UBND TỈNH QUẢNG NAM
TRƯỜNG ĐẠI HỌC QUẢNG NAM
KHOA: CÔNG NGHỆ THÔNG TIN

.....

KHÓA LUẬN TỐT NGHIỆP

Tên đề tài:

TÌM HIỂU TỐI ƯU HÓA CÂU TRUY VẤN TRONG CƠ SỞ DỮ LIỆU QUAN HỆ

Sinh viên thực hiện
NGUYỄN THỊ KIM THU
MSSV: 4111011031
CHUYÊN NGÀNH: CÔNG NGHỆ THÔNG TIN
KHÓA: 2011 – 2014

Cán bộ hướng dẫn :
Th.S. NGUYỄN THỊ MINH CHÂU
MSCS: 1016

Tam Kỳ, tháng 04 năm 2014

LỜI CẢM ƠN

Lời đầu tiên tôi xin gửi lời cảm ơn chân thành nhất đến Cô Giáo: Nguyễn Thị Minh Châu, đã định hướng và nhiệt tình hướng dẫn, giúp đỡ em rất nhiều về mặt chuyên môn trong quá trình làm khóa luận tốt.

Em xin gửi lời biết ơn sâu sắc đến các thầy, các cô giáo đã dạy dỗ và truyền đạt những kinh nghiệm quý báo cho em trong suốt ba năm học tại trường Đại học Quảng Nam.

Bên cạnh đó, để hoàn thành báo cáo khóa luận này, em cũng đã nhận được rất nhiều sự giúp đỡ, những lời động viên của gia đình, bạn bè, em xin cảm ơn.

Tuy nhiên, do thời gian có hạn, mặc dù đã nỗ lực hết sức mình, nhưng bài báo cáo khóa luận khó tránh khỏi thiếu sót. Em rất mong nhận được sự thông cảm và chỉ bảo tận tình của quý Thầy cô và các bạn.

Em xin gửi lời chúc đến các thầy cô trong trường, khoa lời chúc sức khỏe và luôn thành công trong công việc cũng như trong cuộc sống.

Sinh viên thực hiện

Nguyên Thị Kim Thu

MỤC LỤC

Phần 1. MỞ ĐẦU

1.1. Lý do chọn đề tài.

Ngày nay không ai còn nghi ngờ vai trò của công nghệ thông tin trong đời sống, trong khoa học kỹ thuật, kinh doanh, cũng như mọi mặt vận động của xã hội, dưới mọi quy mô, từ xí nghiệp công ty cho đến cả quốc gia và quốc tế. Việc nắm bắt thông tin nhanh, nhiều, chính xác và kịp thời ngày càng đóng vai trò quan trọng trong quản lý, điều hành. Nói cách khác, quản lý thực chất là quản lý thông tin. Song mọi thông tin cần quản lý trên máy tính cũng đều phải được thể hiện bằng các dữ liệu được lưu trong một dạng cơ sở dữ liệu nào đó. Ở nước ta trong những năm gần đây, thuật ngữ *Cơ sở dữ liệu* không còn mấy xa lạ với người làm tin học. Các ứng dụng tin học trong quản lý đang ngày một nhiều hơn. Do đó ngày càng có đông đảo người quan tâm hơn đến thiết kế, xây dựng các cơ sở dữ liệu. Người dùng máy tính có thể thấy có nhiều hệ quản trị cơ sở dữ liệu được các hãng phần mềm lớn trên Thế Giới phát triển thành thương phẩm. Để làm được điều này lập trình viên ngoài những hiểu biết một loạt các kiến thức về các mức mô hình hóa, các cách tiếp cận để thiết kế cơ sở dữ liệu, các ngôn ngữ mô tả, thao tác dữ liệu và biết cách lựa chọn sơ đồ thực thi hiệu quả nhất tức là tối ưu hóa câu truy vấn trong cơ sở dữ liệu quan hệ.

Tuy nhiên vấn đề đang đặt ra đối với các nhà lập trình là làm sao để quá trình xử lý truy vấn đạt được kết quả với việc sử dụng nguồn tài nguyên là thấp nhất và tổng thời gian thực hiện tất cả các quá trình thành phần của truy vấn là tối ưu nhất. Vì vậy, em chọn đề tài *Tìm hiểu tối ưu hóa câu truy vấn trong cơ sở dữ liệu quan hệ* để làm đề tài nghiên cứu trong khóa luận tốt nghiệp của mình.

1.2. Tính mới của đề tài.

Lĩnh vực Cơ sở dữ liệu là một lĩnh vực truyền thống, và là một trong những lĩnh vực chính trong ngành Công nghệ thông tin. Tuy nhiên, việc nghiên cứu vấn đề tối ưu hóa câu truy vấn lại là một vấn đề cấp thiết. Hiện nay, ở Việt Nam chưa có nhiều sách và tài liệu nói về vấn đề này. Đây là một lĩnh vực,

đòi hỏi nhiều kiến thức thuộc nhiều lĩnh vực khác nhau như kiến thức về CSDL, toán học, tối ưu hóa cũng như kiến thức về trí tuệ nhân tạo (xử lý ngôn ngữ), … Vấn đề tối ưu hóa câu truy vấn ngày càng được quan tâm nhiều hơn còn bởi một lý do, đó là dữ liệu chúng ta lưu trữ và xử lý ngày càng nhiều, độ phức tạp trong truy vấn ngày càng cao thì thời gian để xử lý câu truy vấn càng chậm. Vì vậy vấn đề tối ưu hóa câu truy vấn là hết sức cần thiết.

1.3. Mục tiêu của đề tài.

Tối ưu hóa câu truy vấn trong cơ sở dữ liệu quan hệ là tiến trình lựa chọn kế hoạch thực thi câu truy vấn một cách hiệu quả nhất, ít tốn thời gian nhất. Sau khi hoàn tất được đề tài này, em sẽ hiểu được cách tối ưu hóa câu truy vấn trong cơ sở dữ liệu quan hệ và bản chất của nó. Tìm hiểu được một số ví dụ đơn giản để giảm thời gian thực hiện câu truy vấn bằng các phép toán đại số quan hệ.

1.4. Đối tượng và phạm vi nghiên cứu.

Nghiên cứu bằng cách đọc tài liệu, tìm kiếm trên mạng, tìm hiểu các vấn đề liên quan đến truy vấn bằng biểu thức đại số quan hệ và tối ưu hóa câu truy vấn trong cơ sở dữ liệu quan hệ.

Đối tượng nghiên cứu chủ yếu là biểu diễn câu truy vấn bằng các phép toán của đại số quan hệ.

1.5. Phương pháp nghiên cứu.

- Đọc và nghiên cứu các tài liệu, sách báo liên quan đến đề tài.
- Tham khảo từ Internet.

CHƯƠNG 1: TỔNG QUAN VỀ CƠ SỞ DỮ LIỆU

1.1. Một số khái niệm cơ bản.

1.1.1. Định nghĩa về cơ sở dữ liệu.

Dữ liệu được lưu trữ trên các thiết bị lưu trữ theo một cấu trúc nào đó để có thể phục vụ cho nhiều người sử dụng với nhiều mục đích khác nhau gọi là cơ sở dữ liệu.

Cơ sở dữ liệu là tập các quan hệ (bảng) có cấu trúc nằm trong cùng một thể thống nhất.

Hiện nay, trong phần lớn các ứng dụng người ta thường sử dụng CSDL quan hệ. Một CSDL gồm một hoặc nhiều tập tin được thiết kế theo một cấu trúc nhất định và có quan hệ chặt chẽ với nhau.

1.1.2. Đặc điểm của thiết kế cơ sở dữ liệu.

- Giảm sự trùng lặp thông tin xuống mức thấp nhất và do đó bảo đảm được tính nhất quán và toàn vẹn dữ liệu.
- Đảm bảo dữ liệu có thể truy xuất theo nhiều cách khác nhau.
- Khả năng chia sẻ thông tin cho nhiều người sử dụng.

1.1.3. Những vấn đề mà cơ sở dữ liệu cần phải giải quyết.

- Tính chủ quyền của dữ liệu:

+ Tính chủ quyền của dữ liệu được thể hiện ở phương diện an toàn dữ liệu, khả năng biểu diễn các mối liên hệ ngữ nghĩa của dữ liệu và tính chính xác của dữ liệu. Điều này có nghĩa là người khai thác CSDL phải có nhiệm vụ cập nhật các thông tin mới nhất của CSDL.

- Tính bảo mật và quyền khai thác thông tin của người sử dụng:

+ Do có nhiều người được phép khai thác dữ liệu một cách đồng thời, nên cần thiết phải có một cơ chế bảo mật và phân quyền hạn khai thác CSDL. Các hệ điều hành nhiều người sử dụng hay hệ điều hành mạng cục bộ đều có cung cấp cơ chế này.

1.1.4. Các đối tượng sử dụng cơ sơ dữ liệu.

- Những người sử dụng CSDL không chuyên về lĩnh vực tin học và CSDL.

- Các chuyên viên CSDL biết khai thác CSDL, những người này có thể xây dựng các ứng dụng khác nhau, phục vụ cho các mục đích khác nhau trên CSDL.

- Những người quản trị CSDL, đó là những người hiểu biết về tin học, về các hệ quản trị CSDL và hệ thống máy tính. Họ là người tổ chức CSDL, do đó họ phải nắm rõ các vấn đề kỹ thuật về CSDL để có thể phục hồi CSDL khi có sự cố. Họ là những người cấp quyền hạn khai thác CSDL, do vậy họ có thể giải quyết được các vấn đề tranh chấp dữ liệu nếu có.

1.1.5. Định nghĩa về hệ quản trị cơ sở dữ liệu.

Để giải quyết tốt những vấn đề mà cách tổ chức CSDL đặt ra như đã nói ở trên, cần thiết phải có những phần mềm chuyên dùng để khai thác chúng. Những phần mềm này được gọi là các hệ quản trị CSDL. Các hệ quản trị CSDL có nhiệm vụ hỗ trợ cho các nhà phân tích thiết kế CSDL cũng như những người khai thác CSDL. Hiện nay trên thị trường phần mềm đã có những hệ quản trị CSDL hỗ trợ được nhiều tiện ích như: MS Access, Visual Foxpro, SQL Server Oracle, …

1.1.5.1. Các đặc tính cơ bản của hệ CSDL.

- Hệ CSDL phải đảm bảo các đặc tính sau:

+ Tính cấu trúc: Thông tin trong CSDL phải được lưu trữ theo cấu trúc nhất định.

+ Tính toàn vẹn: Các giá trị dữ liệu được lưu giữ trong CSDL phải thỏa mãn một số ràng buộc tùy thuộc vào hoạt động tổ chức mà CSDL phản ánh.

+ Tính nhất quán: Sau những thao tác cập nhật dữ liệu và ngay cả khi xảy ra sự cố trong quá trình cập nhật, dữ liệu trong CSDL phải đảm bảo đúng đắn.

+ Tính an toàn và bảo mật thông tin: CSDL cần được bảo vệ an toàn, phải ngăn chặn được những truy cập không cho phép và phải khôi phục CSDL khi có sự cố phần cứng hay phần mềm.

+ Tính độc lập: Vì một CSDL phải phục vụ cho nhiều mục đích khác nhau nên dữ liệu phải độc lập với các ứng dụng, không phụ thuộc vào một

bài toán cụ thể, đồng thời dữ liệu cũng phải độc lập với phương tiện lưu trữ và xử lý.

+ Tính không dư thừa: CSDL thường không lưu trữ những thông tin trùng lặp hoặc những thông tin có thể dễ dàng suy diễn hay tính toán từ những dữ liệu đã có.

1.1.5.2. Các chức năng của hệ quản trị CSDL.

Một hệ quản trị CSDL có các chức năng sau:

- Cung cấp cách tạo lập CSDL: Thông qua ngôn ngữ định nghĩa dữ liệu, người cung cấp khai báo kiểu và các cấu trúc thể lệ thông tin, khai báo các ràng buộc trên dữ liệu được lưu trữ trong CSDL.

- Cung cấp cách cập nhật dữ liệu, tìm kiếm, kết xuất thông tin: Ngôn ngữ để người dùng diễn tả cập nhật hay tìm kiếm, kết xuất thông tin gọi là thao tác dữ liệu. Thao tác dữ liệu bao gồm:

+ Cập nhật: Nhập, sửa, xóa dữ liệu.

+ Tìm kiếm và kết xuất dữ liệu.

+ Cung cấp công cụ kiểm soát, điều khiển việc truy cập vào CSDL.

1.1.6. Các ứng dụng của cơ sở dữ liệu.

Hiện nay, hầu như CSDL gắn liền với mọi ứng dụng của tin học, chẳng hạn như việc quản lý hệ thống thông tin trong các cơ quan nhà nước, việc lưu trữ và xử lý thông tin trong các doanh nghiệp, trong các lĩnh vực nghiên cứu khoa học, trong công tác giảng dạy, cũng như trong việc tổ chức thông tin đa phương tiện…

1.2. Mô hình dữ liệu.

Nền tảng của cấu trúc cơ sở dữ liệu là mô hình dữ liệu. Mô hình dữ liệu được định nghĩa là một sưu tập các công cụ khái niệm dùng cho việc mô tả dữ liệu, các mối quan hệ dữ liệu, các ngữ nghĩa dữ liệu và các ràng buộc dữ liệu.

1.2.1. Một số mô hình dữ liệu thông dụng.

- Mô hình thực thể liên kết: là mô hình cho phép mô tả các thực thể thông qua các thuộc tính và mối liên hệ giữa các thực thể. Một trong các cách biểu thị mô hình thực thể là dùng đồ thị, sơ đồ khối.

- Mô hình mạng là mô hình thực thể liên kết, trong đó có các mối liên kết bị hạn chế trong kiểu nhị phân (hai thực thể) và nhiều-một hoặc một-một và được biểu diễn bởi một đồ thị có hướng.

- Mô hình hướng đối tượng là mô hình cung cấp các đặt tính nhận dạng đối tượng. Trong đó, mỗi lớp đối tượng được đặc trưng bởi hai yếu tố:

+ Tập các thuộc tính (properties) để nhận dạng đối tượng.

+ Tập các phương thức (methods) để thao tác với đối tượng.

- Mô hình dữ liệu phân cấp (Hieracical Data Model), còn gọi là mô hình phân cấp (Hieracical Model), được thực hiện thông qua sự kết hợp giữa IBM và North American Rockwell vào khoảng năm 1965. Mô hình là một cây, trong đó mỗi nút của cây biểu diễn một thực thể, giữa nút con với nút cha được liên hệ với nhau theo một mối quan hệ xác định.

- Mô hình dữ liệu quan hệ là mô hình dựa vào ký hiệu là tập các tên và cơ sở toán học của nó là các phép toán tập hợp và ánh xạ. Nó là mô hình phổ biến hiện nay. Tập các phép toán trong mô hình này dựa trên hai hệ ký hiệu: ký hiệu đại số và ký hiệu logic. Trong đề tài này em tập trung vào mô hình dữ liệu quan hệ.

1.3. Mô hình dữ liệu quan hệ.

Mô hình dữ liệu quan hệ lần đầu tiên được đề nghị bởi Edgar F. Codd vào năm 1970. Hiện nay mô hình quan hệ là mô hình ưu thế đối với các ứng dụng xử lý dữ liệu thương mại.

1.3.1. Các khái niệm cơ bản.

1.3.1.1. Thuộc Tính(attribute):

- Thuộc tính (attribute) là các đặc điểm riêng của một đối tượng (đối tượng được hiểu như là một thực thể ở mô hình thực thể kết hợp), mỗi thuộc tính có một tên gọi và phải thuộc về một kiểu dữ liệu nhất định.

- Kiểu dữ liệu (data type) là các thuộc tính được phân biệt qua tên gọi và phải thuộc một kiểu dữ liệu nhất định (số, chuỗi, ngày tháng, logic, hình ảnh,...). Kiểu dữ liệu ở đây có thể là kiểu vô hướng hoặc là kiểu có cấu trúc. Nếu thuộc tính có kiểu dữ liệu là vô hướng thì nó được gọi là thuộc tính đơn

hay thuộc tính nguyên tố, nếu thuộc tính có kiểu dữ liệu có cấu trúc thì ta nói rằng nó không phải là thuộc tính nguyên tố.

- Miền giá trị (domain of values) là thông thường mỗi thuộc tính chỉ chọn lấy giá trị trong một tập con của kiểu dữ liệu và tập hợp con đó gọi là miền giá trị của thuộc tính đó.

Trong nhiều hệ quản trị cơ sở dữ liệu, người ta thường đưa thêm vào miền giá trị của các thuộc tính một giá trị đặc biệt gọi là giá trị rỗng (NULL). Tuỳ theo ngữ cảnh mà giá trị này có thể đặc trưng cho một giá trị không thể xác định được hoặc một giá trị chưa được xác định ở vào thời điểm nhập tin nhưng có thể được xác định vào một thời điểm khác.

1.3.1.2. Lược Đồ Quan Hệ (relation schema).

Lược đồ quan hệ (Relation schema) là sự trừu tượng hóa của quan hệ, một sự trừu tượng hóa ở mức cấu trúc của một bảng hai chiều. Khi nói đến lược đồ quan hệ tức là đề cập tới cấu trúc tổng quát của một quan hệ, khi nói đến một quan hệ thì hiểu rằng đó là một bảng có cấu trúc cụ thể trên một lược đồ quan hệ với các bộ giá trị của nó.

Ví dụ: SACH (maSach, tensach, sotrang, tacgia)

- Thể hiện của một lược đồ quan hệ là một quan hê :

Ví dụ: SACH (maSach, tensach, sotrang, tacgia)

maSach	tensach	sotrang	tacgia
s01	Toán	20	Trần Thu

1.2.4. Bộ (Tuple).

Mỗi bộ là những thông tin về một đối tượng thuộc một quan hệ, bộ cũng còn được gọi là mẫu tin.

Thường người ta dùng các chữ cái thường (như t,μ,...) để biểu diễn bộ trong quan hệ, chẳng hạn để nói t là một bộ của quan hệ r thì ta viết t $\in$ r.

1.2.5. Khóa của quan hệ.

- Siêu khóa: là một tập hợp một hay nhiều thuộc tính của quan hệ (table) có tính chất xác định duy nhất một bộ trong mỗi thể hiện của quan hệ (table)

- Khóa: Mỗi thuộc tính hoặc một tập các thuộc tính dùng để xác định một cách duy nhất mỗi thực thể trọng một tập thực thể gọi là khóa đối với tập thực thể đó. Về nguyên tắc, mỗi thực thể có một khóa, bởi vì mỗi thực thể đều có thể phân biệt được với thực thể khác. Nếu không chọn được một tập các thuộc tính có chứa một khóa cho một tập thực thể thì không có khả năng phân biệt được thực thể nay với thực thể kia trong tập thực thể đó. Trong trường hợp này thì các số đếm thường được gán làm thuộc tính khóa.

- Khóa ngoại: là một tập hợp gồm một hay nhiều thuộc tính là khóa của một lược đồ quan hệ (table) khác.

Ví dụ :

SACH (maSach, **tensach, tacgia, sotrang, maNXB): Quan hệ sách**

NXB (maNXB, **tenNXB,diachiNXB):Quan hệ nhà xuất bản**

maSach, maNXB: là khóa

maNXB: là khóa chính trong bản NXB nhưng lại là khóa ngoại trong bản sách.

CHƯƠNG 2: ĐẠI SỐ QUAN HỆ

2.1. Tổng quan về đại số quan hệ.

Đại số quan hệ (ĐSQH) có nền tảng toán học (cụ thể là lý thuyết tập hợp) để mô hình hóa CSDL quan hệ.

Đại số quan hệ được trình bày xem như một phương pháp để mô hình hoá các phép toán trên CSDL quan hệ. Đồng thời đây cũng là một trong những ưu điểm của mô hình dữ liệu quan hệ, đó là việc tiếp nhận các kết quả của công cụ toán học trong việc xây dựng ngôn ngữ khai thác, xử lý dữ liệu. Nhìn chung, các phép toán của đại số quan hệ là khá đơn giản, nhưng nó khá mạnh và là một đại số có tính đầy đủ, phi thủ tục. Tuy nhiên đây là một cơ sở cho việc thiết lập các ngôn ngữ con dữ liệu bậc cao hơn.

Các phép toán cơ bản:

Phép hợp(Union): ∪

Phép giao (Intersect): ∩

Phép trừ (Minus): - hay \

Phép chia (Division): hay /

Phép chọn(Selection): σ

Phép chiếu (Projection): π

Phép tích Đề-Các (Cartesian Product) x

Phép kết nối (Join):⋈

Phép kết nối tự nhiên *

2.2. Biểu thức đại số quan hệ.

- Biểu thức ĐSQH là một biểu thức gồm các phép toán ĐSQH.
- Biểu thức ĐSQH được xem như là một quan hệ (không có tên).
- Có thể đặt tên cho quan hệ được tạo từ một biểu thức ĐSQH.
- Có thể đổi tên các thuộc tính của quan hệ được tạo từ một biểu thức ĐSQH.

2.3. Các phép toán đại số quan hệ.

Đại số quan hệ là một trong những ngôn ngữ thao tác dữ liệu, bao gồm các phép toán trên các quan hệ của một cơ sở dữ liệu cho trước. Đó là phép toán

hợp, giao, chiếu và chọn…Tập hợp các phép toán quan hệ tạo nên một cơ chế truy nhập dữ liệu khá linh hoạt và mềm dẻo. Vì vậy người ta thường lấy đại số quan hệ làm đơn vị đo công suất của hệ quản trị cơ sở dữ liệu quan hệ.

2.3.1. Phép hợp (Union).

Định nghĩa: Hợp của hai quan hệ r và s, ký hiệu là r ∪ s là tập tất cả các bộ thuộc r hoặc s hoặc thuộc cả hai quan hệ.

Biểu diễn: u ∪ s = {t| t∈ r hoặc t ∈ s hoặc t ∈ r và t ∈ s)

Ví dụ:

r (A	B	C)
a1	b1	c1
a2	b1	c2
a2	b2	c1

s (A	B	C)
a1	b1	c1
a2	b2	c2

Ta có:

r∪ s(A	B	C)
a1	b1	c1
a2	b1	c2
a2	b2	c2
a2	b2	c1

Nhận xét:

Thực chất của phép hợp:

- Chỉ thực hiện trên các quan hệ khả hợp (có cùng tập thuộc tính).

- Phép hợp hai quan hệ khả hợp r và s thực chất là việc gộp các bản ghi trong hai quan hệ thành 1 quan hệ nhưng các bản ghi trùng nhau thì chỉ giữ lại một bản ghi.

2.3.2. Phép giao (Intersect).

Định nghĩa: Giao của hai quan hệ r và s, kí hiệu là r ∩ s là tập các bộ thuộc cả hai quan hệ r và s.

Biểu diễn: r ∩ s = {t | t ∈ r và t ∈ s }

Ví dụ:

r (A	B	C)
a1	b1	c1
a2	b1	c2
a2	b2	c1

s (A	B	C)
a1	b1	c1
a2	b2	c2

Ta có:

r∩s(A B C)

a1 b1 c1

Nhận xét :

Thực chất của phép giao là:

- Chỉ thực hiện trên hai quan hệ khả hợp.

- Phép giao giữa r và s thực chất là việc chọn ra trong hai quan hệ r và s những bản ghi trùng nhau.

2.3.3. Phép hiệu (Minus).

Định nghĩa: Hiệu của hai quan hệ r và s, ký hiệu r – s, là tập các bộ thuộc r nhưng không thuộc s.

Biểu diễn: $r - s = \{ t \mid t \in r \text{ và } t \notin s \}$.

Chú ý: Ta có thể biểu diễn phép giao của hai quan hệ qua phép trừ như sau: $r \cap s = r - (r - s)$.

Ví dụ:

r (A B C)	s (A B C)
a1 b1 c1	a1 b1 c1
a2 b1 c2	a2 b2 c2
a2 b2 c1	

Ta có:

r - s(A B C)

a2 b2 c1

a2 b1 c2

Nhận xét:

Thực chất của phép trừ là:

- Chỉ thực hiện trên các quan hệ khả hợp.

- Phép trừ r - s thực chất là việc chọn ra các bản ghi chỉ có ở r mà không có ở s.

2.3.4. Phép chia (Division).

Định nghĩa: Cho quan hệ r với lược đồ quan hệ R (A1, A2,….,An) và quan hệ s với lược đồ quan hệ S (A1, A2,….., Am) trong đó m < n, s # Ø. Khi đó phép chia r ÷ s là tập của tất cả (n – m) – bộ t sao cho với mọi v ∈ s thì t ghép với v thuộc r.

Biểu diễn: r ÷ s = { t | ∀ v ∈ s => (t,v) ∈ r}.

Ví dụ:

r (A	B	C	D)
a	b	c	d
a	b	e	f
b	c	e	f
e	d	c	d
a	b	d	e

s (C	D)
c	d
e	f

Ta có:

r ÷ s = (A	B)
a	b

2.3.5. Phép tích đề - các (Cartesian Product).

Định nghĩa: Gọi r là một quan hệ xác định trên tập thuộc tính {A1,…,An} và s là quan hệ xác định trên tập thuộc tính {B1,…..,Bn}. Tích Đề-các của r và s ký hiệu là r x s là tập (n+ m) bộ với n thành phần đầu có dạng một bộ thuộc r và m thành hàng đầu có dạng của bộ thuộc s.

Biểu diễn: r x s = { t | t có dạng (a1, a2,…,b1,b2,…,bm), trong đó (a1,…,an) ∈ r và (b1,…, bm) ∈ s}.

Ví dụ:

r (A	B	C)
a1	b1	c1
a2	b2	c2
a3	b3	c3

s (D	E	F)
d	e	f
d'	e'	f'

Ta có:

r x s (A	B	C	D	E	F)
a1	b1	c1	d	e	f
a1	b1	c1	d'	e'	f'
a2	b2	c2	d	e	f
a2	b2	c2	d'	e'	f'
a3	b3	c3	d	e	f
a3	b3	c3	d'	e'	f'

2.3.6. Phép chiếu (Projection).

Định nghĩa: Cho quan hệ r với lược đồ quan hệ R(U),U =(A1,…,An). X là tập con của các thuộc tính X ⊆ U.

Phép chiếu của quan hệ r trên tập thuộc tính X, kí hiệu $\pi_x(r)$ là tập các bộ của r xác định trên tập thuộc tính X.

Biểu diễn: $\pi_x(r) = \{t[X] \mid t \in r\}$.

Ví dụ: R ={ A, B, C, D } X={A, B} Y={A, C}.

R(A	B	C	D)
a1	b1	c1	d1
a1	b1	c1	d2
a2	b2	c2	d2
a2	b2	c3	d3

Ta có :

$\pi_X(r)$ = (A B)

a1	b1
a2	b2

$\pi_Y(r)$ =(A C)

a1	c1
a2	c2
a2	c3

2.3.7. Phép chọn (Union).

Định nghĩa: cho một quan hệ r với lược đồ quan hệ R(A1, A2,….,An). F là một biểu thức lô-gíc.

Biểu thức lô-gíc thường là các phép so sánh giữa một thuộc tính và các hằng. Các phép so sánh: <, <=, >=, =, #.

Trong biểu thức lô-gíc có thể dùng các phép toán lô-gíc: ∧ (và) , ∨ (hoặc), ¬ (phủ định) để nối các biểu thức với nhau.

Phép chọn trên r với biểu thức chọn F, ký hiệu: $\sigma_F(r)$, là tập tất cả các bộ của r thõa mãn F.

Biểu diễn : $\sigma_F(r) = \{t \mid t \in r \text{ và } F(t) \text{ đúng}\}$

Ví dụ:

R (A B C)
a1 b1 c1
a1 b2 c2
a1 b2 c1

Ta có:

chọn F = (B=b2) ∧ (C= c2)
$\delta_F(r)$ (A B C)
a1 b2 c2

chọn F = (B=b2) ∨ (C= c1)
$\delta_F(r)$ (A B C)
a1 b1 c1
a1 b2 c2
a1 b2 c1

2.3.8. Phép kết nối (Join).

2.3.8.1. Phép nối có điều kiện.

Định nghĩa: Cho quan hệ r với lược đồ quan hệ R(A1,A2,…,An), cho quan hệ s với lược đồ quan hệ S(B1, B2,…..,Bm).

Ghép các bộ: cho hai bộ u =(a1, a2, ….,an) và v=(v1, v2,….,vm) là một bộ (u, v)=(a1, a2,…., an, v1, v2, …,vm).

Cho biểu thức kết nối F giữa hai quan hệ r và s: Ghép các bộ của hai quan hệ thỏa mãn điều kiện θ : cú pháp của θ :< thuộc tính của r> <toán tử so sánh> < thuộc tính của s>.

Phép kết nối giữa r và s với biểu thức kết nối F, ký hiệu: r ⋈ s, được định nghĩa như sau: r ⋈ s = {t | t=(u,v) và u ∈ r và v ∈ s và F(t) = đúng}.

Ví dụ: kết nối với điều kiện B>=c.

r (A B C)
a1 1 1
a2 2 2
a1 2 2

s(C B E)
1 d1 e1
2 d2 e2
3 d3 e3

Ta có:

r⋈s=(A B C C D E)

a1	1	1	1	d1	e1
a2	2	1	1	d1	e1
a2	2	1	2	d2	e2
a1	2	2	1	d1	e1
a1	2	2	2	d2	e2

2.3.8.2. phép nối tự nhiên

Định nghĩa: Phép kết nối tự nhiên là phép kết nối với phép so sánh bằng các thuộc tính cùng tên ở 2 quan hệ và sau kết nối cắt đi một cột cùng tên bằng phép chiếu thì gọi là phép kết nối tự nhiên.

Ký hiệu phép kết nối tự nhiên: r*s

Ví dụ:

r (A B C)

a1	1	1
a2	2	2
a1	2	2

s(C B E)

1	d1	e1
2	d2	e2
3	d3	e3

Ta có:

r(ABC)*s(CDE)=(A B C D E)

a1	1	1	d1	e1
a2	2	1	d1	e1
a1	2	2	d2	e2

2.2.9. Hàm tính toán và gom nhóm.

Dùng để tính toán các giá trị mang tính chất tổng hợp trong đại số quan hệ. Trong đó:

Hàm tính toán: Đầu vào là một tập giá trị và trả về một giá trị đơn.

Avg(): giá trị trung bình.

Min(): giá trị nhỏ nhất.

Max(): giá trị lớn nhất.

Sum(): tính tổng.

Count(): đếm số mẫu tin.

Gom nhóm: công thức như sau:

G1,G2,..Gnℑ F1(A1),F2(A2)…FN(An)(E), với:

E là biểu thức đại số quan hệ.

G_i là tên thuộc tính gom nhóm (có thể không có).

Fi là Các hàm gom nhóm.

A_i là tên thuộc tính toán trong hàm gom nhóm

Ví dụ:

r (A B)

a1 1

a1 3

a2 1

a2 2

a1 2

a2 4

Ta có: $\Im_{SUM(B)}$ (r)

$_A\Im_{SUM(C)}$ (r)

2.4. Xây dựng chiến lược đại số quan hệ.

1. Phân rã thành các câu hỏi con.
2. Nhìn lược đồ cơ sở sữ liệu dưới cấu trúc đồ thị để thấy các phép kêt.
3. Xác định loại so sánh:
   - Giữa hai giá trị.
   - Giữa giá trị và quan hệ.
   - Giữa hai quan hệ.
4. xây dựng biểu thức đại số quan 4 bước:
   - xác định lược đồ kết quả.
   - Xác định con đường truy vấn (dùng chiến lược 2).
   - Xác định điều kiện chọn (dùng chiến lược 3).
   - Tính toán (thêm thuộc tính hoặc giá trị của các nhóm.
5. Phát biểu các mệnh đề chỉ dùng lượng từ tồn tại.

Trong nhiều trường hợp, khi mà biểu thức phức tạp, chúng ta sẽ dùng các biể thức trung gian. Ngoài ra, trong trường hợp tổng quát, chúng ta có thể phải cần đến các cấu trúc tuần tự, rẽ nhánh và lặp.

2.5. Ví dụ ứng dụng biểu diễn câu truy vấn trên CSDL:

2.5.1. Xây dựng cơ sở dữ liệu:

Cho cơ sở dữ liệu sau áp dụng cho tất cả các ví dụ của đại số quan hệ:

SACH (maSach, tensach, tacgia, sotrang, maNXB).

NXB (maNXB, tenNXB,diachiNXB).

DOCGIA(maGD, tenDG, diachi, ngaysinh, gioitinh).

MUON (maSach, maDG, ngaymuon, ngaytra).

2.5.2. Biểu diễn các phép toán trên ĐSQH:

2.5.2.1. Phép chọn.

Ví dụ 1: Cho biết các độc giả có địa chỉ ở điện bàn.

Quan hệ: DOCGIA.

Thuộc tính:diachi.

Điều kiện: diachi ='Điện Bàn'

Ta có:

$$\delta_{\text{diachi= 'điện bàn'}}(\text{DOCGIA}).$$

Ví dụ 2: Tìm các các cuốn sách tiếng việt có số trang bằng 20 hoặc sách toán có số trang dưới 20 tran.

Quan hệ:SACH.

Thuộc tính: tensach, Trang.

Điều kiện:

Trang=20 và tensach=tiếng việt hoặc

Trang<20 và tensach=toán.

Ta có:

$$\delta_{(\text{tensach= 'tiếng việt' AND trang=20tr)or (tensach= 'toán' AND trang<20}}(\text{SACH}).$$

Nhận Xét:

Phép chọn có tính giao hán.

Kếp hợp nhiều phép chọn thành một phép chọn.

Thực chất của phép chọn là:

- Chọn ra các bản ghi thoả mãn điều kiện chọn.

- Nếu không có bản ghi nào thoả mãn điều kiện chọn thì bảng kết quả thu được là rỗng (∅).

- Điều kiện chọn là một biểu thức Logic.

2.5.2.2. Phép chiếu.

Ví dụ 1: Cho biết tên địa chỉ độc giả.

Quan hệ: DOCGIA

Thuộc tính: ten,diachi

Ta có:

$\pi_{ten,diachi}(DOCGIA)$.

Ví dụ 2: Cho biết mã độc giả có tham gia mượn sách.

Quan hệ: DOCGIA, MUON.

Thuộc tính: maDG.

Ta có:

$\pi_{maDG}(DOCGIA)$.

$\pi_{maDG}(MUON)$.

$\pi_{maDG}(DOCGIA) \cup \pi_{maDG}(MUON)$.

Nhận xét:

Thực chất của phép chiếu là loại bỏ một số thuộc tính của R mà chỉ giữ lại một số thuộc tính X còn lại của quan hệ.

2.5.2.3. Phép tích Đề-Các.

Chú ý: Thông thường theo sau các phép tích Đề-Các là phép chọn.

Ví dụ 1: với mỗi cuốn sách, cho biết thông tin cuốn sách có NXB kim đồng.

Quan hệ: SACH, NXB.

Thuộc tính: tensach,tenNXB,…..

Ta có:

$\delta_{(tenNXB='Kim\ đồng')}(SACH \times NXB)$

Ví dụ 2: cho biết các phòng ban có cùng địa điểm với phòng số 5.

2.5.2.4. Phép nối:

Ví dụ: Cho biết danh sách các cuốn sách có NXB kim đồng

Quan hệ: SACH, NXB.

Thuộc tính: tensach.

Điều kiện: NXB kim đồng.

$$\pi_{\text{tensach}}(\delta_{(\text{tenNXB}=\text{'kim đồng'})}(\text{SACH x NXB})).$$

Nhận xét:

Thực chất phép kết nối là:

- Phép kết nối giữa hai quan hệ r và s thực chất là việc lấy một bản ghi của r "gắn" với một bản ghi của s sao cho bản ghi kết quả thoả mãn điều kiện kết nối.

- Điều kiện kết nối có dạng: A θ B trong đó A, B là hai thuộc tính của r hoặc s. θ là một phép so sánh.

- Phép kết nối như vậy, nói chung gọi là kết nối thường.

2.5.5. Phép chia.

VD1: cho mã độc giả đã mượn tất cả các cuốn sách.

Quan hệ: DOCGIA, MUON.

B1: DG $\leftarrow \pi_{\text{maDG}}$(DOCGIA).

B2: DG_SACH $\leftarrow \pi_{\text{maDG, maSach}}$(muon).

B3 MA_DG $\leftarrow \pi_{\text{maDG}}$(DG_SACH ÷DG).

2.5.6. Hàm tính toán.

Ví Dụ:

R	A	B
	1	2
	3	4
	1	2
	1	2

Ta có:

SUM(B)=10

AVG(A)=1.5

MIN(A) =1

COUNT(A) =4

Nhận xét:

Thực chất của phép chia là:

- Quan hệ R có tập thuộc tính U và quan hệ S có tập thuộc tính V. Phép chia R,S chỉ có thể thực hiện được nếu: V là tập con thực sự của U.

- Quan hệ kết quả thu được có tập thuộc tính là U - V.

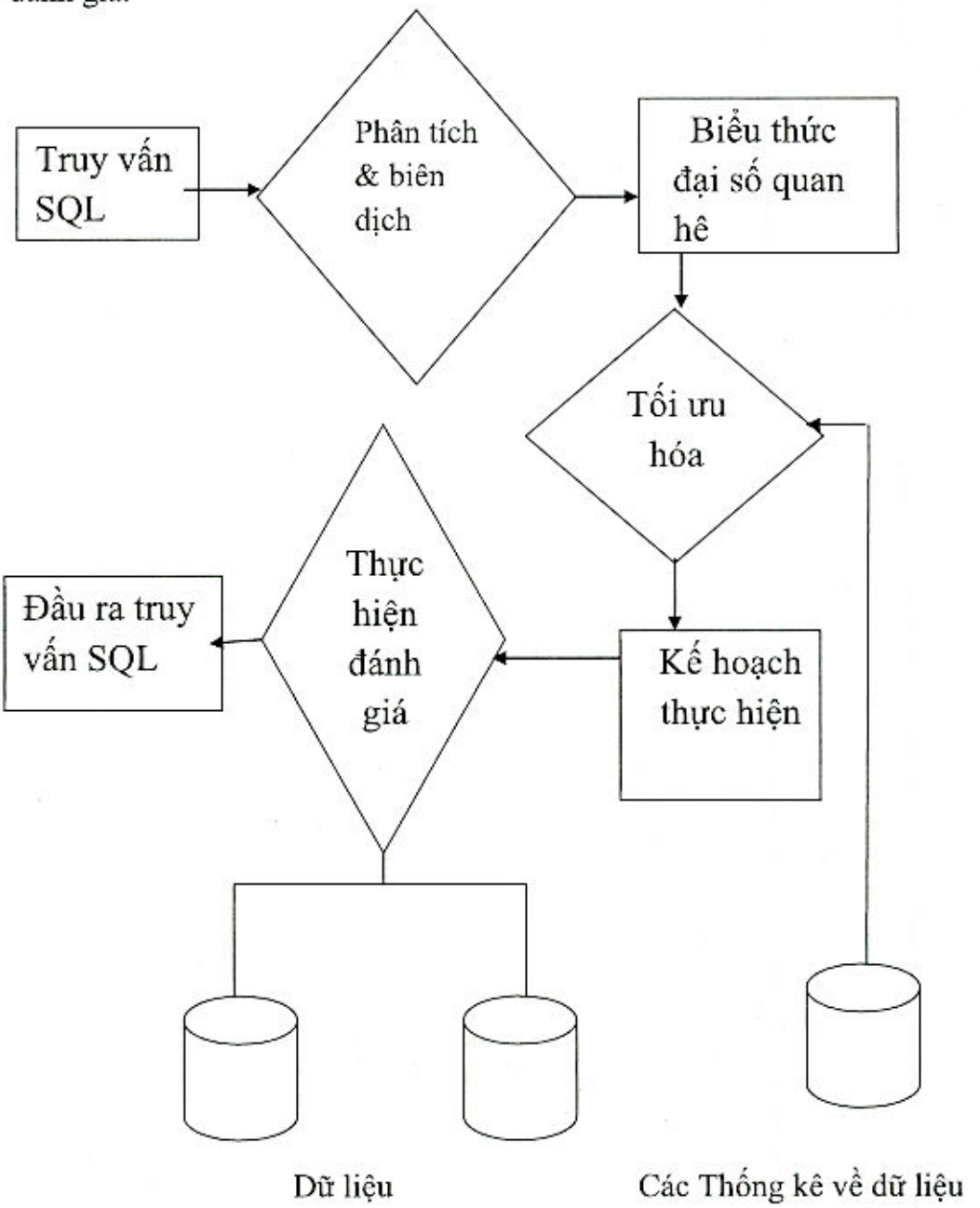
- Bản ghi t nằm trong quan hệ kết quả nếu và chỉ nếu: với mọi bản ghi t' thuộc S thì t^t' là một bản ghi thuộc R. (t^t' là phép lấy t xếp cạnh t').

Chương 3: TỐI ƯU HÓA CÂU TRUY VẤN

3.1. Tổng quan về xử lý truy vấn.

Các bước liên quan đến việc xử lý một truy vấn được minh họa trong hình 1 bao gồm:

* phân tích và dịch.
* tối ưu hóa.
* đánh giá.



Hình 3.1: Các bước trong xử lý truy vấn

*** Phân tích và biên Dịch**

Trước khi việc xử lý truy vấn có thể bắt đầu, hệ thống phải dịch truy vấn này thành một dạng có thể sử dụng được. Một ngôn ngữ chẳng hạng như SQL có thể thích hợp đối với con người khi sử dụng nhưng nó không thích hợp là sự biểu diễn truy vấn bên trong hệ thống. Một sự biểu diễn bên trong thích hợp và có ích hơn là sự biểu diễn dựa trên đại số quan hệ mở rộng. Do vậy, hành động đầu tiên hệ thống này phải thực hiện trong việc xử lý truy vấn là dịch một truy vấn đã cho sang dạng bên trong của nó. Quá trình dịch này là tương tự với công việc được thực hiện với bộ phân tích của một chương trình dịch. Để sinh ra dạng bên trong của truy vấn này, bộ phân tích phải kiểm tra cú pháp của truy vấn được biểu diễn bởi người sử dụng, kiểm tra các tên quan hệ xuất hiện trong truy vấn này có phải là các tên của các quan hệ trong cơ sở dữ liệu đó hay không.... Một biểu diễn cây cú pháp của truy vấn này được xây dựng và sau đó nó được dịch thành một biểu thức đại số quan hệ. Nếu một truy vấn được biểu diễn thông qua các khung nhìn, giai đoạn dịch sẽ thay thế các khung nhìn được sử dụng bởi biểu thức đại số xác định các khung nhìn này.

**** Tối ưu hóa câu truy vấn***

Trong các mô hình mạng và mô hình phân cấp, vấn đề tối ưu hóa câu hỏi là nhiệm vụ của người viết chương trình ứng dụng, bởi vì các lệnh DML của hai mô hình này thường được nhúng trong một ngôn ngữ lập trình chủ và không dễ dàng biến đổi một truy vấn mạng hay phân cấp thành một truy vấn tương đương mà không có tri thức về toàn bộ chương trình ứng dụng. Trái lại, trong mô hình quan hệ, các ngôn ngữ truy vấn hoặc là ngôn ngữ truy vấn khai báo

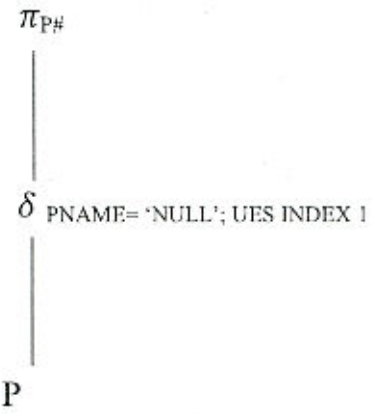
(phí thủ tục) hoặc là ngôn ngữ đại số (thủ tục). các ngôn ngữ khai báo cho phép người sử dụng đặt tả một truy vấn sẽ sinh ra kết quả gì mà không nói hệ thống phải sản sinh kết quả đó như thế nào. Các ngôn ngữ đại số có tính đến sự biến đổi đại số các truy vấn của người sử dụng. Dựa trên sự đặt tả truy vấn này, bộ tối ưu hóa sản sinh các kế hoạch tương đương khác nhau với truy vấn để lựa chọn kế hoạch thực hiện ít tốn kém nhất.

Trong chương trình này chúng ta sẽ thấy rằng đại số quan hệ được cung cấp bởi mô hình quan hệ là một sự giúp đỡ có thể xem xét trong vấn đề tối ưu hóa câu hỏi. Với một câu hỏi cho trước, sẽ có một số phương pháp khác nhau hỏi với việc tính toán câu trả lời. Hơn nữa, chúng ta có thể thực hiện mỗi một phép toán đại số quan hệ sử dụng một số thuộc toán khác nhau. Ví dụ để cài đặt phép chọn trên quan hệ P, chẳng hạn $\delta_{\text{PNAME = 'NULL'}}$ (p), chúng ta có thể diệt cây trong quan hệ P để tìm ra các bộ thõa mãn điều kiện PNAME = 'NULL'. Nếu mỗi chỉ dẫn B- cây là có sẵn trên thuộc tính PNAME này, chúng ta có thể dụng chỉ dẫn này định vị các bộ một các nhanh chóng.

Để đặt tả một cách đầy đủ đánh giá một câu hỏi như thế nào, chúng ta cần cung cấp không chỉ biểu thức đại số quan hệ mà cả chú thích các chỉ thị đặc tả đánh giá mỗi phép toán như thế nào. Các chú thích có thể chỉ ra thuộc toán được sử dụng đối với một phép toán, một chỉ dẫn đặt biệt hay các chỉ dẫn được sử dụng. Một phép toán đại số quan hệ được chú thích với các chỉ thị về đánh giá như thế nào được gọi là một thao tác đánh giá. Một số thao tác đánh giá có được nhóm lại vào một đường ống, trong đó mỗi thao tác bắt đầu làm việc trên các bộ đầu vào của nó thậm chí khi chúng đang được sản sinh bởi một thao tác khác. Một dãy các thao tác sử dụng để đánh giá một câu hỏi được gọi là kế hoạch thực hiện câu hỏi hay kế hoạch đánh giá câu hỏi. Hình 3.2 minh họa một kế hoạch đánh giá câu hỏi $\pi_{P\#}\, \delta_{\text{PNAME='NULL'}}$ (P), trong đó phân biệt trong hình này là index I được đặc tả đối với phép chọn. Bộ thực hiện câu hỏi lấy một kế hoạch thực hiện câu hỏi, thực hiện chúng và trả lại câu trả lời với câu hỏi.

*** Đánh giá.**

Các kế hoạch đánh giá khác nhau đối với một câu hỏi đã cho có thể có các chi phí khác nhau. Chúng ta không mong chờ người sử dụng viết các truy vấn của họ trong đó đã gợi ý kế hoạch đánh giá hiệu quả nhất. Đúng hơn, đó là nhiệm vụ của hệ thống đối với việc xác định một kế hoạch đánh giá câu hỏi để tối ưu hóa chi phí đánh giá câu hỏi.



Hình 3.2: Một kế hoạch đánh giá câu hỏi

Vấn đề tối ưu hóa câu hỏi là quá trình lựa chọn kế hoạch đánh giá hiệu quả nhất đối với một câu hỏi. Một khía cạnh của vấn đề tối ưu hóa xảy ra tại mức đại số quan hệ, ở đó hệ sẽ cố gắng thực hiện các phép biến đổi để tìm ra biểu thức tương đương với một biểu thức đã cho nhưng được thực hiện hiệu quả hơn. Một khía cạnh khác của vấn đề tối ưu hóa liên quan đến việc lựa chọn một chiến lược chi tiết đối với việc xử lý câu hỏi, chẳng hạn lựa chọn thuật toán để thực hiện một phép toán, lựa chọn các chỉ dẫn đặt biệt...

Để lựa chọn một kế hoạch ít tốn kém nhất trong số các kế hoạch đánh giá câu hỏi khác nhau, bộ tối ưu hóa phải đánh giá chi phí của mỗi kế hoạch thực hiện. Việc tính toán chi phí một cách chính xác của một kế hoạch thường là không thể nếu không đánh giá thực sự kế hoạch này. Thay vào đó bộ tối ưu phải sử dụng những thông tin thống kê về các quan hệ, chẳng hạn như kích thước quan hệ, các chỉ dẫn, độ sâu chỉ dẫn để thu được một cách đánh giá tốt đối với chi phí của một kế hoạch.

Xem xét ví dụ trước đây đối với phép chọn được áp dụng trên quan hệ P. Bộ tối ưu hóa ước lượng chi phí của các kế hoạch đánh giá khác nhau giữa một chỉ dẫn là có sẵn trên thuộc tính PNAME của P thì kế hoạch đánh giá được chỉ ra trong hình 3.2, trong đó phép chọn được thực hiện sử dụng chỉ dẫn này được coi như có chi phí thấp nhất và do vậy được lựa chọn.

Khi có một kế hoạch thực hiện truy vấn được lựa chọn, truy vấn được đánh giá với kế hoạch này và kết quả của truy vấn được đưa ra.

3.2. Mô hình chi phí.

3.2.1. Thông tin thư mục đối với đánh giá chi phí.

Chiến lược mà chúng ta lựa chọn để đánh giá câu hỏi phụ thuộc vào chi phí ước lượng của chiến lược này. Bộ tối ưu hóa câu hỏi sử dụng các thông tin thống kê được lưu trữ trong thư mục của hệ quản trị cơ sở dữ liệu để ước lượng chi phí của một kế hoạch. Thông thường mục đích về các quan hệ bao gồm:

N_r là số bộ trong hệ r.

B_r là số khối chứa các bộ của quan hệ r.

S_r là kích thước một bộ của quan hệ r theo số byte.

F_r là số bộ của quan hệ r mà một khối có thể chứa.

V (A ,r) là số các giá trị phân biệt đối với thuộc tính A xuất hiện trong quan hệ r. Giá trị này trùng với lực lượng của $\pi_A(r)$. Nếu A là một khóa của quan hệ r thì V (A,r) là n_r.

SC (A,r) là lực lượng của kết quả chọn trên thuộc tính A của quan hệ r. Cho quan hệ r và một thuộc tính A của quan hệ, SC(A,r) là số trung bình các bộ của quan hệ thõa mãn một điều kiện chọn với phép so sánh = trên thuộc tính A, với giả thiết cho rằng có ít nhất một bộ thỏa mãn điều kiện so sánh = ví dụ, SC (A,r) = 1. Nếu A là khóa của r, với một thuộc tính không khoá, chúng tôi ước lượng rằng V(A,r) các giá trị phân biệt được phân bố đều trong các bộ của quan hệ, do đó SC(A,r) = (n_r/V(A,r)).

Hai thống kê cuối cùng, V(A,r) và SC(A,r) cũng có thể được bảo trì đối với các tập thuộc tính nếu muốn, thay cho các thuộc tính cá thể. Do vậy, với một tập các thuộc tính đã cho A, V(A,r) là lực lượng của $\pi_A(r)$.

Nếu chúng ta giả sử các bộ của quan hệ r được lưu trữ cùng nhau về vật lý trong một tệp, đẳng thức sau là có hiệu lực:

$$B_r = [\, n_r / f_r].$$

Bổ sung thêm các thông tin thư mục về các quan hệ , một số thông tin thư mục về các chỉ dẫn cùng được sử dụng như số mức chỉ dẫn hay chiều cao của chỉ dẫn đối với tổ chức B-cây, số khối trong chỉ dẫn đối với tổ chức băm.

Chúng ta có thể sử dụng các biến thống kê này để ước lượng kích thước của kết quả và chi phí đối với các phép toán và các thuật toán khác nhau.

Nếu chúng ta muốn bảo trì các thống kê này một cách chính xác thì mỗi khi một quan hệ được cập nhập chúng ta cũng phải cập nhật các thống kê này. Cập nhập này phải chịu một tổng chi phí thực sự. Do vậy, hầu hết các hệ thống không cập nhật các thống kê này đối với mỗi cập nhật quan hệ. Thay vào đó, các cập nhật được thực hiện vào thời kỳ tải hệ thống thấp.

Do vậy, các thống kê được sử dụng để lựa chọn một chiến lược xử lý câu hỏi có thể không hoàn toàn chính xác. Tuy nhiên, nếu không có quá nhiều cập nhật xảy ra trong các khoảng thời gian giữa các lần cập nhật ca thống kê này, các thống kê là đủ chính xác đối với việc cung cấp một ước lượng tốt các chi phí liên quan của các kế hoạch khác nhau.

Nhưng thông tin thống kê được đưa ra ở đây đã được đơn giản hóa. Các bộ tối ưu hóa trong thực tế thường bảo trì nhiều thông tin thống kê hơn nữa để tăng cường độ chính xác của các ước lượng chi phí các kế hoạch đánh giá của họ.

3.3. Tối ưu hóa các biểu thức đại số quan hệ.

Tối ưu bằng biến đổi biểu thức đại số quan hệ: Biến đổi thứ tự thực hiện các phép toán của biểu thức đại số quan hệ sao cho các phép toán một ngôi được thực hiện trước các phép toán hai ngôi, do các phép toán chiếu, chọn thì có chi phí nhỏ hơn so với các phép kết nối, tích đề các.

Do vậy, việc tổ chức lại các biểu thức đại số biểu diễn một câu hỏi là cần thiết để giảm chi phí thực hiện các câu hỏi. Trình tự thực hiện các phép toán trong một biểu thức đại số đóng vai trò quan trọng trong quá trình tổ chức câu hỏi.

3.3.1. Đánh giá biểu thức ĐSQH.

Sau các bước phân tích và dịch một truy vấn dựa vào người sử dụng là một biểu thức đại số quan hệ. Với các câu hỏi phức tạp, biểu thức đại số này thường bao gồm nhiều phép toán và tác động trên nhiều quan hệ khác nhau. Việc thực hiện tính toán kết quả của biểu thức đại số này thường rất tốn kém

thời gian và bộ nhớ. Bây giờ, chúng ta sẽ xem xét việc đánh giá một biểu thức chứa nhiều phép toán như thế nào. Một cách hiển nhiên để đánh giá một biểu thức đơn giản đánh giá một phép toán tại một thời điểm theo một thứ tự thích hợp. Kết quả của mỗi đánh giá được vật chất hóa trong một quan hệ tạm thời để sử dụng làm toán hạng cho phép toán tiếp theo. Một bất biến đổi với cách tiếp cận này là cần xây dựng các quan hệ tạm thời mà phải ghi vào đĩa. Một các tiếp cận có thể lựa chọn khác là đánh giá một vài phép toán đồng thời trong một đường ống với các kết quả của các phép toán được chuyển cho phép toán tiếp theo không cần lưu trữ một quan hệ trung gian tạm thời. Ta sẽ phải tiến hành đánh giá biểu thức này. Có 2 hướng tiếp cận để thực thi quá trình đánh giá biểu thức ĐSQH:

+ Vật chất hóa (Materialize).

+ Đường ống (Pipeline).

🎬 Vật chất hóa:

Trong cách tiếp cận này thì ta lần lược đánh giá một biểu thức bằng cách nhìn vào một sự biểu diễn của một biểu thức đại số thông qua một cây toán tử là dễ hiểu nhất.

Nếu chúng ta áp dụng cách tiếp cận vật chất hóa, chúng ta bắt đầu từ các phép toán mức thấp nhất trong biểu thức (tại đáy của cây). Các đầu vào đối với các phép toán mức thấp nhất này là các quan hệ được lưu trữ trong cơ sở dữ liệu. Chúng ta thực hiện các phép toán này sử dụng các thuật toán này trong các quan hệ trung gian tạm thời. Tiếp theo, ta có thể sử dụng các quan hệ trung gian này để thực hiện các phép toán mức trên tiếp theo trong cây toán tử, ở đó các đầu vào bây giờ hoặc là các quan hệ trung gian hoặc là các quan hệ được lưu trữ trong cơ sở dữ liệu. Phép kết nối bây giờ có thể được đánh giá và tạo lập một quan hệ trung gian khác.

Lặp lại quá trình này, ta sẽ đánh giá thực sự phép toán tại gốc của cây và cho kết quả cuối cùng của biểu thức.

Sự đánh giá như vừa được mô tả được gọi là đánh giá vật chất hóa do các kết quả của một phép toán trung gian được tạo lập (được vật chất hóa) và sau đó được sử dụng để đánh giá các phép toán tiếp theo.

Chi phí của một đánh giá vật chất hóa không đơn giản là tổng chi phí các phép toán kéo theo. Khi chúng ta tính toán chi phí ước lượng của các thuật toán cài đặt đối với từng phép toán, chúng ta đã bỏ qua chi phí ghi kết quả của phép toán này vào đĩa. Để tính toán chi phí một biểu thức được thực hiện ở đây, chúng ta phải bổ sung thêm các chi phí của tất cả các phép toán này cũng như chi phí của các thao tác ghi các kết quả trung gian vào đĩa. Chúng ta giả sử rằng, các bản ghi của các kết quả này được chất đống vào một vùng đệm và khi vùng đệm bị đầy chúng sẽ được ghi vào đĩa. Chi phí của thao tác ghi kết quả có thể được ước lượng là n_i / f_i, ở đây n_i là số ước lượng các bộ trong quan hệ kết quả, f_i là số bộ của quan hệ kết quả mà một khối có thể chứa. Vùng đệm đúp (sử dụng hai vùng đệm với một vùng đệm để tiếp tục thực hiện thuật toán trong khi vùng đệm kia được sử dụng đối với thao tác ghi ra đĩa) cho phép thuật toán thực hiện nhanh chóng hơn bởi thực hiện hoạt động CPU song song với hoạt động vào/ra.

• Điểm bất lợi của cách tiếp cận này là việc cần thiết phải xây dựng các quan hệ trung gian tạm thời nhất là khi các quan hệ này thường phải được ghi ra đĩa (trừ khi chúng có kích thước rất nhỏ). Mà việc đọc và ghi ra đĩa có chi phí khá lớn.

Đường ống: Chúng ta có thể cải thiện hiệu quả đánh giá truy vấn bằng cách làm giảm bớt số lượng các quan hệ trung gian tạm thời được tạo ra. Điều này có thể đạt được nhờ việc kết hợp một vài phép toán quan hệ vào một đường ống của các phép toán. Trong đường ống thì kết quả của một phép toán được chuyển trực tiếp cho phép toán tiếp theo mà không cần phải lưu lại trong quan hệ trung gian.

Nếu áp dụng cách tiếp cận vật chất hóa, sự đánh giá sẽ lôi kéo theo tạo lập một quan hệ trung gian để lưu giữ kết quả của phép nối, sau đó đọc kết quả này để thực hiện phép chiếu tiếp theo. Các phép toán này có thể được tổ hợp

như nhau. Khi phép kết nối sinh ra một bộ trong kết quả của nó, bộ này được chuyển trực tiếp đến phép chiếu để xử lý. Việc tổ hợp phép kết nối và phép chiếu cho phép tránh được việc tạo lập kết quả trung gian, thay vào đó tạo lập kết quả cuối cùng một cách trực tiếp.

Chúng ta có thể cài đặt một đường ống bằng cách xây dựng một phép toán tổ hợp các phép toán cấu thành đường ống này, Mặt dù cách tiếp cận này có thể thực hiện được với các tình huống xảy ra khá thường xuyên khác nhau. Nói chung, người ta vẫn hay sử dụng lại đoạn mã đối với từng phép toán cá thể trong việc xây dựng một đường ống. Do vậy mỗi phép toán trong đường ống được mô hình hóa như một tiến trình hay một luồng riềng trong hệ thống mà nó lấy một dòng các bộ tại đầu ra của nó. Với mỗi cặp các phép toán kề nhau trong đường ống, một vùng đệm được tạo lập để lưu giữ các bộ dạng được chuyển tới từ một phép toán đến một phép toán tiếp theo. Đến lược nó, các kết quả của phép kết nối được chuyển đến phép chiếu khi chúng được sinh ra. Bộ nhớ đòi hỏi không cao do các kết quả của một phép toán không được lưu trữ lâu dài. Tuy nhiên, theo cách tiếp cận đường ống, các đầu vào đối với các phép toán là không có sẵn tất cả tại một lần xử lý.

Rõ ràng, cách tiếp cận thứ hai sẽ hạn chế được nhược điểm của cách tiếp cận đầu tiên, nhưng có những trường hợp, ta bắt buộc phải vật chất hóa chứ không dùng đường ống được.

3.3.2. Các chiến lược tối ưu tổng quát.

- Thực hiện các phép chọn và phép chiếu sớm nhất có thể.

Phép biến đổi với chiến lược này nhằm làm giảm bớt kích cỡ của kết quả trung gian và do vậy chi phí phải trả cho việc truy nhập bộ nhớ ngoài cũng như lưu trữ trong bộ nhớ chính sẽ nhỏ đi. Đây là chiến lược quan trọng hơn cả các chiến lược cho phép giảm thời gian thực hiện.

- Tổ hợp các phép chọn xác định với tích Đề-Các thành phép kế nối.

Phép kết nối, đặc biệt là phép kết nối bằng có thể thực hiện "rẻ" hơn so với thực hiện phép tích Đề-Các trên cùng quan hệ. Nếu kết quả của phép tích Đề-Các các RxS là đối số của phép chọn và phép chọn kéo theo phép hội của

các phép so sánh giữa một thuộc tính của R và một thuộc tính của S thì chúng ta có thể áp dụng chiến lược này tổ hợp cả hai phép toán, phép chọn và phép tích Đề- Các thành một phép kết nối.

- Tổ hợp dãy các phép toán một ngôi như các phép chọn và phép chiếu.

Một dãy các phép một ngôi như phép chọn hay phép chiếu mà kết quả của chúng chỉ phụ thuộc vào các bộ của một quan hệ độc lập thì chúng ta có thể nhóm các phép toán đó lại và có thể thực hiện tất cả các phép toán đó đồng thời khi chúng ta quét tệp dữ liệu đối với quan hệ toán hạng.

- Xác định các biểu thức con chung trong một biểu thức.

Nếu kết quả của một biểu thức con chung (biểu thức xuất hiện nhiều hơn một lần) là một quan hệ không lớn lắm và có thể được đọc từ bộ nhớ ngoài với ít thời gian thì nên tính toán trước biểu thức đó chỉ một lần. Biểu thức con chúng có liên quan tới một phép tích Đề-Các hay một phép kết nối thì trong trường hợp tổng quát không thể thay đổi biểu thức tổng thể nhờ việc đẩy phép chọn vào trong.

Điều đáng quan tâm là các biểu thức con chung có tần số xuất hiện được biểu diễn trong các khung nhìn (view) của người sử dụng. Do vậy, khi thực hiện các truy vấn được biểu diễn thông qua các khung nhìn, chúng ta phải thay thế một biểu thức con tương đương ứng với khung nhìn này.

- Xử lý các tệp lệnh trước khi tính toán một phép kết nối (hay tương đương với nó là một phép tích Đề-Các đi theo sau một phép chọn)

Có hai ý tưởng tiền xử lý quan trọng đối với các tệp dữ liệu là sắp xếp trước các tệp và thiết lập các tệp chỉ dẫn. Như vậy, rõ ràng khi thực hiện các phép toán kéo theo hai tệp này (phép toán hai ngôi), các giá trị chung trong hai tệp được kết hợp một các hiệu quả và phép toán được thực hiện nhanh chóng hơn đỡ tốn kém hơn.

- Ước lượng chi phí và lựa chọn thứ tự thực hiện thích hợp.

Một khi cần lựa chọn trình tự thực hiện các phép toán trong một biểu thức hay lựa chọn một trong hay toán hạng của một phép toán hai ngôi cần ước

lượng chi phí thực hiện các phép toán trong đó (số phép tính, thời gian, dung tích bộ nhớ theo một tỷ lệ với kích cỡ của quan hệ...). Từ đó sẽ có được các chi phí phải trả cho các cách khác nhau được thực hiện các truy vấn. Thứ tự các phép toán được chọn ứng với cách thực hiện có chi phí nhỏ hơn cả.

3.3.3. Biểu thức tương đương.

Hầu hết các chiến lược ở trên đều kéo theo một phép chiếu đổi các biểu thức đại số. Trước khi có thể "tối ưu" các biểu thức, chúng ta cần là rõ khái niệm khi nào thì hai biểu thức được gọi là tương đương. Với các định nghĩa quan hệ khác nhau cho các tính chất toán học cũng khác nhau. Nếu quan niệm quan hệ là một tập các n- bộ được sắp với n cố định và khi đó hai quan hệ là tương đương khi và chỉ khi chúng có cùng một tập các bộ. Với quan niệm quan hệ là một tập các ánh xạ từ tập các tên thuộc tính vào tập các giá trị, khi đó hai quan hệ là bằng nhau nếu chúng có cùng một tập ánh xạ. Định nghĩa quan hệ theo quan niệm đầu có thể biến đổi sang định nghĩa quan hệ theo quan niệm thứ hai bằng cách đặt tên thuộc tính cho mỗi cột của bảng biểu diễn quan hệ và ngược lại biến đổi từ định nghĩa hai sang định nghĩa quan hệ theo quan niệm thứ hai bằng cách đặt tên thuộc tính cho mỗi cột của bảng biểu diễn quan hệ và ngược lại biến đổi từ định nghĩa hai sang định nghĩa đầu bằng các cố định thứ tự các thuộc tính. Sau đây, chúng ta sẽ sử dụng định nghĩa thứ hai, nghĩa là quan hệ là một tập ánh xạ từ tập thuộc tính vào tập các giá trị. Các ngôn ngữ truy vấn hiện đang tồn tại, tất cả đều cho phép, thậm chí đòi hỏi tến đối với các cột của bảng biểu diễn một quan hệ. Hơn nữa, trong một ứng dụng cụ thể nào đó, thứ tự các cột của bảng được in ra là không quan trọng khi mỗi cột được gán một tên phản ánh rõ ý nghĩa của nó đối với kết quả đưa ra. Và vì vậy, chúng ta thường lấy tên các thuộc tính cho quan hệ kết quả của một biểu thức đại số là các tên thuộc tính của các đối số của biểu thức.

Một biểu thức trong đại số quan hệ có các toán hạng là biến quan hệ R_1 ,..., R_n và các quan hệ hằng có thể được xác định như ánh xạ từ các k-bộ của

quan hệ (r_1,….,r_k) trong đó r_1 là quan hệ trên cơ đồ quan hệ R_1 đến một quan hệ kết quả đơn khi thay thế mỗi r_1 vào R_1 và đánh giá biểu thức này.

Hai biểu thức đại số E1 và E2 được gọi là tương đương, viết tắt là: E1≡ E2 nếu chúng ta biểu diễn cùng một ánh xạ, nghĩa là khi thay thế cùng các quan hệ và các tên biến giống nhau trong hai biểu thức, chúng ta nhận được cùng một kết quả. Với định nghĩa tương đương này cúng ta có thể liệt kê một số phép biến đổi đại số có ích đối với mục đích tối ưu hóa các biểu thức đại số quan hệ.

3.3.4. Các phép biến dổi tương đương của đại số quan hệ.

Nói rằng 2 biểu thức tương đương nếu thay thế một biểu thức của dạng thứ nhất bởi một biểu thức của dạng thứ hai, và ngược lại ta có thể thay thế biểu thức của dạng thứ hai bởi biểu thức của dạng thứ nhất thì hai biểu thức cùng tạo ra kết quả giống nhau trên bất kỳ hệ CSDL. Các quy tắc sau được sử dụng để biến đổi các biểu thức tương đương với nhau.

Ký hiệu :

$F_1, F_2, \ldots\ldots$ Là các điều kiện.

$L_1, L_2, L_3, \ldots\ldots$ Là tập các thuộc tính.

$E, E_1, E_2, \ldots\ldots$ là các biểu thức đại số quan hệ.

σ : phép chọn.

Π: là phép chiếu.

⋈: Phép kết nối có điều kiện.

∗ : phép kết nối tự nhiên.

X: tích Đề- Các.

3.3.4.1 Các quy tắc cho các phép kết nối và tích Đề các.

a) Quy tắc giao hoán kết nối và tích:

Nếu E1 và E2 là các biểu thức quan hệ, F là một điều kiện trên các thuộc tính của E1 và E2, thì:

$$E_1 \bowtie_F E_2 = E_2 \bowtie_F E_1$$
$$E_1 \bowtie E_2 = E_2 \bowtie E_1$$
$$E_1 \times E_2 = E_2 \times E_1$$

b) Quy tắc kết hợp phép kết nối và tích: Nếu E1, E2 và E3 là các biểu thức quan hệ, F1 và F2 là các điều kiện thì:

$$(E_1 \bowtie_{F1} E_2) \bowtie_{F2} E_3 = E_1 \bowtie_{F1} (E_2 \bowtie_{F2} E_3)$$

$$(E_1 \bowtie E_2) \bowtie E_3 = E_1 \bowtie (E_2 \bowtie E_3)$$

$$(E_1 \times E_2) \times E_3 = E_1 \times (E_2 \times E_3)$$

Định nghĩa quan hệ có 2 định nghĩa tương đương. Định nghĩa thứ nhất phát biểu là quan hệ là một tập con của tích Đề các của các thuộc tính, tức là quan hệ là tập các n_bộ. Hai quan hệ trùng nhau là 2 quan hệ có các bộ trùng nhau. Định nghĩa thứ hai nói rằng, quan hệ là một tập ánh xạ từ các thuộc tính vào tập các giá trị. Và 2 quan hệ trùng nhau nếu tập 2 ánh xạ như nhau. Các quy tắc liên quan đến phép kết nối và tích Đề các được sử dụng theo định nghĩa thứ hai, tập các ánh xạ. $E_1 \times E_2 = E_2 \times E_1$ sẽ đúng vì gọi R và S là các quan hệ có các thuộc tính có chứa tương ứng trong các biểu thức E_1, E_2. Khi đó R.A sẽ được hiểu là thuộc tính của quan hệ R, và S.A cũng sẽ được hiểu là một thuộc tính của S. Gọi μ là một bộ trong $E_1 \times E_2$, khi đó tồn tại một bộ *r* của R và *s* là một bộ của S sao cho μ [R.A] = r[A] và μ [S.A] = s[A]. Tương tự xét $E_2 \times E_1$ sẽ có một bộ *t* sao cho t[R.A] = r[A] và t[S.A] = s[A]. Như vậy μ trùng với *t*. Suy ra $E_1 \times E_2 \subseteq E_2 \times E_1$. Hiển nhiên $E_1 \times E_2 = E_2 \times E_1$.

3.3.4.2. Các quy tắc cho các phép chọn và phép chiếu.

a) Nhóm các phép chiếu thành một phép chiếu duy nhất:

Nếu E là một biểu thức quan hệ và A1, A2,.., An là các thuộc tính có mặt trong B1, A2,.., Bk. Khi đó:

$$\pi_{A1, A2,.., An} (\pi_{B1, A2,.., Bk} (E)) = \pi_{A1, A2,.., An} (E).$$

Nghĩa là thực hiện các phép chiếu liên tiếp trên các thuộc tính của B1, A2,.., Bk, sau đó quan hệ kết quả lại được chiếu trên các thuộc tính của A1, A2,.., An.

b) Nhóm các phép chọn thành một chuỗi phép chọn:

Nếu E là một biểu thức quan hệ và một điều kiện F = F1 ∧ F2 ∧.. ∧ Fn, khi đó:

$$\sigma_{F1 \wedge F2 \wedge .. \wedge Fn}(E) = \sigma_{F1}(\sigma_{F2}(...(\sigma_{Fn}(R))...)).$$

c) Giao hoán các phép chọn.

$$\sigma_{F1}(\sigma_{F2}(E)) = \sigma_{F2}(\sigma_{F1}E)).$$

d) Giao hoán các phép chiếu và phép:

Nếu điều kiện F chỉ chứa các thuộc tính A1, A2,.., An, khi đó:

$$\pi_{A1, A2,.., An}(\sigma F(E)) = \sigma F(\pi_{A1, ., An}(E).$$

Nếu điều kiện F có các thuộc tính B1, A2,.., Bk không chứa các thuộc tính A1, A2,.., An, khi đó:

$$\pi_{A1, A2,.., An}(\sigma F(E)) = \pi_{A1, A2,..,} An(\sigma F(\sigma_{A1, ., An, B1,..Bk}(E)).$$

e) Giao hoán phép chọn và tích Đề các:

Nếu các thuộc tính có mặt trong điều kiện F là các thuộc tính của E1, khi đó:

$$\sigma F(E1 \times E2) = \sigma F(E1) \times E2.$$

Nếu điều kiện F= F1 ^ F2, F1 chứa các thuộc tính của E1 và F2 chứa các thuộc tính E2, khi đó:

$$\sigma F(E1 \times E2) = \sigma F1(E1) \times \sigma F2(E2).$$

Nếu F = F1 ^ F2 , F1 chỉ chứa các thuộc tính của E1 và F2 chứa các thuộc tính của E1 và F2, khi đó:

$$\sigma F(E1 \times E2) = \sigma F2(\sigma F1(E1) \times E2)$$

f) Giao hoán phép chọn và phép hợp.

Nếu biểu thức có dạng E = E1 □ E2 và giả sử các thuộc tính của E1 và E2 có cùng tên với các thuộc tính của E, F là một điều kiện, khi đó:

$$\sigma F(E1 \cup E2) = \sigma F(E1) \cup \sigma F(E2).$$

g) Giao hoán phép chọn và phép trừ.

$$\sigma F(E1 - E2) = \sigma F(E1) - \sigma F(E2).$$

h) Giao hoán phép chọn và phép kết nối tự nhiên.

Nếu điều kiện F chỉ chứa các thuộc tính chung biểu thức E1 và E2, khi đó:

$\sigma F (E1 \bowtie E2) = \sigma F (E1) \bowtie \sigma F (E2)$.

Như vậy, phép chọn được đẩy xuống trong 2 nhánh cây biểu thức (expression tree). Phép chọn làm giảm kích thước của quan hệ kết quả trong cả 2 nhánh.

i) Giao hoán phép chiếu và phép tích Đề các.

Nếu E1 và E2 là các biểu thức, gọi A1, A2,.., An là danh sách các thuộc tính và B1, B2,.., Bk là các thuộc tính của biểu thức E1 và các thuộc tính còn lại, ký hiệu là C1, C2,.., Cj là các thuộc tính của E2 , khi đó:

$\pi A1, A2,.., An (E1 \times E2) = \pi B1, B2,.., Bk (E1) \times \pi C1, C2,.., Cj (E2)$

$\sigma F (E1 \bowtie E2) = \sigma F (E1) \bowtie \sigma F (E2)$.

j) Giao hoán phép chiếu và phép hợp.

$\pi A1, A2,.., An (E1 \cup E2) = \pi A1, A2,.., An (E1) \cup \pi A1, A2,.., An (E2)$.

k) Kết hợp phép giao và phép hợp.

$(E1 \cup E2) \cup E3) = E1 \cup (E2 \cup E3)$

$(E1 \cap E2) \cap E3) = E1 \cap (E2 \cap E3)$

CHƯƠNG 4 : ỨNG DỤNG TỐI ƯU HÓA CÂU TRUY VẤN TRONG CƠ SỞ DỮ LIỆU

4.1. Cơ sở dữ liệu ứng dụng.

Xét cơ sở dữ liệu cho hệ thống quản lý thư viện bao gồm các quan hệ sau đây:

SACH (maSach, tensach, tacgia, sotrang, maNXB) :Quan hệ sách.

NXB (maNXB, tenNXB,diachiNXB):Quan hệ nhà xuất bản.

DOCGIA(maGD, tenDG, diachi, ngaysinh, gioitinh) Quan hệ độc giả.

MUON (maSach, maDG, ngaymuon, ngaytra) Quan hệ mượn.

Trong đó các thuộc tính là:

mSach: mã sách.

tensach: tên sách.

tacgia: tác giả.

sotrang: số trang.

maNXB: mã nhà xuất bản.

tenNXB: tên nhà xuất bản.

diachiNXB: địa chỉ nhà xuất bản.

maDG: mã độc giả.

tenDG: tên độc giả.

diachi: địa chỉ.

ngaysinh: ngày sinh.

gioitinh: giới tính.

ngaymuon: ngày mượn.

ngaytra: ngày trả.

Giả sử :

Quan hệ sách có 100 bộ.

Quan hệ nhà xuất bản 10 bộ.

Quan hệ độc giả có 50 bộ.

Quan hệ mượn có 50 bộ.

4.2. Các quy tắt biến đổi tương đương trong ĐSQH.

QT1: Xử lý toán tử AND trong điều kiện.

$\sigma_{c1\ AND\ c2\ ....AND\ cn}(R) \equiv \sigma_{c1}(\sigma_{c2}(....\sigma_{cn}(R))...)$

SACH (masach, tensacnh, tacgia, maxb, sotrang)

$\sigma_{tensach='văn\ học'\ AND\ tacgia="trần\ đăng\ khoa"}(SACH)$

$\equiv$

$\sigma_{tensach='văn\ học'}(\sigma_{tacgia="trần\ đăng\ khoa"}(SACH))$

QT2: Thay đổi thứ tự của các phép chọn.

$\sigma_{c1}(\sigma_{c2}(R)) \equiv \sigma_{c2}(\sigma_{c1}(R))$

SACH (masach, tensacnh, tacgia, maxb, sotrang)

$\sigma_{tensach='văn\ học'}(\sigma_{tacgia="trần\ đăng\ khoa"}(SACH))$

$\equiv$

$\sigma_{tacgia="trần\ đăng\ khoa"}(\sigma_{tensach='văn\ học'}(SACH))$

QT3: Xử lý các phép chiếu.

SACH (masach, tensacnh, tacgia, maxb, sotrang)

$\pi_{masach,\ tensach}(\pi_{masach,\ tensach,\ tentacgia}(SACH)$

$\equiv$

$\pi_{masach,\ tensach}(SACH)$

QT4: Thay đổi thứ tự các phép chọn và phép chiếu

$\pi_{A1,\ A2,\ ....An}(\sigma_c(R)) \equiv \sigma_c(\pi_{A1,\ A2,...,An}(R))$

SACH (masach, tensacnh, tacgia, maxb, sotrang)

$\pi_{masach,\ tensach}(\sigma_{sotrang='20'}(SACH))$

$\equiv$

$\sigma_{sotrang='20'}(\pi_{masach,\ tensach}(SACH))$

QT5: Tính giao hoán của phép kết và phép tích Descartes

$(R \bowtie_c S) = (S \bowtie_C R)$ $\quad$ $(R \times S) = (S \times R)$

SACH (masach, tensach, tacgia, manxb, sotrang)

NXB (manxb,tennxb,diachinxb)

(SACH ⋈ NXB)

≡

(NXB ⋈ SACH)

QT6.1: Thay đổi thứ tự giữa phép chọn và phép kết

$\sigma_c(R \bowtie S) = (\sigma_c(R)) \bowtie S)$

SACH (masach, tensach, tacgia, manxb, sotrang)

NXB (manxb,tennxb,diachinxb)

$\sigma_{\text{tensach= 'toán'}}$(SACH ⋈ NXB)

≡

$(\sigma_{\text{tensach= 'toán'}}$ (NXB))⋈ SACH)

QT6.2: Phân Phối giữa phép chọn và phép kết

$\sigma(R \bowtie S) \equiv (\sigma_{c1}(R)) \bowtie (\sigma_{c2}(S))$

SACH (masach, tensach, tacgia, manxb, sotrang)

NXB (manxb,tennxb,diachinxb)

$\sigma_{\text{tensach= 'toán' AND tenNXB = ' kim Đồng'}}$(SACH ⋈ NXB)

≡

$(\sigma_{\text{tensach= 'toán'}}$ (SACH)) ⋈ $(\sigma_{\text{tenNXB = ' kim Đồng'}}$ (NXB))

QT 7.1: Phân Phối giữa phép chiếu và phép kết

$\pi_L(R \bowtie_c S) \equiv (\pi_{A1, A2, A3,.....,AN}(R)) \bowtie_C (\pi_{B1, B2, B3,....BM}(S))$

L={A1,…,AN,B1….,BM}; R (A1,…..AN); S(B1,…,BM) Với c ⊂

SACH (masach, tensach, tacgia, manxb, sotrang)

NXB (manxb,tennxb,diachinxb)

$\pi_{\text{masach,tensach,manxb,tennxb}}$ (SACH ⋈ NXB)

≡

$(\pi_{\text{masach,tensach,manxb}}$ (SACH)) ⋈ $((\pi_{\text{tennxb, manxb}}$ (NXB))

QT7.2: Phân phối giữa phép chiếu và phép kết

$\pi_L (R \bowtie_c S) \equiv (\pi_{A1, A2, A3,.....,AN, AN+1, AN+2.....AN+K}^{(R)} \bowtie_C (\pi_{B1, B2, B3,.....,BN, BN+1, BN+2.....BN+P}^{(S)})$

Với c ∉ L ,R $(A_1,.....A_N, A_N, A_{N+1},....A_{N+K})$ S$(B_1,.....B_N, B_N, B_{N+1},....B_{N+K})$

SACH (masach, tensach, tacgia, manxb, sotrang)

NXB (manxb,tennxb,diachinxb)

$\pi_{masach,tensach,manxb}$ (SACH ⋈ NXB)

≡

($\pi_{masach,tensach,manxb}$ (SACH)) ⋈ (($\pi_{tensach,\ manxb}$ (NXB))

QT8: Giao hoán của phép hội và phép giao

R∪ S ≡ S ∪R

R∩S ≡ S ∩ R

QT9: Kết hợp giữa phép kết, tích Descartes, hội và giao

(R θ S) θT = R θ (S θ T)

Trong đó θ là 1 trong các phép toán ⋈ , x, ∪ , ∩

QT 10: Phân phối của phép chọn đối với các phép toán

σ_c(R θ S) = (σ_c(R)) θ(σ_c((S))

Nếu θ là 1 trong các phép toán ∪ , ∩ ,-

QT 11: Phân phối của phép chiếu đối với các phép toán

Nếu θ là 1 trong các phép toán ∪ , ∩ ,-

π_L(R θ S) = (π_L(R)) π_L((S))

QT12: Chuyển các phép (σ, x) thành phép kết

σ_c(R *X* S) = R ⋈$_c$ C

4.3. Giải thuật heuristic.

1. Áp dụng QT1, tách các phép chọn liên kiện thành 1 dãy các phép chọn.

2. Áp dụng QT2,3,4,6,và QR10, để đẩy phép chọn xuống càng sâu càng tốt.

3. Áp dụng QT 9, để tái tổ chức cây cú pháp sao cho phép chọn được thực hiện có lợi nhất (chọn ít nhất) -> heuristic.

4. Phối hợp tích Decartes với các phép chiếu thích hợp theo sau.

5. Áp dụng QT 3,4,7 và 11 để đẩy phép chiếu xuống càng sâu càng tốt (có thể phát sinh phép chiếu mới).

6. Tập trung các phép chọn.

7. Áp dụng QT3 để loại những phép chiếu vô ích.

4.4. Một số câu truy vấn .

1) Cho biết mã độc giả, tên độc giả có địa chỉ ở Hà Nội đã mượn sách vào năm 2014.

2) Cho biết những cuốn sách có tên NXB = Quảng Nam.

3) Cho biết các độc giả chưa trả sách.

4) Cho biết thông tin những độc giả mượn sách trước ngày 13/2/2014 có tên sách toán.

5) Cho biết những cuốn sách có tên tác giả là trần đăng khoa và có số trang >=20 trang.

6) Cho biết các độc giả chưa bao giờ mượn bất kỳ một cuốn sách nào.

4.5. Biểu diễn câu truy vấn bằng ĐSQH.

1) $\pi_{\text{maDG,tenDG}} (\sigma_{\text{(diachi= 'Hà Nội' ^ year(ngay muon= '2014'))}} (\text{DOCGIA} * \text{MUON}))$

2) $\pi_{\text{maSach, tenSach}} (\sigma_{\text{(NXB = 'Quảng Nam')}} (\text{SACH} * \text{NXB}))$

3) $\pi_{\text{maDG,tenDG}} (\sigma_{\text{(ngaytra= 'null')}} (\text{DOCGIA} * \text{MUON}))$

4) $\pi_{\text{maDG,tenDG, maSach, tenSach}} (\sigma_{\text{(ngaymuon < '13/2/2014 '^ tensach= 'Toán')}} (\text{DOCGIA} * \text{MUON*SACH}))$

5) $\pi_{\text{maDG,tenDG}} (\sigma_{\text{(tacgia= 'trần đăng khoa' ^ sotrang>= 20)}} (\text{SACH}))$

6) $\pi_{\text{maDG,tenDG}} (\pi_{\text{maDG}} (\text{DOCGIA}) - \pi_{\text{maDG}} (\text{MUON}))$

4.6 Tối ưu hóa câu truy vấn.

** Tối ưu hóa (1):*

- Sử dụng quy tắt 1 để tách.

$\pi_{\text{maDG,tenDG}} (\sigma_{\text{(diachi= 'Hà Nội'}} (\sigma_{\text{year(ngay muon= '2014'))}} (\text{DOCGIA} * \text{MUON}))$

- Sử dụng quy tắt 6.2

$\pi_{\text{maDG,tenDG}} (\sigma_{\text{(diachi= 'Hà Nội'}} (\sigma_{\text{year(ngay muon= '2014'))}} (\text{MUON}))*\text{DOCGIA})$

- Sử dụng quy tắt 6.2

$\pi_{\text{maDG,tenDG}} (\sigma_{\text{(diachi= 'Hà Nội'}} (\text{DOCGIA}) * \sigma_{\text{year(ngay muon= '2014'))}} (\text{MUON}))$

** Tối ưu hóa (2)*

- Sử dụng quy tắt 6.2

$\pi_{\text{maSach, tenSach}} (\sigma_{\text{(NXB = 'Quảng Nam')}} (\text{NXB}) * (\text{SACH}))$

** Tối ưu hóa (3)*

- Sử dụng quy tắt 6.2

$\pi_{maDG,tenDG}(\sigma_{(ngaytra='null')}(MUON)*(DOCGiA))$

** Tối ưu hóa (4)*

- Sử dụng quy tắt 1 để tách

$\pi_{maDG,tenDG,\ maSach,\ tenSach}(\sigma_{(ngaymuon<'13/2/2014'}(\sigma_{tensach='Toán')}(DOCGIA * MUON*SACH)))$

- Sử dụng quy tắt 6.2

$\pi_{maDG,tenDG,\ maSach,\ tenSach}(\sigma_{(ngaymuon<'13/2/2014'}(MUON))*(\sigma_{(tensach='Toán')}(*SACH))*DOCGIA))$

** Tối ưu hóa (5)*

- Sử dụng quy tắt 1 để tách

$\pi_{maDG,tenDG}(\sigma_{(tacgia='trần đăng khoa'}(\sigma_{(sotrang>=20)}(SACH))$

- Sử dụng quy tắt 6.2

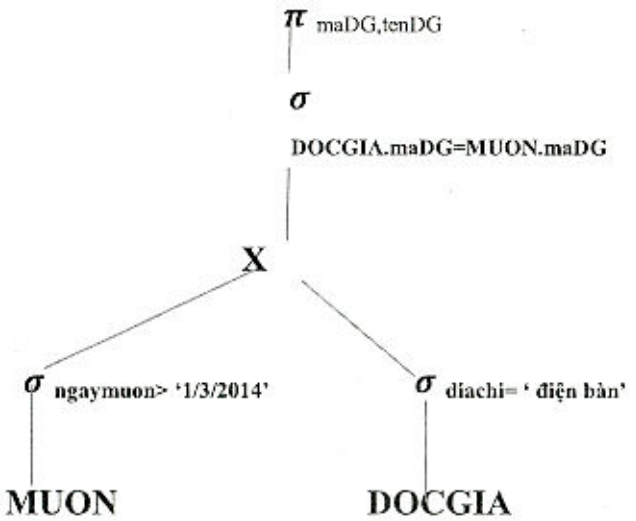
$\pi_{maDG,tenDG}(\sigma_{(tacgia='trần đăng khoa')}(SACH)*(\sigma_{(sotrang>=20)}(SACH))$

* Tối ưu hóa (6)

$\pi_{maDG,tenDG}(\pi_{maDG}(DOCGIA) - \pi_{maDG}(MUON))$

4.7 Vẽ cây truy vấn.

*** Biểu diễn cây truy vấn (1)**



Nhận Xét: Theo cách đánh giá theo mô hình chi phí

Trong đó :

Thỏa điều kiện 1:có 10 bộ

Thỏa điều kiện 2: có 5 bộ

Theo truy vấn ban đầu:

$\pi_{maDG,tenDG}\ (\ \sigma_{(diachi='Hà Nội' \wedge year(ngay\ muon='2014'))}\ (DOCGIA * MUON)\ )$

- Tổng chi phí theo ban đầu

Thì tổng chi phí bằng thực thi yêu cầu thực hiện phép nối của bản mượn với bản độc giả -> 2500 lần. Sau đó ta đi thực hiện phép chọn. Suy ra chi phí thực hiện cao nhất là 2500 lần.

Theo truy vấn đã tối ưu:

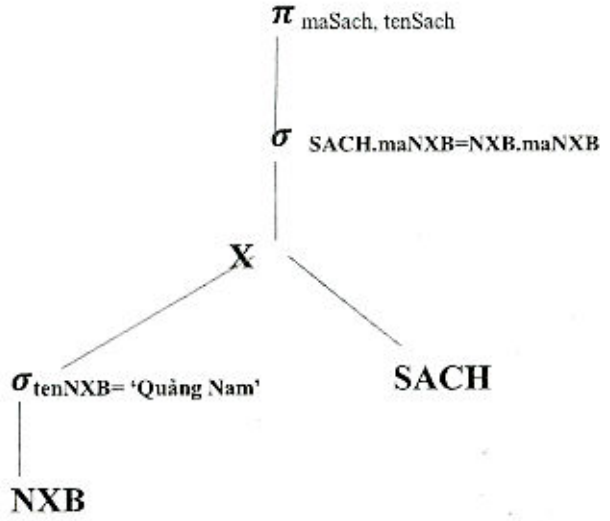
$\pi_{maDG,tenDG}\ (\ \sigma_{(diachi='Hà Nội'}\ (DOCGIA) * \sigma_{year(ngay\ muon='2014'))}\ (MUON)\ )$

- Tổng chi phí đã được tối ưu

Thì tổng chi phí bằng thực thi phép chọn ở điều kiện 1 ->10 bộ thực thi phép chọn ở điều kiện 2 -> 5 bộ nối hai bản lại với nhau -> 50 lần . Suy ra chi phí thực hiện cao nhất là 50 lần.

Thời gian ngắn hơn so với truy vấn ban đầu.

*** Biểu diễn cây truy vấn (2):**



Nhận Xét: Theo cách đánh giá theo mô hình chi phí

Trong đó :

Thỏa điều kiện: có 10 bộ

Theo truy vấn ban đầu:

$\pi_{maSach, tenSach}$ ($\sigma_{(NXB = 'Quảng Nam')}$ (SACH * NXB))

- Tổng chi phí theo ban đầu:

Thì tổng chi phí bằng thực thi yêu cầu thực hiện phép nối của bản NXB với bản SACH ≤ 1500 lần. Sau đó ta đi thực hiện phép chọn. Suy ra chi phí thực hiện cao nhất là 1500 lần

Theo truy vấn đã tối ưu:

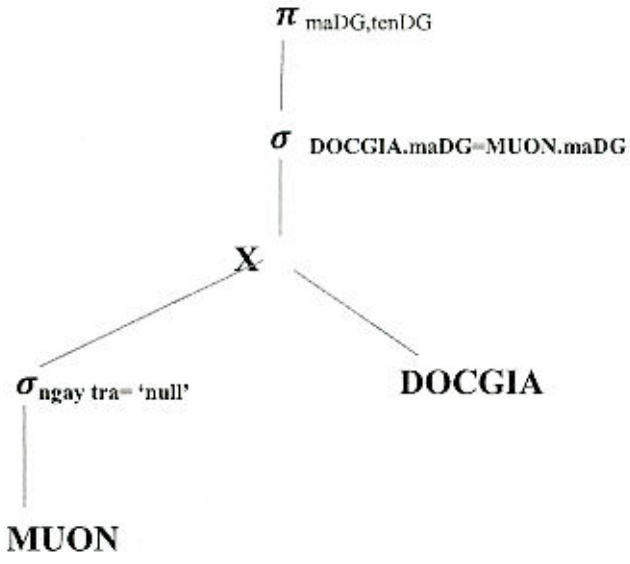
$\pi_{maSach, tenSach}$ ($\sigma_{(NXB = 'Quảng Nam')}$ (NXB) *(SACH))

- Tổng chi phí đã được tối ưu:

Thì tổng chi phí bằng thực thi phép chọn ở điều kiện ->10 bộ. Sau đó nối hai bản lại với nhau ≤ 1000 lần. Suy ra chi phí thực hiện cao nhất là 1000 lần.

Thời gian ngắn hơn so với truy vấn ban đầu.

*** Biểu diễn cây truy vấn (3):**



Nhận Xét: Theo cách đánh giá theo mô hình chi phí

Trong đó :

Thỏa điều kiện: có 5 bộ

Theo truy vấn ban đầu:

$\pi_{maDG,tenDG}$ ($\sigma_{(ngaytra= 'null')}$ (DOCGIA * MUON))

- Tổng chi phí theo ban đầu:

Thì tổng chi phí bằng thực thi yêu cầu thực hiện phép nối của bản DOCGIA với bản MUON ≤ 2500 lần. Sau đó ta đi thực hiện phép chọn. Suy ra chi phí thực hiện cao nhất là 2500 lần.

Theo truy vấn đã tối ưu:

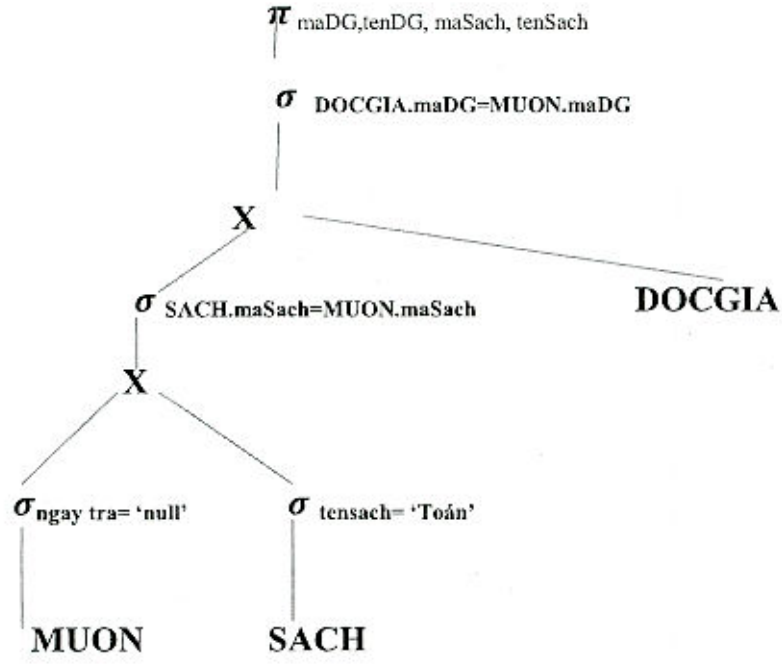
$\pi_{maDG,tenDG}$ ($\sigma_{(ngaytra='null')}$ (MUON)*(DOCGiA))

- Tổng chi phí đã được tối ưu:

Thì tổng chi phí bằng thực thi phép chọn ở điều kiện ->10 bộ. Sau đó nối hai bản lại với nhau -> 500 lần. Suy ra chi phí thực hiện cao nhất là 500 lần.

Vậy thời gian câu truy vấn thực hiện sau khi đã tối ưu ngắn hơn so với truy vấn ban đầu.

*** Biểu diễn cây truy vấn (4):**



Nhận Xét: Theo cách đánh giá theo mô hình chi phí

Trong đó :

Thỏa điều kiện 1: có 5 bộ

Thỏa điều kiện 2: có 5 bộ

Theo truy vấn ban đầu:

π maDG,tenDG, maSach, tenSach (σ (ngaymuon < '13/2/2014 '^ tensach= 'Toán') (DOCGIA * MUON*SACH))

- Tổng chi phí theo ban đầu:

Thì tổng chi phí bằng thực thi yêu cầu thực hiện phép nối của bản DOCGIA, MUON, SACH -> 250000 lần. Sau đó ta đi thực hiện phép chọn .Suy ra chi phí thực hiện cao nhất là 250000 lần

Theo truy vấn đã tối ưu:

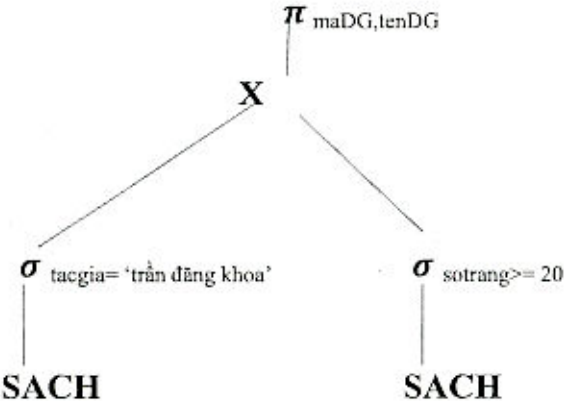
π maDG,tenDG, maSach, tenSach (σ (ngaymuon < '13/2/2014 ' (MUON))* (σ (tensach= 'Toán') (*SACH))*DOCGIA))

- Tổng chi phí đã được tối ưu:

Thì tổng chi phí bằng thực thi phép chọn ở điều kiện 1 ->5 bộ, phép chọn ở điều kiện 2 -> 5. Sau đó nối ba bản lại với nhau -> 1250 lần . Suy ra chi phí thực hiện cao nhất là 1250 lần.

Vậy thời gian câu truy vấn thực hiện sau khi đã tối ưu ngắn hơn so với truy vấn ban đầu

*** Biểu diễn cây truy vấn (5):**



Nhận Xét: Theo cách đánh giá theo mô hình chi phí

Trong đó :

Thỏa điều kiện 1: có 15 bộ

Thỏa điều kiện 2: có 5 bộ

Theo truy vấn ban đầu:

π maDG,tenDG (σ (tacgia= 'trần đăng khoa' ^ sotrang>= 20) (SACH))

- Tổng chi phí theo ban đầu:

Thì tổng chi phí bằng thực thi yêu bảng sách -> 100 lần. Sau đó ta đi thực hiện phép chọn. Suy ra chi phí thực hiện cao nhất là 1000 lần.

Theo truy vấn đã tối ưu:

$\pi_{maDG,tenDG}(\sigma_{(tacgia='trần đăng khoa')}(SACH) * (\sigma_{(sotrang>=20)}(SACH))$

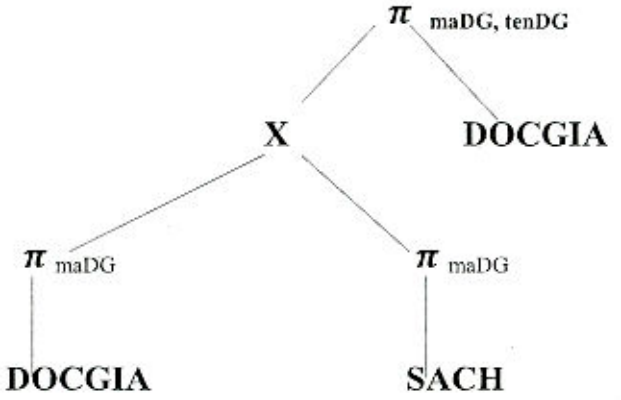
- Tổng chi phí đã được tối ưu:

Thì tổng chi phí bằng thực thi phép chọn ở điều kiện 1 ->15 bộ, phép chọn ở điều kiện 2 -> 5. Sau đó nối lại với nhau -> 75 lần . Suy ra chi phí thực hiện cao nhất là 75 lần.

Vậy thời gian câu truy vấn thực hiện sau khi đã tối ưu ngắn hơn so với truy vấn ban đầu.

*** Biểu diễn cây truy vấn (6):**

$\pi_{maDG,tenDG}(\pi_{maDG}(DOCGIA) - \pi_{maDG}(MUON))$



Nhận Xét: Theo cách đánh giá theo mô hình chi phí

Trong đó :

Thỏa điều kiện 1: có 5 bộ

Thỏa điều kiện 2: có 5 bộ

$\pi_{maDG,tenDG}(\pi_{maDG}(DOCGIA) - \pi_{maDG}(MUON))$

- Tổng chi phí đã được tối ưu:

Thì tổng chi phí bằng thực thi phép chọn ở điều kiện 1 ->5 bộ, phép chọn ở điều kiện 2 -> 5. Sau đó nối lại với nhau -> 25 lần. Suy ra chi phí thực hiện cao nhất là 25 lần.

Vậy thời gian câu truy vấn thực hiện sau khi đã tối ưu ngắn hơn so với truy vấn ban đầu.

PHẦN 3: KẾT LUẬN VÀ KIẾN NGHỊ

3.1. Kết luận.

3.1.1. Kết quả đạt được.

Khóa luận đã trình bày được đầy đủ, chính xác một số kiến thức cơ bản có liên quan đến nội dung nghiên cứu của đề tài như: tổng quan về cơ sở dữ liệu, đại số quan hệ, tối ưu hóa câu truy vấn.

- Trình bày, so sánh độ nhanh chậm của việc xử lý các câu truy vấn trong cơ sở dữ liệu.

- Một số ví dụ minh họa trên cho ta thấy một minh họa về việc chuyển đổi một câu hỏi bằng ngôn ngữ đại số quan hệ về dạng tương đương tốt hơn (hay tối ưu hơn). Phương pháp trên tập trung chủ yếu vào các phép chiếu, phép chọn, với mục đích là sao "đẩy" được phép chọn, phép chiếu xuống mức thất nhất, tức là thi hành các phép toán này càng sớm càng tốt, nếu có thể. Tiếp theo, kết hợp các phép chọn với tích Đề - Các thành phép kết nối tự nhiên để giảm các kết quả trung gian. Cốt lõi của vấn đề tối ưu hóa chính là việc làm giảm thiểu lưu trữ trung gian và từ đó làm tăng nhanh tốc độc xử lý câu hỏi.

3.1.3. Hạn chế của đề tài.

- Đề tài chỉ nghiên cứ tối ưu hóa câu truy vấn trên đại số quan hệ.

- Đề tài chỉ dừng lại ở một số câu truy vấn cơ bản chưa nghiên cứu được tất cả các câu truy vấn phức tạp.

3.2. Kiến nghị.

Vì thời gian và kiến thức có hạn nên dù đề tài có hoàn chỉnh đến đâu cũng không tránh khỏi những hạn chế, sai sót. Vì vậy em rất mong nhận được những đóng góp ý kiến cũng như phản hồi từ các bạn và các quý thầy cô để đề tài này được hoàn chỉnh hơn và có cơ hội được phát triển sau này.

Phần 4: TÀI LIỆU THAM KHẢO

Sách:

[1] *Giáo Trình "cơ sở dữ liệu"* khoa CNTT Đại Học Khoa Học TPHCM

[2] *Các Hệ Quản Trị Cơ Sở Dữ Liệu Và Cơ Sở Tri Thức* – J.Ullman – Biên Dịch Trần Đức Quan.

Tài liệu từ website:

[1] *http://tailieu.vn*

[2] *http://www.doko.vn*