**Nội dung làm slide ngày 26/8**:

* **Phần mở đầu**

Khai phá dữ liệu là một ngành học có tính ứng dụng, áp dụng vào thực tiễn cao, khai thác dữ liệu đã đạt được thành công lớn trong nhiều các ứng dụng. Không thể liệt kê tất cả các ứng dụng mà khai thác dữ liệu đóng vai trò quan trọng. Các bài thuyết trình về khai thác dữ liệu trong các lĩnh vực ứng dụng chuyên sâu về tri thức chẳng hạn như tin sinh học và kỹ thuật phần mềm, yêu cầu điều trị chuyên sâu hơn và nằm ngoài phạm vi của cuốn sách này. Để chứng minh tầm quan trọng của các ứng dụng như một khía cạnh chính trong nghiên cứu và phát triển khai thác dữ liệu, chúng tôi thảo luận ngắn gọn về hai ứng dụng rất thành công và phổ biến của khai thác dữ liệu: kinh doanh thông minh và các công cụ tìm kiếm.

* **Phần 1 : Kinh doanh thông minh**

Điều quan trọng là các doanh nghiệp phải hiểu rõ hơn về tác nhân thương mại của tổ chức của họ, chẳng hạn như khách hàng, thị trường, nguồn cung cấp và nguồn lực và đối thủ cạnh tranh. **Business intelligence (BI)** cung cấp lịch sử, hiện tại và các dự báo về hoạt động kinh doanh. Ví dụ bao gồm báo cáo, phân tích trực tuyến xử lý, quản lý hiệu suất kinh doanh, thông minh cạnh tranh, nhập điểm chuẩn và phân tích dự đoán.

"Kinh doanh thông minh quan trọng như thế nào?" Nếu không có khai thác dữ liệu, nhiều doanh nghiệp có thể không có khả năng thực hiện phân tích thị trường hiệu quả, so sánh phản hồi của khách hàng về các sản phẩm tương tự, phát hiện ra điểm mạnh và điểm yếu của đối thủ cạnh tranh, giữ chân các khách hàng có giá trị cao và đưa ra quyết định kinh doanh thông minh.

Rõ ràng, khai thác dữ liệu là cốt lõi của kinh doanh thông minh. Các công cụ xử lý phân tích trực tuyến trong lĩnh vực kinh doanh thông minh dựa vào kho dữ liệu và khai thác dữ liệu đa chiều. Kỹ thuật phân loại và dự đoán là cốt lõi của phân tích dự đoán trong kinh doanh thông minh, trong đó có nhiều ứng dụng trong việc phân tích thị trường, nguồn cung cấp và bán hàng. Hơn nữa, phân cụm đóng vai trò trung tâm trong mối quan hệ với khách hàng quản lý, phân nhóm khách hàng dựa trên những điểm tương đồng của họ. Sử dụng các kỹ thuật khai thác đặc trưng, ​​chúng tôi có thể hiểu rõ hơn về các đặc điểm của từng nhóm khách hàng và phát triển các chương trình thưởng khách hàng tùy chỉnh.

* **Phần 2 : Công cụ tìm kiếm**

Công cụ tìm kiếm Web là một máy chủ máy tính chuyên dụng để tìm kiếm thông tin trên Web. Kết quả tìm kiếm của truy vấn người dùng thường được trả về dưới dạng danh sách (đôi khi được gọi là lần truy cập). Các lần truy cập có thể bao gồm các trang web, hình ảnh và các loại tệp khác. Một số công cụ tìm kiếm cũng tìm kiếm và trả về dữ liệu có sẵn trong cơ sở dữ liệu công cộng hoặc mở trực tiếp.

Công cụ tìm kiếm khác với thư mục web ở chỗ các thư mục web được duy trì bởi người biên tập trong khi các công cụ tìm kiếm hoạt động theo thuật toán hoặc bởi sự kết hợp của đầu vào của con người và thuật toán còn các công cụ tìm kiếm web về cơ bản là các ứng dụng khai thác dữ liệu rất lớn. Dữ liệu khác nhau, kỹ thuật khai thác được sử dụng trong tất cả các khía cạnh của công cụ tìm kiếm

* Thu thập thông tin (ví dụ: quyết định trang nào nên được thu thập thông tin và tần suất thu thập thông tin)
* Lập chỉ mục (ví dụ: chọn các trang được lập chỉ mục và quyết định mức độ chỉ mục nên được xây dựng)
* Tìm kiếm (ví dụ: quyết định cách xếp hạng các trang, phần quảng cáo nào nên được thêm vào và cách kết quả tìm kiếm có thể được cá nhân hóa hoặc thực hiện "Nhận biết ngữ cảnh").

Các công cụ tìm kiếm đặt ra những thách thức lớn đối với việc khai thác dữ liệu. Đầu tiên, họ phải xử lý một lượng dữ liệu khổng lồ và ngày càng tăng. Thông thường, dữ liệu đó không thể được xử lý bằng cách sử dụng một hoặc một vài máy. Thay vào đó, các công cụ tìm kiếm thường cần sử dụng các đám mây máy tính, bao gồm hàng nghìn hoặc thậm chí hàng trăm nghìn máy tính cộng tác khai thác lượng dữ liệu khổng lồ. Mở rộng quy mô các phương pháp khai thác dữ liệu trên các đám mây máy tính và các tập dữ liệu phân tán lớn là một lĩnh vực cần nghiên cứu thêm.

*Thứ hai*, các công cụ tìm kiếm web thường phải xử lý dữ liệu trực tuyến. Một công cụ tìm kiếm có thể đủ khả năng xây dựng một mô hình ngoại tuyến trên các tập dữ liệu khổng lồ. Để làm điều này, nó có thể xây dựng bộ phân loại truy vấn chỉ định truy vấn tìm kiếm cho các danh mục được xác định trước dựa trên chủ đề truy vấn (tức là truy vấn tìm kiếm “apple” có nhằm truy xuất thông tin hay không về một loại trái cây hoặc một nhãn hiệu máy tính). Cho dù một mô hình được xây dựng ngoại tuyến, ứng dụng của mô hình trực tuyến phải đủ nhanh để trả lời các truy vấn của người dùng trong thời gian thực.

Một thách thức khác là duy trì và từng bước cập nhật một mô hình trên các luồng dữ liệu đang phát triển nhanh chóng. Ví dụ: một bộ phân loại truy vấn có thể cần phải được tăng dần được duy trì liên tục vì các truy vấn mới giữ các danh mục mới nổi và được xác định trước và phân phối dữ liệu có thể thay đổi. Hầu hết các phương pháp đào tạo mô hình hiện có là ngoại tuyến và tĩnh và do đó không thể được sử dụng trong trường hợp như vậy.

*Thứ ba*, các công cụ tìm kiếm web thường phải đối phó với các truy vấn chỉ được hỏi số lần nhỏ. Giả sử một công cụ tìm kiếm muốn cung cấp truy vấn nhận biết ngữ cảnh khuyến nghị. Nghĩa là, khi người dùng đặt ra một truy vấn, công cụ tìm kiếm sẽ cố gắng suy ra ngữ cảnh của truy vấn bằng cách sử dụng hồ sơ của người dùng và lịch sử truy vấn của anh ta để trả về nhiều câu trả lời tùy chỉnh hơn trong một phần nhỏ của giây. Tuy nhiên, mặc dù tổng số truy vấn được hỏi có thể rất lớn, hầu hết các truy vấn chỉ có thể được hỏi một lần

hoặc một vài lần. Dữ liệu bị sai lệch nghiêm trọng như vậy là thách thức đối với nhiều hoạt động khai thác dữ liệu và phương pháp học máy.