Khái niệm về khai phá dữ liệu (Data Mining) hay phát hiện tri thức (Knowledge Discovery) có rất nhiều cách diễn đạt khác nhau nhưng về bản chất đó là quá trình  tự động trích xuất thông tin có giá trị (thông tin dự đoán – Predictive Information) ẩn chứa trong khối lượng dữ liệu khổng lồ trong thực tế.

Data mining nhấn mạnh 2 khía cạnh chính đó là khả năng trích xuất thông tin có ích  **Tự động (Automated)** và  thông tin mang tính **dự đoán (Predictive).**

 Data Mining liên quan chặt chẽ đến các lĩnh vực sau:

Statistics (Thống kê): Kiểm định model và đánh giá tri thức phát hiện được

Machine Learning (Máy học): Nghiên cứu xây dựng các giải thuật trên nền tảng của trí tuệ nhân tạo giúp cho máy tính có thể suy luận (dự đoán) kết quả tương lai thông qua quá trình huấn luyện (học) từ dữ liệu lịch sử.

Databases (Cơ sở dữ liệu): Công nghệ quản trị dữ liệu nhất là kho dữ liệu

Visualization (Trực quan hóa): Giúp dữ liệu dễ hiểu, dễ sử dụng như chart, map

**Nhiệm vụ của Data Mining**

Nhiệm vụ của data mining có thể phân thành 2 loại chính đó là **dự đoán (Predictive)** và**mô tả (Descriptive).**

**Predictive:**

       Classification : Phân lớp

                  Regression : Hồi qui

       Deviation Detection: Phát hiện độ lệch

**Descriptive:**

       Clustering: Phân cụm

       Association Rule Discovery: Phát hiện luật kết hợp

*Dưới đây là một số thuật toán phổ biến được dùng trong Data Mining*

Decision tree: Cây quyết định (Classification Task)

Nearest Neighbor: Láng giềng gần nhất (Classification Task)

Neural Network: Mạng Neural (Classification and Clustering Task)

Rule Induction: Luật qui nạp (Classification Task)

K-Means: Thuật toán K-Means ( Clustering Task)

Data mining là một quá trình trích xuất thông tin có mối quan hệ hoặc có mối tương quan nhất định từ một kho dữ liệu lớn (cực lớn) nhằm mục đích dự đoán các xu thế, các hành vi trong tương lai, hoặc tìm kiếm những tập thông tin hữu ích mà bình thường không thể nhận diện được.

Ứng dụng của nó rất đa dạng và rộng tới, từ marketing, chống gian lận, giảm giá thành sản xuất, tăng doanh thu, phân tích hành vi sử dung người dùng internet để target đúng nhu cầu, đúng đối tượng hay ứng dụng hỗ trợ ra quyết định, nghiên cứu khoa học đến việc chống khủng bố v.v..

Các công cụ, kỹ thuật data mining có thể trả lời các câu hỏi mà các công cụ truyền thống đòi hỏi rất nhiều thời gian cần thiết để có thể giải đáp được (thậm chí các cách truyền thống không thể giải được). Nó có thể tìm thấy được những thông tin cực kỳ hữu ích mà rất dễ bị bỏ qua hoặc không xem xét đến để có thể dự đoán những xu thế/hành động xảy ra trong tương lai.

Để có thể data mining một cách hiệu quả, điều đầu tiên cần phải thu thập dữ liệu và định nghĩa lại theo các tiêu chí cần phân tích. Các kỹ thuật data mining có thể cài đặt rất nhanh chóng trên các nền tảng phần mềm, phần cứng phổ thông mà không cần đòi hỏi quá phức tạp, tuy vậy data mining thường gắn liền với việc phân tích một khối lượng dữ liệu cực lớn nên cần ứng dụng các công nghệ high performance client/server hoặc xử lý song song (parallel programming).

Công nghệ data mining là kết quả của quá trình lâu dài nghiên cứu và phát triển sản phẩm. Quá trình này bắt đầu từ khi những dữ liệu đầu tiên lưu vào máy tính, tiếp đến là quá trình cải tiến, nâng câp cách thức truy xuất dữ liệu và dần dần cho phép người dùng có thể duyệt dữ liệu theo thời gian thực. Kết quả của quá trình này cho phép truy xuất các nguồn dữ liệu quá khứ, tính toán và định vị lại việc truyền tải, cung cấp thông tin cho hiện tại, tương lai hay đưa ra các dự đoán tình huống có thể sắp diễn ra.

Có ba kỹ thuật cơ bản nhưng rất tổng quát và đầy đủ trong công nghệ data mining:  
Thu thập dữ liệu lớn. (Massive data collective).  
Nền tảng sức mạnh tính toán. Đây chính là hạ tầng bao gồm cả phần cứng và phần mềm lõi. (Powerful computing)

Như đã trình bày ở trên, pham vi ứng dụng của data mining rất rộng, bao phủ hầu hết các lĩnh vực. Tuy vậy có thể tóm lược lại phạm vi ứng dụng của data ming trong các bài toán yêu cầu có cần có các chức năng/nghiệp vụ sau:  
Tự động hóa việc dự doán các xu thế và hành vi sẽ diễn ra trong tương lai: Data mining tự động hóa quá trình tìm kiếm và trích xuất các tập thông tin có mối quan hệ hoặc tương quan trong một tập dữ liệu cực lớn. Những vấn đề/câu hỏi đặt ra với các cách truyền thống đòi hỏi một quá trình rất phức tạp và tốn kém cả về tài chính và thời gian để giải đáp thì giờ có thể trả lời một cách nhanh chóng trong khi giá thành là thấp nhất. Một ví dụ đơn giản trong quảng cáo online là quá trình targeting, khi một người dùng click vào một banner quảng cáo bán giầy dành cho nữ giới, thì khả năng người đó là nữ và họ đang quan tâm đến đôi giầy đó, khi hệ thống nhận diện được điều này thì có thể target thêm các kiểu dáng giầy khác nhau, kích cỡ khác nhau, giá thành khác nhau với mục tiêu là người dùng sẽ kết thúc quá trình là mua một đôi giày, ngoài ra bởi vì hệ thống nhận diện người dùng là phụ nữ nên nó có thể quảng cáo thêm các sản phẩm khác như mỹ phẩm, quần áo v.v..

Tự động hóa việc khám phá/nhận diện ra những tập thông tin quá khứ mà hệ thống không/chưa biết: Công cụ data mining trong quá trình phân tích dữ liệu sẽ nhận diện ra nhưng tập thông tin/dữ liệu không có hoặc bị ẩn bởi những mối liên hệ mà được xác định trước. Ví dụ như trong việc bán hàng, có những sản phẩm không có mối liên hệ gì với nhau, nhưng lại hay bán được cùng nhau, từ đó đưa ra cơ chế recommendation. Hay cơ chế nhận diện giao dịch gian lận trong thanh toán điện tử dựa vào việc phân tích các giao dịch bất thường…

Ngày nay, các công nghệ data mining được ứng dụng rộng rãi trong các công ty lấy khách hàng làm trung tâm như truyền thông, tài chính, marketing, bán hàng, các nghành công nghiệp sản xuất v.v… Nó cho phép các công ty xác định được các mối quan hệ giữa các yếu tố nội tại như giá thành, mẫu mã, cách thức quảng cáo, thậm chí là kỹ năng của nhân viên công ty… các yếu tố bên ngoài như đối thủ cạnh tranh, chính sách kinh tế hay nhu cầu thị trường v.v.. Và nó còn hỗ trợ việc xác định được sự tác động của các chính sách khuyến mại, giảm giá, độ hài lòng của khách hàng và lợi nhuận của doanh nghiệp…

Từ đầu đến giờ toàn nói data mining ứng dụng vào cái này cái kia mà chưa nói gì đến cách thức nó hoạt động thế nào. Phần tiếp theo chúng ta sẽ tìm hiểu kỹ hơn về vấn đề này.

Vậy chính xác thì data mining giúp bạn dự đoán được hoặc nhận biết nhưng điều chưa biết dựa trên tập dữ liệu quá khứ như thế nào? Các kỹ thuật ứng dụng để giải các câu hỏi trên trong data mining được gọi là quá trình mô hình hóa dữ liệu. Có nghĩa là xây dựng các mô hình trong các tình huống mà bạn đã biết/nhận diện được câu trả lời/kết quả và sau đó là áp dụng vào trong các tình huống khác mà bạn chưa biết/nhận diện được.

Vài công nghệ và kỹ thuật data mining thường áp dụng như: thống kê (statistical), máy học (machine learning), mạng nơ ron (neural network).

1. Thu thập, bóc tách, chuẩn hóa dữ liệu và nhập dữ liệu vào hệ thống kho dữ liệu (Datawarehouse).

2. Lưu trữ và quản lý dữ liệu dưới dạng đa chiều.

3. Đưa ra các cơ chế truy xuất cho các ứng dụng phân tích dữ liệu .

4. Sử dụng các phần mềm phân tích để tính toán.

5. Kết xuất dữ liệu dưới dạng dễ hiểu, như biểu đồ hoặc dạng report.

Để ví dụ cho quá trình này, chúng ta quay trở lại trả lời các câu hỏi ở mục Mở đầu.  
Đầu tiên là câu hỏi tại sao kết quả tìm kiếm của Google lại khác nhau khi đăng nhập và khi không đăng nhập? Tôi nghĩ đọc đến đây, có lẽ bạn đã có thể hình dung được câu trả lời. Câu trả lời ngắn gọn là: Khi bạn đăng nhập Google biết bạn là ai, biết hành vi trong quá khứ của bạn như thế nào (thông qua quá trình thu thập dữ liệu hành vi của bạn trên các site mà Google hiện diện), do đó nó dựa vào sự hiểu biết này nên kết quả trả về bao giờ cũng là kết quả có độ chính xác với nhu cầu của bạn cao nhất và ngược lại, khi không đăng nhập, Google chỉ trả về kết quả ranking bình thường mà không có kèm theo tham số ranking theo user. Và dĩ nhiên để biết được hành vi trong quá khứ của bạn, Google phải sử dụng đến các công nghệ data mining.

Thứ hai, tại sao Facebook lại có thể gợi ý cho tôi toàn những người mà tôi đã quen biết? Để có thể đưa ra gợi ý này đòi hỏi Facebook kết hợp rất nhiều các tham số, cụ thể là những tham số gì thì là bí mật công nghệ riêng của FB. Tuy nhiên có thể giải thích tóm lược như sau: Khi bạn đăng ký tài khoản Facebook, thông thường bạn nhập them các thông tin khác như: quá trình làm việc ở các công ty, công ty bạn đang làm, trường/lớp bạn đã/đang học, các hội bạn tham gia, bạn gái/trai/vợ/chồng của bạn v.v… Đây chính là các thông tin mà Facebook có thể dựa vào để xác định ai là bạn của bạn. Ở đây, Facebook đã mô hình hóa các tham số có độ liên quan nhất định để đưa ra những gợi ý mà có xác xuất đúng rất cao, và bạn thấy đó, thực sự là rất đúng:D. News stream của Facebook còn phức tạp và hay ho hơn nữa. Tất các điều đó Facebook làm được là quá trình thu thập dữ liệu hoạt động của người dùng trên site, sau đó sử dụng các công nghệ data mining để cung cấp nội dung, tính năng phù hợp nhất cho người dùng.

Thứ ba, tại sao Amazon lại có thể đưa ra danh sách các quyển sách bán kèm có tỷ lệ bán được cao như vậy? Để làm được điều này Amazon đã đầu tư phát triển hệ thống recommendation trong hàng chục năm, thời gian dài đó không chỉ là phát triển thuần túy công nghệ mà còn là quãng thời gian thu thập và phân tích hành vi người sử dụng trên site của Amazon, thời gian càng dài, dữ liệu thu thập càng lớn, dẫn đến tập thông tin có độ tương quan càng nhiều và càng chính xác. Nôm na bạn có thể hiểu là 100 khách hàng mua quyển sách ABC thì có tới 40 khách hàng đồng thời mua quyển sách DEF, vì thế với khách hàng 101 xem quyển ABC, hệ thống nhận diện và đưa ra gợi ý mua them quên DEF, đơn giản là vậy, tuy nhiên bên dưới còn có nhiều tham số khác để tăng độ tương quan lên nữa. Và đây cũng là một quá trình mining dữ liệu.

Mạng trí tuệ nhân tạo (Artificial neural networks): Đây là mô hình mà hệ thống có thể tự học thông qua đào tạo với tập dữ liệu ban đầu, từ đó suy đoán ra các tập kết quả từ tập dữ liệu mà nó khai thác.

Cây quyết định (Decisions Trees):  Một tập các decisions biểu diễn dưới dạng cây, các decisions này tạo ra các luật cho việc phân loại tập dữ liệu. Nôm na là, nếu tập thông tin A thõa mãn các luật B thì quyết định C.

Giải thuật di truyền (Generic Algorithms): Kỹ thuật này sử dụng trong các quá trình phối hợp, biến đổi, chọn lọc tự nhiên kế thừa từ khái niệm tiến hóa.  
Phương pháp ông hàng xóm gần nhất (Nearest neighbor method): Đây là kỹ thuật phân loại từng bản ghi/thông tin trong tập dữ liệu dựa trên sự kết hợp của k records có độ giống nhau nhất trong tập dữ liệu quá khứ.  
Nguyên tắc suy diễn (Rule induction): Kỹ thuật bóc tác dữ liệu dựa trên nguyên tắc Nếu-Thì từ các tập dữ liệu thống kê.

Các bài toán data mining hầu hết là các bài toán xử lý lượng dữ liệu lớn, và phân tích phức tạp. Điều đó đòi hỏi phải có hạ tầng đủ lớn và có khả năng mở rộng mềm dẻo, linh hoạt (scale out) phù hợp với sự tăng trưởng của dữ liệu.  
Mô hình hệ thống thông thường được xây dựng trên cơ sở phân tán, cho phép lưu trữ và tính toán song song, khả năng chịu lỗi cao.

Kích cỡ không gian lưu trữ dữ liệu tính bằng TB, băng thông giữa các node là Gb, tùy thuộc vào độ phức tạp của các công thức tính toán mà đòi hỏi hệ thống phải có lượng RAM đủ lớn, single CPU hay multi CPU v.v…  
Hệ thống thu thập dữ liệu phải có khả năng chịu tải cao và đảm bảo tính sẵn sang cao (High available), tốc độ xử lý nhanh (High

Bài viết này mới chỉ là bước dạo đầu để chúng ta có thể hình dung được data mining là gì. Các khái niệm, định nghĩa còn khá chung chung. Tuy vậy cũng đủ để các bạn có thể nắm được các điểm cốt lõi của một hệ thống data mining cần có những gì.

Ngày nay, việc xây dựng ra một sản phẩm (đặc biệt là sản phẩm internet) ngày càng trở nên đơn giản hóa, tuy vậy để tồn tại, phát triển và mở rộng cần có những tính năng thông minh hỗ trợ cho người dùng đơn giản hóa việc truy xuất, đúng thông tin họ cần… Tất cả điều này đều cần đến nền tảng data mining bên dưới.  
Các công nghệ data mining thực ra không mới, nó đã hình thành cách đây hơn 20 năm, từ khi các công nghệ điện toán xử lý song song phát triển. Tuy nhiên việc ứng dụng nó và các hệ thống web là điều còn khá mới mẻ, đặc biệt là các sản phẩm trực tuyến ở Việt nam chúng ta, hầu như chưa trang web nào ứng dụng thực sự các công nghệ data mining, lý do là quy mô chưa đủ lớn, lượng dữ liệu cần phân tích còn nhỏ

**Khai thác dữ liệu** (*data mining*) là quá trình khám phá các [tri thức](http://vi.wikipedia.org/wiki/Tri_th%E1%BB%A9c) mới và các tri thức có ích ở dạng [tiềm năng](http://vi.wikipedia.org/w/index.php?title=Ti%E1%BB%81m_n%C4%83ng&action=edit&redlink=1) trong nguồn [dữ liệu](http://vi.wikipedia.org/wiki/D%E1%BB%AF_li%E1%BB%87u) đã có.

Khai phá dữ liệu là một bước của quá trình [khai thác tri thức](http://vi.wikipedia.org/w/index.php?title=Khai_th%C3%A1c_tri_th%E1%BB%A9c&action=edit&redlink=1) (*Knowledge Discovery Proces*s), bao gồm:

* Xác định vấn đề và không gian dữ liệu để giải quyết vấn đề (*Problem understanding and data understanding*).
* Chuẩn bị dữ liệu (*Data preparation*), bao gồm các quá trình làm sạch dữ liệu (*data cleaning*), tích hợp dữ liệu (*data integration*), chọn dữ liệu (*data selection*), biến đổi dữ liệu (*data transformation*).
* Khai thác dữ liệu (*Data mining*): xác định *nhiệm vụ khai thác dữ liệu* và lựa chọn *kỹ thuật khai thác dữ liệu*. Kết quả cho ta một *nguồn tri thức thô*.
* Đánh giá (*Evaluation*): dựa trên một số tiêu chí tiến hành *kiểm tra* và *lọc* nguồn tri thức thu được.
* Triển khai (*Deployment*).

Quá trình khai thác tri thức không chỉ là một quá trình tuần tự từ bước đầu tiên đến bước cuối cùng mà là một quá trình lặp và có quay trở lại các bước đã qua.

* Phân loại (Classification): Discovery of a predictive learning function that classifies a data item into one of several predefined classes.
* Hồi qui (Regression): Discovery of a prediction learning function, which maps a data item to a real-value prediction variable.
* Phân nhóm (Clustering): A common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data.
* Tổng hợp (Summarization): An additional descriptive task that involves methods for finding a compact description for a set (or subset) of data.
* Mô hình ràng buộc (Dependency modeling): Finding a local model that describes significant dependencies between variables or between the values of a feature in a data set or in a part of a data set.
* Dò tìm biến đổi và độ lệch (Change and Deviation Dectection): Discovering the most significant changes in the data set.
* Vẫn có các mối lo ngại về [tính riêng tư](http://vi.wikipedia.org/w/index.php?title=T%C3%ADnh_ri%C3%AAng_t%C6%B0&action=edit&redlink=1) gắn với việc khai thác dữ liệu. Ví dụ, nếu một ông chủ có quyền truy xuất vào các hồ sơ y tế, họ có thể loại những người có bệnh tiểu đường hay bệnh tim. Việc loại ra những nhân viên như vậy sẽ cắt giảm chi phí bảo hiểm, nhưng tạo ra các vấn đề về tính hợp pháp và đạo đức.
* Khai thác dữ liệu các tập dữ liệu thương mại hay chính phủ cho các mục đích áp đặt luật pháp và an ninh quốc gia cũng là những mối lo ngại về tính riêng tư đang tăng cao. [5](http://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u#fn_5)
* Có nhiều cách sử dụng hợp lý với khai thác dữ liệu. Ví dụ, một CSDL các mô tả về thuốc được thực hiện bởi một nhóm người có thể được dùng để tìm kiếm sự kết hợp của các loại thuốc tạo ra các phản ứng (hóa học) khác nhau. Vì việc kết hợp có thể chỉ xảy ra trong 1 phần 1000 người, một trường hợp đơn lẻ là rất khó phát hiện. Một dự án liên quan đến y tế như vậy có thể giúp giảm số lượng phản ứng của thuốc và có khả năng cứu sống con người. Không may mắn là, vẫn có khả năng lạm dụng đối với một CSDL như vậy.
* Về cơ bản, khai thác dữ liệu đưa ra các thông tin mà sẽ không có sẵn được. Nó phải được chuyển đổi sang một dạng khác để trở nên có nghĩa. Khi dữ liệu thu thập được liên quan đến các cá nhân, thì có nhiều câu hỏi đặt ra liên quan đến tính riêng tư, tính hợp pháp, và đạo đức.

Các lĩnh vực hiện tại có ứng dụng *Khai thác dữ liệu* bao gồm:

* [Thiên văn học](http://vi.wikipedia.org/wiki/Thi%C3%AAn_v%C4%83n_h%E1%BB%8Dc)
* [Tin sinh học](http://vi.wikipedia.org/wiki/Tin_sinh_h%E1%BB%8Dc)
* Bào chế thuốc
* [Thương mại điện tử](http://vi.wikipedia.org/wiki/Th%C6%B0%C6%A1ng_m%E1%BA%A1i_%C4%91i%E1%BB%87n_t%E1%BB%AD)
* Phát hiện lừa đảo
* [Quảng cáo](http://vi.wikipedia.org/wiki/Qu%E1%BA%A3ng_c%C3%A1o)
* Marketing
* [Quản lý quan hệ khách hàng](http://vi.wikipedia.org/wiki/Qu%E1%BA%A3n_l%C3%BD_quan_h%E1%BB%87_kh%C3%A1ch_h%C3%A0ng) (CMR - Customer Relationship Management)
* Chăm sóc sức khỏe
* [Viễn thông](http://vi.wikipedia.org/wiki/Vi%E1%BB%85n_th%C3%B4ng)
* [Thể thao](http://vi.wikipedia.org/wiki/Th%E1%BB%83_thao), [giải trí](http://vi.wikipedia.org/wiki/Gi%E1%BA%A3i_tr%C3%AD)
* [Đầu tư](http://vi.wikipedia.org/wiki/%C4%90%E1%BA%A7u_t%C6%B0)
* [Máy tìm kiếm](http://vi.wikipedia.org/wiki/M%C3%A1y_t%C3%ACm_ki%E1%BA%BFm) (web)