Web Search Engines là tên gọi dành cho các website có chức năng tìm kiếm dữ liệu trên Internet hiện nay. Nổi bật nhất về Web Search Engines có thể kể đến Google Chrome, Cốc Cốc.

Kết quả tìm kiếm của truy vấn người dùng thường được trả về dưới dạng danh sách. các công cụ tìm kiếm hoạt động theo thuật toán hoặc bởi sự kết hợp của thuật toán và đầu vào của con người.

Các công cụ tìm kiếm web về cơ bản là các ứng dụng khai thác dữ liệu rất lớn. Các kỹ thuật khai thác dữ liệu khác nhau được sử dụng trong tất cả các khía cạnh của công cụ tìm kiếm, từ thu thập dữ liệu (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 được xây dựng), và tìm kiếm (ví dụ: quyết định cách xếp hạng các trang, quảng cáo nào nên được thêm vào và cách kết quả tìm kiếm có thể được cá nhân hóa hoặc được thực hiện "theo 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 không ngừng tăng lên. Thông thường, dữ liệu đó không thể được xử lý bằng một hoặc một vài máy. Thay vào đó, các công cụ tìm kiếm thường cần sử dụng các điện toán đám mây, bao gồm hàng nghìn hoặc thậm chí hàng trăm nghìn máy tính cùng hợp tác khai thác một lượng lớn dữ liệu. Mở rộng quy mô các phương pháp khai thác dữ liệu trên các điện toán đám mây 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 để xây dựng một mô hình ngoại tuyến trên các tập dữ liệu khổng lồ. Để thực hiện điều này, nó có thể tạo 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” nhằm truy xuất thông tin về trái cây hay 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. Ví dụ: một bộ phân loại truy vấn có thể cần được duy trì liên tục vì các truy vấn mới tiếp tục nổi lên và các danh mục 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 trên web thường phải đối phó với các truy vấn chỉ được hỏi một số lần rất nhỏ. Giả sử một công cụ tìm kiếm muốn cung cấp các đề xuất truy vấn nhận biết ngữ cảnh. 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 họ để trả lại 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 có thể chỉ được hỏi một lần hoặc một vài lần. Dữ liệu bị sai lệch nghiêm trọng như vậy là thách thức đối với nhiều phương pháp khai thác dữ liệu và học máy.