TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

0.1. NHU CẦU KHAI PHÁ DỮ LIỆU

0.2. KHAI PHÁ DỮ LIỆU LÀ GÌ?

0.3. KHÁI NIỆM VỀ DỮ LIỆU, MẪU VÀ TRI THỨC

0.4. CÁC BÀI TOÁN KHAI PHÁ DỮ LIỆU CƠ BẢN

0.5. CÁC GIAI ĐOẠN TRONG KHAI PHÁ DỮ LIỆU

0.6. KIẾN TRÚC ĐIỂN HÌNH CỦA MỘT HỆ THỐNG KPDL

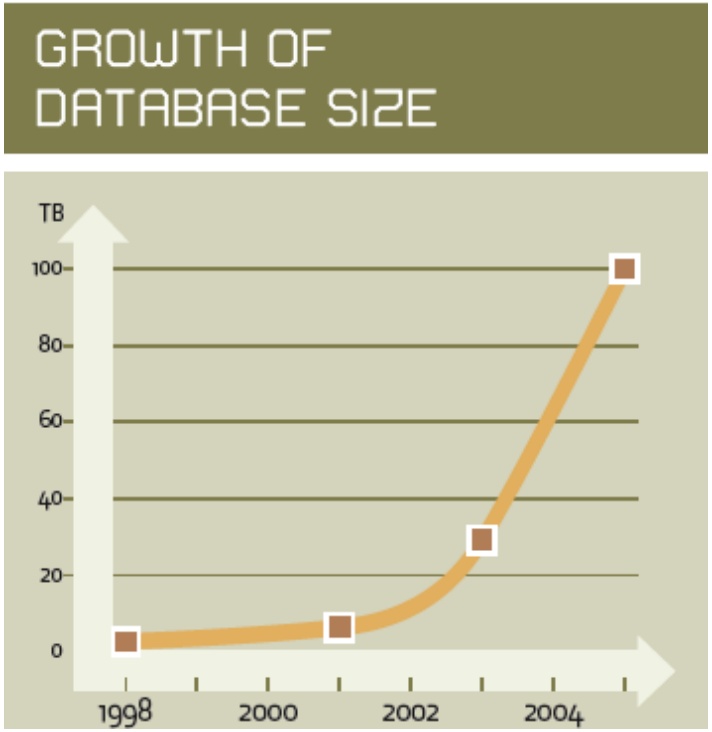
0.7. CÁC NGUỒN DỮ LIỆU PHỤC VỤ CHO KHAI PHÁ

0.8. ỨNG DỤNG CỦA KHAI PHÁ DỮ LIỆU

0.1. NHU CẦU KHAI PHÁ DỮ LIỆU

SỰ BÙNG NỔ THÔNG TIN!

- **Nhiều dữ liệu được sinh thêm**:
  - Web, văn bản, ảnh …
  - Giao dịch thương mại, cuộc gọi, ...
  - DL khoa học: thiên văn, sinh học …
- **Thêm nhiều dữ liệu được nắm giữ**:
  - Công nghệ lưu giữ nhanh hơn và rẻ hơn.
  - Hệ quản trị CSDL có thể quản lý các cơ sở dữ liệu với kích thước lớn hơn.




Google


YAHOO!


bing

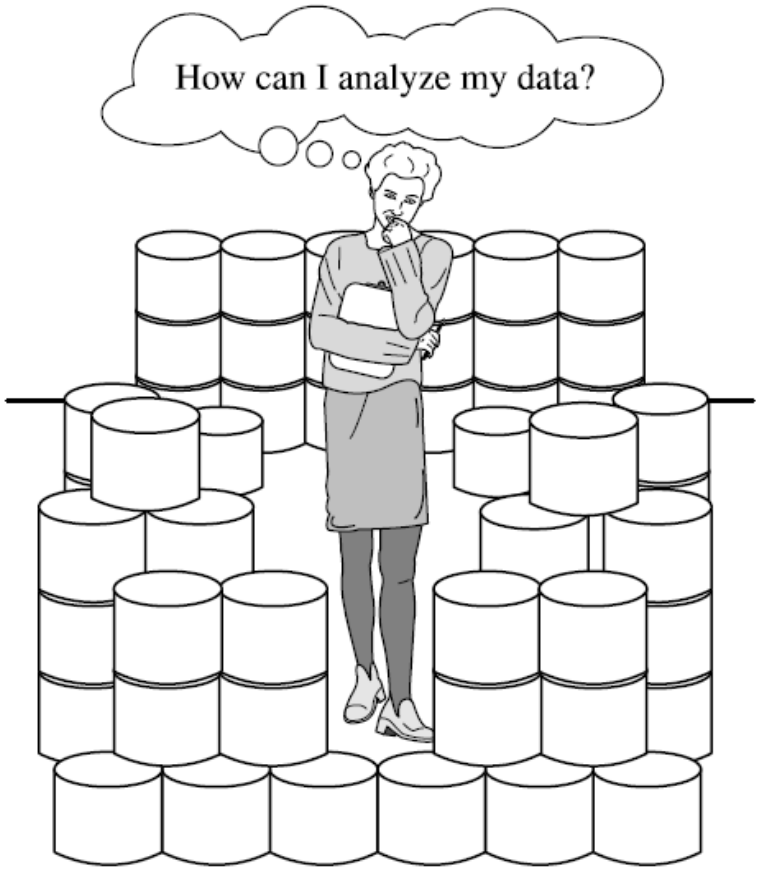

Alexa
The Web Information Company


YouTube


twitter


facebook

- **Vấn đề bùng nổ dữ liệu**
  - Các tiện ích thu thập dữ liệu tự động và công nghệ cơ sở dữ liệu lớn mạnh dẫn tới một lượng lớn dữ liệu được tích lũy và/hoặc cần được phân tích trong cơ sở dữ liệu, kho dữ liệu và trong các nguồn chứa dữ liệu khác.
- **Chúng ta bị ngập lụt trong dữ liệu mà khát tri thức!**
- **Giải pháp: Kho dữ liệu và Khai phá dữ liệu (mining)**
  - Tạo lập kho dữ liệu và quá trình phân tích dữ liệu trực tuyến OLAP.
  - Khai phá tri thức hấp dẫn (luật, quy luật, mẫu, ràng buộc) từ dữ liệu trong CSDL lớn.



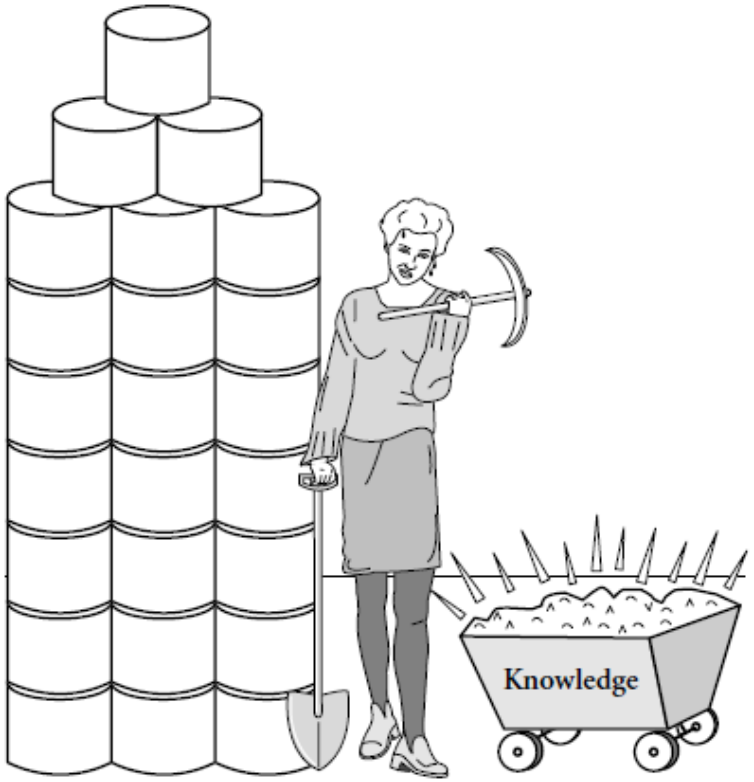
0.2. KHAI PHÁ DỮ LIỆU LÀ GÌ?

Theo J.Han và M.Kamber (2006) [1]:

<u>Quan niệm 1</u>:

Khai phá dữ liệu (Data Mining) là quá trình trích chọn ra tri thức từ trong một tập hợp rất lớn dữ liệu.

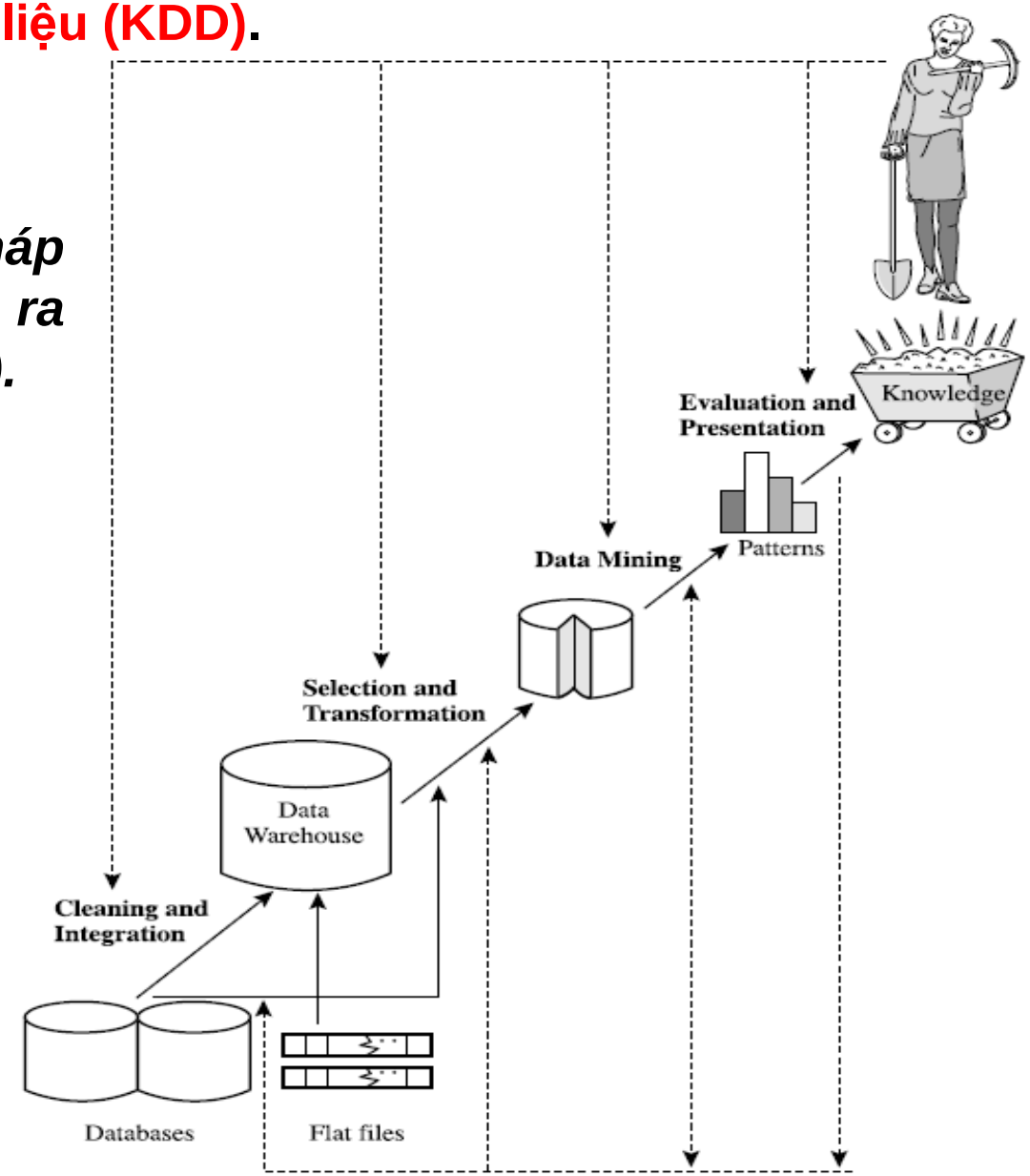
Khai phá dữ liệu = Phát hiện tri thức từ dữ liệu (KDD: Knowledge Discovery From Data).



Quan niệm 2:

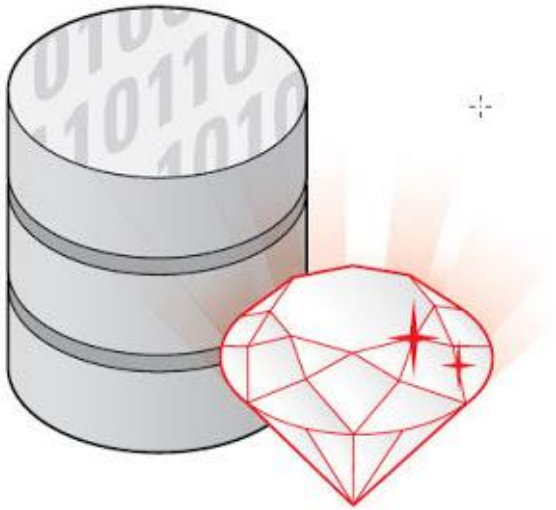
Khai phá dữ liệu (Data Mining) chỉ là một bước quan trọng trong quá trình phát hiên tri thức từ dữ liệu (KDD).

Áp dụng các phương pháp "thông minh" để trích chọn ra các mẫu dữ liệu (data pattern).



Theo Hà Quang Thụy và các tác giả (2009) [4] (trang 11 và 16):

***Khái niệm 1*: Phát hiện tri thức trong cơ sở dữ liệu (đôi khi còn được gọi là khai phá dữ liệu) là một quá trình không tầm thường nhằm phát hiện ra những mẫu có giá trị, mới, hữu ích tiềm năng và có thể thể hiểu được từ dữ liệu.**

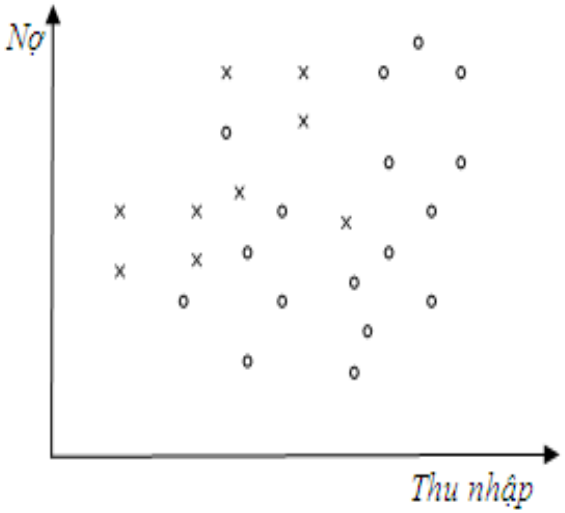


***Khái niệm 2*: Khai phá dữ liệu là một bước trong quá trình phát hiện tri thức trong cơ sở dữ liệu, thi hành một thuật toán khai phá dữ liệu để tìm ra các mẫu từ dữ liệu theo khuôn dạng thích hợp**

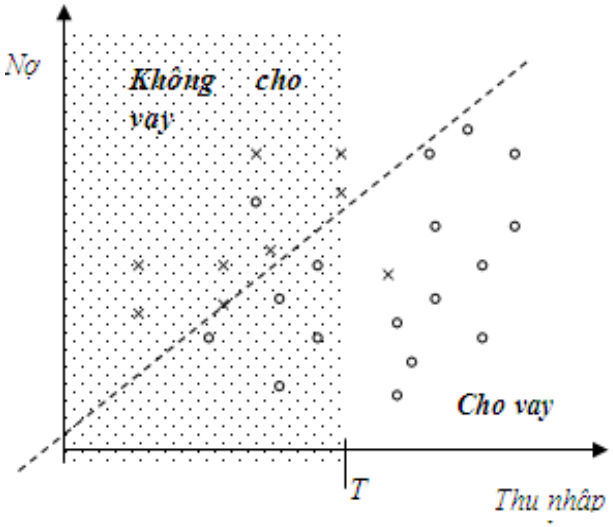
0.3. KHÁI NIỆM VỀ DỮ LIỆU, MẪU VÀ TRI THỨC

A. Khái niệm về dữ liệu và mẫu

- ***Dữ liệu (tập dữ liệu)***
  - Là một tập F gồm hữu hạn các *trường hợp* (*sự kiện*).
  - Trong khai phá dữ liệu, tập dữ liệu F thường phải gồm rất nhiều trường hợp.



- ***Mẫu***
  - Trong quá trình khai phá, người ta sử dụng ngôn ngữ L để biểu diễn các tập con các sự kiện (dữ liệu) thuộc vào tập sự kiện F.
  - Mỗi biểu thức E trong ngôn ngữ L biểu diễn tập con F_E tương ứng các sự kiện trong F.

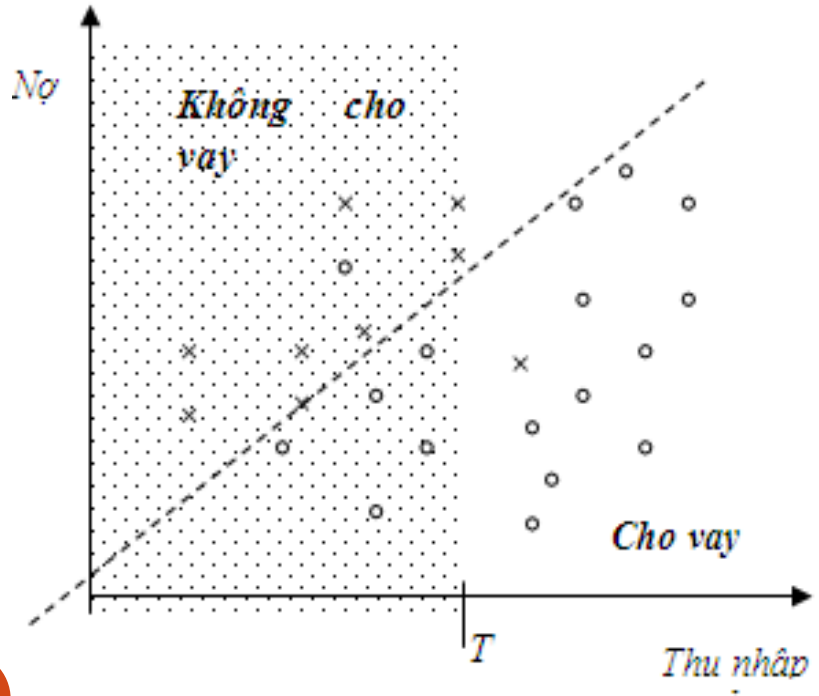


⟹ E được gọi là *mẫu* nếu nó đơn giản hơn so với việc liệt kê các sự kiện thuộc F_E.

Ví dụ: Mẫu "Thu nhập < T"

B. Tính có giá trị của mẫu

- Mẫu được phát hiện phải có *giá trị* đối với các dữ liệu mới (xuất hiện trong tương lai) theo một mức 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. Một biểu thức E trong L biểu diễn một tập con $F_E \subset F$ có thể được gán một độ đo chân thực c = C(E,F).



Với mẫu "THUNHẬP < $t"*: đường biên xác định mẫu 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.*

Với mẫu "a*THUNHẬP + b*NỢ < 0"*: tình trạng người vay nợ rơi vào tình trạng không thể chi trả tương ứng với nửa mặt phẳng trên ⟹ cho độ chân thực cao hơn.*

C. 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 khi quan tâm tới sự thay đổi trong:

- Dữ liệu: so sánh giá trị hiện tại với giá trị quá khứ hoặc giá trị kỳ vọng
- 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*.

Chẳng hạn: Hàm U ánh xạ các biểu thức trong L tới một không gian đo có thứ tự (bộ phận hoặc toàn bộ) M_U theo đó u = U (E,F).

D. Tính hiểu được, tính hấp dẫn và khái niệm về tri thức

- ***Tính hiểu được***: Mẫu phải hiểu được
  - Mục tiêu của khai phá dữ liệu là tạo ra các *mẫu mà con người hiểu chúng dễ dàng hơn* các dữ liệu nền (dữ liệu sẵn có trong hệ thống).
  - "Có thể hiểu được" là tiêu chí khó đo được một cách chính xác ⟹ Đưa ra một số độ đo về sự dễ hiểu và các độ đo như vậy được 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 theo đó s = S(E,F).
- ***Tính hấp dẫn***: Độ hấp dẫn (được coi là *độ đo tổng thể về mẫu)* là sự kết hợp của các tiêu chí *giá trị*, *mới*, *hữu ích* và *dễ hiểu*. Các hệ thống KPDL thường:
  - 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 thông qua thứ tự của các mẫu được phát hiện.

- ***Tri thức***: Một mẫu E $\in$ L được gọi là *tri thức* nếu như đối với một lớp người sử dụng nào đó, chỉ ra được một ngưỡng i $\in$ M_i mà độ hấp dẫn I(E,F,C,N,U,S) > i.

0.4. CÁC BÀI TOÁN KHAI PHÁ DỮ LIỆU ĐIỂN HÌNH

Mục tiêu tổng quát của khai phá dữ liệu là mô tả và dự báo

- ***Bài toán mô tả*: hướng tới việc tìm ra các mẫu mô tả dữ liệu.**
- ***Bài toán dự báo*: sử dụng một số biến (hoặc trường) trong cơ sở dữ liệu để dự đoán về giá trị chưa biết hoặc giá trị sẽ có trong tương lai của các biến.**

⟹ **Thể hiện thông qua các bài toán cụ thể:**

1. ***Mô tả khái niệm***
2. ***Quan hệ kết hợp***
3. ***Phân cụm***
4. ***Phân lớp***
5. ***Hồi quy***
6. ***Mô hình phụ thuộc***
7. ***Phát hiện thay đổi và độ lệch***

0.4.1. Mô tả khái niệm

- Nhằm tìm ra các đặc trưng và tính chất của khái niệm.
- Các bài toán điển hình bao gồm: tổng quát hóa, tóm tắt, phát hiện các đặc trưng dữ liệu ràng buộc,…

Bài toán tóm tắt là một trong những bài toán mô tả điển hình, áp dụng các phương pháp để tìm ra một mô tả cô đọng đối với một tập con dữ liệu. Ví dụ: xác định kỳ vọng và độ lệch chuẩn của một dãy các giá trị.

0.4.2. Tìm quan hệ kết hợp

- Phát hiện mối quan hệ kết hợp trong tập dữ liệu là bài toán quan trọng trong khai phá dữ liệu.
- Một trong những mối quan hệ kết hợp điển hình là quan hệ kết hợp giữa các biến dữ liệu trong đó ***bài toán khai phá luật kết hợp*** là một bài toán tiêu biểu.

Bài toán khai phá luật kết hợp thực hiện việc phát hiện ra mối quan hệ kết hợp giữa các tập thuộc tính (các tập biến) có dạng X⟶Y, trong đó X và Y là hai tập thuộc tính.

"Sự xuất hiện của X kéo theo sự xuất hiện của Y như thế nào?"

0.4.3. Phân lớp

- Thực hiện việc xây dựng (mô tả) các mô hình (hàm) dự báo nhằm mô tả hoặc phát hiện các lớp hoặc khái niệm cho các dự báo tiếp theo.
- Một số phương pháp điển hình là: cây quyết định, luật phân lớp, mạng neuron,…
- Nội dung của phân lớp chính là một hàm ánh xạ các dữ liệu vào trong một số các lớp (nhóm) đã biết.
- Phân lớp còn được gọi là “học máy có giám sát” (supervised learning).

0.4.4. Phân cụm

- Thực hiện việc nhóm dữ liệu thành các “cụm” (có thể coi là một lớp mới) để có thể phát hiện được các mẫu phân bố dữ liệu trong miền ứng dụng.
- Hướng tới việc nhận biết một tập hữu hạn các cụm hoặc các lớp để mô tả dữ liệu.
- Mục tiêu của phân cụm là cực đại hóa tính tương đồng giữa các phần tử trong cùng cụm và cực tiểu hóa tính tương đồng giữa các phần tử khác cụm.
- Phân cụm còn được gọi là “học máy không có giám sát” (unsupervised learning).

0.4.5. Hồi quy

- Là bài toán điển hình trong phân tích thống kê và dự báo.
- Tiến hành việc dự đoán các giá trị của một hoặc một số biến phụ thuộc vào giá trị của một tập hợp các biến độc lập.
- Có thể quy về việc học một hàm ánh xạ dữ liệu nhằm xác định giá trị thực của một biến theo một số biến khác.

0.4.6. Mô hình phụ thuộc

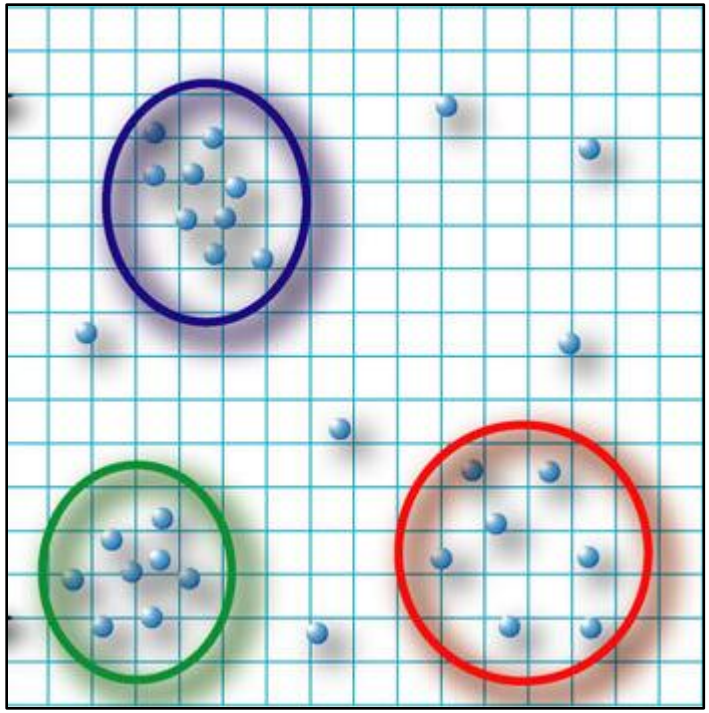
- Hướng tới việc tìm ra một mô hình mô tả sự phụ thuộc có ý nghĩa giữa các biến.
- Bao gồm 2 mức:
  - ✓ *Mức cấu trúc của mô hình*: thường dưới dạng đồ thị trong đó các biến là phụ thuộc bộ phân vào các biến khác.
  - ✓ *Mức định lượng của mô hình*: mô tả sức mạnh của tính phụ thuộc khi sử dụng việc đo tính theo giá trị số.

0.4.7. Phát hiện biến đổi và độ lệch

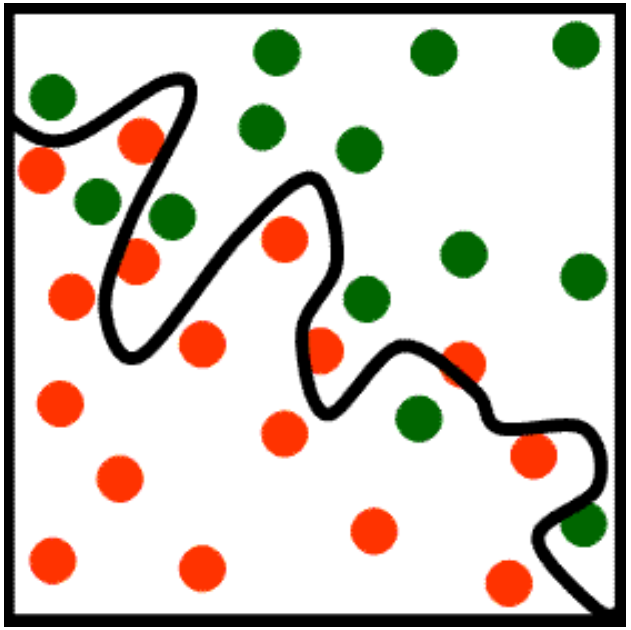
- Tập trung phát hiện hầu hết sự thay đổi có ý nghĩa dưới dạng độ đo đã biết trước hoặc giá trị chuẩn, cung cấp những tri thức về sự biến đổi và độ lệch cho người dùng. Thường được ứng dụng trong bước tiền xử lý.

{Milk, Coke} ⟶ {Sweet} (sup=30%, conf=70%)
{Beer} ⟶ {Cigar, Coffee} (sup=35%, conf = 65%)
{Coffee} ⟶ {Tea, Biscuit} (sup=22%, conf = 75%)
. . .

Khai phá Luật kết hợp

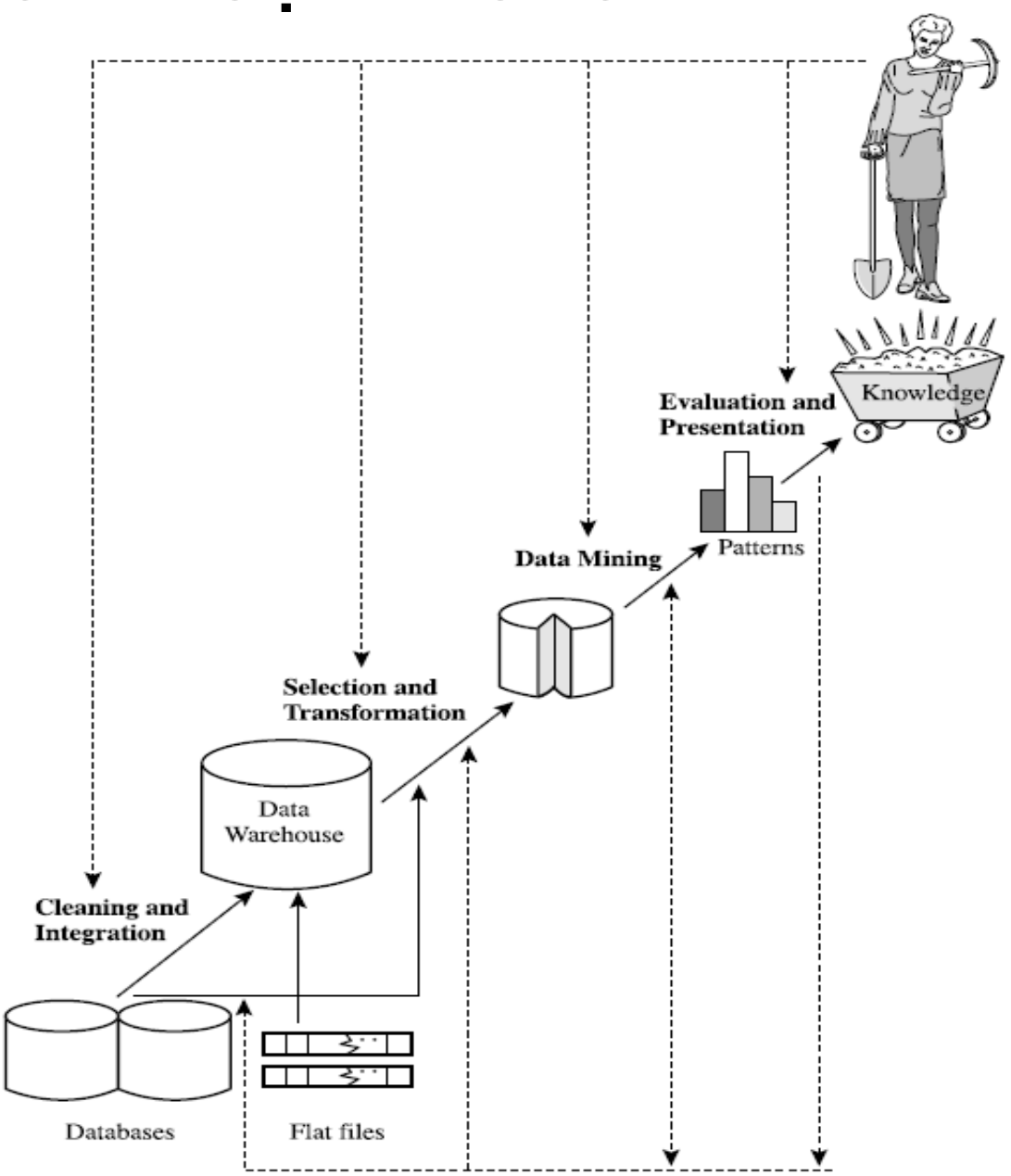


Phân cụm dữ liệu



Phân lớp dữ liệu

0.5. CÁC GIAI ĐOẠN TRONG KHAI PHÁ DỮ LIỆU



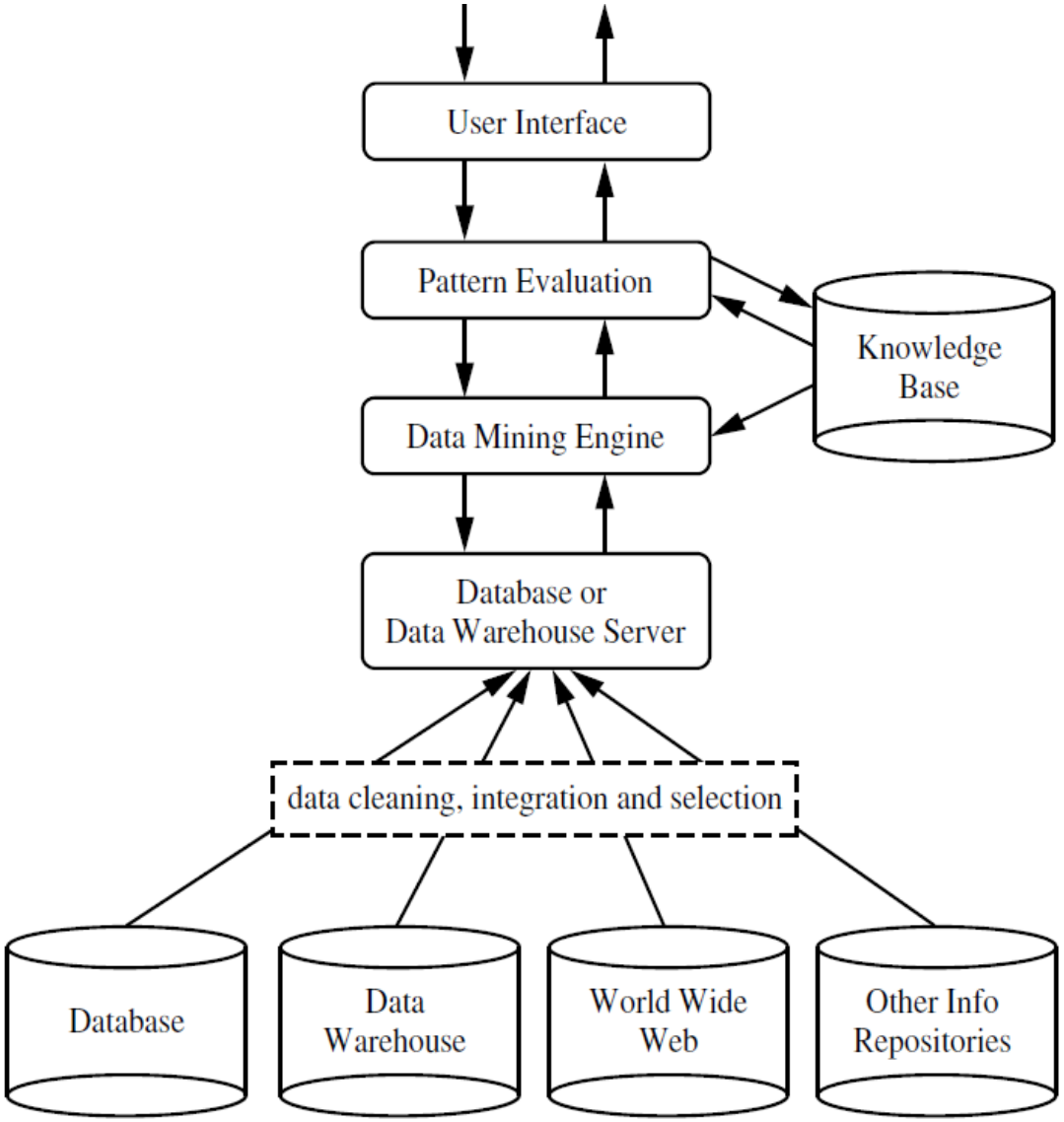
1. **Làm sạch dữ liệu (Data Cleaning):** Loại bỏ nhiễu (noisy) và các dữ liệu không nhất quán.
2. **Tích hợp dữ liệu (Data Integration):** Kết hợp dữ liệu từ các nguồn dữ liệu khác nhau.
3. **Lựa chọn dữ liệu (Data Selection):** Dữ liệu phù hợp cho thao tác phân tích được lấy về từ cơ sở dữ liệu.
4. **Chuyển dạng dữ liệu (Data Transformation):** Dữ liệu được chuyển dạng hoặc hợp nhất thành những dạng phù hợp cho quá trình khai phá bằng cách thực hiện các thao tác như tóm tắt (summary) hoặc gộp nhóm dữ liệu (aggregation).
5. **Trích chọn mẫu (Data Patterns Extracting):** Áp dụng các phương pháp "thông minh" để trích chọn ra các mẫu thực sự đáng quan tâm từ dữ liệu. Đôi khi chính bản thân bước này cũng được gọi là khai phá dữ liệu (Data Mining) (hiểu theo nghĩa hẹp).

6. **Đánh giá mẫu (Pattern Evaluation):** Dựa trên các độ đo đặc trưng, xác định ra các mẫu đáng quan tâm biểu diễn tri thức.
7. **Biểu diễn tri thức (Knowledge Presentation):** Các kỹ thuật biểu diễn tri thức và trực quan hóa (visualization) được sử dụng để biểu diễn các tri thức khai phá được đến với người dùng.

Chú ý:

Các giai đoạn từ 1. đến 4. được gọi là các giai đoạn tiền xử lý dữ liệu (data preprocessing) nhằm chuẩn bị dữ liệu cho quá trình khai phá (trích chọn mẫu).

0.6. KIẾN TRÚC ĐIỂN HÌNH CỦA MỘT HỆ THỐNG KHAI PHÁ DỮ LIỆU



1. **Cơ sở dữ liệu (Database), kho dữ liệu (Data Warehouse), World Wide Web và các nguồn chứa thông tin khác:**
   - ❖ Đây có thể là một hoặc một nhóm các cơ sở dữ liệu/kho dữ liệu hoặc các nguồn chứa thông tin (information repositories).
   - ❖ Các kỹ thuật làm sạch dữ liệu và tích hợp dữ liệu có thể được thực hiện trên các dữ liệu này.
2. **Máy chủ cơ sở dữ liệu hoặc kho dữ liệu (Database or Data Warehouse Server):**
   - ❖ Chịu trách nhiệm lấy về các dữ liệu phù hợp dựa trên yêu cầu khai phá của người dùng.
3. **Cơ sở tri thức (Knowledge Base):**
   - ❖ Đây là tri thức miền (domain knowledge) được sử dụng để dẫn hướng quá trình tìm kiếm hoặc đánh giá độ hấp dẫn của các mẫu tìm thấy.
   - ❖ Tri thức như vậy có thể bao gồm cả sự phân cấp khái niệm (concept hierarchies) (được sử dụng để tổ chức các thuộc tính và giá trị thuộc tính thành các mức trừu tượng khác nhau).

4. Engine khai phá dữ liệu (Data Mining Engine):

- Đây là thành phần chủ yếu của một hệ thống KPDL.
- Bao gồm các module thực hiện các tác vụ như phân tích đặc trưng (characterization) và quan hệ kết hợp (association/correlation analysis), phân lớp (classification), dự đoán (prediction), phân tích cụm (cluster analysis),…

5. Module đánh giá mẫu (Pattern Evaluation Module):

- Sử dụng các độ đo hấp dẫn và có sự tương tác với engine khai phá dữ liệu nhằm tập trung vào việc tìm ra các mẫu đáng quan tâm. Có thể sử dụng ngưỡng độ hấp dẫn để lọc bớt các mẫu tìm được.
- Có thể được tích hợp với module khai phá tùy thuộc vào phương pháp khai phá được sử dụng và cách thức cài đặt.
- Khuyến khích: Thao tác đánh giá mẫu cần được tích hợp càng chặt chẽ càng tốt với tiến trình khai phá nhằm nâng cao hiệu quả khai phá (giới hạn việc tìm kiếm chỉ với các mẫu đáng quan tâm).

4. **Giao diện người sử dụng (User Interface):** Module này làm nhiệm vụ giao tiếp giữa người dùng và hệ thống KPDL:

- ❖ Cho phép người dùng tương tác với hệ thống bằng cách chỉ ra truy vấn hoặc tác vụ khai phá mong muốn.
- ❖ Cung cấp thông tin giúp cho thao tác tìm kiếm được tập trung.
- ❖ Thực hiện khai phá thăm dò (Exploratory Data Mining) dựa trên các kết quả khai phá trung gian.
- ❖ Cho phép người dùng duyệt cơ sở dữ liệu, lược đồ kho dữ liệu và các cấu trúc dữ liệu, đánh giá các mẫu được khai phá và biểu diễn trực quan mẫu dưới các dạng thức khác nhau.

0.7. CÁC NGUỒN DỮ LIỆU PHỤC VỤ CHO KHAI PHÁ

1. CƠ SỞ DỮ LIỆU QUAN HỆ (RELATIONAL DATABASE)

customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	. . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*cost*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
. . .	. . .	. . .	. . .	. . .	. . .

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
. . .	. . .	. . .

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
. . .	. . .	. . .	. . .	. . .	. . .	. . .

items_sold

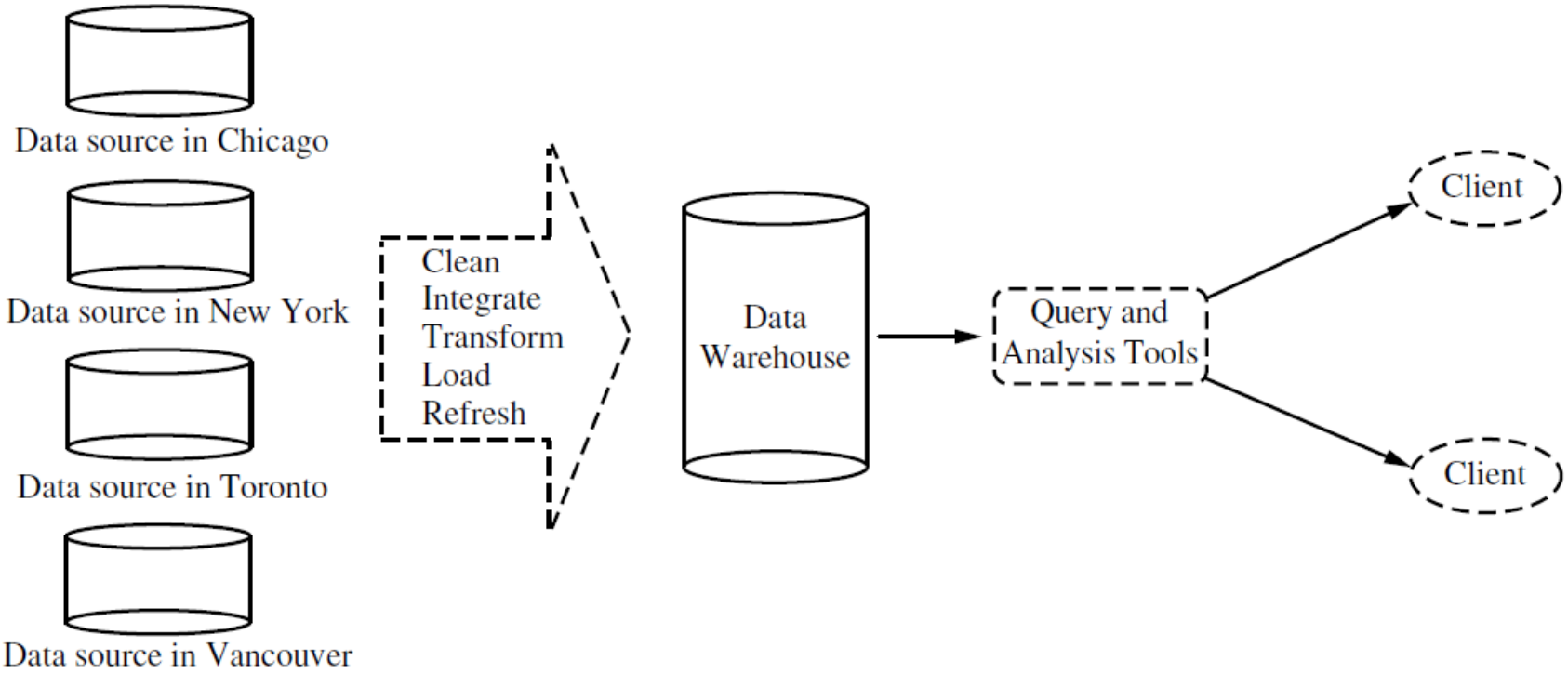
trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
. . .	. . .	. . .

works_at

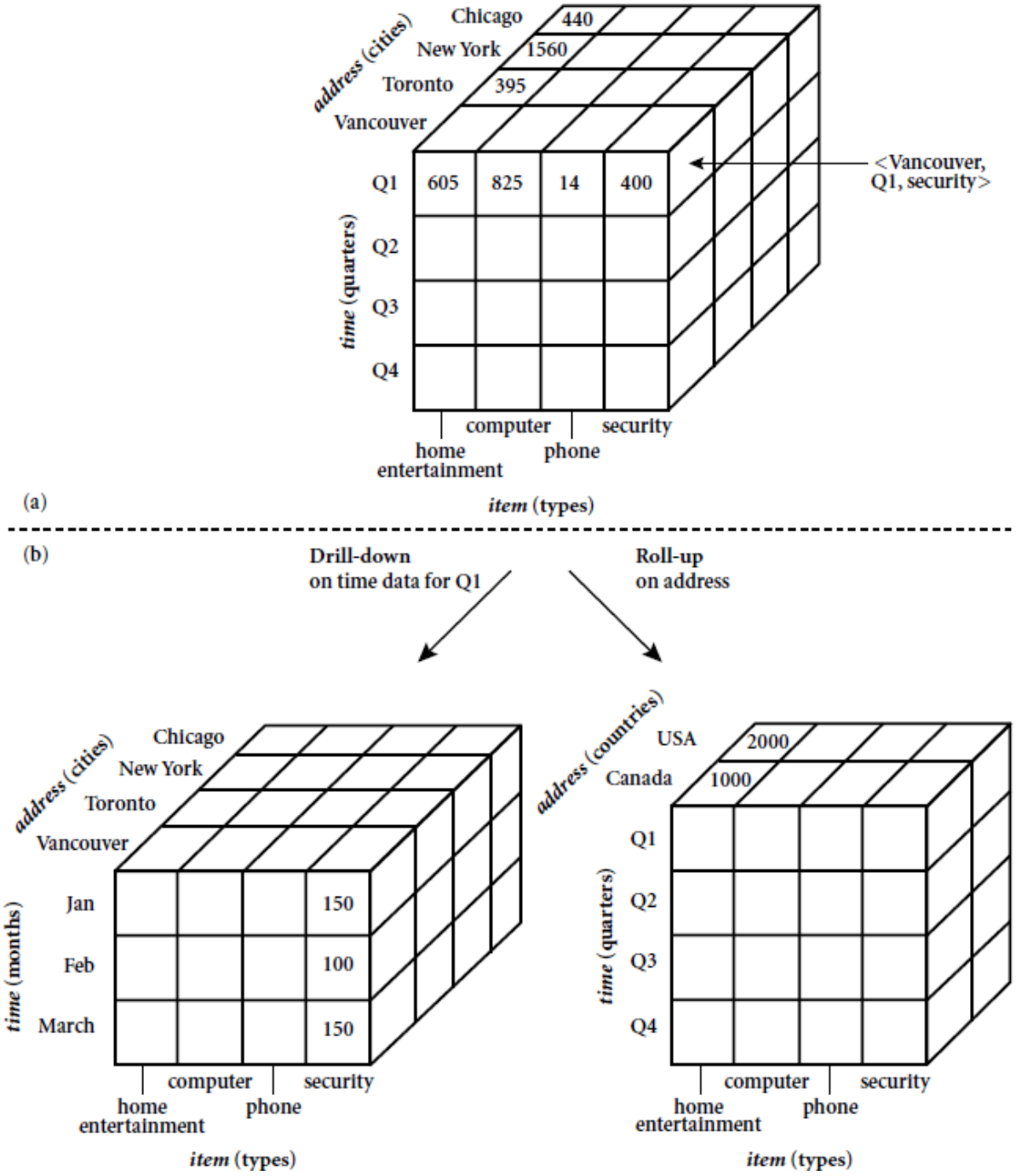
empl_ID	*branch_ID*
E55	B1
. . .	. . .

2. KHO DỮ LIỆU (DATA WAREHOUSE)

- Là nơi tập trung dữ liệu từ nhiều nguồn khác nhau (multiple sources) được lưu trữ dưới một lược đồ thống nhất (unified shema) và được tập trung tại một nơi.
- Được xây dựng thông qua các tiến trình *làm sạch dữ liệu* (data cleaning), *tích hợp dữ liệu* (data integration), *chuyển dạng dữ liệu* (data transformation), *tải dữ liệu* (data loading) và *làm tươi dữ liệu định kỳ* (periodic data refreshing).



- Để thuận tiện cho việc ra quyết định, dữ liệu trong kho dữ liệu thường được tổ chức xoay quanh các chủ đề chính đáng quan tâm như khách hàng (customer), hàng hóa (item), nhà cung cấp (supplier),…
- Dữ liệu được lưu trữ nhằm cung cấp thông tin dựa trên một cái nhìn toàn cảnh về dữ liệu tác nghiệp của doanh nghiệp trong khoảng từ 5 -10 năm và thường được tóm tắt (summarized) để thuận tiện cho xử lý.
- Kho dữ liệu thường được mô hình hóa dưới dạng một cấu trúc cơ sở dữ liệu đa chiều (multidimensional database structure), ở đó mỗi chiều tương ứng với một thuộc tính hoặc tập thuộc tích của lược đồ và mỗi ô (cell) lưu trữ giá trị của một số đại lượng được gộp nhóm.
- Cấu trúc vật lý thực sự của kho dữ liệu có thể là dưới dạng một cơ sở dữ liệu quan hệ hoặc một data cube đa chiều. Một data cube cung cấp cái nhìn đa chiều về dữ liệu và cho phép thực hiện các thao tác tiền tính toán (precomputation) và truy cập nhanh tới dữ liệu đã được tóm tắt.


address (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Q1
Q2
Q3
Q4
605
825
14
400
<Vancouver, Q1, security>
time (quarters)
computer
security
home entertainment
phone
item (types)
(a)
(b)
Drill-down on time data for Q1
Roll-up on address
address (cities)
Chicago
New York
Toronto
Vancouver
Jan
Feb
March
150
100
150
time (months)
computer
security
home entertainment
phone
item (types)
address (countries)
USA
Canada
2000
1000
Q1
Q2
Q3
Q4
time (quarters)
computer
security
home entertainment
phone
item (types)

3. CƠ SỞ DỮ LIỆU GIAO DỊCH (TRANSACTION DATABASE)

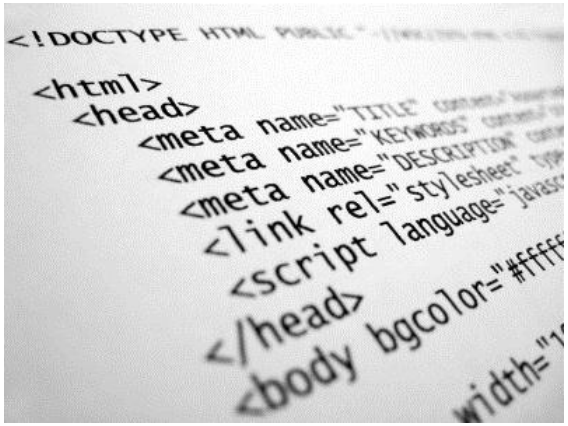
- Cơ sở dữ liệu giao dịch là một tập hợp các giao dịch. Mỗi giao dịch bao gồm một *số hiệu giao dịch* (trans_ID) và danh sách các mục (item) cấu thành giao dịch.

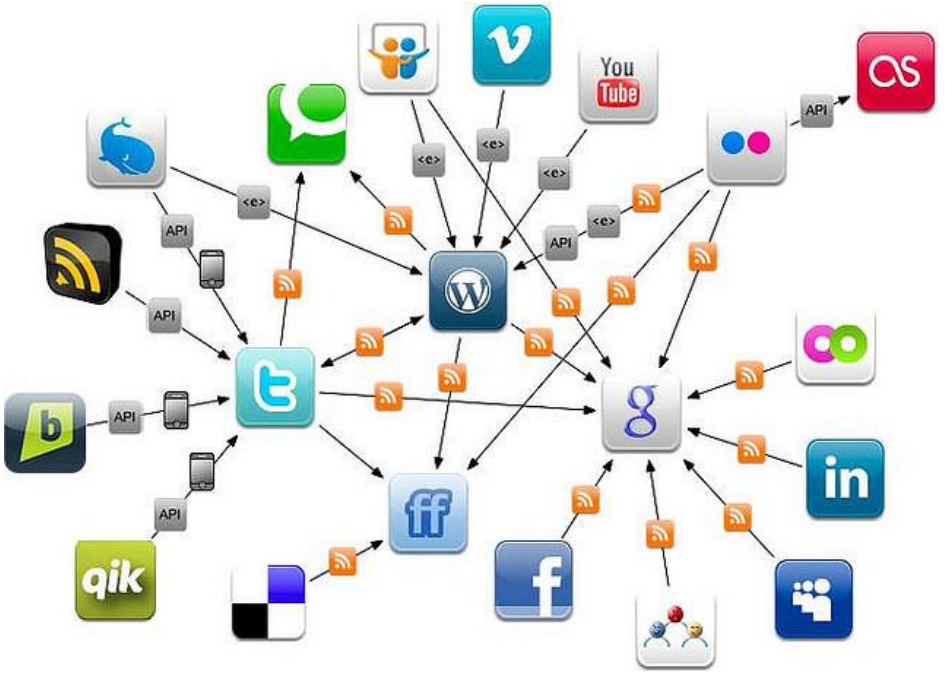
Trans_ID	Item List
T1	Milk, Bread, Coke
T2	Beer, Bread
T3	Beer, Milk, Diaper, Coke
T4	Beer, Milk, Diaper, Bread
T5	Milk, Diaper, Coke

TID	Beer	Milk	Diaper	Bread	Coke
T1	0	1	0	1	1
T2	1	0	0	1	0
T3	1	1	1	0	1
T4	1	1	1	1	0
T5	0	1	1	0	1

4. CÁC DẠNG DỮ LIỆU NÂNG CAO

- **Dữ liệu văn bản:** bao gồm các dạng có cấu trúc, bán cấu trúc hoặc không có cấu trúc.
- **Dữ liệu Multimedia:** hình ảnh, âm thanh, video,…
- **Dữ liệu World Wide Web:** dữ liệu nội dung web, dữ liệu cấu trúc web, dữ liệu sử dụng web.





0.6. ỨNG DỤNG CỦA KHAI PHÁ DỮ LIỆU

- **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.***

Q & A