Chương 1 : Giới thiệu chung về khai phá dữ liệu

Trịnh Anh Phúc [1]

[1]Bộ môn Khoa Học Máy Tính, Viện CNTT & TT,
Trường Đại Học Bách Khoa Hà Nội

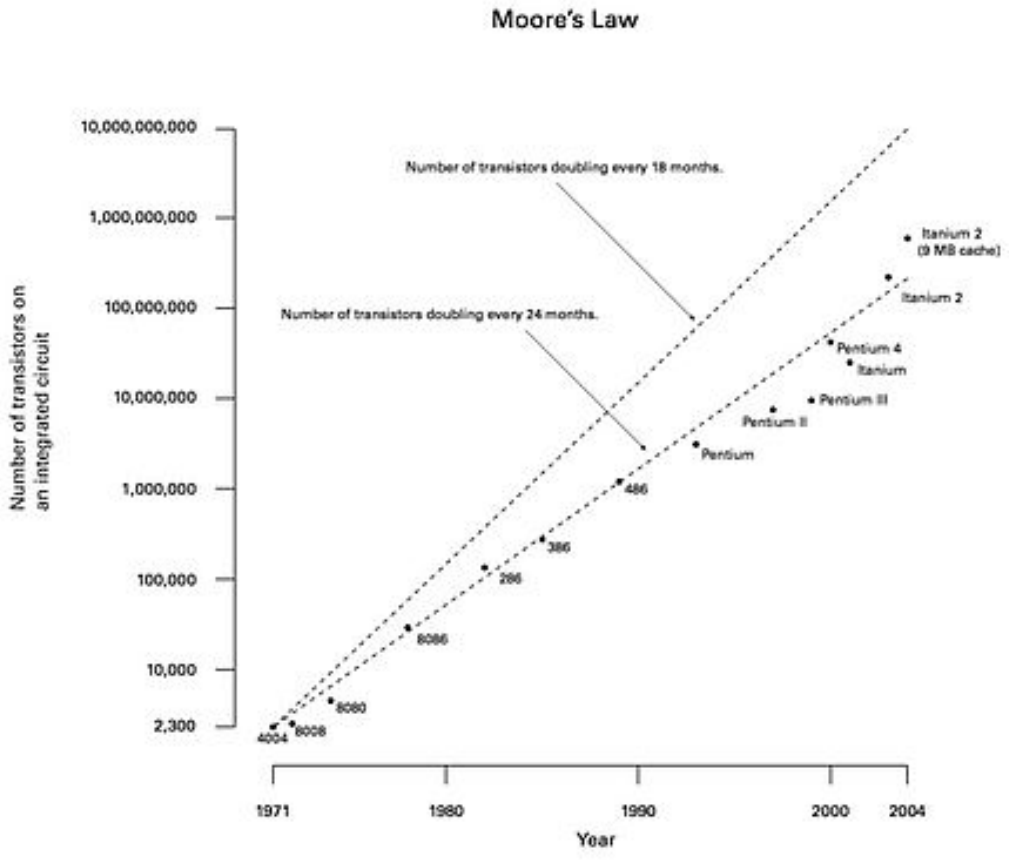
Ngày 2 tháng 5 năm 2014

Giới thiệu

1. Nhu cầu của khai phá dữ liệu (KPDL)
   - Luật Moore còn đúng ?
   - Thông tin trở nên quá tải
   - Sự đa dạng của các thiết bị số hóa
   - Xã hội hóa thông tin và kinh tế tri thức
2. Định nghĩa về khai phá dữ liệu
   - Tri thức là gì ?
   - Khai phá dữ liệu
   - Sự khác biệt HQTCSDL và hệ khai phá dữ liệu
   - Các kiểu dữ liệu trong HKPDL
3. Các bài toán thường gặp trong khai phá dữ liệu
4. Các lĩnh vực chính ảnh hưởng đến KPDL
5. Các ứng dụng cơ bản của hệ KPDL
6. Tổng kết
7. Chương trình Weka

Nhu cầu của khai phá dữ liệu

Luật Moore : "Số lượng transistor trên mỗi đơn vị inch vuông sẽ tăng lên gấp đôi sau mỗi năm"



Nhu cầu của Khai Phá Dữ liệu

- Định luật Moore không còn đúng

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

Nhu cầu của khai phá dữ liệu

- Sự bùng nổ của thông tin

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}	0.0000000000001	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

Nhu cầu của khai phá dữ liệu

- Sự bùng nổ của thông tin

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

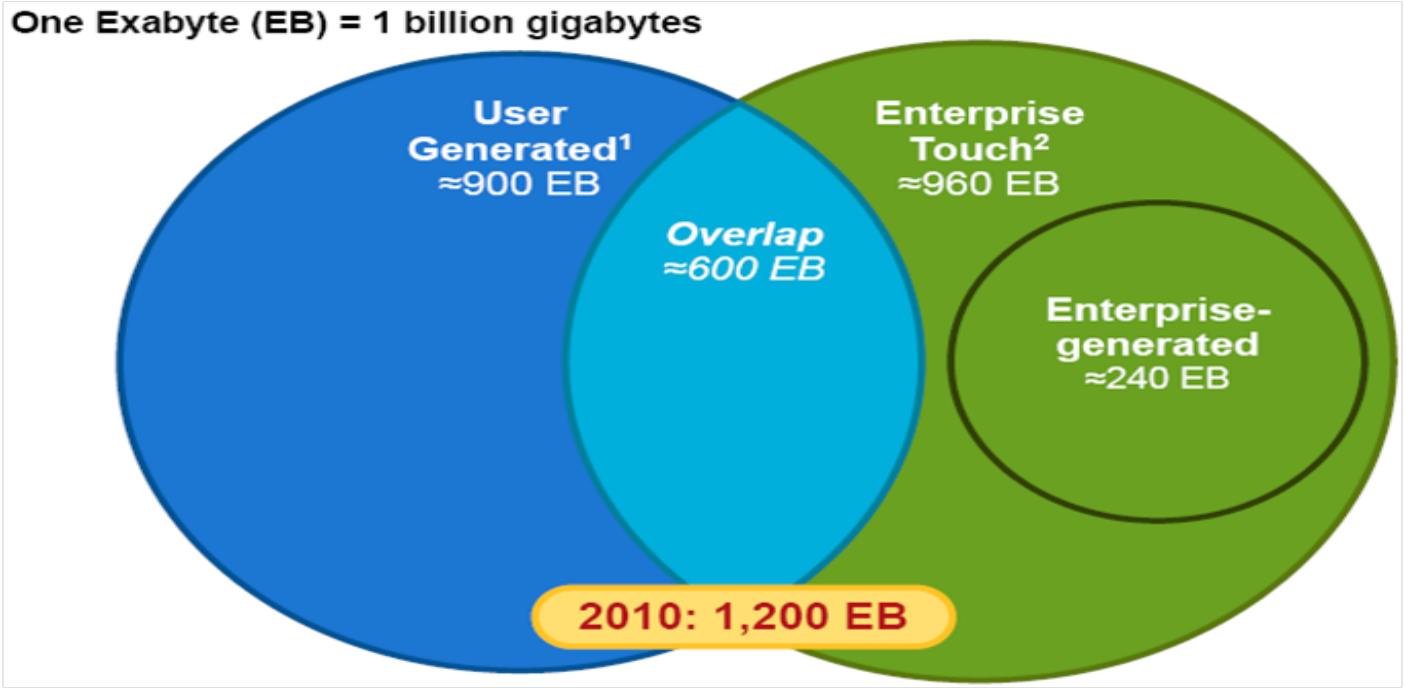
Nhu cầu của khai phá dữ liệu

- Năng lực số hóa các thiết bị điện tử trong mọi lĩnh vực cuộc sống

Nhu cầu của khai phá dữ liệu

Có nhiều đối tượng tạo thông tin trên mạng

- Hệ thống trực tuyến người dùng, Mạng xã hội...
- Mạng xã hội Facebook chứa tới 40 tỷ ảnh

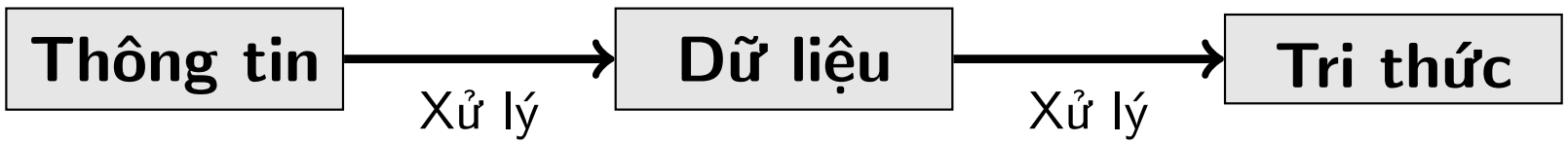


Nhu cầu của khai phá dữ liệu

Yêu cầu mới đặt ra

Do tính chất của dữ liệu lớn, đa dạng, nhiều nguồn nảy sinh nhu cầu xử lý từ dữ liệu lớn để có được tri thức - knowledge

- Khả năng tự động sinh tri thức
- Công cụ máy tính phục vụ nhu cầu kiết xuất tri thức
- Sử dụng tri thức nhằm phát triển xã hội, con người

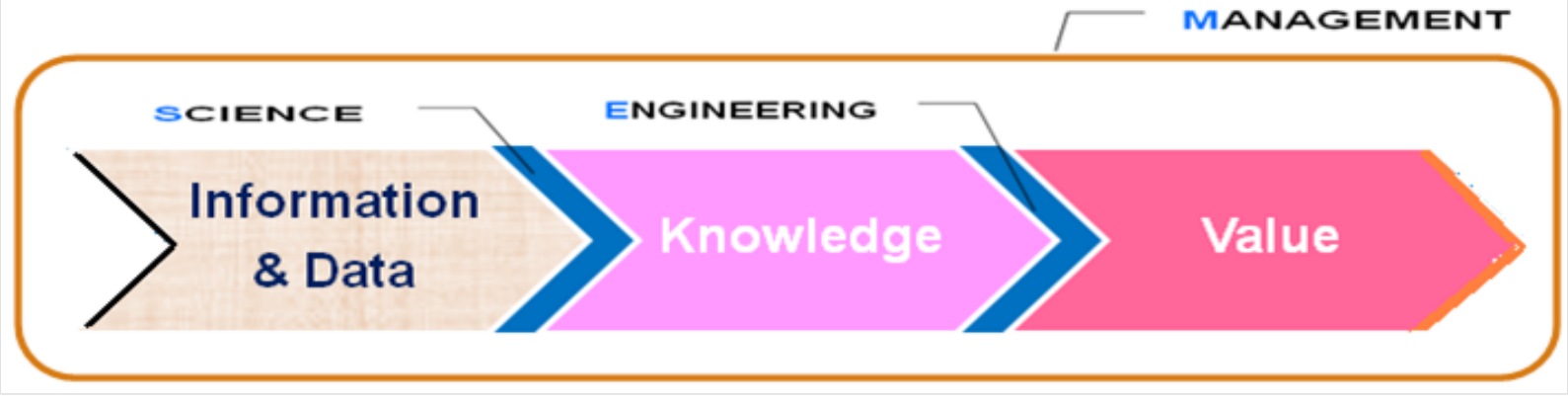


Mô hình chung của tiến trình xử lý thông tin

Nhu cầu của khai phá dữ liệu

Nền kinh tế dựa trên tri thức

- Khoa học : Thông tin và Dữ liệu
- Công nghệ : Tạo nên tri thức từ khoa học
- Quản lý : Tạo nên giá trị gia tăng cho các sản phẩm dựa trên tri thức



Dành cho trả lời câu hỏi

Định nghĩa về khai phá dữ liệu

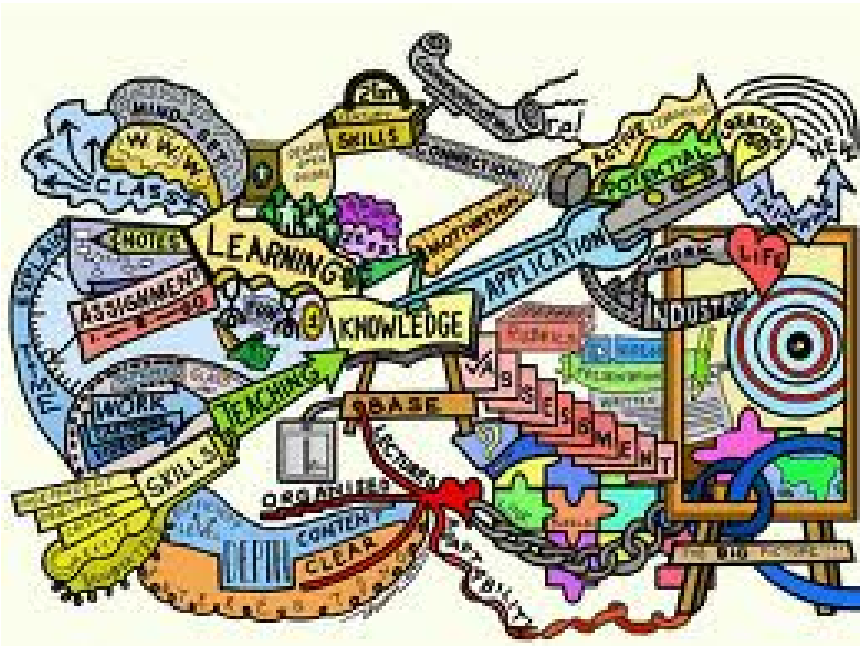
What is knowledge ?

Knowledge is a familiarity, awareness or *understanding of someone or something*, such as facts, information, descriptions, or skills, which is *acquired through experience or education* by perceiving, discovering, or learning. (Wikipedia)

Định nghĩa về khai phá dữ liệu

Định nghĩa về tri thức

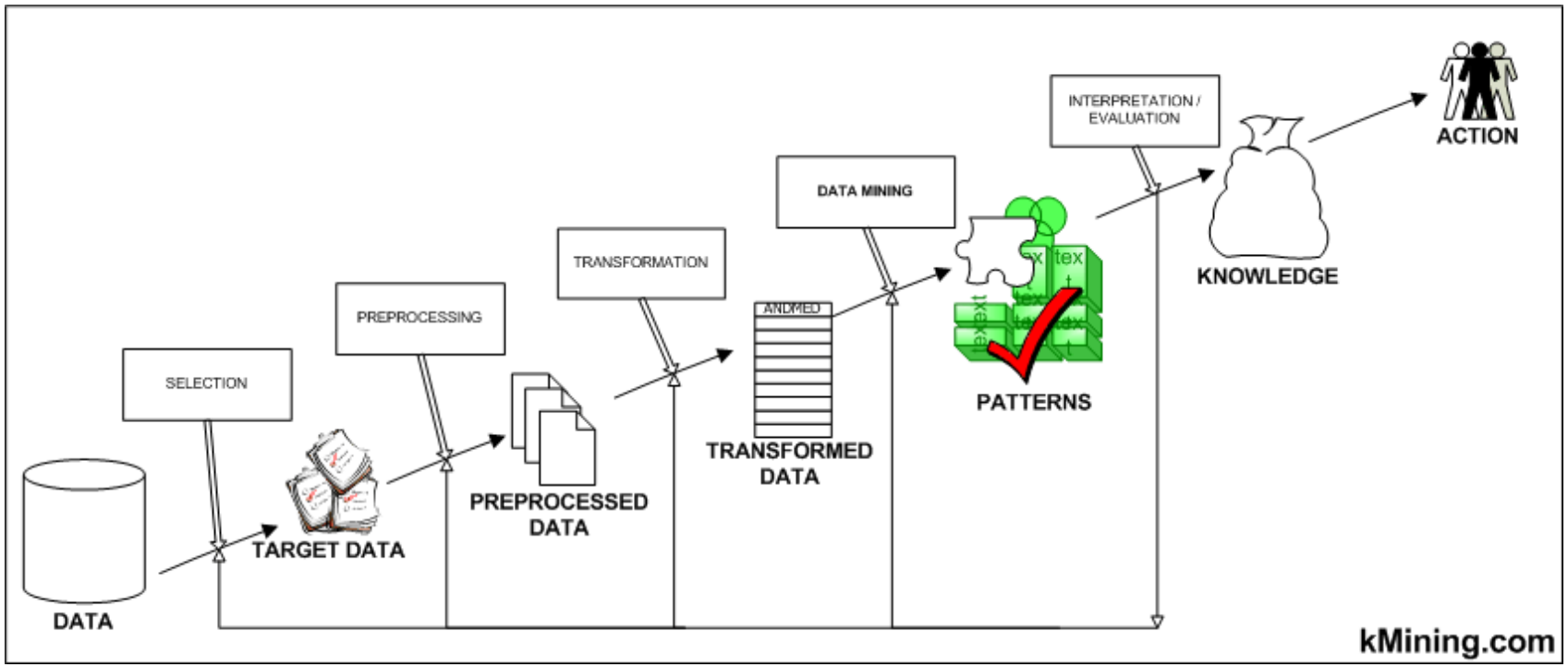
Tri thức bao gồm những *dữ kiện, thông tin*, sự mô tả, hay kỹ năng có được nhờ trải nghiệm hay thông qua giáo dục. Sự hình thành của tri thức liên quan đến những quá trình nhận thức, khám phá và học hỏi của mỗi người.



Định nghĩa về khai phá dữ liệu

What is Datamining ?

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information or knowledge



Định nghĩa về khai phá dữ liệu

Định nghĩa về khai phá dữ liệu

Khai phá dữ liệu (đôi khi còn gọi là khám phá tri thức) là một qua trình phân tích dữ liệu theo nhiều khía cạnh và tổng hợp nó lại để có được thông tin hữu ích hay tri thức. Như vậy có thể coi nó là bước *quan trọng nhất* trong quá trình phát hiện tri thức.

Định nghĩa về khai phá dữ liệu

Các bước của quá trình phát hiện tri thức gồm

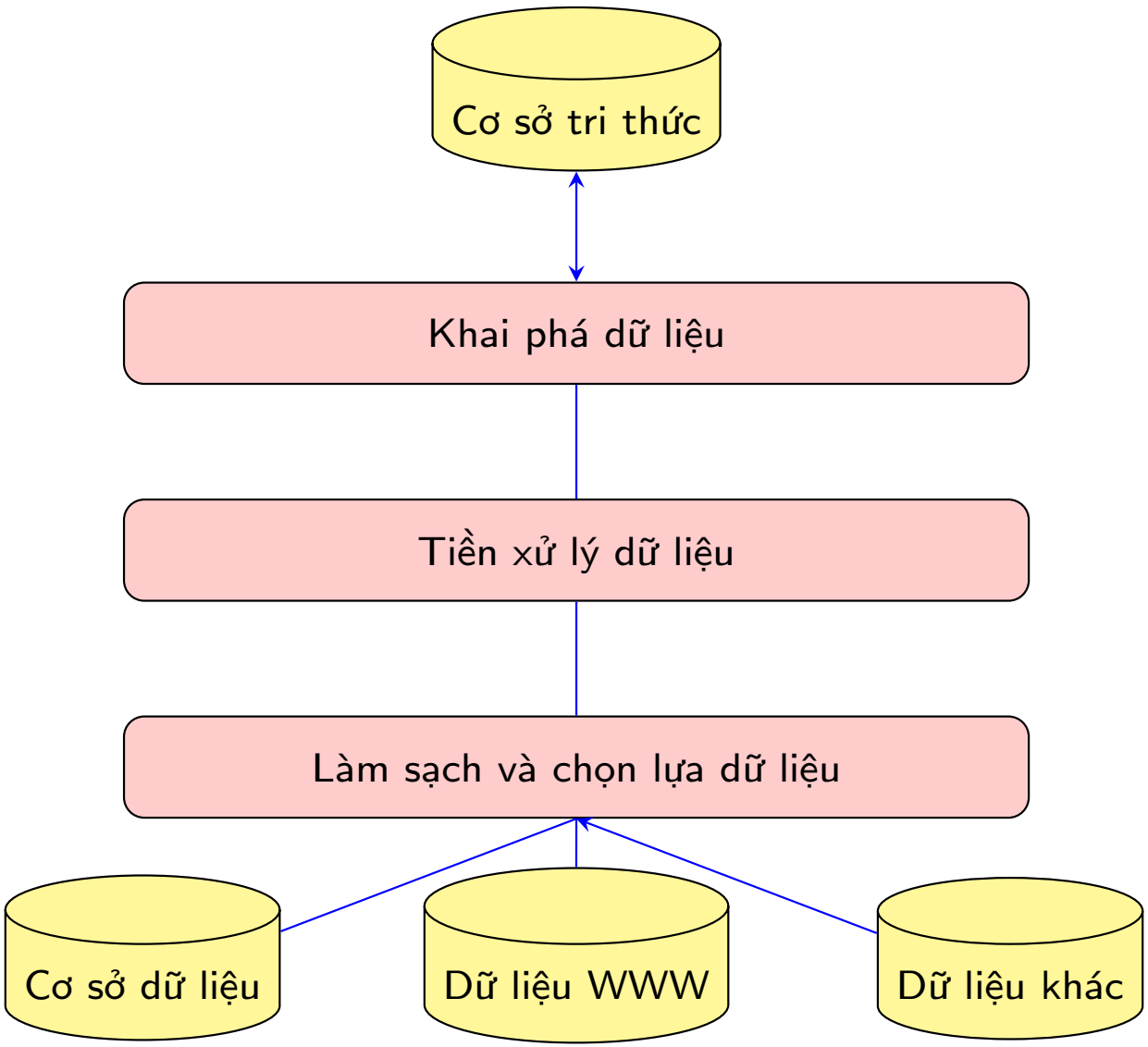
1. Lựa chọn (Selection)
   - Khởi tạo tập dữ liệu đích liên quan đến tri thức muốn phát hiện
2. Tiền xử lý (Preprocessing)
   - Tìm các đặc trưng hữu dụng, rút gọn chiều/biến,
3. Chuyển đổi (Transformation)
   - Xem xét các khía cạnh khác nhau của dữ liệu
4. Khai phá dữ liệu (Data Mining) *
5. Giải thích/Đánh giá (Interpretation/Evaluation)

Định nghĩa về khai phá dữ liệu

Bước 4 : Khai phá dữ liệu

1. Lựa chọn chức năng của ứng dụng : phân lớp, hồi quy, phân cụm, phân hạng (ranking)
2. Lựa chọn thuật toán tương ứng từng chức năng
3. Tìm mẫu (pattern) thích hợp tương ứng chức năng để giải thích/đánh giá thường có tính chất sau
   - Tính hiểu được
   - Tính so sánh
   - Tính phát hiện tri thức

Định nghĩa về khai phá dữ liệu



Định nghĩa về khai phá dữ liệu

Thông tin xung quanh tài khoản ngân hàng của bạn

Truy vấn HQTCSDL

1. Số dư trong tài khoản bạn là bao nhiêu ?
2. Bạn đã chi bao nhiêu trong tháng trước ?
3. Hiển thị tổng thu nhập từ lương bạn trong năm ngoái ?

Truy vấn Hệ khai phá dữ liệu

1. Làm sao để tăng số dư tài khoản của bạn ?
2. Số chi của bạn có xu hướng tăng trong các tháng nào của năm ?
3. Lương bạn có "đủ sống" trong khoảng năm năm tới ?

Định nghĩa về khai phá dữ liệu

- Cơ sở dữ liệu : bảng dữ liệu (Excel), dữ liệu văn bản (Text), dữ liệu quan-hệ (ER relationship), dữ liệu đối tượng (Object)
- Dữ liệu WWW : dữ liệu bán cấu trúc XML, HTML, nội dung web ....
- Dữ liệu khác
  - Dữ liệu không gian và thời gian
  - Dữ liệu chuỗi thời gian
  - Dữ liệu mạng xã hội
  - Dữ liệu đa phương tiện
  - Dữ liệu động không cùng khuôn dạng

Định nghĩa về khai phá dữ liệu

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]

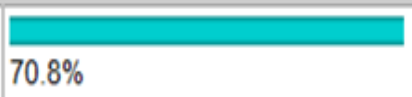


table data (fixed # of columns) (102)	70.8%
time series (56)	38.9%
itemsets / transactions (52)	36.1%
text (free-form) (43)	29.9%
anonymized data (38)	26.4%
social network data (28)	19.4%
other (22)	15.3%
web content (19)	13.2%
XML data (17)	11.8%
web clickstream (15)	10.4%
email (15)	10.4%
images /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

Dành cho trả lời câu hỏi

Các bài toán thường gặp trong khai phá dữ liệu

Các bài toán thường gặp bao gồm

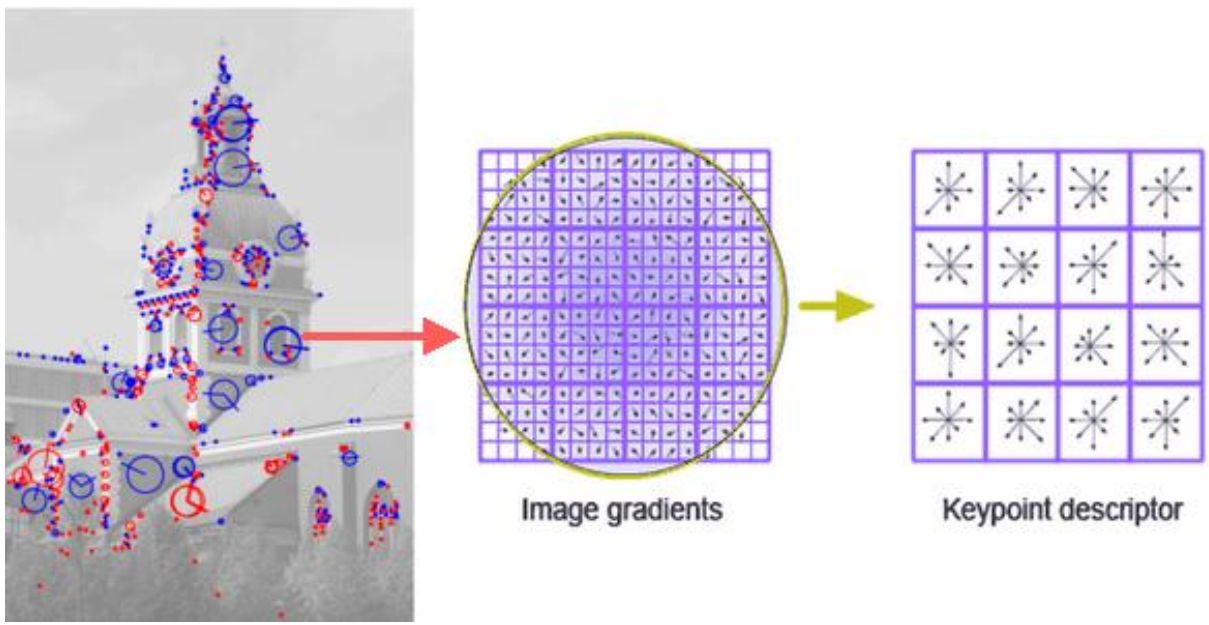
- Trích chọn đặc trưng
- Quan hệ kết hợp
- Phân lớp
- Phân cụm
- Hồi quy
- Phát hiện biến đổi và độ lệch

Ngoài ra 10 thách thức khác nữa
`http://www.cs.uvm.edu/~icdm/10Problems/index.shtml`

Các bài toán thường gặp trong khai phá dữ liệu

Trích chọn đặc trưng

- Tìm các đặc trưng và tính chất của dữ liệu, đối tượng
- Tổng quát hóa, tóm tắt, phát hiện đặc trưng ràng buộc, tương phản



Các bài toán thường gặp trong khai phá dữ liệu

Quan hệ kết hợp

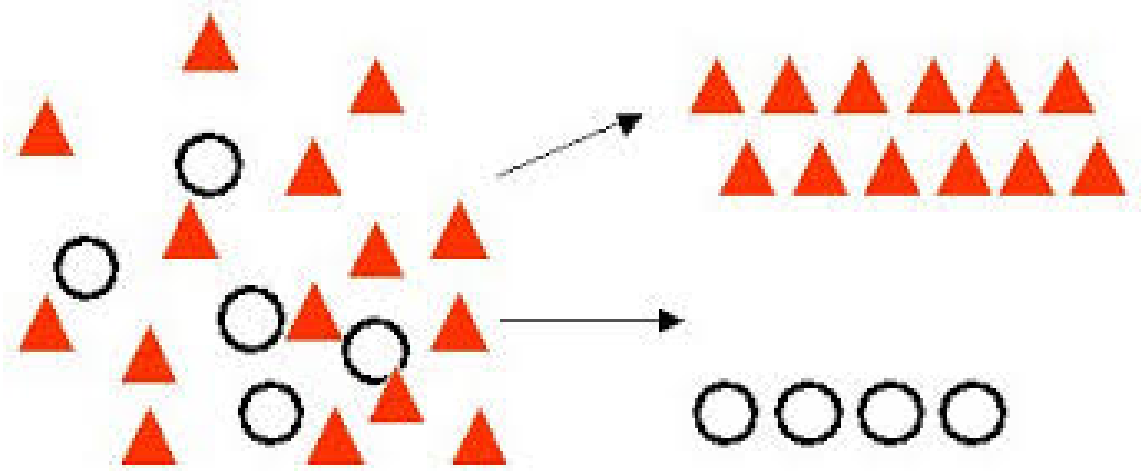
- Quan hệ kết hợp giữa các biến dữ liệu
- Dùng luật kết hợp (Association Rule Mining)

outlook	temperature	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no

Các bài toán thường gặp trong khai phá dữ liệu

Phân lớp

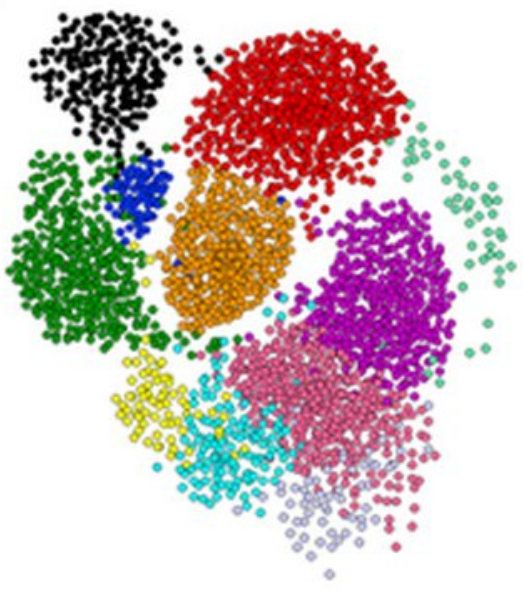
- Xây dựng các mô hình để mô tả và phân biệt khái niệm cho các lớp hoặc khái niệm để dự đoán trong tương lai
- Dự đoán giá trị số chưa biết hoặc đã mất



Các bài toán thường gặp trong khai phá dữ liệu

Phân cụm

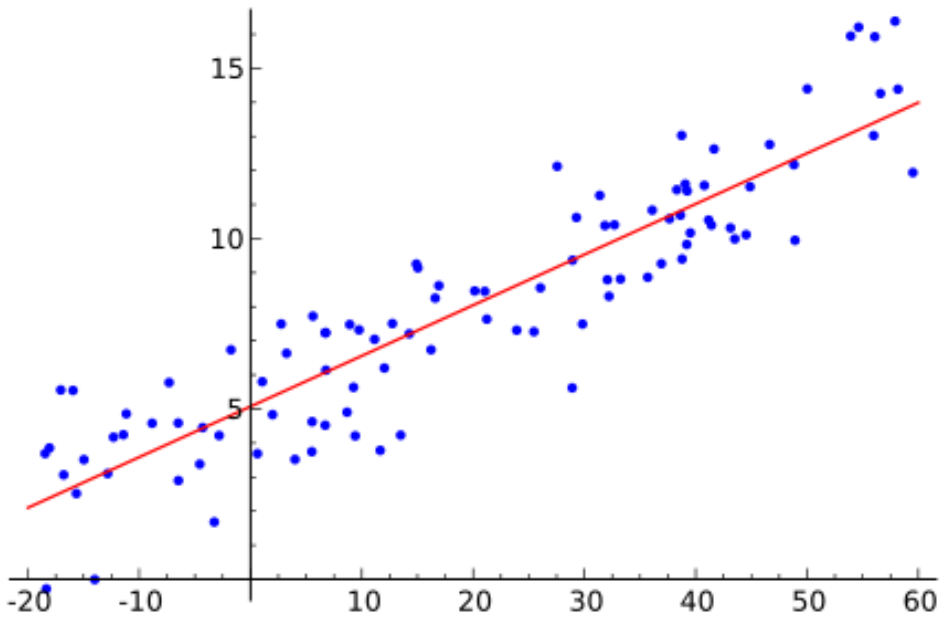
- nhóm dữ liệu thành các "cụm" để phát hiện được mẫu phân bố dữ liệu.
- tính tương tự của dữ liệu trong cùng "cụm"



Các bài toán thường gặp trong khai phá dữ liệu

Hồi quy

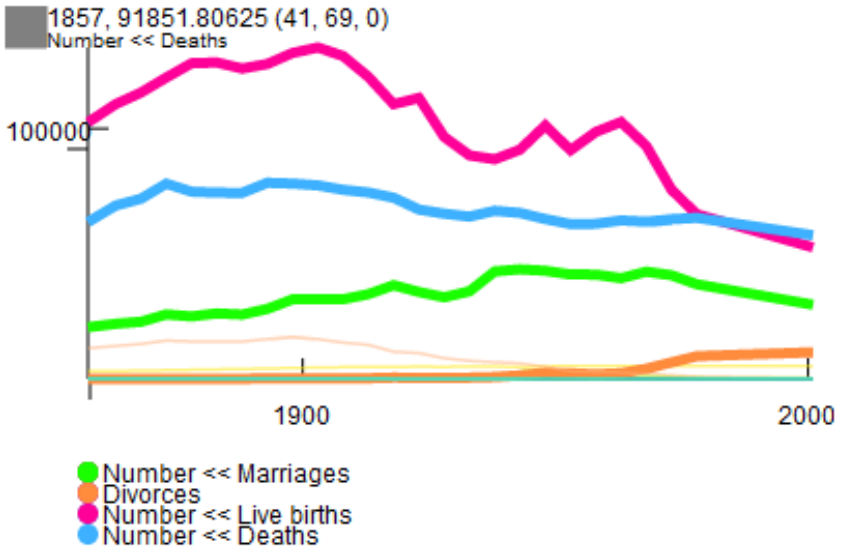
- xây dựng ánh xạ biểu diễn tập dữ liệu cùng nguồn gốc
- sử dụng ánh xạ để điền đầy dữ liệu



Các bài toán thường gặp trong khai phá dữ liệu

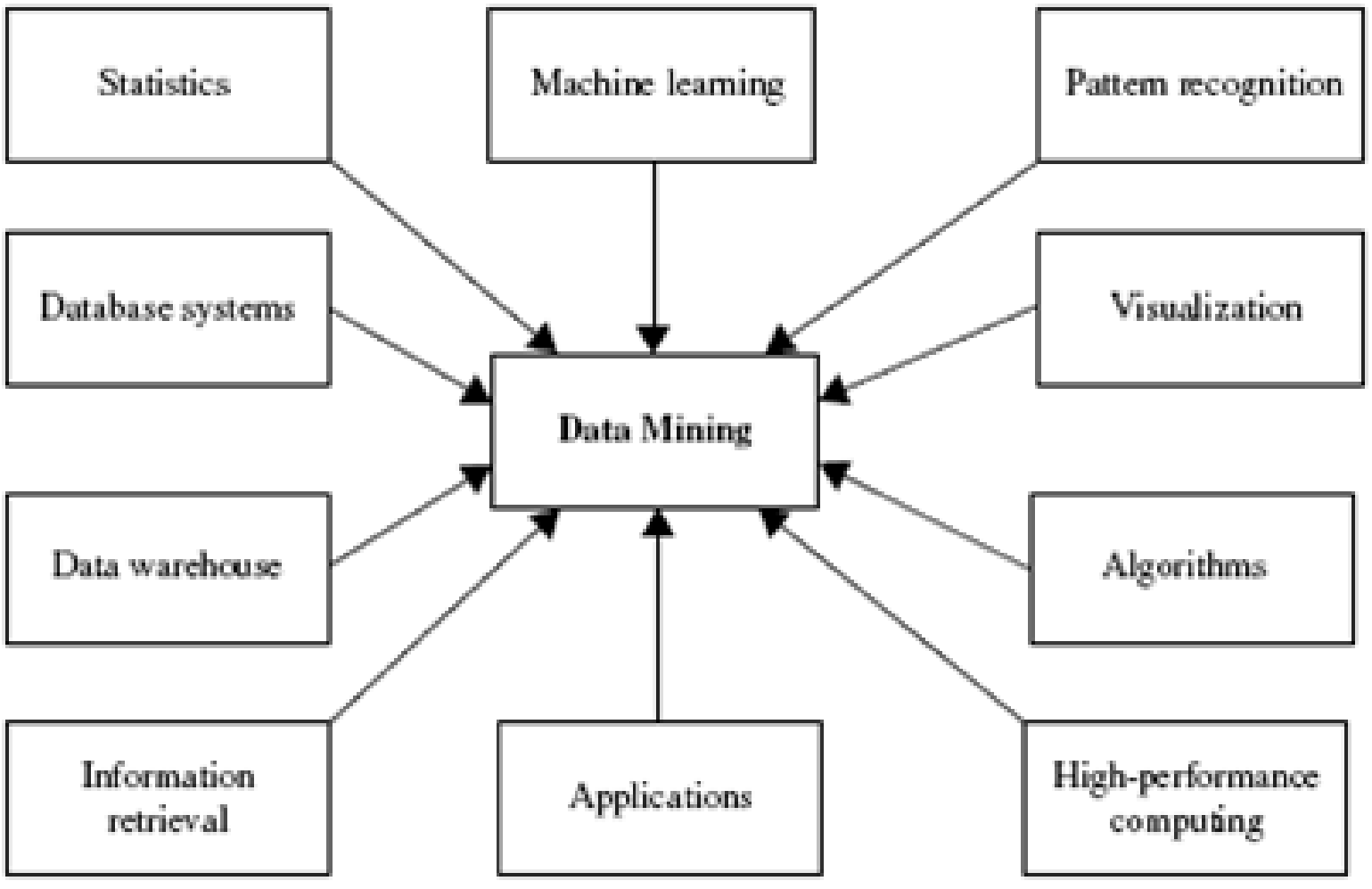
Phát hiện biến đổi và độ lệch

- Xu hướng và độ lệch: phân tích hồi quy
- Khai phá dữ liệu theo thời gian : phân tích chu kỳ
- Phân tích dựa trên tương tự



Dành cho trả lời câu hỏi

Các lĩnh vực chính ảnh hưởng đến KPDL



Các lĩnh vực chính ảnh hưởng đến KPDL

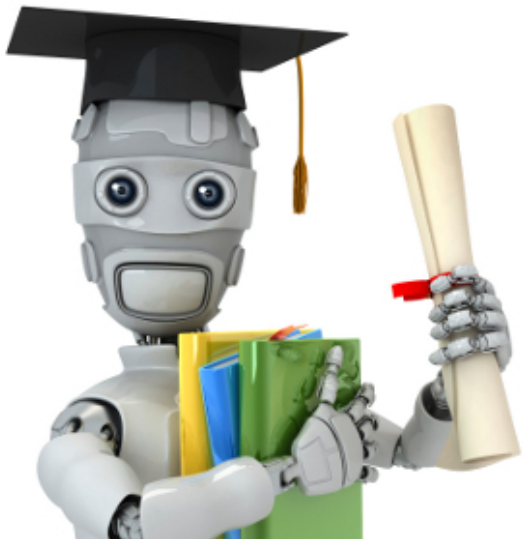
Thống kê & Xác suất (Statistics, Probability)

- Mô tả (Descritive Statistics) : sử dụng chính trong các tiến trình miêu tả sự biến đổi, lựa chọn dữ liệu
- Dự đoán (Predictive Statistics) : sử dụng chính trong các phần dự đoán, suy diễn của quá trình khai phá

Các lĩnh vực chính ảnh hưởng đến KPDL

Học máy (Machine Learning)

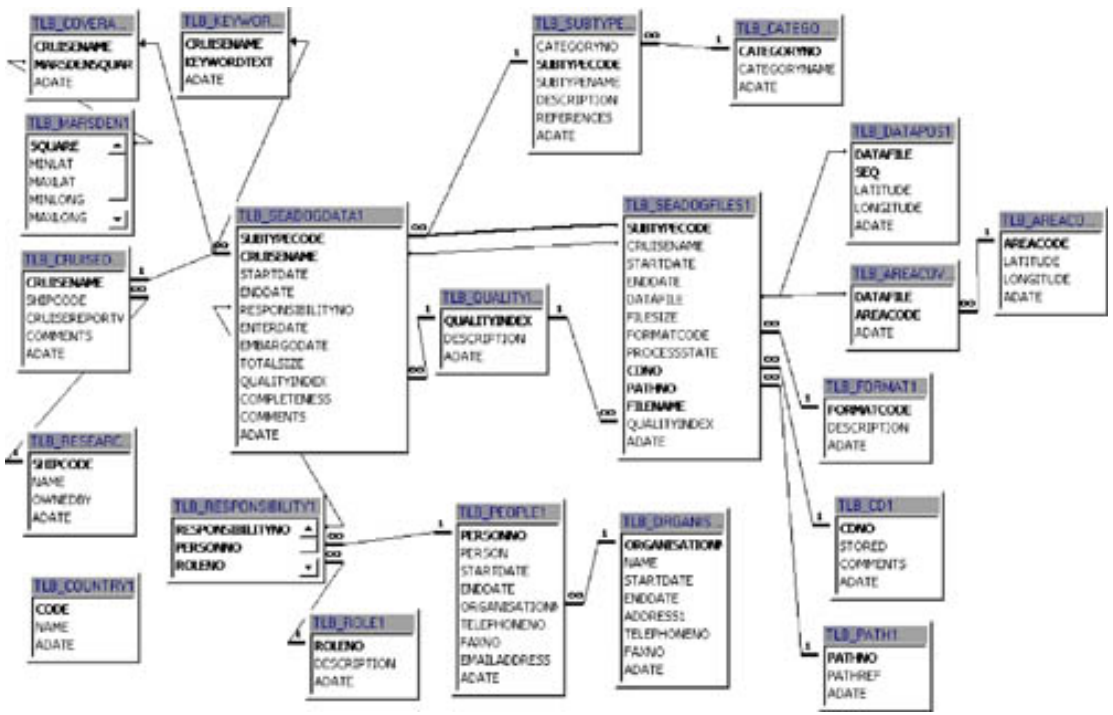
- Các chức năng của KPDL đều thuộc các bài toán trong lĩnh vực học máy
- Các bước lựa chọn, tiền xử lý cũng qui được về các bài toán trong lĩnh vực này



Các lĩnh vực chính ảnh hưởng đến KPDL

Các hệ cơ sở dữ liệu (Database Systems)

- Lưu trữ, tổ chức, hồi phục, sao lưu, truy vấn dữ liệu kích thước lớn. Dùng bước lựa chọn dữ liệu đích



Các lĩnh vực chính ảnh hưởng đến KPDL

Tìm kiếm (Information Retrieval)

- Tìm ra các thông tin liên quan đến tri thức

Các lĩnh vực chính ảnh hưởng đến KPDL

Giải thuật (Algorithms)

- Thiết kế giải thuật giải bài toán KPDL

```
function TARJAN(Node* node)
    node.visited ← true
    node.index ← indexCounter
    s.push(node)
    for all successor in node.successors do
        if !node.visited then TARJAN(successor)
        end if
        node.lowlink ← MIN(node.lowlink, successor.lowlink)
    end for
    if node.lowlink == node.index then
        repeat
            successor ← stack.pop()
        until successor == node
    end if
end function
```

Các lĩnh vực chính ảnh hưởng đến KPDL

Minh họa (Visualization)

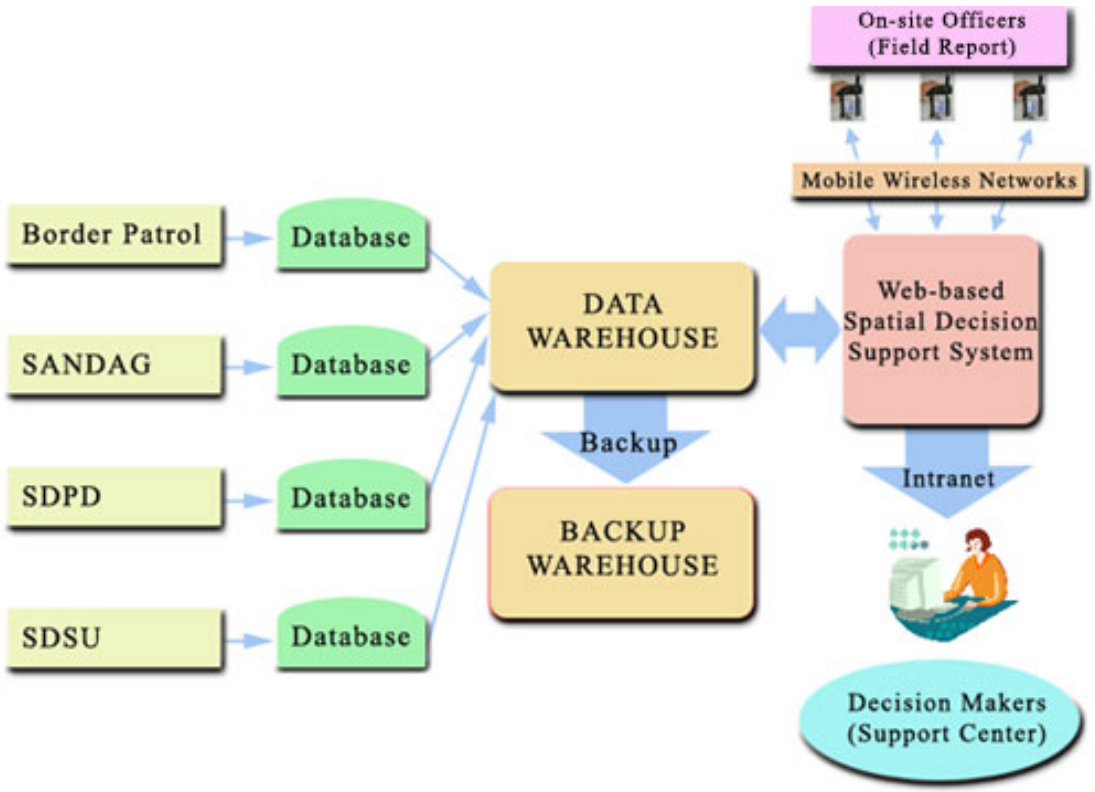
- Vẽ lại những tri thức giúp người dùng dễ hình dung

Dành cho trả lời câu hỏi

Các ứng dụng cơ bản của hệ KPDL

Hệ trợ giúp quyết định

- Phân tích và quản lý thị trường



Các ứng dụng cơ bản của hệ KPDL

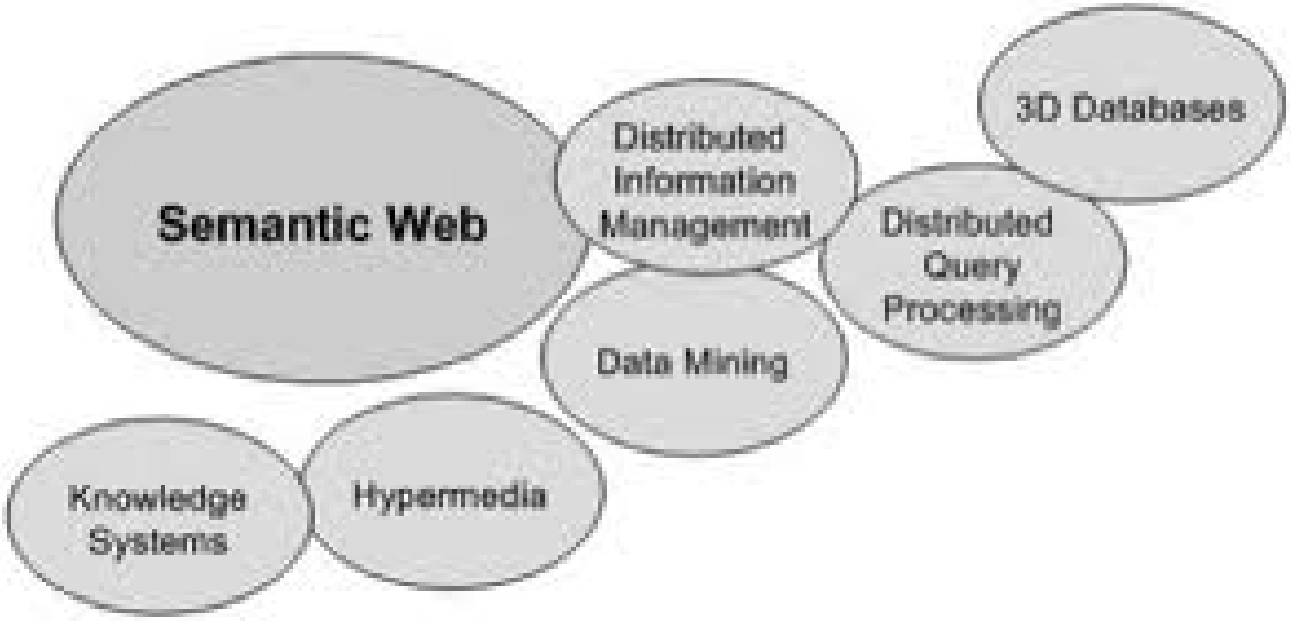
Tìm kiếm thông tin

- Tìm kiếm tài liệu hoặc tìm kiếm thông tin trong tài liệu theo một truy vấn. Tài liệu: văn bản, đa phương tiện, web...
- Hai giả thiết: (i) Dữ liệu tìm kiếm là không cấu trúc; (ii) Truy vấn dưới dạng từ khóa/cụm từ khóa mà không phải cấu trúc phức tạp

Các ứng dụng cơ bản của hệ KPDL

Khai phá Web

- Sử dụng Web ngữ nghĩa để lựa chọn dữ liệu
- Sử dụng các chức năng KPDL để thu nhập tri thức



Tổng kết

- Nhu cầu xuất hiện KPDL
- Định nghĩa KPDL
- Các bài toán thường gặp
- Các lĩnh vực ảnh hưởng
- Những ứn dụng cơ bản

Chương trình Weka

Yêu cầu

- Tải chương trình cài đặt Weka
- Tải các tập dữ liệu
- Chạy chương trình giải thích các phần giao diện