**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**🙤🙧🟍🙥🙦**

****

**BÁO CÁO ĐỒ ÁN**

**CHỦ ĐỀ:**

**XÂY DỰNG MÔ HÌNH TRUY XUẤT THÔNG TIN**

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

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

**19522351 – Trần Trung Tín**

**19522390 – Phạm Minh Trí**

**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. Đối với hình ảnh, cũng đã có các hệ thống tương tự. Với hệ thống này, bằng cách lấy một ảnh đầu vào từ người dùng, hệ thống cố gắng tìm kiếm các ảnh giống nhất trong dữ liệu rồi trả lại cho người sử dụng. Đây là hệ thống tra cứu ảnh theo nội dung hay đơn giản là tra cứu ảnh. Trong bài viết này sẽ giới thiệu về 2 mô hình là *“truy vấn dữ liệu text”* và *“truy vấn dữ liệu ảnh”.*

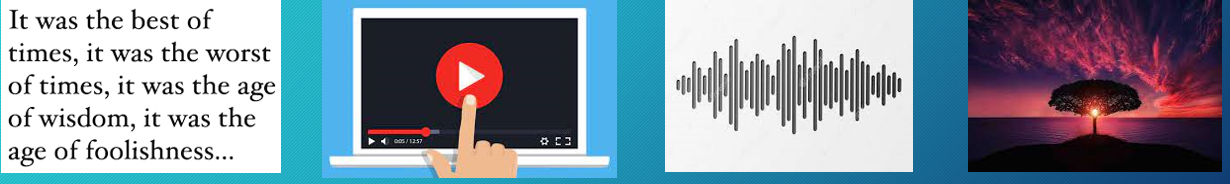
1. **Truy xuất thông tin.**
2. **Truy xuất thông tin là gì?**

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).

1. **Khái niệm Documents**

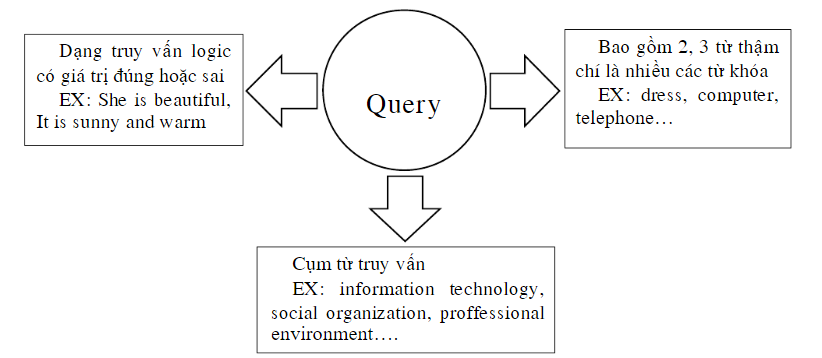
Trong truy xuất thông tin, còn có các khái niệm liên quan trong đó tài liệu là bất kì nội dung chứa thông tin và mang ý nghĩa. Các dạng tài liệu gồm các tài liệu văn bản và phi văn bản.

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,….



1. **Khái niệm Query**

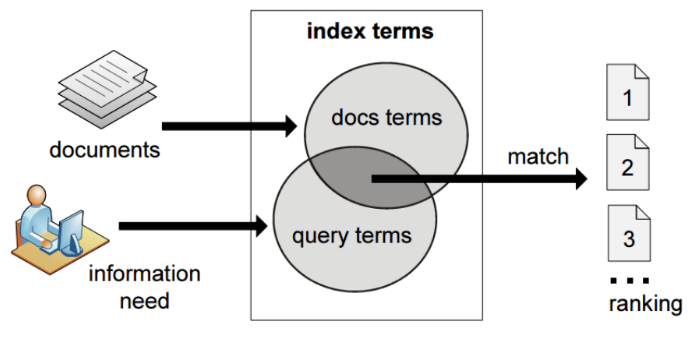
Là thông tin cần thiết do người dùng nhập vào hệ thống.



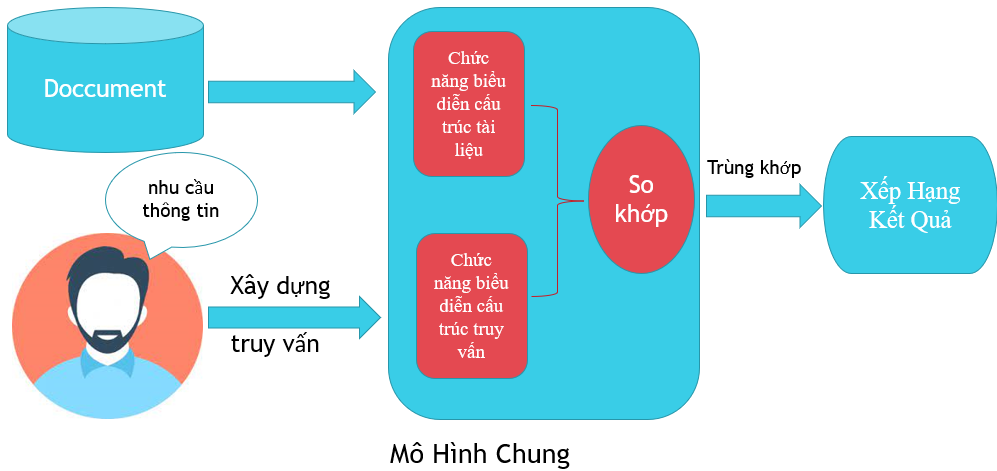
1. **Best-Match Retrieval**

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:



1. **Truy vấn dữ liệu text**
2. **Giới thiệu**

Trong tìm kiếm thông tin, Okapi BM25 là hàm tính thứ hạng được các công cụ tìm kiếm sử dụng để xếp hạng các văn bản theo độ phù hợp với truy vấn nhất định. Hàm xếp hạng này dựa trên mô hình xác suất, được phát minh ra vào những năm 1970 – 1980. Phương pháp có tên là BM25 (BM – Best Match 25), nhưng người ta thường gọi là *“Okapi BM25”*, vì lần đầu tiên công thức được sử dụng trong hệ thống tìm kiếm Okapi, được sáng lập tại trường đại học London những năm 1980 và 1990.

BM25 dựa trên nền tảng của TF/IDF, được cải tiến dễ sử dụng và được sử dụng rộng rãi trong tìm kiếm. Trong Web search những hàm xếp hạng này thường được sử dụng như một phần của các phương pháp tích hợp để dùng trong machine learning, xếp hạng.

1. **Phương pháp xếp hạng**

Ví dụ: Cho một query Q, có chứa những từ q1, q2, q3,…., qn BM25 sẽ đưa ra đánh giá phù hợp của tài liệu D đối với truy vấn Q như sau:

BM25(D, Q) =

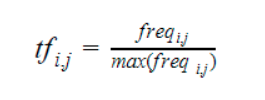
Trong đó:

* f(qi, D) : là TF tần suất xuất hiện của term qi trong tài liệu D
* |D| : là số từ trong tất cả tài liệu D
* davg : là số lượng từ trung bình trong mỗi tài liệu
* b và k1 : là hằng số của mô hình BM25, có giá trị k1 = 1.2, b = 0.75
* IDF(qi) : IDF cho term qi với tập gồm N tài liệu được tính bằng:

IDF(qi) = ln()

1. **TF**

TF – Term frequency: là tần số xuất hiện của 1 từ trong 1 văn bản. Từ nào xuất hiện nhiều trong câu thì quan trọng. TF được tính bằng công thức dưới đây:



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

1. **IDF**

IDF là tần suất tài liệu ngịch đảo. Từ nào xuất hiện nhiều trong mọi câu thì không mang nhiều ý ngĩa (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. IDF được tính bằng công thức dưới đây:

Trong đó: N là tổng số văn bản

ni là số các văn bản có từ hoặc cụm từ i

1. **TF–IDF**

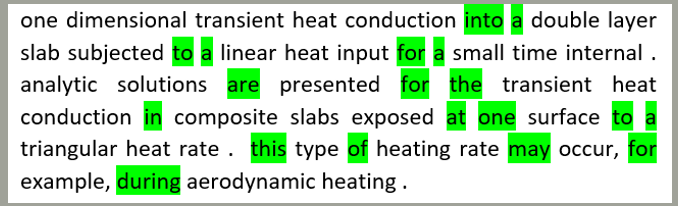
TF–IDF là tần suất thuật ngữ - Tần suất tài liệu nghịch đảo. Phép nhân giữa TF và IDF cho phép ta kết hợp cả 2 độ đo trên, từ vừa xuất hiện nhiều trong câu, vừa không phải là từ phổ biến xuất hiện trong mọi câu. TF–IDF được tính bằng công thức dưới đây:

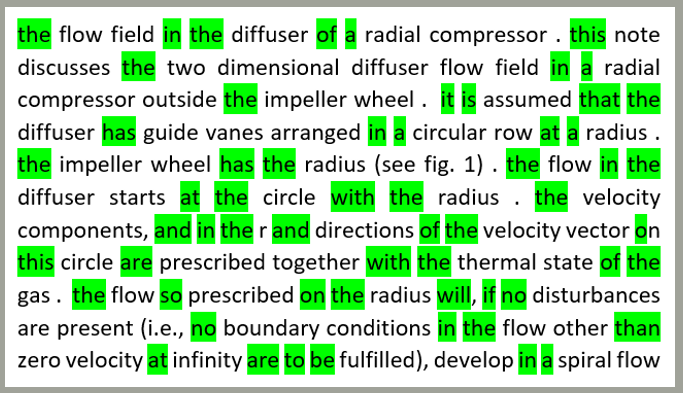
TF–IDF = TF \* IDF

1. **Phân tích tài liệu**

Trong một đoạn văn bản sẽ tồn tại từ dư thừa không cần thiết, không mang ý nghĩa trọng tâm mà chỉ bổ nghĩa cho câu.

Sử dụng nghững từ này làm term sẽ gây phân tán thông tin bởi văn bản nào cũng tồn tại những từ này thì khả năng tập trung vào những văn bản có nội dung thật sự liên quan sẽ trở nên kém đi.



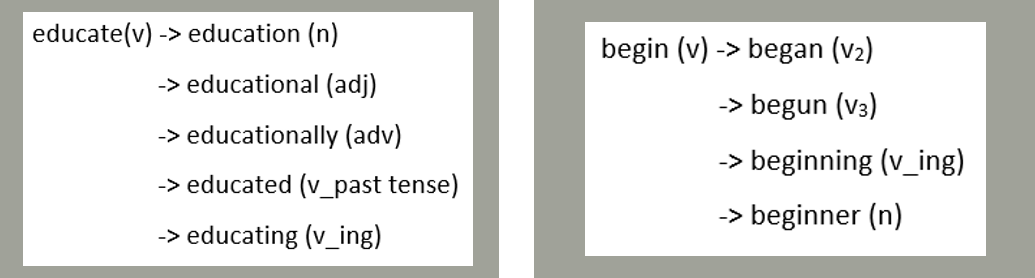


Với tất cả các tập tài liệu và câu truy vấn, ta sẽ có quá trình tiền xử lý như sau:

* Loại bỏ stopword
* Loại bỏ các kí tự đặc biệt
* Tokenize
* Stemming (lấy gốc từ)

Kết quả nhận được là tập các term. Term được trích xuất sau khi văn bản hoặc câu truy vấn trải qua ba giai đoạn xử lý:

* **B1: Xử lý kí tự**
* Kí tự “-” được giữ nguyên để tránh trường hợp các từ nối (vd: time-consuming, co-operate,….)
* Lượt bỏ các kí tự đặc biệt và các từ đi kèm, thường là những từ viết tắt (vd: i.e, Ex,…)
* Loại bỏ stopword thông qua từ điển các stopword thông dụng
* **B2: Tokenize**
* Tất cả các từ trong câu sẽ chuyển về dạng viết thường và tách ra để sử dụng như một feature để tính toán đặc trưng
* **B3: Stemming**
* Sau khi Tokenize ta đã có thể tính toán độ liên quan tuy nhiên ta còn một trường hợp cần quan tâm đó là những từ được chia dạng từ một từ gốc



* Nếu không chuyển về dạng gốc một câu có nhiều từ cùng chủ đề với một câu chỉ cố duy nhất 1 từ sẽ được xem là ngang nhau gây sai lệch truy xuất

1. **Lập chỉ mục**

Mục đích của việc lập chỉ mục là đánh chỉ mục các tập tin văn bản dùng để phục vụ cho việc tìm kiếm và đánh giá.

Chỉ mục bao gồm những thành phần:

* Tần số của từ trong mỗi tài liệu
* Chỉ số tài liệu có chứa từ khóa
* Doc\_len: độ dài mỗi tài liệu
* IDF của mỗi từ có trong tài liệu