**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**🙤🙧🟍🙥🙦**

****

**BÁO CÁO ĐỒ ÁN**

**CHỦ ĐỀ:**

**VECTOR SPACE MODEL**

**& LATENT SEMANTIC ANALYSIS**

Giảng viên hướng dẫn:**ThS. Nguyễn Trọng Chỉnh**

Sinh viên thực hiện:**19522093 – Lê Vinh Quang**

**19522351 – Trần Trung Tín**

**18520634 – Mai Viết Dũng**

**TP. HỒ CHÍ MINH – NĂM 2021 – 2022**

**LỜI MỞ ĐẦU**

Trong vài năm trở lại đây, với sự xuất hiện của Internet đã thay đổi hoàn toàn cách thức chúng ta tìm kiếm thông tin. Ví dụ khi cần tìm kiếm, đơn giản chỉ cần gõ một vài từ khóa vào máy tìm kiếm như Google hay Bing, ngay lập tức có được một danh sách tương đối chính xác các trang web có liên quan đến thông tin cần tìm. Với các hệ thống này, bằng cách lấy câu truy vấn từ người dùng, hệ thống cố gắng tìm kiếm các thông tin giống nhất trong cơ sở dữ liệu rồi trả lại cho người sử dụng. Đây là hệ thống tra cứu thông tin và trong bài viết này sẽ giới thiệu về 2 mô hình truy xuất thông tin là *“VECTOR SPACE MODEL”* và *”LATENT SEMANTIC INDEXING”.*

**I. Truy xuất thông tin.**

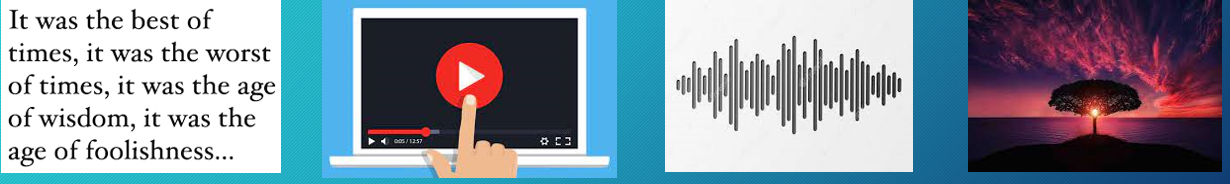
**1. Truy xuất thông tin là gì?**

Truy xuất thông tin là tìm kiếm thông tin (thường là các tài liệu) ở một dạng phi cấu trúc (thông thường là văn bản) thỏa mãn nhu cầu tìm kiếm thông tin từ trong những nguồn thông tin lớn (được lưu trữ trên máy tính).

**2. Khái niệm Documents**

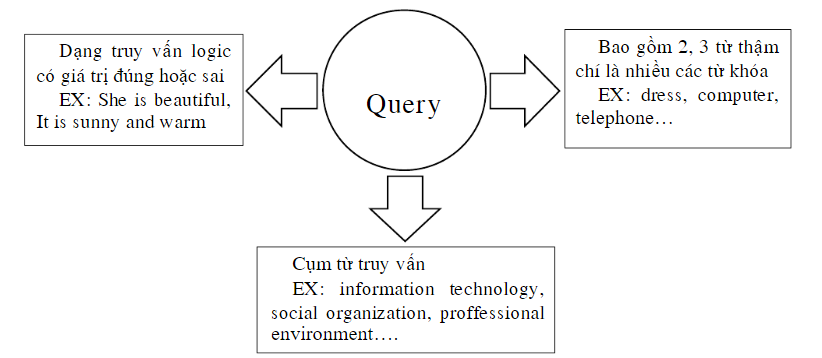
Trong truy xuất thông tin, còn có các khái niệm liên quan trong đó tài liệu là bất kì nội dung chứa thông tin và mang ý nghĩa. Các dạng tài liệu gồm các tài liệu văn bản và phi văn bản.

Tài liệu văn bản là các tài liệu dạng text, còn tài liệu phi văn bản là các tài liệu ở dạng: Video, âm thanh, hình ảnh,….



**3. Khái niệm Query**

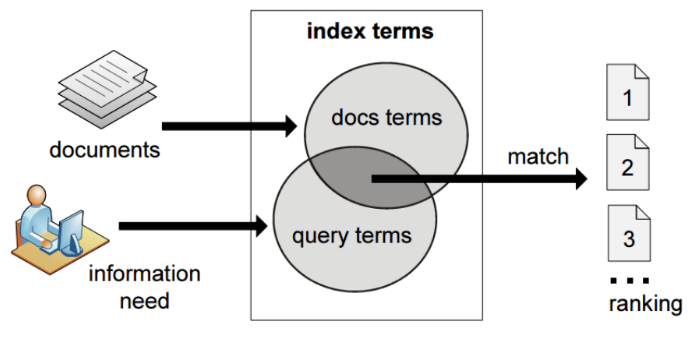
Là thông tin cần thiết do người dùng nhập vào hệ thống.



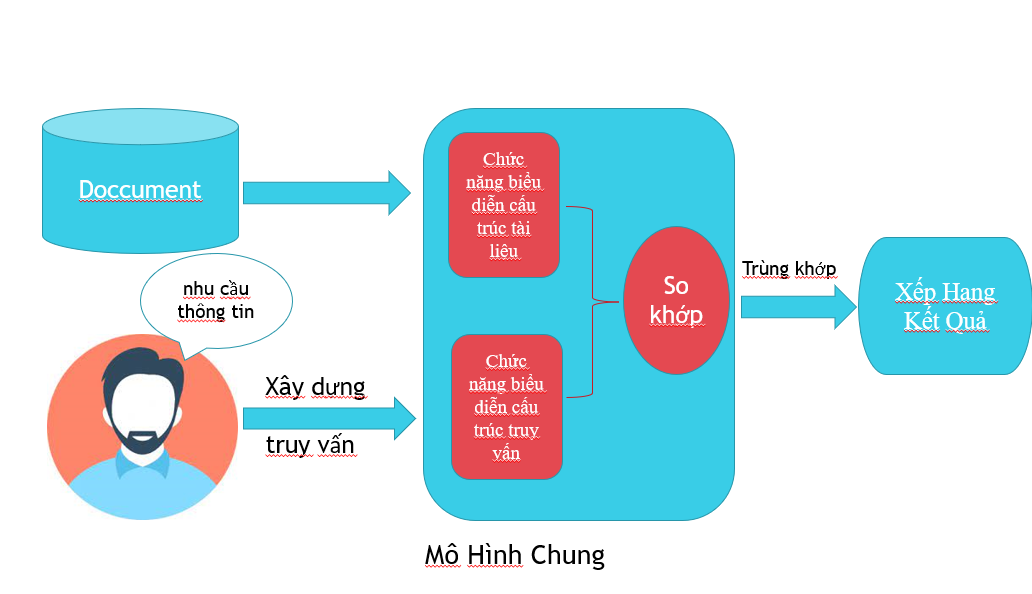
**4. Best-Match Retrieval**

Best-Match Retrieval là so sánh thuật ngữ trong một documents và query. Tính độ tương quan giữa mỗi documents trong kho tài liệu và query dựa trên thuật ngữ mà chúng có điểm chung.

Sắp xếp các documents theo thứ tự giảm dần độ tương quan với query. Kết quả đầu ra là một danh sách sắp xếp các documents và hiện thị đến người dùng mà các documents có độ liên quan cao được đánh giá bởi hệ thống.



Mô hình chung của truy xuất thông tin:



**II. Vector Space Model**

**1. Giới thiệu**

Mô hình Vector Space Model là một mô hình đại số thể hiện thông tin văn bản như một vector, các phần tử của vector này thể hiện mức độ quan trọng của một từ và cả sự xuất hiện hay không xuất hiện của nó trong tài liệu.

Vector Space Model được phát triển dựa trên kĩ thuật tf-idf và được phát triển bởi Gerard Salton vào đầu thập niên 1960s.

**2. Ý tưởng**

Ý tưởng của Vector Space Model là biểu diễn văn bản và các câu truy vấn dưới dạng Vector, Rep(d) của docs và Rep(q) của query sẽ cho kết quả là các vector. Sau đó tính độ tương đồng của query với từng documents theo công thức Sim(Rep(q), Rep(d)) để tìm ra docs nào phù hợp nhất với query.

**3. Ma trận Term-doc**

Để biểu diễn các tài liệu và câu truy vấn dưới dạng vector thì ta sẽ dùng các trọng số. Trong đó:

* Mỗi trọng số biểu diễn một chiều.
* K trọng số biểu diễn một không gian nhiều chiều.
* Ta cần định nghĩa trọng số (weight) cho từng chiều của Vector.

Ví dụ:

|  |  |  |  |
| --- | --- | --- | --- |
|  | D1 | D2 | D3 |
| computer | 1 | 0 | 0 |
| sreen | 1 | 1 | 0 |
| phone | 0 | 1 | 1 |

Ở ví dụ trên ta có 1 cái bảng với các cột đầu tiên là các term, còn hàng đầu tiên tương ứng với các tài liệu, những ô giá trị bên trong chính là giá trị trọng số giữa term và tài liệu. Trong trường hợp ở đây thì ta sử dụng trọng số có giá trị là 1 cho term có xuất hiện và trọng số có giá trị là 0 cho term không có xuất hiện trong tài liệu. Ở ví dụ trên thì các tài liệu sẽ được biểu diễn dưới dạng một vector 3 chiều, với mỗi chiều là một trọng số tương ứng giữa tài liệu với các term.

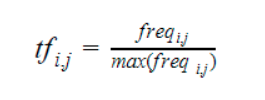
**4. Tìm trọng số cho Vector**

Cách xác định và tính trọng số cho Vector là hết sức quan trọng, ảnh hưởng đến độ chính xác của các thuật toán xếp hạng. Việc các từ có trọng số khác nhau là do không phải các từ đều có sự quan trọng giống nhau, sử dụng số lần xuất hiện của các từ làm Vector không phải là một cách tối ưu. Ở phương diện documents, một vài từ có thể mang nhiều thông tin hơn các từ còn lại.

Có nhiều kỹ thuật tinh trọng số: TF, IDF, TF-IDF,….Các kỹ thuật được giới thiệu giới đây sẽ được áp dụng tính cho cả 2 mô hình *“Vector Space Model”* và *“Latent Semantic Indexing”.*

**a. TF: Term frequency**

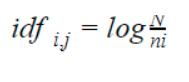
Từ nào xuất hiện nhiều trong câu thì quan trọng, công thức này sẽ đếm tần suất xuất hiện các từ trong câu.



Trong đó: freqi,j là số cụm từ i xuất hiện trong tài liệu dj.

**b. IDF: Tần suất tài liệu nghịch đảo**

Từ nào xuất hiện nhiều trong mọi câu thì không mang nhiều ý nghĩa (ví dụ như: a, the, are, thì, là,…). Vì vậy trọng số IDF là nghịch đảo của tần suất xuất hiện của các từ trong các documents.



Trong đó: ni là số các văn bản có từ hoặc cụm từ i.

N là tổng số văn bản.

**c. TF-IDF: Tần suất thuật ngữ - Tần suất tài liệu nghịch đảo**

Phép nhân giữa TF và IDF cho phép ta kết hợp cả 2 độ đo trên, từ vừa xuất hiện nhiều lần trong câu, vừa không phải là từ phổ biến xuất hiện trong mọi câu.

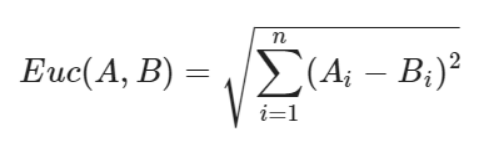
TF-IDF được tính bởi:

TF-IDF = TF \* IDF

**5. Các độ đo similarity**

Sau khi có được các vector cho query và docs, ta tính được similarity bằng cách tính khoảng cách giữa các vector.

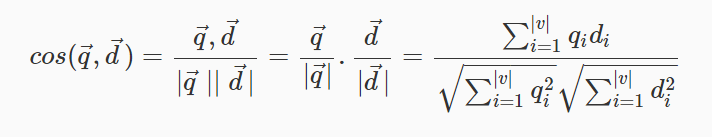
**a. Khoảng cách Euclide**



Nhược điểm: Docs dài hơn sẽ bị giảm hạn do chứa nhiều từ khác gây nhiễu, khiên khoảng cách xa hơn.

**b. Độ đo góc**

Việc tính xem các vector xem các vector overlappend với nhau như thế nào cũng là một phép so sánh hay sử dụng, độ đo góc phổ biến là Cosine Similarity.



Trong đó: qi là trọng số TF-IDF của từ i trong câu truy vấn.

di là trọng số TF-IDF của từ i trong tài liệu.

cos(, ) là sự tương đồng giữa và hay là cosine của góc giữa và .

Đối với các vector đã được chuẩn hóa về độ dài, sự tương đồng cosine chỉ đơn giản là tích vô hướng của hai vector.



Trong đó: q, d là độ dài được chuẩn hóa.

**6. Ưu và nhược điểm của Vector Space Model**

**a. Ưu điểm**

* Dễ hiểu và dễ cài đặt.
* Đã được nghiên cứu từ lâu và thực nghiệm cho kết quả tốt nhất và khả thi.
* Hiện tại là phương pháp được sử dụng rộng rãi.
* Có nhiều công thức tính TF-IDF khác nhau.

**b. Nhược điểm**

* Để cho kết quả đúng đắn, ta giả định:
* Các câu là độc lập với nhau.
* Các query và documents cùng loại với nhau.
* Có nhiều tham số để hiệu chỉnh:
* Bộ từ điển**.**
* Tham số trong các hàm Normalize.
* Threshold để chọn ra top kết quả.

**II. Latent Semantic Indexing**

**1. Đặt vấn đề**

Trong ngôn ngữ sẽ luôn có trường hợp những từ mang một lúc nhiều ý nghĩa và cũng có trường hợp mà nhiều từ khác nhau cùng mang trùng 1 ý nghĩa. Điều này sẽ gây sự mơ hồ nhọc nhằng cho người đọc, khó có thể hiểu được ý nghĩa mà từ ngữ hướng đến.



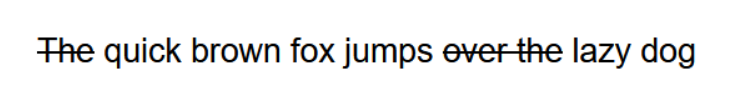
**2. Giới thiệu**

Latent Semantic Indexing được phát triển vào những năm 1980, LSI (Latent Semantic Indexing) sử dụng một phương pháp toán học giúp cho việc truy xuất thông tin chính xác hơn. Phương pháp này hoạt động bằng cách xác định các mối quan hệ ngữ cảnh ẩn giữa các từ.

**3. LSI hoạt động như thế nào?**

LSI hoạt động bằng cách tạo không gian khái niệm bằng kỹ thuật tách các giá trị đơn (SVD – Singular Value Decomposition) làm giảm kích thước của ma trận term-doc. SVD là một kỹ thuật làm giảm ma trận thành các phần cấu thành của nó để có các phép tính đơn giản và hiệu quả. Từ đó xác định được tọa độ vector tài liệu và truy vấn mới trong không gian khải niệm.

Khi phân tích một chuỗi từ, LSI loại bỏ các liên từ, đại từ và động từ thông dụng, còn được gọi là các stop word. Điều này tách biệt các từ bao gồm “nội dụng” chính cảu một cụm từ. Dưới đây là một ví dụ nhanh về cách điều này có thể trong như thế nào:



Các bước thực hiện của LSI:

* B1: Ta đánh trọng số cho documents và query. Sau khi đánh trọng số ta sẽ thu được ma trận doc A và ma trận truy vấn Q.
* B2: Áp dụng kỹ thuật SVD ta thu được các ma trận S, U,V và tính ma trận A = S\*U\*VT.
* B3: Sau khi áp dụng kỹ thuật SVD, ta giảm chiều ma trận và thu được những ma trận mới SS, VS, US.
* B4: Tìm các vector tài liệu và truy vấn, ta sẽ có các vector tài liệu và truy vấn mới.
* B5: Tính độ tương đồng và xếp hạng, ta sẽ có được các tài liệu có liên quan.

