Bộ Giáo dục và Đào tạo

**Trường Đại Học Sài Gòn**

Khoa Công Nghệ Thông Tin

Bộ môn PPNC Khoa học trong CNTT

**0B**

ĐỀ CƯƠNG CHI TIẾT

Ngày thực hiện đề cương: 04/05/2015

|  |  |  |
| --- | --- | --- |
| **TÊN ĐỀ TÀI** | MỘT HỆ QUẢN TRỊ CSDL QUAN HỆ CHO CÁC THUỘC TÍNH CÓ GIÁ TRỊ KHOẢNG XÁC SUẤT | |
| **LĨNH VỰC CHUYÊN NGÀNH** | Khoa học máy tính | |
| **LOẠI HÌNH NGHIÊN CỨU** | Sản phẩm | |
| **NGƯỜI HƯỚNG DẪN** | GV. Đỗ Như Tài | |
| ***SINH VIÊN THỰC HIỆN 1*** | Nguyễn Đào Linh Đan | 3121410137 |

GIỚI THIỆU

|  |  |
| --- | --- |
| **CÂU HỎI GỢI Ý** | **Lĩnh vực và nghiên cứu liên quan đã và đang phát triển như thế nào? Các vấn đề, và bài toán đặt ra cần giải quyết là gì?** |
| **HƯỚNG DẪN** | Giới thiệu tổng quan về đề tài – những vấn đề và lĩnh vực liên quan đến đề tài. |
| Ngày nay, dữ liệu đã trở thành một nguồn tài nguyên quan trọng, đóng vai trò không thể thiếu trong việc vận hành, phân tích và đưa ra quyết định trong hầu hết các lĩnh vực như kinh tế, y tế, tài chính, giáo dục, và đặc biệt là công nghệ thông tin. Với sự phát triển mạnh mẽ của khoa học dữ liệu và trí tuệ nhân tạo, việc xử lý dữ liệu không chắc chắn, không đầy đủ đã và đang trở thành một nhu cầu thực tế cấp thiết trong nhiều hệ thống ứng dụng hiện đại.  Trong thời đại số, thông tin mà con người thu thập được không phải lúc nào cũng đầy đủ hoặc chính xác tuyệt đối. Các mô hình cơ sở dữ liệu quan hệ truyền thống vốn giả định dữ liệu là chắc chắn và xác định đã bộc lộ những hạn chế khi đối mặt với dữ liệu mơ hồ hoặc mang tính xác suất. Từ thực tiễn đó, các nhà nghiên cứu trong và ngoài nước đã và đang nỗ lực mở rộng mô hình dữ liệu truyền thống, tích hợp lý thuyết xác suất nhằm tạo ra những mô hình cơ sở dữ liệu mới, phù hợp hơn với yêu cầu của thời đại. Một trong những hướng đi đầy triển vọng là mô hình Cơ sở dữ liệu quan hệ xác suất (Probabilistic Relational Database - PRDB).  Trên thế giới, nhiều công trình đã đề xuất các cách tiếp cận để biểu diễn thông tin không chắc chắn như sử dụng giá trị NULL, xác suất gán cho bộ dữ liệu, hay xác suất gán cho giá trị thuộc tính. Tuy nhiên, các mô hình này vẫn còn nhiều hạn chế về mặt biểu diễn, hiệu năng tính toán hoặc tính khả thi khi áp dụng thực tế. Mới đây, một số nghiên cứu đã đề xuất mô hình sử dụng khoảng xác suất cho giá trị thuộc tính, mở ra một hướng tiếp cận mềm dẻo và linh hoạt hơn.  Tại Việt Nam, lĩnh vực này vẫn còn mới mẻ và ít được khai thác sâu, đặc biệt là trong việc hiện thực hóa các mô hình lý thuyết thành hệ thống quản trị CSDL thực tế. Điều này đặt ra một bài toán lớn: **Liệu có thể xây dựng một hệ quản trị CSDL quan hệ cho các thuộc tính có giá trị khoảng xác suất, có thể truy vấn với cú pháp gần SQL, và phù hợp để áp dụng vào các tình huống dữ liệu thực tế hay không?**  Từ đó, đề tài “Một hệ quản trị CSDL quan hệ cho các thuộc tính có giá trị khoảng xác suất (IPRDB)” ra đời, với mục tiêu không chỉ đề xuất mô hình lý thuyết mà còn hiện thực hóa thành một hệ thống cụ thể – IPRDB-SQLite, mở rộng từ SQLite mã nguồn mở. Đây là một bước tiến đáng kể trong việc tích hợp lý thuyết xác suất vào cơ sở dữ liệu quan hệ, giúp biểu diễn, truy vấn và xử lý thông tin không chắc chắn một cách hiệu quả hơn. | |

|  |
| --- |
|  |
|  |
| **GHI CHÚ** |
|  |

TÌNH HÌNH NGHIÊN CỨU TRONG VÀ NGOÀI NƯỚC

|  |  |
| --- | --- |
| **CÂU HỎI GỢI Ý** | **Lĩnh vực và nghiên cứu liên quan đã và đang phát triển như thế nào? Các vấn đề, và bài toán đặt ra cần giải quyết là gì?** |
| **HƯỚNG DẪN** | Tìm hiểu các nghiên cứu đã công bố gần nhất (5 năm trở lại) về lĩnh vực liên quan trong và ngoài nướcPhân tích các kết quả đóng góp, nhận xét các hạn chế còn tồn tại. |
| Ra đời từ những năm cuối thế kỉ 20, các mô hình cơ sở dữ liệu quan hệ xác suất (Probabilistic Relational Databases - PRDB) đã dần trở thành hướng nghiên cứu quan trọng trong lĩnh vực cơ sở dữ liệu. Bắt đầu từ các mở rộng sử dụng giá trị NULL, cho đến các mô hình tiên tiến hơn cho phép gán xác suất hoặc khoảng xác suất vào giá trị thuộc tính, các nghiên cứu trong lĩnh vực này ngày càng chú trọng hơn đến việc biểu diễn và xử lý thông tin không chắc chắn, điều vốn rất phổ biến trong dữ liệu thực tế ngày nay.  Trong khoảng 5 năm trở lại đây, nhiều công trình quốc tế đã tập trung vào việc tối ưu mô hình dữ liệu xác suất theo hai hướng chính: (1) giảm độ phức tạp của biểu diễn và tính toán phân bố xác suất, và (2) phát triển ngôn ngữ truy vấn xác suất thân thiện hơn với người dùng, gần giống SQL. Một số công trình đáng chú ý như [23] đã đề xuất mô hình sử dụng khoảng xác suất trên tập giá trị thuộc tính, giúp tăng khả năng mô hình hóa và hiệu năng truy vấn. Tuy nhiên, phần lớn các mô hình hiện nay vẫn chưa được hiện thực hóa thành hệ quản trị CSDL hoàn chỉnh, gây khó khăn trong việc đánh giá và áp dụng thực tế.  Tại Việt Nam, lĩnh vực cơ sở dữ liệu xác suất vẫn còn khá mới. Một số nghiên cứu lý thuyết đã bước đầu đề cập đến mô hình dữ liệu không chắc chắn, nhưng việc phát triển hệ quản trị thực nghiệm còn chưa được quan tâm đúng mức. Hiện chưa có hệ thống nào đủ hoàn chỉnh để cho phép người dùng xây dựng, truy vấn và thao tác dữ liệu xác suất một cách trực quan và hiệu quả.  Từ đó, bài toán đặt ra là làm sao xây dựng được một hệ quản trị cơ sở dữ liệu quan hệ cho các thuộc tính có giá trị khoảng xác suất, vừa đảm bảo đúng mô hình lý thuyết, vừa đáp ứng được yêu cầu ứng dụng thực tiễn, đồng thời hỗ trợ ngôn ngữ truy vấn gần SQL để người dùng dễ tiếp cận.  Đề tài này ra đời nhằm lấp đầy khoảng trống đó, với mục tiêu thiết kế mô hình IPRDB và phát triển hệ quản trị IPRDB-SQLite dựa trên mã nguồn mở SQLite, hỗ trợ đầy đủ các phép toán đại số quan hệ xác suất, qua đó mở rộng khả năng xử lý dữ liệu không chắc chắn trong các hệ thống thực tế như y tế, chẩn đoán bệnh, dự báo và quản lý rủi ro. | |
| **GHI CHÚ** | |
|  | |

|  |  |
| --- | --- |
| **MỤC TIÊU VÀ PHẠM VI NGHIÊN CỨU** | |
| **CÂU HỎI GỢI Ý** | **Mục tiêu nghiên cứu chính của đề tài là gì? Phạm vi nghiên cứu là gì?** |
| **HƯỚNG DẪN** | Đặt bài toán giải quyết và trình bày mục tiêu nghiên cứu chính của đề tài.  Nêu phạm vi nghiên cứu của đề tài, bao gồm việc giới hạn phạm vi nghiên cứu và triển khai, các giả định ban đầu đối với nghiên cứu. |
| Mục tiêu chính của đề tài là xây dựng một hệ quản trị cơ sở dữ liệu quan hệ IPRDB, với các chức năng trọng tâm sau:   * Thiết kế mô hình dữ liệu quan hệ mới hỗ trợ thuộc tính có giá trị khoảng xác suất (Probability Interval Valued Attributes); * Phát triển các phép toán đại số quan hệ xác suất, mở rộng từ các phép toán trong CSDL quan hệ truyền thống; * Xây dựng một hệ quản trị thực nghiệm IPRDB-SQLite dựa trên SQLite mã nguồn mở, cho phép truy vấn dữ liệu không chắc chắn bằng cú pháp gần giống SQL; * Minh họa ứng dụng của hệ thống trong một bài toán thực tế (ví dụ: quản lý bệnh án với dữ liệu y tế không chắc chắn). | |
|  | |
| **GHI CHÚ** | |
|  | |

|  |  |
| --- | --- |
| **CÁC GIẢ THUYẾT VÀ CÁCH TIẾP CẬN** | |
| **CÂU HỎI GỢI Ý** | **Các giả thuyết đặt ra để giải quyết bài toán chính? Các cách tiếp cận để giải quyết bài toán đặt ra?** |

|  |  |
| --- | --- |
| **HƯỚNG DẪN** | Đặt ra những giả thuyết, hay vấn đề-bài toán con cần phải giải quyết để đạt được mục tiêu nghiên cứu đề tài.  Các cách tiếp cận (dự kiến) để giải quyết các giả thuyết, bài toán con đã đặt. |
| **Giả thuyết nghiên cứu**  Chúng tôi đặt ra giả thuyết: Có thể xây dựng được một hệ quản trị cơ sở dữ liệu quan hệ mở rộng (IPRDB) có khả năng:   * Biểu diễn các thuộc tính có giá trị khoảng xác suất để mô hình hóa thông tin không chắc chắn trong thực tế; * Thực hiện được các phép toán đại số xác suất một cách hiệu quả và đúng đắn, tương thích với mô hình dữ liệu truyền thống; * Hỗ trợ ngôn ngữ truy vấn tương tự SQL để người dùng dễ dàng thao tác dữ liệu xác suất; * Hiện thực thành công trên nền tảng SQLite mã nguồn mở, từ đó chứng minh tính khả thi và khả năng ứng dụng của mô hình.   **Bài toán con cần giải quyết**  Từ mục tiêu trên, đề tài đặt ra ba bài toán chính:   1. Bài toán 1: Mở rộng các khái niệm trong CSDL quan hệ truyền thống (kiểu, giá trị, lược đồ, quan hệ, phụ thuộc hàm, khóa) để phù hợp với dữ liệu có tính không chắc chắn. 2. Bài toán 2: Xây dựng đầy đủ hệ phép toán đại số quan hệ xác suất, đảm bảo tính đúng đắn và tương thích với mô hình quan hệ truyền thống. 3. Bài toán 3: Thiết kế và hiện thực một hệ quản trị CSDL thực nghiệm (IPRDB-SQLite) có thể lưu trữ, truy vấn và thao tác dữ liệu xác suất.   **Các cách tiếp cận đề xuất**   1. Thiết kế mô hình dữ liệu IPRDB dựa trên cơ sở lý thuyết của xác suất và đại số quan hệ. 2. Phát triển các phép toán đại số như chọn, chiếu, kết, giao, hợp, trừ… với các chiến lược kết hợp khoảng xác suất (⊗, ⊕, ⊖). 3. Tích hợp mô hình dữ liệu IPRDB vào hệ quản trị SQLite, tạo thành hệ thống IPRDB-SQLite. 4. Xây dựng giao diện và bộ xử lý truy vấn cho người dùng, hỗ trợ cú pháp truy vấn gần giống SQL. 5. Thử nghiệm mô hình với dữ liệu thực tế, ví dụ dữ liệu y tế tại phòng khám, để kiểm chứng tính hiệu quả, trực quan và khả năng áp dụng. 6. Đánh giá và so sánh kết quả, từ đó rút ra những ưu điểm và hạn chế của hệ thống, định hướng cải tiến trong tương lai. | |
|  | |
| **GHI CHÚ** | |
|  | |

TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| **CÂU HỎI GỢI Ý** | **Các tài liệu tham khảo thuộc lĩnh vực nghiên cứu liên quan là gì?** |
| **HƯỚNG DẪN** | 1. Anderson, R.E. ***Social impacts of computing: Codes of professional ethics.*** Social Science Computing Review, 2 (Winter 1992), 453-469. 2. ACM SIG PROCEEDINGS template. [http://www.acm.org/sigs/pubs/proceed/template.html.](http://www.acm.org/sigs/pubs/proceed/template.html) 3. Conger., S., and Loch, K.D. (eds.). ***Ethics and computer use. Commun***. ACM 38, 12 (entire issue). 4. Mackay, W.E. ***Ethics, lies and videotape***. In Proceedings of CHI '95 (Denver CO, May 1995), ACM Press, 138-145. 5. Schwartz, M., and Task Force on Bias-Free Language. ***Guidelines for Bias-Free Writing***. Indiana University Press, Bloomington, IN, 1995. |
| * **Tiếng Anh**   [1] E.F. Codd, “A relational model of data for large shared data banks,” Communications of the ACM, vol.13, no.6, pp.377-387, 1970.  [2] R. Elmasri and S.B. Navathe, Fundamentals of Database Systems, 6th edition, Addison-Wesley, 2015.  [3] D. Dey and S. Sarkar, “A probabilistic relational model and algebra,” ACM Transactions on Database Systems, vol.21, no.3, pp.339-369, 1996.  [4] T. Imielinski, W. J.R. Lipski, “Incomplete information in relational databases,” Journal of the Association for Computing Machinery, vol.31, no.4, pp.761-791, 1984.  [5] D. Barbara, H. Garcia-Molina, and D. Porter, “The management of probabilistic data,” IEEE Transactions on Knowledge and Data Engineering, vol.4, no.5, pp.487-502, 1992.  [6] V.V. Kheradkar and S. K. Shirgave, “Query processing over relational cross model in uncertain and probabilistic databases,” Proceedings of 3Th International Conference on Artificial Intelligence and Smart Energy, Coimbatore, India, pp.763-769, 2023.  [7] N. Fuhr and T. Rolleke, “A probabilistic relational algebra for the integration of information retrieval and database systems,” ACM Transactions on Information Systems, vol.15, no.1, pp.32-66, 1997.  [8] S. Zhang and C. Zhang, “A probabilistic data model and its semantics,” Journal of Research and Practice in Information Technology” vol.35, no.4, pp.237-256, 2003.  [9] Z. Ma and L. Yan, Advances in probabilistic databases for uncertain information management, Springer-Verlag Berlin Heidelberg, 2013.  [10] Y. Li, J. Chen, and L. Feng, “Dealing with uncertainty: A survey of theories and practices,” IEEE Transactions on Knowledge and Data Engineering, vol. 25, no.11, pp.2463-2482, 2013.  [11] I.I. Ceylan, A. Darwiche, and G.V.D. Broeck, “Open-world probabilistic databases: Semantics, algorithms, complexity,” Journal of Artificial Intelligence,  vol.295, no.11, pp.103474-103513, 2021.  [12] H. Debbi, “Explaining query answers in probabilistic databases,” International Journal of Interactive Multimedia and Artificial Intelligence, vol.8, no.4, pp.140-152, 2023.  [13] L.V.S. Lakshmanan, N. Leone, R. Ross, and V.S. Subrahmanian, “Probview: A flexible probabilistic database system,” ACM Transactions on Database Systems, vol.22, no.3, pp.419-469, 1997.  [14] H. Nguyen, “Extending relational database model for uncertain information,” Journal of Computer Science and Cybernetics, vol.35, no.4, pp.355-372, 2019.  [15] W. Zhao, A. Dekhtyar, and J. Goldsmith, “Databases for interval probabilities,” International Journal of Intelligent Systems, vol.19, no.9, pp.789-815, 2004.  [16] R. Ross and V.S. Subrahmanian, “Aggregate operators in probabilistic databases,” Journal of the ACM, vol.52, no.1, pp.54-101, 2005.  [17] D. Dey and S. Sarkar, “Generalized normal forms for probabilistic relational data,” IEEE Transactions on Knowledge and Data Engineering, vol.14, no.3, pp.485-497, 1992.  [18] T. Eiter, T. Lukasiewicz, and M. Walter, “A data model and algebra for probabilistic complex values,” Annals of Mathematics and Artificial Intelligence, vol.33, pp.205-252, 2001.  [19] S.K. Lee, “An extended relational database model for uncertain and imprecise information,” Proceedings of 18th Conference on Very Large Data Bases, Vancouver, Canada, pp.211-220, 1992.  [20] H. Nguyen, “A probabilistic relational database model and algebra,” Journal of Computer Science and Cybernetics, vol.31, no.4, pp.305-321, 2015.  [21] H. Nguyen, T.N. Nguyen, and T.T.N. Tran, “A probabilistic relational database model with uncertain multivalued attributes,” ICIC Express Letters, vol. 16, no.3, pp.241-248, 2022.  [22] H. Nguyen, “Extending probabilistic relational database model with uncertain multivalued attributes,” International Journal of Innovative Computing, Information and Control, vol.18, no.5, pp.1477–1492, 2022.  [23] H. Nguyen, D.N. Le, 2024, “A Relational Database Model with Interval Probability Valued Attributes for Uncertain and Imprecise Information,” ECTI  Transactions on Computer and Information Technology, vol.18, no.3, pp.307- 318.  [24] SQLite, “SQLite Home Page,” http://www.sqlite.org, accessed Jan. 10, 2025. | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | LCMS nguồn mở Moodle.   * **Trang web**   [39] Trang chủ Moodle: [http://moodle.org](http://moodle.org/) | |  | | **GHI CHÚ** | |  |  |  |  |  | | --- | --- | --- | | **NGƯỜI THỰC HIỆN ĐỀ CƯƠNG** | | | | **SINH VIÊN THỰC HIỆN 1** | | Nguyễn Đào Linh Đan | 3121410137 |  |  |  |  | | --- | --- | --- | | **XÁC NHẬN CỦA GIÁO VIÊN HƯỚNG DẪN** | | | | **NGƯỜI HƯỚNG DẪN** | GV. Đỗ Như Tài | | | **Ý KIẾN** |  | | |  | * đồng ý hướng dẫn | * không đồng ý hướng dẫn | |