Bài Tập Lớn Môn Học

Nhập Môn Cơ Sở Dữ Liệu

1. *Mục đích*

Thông qua quá trình thực hiện bài tập lớn Môn học "Nhập môn Cơ sở dữ liệu", sinh viên được củng cố phần kiến thức lý thuyết (mô hình quan hệ; ngôn ngữ SQL; các nguyên lý về thiết kế hệ CSDL; tối ưu hóa câu hỏi; an toàn và toàn vẹn dữ liệu).

Sinh viên làm quen với việc nghiên cứu: đọc, tìm hiểu và phân tích những kết quả thông qua các bài báo.

Sinh viên có nhìn nhận về các ứng dụng thực tiễn của cơ sở dữ liệu, cũng như có các kỹ năng cần thiết về lập trình, thao tác trên hệ cơ sở dữ liệu.

2. *Yêu cầu bài tập lớn*

Bài tập lớn gồm có 2 phần: tìm hiểu lý thuyết và cài đặt thử nghiệm. *Phần cài đặt thử nghiệm tùy theo đề bài có mức độ khác nhau (được chỉ ra cụ thể trong từng đề bài), tuy nhiên đây là phần khuyến khích nên có.*

Phần tìm hiểu lý thuyết: dựa trên những gợi ý về tài liệu tham khảo, sinh viên tìm hiểu theo diện rộng các phương pháp, các giải pháp tiếp cận xử lý bài toán. Trên cơ sở đó đưa ra những phân tích, đánh giá ưu nhược điểm của mỗi phương pháp; khả năng áp dụng thực tế và ngữ cảnh sử dụng; lựa chọn và đưa ra phương pháp phục vụ cho việc cài đặt. *Nhằm nâng cao kỹ năng nghiên cứu, sinh viên cần thể hiện được tính khoa học, tính logic trong bố cục trình bày, phát triển vấn đề.*

Phần cài đặt thử nghiệm: sinh viên cài đặt thuật toán đã lựa chọn. Với chương trình cài đặt, sinh viên cần thử nghiệm ít nhất với một cơ sở dữ liệu.

Sinh viên làm báo cáo theo nhóm. Báo cáo trình bày những phần sau:

- Mô tả bài toán.
- Phương pháp thực hiện: trình bày các phương pháp tìm hiểu được, tập trung vào giải pháp lựa chọn để giải quyết bài toán.
- Cài đặt chương trình, thử nghiệm và đánh giá, xây dựng CSDL dùng thử nghiệm.
- Tài liệu tham khảo.

Chương trình có thể bằng ngôn ngữ tùy chọn, nếu có sử dụng thư viện và phần mềm hỗ trợ thì cần ghi hướng dẫn cài đặt và sử dụng.

3. *Tổ chức thực hiện*

Các lớp tổ chức thành nhóm 5 người, mỗi nhóm nhận một đề bài tập lớn.

Nộp báo cáo, chương trình và tài liệu tham khảo *theo nhóm*.

Bài tập lớn được giao trong quá trình học lý thuyết.

Thời gian bảo vệ bài tập lớn: 8 tuần cuối của môn học.

Kế hoạch bảo vệ bài tập lớn K52 (tín chỉ): 209 sv→42 nhóm, 6 nhóm 1 đề.

- Tuần 10: hướng dẫn sinh viên về BTL.
- Tuần 11: bảo vệ các nhóm đề 1.
- Tuần 12: bảo vệ các nhóm đề 2.
- Tuần 13: bảo vệ các nhóm đề 3.
- Tuần 14: bảo vệ các nhóm đề 4.
- Tuần 15: bảo vệ các nhóm đề 5.
- Tuần 16: bảo vệ các nhóm đề 6.
- Tuần 17: bảo vệ các nhóm đề 7.

4. . Đề bài tập lớn

4.1 Dạng 1 – Kiểm thử ứng dụng database

4.1.1 Đặt vấn đề

Trong quá trình xây dựng các ứng dụng phần mềm sử dụng CSDL, người lập trình thường xuyên phải xây dựng các truy vấn phục vụ cho việc tìm kiếm và xử lý dữ liệu. Bên cạnh việc đảm bảo truy vấn thực hiện nhanh, tối ưu thì việc đảm bảo cho truy vấn trả lại kết quả chính xác theo ý định của người dùng là việc rất quan trọng. Quá trình này được thực hiện bằng kiểm thử phần mềm, dựa trên tập CSDL thử nghiệm. Vấn đề ở đây là cần có phương pháp đánh giá chất lượng của quá trình kiểm thử này, đánh giá tập CSDL được dùng đã kiểm thử được cho truy vấn ở mức độ nào.

Ở khía cạnh khác, trong quá trình kiểm thử đối với các ứng dụng CSDL, đòi hỏi có CSDL phù hợp cho việc kiểm thử. Theo cách truyền thống, dữ liệu được tạo ra trước truy vấn, sau đó thực hiện dãy các truy vấn trên dữ liệu để so sánh kết quả thu nhận được với kết quả mong đợi nhằm đánh giá hoạt động của hệ thống. Tuy nhiên trong mỗi trường hợp kiểm thử, mỗi truy vấn thường gồm nhiều phép toán và có những yêu cầu riêng về kích thước của bộ kết quả sau mỗi phép toán. Cách tạo dữ liệu truyền thống không đáp ứng được yêu cầu này, đòi hỏi cần có cách thức tạo ra bộ dữ liệu cho mỗi trường hợp kiểm thử, nhằm thỏa mãn các yêu cầu của quá trình kiểm thử và thỏa mãn các ràng buộc về dữ liệu.

4.1.2 Bài tập lớn dạng 1

Đề 1: Sử dụng độ đo SQL trong đánh giá việc kiểm thử các ứng dụng database

Tìm hiểu các độ đo SQL dùng để đánh giá việc kiểm thử các ứng dụng database, cấu trúc biểu diễn câu truy vấn SQL, các phương pháp và các kỹ thuật tự động tính toán độ đo nhằm đánh giá mức độ kiểm thử đối với câu truy vấn (có chứa mệnh đề WHERE, GROUP, HAVING…); lựa chọn phương pháp khả thi và xây dựng chương trình thử nghiệm (nếu có). Các kỹ thuật nên chỉ rõ ngữ cảnh áp dụng, có sự so sánh, đánh giá bằng thực nghiệm.

Gợi ý tài liệu tham khảo:

Using an SQL Coverage Measurement for Testing Database Applications, María José Suárez-Cabal, Javier Tuya, ACM SIGSOFT Software Engineering Notes archive Volume 29, Issue 6 (November 2004).

Structural Coverage Criteria for Testing SQL Queries, María José Suárez-Cabal, Javier Tuya, 2009.

Đề 2: Xây dựng dữ liệu kiểm thử hướng truy vấn

Tìm hiểu các phương pháp và kỹ thuật xây dựng dữ liệu kiểm thử hướng truy vấn: kỹ thuật phân tích truy vấn, xử lý các phép toán truy vấn, sinh dữ liệu cho việc kiểm thử. Lựa chọn ra phương pháp khả thi và xây dựng chương trình thử nghiệm (nếu có). Các kỹ thuật nên chỉ rõ ngữ cảnh áp dụng, có sự so sánh, đánh giá bằng thực nghiệm.

Gợi ý tài liệu tham khảo:

QAGen: Generating Query-Aware Test Databases, Carsten Binnig, Donald Kossmann, Eric Lo, Marc Nunkesser, Tamer Ozsu Proceedings of the ACM SIGMOD International Conference on Management of Data, Beijing, China, June 12-14, 2007.

4.1.3 Yêu cầu

Sinh viên tìm hiểu chú trọng về mặt lý thuyết, phương pháp luận, thuật toán, so sánh các phương pháp.

Cài đặt được khuyến khích.

4.2 Dạng 2– Phân hạng, phân loại kết quả truy vấn

4.2.1 Đặt vấn đề

Trong cuộc sống hiện nay, các ứng dụng, dịch vụ sử dụng CSDL ngày càng phổ biến, từ việc thanh toán qua ngân hàng, lưu trữ dữ liệu trong doanh nghiệp, đặt vé máy bay, tổ chức thông tin thư viện... Một điểm chung là các các CSDL này ngày càng tăng về kích thước nhằm làm phong phú dữ liệu, đáp ứng nhu cầu tìm kiếm thông tin của người dùng, cũng như phục vụ một số lượng đông đảo người dùng. Tăng kích thước CSDL dẫn đến việc khi người dùng nhập truy vấn tìm kiếm dữ liệu và thực hiện truy vấn, kết quả thỏa mãn trả về là rất nhiều, đòi hỏi cần có phương pháp xếp hạng, phân loại các kết quả. Thông thường, người dùng chỉ quan tâm đến một số K kết quả truy vấn đầu tiên hoặc chỉ thu nhận kết quả theo những nhóm quan tâm. Do đó, việc phân hạng, phân loại là việc làm có ý nghĩa quan trọng, đặc biệt trong các ứng dụng tìm kiếm lớn trong doanh nghiệp, tìm kiếm trên Web.

4.2.2 Bài tập lớn dạng 2

Đề 3: Các kỹ thuật phân hạng kết quả truy vấn

Tìm hiểu các kỹ thuật tự động phân hạng (ranking) kết quả trong truy vấn cơ sở dữ liệu, nhằm phân hạng kết quả theo mức độ thỏa mãn truy vấn của người dùng; lựa chọn kỹ thuật khả thi và xây dựng chương trình thử nghiệm (nếu có). Các kỹ thuật nên chỉ rõ ngữ cảnh áp dụng, có sự so sánh, đánh giá bằng thực nghiệm.

Gợi ý tài liệu tham khảo:

Probabilistic ranking of database query results, Surajit Chaudhuri, Gautam Das, Vagelis Hristidis, Gerhard Weikum, Microsoft Research, One Microsoft Way, Redmond, WA, 2004.

Automated ranking of database query results, Surajit Chaudhuri , Gautam Das, Microsoft Research, One Microsoft Way, Redmond, WA, 2003.

Đề 4: Các kỹ thuật phân loại kết quả truy vấn

Tìm hiểu các kỹ thuật tự động phân loại kết quả trong truy vấn cơ sở dữ liệu, tạo điều kiện thuận lợi cho người dùng xem những nhóm kết quả mà mình quan tâm; lựa chọn kỹ thuật khả thi và xây dựng chương trình thử nghiệm (nếu có). Các kỹ thuật nên chỉ rõ ngữ cảnh áp dụng, có sự so sánh, đánh giá bằng thực nghiệm.

Gợi ý tài liệu tham khảo:

Automatic categorization of query results, Kaushik Chakrabarti Surajit Chaudhuri, Seung-won Hwang, University of Illinois, 2004.

4.2.3 Yêu cầu

Sinh viên tìm hiểu các kỹ thuật, đồng thời nên có cài đặt cụ thể.

4.3 Dạng 3 - Phát hiện lỗi ngữ nghĩa trong truy vấn

4.3.1 Đặt vấn đề

Trong thao tác truy vấn CSDL, truy vấn có thể sai về mặt cú pháp như nhập sai từ khóa, thiếu điều kiện truy vấn, sai cấu trúc truy vấn…; trong những trường hợp như vậy, hệ CSDL sẽ báo cho người dùng biết truy vấn có lỗi, không thực hiện được. Những lỗi như trên được gọi là lỗi về cú pháp. *Tuy nhiên, một câu truy vấn đã đúng về cú pháp thì vẫn có thể có lỗi về ngữ nghĩa như có mâu thuẫn giữa các mệnh đề làm cho kết quả truy vấn là rỗng hoặc câu truy vấn phức tạp không cần thiết. Những lỗi này gọi là lỗi ngữ nghĩa, hay lỗi logic, rất thường gặp khi sinh viên bắt đầu làm quen với hệ CSDL*.

Mục đích bài tập lớn dạng này rèn luyện cho sinh viên viết câu truy vấn đúng cú pháp, đúng ngữ nghĩa, tránh được các lỗi thường gặp về ngữ nghĩa, thành thạo các thao tác đối với hệ CSDL.

Phát hiện ra mâu thuẫn về mặt ngữ nghĩa trong biểu thức truy vấn là không đơn giản, và rất khó có thể phát hiện đầy đủ các trường hợp. Mục đích của xây dựng chương trình nhằm đưa ra được một công cụ hỗ trợ, giảm bớt các lỗi có thể mắc phải trong quá trình người dùng tạo câu truy vấn.

4.3.2 Đề bài tập lớn dạng 3

Đề 5: Phát hiện lỗi ngữ nghĩa trong truy vấn

Tìm hiểu các trường hợp lỗi ngữ nghĩa có thể mắc phải khi viết câu truy vấn (lỗi trong mệnh đề SELECT, WHERE, FROM, GROUP BY…), thuật toán để phát hiện lỗi ngữ nghĩa đó, cần đưa ra các minh họa ví dụ. Xây dựng chương trình thử nghiệm phát hiện lỗi ngữ nghĩa.

Gợi ý tài liệu tham khảo:

Semantic Errors in SQL Queries: A Quite Complete List, Stefan Brass, Christian Goldberg.

Proving the Safety of SQL Queries, Stefan Brass, Christian Goldberg.

Detecting Logical Errors in SQL Queries, Stefan Brass, Christian Goldberg.

Hệ Sqllint: http://dbs.informatik.uni-halle.de/sqllint/

4.3.3 Yêu cầu

Bên cạnh mặt lý thuyết của phương pháp, sinh viên nên cài đặt chương trình để phát hiện các lỗi ngữ nghĩa truy vấn, với yêu cầu là phát hiện ra các lỗi ngữ nghĩa hay gặp và không đòi hỏi phương pháp quá phức tạp.

Sinh viên có thể thay thế việc lập trình bằng cách khai thác phần mềm Sqllint, phân tích về các trường hợp lỗi mà hệ bắt được, tính hữu dụng của phần mềm này, đưa ra các ví dụ và phân tích kết quả.

4.4 Dạng 4 – SQL Injection trong ứng dụng Web

4.4.1 Đặt vấn đề

Việc thiết kế và đưa vào hoạt động một website luôn đòi hỏi các nhà phát triển phải quan tâm đến các vấn đề về an toàn, bảo mật nhằm giảm thiểu tối đa khả năng bị tấn công từ các tin tặc. Thông thường, các nhà phát triển đa số tập trung vào các vấn đề an toàn trong việc chọn hệ điều hành, hệ quản trị CSDL, webserver sẽ chạy ứng dụng… mà quên mất rằng ngay cả bản thân ứng dụng chạy trên đó cũng tiềm ẩn một lỗ hổng bảo mật rất lớn. Một trong số các lỗ hổng này đó là SQL injection. *SQL injection là một kĩ thuật cho phép những kẻ tấn công lợi dụng lỗ hổng trong việc kiểm tra dữ liệu nhập trong các ứng dụng web và các thông báo lỗi của hệ quản trị cơ sở dữ liệu để "tiêm vào" (inject) và thi hành các câu lệnh SQL bất hợp pháp (không được người phát triển ứng dụng lường trước).* Hậu quả của nó rất tai hại vì nó cho phép những kẻ tấn công có thể thực hiện các thao tác xóa, hiệu chỉnh…do có toàn quyền trên cơ sở dữ liệu của ứng dụng, thậm chí là server mà ứng dụng đó đang chạy. Lỗi này thường xảy ra trên các ứng dụng web có dữ liệu

được quản lí bằng các hệ quản trị cơ sở dữ liệu như SQL Server, MySQL, Oracle, DB2, Sysbase.

4.4.2 Bài tập lớn dạng 4

Đề 6: SQL Injection trong ứng dụng Web

Tìm hiểu các kỹ thuật tấn công SQL injection, đưa ra các ví dụ minh họa, phân tích nguyên nhân mất an toàn, các giải pháp để phòng tránh các tấn công trên. Khuyến khích tìm hiểu các kỹ thuật tấn công mới, các lỗi SQL injection mà người dùng hay mắc phải. Xây dựng chương trình mô phỏng một số trường hợp.

Gợi ý tài liệu tham khảo:

Advanced SQL Injection In SQL Server Applications, Chris Anley, An NGSSoftware Insight Security Research (NISR) Publication.

(More) Advanced SQL Injection, Chris Anley, An NGSSoftware Insight Security Research (NISR) Publication.

The Essence of Command Injection Attacks in Web Applications, Zhendong Su, Gary Wassermann Conference record of the 33rd ACM SIGPLAN-SIGACT symposium on Principles of programming languages.

CANDID: Preventing SQL Injection Attacks using dynamic Candidate Evaluations, Sruthi Bandhakavi Prithvi Bisht P. Madhusudan V. N. Venkatakrishnan Proceedings of the 14th ACM conference on Computer and communications security.

4.4.3 Yêu cầu

Bên cạnh mặt lý thuyết của phương pháp, phân tích các trường hợp lỗi SQL injection, sinh viên nên có demo chương trình để phát hiện các lỗi SQL injection hay gặp.

4.5 Dạng 5: Biểu diễn truy vấn nâng cao

4.5.1 Đặt vấn đề

Bên cạnh việc sử dụng thành thạo các cấu trúc ngôn ngữ QBE, SQL có sẵn của các hệ CSDL; trong thực tế, cần có những phương pháp biểu diễn truy vấn mở rộng hơn để đáp ứng được nhu cầu thực hiện truy vấn phức tạp của người dùng.

Trong phần này tập trung vào các kỹ thuật biểu diễn truy vấn nâng cao nhằm đáp ứng nhu cầu người dùng, đem lại giao diện truy vấn gần gũi, dễ dùng, tự nhiên; câu truy vấn thông minh hơn và đem lại kết quả sát với ý người dùng hơn.

4.5.2 Bài tập lớn dạng 5

Đề 7: Biểu diễn và tối ưu truy vấn trong tìm kiếm xấp xỉ

Tìm hiểu phương pháp biểu diễn và tối ưu các truy vấn SQL tìm kiếm xấp xỉ. Truy vấn tìm kiếm xấp xỉ cho phép người dùng tìm ra các kết quả truy vấn có giá trị xấp xỉ giá trị truy vấn đầu vào và có độ sai số chấp nhận được. Tìm hiểu các mô hình biểu diễn truy vấn SQL xấp xỉ, đồng thời tìm hiểu các toán tử và thuật toán tối ưu đối với từng mô hình biểu diễn, phân tích và đánh giá các phương pháp.

Gợi ý tài liệu tham khảo:

Expressing and optimizing similarity-based queries in SQL, Like Gao, Min Wang, X. Sean Wang, Sriram Padmanabhan, 2006.

4.5.3 Yêu cầu

Sinh viên tìm hiểu chú trọng về mặt lý thuyết, phương pháp luận, thuật toán, so sánh các phương pháp.

Cài đặt được khuyến khích.