**Đề cương môn Tìm kiếm và truy xuất thông tin**

1. **Lý do cần tìm kiếm trong HTTT**

* *Cần lưu trữ*
* *Tìm để biết có hay không*
* *Lọc dữ liệu là gì?* - Lọc dữ liệu là chọn và hiển thị dữ liệu thoả mãn yêu cầu đặt ra và ẩn những dữ liệu không thoả mãn yêu cầu
* Hệ thống thông tin là hệ thống mà I/O là thông tin
* Một hệ thống thông tin được thiết kế và tổ chức tốt có thể cung cấp các nguồn thông tin đầy đủ, tiện lợi với chi phí thấp. Năng lực lưu trữ và truy cập thông tin của hệ thống cho phép xử lý và lưu trữ một số lượng lớn các thông tin.
* Tìm kiếm là gì: Tìm kiếm 🡪 truy hồi –RETRIEVAL(us), SEARCH(uk)

+” Tìm X sau khi lưu X”

+ Khi tìm 🡪 Không thấy 🡪 thông báo

🡪Thấy 🡪 bổ xung vào csdl

Khi thấy 🡪 chỉ ra 1 trường hợp

🡪 chỉ ra toàn bộ trường hợp

1. **Vai trò của dữ liệu, thông tin trong hệ thống tìm kiếm thông tin**

* *Dữ liệu là gì?*: là thông tin đã được mã hoá trong máy tính
* *Thông tin là gì?*: được hiểu như là sự thông báo, trao đổi, giải thích về một đối tượng nào đó và thường được thể hiện dưới dạng các tín hiệu như: chữ số, chữ viết, âm thanh, …

🡺 ***Vai trò:*** DL và TT đều là yếu tố cơ bản trong hệ thống thông tin nên phải tìm kiếm, nó là đối tượng của quá trình tìm kiếm thông tin

1. **Đối với văn bản thì TF là gì?**

* TF tiếng anh là gì: Term Frequency
* Tần suất là gì? Là số lần xuất hiện
* Cho Document, Xuất hiện từ khóa Term
* Đ/n: TF (Term Frequency) là tần suất xuất hiện của từ khóa (Term) trong dữ liệu (Document) cho trước.

1. **Lý do dùng IDF**

* TF (Term Frequency): Tần suất xuất hiện của một từ trong văn bản
* IDF (Inverse Document Frequency) là chỉ số biểu hiện cho tần suất xuất hiện của từ khóa (term) trong tất cả các dữ liệu (documents). Từ khóa xuất hiện càng nhiều thì chỉ số này càng thấp, vì xuất hiện nhiều đồng nghĩa với việc độ quan trọng rất thấp.

🡺 Lý do dùng IDF: ta dùng IDF để đánh giá “từ hiếm” nhưng quan trọng

1. **Trình bày bài toán đối với các tư liệu, sưu tập và thuật ngữ?**

* Một Bộ sưu tập C (Collection) gồm nhiều tư liệu Di (Document). Mỗi tư liệu Di lại chứa các thuật ngữ Tj (Term). Bài toán là tìm kiếm trong C thuật ngữ Tj nào xuất hiện với tần xuất nhiều nhất.
* Nêu một ví dụ tìm kiếm Term = “Bác Hồ” trong một tập C.

1. **\* Vai trò của chỉ mục trong hệ thống tìm kiếm thông tin**

* *Chỉ mục là gì?*: Chỉ mục (Index) là bảng tra cứu đặc biệt mà Công cụ tìm kiếm có thể sử dụng để tăng nhanh thời gian và hiệu suất thu thập dữ liệu. Hiểu đơn giản, một chỉ mục là một con trỏ tới dữ liệu trong một bảng. Chỉ mục hoạt động tương tự như Mục lục của cuốn sách
* Lợi ích của chỉ mục: (?)
* Tìm theo chỉ mục: (?)
* Ví dụ về tìm kiếm theo chỉ mục (?)

1. **Cấu trúc dữ liệu chỉ mục**

* Chỉ mục là gì? (?)
* Cây là tập hợp các nút liên kết với nhau thành các cụm, cây là đồ thị liên thông
  + Cây k-d:
  + Cây B (Balance) là cây có các nhánh con có chiều sâu như nhau hay còn gọi là cây tương đối, có nhiều nhánh tạo thành khối ứng với lá của cây
  + Cây tứ phân (Bản đồ)
  + Cây chỉ mục R
  + Cây TV (cho văn bản)
* Cây B và B+ khác nhau ở điểm nào?

1. **Định ng hĩa cây 2-d**

* Cây 2-d là cây có các nút tương ứng với tọa độ (x,y), gốc là nút 1 (mức 1)
  + Mức lẻ có 2 cây con
    - Cây con trái: nếu khóa k < x
    - Cây cong phải: nếu khóa k ≥ x
  + Mức chẵn có 2 cây con
    - Cây con trái: nếu khóa k < y
    - Cây cong phải: nếu khóa k ≥ y

1. **Định nghĩa cây k-d (k mức khác nhau, k chiều)**

* Cây k-d dùng trong trường hợp nào?
* Trong [khoa học máy tính](https://en.wikipedia.org/wiki/Computer_science), một cây k-d (viết tắt là cây k-chiều) là một [không gian phân vùng](https://en.wikipedia.org/wiki/Space_partitioning)[cấu trúc dữ liệu](https://en.wikipedia.org/wiki/Data_structure) để tổ chức [điểm](https://en.wikipedia.org/wiki/Point_(geometry)) trong một [không gian](https://en.wikipedia.org/wiki/Euclidean_space) k chiều.
* K-d cây là cấu trúc dữ liệu hữu ích cho một số ứng dụng, chẳng hạn như các tìm kiếm liên quan đến khóa tìm kiếm đa chiều
* Ví dụ: [tìm kiếm phạm vi](https://en.wikipedia.org/wiki/Range_search) và [tìm kiếm lân cận gần nhất](https://en.wikipedia.org/wiki/Nearest_neighbor_search). Cây k-d là trường hợp đặc biệt của cây [phân vùng không gian nhị phân](https://en.wikipedia.org/wiki/Binary_space_partitioning)

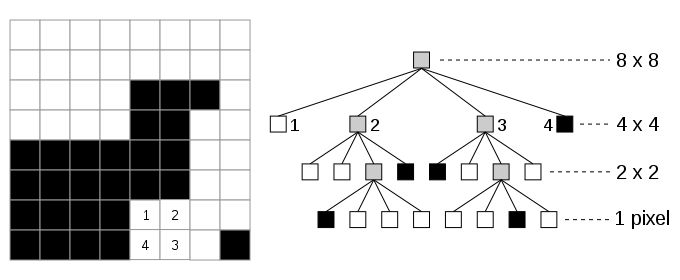
1. **Cấu trúc dữ liệu cây tứ phân (dùng cho bản đồ, xử lý ảnh)**

* Cây tứ phân là một [cấu trúc dữ liệu cây](https://en.wikipedia.org/wiki/Tree_data_structure) trong đó mỗi nút bên trong có bốn con. [Cây](https://en.wikipedia.org/wiki/Octree) tứ phân thường được sử dụng để phân vùng một không gian hai chiều bằng cách đệ quy phân chia nó thành bốn góc phần tư hoặc vùng. (Các vùng được chia nhỏ có thể là hình vuông hoặc hình chữ nhật hoặc có thể có hình dạng tùy ý)

Dữ liệu liên kết với một ô lá khác nhau tùy theo ứng dụng, nhưng ô lá đại diện cho một "đơn vị không gian thông tin ".

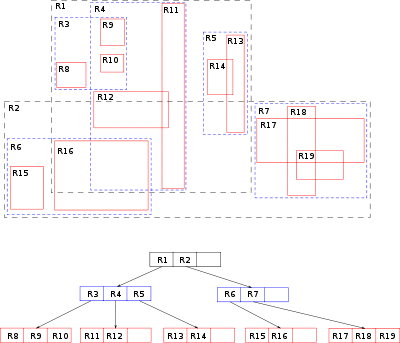
* Ví dụ minh họa:

Một cây-kim tự tháp (T-pyramid) là một cây "hoàn chỉnh"; mỗi nút của kim tự tháp T có bốn nút con ngoại trừ các nút lá; tất cả các lá trên cùng một mức, mức tương ứng với từng điểm ảnh trong hình ảnh.



1. **Cấu trúc dữ liệu cây chỉ mục R (Regtangle )**

* Ý tưởng chính của cấu trúc dữ liệu là nhóm các đối tượng lân cận và đại diện cho chúng với hình chữ nhật giới hạn tối thiểu của chúng ở cấp cao hơn tiếp theo của cây; "R" trong R-tree là hình chữ nhật. Vì tất cả các đối tượng nằm trong hình chữ nhật giới hạn này, một truy vấn không giao nhau với hình chữ nhật giới hạn cũng không thể giao nhau với bất kỳ đối tượng chứa nào. Ở cấp độ lá, mỗi hình chữ nhật mô tả một đối tượng duy nhất; ở cấp độ cao hơn, tập hợp của một số lượng ngày càng tăng của các đối tượng. Điều này cũng có thể được xem như là một xấp xỉ gần đúng thô của tập dữ liệu.
* Tương tự như cây B, cây R cũng là cây tìm kiếm cân bằng (vì vậy tất cả các nút lá có cùng chiều cao), tổ chức dữ liệu trong các trang và được thiết kế để lưu trữ trên đĩa (như được sử dụng trong cơ sở dữ liệu. Giống như hầu hết các cây, thuật toán tìm kiếm (ví dụ: giao lộ, ngăn chặn, tìm kiếm lân cận gần nhất) khá đơn giản. Ý tưởng chính là sử dụng các hộp giới hạn để quyết định có nên tìm kiếm trong một cây con hay không. Bằng cách này, hầu hết các nút trong cây không bao giờ được đọc trong khi tìm kiếm. Giống như cây B, điều này làm cho cây R phù hợp với các tập dữ liệu và cơ sở dữ liệu lớn, nơi các nút có thể được phân trang vào bộ nhớ khi cần thiết và toàn bộ cây không thể được giữ trong bộ nhớ chính.
* Ví dụ minh họa



1. **\* Đặc trưng của đối tượng trong hệ thống tìm kiếm**

* Meta: con chó, Tên: Toto, chủ: Mĩ
* Ghi chú: chú thích kèm theo mô tả con vật
* Đặc trưng về mức thấp vật lý, vd: dữ liệu ảnh, nét, đen trắng
* Mô tả đặc trưng thấp: Nội dung của đối tượng

1. **Tìm kiếm theo logic**

* Tìm kiếm là gì? là sử dụng những thông tin đã có để thực hiện truy suất trong tập dữ liệu lấy những thông tin cần thiết thỏa mãn câu hỏi tìm kiếm (Tiến chém)
* Định nghĩa biểu thức logic: là các khẳng định A, B. C nối với nhau bằng các phép toán logic (And, Or, Not, kéo theo, tương đương)
* Tìm kiếm logic: là tìm trong hệ thống các đối tượng có đặc trưng thỏa mãn biểu thức logic (yes/no) điều kiện logic

1. **Vai trò của không gian vector trong hệ thống tìm kiếm**

* Đ/n vector là một phần tử trong không gian vector
* ***Không gian vectơ*** là một tập hợp mà trên đó hai [phép toán](https://vi.wikipedia.org/wiki/Ph%C3%A9p_to%C3%A1n), [phép cộng vectơ](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A9p_c%E1%BB%99ng_vect%C6%A1&action=edit&redlink=1) và [phép nhân vectơ](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A9p_nh%C3%A2n_vect%C6%A1&action=edit&redlink=1) với một số, được [định nghĩa](https://vi.wikipedia.org/wiki/%C4%90%E1%BB%8Bnh_ngh%C4%A9a) và thỏa mãn các [tiên đề](https://vi.wikipedia.org/wiki/Ti%C3%AAn_%C4%91%E1%BB%81):
  + Phép cộng vectơ (có tính kết hợp, tính giao hoán, phân tử trung hòa phân tử đối)
  + Phép nhân vô hướng phân phối với phép cộng vectơ/phép cộng vô hướng
  + Phép nhân vô hướng tương thích với phép nhân trong trường các số vô hướng
  + …

Cho tập x= (x1, x2,……xn) thuộc Rn

Vector x có n chiều, trong đó mỗi xi ứng với 1 đặc trưng i

* So sánh X và Y để kiểm traxem X, Y có giống không
* Vai trò: là công cụ để biểu diễn các đặc trưng của đối tượng

1. **Tệp là gì?**

* Đ/n: Tệp là 1 đơn vị ghi thông tin, có tên phân biệt
* Hệ quản trị tệp là gì? là một thành phần của hệ điều hành, có nhiệm vụ tổ chức thông tin trên bộ nhớ ngoài, cung cấp các dịch vụ để người dùng có thể dễ dàng đọc ghi thông tin trên bộ nhớ ngoài và đảm bảo cho các chương trình hoạt động trong hệ thong có thể đồng thời truy cập tói các tệp.
* Truy cập tệp bằng ngôn ngữ lập trình

Ngôn ngữ lập trình là 1 tập con của ngôn ngữ máy tính, là một hệ thống được ký hiệu hóa để miêu tả những tính toán (qua máy tính) trong một dạng mà cả con người và máy đều có thể đọc và hiểu được

Ngôn ngữ lập trình đầu tiên trên thế giới là: (?)

Ngôn ngữ lập trình đầu tiên được đề cử là [Plankalkül](https://vi.wikipedia.org/w/index.php?title=Plankalk%C3%BCl&action=edit&redlink=1), do [Konrad Zuse](https://vi.wikipedia.org/w/index.php?title=Konrad_Zuse&action=edit&redlink=1) thiết kế vào những năm 194x, nhưng không được công chúng biết đến mãi cho đến năm 1972 (và không được thực hiện cho đến năm 1998). Ngôn ngữ lập trình đầu tiên được biết đến rộng rãi và thành công là [Fortran](https://vi.wikipedia.org/wiki/Fortran), được phát triển từ năm 1954 đến năm 1957 bởi một nhóm nhà nghiên cứu [IBM](https://vi.wikipedia.org/wiki/IBM) được dẫn dắt bởi [John Backus](https://vi.wikipedia.org/w/index.php?title=John_Backus&action=edit&redlink=1).

1. **Các cách tổ chức tệp**

* Định nghĩa: Tổ chức tệp là cách sắp xếp các bản ghi trong tệp

Bản ghi là đơn vị dữ liệu ứng với 1 đối tượng

* Có 3 cách chính để tổ chức tệp (IBM liệt kê 6 cách): tuần tự, index và hàm địa chỉ
  + Dãy tuần tự các byte không cấu trúc: HĐH không biết nội dung của tập tin: MS-DOS và UNIX sử dụng loại này. Dãy các record có chiều dài cố định
  + Index (Cấu trúc cây): Gôm cây của những record, không cần thiết có dùng độ dài, mỗi record có một trường khóa giúp cho việc tìm kiêm nhanh hơn.
  + Kiểu tập tin: Nếu hđh nhận biết được loại tập tin, nó có thể thao tác một cách hợp lý trên tập tin đo. Các hđh hỗ trợ nhiều loại tập tin khác nhau bao gồm các kiểu như: tập tin thường, thư mục, tập tin có ký tự đặc biệt. tập tin khối.

1. **Tệp tuần tự và tìm kiếm trong tệp tuần tự**

|  |  |
| --- | --- |
| * Định nghĩa: tổ chức tệp tuần tự là cách sắp xếp tệp sao cho muốn đến bản ghi thứ i thì ta phải đi qua (i-1) bản ghi trước đó * Tìm kiếm trên tệp tuần tự:   + Là phương pháp tìm kiếm một phần tử cho trước trong một danh sách bằng cách duyệt lần lượt từng phần tử của danh sách đó cho đến lúc tìm thấy giá trị mong muốn hay đã duyệt qua toàn bộ danh sách.   + Tìm kiếm vét cạn: Duyệt hết toàn bộ d/s đến khi tìm thấy giá trị cần tìm   + Tìm kiếm chia đôi: Với tập tuần tự đã sắp xếp 🡪 Chia đôi tập d/s tìm kiếm và so sánh với bản ghi ở giữa để thu gọn tập danh sách tìm kiếm. Lặp lại đến khi tìm thấy giá trị |  |

1. **Tệp có chỉ mục và tìm kiếm trên tệp có chỉ mục**

* Định nghĩa:
  + Chỉ mục (index) là một con trỏ tới một bản ghi dữ liệu.
  + Tổ chức tệp theo chỉ mục là cách tổ chức dùng bảng index để truy cập đến địa chỉ của bản ghi.
* Tìm kiếm trên tệp có chỉ mục (?)
* Ví dụ (?)

1. **Tổ chức theo hàm địa chỉ và tìm kiếm trên hàm địa chỉ**

|  |  |
| --- | --- |
| * Định nghĩa: f(k) = địa chỉ (?)   + Hàm địa chỉ f(k) là hàm được sử dụng để tính ra địa chỉ của bản ghi k.   + Yêu cầu đối với hàm địa chỉ f(k): phân bố đều và ít trùng   + Nếu trùng địa chỉ 🡪 có 4 phương pháp xử lý: * Tìm kiếm trên hàm địa chỉ. * Ví dụ: Tìm kiếm bản ghi X có hàm địa chỉ tương tứng y = f(x) |  |

Câu hỏi (t k hiểu ý nghĩa câu hỏi này như nào nữa :D): Nếu hàm f “kém” mà bài toán đặt ra cần xếp 2 người vào 1 ngôi nhà thì phải xử lý thế nào?

1. **Mô tả đặc trưng cho đối tượng đa phương tiện**

* Đa phương tiện: là các mô phỏng và các ứng dụng trên các dữ liệu đa phương tiện
* Có 4 loại dữ liệu đa phương tiện: text, image, sound, animation
* Đặc trưng của đối tượng:
  + ***Metadata :*** là thông tin mô tả nội dung của cơ sở dữ liệu (CSDL). Metadata báo cho người dùng biết thời gian cập nhật sau cùng của dữ liệu, định dạng và mục đích sử dụng của nó. Những thông tin này có thể hướng dẫn người dùng duyệt qua CSDL và giúp họ hiểu được ý nghĩa và ngữ cảnh của dữ liệu tài chính, các bản ghi về khách hàng và các giao dịch kinh doanh
  + Ghi chú, mô tả
  + Đặc trưng dữ liệu mức thấp (dữ liệu ảnh)
  + Đặc trưng thấp (nội dung của đối tượng)

1. **Tìm kiếm trên website có đặc điểm**

* Số lượng thông tin lớn
* Các thông tin không tin cậy
* Các kết quả cần kiểm chứng lại

1. **Qui trình tìm kiếm tương tác**

* Tìm kiếm là gì?
  + Tìm kiếm là xác định THẤY hay KHÔNG THẤY một bản ghi trong một tập dữ liệu tìm kiếm.
  + Khi thấy 🡪 Chỉ ra 1 TH hay toàn bộ TH
* Quy trình tìm kiếm:
  + Là quá trình đặt câu hỏi – trả lời Question – Answer và lặp đi lặp lại đến khi có nhận được kết quả thỏa mãn.
  + Mỗi vòng lặp là một lần điều chỉnh câu hỏi Q
  + Q1 🡪 A1 🡪 Q2 🡪 A2 🡪 Q3 🡪 A3

1. **Độ chính xác của hệ thống tìm kiếm**

* Hiệu quả của một hệ thống IR thể hiện khả năng đáp ứng của kết quả truy xuất đối với một yêu cầu thông tin. Các độ đo hiệu quả truy xuất phổ biến là độ chính xác (precision) và độ bao phủ (recall). Cả hai độ đo đều dựa trên các đánh giá của người dùng theo quá trình truy xuất.
* Độ chính xác (*P-* Precision) là tỷ lệ tài liệu được thu thập có liên quan, nó cho biết khả năng thu thập những tài liệu được xếp hạng trên cùng mà phần lớn là có liên quan.

*P* = A/B (trong đó, A là số lượng tài liệu có liên quan được thu thập, B là số lượng tài liệu được thu thập).

1. **Xác định độ nhạy (phủ)**

* Định nghĩa độ R (Recall) là tỷ lệ tài liệu có liên quan được thu thập, nó cho biết khả  
  năng tìm kiếm tất cả các tài liệu có liên quan trong tập hợp.

*R* = A/C (rong đó, A là số lượng tài liệu có liên quan được thu thập, C là số lượng tài liệu thực sự có liên quan)

1. **Tìm kiếm theo siêu dữ liệu (meta – data)**

* Trình bày câu 12
* Tìm kiếm theo mô tả đặc trưng đầu tiên
* [Siêu dữ liệu](http://it.die.vn/la-gi/sieu-du-lieu/) ([metadata](http://it.die.vn/la-gi/metadata/)) là dạng [dữ liệu](http://it.die.vn/la-gi/du-lieu/) miêu tả về dữ liệu/tài nguyên thông tin. Vì vậy, có thể hiểu metadata là dữ liệu về dữ liệu hoặc thông tin về thông tin. Cụ thể trong tài liệu thì siêu dữ liệu được xác định là “dữ liệu mô tả các thuộc tính của đối tượng thông tin và trao cho các thuộc tính này ý nghĩa, khung cảnh và tổ chức. Siêu dữ liệu còn có thể được định nghĩa là dữ liệu có cấu trúc về dữ liệu”.

Siêu dữ liệu tạo thuận lợi cho phát hiện các thông tin thích hợp và tìm kiếm và truy xuất tài nguyên. Được gắn thẻ với siêu dữ liệu, bất kỳ đối tượng số nào cũng có thể được tự động liên kết với các yếu tố thích hợp khác và vì thế dễ dàng tổ chức và phát hiện.

Với siêu dữ liệu bạn có thể:

* + Tìm kiếm tài nguyên theo tất cả các dạng tiêu chí;
  + Nhận diện các tài nguyên khác nhau;
  + Thu thập các tài nguyên theo chủ đề;
  + Lần vết các tài nguyên.

1. **Tìm kiếm theo nội dung**

* Trình bày theo câu 12
* Tìm kiếm theo mô tả đặc trưng 4 (chụp ảnh 🡪 tìm kiếm)

1. **Biểu diến cây 2-d Cho A(5,5), B(6,6), C(7,7)**

* Cây hai chiều là một cấu trúc dữ liệu động, rất giống cây nhị phân ngoại trừ điều là nó chia không gian hình học theo một cách thuận tiện cho việc tìm kiếm vùng.
* Phương pháp cây hai chiều mang ý tưởng là xây dựng cây tìm kiếm nhị phân với các điểm được chứa trong các nút, dùng các toạ độ x và y của các điểm như là các khoá theo một trình tự thay đổi nghiêm ngặt.
* Ở nút gốc, ta dùng tung độ y (nếu điểm được chèn có tung độ y nhỏ hơn điểm ở nút gốc, thì sang trái; nếu không sẽ sang phải), sau đó ở mức tiếp theo ta dùng hoành độ x, rồi ở mức tiếp theo nữa thì dùng tung độ y,… cứ luân phiên như vậy cho đến khi gặp nút ngoài cùng.

Xét gốc ở A(5,5) 🡪 Mức 1 lẻ xét theo x

Xét B(6,6) so sánh Bx với Ax có (6>5) 🡪 B nằm bên phải A 🡪 Mức 2 chẵn xét theo Y

Xét C(7,7) so sánh Cx với Ax có (7>5) 🡪 C nằm bên phải A

so sánh Cy với By có (7>6) 🡪 C nằm bên phải B

Biểu diễn:



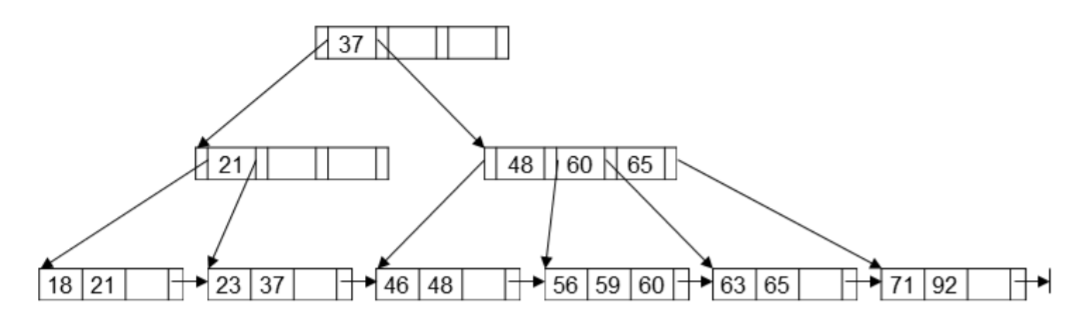
1. **Sơ đồ khối tìm kiếm tuần tự trên file**

* F (Toto, 16t, HN)

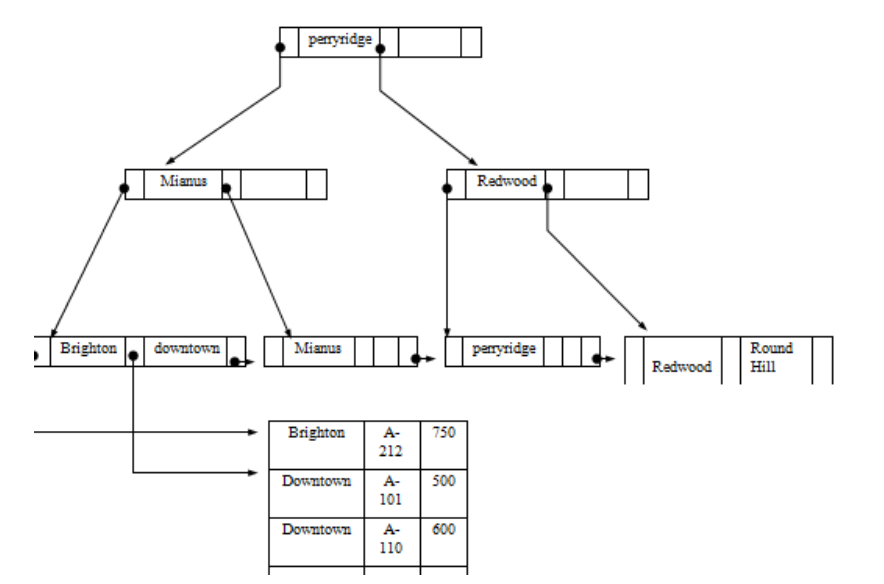


1. **Tổ chức cây B+**

* Là cây Balance có số nút, chiều sâu đến 3 tầng.
* Các record chỉ được lưu ở nút lá của cây. Khoảng giá trị mà mỗi nút lá chứa là không chồng chéo
* Tất cả các nút (ngoại trừ nút gốc) có tối thiểu là m/2 và tối đa là m cây(nhánh) con.
* Nút gốc được phép chứa ít nhất 2 con trỏ. Số con trỏ trong một nút được gọi là số nan (fanout) của nút.
* Các nút không là lá của một B+-cây tạo ra một chỉ mục nhiều mức trên các nút lá.
* Cấu trúc của các nút không là lá tương tự như cấu trúc nút lá ngoại trừ tất cả các con trỏ đều trỏ đến các nút của cây
* Ví dụ 1:



* Ví dụ 2:



1. **Lý do cần tìm kiếm tương tác**

* Q 🡪 A
* Do A không thích hợp 🡪 Q := Q1

1. **HTTT là gì?**

* HTTT là hệ thống mà I/O là thông tin
* I 🡪 xử lý 🡪 O

1. **KN về dữ liệu, thông tin, tri thức**

* Dữ liệu là thông tin đã được mã hoá trong máy tính
* Thông tin được hiểu như là sự thông báo, trao đổi, giải thích về một đối tượng nào đó và thường được thể hiện dưới dạng các tín hiệu như: chữ số, chữ viết, âm thanh, dòng điện…
* Tri thức hay kiến thức: bao gồm những dữ kiện, thông tin, sự mô tả, hay kỹ năng có được nhờ [trải nghiệm](https://vi.wikipedia.org/wiki/Tr%E1%BA%A3i_nghi%E1%BB%87m) hay thông qua [giáo dục](https://vi.wikipedia.org/wiki/Gi%C3%A1o_d%E1%BB%A5c).

1. **Sử dụng tri thức trong hệ thống tìm kiếm**

* ĐN tri thức là gì?
* Tri thức để đặt Q cho tốt Ø

Ví dụ : VNG <=> ĐBP

1. **Tìm kiếm nút A trên cây nhị phân**



1. **Tìm kiếm trong csdl**

* Csdl: là một tập hợp thông tin có cấu trúc (một tập hợp liên kết các [dữ liệu](http://it.die.vn/la-gi/du-lieu/) được lưu trên một thiết bị lưu trữ như đĩa hay băng. Dữ liệu này được duy trì dưới dạng một tập hợp các tập tin trong [hệ điều hành](http://it.die.vn/la-gi/he-dieu-hanh/) hay được lưu trữ trong các [hệ quản trị cơ sở dữ liệu](http://it.die.vn/la-gi/he-quan-tri-co-so-du-lieu/))
* Hệ thống CSDL: một hệ thống các thông tin có cấu trúc, được lưu trữ trên các thiết bị lưu trữ nhằm thõa mãn yêu cầu khai thác thông tin đồng thời của nhiều người sử dụng hay nhiều chương trình ứng dụng chạy cùng một lúc với những mục đích khác nhau
* Ngôn ngữ:
  + Đại số quan hệ
  + Phép toán quan hệ trên miền
  + Phép toán quan hệ trên bộ

1. **Tìm kiếm trong văn bản thực hiện công việc gì**

* Đưa ra TERM (chọn TERM)
* Index theo TERM
* Câu hỏi Q theo TERM
* Tìm kiếm trên chỉ mục(2)

1. **Vai trò của xếp thứ tự trong tìm kiếm**

* Xếp thứ tự? tăng dần hoặc giảm dần theo tiêu chuẩn
* Tìm kiếm nhanh hơn (phương pháp chia đôi)

2. **SQL là gì?**

* Ngôn ngữ truy vấn trong csdl (Tập các danh từ, động từ cho phép câu tìm kiếm/thêm mới và sửa đổi dữ liệu)

Vd: SELECT \* FROM \_ WHERE (Biểu thức logic);