* + 1. **Một số mô hình của máy tìm kiếm**
       1. **Mô hình Boolean**

Mô hình Boolean là mô hình cổ điển và đơn giản đã được sử dụng trước đây và cho đến nay vẫn còn được sử dụng trong các hệ thống IR. Mô hình Boolean dựa trên lý thuyết tập hợp (*set theory*) và đại số Boolean (*Boolean algebra*). Mô hình Boolean phổ biến bởi vì cả lý thuyết tập hợp và đại số Boolean có mối quan hệ đơn giản và dễ hiểu, vì vậy các hệ IR được xây dựng trên mô hình nầy, người dùng dễ dàng sử dụng.

Với mô hình Boolean văn bản được biểu diễn bởi một vector nhị phân, tức là các vector có các phần tử thuộc {0, 1}. Từ chỉ mục thứ *ki* xuất hiện trong văn bản *dj* thì trọng số *wij* = 1, ngược lại *wij* = 0.

Tất cả các truy vấn được biểu diễn bởi các biểu thức Boolean, sử dụng ba phép toán cơ bản: *not, and, or*, được biểu diễn trong hình 1.2. Văn bản truy vấn sử dụng mô hình này được xem như: hoặc liên quan đến nội dung truy vấn hoặc không, ở đây không có cách để để tìm các văn bản chỉ liên quan cục bộ hay còn gọi là liên quan một phần (*partially relevant*) của câu truy vấn. Ví dụ cho văn bản *d*, *d* có từ chỉ mục *kb*, tuy nhiên *d* được xem như không liên quan tới câu truy vấn *q = ka AND (kb or kc)*. Bởi vì *d* không có từ chỉ mục *kA* nên không liên quan (*irrelevant*) đến câu truy vấn.

****

*Hình 1.3 trình bày kết quả truy vấn q = ka AND (kb or kc)*.

* Ưu điểm của mô hình Boolean:
* Đơn giản và dễ sử dụng.
* Nhược điểm của mô hình Boolean:
* Vì dựa trên phép toán logic nhị phân nên một văn bản được tìm kiếm chỉ xác định hai trạng thái: liên quan hoặc không với câu truy vấn.
* Việc chuyển một câu truy vấn của người dùng sang dạng biểu thức Boolean không đơn giản.
  + - 1. **Mô hình không gian vector**

Mô hình không gian vector khắc phục những nhược điểm của mô hình boolean là việc sử dụng trọng số cho từ chỉ mục khác trọng số nhị phân (*non-binary*). Trọng số từ chỉ mục không giới hạn bởi hai trị 0 hoặc 1, các trọng số này được sử dụng để tính toán độ đo tương tự của mỗi văn bản với câu truy vấn. Với mô hình không gian vector, các văn bản, câu truy vấn và từ chỉ mục được biểu diễn thành các vector trong không gian vector. Sử dụng các phép toán trên không gian vector để tính toán độ đo tương tự giữa câu truy vấn và các văn bản hoặc các từ chỉ mục, kết quả sau khi tính toán có thể được xếp hạng theo độ đo tương tự với vector truy vấn. Ngoài ra, mô hình không gian vector còn hướng dẫn người dùng biết được những văn bản độ tương tự cao hơn có nội dung gần với nội dung họ cần hơn so với các văn bản khác.

Mô hình không gian vector dựa trên giả thiết là nội dung của văn bản có thể được hiểu như sự kết hợp của các từ chỉ mục. Một văn bản *d* được biểu diễn như một vector của các từ chỉ mục  với *ti* là từ chỉ mục thứ *i* (*1≤ i ≤ n*) ( *các giá trị có thể là số lần xuất hiện của term ti trong văn bản d*). Mỗi từ chỉ mục trong văn bản biểu diễn một chiều (*dimension*) trong không gian. Tương tự, câu truy vấn cũng được biểu diễn như một vector  .

Sau khi đã biểu diễn tập văn bản và câu truy vấn thành các vector trong không gian vector, ta có thể sử dụng độ đo *cosines* để tính độ đo tương tự giữa các vector văn bản và vector truy vấn.

* Ưu điểm của mô hình không gian vector:
* Đơn giản, dễ hiểu
* Cài đặt đơn giản
* Khắc phục các hạn chế trên mô hình Boolean
* Nhược điểm mô hình không gian vector:
* Số chiều biểu diễn cho tập văn bản có thể rất lớn nên tốn nhiều không gian lưu trữ.
  + - 1. **Mô hình xác suất**

Cho câu truy vấn của người dùng *q* và văn bản *d* trong tập văn bản. Mô hình xác suất tính xác suất mà văn bản *d* liên quan đến cấu truy vấn của người dùng. Mô hình giả thiết xác suất liên quan của một văn bản với câu truy vấn phụ thuộc cách biểu diễn chúng. Tập văn bản kết quả được xem là liên quan và có tổng xác suất liên quan với câu truy vấn lớn nhất.

* Ưu điểm của mô hình xác suất:
* Văn bản được sắp xếp dựa vào xác suất liên quan đến câu truy vấn
* Nhược điểm mô hình xác suất:
* Mô hình không quan tâm đến số lần xuất hiện của từ chỉ mục trong văn bản
* Việc tính toán xác suất khá phức tạp và tốn nhiều chi phí.