

Khai phá dữ liệu: Giới thiệu môn học

(Data Mining: Course Introduction)

Phan Xuân Hiếu

Khoa Công nghệ Thông tin
Trường ĐH Công nghệ (UET), ĐHQG Hà Nội (VNU)
hieupx@vnu.edu.vn

(last updated: 19–01–2016)

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPDL)

- Động lực và mục đích của KPDL
- Định nghĩa và quá trình KPDL
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPDL

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPDL)

- Động lực và mục đích của KPDL
- Định nghĩa và quá trình KPDL
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPDL

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPDL)

- Động lực và mục đích của KPDL
- Định nghĩa và quá trình KPDL
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPDL

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

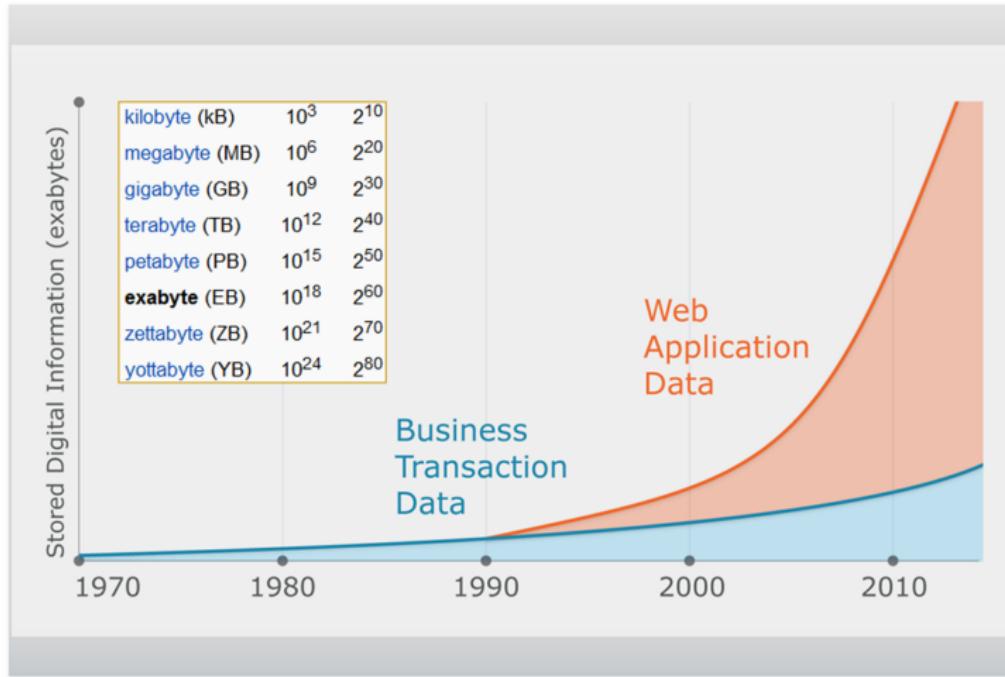
4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Kỷ nguyên của dữ liệu lớn, đa dạng, phức hợp



Tốc độ bùng nổ dữ liệu



Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPD)

- Động lực và mục đích của KPD
- **Định nghĩa và quá trình KPD**
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPD

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

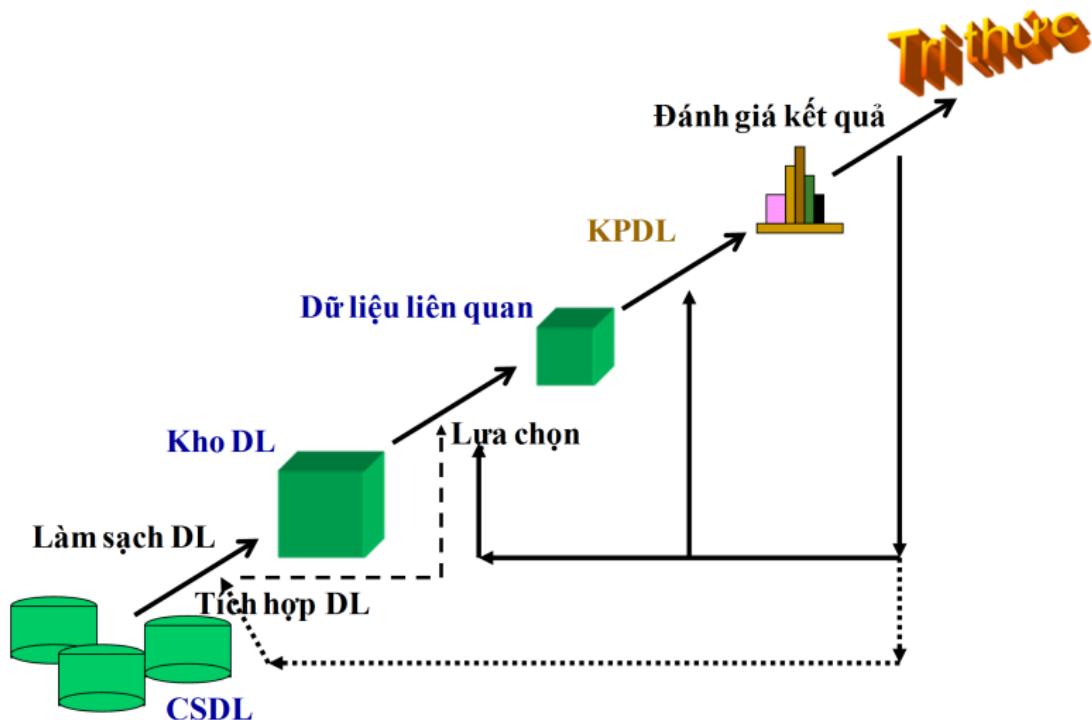
- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Khai phá dữ liệu (KPDL)

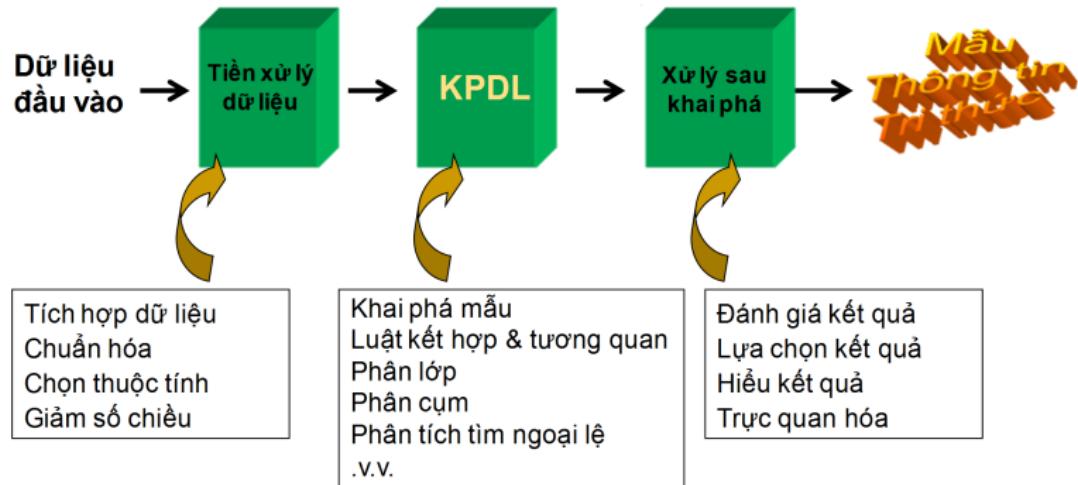
Khai phá dữ liệu là quá trình phân tích và phát hiện các *quy luật* hoặc các *mẫu thông tin hay tri thức hấp dẫn, hữu ích, chưa được biết đến ẩn chứa* trong các tập dữ liệu (lớn).

- Các điểm cần lưu ý trong định nghĩa:
 - ▶ *Các quy luật:* laws
 - ▶ *Các mẫu thông tin hay tri thức:* information/knowledge patterns
 - ▶ *Hấp dẫn:* interesting
 - ▶ *Hữu ích:* useful
 - ▶ *Chưa được biết đến:* previously unknown
 - ▶ *Ẩn chứa trong dữ liệu:* hidden in data
- Tên gọi và một số thuật ngữ liên quan:
 - ▶ Data Mining
 - ▶ Knowledge Discovery in Databases (KDD)
 - ▶ Data and Knowledge Engineering
 - ▶ Big Data, Data Analytics, Data Science, Data Scientist

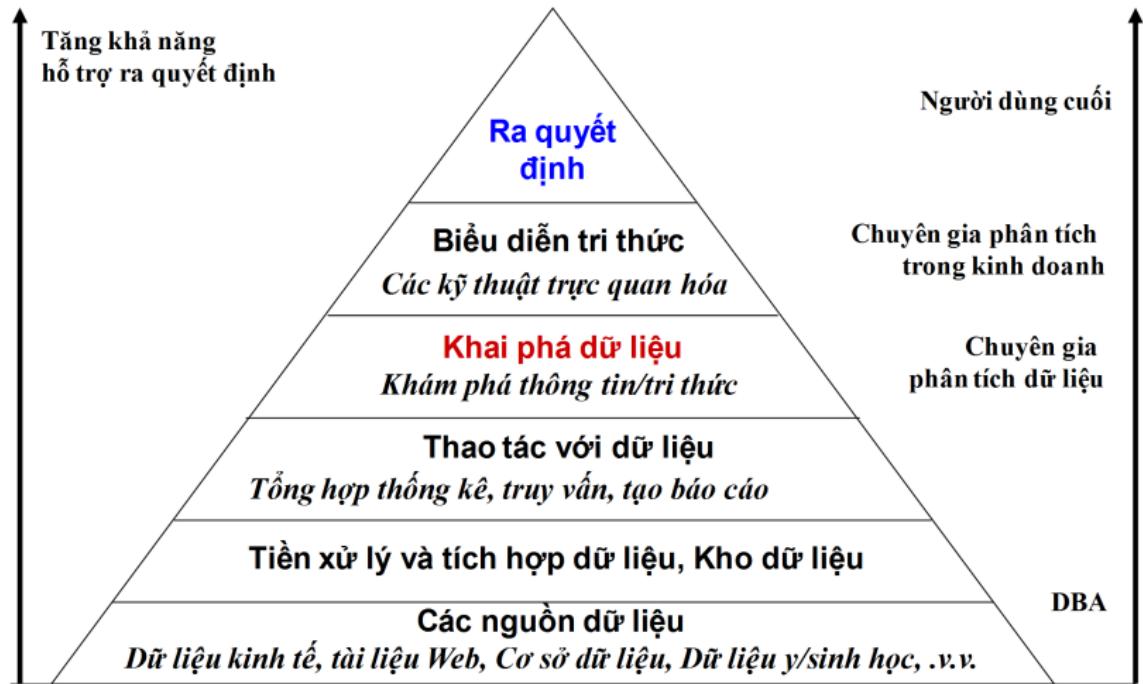
Quá trình khai phá dữ liệu



Quá trình KPDL: tiền xử lý, khai phá, hậu xử lý



KPDL và thông minh doanh nghiệp (business intelligence)



Khai phá dữ liệu và các lĩnh vực liên quan



Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPDL)

- Động lực và mục đích của KPDL
- Định nghĩa và quá trình KPDL
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPDL

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

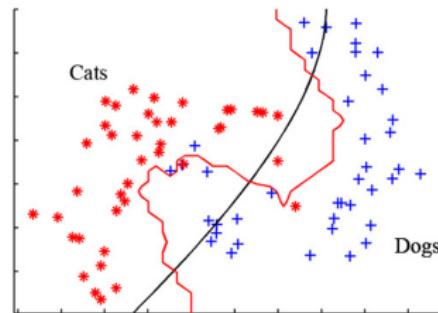
Dữ liệu

- Dữ liệu giao dịch bán lẻ (retail).
- Dữ liệu tài chính, ngân hàng, chứng khoán.
- Dữ liệu viễn thông (telecommunication data).
- Dữ liệu văn bản, ngôn ngữ tự nhiên.
- Dữ liệu chuỗi (time-series, temporal, sequential, ...).
- Dữ liệu đa phương tiện (image, audio, video).
- Dữ liệu bản đồ, không gian (map and spatial data).
- Dữ liệu y/sinh học.
- Dữ liệu Web (Web pages, Web logs, online user behaviors, ...).
- Dữ liệu quảng cáo trực tuyến và thương mại điện tử.
- Dữ liệu đồ thị, mạng liên kết, mạng xã hội (social network data).
- Dữ liệu môi trường, hệ sinh thái, thủy lợi, thủy văn.
- ...

Các dạng vấn đề hay bài toán lớn trong KPDL

- Phân lớp/phân loại (Classification/Categorization).
- Hồi quy (Regression).
- Phân cụm (Clustering).
- Phân tích quan hệ tương quan và kết hợp (Correlation & Association).
- Dự đoán, dự báo (Prediction, Forecast).
- Xác định ngoại lệ, gian lận (Anomaly, Outlier, and Fraud Detection).
- Phân tích chủ đề, xu hướng (Topic/Trend Analysis).
- Phân tích mạng xã hội (Social Network Analysis).
- Phân tích quan điểm (Opinion Mining & Sentiment Analysis).
- Các hệ thống tư vấn/khuyến nghị (Recommender Systems).
- Dự đoán và tối ưu trong quảng cáo trực tuyến (Computational Advertising).
- ...

Phân lớp/phân loại (classification/categorization)



- Thuật ngữ khác: supervised learning (học có giám sát).
- Phát biểu bài toán:
 - ▶ $C = \{c_1, c_2, \dots, c_K\}$: tập K nhãn (label) tương ứng với K lớp (class).
 - ▶ $\mathbf{X} = \{\mathbf{x}_i\}$ ($i = 1, 2, \dots$): không gian đối tượng cần phân lớp.
 - ▶ Mô hình phân lớp là ánh xạ $f : \mathbf{X} \rightarrow C$.
 - ▶ Tập dữ liệu có gắn nhãn (labeled/annotated data):
 $\mathbf{D} = \{(\mathbf{x}^1, c^1), (\mathbf{x}^2, c^2), \dots, (\mathbf{x}^N, c^N)\}$, $\mathbf{x}^i \in \mathbf{X}$, $c^i \in C$.
 - ▶ Xây dựng mô hình f bằng học giám sát dựa trên \mathbf{D} .
 - ▶ Tiêu chí của f : phân lớp chính xác, nhỏ gọn, thực hiện nhanh.

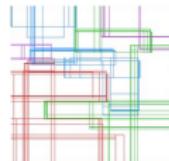
Các ứng dụng của bài toán phân lớp/phân loại

- Phân lớp tài liệu, văn bản (document classification/categorization).
- Lọc spam (spam filtering).
- Thị giác máy (computer vision): robot, ô tô tự hành, ...
- OCR (optical character recognition).
- Nhận dạng chữ viết tay (handwriting recognition).
- Các bài toán trong xử lý ngôn ngữ tự nhiên.
- Ứng dụng trong y/sinh học: phân loại bệnh, thuốc, hình ảnh MRI, dữ liệu di truyền, ...
- Phân loại và phát hiện gian lận (outlier and fraud detection).
- Phân lớp trong tài chính: khách hàng, tín dụng, rủi ro, ...
- ... và rất nhiều các ứng dụng phân loại, dự báo khác trong thực tế.

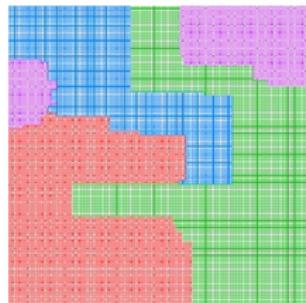
Xây dựng mô hình phân lớp f



Dữ liệu & dữ liệu huấn luyện D



Huấn luyện mô hình



Mô hình phân lớp f

- Một số thuật ngữ:
 - ▶ Ước lượng mô hình (model estimation).
 - ▶ Huấn luyện mô hình (model training).
 - ▶ Lựa chọn mô hình (model selection).
 - ▶ Đánh giá mô hình (model evaluation).

Các phương pháp xây dựng mô hình phân lớp

- Case-based reasoning.
- Nearest neighbor classification.
- Rule-based classification (phân lớp dựa trên luật).
- Decision trees (cây quyết định).
- Naive Bayes classification.
- Random forest.
- Maximum entropy classification (MaxEnt – cực đại hóa entropy).
- Artificial neural networks (ANNs – mạng nơ ron nhân tạo).
- Support vector machines (SVMs – máy véc tơ hỗ trợ).
- Boosting (meta-algorithm).
- ...

Phân cụm (clustering)



Phân cụm dữ liệu là bài toán gom các đối tượng dữ liệu vào thành từng nhóm/cụm (group/cluster) sao cho các đối tượng trong cùng một cụm có sự tương đồng theo một tiêu chí nào đó.

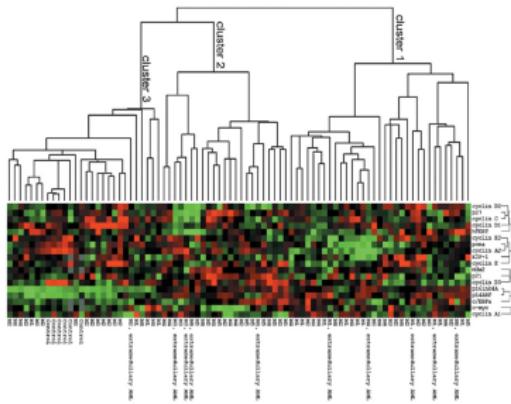
Phân cụm còn được gọi là học không giám sát (unsupervised learning) hoặc phân lớp không giám sát (unsupervised classification).

Ứng dụng của phân cụm

- Phân cụm tài liệu/văn bản.
 - Phân cụm dữ liệu giao dịch mua bán.
 - Phân cụm/nhóm người dùng.
 - Phân đoạn/vùng ảnh (image segmentation).
 - Phân cụm dữ liệu gen di truyền (gene expression data).
 - Phân cụm tìm các cộng đồng (communities) trong mạng xã hội.
 - ...



[nguồn: buffy.eecs.berkeley.edu]



[nguồn: www.nature.com]

Các cách tiếp cận phân cụm (clustering approaches)

- **Phân cụm phân cấp (hierarchical methods):**
 - ▶ Còn gọi là phân cụm dựa vào kết nối (connectivity-based).
 - ▶ Phương pháp: top-down (divisive) hoặc bottom-up (agglomerative).
- **Phân cụm phân hoạch (partitioning/centroid-based methods):**
 - ▶ Khởi tạo với k cụm. Lặp đi lặp lại việc gán mỗi đối tượng vào cụm có tâm gần nhất, sau đó điều chỉnh lại các tâm cụm.
 - ▶ Phương pháp: K -means, K -medoids.
- **Phân cụm dựa trên phân bố (distribution-based methods):**
 - ▶ Giả định mỗi cụm dữ liệu được sinh ra từ một phân bố xác suất.
 - ▶ Phương pháp: mô hình trộn Gaussian với thuật toán EM.
- **Phân cụm dựa trên mật độ (density-based methods):**
 - ▶ Các cụm là các vùng có mật độ cao. Phương pháp DBSCAN, OPTICS.
- **Phân cụm dựa trên lưới (grid-based methods):**
 - ▶ Phân chia dữ liệu thành các ô lưới, phù hợp với dữ liệu không gian.
- **Phân cụm dựa trên đồ thị (graph-based methods):**
 - ▶ Các cụm là các vùng đồ thị con dày đặc (đồ thị dày đủ hoặc gần dày đủ – cliques/quasi-cliques). Phương pháp HCS.

Khái niệm luật kết hợp (association rule)



- Luật kết hợp: *Mỗi quan hệ kết hợp giữa các tập thuộc tính trong cơ sở dữ liệu.*
- Ví dụ:
 - ▶ $\{bánh mỳ, bơ, mứt dâu\} \rightarrow \{sữa tươi\}$ (phổ biến: 3%, tin cậy: 80%)
 - ▶ $\{tuổi > 45, gia đình có lịch sử tiểu đường, huyết áp cao\} \rightarrow \{mắc bệnh tiểu đường\}$ (phổ biến: 1.5%, tin cậy: 76%)

Ứng dụng của mẫu phổ biến và luật kết hợp

- Phân tích dữ liệu giao dịch bán lẻ (market basket analysis).
 - ▶ Tối ưu việc nhập các ngành hàng.
 - ▶ Sắp xếp vị trí các ngành hàng hợp lý (store layout).
 - ▶ Marketing và khuyến mại.
 - ▶ Gợi ý và khuyến nghị trực tuyến (online recommendation), ví dụ: “frequently bought together products” và “bought this also bought ...”
- Hiểu người dùng thông qua phân tích các mẫu phổ biến từ nhật ký duyệt web (web logs).
- Phân tích tìm ngoại lệ (outlier detection).
- Phân tích về tội phạm và an ninh.
- Khai phá các cấu trúc mạng xã hội (mẫu đồ thị phổ biến).
- Ứng dụng trong phân lớp, phân loại (decision rules).
- Ứng dụng trong khai phá dữ liệu text (text mining).
- Ứng dụng trong khai phá dữ liệu y/sinh học (biomedical data mining).
- Ứng dụng trong khai phá dữ liệu không/thời gian và dữ liệu dòng (stream data).
- ...

Các lĩnh vực ứng dụng của khai phá dữ liệu



Ứng dụng trong hệ tư vấn

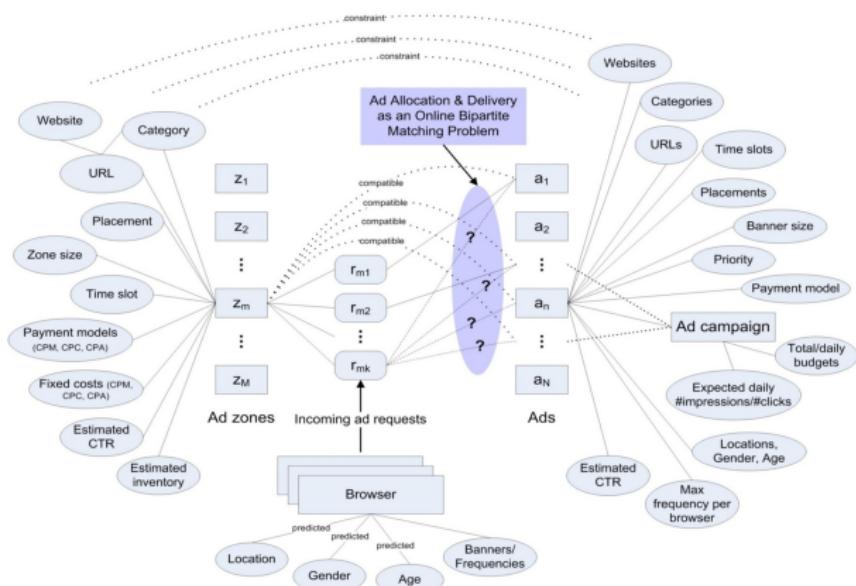
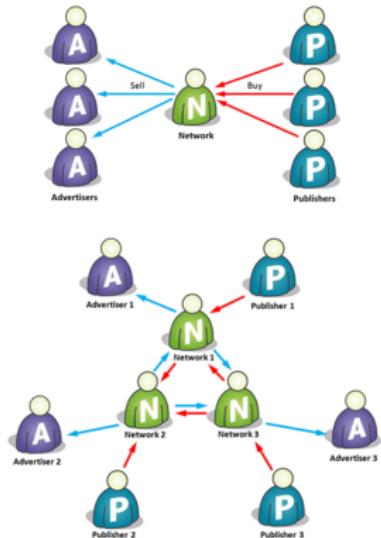
(recommender systems)

The Science Behind the Netflix Algorithms That Decide What You'll Watch Next

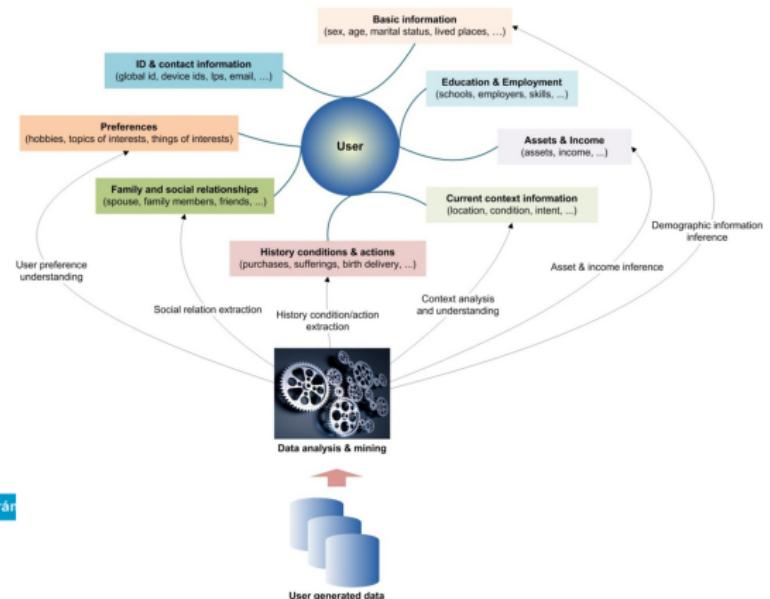


If you liked 1960s *Star Trek*, the first non-Trek title that Netflix is likely to suggest to you is the original *Mission: Impossible* series (the one with the cool Lalo Schifrin soundtrack). Streaming the latest *Doctor Who* is likely to net you the supernatural TV drama *Being Human* (the UK version). Watch *From Dusk Till Dawn* and *300* and say hello to a new row on your homepage: Visually Striking Violent Action & Adventure. Trying to understand the invisible array of algorithms that power your Netflix suggestions has long been a favorite sport, but what's actually going on in that galaxy of big data, those billions and billions of ratings stars? Turns out there are 800 Netflix engineers working behind the scenes at their Silicon Valley HQ. The company estimates that **75 percent of viewer activity is driven by recommendation**. This summer it's unveiling a profile feature enabling family members to demarcate their preferences with individual queues. In March the company shipped its 4 billionth DVD, but in the first quarter of 2013 alone, it streamed more than 4 billion hours. We spoke with Netflix's recommendation dynamos—Carlos Gomez-Uribe, VP of product innovation and personalization algorithms (right), and Xavier Amatriain, engineering director—about how they control what you watch.

Ứng dụng cho tính toán trong quảng cáo trực tuyến (computational advertising)

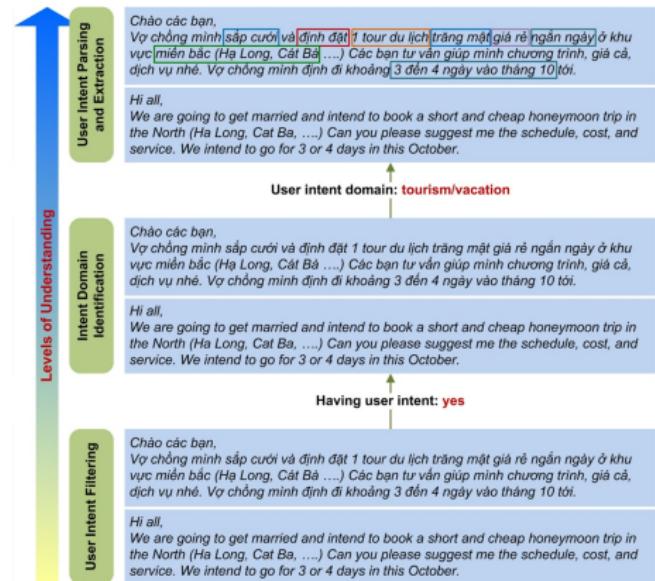


Ứng dụng trong phân tích và hiểu người dùng trực tuyến (online user behavior analysis & understanding)

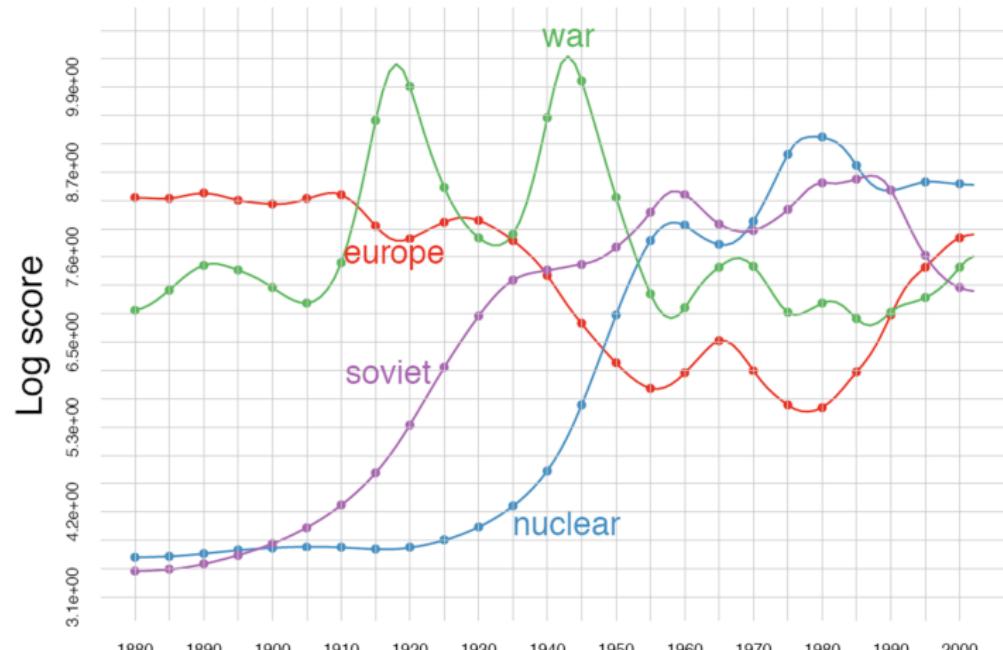


Ứng dụng trong nhận diện ý định người dùng (online user intent identification)

Online Vietnamese social media texts	Intent?
Tinh hình là mình đang cần thuê nhà quanh khu vực Phuong Mai, Bach Khoa hoặc Tôn Thất Tùng cho ba người lớn và một cháu nhỏ. Tầm tiền khoảng 3 triệu. Bạn nào có thông tin gì xin liên hệ với mình theo số 0905231880. Cám ơn nhiều. (I am looking for a house to rent near Phuong Mai, Bach Khoa or Ton That Tung street for three adults and one child. The price is about 3 million vnd. Please contact me at 0905231880 if you have any information. Thank you a lot.)	Explicit intent
Thế tí thi bây giờ nếu bạn vay tiền ở bất kỳ ngân hàng nào bạn cũng phải chịu lãi suất cao. (Actually, if you borrow money from any banks at this time, you have to pay high loan interest rate)	Non-intent
Mình đang định vay ngân hàng một khoản bằng bảng lương của mình. Không biết có mẹ nào ở đây có kinh nghiệm về việc này có thể tư vấn cho mình được không a. Mặc dù mình biết không thể vay được nhiều tiền theo cách này nhưng mình thấy nó đơn giản và hơn nữa có quan niệm lại trả lương qua tài khoản ATM. (I intend to borrow an amount of money from any bank using my payroll. If any mom here has experience about this, please give me a tip ...)	Explicit intent
Với số tiền bạn có thì khó có thể mua được một căn hộ tại ở khu vực Cầu Giấy hoặc Thanh Xuân. (It is impossible to buy an apartment in Cau Giay or Thanh Xuan areas with your amount of money.)	Non-intent
Mình đang tìm một lớp luyện IELTS 6.5, học 2 ngày một tuần (trong đó một ngày là thứ 7 hoặc chủ nhật), từ 16h30 đến 18h30. Nhà mình ở Long Biên, mình đi làm ở Lô Đức. Mẹ nào biết lớp học nào gần khu vực này thi cho mình xin thông tin với nhé. Mình cảm ơn nhiều. (I am looking for an IELTS 6.5 class, studying 2 days a week (one is Saturday or Sunday), from 16:30 to 18:30. I live in Long Bien and work at Lo Duc. If any mom knows any class in these areas, please let me know. Thanks a lot.)	Explicit intent



Ứng dụng trong phân tích xu hướng thông tin (topic/trend analysis)



Source: <http://www.cs.princeton.edu/~blei/modeling-science.pdf>

Ứng dụng trong phân tích truyền thông xã hội trực tuyến (online social media analysis & monitoring)

[CÀNH BÁO: Bác sĩ LONG \(phòng khám 70 Nguyễn Chí Thanh - HN\) Đừng...](#) - [Translate this page]
www.webtretho.com/.../canh-bao-bac-si-long-phong-kham-70-ngu... - Cached

1 post - 1 author

Chào các mẹ, Em là thành viên quen thuộc của diễn đàn nhà mình từ lâu, em có chuyện này cần báo với các mẹ 1 việc.

[Cành báo về Trường mầm non Bibihome Hà Nội](#) - [Translate this page]
www.webtretho.com/.../canh-bao-ve-truong-mam-non-bibihome-ha... - Cached

10 posts - 9 authors - Last post: 11 Jul

Tôi xin **cành báo** với các mẹ nào có ý định gửi con tại trường mầm non bibihome - phố Đặng Văn Ngữ - Đông Đa - HN Cơ.

[Cành báo về phòng khám Maria 65 Thái Thịnh I](#) - [Translate this page]
www.webtretho.com/.../canh-bao-ve-phong-kham-maria-65-thai-thi... - Cached

10 posts - 7 authors - Last post: 30 Jun

Em hôm nay đi khám ở Maria, nhưng riêng chi phí khám và xét nghiệm đã mất toi gần 2 triệu rồi, em thấy chi phí đắt quá.

[Cành báo các mẹ đang ăn sữa chua Ba Vì, nguy hiểm quá, ăc ăc ...](#) - [Translate this page]
www.webtretho.com/.../canh-bao-cac-me-dang-an-sua-chua-ba-vi-n... - Cached

10 posts - 10 authors - Last post: 11 May

Em ko biết nên gửi bài này vào đâu, nên viết vào đây, mod thông cảm nhé, tại ở đây nhiều mẹ tham gia mà. Tình hình là.

[Cành báo về phòng khám Maria 65 Thái Thịnh I - Trang 16](#) - [Translate this page]
www.webtretho.com/forum/f77/canh-bao-ve-phong-kham-maria-65-thai-thin... - Cached

Có người của phòng khám M vào đây bênh vực rồi kia.

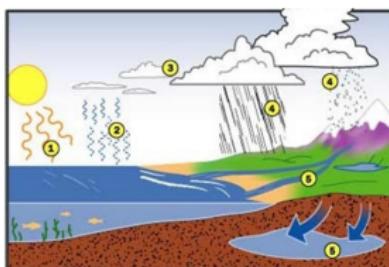
[Cành báo: Phòng chẩn trị y học cổ truyền tư nhân trung tâm cây chỉ ...](#) - [Translate this page]
www.webtretho.com/.../canh-bao-phong-cha-n-tri-y-hoc-co-truyen... - Cached

6 posts - 5 authors - Last post: 14 Aug

Các mẹ ạ, có mẹ nào đưa người thân hay biết bạn bè đến khám và chẩn trị ở đây chưa ạ?
Hôm nay nhà em lao đao vì.

Ứng dụng trong dự báo thủy lợi, thủy văn

(prediction/forecast in hydrology)



Thu thập dữ liệu thủy văn

(và tiền xử lý dữ liệu như khử nhiễu, rời rạc hóa, chuẩn hóa, biến đổi dữ liệu)

Mô hình hóa dựa trên dữ liệu

- Phân tích dữ liệu
- Xây dựng các mô hình
(dự báo, dự đoán, ước lượng, phân loại, hỗ trợ quyết định, ...)

Dự báo mực nước, triều cường

Dự báo lưu lượng nước
(sông, hồ, kênh, rạch)

Dự báo hạn hán, lũ lụt

Ước lượng lở đất, xói mòn, bồi tụ

Hỗ trợ ra quyết định trong quản lý nguồn nước và phòng chống thiên tai

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPD)

- Động lực và mục đích của KPD
- Định nghĩa và quá trình KPD
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPD

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPD)

- Động lực và mục đích của KPD
- Định nghĩa và quá trình KPD
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPD

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Các môn học và kỹ năng tiên quyết

Các môn học, kiến thức tiên quyết:

- Cơ sở dữ liệu (database) và kho dữ liệu (data warehouse).
- Cấu trúc dữ liệu và giải thuật.
- Đại số tuyến tính.
- Xác suất thống kê.
- Tối ưu toán học (optional).
- Internet và Web.

Các kỹ năng cần có:

- Thành thạo ít nhất một trong những ngôn ngữ sau: C/C++, Java, Python, Ruby, Perl, ...
- Các ngôn ngữ và môi trường phân tích dữ liệu: Matlab, R, Mathematica, ...

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPD)

- Động lực và mục đích của KPD
- Định nghĩa và quá trình KPD
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPD

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Thi và tính điểm

- 40% điểm giữa kỳ:
 - ▶ Điểm bài tập lớn (dự án - projects) làm theo nhóm, hoặc
 - ▶ Điểm trình bày seminar theo nhóm.
- 60% điểm thi cuối kỳ:
 - ▶ 10% điểm góp mặt và tham gia góp ý, phát biểu ở lớp.
 - ▶ 50% điểm thi vấn đáp cuối kỳ.

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPD)

- Động lực và mục đích của KPD
- Định nghĩa và quá trình KPD
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPD

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPD)

- Động lực và mục đích của KPD
- Định nghĩa và quá trình KPD
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPD

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Các nội dung giảng ở lớp

- Tiền xử lý dữ liệu.
- Phân lớp và các phương pháp phân lớp.
- Lựa chọn và đánh giá mô hình phân lớp.
- Các độ đo tương tự.
- Phân cụm và các phương pháp phân cụm dữ liệu.
- Biểu diễn và lựa chọn thuộc tính trong phân lớp, phân cụm.
- Khai phá mẫu phổ biến và luật kết hợp.
- Các hệ thống tư vấn/khuyến nghị.
- Phân tích chủ đề và các mô hình phân tích chủ đề.

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPDL)

- Động lực và mục đích của KPDL
- Định nghĩa và quá trình KPDL
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPDL

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- **Các dự án (projects)**
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Danh sách các dự án bài tập lớn (projects)

- ① Thu thập và trích chọn thông tin về lịch chiếu phim (rạp) và chương trình tivi.
- ② Thu thập và trích chọn thông tin y tế, sức khỏe từ Internet.
- ③ Thu thập và trích chọn thông tin từ Facebook.
- ④ Phân loại tin tức trực tuyến.
- ⑤ Phân cụm tin tức trực tuyến.
- ⑥ Đoán nhận giới tính dựa vào họ tên.
- ⑦ Đoán nhận giới tính của người dùng đọc báo trực tuyến.
- ⑧ Đoán nhận tuổi của người dùng đọc báo trực tuyến.
- ⑨ Phát hiện tự động các chủ đề nóng trên Internet.
- ⑩ Chuẩn hóa ngôn ngữ teen trên Internet.

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPD)

- Động lực và mục đích của KPD
- Định nghĩa và quá trình KPD
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPD

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Danh sách các chủ đề seminar (seminar topics)

- ① Kho dữ liệu (Data warehouse).
- ② Khai phá luật kết hợp (Association rule mining).
- ③ Phân tích chủ đề và ứng dụng (Topic analysis & applications).
- ④ Trích chọn thông tin trên Web (Information extraction on the Web).
- ⑤ Quảng cáo trực tuyến (Online advertising).
- ⑥ Phân tích quan điểm (Sentiment analysis & opinion mining).
- ⑦ Phân tích mạng xã hội (Social network analysis).
- ⑧ Thị trường & đấu giá (Market & auction).
- ⑨ Các hệ thống gợi ý trực tuyến (Online recommender systems).
- ⑩ Các nền tảng phân tích dữ liệu lớn (Big data platforms).

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPD)

- Động lực và mục đích của KPD
- Định nghĩa và quá trình KPD
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPD

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPD)

- Động lực và mục đích của KPD
- Định nghĩa và quá trình KPD
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPD

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Giáo trình và sách tham khảo



J. Han, M. Kamber, and J. Pei.

Data Mining: Concepts and Techniques.

The 3rd Edition, Morgan Kaufmann, Elsevier, 2012.



A. Rajaraman, J. Leskovec, and J. D. Ullman.

Mining of Massive Datasets.

The 2nd Edition, Cambridge University Press, 2013.



M. J. Zaki and W. M. Jr.

Data Mining and Analysis: Fundamental Concepts and Algorithms.

Cambridge University Press, 2013.



C. M. Bishop.

Pattern Recognition and Machine Learning.

Springer, 2006.

Giáo trình và sách tham khảo (2)



I. H. Witten and E. Frank.

Data Mining: Practical Machine Learning Tools and Techniques.

Morgan Kaufmann Publishers, 2nd edition, 2005.



D. Easley and J. Kleinberg.

Networks, Crowds, and Markets: Reasoning About a Highly Connected World.

Cambridge University Press, 2010.



P. Baldi and P. Frasconi and P. Smyth.

Modeling the Internet and the Web: Probabilistic Methods and Algorithms.

Wiley, 2003.

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPD)

- Động lực và mục đích của KPD
- Định nghĩa và quá trình KPD
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPD

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- **Các nền tảng, thư viện và công cụ phần mềm**
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Nền tảng phần mềm thu thập (crawling) dữ liệu

- Scrapy: A Fast and Powerful Scraping and Web Crawling Framework (Python).
- Nutch: Highly Extensible, Highly Scalable Web Crawler (Java).
- Norconex HTTP Collector: Open-Source Enterprise Web Crawler (Java).
- Heritrix Web Crawler (Java).
- The Xapian Project (C++).
- Sphinx Search Server (C++).
- ...

Nền tảng phần mềm học máy (machine learning) và KPDL

- Weka: Data mining software in Java.
- Scikit-learn: Machine learning in Python.
- H2O: Fast scalable machine learning.
- Apache Mahout: Scalable machine learning and data mining.
- TensorFlow: Open source software library for machine intelligence.
- Dato Open Source (GraphLab).
- MLPack: A scalable C++ machine learning library.
- NLTK: Natural language toolkit.
- OpenCV: Open source for computer vision.
- R: The R project for statistical computing.
- KNIME: The Konstanz Information Minder.
- RapidMiner: Open source predictive analytics platform.
- SAS: Statistical Analysis System.
- Oracle Data Mining.
- IBM BigInsights.
- ...

Nội dung

1 Giới thiệu về Khai phá dữ liệu (KPDL)

- Động lực và mục đích của KPDL
- Định nghĩa và quá trình KPDL
- Dữ liệu, các bài toán và lĩnh vực ứng dụng của KPDL

2 Yêu cầu và đánh giá môn học

- Các môn học và kỹ năng tiên quyết
- Thi và tính điểm

3 Nội dung chuyên môn ở lớp

- Các nội dung sẽ giảng
- Các dự án (projects)
- Các chủ đề seminar

4 Tài liệu tham khảo

- Giáo trình và sách tham khảo
- Các nền tảng, thư viện và công cụ phần mềm
- Các diễn đàn, hội nghị, tạp chí chuyên ngành

Các tổ chức, diễn đàn về KPDL

- ACM SIGKDD: ACM Special Interest Group on Knowledge Discovery and Data Mining (www.sigkdd.org).
- KDnuggets: Data Mining, Analytics, Big Data, and Data Science (www.kdnuggets.com).
- IEEE Big Data: bigdata.ieee.org.
- Data Science Central: www.datasciencecentral.com.
- Big Data University: bigdatauniversity.com.
- Data Science Community: datascience.community.
- ...

Các hội nghị và tạp chí về khai phá dữ liệu

- Các hội nghị chuyên ngành:
 - ▶ ACM SIGKDD: International Conference on Knowledge Discovery and Data Mining.
 - ▶ WSDM: ACM International Conference on Web Search and Data Mining.
 - ▶ SIAM International Conference on Data Mining.
 - ▶ IEEE ICDM: IEEE International Conference on Data Mining.
 - ▶ Big Data: International Conference on Big Data.
 - ▶ PKDD: Principles and Practice of Knowledge Discovery in Databases.
 - ▶ PAKDD: Pacific-Asia Conference on Knowledge Discovery and Data Mining.
 - ▶ ...
- Các tạp chí chuyên ngành:
 - ▶ Data Mining and Knowledge Discovery (Springer).
 - ▶ ACM Transactions on Knowledge Discovery from Data (ACM TKDD).
 - ▶ IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE).
 - ▶ KDD Explorations.
 - ▶ ...

Các hội nghị thuộc các lĩnh vực liên quan

- Databases và Data Engineering: ACM SIGMOD, VLDB, ICDE, ICKM, ...
- Machine Learning: NIPS, ICML, COLT, ECML, ACML, ...
- Information Retrieval và Web: ACM SIGIR, WWW, ICEC, ECIR, Web Intelligence, ...
- Natural language processing và Text/Web Mining: ACL, NAACL, EMNLP, HLT, CoNLL, COLING, EACL, MT Summit, IJCNLP, PAACL, PALIC, ...
- Artificial Intelligence: IJCAI, AAAI, AISTATS, UAI, ...