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).

Precision cao đồng nghĩa với việc độ chính xác của các document tìm được là cao. Recall cao đồng nghĩa với việc True Positive cao, tức tỉ lệ bỏ sót các document thực sự positive là thấp.

2 tỉ lệ này không âm, nhỏ hơn hoặc bằng 1.

Hình 15.2. Tuy nhiên, giả sử có 3 tập được truy xuất về, cùng số document liên quan và không liên quan. Khi đó nếu tính đơn giản precision, thì cả 3 đều là 50%. Nhưng ở kết quả 1 sẽ giúp người dùng dễ dàng tìm thấy hơn kết quả 2, và nó cũng rõ ràng hơn kết quả 3. Do đó, kết hợp precision với xếp hạng. Đó là những gì người dùng làm trong nhiều tình huống thực tế, ví dụ, khi tìm kiếm trên web.

Một biện pháp được sử dụng là độ chính xác tại một điểm cắt cụ thể, ví dụ 5 hoặc 10 tài liệu (các giới hạn điển hình khác là 20 và 100).

Bằng cách xem xét độ chính xác cho một số phân đoạn ban đầu của danh sách xếp hạng, người ta có thể có được ấn tượng tốt về phương thức xếp hạng các tài liệu có liên quan trước các tài liệu không liên quan.

15.1.3 The probability ranking principle (PRP)

Nguyên tắc xếp hạng xác suất.

Xếp hạng các documents có liên quan với query theo thứ tự giảm dần xác suất.

Ý tưởng cơ bản là truy xuất dưới dạng tìm kiếm tham lam nhằm xác định documents có giá trị nhất tại thời điểm bất kỳ.

Document d được cho là có giá trị nhất khi ước lượng xác suất liên qua của nó là cao nhất.

15.2 The Vector Space Model.

Mô hình không gian vector.

Được sử dụng rộng rãi cho truy xuất ad-hoc. Chủ yếu là vì khái niệm đơn giản và việc ẩn dụ khoảng cách không gian với sự gần gũi ngữ nghĩa.

Documents và query cùng được biểu diễn trong không gian nhiều chiều, trong đó mỗi chiều của không gian tương ứng với một từ trong tập các documents.

Các documents liên quan nhất cho một truy vấn được mong đợi là những documents đại diện bởi các vector gần nhất với truy vấn, nghĩa là các documents sử dụng các từ giống với truy vấn. Thay vì xem xét độ lớn của các vector, sự gần gũi thường được tính bằng cách chọn document có vector tạo với vector truy vấn một góc nhỏ nhất.

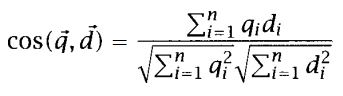
Tọa độ hay term *weights* được xác định bằng số lần xuất hiện. Ví dụ: hình 15.3 *insurance* chỉ có một tham chiếu tới d1, trong khi *car* có nhiều.

(Trong truy xuất thông tin, *term* sử dụng cho cả từ và cụm từ.)

15.2.1 Vector similarity.

Trong mô hình không gian vector, để truy xuất, documents được xếp hạng là giống với query bằng phương pháp *cosine* hoặc *normalized correlation coefficient*. (Hệ số tương quan).

Cosin: section 8.5.1.



Độ similarity của hai vector là 1 số trong đoạn [-1, 1]. Giá trị bằng 1 thể hiện hai vector hoàn toàn similar nhau. Hàm số cos của một góc bằng 1 nghĩa là góc giữa hai vector bằng 0, tức một vector bằng tích của một số dương với vector còn lại. Giá trị cos bằng -1 thể hiện hai vector này hoàn toàn trái ngược nhau.

15.2.2 Term weighting.

Có thể đếm số từ trong documents. Tuy nhiên, có nhiều phương pháp để đánh trọng số cho từ. Để tính trọng số cho term, cần sử dụng:

1frequency,

2document frequency,

3collection frequency.

Lưu ý: 2,3 chỉ có thể được sử dụng nếu có 1 collection.

Giả định này không phải lúc nào cũng đúng, ví dụ nếu các bộ sưu tập được tạo động bằng cách chọn một số cơ sở dữ liệu từ một tập lớn (như trường hợp trên một trong các dịch vụ thông tin trực tuyến lớn) và kết hợp chúng thành một bộ sưu tập tạm thời.

Một thông tin được lấy theo term frequency nó cho biết từ đó nổi bật trong documents như thế nào. Tần suất từ/cụm từ càng cao (từ thường xảy ra càng nhiều) càng có nhiều khả năng từ đó là mô tả tốt về nội dung của tài liệu.

Term frequency cần hàm để đưa về số nhỏ hơn.

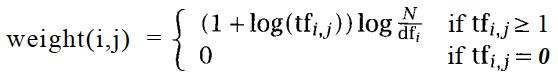
Document frequency, có thể được hiểu là một chỉ báo về tính thông tin.

Một từ có nghĩa thường xuất hiện nhiều lần trong một document.

Các từ có nghĩa thường được trải đều ra đồng nhất trên tất cả documents.

Ví dụ: trong một corpus tựa đề trong các tờ New York Times có từ “insurance” và “try”. Cả 2 cùng có chung collection frequency, nhưng “insurance” chỉ còn gần một nửa so với “try” trong document frequency.

Kết hợp term frequency và document frequency vào một trọng số.



Document frequency cũng được tỷ lệ theo logarit

Inverse Document Frequency. Ước lượng mức độ quan trọng của một từ, loại bỏ các trưởng hợp các từ phổ biến (không mang ý nghĩa thực sự).

TFIDF.

15.3 Term Distribution Models.

Một thay thế cho tf.idf là phát triển mô hình cho sự phân bố của một từ và sử dụng mô hình để mô tả tầm quan trọng của nó trong việc truy xuất. Tức là ước tính tỉ lệ Pi(k) số lần từ wi xuất hiện k lần trong một document.

Hầu hết các mô hình phân phối cố gắng mô tả làm thế nào mà một từ có tính thông tin. Cũng chính là thông tin mà inverse document frequency lấy. Người ta có thể bỏ vấn đề như là một trong những cách phân biệt các từ có nội dung với các từ không có nội dung (hoặc hàm), nhưng hầu hết các mô hình đều có một khái niệm được phân loại về mức độ thông tin của một từ.

Giới thiệu một số mô hình chính thức hóa các khái niệm về tính thông tin.

15.3.1 The Poisson distribution.

Phân phối Poisson. Phân phối xác suất rời rạc. Thông tin cho biết là trung bình số lần xảy ra thành công của một sự kiện trong một khoảng thời gian nhất định (Lamda).

Mô hình xác suất chuẩn để phân phối một loại sự kiện nhất định trên các đơn vị có kích thước cố định (chẳng hạn như khoảng thời gian hoặc khối lượng chất lỏng).

Ví dụ cổ điển về phân phối Poisson là số lượng các mục sẽ được trả về dưới dạng lỗi trong một khoảng thời gian nhất định, số lỗi nhập trên trang và số lượng vi khuẩn xảy ra trong một lượng nước nhất định.

Công thức Poisson:

Lamda\_i > 0 là trung bình số các w\_i xuất hiện trên 1 document. (w / document)

Trong trường hợp này, sự kiện cần quan tâm là sự xuất hiện của một từ cụ thể và đơn vị cố định document. Sử dụng PPPoission để ước lượng câu trả lời cho câu hỏi: Xác suất mà một từ xuất hiện một số lần cụ thể trong một document là gì?

Có thể gọi Pi(k) = p(k;lamda\_i) là xác suất của một document có chính xác k lần wi xuất hiện, trong đó lamda\_i được tính một cách thích hợp cho mỗi từ.

Ví dụ: hình 15.4.

Phân bố Poisson là giới hạn của phân phối nhị thức.

Kiểm tra tần suất tài liệu là cách dễ nhất để kiểm tra xem một thuật ngữ có được phân phối Poisson hay không. Số lượng tài liệu được dự đoán có ít nhất một lần xuất hiện của một cụm từ có thể được tính là phần bổ sung của số dự đoán không có lần xuất hiện.

15.3.2 The two-Poisson model.

Mô hình giả định rằng có 2 lớp các documents liên kết với một term, một lớp có số lần xuất hiện trung bình thấp (lớp không có đặc quyền) và một lớp với số lần xuất hiện trung bình cao (lớp có đặc quyền). Pi là xác suất một document nằm trong lớp đặc quyền, (1-pi) ngược lại.

Mô hình 2 Poisson phỏng rằng một từ có nội dung (a content word) đóng 2 vai trò khác nhau trong documents.

Trong lớp không đặc quyền, nó xuất hiện ngẫu nhiên, do đó không được sử dụng như một thuật ngữ chỉ mục, chỉ là một từ không có nội dung. (non-content). Trung bình số lần xuất hiện của từ trong lớp này là thấp.

Trong lớp đặc quyền, từ này là một từ có nội dung trọng tâm. Số lần xuất hiện trung bình của từ trong lớp này cao, do đó có thể sử dụng làm thuật ngữ chỉ mục.

15.3.3 The K mixture.

Được biết đến là một trong những mô hình phân phối cụm từ tốt nhất để nắm bắt các thuộc tính ngôn ngữ của các lần xuất hiện cụm từ.

Bản phân phối đơn giản hơn phù hợp với phân phối từ thực nghiệm cũng như nhị thức âm là Kats, K mixture:

Công thức:

15.3.4 Inverse document frequency.

15.3.5 Residual inverse document frequency.

Một thay thế cho IDF là RIDF. (residual = dư).

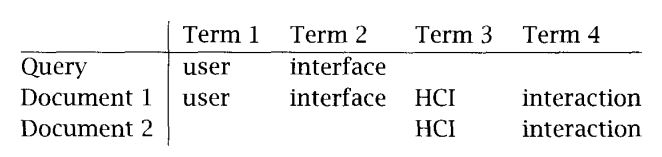
15.3.6 Usage of term distribution models.

Chúng ta có thể khai thác các mô hình phân phối hạn trong việc thu thập thông tin bằng cách sử dụng các tham số của mô hình phù hợp với một thuật ngữ cụ thể như các chỉ số. Ví dụ: chúng ta sử dụng RIDF hoặc beta trong K mixture thay thế cho trọng số IDF

sự liên quan.

15.4 Latent Semantic Indexing. Đánh chỉ mục ngữ nghĩa tìm ẩn.

Trong phần trước, chúng tôi đã xem xét các mẫu xuất hiện của từng từ riêng lẻ. Một nguồn thông tin khác về các thuật ngữ có thể được khai thác trong truy xuất thông tin là sự xuất hiện đồng thời: thực tế là hai hoặc nhiều cụm từ xuất hiện trong cùng một document thường xuyên.



Trong hình 15.8. Document1 có khả năng liên quan đến query vì nó chứa tất cả các từ trong query. Document 2 cũng mang lại kết quả tốt cho việc truy xuất. Its terms HCI and interaction co-occur with and interface, which can be evidence for semantic relatedness.

LSI là một kỹ thuật đưa ra các query và document vào một không gian với các tham số ngữ nghĩa tiềm ẩn.

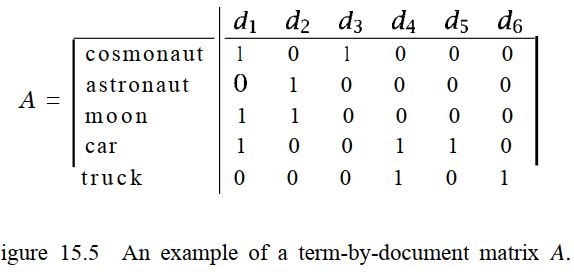
Những term xuất hiện cùng nhau, được chiếu lên cùng một chiều.

Những term không xuất hiện cùng nhau, được chiếu lên một chiều khác.

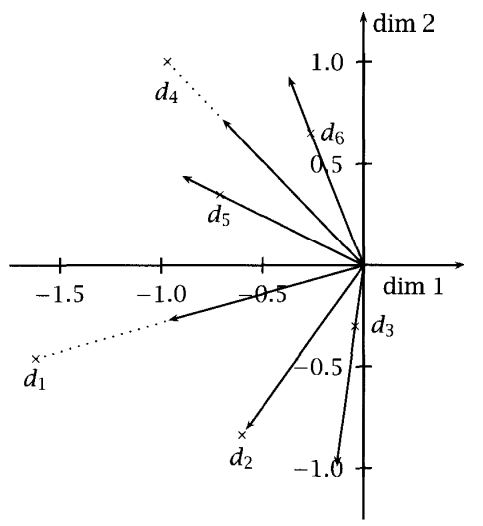
Trong không gian ngữ nghĩa tìm ẩn, một query và một document có thể có độ tương tự cosine cao ngay cả khi cả 2 không cùng chung term nào miễn là khi phân tích sự xuất hiện cùng nhau các term của chúng tương tự ngữ nghĩa.

Có thể coi LSI như một dạng chỉ số tương tự, thay thế cho các biện pháp chồng chéo từ như tf.idf.

LSI giảm bớt số chiều. Kỹ thuật giảm số chiều, thường đưa về không gian hai hoặc ba chiều cho mục đích trực quan hóa.



Biểu diễn không gian 5 chiều (5 từ bên trái) và 6 đối tượng trong không gian.



Sau khi áp dụng SVD chuyển từ 5 chiều về 2 chiều, những document có những từ có liên quan về mặt ngữ nghĩa sẽ gần nhau. (góc nhỏ).

D2, d3 không có term giống nhau.\*

Đầu tiên, xét trục mà trong không gian ban đầu lấy được nhiều biến thể của dữ liệu nhất. Chiều thứ 2, cũng lấy trục có lấy được nhiều biến thể của dữ liệu sau khi trừ trục 1 ra...

Điều này cho thấy tại sao LSI giống với Principle component Analysis (PCA) (reduction khác, PCA chỉ làm với ma trận vuông).

LSI là ứng dụng của SVD – Singular Value Decomposition.

15.4.1 Least-squares methods.

Tìm hàm số thỏa mãn các cặp giá trị cho trước.

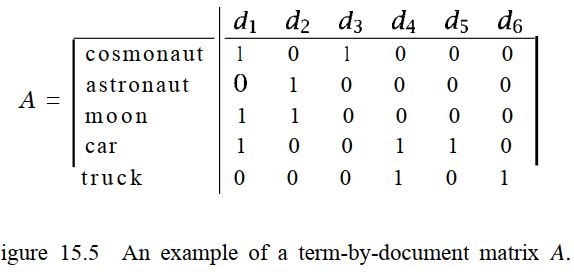
15.4.2 Singular Value Decomposition.

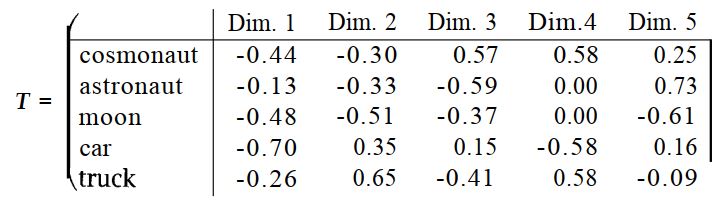
Xem SVD là phương pháp phân tích theo sự xuất hiện cùng nhau.

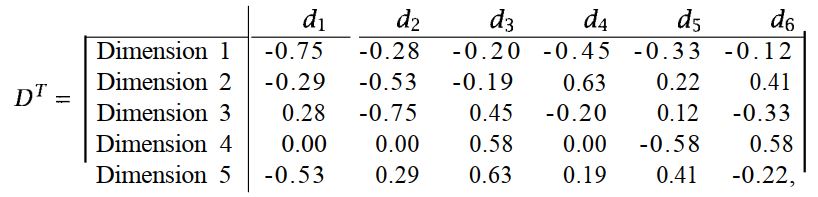


N = min(t,d)

Ma trận D^T là ma trận chuyển vị của D, D xoay quanh đường chéo của nó tạo thành D^T.







T, D là ma trận trực giao. (T^TxT=I, I là ma trận đơn vị).

Chúng ta có thể xem SVD như một phương pháp để xoay các trục của n-chiều không gian sao cho trục đầu tiên chạy dọc theo hướng biến thiên lớn nhất giữa các document, chiều thứ hai chạy dọc theo hướng với biến thể lớn thứ hai và vv.

Ma trận T, D biểu diễn terns, documents trong một không gian mới. Ví dụ:

Cột đầu tiên của T biểu diễn dòng đầu tiên của A, cột đầu tiên của D là cột đầu tiên của A.

15.4.3 Latent Semantic Indexing in IR.

15.5 Discourse Segmentation.

15.5.1 TextTiling.

15.6