# **Lý do cần tìm kiếm trong HTTT**

* Cần lưu trữ
* Tìm để biết có hay không
* Lọc dữ liệu

# **Vai trò của dữ liệu & thông tin trong HTTKTT**

* Dữ liệu: số, chữ, hình ảnh, video,… phục vụ người dùng
* Thông tin: là dữ liệu đã được tinh chế
* DL & TT là cơ bản trong HTTT dùng để tìm kiếm. Đây là 2 đối tượng chính của tìm kiếm.

# **Với văn bản thì *tf* là gì?**

* Cho DOC xuất hiện từ khóa TERM: *tf* là số lần xuất hiện của TERM

# **Lý do sử dụng IDF**

* DF: tần suất tư liệu của TERM, tức là số document chứa TERM
* IDF: ngược lại của DF
* IDF = log10(N / DF)

# **Trình bày bài toán đối với tư liệu (document), sưu tập (collection), thuật ngữ (term)**

* Sưu tập (collection – C): tập hợp các document {D}
* Tài liệu (document – D): tập hợp các term {T}
* Bài toán:
  + Tìm trong C, Ti nào xuất hiện nhiều nhất
  + Tìm Di có Ti xuất hiện

# **Vai trò của chỉ mục (index) trong HTTKTT**

* Chỉ mục (index): chỉ mục là con trỏ trỏ tới địa chỉ của bản ghi dữ liệu, dùng bảng index để chỉ vị trí của bản ghi
* Lợi ích: tìm kiếm nhanh
* Tìm theo index (ví dụ):

|  |  |
| --- | --- |
| Max của khóa | Rãnh |
| 20 | #2 |
| 40 | #4 |

Tuổi là khóa

Toto, 16t, Hà Nội

Titi, 30t, Bắc Giang

Bourn, 21, Vientian

Ấm, 25t, Sài Gòn

# **Cấu trúc dữ liệu (CTDL) index (cây index)**

* Index là gì: xem câu 6
* Bao gồm: cây k-d, B, B+, tứ phân, R, TV

# **Định nghĩa cây 2-d**

* Là cây có các nút ứng với tọa độ (x, y) hai chiều. Gốc là nút 1, tại mức lẻ có hai cây con: cây trái index cho các nút < x của gốc, cây phải index cho các nút > x của gốc. Nút mức chẵn tách hai cây con theo y: sang trái khi k < y, sang phải khi k > y
* Ví dụ:



# **Định nghĩa cây k-d**

* Tương tự cây 2-d, k tương ứng với k chiều

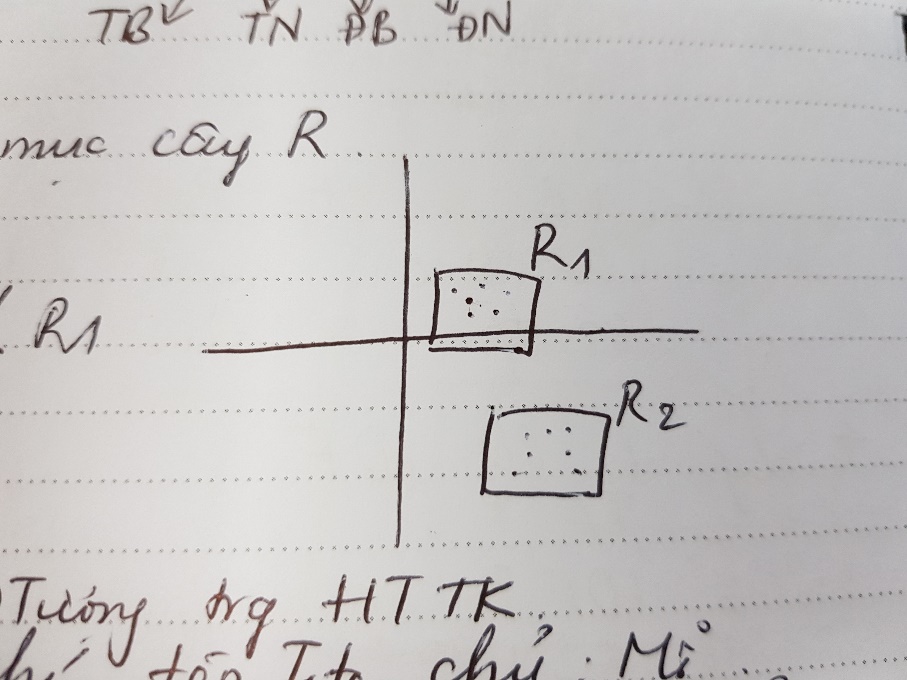
# **CTDL cây tứ phân**

* Dùng cho bản đồ. Chia thành 4 phần
* Nút của cây có dạng như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dữ liệu (x, y) | | | | | |
|  | |  |  |  | |
| TB | | TN | | ĐB | | ĐN | |

# **Cấu trúc chỉ mục cây R**

* Sử dụng hình chữ nhật
* Ví dụ:



# **Đặc trưng của đối tượng trong HTTK**

* Meta-data: con chó, tên Toto, chủ Mỗ
* Ghi chú: chú thích kèm theo (vd: mô tả con vật)
* Đặc trưng về dữ liệu mức thấp: dữ liệu ảnh, nét, đen trắng
* Đặc trưng thấp: tìm theo nội dung của đối tượng

# **Tìm kiếm theo logic là gì?**

* Biểu thức logic: khẳng định A, B, C nối nhau = các phép logic ()
* Tìm kiếm: truy xuất thông tin từ tập dữ liệu đã có để lấy ra TERM (chém)
* Tìm theo tiêu chí thỏa mãn biểu thức logic: YES or NO

# **Vai trò của không gian vector trong HTTK**

* Vector X = (x1, x2, x3, …, xn) Rn
* Mỗi xi tương ứng với đặc trưng i. So sánh 2 vector X, Y để kiểm tra X, Y có giống nhau không
* Đây là công cụ thể hiện đặc trưng của đối tượng

# **Tệp là gì?**

* Là đơn vị ghi thông tin có tên để phân biệt
* Hệ quản trị tệp: là một thành phần của HĐH, có nhiệm vụ tổ chức thông tin trên bộ nhớ ngoài và đảm bảo cho các chương trình hoạt động có thể truy cập tới các tệp đồng thời
* Truy cập tệp bằng ngôn ngữ lập trình. Ngôn ngữ lập trình cao cấp đầu tiên: FORTRAN (1958)

# **Các cách tổ chức tệp**

* Tổ chức tệp là cách sắp xếp các bản ghi
* Theo IBM có 6 cách
* Thông thường có 3 cách:
  + Tuần tự
  + Index
  + Hàm địa chỉ

# **Tệp tuần tự & tìm kiếm trên tệp tuần tự**

* Cách sắp xếp qua i – 1 bản ghỉ trước khi đến bản ghi i
* Tìm như thế nào: có 2 cách vét cạn hoặc chia đôi
* Thường dùng trong băng cassette

# **Tệp có chỉ mục & tìm kiếm**

* Chỉ mục: xem câu 6
* Tìm kiếm:

Index table

Vùng nhớ

# **Tổ chức theo hàm địa chỉ & tìm kiếm**

* Hàm địa chỉ: f(k) = địa chỉ
* Dùng hàm đ/c để tính đ/c của bản ghi, yêu cầu hàm f(k) rải đều, không được trùng
* Nếu trùng địa chỉ: 4 cách

# **Đặc trưng cho đối tượng ĐPT**

* ĐPT: văn bản, hình ảnh, âm thanh, hình động
* Đặc trưng: xem câu 12

# **Tìm kiếm trên website có đặc điểm gì?**

* Nhiều thông tin
* Thông tin không tin cậy
* Cần được kiểm chứng lại

# **Quy trình tìm kiếm tương tác**

* Tìm kiếm: truy xuất thông tin từ tập dữ liệu đã có để lấy ra TERM (chém)
* Question – Answer: Q1 → A1 → Q2 → A2 → Q3 → A3 → …
* Điều chỉnh câu hỏi Q

# **Độ chính xác của HTTK**

* Định nghĩa:

# **Độ nhạy**

* Định nghĩa:
* Hình vẽ tương tự P

# **Tìm kiếm theo siêu dữ liệu (meta-data)**

* Siêu dữ liệu: xem câu 12
* Tìm kiếm theo mô tả đối tượng đầu tiên

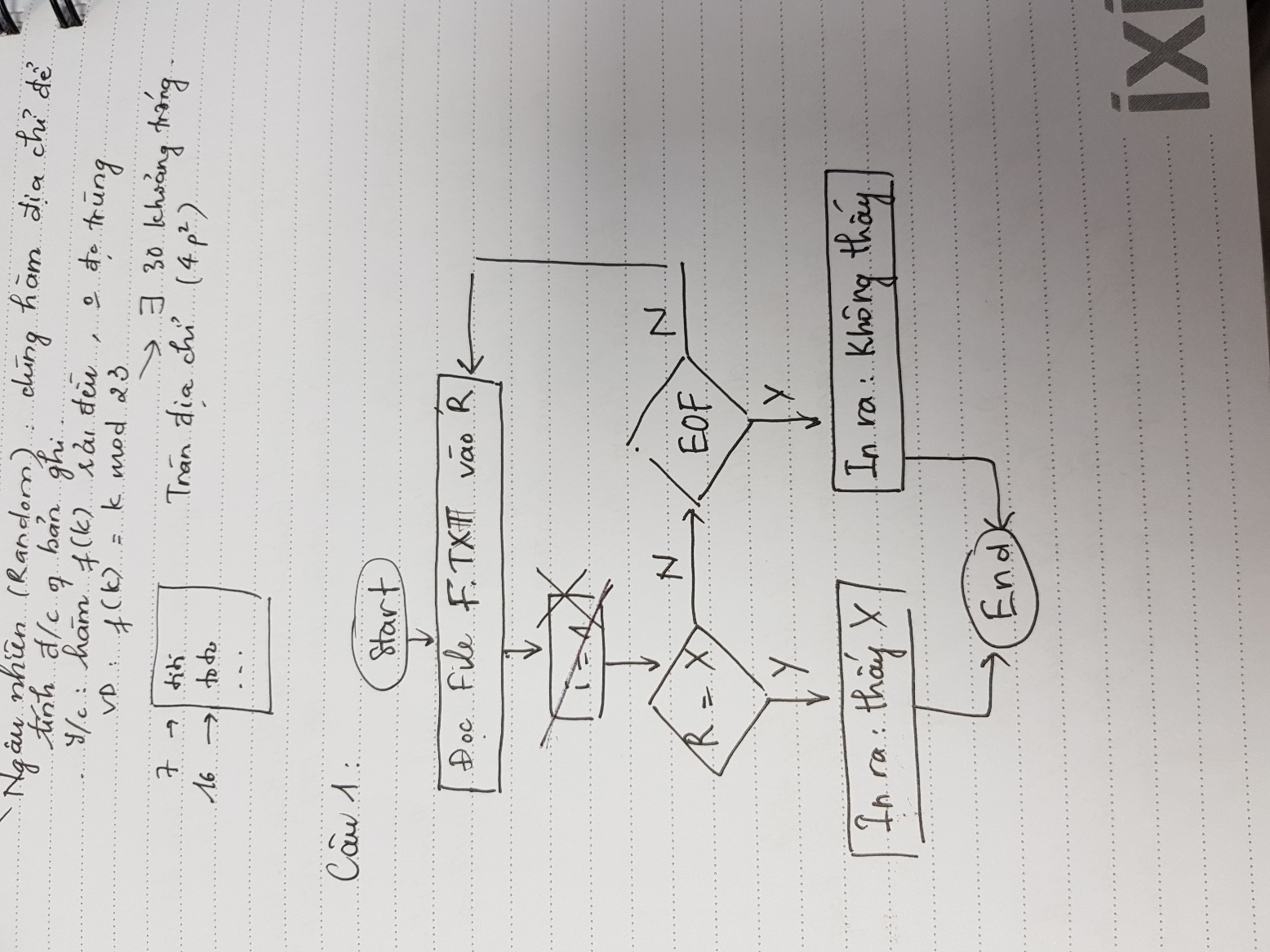
# **Tìm kiếm theo nội dung**

* Nội dung: xem câu 12
* Đối tượng tìm kiếm: bản chất (chụp ảnh để tìm kiếm)

# **Cho A(5,5) B(6,6) C(7,7). Vẽ cây 2-d**

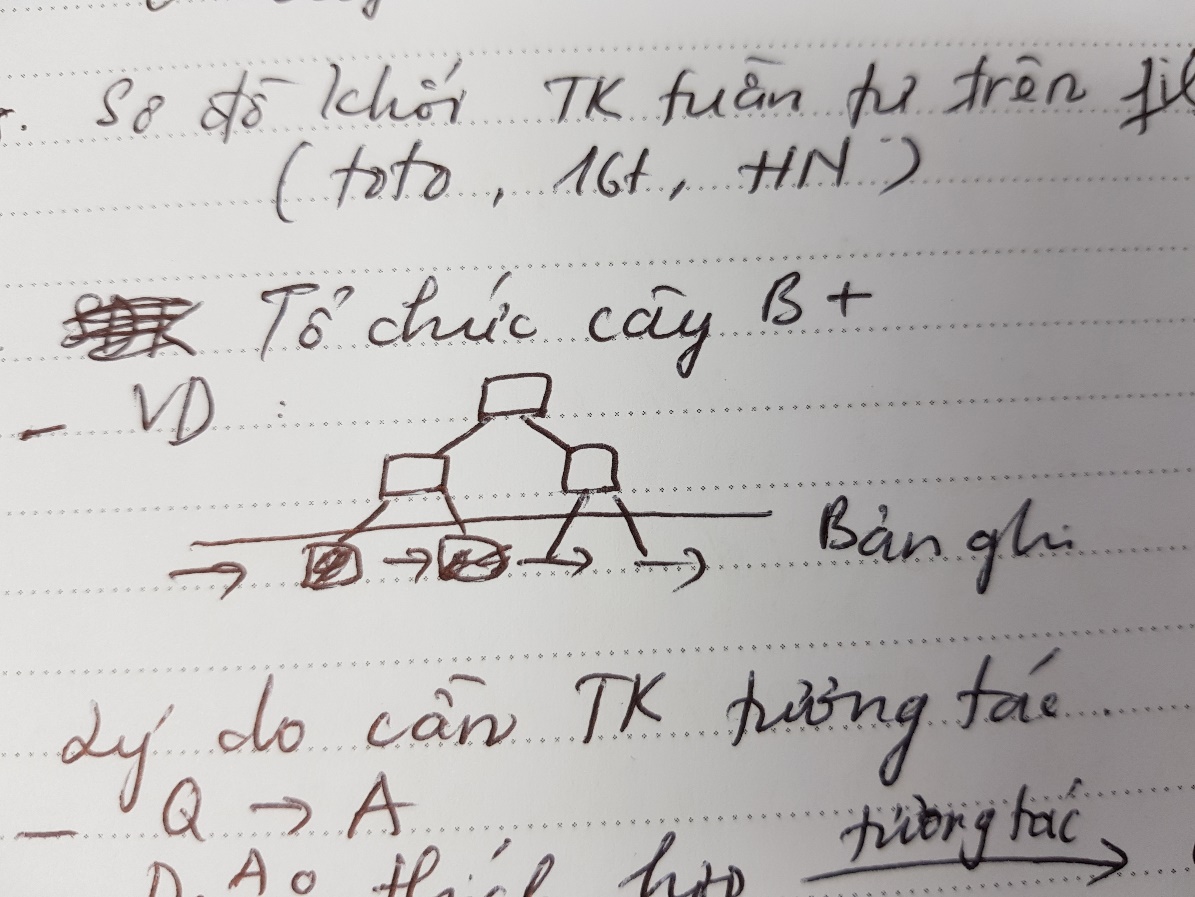


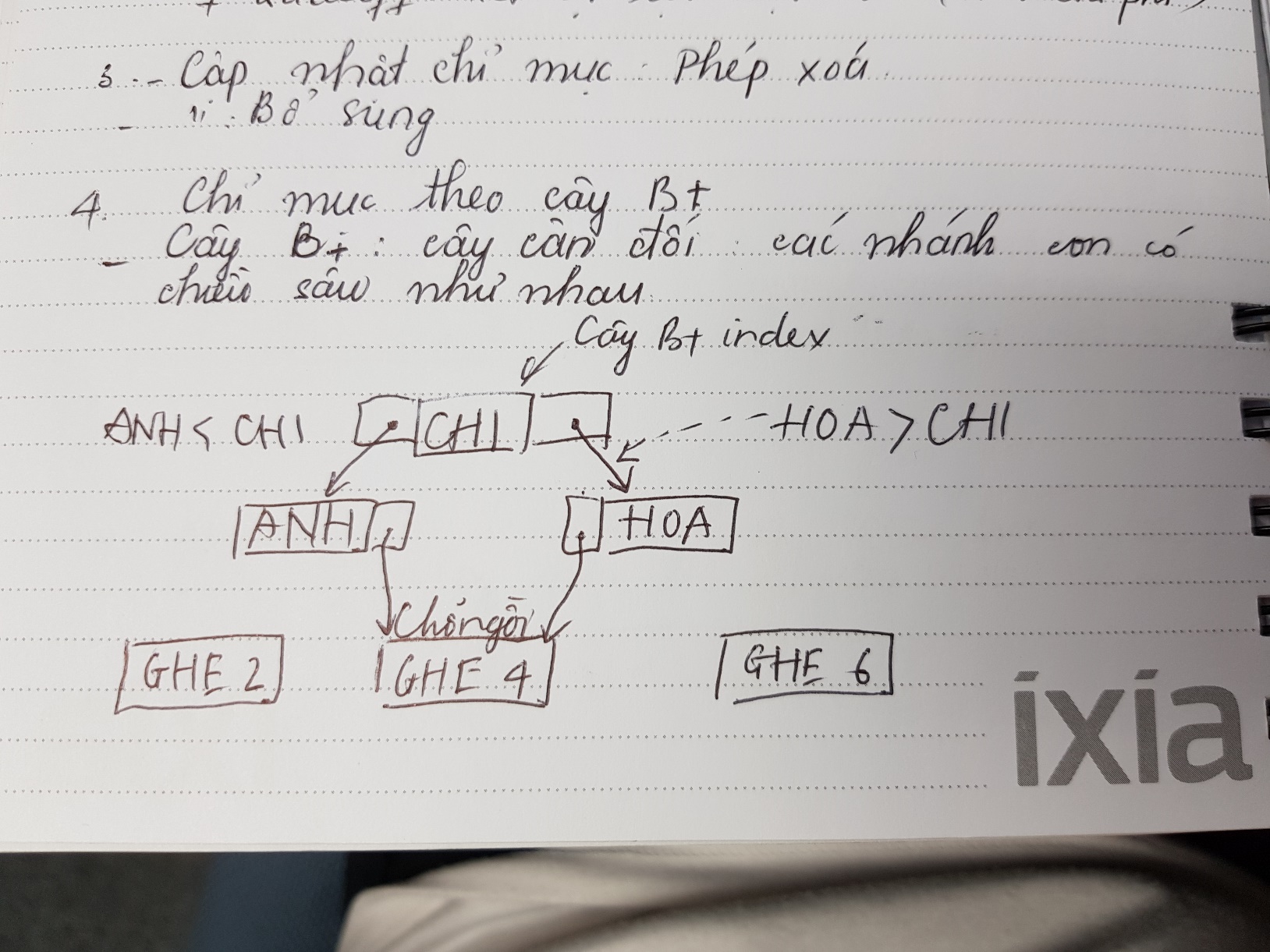
# **Sơ đồ khối tìm kiếm tuần tự trên file F**



# **Tổ chức cây B+**

* Là cây cân đối (cân bằng, BALANCE): các nhánh con có chiều sâu như nhau



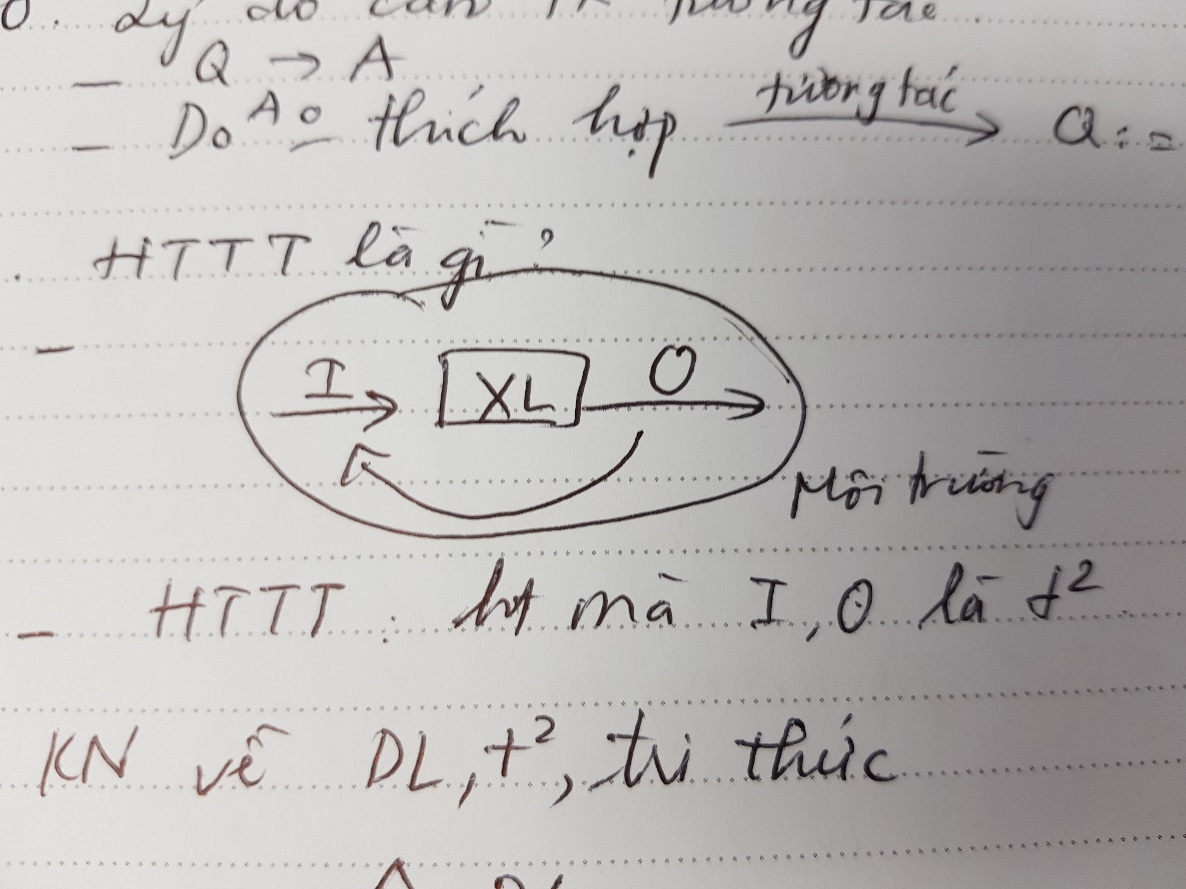


# **Lý do cần tìm kiếm tương tác**

* Q → A
* Do A không thích hợp

# **HTTT là gì?**

* Là hệ thống mà I, O là thông tin



# **Khái niệm về dữ liệu, thông tin, tri thức**

* Dữ liệu: là thông tin đc mã hóa trong máy tính, bao gồm số, chữ, hình ảnh, video,… phục vụ người dùng
* Thông tin: là dữ liệu đã được tinh chế, xử lý để làm rõ
* Tri thức: là hiểu biết về một lĩnh vực, có được qua trải nghiệm hoặc giáo dục

Tri thức

Thông tin

Dữ liệu

# **Người ta sử dụng tri thức trong HTTK như thế nào?**

* Tri thức: là hiểu biết về một lĩnh vực, có được qua trải nghiệm hoặc giáo dục
* Tri thức để đặt câu hỏi cho tốt
* Vd: Võ Nguyên Giáp ↔ Điện Biên Phủ

# **Sơ đồ khối tìm kiếm nút A trên cây nhị phân**



\*\*

# **Tìm kiếm trong CSDL**

* CSDL: tập hợp thông tin có cấu trúc, = chứa dữ liệu
* Đại số quan hệ: SQL
* Phép toán quan hệ trên miền
* Phép toán quan hệ trên bộ

# **Tìm kiếm trong văn bản thực hiện các công việc gì?**

* Đưa ra TERM (chọn TERM)
* Index theo TERM
* Câu hỏi Q theo TERM
* Tìm trên chỉ mục

# **Vai trò của xếp thứ tự trong tìm kiếm**

* Xếp thứ tự: tăng dần or giảm dần theo tiêu chuẩn nào đó
* Tìm kiếm nhanh hơn (Phương pháp chia đôi)

# **SQL là gì?**

* Là ngôn ngữ hỏi trong CSDL
* VD: SELECT …… FROM …… WHERE (biểu thức logic);