TRƯỜNG ĐẠI HỌC TRÀ VINH
BỘ MÔN CÔNG NGHỆ THÔNG TIN

BÀI GIẢNG MÔN HỌC

KHAI KHOÁNG DỮ LIỆU

CHƯƠNG 1. TỔNG QUAN VỀ KHAI KHOÁNG DỮ LIỆU

Giảng viên: Ths. Hà Thị Thúy Vi

Email: hattvi201084@tvu.edu.vn

SĐT: 0983.001084

TÀI LIỆU HỌC TẬP

1. **Giáo trình/Tài liệu học tập chính**

   Jiawei Han, Micheline Kamber, *Data mining: Concepts and Techniques 3rd Edition*, Morgan Kaufmann Publishers, 2011.

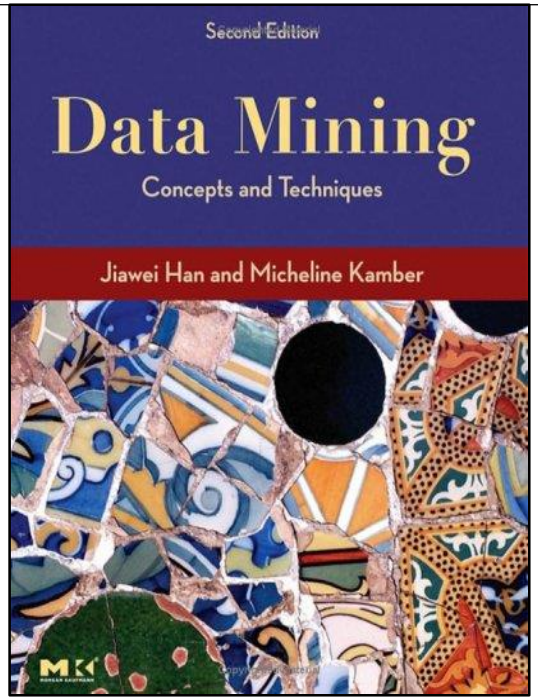
2. **Tài liệu tham khảo thêm**

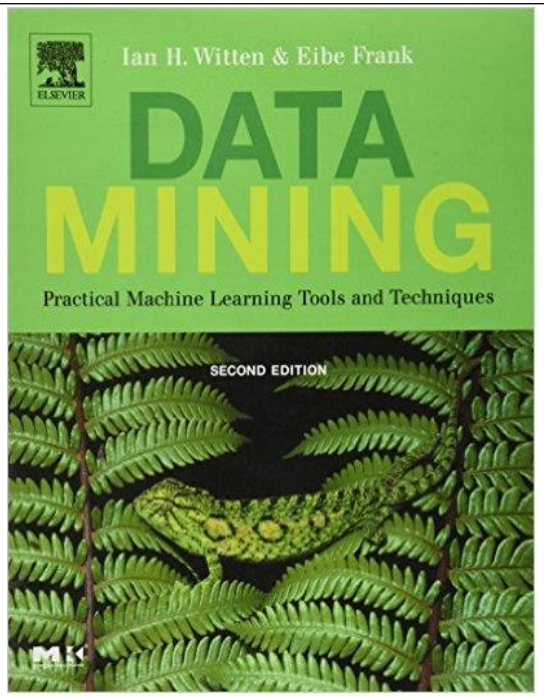
[1]. Jamie MacLennan, ZhaoHui Tang & Bogdan Crivat, *Data Mining with Microsoft SQL Server 2008*, Wiley Publishing, 2008.

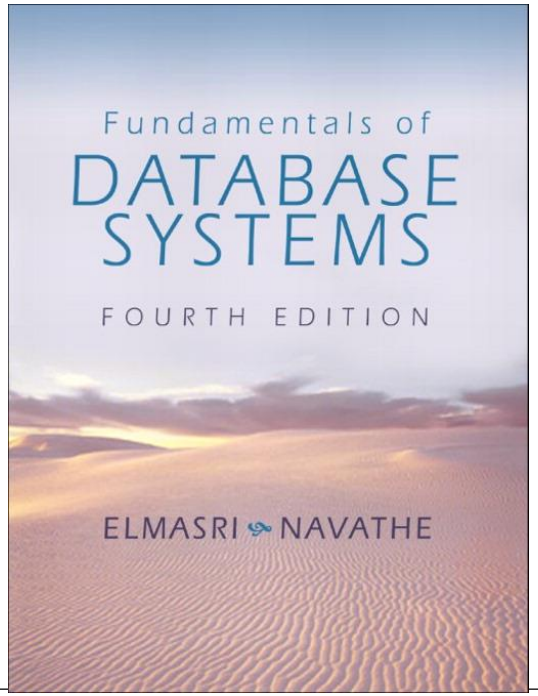
[2]. Đỗ Thanh Nghị - *Khai mỏ dữ liệu bằng ngôn ngữ R* – NXB Đại học Cần Thơ.

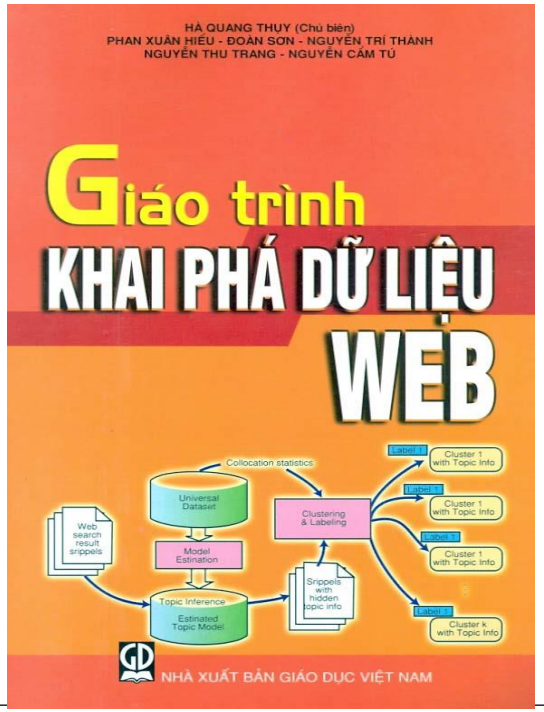
[3]. Đỗ Phúc – *Giáo Trình Khai Thác Dữ Liệu* – NXB Đại học Quốc Gia TP.HCM.

[4]. Hà Quang Thụy, Phan Xuân Hiếu, Đoàn Sơn, Nguyễn Trí Thành, Nguyễn Thu Trang, Nguyễn Cẩm Tú, *Giáo trình Khai khoáng dữ liệu Web*, NXB Giáo dục, 2009


Second Edition
Data Mining
Concepts and Techniques
Jiawei Han and Micheline Kamber


Ian H. Witten & Eibe Frank
DATA MINING
Practical Machine Learning Tools and Techniques
SECOND EDITION


Fundamentals of
DATABASE SYSTEMS
FOURTH EDITION
ELMASRI NAVATHE


HÀ QUANG THỤY (Chủ biên)
PHAN XUÂN HIẾU - ĐOÀN SƠN - NGUYỄN TRÍ THÀNH
NGUYỄN THU TRANG - NGUYỄN CẨM TÚ
Giáo trình
KHAI PHÁ DỮ LIỆU WEB
Collocation statistics
Universal Dataset
Model Estimation
Topic Inference
Estimated Topic Model
Web search result snippets
Clustering & Labeling
Snippets with hidden topic info
Cluster 1 with Topic Info
Cluster 1 with Topic Info
Cluster 1 with Topic Info
Cluster k with Topic Info
NHÀ XUẤT BẢN GIÁO DỤC VIỆT NAM

NỘI DUNG CHÍNH

1.1. NHU CẦU KHAI KHOÁNG DỮ LIỆU

1.2. KHAI KHOÁNG DỮ LIỆU LÀ GÌ?

1.3. CÁC KỸ THUẬT KHAI KHOÁNG DỮ LIỆU CƠ BẢN

1.4. CÁC GIAI ĐOẠN TRONG KHAI KHOÁNG DỮ LIỆU

1.5. KIẾN TRÚC ĐIỂN HÌNH CỦA MỘT HỆ THỐNG KKDL

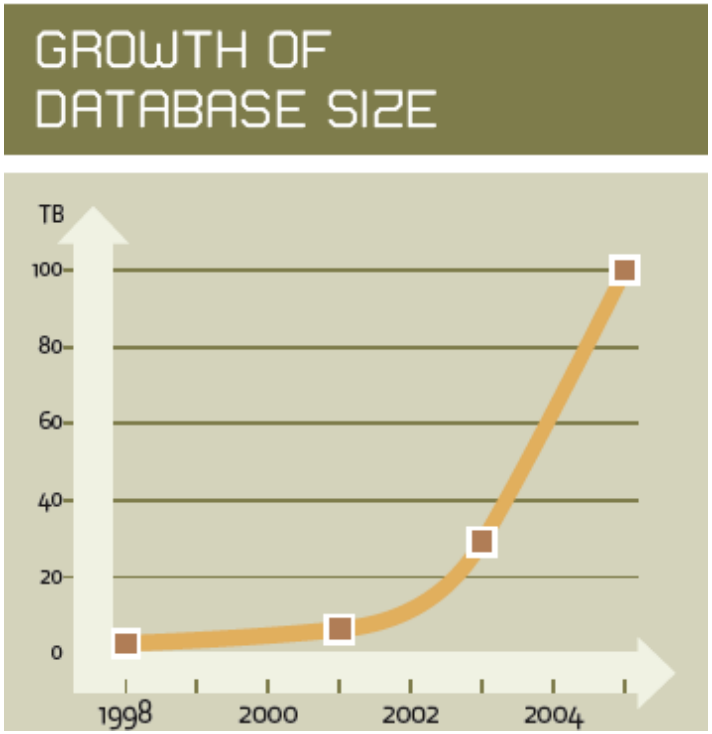
1.6. CÁC NGUỒN DỮ LIỆU PHỤC VỤ CHO KHAI KHOÁNG

1.7. ỨNG DỤNG CỦA KHAI KHOÁNG DỮ LIỆU

1.1. NHU CẦU KHAI KHOÁNG 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

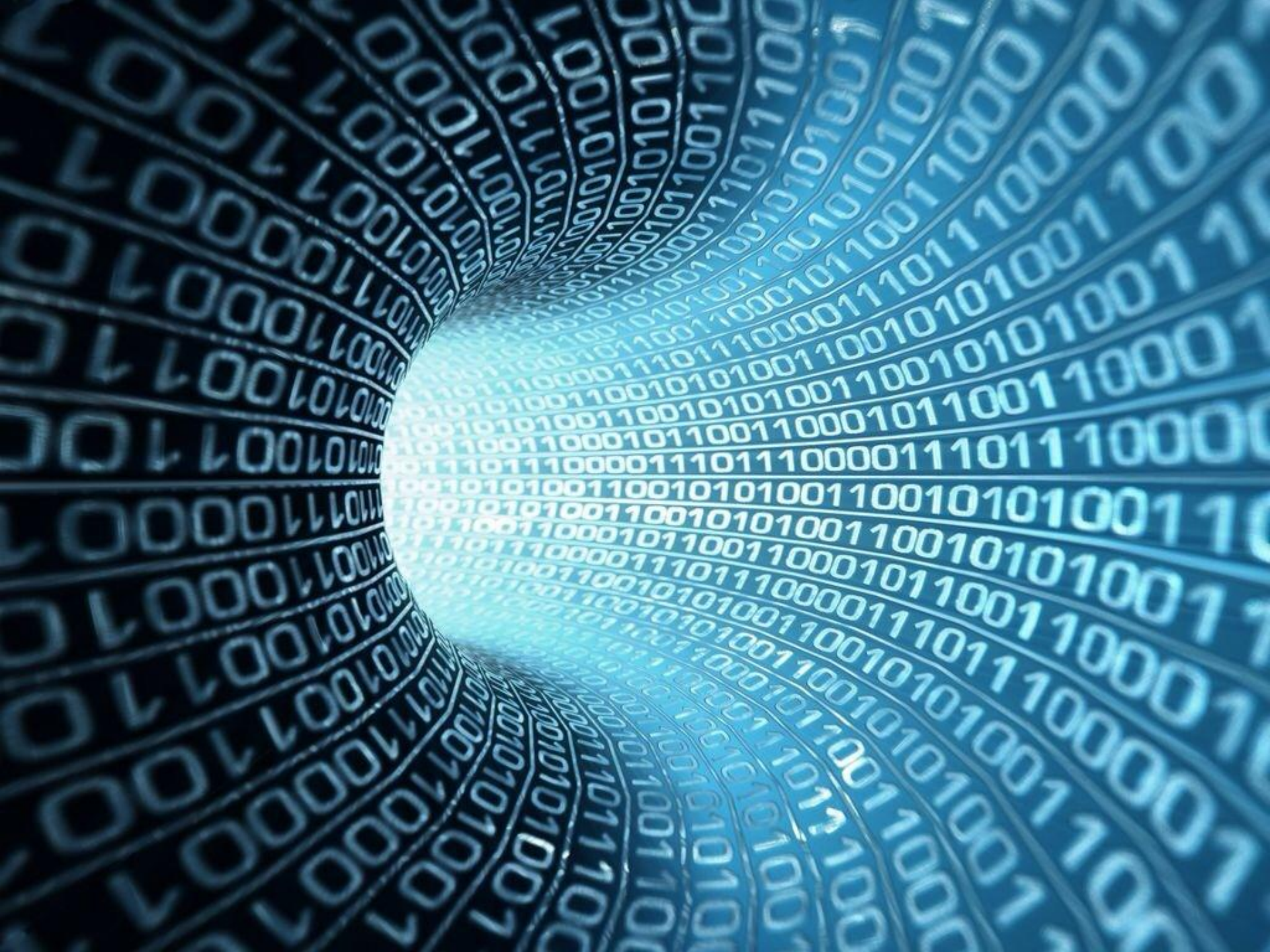

Alexa
The Web Information Company


YouTube

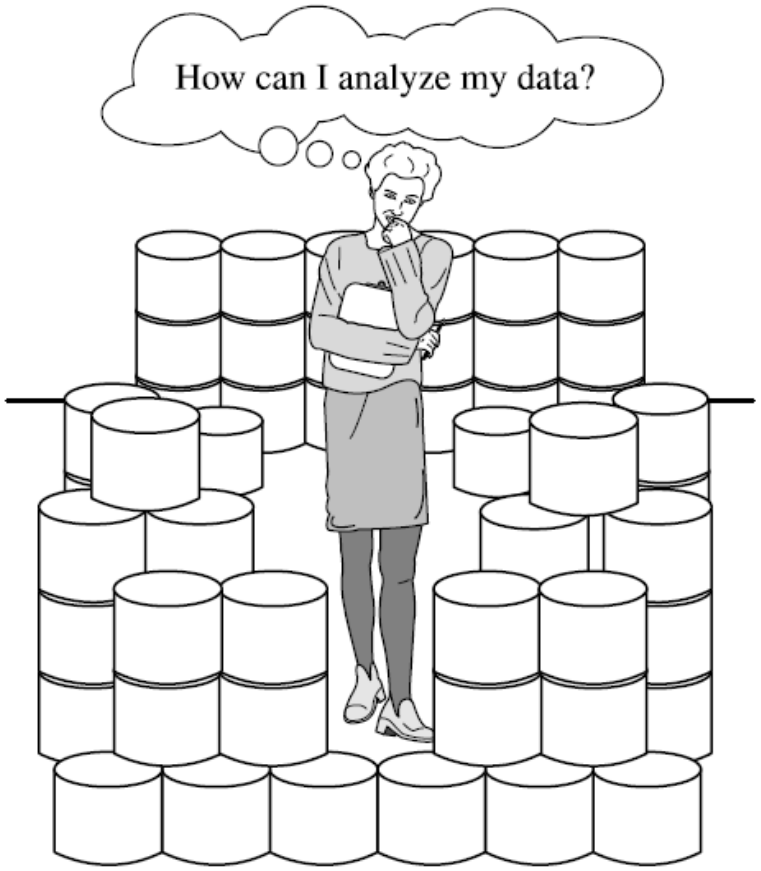

altavista

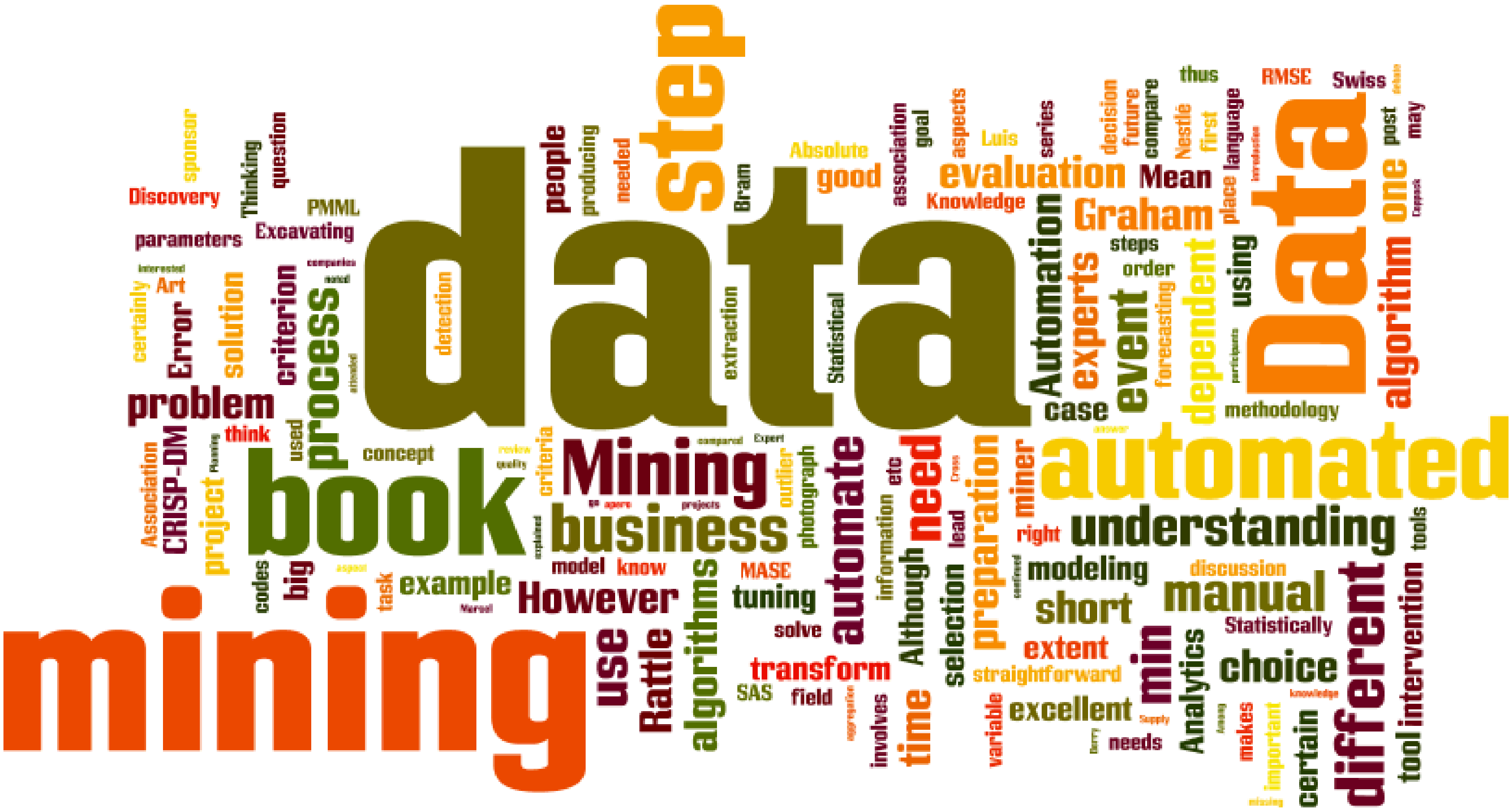

twitter


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 khoáng dữ liệu (mining)**
  - Tạo lập kho dữ liệu và quá trình phân tích dữ liệu trực tuyến OLAP.
  - Khai khoáng tri thức hấp dẫn (luật, quy luật, mẫu, ràng buộc) từ dữ liệu trong CSDL lớn.




data
mining
Data
automated
step
book
Mining
business
understanding
automate
manual
different
algorithm
Automation
evaluation
Graham
problem
process
need
preparation
However
algorithms
Rattle
use
choice
short
modeling
excellent
tuning
transform
event
expert
dependent
extent
straightforward
min
Analytics
certain
intervention
tool
tools
methodology
criterion
solution
Error
CRISP-DM
project
Association
Discovery
parameters
Excavating
PMML
Thinking
question
sponsor
Art
certainly
think
concept
example
model
know
MASE
solve
SAS
field
time
involves
Although
information
selection
lead
right
miner
variable
needs
makes
important
Statistically
discussion
case
order
steps
using
forecasting
Mean
Knowledge
good
Absolute
Luis
series
aspects
goal
association
Bram
people
producing
needed
detection
extraction
Statistical
criteria
outlier
photograph
etc
big
codes
task
used
decision
future
compare
Nestlé
first
language
place
thus
RMSE
Swiss
post
may
one

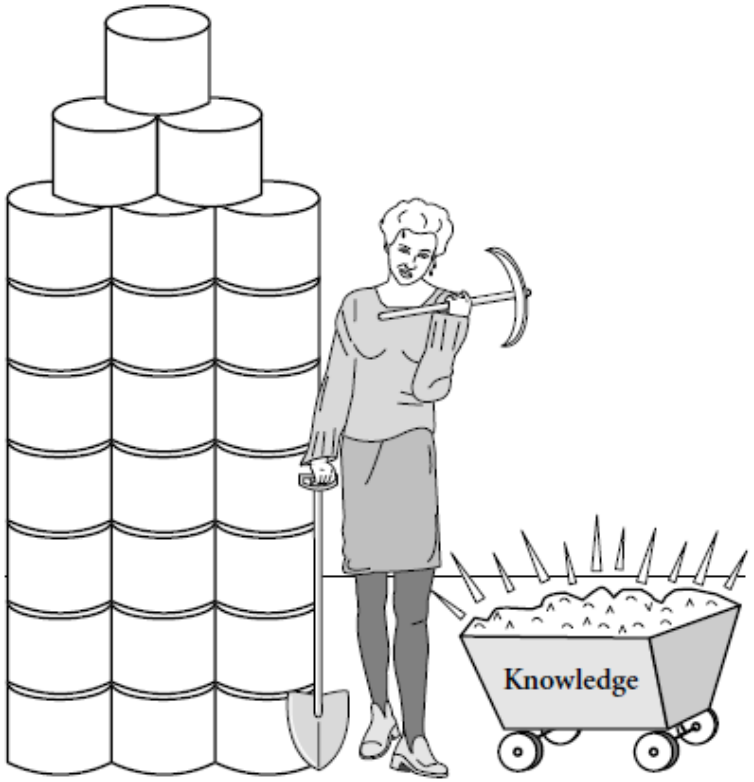
1.2. KHAI KHOÁNG DỮ LIỆU LÀ GÌ?

Theo J.Han và M.Kamber (2006) [1]:

Thứ nhất:

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

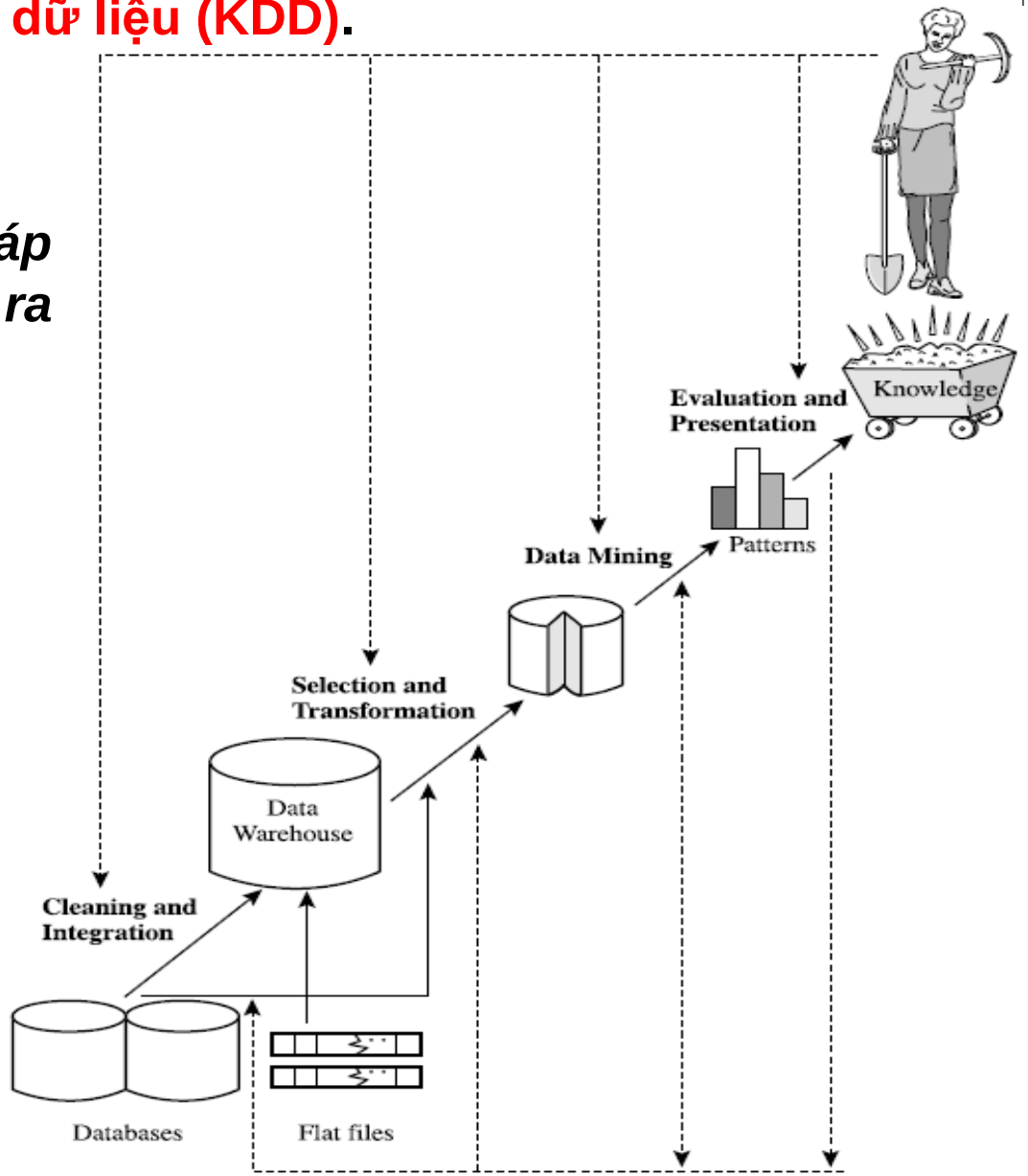
Khai khoáng dữ liệu = Phát hiện tri thức từ dữ liệu (KDD: Knowledge Discovery From Data).



Thứ 2:

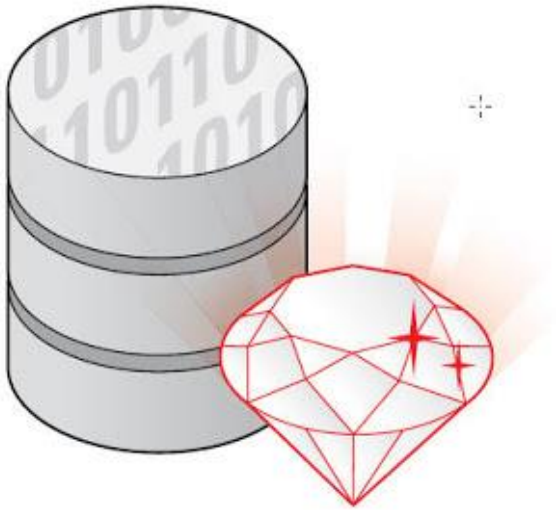
Khai khoáng dữ liệu (Data Mining) chỉ là một bước quan trọng trong quá trình phát hiên tri thức từ dữ liệu (KDD).

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 khoáng 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 khoáng dữ liệu là một bước trong quá trình phát hiện tri thức trong cơ sở dữ liệu, thi hành một thuật toán khai khoáng dữ liệu để tìm ra các mẫu từ dữ liệu theo khuôn dạng thích hợp

1.3. CÁC PHƯƠNG PHÁP, KỸ THUẬT CHÍNH TRONG KHAI KHOÁNG DỮ LIỆU

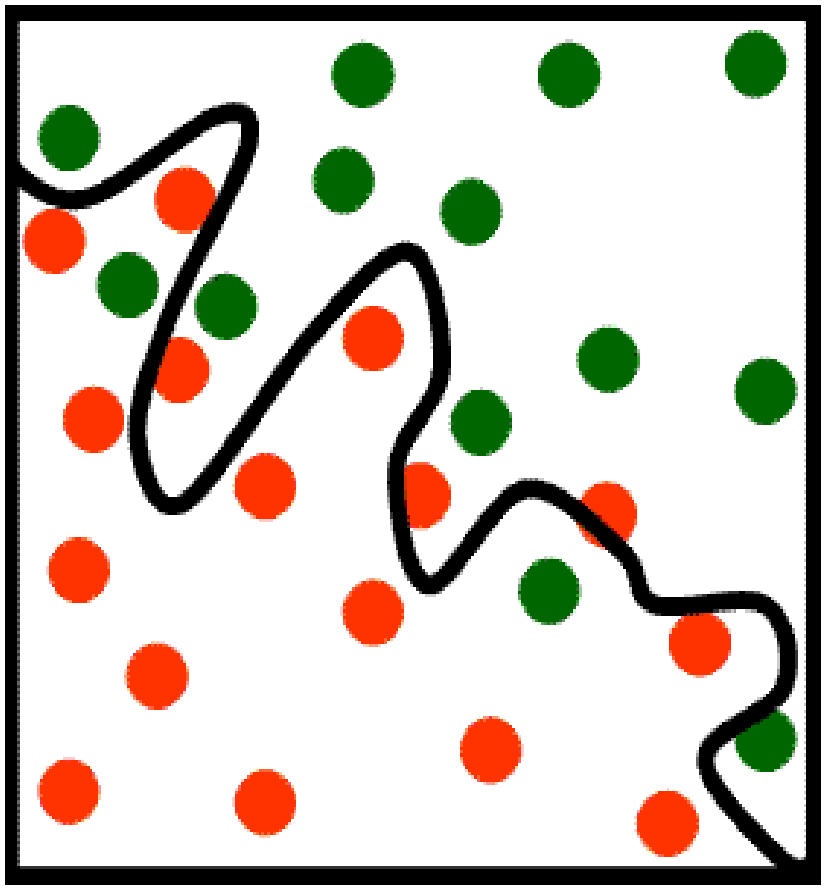
Các kỹ thuật KKDL được có thể chia làm 2 nhóm chính:

- ***Kỹ thuật KKDL mô tả:*** *có nhiệm vụ mô tả về các tính chất hoặc các đặc tính chung của dữ liệu trong CSDL hiện có. Nhóm kỹ thuật này gồm các phương pháp: phân nhóm (Clustering), tổng hợp hóa (Summerization), phát hiện sự biến đổi và độ lệch (Change and deviation detection), phân tích luật kết hợp (Association Rules), ...*
- ***Kỹ thuật KKDL dự đoán:*** *có nhiệm vụ đưa ra các dự đoán dựa vào các suy diễn trên dữ liệu hiện thời. Nhóm kỹ thuật này gồm các phương pháp: phân lớp (Classification), hồi quy (Regression), ...*

1.3.1. PHƯƠNG PHÁP PHÂN LỚP

Phân lớp(<u>Classification):</u> *là đặt các mẫu vào các lớp được xác định trước. Nhiệm vụ chính là tìm các hàm ánh xạ các mẫu dữ liệu một cách chính xác vào trong các lớp.*

Ví dụ*:* một ngân hàng muốn phân loại các khách hàng của họ vào trong hai nhóm có nợ hay không nợ, từ đó giúp họ ra quyết định cho vay hay không cho vay.



1.3.2. PHƯƠNG PHÁP HỒI QUY

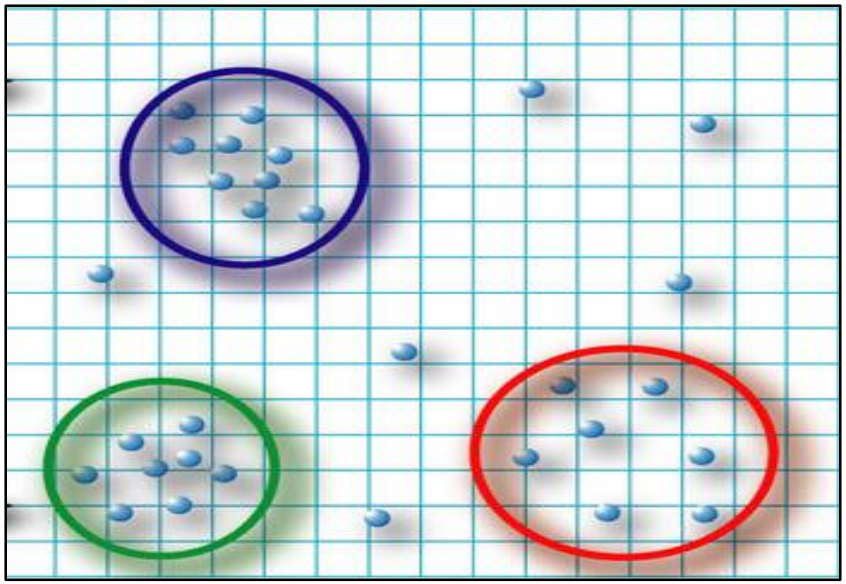
Hồi quy (Regression): *là việc học một hàm ánh xạ từ một mẫu dữ liệu thành một biến dự đoán có giá trị thực. Nhiệm vụ của hồi quy tương tự như phân lớp, điểm khác nhau chính là ở chỗ thuộc tính để dự báo là liên tục chứ không rời rạc.*

Việc dự báo các giá trị số thường được làm bởi các phương pháp thống kê cổ điển chẳng hạn như hồi quy tuyến tính. Tuy nhiên phương pháp mô hình hóa cũng có thể được sử dụng như cây quyết định.

1.3.3. PHƯƠNG PHÁP PHÂN NHÓM – ĐOẠN

Phân nhóm – đoạn (Clustering / Segmentation): là nhóm các đối tượng tương tự nhau trong tập dữ liệu vào các nhóm sao cho mức độ tương tự giữa các đối tượng trong cùng một nhóm là lớn nhất và mức độ tương tự giữa các đối tượng nằm trong các nhóm khác nhau là nhỏ nhất. Các nhóm có thể tách nhau hoặc phân cấp gối lên nhau và số lượng các nhóm là chưa biết trước.

Phân nhóm dữ liệu được sử dụng nhiều trong các ứng dụng về phân đoạn thị trường, phân đoạn khách hàng, nhận dạng mẫu, phân loại trang Web, …



1.3.4. LUẬT KẾT HỢP (ASSOCIATION RULES)

Luật kết hợp (Association Rules): là dạng luật biểu diễn tri thức ở dạng tương đối đơn giản. Mục tiêu của phương pháp này là phát hiện và đưa ra các mối liên hệ giữa các giá trị dữ liệu trong CSDL. Mẫu đầu ra của giải thuật KKDL là tập luật kết hợp tìm được.

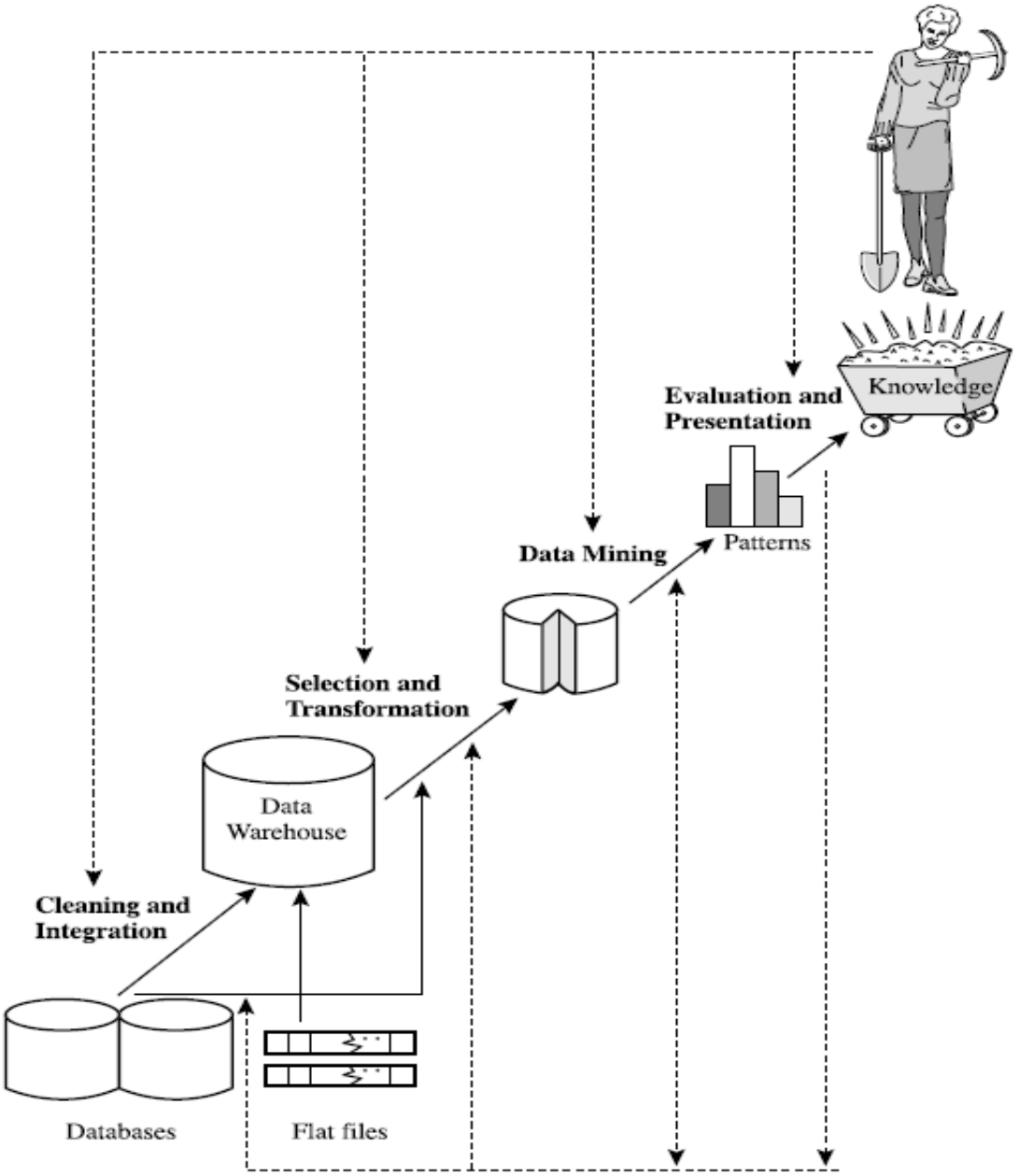
{Milk, Coke} ⟶ {Sweet} (sup=30%, conf=70%)

{Beer} ⟶ {Cigar, Coffee} (sup=35%, conf = 65%)

{Coffee} ⟶ {Tea, Biscuit} (sup=22%, conf = 75%)

. . .

1.4. CÁC GIAI ĐOẠN TRONG KHAI KHOÁNG DỮ LIỆU



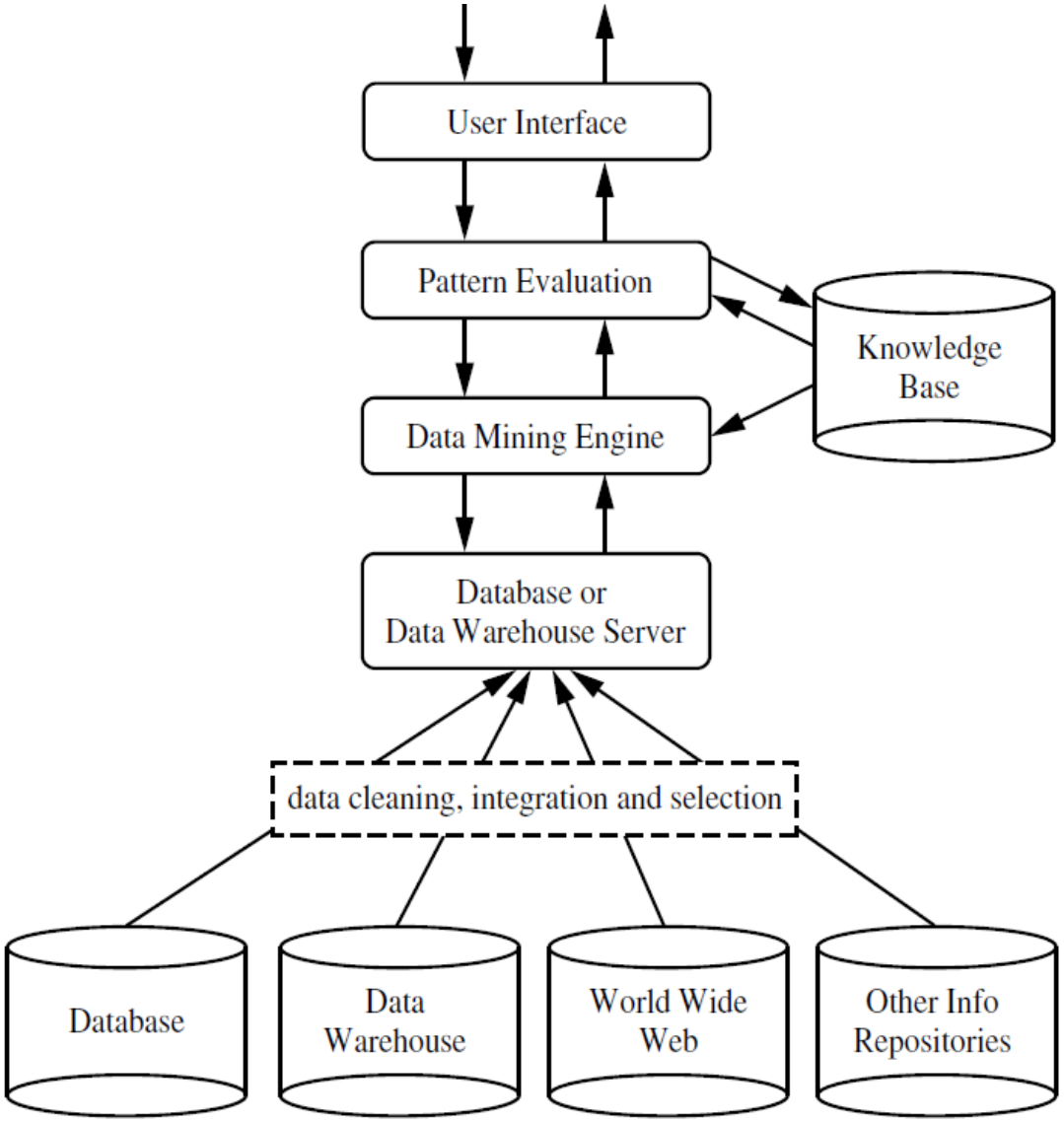
1. **Làm sạch dữ liệu (Data Cleaning):** Loại bỏ nhiễu (noisy) và các dữ liệu không nhất quán.
2. **Tích hợp dữ liệu (Data Integration):** Kết hợp dữ liệu từ các nguồn dữ liệu khác nhau.
3. **Lựa chọn dữ liệu (Data Selection):** Dữ liệu phù hợp cho thao tác phân tích được lấy về từ cơ sở dữ liệu.
4. **Chuyển dạng dữ liệu (Data Transformation):** Dữ liệu được chuyển dạng hoặc hợp nhất thành những dạng phù hợp cho quá trình khai khoáng 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 khoáng 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 khoáng 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 khoáng (trích chọn mẫu).

1.5. KIẾN TRÚC ĐIỂN HÌNH CỦA MỘT HỆ THỐNG KHAI KHOÁNG 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 khoáng 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 khoáng dữ liệu (Data Mining Engine):

- Đây là thành phần chủ yếu của một hệ thống KKDL.
- 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 khoáng 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 khoáng nhằm nâng cao hiệu quả khai khoáng (giới hạn việc tìm kiếm chỉ với các mẫu đáng quan tâm).

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

- Cho phép người dùng tương tác với hệ thống bằng cách chỉ ra truy vấn hoặc tác vụ khai khoáng mong muốn.
- Cung cấp thông tin giúp cho thao tác tìm kiếm được tập trung.
- Thực hiện khai khoáng thăm dò (Exploratory Data Mining) dựa trên các kết quả khai khoáng 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 khoáng và biểu diễn trực quan mẫu dưới các dạng thức khác nhau.

1.6. CÁC NGUỒN DỮ LIỆU PHỤC VỤ CHO KHAI KHOÁNG

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

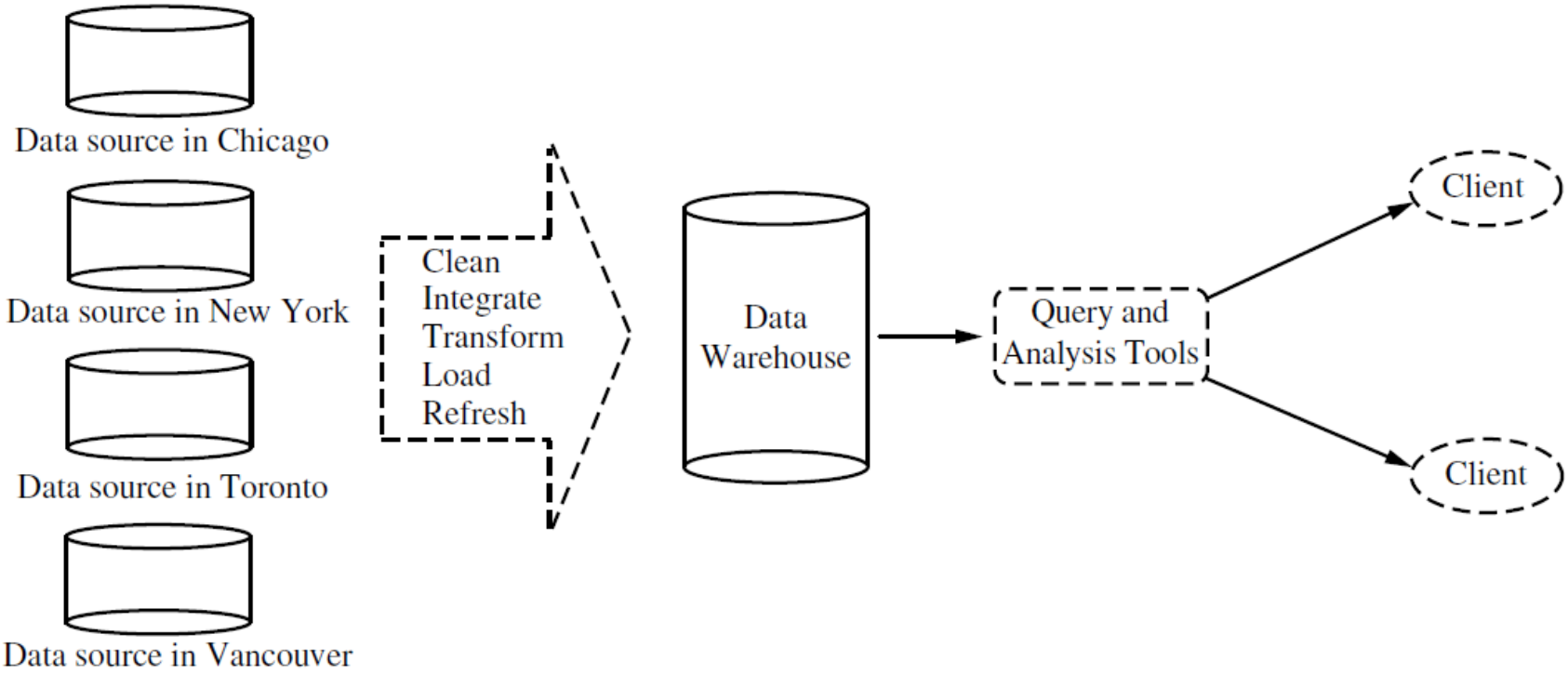
trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
. . .	. . .	. . .

works_at

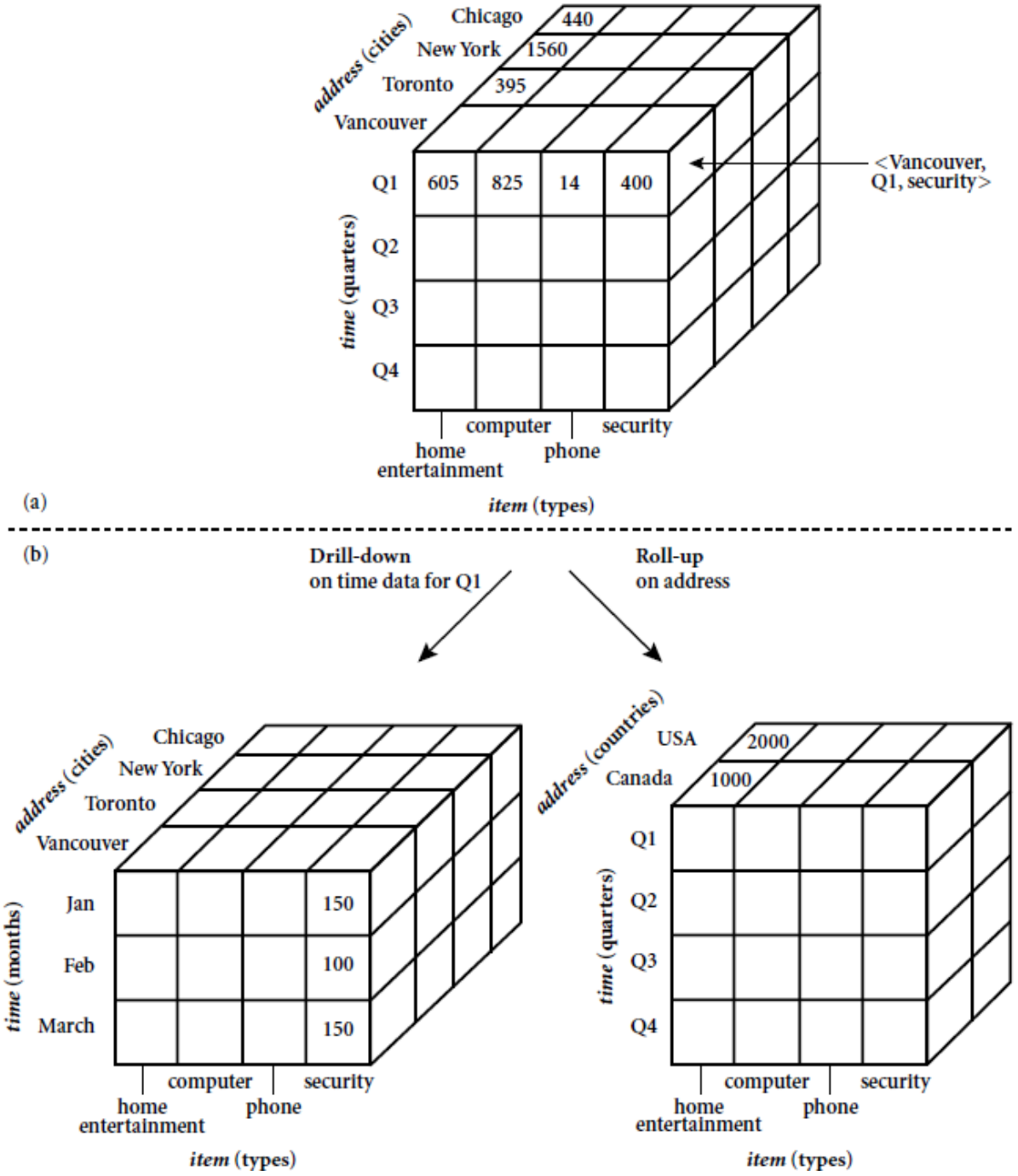
empl_ID	*branch_ID*
E55	B1
. . .	. . .

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
time (quarters)
Q1
Q2
Q3
Q4
605
825
14
400
<Vancouver, Q1, security>
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
time (months)
Jan
Feb
March
150
100
150
computer
security
home entertainment
phone
item (types)
address (countries)
USA
Canada
2000
1000
time (quarters)
Q1
Q2
Q3
Q4
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

Milk, Bread, Coke
10:05

Beer, Bread
10:12

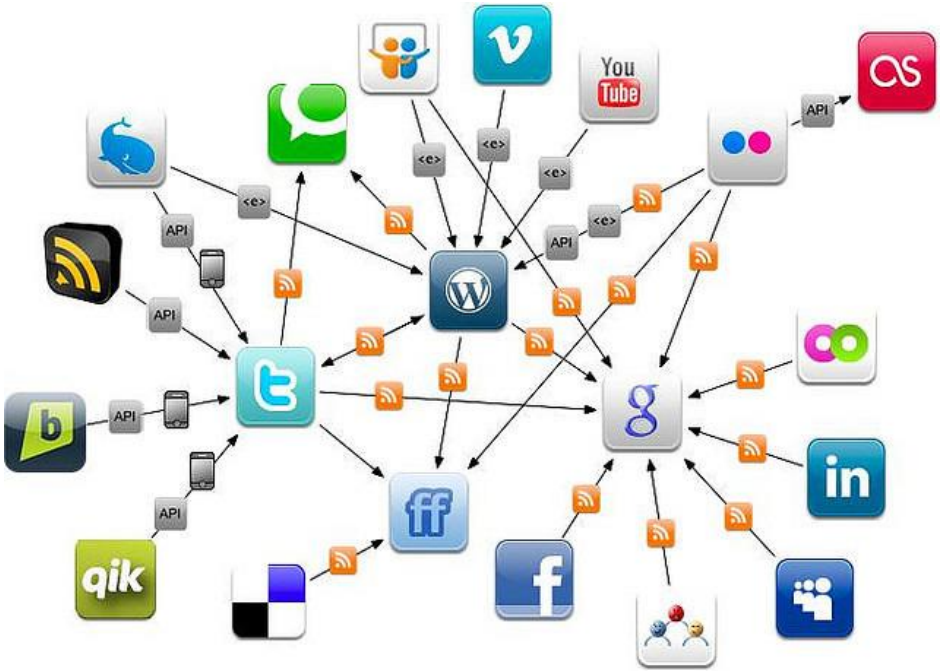
Beer, Milk, Diaper, Coke
10:15

Beer, Milk, Diaper, Bread
10:33

Milk, Diaper, Coke
10:30

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.



1.7. ỨNG DỤNG CỦA KHAI KHOÁNG 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 khoáng Text (nhóm mới, email, tài liệu) và khai khoáng Web.***
  - ***Khai khoáng dữ liệu dòng.***
  - ***Phân tích DNA và dữ liệu sinh học.***

CÁC CÔNG CỤ KHAI KHOÁNG DỮ LIỆU

- ***RapidMiner:*** Là một trong những công cụ phổ biến nhất để khai khoáng dữ liệu. Nó cung cấp các chức năng khai khoáng dữ liệu khác nhau như tiền xử lý dữ liệu, biểu diễn dữ liệu, lọc, phân cụm, v.v.
- **Weka:** là một phần mềm khai thác dữ liệu mã nguồn mở. Giống như RapidMiner, Weka không có mã hóa và sử dụng GUI đơn giản. cung cấp một loạt các công cụ như trực quan hóa, tiền xử lý, phân loại, phân cụm, v.v.
- **Orange:** Phần mềm Orange được biết đến bởi việc tích hợp các công cụ khai khoáng dữ liệu và học máy. Nó được viết bằng Python và cung cấp trực quan tương tác và thẩm mỹ cho người dùng.
- **Một số công cụ khác như:** TeraData, Oracle DataMining, Knime, …

Q & A