BÀI 3

GIỚI THIỆU VỀ KHOA HỌC DỮ LIỆU

(Tiếp theo)

Học xong bài này, em sẽ:

- Biết được dữ liệu lớn là gì và các đặc trưng của dữ liệu lớn.
- Biết được vai trò của máy tính đối với sự phát triển của khoa học dữ liệu.
- Biết được tính ưu việt trong việc sử dụng máy tính và thuật toán hiệu quả để xử lí dữ liệu lớn, nêu được ví dụ minh hoạ.

Hiện nay người ta nói nhiều đến "Dữ liệu lớn". Em hãy lấy một ví dụ về dữ liệu lớn mà em biết.

1 Các đặc trưng của dữ liệu lớn

Dữ liệu lớn (Big Data) đề cập đến nguồn dữ liệu có khối lượng rất lớn, có tính đa dạng và phức tạp đến mức các công cụ truyền thống khó có thể lưu trữ và xử lí một cách hiệu quả.

Dữ liệu lớn có các đặc trưng thường được nêu tóm tắt bằng các chữ V, từ "3V" đến "5V", thậm chí đến "10V". Sau đây là năm chữ V nói về những đặc trưng thường được đề cập của dữ liệu lớn (*Hình 2*):

Hình 2. Năm chữ V của dữ liệu lớn

Khối lượng (Volume): Tập dữ liệu được coi là "dữ liệu lớn" có khối lượng ở mức nhiều petabyte hoặc exabyte. Ví dụ: Tập dữ liệu về hàng triệu khách hàng của một doanh nghiệp lớn có thể gồm hàng tỉ tệp, mỗi tệp nhiều megabyte.

Tốc độ (Velocity): Dữ liệu được tạo thêm rất nhanh và có thể cần xử lí hàng loạt, nhanh chóng theo thời gian thực để đáp ứng việc ra quyết định kịp thời. Ví dụ như quyết định về mua bán chứng khoán,... Các nguồn dữ liệu như: thiết bị cảm biến, mạng xã hội và các trang web,... tạo ra luồng dữ liệu lớn và liên tục. Lưu trữ và quản lí một lượng dữ liệu lớn, không ngừng tăng lên hằng ngày, liên quan đến một phạm vi rộng trên khắp thế giới là một thách thức.

Tính đa dạng (Variety): Dữ liệu đến từ nhiều nguồn khác nhau, dưới các dạng khác nhau như văn bản, hình ảnh, âm thanh, video,.... Ví dụ: Facebook mỗi ngày có thể tạo ra khoảng 500 terabyte dữ liệu. Tính đa dạng làm tăng độ phức tạp trong việc tổ chức lưu trữ, tìm kiếm, chuyển đổi khuôn dạng,... để các phần mềm phân tích dữ liệu có thể xử lí được.

Tính xác thực (Veracity): Đề cập đến độ tin cậy và độ chính xác của dữ liệu, bao gồm các yếu tố như: chất lượng dữ liệu, tính toàn vẹn, tính nhất quán và tính đầy đủ. Tính xác thực rất quan trọng trong việc đảm bảo rằng những hiểu biết sâu sắc được tạo ra từ dữ liệu lớn là chính xác và đáng tin cậy. Dữ liệu lớn đến từ nhiều nguồn khác nhau làm cho việc đảm bảo tính xác thực là một thách thức.

Giá trị (Value): Dữ liệu lớn có tiềm năng mang lại những thông tin và tri thức có giá trị, từ đó đưa ra những quyết định mang lại hiệu quả cao. Xử lí dữ liệu lớn để khai thác được các giá trị tiềm năng cũng là một thách thức. Ví dụ: Dự án Bộ gen người HGP có thể coi là một dự án dữ liệu lớn. Kết quả của dự án là vô giá vì nó mở ra một kỉ nguyên mới trong lĩnh vực y tế và chăm sóc sức khoẻ con người.

Quản lí và khai phá lượng lớn dữ liệu mang lại các lợi ích tầm chiến lược nhưng có nhiều thách thức.

2 Phân tích dữ liệu, phát hiện tri thức

a) Phân tích dữ liệu

Trong môn Toán, nội dung "Thống kê và xác suất" có phần "Phân tích và xử lí dữ liệu" với yêu cầu vận dụng các kiến thức để giải quyết một số bài toán thực tiễn. Em hãy nêu một số vấn đề thực tế có thể giải quyết bằng phân tích và xử lí dữ liệu thống kê. Theo em, đây có phải là phát hiện tri thức không?

Phân tích dữ liệu là quá trình kiểm tra, làm sạch, chuyển đổi và lập mô hình dữ liệu với mục đích tìm ra các thông tin hữu ích từ dữ liệu để đưa ra kết luận hoặc dự đoán. Phân tích dữ liệu có thể chia thành hai loại tuỳ theo mục đích là phân tích mô tả và phân tích dự đoán.

– *Phân tích mô tả* là tóm tắt dữ liệu quá khứ và trình bày trực quan, giúp người sử dụng dễ dàng nắm bắt được những thông tin quan trọng cần biết. Các thông tin rút ra từ tập dữ liệu được biểu diễn bằng sơ đồ, biểu đồ, đồ thị,... giúp người sử dụng dễ nhận ra các mẫu hoặc xu hướng, có cái nhìn rõ ràng, tổng thể về vấn đề cần giải quyết.

– *Phân tích dự đoán* nhằm đưa ra dự đoán (dự báo) hoặc phân loại dữ liệu mới.

Ví dụ: Nhằm điều chỉnh giá bán hàng sao cho lợi nhuận thu được nhiều hơn, từ phân tích dữ liệu có thể đưa ra phỏng đoán "quá mức ngưỡng X đồng, giá bán càng cao thì doanh số càng giảm". Đây là một giả thuyết thống kê. Kiểm định giả thuyết thống kê nhằm ra quyết định có thể chấp nhận hay phải bỏ một giả thuyết. Nếu giả thuyết thống kê nêu trên được chấp nhận thì "mức ngưỡng X đồng là giá bán tốt nhất" là hiểu biết mới được rút ra từ dữ liệu đã có.

Dữ liệu chuỗi thời gian (time series) là chuỗi các điểm dữ liệu được ghi lại theo chu kì thời gian (ví dụ: hằng ngày, hằng tuần, hằng tháng). Phân tích chuỗi thời gian cho phép dự đoán các điểm dữ liệu trong tương lai, trước khi sự việc xảy ra.

Phân tích hồi quy là một kĩ thuật cho phép xác định mối quan hệ phụ thuộc của một giá trị muốn biết với các giá trị một số thuộc tính khác và cho phép dự đoán giá trị muốn biết khi có dữ liệu mới.

b) Khai phá dữ liệu, phát hiện tri thức

Phát hiện hay khám phá tri thức đề cập đến toàn bộ quy trình trích xuất tri thức từ dữ liệu. Khai phá dữ liệu là một bước trong quy trình này. *Khai phá dữ liệu* là phát hiện các mẫu, các xu hướng trong tập dữ liệu. Trong khai phá dữ liệu thường dùng các phương pháp giao thoa giữa học máy và thống kê.

Để trích xuất thông tin hữu ích từ các tập dữ liệu lớn có nhiều kĩ thuật khai phá dữ liệu khác nhau, trong đó phân loại, phân cụm là hai kĩ thuật khai phá dữ liệu đã được trình bày trong bài Giới thiệu về Học máy.

3 Vai trò của máy tính và thuật toán ưu việt với khoa học dữ liệu

a) Máy tính là công cụ quan trọng trong khoa học dữ liệu

Từ đầu thế kỉ XXI, sự phát triển của mạng xã hội, thiết bị di động, cảm biến,... đã tạo ra lượng lớn dữ liệu hằng ngày. Máy tính và thiết bị số là các công cụ thiết yếu để lưu trữ và xử lí lượng dữ liệu lớn này. Nhu cầu phân tích dữ liệu, trích xuất các giá trị từ dữ liệu, phát hiện tri thức từ dữ liệu để ra quyết định và lập kế hoạch đã thúc đẩy sự phát triển khoa học dữ liệu.

Trước đây, việc phân tích dữ liệu, trích rút thông tin và tri thức chủ yếu do chuyên gia trực tiếp thực hiện. Hiện nay, máy tính đóng vai trò quan trọng trong việc xử lí và phân tích dữ liệu để đạt các mục tiêu của khoa học dữ liệu. Các giai đoạn của dự án khoa học dữ liệu như thu thập dữ liệu, chuẩn bị dữ liệu, phân tích dữ liệu đều cần đến máy tính.

Máy tính mang lại khả năng lưu trữ và quản lí dữ liệu hiệu quả. Trí tuệ nhân tạo nói chung và học máy nói riêng nghiên cứu phát triển các công cụ, quy trình, thuật toán để mô hình hoá dữ liệu, tự động phát hiện tri thức trong dữ liệu. Khoa học dữ liệu đang phát triển mạnh mẽ nhờ có học máy và trí tuệ nhân tạo. Các công ty lớn như

Uber, Google, Facebook và nhiều doanh nghiệp khác đã lập các nhóm nghiên cứu về khoa học dữ liệu để tăng hiệu quả hoạt động kinh doanh. Theo Harvard Business Review (tháng 1 năm 2012), làm nhà khoa học dữ liệu là công việc hấp dẫn nhất của thế kỉ XXI. Nhà khoa học dữ liệu xếp hạng 3 trong các nghề công nghệ theo danh sách xếp hạng việc làm năm 2022 của US News & World Report. Cục thống kê lao động Mỹ (U.S. Bureau of Labor Statistics) dự đoán từ năm 2020 – 2023 mức tăng trưởng việc làm này là 30 – 35%.

b) Máy tính và thuật toán ưu việt giúp phân tích dữ liệu hiệu quả

Máy tính chạy các phần mềm phân tích dữ liệu để mô hình hoá dữ liệu, phát hiện tri thức trong dữ liệu. Các chuyên gia trong mỗi lĩnh vực ứng dụng sử dụng những phần mềm này để phát hiện vấn đề, lựa chọn và đề xuất giải pháp giải quyết vấn đề cho lãnh đạo tổ chức, doanh nghiệp.

Các siêu máy tính có tốc độ hàng nghìn tỉ phép tính một giây, có bộ nhớ và các ổ đĩa dung lượng rất lớn cho phép quản lí, lưu trữ dữ liệu lớn; các thuật toán ưu việt giúp phân tích, xử lí dữ liệu lớn để phát hiện được tri thức hữu ích. Khoa học máy tính và công nghệ thông tin đã phát triển các giải pháp ưu việt và tạo ra các công cụ hiệu quả để giải quyết những vấn đề mà dữ liệu lớn đặt ra.

Điện toán đám mây có nhiều ưu việt, mang lại những lợi ích rõ ràng cho người dùng. Dữ liệu lớn lưu trữ trên đám mây tiện lợi cho truy cập và sử dụng mọi lúc mọi nơi, chỉ cần có thiết bị kết nối Internet. Sử dụng dịch vụ điện toán đám mây, doanh nghiệp có thể tiết kiệm chi phí, không cần đầu tư vào cơ sở hạ tầng. Điện toán đám mây rất linh hoạt, thích nghi với các thay đổi mở rộng hoặc thu hẹp triển khai các tài nguyên số phù hợp với nhu cầu của tổ chức doanh nghiệp. Có thể kể tên một số dịch vụ điện toán đám mây phổ biến như Amazon Web Services, Microsoft Azure,...

Cơ sở dữ liệu NoSQL đề cập đến các giải pháp cơ sở dữ liệu bổ sung để làm việc với dữ liệu không cấu trúc, không được tổ chức để truy vấn theo SQL. Các thuật toán như: sắp xếp ngoài, tìm kiếm cho phép tổ chức lưu trữ linh hoạt, dễ dàng mở rộng cho lượng dữ liệu lớn và lượng người dùng cao, phù hợp để quản lí và phân tích dữ liệu lớn. Có thể kể tên một số hệ quản trị cơ sở dữ liệu NoSQL được sử dụng nhiều như: Amazon DynamoDB, Google MongoDB, IBM Cloudant hay nguồn mở như Apache Hadoop,...

Máy tính cụm (Cluster) là một tập hợp các máy tính tích hợp để hoạt động như một máy tính đơn nhất. Máy tính cụm có các tính năng ưu việt như: tính sẵn sàng cao, dễ mở rộng, dễ quản lí, tiết kiệm chi phí hơn so với các máy tính lớn có sức mạnh tương đương.

Các thuật toán song song thực hiện nhiều phép tính đồng thời, tiến hành nhiều tiến trình cùng lúc, có thể triển khai trên máy tính cụm. Ví dụ thuật toán sắp xếp nhanh song song chia mảng đầu vào thành các mảng con và thực hiện song song việc sắp xếp những mảng con này. Việc chia một bài toán lớn đòi hỏi tính toán rất nhiều thành các bài toán nhỏ hơn có thể giải đồng thời bằng các thuật toán song song làm giảm đáng kể khoảng thời gian xử lí lượng lớn dữ liệu và có kết quả kịp thời theo yêu cầu sử dụng. Các thuật toán song song có tính ưu việt giúp tăng tốc độ tính toán nhiều lần với chi phí thấp hơn so với sử dụng hệ thống phần cứng tương đương.

Trong buổi thảo luận nhóm, một số bạn có những phát biểu sau. Em hãy cho biết mỗi phát biểu là đúng hay sai:

a) Dữ liệu lớn có khuôn dạng xác định, ý nghĩa rõ ràng.

b) Phân tích dữ liệu nhằm rút ra các thông tin hữu ích còn tiềm ẩn.

c) Khai phá dữ liệu có mục đích tìm ra dữ liệu mới.

d) Học máy thúc đẩy việc phát triển những phương pháp mới để khai phá dữ liệu.

Trong bài học đã có ví dụ cho từng chữ V, em hãy nêu một ví dụ khác cho một trong năm chữ V về đặc trưng của dữ liệu lớn.

Câu 1. Dữ liệu lớn có những đặc trưng gì?

Câu 2. Điều gì thể hiện máy tính là công cụ quan trọng trong khoa học dữ liệu?

Câu 3. Các thuật toán song song thể hiện tính ưu việt ở những điểm nào?

Tóm tắt bài học

- Dữ liệu lớn là nguồn dữ liệu với các đặc trưng như: khối lượng, tốc độ, tính đa dạng, tính xác thực, giá trị.
- Máy tính và thuật toán ưu việt đóng vai trò quan trọng trong việc xử lí và phân tích dữ liệu hiệu quả.
- Phân tích dữ liệu và khai phá dữ liệu đều có mục đích chung là rút ra tri thức tiềm ẩn từ dữ liệu, hiểu biết sâu sắc hơn về dữ liệu, có thể giúp giải quyết vấn đề hay đưa ra các dự đoán.