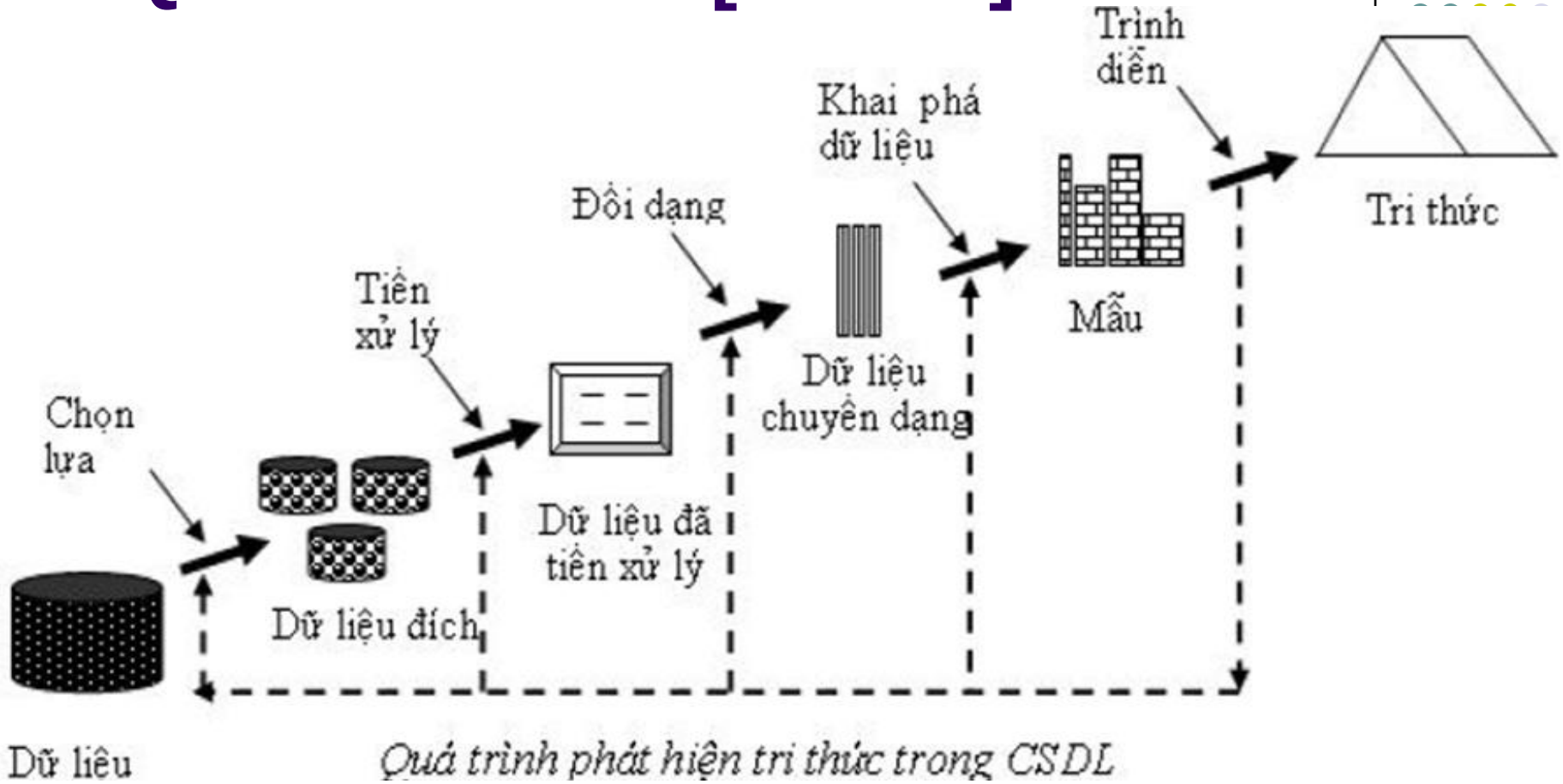
Khái niệm KDD



- Knowledge discovery from databases
  - Trích chọn các mẫu hoặc tri thức hấp dẫn (không tầm thường, ẩn, chưa biết và hữu dụng tiềm năng) từ một tập hợp lớn dữ liệu
  - KDD và KPDL: tên gọi lẫn lộn? *theo hai tác giả*|*Khai phá dữ liệu*
  - **Data Mining là một bước trong quá trình KDD**

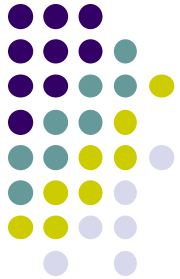


Quá trình KDD [FPS96]



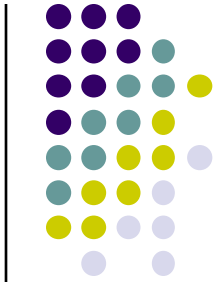
[FPS96] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth (1996). From Data Mining to Knowledge Discovery: An Overview, *Advances in Knowledge Discovery and Data Mining 1996*: 1-34

Các bước trong quá trình KDD



- Học từ miền ứng dụng
  - Tri thức sẵn có liên quan và mục tiêu của ứng dụng
- Khởi tạo một tập dữ liệu đích: chọn lựa dữ liệu
- Chuẩn bị dữ liệu và tiền xử lý: (huy động tới 60% công sức!)
- Thu gọn và chuyển đổi dữ liệu
  - Tìm các đặc trưng hữu dụng, rút gọn chiều/biến, tìm các đại diện bất biến.
- Chọn lựa chức năng (hàm) KPDL
  - Tóm tắt, phân lớp, hồi quy, kết hợp, phân cụm.
- Chọn (các) thuật toán KPDL
- Bước KPDL: tìm mẫu hấp dẫn
- Đánh giá mẫu và trình diễn tri thức
  - Trực quan hóa, chuyển dạng, loại bỏ các mẫu dư thừa, v.v.
- Sử dụng tri thức phát hiện được

Các khái niệm liên quan

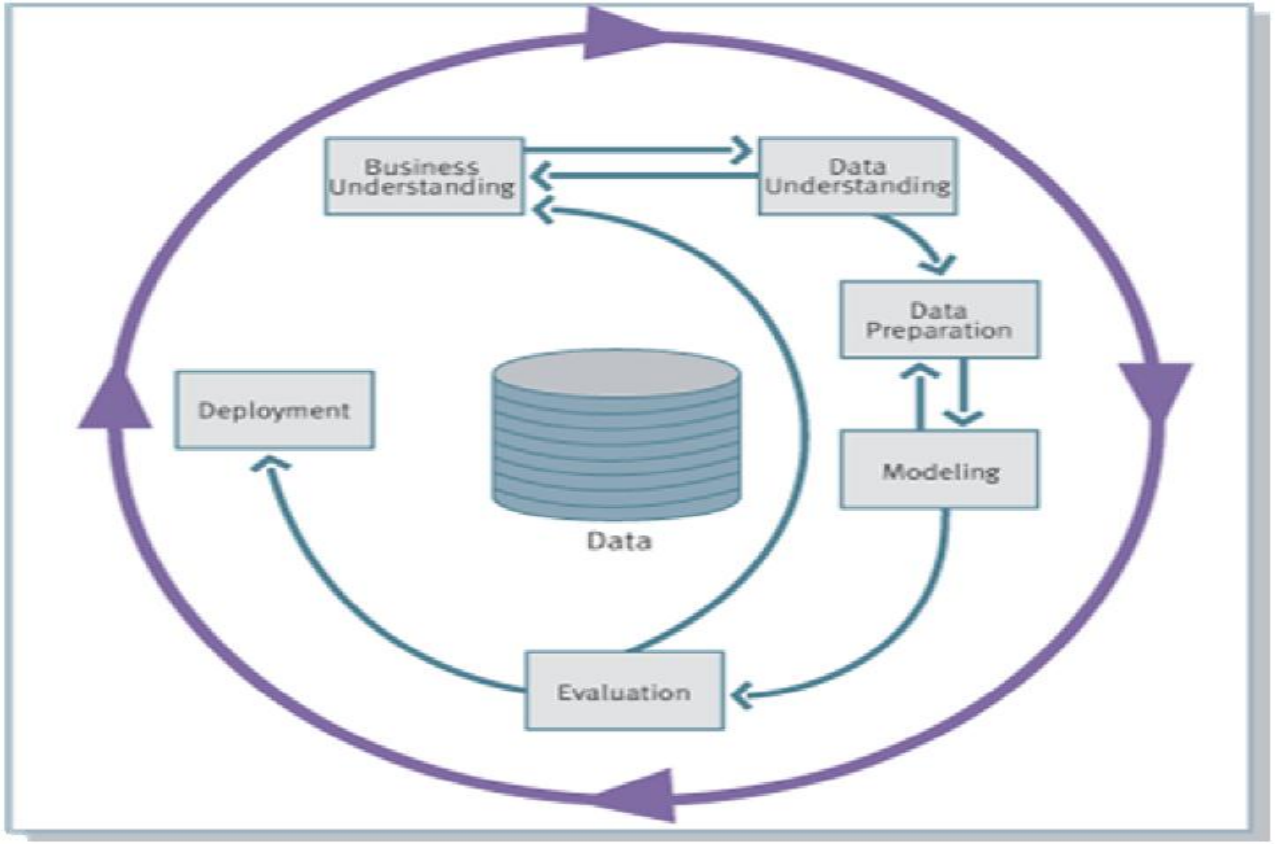


- Các tên thay thế
  - chiết lọc tri thức (knowledge extraction),
  - phát hiện thông tin (information discovery),
  - thu hoạch thông tin (information harvesting),
  - khai quật/nạo vét dữ liệu (data archaeology/ dredging),
  - Phân tích/xử lý mẫu/dữ liệu (data/pattern analysis/processing)
  - Thông minh doanh nghiệp (business intelligence -BI)
  - …
- Phân biệt: Phải chăng mọi thứ là DM?
  - Xử lý truy vấn suy diễn.
  - Hệ chuyên gia hoặc chương trình học máy/thống kê nhỏ

Mô hình quá trình KDD lặp [CCG98]

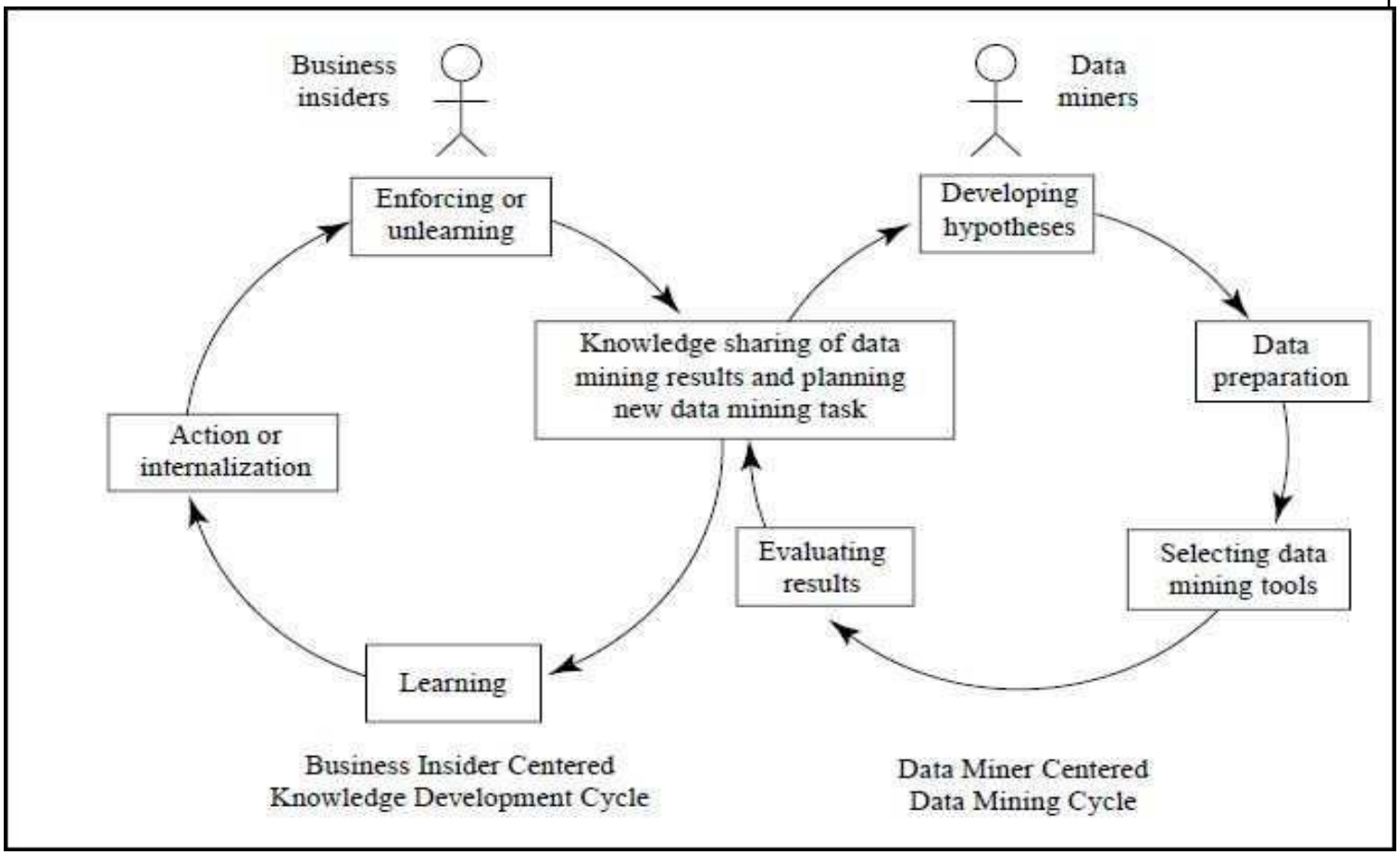
- Một mô hình cải tiến quá trình KDD
  - Định hướng kinh doanh: Xác định 1-3 câu hỏi hoặc mục đích hỗ trợ đích KDD
  - Kết quả thi hành được: xác định tập kết quả thi hành được dựa trên các mô hình được đánh giá
  - Lặp kiểu vòng đời phát triển phần mềm
  - [CCG98] Kenneth Collier, Bernard Carey, Ellen Grusy, Curt Marjaniemi, Donald Sautter (1998). A Perspective on Data Mining, *Technical Reporrt*, Northern Arizona University

Mô hình CRISP-DM 2000



- Quy trình chuẩn tham chiếu công nghiệp KPDL
  - Các pha trong mô hình quy trình CRISP-DM (Cross-Industry Standard Process for Data Mining). “Hiểu kinh doanh”: hiểu bài toán và đánh giá
  - Thi hành chỉ sau khi tham chiếu kết quả với “hiểu kinh doanh”
  - CRISP-DM 2.0 SIG WORKSHOP, LONDON, 18/01/2007
  - Nguồn: http://www.crisp-dm.org/Process/index.htm (13/02/2011)

Mô hình tích hợp DM-BI [WW08]



Chu trình phát triển tri thức thông qua khai phá dữ liệu

Wang, H. and S. Wang (2008). A knowledge management approach to data mining process for business intelligence, *Industrial Management & Data Systems*, 2008. **108**(5): 622-634. **[Oha09]**

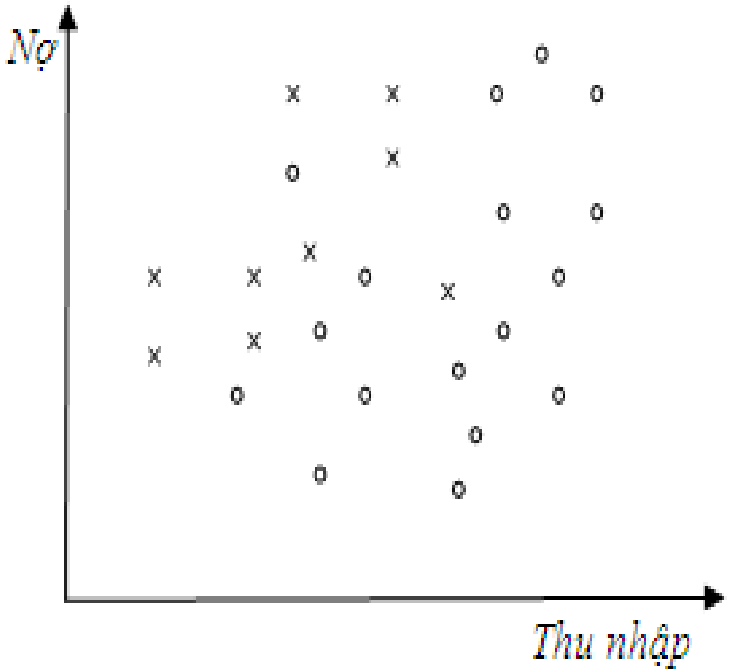
Dữ liệu và Mẫu

Dữ liệu (tập dữ liệu)

- tập F gồm hữu hạn các *trường hợp* (*sự kiện*).
- KDD:phải gồm rất nhiều trường hợp

Mẫu

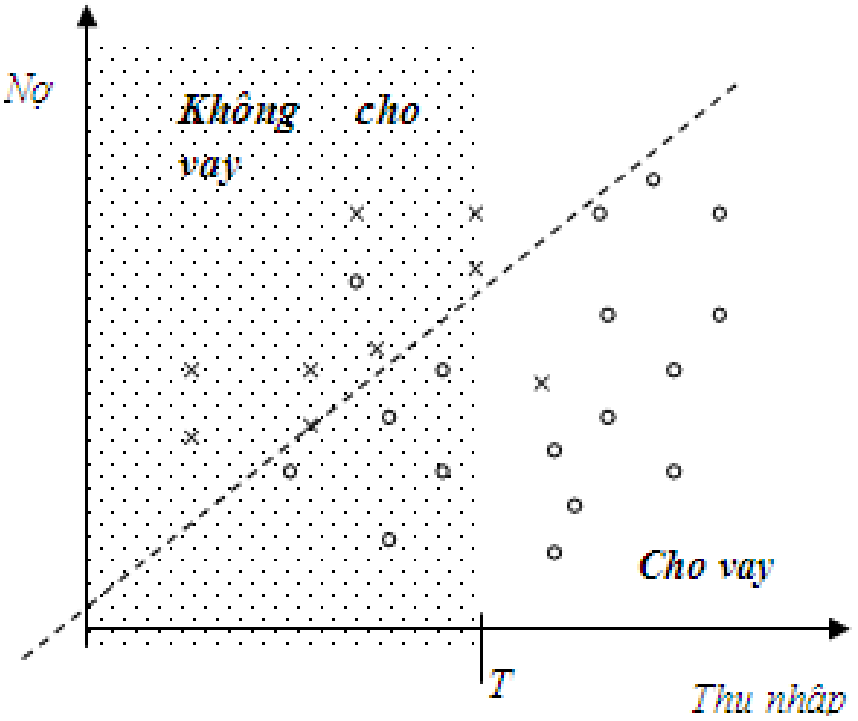
- Trong KDD: ngôn ngữ L để biểu diễn các tập con các sự kiện (dữ liệu) thuộc vào tập sự kiện F,
- Mẫu: biểu thức E trong ngôn ngữ L ⇔ tập con F_E tương ứng các sự kiện trong F. E được gọi là *mẫu* nếu nó đơn giản hơn so với việc liệt kê các sự kiện thuộc F_E.
- Chẳng hạn, biểu thức "THUNHẬP < $t" (mô hình chứa một biến THUNHẬP)



Hình 1.2. Tập dữ liệu có hai lớp x và o

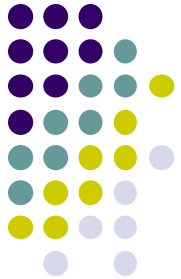
Tính có giá trị

- Mẫu được phát hiện: phải có *giá trị* đối với các dữ liệu mới theo độ chân thực nào đấy.
- Tính "có giá trị" : một *độ đo tính có giá trị (chân thực)* là một hàm C ánh xạ một biểu thức thuộc ngôn ngữ biểu diễn mẫu L tới một không gian đo được (bộ phận hoặc toàn bộ) M_C.
- Chẳng hạn, đường biên xác định mẫu "THUNHẬP < $t“ dịch sang phải (biến THUNHẬP nhận giá trị lớn hơn) thì độ chân thực giảm xuống do bao gói thêm các tình huống vay tốt lại bị đưa vào vùng không cho vay nợ.
- Nếu *a**THUNHẬP + *b**NỢ < 0 mẫu có giá trị hơn.



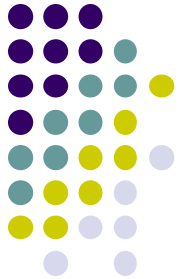
Hình 1.3. Ngưỡng đơn T theo thu nhập để phân lớp cho vay (Lưu ý, đường nghiêng rời nét cho quyết định tốt hơn).

Tính mới và hữu dụng tiềm năng



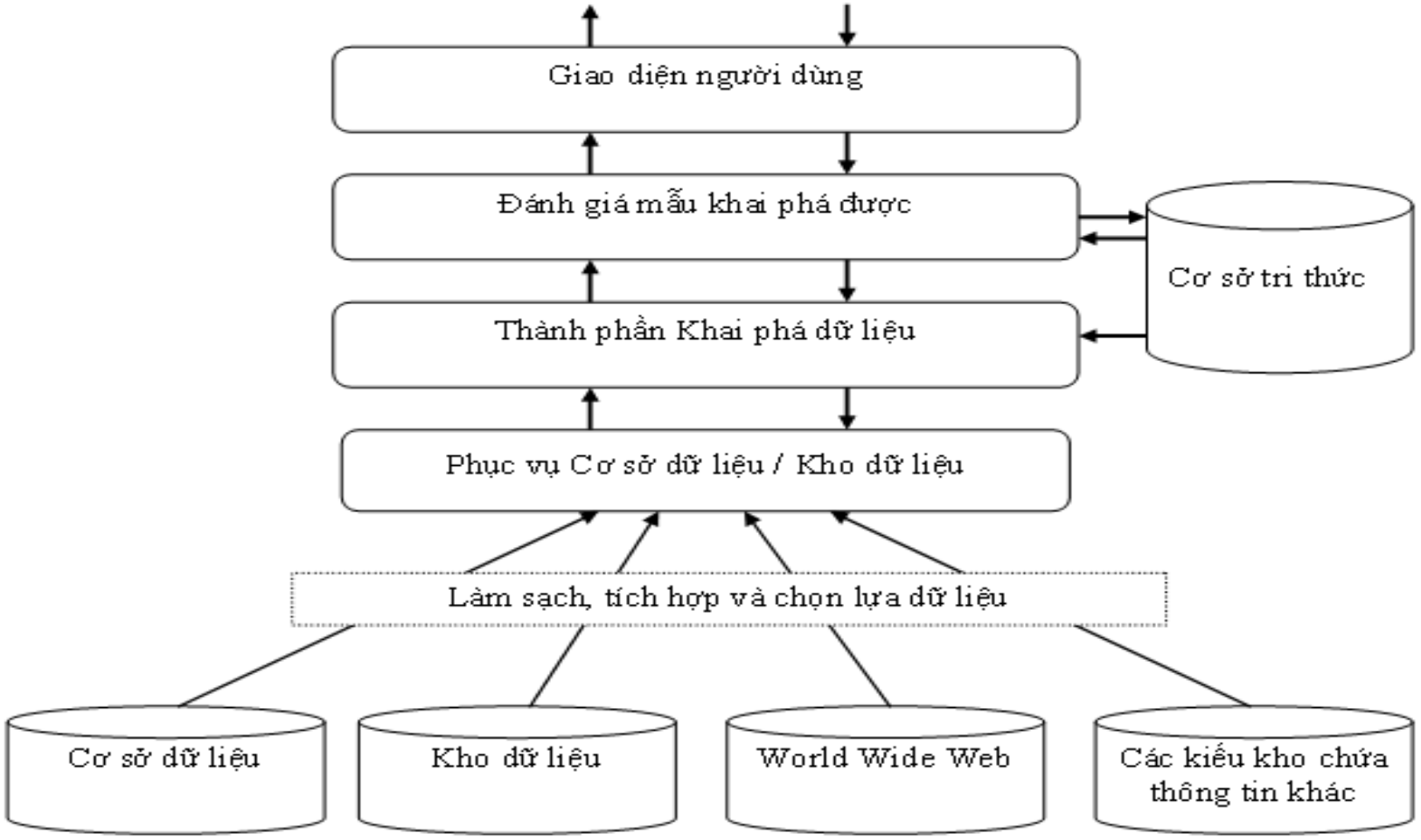
- *Tính mới*: Mẫu phải là mới trong một miền xem xét nào đó, ít nhất là hệ thống đang được xem xét.
  - *Tính mới có thể đo được* :
    - sự thay đổi trong dữ liệu: so sánh giá trị hiện tại với giá trị quá khứ hoặc giá trị kỳ vọng
    - hoặc tri thức: tri thức mới quan hệ như thế nào với các tri thức đã có.
    - Tổng quát, điều này có thể được đo bằng một hàm N(E,F) hoặc là độ đo về tính mới hoặc là độ đo kỳ vọng.
- *Hữu dụng tiềm năng*: Mẫu cần có khả năng chỉ dẫn tới các tác động hữu dụng và *được đo bởi một hàm tiện ích*.
  - Hàm U ánh xạ các biểu thức trong L tới một không gian đo có thứ tự (bộ phận hoặc toàn bộ) M_U: u = U (E,F).
  - Ví dụ, trong tập dữ liệu vay nợ, hàm này có thể là *sự tăng hy vọng theo sự tăng lãi của nhà băng* (tính theo đơn vị tiền tệ) kết hợp với quy tắc quyết định được trình bày trong Hình 1.3.

Tính hiểu được, tính hấp dẫn và tri thức



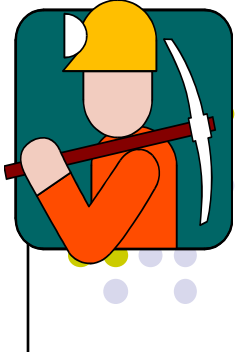
- *Tính hiểu được*: Mẫu phải hiểu được
  - KDD: *mẫu mà con người hiểu chúng dễ dàng hơn* các dữ liệu nền.
  - Khó đo được một cách chính xác: "có thể hiểu được“ [redacted]ễ hiểu.
  - Tồn tại một số độ đo dễ hiểu:
    - Sắp xếp từ cú pháp (tức là cỡ của mẫu theo bit) tới ngữ nghĩa (tức là dễ dàng để con người nhận thức được theo một tác động nào đó).
    - Giả định rằng tính hiểu được là *đo được* bằng một hàm S ánh xạ biểu thức E trong L tới một không gian đo được có thứ tự (bộ phận /toàn bộ) M_S: s = S(E,F).
- *Tính hấp dẫn*: *độ đo tổng thể về mẫu* là sự kết hợp của các tiêu chí *giá trị*, *mới*, *hữu ích* và *dễ hiểu*.
  - Hoặc dùng một hàm hấp dẫn: i = I (E, F, C, N, U, S) ánh xạ biểu thức trong L vào một không gian đo được M_i.
  - Hoặc xác định độ hấp dẫn trực tiếp: thứ tự của các mẫu được phát hiện.
- *Tri thức*: Một mẫu E [redacted] được gọi là *tri thức* nếu như đối với một lớp người sử dụng nào đó, chỉ ra được một ngưỡng i [redacted]M_i mà độ hấp dẫn I(E,F,C,N,U,S) > i.

Kiến trúc điển hình hệ thống KPDL



Hình 1.6. Kiến trúc điển hình của hệ thống khai phá dữ liệu

Khai phá dữ liệu và quản trị CSDL



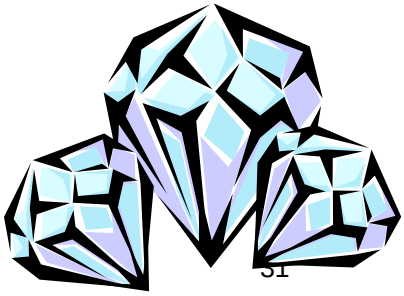
Câu hỏi thuộc hệ quản trị CSDL (DBMS)

Hãy hiển thị số tiền Ông Smith trong ngày 5 tháng Giêng ? *ghi nhận riêng lẻ do xử lý giao dịch trực tuyến* (on-line transaction processing – OLTP).

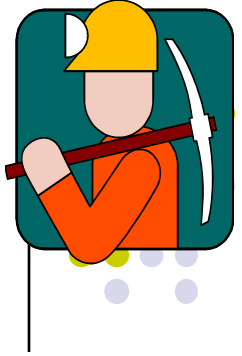
Có bao nhiêu nhà đầu tư nước ngoài mua cổ phiếu X trong tháng trước ? *ghi nhận thống kê do hệ thống hỗ trợ quyết định thống kê* (stastical decision suppport system - DSS)

Hiển thị mọi cổ phiếu trong CSDL với mệnh giá tăng ? *ghi nhận dữ liệu đa chiều do xử lý phân tích trực tuyến* (on-line analytic processing - OLAP).

Cần có một *giả thiết "đầy đủ" về tri thức miền phức tạp*!



Khái niệm KPDL: câu hỏi DMS



Câu hỏi thuộc hệ thống khai phá dữ liệu (DMS)

Các cổ phiếu tăng giá có đặc trưng gì ?

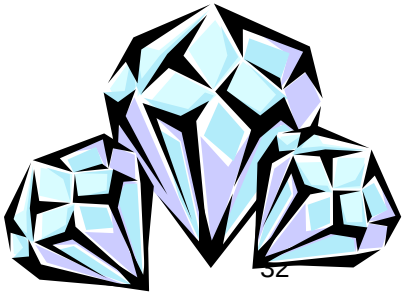
Tỷ giá US$ - DMark có đặc trưng gì ?

Hy vọng gì về cổ phiếu X trong tuần tiếp theo ?

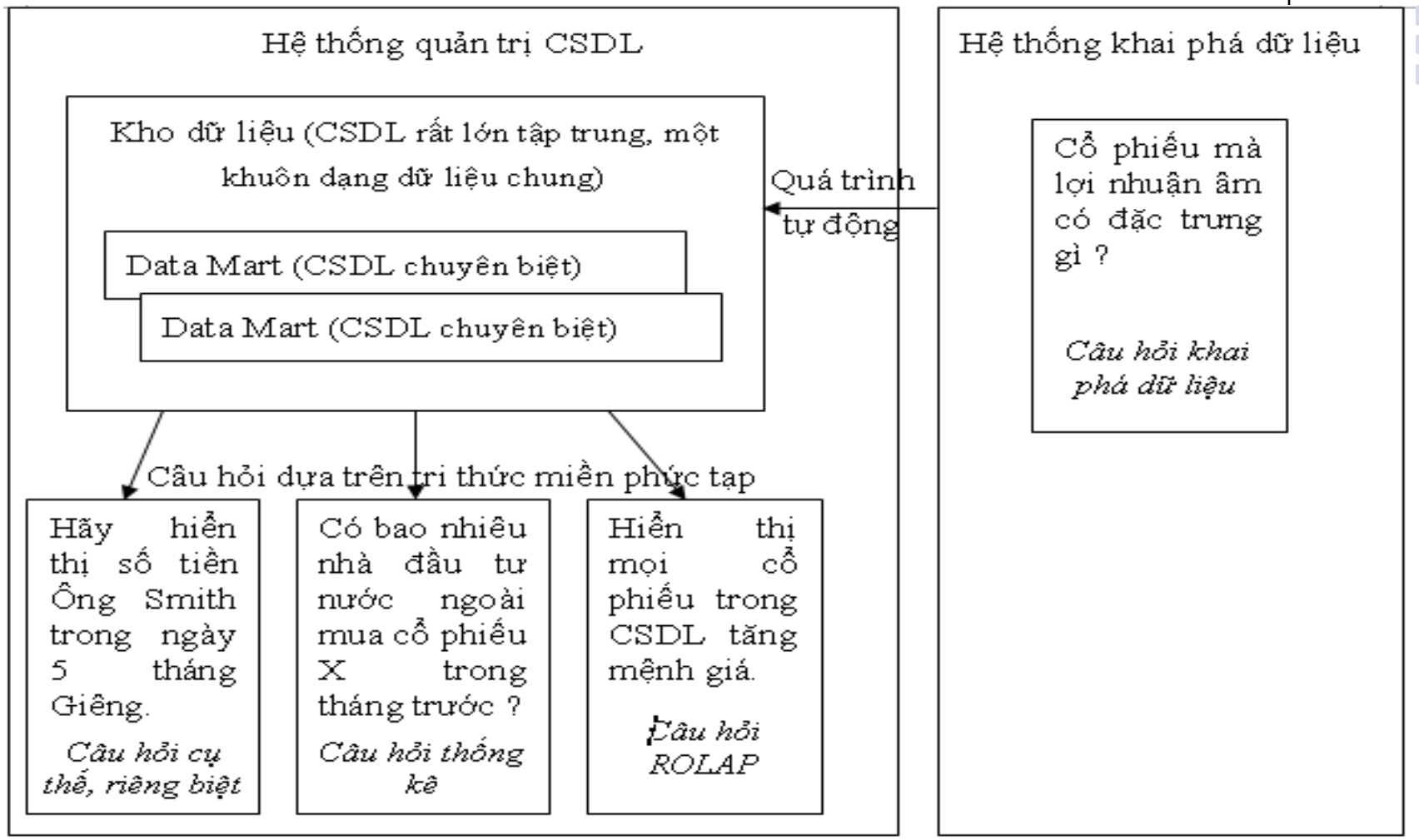
Trong tháng tiếp theo, sẽ có bao nhiêu đoàn viên công đoàn không trả được nợ của họ ?

Những người mua sản phẩm Y có đặc trưng gì ?

Giả thiết tri thức "đầy đủ" không còn có tính cốt lõi, cần bổ sung tri thức cho hệ thống [redacted]ải tiến (nâng cấp) miền tri thức !

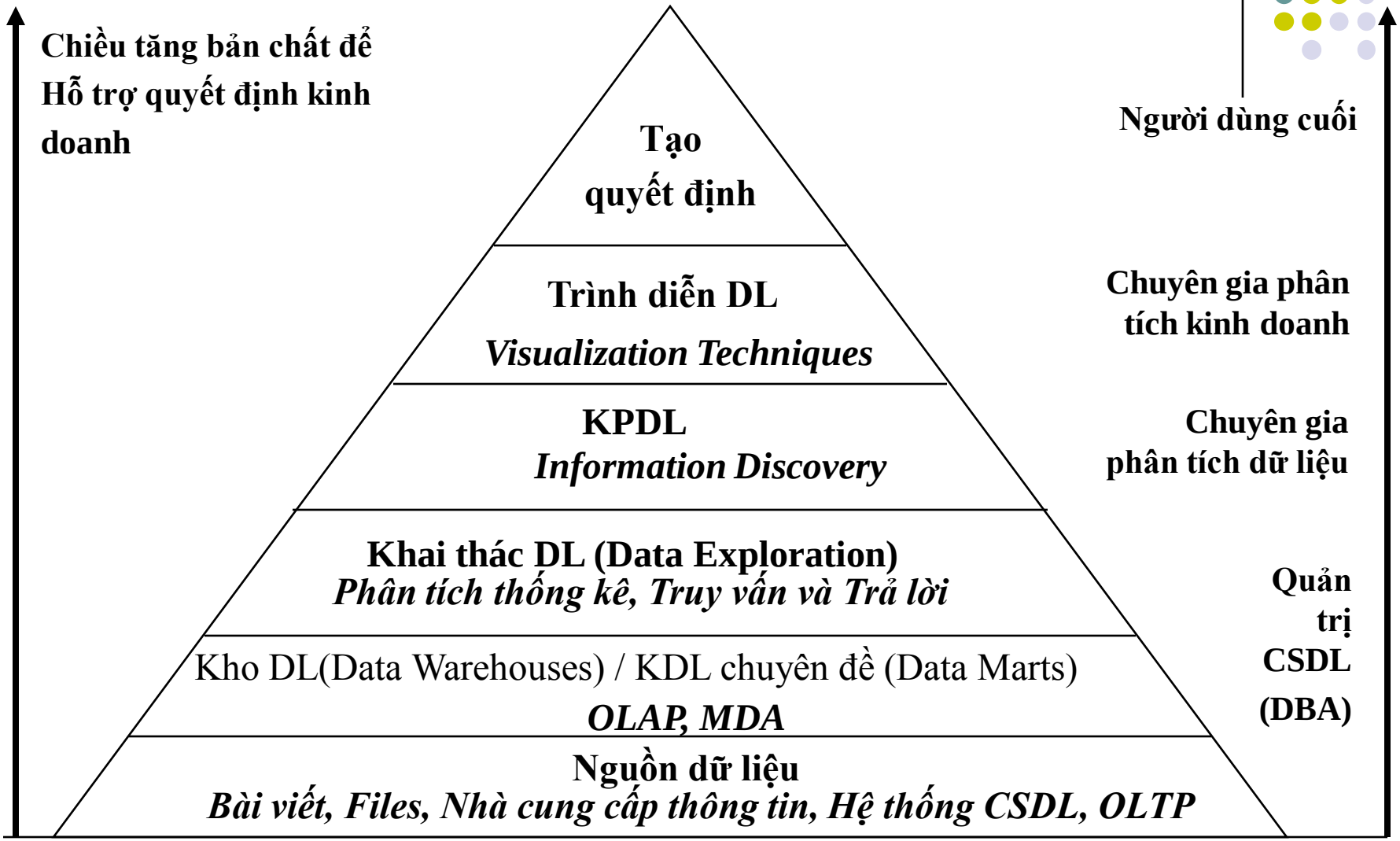


Hệ thống CSDL và Hệ thống Khai phá dữ liệu

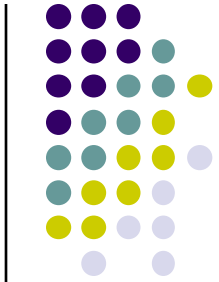


Hình 1.7. Mối quan hệ giữa hệ thống CSDL và hệ thống khai phá dữ liệu

KHAI PHÁ DỮ LIỆU VÀ THÔNG MINH KINH DOANH



Ứng dụng cơ bản của KPDL

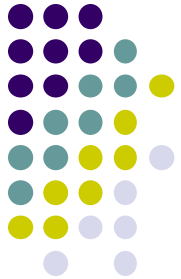


- Phân tích dữ liệu và hỗ trợ quyết định
  - Phân tích và quản lý thị trường
    - Tiếp thị định hướng, quản lý quan hệ khách hàng (CRM), phân tích thói quen mua hàng, bán hàng chéo, phân đoạn thị trường
  - Phân tích và quản lý rủi ro
    - Dự báo, duy trì khách hàng, cải thiện bảo lãnh, kiểm soát chất lượng, phân tích cạnh tranh
  - Phát hiện gian lận và phát hiện mẫu bất thường (ngoại lai)
- Ứng dụng khác
  - Khai phá Text (nhóm mới, email, tài liệu) và khai phá Web
  - Khai phá dữ liệu dòng
  - Phân tích DNA và dữ liệu sinh học

Phân tích và quản lý thị trường

- Nguồn dữ liệu có từ đâu ?
  - Giao dịch thẻ tín dụng, thẻ thành viên, phiếu giảm giá, các phàn nàn của khách hàng, các nghiên cứu phong cách sống (công cộng) bổ sung
- Tiếp thị định hướng
  - Tìm cụm các mô hình khách hàng cùng đặc trưng: sự quan tâm, mức thu nhập, thói quen chi tiêu...
  - Xác định các mẫu mua hàng theo thời gian
- Phân tích thị trường chéo
  - Quan hệ kết hợp/đồng quan hệ giữa bán hàng và sự báo dựa theo quan hệ kết hợp
- Hồ sơ khách hàng
  - Kiểu của khách hàng mua sản phẩm gì (phân cụm và phân lớp)
- Phân tích yêu cầu khách hàng
  - Định danh các sản phẩm tốt nhất tới khách hàng (khác nhau)
  - Dự báo các nhân tố sẽ thu hút khách hàng mới
- Cung cấp thông tin tóm tắt
  - Báo cáo tóm tắt đa chiều
  - Thông tin tóm tắt thống kê (xu hướng trung tâm dữ liệu và biến đổi)

Phân tích doanh nghiệp & Quản lý rủi ro

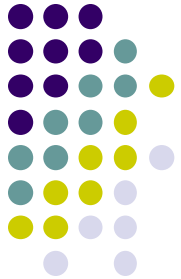


- Lên kế hoạch tài chính và đánh giá tài sản
  - Phân tích và dự báo dòng tiền mặt
  - Phân tích yêu cầu ngẫu nhiên để đánh giá tài sản
  - Phân tích lát cắt ngang và chuỗi thời gian (tỷ số tài chính, phân tích xu hướng…)
- Lên kế hoạch tài nguyên
  - Tóm tắt và so sánh các nguồn lực và chi tiêu
- Cạnh tranh
  - Theo dõi đối thủ cạnh tranh và định hướng thị trường
  - Nhóm khách hàng thành các lớp và định giá dựa theo lớp khách
  - Khởi tạo chiến lược giá trong thị trường cạnh tranh cao

Phát hiện gian lận và khai phá mẫu hiếm

- Tiếp cận: Phân cụm & xây dựng mô hình gian lận, phân tích bất thường
- Ứng dụng: Chăm sóc sức khỏe, bán lẻ, dịch vụ thẻ tín dụng, viễn thông.
  - Bảo hiểm tự động: vòng xung đột
  - Rửa tiền: giao dịch tiền tệ đáng ngờ
  - Bảo hiểm y tế
    - Bệnh nghề nghiệp, nhóm bác sỹ, và nhóm chỉ dẫn
    - Xét nghiệm không cần thiết hoặc tương quan
  - Viễn thông: cuộc gọi gian lận
    - Mô hình cuộc gọi: đích cuộc gọi, độ dài, thời điểm trong ngày hoặc tuần. Phân tích mẫu lệch một dạng chuẩn dự kiến
  - Công nghiệp bán lẻ
    - Các nhà phân tích ước lượng rằng 38% giảm bán lẻ là do nhân viên không trung thực
  - Chống khủng bố

Ứng dụng khác

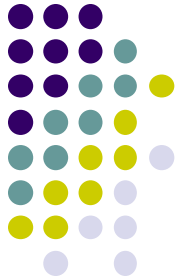


- Thể thao
  - IBM Advanced Scout phân tích thống kế môn NBA (chặn bóng, hỗ trợ và lỗi) để đưa tới lợi thế cạnh trang cho New York Knicks và Miami Heat
- Thiên văn học
  - JPL và Palomar Observatory khám phá 22 chuẩn tinh (quasar) với sự trợ giúp của KPDL
- Trợ giúp lướt web Internet
  - Trợ giúp IBM áp dụng các thuật toán KPDL biên bản truy nhập Web đối với các trang liên quan tới thị trường để khám phá ưu đãi khách hàng và các trang hành vi, phân tích tính hiệu quả của tiếp thị Web, cải thiệ cách tổ chức Website …

Bảng 1. Xu thế phát triển của các lĩnh vực ứng dụng khai phá dữ liệu điển hình [Pia06]

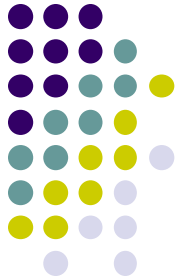
Field	*Y03*	*Y04*	*Y06*	*Growth*	*Trend*
e-commerce & Web mining	**11%**	**14%**	**30%**	**137%**	**up**
Security / Anti-terrorism & Government /Military	4.6%	9.3%	11%	58%	up
Travel/Hospitality	3.8%	2.3%	4.5%	47%	up
Investment / Stocks	6.1%	10%	10%	21%	~
Fraud Detection	18%	22%	22%	8%	~
Direct Marketing/ Fundraising	22%	22%	20%	-9%	~
Manufacturing	3.8%	10%	6.4%	-10%	~
Other	9.2%	22%	14%	-13%	—
Retail	13%	10%	10%	-15%	~
Telecom	16%	14%	13%	-15%	~
Biotech/Genomics	21%	21%	16%	-25%	down
Banking / Credit Scoring	28%	34%	20%	-35%	down
Insurance	18%	17%	11%	-38%	down
Science	18%	23%	11%	-47%	down
Medical/ Pharma	12%	17%	7.3%	-51%	down

KPDL: Sơ đồ phân loại (Chức năng)



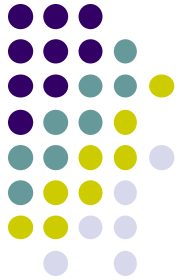
- Chức năng chung
  - KPDL mô tả: tóm tắt, phân cụm, luật kết hợp…
  - KPDL dự đoán: phân lớp, hồi quy…
- Các bài toán điển hình
  - Mô tả khái niệm
  - Quan hệ kết hợp
  - Phân lớp
  - Phân cụm
  - Hồi quy
  - Mô hình phụ thuộc
  - Phát hiện biến đổi và độ lệch
  - Phân tích định hướng mẫu, các bài toán khác

KPDL: Sơ đồ phân loại (Chức năng)



- Mô tả khái niệm: Đặc trưng và phân biệt
  - Tìm các đặc trưng và tính chất của khái niệm
  - Tổng quát hóa, tóm tắt, phát hiện đặc trưng ràng buộc, tương phản, chẳng hạn, các vùng khô so sánh với ướt
  - Bài toán mô tả điển hình: Tóm tắt (tìm mô tả cô đọng)
    - Kỳ vọng, phương sai
    - Tóm tắt văn bản
- Quan hệ kết hợp
  - Quan hệ kết hợp giữa các biến dữ liệu: Tương quan và nhân quả)
  - Diaper → Beer [0.5%, 75%]
  - Luật kết hợp: X[redacted]
  - Ví dụ, trong khai phá dữ liệu Web
    - Phát hiện quan hệ ngữ nghĩa
    - Quan hệ nội dung trang web với mối quan tâm người dùng

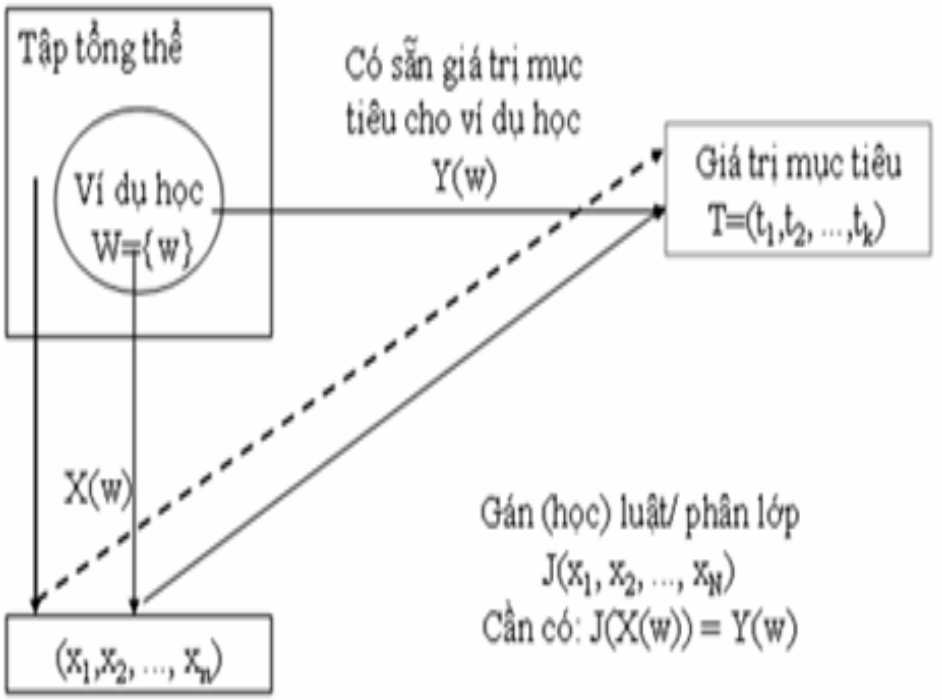
Các bài toán KPDL: Chức năng KPDL



- Phân lớp và Dự báo
  - Xây dựng các mô hình (chức năng) để mô tả và phân biệt khái niệm cho các lớp hoặc khái niệm để dự đoán trong tương lai
    - Chẳng hạn, phân lớp quốc gia dựa theo khí hậu, hoặc phân lớp ô tô dựa theo tiêu tốn xăng
  - Trình diễn: cây quyết định, luật phân lớp, mạng nơron
  - Dự đoán giá trị số chưa biết hoặc đã mất

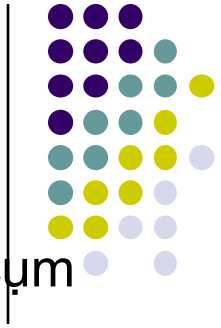
KPDL: Sơ đồ phân loại (Chức năng)

- Phân lớp
  - xây dựng/mô tả mô hình/ hàm dự báo để mô tả/phát hiện lớp/khái niệm cho dự báo tiếp
  - học một hàm ánh xạ dữ liệu vào một trong một số lớp đã biết
- Phân cụm
  - nhóm dữ liệu thành các "cụm" (lớp mới) để phát hiện được mẫu phân bố dữ liệu miền ứng dụng.
  - Tính tương tự



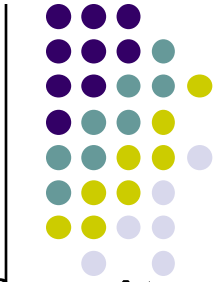
Hình 1.8. Sơ đồ biểu diễn mô hình học máy: cần học đường rời nét (Lưu ý, học máy không giám sát (phân cụm) không có giá trị mục tiêu cho ví dụ học (không có hai đường liền nét hướng tới giá trị mục tiêu)

Chức năng KPDL (2)



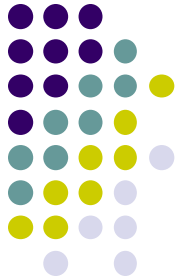
- Phân tích cụm
  - Nhãn lớp chưa biết: Nhóm dữ liệu thành các lớp mới: phân cụm các nhà để tìm mẫu phân bố
  - Cực đại tương tự nội bộ cụm & cực tiểu tương tự giữa các cụm
- Phân tích bất thường
  - Bất thường: đối tượng dữ liệu không tuân theo hành vi chung của toàn bộ dữ liệu. *Ví dụ, sử dụng kỳ vọng mẫu và phương sai mẫu*
  - Nhiễu hoặc ngoại lệ? Không phải! Hữu dụng để phát hiện gian lận, phân tích các sự kiện hiếm
- Phát hiện biến đổi và độ lệch
  - Hầu như sự thay đổi có ý nghĩa dưới dạng độ đo đã biết trước/giá trị chuẩn, cung cấp tri thức về sự biến đổi và độ lệch
  - Phát hiện biến đổi và độ lệch <> tiền xử lý

KPDL: Sơ đồ phân loại (Chức năng)



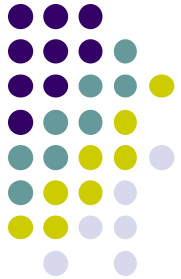
- Hồi quy
  - học một hàm ánh xạ dữ liệu nhằm xác định giá trị thực của một biến theo một số biến khác
  - điển hình trong phân tích thống kê và dự báo
  - dự đoán giá trị của một/một số biến phụ thuộc vào giá trị của một tập biến độc lập.
- Mô hình phụ thuộc
  - xây dựng mô hình phụ thuộc: tìm một mô hình mô tả sự phụ thuộc có ý nghĩa giữa các biến
  - mức cấu trúc:
    - dạng đồ thị
    - biến là phụ thuộc bộ phận vào các biến khác
  - mức định lượng: tính phụ thuộc khi sử dụng việc đo tính theo giá trị số

KPDL: Sơ đồ phân loại (Chức năng)



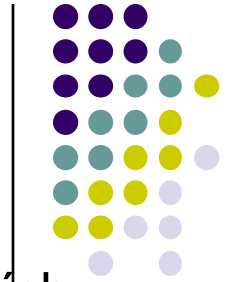
- Phân tích xu hướng và tiến hóa
  - Xu hướng và độ lệch: phân tích hồi quy
  - Khai phá mẫu tuần tự, phân tích chu kỳ
  - Phân tích dựa trên tương tự
- Phân tích định hướng mẫu khác hoặc phân tích thống kê

KPDL: Sơ đồ phân loại (2)



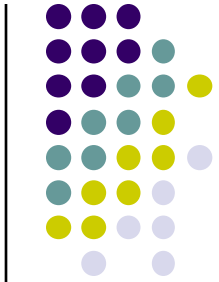
- Phân loại theo khung nhìn
  - Kiểu dữ liệu được KP
  - Kiểu tri thức cần phát hiện
  - Kiểu kỹ thuật được dùng
  - Kiểu miền ứng dụng

Khung nhìn đa chiều của KPDL



- **Dữ liệu được khai phá**
  - Quan hệ, KDL, giao dịch, dòng, hướng đối tượng/quan hệ, tích cực, không gian, chuỗi thời gian, văn bản, đa phương tiện, không đồng nahats, kế thừa, WWW
- **Tri thức được khai phá**
  - Đặc trưng, phân biệt, kết hợp, phân lớp, phân cụm, xu hướng/độ lệch, phân tích bất thường,…
  - Các chức năng phức/tích hợp và KPDL các mức phức hợp
- **Kỹ thuật được dùng**
  - Định hướng CSDL, KDL (OLAP), học máy, thống kê, trực quan hóa, ….
- **Ứng dụng phù hợp**
  - Bán lẻ, viễn thông, ngân hàng, phân tích gian lận, KPDL sinh học, phân tích thị trường chứng khoán, KP văn bản, KP Web, …

KPDL: các kiểu dữ liệu



- CSDL quan hệ
- Kho dữ liệu
- CSDL giao dịch
- CSDL mở rộng và kho chứa thông tin
  - CSDL quan hệ-đối tượng
  - Dữ liệu không gian và thời gian
  - Dữ liệu chuỗi thời gian
  - Dữ liệu dòng
  - Dữ liệu đa phương tiện
  - Dữ liệu không đồng nhất và thừa kế
  - CSDL Text & WWW

Kiểu dữ liệu được phân tích/khai phá 8/2009

http://www.kdnuggets.com/polls/2010/data-types-analyzed.html

KDnuggets Home » Polls » Data types analyzed/mined (Aug 2010)

Data types analyzed/mined in the past 12 months

Types of Data Analyzed/Mined in the past 12 months
[144 voters]

table data (fixed # of columns) (102)	70.8%
time series (56)	38.9%
itemsets / transactions (52)	36.1%
text (free-form) (43)	29.9%
anonymized data (38)	26.4%
social network data (28)	19.4%
other (22)	15.3%
web content (19)	13.2%
XML data (17)	11.8%
web clickstream (15)	10.4%
email (15)	10.4%
images / video (11)	7.6%
music / audio (3)	2.1%

Comparing with a similar 2009 KDnuggets Poll: Types of Data Analyzed/Mined in the past 12 months , we see that the top 3 data types are still

1. table data
2. time series
3. itemsets / transactions

Ignoring Spatial data, which was accidentally excluded from the 2010 poll, Data types with the highest increase in popularity measured by *(pct_usage2010 - pct_usage2009)/pct_usage2009* were

1. other , up 61.3%
2. social network data, up 53.9%
3. anonymized data, up 39.3%

Largest decrease in popularity was for

1. text free-form, down 21.2%
2. images / video, down 39.5%
3. music / audio , down 71.7%

http://www.kdnuggets.com/polls/2009/largest-database-data-mined.htm

http://www.kdnuggets.com/polls/2010/data-miner-salary.html

Largest database data-mined Poll

What was the largest database or dataset you data-mined? [185 votes total]

less than 1 MB (5)	3%
1.1 to 10 MB (9)	5%
11 to 100 MB (12)	6%
101 MB to 1 GB (19)	10%
1.1 to 10 GB (46)	25%
11 to 100 GB (35)	19%
101 GB to 1 Terabyte (30)	16%
1.1 to 10 Terabytes (17)	9%
over 10 Terabytes (12)	6%

The median database size is 10-20 GB, same as in 2008 KDnuggets Poll: largest database or dataset you data-mined.

We note that 2009 poll run for longer period of time, so it got more votes.

We observe that there are the same number of votes in over 10 Terabytes range, but significantly more votes in 101 GB to 1 Terabyte, and especially 1.1 to 10 Terabytes ranges.

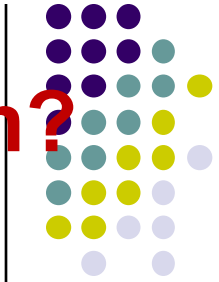
Region	2010 Salary	2009 Salary	% Change	2010 Count
US/Canada	$102,900	110.1	-6.5%	123
Australia/NZ	$95,600	83.6	14.4%	8
W. Europe	$67,900	74.8	-9.2%	37
Asia	$56,700	31.1	82.1%	14
Africa/MidEast	$52,500	20.0	162.5%	4
E. Europe	$46,600	56.0	-16.8%	17
Latin America	$39,400	74.3	-47.0%	5

The following figure gives a breakdown by employer.

Mọi mẫu khai phá được đều hấp dẫn?

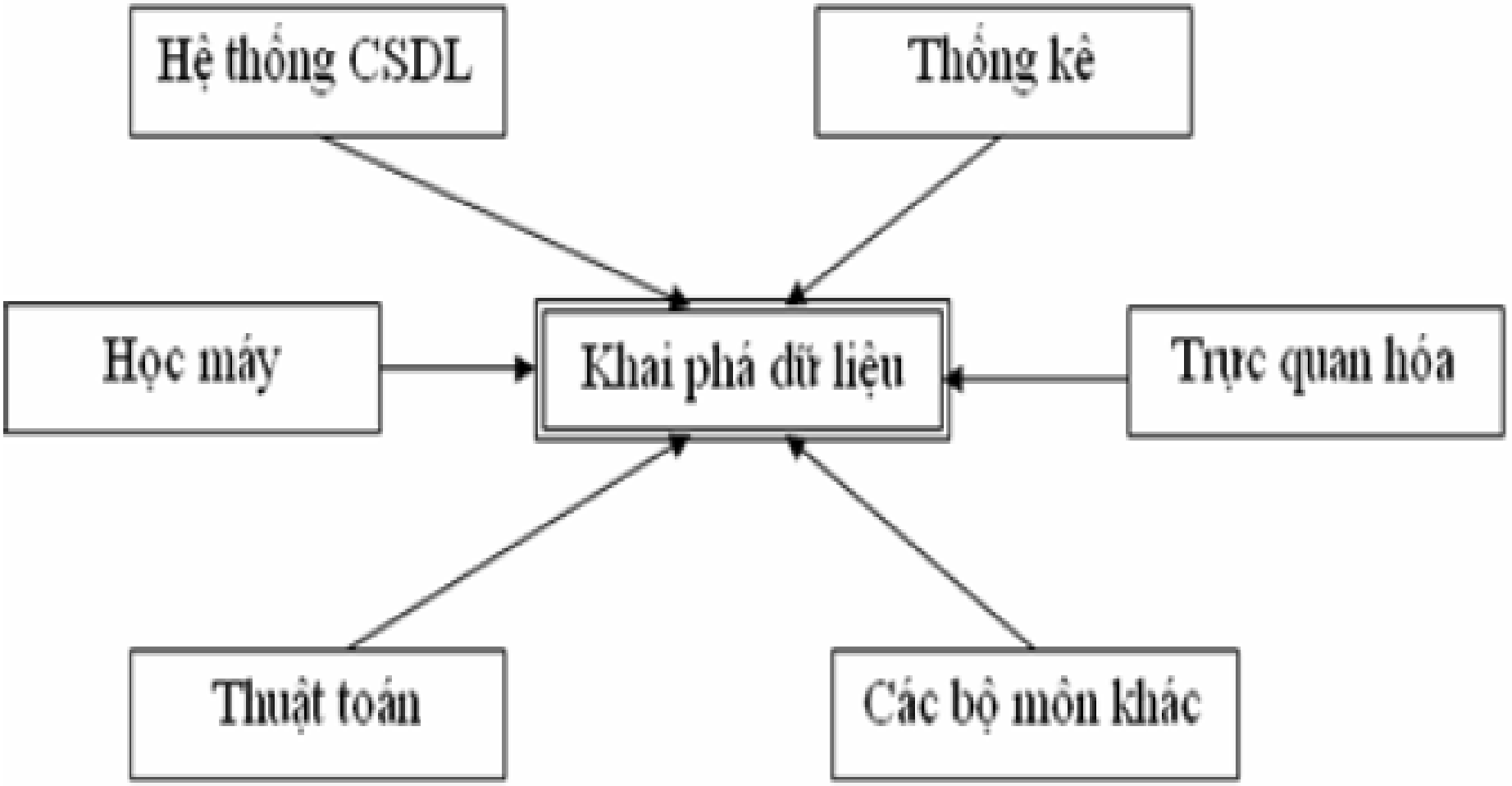
- KPDL có thể sinh ra tới hàng nghìn mẫu: Không phải tất cả đều hấp dẫn
  - Tiếp cận gợi ý: KPDL hướng người dùng, dựa trên câu hỏi, hướng đích
- **Độ đo hấp dẫn**
  - Mẫu là hấp dẫn nếu dễ hiểu, có giá trị theo dữ liệu mới/kiểm tra với độ chắc chắn, hữu dụng tiềm năng, mới lạ hoặc xác nhận các giả thiết mà người dùng tìm kiếm để xác thực.
- **Độ đo hấp dẫn khách quan và chủ quan**
  - Khách quan: dựa trên thống kê và cấu trúc của mẫu, chẳng hạn, độ hỗ trợ, độ tin cậy, …
  - Chủ quan: dựa trên sự tin tưởng của người dùng đối với dữ liệu, chẳng hạn, sự không chờ đón, tính mới mẻ, tác động được...

Tìm được tất cả và chỉ các mẫu hấp dẫn?



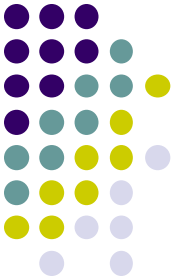
- Tìm được mọi mẫu hấp dẫn: Bài toán về tính đầy đủ
  - Hệ thống KHDL có khả năng tìm **mọi** mẫu hấp dẫn?
  - Tìm kiếm mày mò (heuristic) <> tìm kiếm đầy đủ
  - Kết hợp <> phan lớp <> phân cụm
- Tìm chỉ các mẫu hấp dẫn: Bài toán tối ưu
  - Hệ thống KPDL có khả năng tìm ra **đúng** các mẫu hấp dẫn?
  - Tiếp cận
    - Đầu tiên tìm tổng thể tất cả các mẫu sau đó lọc bỏ các mẫu không hấp dẫn.
    - Sinh ra chỉ các mẫu hấp dẫn—tối ưu hóa câu hỏi khai phá

KPDL: Hội tụ của nhiều ngành phức



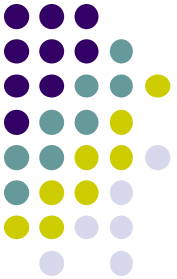
Hình 1.9. Tính đaliên ngành của Khai phá dữ liệu

Thống kê toán học với Khai phá dữ liệu



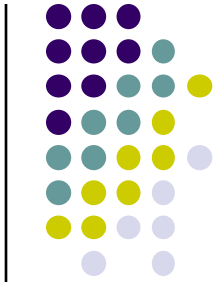
- Nhiều điểm chung giữa KPDL với *thống kê*:
  - Đặc biệt như phân tích dữ liệu thăm dò (EDA: Exploratory Data Analysis) cũng như dự báo [Fied97, HD03].
  - Hệ thống KDD thường gắn kết với các thủ tục thống kê đặc biệt đối với mô hình dữ liệu và nắm bắt nhiễu trong một khung cảnh phát hiện tri thức tổng thể.
  - Các phương pháp KPDL dựa theo thống kê nhận được sự quan tâm đặc biệt.

Thống kê toán học với Khai phá dữ liệu



- Phân biệt giữa bài toán thống kê và bài toán khai phá dữ liệu
  - Bài toán kiểm định giả thiết thống kê: cho trước một giả thiết + tập dữ liệu quan sát được. Cần kiểm tra xem tập dữ liệu quan sát được có phù hợp với giả thiết thống kê hay không/ giả thiết thống kê có đúng trên toàn bộ dữ liệu quan sát được hay không.
  - Bài toán học khai phá dữ liệu: mô hình chưa có trước. Mô hình kết quả phải phù hợp với tập toàn bộ dữ liệu -> cần đảm bảo các tham số mô hình không phụ thuộc vào cách chọn tập dữ liệu học. Bài toán học KPDL đòi hỏi tập dữ liệu học/tập dữ liệu kiểm tra cần "đại diện" cho toàn bộ dữ liệu trong miền ứng dụng và cần độc lập nhau. Một số trường hợp: hai tập dữ liệu này (hoặc tập dữ liệu kiểm tra) được công bố dưới dạng chuẩn.
  - Về thuật ngữ: KPDL: *biến ra*/*biến mục tiêu*, *thuật toán khai phá dữ liệu*, *thuộc tính*/*đặc trưng*, *bản ghi*... XLDLTK: *biến phụ thuộc*, *thủ tục thống kê*, *biến giải thích*, *quan sát*...
  - Tham khảo thêm từ Nguyễn Xuân Long

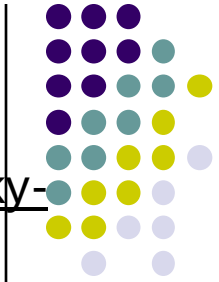
Nguồn chỉ dẫn về KPDL



- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations
- Database systems (SIGMOD: CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: ACM-TODS, IEEE-TKDE, JIIS, J. ACM, etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), etc.
  - Journals: Machine Learning, Artificial Intelligence, etc.
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.
- Một số tham khảo khác
  - http://www.kdnuggets.com/
  - Danh sách tài liệu tham khảo
  - Future Directions in Computer Science


http://www.kdnuggets.com/
KDnuggets™
Data Mining Community's Top Resource Since 1997
for Data Mining and Analytics Software, Jobs, Consulting,
Courses, Education, News, and more.
search KDnuggets
Search
advanced search
help
BAYESIA
BayesiaLab 4.5
Bayesian Networks for
Download the BayesiaLab FREE trial version! Have a look at the dynamic presentations and the case studies!
You are here: KDnuggets Home
Data Mining and Analytics Resources
Software
Suites, Text,
Classification, Visualization
• Data Mining / Analytic Jobs
Latest: Cars.com, LYZ Capital
• Academic / Research positions
Latest: Cornell, Nat. U. Ireland, INRIA
• KDnuggets News, the leading newsletter on Data Mining, Analytics & Knowledge Discovery topics
Latest: KDnuggets News 09:n07
Subscribe to KDnuggets!
KDnuggets News Schedule
KDnuggets News Archive | Submit
Miner3D
Miner3D 7.1 - Free Evaluation
MAGNUM OPUS
Find the real associations
Interesting associations: Free Trial!
• Companies
Consulting, Products
• Gregory Piatetsky-Shapiro
Data Mining Consulting
• Domain-specific Solutions
CRM, NEW Twitter, Web
• Datasets
Competitions, KDD Cup
• Useful Data Mining / Analytics sites
Blogs, NEW Twitters, Social
• KDnuggets Polls
NEW Data Mining Salary by Region
• FAQ
DM Tool Comparison
ACM SIGKDD: The Knowledge Discovery and Data Mining Society
• Webcasts: live, on-demand
May 7: Data Mining: Failure to Launch
• Courses
Data Mining, Las Vegas, Apr 27-30
• Data Mining Forums
Open Issues, Beginners, Experts
• Meetings, Conferences
NEW KDD-09 - register here!
• Education:
on-line, USA, Europe
• Data Mining Course
lectures and teaching materials
• Publications
Textbooks, Professional books
Data Mining Crossword
Can you solve it?
About KDnuggets | Site Map | Advertise on KDnuggets | Contact Us
Copyright © 2009 KDnuggets. All Rights Reserved.
Current KDnuggets News
KDN 09:n07: Data Miner Salary Survey; SIGKDD Election; Seth Grimes and Text Analytics
Most popular:
#1 Data Mining Salary survey
#2 Interview with Seth Grimes
#3 SPSS Renames software
NEW KDnuggets on twitter
KDnuggets
Data Mining Forums
Poll
ACM SIGKDD Members can vote at ACM Election site
Everyone can participate in this SIGKDD Election Poll (candidates listed in random order)
For SIGKDD Chair
(choose one)
Bing Liu
Usama Fayyad
For SIGKDD Treasurer
(choose one)
Ismail Parsa
Osmar Zaiane
For SIGKDD Directors
Try
Data Mining
Today!
FREE
Trial Version
sas
THE POWER TO KNOW.
Justification for analytics versus spreadsheets is finally in.
Click here for Aberdeen analyst report.
sas
THE POWER TO KNOW.

Sơ lược lịch sử phát triển cộng đồng KPDL

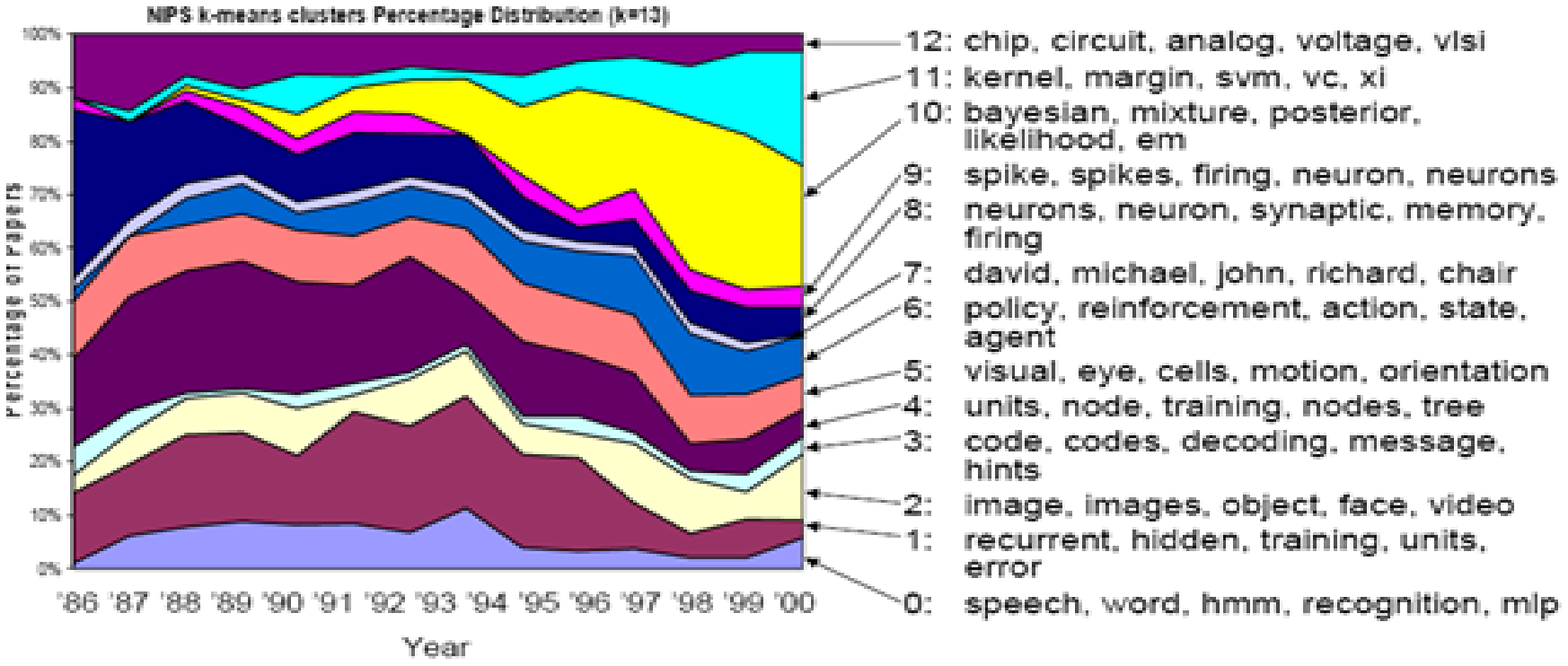


- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, 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 (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.

Khai phá dữ liệu: tốp 20 từ khóa hàng đầu

Các chủ đề liên quan DM đang là thời sự !

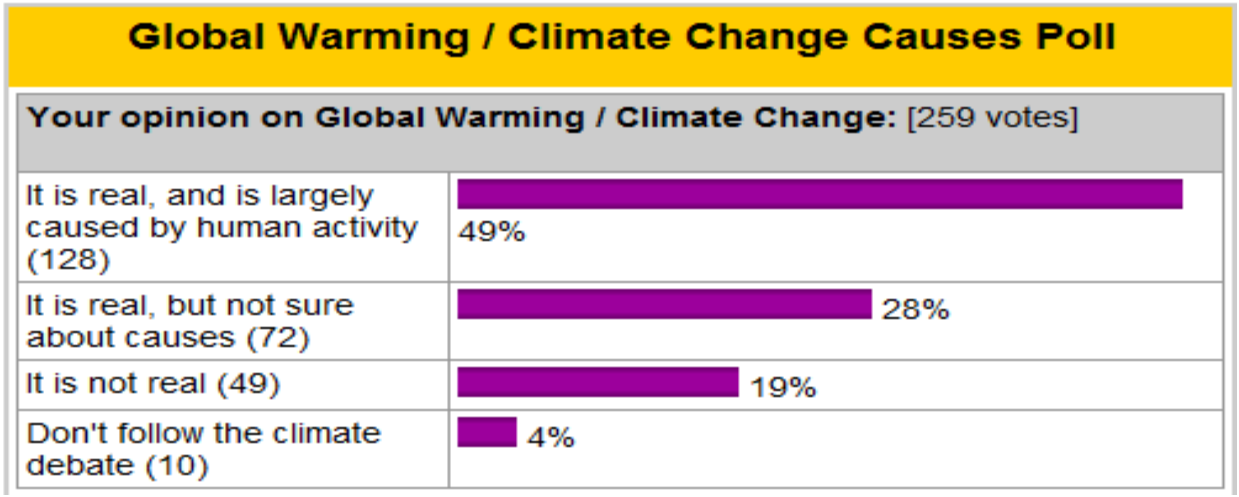
Temporal Cluster Histograms: Results



Hình 1.10. Tình hình phát triển một số nhóm chủ đề trong khoa học máy tính qua phân cụm tài liệu khoa học [CJG06]

Trang web KDD; KPDL & biến đổi khí hậu

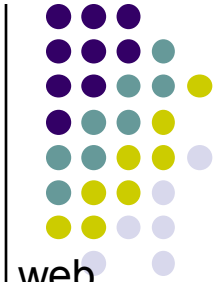
KDnuggets Home » Polls » Global Warming / Climate Change Causes Poll (May 2010)



Nguyên nhân gây biến đổi khí hậu:

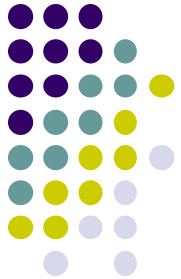
- Gần 50% độc giả KDnuggets tin rằng thay đổi khí hậu hiện nay phần lớn là do hoạt động của con người, một số đáng kể số người nghi ngờ.
- Khí hậu rất phức tạp và các nhà khoa học không phải là tuyên bố rằng hoạt động của con người là nguyên nhân duy nhất của thay đổi khí hậu.
- Đồng thuận với Hội đồng liên chính phủ về Biến đổi khí hậu: hoạt động của con người là một trong những nguyên nhân chính.
- Khai phá nhận định: Opinion Mining / Sentiment Mining

Vấn đề hiện tại trong KPDL



- Phương pháp luận khai phá
  - Khai phá các kiểu tri thức khác nhau từ dữ liệu hỗn tạp như sinh học, dòng, web…
  - Hiệu năng: Hiệu suất, tính hiệu quả, và tính mở rộng
  - Đánh giá mẫu: bài toán về tính hấp dẫn
  - Kết hợp tri thức miền: ontology
  - Xử lý dữ liệu nhiễu và dữ liệu không đầy đủ
  - Tính song song, phân tán và phương pháp KP gia tăng
  - Kết hợp các tri thức được khám phá với tri thức hiện có: tổng hợp tri thức
- Tương tác người dùng
  - Ngôn ngữ hỏi KPDL và khai phá “ngẫu hứng”
  - Biểu diễn và trực quan kết quả KPDL
  - Khai thác tương tác tri thức ở các cấp độ trừu tượng
- Áp dụng và chỉ số xã hội
  - KPDL đặc tả miền ứng dụng và KPDL vô hình
  - Bảo đảm bí mật dữ liệu, toàn vẹn và tính riêng tư

Một số yêu cầu ban đầu



- Sơ bộ về một số yêu cầu để dự án KPDL thành công
  - Cần có kỳ vọng về một lợi ích đáng kể về kết quả KPDL
    - Hoặc trực tiếp nhận được "trái cây treo thấp" ("low-hanging fruit") dễ thu lượm (như Mô hình mở rộng khách hàng qua tiếp thị và bán hàng)
    - Hoặc gián tiếp tạo ra đòn bẩy cao khi tác động vào quá trình sống còn có ảnh hưởng sóng ngầm mạnh (Giảm các nợ khoản khó đòi từ 10% còn 9,8% có số tiền lớn).
  - Cần có một đội dự án thi hành các kỹ năng theo yêu cầu: chọn dữ liệu, tích hợp dữ liệu, phân tích mô hình hóa, lập và trình diễn báo cáo. Kết hợp tốt giữ người phân tích và người kinh doanh
  - Nắm bắt và duy trì các dòng thông tin tích lũy (chẳng hạn, mô hình kết quả từ một loạt chiến dịch tiếp thị)
  - Quá trình học qua nhiều chu kỳ, cần "chạy đua với thực tiễn" (mô hình mở rộng khách hàng ban đầu chưa phải đã tối ưu).
- Một tổng hợp về các bài học KPDL thành công, thất bại

  [NEM09] Robert Nisbet, John Elder, and Gary Miner (2009). Handbook of Statistical Analysis and Data Mining, *Elsevier*, 2009.

KHAI PHÁ DỮ LIỆU

CHƯƠNG 2. PHÁT HIỆN TRI THỨC TỪ DỮ LIỆU

Chapter 2: Phát hiện tri thức từ dữ liệu

- Công nghệ tri thức
- Quản lý tri thức
- Cơ sở của phát hiện tri thức từ dữ liệu
- Bài toán phát hiện tri thức từ dữ liệu
- Một số nội dung liên quan

Công nghệ tri thức

- Vai trò của CNTT trong kinh tế
  - Nghịch lý về tính hiệu quả của CNTT
  - Luận điểm của CARR
  - Bản chất vai trò của CNTT trong kinh tế
- Kinh tế tri thức
  - Khái niệm kinh tế tri thức
  - Bốn cột trụ của nền kinh tế tri thức
  - Các yếu tố đầu vào cốt lõi của kinh tế tri thức: R&D, giáo dục đại học, phần mềm
- Cơ bản về Công nghệ tri thức
  - Khái niệm công nghệ tri thức
  - Nội dung cơ bản của công nghệ tri thức

Vai trò của CNTT

- Nghịch lý hiệu quả của CNTT
  - Robert Solow, nhà kinh tế được giải thưởng Nobel, có nhận định "chúng ta nhìn thấy máy tính ở mọi nơi ngoại trừ trong thống kê hiệu quả statistics." (1987)
  - Căn cứ: Thống kê hiệu quả kinh tế (theo lý thuyết kinh tế cổ điển) và đầu tư CNTT
- Luận điểm của CARR
  - "CNTT không quan trọng": IT does not matter !
  - Nhận đinh về luận điểm của CARR
- Vai trò bản chất của CNTT trong kinh tế
  - Hệ thống tác nghiệp, điều hành
  - Hệ thống phát hiện tri thức

Nghịch lý hiệu quả

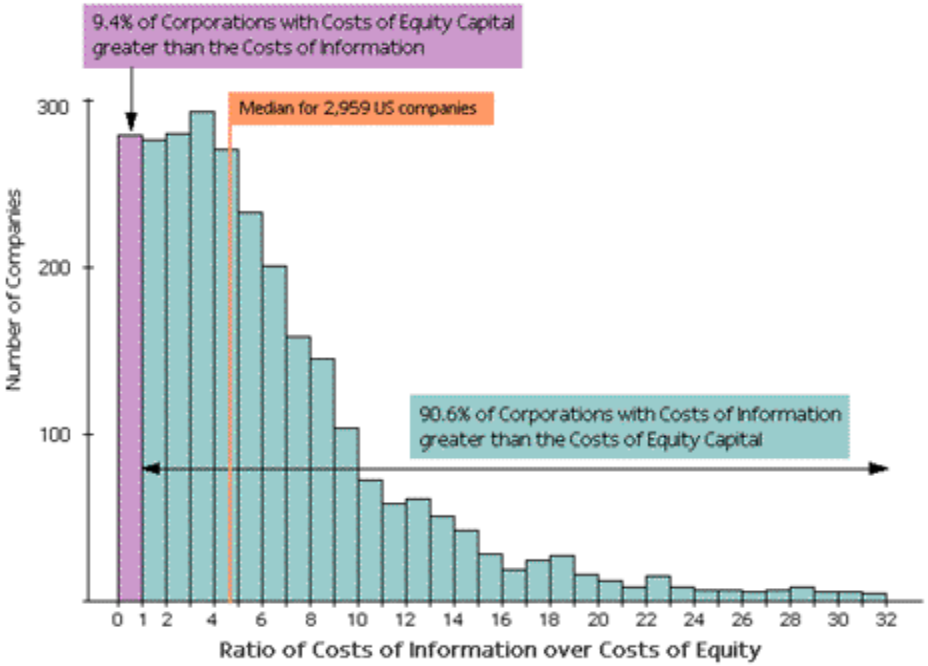
- **"Nghịch lý hiệu quả": Một xung đột của kỳ vọng với thống kê**
  - Mối quan hệ giữa IT và hiệu quả: nhiều tranh luận song hiểu biết vẫn còn rất hạn chế.
    - Năng lực máy tính được đưa vào kinh tế Mỹ đã tăng hơn bậc hai về độ lớn từ năm 1970
    - Hiệu quả, đặc biệt trong khu vực dịch vụ có vẻ đình trệ.
  - Cho một hứa hẹn khổng lồ của IT tới mở ra trong "cuộc cách mạng công nghệ lớn nhất mà loài người từng có" (Snow, 1966),
    - Sự vỡ mộng, thâm chí làm thất vọng với công nghệ gia tăng một cách hiển nhiên: "Không, máy tính không làm tăng hiệu quả, ít nhất không hầu hết thời gian" (Economist, 1990).

Erik Brynjolfsson , The Productivity Paradox of Information Technology: Review and Assessment , Published in Communications of the ACM, December, 1993; and Japan Management Research, June, 1994 (in Japanese)

Toàn nền kinh tế Mỹ: nghịch lý hiệu quả

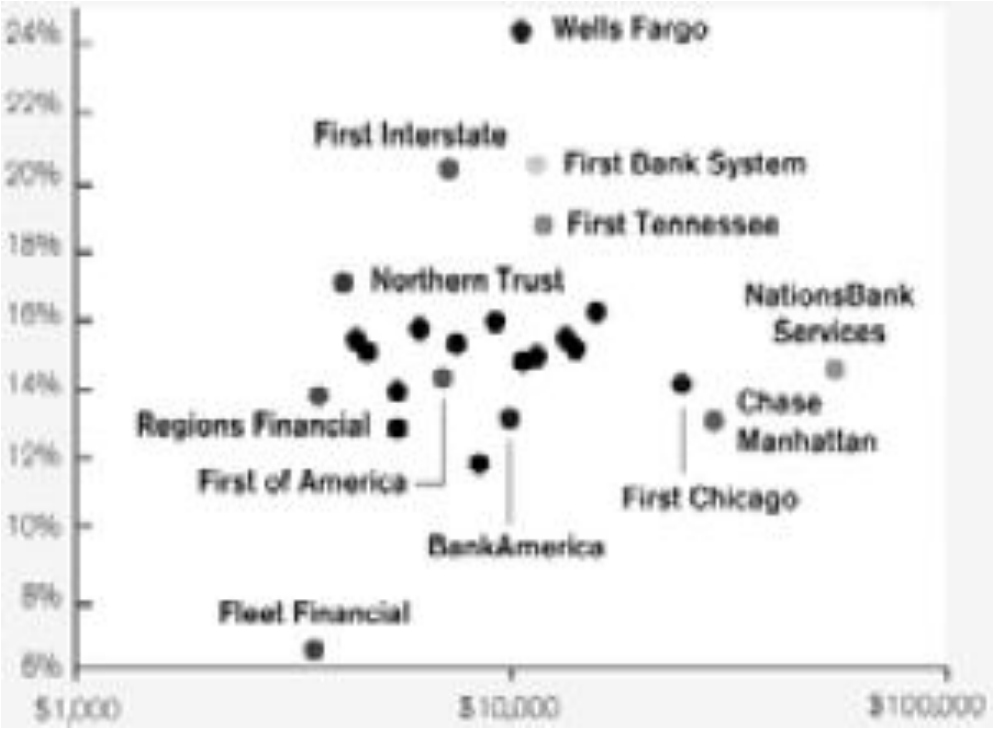
Sự không tương quan trong tăng GNP		
Giai đoạn	Chi phí cho máy tính (%GNP)	Tăng GNP hàng năm
1960s	0.003	4.50%
1970s	0.05	2.95%
1980s	0.3	2.75%
1990s	3.1	2.20%

Nghịch lý hiệu quả: mức công ty



- Trái: Không có quan hệ giữa đầu tư CNTT/nhân viên (trục hoành) với thu hồi vốn (trục tung): tỷ lệ đầu tư nhiều cũng như ít !
- Phải: Có 90,6 % số công ty giá thành CNTT lớn hơn giá thu hồi vốn: đầu tư CNTT lãng phí ? Thu hồi vốn chậm ?
- http://www.strassmann.com/pubs/cf/cf970603.html

Nghịch lý hiệu quả: mức công ty tài chính



- Có quan hệ "tỷ lệ thuận" giữa đầu tư CNTT/nhân viên (trục hoành) với thu hồi vốn (trục tung) tại các công ty tài chính

Luận điểm của G. Carr: IT does'n matter !

- Nicholas G. Carr. IT does'n matter! *HBR at Large*, May 2003: 41-49
  - CNTT xuất hiện khắp nơi và tầm quan trọng chiến lược của nó đã giảm. Cách tiếp cận đầu tư và quản lý CNTT cần phải thay đổi đáng kể !
  - Khi một tài nguyên trở thành bản chất để cạnh tranh nhưng không quan trọng cho chiến lược, rủi ro nó tạo ra trở thành quan trọng hơn các lợi thế mà nó cung cấp.
  - Với các cơ hội đạt được lợi thế chiến lược từ CNTT nhanh chóng biến mất, nhiều công ty sẽ cần có một cái nhìn nghiêm khắc đầu tư vào CNTT và quản lý các hệ thống của họ. Carr đưa ra ba quy tắc hướng dẫn cho tương lai: phủ nhận vai trò chiến lược của CNTT !
- Nicholas G. Carr. The end of corporate computing, *MIT Sloan Management Review*, Spring 2005: 67-73.
- Thuộc 100 người có tên được nhắc đến nhiều nhất !

John Seely Brown, *Former Chief Scientist, Xerox, Palo Alto, California*

John Hagel III, *Management Consultant and Author, Burlingame, California*

If we've learned one thing from the 1990s, it's that big bang, IT-driven initiatives rarely produce expected returns.

Một điều chúng ta học được từ những năm 1990, nó như một vụ nổ vũ trụ, là khởi đầu dựa theo IT hiếm khi tạo ra một đền đáp như được kỳ vọng

Rather than help companies understand that IT is only a tool, technology vendors have tended to present it as a panacea. "Buy this technology and all your problems will be solved."

Nhẽ ra phải giúp các công ty hiểu rằng IT chỉ là một công cụ, các nhà cung cấp công nghệ lại nhằm tới nó như một thuốc bách bệnh "Mua công nghệ này đi và các vấn đề của anh sẽ được giải quyết".

John Seely Brown, *Former Chief Scientist, Xerox, Palo Alto, California*

John Hagel III, *Management Consultant and Author, Burlingame, California*

If we've learned one thing from the 1990s, it's that big bang, IT-driven initiatives rarely produce expected returns.

Một điều chúng ta học được từ những năm 1990, nó như một vụ nổ vũ trụ, là khởi đầu dựa theo IT hiếm khi tạo ra một đền đáp như được kỳ vọng

Rather than help companies understand that IT is only a tool, technology vendors have tended to present it as a panacea. "Buy this technology and all your problems will be solved."

Nhẽ ra phải giúp các công ty hiểu rằng IT chỉ là một công cụ, các nhà cung cấp công nghệ lại nhằm tới nó như một thuốc bách bệnh "Mua công nghệ này đi và các vấn đề của anh sẽ được giải quyết".

F. Warren McFarlan, *Albert H. Gordon Professor of Business Administration, Harvard Business School, Boston*

Richard L. Nolan, *William Barclay Harding Professor of Business Administration, Harvard Business School, Boston*

The jobs of CTO and CIO are and will be of unparalleled importance in the decades ahead.

The package of skills needed inside an organization is changing very fast for competition in the information age.

Công việc của CTO (người đứng đầu bộ phận công nghệ) và CIO (người đứng đầu về TT) sẽ quan trọng chưa từng có trong các thập niên tiếp theo. Gói kỹ năng cần trong một tổ chức sẽ thay đổi rất nhanh để cạnh tranh trong thời đại thông tin.

We wish Carr were right, because everyone's golf handicap could then improve. Unfortunately, the evidence is all to the contrary.

Chúc Carr đúng vì điều bất lợi của mọi người có thể tăng lên. Không may, tất cả minh chứng đều ngược lại !

Jason Hittleman, *IT Director, RKA Petroleum Companies, Romulus, Michigan*

I largely agree with Nicholas Carr's suggestions on how companies should respond to the unbearable reality that IT is becoming more of a commodity. But why does Carr suggest that IT management should become boring? Are leadership tasks such as managing risk and reining in costs any less engaging or challenging than seeking competitive advantage is?

Tôi đồng tình nhiều với khuyến cáo của Nicholas Carr về cách thức các công ty nên phản ứng với thực tế không thể chịu đựng được là IT đã trở thành một loại hàng hóa.
Nhưng tại sao Carr lại khuyến cáo các nhà quản lý IT sẽ trở nên buồn rầu ? Phải chăng là vì các bài toán lãnh đạo như quản lý và kiểm soát rủi ro về kinh phí ít hứa hẹn hoặc thách thức hơn so với theo đuổi lợi thế cạnh tranh ?

IT will always matter—it will just matter in different ways now. IT must continue to support the business—not just through the logical application of technologies but also through the logical application of common sense.

IT luôn luôn quan trọng – là vấn đề trong mọi quan niệm. **IT bắt buộc hỗ trợ kinh doanh** – không chỉ bằng áp dụng lôgíc về công nghệ mà còn bằng áp dụng lôgic về bản chất chung.

Tri thức và kinh tế tri thức

Tri thức

- Khái niệm
  - Từ điển *Compact Oxford English Dictionary*
    - sự hiểu biết tinh thông cùng với các kỹ năng mà con người thu nhận được qua kinh nghiệm hoặc giáo dục
    - tổng hợp những gì mà con người biết rõ
    - nhận thức và hiểu biết tường minh về một sự việc hoặc một hiện tượng mà thu nhận được nhờ kinh nghiệm
  - http://en.wikipedia.org/wiki/Knowledge hoặc http://vi.wikipedia.org/wiki/Tri_thức
  - Nội dung khái niệm còn phụ thuộc vào từng lĩnh vực:
    - Ở đây: *Compact Oxford English Dictionary*
    - Khai phá dữ liệu: *mẫu có độ hấp dẫn vượt qua ngưỡng*
- Hình thức thu nhân tri thức: giáo dục, kinh nghiệm qua hoạt động thực tiễn

Phân loại tri thức

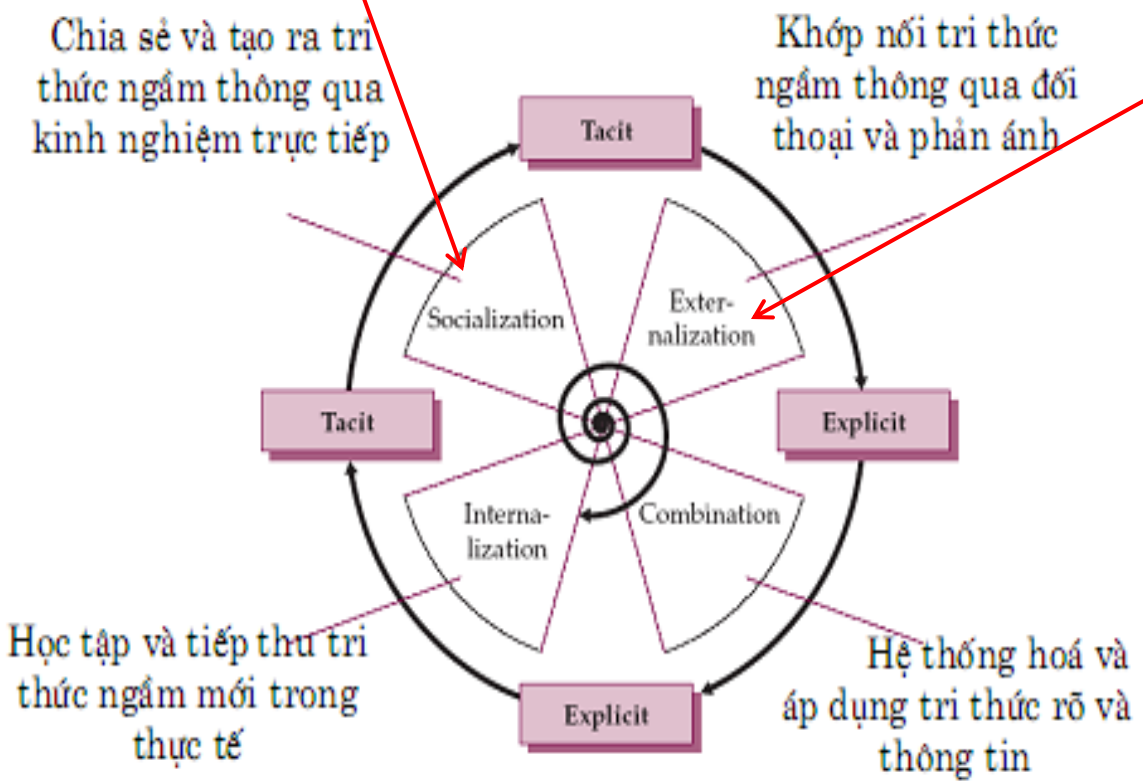
	Tri thức hiện	*Tri thức ẩn*
Tri thức "that"	lý thuyết, khái niệm…	nhận thức, sự phán đoán…
Tri thức "how"	phương pháp, thủ tục…	khả năng, kỹ năng…

	Tri thức hiện	*Tri thức ẩn*
Tri thức chủ quan	sự kiện, chân thực, quan sát...	hiểu trực giác về sự kiện…
Tri thức khách quan	quan điểm rõ ràng, niềm tin…	giả định ẩn, thế giới quan…

- tri thức hiện – tri thức ẩn (Explicit knowledge – Tacit knowledge), tri thức chủ quan – tri thức khách quan (Objective knowledge – Subjective knowledge), tri thức biết – tri thức hành động (Knowing that – Knowing how). Ví dụ tri thức ẩn ⇨ tri thức hiện: **ngành CNPM**
- "know what": tri thức về sự vật, sự kiện, hiện tượng
- "know why": tri thức về thế giới, xã hội và trí tuệ con người,
- "know who": tri thức về ai và họ làm được gì,
- "know how": tri thức về kỹ năng và kinh nghiệm thực tiễn.
- "know where", "know when": tri thức quan trọng cho một nền kinh tế mềm dẻo và động,

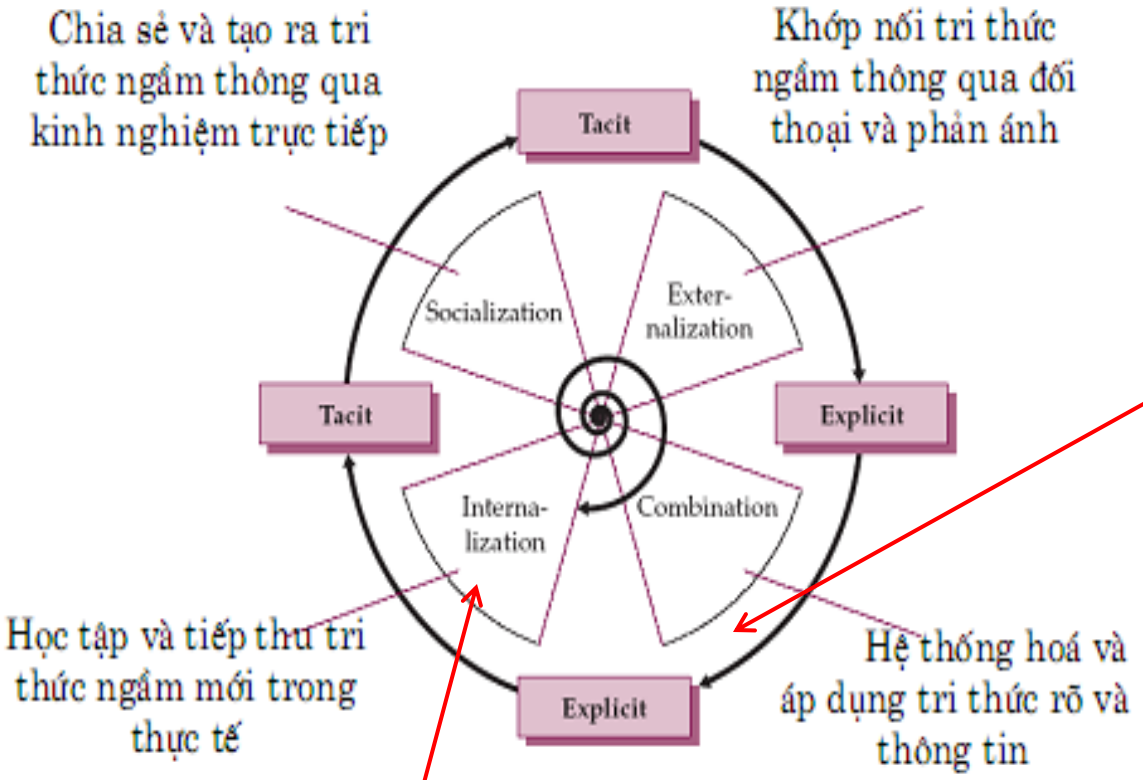
Chuyển hóa tri thức

- *Xã hội hóa* (Socialization): quá trình chia sẻ kinh nghiệm và do đó tạo ra tri thức ẩn (tri thức của cá nhân bao gồm nhận định, sự hiểu biết, niềm tin và trực giác, tri thức tiềm ẩn, cá nhân hóa sâu sắc và trình bày khuếch tán trong phạm vi tổ chức). Một cá nhân có thể tiếp thu tri thức ẩn của cá nhân khác mà không cần sử dụng ngôn ngữ. Bắt chước được coi là một phương tiện đào tạo xã hội.



- *Ngoại diên hóa* (Externalization) là quá trình chuyển đổi tri thức ẩn thànhtri thức hiện (tri thức hình thức, dễ tiếp cận, tương đối dễ dàng lây truyền giữa các cá nhân và nhóm) bằng cách sử dụng phép ẩn dụ, tương tự hóa và các mô hình.
- Ngoại diên hóa tri thức ẩn là hoạt động quan trọng nhất liên quan đến việc tạo ra tri thức, nhưng cũng là khó khăn nhất.

Chuyển hóa tri thức



- *Kết hợp / trộn* (Combination/mixing) là quá trình tạo ra tri thức rõ bằng cách kết hợp tri thức từ các nguồn khác nhau. Vì vậy, các cá nhân thay đổi và kết hợp tri thức rõ của họ bằng cách chuyển đổi các cuộc họp qua điện thoại.
- **Thông tin có trong CSDL có thể được xử lý để tạo tri thức rõ mới (KDD)**

- *Chủ quan hóa* (Internalization) là quá trình bao chứa các tri thức rõ ràng vào tri thức ẩn. Điều này tạo điều kiện thuận lợi nếu cá nhân có thể lại trải nghiệm những kinh nghiệm của những người khác, gián tiếp.

Trong công ty sáng tạo tri thức, bốn mô hình chuyển đổi tri thức làm việc theo tương tác động, trong một xoắn ốc tri thức.

Kinh tế tri thức

Khái niệm

- Knowledge Economy/Knowledge-Based Economy
- [WB06] *nền kinh tế mà việc sử dụng tri thức là động lực chủ chốt cho tăng trưởng kinh tế*. Trong nền kinh tế tri thức, tri thức được yêu cầu, được phát sinh, được phổ biến và được vận dụng một cách hiệu quả cho tăng trưởng kinh tế.
- [UN00] nền kinh tế mà các yếu tố then chốt cho sự phát triển là tri thức, năng lực trí tuệ, một thiết chế xã hội cho một hạ tầng thông tin hữu hiệu và truy nhập được.
- Hai định nghĩa trên là tương tự nhau: ở đây sử dụng định nghĩa [WB06].

Kinh tế tri thức: đặc trưng

Bốn cột trụ của một nền kinh tế tri thức

- Một thiết chế xã hội pháp quyền và khuyến khích kinh tế (*An economic incentive and institutional regime*)

  Cột trụ này bao gồm các chính sách và thể chế kinh tế tốt, khuyến khích phân phối hiệu quả tài nguyên, kích thích cách tân và thúc đẩy phát kiến, phổ biến và sử dụng các tri thức đang có.

- Một lực lượng lao động được giáo dục và lành nghề (*An educated and skilled labor force*)

  Cột trụ này bao gồm các yếu tố về năng lực tri thức của nguồn nhân lực trong nền kinh tế. Các thông số về giáo dục và sáng tạo được lựa chọn nhằm thể hiện tiềm năng nói trên. Xã hội học tập và hoạt động học tập suốt đời cũng là các yếu tố đảm bảo tăng cường tiềm năng tri thức của nền kinh tế.

Kinh tế tri thức: đặc trưng

Bốn cột trụ của một nền kinh tế tri thức

- Một hệ thống cách tân hướng tri thức hiệu quả (a *effective innovation system)*

  Nền kinh tế tri thức cần là một nền kinh tế cách tân hiệu quả của các tập đoàn, trung tâm nghiên cứu, trường đại học, các chuyên gia và các tổ chức khác , trong đó, tri thức khi mà đã trở nên lỗi thời - lạc hậu cần liên tục được thay thế bằng tri thức mới - tiến bộ phù hợp với trình độ phát triển của nền kinh tế tri thức. Trong nền kinh tế tri thức, hoạt động không ngừng cách tân tri thức, phát huy sáng kiến mang tính xã hội.

- Một hạ tầng thông tin hiện đại và đầy đủ (a *modern and adequate information infrastructure*) là phương tiện hiệu quả để truyền thông, phổ biến và xử lý thông tin và tri thức

  Hạ tầng thông tin hiện đại và đầy đủ đảm bảo hoạt động thu nhận, cách tân tri thức cũng như để đảm bảo xã hội học tập và hoạt động học tập suốt đời.

Kinh tế tri thức: đo lường

- Là một công việc khó khăn: Từ chính Khái niệm tri thức và nội dung 4 cột trụ [OEC96, RF99, CD05]
- [Ram08] nhận định "*người ta ngày càng nhận thức rõ hơn rằng tri thức về tăng trưởng kinh tế không hoàn toàn rõ ràng như ta vẫn tưởng*".
- [OEC96] xác định 4 khó khăn nguyên tắc (trang sau)
- Thông qua hệ thống tiêu chí: Đầu ra của kinh tế tri thức
- Đang trong quá trình hình thành và cải tiến:
  - Hệ thống tiêu chí
  - Đo lường từng tiêu chí
  - Tổng hợp các tiêu chí

Kinh tế tri thức: đo lường

BỐN NGUYÊN TẮC [OEC96]

- Không có một công thức hoặc một cách làm ổn định để chuyển dịch các đầu vào của nguồn tạo tri thức thành đầu ra tri thức.

  tính phức tạp của quá trình nhận thức cho nên không thể có một công thức hay cách làm nói trên.

  hệ thống các tiêu chí thể hiện được tiềm năng tạo ra tri thức cho nền kinh tế ? công thức định lượng đúng tuyệt đối

  Ví dụ, đầu tư cho khoa học – công nghệ <=> kinh tế tri thức

- Việc lên sơ đồ cho đầu vào của bộ tạo tri thức là rất khó khăn vì chưa có cách thức thống kê tri thức tương tự như cách thức thống kê quốc dân truyền thống.

  Việc chọn các tiêu chí trong hệ thống đánh giá kinh tế tri thức vẫn đang được nghiên cứu đề xuất, chẳng hạn hệ thống đo lường kinh tế tri thức của Ngân hàng thế giới (KAM) được đổi mới theo thời gian

Kinh tế tri thức: đo lường

BỐN NGUYÊN TẮC [OEC96]

- Thiếu tri thức về một hệ thống định giá có tính phương pháp luận để làm cơ sở kết hợp các phần tử tri thức thành một thành phần bản chất duy nhất.

  Thành phần bản chất duy nhất được đề cập ở đây là được dùng để làm giá trị đo mức độ "tri thức" của một nền kinh tế. Chẳng hạn, trong hệ thống KAM, việc "đo" cho từng tiêu chí cũng như tổng hợp các giá trị đó thành giá trị "đo" mức độ kinh tế tri thức của một quốc gia vẫn chưa có tính phương pháp luận hoàn toàn.

- Việc tạo tri thức mới không cần phải bổ sung mạng vào kho tri thức và sự lạc hậu của các phần tử trong kho tri thức là không được văn bản hóa.

Kinh tế tri thức: đo lường

CÁC BÀI TOÁN CẦN GiẢI QUYẾT [OEC96]

- Đo lường tri thức của đầu vào.
- Đo lường kho tri thức và tri thức trong kho.
- Đo lường tri thức của đầu ra
- Đo lường mạng tri thức
- Đo lường tri thức thông qua học tập

Yogesh Malhotra [Mal03] trình bày hệ thống về mô hình đánh giá kinh tế tri thức của một quốc gia.

- phân tích nội dung, điểm mạnh và điểm hạn chế của một số hệ thống đánh giá điển hình.
- đề xuất một mô hình đánh giá kinh tế tri thức của một quốc gia
- hệ thống đo lường kinh tế tri thức phổ biến: có KAM của WB

Đo lường kinh tế tri thức: KAM

KAM - Knowledge Assessment Methodology [CD05]

- Đo lường điển hình KTTT
  - Chi tiết hóa 4 cột trụ bằng hệ thống tiêu chí
  - Đang được cải tiến
  - 2005: 80 tiêu chí; 2008: 83 tiêu chí; 2009: 109 tiêu chí

Stt	*Cột trụ*	*Nhóm tiêu chí con*	*Số lượng tiêu chí của KAM*		
			KAM-2005 (80)	*KAM-2008 (83)*	*KAM-2009 (109)*
1	Thiết chế xã hội pháp quyền và khuyến khích kinh tế	Năng lực kinh tế	9	9	6
		Thể chế kinh tế	10	12	12
		Điều hành chính quyền	7	7	7
2	Lực lượng lao động được giáo dục và có kỹ năng	Giáo dục	14	14	15
		Giới tính	6	5	5
		Lao động	×	×	24
3	Hệ thống cách tân hướng tri thức hiệu quả	×	22	24	28
4	Hạ tầng thông tin hiện đại và đầy đủ	×	12	12	12

Đo lường kinh tế tri thức: KAM

Một số giải thích

- Tiêu đề *Điều hành chính quyền* được chuyển từ các tiêu đề tiếng Anh là *Institutions* (KAM-2005) và *Governance* (KAM-2008, KAM-2009)
- Hệ thống KAM chứa một số tiêu chí có nội dung liên quan trực tiếp tới kinh tế dịch vụ, chẳng hạn như các tiêu chí *Employment in Services* (%), *Local availability of specialized research and training services*,
- Cột trụ *Hệ thống cách tân* được thi hành trong các tập đoàn, trung tâm nghiên cứu, trường đại học, các chuyên gia và các tổ chức khác nhằm đảm bảo sự tiến hóa tri thức, chuyển đổi thành dòng tăng trưởng tri thức tổng thể, đồng hóa và làm phù hợp tri thức mới cho nhu cầu địa phương
- Cột trụ *Hạ tầng thông tin hiện đại và đầy đủ* đảm bảo hệ thống phương tiện hiệu quả để truyền thông, phổ biến và xử lý thông tin và tri thức

Cơ bản về công nghệ tri thức

- Khái niệm công nghệ tri thức
  - Công nghệ tri thức là một quá trình về cơ bản bao gồm việc thu nhận và biểu diễn tri thức, và xây dựng cơ chế suy luận và giải thích.
- Bốn bước thi hành
  - thu nhận tri thức, biểu diễn tri thức, xây dựng một cơ chế suy luận, và thiết kế các công cụ giải thích.
- Một số khái niệm liên quan
  - Metaknowledge: tri thức về tri thức. Một số ví dụ: làm thế nào để sử dụng tri thức trong các tình huống cụ thể, làm thế nào để xác định những tri thức có liên quan, và khi nào tri thức là chưa đầy đủ. So sánh với metadata (dữ liệu về dữ liệu): dữ liệu mô tả file
  - Metaknowledge: YKYN, YDYK, YKYD, YDYD

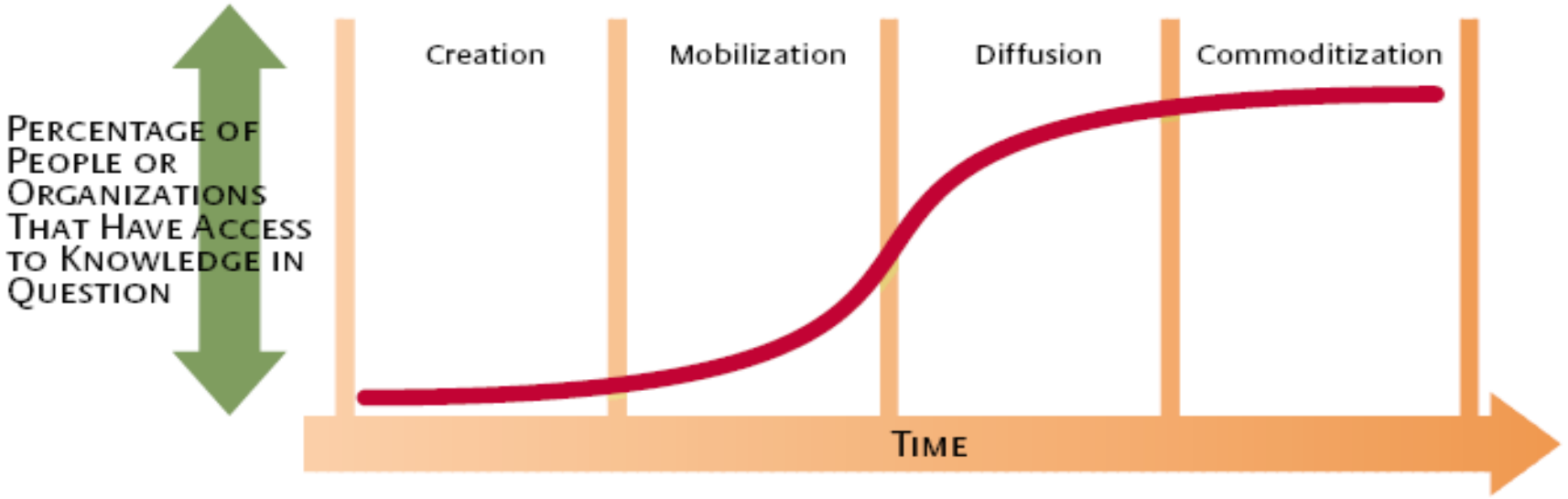
Cơ bản về công nghệ tri thức

- Một số khái niệm
  - Thu nhận tri thức là việc khai thác tri thức từ nguồn (chuyên gia) đã văn bản hóa và chưa văn bản hóa và chuyển nó vào máy tính. Sử dụng 3 kỹ thuật: quy nạp, lập luận dựa trên trường hợp, tính toán neuron.
  - Biểu diễn tri thức liên quan đến việc tổ chức các tri thức trong các cơ sở tri thức.
  - Tri thức văn bản hóa có trong sách vở, đĩa máy tính, báo cáo, phim… Tri thức không văn bản hóa có trong tâm trí con người. Tri thức văn bản hóa là mục tiêu (dù có thể được diễn giải một cách chủ quan)
- Một số nguồn tri thức
  - Chuyên gia, **sách hướng dẫn, phim ảnh, sách, cơ sở dữ liệu**, **tập tin văn bản, hình ảnh, băng hình, cảm biến, và các bức ảnh chụp**.

Chapter 2: Phát hiện tri thức từ dữ liệu

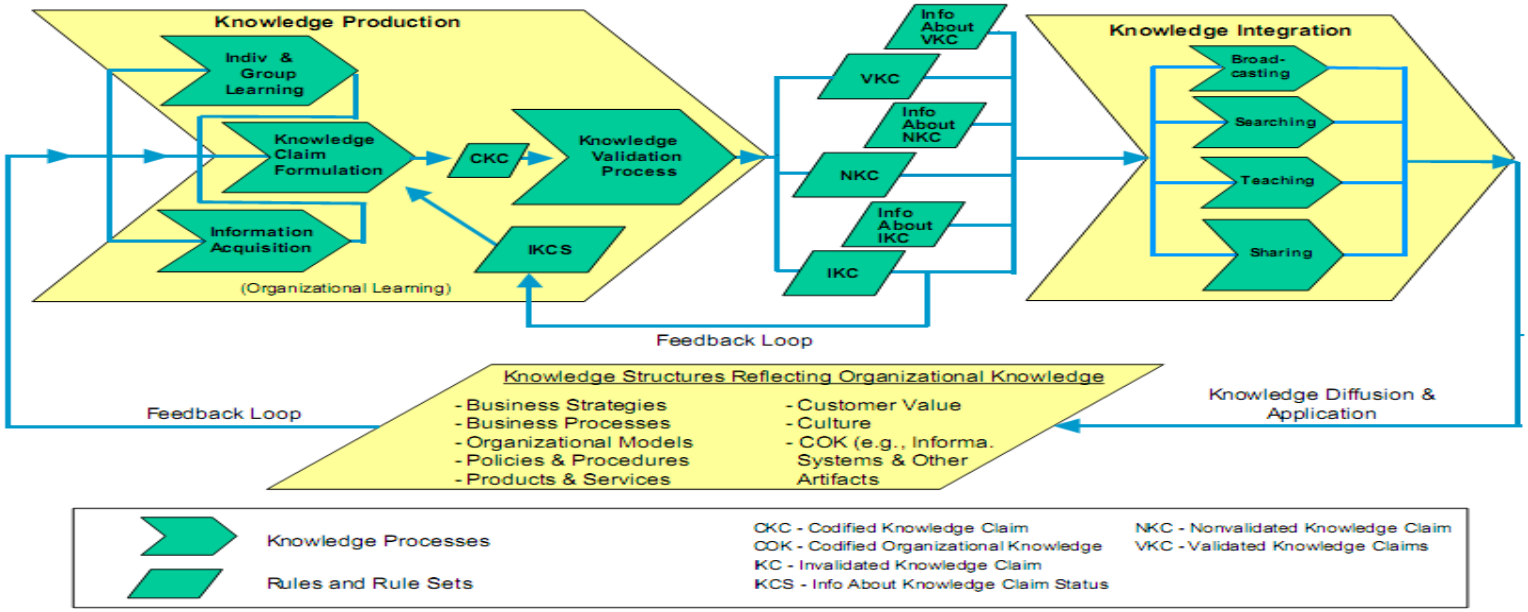
- Công nghệ tri thức
- Quản lý tri thức
- Cơ sở của phát hiện tri thức từ dữ liệu
- Bài toán phát hiện tri thức từ dữ liệu
- Một số nội dung liên quan

Quản lý tri thức trong tổ chức



- Tri thức tiến bộ thông qua bốn giai đoạn là nó phát triển theo thời gian: khởi tạo, huy động, phổ biến, hàng hóa hóa
- Khi nó trở nên truy cập vào nhiều hơn và nhiều người - đầu tiên trong một tổ chức, sau đó tại nhiều tổ chức, và cuối cùng cho công chúng nói chung - các công ty phải sử dụng chiến lược khác nhau để nhận ra giá trị lớn nhất của nó.

The Knowledge Life Cycle

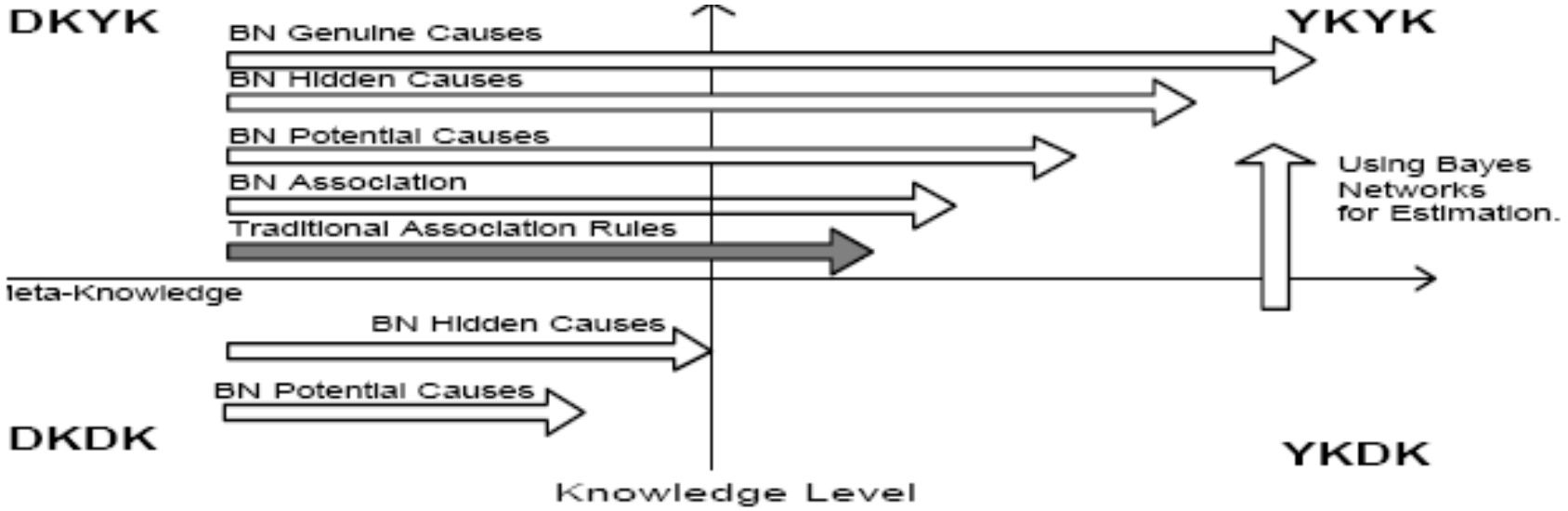


- CKC - Codified Knowledge Claim: Yêu cầu tri thức hệ thống hóa
- UKC - Unvalidated Knowledge Claim: Yêu cầu tri thức không hợp lệ
- VKC - Validated Knowledge Claim: Yêu cầu tri thức hợp lệ
- IK - Invalidated Knowledge: Tri thức hết hiệu lực
- IKC - Invalidated Knowledge Claim: Yêu cầu tri thức hết hiệu lực
- OK - Organizational Knowledge: Tri thức của tổ chức

Chapter 2: Phát hiện tri thức từ dữ liệu

- Công nghệ tri thức
- Quản lý tri thức
- Cơ sở của phát hiện tri thức từ dữ liệu
- Bài toán phát hiện tri thức từ dữ liệu
- Một số nội dung liên quan

Chuyển đổi meta-knowledge



Knowledge/Meta-knowledge Diagram

- Hầu hết kỹ thuật khai phá dữ liệu chuyển hóa DKYK [redacted]KYK.
- Cựu giám đốc điều hành HP, Lew Platt đã từng nói, "Nếu HP biết được những gì HP biết, chúng tôi sẽ có ba lần lợi nhuận"

Tiếp cận truyền thống và tiếp cận KPDL

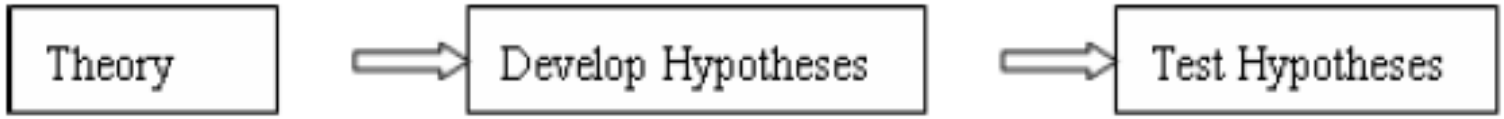


Figure 1. Traditional research approach

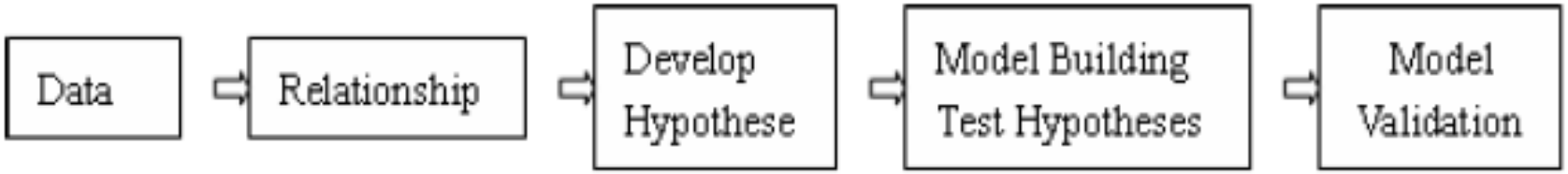


Figure 2. Data mining approach

- Tiếp cận truyền thống
  - Từ lý thuyết (hệ toán mệnh đề) [redacted]hát triển các giả thuyết [redacted] kiểm định (chứng minh) giả thuyết. Ngô Bảo Châu: Bổ đề cơ bản
- Tiếp cận khai phá dữ liệu
  - Từ dữ liệu [redacted]hát hiện quan hệ [redacted]hát triển giả thuyết [redacted]ây dựng mô hình và kiểm định giả thuyết [redacted]ánh giá mô hình [redacted]ử dụng mô hình.

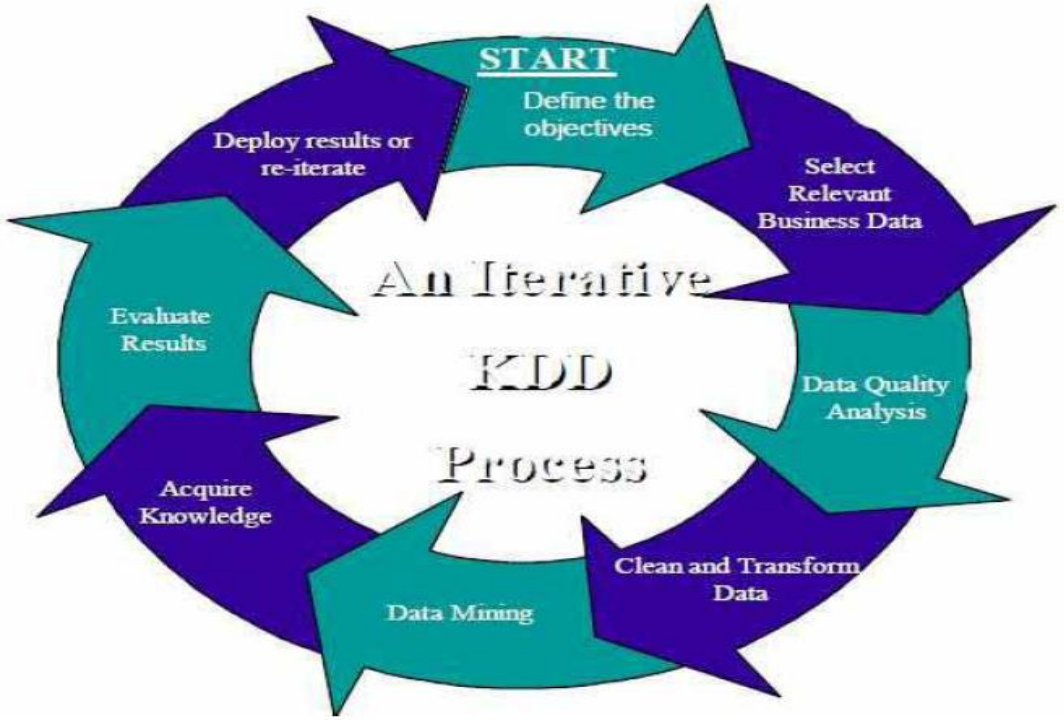
Chapter 2: Phát hiện tri thức từ dữ liệu

- Công nghệ tri thức
- Quản lý tri thức
- Cơ sở của phát hiện tri thức từ dữ liệu
- Bài toán phát hiện tri thức từ dữ liệu
- Một số nội dung liên quan

Bài toán phát hiện tri thức

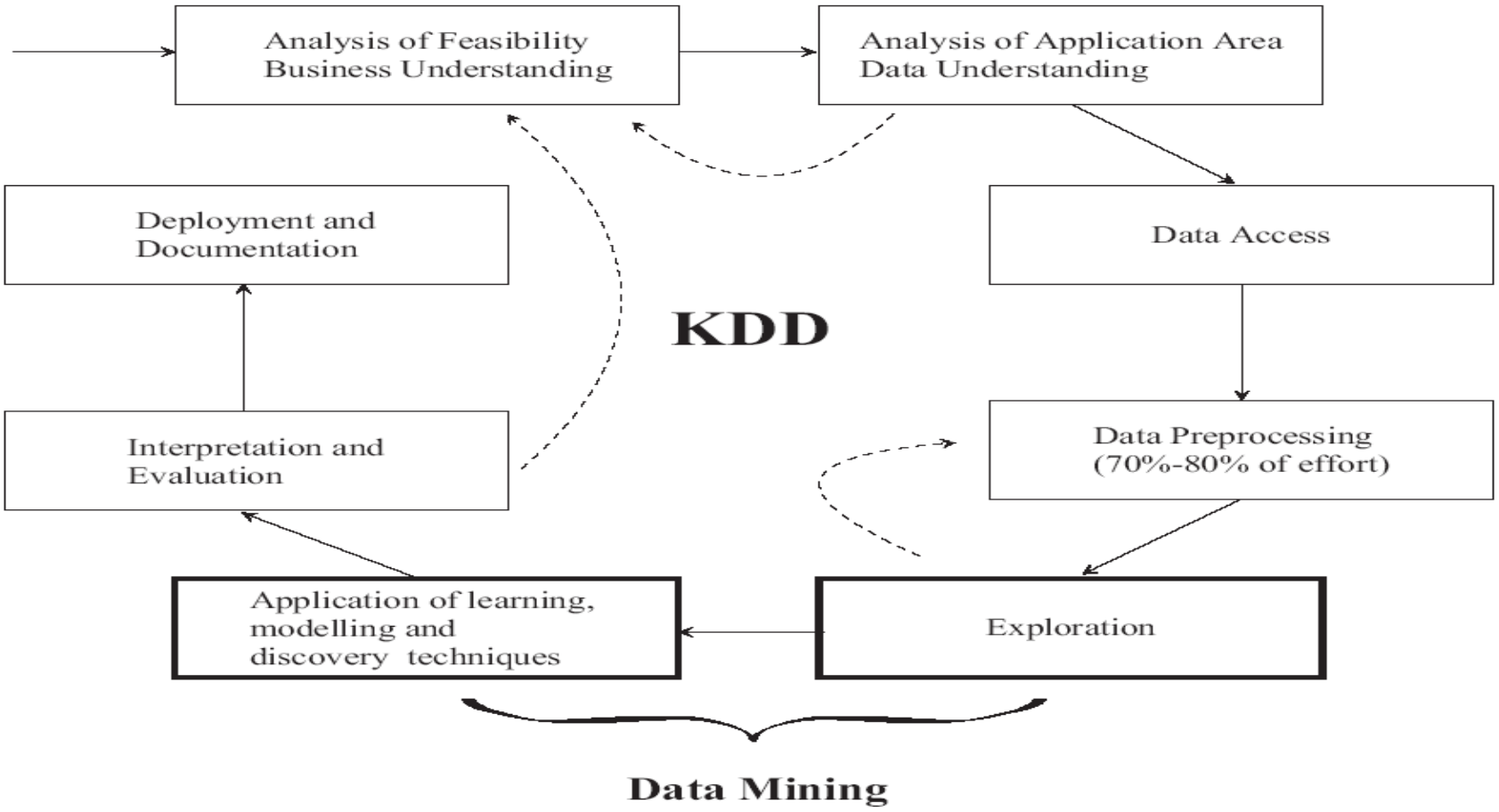
- Nội dung cơ bản của KDD và DM
  - Khai phá dữ liệu và phát hiện tri thức trong CSDL là bài toán "kinh doanh", bài toán "chiến lược" mà không phải là bài toán công nghệ.
- Khi nào nên khai phá dữ liệu
  - Ví dụ: Chương 3 sách Data Mining: Methods and Tools, 1998.

Mô hình vòng khai phá dữ liệu DN'98



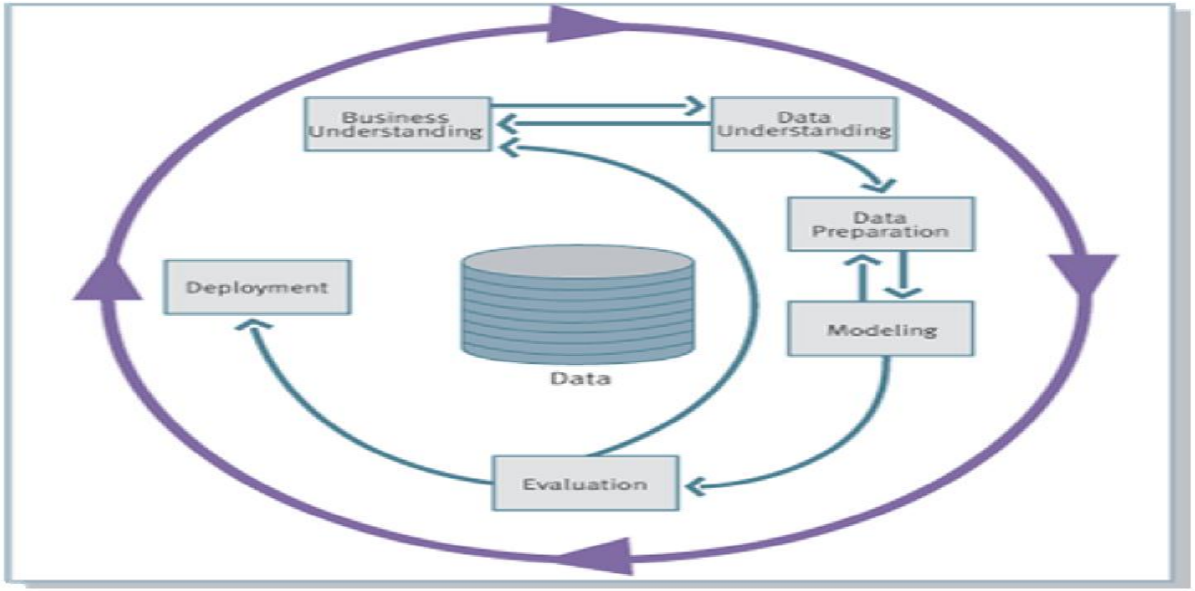
- Mô hình năm 1998

Một mô hình khai phá dữ liệu DN'00



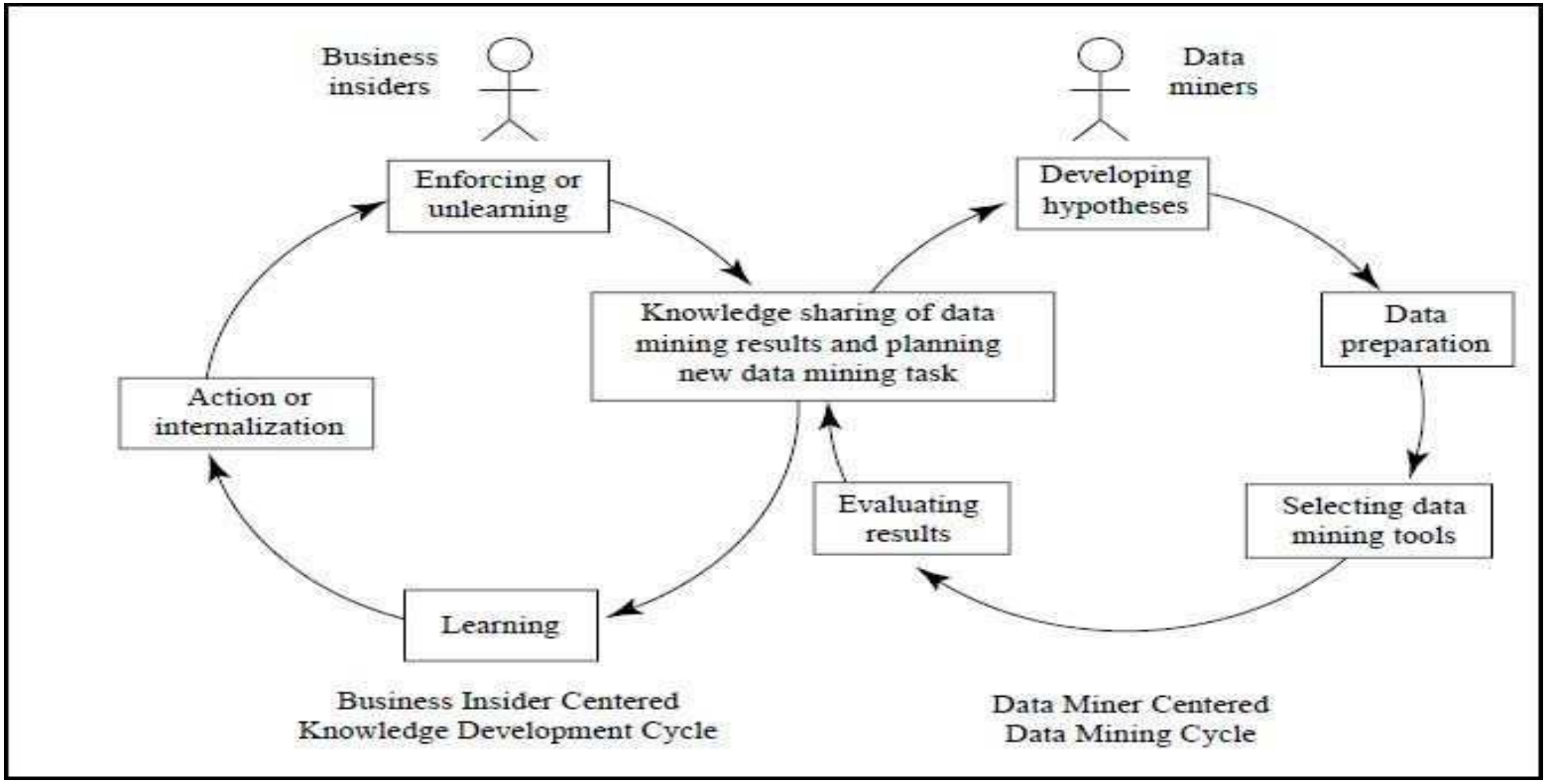
- Một mô hình KDD năm 2000 [Nac00]

Chuẩn công nghiệp khai phá dữ liệu CRISP-DM



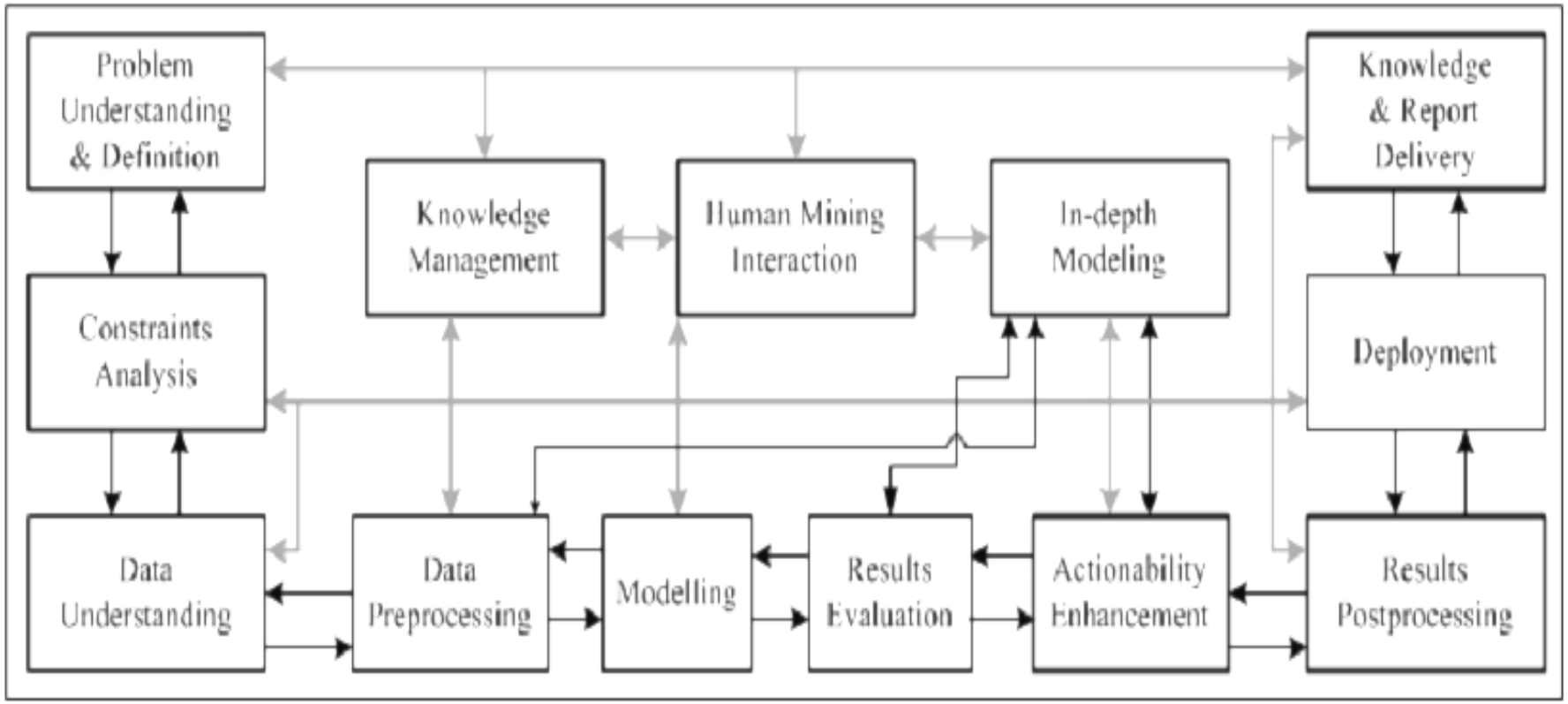
- Các pha trong mô hình quy trình CRISP-DM (Cross-Industry Standard Process for Data Mining). “Hiểu kinh doanh”: hiểu bài toán và đánh giá
- Thi hành chỉ sau khi tham chiếu kết quả với “hiểu kinh doanh”
- CRISP-DM 2.0 SIG WORKSHOP, LONDON, 18/01/2007
- Nguồn: http://www.crisp-dm.org/Process/index.htm (13/02/2011)

Mô hình KPDL và mô hình kinh doanh'08



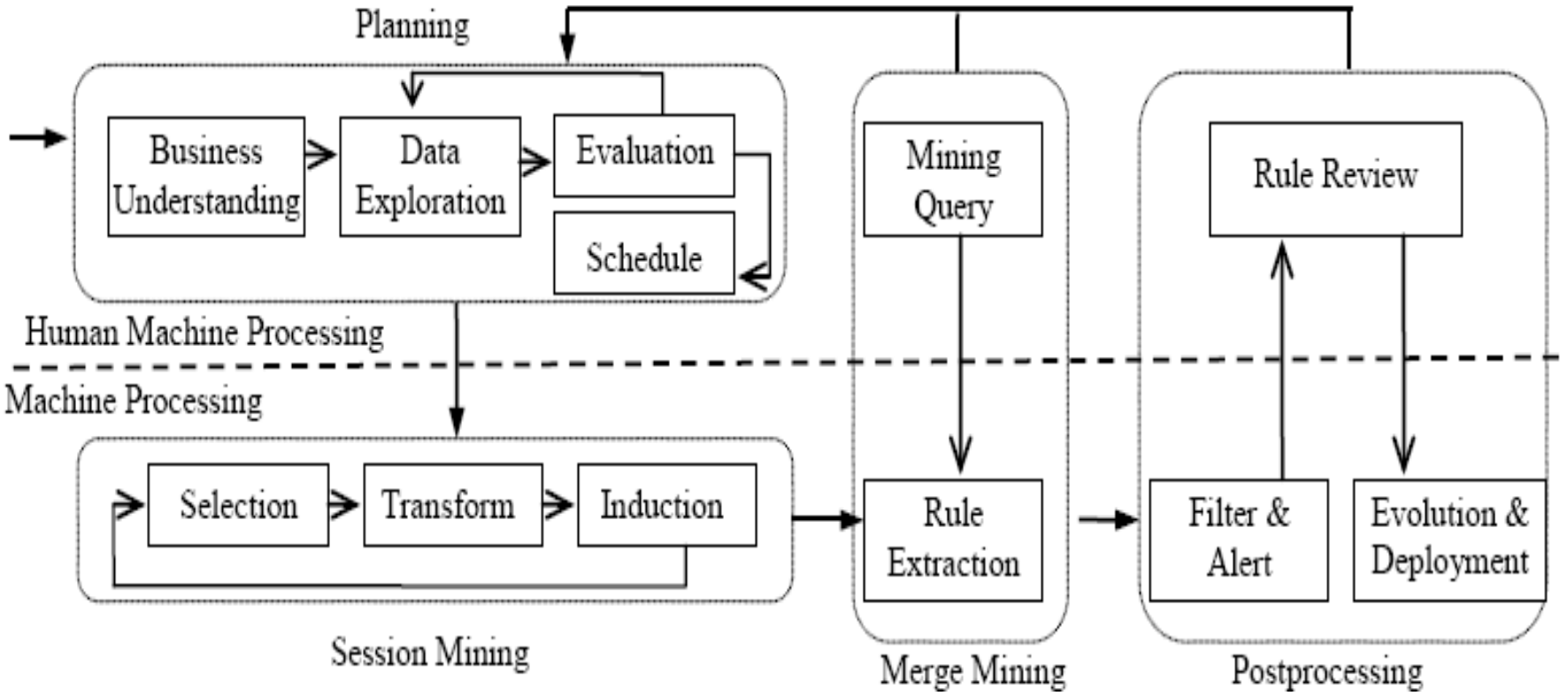
- Wang, H. and S. Wang (2008). A knowledge management approach to data mining process for business intelligence, *Industrial Management & Data Systems*, 2008. **108**(5): 622-634. **[Oha09]**

Một mô hình KPDL hướng ứng dụng



- Mô hình quá trình khai phá dữ liệu hướng miền ứng dụng [CYZ10]

Tương tác người-máy trong KPDL'10



- Mô hình quá trình C-KDD

Chapter 2: Phát hiện tri thức từ dữ liệu

- Công nghệ tri thức
- Quản lý tri thức
- Cơ sở của phát hiện tri thức từ dữ liệu
- Bài toán phát hiện tri thức từ dữ liệu
- Một số nội dung liên quan

Một số vấn đề liên quan

- Đô đo "tri thức"
  - Tri thức [redacted]mẫu có giá trị"
  - Mỗi bài toán KPDL thường đi kèm độ đo: phân lớp có độ đo đánh giá (chính xác + hồi tưởng, chính xác + lỗi), phân cụm: đo theo từng phương pháp, luật kết hợp (độ hỗ trợ + độ tin cậy)…
  - Độ đo là nội dung nghiên cứu trong KPDL
- Lựa chọn thuật toán
  - Không có thuật toán "tốt nhất" cho mọi bài toán khai phá dữ liệu.
  - Kết hợp giải pháp
- Vai trò dữ liệu mẫu
  - Dữ liệu học, dữ liệu kiểm tra.
- Vai trò của người sử dụng.

Bài giảng môn học

KHAI PHÁ DỮ LIỆU

CHƯƠNG 3. TIỀN XỬ LÝ DỮ LIỆU

Tài liệu tham khảo

- [HK06] J. Han and M. Kamber (2006). Data Mining-Concepts and Techniques (Second Edition), *Morgan Kaufmann*. **Chapter 2. Data Preprocessing**
- [NEM09] Robert Nisbet, John Elder, and Gary Miner (2009). Handbook of Statistical Analysis and Data Mining, Elsevier, 6/2009. **Chapter 4. Data Understanding and Preparation; Chapter 5. Feature Selection.**
- [Chap05] Chapman, A. D. (2005). Principles of Data Cleaning, *Report for the Global Biodiversity Information Facility,* Copenhagen
- [Chap05a] Chapman, A. D. (2005a). Principles and Methods of Data Cleaning – Primary Species and Species- Occurrence Data (version 1.0), *Report for the Global Biodiversity Information Facility,* Copenhagen
- [Hai02] Đoàn An Hải (2002). Learning to Map between Structured Representations of Data, *PhD Thesis*, The University of Washington, **ACM 2003 Award Winners and Fellows (Doctoral Dissertation Award)**.
- [RD00] Erhard Rahm, Hong Hai Do (2000). Data Cleaning: Problems and Current Approaches, *IEEE Data Eng. Bull.*, **23**(4): 3-13 (2000)
- và một số tài liệu khác

Chapter 3: Tiền xử lý dữ liệu

- Hiểu dữ liệu và chuẩn bị dữ liệu
- Vai trò của tiền xử lý dữ liệu
- Làm sạch dữ liệu
- Tích hợp và chuyển dạng dữ liệu
- Rút gọn dữ liệu
- Rời rạc và sinh kiến trúc khái niệm

Những vấn đề cơ bản để hiểu dữ liệu

- Cách thu thập được dữ liệu cần thiết để mô hình hóa:
  - Data Acquisition
- Cách kết hợp dữ liệu tìm được từ các nguồn dữ liệu khác nhau
  - Data Integeation.
- Mô tả dữ liệu
  - Data Description
- Đánh giá chất lượng (sự sạch sẽ) của dữ liệu
  - Data Assessment

Thu thập dữ liệu

- Cách thu thập dữ liệu cần thiết để mô hình hóa Data Acquisition:
  - Trích chọn dữ liệu theo câu hỏi từ CSDL tới tập tin phẳng
  - Ngôn ngữ hỏi bậc cao truy nhập trực tiếp CSDL
  - Kết nối mức thấp để truy nhập trực tiếp CSDL
    - Loại bỏ ràng buộc không gian/thời gian khi di chuyển khối lượng lớn dữ liệu
    - Hỗ trợ việc quản lý và bảo quản dữ liệu tập trung hóa
    - Rút gọn sự tăng không cần thiết của dữ liệu
    - Tạo điều kiện quản trị dữ liệu tốt hơn để đáp ứng mối quan tâm đúng đắn

Tích hợp dữ liệu

- Cách kết hợp dữ liệu tìm được từ các nguồn dữ liệu khác nhau Data Integeation.

File #1: Name & Address

Name	Address	City	State	Zipcode
John Brown	1234 E St.	Chicago	IL	60610
Jean Blois	300 Day St.	Houston	TX	77091
Neal Smith	325 Clay St.	Portland	OR	97201

File #2: Product

Name	Address	Product	Sales Date
John Brown	1234 E. St.	Mower	1/3/2007
John Brown	1234 E. St.	Rake	4/16/2006
Neal Smith	325 Clay St.	Shovel	8/23/2005
Jean Blois	300 Day St.	Hoe	9/28/2007

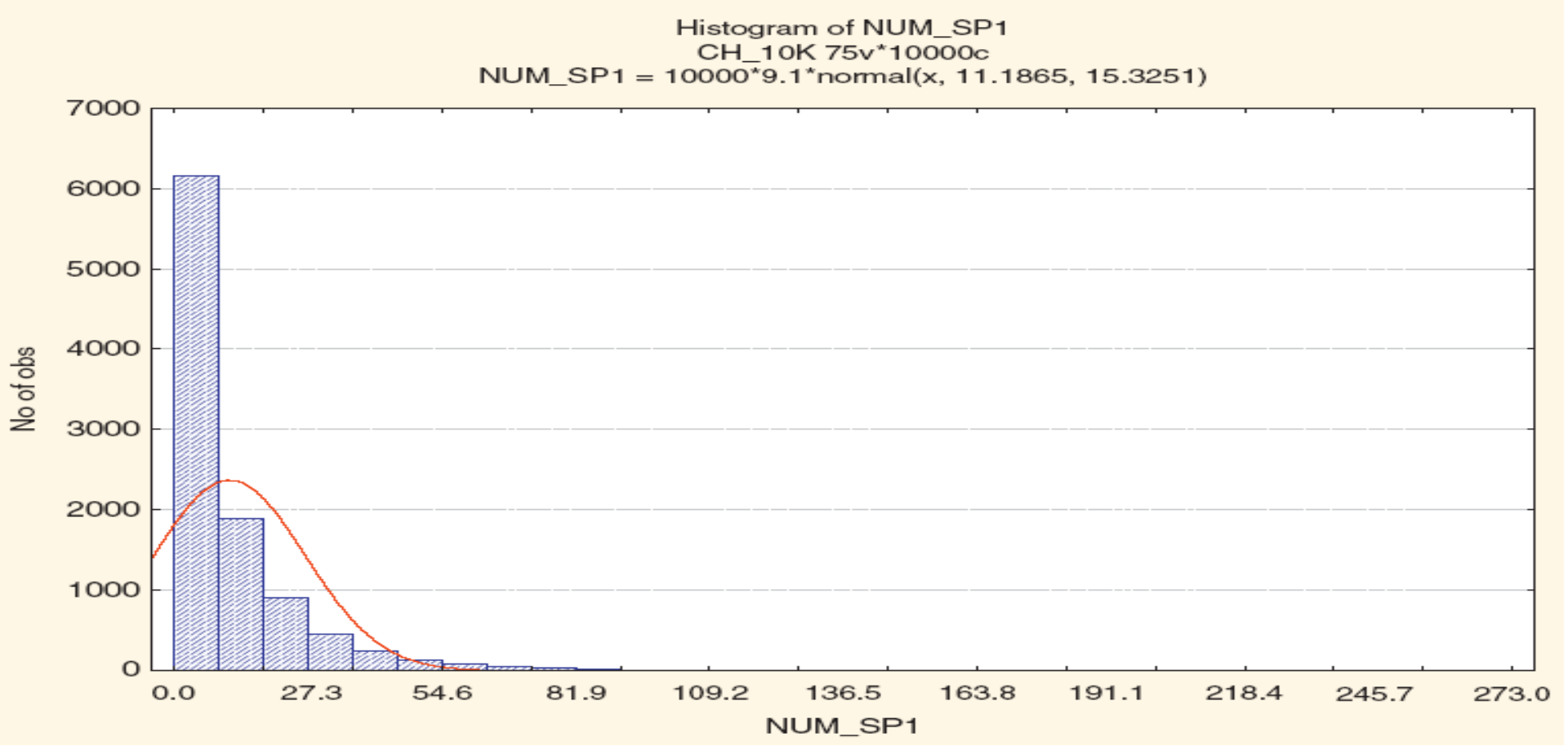
Name	Address	City	State	Zipcode	Product1	Product2
John Brown	1234 E. St.	Chicago	IL	60610	Mower	Rake
Neal Smith	325 Clay St.	Portland	OR	97201	Shovel	
Jean Blois	300 Day St.	Houston	TX	77091	Hoe	

Mô tả dữ liệu

- Giá trị kỳ vọng (mean)
  - Xu hướng trung tâm của tập dữ liệu
- Độ lệch chuẩn (Standard deviation)
  - Phân bố dữ liệu xung quanh kỳ vọng
- Cực tiểu (Minimum)
  - Giá trị nhỏ nhất
- Cực đại (Maximum)
  - Giá trị lớn nhất
- Bảng tần suất (Frequency tables)
  - Phân bố tần suất giá trị của các biến
- Lược đồ (Histograms)
  - Cung cấp kỹ thuật đồ họa biểu diễn tần số giá trị của một biến

Mô tả dữ liệu, so sánh với phân bố chuẩn (chủ yếu trong miền [0,10])

N	Mean	Min	Max	StDev
10000	9.769700	0.00	454.0000	15.10153



Đánh giá và lập hồ sơ dữ liệu

- Đánh giá dữ liệu
  - Định vị một vấn đề trong dữ liệu cần giải quyết: Tìm ra và quyết định cách nắm bắt vấn đề
  - Mô tả dữ liệu sẽ làm hiện rõ một số vấn đề
  - Kiểm toán dữ liệu: lập hồ sơ dữ liệu và phân tích ảnh hưởng của dữ liệu chất lượng kém.
- Lập hồ sơ dữ liệu (cơ sở căn cứ: phân bố dữ liệu)
  - Tâm của dữ liệu
  - Các ngoại lai tiềm năng bất kỳ
  - Số lượng và phân bố các khoảng trong trong mọi trường hợp
  - Bất cứ dữ liệu đáng ngờ, như mã thiếu (miscodes), dữ liệu học, dữ liệu test, hoặc chỉ đơn giản dữ liệu rác
  - Những phát hiện nên được trình bày dưới dạng các báo cáo và liệt kế như các mốc quan trọng của kế hoạch

Những vấn đề cơ bản để chuẩn bị dữ liệu

- Cách thức làm sạch dữ liệu:
  - Data Cleaning
- Cách thức diễn giải dữ liệu:
  - Data Transformation
- Cách thức nắm bắt giá trị thiếu:
  - Data Imputation
- Trọng số của các trường hợp:
  - Data Weighting and Balancing
- Xử lý dữ liệu ngoại lai và không mong muốn khác:
  - Data Filtering
- Cách thức nắm bắt dữ liệu thời gian/chuỗi thời gian:
  - Data Abstraction
- Cách thức rút gọn dữ liệu để dùng: Data Reduction
  - Bản ghi : Data Sampling
  - Biến: Dimensionality Reduction
  - Giá trị: Data Discretization
- Cách thức tạo biến mới: Data Derivation

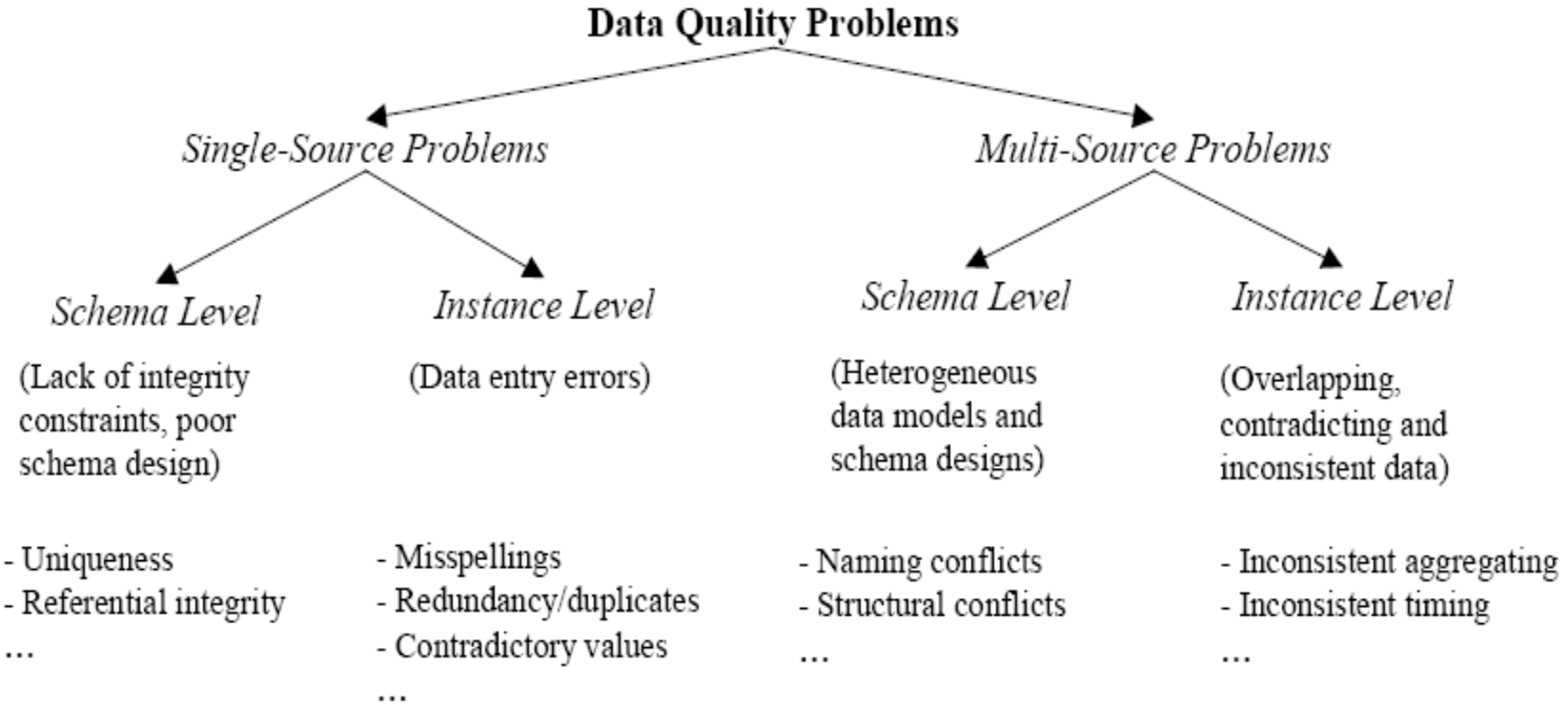
Chapter 3: Tiền xử lý dữ liệu

- Hiểu dữ liệu và chuẩn bị dữ liệu
- Vai trò của tiền xử lý dữ liệu
- Làm sạch dữ liệu
- Tích hợp và chuyển dạng dữ liệu
- Rút gọn dữ liệu
- Rời rạc và sinh kiến trúc khái niệm

Tính quan trọng của tiền xử lý

- Không có dữ liệu tốt, không thể có kết quả khai phá tốt!
  - Quyết định chất lượng phải dựa trên dữ liệu chất lượng
    - Chẳng hạn, dữ liệu bội hay thiếu là nguyên nhân thống không chính xác, thậm chí gây hiểu nhầm.
  - Kho dữ liệu cần tích hợp nhất quán của dữ liệu chất lượng
- Phân lớn công việc xây dựng một kho dữ liệu là trích chọn, làm sạch và chuyển đổi dữ liệu —Bill Inmon .
- Dữ liệu có chất lượng cao nếu như phù hợp với mục đích sử dụng trong điều hành, ra quyết định, và lập kế hoạch

Các vấn đề về chất lượng dữ liệu [RD00]



- (Thiếu lược đồ toàn vẹn, thiết kế sơ đồ sơ sài) đơn trị, toàn vẹn tham chiếu…
- (Lỗi nhập dữ liệu) sai chính tả, dư thừa/sao, giá trị mâu thuẫn…
- (Mô hình dữ liệu và thiết kế sơ đồ không đồng nhất) xung đột tên, cấu trúc
- (Dữ liệu chồng chéo, mâu thuẫn và không nhất quán) không nhất quán tích hợp và thời gian

[RD00] Erhard Rahm, Hong Hai Do (2000). Data Cleaning: Problems and Current Approaches, *IEEE Data Engineering Bulletin*, **23**(4): 3-13, 2000.

Độ đo đa chiều chất lượng dữ liệu

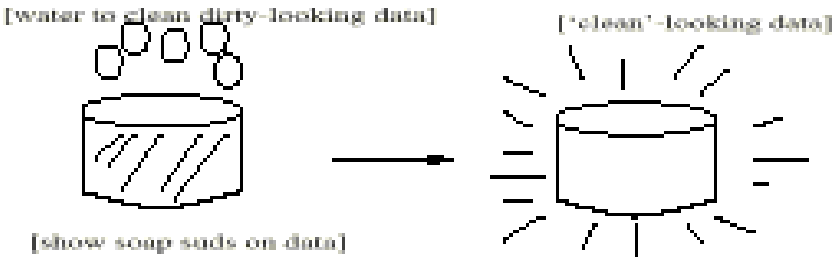
- Khung đa chiều cấp nhận tốt:
  - Độ chính xác (Accuracy)
  - Tính đầy đủ (Completeness)
  - Tính nhất quán (Consistency)
  - Tính kịp thời (Timeliness)
  - Độ tin cậy (Believability)
  - Giá trị gia tăng (Value added)
  - Biểu diễn được (Interpretability)
  - Tiếp cận được (Accessibility)
- Phân loại bề rộng (Broad categories):
  - Bản chất (intrinsic), ngữ cảnh (contextual), trình diễn (representational), và tiếp cận được (accessibility).

Các bài toán chính trong tiền XL DL

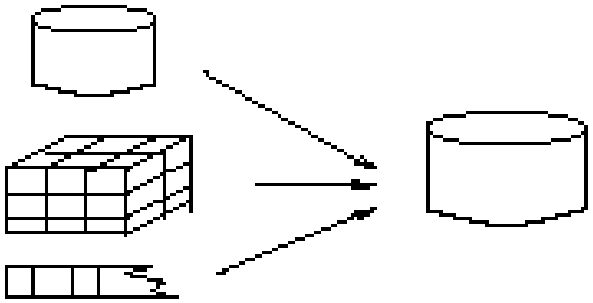
- Làm sạch dữ liệu
  - Điền giá trị thiếu, làm trơn dữ liệu nhiễu, định danh hoặc xóa ngoại lai, và khử tính không nhất quán
- Tích hợp dữ liệu
  - Tích hợp CSDL, khối dữ liệu hoặc tập tin phức
- Chuyển dạng dữ liệu
  - Chuẩn hóa và tổng hợp
- Rút gọn dữ liệu
  - Thu được trình bày thu gọn về kích thước những sản xuất cùng hoặc tương tự kết quả phân tích
- Rời rạc dữ liệu
  - Bộ phận của rút gọn dữ liệu nhưng có độ quan trọng riêng, đặc biệt với dữ liệu số

Các thành phần của tiền xử lý dữ liệu (Bảng 2.1)

Data Cleaning



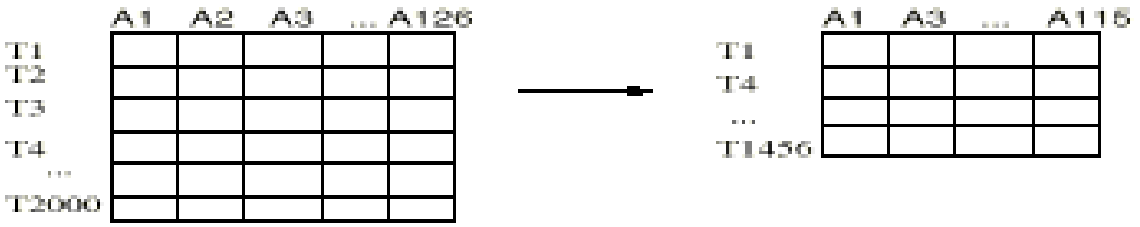
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Chapter 3: Tiền xử lý dữ liệu

- Hiểu dữ liệu và chuẩn bị dữ liệu
- Vai trò của tiền xử lý dữ liệu
- Làm sạch dữ liệu
- Tích hợp và chuyển dạng dữ liệu
- Rút gọn dữ liệu
- Rời rạc và sinh kiến trúc khái niệm

Làm sạch dữ liệu

- Là quá trình
  - xác định tính không chính xác, không đầy đủ/tính bất hợp lý của dữ liệu
  - chỉnh sửa các sai sót và thiếu sót được phát hiện
  - nâng cao chất lượng dữ liệu.
- Quá trình bao gồm
  - kiểm tra định dạng, tính đầy đủ, tính hợp lý, miền giới hạn,
  - xem xét dữ liệu để xác định ngoại lai (địa lý, thống kê, thời gian hay môi trường) hoặc các lỗi khác,
  - đánh giá dữ liệu của các chuyên gia miền chủ đề.
- Quá trình thường dẫn đến
  - loại bỏ, lập tài liệu và kiểm tra liên tiếp và hiệu chỉnh đúng bản ghi nghi ngờ.
  - Kiểm tra xác nhận có thể được tiến hành nhằm đạt tính phù hợp với các chuẩn áp dụng, các quy luật, và quy tắc.

Nguồn dữ liệu đơn: mức sơ đồ (Ví dụ)

Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Illegal values	bdate=30.13.70	values outside of domain range
Record	Violated attribute dependencies	age=22, bdate=12.02.70	age = (current date – birth date) should hold
Record type	Uniqueness violation	emp_1=(name="John Smith", SSN="123456") emp_2=(name="Peter Miller", SSN="123456")	uniqueness for SSN (social security number) violated
Source	Referential integrity violation	emp=(name="John Smith", deptno=127)	referenced department (127) not defined

Nguồn dữ liệu đơn: mức thể hiện (Ví dụ)

Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Missing values	phone=9999-999999	unavailable values during data entry (dummy values or null)
	Misspellings	city="Liipzig"	usually typos, phonetic errors
	Cryptic values, Abbreviations	experience="B"; occupation="DB Prog."	
	Embedded values	name="J. Smith 12.02.70 New York"	multiple values entered in one attribute (e.g. in a free-form field)
	Misfielded values	city="Germany"	
Record	Violated attribute dependencies	city="Redmond", zip=77777	city and zip code should correspond
Record type	Word transpositions	$name_1$= "J. Smith", $name_2$="Miller P."	usually in a free-form field
	Duplicated records	emp_1=(name="John Smith",...); emp_2=(name="J. Smith",....)	same employee represented twice due to some data entry errors
	Contradicting records	emp_1=(name="John Smith", bdate=12.02.70); emp_2=(name="John Smith", bdate=12.12.70)	the same real world entity is described by different values
Source	Wrong references	emp=(name="John Smith", deptno=17)	referenced department (17) is defined but wrong

Nguồn dữ liệu phức: mức sơ đồ và thể hiện (Ví dụ)

Customer (source 1)

CID	*Name*	*Street*	*City*	*Sex*
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Client (source 2)

Cno	*LastName*	*FirstName*	*Gender*	*Address*	*Phone/Fax*
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

Customers (integrated target with cleaned data)

No	*LName*	*FName*	*Gender*	*Street*	*City*	*State*	*ZIP*	*Phone*	*Fax*	*CID*	*Cno*
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	Smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

Làm sạch dữ liệu

- Nguyên lý chất lượng dữ liệu cần được áp dụng ở mọi giai đoạn quá trình quản lý dữ liệu (nắm giữ, số hóa, lưu trữ, phân tích, trình bày và sử dụng).
  - hai vấn đề cốt lõi để cải thiện chất lượng - phòng ngừa và chỉnh sửa
  - Phòng ngừa liên quan chặt chẽ với thu thập và nhập dữ liệu vào CSDL.
  - Tăng cường phòng ngừa lỗi, vẫn/tồn tại sai sót trong bộ dữ liệu lớn (Maletic và Marcus 2000) và không thể bỏ qua việc xác nhận và sửa chữa dữ liệu
- Vai trò quan trọng
  - "là một trong ba bài toán lớn nhất của kho dữ liệu"—Ralph Kimball
  - "là bài toán "number one" trong kho dữ liệu"—DCI khảo sát
- Các bài toán thuộc làm sạch dữ liệu
  - Xử lý giá trị thiếu
  - Dữ liệu nhiễu: định danh ngoại lai và làm trơn.
  - Chỉnh sửa dữ liệu không nhất quán
  - Giải quyết tính dư thừa tạo ra sau tích hợp dữ liệu.

Xử lý thiếu giá trị

- Bỏ qua bản ghi có giá trị thiếu:
  - Thường làm khi thiếu nhãn phân lớp (giả sử bài toán phân lớp)
  - không hiểu quả khi tỷ lệ số giá trị thiếu lớn (bán giám sát)
- Điền giá trị thiếu bằng tay:
  - tẻ nhạt
  - tính khả thi
- Điền giá trị tự động:
  - Hằng toàn cục: chẳng hạn như“chưa biết”, có phải một lớp mới
  - Trung bình giá trị thuộc tính các bản ghi hiện có
  - Trung bình giá trị thuộc tính các bản ghi cùng lớp: tinh hơn
  - Giá trị khả năng nhất: dựa trên suy luận như công thức Bayes hoặc cây quyết định

Dữ liệu nhiễu

- Nhiễu:
  - Lỗi ngẫu nhiên
  - Biến dạng của một biến đo được
- Giá trị không chính xác do
  - Lỗi do thiết bị thu thập dữ liệu
  - Vấn đề nhập dữ liệu: người dùng hoặc máy có thể sai
  - Vấn đề truyền dữ liệu: sai từ thiết bị gửi/nhận/truyền
  - Hạn chế của công nghệ: ví dụ, phần mềm có thể xử lý không đúng
  - Thiết nhất quán khi đặt tên: cũng một tên song cách viết khác nhau
- Các vấn đề dữ liệu khác yêu cầu làm sạch dữ liệu
  - Bộ bản ghi
  - Dữ liệu không đầy đủ
  - Dữ liệu không nhất quán

Nắm bắt dữ liệu nhiễu

- Phương pháp đóng thùng (Binning):
  - Sắp dữ liệu tăng và chia "đều" vào các thùng
  - Làm trơn: theo trung bình, theo trung tuyến, theo biên…
- Phân cụm (Clustering)
  - Phát hiện và loại bỏ ngoại lai (outliers)
- Kết hợp kiểm tra máy tính và con người
  - Phát hiện giá trị nghi ngờ để con người kiểm tra (chẳng hạn, đối phó với ngoại lai có thể)
- Hồi quy
  - Làm trơn: ghép dữ liệu theo các hàm hồi quy

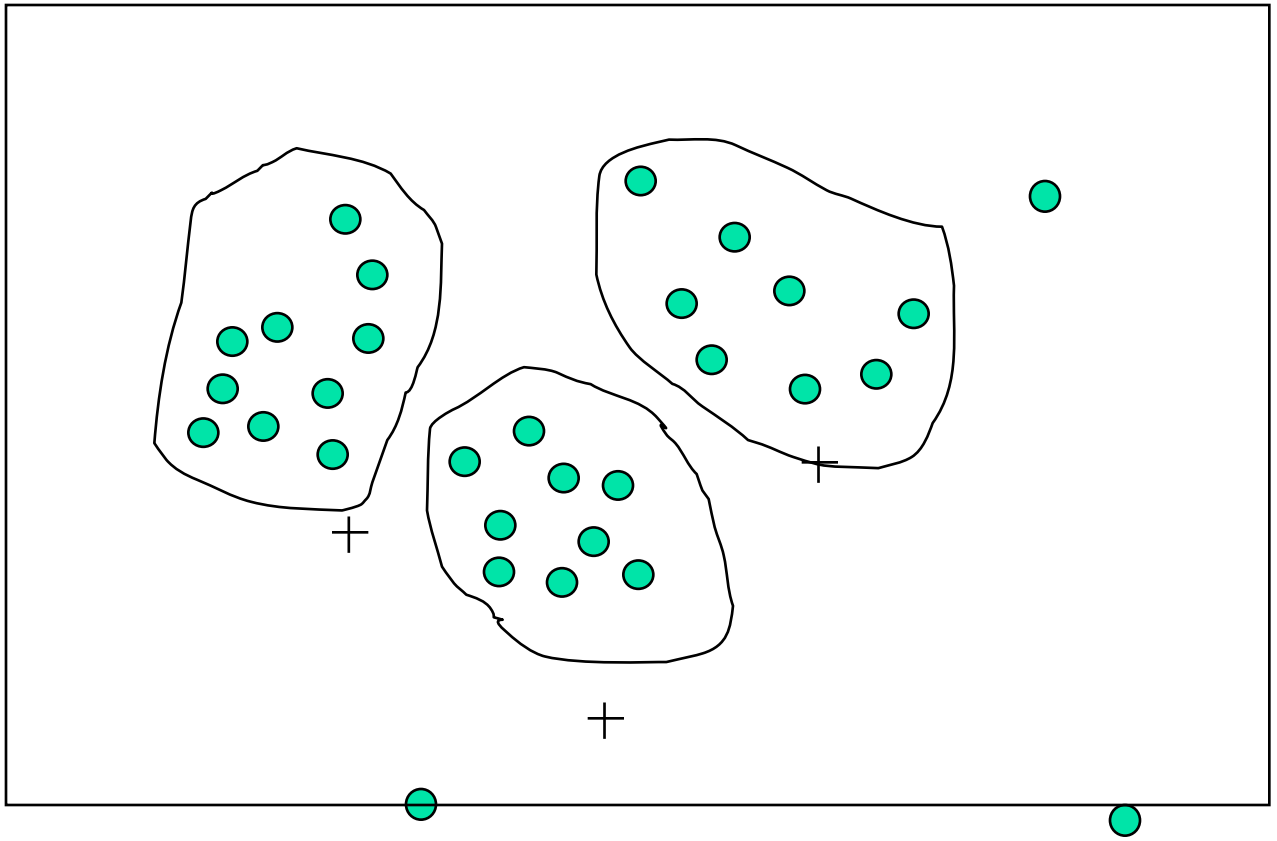
Phương pháp rời rạc hóa đơn giản: Xếp thùng (Binning)

- Phân hoạch cân bằng bề rộng Equal-width (distance) partitioning:
  - Chia miền giá trị: *N* đoạn dài như nhau: uniform grid
  - Miền giá trị từ A (nhỏ nhất) tới *B* (lớn nhất) -> $W = (B - A)/N$.
  - Đơn giản nhất song bị định hướng theo ngoại lai.
  - Không xử lý tốt khi dữ liệu không cân bằng (đều).
- Phân hoạch cân bằng theo chiều sâu Equal-depth (frequency) partitioning:
  - Chia miền xác định thành N đoạn "đều nhau về số lượng", các đoạn có xấp xỉ số ví dụ mẫu.
  - Khả cỡ dữ liệu: tốt.
  - Việc quản lý các thuộc tính lớp: có thể "khôn khéo".

Phương pháp xếp thùng làm trơn dữ liệu (Data Smoothing)

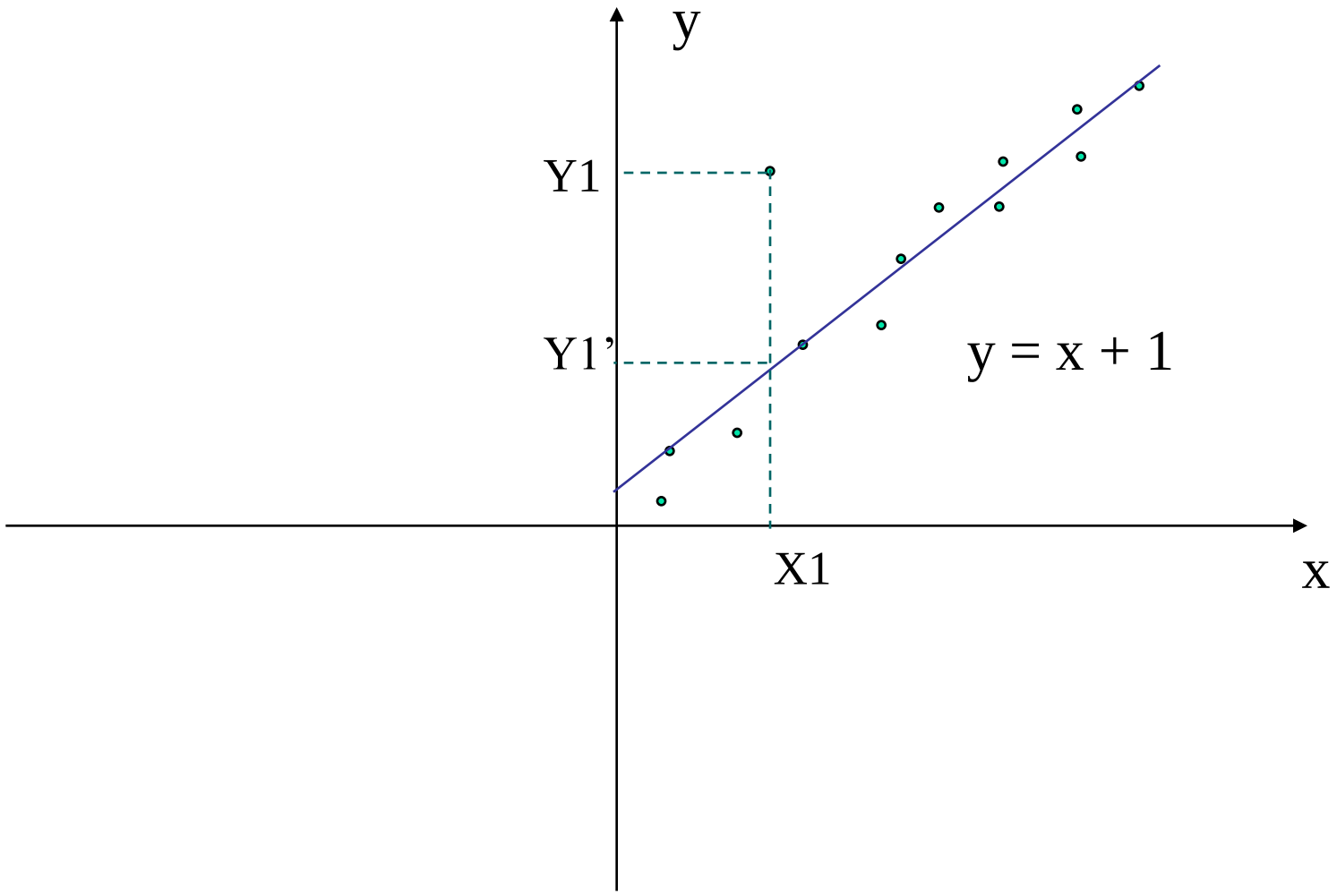
* Dữ liệu được xếp theo giá: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Chia thùng theo chiều sâu:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Làm trơn thùng theo trung bình:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Làm trơn thùng theo biên:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Phân tích cụm (Cluster Analysis)



Cụm: Các phần tử trong cụm là "tương tự nhau"
Làm trơn phần tử trong cụm theo đại diện.
Thuật toán phân cụm: Chương 6.

Hồi quy (Regression)



Chapter 3: Tiền xử lý dữ liệu

- Hiểu dữ liệu và chuẩn bị dữ liệu
- Vai trò của tiền xử lý dữ liệu
- Làm sạch dữ liệu
- Tích hợp và chuyển dạng dữ liệu
- Rút gọn dữ liệu
- Rời rạc và sinh kiến trúc khái niệm

Tích hợp dữ liệu

- Tích hợp dữ liệu (Data integration):
  - Kết hợp dữ liệu từ nhiều nguồn thành một nguồn lưu trữ chung
- Tích hợp sơ đồ
  - Tích hợp sieu dữ liệu từ các nguồn khác nhau
  - Vấn đề định danh thực thế: xác định thực thể thực tế từ nguồn dữ liệu phức, chẳng hạn, A.cust-id [redacted]B.cust-#
- Phát hiện và giải quyết vấn đề thiết nhất quá dữ liệu
  - Cùng một thực thể thực sự: giá trị thuộc tính các nguồn khác nhau là khác nhau
  - Nguyên nhân: trình bày khác nhau, cỡ khác nhau, chẳng hạn, đơn vị quốc tế khác với Anh quốc

Nắm bắt dư thừa trong tích hợp dữ liệu (Handling Redundancy in Data Integration)

- Dư thừa dữ liệu: thường có khi tích hợp từ nhiều nguồn khác nhau
  - Một thuộc tính có nhiều tên khác nhau ở các CSDL khác nhau
  - Một thuộc tính: thuộc tính "nguồn gốc" trong CSDL khác, chẳng hạn, doanh thu hàng năm
- Dữ liệu dư thừa có thể được phát hiện khi phân tích tương quan
- Tích hợp cẩn trọng dữ liệu nguồn phức giúp giảm/tránh dư thừa, thiếu nhất quán và tăng hiệu quả tốc độ và chất lượng

Chuyển dạng dữ liệu

- Làm trơn (Smoothing): loại bỏ nhiễu từ dữ liệu
- Tổng hợp (Aggregation): tóm tắt, xây dựng khối dữ liệu
- Tổng quát hóa (Generalization): leo kiến trúc khái niệm
- Chuẩn hóa (Normalization): thu nhỏ vào miền nhỏ, riêng
  - Chuẩn hóa min-max
  - Chuẩn hóa z-score
  - Chuẩn hóa tỷ lệ thập phân
- Xây dựng thuộc tính/đặc trưng
  - Thuộc tính mới được xây dựng từ các thuộc tính đã có

Chuyển đổi dữ liệu: Chuẩn hóa

- Chuẩn hóa min-max

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- Chuẩn hóa z-score

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- Chuẩn hóa tỷ lệ thập phân

$$v' = \frac{v}{10^j}$$

j : số nguyên nhỏ nhất mà Max(| |)<1 v'

Chapter 3: Tiền xử lý dữ liệu

- Hiểu dữ liệu và chuẩn bị dữ liệu
- Vai trò của tiền xử lý dữ liệu
- Làm sạch dữ liệu
- Tích hợp và chuyển dạng dữ liệu
- Rút gọn dữ liệu
- Rời rạc và sinh kiến trúc khái niệm

Chiến lược rút gọn dữ liệu (Data Reduction Strategies)

- Kho dữ liệu chứa tới hàng TB
  - Phân tích/khai phá dữ liệu phức mất thời gian rất dài khi chạy trên tập toàn bộ dữ liệu
- Rút gọn dữ liệu
  - Có được trình bày gọn của tập dữ liệu mà nhỏ hơn nhiều về khối lượng mà sinh ra cùng (hoặc hầu như cùng) kết quả.
- Chiến lược rút gọn dữ liệu
  - Tập hợp khối dữ liệu
  - Giảm đa chiều – loại bỏ thuộc tính không quan trọng
  - Nén dữ liệu
  - Giảm tính số hóa – dữ liệu thành mô hình
  - Rời rạc hóa và sinh cây khái niệm

Kết hợp khối dữ liệu (Data Cube Aggregation)

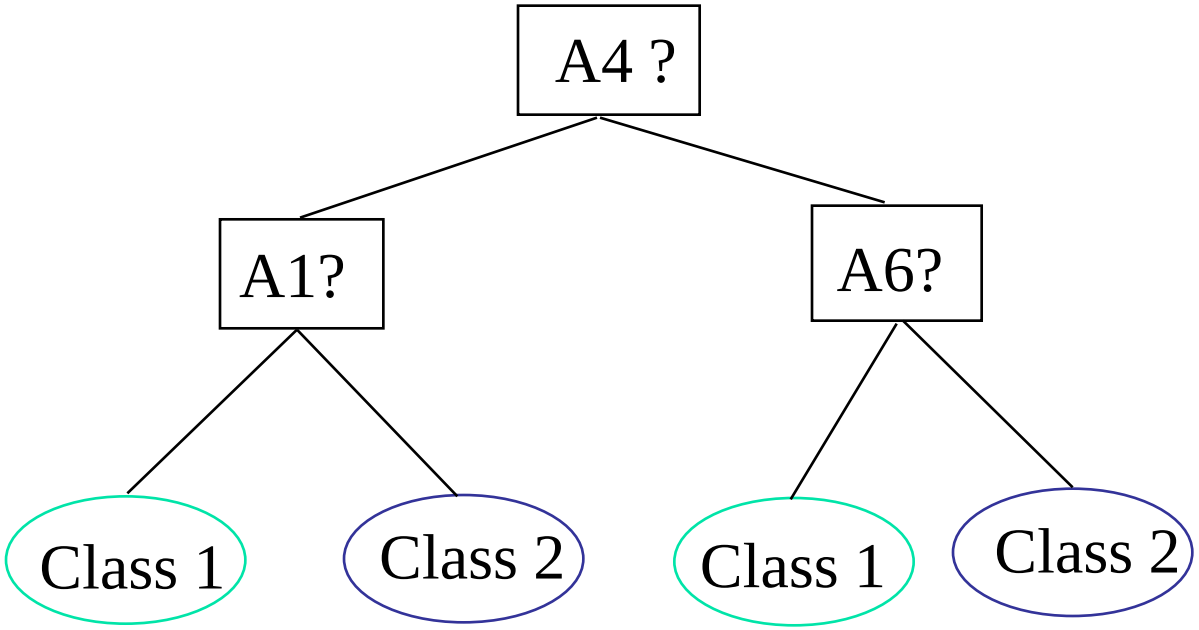
- Mức thấp nhất của khối dữ liệu
  - Tổng hợp dữ liệu thành một cá thể quan tâm
  - Chẳng hạn, một khách hàng trong kho dữ liệu cuộc gọi điện thoại.
- Các mức phức hợp của tích hợp thành khối dữ liệu
  - Giảm thêm kích thước dữ liệu
- Tham khảo mức thích hợp
  - Sử dụng trình diễn nhỏ nhất đủ để giải bài toán
- Nên sử dụng dữ liệu khối lập phương khi trả lời câu hỏi tổng hợp thông tin

Rút gọn chiều

- Rút gọn đặc trưng (như., lựa chọn tập con thuộc tính):
  - Lựa chọn tập nhỏ nhất các đặc trưng mà phân bố xác suất của các lớp khác nhau cho giá trị khi cho giá trị của các lớp này gần như phân bổ vốn có đã cho giá trị của các đặc trưng
  - Rút gọn # của các mẫu trong tập mẫu dễ dàng hơn để hiểu dữ liệu
- Phương pháp Heuristic (có lực lượng mũ # phép chọn):
  - Khôn ngoan chọn chuyển tiếp từ phía trước
  - Kết hợp chon chuyển tiếp và loại bỏ lạc hậu.
  - Rút gọn câu qyuyết định

Ví dụ rút gọn cây quyết định

Tập thuộc tính khởi tạo:
{A1, A2, A3, A4, A5, A6}

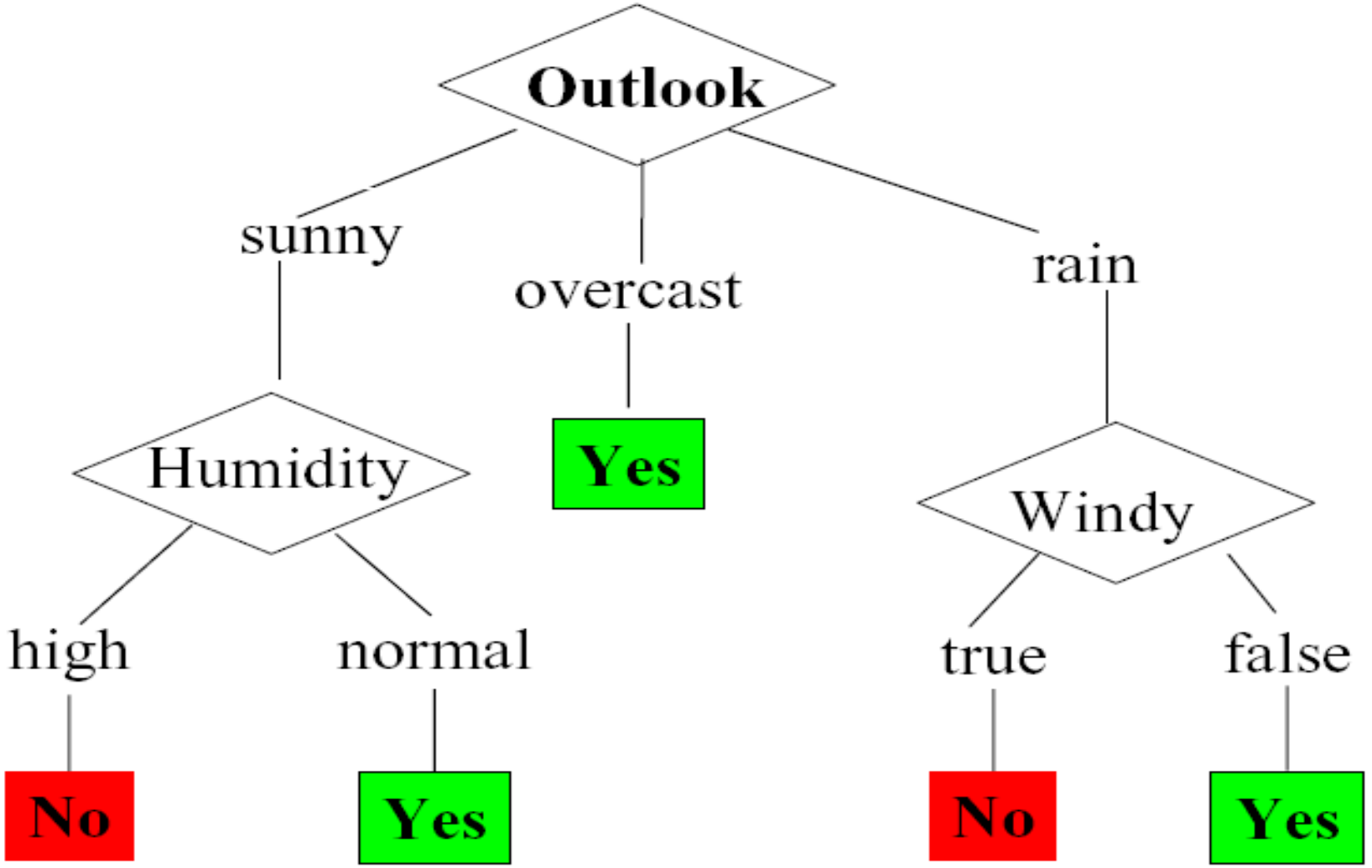


-----> Tập thuộc tinh rút gọn: {A1, A4, A6}

Phân lớp cây quyết định

- Đồ thị dạng cây
- Đỉnh trong là một hàm test
- Các nhánh tương ứng với kết quả kiểm tra tại đỉnh trong
- Các lá là các nhãn, hoặc các lớp.
- Xem Chương 5

Phân lớp cây quyết định



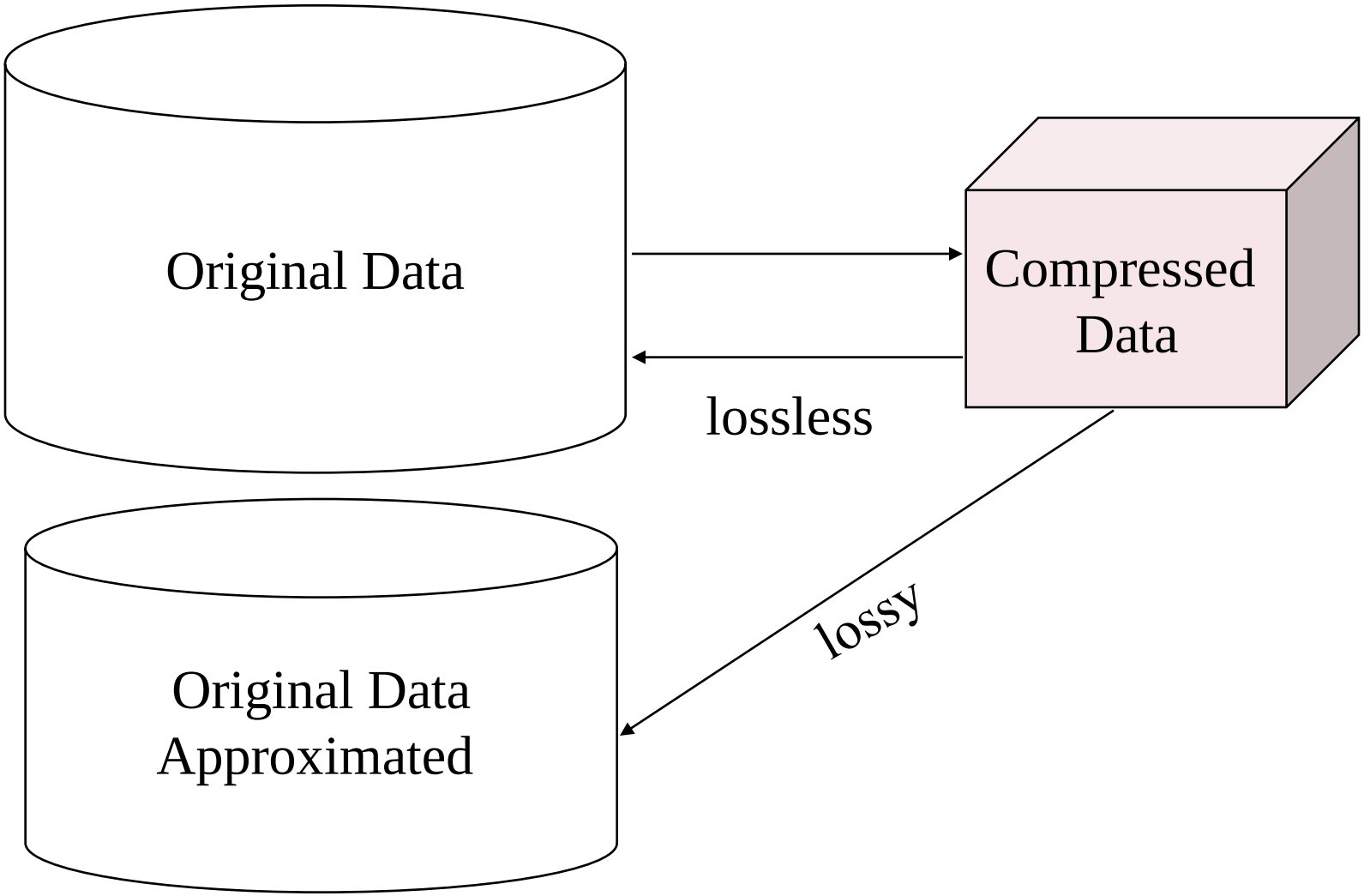
Phân lớp cây quyết định

- Xây dựng cây quyết định:
  - Xây dựng cây quyết định
    - Phương pháp top-down
  - Cắt tỉa cây (pruning)
    - Phương pháp bottom-up: xác định và loại bỏ những nhánh rườm rà tăng độ chính xác khi phân lớp những đối tượng mới
- Sử dụng cây quyết định: phân lớp các đối tượng chưa được gán nhãn

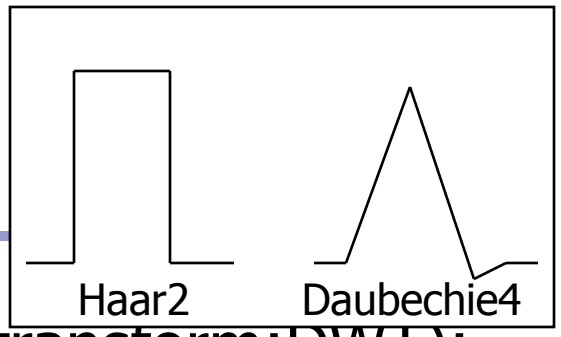
Nén dữ liệu (Data Compression)

- Nén xâu văn bản
  - Tồn tại lý thuyết phong phú và thuật toán điển hình
  - *Không tổn thất điển hình*
  - Nhưng chỉ các thao tác hạn hẹp mà không mở rộng
- Nén Audio/video
  - Nén *tổn thất điển hình*, với tinh lọc cải tiến
  - Vài trường hợp mảnh tín hiệu nhỏ được tái hợp không cần dựng toàn bộ
- Dãy thời gian mà không là audio
  - Ngắn điển hình và thây đổi chậm theo thời gian

Nén dữ liệu (Data Compression)

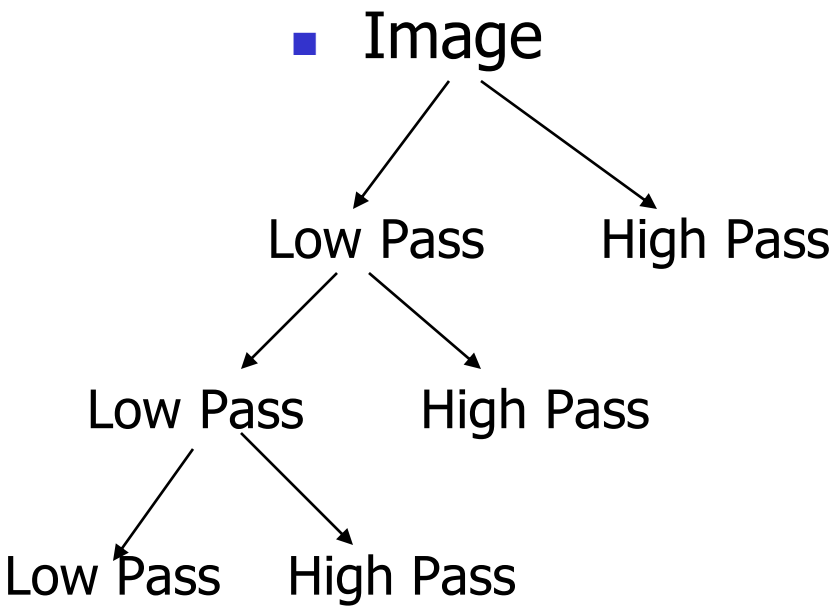


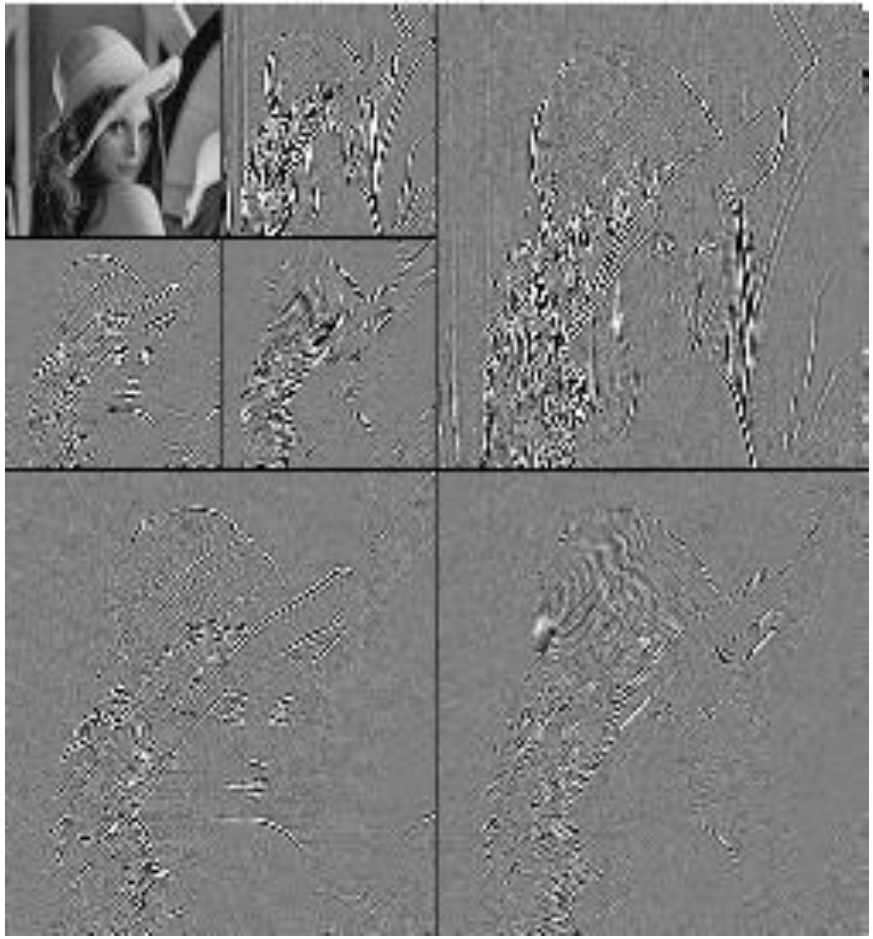
Chuyển dạng sóng (Wavelet Transformation)



- Biến dạng sóng rời rạc (Discrete wavelet transform:DWT): XL tín hiệu tuyến tính, phân tích đa giải pháp
- Xấp xỉ nén: chỉ lưu một mảnh nhỏ các hệ số sóng lớn nhất
- Tương tự như biến đổi rời rạc Fourier (DFT), nhưng nén tổn thất tốt hơn, bản địa hóa trong không gian
- Phương pháp:
  - Độ dài, L, buộc là số nguyên lũy thừa 2 (đệm thêm các chữ số 0, khi cần)
  - Mỗi phép biến đổi có 2 chức năng: làm mịn, tách biệt
  - Áp dụng cho các cặp DL, kết quả theo 2 tập DL độ dài L/2
  - Áp dụng đệ quy hai chức năng đến độ dài mong muốn

DWT cho nén ảnh

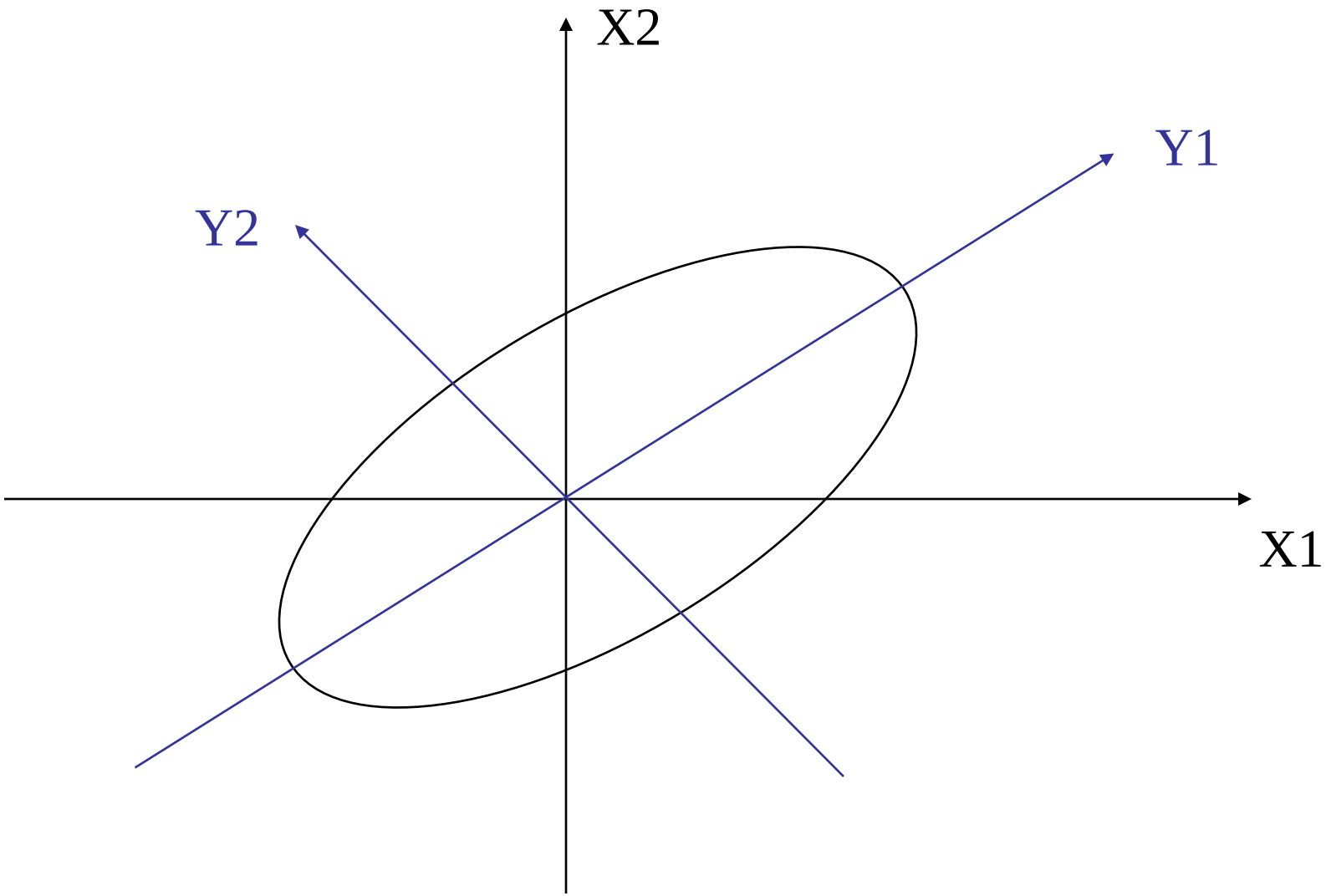




Phân tích thành phần chính (Principal Component Analysis)

- Cho *N* vector dữ liệu *k*-chiều, tìm *c (<= k)* vector trực giao tốt nhất để trình diễn dữ liệu.
  - Tập dữ liệu gốc được rút gọn thành N vector dữ liệu c chiều: *c thành phần chính* (chiều được rút gọn).
- Mỗi vector dữ liệu là tổ hợp tuyến tính của các vector thành phần chính.
- *Chỉ áp dụng cho dữ liệu số*.
- Dùng khi số chiều vector lớn.

Phân tích thành phần chính (PCA)



Rút gọn kích thước số

- Phương pháp tham số
  - *Giả sử dữ liệu phù hợp với mô hình nào đó*, ước lượng tham số mô hình, lưu chỉ các tham số, và không lưu dữ liệu (ngoại trừ các ngoại lai có thể có)
  - Mô hình tuyến tính loga (Log-linear models): lấy giá trị tại một điểm trong không gian M-chiều như là tích của các không gian con thích hợp
- Phương pháp không tham số
  - Không giả thiết mô hình
  - Tập hợp chính: biểu đồ (histograms), phân cụm (clustering), lấy mẫu (sampling)

Hồi quy và mô hình logarit tuyến tính

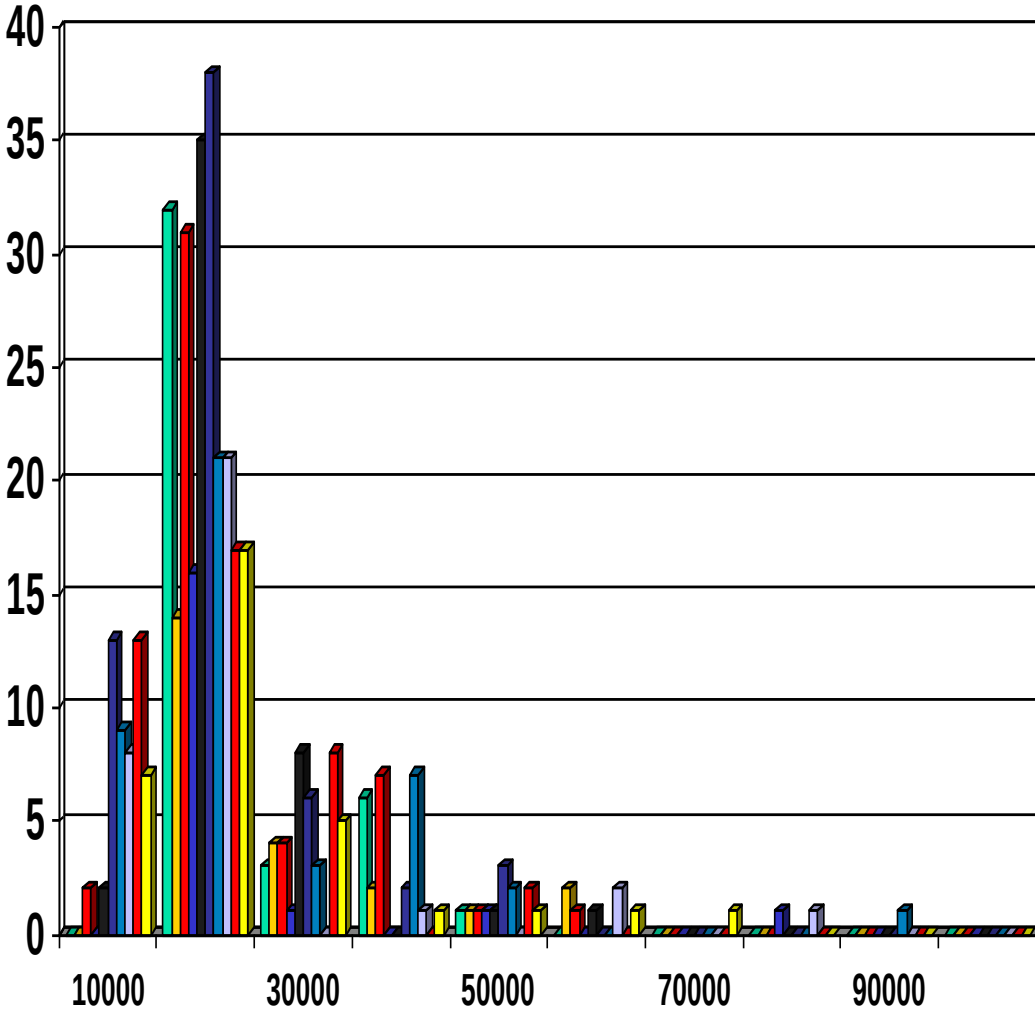
- Hồ quy tuyến tính: DL được mô hình hóa phù hợp với 1 đường thẳng
  - Thường dùng phương pháp bình phương tối thiểu để khớp với đường
- Hồ quy đa chiều: Cho một biến đích Y được mô hình hóa như ột hàm tuyến tính của vector đặc trưng đa chiều
- Mô hình tuyến tính loga: rời rạc hóa xấp xỉ các phân bố xác suất đa chiều

Phân tích hồi quy và mô hình logarit tuyến tính

- Hồi quy tuyến tính: $Y = \alpha + \beta X$
  - Hai tham số, α và β đặc trưng cho đường và được xấp xỉ qua dữ liệu đã nắm bắt được.
  - Sử dụng chiến lược BP tối thiểu tới các giá trị đã biết $Y_1, Y_2, \ldots, X_1, X_2, \ldots$.
- Hồ quy đa chiều: $Y = b0 + b1\ X1 + b2\ X2.$
  - Nhiều hàm không tuyến tính được chuyển dạng như trên.
- Mô hình tuyến tính loga:
  - Bảng đa chiều của xác suất tích nối được xấp xỉ bởi tích của các bảng bậc thấp hơn
  - Xác suất: $p(a, b, c, d) = \alpha_{ab}\beta_{ac}\delta_{ad}\gamma_{bcd}$

Lược đồ (Histograms)

- Kỹ thuật rút gọn dữ liệu phổ biến
- Phân dữ liệu vào các thùng và giữ trunh bình (tổng) của mỗi thùng
- Có thể được dựng tối ưu hóa theo 1 chiều khi dùng quy hoạch động
- Có quan hệ tới bài toán lượng tử hóa.



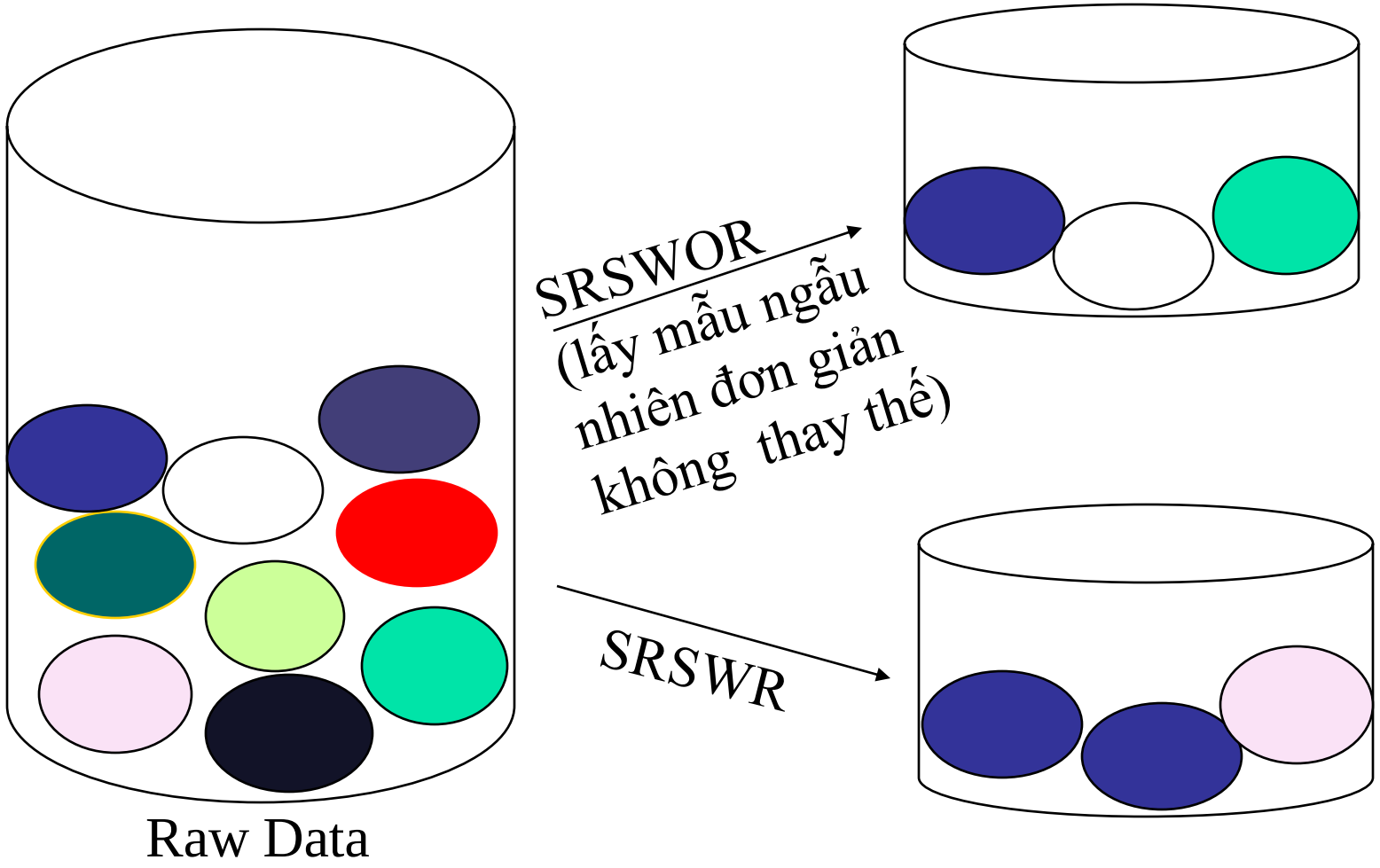
Phân cụm

- Phân tập DL thành các cụm, và chỉ cần lưu trữ đại diện của cụm
- Có thể rất hiệu quả nếu DL là được phân cụm mà không chứa dữ liệu "bẩn"
- Có thể phân cụm phân cấp và được lưu trữ trong cấu trúc cây chỉ số đa chiều
- Tồn tài nhiều lựa chọn cho xác định phân cụm và thuật toán phân cụm

Rút gọn mẫu (Sampling)

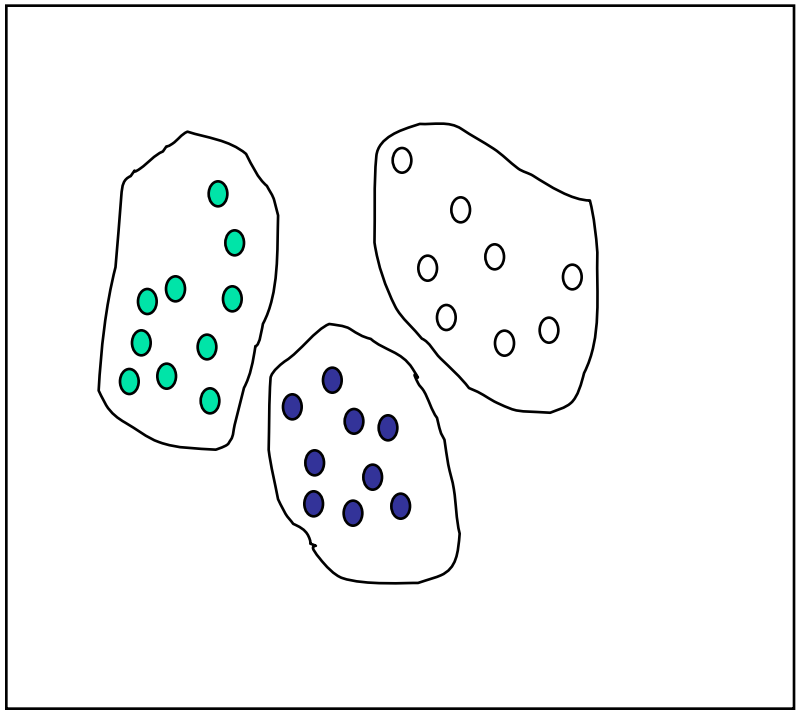
- Cho phép một thuật toán khai phá chạy theo độ phức tạp tựa tuyến tính theo cỡ của DL
- Lựa chọn một tập con trình diễn dữ liệu
  - Lấy mẫu ngẫu nhiên đơn giản có hiệu quả rất tồi nếu có DL lệch
- Phát triển các phương pháp lấy mẫu thích nghi
  - Lấy mẫu phân tầng:
    - Xấp xỉ theo phần trăm của mỗi lớp (hoặc bộ phận nhận diện được theo quan tâm) trong CSDL tổng thể
    - Sử dụng kết hợp với dữ liệu lệch
- Lẫy mẫu có thể không rút gọn được CSDL.

Rút gọn mẫu (Sampling)

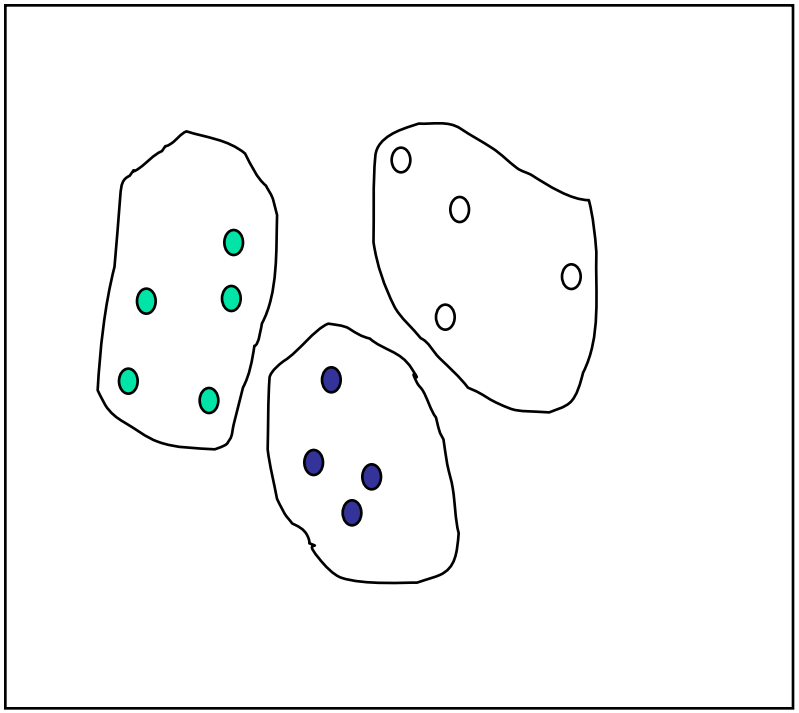


Rút gọn mẫu (Sampling)

Raw Data



Mẫu cụm/phân tầng



Rút gọn phân cấp

- Dùng cấu trúc đa phân giải với các mức độ khác nhau của rút gọn
- Phân cụm phân cấp thường được thi hành song có khuynh hướng xác định phân vùng DL hớn là "phân cụm"
- Phương pháp tham số thường không tuân theo trình bày phân cấp
- Tích hợp phân cấp
  - Một cấy chỉ số được chia phân cấp một tập DL thành các vùng bởi miền giá trị của một vài thuộc tính
  - Mỗi vùng được coi như một thùng
  - Như vậy, cây chỉ số với tích hợp lưu trữ mỗi nút là một sơ đồ phân cấp

Chapter 3: Tiền xử lý dữ liệu

- Hiểu dữ liệu và chuẩn bị dữ liệu
- Vai trò của tiền xử lý dữ liệu
- Làm sạch dữ liệu
- Tích hợp và chuyển dạng dữ liệu
- Rút gọn dữ liệu
- Rời rạc và sinh kiến trúc khái niệm

Rời rạc hóa

- Ba kiểu thuộc tính:
  - Định danh — giá trị từ một tập không có thứ tự
  - Thứ tự — giá trị từ một tập được sắp
  - Liên tục — số thực
- Rời rạc hóa:
  - Chia miền thuộc tính liên tục thành các đoạn
  - Một vài thuật toán phân lớp chỉ chấp nhận thuộc tính phân loại.
  - Rút gọn cỡ DL bằng rời rạc hóa
  - Chuẩn bị cho phân tích tiếp theo

Rời rạc hóa và kiến trúc khái niệm

- Rời rạc hóa
  - Rút gọn số lượng giá trị của thuộc tính liên tục bằng cách chia miền giá trị của thuộc tính thành các đoạn. Nhãn đoạn sau đó được dùng để thay thế giá trị thực.

- Phân cấp khái niệm
  - Rút gọn DL bằng tập hợp và thay thế các khái niệm mức thấp (như giá trị số của thuộc tính tuổi) bằng khái niệm ở mức cao hơn (như trẻ, trung niên, hoặc già)

Rời rạc hóa và kiến trúc khái niệm với dữ liệu số

- Phân thùng (xem làm trơn khử nhiễu)
- Phân tích sơ đồ (đã giới thiệu)
- Phân tích cụm (đã giới thiệu)
- Rời rạc hóa dựa theo Entropy
- Phân đoạn bằng phân chia tự nhiên

Rời rạc hóa dựa trên Entropy

- Cho tập ví dụ S, nếu S được chia thành 2 đoạn S1 và S2 dùng biên T, thì entropy sau khi phân đoạn là

$$E(S,T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

- Biên làm cực tiểu hàm entropy trên tất cả các biên được chọn như một rời rạc hóa nhị phân.
- Quá trình đệ quy tới các vùng cho tới khi đạt điều kiện dừng nào đó, như

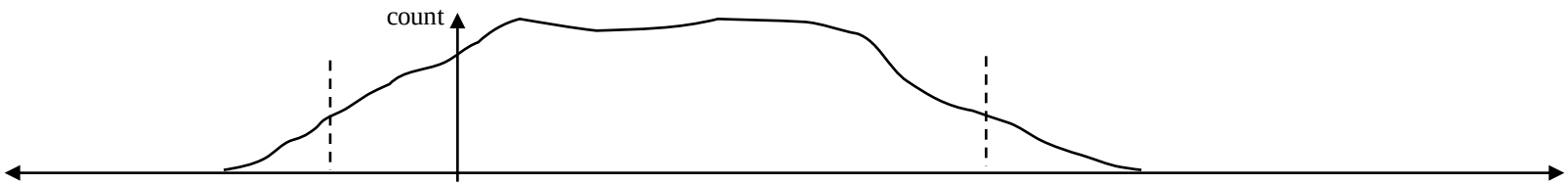
$$Ent(S) - E(T,S) > \delta$$

- Thực nghiệm chỉ ra rằng cho phép rút gọn cỡ DL và tăng độ chính xác phân lớp

Phân đoạn bằng phân hoạch tự nhiên

- Quy tắc đơn giản 3-4-5 được dùng để phân đoạn dữ liệu số thành các đoạn tương đối thống nhất, "tự nhiên".
  - Hướng tới số giá trị khác biệt ở vùng quan trọng nhất
  - Nếu 3, 6, 7 hoặc 9 giá trị khác biệt thì chia miền thành 3 đoạn tương đương.
  - Nếu phủ 2, 4, hoặc 8 giá trị phân biệt thì chia thành 4.
  - Nếu phủ 1, 5, hoặc 10 giá trị phân biệt thì chia thành 5.

Ví dụ luật 3-4-5

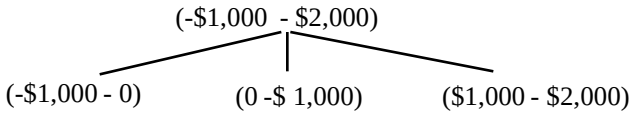


Step 1: -$351 -$159 profit $1,838 $4,700

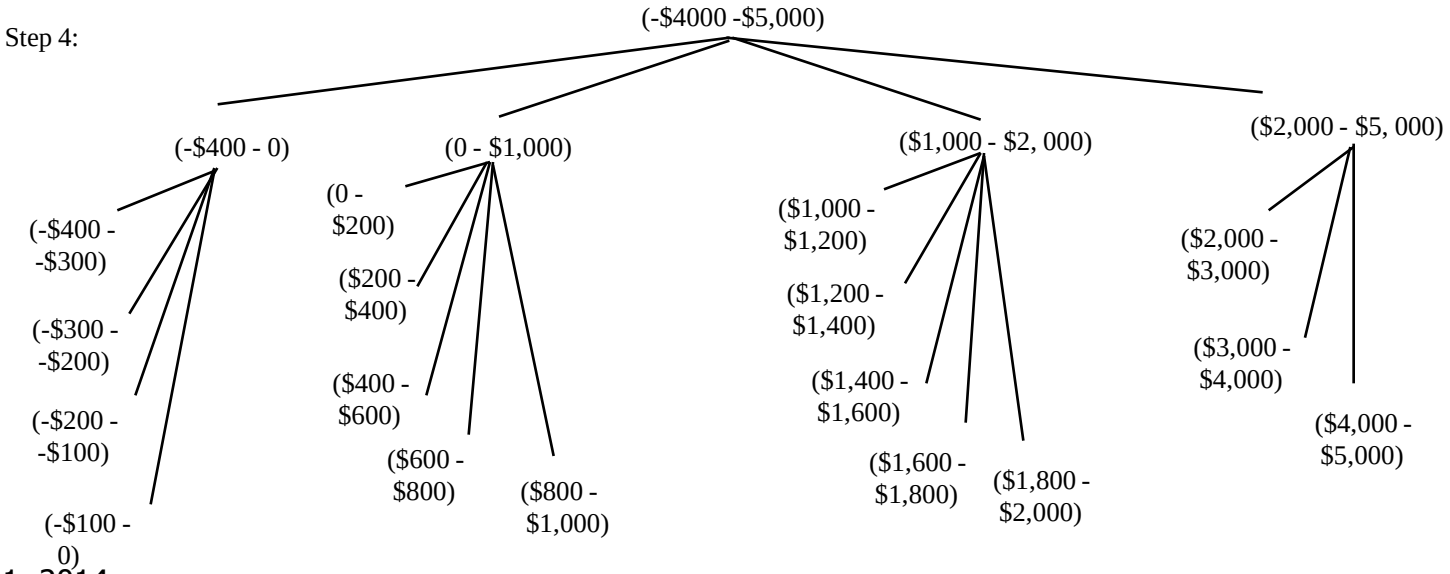
Min Low (i.e, 5%-tile) High(i.e, 95%-0 tile) Max

Step 2: msd=1,000 Low=-$1,000 High=$2,000

Step 3:



Step 4:

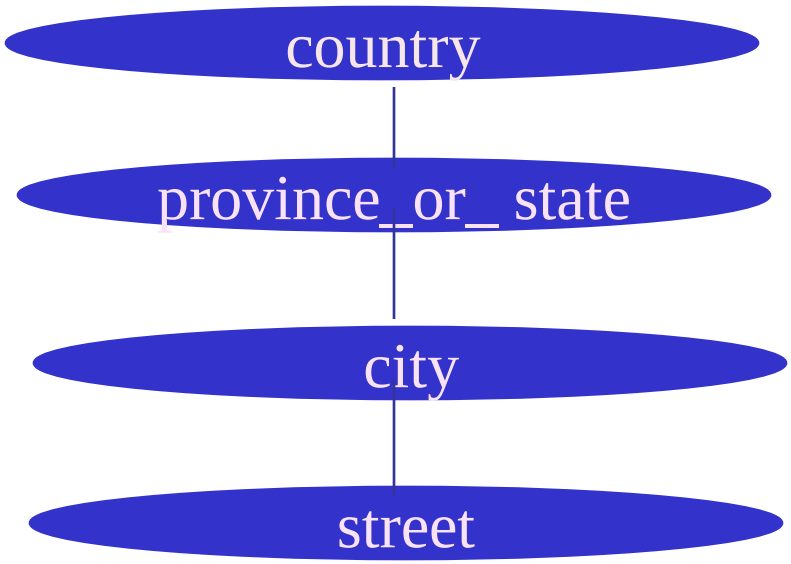


Sinh kiến trúc khái niệm cho dữ liệu phân loại

- Đặc tả một thứ tự bộ phận giá trị thuộc tính theo mức sơ đồ do người dùng hoặc chuyên gias
  - street<city<state<country
- Đặc tả thành cấu trúc phân cấp nhờ nhóm dữ liệu
  - {Urbana, Champaign, Chicago}<Illinois
- Đặc tả theo tập các thuộc tính.
  - Tự động sắp xếp một phần bằng cách phân tích số lượng các giá trị khác biệt
  - Như, street < city <state < country
- Đặc tả một phần thứ tự bộ phận
  - Như, chỉ street < city mà không có cái khác

Sinh kiến trúc khái niệm tự động

- Một vài kiến trúc khái niệm có thể được sinh tự động dựa trên phân tích số lượng các giá trị phân biệt theo thuộc tính của tập DL đã cho
  - Thuộc tính có giá trị phân biệt nhất được đặt ở cấp độ phân cấp thấp nhất
  - Lưu ý: Ngoài trừ, các ngày trong tuần, tháng, quý, năm



15 giá trị phân biệt

65 giá trị phân biệt

3567 giá trị phân biệt

674,339 giá trị phân biệt

Chương 4:
Khai phá luật kết hợp

Dựa theo "Data Mining: Concepts and Techniques"

Chapter 6. Mining Association Rules in Large Databases

www.cs.uiuc.edu/~hanj

Chương 4: Khai phá luật kết hợp

- Khai phá luật kết hợp (Association rule)
- Các thuật toán khai phá vô hướng luật kết hợp (giá trị lôgic đơn chiều) trong CSDL giao dịch
- Khai phá kiểu đa dạng luật kết hợp/tương quan
- Khai phá kết hợp dựa theo ràng buộc
- Khai phá mẫu dãy
- Ứng dụng/mở rộng khai phá mẫu phổ biến

Khái niệm cơ sở: Tập phổ biến và luật kết hợp

Một số ví dụ về *"luật kết hợp" (associate rule)*

- *"*98% khách hàng *mà* mua tạp chí thể thao *thì đều* mua các tạp chí về ôtô*"* ⇨ sự ***kết hợp*** giữa *"tạp chí thể thao"* với *"tạp chí về ôtô"*
- *"*60% khách hàng *mà* mua bia tại siêu thị *thì đều* mua bỉm trẻ em" ⇨ sự ***kết hợp*** giữa *"bia"* với *"bỉm trẻ em"*
- *"*Có tới 70% người truy nhập Web vào địa chỉ Url 1 thì cũng vào địa chỉ Url 2 trong một phiên truy nhập web" ⇨ sự ***kết hợp*** giữa *"Url 1"* với *"Url 2"*. Khai phá dữ liệu sử dụng Web (Dữ liệu từ file log của các site, chẳng hạn được MS cung cấp).
- Các Url có gắn với nhãn "lớp" là các đặc trưng thì có luật kết hợp liên quan giữa các lớp Url này.

Khái niệm cơ sở: Tập phổ biến và luật kết hợp

opinion & misc & travel	-->	on-air	90.26%
news & misc & business & bbs	-->	frontpage	90.24%
living & business & sports & bbs	-->	frontpage	90.00%
news & misc & business & sports	-->	frontpage	89.68%
news & tech & living & business & sports	-->	frontpage	89.00%
news & living & business & bbs	-->	frontpage	88.01%
frontpage & tech & living & business & sports	-->	news	87.87%
frontpage & opinion & living & sports	-->	news	87.81%
frontpage & tech & opinion & living	-->	news	87.60%
frontpage & tech & on-air & business & sports	-->	news	87.59%
news & misc & sports & bbs	-->	frontpage	87.56%
news & tech & on-air & business & sports	-->	frontpage	87.43%
news & living & business & sports	-->	frontpage	87.18%
news & business & sports & bbs	-->	frontpage	86.70%
misc & living & travel	-->	on-air	86.56%
tech & living & sports & bbs	-->	frontpage	86.52%
tech & business & sports & bbs	-->	frontpage	86.40%
news & misc & living & business	-->	frontpage	86.22%
on-air & business & sports & bbs	-->	frontpage	86.22%
news & tech & misc & bbs	-->	frontpage	86.18%
on-air & misc & business & sports	-->	frontpage	86.16%
tech & misc & travel	-->	on-air	86.09%
tech & living & business & sports	-->	frontpage	86.08%
news & living & sports & bbs	-->	frontpage	85.99%
misc & business & sports	-->	frontpage	85.79%
frontpage & tech & opinion & sports	-->	news	85.78%
news & opinion & living & sports	-->	frontpage	85.69%
misc & business & travel	-->	on-air	85.65%
news & tech & misc & business	-->	frontpage	85.63%
misc & business & bbs	-->	frontpage	85.57%
tech & living & sports & bbs	-->	news	85.49%
local & misc & business & sports	-->	frontpage	85.43%
news & opinion & business & bbs	-->	frontpage	85.32%
news & misc & living & sports	-->	frontpage	85.19%
news & on-air & business & sports	-->	frontpage	85.01%

(a)

misc → local	2.07%
frontpage → frontpage → sports	2.02%
local → frontpage	1.83%
on-air → misc → on-air	1.72%
on-air → frontpage	1.69%
on-air → news	1.51%
news → frontpage → news	1.49%
local → news	1.46%
frontpage → frontpage → business	1.35%
news → sports	1.33%
news → bbs	1.23%
health → local	1.16%
misc → frontpage → frontpage	1.16%
on-air → local	1.15%
misc → on-air → misc	1.15%
frontpage → frontpage → living	1.14%
local → frontpage → frontpage	1.13%
health → misc	1.12%
misc → on-air → on-air	1.10%
local → misc → local	1.09%
misc → news	1.06%
news → living	1.06%
on-air → misc → on-air → misc	1.00%

(b)

[IV06] Renáta Iváncsy, István Vajk (2006). Frequent Pattern Mining in Web Log Data, *Acta Polytechnica Hungarica*, 3(1):77-90, 2006

Khái niệm cơ sở: Tập phổ biến và luật kết hợp

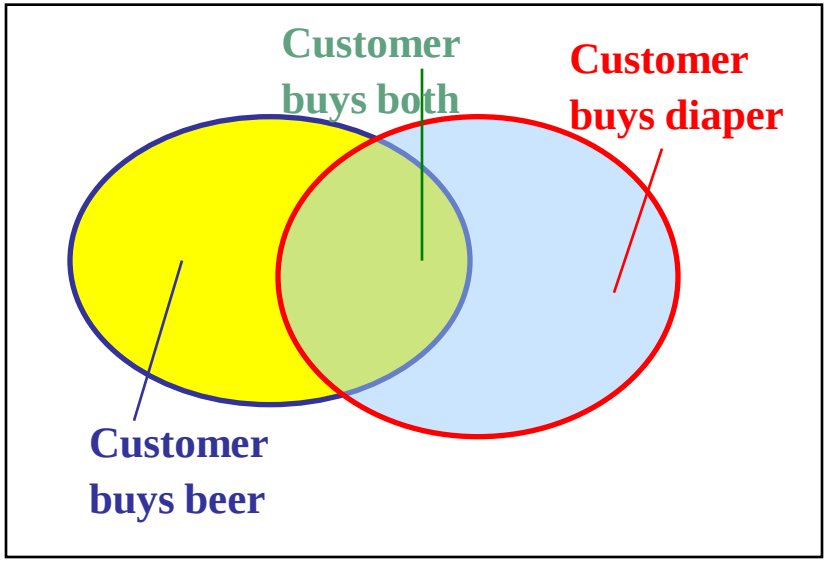
Cơ sở dữ liệu giao dịch (transaction database)

- *Giao dịch*: danh sách các mặt hàng (mục: item) trong một phiếu mua hàng của khách hàng. Giao dịch T là một tập mục.
- Tập toàn bộ các mục I = {i_1, i_2, …, i_k} "tất cả các mặt hàng". Một giao dịch T là một tập con của I: T [redacted] Mỗi giao dịch T có một định danh là T_{ID}.
- A là một tập mục A [redacted] và T là một giao dịch: Gọi T chứa A nếu A [redacted].

- Luật kết hợp
  - Gọi A [redacted] là một "luật kết hợp" nếu A [redacted] B [redacted] và A [redacted] =[redacted]
  - Luật kết hợp A [redacted] có độ hỗ trợ (support) *s* trong CSDL giao dịch D nếu trong D có s% các giao dịch T chứa AB: chính là xác suất P(AB). Tập mục A có P(A) [redacted]s>0 (với s cho trước) được gọi là *tập phổ biến (frequent set)*. Luật kết hợp A [redacted] có độ tin cậy (confidence) *c* trong CSDL D nếu như trong D có c% các giao dịch T chứa A thì cũng chứa B: chính là xác suất P(B|A).
  - Support (A [redacted]) = P(A [redacted]) : 1 [redacted] (A [redacted]) [redacted]
  - Confidence (A [redacted]) = P(B|A) : 1 [redacted] (A [redacted]) [redacted]
  - Luật A [redacted] được gọi là đảm bảo độ hỗ trợ s trong D nếu s(A [redacted]) [redacted]s. Luật A [redacted] được gọi là đảm bảo độ tin cậy c trong D nếu c(A [redacted]) [redacted]. Tập mạnh.

Khái niệm cơ bản: Mẫu phổ biến và luật kết hợp

Transaction-id	Items bought
10	A, B, C
20	A, C
30	A, D
40	B, E, F



- Tập mục I={i_1, …, i_k}. CSDL giao dịch D = {d
- A, B A = A → B là luật kết hợp
- Bài toán tìm luật kết hợp.

  Cho trước độ hỗ trợ tối thiểu s>0, độ tin cậy tối thiểu c>0. Hãy tìm mọi luật kết hợp mạnh *X→Y.*

Giả sử min_support = 50%,
min_conf = 50%:

$A \rightarrow C$ (50%, 66.7%)

$C \rightarrow A$ (50%, 100%)

- Hãy trình bày các nhận xét về khái niệm luật kết hợp với khái niệm phụ thuộc hàm.
- Các tính chất Armstrong ở đây.

Một ví dụ tìm luật kết hợp

Transaction-id	Items bought
10	A, B, C
20	A, C
30	A, D
40	B, E, F

Min. support 50%
Min. confidence 50%

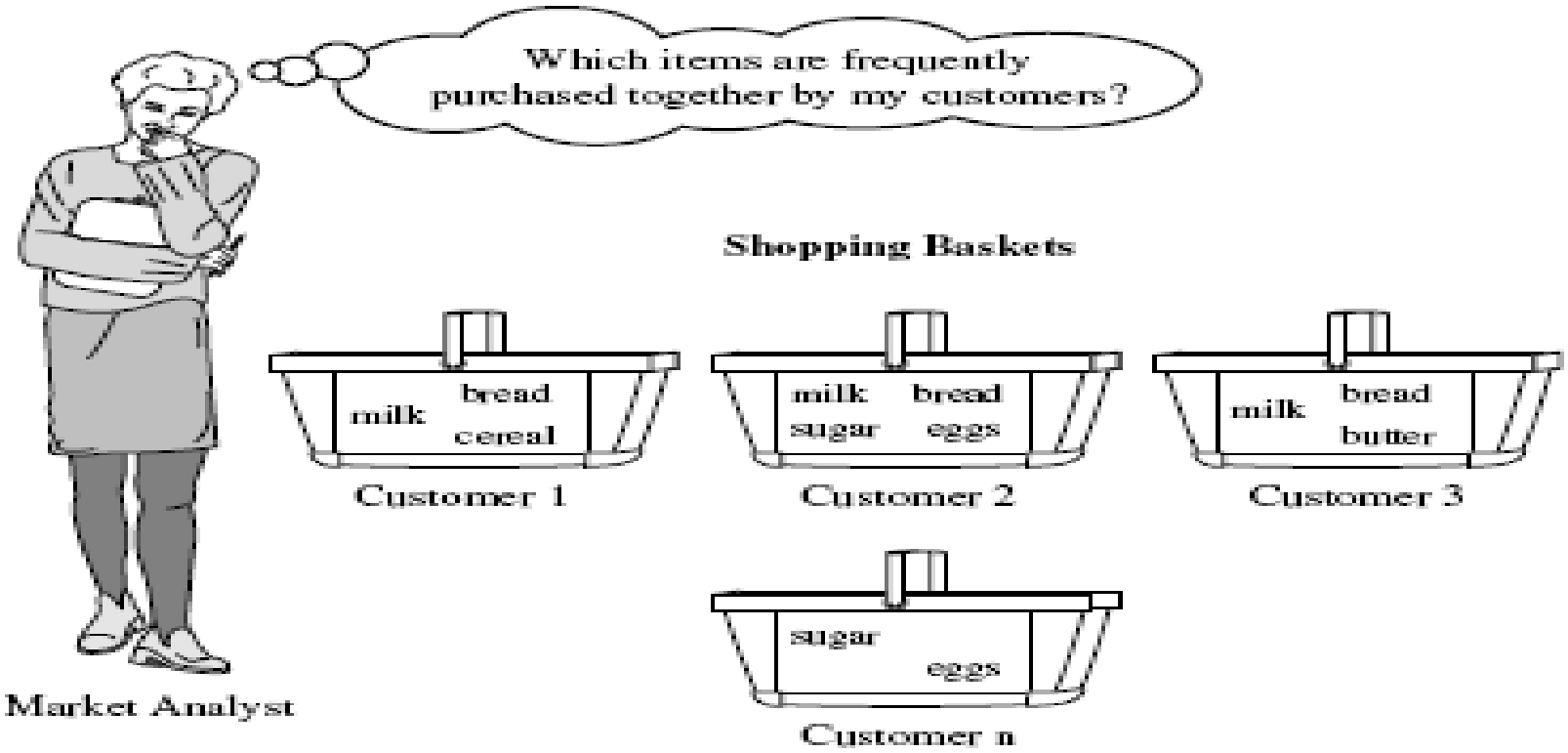
Frequent pattern	Support
{A}	75%
{B}	50%
{C}	50%
{A, C}	50%

For rule $A \Rightarrow C$:

support = support($\{A\} \Rightarrow C\}$) = 50%

confidence = support($\{A\} \Rightarrow C\}$)/support($\{A\}$) = 66.6%

Khai niệm khai phá kết hợp



$$computer \Rightarrow antivirus\_software \ [support = 2\%, confidence = 60\%]$$

Khái niệm khai phá luật kết hợp

- Khai phá luật kết hợp:
  - Tìm tất cả mẫu phổ biến, kết hợp, tương quan, hoặc cấu trú nhan-quả trong tập các mục hoặc đối tượng trong CSDL quan hệ hoặc các kho chứa thông tin khác.
  - Mẫu phổ biến (Frequent pattern): là mẫu (tập mục, dãy mục…) mà xuất hiện phổ biến trong 1 CSDL [AIS93]
- Động lực: tìm mẫu chính quy (regularities pattern) trong DL
  - Các mặt hàng nào được mua cùng nhau? — Bia và bỉm (diapers)?!
  - Mặt hàng nào sẽ được mua sau khi mua một PC ?
  - Kiểu DNA nào nhạy cảm với thuộc mới này?
  - Có khả năng tự động phân lớp Web hay không ?

Mẫu phổ biến và khai phá luật kết hợp là một bài toán bản chất của khai phá DL

- Nền tảng của nhiều bài toán KPDL bản chất
  - Kết hợp, tương quan, nhân quả
  - Mẫu tuần tự, kết hợp thời gian hoặc vòng, chu kỳ bộ phận, kết hợp không gian và đa phương tiện
  - Phân lớp kết hợp, phân tích cụm, khối tảng băng, tích tụ (nén dữ liệu ngữ nghĩa)
- Ứng dụng rộng rãi
  - Phân tích DL bóng rổ, tiếp thị chéo (cross-marketing), thiết kế catalog, phân tích chiến dịch bán hàng
  - Phân tích Web log (click stream), Phân tích chuỗi DNA v.v.

Chương 4: Khai phá luật kết hợp

- Khai phá luật kết hợp (Association rule)
- Các thuật toán khai phá vô hướng luật kết hợp (giá trị lôgic đơn chiều) trong CSDL giao dịch
- Khai phá kiểu đa dạng luật kết hợp/tương quan
- Khai phá kết hợp dựa theo ràng buộc
- Khai phá mẫu dãy
- Ứng dụng/mở rộng khai phá mẫu phổ biến

Apriori: Một tiếp cận sinh ứng viên và kiểm tra

- Khái quát: Khai phá luật kết hợp gồm hài bước:
  - Tìm mọi tập mục phổ biến: theo min-sup
  - Sinh luật mạnh từ tập mục phổ biến
- Mọi tập con của tập mục phổ biến cũng là tập mục phổ biến
  - Nếu {*bia*, *bỉm*, *hạnh nhân*} là phổ biến thì {*bia*, *bỉm*} cũng vậy: Mọi giao dịch chứa {*bia*, *bỉm*, *hạnh nhân*} cũng chứa {*bia*, *bỉm*}.
- Nguyên lý tỉa Apriori: Với mọi tập mục không phổ biến thì mọi tập bao không cần phải sinh ra/kiểm tra!
- Phương pháp:
  - Sinh các tập mục ứng viên dài (k+1) từ các tập mục phổ biến có độ dài k (Độ dài tập mục là số phần tử của nó),
  - Kiểm tra các tập ứng viên theo CSDL
- Các nghiên cứu hiệu năng chứng tỏ tính hiệu quả và khả năng mở rộng của thuật toán
- Agrawal & Srikant 1994, Mannila, và cộng sự 1994

Thuật toán Apriori

Trên cơ sở tính chất (nguyên lý tỉa) Apriori, thuật toán hoạt động theo quy tắc quy hoạch động

- Từ các tập $F_i = \{c_i | \ c_i$ tập phổ biến, $|c_i| = i\}$ gồm mọi tập mục phổ biến có độ dài i với 1 [redacted] k,
- đi tìm tập F_{k+1} gồm mọi tập mục phổ biến có độ dài k+1.

Trong thuật toán, các tên mục $i_1, i_2, \ldots i_n$ (n = |I|) *được sắp xếp theo một thứ tự cố định* (thường được đánh chỉ số 1, 2, ..., n).

Thuật toán Apriori

Thuật toán Apriori [WKQ08]:

Input: - Cơ sở dữ liệu giao dịch D = {t| t giao dịch}

- Độ hỗ trợ tối thiểu minsup > 0

Output: - Tập hợp tất cả các tập phổ biến.

0: mincount = minsup * |D|;
1. F_1 = {các tập phổ biến có độ dài 1}
2. for (k=1; $F_k \neq \emptyset$; k++) do begin
3. C_{k+1} = apriori-gen (F_k); // sinh mọi ứng viên độ dài k+1
4. for t ∈ D do begin
5. C_t = {c ∈ C_{k+1} | c ⊆ t}; //mọi ứng viên chứa trong t
6. for c ∈ C_t do
7. c.count ++;
8. end
9. F_{k+1} = {c ∈ C_{k+1} | c.count ≥ mincount} ;
10. end
11. Answer $\cup_k F_k$;

Thuật toán Apriori: Thủ tục con Apriori-gen

Trong mỗi bước k, thuật toán Apriori đều phải duyệt CSDL D.

Khởi động, duyệt D để có được F_1.

Các bước k sau đó, duyệt D để tính số lượng giao dịch t thoả từng ứng viên c của C_{k+1}: mỗi giao dịch t chỉ xem xét một lần cho mọi ứng viên c thuộc C_{k+1}.

Thủ tục con Apriori-gen sinh tập phổ biến: *tư tưởng*

Bước nối: Sinh các tập mục R_{k+1} là ứng viên tập phổ biến có độ dài k+1 bằng cách kết hợp hai tập phổ biến P_k và Q_k có độ dài k và trùng nhau ở k-1 mục đầu tiên:

$$R_{k+1} = P_k \cup Q_k = \{i_1, i_2, ..., i_{k-1}, i_k, i_{k'}\} \text{ với}$$

$$P_k = \{i_1, i_2, ..., i_{k-1}, i_k\} \text{ và } Q_k = \{i_1, i_2, ..., i_{k-1}, i_{k'}\}$$

trong đó $i_1 \le i_2 \le ... \le i_{k-1} \le i_k \le i_{k'}$.

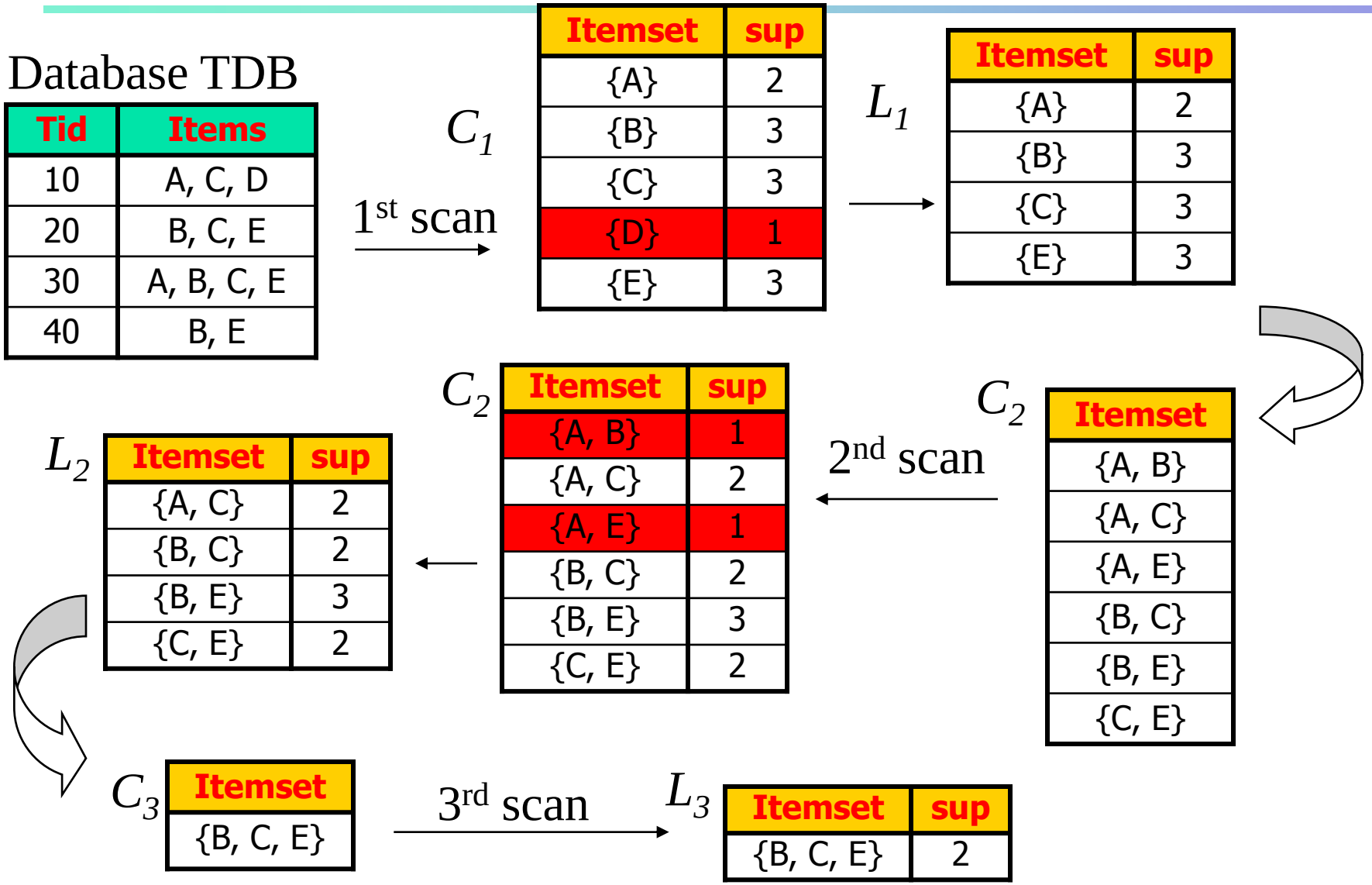
Bước tỉa: Giữ lại tất cả các R_{k+1} thỏa tính chất Apriori ($\forall X \subseteq R_{k+1}$ và $|X|=k \Rightarrow X \in F_k$), nghĩa là đã loại (tia) bớt đi mọi ứng viên R_{k+1} không đáp ứng tính chất này.

Thủ tục con Apriori-gen

(1) for mọi tập mục phổ biến $l_1 \in L_k$
(2) for mọi tập mục phổ biến $l_2 \in L_k$
(3) if $(l_1[1]=l_2[1]) \wedge (l_1[2]=l_2[2]) \wedge \ldots \wedge (l_1[k-1]=l_2[k-1]) \wedge (l_1[k]<l_2[k])$
then {
$c = l_1 \Leftrightarrow l_2$; // join step: generate candidates
// c = $\{l_1[1], l_1[2], \ldots, l_1[k-1], l_1[k], l_2[k]\}$
(5) if has_infrequent_subset(c, L_k) then
(6) delete c; // bước tỉa: bỏ ứng viên không đúng
else add c to C_{k+1};
(8) }
(9) return Ck;

procedure has_infrequent_subset(c: tập ứng viên độ dài $k+1$;
L_k: tập các tập mục phổ biến độ dài k); // tri thức đã có
(1) for mỗi tập con s độ dài k *của* c
(2) if $s \notin L_k$ then
(3) return TRUE;
(4) return FALSE;

Một ví dụ thuật toán Apriori (s=0.5)

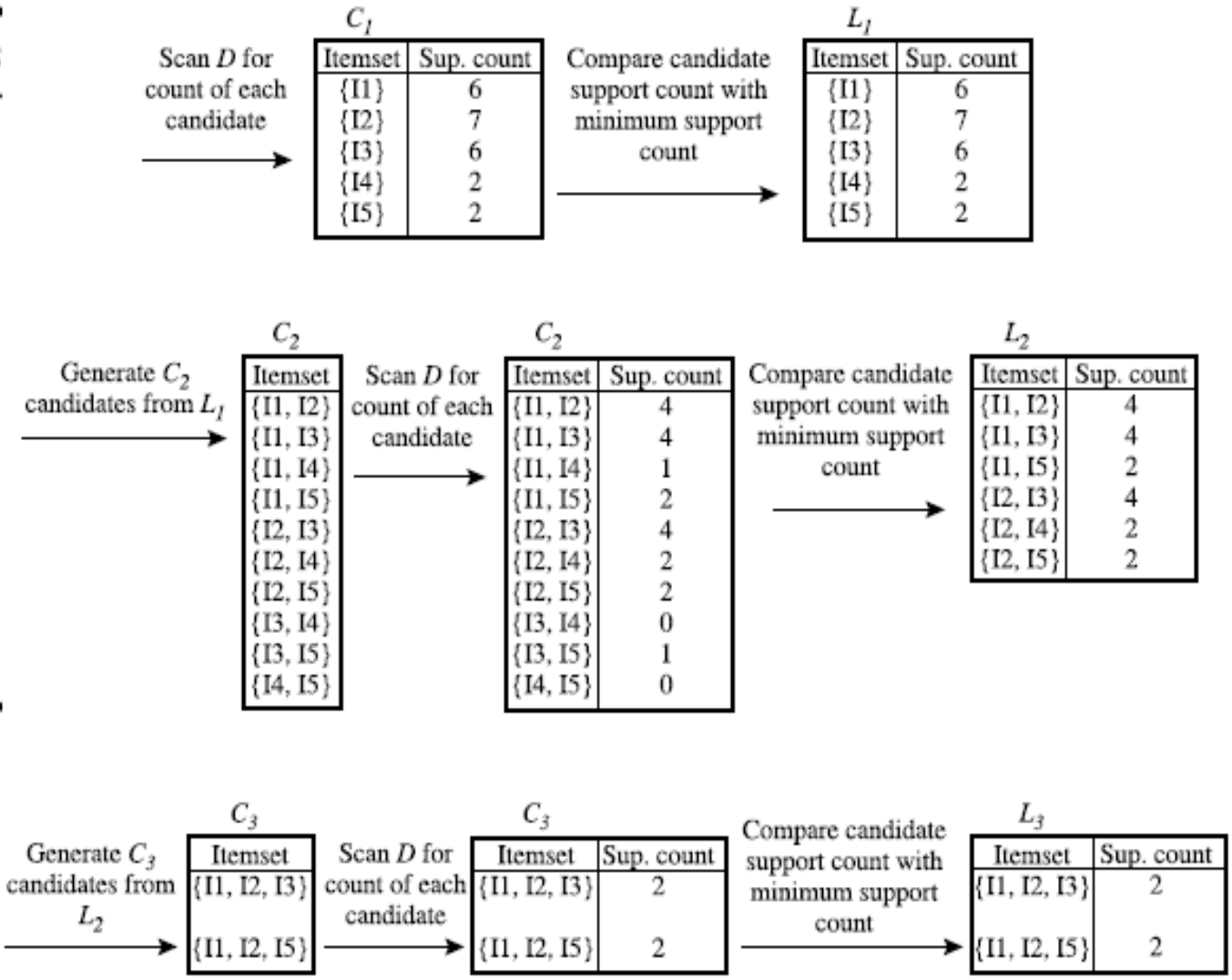


Chi tiết quan trọng của Apriori

- Cách thức sinh các ứng viên:
  - Bước 1: Tự kết nối L_k
  - Step 2: Cắt tỉa
- Cách thức đếm hỗ trợ cho mỗi ứng viên.
- Ví dụ thủ tục con sinh ứng viên
  - L_3={*abc, abd, acd, ace, bcd*}
  - Tự kết nối: L_3*L_3
    - *abcd* từ *abc* và *abd*
    - *acde* từ *acd* và *ace*
  - Tỉa:
    - *acde* là bỏ đi vì *ade* không thuộc L_3
  - C_4={*abcd*}

Ví dụ: D, min_sup*|D| = 2 (C_4 =

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3



Sinh luật kết hợp

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Việc sinh luật kết hợp gồm hai bước

- Với mỗi tập phổ biến W tìm được hãy sinh ra mọi tập con thực sự X khác rỗng của nó.
- Với mỗi tập phố biến W và tập con X khác rỗng thực sự của nó: sinh luật X [redacted] W – X) nếu P(W-X|X) [redacted].

Như ví dụ đã nêu có L3 = {{I1, I2, I3}, {I1, I2, I5}}

Với độ tin cậy tối thiểu 70%, xét tập mục phổ biến {I1, I2, I5} có 3 luật như dưới đây:

$$I1 \wedge I2 \Rightarrow I5, \quad confidence = 2/4 = 50\%$$
$$I1 \wedge I5 \Rightarrow I2, \quad confidence = 2/2 = 100\%$$
$$I2 \wedge I5 \Rightarrow I1, \quad confidence = 2/2 = 100\%$$
$$I1 \Rightarrow I2 \wedge I5, \quad confidence = 2/6 = 33\%$$
$$I2 \Rightarrow I1 \wedge I5, \quad confidence = 2/7 = 29\%$$
$$I5 \Rightarrow I1 \wedge I2, \quad confidence = 2/2 = 100\%$$

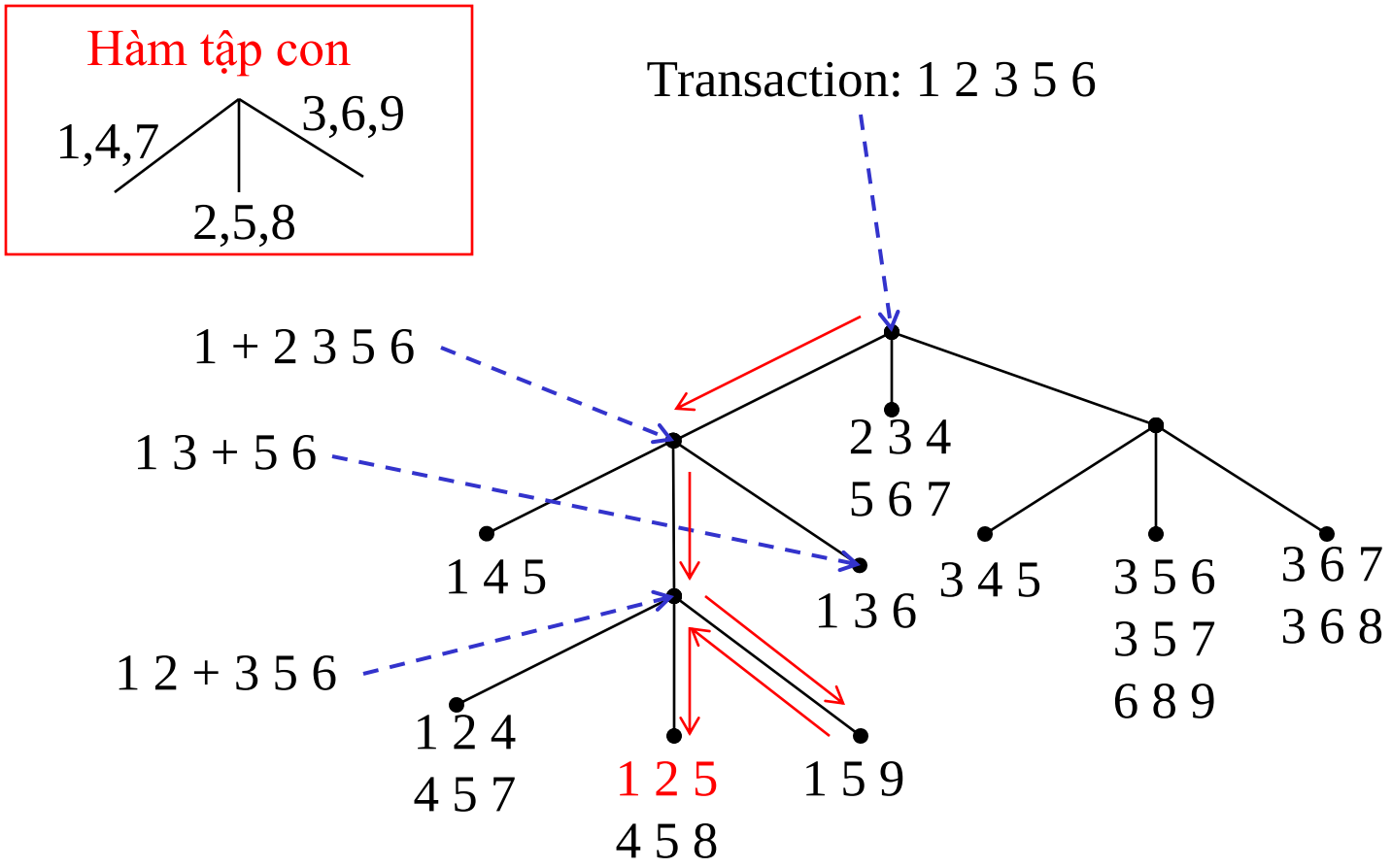
Cách thức tính độ hỗ trợ của ứng viên

- Tính độ hỗ trợ ứng viên là vấn đề cần quan tâm
  - Số lượng ứng viên là rất lớn
  - Một giao dịch chứa nhiều ứng viên
- Phương pháp:
  - Tập mục ứng viên được chứa trong một cây-băm (*hash-tree*)
  - *Lá* của cây băm chứa một danh sách các tập mục và bộ đếm
  - *Nút trong* chứa bảng băm
  - *Hàm tập con*: tìm tất cả các ứng viên

Cách thức tính độ hỗ trợ của ứng viên

- Tập các ứng viên C_k được lưu trữ trong một cây-băm.
  - Gốc của cây băm ở độ sâu 1. Lá chứa một danh sách tập mục
  - Nút trong chứa một bảng băm: mỗi thùng của bảng trỏ tới một nút khác (Nút ở độ sâu d trỏ tới các nút ở độ sâu d+1).
  - Khi khởi tạo, tất cả các nút là lá.
- Khi thêm một tập mục c:
  - bắt đầu từ gốc đi xuống theo cây cho đến khi gặp một lá.
  - Tại một nút trong độ sâu d:
    - quyết định theo nhánh nào bằng cách áp dụng hàm băm tới mục thứ d của tập mục này.
    - Khi số lượng tập mục tại một lá vượt quá ngưỡng quy định, nút lá được chuyển thành một nút trong.
- Bắt đầu từ gốc, tìm tất cả các ứng viên thuộc giao dịch t:
  - Nếu ở nút gốc: băm vào mỗi mục trong t.
  - Nếu ở một lá: tìm các tập mục ở lá này thuộc t và bổ sung chỉ dẫn tới các tập mục này tới tập trả lời.
  - Nếu ở nút trong và đã đạt được nó bằng cách băm mục i, trên từng mục đứng sau i trong t và áp dụng đệ quy thủ tục này sang nút trong thùng tương ứng.

Ví dụ: Tính hỗ trợ các ứng viên



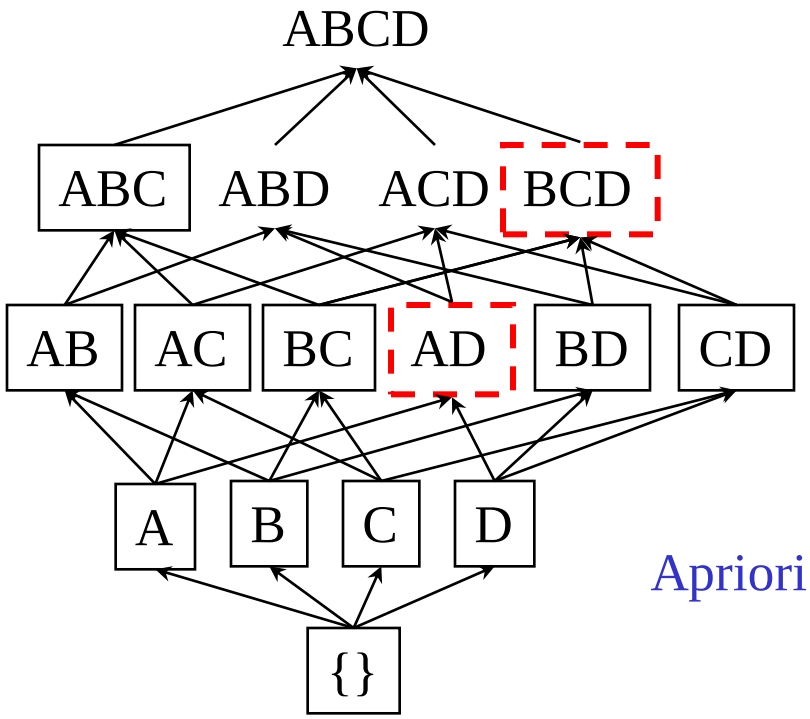
Thi hành hiệu quả thuật toán Apriori trong SQL

- Khó để có thể có một hiệu quả tốt nếu chỉ tiếp cận thuần SQL (SQL-92)
- Sử dụng các mở rộng quan hệ - đối tượng như UDFs, BLOBs, hàm bảng v.v.
  - Nhận được các thứ tự tăng quan trọng
- Xem bải: S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database systems: Alternatives and implications. In SIGMOD'98

Thách thức khai phá mẫu phổ biến

- Thách thức
  - Duyệt phức CSDL giao dịch
  - Lượng rất lớn các ứng viên
  - Tẻ nhạt việc tính toán độ hỗ trợ
- Cải tiến Apriori: tư tưởng chung
  - Giảm số lần duyệt CSDL giao dịch
  - Rút số lượng các ứng viên
  - Giảm nhẹ tính độ hỗ trợ của các ứng viên

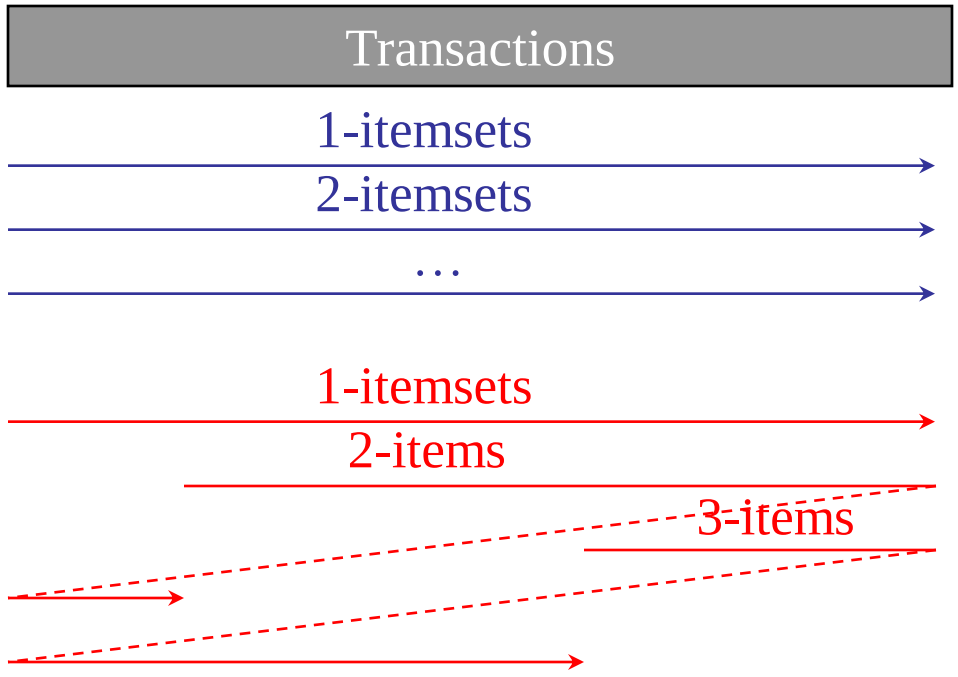
DIC (Đếm tập mục động): Rút số lượng duyệt CSDL



Itemset lattice

S. Brin R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD'97*

- Mỗi khi A và D được xác định là phổ biến thì việc tính toán cho AD được bắt đầu
- Mỗi khi mọi tập con độ dài 2 của BCD được xác định là phổ biến: việc tính toán cho BCD được bắt đầu.



Giải pháp Phân hoạch (Partition): Duyệt CSDL chỉ hai lần

- Mọi tập mục là phổ biến tiềm năng trong CSDL bắt buộc phải phổ biến ít nhất một vùng của DB
  - Scan 1: Phân chia CSDL và tìm các mẫu cục bộ
  - Scan 2: Hợp nhất các mẫu phổ biến tổng thể
- A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association in large databases. In *VLDB'95*

Ví dụ về mẫu phổ biến

- Chọn một mẫu của CSDL gốc, khai phá mẫu phổ biến nội bộ mẫu khi dùng Apriori
- Duyệt CSDL một lần để kiểm tra các tập mục phổ biến tìm thấy trong ví dụ, chỉ có bao (*borders*) *đóng* của các mẫu phổ biến được kiểm tra
  - Ví dụ: kiểm tra *abcd* thay cho *ab, ac, …, v.v.*
- Duyệt CSDL một lần nữa để tìm các mẫu phổ biến bị mất (bỏ qua)

H. Toivonen. Sampling large databases for association rules. In *VLDB'96*

DHP: Rút gọn số lượng các ứng viên

- Một *k*-tập mục mà bộ đếm trong lô băm tương ứng dưới ngưỡng thì không thể là tập mục phổ biến
  - Ứng viên: a, b, c, d, e
  - Điểm vào băm: {ab, ad, ae} {bd, be, de} …
  - 1-tập mục phổ biến: a, b, d, e
  - ab không là một ứng viên 2-tập mục nếu tổng bộ đếm của {ab, ad, ae} là dưới ngưỡng hỗ trợ

J. Park, M. Chen, and P. Yu. An effective hash-based algorithm for mining association rules. In *SIGMOD'95*

Eclat/MaxEclat và VIPER: Thăm dò dạng dữ liệu theo chiều ngang

- Dùng danh sách tid của giáo dịch trong một tập mục
- Nén danh sách tid
  - Tập mục A: t1, t2, t3, sup(A)=3
  - Tập mục B: t2, t3, t4, sup(B)=3
  - Tập mục AB: t2, t3, sup(AB)=2
- Thao tác chính: lấy giáo của các danh sách tid
- M. Zaki et al. New algorithms for fast discovery of association rules. In KDD'97
- P. Shenoy et al. Turbo-charging vertical mining of large databases. In SIGMOD'00

Thắt cơ chai của khai phá mẫu phổ biến

- Duyệt CSDL nhiều là tốn kém
- KP mẫu dài cần nhiều bước để duyệt và sinh nhiều ứng viên
  - Để tìm các tập mục phổ biến $i_1 i_2 \ldots i_{100}$
    - # duyệt: 100
    - # ứng viên: $\binom{100}{1} + \binom{100}{2} + \ldots + \binom{1}{1}\binom{0}{0}\binom{0}{0} = 2^{100}-1 = 1.27*10^{30}$!
- Thắt cổ chai: sinh ứng viên và kiểm tra
- Tránh sinh ứng viên?

KP mẫu phổ biến không cần sinh ƯV

- Dùng các mục phổ biến để tăng độ dài mẫu từ các mẫu ngắn hơn
  - "abc" là một mẫu phổ biến
  - Nhận mọi giao dịch có "abc": DB|abc
  - "d" là một mục phổ biến trong DB|abc → abcd là một mẫu phổ biến

Xây dựng cây FP từ một CSDL giao dịch

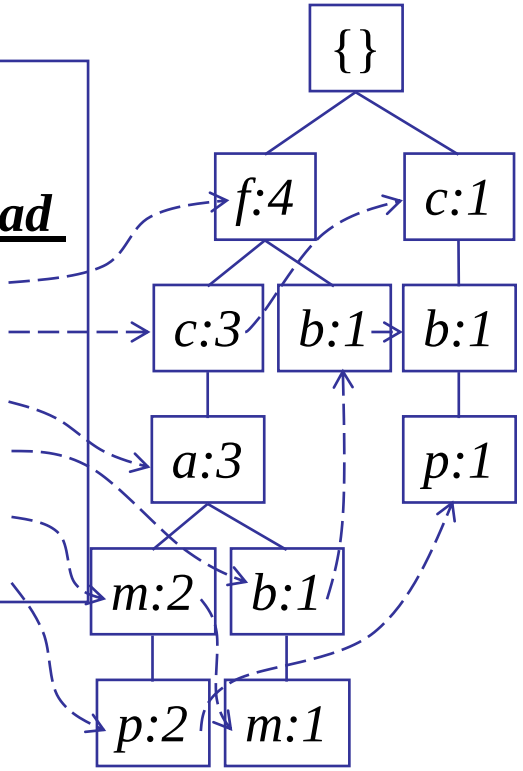
TID	Items bought	(ordered) frequent items
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o, w}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

min_support = 3

1. Duyệt CSDL một lần, tìm các 1-tập mục phổ biến (mẫu mục đơn)
2. Sắp xếp các mục phổ biến theo thứ tự giảm dần về bậc, f-list
3. Duyệt CSDL lần nữa, xây dựng FP-tree

Header Table

Item	frequency	head
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	



F-list=f-c-a-b-m-p

Lợi ích của cấu trúc FP-tree

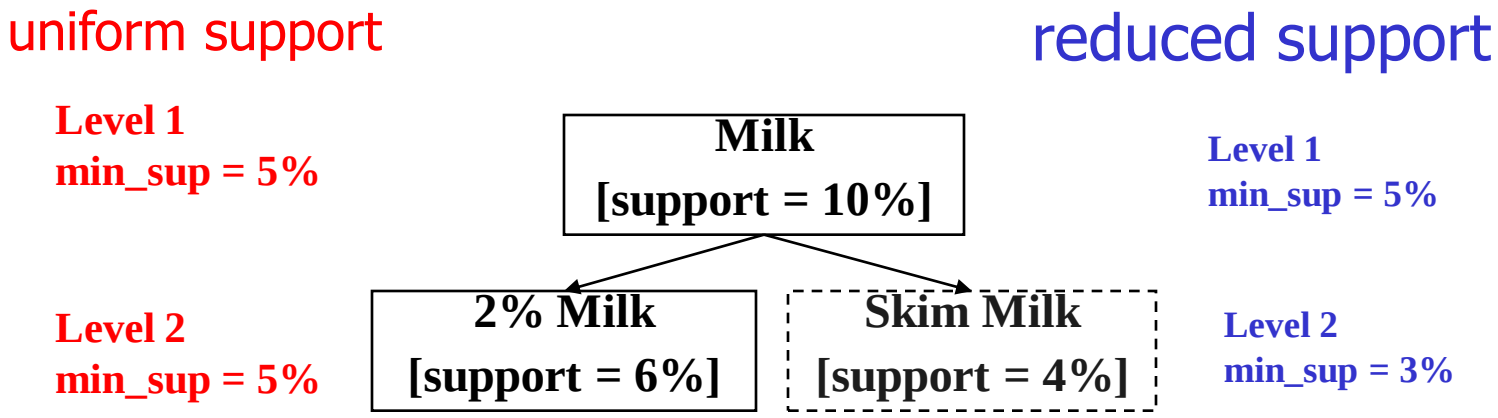
- Tính đầy đủ
  - Duy trì tính đầy đủ thông tin để khai phá mẫu phổ biến
  - Không phá vỡ mẫu dài bởi bất kỳ giao dich
- Tính cô đọng
  - Giảm các thông tin không liên quan: mục không phổ biến bỏ đi
  - Sắp mục theo tần số giảm: xuất hiện càng nhiều thì cành hiệu quả
  - Không lớn hơn so với CSDL thông thường

Chương 4: Khai phá luật kết hợp

- Khai phá luật kết hợp (Association rule)
- Các thuật toán khai phá vô hướng luật kết hợp (giá trị lôgic đơn chiều) trong CSDL giao dịch
- Khai phá kiểu đa dạng luật kết hợp/tương quan
- Khai phá kết hợp dựa theo ràng buộc
- Khai phá mẫu dãy
- Ứng dụng/mở rộng khai phá mẫu phổ biến

Luật kết hợp đa mức

- Các mục có thể đa phân cấp
- Đặt hỗ trợ linh hoạt: Mục cấp thấp hơn là kỳ vọng hỗ trợ thấp hơn.
- CSDL giao dịch có thể được mã hóa theo chiều và mức
- Thăm dò KP đa mức chia sẻ



Kết hợp đa chiều

- Luật đơn chiều:

  buys(X, “milk”) ⇒ buys(X, “bread”)

- Luật đa chiều: ≥ 2 chiều hoặc thuộc tính
  - Luật kết hợp liên chiều (không có thuộc tính lặp)

    age(X,”19-25”) ∧ occupation(X,“student”) ⇒ buys(X,“coke”)
  - Luật KH chiều-kết hợp (lai/hybrid) (lặp thuộc tính)

    age(X,”19-25”) ∧ buys(X, “popcorn”) ⇒ buys(X, “coke”)
- Thuộc tính phân lớp
  - Tìm số lượng các giá trị khả năng không được sắp
- Thuộc tính định lượng
  - Số, thứ tự ngầm định trong miền giá trị

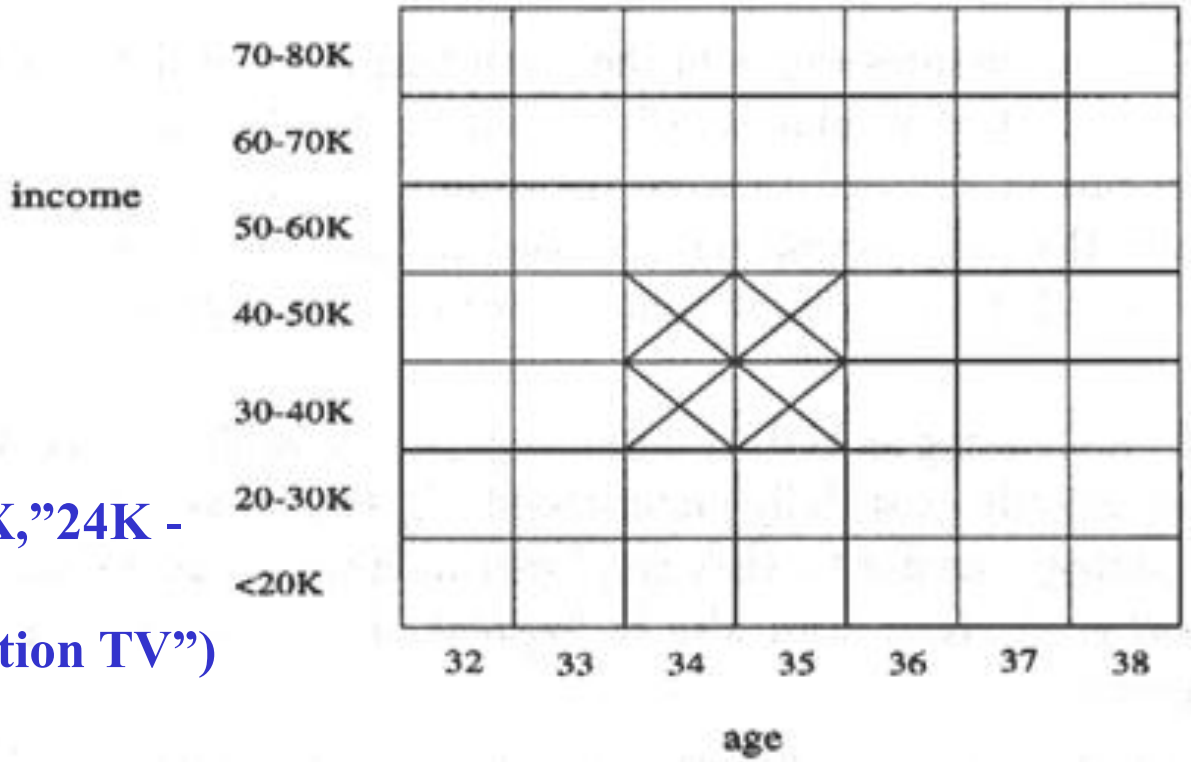
Kết hợp đa mức: Rút gọn lọc

- Một vài luật có thể dư thừa do có quan hệ "tổ tiên" giữa các mục.
- Ví dụ
  - milk [redacted]heat bread [support = 8%, confidence = 70%]
  - 2% milk [redacted]heat bread [support = 2%, confidence = 72%]
- Nói rằng: luật đầu tiên là tổ tiên luật thứ hai.
- Một luật là dư thừa nếu độ hỗ trợ của nó là khít với giá trị "mong muốn", dựa theo tổ tiên của luật.

Luật kết hợp định lượng

- Numeric attributes are *dynamically* discretized
  - Such that the confidence or compactness of the rules mined is maximized
- 2-D quantitative association rules: A_{quan1} ∧ A_{quan2} ⇒ A_{cat}
- Cluster "adjacent"
  association rules to form general rules using a 2-D grid
- Example

age(X,"30-34") ∧ income(X,"24K - 48K") ⇒ buys(X,"high resolution TV")



Khai phá luật KH dựa theo khoảng cách

- Binning methods do not capture the semantics of interval data

Price($)	Equi-width (width $10)	Equi-depth (depth 2)	Distance-based
7	[0,10]	[7,20]	[7,7]
20	[11,20]	[22,50]	[20,22]
22	[21,30]	[51,53]	[50,53]
50	[31,40]		
51	[41,50]		
53	[51,60]		

- Distance-based partitioning, more meaningful discretization considering:
  - density/number of points in an interval
  - "closeness" of points in an interval

Độ đo hấp dẫn: Tương quan (nâng cao)

- *play basketball* [redacted] *at cereal* [40%, 66.7%] là lạc
  - Phần trăm chung của sinh viên ăn ngũ cố là 75% cao hơn so với 66.7%.
- *play basketball* [redacted] *ot eat cereal* [20%, 33.3%] là chính xác hơn, do độ hỗ trợ và tin cậy thâos hơn
- Độ đo sự kiện phụ thuộc/tương quan: lift (nâng cao)

$$corr_{A,B} \; [redacted] \frac{P(A[redacted])}{P(A)P(B)}$$

	Basketball	Not basketball	Sum (row)
Cereal	2000	1750	3750
Not cereal	1000	250	1250
Sum(col.)	3000	2000	5000

Chương 4: Khai phá luật kết hợp

- Khai phá luật kết hợp (Association rule)
- Các thuật toán khai phá vô hướng luật kết hợp (giá trị lôgic đơn chiều) trong CSDL giao dịch
- Khai phá kiểu đa dạng luật kết hợp/tương quan
- Khai phá kết hợp dựa theo ràng buộc
- Khai phá mẫu dãy
- Ứng dụng/mở rộng khai phá mẫu phổ biến

KPDL dựa trên ràng buộc

- Tìm tất cả các mẫu trong CSDL tự động? — phi hiện thực!
  - Mẫu có thể quá nhiều mà không mục đích!
- KPDL nên là quá trình tương tác
  - Người dùng trực tiếp xác định KPDL gì khi dùng ngôn ngữ hỏi KPDL (hoặc giao diện đồ họa)
- KP dựa theo ràng buộc
  - Linh hoạt người dùng: cung cấp ràng buộc trên cái mà KP
  - Tối ưu hệ thống: thăm dò các ràng buộc để hiệu quả KP: KP dựa theo ràng buộc

Ràng buộc trong KPDL

- Ràng buộc kiểu tri thức:
  - classification, association, etc.
- Ràng buộc dữ liệu — dùng câu hỏi kiểu SQL
  - Tìm các cặp sản phẩn mua cùngnhau trong Vancouver vào Dec.'00
- Ràng buộc chiều/cấp
  - Liên quan tới vùng, giá, loại hàng, lớp khách hàng
- Ràng buộc luật (mẫu)
  - Mua hàng nhỏ (price < $10) nhanh hơn mua hàng lớn (sum > $200)
- Ràng buộc hấp dẫn
  - Luật mạng: min_support ■ 3%, min_confidence ■ 60%

KP ràng buộc <> tìm kiếm dựa theo ràng buộc

- KP ràng buộc <> tìm/lập luận dựa theo ràng buộc
  - Cả hai hướng tới rút gọn không gian tìm kiếm
  - Tìm mọi mẫu bảm đảm ràng buộc <> tìm một vài (một_ câu trả lời của tìm dựa theo ràng buộc trong AI (TTNT)
  - Cố tìm theo ràng buộc <> tìm kiếm heuristic
  - *Tích hợp hai cái cho một bài toán tìm kiếm thú vịi*
- KP ràng buộc <> quá trình hỏi trong hệ CSDL quan hệ
  - Quá trình hỏi trong CSDL quan hệ đòi hỏi tìm tất cả
  - KP mẫu ràng buộc chung một triết lý tương tựng như cố gắng chọn về chiều sâu của câu hỏi

KP mẫu phổ biến ràng buộc: vấn đề tối ưu hóa câu hỏi

- Cho một câu hỏi KP Mẫu phổ biến với một tập ràng buộc C, thì thuật toán nên là
  - Mạnh mẽ: chỉ tìm các tập phố biến bảo đảm ràng buộc *C*
  - đầy đủ: Tìm tất cả tập phổ biến bảo đảm ràng buộc *C*
- Giải pháp "thơ ngây/hồn nhiên" (naïve)
  - Tìm tất cát tập PB sau đó kiểm tra ràng buộc
- Tiếp cận hiệu quả hơn
  - Phân tích tính chất các ràng buộc một cách toàn diện
  - Khai thác chúng sâu sắc có thể nhất trong tính toán mẫu PB.

Chống đơn điệu trong KP theo ràng buộc

- Chống đon điệu (Anti-monotonicity)
  - *Một tập mục S **vi phạm** ràng buộc, mọi tập lớn hơn nó cũng vi phạm*
  - *sum(S.Price)* [redacted] *v* là chống đơn điệu
  - *sum(S.Price)* [redacted] *v* là không chống đơn điệu
- Ví dụ. C: range(S.profit) [redacted]15 là chống đơn điệu
  - Tập mục *ab* vi phạm C
  - Cũng vậy mọi tập chưa *ab*

TDB (min_sup=2)

TID	Transaction
10	a, b, c, d, f
20	b, c, d, f, g, h
30	a, c, d, e, f
40	c, e, f, g

Item	Profit
a	40
b	0
c	-20
d	10
e	-30
f	30
g	20
h	-10

Ràng buộc nào là chống đơn điệu

Ràng buộc	Chống đơn điệu
v [redacted]	No
S [redacted]	no
S [redacted]	yes
min(S) [redacted]	no
min(S) [redacted]	yes
max(S) [redacted]	yes
max(S) [redacted]	no
count(S) [redacted]	yes
count(S) [redacted]	no
sum(S) [redacted] (a [redacted], a [redacted])	yes
sum(S) [redacted] (a [redacted], a [redacted])	no
range(S) [redacted]	yes
range(S) [redacted]	no
avg(S) [redacted], [redacted]	convertible
support(S) [redacted]	yes
support(S) [redacted]	no

Tính đơn điệu trong KP dựa theo ràng buộc

TDB (min_sup=2)

TID	Transaction
10	a, b, c, d, f
20	b, c, d, f, g, h
30	a, c, d, e, f
40	c, e, f, g

- Tính đơn điệu
  - *Khi một tập mục S **thỏa mãn** ràng buộc, thì mọi tập lớn hơn của nó cung thỏa mãn*
  - *sum(S.Price)* [redacted] là đơn điệu
  - *min(S.Price)* [redacted] *là đơn điệu*
- Ví dụ. C: range(S.profit) [redacted]15
  - Tập mục *ab đảm bảo* C
  - Cúng vậy mọi tập chứa *ab*

Item	Profit
a	40
b	0
c	-20
d	10
e	-30
f	30
g	20
h	-10

Ràng buộc đơn điệu

Ràng buộc	Đơn điệu
v [redacted]	yes
S [redacted]	yes
S [redacted]	no
min(S) [redacted]	yes
min(S) [redacted]	no
max(S) [redacted]	no
max(S) [redacted]	yes
count(S) [redacted]	no
count(S) [redacted]	yes
sum(S) [redacted] (a [redacted], a [redacted])	no
sum(S) [redacted] (a [redacted], a [redacted])	yes
range(S) [redacted]	no
range(S) [redacted]	yes
avg(S) [redacted], [redacted]	convertible
support(S) [redacted]	no
support(S) [redacted]	yes

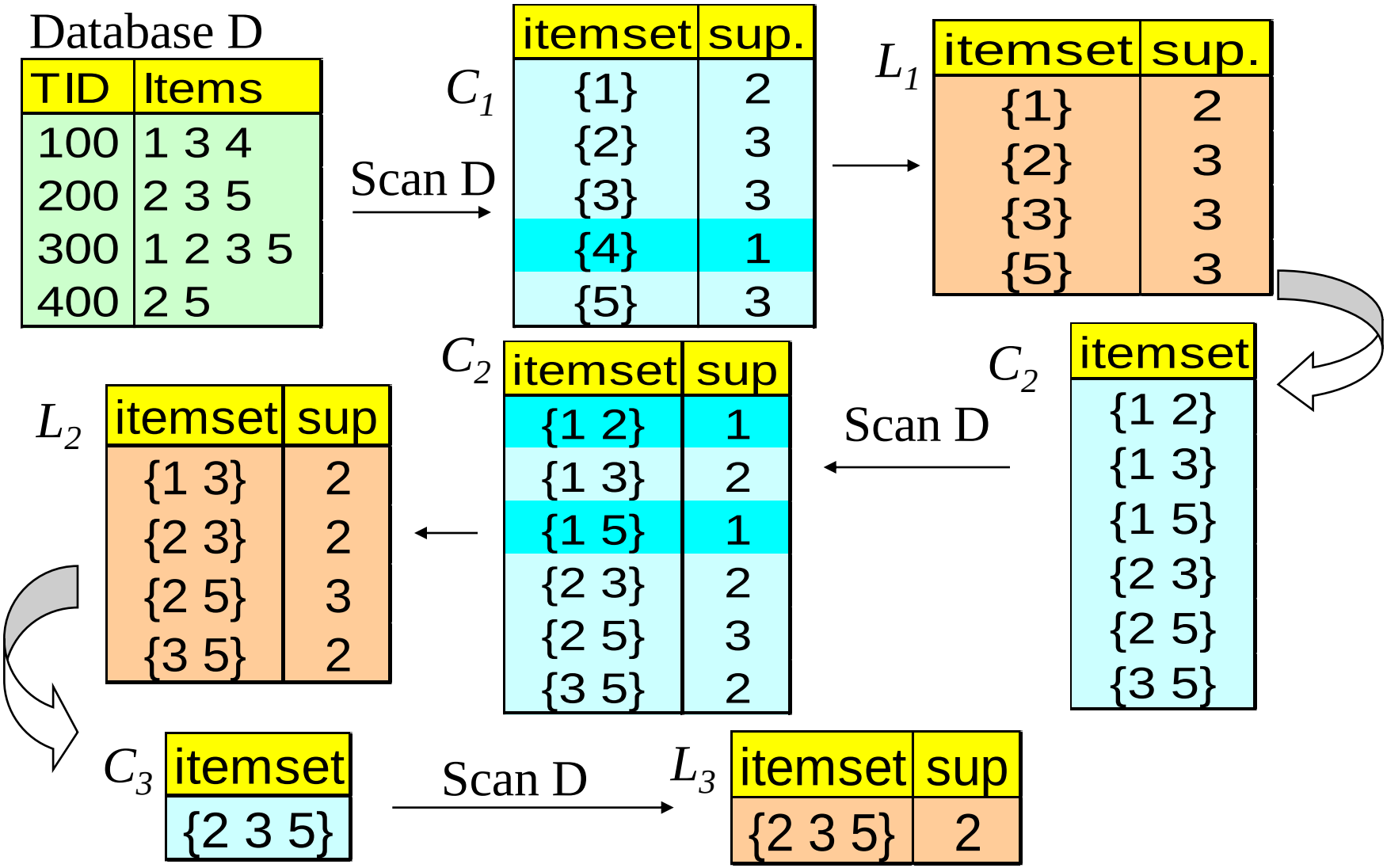
Tính cô đọng

- Tính cô đọng:
  - Cho A_1, là tập mục bảo đảm một ràng buộc cô đọng C, thì mọi S bảm đảm C là dựa trên A_1, chẳng hạn., S chưa một tập con thuộc A_1
  - Tư tưởng: Bỏ qua xem xét CSDL giao dịch, có chăng một tập mục *S bảo đảm ràng buộc* C có thể được xác định dựa theo việc chọn các mục
  - *min*(*S.Price*) [redacted] là cô đọng
  - *sum*(*S.Price*) [redacted] không cô đọng
- Tối ưu hóa: Nếu C là cô đọng có thể đẩy đếm trước

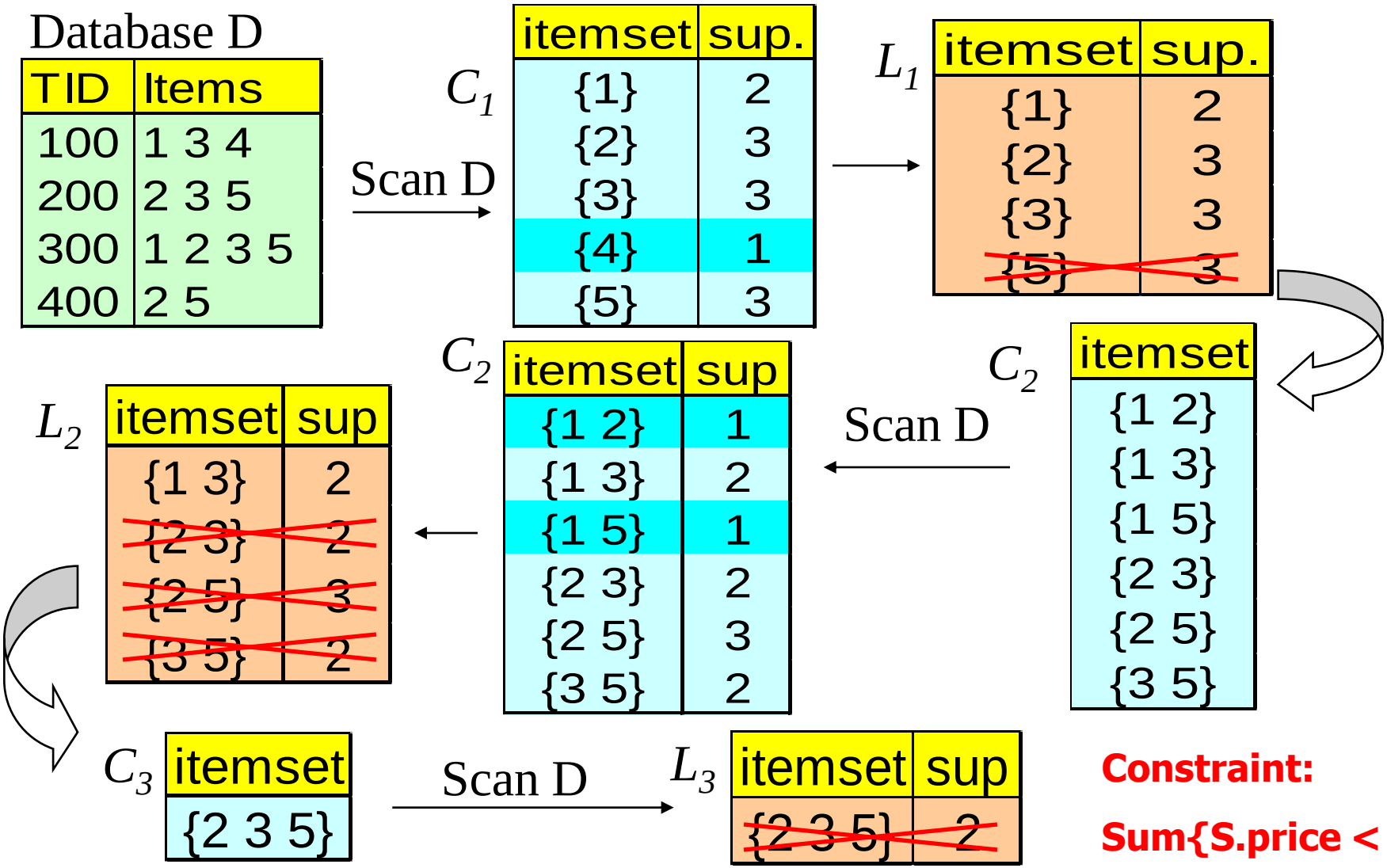
Ràng buộc cô đọng

Ràng buộc	Cô đọng
v [redacted]	yes
S [redacted]	yes
S [redacted]	yes
min(S) [redacted]	yes
min(S) [redacted]	yes
max(S) [redacted]	yes
max(S) [redacted]	yes
count(S) [redacted]	weakly
count(S) [redacted]	weakly
sum(S) [redacted] (a [redacted], a [redacted])	no
sum(S) [redacted] (a [redacted], a [redacted])	no
range(S) [redacted]	no
range(S) [redacted]	no
avg(S) [redacted], [redacted]	no
support(S) [redacted]	no
support(S) [redacted]	no

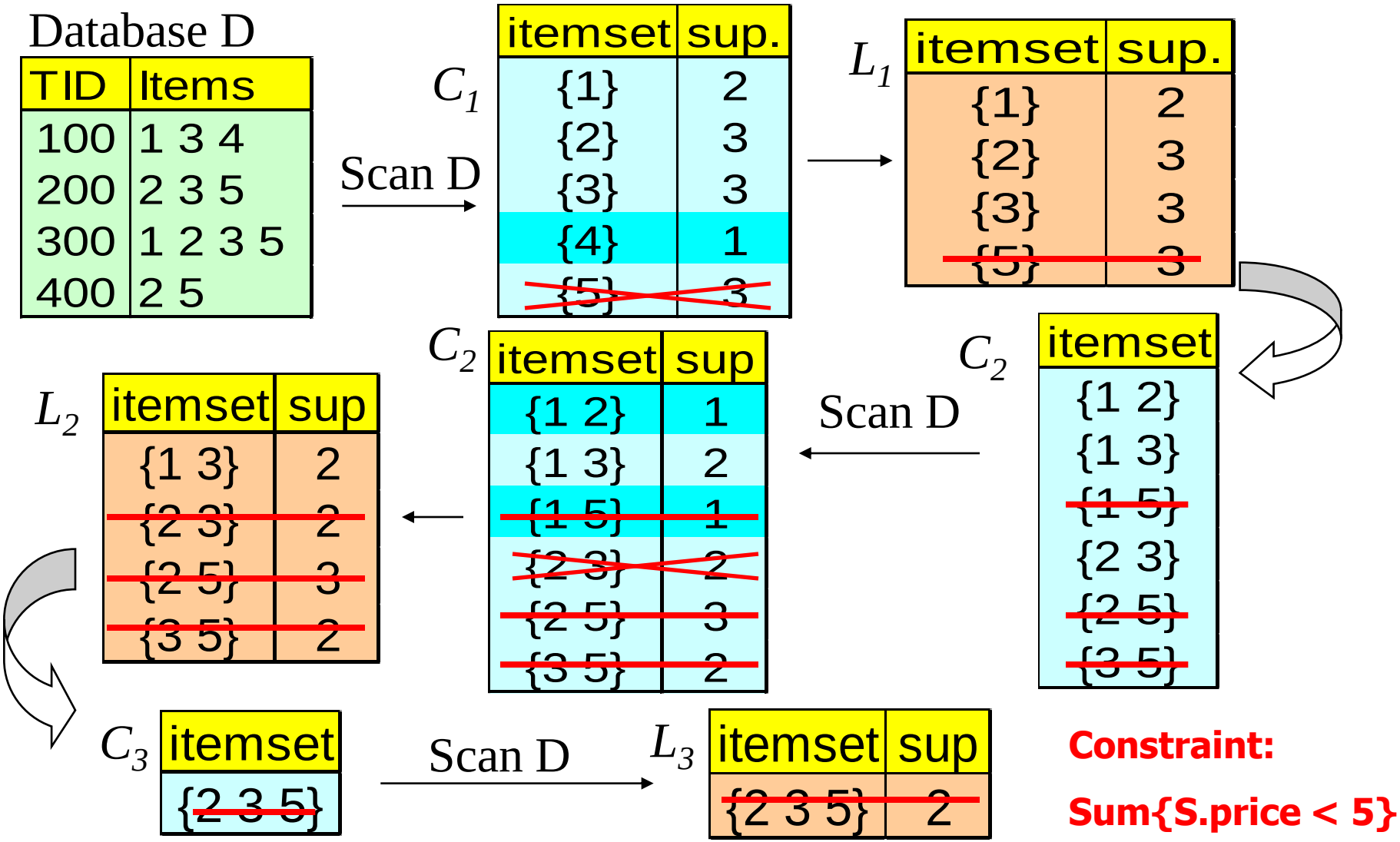
Thuật toán Apriori— Ví dụ



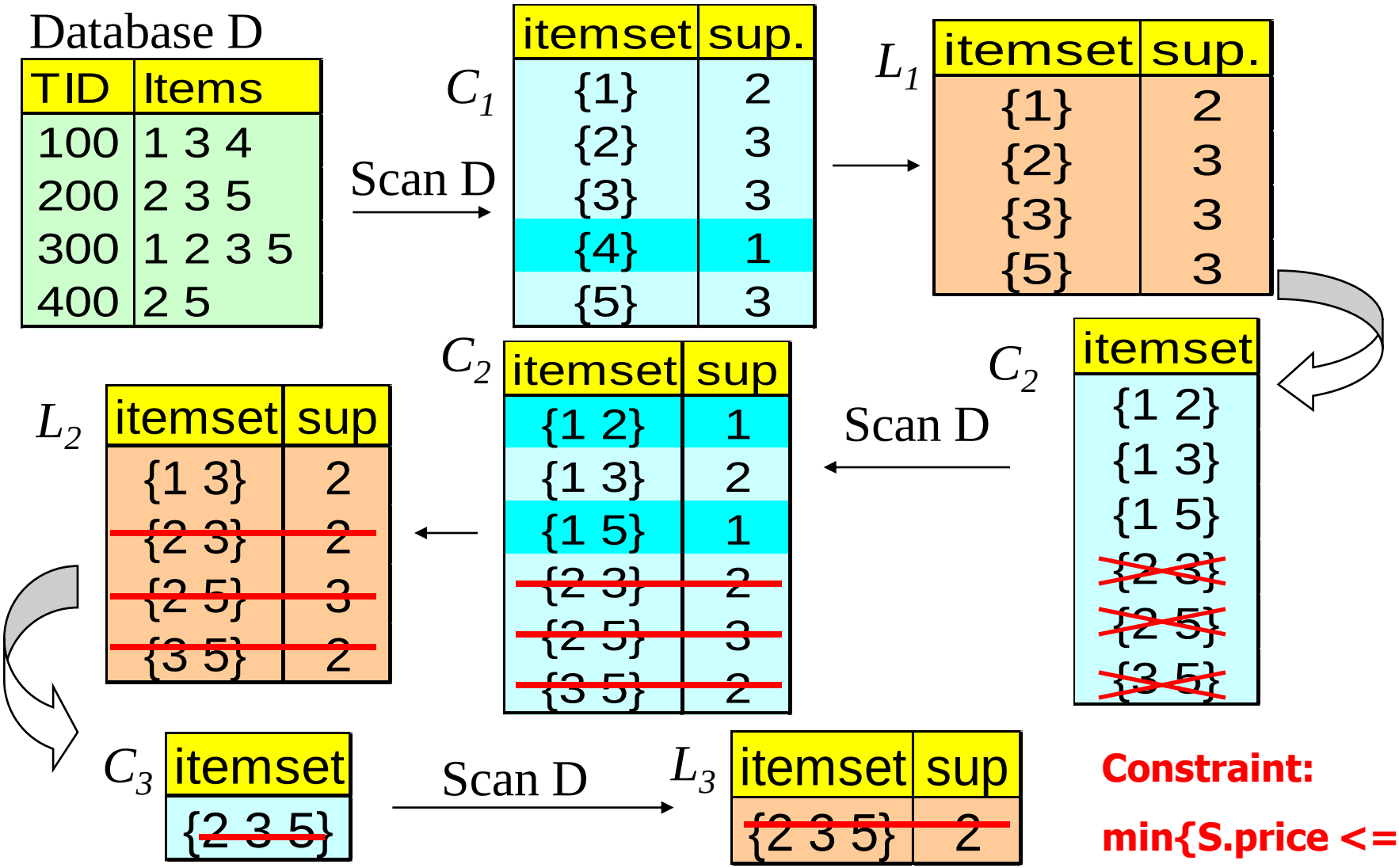
Thuật toán Naïve: Apriori +ràng buộc



Thuật toán Apriori ràng buộc: Đẩy ràng buộc chống đơn điệu xuống sâu



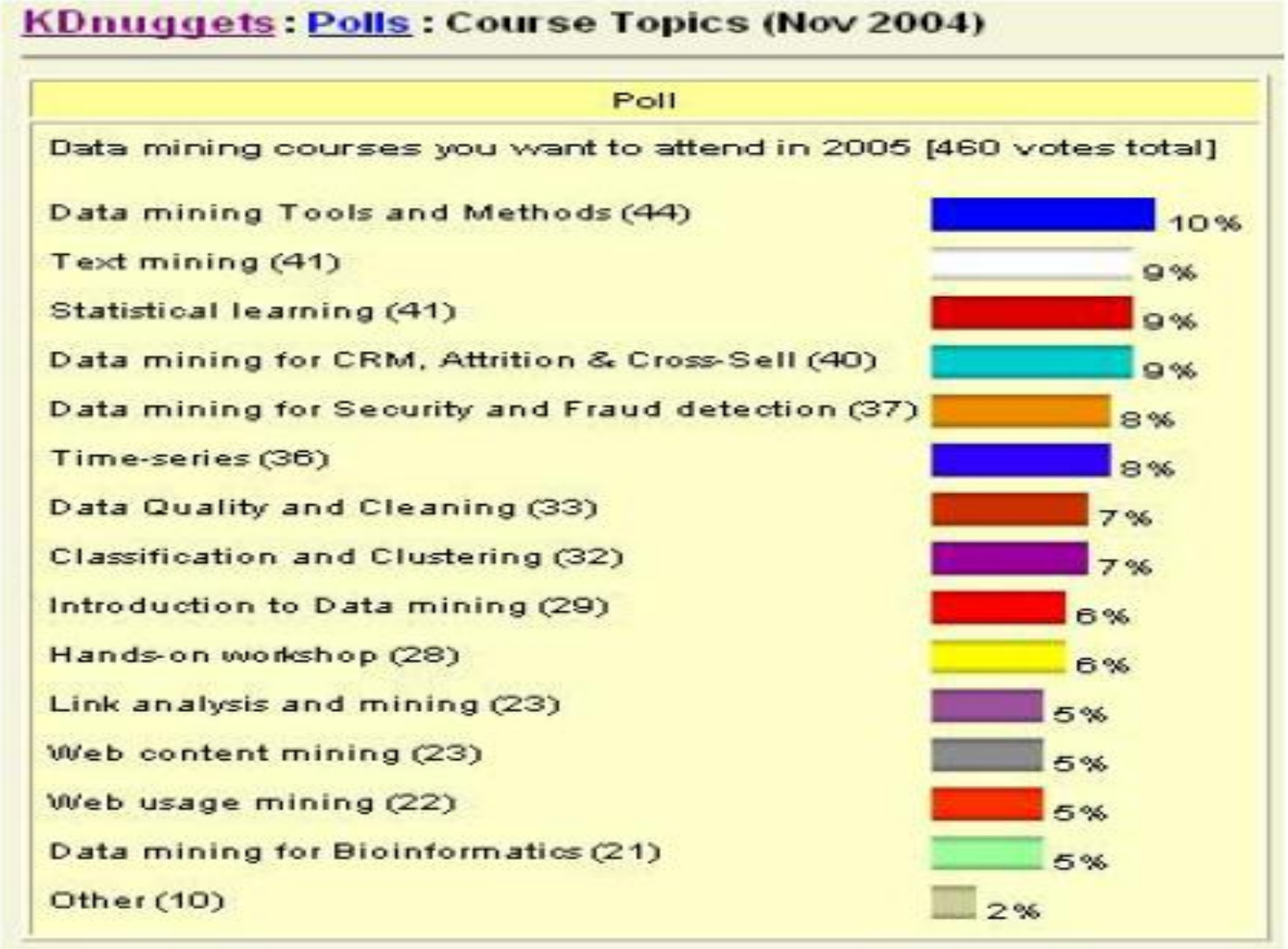
The Constrained Apriori Algorithm: Push a Succinct Constraint Deep



Chương 4: Khai phá luật kết hợp

- Khai phá luật kết hợp (Association rule)
- Các thuật toán khai phá vô hướng luật kết hợp (giá trị lôgic đơn chiều) trong CSDL giao dịch
- Khai phá kiểu đa dạng luật kết hợp/tương quan
- Khai phá kết hợp dựa theo ràng buộc
- Khai phá mẫu dãy
- Ứng dụng/mở rộng khai phá mẫu phổ biến

CSDL tuần tự và Phân tích mẫu tuần tự



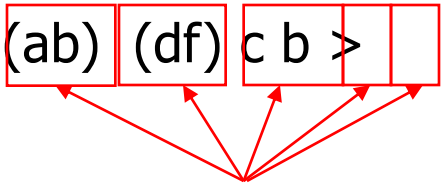
CSDL TT và PT MTT (2)

- ***CSDL giao dịch, CSDL chuỗi thời gian <> CSDL tuần tự***
- Mẫu PB <> mẫu TT (PB)
- Ứng dụng của KP Mẫu TT
  - Tuần tự mua của khách hàng:
    - *Đầu tiên mua máy tính, sau đó CD-ROM, và sau đó là máy ảnh số, trong vòng 3 tháng.*
  - Phẫu thuật y tế, thảm họa tự nhiên (động đất…), quá trình KH và kỹ nghệ, chứng khoán và thị trường….
  - Mẫu gọ điện thoại, dòng click tại Weblogs
  - Dãy DNA và cấu trúc gene

Khái niệm KP mẫu TT

- Cho một tập các dãy, tìm tập đầy đủ các dãy con phổ biến

dãy TT : < (ef) (ab) (df) c b >



CSDL dãy TT

SID	sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

Một phần tử chứa một tập mục.
Tập mục trong một phần tử là không thứ tự , và viết chúng theo ABC.

<a(bc)dc> là *dãy con* của <a(abc)(ac)d(cf)>

Cho độ hỗ trợ *min_sup* =2, <(ab)c> là mẫu tuần tự *sequential pattern*

BÀI GIẢNG NHẬP MÔN KHAI PHÁ DỮ LIỆU

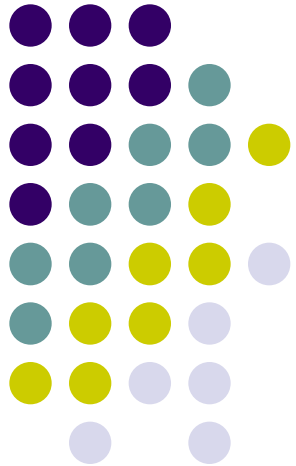
CHƯƠNG 5. PHÂN LỚP

PGS. TS. HÀ QUANG THỤY

HÀ NỘI 9-2011

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐẠI HỌC QUỐC GIA HÀ NỘI

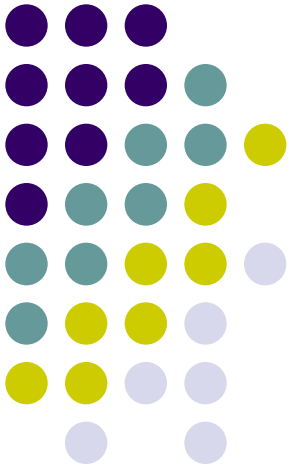


Nội dung

Giới thiệu phân lớp

Phân lớp học giám sát

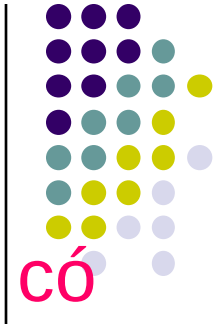
Phân lớp học bán giám sát



Bài toán phân lớp

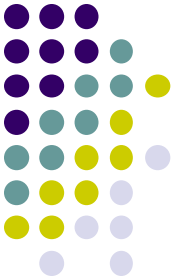
- Đầu vào
  - Tập dữ liệu D = $\{d_i\}$
  - Tập các lớp $C_1, C_2, \ldots, C_k$ mỗi dữ liệu d thuộc một lớp C_i
  - Tập ví dụ $D_{exam} = D_1 + D_2 + \ldots + D_k$ với $D_i = \{d \in D_{exam}$: d thuộc $C_i\}$
  - Tập ví dụ D_{exam} đại diện cho tập D
- Đầu ra
  - Mô hình phân lớp: ánh xạ từ D sang C
- Sử dụng mô hình
  - d $\in D \setminus D_{exam}$: xác định lớp của đối tượng d

Phân lớp: Quá trình hai pha



- Xây dựng mô hình: Tìm mô tả cho tập lớp đã có
  - Cho trước tập lớp C = {C_1, C_2, …, C_k}
  - Cho ánh xạ (chưa biết) từ miền D sang tập lớp C
  - Có tập ví dụ $D_{exam}=D_1+D_2+ …+ D_k$ với $D_i=\{d[redacted]_{exam}: d[redacted]_i\}$
    D_{exam} được gọi là tập ví dụ mẫu.
  - Xây dựng ánh xạ (mô hình) phân lớp trên: Dạy bộ phân lớp.
  - Mô hình: Luật phân lớp, cây quyết định, công thức toán học…
- Pha 1: Dạy bộ phân lớp
  - Tách D_{exam} thành D_{train} (2/3) + D_{test} (1/3). D_{train} và D_{test} "tính đại diện" cho miền ứng dụng
  - D_{train} : xây dựng mô hình phân lớp (xác định tham số mô hình)
  - D_{test} : đánh giá mô hình phân lớp (các độ đo hiệu quả)
  - Chọn mô hình có chất lượng nhất
- Pha 2: Sử dụng bộ phân lớp
  - d [redacted] \ D_{exam} : xác định lớp của d.

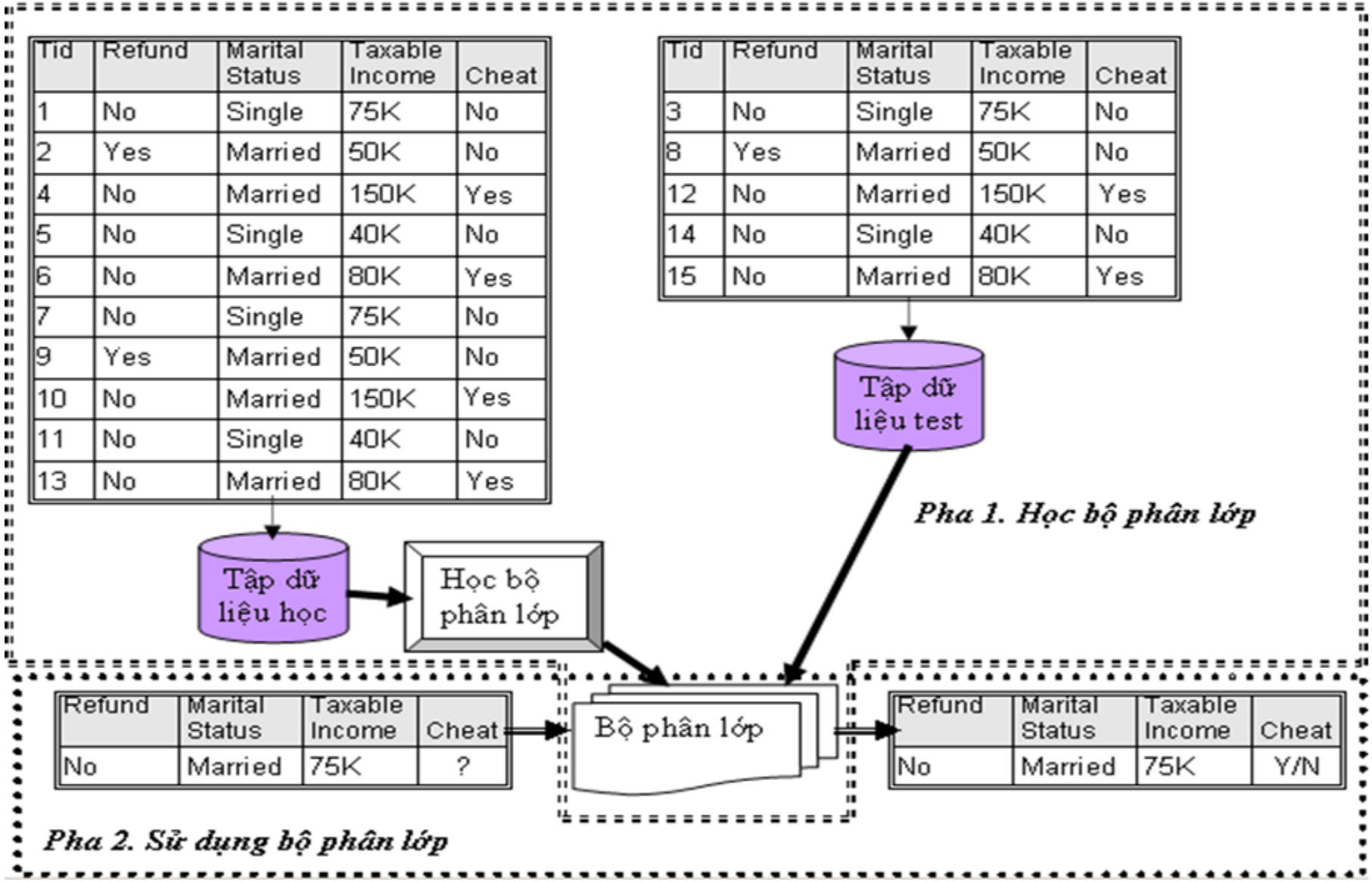
Ví dụ phân lớp: Bài toán cho vay



Tid	*Refund*	*Marital Status*	*Taxable Income*	*Cheat*
1	No	Single	75K	No
2	Yes	Married	50K	No
3	No	Single	75K	No
4	No	Married	150K	Yes
5	No	Single	40K	No
6	No	Married	80K	Yes
7	No	Single	75K	No
8	Yes	Married	50K	No
9	Yes	Married	50K	No
10	No	Married	150K	Yes
11	No	Single	40K	No
12	No	Married	150K	Yes
13	No	Married	80K	Yes
14	No	Single	40K	No
15	No	Married	80K	Yes

B

Phân lớp: Quá trình hai pha



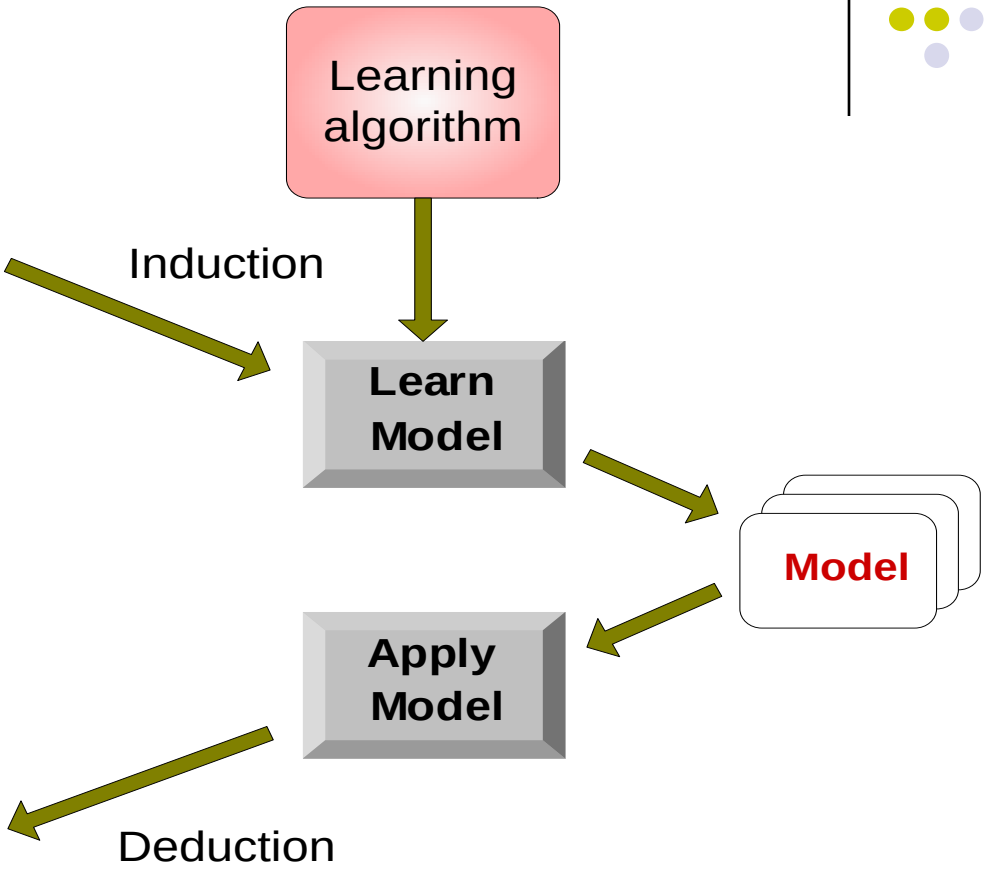
Phân lớp: Quá trình hai pha

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

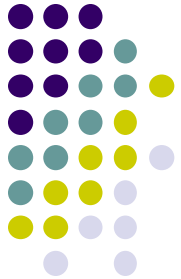
Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

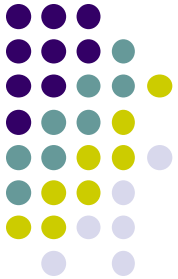


Các loại phân lớp



- Phân lớp nhị phân/ đa lớp:
  - |C|=***2****: phân lớp nhị phân.*
  - |C|>***2:*** *phân lớp đa lớp.*
- Phân lớp đơn nhãn/ đa nhãn:
  - Đơn nhãn: mỗi tài liệu được gán vào chính xác một lớp.
  - Đa nhãn: một tài liệu có thể được gán nhiều hơn một lớp.
  - Phân cấp: lớp này là cha/con của lớp kia

Các vấn đề đánh giá mô hình



- Các phương pháp đánh giá hiệu quả

  Câu hỏi: Làm thế nào để đánh giá được hiệu quả của một mô hình?

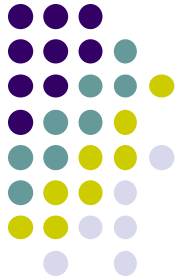
- Độ đo để đánh giá hiệu quả

  Câu hỏi: Làm thế nào để có được ước tính đáng tin cậy?

- Phương pháp so sánh mô hình

  Câu hỏi: Làm thế nào để so sánh hiệu quả tương đối giữa các mô hình có tính cạnh tranh?

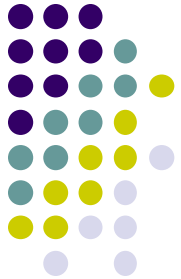
Đánh giá phân lớp nhị phân



- Theo dữ liệu test
- Giá trị thực: P dương / N âm; Giá trị qua phân lớp: T đúng/F sai. : còn gọi là *ma trận nhầm lẫn*
- Sử dụng các ký hiệu TP (true positives), TN (true negatives), FP (false positives), FN (false negatives)
  - TP: số ví dụ dương P mà thuật toán phân lớp cho giá trị đúng T
  - TN: số ví dụ âm N mà thuật toán phân lớp cho giá trị đúng T
  - FP: số ví dụ dương P mà thuật toán phân lớp cho giá trị sai F
  - FN: số ví dụ âm N mà thuật toán phân lớp cho giá trị sai F
- Độ hồi tưởng [redacted] độ chính xác [redacted] các độ đo F_1 và F[redacted]

[redacted] $\frac{TP}{TP \; [redacted] \; FP}$ [redacted] $\frac{TP}{TP \; [redacted] \; TN}$ $f_\beta = \frac{(\beta^2+1)\pi\rho}{\beta^2\pi+\rho}$ $f_1 = \frac{2\pi\rho}{\pi+\rho}$

Đánh giá phân lớp nhị phân



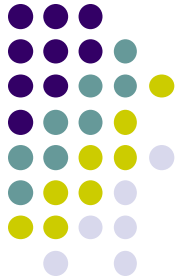
- Phương án khác đánh giá mô hình nhị phân theo độ chính xác (accuracy) và hệ số lỗi (Error rate)
- *Ma trận nhầm lẫn*

		Lớp dự báo	
		Lớp = 1	Lớp = 0
Lớp thực sự	Lớp = 1	f_{11}	f_{10}
	Lớp = 0	f_{01}	f_{00}

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$\text{Error rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

So sánh hai phương án



- Tập test có 9990 ví dụ lớp 0 và 10 ví dụ lớp 1. Kiểm thử: mô hình dự đoán cả 9999 ví dụ là lớp 0 và 1 ví dụ lớp 1 (chính xác: TP)
  - Theo phương án (precision, recall) có

    ■ 1/10=0.1; ■1/1=1; f_1 = 2*0.1/(0.1+1.0)= 0.18
  - Theo phương án (accurary, error rate) có

    accurary=0.9991; error rate = 9/10000 = 0.0009

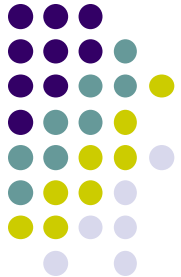
    Được coi là rất chính xác !
  - f_1 thể hiện việc đánh giá nhạy cảm với giá dữ liệu

Đánh giá phân lớp đa lớp

- Bài toán ban đầu: C gồm có k lớp
  - Đối với mỗi lớp C_i , cho thực hiện thuật toán với các dữ liệu thuộc D_{test} nhận được các đại lượng TP_i, TF_i, FP_i, FN_i (như bảng dưới đây)

Lớp C_i		Giá trị thực	
		Thuộc lớp C_i	Không thuộc lớp C_i
Giá trị qua bộ phân lớp đa lớp	Thuộc lớp C_i	TP_i	TN_i
	Không thuộc lớp C_i	FP_i	FN_i

Đánh giá phân lớp đa lớp



- Tương tự bộ phân lớp hai lớp (nhị phân)
  - Độ chính xác Pr_i của lớp C_i là tỷ lệ số ví dụ dương được thuật toán phân lớp cho giá trị đúng trên tổng số ví dụ được thuật toán phân lớp vào lớp C_i :

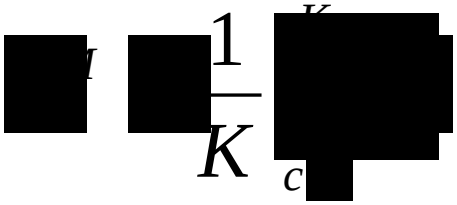
$$\mathrm{Pr}_i \quad \frac{TP_i}{TP_i \quad TN_i}$$

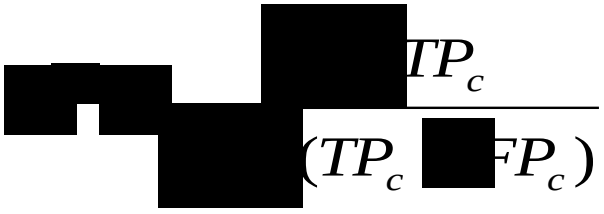
  - Độ hồi tưởng Re_i của lớp C_i là tỷ lệ số ví dụ dương được thuật toán phân lớp cho giá trị đúng trên tổng số ví dụ dương thực sự thuộc lớp C_i:

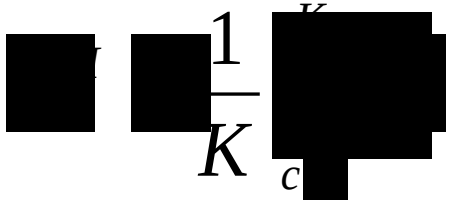
$$\mathrm{Re}_i \quad \frac{TP_i}{TP_i \quad FP_i}$$

Đánh giá phân lớp đa lớp

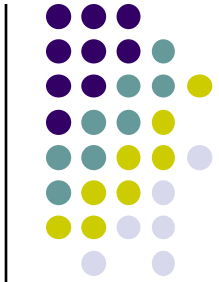
- Các giá trị [redacted] và [redacted]: độ hồi phục và độ chính xác đối với lớp C_i.
- Đánh giá theo các độ đo
  - vi trung bình-microaveraging (được ưa chuộng) [redacted] và [redacted]
  - trung bình lớn-macroaveraging [redacted] và [redacted]







Các kỹ thuật phân lớp



- Các phương pháp cây quyết định
  Decision Tree based Methods
- Các phương pháp dựa trên luật
  Rule-based Methods
- Các phương pháp Bayes «ngây thơ» và mạng tin cậy Bayes
  Naïve Bayes and Bayesian Belief Networks
- Các phương pháp máy vector hỗ trợ
  Support Vector Machines
- Lập luận dưa trên ghi nhớ
  Memory based reasoning
- Các phương pháp mạng nơron
  Neural Networks
- Một số phương pháp khác

Phân lớp cây quyết định

- Mô hình phân lớp là cây quyết định
- Cây quyết định
  - Gốc: tên thuộc tính; không có cung vào + không/một số cung ra
  - Nút trong: tên thuộc tính; có chính xác một cung vào và một số cung ra (gắn với điều kiện kiểm tra giá trị thuộc tính của nút)
  - Lá hoặc nút kết thúc: giá trị lớp; có chính xác một cung vào + không có cung ra.
  - Ví dụ: xem trang tiếp theo
- Xây dựng cây quyết định
  - Phương châm: “chia để trị”, “chia nhỏ và chế ngự”. Mỗi nút tương ứng với một tập các ví dụ học. Gốc: toàn bộ dữ liệu học
  - Một số thuật toán phổ biến: Hunt, họ ID3+C4.5+C5.x
- Sử dụng cây quyết định
  - Kiểm tra từ gốc theo các điều kiện

Ví dụ cây quyết định và sử dụng

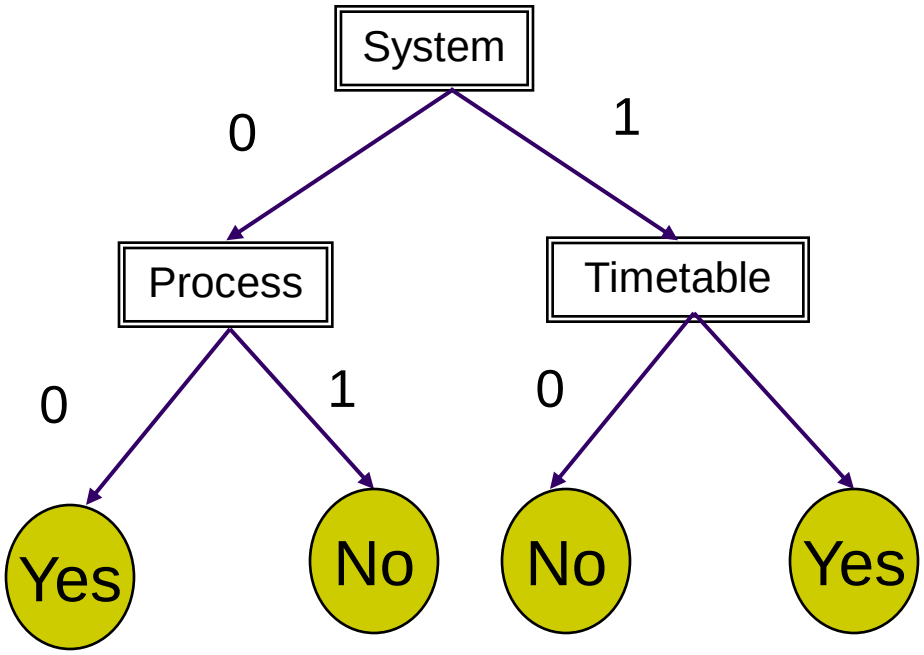
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Kết luận: Gán giá trị YES vào trường Cheat cho bản ghi

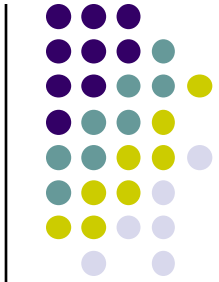
Ví dụ cây quyết định phân lớp văn bản

- Phân lớp văn bản vào lớp AI : trí tuệ nhân tạo
- Dựa vào các từ khóa có trong văn bản: System, Process, Timetable (Phân tích miền ứng dụng)



1. **If** System=0 **and** Process=0 **then** Class AI = Yes.
2. **If** System=0 **and** Process=1 **then** Class AI = No.
3. **If** System=1 **and** Timetable=1 **then** Class AI = Yes.
4. **If** System=1 **and** Timetable=0 **then** Class AI = No.

Dựng cây quyết định: thuật toán Hunt



- Thuật toán dựng cây quyết định sớm nhất, đệ quy theo nút của cây, bắt đầu từ gốc
- Input
  - Cho nút t trên cây quyết định đang được xem xét
  - Cho tập các ví dụ học D_t.
  - Cho tập nhãn lớp (giá trị lớp) y_1, y_1, … y_k. (k lớp)
- Output
  - Xác định nhãn nút t và các cung ra (nếu có) của t
- Nội dung

  1: Nếu mọi ví dụ trong D_t đều thuộc vào một lớp y thì nút t là một lá và được gán nhãn y.

  2: Nếu D_t chứa các ví dụ thuộc nhiều lớp thì

  2.1. Chọn 1 thuộc tính A để phân hoạch D_t và gán nhãn nút t là A

  2.2. Tạo phân hoạch D_t theo tập giá trị của A thành các tập con

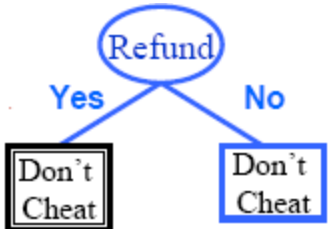
  2.3. Mỗi tập con theo phân hoạch của D_t tương ứng với một nút con u của t: cung nối t tới u là miền giá trị A theo phân hoạch, tập con nói trên được xem xét vơi u tiếp theo. Thực hiện thuật toán với từng nút con u của t.

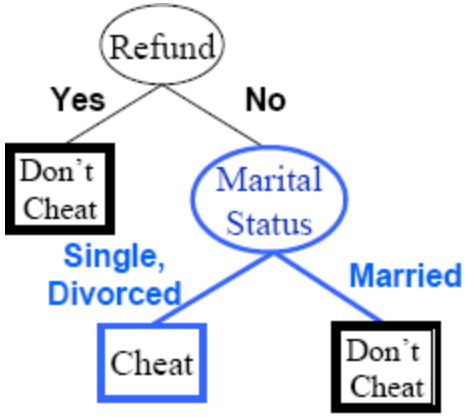
Ví dụ: thuật toán Hunt

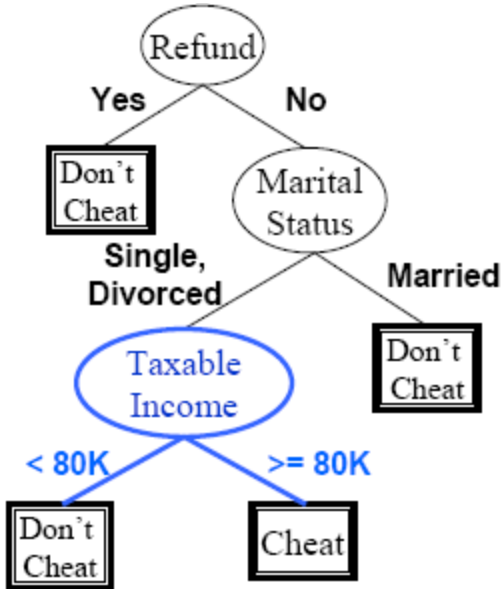
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Giải thích

- Xuất phát từ gốc với 10 bản ghi
- Thực hiện bước 2: chọn thuộc tính Refund có hai giá trị Yes, No. Chia thành hai tập gồm 3 bản ghi có Refund = Yes và 7 bản ghi có Refund = No
- Xét hai nút con của gốc từ trái sang phải. Nút trái có 3 bản ghi cùng thuộc lớp Cheat=No (Bước 1) nên là lá gán No (Don't cheat). Nút phải có 7 bản ghi có cả No và Yes nên áp dụng bước 2. Chọn thuộc tính Marital Status với phân hoạch Married và hai giá trị kia…







Thuật toán cây quyết định ID3

ID3 (*Examples, Target_attribute, Attributes*)

Ở đây: *Examples* là tập ví dụ học; *Target_attribute* là các thuộc tính đầu ra (lớp) cho cây quyết định dự đoán; *Attributes* là danh sách các thuộc tính khác tham gia trong quá trình học của cây quyết định. Kết quả thủ tục trả về cây quyết định phân lớp đúng các mẫu ví dụ đưa ra.

1. Tạo một nút gốc *Root* cho cây quyết định.
2. Nếu toàn bộ *Examples* đều là các ví dụ thuộc cùng một lớp thì trả lại cây *Root* một nút đơn với nhãn + (nếu các ví dụ thuộc lớp +) hoặc với nhãn - (nếu các ví dụ thuộc lớp -).
3. Nếu *Attributes* là rỗng thì trả lại cây *Root* một nút đơn với nhãn gán bằng giá trị phổ biến nhất của *Target_attribute* trong *Examples*.
4. Còn lại

   Begin

   4.1. Gán $A \leftarrow$ thuộc tính từ tập *Attributes* mà phân lớp tốt nhất tập *Examples*.

   4.2. Thuộc tính quyết định cho *Root* $\leftarrow A$

   4.3. Lặp với các giá trị có thể v_i của A,

   - Cộng thêm một nhánh cây con ở dưới *Root*, phù hợp với biểu thức kiểm tra $A = v_i$.
   - Đặt $Examples_{v_i}$ là một tập con của tập các ví dụ có giá trị v_i cho A
   - Nếu $Examples_{v_i}$ rỗng
     + Thì dưới mỗi nhánh mới thêm một nút lá với nhãn = giá trị phổ biến nhất của *Target_attribute* trong tập *Examples*.
     + Ngược lại thì dưới nhánh mới này thêm một cây con

       ID3($Examples_{v_i}$, Target_attribute, Attribute - {A}).

   End

5. Return *Root*.

Thuộc tính tốt nhất: Độ đo Gini

- Bước 4.1. chọn thuộc tính A tốt nhất gán cho nút t.
- Tồn tại một số độ đo: Gini, Information gain…
- Độ đo Gini
  - Đo tính hỗn tạp của một tập ví dụ mẫu
  - Công thức tính độ đo Gini cho nút t:

$$Gini(t) = \blacksquare \sum_{j} \blacksquare (j \mid t) \blacksquare$$

Trong đó p(j|t) là tần suất liên quan của lớp j tại nút t

  - Gini (t) lớn nhất = $1-1/n_c$ (với n_c là số các lớp tại nút t): khi các bản ghi tại t phân bố đều cho n_c lớp; tính hỗn tạp cao nhất, không có phân biệt giữa các lớp
  - Gini (t) nhỏ nhất = 0 khi tất cả các bản ghi thuộc một lớp duy nhất.
- Ví dụ: Bốn trường hợp

C1	**0**
C2	**6**
Gini=0.000	

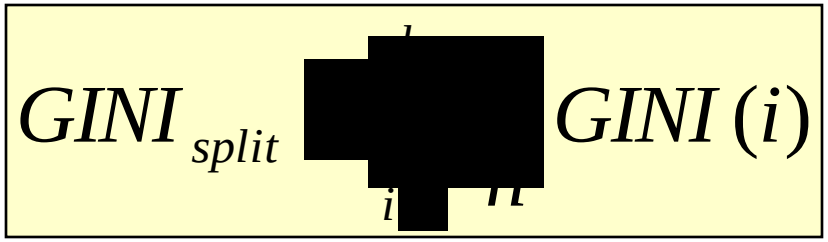
C1	**1**
C2	**5**
Gini=0.278	

C1	**2**
C2	**4**
Gini=0.444	

C1	**3**
C2	**3**
Gini=0.500	

Chia tập theo độ đo Gini

- Dùng trong các thuật toán CART, SLIQ, SPRINT
- Khi một nút t được phân hoạch thành k phần (k nút con của t) thì chất lượng của việc chia tính bằng



trong đó

- n là số bản ghi của tập bản ghi tại nút t,
- .n_i là số lượng bản ghi tại nút con I (của nút t).

Chia tập theo độ đo Gini: Ví dụ

- Tính toán GINI cho Refund (Yes, No), Marital Status (Single&Divorced, Married) và Taxable Income (<80K, [redacted]80K).
- Refund: 3/10 * (0) + 7/10 * (1-(3/7)^2 – (4/7)^2) = 7/10*(24/49) = 24/70
- Marital Status: 4/10 * 0 + 6/10 * (1- (3/6)^2 – (3/6)^2) = 6/10 * ½ = 3/10
- Taxable Income: thuộc tính liên tục cần chia khoảng (tồn tại một số phương pháp theo Gini, kết quả 2 thùng và 80K là mốc)

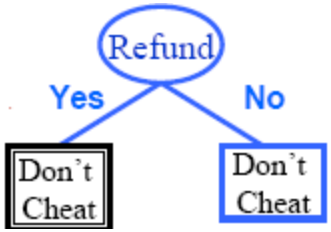
  3/10 * (0) + 7/10 * (1-(3/7)^2 – (4/7)^2) = 7/10*(24/49) = 24/70

Như vậy, Gini của Refund và Taxable Income bằng nhau (24/70) và lớn hơn Gini của Marital Status (3/10) nên chọn Refund cho gốc cây quyết định.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$GINI_{split} = \sum_{i} \frac{n_i}{n} GINI(i)$$

$$Gini(t) = 1 - \sum_{j} [p(j \mid t)]^2$$



Chọn thuộc tính: Information Gain

- Độ đo Information Gain
  - Thông tin thu được sau khi phân hoạch tập ví dụ
  - Dùng cho các thuật toán ID3, họ C4.5
- Entropy
  - Công thức tính entropy nút t:
    $$Entropy(t) = -\sum_j p(j|t)\log p(j|t)$$
    Trong đó p(j|t) là tần suất liên quan của lớp j tại nút t
    độ không đồng nhất tại nút t.
  - Entropy (t) lớn nhất = log (n_c) (với n_c là số các lớp tại nút t): khi các bản ghi tại t phân bố đều cho n_c lớp; tính hỗn tạp cao nhất, không có phân biệt giữa các lớp
  - Entropy (t) nhỏ nhất = 0 khi tất cả các bản ghi thuộc một lớp duy nhất.
  - Lấy loga cơ số 2 thay cho loga tự nhiên
- Tính toán entropy (t) cho một nút tương tự như Gini (t)

Chọn thuộc tính: Information Gain

- Độ đo Information Gain

$$Gain_{chia} = entropy(t) - \sum_{i=1}^{k} \frac{n_i}{n} entropy(i)$$

Trong đó, n là số lượng bản ghi tại nút t, k là số tập con trong phân hoạch, n_i là số lượng bản ghi trong tập con thứ i.

Độ đo giảm entropy sau khi phân hoạch: chọn thuộc tính làm cho Gain đạt lớn nhất.

C4.5 là một trong 10 thuật toán KPDL phổ biến nhất.

  - *Hạn chế*: Xu hướng chọn phân hoạch chia thành nhiều tập con

- Cải tiến

$$GainRATIO = \frac{Gain_{chia}}{SplitINFO} \qquad SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

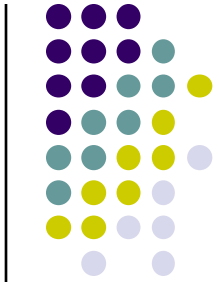
  - Dùng GainRatio để khắc phục xu hướng chọn phân hoạch nhiều tập con

- Áp dụng: Tự tiến hành

Phân lớp dựa trên luật

- Giới thiệu
  - Phân lớp các bản ghi dựa vào tập các luật "kiểu" if … then
- Luật
  - Luật: <điều kiện> [redacted]
    Trong đó:
    <điều kiện> là sự kết nối các thuộc tính (còn gọi là tiên đề/điều kiện của luật: LHS bên trái)
    y là nhãn lớp (còn gọi là kết quả của luật: RHS bên phải).
  - Ví dụ
    Refund = 'Yes" [redacted] heat = "No"
    (Refund = "No") [redacted] Marital Status = "Married") [redacted] heat = "No"
- Sử dụng luật
  - Một luật được gọi là "bảo đảm" thể hiện r (bản ghi) nếu các thuộc tính của r đáp ứng điều kiện của luật.
  - Khi đó, vế phải của luật cũng được áp dụng cho thể hiện.

Xây dựng luật phân lớp



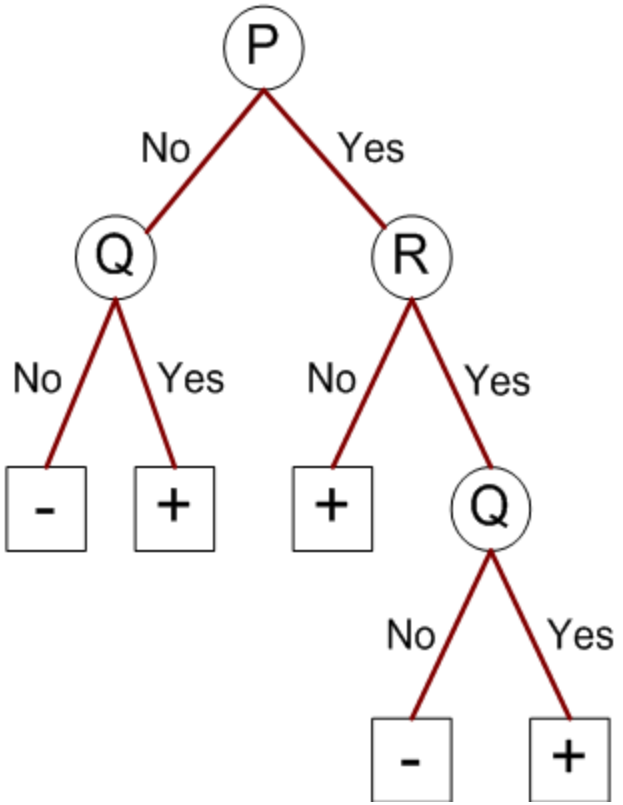
- Giới thiệu
  - Trực tiếp và gián tiếp
- Trực tiếp
  - Trích xuất luật trực tiếp từ dữ liệu
  - Ví dụ: RIPPER, CN2, Holte's 1R
  - Trích xuất luật trực tiếp từ dữ liệu
    1. Bắt đầu từ một tập rỗng
    2. Mở rộng luật bằng hàm Học_một_luật
    3. Xóa mọi bản ghi "bảo đảm" bởi luật vừa được học
    4. Lặp các bước 2-3 cho đến khi gặp điều kiện dừng
- Gián tiếp
  - Trích xuất luật từ mô hình phân lớp dữ liệu khác, chẳng hạn, mô hình cây quyết định, mô hình mạng nơ ron, …
  - Ví dụ:C4.5Rule

Mở rộng luật: một số phương án

- Sử dụng thống kê
  - Thống kê các đặc trưng cho ví dụ
  - Tìm đặc trưng điển hình cho từng lớp
- Thuật toán CN2
  - Khởi đầu bằng liên kết rỗng: {}
  - Bổ sung các liên kết làm cực tiểu entropy: {A}, {A, B}…
  - Xác định kết quả luật theo đa số của các bản ghi đảm bảo luật
- Thuật toán RIPPER
  - Bắt đầu từ một luật rỗng: {} [redacted]p
  - Bổ sung các liên kết làm cực đại lợi ích thông tin FAIL
  - R0: {} => lớp (luật khởi động)
  - R1: {A} => lớp (quy tắc sau khi thêm liên kết)
  - Gain (R0, R1) = t [log (p1 / (p1 + n1)) - log (p0 / (p0 + n0))]

  với t: số thể hiện đúng đảm bảo cả hai R0 và R1
    - p0: số thể hiện đúng được bảo đảm bởi R0
    - n0: số thể hiện sai được đảm bảo bởi R0
    - P1: số thể hiện đúng được bảo đảm bởi R1
    - n 1: số trường hợp sai được đảm bảo bởi R1

Luật phân lớp: từ cây quyết định



Tập luật

Liệt kê các đường đi từ gốc

r1: (P=No,Q=No) ==> -

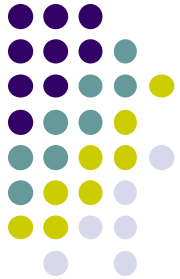
r2: (P=No,Q=Yes) ==> +

r3: (P=Yes,R=No) ==> +

r4: (P=Yes,R=Yes,Q=No) ==> -

r5: (P=Yes,R=Yes,Q=Yes) ==> +

Sinh luật gián tiếp: C4.5rules

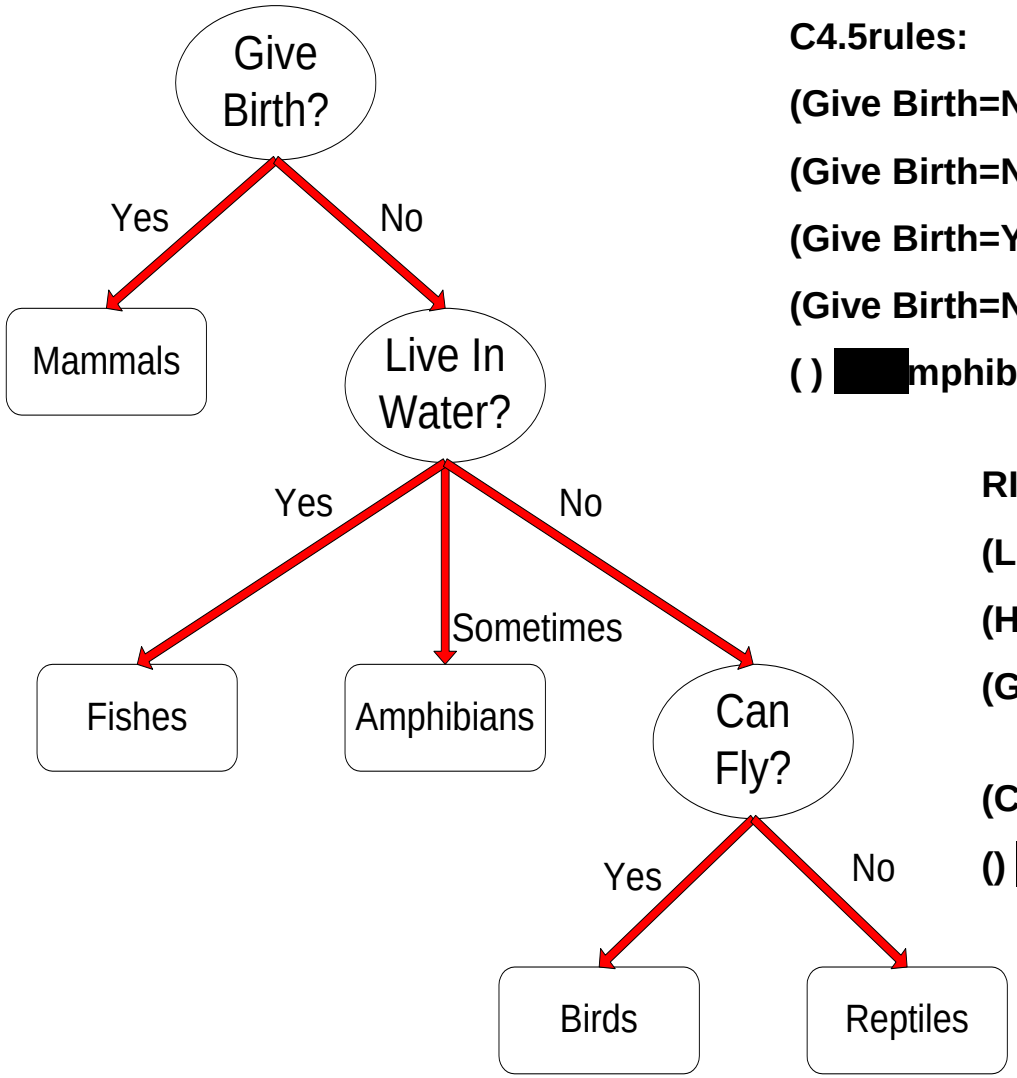


- Trích xuất luật từ cây quyết định chưa cắt tỉa
- Với mỗi luật, r: A → y
  - Xem xét luật thay thế r': A' → y, trong đó A' nhận được từ A bằng cách bỏ đi một liên kết
  - So sánh tỷ lệ lỗi r so với các r'
  - Loại bỏ các r' có lỗi thấp hơn r
  - Lặp lại cho đến khi không cải thiện được lỗi tổng thể
- Thay thế sắp xếp theo luật bằng sắp xếp theo tập con của luật (thứ tự lớp)
  - Mỗi tập con là một tập các luật với cùng một kết quả (lớp)
  - Tính toán độ dài mô tả của mỗi tập con
  - Độ dài mô tả = L(lỗi) + g* L(mô hình)
  - g : tham số đếm sự hiện diện của các thuộc tính dư thừa trong một tập luật (giá trị chuẩn, g=0.5)

C4.5rules: Ví dụ

Name	Give Birth	Lay Eggs	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	no	yes	mammals
python	no	yes	no	no	no	reptiles
salmon	no	yes	no	yes	no	fishes
whale	yes	no	no	yes	no	mammals
frog	no	yes	no	sometimes	yes	amphibians
komodo	no	yes	no	no	yes	reptiles
bat	yes	no	yes	no	yes	mammals
pigeon	no	yes	yes	no	yes	birds
cat	yes	no	no	no	yes	mammals
leopard shark	yes	no	no	yes	no	fishes
turtle	no	yes	no	sometimes	yes	reptiles
penguin	no	yes	no	sometimes	yes	birds
porcupine	yes	no	no	no	yes	mammals
eel	no	yes	no	yes	no	fishes
salamander	no	yes	no	sometimes	yes	amphibians
gila monster	no	yes	no	no	yes	reptiles
platypus	no	yes	no	no	yes	mammals
owl	no	yes	yes	no	yes	birds
dolphin	yes	no	no	yes	no	mammals
eagle	no	yes	yes	no	yes	birds

C4.5rules: Ví dụ



C4.5rules:

(Give Birth=No, Can Fly=Yes) [redacted]irds

(Give Birth=No, Live in Water=Yes) [redacted]shes

(Give Birth=Yes) [redacted]ammals

(Give Birth=No, Can Fly=No, Live in Water=No) [redacted]eptiles

( ) [redacted]mphibians

RIPPER:

(Live in Water=Yes) [redacted]shes

(Have Legs=No) [redacted]eptiles

(Give Birth=No, Can Fly=No, Live In Water=No) [redacted]eptiles

(Can Fly=Yes,Give Birth=No) [redacted]irds

() [redacted]ammals

Phân lớp Bayes

- Giới thiệu
  - Khung xác suất để xây dựng bộ phân lớp
  - Xác suất có điều kiện

Hai biến cố A và C

$$P(C \mid A) = \frac{P(A,C)}{P(A)}$$

$$P(A \mid C) = \frac{P(A,C)}{P(C)}$$

- Định lý Bayes:

$$P(c|x) = P(x|c).P(c)/P(x)$$

  - P(x) bằng nhau cho tất cả các lớp
  - Tìm c sao cho P(c|x) lớn nhất ⇔Tìm c sao cho P(x|c).P(c) lớn nhất
  - P(c): tần suất xuất hiện của các tài liệu thuộc lớp c
  - Vấn đề: làm thế nào để tính P(x|c)?

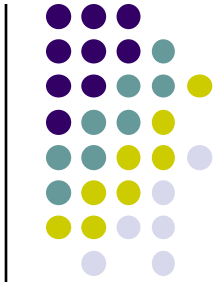
Định lý Bayes: Ví dụ

- Một bác sỹ biết
  - Bệnh nhân viêm màng não có triệu chứng cứng cổ S|M: 50%
  - Xác suất một bệnh nhân bị viêm màng não M là 1/50.000
  - Xác suất một bệnh nhân bị cứng cổ S là 1/20
- Một bệnh nhân bị cứng cổ hỏi xác suất anh/cô ta bị viêm màng não ?

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.5 / 50000}{1/20} = 0.0002$$

Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Introduction to Data Mining (Chapter 5: Classification: Alternative Techniques), *Addison Wesley*, 2005, http://www.cs.uu.nl/docs/vakken/dm/dmhc13.pdf

Phân lớp Bayes



- Các thuộc tính (bao gồm nhãn lớp) là các biến ngẫu nhiên.
- Cho một bản ghi với các giá trị thuộc tính (A_1, A_2, …, A_n)
  - Cần dự báo nhãn c
  - Tìm lớp c để cực đại xác suất $P(C|A_1, A_2, \ldots, A_n)$
- Có thể tính xác suất $P(C|A_1, A_2, \ldots, A_n)$ từ dữ liệu học?

Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Introduction to Data Mining (Chapter 5: Classification: Alternative Techniques), *Addison Wesley*, 2005, http://www.cs.uu.nl/docs/vakken/dm/dmhc13.pdf

Phân lớp Naïve Bayes

- Giả thiết Naïve Bayes:
  - giả thiết độc lập: xác suất xuất hiện của thuộc tính trong đối tượng độc lập với ngữ cảnh và vị trí của nó trong đối tượng:

$$p(c \mid x, \blacksquare\blacksquare\blacksquare) \propto \sum_{T \text{ in } \blacksquare} p(c \mid x, T)\, p(T \mid \overline{x})$$

$$\mathrm{P}(\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{k} \mid \mathrm{C}) = \mathrm{P}(\mathrm{x}_1 \mid \mathrm{C}) \cdot \ldots \cdot \mathrm{P}(\mathrm{x}_\mathrm{k} \mid \mathrm{C})$$

Phân lớp văn bản Naïve Bayes

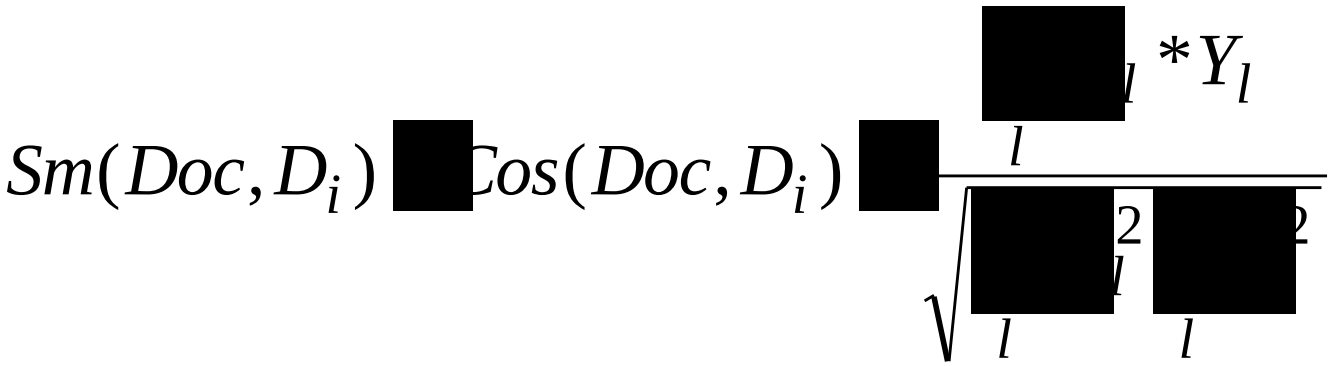
- Cho
  - Tập ví dụ $D_{exam} = D_{learn} + D_{test}$
  - Tập từ vựng $V = \{f_1, f_2, \ldots, f_{||V||}\}$
  - Tập lớp C= {C1, C2, …, Cn} với mỗi C_i một ngưỡng [redacted] > 0
- Tính xác suất tiên nghiệm
  - Trên tập ví dụ học D_{learn}
  - $p(C_i) = M_i/M$, $M = ||D_{learn}||$, $M_i = ||\text{Doc } [\text{redacted}]_{learn} / \text{Doc } [\text{redacted}]_i||$
  - Xác suất một đặc trưng (từ) f_j thuộc lớp C:

$$P(f_j \mid C) = \frac{1 + [\text{redacted}]F(f_j, C)}{|V| + [\text{redacted}]_i F(j_l, C_i)}$$

- Cho tài liệu Doc mới
  - Tính xác suất hậu nghiệm
  - Nếu P(C|Doc) > [redacted] thì Doc [redacted]C!

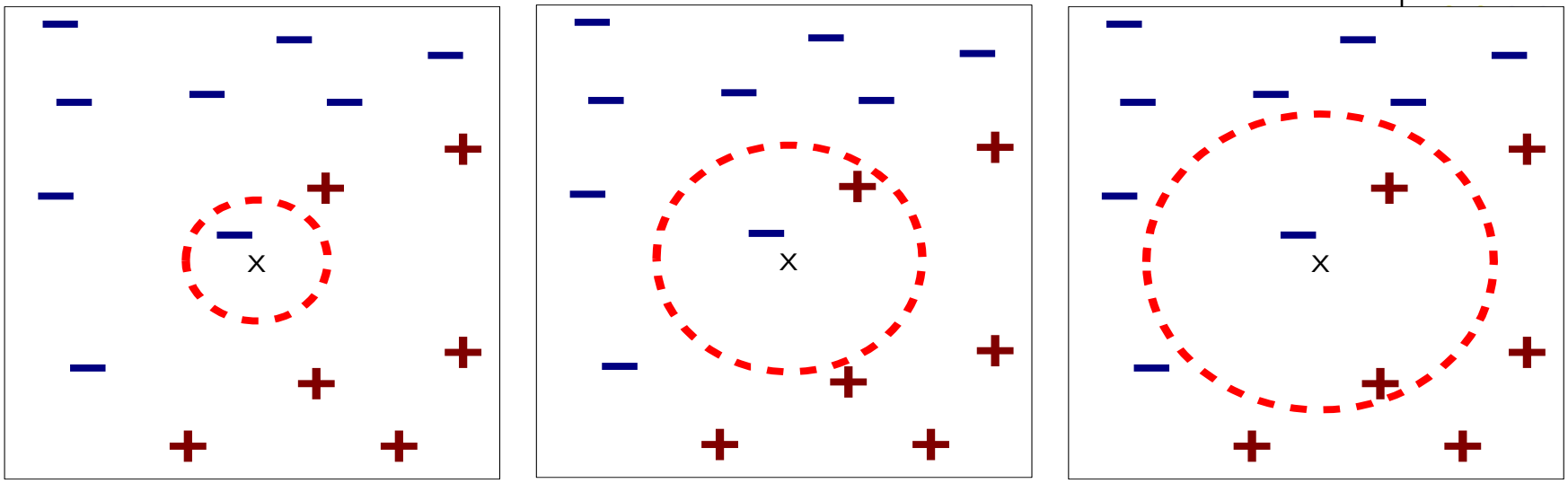
$$P(C \mid Doc) = [\text{redacted}] \frac{p(C) * [\text{redacted}]_{F_j} (F_j \mid C)^{TF(F_j, Doc)}}{[\text{redacted}]_i C_i) * [\text{redacted}]_{F_j} (F_j \mid C_i)^{TF(F_j, Doc)}}$$

Phân lớp k-NN



- Cho trước
  - Một tập D các đối tượng dữ liệu biểu diễn bản ghi các đặc trưng
  - Một đo đo khoảng cách (Ơcơlit) hoặc tương tự (như trên)
  - Một số k > 0 (láng giềng gần nhất
- Phân lớp đối tượng mới Doc được biểu diễn
  - Tính khoảng cách (độ tương tự) từ Doc tới tất cả dữ liệu thuộc D
  - Tìm k dữ liệu thuộc D gần Doc nhất
  - Dùng nhãn lớp của k-láng giềng gần nhất để xác định nhãn lớp của Doc: nhãn nhiều nhất trong k-láng giềng gần nhất

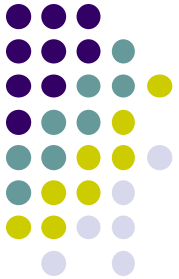
Phân lớp k-NN: Ví dụ



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

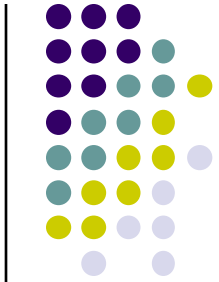
- Ba trường hợp như hình vẽ
  - 1-NN: Chọn lớp “-”: láng giềng có nhãn “-” là nhiều nhất
  - 2-NN: Chọn lớp “-”: hai nhãn có số lượng như nhau, chọn nhãn có tổng khoảng cách gần nhất
  - 3-NN: Chọn lớp “+”: láng giềng có nhãn “+” là nhiều nhất

Thuật toán SVM



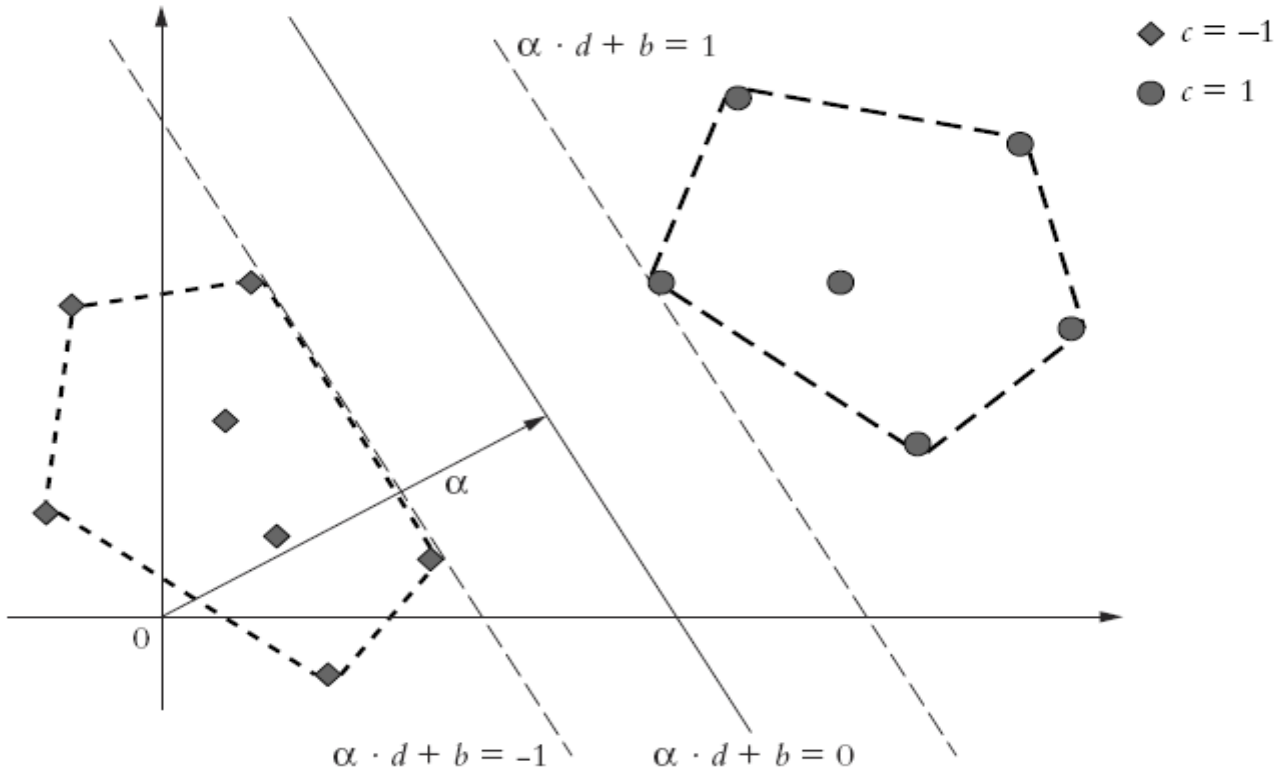
- Thuật toán máy vector hỗ trợ (Support Vector Machine – SVM): được Corters và Vapnik giới thiệu vào năm 1995.
- SVM rất hiệu quả để giải quyết các bài toán với dữ liệu có số chiều lớn (như các vector biểu diễn văn bản).

Thuật toán SVM



- Tập dữ liệu học: D= $\{(X_i, C_i), i=1,\ldots n\}$
  - $C_i \in \{-1,1\}$ xác định dữ liệu dương hay âm
- Tìm một siêu phẳng: $\alpha_{SVM} \mathbf{.d + b}$ phân chia dữ liệu thành hai miền.
- Phân lớp một tài liệu mới: xác định dấu của
  - $f(d) = \alpha_{SVM} \mathbf{.d + b}$
  - Thuộc lớp dương nếu $f(d) > 0$
  - Thuộc lớp âm nếu $f(d) < 0$

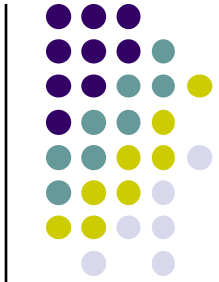
Thuật toán SVM



Thuật toán SVM

- Nếu dữ liệu học là tách rời tuyến tính:
  - Cực tiểu: $\frac{1}{2}$ [redacted]
  - Thỏa mãn: c_i [redacted] (2)
- Nếu dữ liệu học không tách rời tuyến tính: thêm biến {ξ1… ξn}:
  - Cực tiểu:
  - Thỏa mãn: $\frac{1}{2}$ [redacted] $i=1$ c_i [redacted] (4)

Phân lớp bán giám sát



- Giới thiệu phân lớp bán giám sát
  - Khái niệm sơ bộ
  - Tại sao học bán giám sát
- Nội dung phân lớp bán giám sát
  - Một số cách tiếp cận cơ bản
  - Các phương án học bán giám sát phân lớp
- Phân lớp bán giám sát trong NLP

Học bán giám sát:Tài liệu tham khảo

1. Xiaojin Zhu (2006 ***). Semi-Supervised Learning Literature Survey, 1-2006. (Xiao Zhu [1])
   http://www.cs.wisc.edu/~jerryzhu/pub/ssl survey.pdf

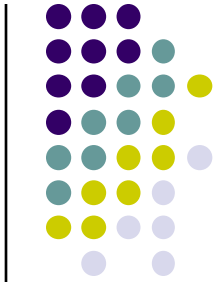
Zhou, D., Huang, J., & Scholkopf, B. (2005). Learning from labeled and unlabeled data on a directed graph. *ICML05, 22nd International Conference on Machine Learning*. Bonn, Germany.

Zhou, Z.-H., & Li, M. (2005). Semi-supervised regression with co-training. *InternationalJoint Conference on Artificial Intelligence (IJCAI)*.

Zhu, X. (2005). *Semi-supervised learning with graphs*. Doctoral dissertation, Carnegie Mellon University (mã số CMU-LTI-05-192).

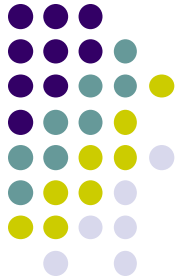
1. Olivier Chapelle, Mingmin Chi, Alexander Zien (2006) A Continuation Method for Semi-Supervised SVMs. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.
   và các tài liệu khác

Sơ bộ về học bán giám sát



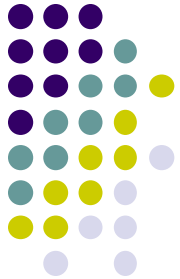
- Học bán giám sát là gì ? Xiao Zhu [1] FQA
  - Học giám sát: tập ví dụ học đã được gán nhãn (ví dụ gắn nhãn) là tập các cặp (tập thuộc tính, nhãn)
  - ví dụ gắn nhãn
    - Thủ công: khó khăn [redacted]huyên gia [redacted]n thời gian, tiền
    - Tự động: như tự động sinh corpus song hiệu quả chưa cao
  - ví dụ chưa gắn nhãn
    - Dễ thu thập [redacted]hiều
      - xử lý tiếng nói: bài nói nhiều, xây dựng tài nguyên đòi hỏi công phu
      - xử lý văn bản: trang web vô cùng lớn, ngày càng được mở rộng
    - Có sẵn [redacted]ó điều kiện tiến hành tự động gắn nhãn
  - Học bán giám sát: dùng cả ví dụ có nhãn và ví dụ chưa gắn nhãn
    - Tạo ra bộ phân lớp tốt hơn so với chỉ dùng học giám sát: học bán giám sát đòi hỏi điều kiện về dung lượng khối lượng

Cơ sở của học bán giám sát



- Biểu diễn dữ liệu chưa mô tả hết ánh xạ gán nhãn trên dữ liệu
  - chẳng hạn, nghịch lý “hiệu quả như nhau” trong biểu diễn văn bản
- Ánh xạ gán nhãn có liên quan mô hình dữ liệu (mô hình / đặc trưng/ nhân / hàm tương tự) mô hình đã có theo tự nhiên hoặc giả thiết dữ liệu tuân theo.

Hiệu lực của học bán giám sát

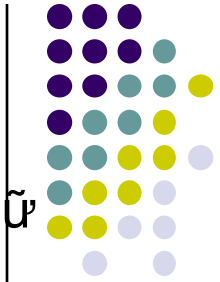


- Dữ liệu chưa nhãn không luôn luôn hiệu quả
  - Nếu giả thiết mô hình không phù hợp [redacted]ảm hiệu quả
  - Một số phương pháp cần điều kiện về miền quyết định: tránh miền có mật độ cao:
    - Transductive SVM (máy hỗ trợ vector lan truyền)
    - Information Regularization (quy tắc hóa thông tin)
    - mô hình quá trinh Gauxơ với nhiễu phân lớp bằng không
    - phương pháp dựa theo đồ thị với trọng số cạnh là khoảng cách
  - "Tồi" khi dùng phương pháp này song lại "tốt" khi dùng phương pháp khác

Phương pháp học bán giám sát

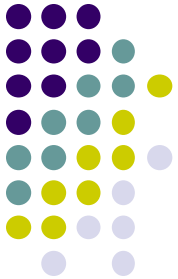
- Các phương pháp học bán giám sát điển hình
  - EM với mô hình trộn sinh
  - Self-training
  - Co-training
  - TSVM
  - Dựa trên đồ thị
  - ...
- So sánh các phương pháp
  - Đòi hỏi các giả thiết mô hình mạnh. Giả thiết mô hình phù hợp cấu trúc dữ liệu: khó kiểm nghiệm
  - Một số định hướng lựa chọn
    - Lớp [redacted]hân cụm tốt: dùng EM với mô hình sinh trộn.
    - Đặc trưng phân thành hai phần riêng rẽ: co-training
    - Nếu hai điểm tương tự hướng tới một lớp: dựa trên đồ thị
    - Đã sử dụng SVM thì mở rộng TSVM
    - Khó nâng cấp học giám sát đã có: dùng self-traning
    -

Phương pháp học bán giám sát



- Dùng dữ liệu chưa gán nhãn
  - Hoặc biến dạng hoặc thay đổi thứ tự giả thiết thu nhờ chỉ dữ liệu có nhãn
  - Mô tả chung
    - Giả thiết dưới dạng p(y|x) còn dữ liệu chưa có nhãn p(x)
    - Mô hình sinh có tham số chung phân bố kết nối p(x, y)
    - Mô hình trộn với EM mở rộng thêm self-training
    - Nhiều phương pháp là phân biệt: TSVM, quy tắc hóa thông tin, quá trình Gauxơ, dựa theo đồ thị
  - Có dữ liệu không nhãn: nhận được xác suất p(x)
- Phân biệt "học lan truyền" với "học bán giám sát"
  - Đa dạng về cách gọi. Hạn chế bài toán phân lớp.
  - "Bán giám sát"
    - dùng ví dụ có / không có nhãn,
    - "học dữ liệu nhãn/không nhãn,
    - "học dữ liệu phân lớp/có nhãn bộ phận".
    - Có cả lan truyền hoặc quy nạp.
  - Lan truyền để thu hẹp lại cho quy nạp: học chỉ dữ liệu sẵn. Quy nạp: có thể liên quan tới dữ liệu chưa có.

Mô hình sinh: Thuật toán EM



- Sơ bộ
  - Mô hình sớm nhất, phát triển lâu nhất
  - Mô hình có dạng p(x,y) = p(y)*p(x|y)
  - Với số lượng nhiều dữ liệu chưa nhãn cho P(x|y) mô hình trộn đồng nhất. Miền tài liệu được phân thành các thành phần,
  - Lý tưởng hóa tính "Đồng nhất": chỉ cần một đối tượng có nhãn cho mỗi thành phần
- Tính đồng nhất
  - Là tính chất cần có của mô hình
  - Cho họ phân bố {p[redacted]} là đồng nhất nếu [redacted] thì p[redacted] cho tới một hoán đổi vị trí các thành phần [redacted]nh khả tách của phân bố tới các thành phần

Mô hình sinh: Thuật toán EM

- Tính xác thực của mô hình
  - Giả thiết mô hình trộn là chính xác [redacted]ữ liệu không nhãn sẽ làm tăng độ chính xác phân lớp
  - Chú ý cấu trúc tốt mô hình trộn: nếu tiêu đề được chia thành các tiêu đề con thì nên mô hình hóa thành đa chiều thay cho đơn chiều
- Cực đại EM địa phương
  - Miền áp dụng
    - Khi mô hình trộn chính xác
  - Ký hiệu
    - D: tập ví dụ đã có (có nhẫn /chưa có nhãn)
    - D^K: tập ví dụ có nhãn trong D ($|D^K| << |D|$)

Mô hình sinh: Thuật toán EM

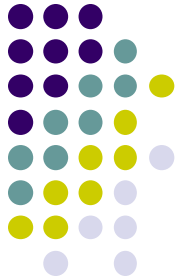
- Nội dung thuật toán

1: Cố định tập tài liệu không nhãn D^U [redacted] $\setminus D^K$ dùng trong E-bước và M-bước
2: dùng D^K xây dựng mô hình ban đầu [redacted]
3: **for** i = 0, 1, 2, . . . cho đến khi kết quả đảm bảo **do**
4: **for** mỗi tài liệu d [redacted] D^U **do**
5: E-bước: dùng phân lớp Bayes thứ nhất xác định P($c|d$,[redacted]
6: **end for**
7: **for** mỗi lớp c và từ khóa t **do**
8: M-bước: xác định [redacted] dùng công thức (*) để xây dựng mô hình i+1
9: **end for**
10: **end for**

$$\mathrm{P}(d|c) = \mathrm{P}(L = \ell_d|c)\binom{\ell_d}{\{n(d,t)\}}\prod_{t\in d}\theta_t^{n(d,t)}$$

$$\theta_{c,t} = \frac{1+\sum_{d\in D}\mathrm{P}(c|d)\,n(d,t)}{|W|+\sum_{d\in D}\sum_{\tau}\mathrm{P}(c|d)\,n(d,\tau)} \qquad \mathrm{P}(c) = \frac{1}{|D|}\sum_{d\in D}\mathrm{P}(c|d)$$

Mô hình sinh: Thuật toán EM

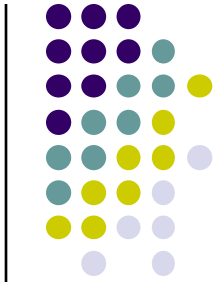


- Một số vấn đề với EM
  - Phạm vi áp dụng: mô hình trộn chính xác
  - Nếu cực trị địa phương khác xa cực trị toàn cục thì khai thác dữ liệu không nhãn không hiệu quả
  - "Kết quả đảm bảo yêu cầu": đánh giá theo các độ đo hồi tưởng, chính xác, F_1...
  - Một số vấn đề khác cần lưu ý:
    - Thuật toán nhân là Bayes naive: có thể chọn thuật toán cơ bản khác
    - Chọn điểm bắt đầu bằng học tích cực

Mô hình sinh: Thuật toán khác

- Phân cụm - và - Nhãn
  - Sử dụng phân cụm cho toàn bộ ví dụ
    - cả dữ liệu có nhãn và không có nhãn
    - dành tập D_{test} để đánh giá
  - Độ chính xác phân cụm cao
    - Mô hình phân cụm phù hợp dữ liệu
    - Nhãn cụm (nhãn dữ liệu có nhãn) làm nhãn dữ liệu khác
- Phương pháp nhân Fisher cho học phân biệt
  - Phương pháp nhân là một phương pháp điển hình
  - Nhân là gốc của mô hình sinh
  - Các ví dụ có nhãn được chuyển đổi thành vector Fisher để phân lớp

Self-Training



- Giới thiệu
  - Là kỹ thuật phổ biến trong SSL
    - EM địa phương là dạng đặc biệt của seft-training
- Nội dung

  Gọi

  L : Tập các dữ liệu gán nhãn.

  U : Tập các dữ liệu chưa gán nhãn

  Lặp (cho đến khi U = [redacted]

  Huấn luyện bộ phân lớp giám sát h trên tập L

  Sử dụng h để phân lớp dữ liệu trong tập U

  Tìm tập con U' [redacted]U có độ tin cậy cao nhất:

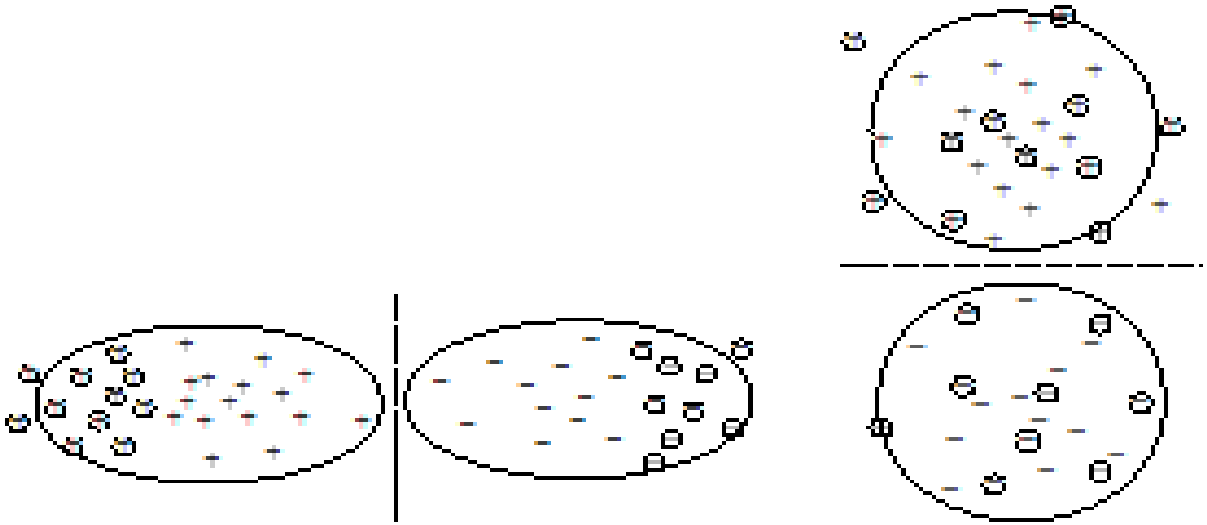
  L + U' [redacted]

  U – U' [redacted]

  Vấn đề tập U' có "độ tin cậy cao nhất"
- Thủ tục "bootstrapping"
- Thường được áp dụng cho các bài toán NLP

Co-Training

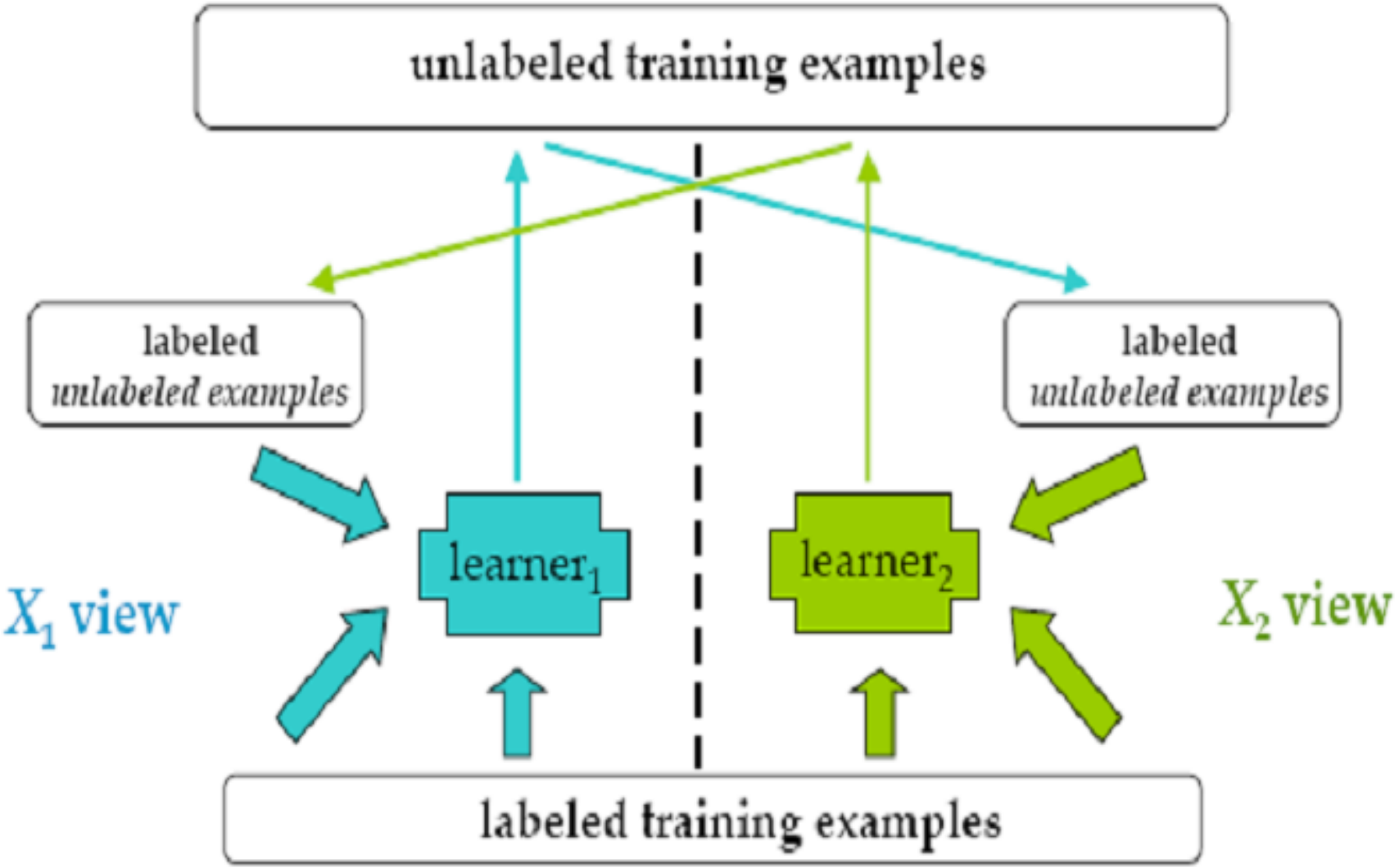
- Tư tưởng
  - Một dữ liệu có hai khung nhìn
  - Ví dụ, các trang web
    - Nội dung văn bản
    - Tiêu đề văn bản



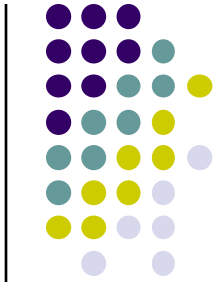
(a) x^1 view (b) x^2 view

Co-Training

- Mô hình thuật toán

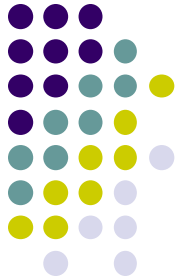


Co-Training



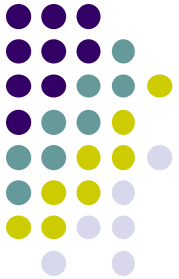
- Điều kiện dừng
  - hoặc tập dữ liệu chưa gán nhãn là rỗng
  - hoặc số vòng lặp đạt tới ngưỡng được xác định trước
- Một số lưu ý
  - Tập dữ liệu gán nhãn có ảnh hưởng lớn đến co-training
    - Quá ít: không hỗ trợ co-training
    - Quá nhiều: không thu lợi từ co-training
  - Cơ sở tăng hiệu quả co-training: thiết lập tham số
    - Kích cỡ tập dữ liệu gán nhãn
    - Kích cỡ tập dữ liệu chưa gán nhãn
    - Số các mẫu thêm vào sau mỗi vòng lặp
  - Bộ phân lớp thành phần rất quan trọng

Chặn thay đổi miền dày đặc



- Transductive SVMs (S3VMs)
  - Phương pháp phân biệt làm việc trên p(y|x) trực tiếp
  - Khi p(x) và p(y|x) không tương thích [redacted]ưa p(x) ra khỏi miền dầy đặc
- Quá trình Gauxơ)

Mô hình đồ thị



- Biểu diễn dữ liệu chưa mô tả hết ánh xạ gán nhãn trên dữ liệu (chẳng hạn, nghịch lý “hiệu quả như nhau” trong biểu diễn văn bản)
- Ánh xạ gán nhãn có liên quan mô hình dữ liệu (mô hình / đặc trưng/ nhân / hàm tương tự) [redacted]ô hình đã có theo tự nhiên hoặc giả thiết dữ liệu tuân theo.

BÀI GIẢNG NHẬP MÔN KHAI PHÁ DỮ LIỆU

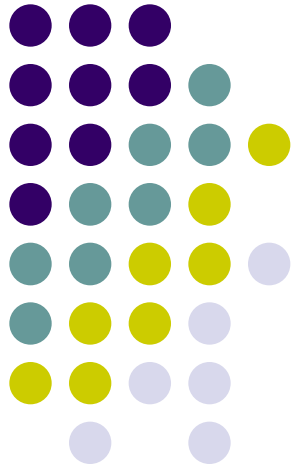
CHƯƠNG 6. PHÂN CỤM DỮ LiỆU

PGS. TS. HÀ QUANG THỤY

HÀ NỘI 9-2011

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐẠI HỌC QUỐC GIA HÀ NỘI



Nội dung

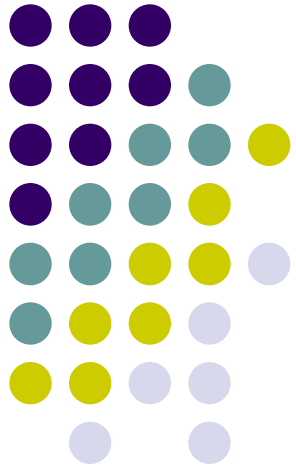
Giới thiệu phân cụm

Thuật toán phân cụm k-min

Thuật toán phân cụm phân cấp

Gán nhãn cụm

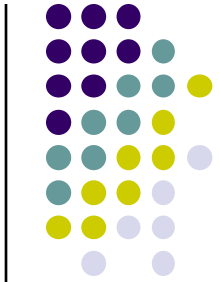
Đánh giá phân cụm



1. Bài toán phân cụm Web

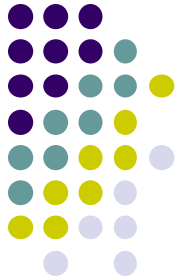
- Bài toán
  - Tập dữ liệu D = $\{d_i\}$
  - Phân các dữ liệu thuộc D thành các cụm
    - Các dữ liệu trong một cụm: "tương tự" nhau (gần nhau)
    - Dữ liệu hai cụm: "không tương tự" nhau (xa nhau)
  - Đo "tương tự" (gần) nhau ?
    - *Tiên đề phân cụm:* Nếu người dùng lựa chọn một đối tượng *d* thì họ cũng lựa chọn các đối tượng cùng cụm với *d*
    - Khai thác "cách chọn lựa" của người dùng
    - Đưa ra một số độ đo "tương tự" theo biểu diễn dữ liệu
- Một số nội dung liên quan
  - Xây dựng độ đo tương tự
  - Khai thác thông tin bổ sung
  - Số lượng cụm cho trước, số lượng cụm không cho trước

Sơ bộ tiếp cận phân cụm



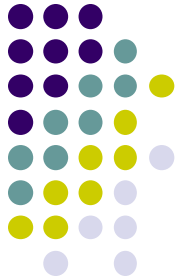
- Phân cụm mô hình và phân cụm phân vùng
  - Mô hình: Kết quả là mô hình biểu diễn các cụm tài liệu
  - Vùng: Danh sách cụm và vùng tài liệu thuộc cụm
- Phân cụm đơn định và phân cụm xác suất
  - Đơn định: Mỗi tài liệu thuộc duy nhất một cụm
  - Xác suất: Danh sách cụm và xác suất một tài liệu thuộc vào các cụm
- Phân cụm phẳng và phân cụm phân cấp
  - Phẳng: Các cụm tài liệu không giao nhau
  - Phân cấp: Các cụm tài liệu có quan hệ phân cấp cha- con
- Phân cụm theo lô và phân cụm tăng
  - Lô: Tại thời điểm phân cụm, toàn bộ tài liệu đã có
  - Tăng: Tài liệu tiếp tục được bổ sung trong quá trình phân cụm

Các phương pháp phân cụm



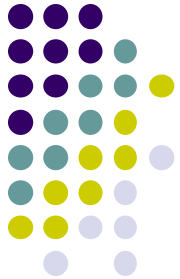
- Các phương pháp phổ biến
  - Phân vùng, phân cấp, dựa theo mật độ, dựa theo lưới, dựa theo mô hình, và mờ
- Phân cụm phân vùng
  - Xây dựng từng bước phân hoạch các cụm và đánh giá chúng theo các tiêu chí tương ứng
  - Độ đo tương tự / khoảng cách
  - K-mean, k-mediod
  - CLARANS, …
- Phân cụm phân cấp
  - Xây dựng hợp (tách) dần các cụm tạo cấu trúc phân cấp và đánh giá theo các tiêu chí tương ứng
  - Độ đo tương tự / khoảng cách
  - HAC: Hierarchical agglomerative clustering
  - CHAMELEON, BIRRCH và CURE, …

Các phương pháp phân cụm



- Phân cụm dựa theo mật độ
  - Hàm mật độ: Tìm các phần tử chính tại nơi có mật độ cao
  - Hàm liên kết: Xác định cụm là lân cận phần tử chính
  - DBSCAN, OPTICS…
- Phân cụm dựa theo lưới
  - Sử dụng lưới các ô cùng cỡ
  - Tạo phân cấp ô lưới theo một số tiêu chí: số lượng đối tượng trong ô
  - STING, CLIQUE, WaweCluster…
- Phân cụm dựa theo mô hình
  - Sử dụng một số mô hình giả thiết được phân cụm
  - Xác định mô hình tốt nhất phù hợp với dữ liệu
  - MCLUST…
- Phân cụm mờ
  - Giả thiết: không có phân cụm “cứng” cho dữ liệu và đối tượng có thể thuộc một số cụm
  - Sử dụng hàm mờ từ các đối tượng tới các cụm
  - FCM (Fuzzy CMEANS),…

Chế độ và đặc điểm phân cụm web



- Hai chế độ
  - Trực tuyến: phân cụm kết quả tìm kiếm người dùng
  - Ngoại tuyến: phân cụm tập văn bản cho trước
- Đặc điểm
  - Chế độ trực tuyến: tốc độ phân cụm
    - Web số lượng lớn, tăng nhanh và biến động lớn
    - Quan tâm tới phương pháp gia tăng
  - Một lớp quan trọng: phân cụm liên quan tới câu hỏi tìm kiếm
    - Trực tuyến
    - Ngoại tuyến

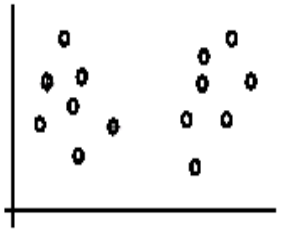
Carpineto C., Osinski S., Romano G., Weiss D. (2009). A survey of web clustering engines, *ACM Comput. Surv.* , **41**(3), Article 17, 38 pages.

Thuật toán K-mean gán cứng

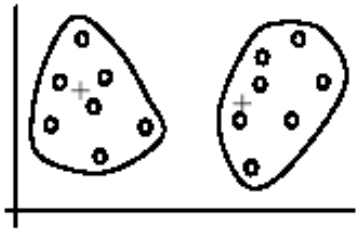
1. Khởi động: Chọn ngẫu nhiên *k* dữ liệu trong *S* làm trọng tâm (đại diện) cho các cụm $S_i = \{c_i: c_i \in S\}, \forall i=1,...,k$
2. Bước lặp:
   - 2.1. $S_i = \varnothing$ // Các cụm mới là rỗng
   - 2.2. $\forall d \in S$:
     - 2.2.1. Tính sim $(d, c_i), \forall i=1,...,k$
     - 2.2.2. $S_t = S_t \cup \{d\}$ nếu sim $(d, c_t) = \max \{\text{sim}(d, c_i) | i=1,...,k\}$
   - 2.3. $\forall i=1,...,k$, tính lại trọng tâm các cụm S_i : $c_i = \frac{1}{\|S_i\|} \sum_{d \in S_i} d$
3. Nếu chưa gặp Điều kiện dừng thì quay lại bước 2, ngược lại Dừng

- Một số lưu ý
  - Điều kiện dừng
    - Sau bước 2 không có sự thay đổi cụm
    - Điều kiện dừng cưỡng bức
      - Khống chế số lần lặp
      - Giá trị mục tiêu đủ nhỏ
  - Vấn đề chọn tập đại diện ban đầu ở bước Khởi động
  - Có thể dùng độ đo khoảng cách thay cho độ đo tương tự

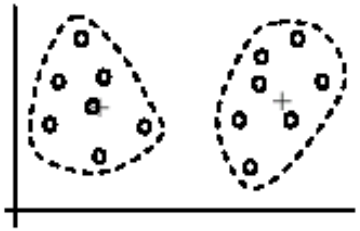
Thuật toán K-mean gán cứng



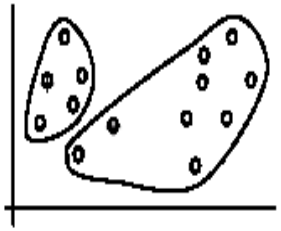
(A). Random selection of k centers



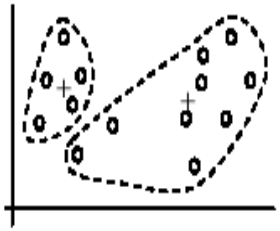
Iteration 2: (D). Cluster assignment



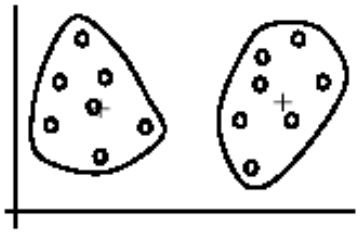
(E). Re-compute centroids



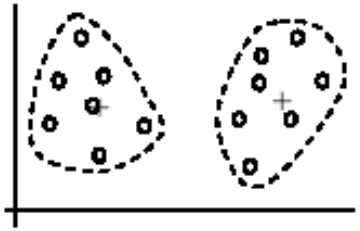
Iteration 1: (B). Cluster assignment



(C). Re-compute centroids



Iteration 3: (F). Cluster assignment



(G). Re-compute centroids

- Một số lưu ý (tiếp) và ví dụ
  - Trong bước 2: các trọng tâm có thể không thuộc S
  - Thực tế: số lần lặp [redacted]0
  - Thi hành k-mean với dữ liệu trên đĩa
    - Toàn bộ dữ liệu quá lớn: không thể ở bộ nhớ trong
    - Với mỗi vòng lặp: duyệt CSDL trên đĩa 1 lần
      - Tính được độ tương tự của d với các c_i.
      - Tính lại c_i mới: bước 2.1 khởi động (tổng, bộ đếm); bước 2.2 cộng và tăng bộ đếm; bước 2.3 chỉ thực hiện k phép chia.

Bing Liu (2007), Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, *Spinger*, 2007.

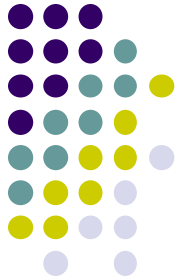
Thuật toán K-mean dạng mềm

- Input
  - Số nguyên k > 0: số cụm biết trước
  - Tập tài liệu D (cho trước)
- Output
  - Tập k "đại diện cụm" [redacted] làm tối ưu lỗi "lượng tử" $\sum_d \min_c |d - \mu_c|^2$
- Định hướng
  - Tinh chỉnh [redacted] dần với tỷ lệ học [redacted] learning rate) $\mu_c \leftarrow \mu_c + \Delta\mu_c$

$$\Delta\mu_c = \sum_d \begin{cases} \eta(d - \mu_c) & \text{nếu } \mu_c \text{ gần } d \text{ nhất} \\ 0 & \text{các trường hợp khác} \end{cases}$$

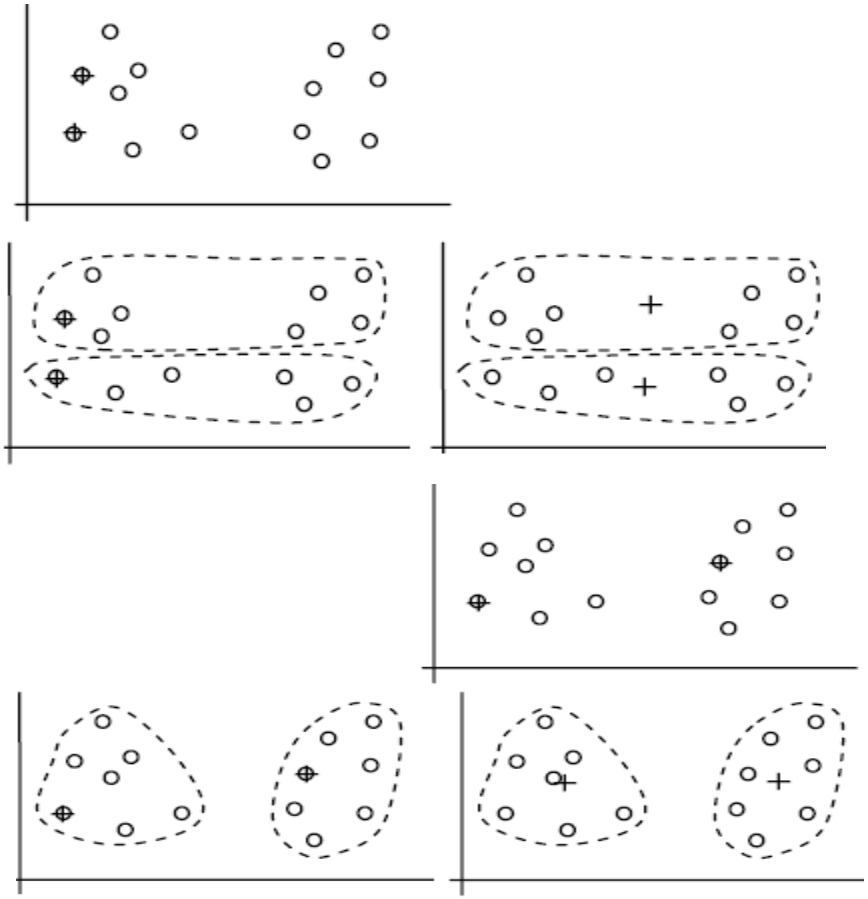
$$\Delta\mu_c = \eta \frac{1/|d-\mu_c|^2}{\sum_\gamma 1/|d-\mu_\gamma|^2}(d-\mu_c) \qquad \Delta\mu_c = \eta \frac{\exp(-|d-\mu_c|^2)}{\sum_\gamma \exp(-|d-\mu_\gamma|^2)}(d-\mu_c)$$

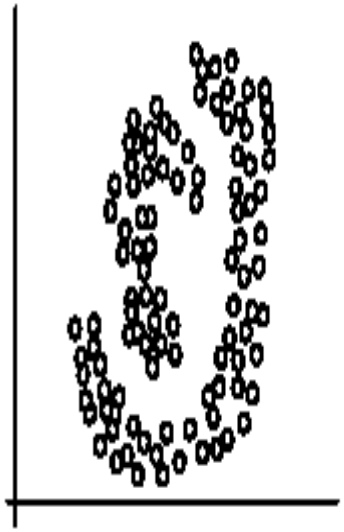
Thuật toán K-mean



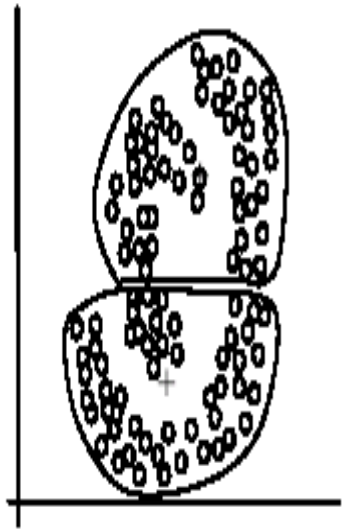
- Ưu điểm
  - Đơn giản, dễ sử dụng
  - Hiệu quả về thời gian: tuyến tính O(tkn), t số lần lặp, k số cụm, n là số phần tử
  - Một thuật toán phân cụm phổ biến nhất
  - Thường cho tối ưu cục bộ. Tối ưu toàn cục rất khó tìm
- Nhược điểm
  - Phải "tính trung bình được": dữ liệu phân lớp thì dựa theo tần số
  - Cần cho trước k : số cụm
  - Nhạy cảm với ngoại lệ (cách xa so với đại đa số dữ liệu còn lại): ngoại lệ thực tế, ngoại lệ do quan sát sai (làm sạch dữ liệu)
  - Nhạy cảm với mẫu ban đầu: cần phương pháp chọn mẫu thô tốt
  - Không thích hợp với các tập dữ liệu không siêu-ellip hoặc siêu cầu (các thành phần con không ellip/cầu hóa)

Thuật toán K-mean





(A): Two natural clusters



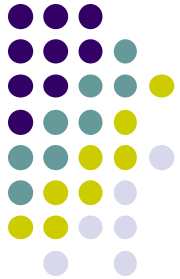
(B): *k*-means clusters

Trái: Nhạy cảm với chọn mẫu ban đầu

Phải: Không thích hợp với bộ dữ liệu không siêu ellip/cầu hóa

Bing Liu (2007), Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, *Spinger*, 2007.

3. Phân cụm phân cấp từ dưới lên



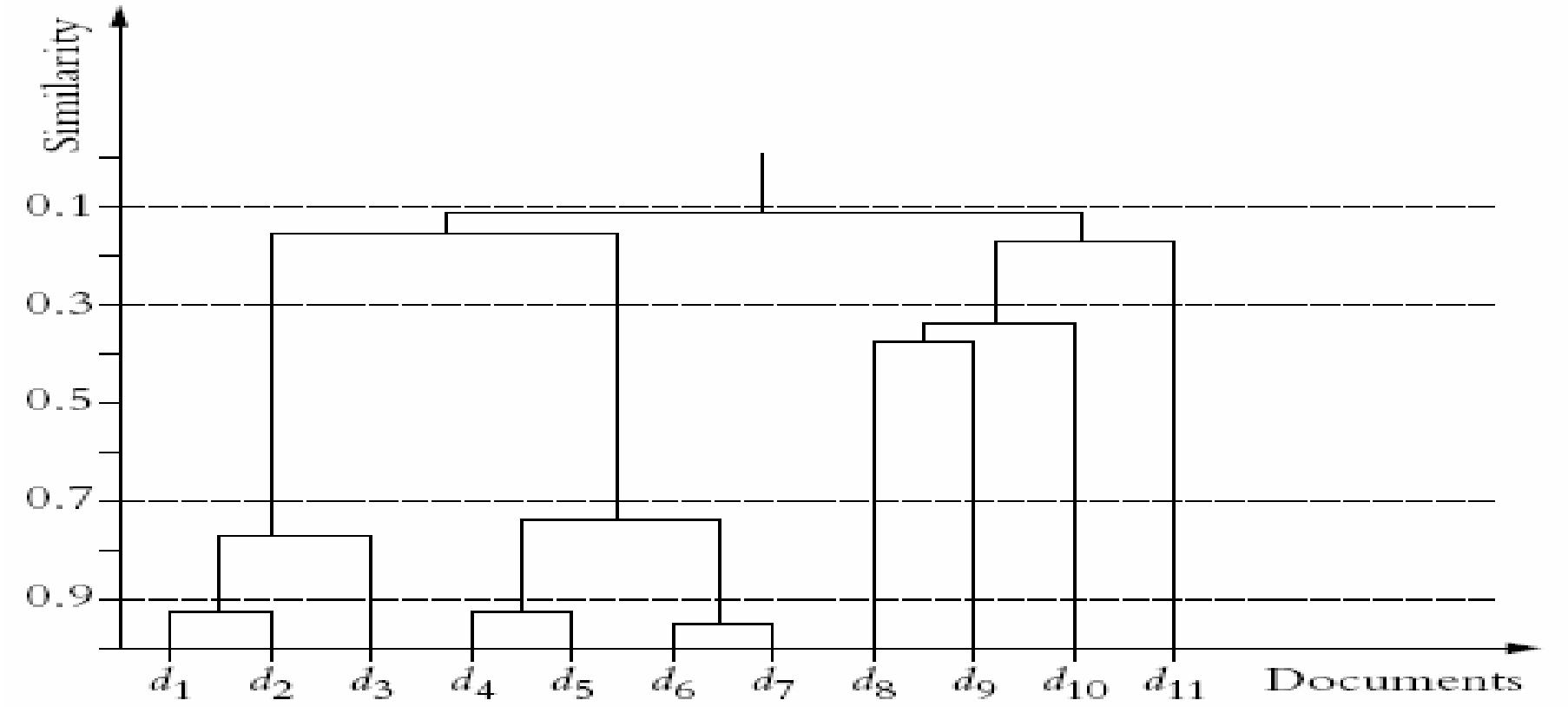
- HAC: Hierarchical agglomerative clustering
- Một số độ đo phân biệt cụm
  - Độ tương tự hai tài liệu
  - Độ tương tư giữa hai cụm
    - Độ tương tự giữa hai đại diện
    - Độ tương tự cực đại giữa hai tài liệu thuộc hai cụm: single-link
    - Độ tương tự cực tiểu giữa hai tài liêu thuộc hai cum: complete-link
    - Độ tương tự trung bình giữa hai tài liêu thuộc hai cum
- Sơ bộ về thuật toán
  - Đặc điểm: Không cho trước số lượng cụm k, cho phép đưa ra các phương án phân cụm theo các giá trị k khác nhau
  - Lưu ý: k là một tham số ⇨ “tìm k tốt nhất”
  - Tinh chỉnh: Từ cụ thể tới khái quát

Phân cụm phân cấp từ dưới lên

1. $G \leftarrow \{ \{d\} \mid d \in S\}$ (khởi tạo G là tập các cụm chỉ gồm một trang web trong tập S).
2. Nếu $|G| < k$ thì dừng thuật toán (đã đạt được số lượng cụm mong muốn).
3. Tìm hai cụm $S_i, S_j \in G$ sao cho $(i,j) = \arg\max_{(i,j)} sim(S_i, S_j)$ (tìm hai cụm có độ tương tự lớn nhất).
4. Nếu $sim(S_i, S_j) < q$ thì dừng thuật toán (độ tương tự của 2 cụm nhỏ hơn ngưỡng cho phép).
5. Loại bỏ S_i, S_j khỏi G.
6. $G = G \cup \{ S_i, S_j \}$ (ghép hai cụm S_i, S_j và đưa vào trong tập G).
7. Nhảy đến bước 2.

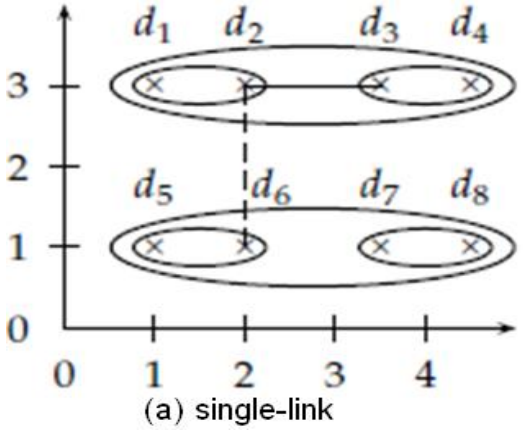
- Giải thích
  - G là tập các cụm trong phân cụm
  - Điều kiện |G| < k có thể thay thế bằng |G|=1

Phân cụm phân cấp từ dưới lên

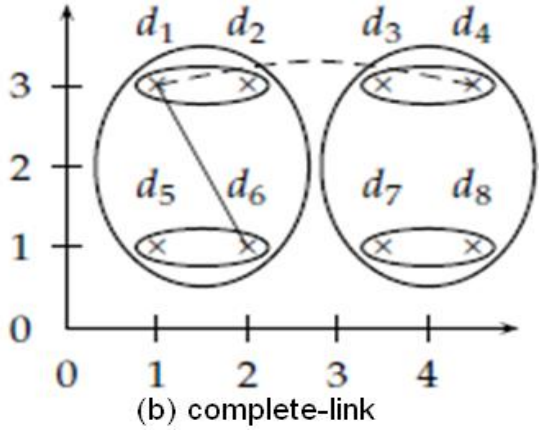


- Hoạt động HAC
  - Cho phép với mọi k
  - Chọn phân cụm theo “ngưỡng” về độ tương tự

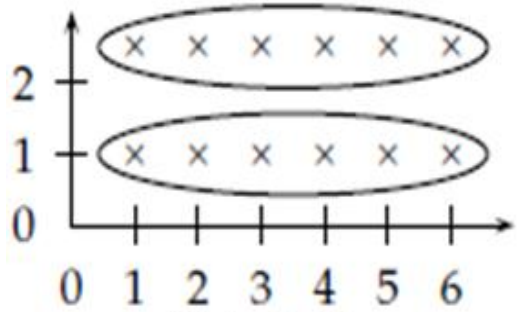
HAC với các độ đo khác nhau



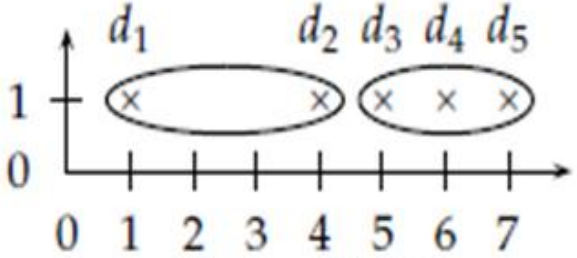
(a) single-link



(b) complete-link



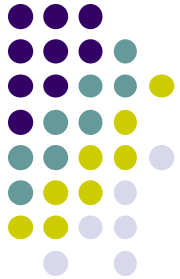
(a) single-link



(b) complete-link

- Ảnh hưởng của các độ đo
  - Trên: Hoạt động thuật toán khác nhau theo các độ đo khác nhau: độ tương tự cực tiểu (complete-link) có tính cầu hơn so với cực đại
  - Dưới: Độ tương tự cực đại (Single-link) tạo cụm chuỗi dòng

4. Biểu diễn cụm và gán nhãn



- Các phương pháp biểu diễn điển dình
  - Theo đại diện cụm
    - Đại diện cụm làm tâm
    - Tính bán kính và độ lệch chuẩn để xác định phạm vi của cụm
    - Cụm không ellip/cầu hóa: không tốt
  - Theo mô hình phân lớp
    - Chỉ số cụm như nhãn lớp
    - Chạy thuật toán phân lớp để tìm ra biểu diễn cụm
  - Theo mô hình tần số
    - Dùng cho dữ liệu phân loại
    - Tần số xuất hiện các giá trị đặc trưng cho từng cụm
- Lưu ý
  - Dữ liệu phân cụm ellip/cầu hóa: đại diện cụm cho biểu diễn tốt
  - Cụm hình dạng bất thường rất khó biểu diễn

Gán nhãn cụm tài liệu

- Phân biệt các cụm (MU)

$$I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}}$$

  - Chọn từ khóa đặc trưng tương quan cụm
  - Nxy (x có từ khóa t, y tài liệu thuộc C)
    - N11 : số tài liệu chứa t thuộc cụm C
    - N10 : số tài liệu chứa t không thuộc cụm C
    - N01 : số tài liệu không chứa t thuộc cụm C
    - N00 : số tài liệu không chứa t không thuộc cụm C
    - N: Tổng số tài liệu
- Hướng "trọng tâm" cụm
  - Dùng các từ khóa tần số cao tại trọng tâm cụm
- Tiêu đề
  - Chon tiêu đề của tài liệu trong cụm gần trọng tâm nhất

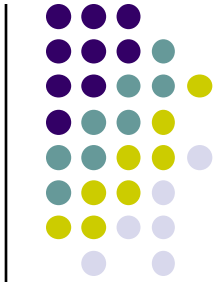
Gán nhãn cụm tài liệu

	# docs	labeling method		
		centroid	mutual information	title
4	622	oil plant mexico production crude **power 000 refinery gas** bpd	plant oil production **barrels** crude bpd mexico **dolly capacity petroleum**	MEXICO: Hurricane Dolly heads for Mexico coast
9	1017	police security **russian** people military peace killed told **grozny court**	police killed military security peace told **troops forces rebels** people	RUSSIA: Russia's Lebed meets rebel chief in Chechnya
10	1259	00 000 tonnes traders futures wheat prices **cents september** tonne	**delivery** traders futures tonne tonnes **desk** wheat prices 000 00	USA: Export Business - Grain/oilseeds complex

- Ví dụ
  - Ba phương pháp chọn nhãn cụm đối với 3 cụm là cụm 4 (622 tài liệu), cụm 9 (1017 tài liệu), cụm 10 (1259 tài liệu) khi phân cụm 10000 tài liệu đầu tiên của bộ Reuters-RCV1
  - centroid: các từ khóa có tần số cao nhất trong trọng tâm; mutual information (MU): thông tin liên quan phân biệt các cụm; title: tiêu đề tài liệu gần trọng tâm nhất.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.

5. Đánh giá phân cụm



- Đánh giá chất lượng phân cụm là khó khăn
  - Chưa biết các cụm thực sự
- Một số phương pháp điển hình
  - Người dùng kiểm tra
    - Nghiên cứu trọng tâm và miền phủ
    - Luật từ cây quyết định
    - Đọc các dữ liệu trong cụm
  - Đánh giá theo các độ đo tương tự/khoảng cách
    - Độ phân biệt giữa các cụm
    - Phân ly theo trọng tâm
  - Dùng thuật toán phân lớp
    - Coi mỗi cụm là một lớp
    - Học bộ phân lớp đa lớp (cụm)
    - Xây dựng ma trận nhầm lẫn khi phân lớp
    - Tính các độ đo: entropy, tinh khiết, chính xác, hồi tưởng, độ đo F và đánh giá theo các độ đo này

Đánh giá theo độ đo tương tự

- Độ phân biệt các cụm
  - Cực đại hóa tổng độ tương tự nội tại của các cụm
  - Cực tiểu hóa tổng độ tương tự các cặp cụm khác nhau
  - Lấy độ tương tự cực tiểu (complete link), cực đại (single link)

$$J_e = \frac{1}{2}\sum_{i=1}^{k}\frac{1}{|S_i|}\sum_{d_j,d_l\in S_i}\|d_j - d_l\|^2$$

$$J_s = \frac{1}{2}\sum_{i=1}^{k}\frac{1}{|S_i|}\sum_{d_j,d_l\in S_i} sim(d_j,d_l) = \frac{1}{2}\sum_{i=1}^{k}|S_i|sim(S_i)$$

- Một số phương pháp điển hình
  - Phân lý theo trọng tâm

$$J_e = \sum_{i=1}^{k}\sum_{d\in S_i}\|d - c_i\|^2$$

Ví dụ

Bảng 7.2 Dữ liệu mẫu dành cho phân cụm phẳng						
Tên trang web	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*
Anthropology	0	0.537	0.477	0	0.673	0.177
Art	0	0	0	0.961	0.195	0.196
Biology	0	0.347	0.924	0	0.111	0.112
Chemistry	0	0.975	0	0	0.155	0.158
Communication	0	0	0	0.78	0.626	0
Computer Science	0	0.989	0	0	0.13	0.067
Criminal Justice	0	0	0	0	1	0
Economics	0	0	1	0	0	0
English	0	0	0	0.98	0	0.199
Geography	0	0.849	0	0	0.528	0
History	0.991	0	0	0.135	0	0
Mathematics	0	0.616	0.549	0.49	0.198	0.201
Modern Languages	0	0	0	0.928	0	0.373
Music	0.97	0	0	0	0.17	0.172
Philosophy	0.741	0	0	0.658	0	0.136
Physics	0	0	0.894	0	0.315	0.318
Political Science	0	0.933	0.348	0	0.062	0.063
Psychology	0	0	0.852	0.387	0.313	0.162
Sociology	0	0	0.639	0.57	0.459	0.237
Theatre	0	0	0	0	0.967	0.254

Bảng 7.6 **Giá trị của hàm đánh giá dựa trên độ đo tương tự với giải thuật k-means**

k=2	*k*=3	*k*=4
1 [8.53381] Anthropology, Biology, Chemistry, Computer Science, Economics, Geography, Mathematics, Physics, Political Science, Psychology, Sociology **2 [6.12743]** Art, Communication, Criminal Justice, English, History, Modern Languages, Music, Philosophy, Theatre **Σ= [12.0253]**	**1 [2.83806]** History, Music, Philosophy **2 [6.09107]** Anthropology, Biology, Chemistry, Computer Science, Geography, Mathematics, Political Science **3 [7.12119]** Art, Communication, Criminal Justice, Economics, English, Modern Languages, Physics, Psychology, Sociology, Theatre **Σ= [12.0253]**	**1 [3.81771]** Art, Communication, English, Modern Languages **2 [5.44416]** Biology, Economics, Mathematics, Physics, Psychology, Sociology **3 [2.83806]** History, Music, Philosophy **4 [5.64819]** Anthropology, Chemistry, Computer Science, Criminal Justice, Geography, Political Science, Theatre **Σ= [12.0253]**