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. Những chức năng chung của các hệ thống truy xuất thông tin.

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

Một chỉ mục ngược là một cấu trúc dữ liệu lưu các documents chứa word nào đó.

VD:

Nâng cao hơn chỉ mục ngược lưu thêm vị trí của word trong document. Từ đó có thể tìm kiếm theo cụm từ (Phrases).

VD:

Cụm từ được đề cập trong sách, được sử dụng với định nghĩa khá cũ. Chỉ có thể tìm kiếm các cụm từ cố định. Đây là một lĩnh vực nghiên cứu thống kê NLP trong tương lai có thể đóng góp quan trọng vào việc truy xuất thông tin. Nghiên cứu gần đây nhất về các cụm từ trong IR đã đưa ra cách tiếp cận thiết kế một mô-đun nhận dạng cụm từ riêng biệt và sau đó lập chỉ mục các documents cho các cụm từ được xác định cũng như các từ. Để xác định cụm từ thì chọn những từ đi chung với nhau thường xuyên nhất (ví dụ ít nhất 25 lần).

Việc chia các cụm từ thành mô-đun, giống với vấn đề phát hiện collocation (Chương 5). Do đó, có thể áp dụng các kỹ thuật trong đó để tìm các cụm từ có hiệu quả cho việc đánh chỉ mục và tìm kiếm.

Collocation (She’s got black skin. She’s got dark skin).

STOP LIST.

Một stop list grammatical hoặc hư từ (function words: đang, với) liệt kê những từ được nghĩ rằng không có lợi cho việc tìm kiếm.

Các từ như the, from, could có chức năng ngữ nghĩa quan trọng trong tiếng Anh. Nhưng hiếm khi đóng góp thông tin khi thực hiện tìm kiếm word-by-word.

Theo luật ZipfV một stop list bao gồm nhiều tá từ (a few dozen words). 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. Tuy nhiên không thể tìm kiếm các cụm từ có chứa các từ trong stop list.

STEMMING. (vs Lemmatization)

Đưa một từ về dạng gốc của nó, ví dụ: laugh có các hình thái laughing, laughed, laughs.

Các bộ xử lý stemming là stemmer. Stemmer thông dụng là Lovins và Porter. Hai vấn đề trong việc chọn vị trí để cắt và sau khi cắt, không thể hiểu được từ đó nữa. Ví dụ: gallery – gall có thể cùng là gall sau khi stem; gallery có thể là gull.

15.1.2 Evaluation measures. Các biện pháp đánh giá.

Khi xây dựng một mô hình Machine Learning, cần một biện pháp đánh giá để xem mô hình sử dụng có hiệu quả không và để so sánh với các mô hình khác.

Một số phương pháp: accuracy score, confusion matrix, ROC curve, Area Under the Curve, Precision and Recall, F1 score, Top R error,...

Với bài toán phân loại mà tập dữ liệu của các lớp là chênh lệch nhau rất nhiều, có một phép đó hiệu quả thường được sử dụng là Precision-Recall.

Chất lượng của hệ thống truy xuất dựa vào cách nó quản lý xếp hạng các documents liên quan trước những documents không liên quan, các nghiên cứu truy xuất thông tin đã phát triển các biện pháp đánh giá thứ hạng. Hầu hết các biện pháp là kết hợp giữa độ chính xác và sự gọi lại (precision and recall). *Precision* là tỉ lệ của những documents có liên quan trong tập được hệ thống phân loại là có liên quan, *recall* là tỉ lệ documents hệ thống chọn là liên quan trong tất cả documents thực sự có liên quan. (Precision là trong tập tìm được bao nhiêu document đúng. Recall trong tất cả những cái đúng, tìm ra được bao nhiêu).

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