BÀI 27 MÁY TÍNH VÀ KHOA HỌC DỮ LIỆU

SAU BÀI HỌC NÀY EM SẼ:

- Biết được vai trò của máy tính đối với sự phát triển của Khoa học dữ liệu cùng tính ưu việt trong việc sử dụng máy tính và thuật toán hiệu quả để xử lí dữ liệu có kích thước lớn.

Những khả năng to lớn nào đã làm cho máy tính ngày nay trở thành một công cụ xử lí thông tin hữu hiệu?

1. VAI TRÒ CỦA MÁY TÍNH ĐỐI VỚI SỰ PHÁT TRIỂN CỦA KHOA HỌC DỮ LIỆU

Hoạt động **Tìm hiểu vai trò của máy tính trong Khoa học dữ liệu**

Hãy thảo luận và cho biết máy tính có vai trò như thế nào đối với sự phát triển của Khoa học dữ liệu.

Trong Khoa học dữ liệu, quy trình Khoa học dữ liệu là một chuỗi các bước được thực hiện để nghiên cứu, phân tích và khám phá tri thức từ dữ liệu. Quy trình này thường bao gồm các giai đoạn như thu thập và tiền xử lí dữ liệu; khám phá tri thức; phân tích, đánh giá, triển khai và báo cáo kết quả,... Quy trình đó có thể được coi như là một khung hành động để triển khai các dự án Khoa học dữ liệu, làm cho việc tương tác với dữ liệu trở nên có hệ thống và hiệu quả hơn, để chuyển đổi dữ liệu thành tri thức và thông tin hữu ích.

Máy tính có vai trò không thể thiếu trong mọi giai đoạn của quy trình khoa học dữ liệu. Nó cung cấp sức mạnh tính toán, khả năng lưu trữ và khả năng tự động hoá cần thiết để xử lí, phân tích và khám phá tri thức từ dữ liệu, góp phần vào sự phát triển và thành công của Khoa học dữ liệu. Vai trò quan trọng của máy tính đối với sự phát triển của Khoa học dữ liệu có thể được nhìn nhận từ nhiều góc độ khác nhau:

- *Xử lí và lưu trữ dữ liệu*: Máy tính cung cấp công cụ và phương tiện để xử lí, lưu trữ và quản lí khối lượng lớn dữ liệu. Nó cung cấp sức mạnh tính toán cần thiết để làm việc với dữ liệu lớn, phức tạp, được lưu trữ với nhiều định dạng khác nhau, từ các cơ sở dữ liệu đến hệ thống tệp phân tán.
- *Phân tích và khai phá dữ liệu*: Khoa học dữ liệu thường liên quan đến việc sử dụng các mô hình thống kê và Học máy để phân tích và khai phá dữ liệu phức tạp. Máy tính là phương tiện không thể thiếu để thực hiện các thuật toán, huấn luyện và kiểm nghiệm các mô hình học máy, nhằm khám phá tri thức từ dữ liệu, đưa ra dự đoán và xác định các mẫu.
- *Trực quan hoá dữ liệu*: Máy tính cho phép tạo ra các biểu diễn dữ liệu trực quan, giúp các nhà khoa học dữ liệu khám phá và trình bày những phát hiện của họ dễ dàng hơn. Các công cụ và thư viện trực quan hoá dữ liệu cho phép tạo nhiều loại biểu đồ, đồ thị và báo cáo tổng quan có khả năng tương tác.

- *Tự động hoá*: Quy trình khoa học dữ liệu thường bao gồm nhiều nhiệm vụ lặp đi lặp lại như làm sạch dữ liệu và huấn luyện mô hình. Nhiều công cụ máy tính có khả năng trợ giúp việc tự động hoá những tác vụ này, giảm thiểu các lỗi nảy sinh do các thao tác thủ công và tăng tốc quá trình xử lí, phân tích.
- *Xử lí song song*: Nhiều nhiệm vụ trong quy trình khoa học dữ liệu có khả năng song song hoá cao. Máy tính với bộ xử lí đa lõi, các siêu máy tính hoặc hệ thống tính toán phân tán có thể xử lí dữ liệu song song, giảm đáng kể thời gian cần thiết để phân tích, đặc biệt là khi xử lí dữ liệu lớn.
- *Điện toán đám mây*: Nền tảng đám mây cung cấp tài nguyên tính toán, bao gồm các dịch vụ và cơ sở hạ tầng đa dạng, cho phép các nhà khoa học có thể thực hiện việc phân tích dữ liệu mà không cần đầu tư vào phần cứng và những cơ sở hạ tầng đắt tiền (Hình 27.1).

Hình 27.1. Một trung tâm cung cấp dịch vụ điện toán đám mây

- *Hợp tác và truyền thông*: Thông qua các công cụ làm việc theo nhóm, làm việc từ xa, cùng các phương tiện chia sẻ thông tin và dữ liệu, máy tính hỗ trợ đắc lực cho việc phối hợp, cộng tác khoa học. Nhờ các công cụ và phương tiện máy tính, các nhà khoa học dữ liệu có thể truyền đạt những phát hiện của họ một cách hiệu quả tới các bên liên quan.

Máy tính có vai trò không thể thiếu trong mọi giai đoạn của quy trình khoa học dữ liệu. Nó cung cấp sức mạnh tính toán, khả năng lưu trữ và khả năng tự động hoá cần thiết để xử lí, phân tích và khám phá tri thức từ dữ liệu, góp phần vào sự phát triển và thành công của Khoa học dữ liệu, mở ra cơ hội làm việc với dữ liệu lớn mà trước đây không thể thực hiện được. Điều này đã thúc đẩy sự phát triển của lĩnh vực Khoa học dữ liệu, giúp tạo lập giá trị và tri thức từ nguồn dữ liệu lớn phong phú và đa dạng.

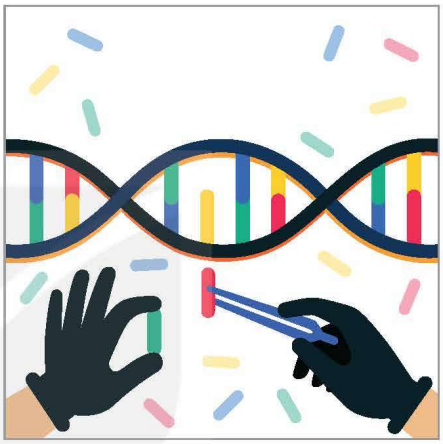
1. Hãy phân tích vai trò của máy tính trong việc thu thập và lưu trữ dữ liệu phục vụ quá trình Khoa học dữ liệu.

2. Các công cụ trực quan hoá dữ liệu của máy tính có vai trò như thế nào trong Khoa học dữ liệu?

2. TÍNH ƯU VIỆT TRONG VIỆC SỬ DỤNG MÁY TÍNH VÀ THUẬT TOÁN HIỆU QUẢ ĐỂ XỬ LÍ DỮ LIỆU LỚN

Tính ưu việt của việc sử dụng máy tính và các thuật toán hiệu quả trong xử lí dữ liệu lớn nói chung và dữ liệu có kích thước lớn nói riêng được thể hiện qua khả năng lưu trữ, xử lí, phân tích, khai phá dữ liệu ấy một cách nhanh chóng, nhất quán và hiệu quả. Để nhận biết được điều này ta sẽ xem xét một ví dụ cụ thể.

Hệ gene người (cũng như các loài khác) là một chuỗi các nucleotide, kí hiệu là A, C, G, T, mang thông tin di truyền quyết định đến hình dáng, sức khoẻ, bệnh tật và thậm chí cả tính cách con người. Nói một cách đơn giản, hệ gene người có thể được xem như là một chuỗi có độ dài khoảng 3 tỉ các kí tự A, C, G, T. Chuỗi kí tự này của hai người bất kì là khác nhau, trừ vài trường hợp đặc biệt, ví dụ sinh đôi từ cùng một trứng. Giải trình tự gene (Hình 27.2) là việc xác định trình tự xuất hiện các kí tự A, C, G, T trong chuỗi kí tự đó. Tuy nhiên các máy giải trình tự gene thường chỉ xác định được các đoạn nucleotide ngắn, có chiều dài vài trăm kí tự và cũng không xác định được các đoạn này nằm ở vị trí nào trên hệ gene. Người ta thu thập rất nhiều đoạn ngắn như vậy và lắp ráp hàng triệu đoạn ngắn này thành một hệ gene hoàn chỉnh. Quá trình này rất phức tạp, cần hệ thống máy tính mạnh, các thuật toán có độ chính xác cao và tốc độ nhanh để thực hiện.



Hình 27.2. Giải trình tự gene

Dự án Hệ gene người (Human Genome Project - HGP) là một nỗ lực khoa học mang tính đột phá, nhằm xác lập hệ gene và giải mã bản thiết kế di truyền hoàn chỉnh của con người. Bằng cách xác định thứ tự của tất cả nucleotide trong hệ gene, Dự án tìm cách khám phá những bí mật về cấu trúc di truyền của con người. HPG tạo ra một lượng dữ liệu khổng lồ và đòi hỏi nguồn lực tính toán hết sức to lớn. Dưới đây là một vài số liệu cụ thể:

- *Kích thước dữ liệu*: Chuỗi kí tự được nói ở trên của hệ gene người có độ dài khoảng 107,8 tỉ km. Việc giải trình tự toàn bộ hệ gene người tạo ra hàng trăm gigabyte dữ liệu thô.
- *Lưu trữ dữ liệu*: Việc lưu trữ dữ liệu từ HGP là một thách thức đáng kể. Tổng dung lượng lưu trữ cho dữ liệu HGP được ước tính chiếm khoảng một trăm nghìn gigabyte.
- *Sức mạnh xử lí*: Phân tích dữ liệu HPG đòi hỏi nguồn lực tính toán mạnh mẽ. Vào thời kì đỉnh cao, HGP dựa vào mạng lưới siêu máy tính trên khắp thế giới. Sức mạnh tính toán được sử dụng trong Dự án tương đương với hàng nghìn máy tính xách tay hiện đại hoạt động đồng thời.

Được thực hiện từ năm 1990 đến năm 2003, sự thành công của HPG đã cung cấp rất nhiều thông tin về gene người và chức năng của chúng, làm thay đổi hiểu biết hiện nay về di truyền học, dẫn tới nhiều tiến bộ y học và khoa học. Nó mở đường cho việc phát triển y học cá nhân hoá, nghiên cứu bệnh tật, đồng thời cho phép đánh giá sâu sắc hơn về sinh học con người. HGP cũng cho thấy tầm quan trọng to lớn của mạng máy tính, các phương pháp và kĩ thuật quản lí dữ liệu,... tiên tiến trong nghiên cứu

bộ gene. Tính ưu việt của việc sử dụng máy tính và các thuật toán hiệu quả trong việc xử lí dữ liệu lớn cho HPG được thể hiện ở nhiều khía cạnh:

- *Tốc độ và hiệu quả*: Máy tính và thuật toán hiệu quả đã đẩy nhanh đáng kể quá trình phân tích lượng dữ liệu di truyền khổng lồ. Những gì có thể phải mất nhiều thập niên theo cách thủ công đã đạt được trong khoảng thời gian ngắn hơn nhiều, giúp Dự án có thể hoàn thành.
- *Độ chính xác*: Các quy trình tự động giúp giảm nguy cơ sai sót của con người trong phân tích dữ liệu và đảm bảo tính chính xác của trình tự bộ gene cuối cùng.
- *Xử lí dữ liệu*: Cơ sở hạ tầng tính toán cho phép quản lí và lưu trữ các bộ dữ liệu gene lớn, giúp tổ chức và truy cập thông tin di truyền mở rộng do Dự án tạo ra.
- *Tích hợp dữ liệu*: Máy tính và thuật toán tích hợp dữ liệu từ các nhóm và tổ chức nghiên cứu khác nhau giúp đảm bảo tính nhất quán trong khám phá tri thức từ dữ liệu và làm tăng thêm hiệu quả hợp tác khoa học.
- *Giải thích dữ liệu*: Các thuật toán phức tạp được sử dụng để giải thích thông tin di truyền, xác định gene, cùng các đặc tính và các vùng chức năng khác trong bộ gene.
- *Phân tích thời gian thực*: Khả năng này của máy tính cho phép đưa ra quyết định nhanh chóng, điều này rất quan trọng đối với tiến độ của Dự án và tác động khoa học của nó.
- *Xử lí song song*: Các kĩ thuật tính toán song song cho phép xử lí đồng thời nhiều luồng dữ liệu, tăng tốc đáng kể việc phân tích dữ liệu di truyền.
- *Khả năng mở rộng*: Cơ sở hạ tầng tính toán được thiết kế để xử lí quy mô và độ phức tạp của dữ liệu bộ gene, điều này rất cần thiết cho các dự án khoa học quy mô lớn như HPG.

Tính ưu việt của việc sử dụng máy tính và các thuật toán hiệu quả trong xử lí khối dữ liệu lớn được thể hiện qua khả năng lưu trữ, xử lí, phân tích, khai phá dữ liệu ấy một cách nhanh chóng, nhất quán và hiệu quả.

1. Để giải quyết những nhiệm vụ trong Dự án hệ gene người cần phải xử lí và lưu trữ khối lượng dữ liệu có quy mô lớn như thế nào?
2. Có thể thực hiện việc phân tích dữ liệu liên quan tới Dự án hệ gene người trên máy tính cá nhân thông thường hay không?

LUYỆN TẬP

1. Nêu ngắn gọn vai trò của máy tính trong sự phát triển của Khoa học dữ liệu.
2. Trong trường hợp xấu nhất, để sắp xếp các đoạn nucleotide ngắn thành hệ gene người hoàn chỉnh, ước tính cần bao nhiêu phép thử?

VẬN DỤNG

Sử dụng công cụ tìm kiếm trên Internet để biết được một số bài toán liên quan tới dữ liệu lớn cần tới tính ưu việt của máy tính và các thuật toán hiệu quả để giải quyết.