



Trần Thị Liên
tranthilien@qnu.edu.vn

KHAI PHÁ DỮ LIỆU



NỘI DUNG

SỐ TÍN CHỈ: 3

Chương 1. GIỚI THIỆU TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

Chương 2. THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU

Chương 3. KHAI PHÁ LUẬT KẾT HỢP

Chương 4. PHÂN LỚP DỮ LIỆU

Chương 5. PHÂN CỤM



Tài liệu tham khảo

- [HK06] J. Han and M. Kamber (2006). [Data Mining-Concepts and Techniques \(Second Edition\)](#), Morgan Kaufmann.
- [NEM09] Robert Nisbet, John Elmer, and Gary Miner (2009). Handbook of Statistical Analysis and Data Mining, Elsevier, 6/2009.
- [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



Chương 1 Giới thiệu tổng quan về khai phá dữ liệu

5



Nội dung

- 0. Tình huống
- 1.1. Sự cần thiết khai phá dữ liệu
- 1.2. Quy trình khám phá tri thức
- 1.3. Khai phá dữ liệu là gì
- 1.4. Các nhiệm vụ của khai phá dữ liệu
- 1.5. Các ứng dụng của khai phá dữ liệu

6

6



0. Tình huống 1



Người đang sử dụng thẻ ID = 1234 thật sự là chủ nhân của thẻ hay là một tên trộm?

7

7



0. Tình huống 2

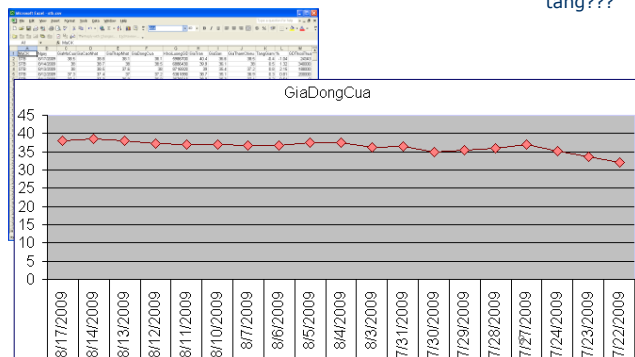
Khóa	MãSV	MônHọc1	MônHọc2	...	TốtNghệp
2004	1	9.0	8.5	...	Có
2004	2	6.5	8.0	...	Có
2004	3	4.0	2.5	...	Không
2004	8	5.5	3.5	...	Không
2004	14	5.0	5.5	...	Có
...
2005	90	7.0	6.0	...	Có (80%)
2006	24	9.5	7.5	...	Có (90%)
2007	82	5.5	4.5	...	Không (45%)
2008	47	2.0	3.0	...	Không (97%)
...

Làm sao xác định được khả năng tốt nghiệp của một sinh viên hiện tại?



0. Tình huống 3

Ngày mai cổ phiếu STB sẽ tăng???



0. Tình huống ...



We are data rich, but information poor.
"Necessity is the mother of invention". - Plato

10



Ai thực ai giả



11

11



Làm thế nào để chia những bức ảnh này vào 3 nhóm



12

12



1.1. Sự cần thiết khai phá dữ liệu

Sự bùng nổ dữ liệu

Khía cạnh công nghệ

Khía cạnh thương mại

Nhu cầu thu nhận tri thức từ dữ liệu

Ngành kinh tế định hướng dữ liệu

13

13



Khía cạnh công nghệ

❖ Ngành công nghệ bán dẫn 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.
- 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)

❖ 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

14

14



Khía cạnh công nghệ

- 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 năm 2000
 - Vài tuần đầu tiên: thu thập dữ liệu thiên văn học = toàn bộ trong quá khứ. Sau 10 năm: 140 TB
- Bùng nổ dữ liệu: Công nghệ mạng
- **Các kỹ thuật truyền thống không đủ khả năng khai thác dữ liệu thô.**

15

15



Bùng nổ dữ liệu:

❖ A huge demand on Data Science

▪ “Data scientist: the sexiest job of the 21 st century” – Harvard

Business Review.

<http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/>

▪ “The Age of Big Data” – The New York Times

http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-theworld.html?pagewanted=all&_r=0

16

16

Data Analyst
San Francisco Bay Area
Posted 18 days ago

Data Analyst
Greater New York City Area
Posted 25 days ago

Statistical Analyst - Data...
Greater New York City Area
Posted 9 hours ago

Data Analyst
Greater New York City
Posted 15 days ago

DATA SCIENTIST
Greater New York City
Posted 25 days ago

Data Scientist
Greater New York City
Posted 14 days ago

Marketing Analytics Associate
Greater New York City Area
Posted 24 days ago

Financial Data Analyst
Greater New York City Area
Posted 20 days ago

Data Analyst...
Greater New York City Area
Posted 13 days ago

Data Analyst
Amazon - Newark, NJ
Posted 24 days ago
Apply on company website Save

Senior Data Analyst - Big Data, Meta Product
TripAdvisor - Newton, MA
Posted 12 days ago
Apply now Save

Data Analyst
Apple - Daly City - California - US
Posted 18 days ago
Apply on company website Save



Khía cạnh thương mại

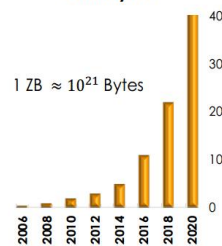
- ❖ Khối lượng lớn dữ liệu được thu thập và lưu trữ
 - Web data, e-commerce
 - Hóa đơn mua hàng tại siêu thị
 - Giao dịch ngân hàng
- ❖ Máy tính mạnh mẽ hơn
- ❖ Áp lực cạnh tranh rất mạnh
 - Cung cấp các dịch vụ đa dạng, chất lượng tốt

18

18

- Data mining, inference, prediction
- ML & DM provides an efficient way to make intelligent systems/services.
- ML provides vital methods and a foundation for Big Data.

All global data in Zettabytes



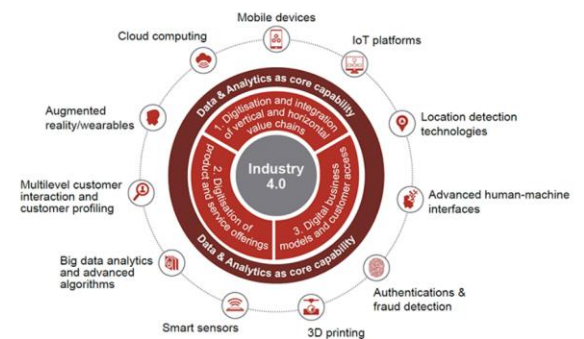
Each day:
230M tweets,
2.7B comments to FB,
86400 hours of video
to YouTube

19

19



Công nghiệp 4.0



<https://www.pwc.com/ca/en/industries/industry-4-0.html>

20

20

❖ Tạo **Trí tưởng tượng** (Imagination)

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log (1 - D(G(z)))]$$

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In *NIPS*, pp. 2672-2680. 2014.

21

■ AlphaGo of Google DeepMind the world champion at Go (cờ vây), 3/2016

- Go is a 2500 year-old game.
- Go is one of the most complex games.

■ AlphaGo learns from 30 millions human moves, and plays itself to find new moves.

■ It beat Lee Sedol (World champion)

- <http://www.wired.com/2016/03/two-redefined-future/>
- <http://www.nature.com/news/google-game-of-go-1.19234>

22

■ Tạo khả năng **Viết** cho máy tính

- Một mô hình khổng lồ được huấn luyện từ dữ liệu khổng lồ
- Nó có thể được dùng vào nhiều **bài toán có ít dữ liệu**

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

he mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that I could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

	Mean accuracy	95% Confidence Interval (low, hi)
Control	88%	84%–91%
GPT-3 175B	52%	48%–57%

Con người không thể nhận diện bài viết 500 từ là do máy hay người viết

23

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 dự 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].

24



Nhu cầu thu nhận tri thức từ dữ liệu

- ❖ **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

25



Ngành kinh tế định hướng dữ liệu

❖ Ngành công nghiệp quản lý và phân tích dữ liệu

- “Chúng ta ngập trong dữ liệu mà đói khát tri thức”
- 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

26



2. Quy trình khám phá tri thức

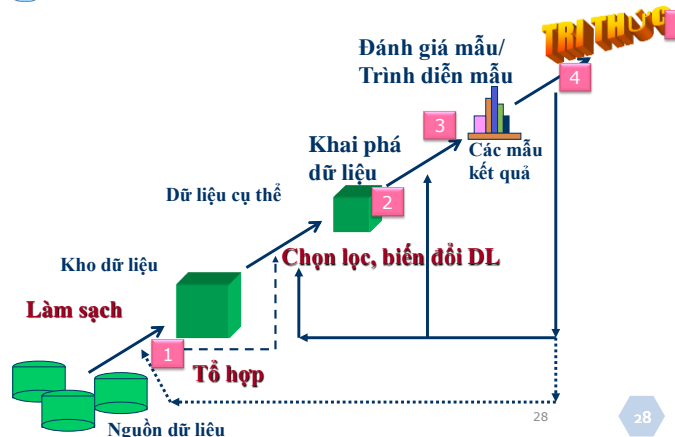
Khám phá tri thức là một quá trình truy xuất tri thức từ cơ sở dữ liệu lớn.

Tri thức này được sử dụng để giải quyết một loạt nhiệm vụ trong những lĩnh vực nhất định

27



1.2. Quy trình khám phá tri thức (tt)



28





1.2. Quy trình khám phá tri thức (tt)

- ❖ Quá trình khám phá tri thức là một chuỗi lặp gồm các bước được thực thi với:
 - Data sources (các nguồn dữ liệu)
 - Data warehouse (kho dữ liệu)
 - Task-relevant data (dữ liệu cụ thể sẽ được khai phá)
 - Patterns (mẫu kết quả từ khai phá dữ liệu)
 - Knowledge (tri thức đạt được)

29

29



1.2. Quy trình khám phá tri thức (tt)

Các giai đoạn của quá trình phát hiện tri thức:

1: Hình thành và định nghĩa bài toán.

* Các mục đích của bài toán, các tri thức cụ thể của lĩnh vực

2. Làm sạch và tiền xử lý dữ liệu

3. Khai phá dữ liệu

4. Phân tích và đánh giá kết quả.

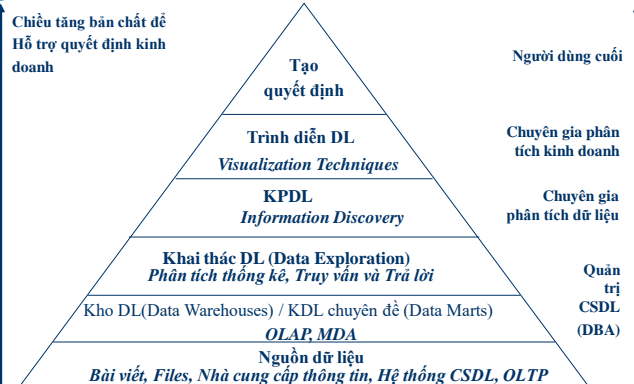
5. Sử dụng các tri thức phát hiện được.

30

30



1.2. Quy trình khám phá tri thức (tt)



31

31



1.3. Khai phá dữ liệu là gì

“Khai phá dữ liệu là quá trình không tầm thường của việc xác định các mẫu *tiềm ẩn có tính hợp lệ, mới lạ, có ích và có thể hiểu được* tối đa trong CSDL lớn” – U.Fayyad, ... (1996)

Quá trình không tầm thường

Đa xử lý

Hợp lệ

Chứng minh tính đúng đắn của mẫu/ Mô hình

Mới lạ

Chưa biết trước

Có ích

Có thể sử dụng được

Có thể hiểu được

Bởi người và máy

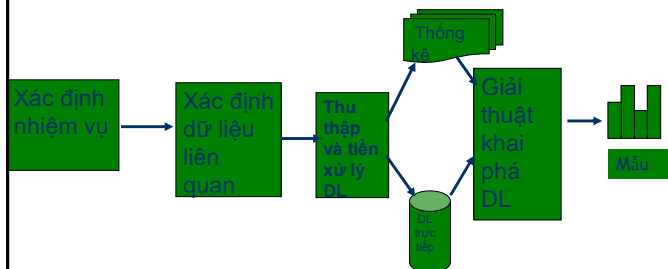
32

32



1.3.Khai phá dữ liệu là gì (tt)

■ Sơ đồ khai phá dữ liệu

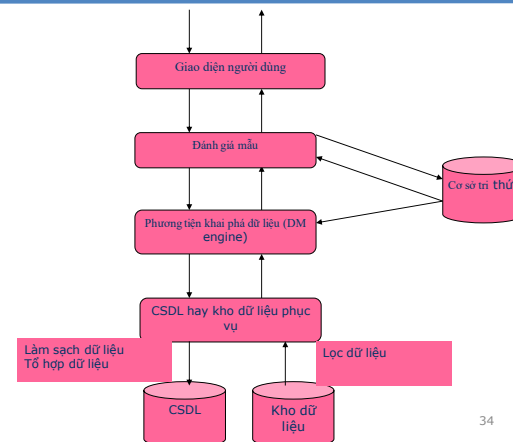


33

33



1.3.1Kiến trúc hệ thống khai phá dữ liệu



34

34



1.3.1Kiến trúc hệ thống khai phá dữ liệu

❖ CSDL, kho dữ liệu

- Thành phần này là các nguồn dữ liệu/thông tin sẽ được khai phá.
- Trong những tình huống cụ thể, thành phần này là nguồn nhập (input) của các kỹ thuật tích hợp và làm sạch dữ liệu.

❖ CSDL hay kho dữ liệu phục vụ

- Thành phần chịu trách nhiệm chuẩn bị dữ liệu thích hợp cho các yêu cầu khai phá dữ liệu.

35

35



1.3.1Kiến trúc hệ thống khai phá dữ liệu

❖ Cơ sở tri thức(Knowledge base)

- Thành phần chứa tri thức miền, được dùng để hướng dẫn quá trình tìm kiếm, đánh giá các mẫu kết quả được tìm thấy.
- Tri thức miền có thể là các phân cấp khái niệm, niềm tin của người sử dụng, các ràng buộc hay các ngưỡng giá trị, siêu dữ liệu, ...

❖ Phương tiện khai phá dữ liệu(Data mining engine)

- Thành phần chứa các khối chức năng thực hiện các nhiệm vụ khai phá dữ liệu.

36

36



1.3.1 Kiến trúc hệ thống khai phá dữ liệu

❖ Đánh giá mẫu (Pattern evaluation module)

- Thành phần này làm việc với các độ đo (và các ngưỡng giá trị) hỗ trợ tìm kiếm và đánh giá các mẫu sao cho các mẫu được tìm thấy là những mẫu được quan tâm bởi người sử dụng.
- Thành phần này có thể được tích hợp vào thành phần Data mining engine.

37

37



1.3.1 Kiến trúc hệ thống khai phá dữ liệu

❖ Giao diện người dùng (User interface)

- Thành phần hỗ trợ sự tương tác giữa người sử dụng và hệ thống khai phá dữ liệu.
 - Người sử dụng có thể chỉ định câu truy vấn hay nhiệm vụ khai phá dữ liệu.
 - Người sử dụng có thể được cung cấp thông tin hỗ trợ việc tìm kiếm, thực hiện khai phá dữ liệu sâu hơn thông qua các kết quả khai phá trung gian.
 - Người sử dụng cũng có thể xem các lược đồ cơ sở dữ liệu/kho dữ liệu, các cấu trúc dữ liệu; đánh giá các mẫu khai phá được; trực quan hóa các mẫu này ở các dạng khác nhau.

38



1.3.2. Các hệ thống khai phá dữ liệu

❖ Một số hệ thống khai phá dữ liệu:

- Intelligent Miner (IBM)
- Microsoft data mining tools (Microsoft SQL Server 2000/2005/2008)
- Oracle Data Mining (Oracle 9i/10g/11g)
- Enterprise Miner (SAS Institute)
- Weka (the University of Waikato, New Zealand,

www.cs.waikato.ac.nz/ml/weka

39

39



1.3.2. Các hệ thống khai phá dữ liệu

❖ Phân biệt các hệ thống khai phá dữ liệu với

- Các hệ thống phân tích dữ liệu thống kê (statistical data analysis systems)
- Các hệ thống học máy (machine learning systems)
- Các hệ thống truy hồi thông tin (information retrieval systems)
- Các hệ cơ sở dữ liệu diễn dịch (deductive database systems)
- Các hệ cơ sở dữ liệu (database systems)
- ...

40

40



Dữ liệu

- ❖ **Bảng ghi**
 - Dữ liệu giao dịch
- ❖ **Dữ liệu thời gian**
 - Dữ liệu chuỗi thời gian
 - Dữ liệu tuần tự
- ❖ **Dữ liệu không gian và thời gian**
 - Các tiêu đề báo
 - Các bình luận trên Twitter
- ❖ **Dữ liệu đồ thị**
- ❖ **Dữ liệu bán cấu trúc**
 - Tài liệu XML
- ❖ **Dữ liệu không có cấu trúc**

41

41



Dữ liệu bảng ghi

- ❖ Giao dịch

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

2



Dữ liệu thời gian

Dữ liệu chuỗi thời gian

year	CPI	Lai_suat	GTSX_CN	PA
2007M1	1.0	11	49,212	12,00
2007M2	3.4	11	35,392	12,00
2007M3	3.1	11	45,154	12,00
2007M4	3.7	11	47,344	12,00
2007M5	4.5	11	47,953	12,00

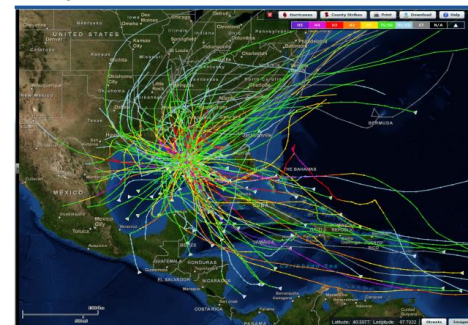
43

43



Dữ liệu không gian và thời gian

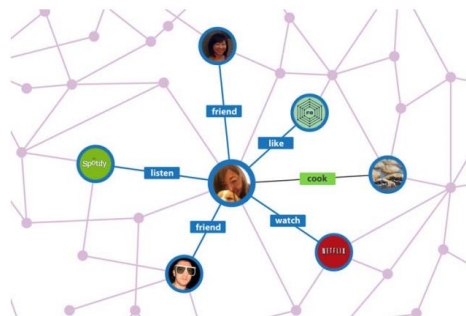
- Tập dữ liệu quỹ đạo của cơn bão



<http://csc.noaa.gov/hurricanes>

44

Dữ liệu đồ thị

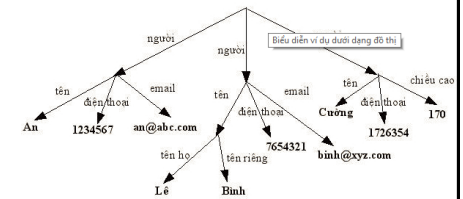


45




Dữ liệu bán cấu trúc

- ❖ người: {tên: "An", điện_thoại: 1234567, email: "an@abc.com"}.
- ❖ người: {tên: {tên_riêng: "Bình", tên_họ: "Lê"}, điện_thoại: 37654321, email: "binh@xyz.com"}.
- ❖ người: {tên: "Cường", điện_thoại: 1726354, chiều_cao: 170}. }



Dữ liệu không có cấu trúc

- ❖ Emails
- ❖ Medial
- ❖




Jason Y.
Reviews
Member 79

Excellent Vietnamese restaurant. Some of my favorites are the sour salmon soup, vietnamese steak, summer rolls, and orange chicken. The tofu wingler and scallion seared is good. I prefer the pho at pho-specific places. They don't give you basil or chilies with the pho here.

[Bookmark](#) [Send to a Friend](#)

09/23/2006




Aaron E.
Reviews
Member 79

They have great fried spring rolls and fresh garden rolls. However, for \$3.50 2 smallish rolls don't seem that great. They make you sit for with a large bowl of Pho soup for around \$6.00. This place is great for food if you just before you have a movie at the Uptown Theater. However, leave plenty of time. When the place is busy, the waitstaff or kitchen starts to stop and get real slow. You will miss your movie. Give yourself 1-1.5 hours before a movie siting. This PHO place is not as ghetto as other PHO places (like PHO 75).

[Bookmark](#) [Send to a Friend](#)

01/18/2006




Jason G.
Reviews
Member 126

The pho is good -- sort of DC -- but not great compared to some of the Pho joints in Seattle who can drop a bowl of food and noodles that will blow your mind for under \$4 bucks. Travelling, the downtown area, learn what you're missing.... That said, this is a decent spot for a late lunch over a hot bowl of the good stuff.

[Bookmark](#) [Send to a Friend](#)

03/06/2006



Fritz J.
Reviews
Member 78

Good pho, reasonable prices, and great for eating dinner before you go to the Uptown for a flick. The wait staff can be very unresponsive, though, which is annoying. Unfortunately, they don't give you basil and chilis along with the sprouts to dump into your pho.

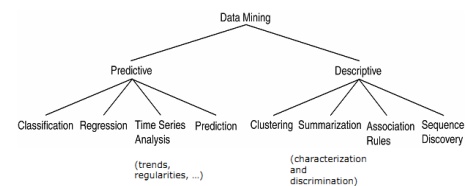
[Bookmark](#) [Send to a Friend](#)

04/14/2006

47



1.4. Các nhiệm vụ khai phá dữ liệu



48



1.4. Các nhiệm vụ khai phá dữ liệu

- Dự đoán (Predictive)
 - Sử dụng một vài biến để dự báo giá trị chưa biết hoặc giá trị tương lai của các biến khác.
 - Phân lớp
 - Hồi quy
 - Xác định sự thay đổi/lạc hướng
- Mô tả (Descriptive) : xác định các mẫu dữ liệu mà con người có thể hiểu được
 - Phân cụm
 - Luật kết hợp
 - Tóm tắt
 - Mô hình hóa phụ thuộc

49

49



Prediction of stock indices

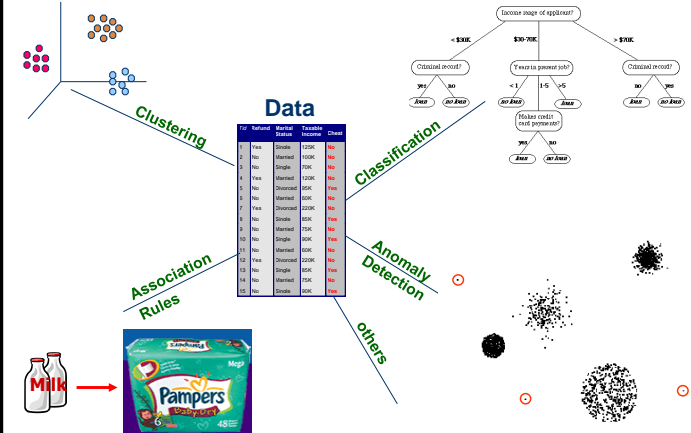


50

50



1.4. Các nhiệm vụ khai phá dữ liệu



1.4. Các nhiệm vụ khai phá dữ liệu

- ❖ Luật kết hợp
 - Quan hệ kết hợp giữa các biến dữ liệu: Tương quan và nhân quả)
 - Diaper và Beer [0.5%, 75%]
 - Luật kết hợp: $X \rightarrow Y$
 - 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
- ❖ Phân lớp: là học một hàm ánh xạ (hay phân loại) một mẫu dữ liệu vào một trong số các lớp đã xác định (Hand 1981; Weiss & Kulikowski 1991.)
- ❖ Gom cụm :Tìm ra một tập xác định các nhóm hay các cụm để mô tả dữ liệu

52

52



1.4. Các nhiệm vụ khai phá dữ liệu

- ❖ **Hồi quy:** Hồi quy là một hàm học mà ánh xạ mục dữ liệu thành một biến dự đoán có giá trị thực
- ❖ **Tóm tắt:** Liên quan đến phương pháp tìm kiếm một mô tả cô đọng cho tập con dữ liệu
- ❖ **Mô hình hóa phụ thuộc:** Bao gồm việc tìm kiếm một mô hình mô tả sự phụ thuộc đáng kể giữa các biến
- ❖ **Phát hiện sự thay đổi và lạc hướng:** Tập trung vào khai thác những thay đổi đáng kể nhất trong dữ liệu từ các giá trị chuẩn hoặc độ đo trước đó.

53

53



1.4. Các nhiệm vụ khai phá dữ liệu

Ứng dụng: Khai phá luật kết hợp

Quản lý bán hàng ở siêu thị

Mục đích: Xác định những mặt hàng được nhiều khách hàng mua chung

Hướng giải quyết:

- ❖ Xử lý dữ liệu bán hàng để tìm ra mối liên hệ giữa các mặt hàng
- ❖ Luật kết hợp cổ điển: Nếu mua tã giấy thì cũng mua bìa.

54

54



1.4. Các nhiệm vụ khai phá dữ liệu

Ứng dụng: phân lớp

Quảng cáo

Mục đích: giảm chi phí thu tín quảng cáo bằng tập trung vào những khách hàng có nhiều khả năng mua sản phẩm điện thoại mới.

Hướng giải quyết:

- ❖ Sử dụng dữ liệu cho sản phẩm tương tự trước đây
- ❖ Dùng quyết định {mua, không mua} làm thuộc tính lớp
- ❖ Thu thập thông tin cá nhân, cách sống và quan hệ của tất cả các khách hàng
- ❖ Sử dụng các dữ liệu trên để xây dựng mô hình phân lớp

55

55



1.4. Các nhiệm vụ khai phá dữ liệu

Ứng dụng: Gom cụm

Gom cụm khách hàng

Mục đích: Chia khách hàng thành các cụm riêng biệt để có thể áp dụng các biện pháp quảng cáo khác nhau

Hướng giải quyết:

- Thu thập thông tin cá nhân, cách sống của tất cả khách hàng
- Xác định cụm các khách hàng giống nhau.

56

56



1.4. Các nhiệm vụ khai phá dữ liệu

Ứng dụng: Gom cụm

Gom cụm tài liệu

Mục đích: Tìm nhóm tài liệu giống nhau dựa trên các từ quan trọng

Hướng giải quyết:

- ❖ Xác định độ đo phổ biến của các từ quan trọng trong tài liệu. Xây dựng độ đo tương tự dựa trên độ phổ biến của các từ để gom cụm
- ❖ Lợi ích: Trong truy vấn thông tin có thể dùng các cụm để liên kết tài liệu mới với tài liệu đã gom cụm

57

57



1.5. Ứng dụng cơ bản của KPD

❖ 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)

58

58



1.5. Ứng dụng cơ bản của KPD

❖ 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

59

59



Phân tích thị trường

❖ Nguồn dữ liệu có từ đâu ?

- Giao dịch thẻ tín dụng, thẻ thành viên, phiếu giảm giá, các phản nản 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 bán hàng thường xuyên

❖ Phân tích thị trường chéo

- Tìm ra các mối quan liên kết /tương quan giữa các sản phẩm bán(hoặc giữa các đợt bán hàng) để đưa ra các bảo dựa theo quan hệ kết hợp

60

60



Phân tích thị trường

❖ Hồ sơ khách hàng

- Những kiểu của khách hàng nào mua sản phẩm gì (phân cụm và phân lớp)

❖ Phân tích yêu cầu khách hàng

- Xác định các sản phẩm phù hợp nhất cho các nhóm khách hàng khác nhau
- Dự đoán những yếu tố nào sẽ thu hút được các 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)

61

61



Phân tích doanh nghiệp & Quản lý rủi ro

❖ Lên kế hoạch tài chính và đánh giá tài sản

- Phân tích và dự đoán luồng tiền mặt
- Phân tích các tuyên bố tài chính của doanh nghiệp để đánh giá tài sản
- Phân tích các chuỗi dữ liệu tài chính

❖ Lên kế hoạch sử dụng tài nguyên

- Tóm tắt và so sánh các nguồn lực và chi tiêu

❖ Cạnh tranh

- Theo dõi đối thủ cạnh tranh và các xu hướng của thị trường
- Nhóm khách hàng thành các lớp và định giá dựa theo lớp khách
- Xây dựng chiến lược giá trong thị trường cạnh tranh cao

62

62



Phát hiện gian lận

❖ 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, sự móc nối giữa bệnh nhân và bác sỹ, các xét nghiệm không cần thiết
 - 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: phát hiện các người làm thuê gian lận
 - Các nhà phân tích ước lượng rằng 38% giảm bán lẻ là do nhân viên không trung thực
- Chống khủng bố

63

63



Ứng dụng khác

❖ Thể thao

- IBM Advanced Scout phân tích thống kê môn NBA (chặn bóng, hỗ trợ và lỗi) để đưa tới lợi thế cạnh tranh cho New York Knicks và Miami Heat

❖ Thiên văn học

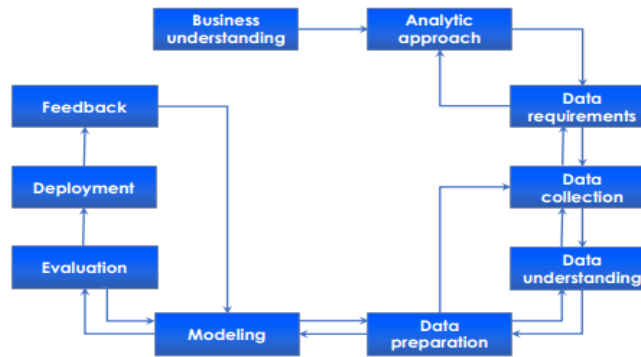
- JPL và Palomar Observatory khám phá 22 chuẩn tinh (quasar) với sự trợ giúp của KPDL

❖ 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 thiện cách tổ chức

64

64



- ❖ Chọn tập dữ liệu phù hợp với yêu cầu
- ❖ Lựa chọn kiểu hàm chức năng trả về 0,1 hay tập nhãn...
- ❖ Chọn mô hình thể hiện
 - Tuyến tính
 - Phân lớp
 - Phân cụm
 -
- ❖ Lựa chọn thuật toán phù hợp