Information Retrieval hiểu một cách cơ bản là tìm những items trong cơ sở dữ liệu có liên quan đến query, thường là chưa có trong cơ sở dữ liệu. Ví dụ như Google Search và Google Search Image.

Bài toán đặt ra là cho một query, bạn phải sắp xếp, hoặc ít nhất là tìm kiếm, những items có liên quan trong cơ sở dữ liệu. Khi cơ sở dữ liệu là các hình ảnh thì nhánh này được gọi là Image Retrieval.

Nghiên cứu truy xuất thông tin liên quan đến việc phát triển các thuật toán và mô hình để lấy thông tin từ kho lưu trữ tài liệu.

15.1 Some Background on Information Retrieval.

Mục đích của nghiên cứu truy xuất thông tin là phát triển các mô hình và thuật toán cho việc truy xuất thông tin từ các tài liệu được lưu trữ, thông tin mang tính văn bản. Vấn đề cổ điển trong IR là truy xuất ad-hoc.

Trong truy xuất ad-hoc, người dùng nhập một truy vấn mô tả thông tin mong muốn, hệ thống trả lại danh sách các tài liệu (documents). Có 2 mô hình chính: Hệ thống kết hợp chuẩn xác (Exact match system) trả về các tài liệu đáp ứng chính xác một số biểu thức truy vấn có cấu trúc, trong đó loại được biết nhiều nhất là các truy vấn Boolean, vẫn được sử dụng rộng rãi trong các hệ thống thông tin thương mại. Nhưng với các bộ tài liệu lớn và không đồng nhất, tập kết quả thường rỗng hoặc lớn và khó sử dụng. Do đó, xuất hiện hệ thống xếp hạng các tài liệu theo mức độ liên quan với truy vấn. Hệ thống này sử dụng các phương pháp xác suất.

Ví dụ của truy xuất ad-hoc: Hình 15.1.

Một vài khía cạnh của truy xuất ad-hoc đã được giải quyết trong nghiên cứu IR, relevance feedback, bằng cách phản hồi liên quan, người dùng cải thiện công thức ban đầu của truy vấn một cách có tính tương tác; kết hợp các văn bản từ các cơ sở dữ liệu khác nhau vào một; mô hình nào chấp nhận dữ liệu bị hỏng một phần, ví dụ: OCRed documents (Optical Character Recognition, văn bản file ảnh).

Một số trường con của việc truy xuất thông tin dựa vào một kho dữ liệu huấn luyện của các tài liệu đã được phân loại là có liên quan hoặc không liên quan đến một truy vấn cụ thể. Trong phân loại văn bản (text categorization), chỉ định tài liệu cho hai hoặc nhiều danh mục được xác định trước. Ví dụ: mã chủ đề được phân công bởi Reuters cho các câu chuyện tin tức của nó (Lewis 1992). Các mã như CORP-NEWS (tin tức của công ty), CRUDE (dầu thô) hoặc ACQ (mua lại) giúp người theo dõi dễ dàng tìm thấy những chuyện họ quan tâm hơn. Nhà phân tích tài chính quan tâm đến việc mua lại có thể yêu cầu tùy chỉnh chỉ cung cấp tài liệu được gắn thẻ với ACQ.

15.1.1 Common design features of IR systems.

INVERTED INDEX (đánh chỉ số đảo ngược): Đa số các hệ thống IR đều có cấu trúc chính là đánh chỉ số đảo ngược.

Một chỉ số đảo ngược là một cấu trúc dữ liệu, chứa các bộ từ-tài liệu trong bộ các tài liệu có chứa nó và tần suất trong mỗi tài liệu đó.

Một phiên bản tinh vi hơn của chỉ mục đảo ngược cũng chứa thông tin vị trí.

Thay vì chỉ liệt kê các tài liệu mà một từ xuất hiện, vị trí của tất cả các lần xuất hiện trong tài liệu cũng được liệt kê.

Vị trí này có thể được mã hóa dưới dạng độ lệch byte tương ứng với bắt đầu của tài liệu.

Một chỉ số đảo ngược với thông tin vị trí cho phép chúng ta tìm kiếm cụm từ (Phrases).

VD: Để tìm car insurance, tạo 2 inverted index car, insurance. Đầu tiên, Lấy phần giao của 2 tập. Sau đó, xem thông tin vị trí và chỉ giữ lại những lần truy cập bởi thông tin vị trí đó.

STOP LIST.

Stop word là các từ không hữu ích trong tìm kiếm, không cần lưu vào inverted index. Mục đích để giảm bớt số inverted index phát sinh.

STEMMING.

Đơn giản hóa một từ, lấy gốc của từ, ví dụ: laugh có các hình thái laughing, laughed, laughs.

15.1.2 Evaluation measures.

15.1.3 The probability ranking principle (PRP)

15.2 The Vector Space Model.

15.2.1 Vector similarity.

15.2.2 Term weighting.

15.3 Term Distribution Models.

15.3.1 The Poisson distribution.

15.3.2 The two-Poisson model.

15.3.4 Inverse document frequency.

15.3.5 Residual inverse document frequency.

15.3.6 Usage of term distribution models.

15.4 Latent Semantic Indexing.

15.4.2 Singular Value Decomposition.

15.4.3 Latent Semantic Indexing in IR.

15.5 Discourse Segmentation.

15.5.1 TextTiling.

15.6