**ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**Nguyễn Đào Linh Đan**

**MỘT HỆ QUẢN TRỊ CSDL QUAN HỆ CHO CÁC THUỘC TÍNH CÓ GIÁ TRỊ KHOẢNG XÁC SUẤT**

**ĐỀ CƯƠNG LUẬN VĂN CỬ NHÂN**

NGÀNH: KHOA HỌC MÁY TÍNH

**Thành phố Hồ Chí Minh, năm 2025**

**ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**Nguyễn Đào Linh Đan**

**MỘT HỆ QUẢN TRỊ CSDL QUAN HỆ CHO CÁC THUỘC TÍNH CÓ GIÁ TRỊ KHOẢNG XÁC SUẤT**

**ĐỀ CƯƠNG LUẬN VĂN cử nhân**

ngành: KHOA HỌC MÁY TÍNH

Mã số NGÀNH: 841324

**Người hướng dẫn**

**Đỗ Như Tài**

**Thành phố Hồ Chí Minh, năm 2025**

# ĐỀ CƯƠNG LUẬN VĂN

## **1. Lý do chọn đề tài**

Trong bối cảnh hiện nay, dữ liệu đã trở thành tài nguyên quan trọng trong mọi lĩnh vực như kinh tế, y tế, tài chính, giáo dục và đặc biệt là Công nghệ Thông tin. Tuy nhiên, thông tin thu thập được từ thực tế thường không đầy đủ, thiếu chính xác hoặc mang tính xác suất. Mô hình cơ sở dữ liệu quan hệ truyền thống (đề xuất bởi Codd năm 1970) giả định rằng dữ liệu được mô tả một cách chắc chắn và xác định, nên bộc lộ hạn chế khi phải xử lý các trường hợp dữ liệu mờ hoặc không chắc chắn. Ví dụ, các hệ CSDL quan hệ hiện có không thể trả lời các truy vấn như “tìm các bệnh nhân có khả năng từ 80–90% bị bệnh X hoặc Y” hay “tìm các gói hàng có xác suất ít nhất 90% được chuyển đến trong 48 giờ”. Những tình huống này đòi hỏi một mô hình dữ liệu có khả năng biểu diễn và xử lý độ không chắc chắn vốn tồn tại trong thông tin thực tế. Trên thế giới, đã có nhiều cách tiếp cận để giải quyết dữ liệu không chắc chắn trong CSDL quan hệ. Một số mô hình sử dụng giá trị NULL để biểu diễn thông tin chưa biết (như trong nghiên cứu của Imielinski và Lipski), một số gán mức xác suất cho mỗi bộ (tuple-level) hoặc cho từng giá trị thuộc tính (attribute-level) trong quan hệ. Tuy nhiên, các phương pháp này đều có hạn chế riêng về khả năng biểu diễn hoặc độ phức tạp tính toán. Gần đây, một hướng tiếp cận mềm dẻo hơn được đề xuất là gán một khoảng xác suất cho giá trị thuộc tính, tức là cho mỗi thuộc tính có một khoảng giá trị xác suất thay vì một giá trị cố định. Việc sử dụng khoảng xác suất giúp mô hình linh hoạt hơn trong việc biểu diễn mức độ không chắc chắn của thông tin. Tại Việt Nam, lĩnh vực CSDL xác suất vẫn còn rất mới và chưa có nhiều công trình hiện thực hóa lý thuyết thành một hệ quản trị cụ thể. Điều này đặt ra câu hỏi nghiên cứu: Liệu có thể xây dựng một hệ quản trị cơ sở dữ liệu quan hệ mở rộng, hỗ trợ các thuộc tính giá trị khoảng xác suất, có thể được truy vấn bằng cú pháp SQL tương tự, và áp dụng hiệu quả vào các tình huống dữ liệu thực tế? Từ đó, đề tài “Một hệ quản trị CSDL quan hệ cho các thuộc tính có giá trị khoảng xác suất (IPRDB)” ra đời. Mục tiêu chính của đề tài là đề xuất một mô hình lý thuyết IPRDB và hiện thực thành một hệ thống cụ thể (IPRDB-SQLite), mở rộng từ SQLite mã nguồn mở. Công trình này hứa hẹn là bước tiến quan trọng trong việc tích hợp lý thuyết xác suất vào CSDL quan hệ, nhằm biểu diễn, truy vấn và xử lý thông tin không chắc chắn một cách hiệu quả hơn so với các mô hình hiện tại.

## **2. Tổng quan vấn đề nghiên cứu**

## **2.1. Tình hình nghiên cứu hiện tại**

Mô hình cơ sở dữ liệu quan hệ (CSDL QH) truyền thống, được đề xuất bởi E.F. Codd năm 1970, đã chứng minh được hiệu quả trong việc mô hình hóa, thiết kế và triển khai các hệ thống thông tin lớn. Tuy nhiên, hạn chế lớn của mô hình này là chỉ xử lý được dữ liệu chắc chắn – tức dữ liệu phải xác định, chính xác và đầy đủ.

Trong thực tế, thông tin thu thập được thường mang tính không chắc chắn, không đầy đủ hoặc mơ hồ, ví dụ như:

* Xác suất bệnh nhân bị viêm gan trong khoảng 80–90%;
* Xác suất gói hàng đến đúng hạn là trên 90%;
* Khả năng một người phù hợp với một điều kiện nào đó trong khoảng xác suất nhất định.

Các mô hình CSDL truyền thống, kể cả khi mở rộng với giá trị NULL, không thể biểu diễn và xử lý được các loại thông tin này một cách chính xác và đầy đủ.

Để khắc phục, nhiều hướng nghiên cứu đã được đề xuất trên thế giới:

* Mô hình sử dụng giá trị NULL (như của Imielinski & Lipski) để biểu diễn thông tin chưa biết, nhưng không thể phản ánh được mức độ tin cậy.
* Mô hình xác suất mức bộ (tuple-level): Mỗi bộ dữ liệu được gán một giá trị xác suất, thể hiện độ tin cậy của toàn bộ bản ghi (ví dụ: Dey & Sarkar, Barbara et al.). Tuy nhiên, không chi tiết đến từng thuộc tính.
* Mô hình xác suất mức thuộc tính (attribute-level): Mỗi giá trị thuộc tính được gán một xác suất duy nhất, nhưng không phản ánh được sự dao động hoặc không rõ ràng trong xác suất đó.
* Mô hình khoảng xác suất (interval-based): Các giá trị thuộc tính được kết hợp với một khoảng xác suất [l, u], cho phép biểu diễn cả sự không chắc chắn về giá trị và về mức độ tin cậy. Hướng tiếp cận này mềm dẻo hơn, gần với thực tiễn, và được thể hiện trong các nghiên cứu gần đây như [19], [23].

Tuy nhiên, phần lớn các nghiên cứu nêu trên chỉ tập trung vào xây dựng mô hình lý thuyết mà chưa có hệ quản trị CSDL cụ thể và hiệu quả để triển khai, đặc biệt trong môi trường ứng dụng thực tế.

## **2.2. Hướng tiếp cận của đề tài**

Trên cơ sở các hạn chế của các mô hình và hệ thống trước đó, đề tài lựa chọn tiếp cận theo hướng sau:

* Mở rộng mô hình CSDL quan hệ truyền thống để tạo ra mô hình IPRDB (Interval Probabilistic Relational Database), cho phép biểu diễn giá trị thuộc tính dưới dạng khoảng xác suất, thay vì một xác suất cố định.
* Phát triển các phép toán đại số quan hệ mở rộng cho IPRDB: chọn, chiếu, kết, giao, hợp, trừ,… phù hợp với dữ liệu có tính không chắc chắn.
* Hiện thực hóa mô hình lý thuyết thành hệ quản trị CSDL cụ thể, đặt tên là IPRDB-SQLite, mở rộng từ hệ quản trị mã nguồn mở SQLite.
* Hỗ trợ truy vấn bằng cú pháp gần giống SQL truyền thống, giúp người dùng dễ dàng tiếp cận và thao tác mà không cần học một ngôn ngữ mới.
* Kiểm chứng bằng thực nghiệm trên dữ liệu mô phỏng, như hồ sơ bệnh nhân với chẩn đoán xác suất, để đánh giá tính khả thi và hiệu quả ứng dụng của hệ thống.

Tóm lại, hướng tiếp cận của đề tài là kết hợp giữa lý thuyết và thực tiễn: vừa phát triển mô hình dữ liệu mới, vừa hiện thực thành một hệ thống cụ thể có thể áp dụng vào các bài toán thực tế liên quan đến thông tin không chắc chắn.

## **3. Mục đích và nhiệm vụ nghiên cứu**

**Mục đích nghiên cứu:** Đề tài nhằm đề xuất và hiện thực một mô hình cơ sở dữ liệu quan hệ mở rộng hỗ trợ giá trị thuộc tính dạng khoảng xác suất (IPRDB), từ đó nâng cao khả năng biểu diễn và truy vấn thông tin không chắc chắn so với các mô hình truyền thống. Cụ thể, nghiên cứu hướng đến việc xây dựng nền tảng lý thuyết cho IPRDB, thiết kế kiến trúc cho một hệ quản trị tương ứng (IPRDB-SQLite), và kiểm chứng tính khả thi của mô hình thông qua việc phát triển và thử nghiệm hệ thống.

**Nhiệm vụ nghiên cứu:** Để đạt được mục tiêu trên, các nhiệm vụ chính bao gồm:

* Nghiên cứu tổng quan và phân tích lý thuyết: Thu thập và tổng hợp kiến thức về mô hình cơ sở dữ liệu quan hệ và các nghiên cứu CSDL xác suất. Đánh giá ưu nhược điểm của các mô hình hiện có (tuple-level, attribute-level, interval-level) để làm cơ sở chọn lựa và phát triển mô hình IPRDB.
* Đề xuất mô hình IPRDB: Xây dựng khái niệm, định nghĩa chính thức cho mô hình dữ liệu IPRDB, bao gồm việc mở rộng khái niệm quan hệ, thuộc tính và giá trị dữ liệu sang dạng khoảng xác suất. Thiết kế ngữ nghĩa toán học cho các giá trị khoảng và mối quan hệ xác suất.
* Xây dựng các phép toán đại số IPRDB: Phát triển các phép toán quan hệ mở rộng (chọn, chiếu, kết tự nhiên, giao, hợp, trừ, v.v.) cho mô hình IPRDB, đồng thời phân tích và chứng minh các tính chất đại số cơ bản của chúng.
* Thiết kế kiến trúc hệ quản trị IPRDB-SQLite: Lựa chọn công nghệ (SQLite) và thiết kế mở rộng để hỗ trợ kiểu dữ liệu khoảng xác suất. Xác định cấu trúc cơ sở dữ liệu, các đối tượng, và ngôn ngữ truy vấn SQL-sửa đổi (nếu cần) để thao tác với dữ liệu xác suất.
* Hiện thực hệ thống: Xây dựng một hệ quản trị CSDL IPRDB-SQLite dựa trên SQLite mã nguồn mở. Triển khai mã để lưu trữ, truy vấn và xử lý quan hệ IPRDB. Cài đặt các thao tác SQL tương ứng với các phép toán mở rộng đã đề xuất.
* Thử nghiệm và đánh giá: Chuẩn bị kịch bản thử nghiệm minh họa (có thể sử dụng tập dữ liệu mô phỏng) để kiểm nghiệm hệ thống. Thực hiện các truy vấn mẫu trên IPRDB-SQLite và so sánh kết quả với CSDL quan hệ truyền thống. Đánh giá hiệu quả và khả năng áp dụng của mô hình dựa trên kết quả thử nghiệm.
* Viết báo cáo và hoàn thiện luận văn: Tổng hợp kết quả nghiên cứu, phân tích, kết luận, và đề xuất hướng phát triển tiếp theo cho IPRDB và các ứng dụng CSDL xác suất.**4. Đối tượng và phạm vi nghiên cứu**

**4. Đối tượng và phạm vi nghiên cứu**

**Đối tượng nghiên cứu:** Đề tài tập trung vào mô hình và hệ quản trị cơ sở dữ liệu quan hệ mở rộng cho phép các thuộc tính có giá trị khoảng xác suất. Cụ thể, đối tượng nghiên cứu bao gồm:

* Mô hình dữ liệu IPRDB với các khái niệm quan hệ, thuộc tính, giá trị ở dạng khoảng xác suất.
* Kiến trúc và thành phần của hệ quản trị CSDL IPRDB (IPRDB-SQLite) mở rộng từ SQLite để hỗ trợ kiểu dữ liệu mới này.
* Các truy vấn trên IPRDB, bao gồm ngôn ngữ truy vấn SQL hoặc biến thể của nó cho dữ liệu xác suất.

**Phạm vi nghiên cứu:** Nội dung đề tài được giới hạn trong những khía cạnh sau:

* Mô hình dữ liệu: Chỉ xét mở rộng trong khuôn khổ CSDL quan hệ, xử lý thông tin không chắc chắn dưới dạng xác suất. Không nghiên cứu các dạng biểu diễn không chắc chắn khác như dữ liệu mờ (fuzzy) hay dữ liệu theo phân phối phức tạp.
* Kiến trúc hệ thống: Tập trung vào việc mở rộng một hệ quản trị CSDL mã nguồn mở (SQLite) để hỗ trợ IPRDB, không nghiên cứu các hệ quản trị khác.
* Ngôn ngữ truy vấn: Sử dụng cú pháp SQL mở rộng với khả năng thao tác trên giá trị khoảng xác suất; không xây dựng một ngôn ngữ truy vấn hoàn toàn mới.
* Dữ liệu thử nghiệm: Sử dụng dữ liệu giả lập hoặc mẫu để minh họa chức năng của hệ thống; không tập trung vào việc thu thập dữ liệu thực tế quy mô lớn.
* Mục tiêu ứng dụng: Chỉ nghiên cứu tính khả thi về mặt lý thuyết và kỹ thuật của IPRDB. Các ứng dụng cụ thể trong nghiệp vụ sâu hơn (như kho dữ liệu kinh doanh, y tế, v.v.) sẽ được đề cập ở góc độ tiềm năng phát triển nhưng không nằm trong phạm vi phát triển hệ thống chính.

## **5. Phương pháp nghiên cứu**

Đề tài được thực hiện theo hướng kết hợp chặt chẽ giữa nghiên cứu lý thuyết và thực nghiệm phát triển hệ thống. Cụ thể, các phương pháp chính bao gồm:

* Phương pháp nghiên cứu tài liệu (tổng quan nghiên cứu): Thu thập, phân tích các bài báo, sách chuyên khảo và các công trình liên quan đến cơ sở dữ liệu quan hệ xác suất, các mô hình hiện có (NULL, probabilistic, interval), cũng như các giải pháp hiện thực hóa tương tự. Đánh giá và tổng hợp các kết quả để làm nền tảng lý thuyết cho đề tài.
* Phương pháp phân tích và tổng hợp: Trên cơ sở lý thuyết thu thập được, thực hiện phân tích các yêu cầu của mô hình IPRDB. Đề xuất và định nghĩa chính thức mô hình dữ liệu, các phép toán đại số mở rộng. Sử dụng phương pháp logic toán học để chứng minh tính nhất quán và tính chất của phép toán.
* Phương pháp thiết kế hệ thống: Áp dụng phương pháp thiết kế phần mềm có hướng đối tượng hoặc có cấu trúc để xây dựng kiến trúc IPRDB-SQLite. Xác định các lớp, đối tượng và cấu trúc dữ liệu cần thiết cho việc lưu trữ và xử lý dữ liệu khoảng xác suất.
* Phương pháp thực nghiệm (thực hành/triển khai): Triển khai mã nguồn của hệ quản trị dựa trên SQLite, bao gồm việc định nghĩa kiểu dữ liệu mới cho khoảng xác suất, mở rộng trình phân tích cú pháp SQL, và cài đặt các phép toán mở rộng. Thực hiện các bài toán thử nghiệm cụ thể để kiểm tra tính đúng đắn và hiệu suất hệ thống.
* Phương pháp đánh giá (định lượng và định tính): Sử dụng các chỉ số đánh giá như độ chính xác của kết quả truy vấn hay thời gian thực thi để đánh giá hiệu quả của hệ thống so với mô hình chuẩn. Kết hợp với phân tích định tính về tính ưu việt của mô hình IPRDB trong việc biểu diễn thông tin không chắc chắn.
* Phương pháp chuyên gia: Trao đổi, thảo luận với người hướng dẫn và chuyên gia trong lĩnh vực CSDL để được hướng dẫn, nhận xét và điều chỉnh hướng nghiên cứu cũng như giải pháp kỹ thuật cho phù hợp.

Kết hợp các phương pháp trên sẽ đảm bảo quá trình nghiên cứu diễn ra có hệ thống, logic và đạt được kết quả khoa học cũng như ứng dụng thực tiễn cao.

## **6. Giả thuyết khoa học**

Đề tài này xây dựng trên các giả thuyết khoa học chính sau:

* Giả thuyết 1: Mở rộng mô hình CSDL quan hệ để cho phép các thuộc tính có giá trị khoảng xác suất (IPRDB) sẽ nâng cao khả năng biểu diễn thông tin không chắc chắn so với mô hình truyền thống. Cụ thể, dữ liệu không chắc chắn và thiếu thông tin có thể được biểu diễn chính xác hơn dưới dạng các khoảng xác suất, giúp truy vấn và xử lý linh hoạt hơn.
* Giả thuyết 2: Các phép toán đại số quan hệ mở rộng cho IPRDB có thể bảo toàn tính chất cơ bản và tính nhất quán của phép toán CSDL quan hệ tiêu chuẩn. Nói cách khác, khi chuyển đổi các phép toán như chọn, chiếu, kết, giao, hợp… sang IPRDB, phép toán mới vẫn cho kết quả đúng đắn về mặt xác suất khoảng và duy trì các tính chất như giao hoán, kết hợp (nếu có) như mô hình cơ sở.
* Giả thuyết 3: Hệ quản trị IPRDB-SQLite được triển khai sẽ chứng minh được tính khả thi và hiệu quả thực tiễn của mô hình IPRDB. Cụ thể, IPRDB-SQLite có thể thực thi các truy vấn chứa dữ liệu khoảng xác suất một cách hiệu quả, gần tương đương với hệ SQLite gốc về mặt hiệu năng xử lý, đồng thời hỗ trợ mở rộng cú pháp SQL để thao tác với dữ liệu xác suất.

Những giả thuyết trên sẽ được kiểm chứng thông qua việc xây dựng mô hình và triển khai hệ thống, sau đó đánh giá kết quả thực nghiệm.

## **7. Những đóng góp mới của đề tài**

Đề tài "IPRDB" hướng tới các đóng góp mới quan trọng sau:

* Mô hình CSDL IPRDB: Đề xuất một mô hình cơ sở dữ liệu quan hệ mở rộng cho các thuộc tính giá trị khoảng xác suất, với định nghĩa toán học đầy đủ cho khái niệm quan hệ và giá trị khoảng xác suất. Mô hình này cho phép tích hợp lý thuyết xác suất vào CSDL quan hệ một cách trực quan và nhất quán.
* Các phép toán đại số mở rộng: Thiết kế và xây dựng các phép toán đại số quan hệ trên IPRDB, bao gồm phép chọn (σ), chiếu (π), kết tự nhiên (⋈), giao, hợp, trừ… Duy trì hoặc chứng minh lại các tính chất cơ bản của đại số quan hệ trong ngữ cảnh dữ liệu xác suất. Đây là nền tảng quan trọng để thực hiện truy vấn trên dữ liệu IPRDB.
* Hệ quản trị IPRDB-SQLite: Thiết kế và hiện thực một hệ quản trị CSDL mở rộng trên nền SQLite mã nguồn mở. Hệ quản trị này cho phép lưu trữ và xử lý trực tiếp các giá trị khoảng xác suất. Việc lựa chọn SQLite giúp đảm bảo tính cơ bản của hệ thống và khả năng mở rộng, đồng thời cung cấp một cơ sở để nghiên cứu về tích hợp xác suất vào ngôn ngữ SQL.
* Hỗ trợ truy vấn SQL cho IPRDB: Mở rộng cú pháp và các thủ tục xử lý SQL để tương thích với dữ liệu IPRDB. Cụ thể, cho phép người dùng viết các truy vấn gần giống SQL quen thuộc nhưng có thể thao tác với các thuộc tính khoảng xác suất, ví dụ như các điều kiện truy vấn về khoảng xác suất.
* Kiểm chứng thực nghiệm: Thực hiện các thử nghiệm và đánh giá hệ thống với các trường hợp minh họa. Kết quả cho thấy IPRDB-SQLite có khả năng xử lý truy vấn xác suất, so sánh được mức độ chính xác và hiệu quả so với các hệ truyền thống. Qua đó chứng minh tính khả thi và lợi ích của đề tài trong việc áp dụng vào các bài toán thực tế liên quan đến dữ liệu không chắc chắn.
* Đề xuất hướng phát triển: Nghiên cứu cung cấp một hướng tiếp tục mở rộng cho các nghiên cứu CSDL xác suất trong tương lai. Ví dụ, có thể phát triển thêm các phương pháp tối ưu hóa truy vấn trong IPRDB, mở rộng sang các hệ quản trị khác, hoặc ứng dụng mô hình IPRDB vào các lĩnh vực cụ thể (y tế, logistic, v.v.).

Những đóng góp trên hứa hẹn thúc đẩy việc nghiên cứu và ứng dụng các mô hình CSDL quan hệ xác suất, đồng thời đáp ứng nhu cầu thực tiễn trong xử lý dữ liệu không chắc chắn ngày càng gia tăng.

## **8. Dự kiến kế hoạch nghiên cứu**

Bảng 8.1 Dự kiến kế hoạch nghiên cứu

|  |  |  |
| --- | --- | --- |
| **STT** | **Nội dung** | **Dự kiến thời gian thực hiện** |
| 1 | Thu thập tài liệu liên quan về cơ sở dữ liệu quan hệ xác suất và các mô hình thông tin không chắc chắn. Xác định yêu cầu, phạm vi nghiên cứu và đề cương chi tiết cho luận văn. | 1 tháng |
| 3 | Xây dựng mô hình lý thuyết IPRDB: Định nghĩa khái niệm cơ bản, các kiểu dữ liệu, lược đồ quan hệ và giá trị khoảng xác suất. Thiết kế sơ bộ các phép toán đại số mở rộng cho IPRDB. | 1 tháng |
| 4 | Hiện thực hệ quản trị IPRDB-SQLite: Mã hóa các lớp và module xử lý giá trị xác suất, tích hợp vào SQLite. Xây dựng ngôn ngữ truy vấn SQL mở rộng nếu cần. Tiến hành kiểm thử ban đầu (unit test) cho các phép toán cơ bản. | 1 tháng |
| 5 | Thử nghiệm hệ thống với các trường hợp mẫu: Chuẩn bị dữ liệu thử nghiệm minh họa, thực hiện các truy vấn IPRDB mẫu. Đánh giá kết quả truy vấn và hiệu năng. Điều chỉnh, tối ưu hệ thống dựa trên phản hồi. | 1 tháng |
| 6 | Tổng hợp kết quả, viết báo cáo, hoàn thiện luận văn. So sánh và phân tích kết quả thu được. Chuẩn bị các phần trình bày, bảo vệ luận văn nếu cần. Nộp bản hoàn chỉnh cho hội đồng đánh giá. | 1 tháng |

## **9. Dự kiến nội dung của luận văn**

**Chương 1:** Giới thiệu

* Lý do chọn đề tài và tính cấp thiết của việc xử lý thông tin không chắc chắn trong CSDL quan hệ.
* Mục tiêu và nhiệm vụ nghiên cứu của đề tài.
* Đối tượng và phạm vi nghiên cứu.
* Phương pháp nghiên cứu chính.
* Những đóng góp mới của đề tài.
* Cấu trúc của luận văn.

**Chương 2:** Cơ sở lý thuyết và các công trình liên quan

* Tổng quan về mô hình CSDL quan hệ truyền thống.
* Các mô hình cơ sở dữ liệu quan hệ xử lý thông tin không chắc chắn: NULL, probabilistic (tuple-level và attribute-level).
* Mô hình thuộc tính giá trị khoảng xác suất (Interval probabilistic model) và các nghiên cứu liên quan (ví dụ Zhao et al. 2004).
* Các cơ sở toán học phục vụ cho việc xử lý xác suất trong CSDL.
* Các phép toán đại số quan hệ cơ bản và yêu cầu mở rộng trong bối cảnh dữ liệu xác suất.

**Chương 3:** Mô hình dữ liệu IPRDB

* Định nghĩa khái niệm quan hệ mở rộng, lược đồ và các kiểu giá trị trong IPRDB.
* Cách biểu diễn giá trị thuộc tính dưới dạng khoảng xác suất.
* Cấu trúc của lược đồ quan hệ trong IPRDB (ProbSchema, ProbTable, v.v.).
* Ví dụ minh họa cấu trúc dữ liệu và quan hệ mẫu trong IPRDB.

**Chương 4:** Các phép toán đại số trên IPRDB

* Phép chọn (selection) và chiếu (projection) mở rộng trong IPRDB.
* Phép kết tự nhiên (join), phép tích Descartes mở rộng với dữ liệu xác suất.
* Các phép giao, hợp và trừ trong IPRDB.
* Các tính chất đại số của các phép toán (giao hoán, kết hợp, phân phối nếu có) và chứng minh tính đúng đắn.
* Ví dụ và phân tích kết quả của các phép toán trên dữ liệu IPRDB mẫu.

**Chương 5:** Thiết kế và hiện thực hệ quản trị IPRDB-SQLite

* Giới thiệu tổng quan về hệ quản trị SQLite và các khả năng mở rộng.
* Kiến trúc hệ thống IPRDB-SQLite: thành phần chính, luồng xử lý (từ nhận truy vấn SQL đến trả kết quả).
* Chi tiết thiết kế cơ sở dữ liệu: cách lưu trữ quan hệ xác suất, cấu trúc dữ liệu cho khoảng xác suất.
* Mở rộng ngôn ngữ SQL: cú pháp hỗ trợ giá trị khoảng xác suất và các hàm xử lý liên quan.
* Ví dụ về truy vấn thực thi trên IPRDB-SQLite và kết quả mẫu.
* (Nếu cần thiết) Kết quả thử nghiệm hiệu suất hệ thống: đo đạc thời gian thực thi, đánh giá khả năng mở rộng.
* Hạn chế của giải pháp hiện tại và đề xuất hướng phát triển.

**Chương 6:** Kết luận và hướng phát triển

* Tóm tắt các kết quả đạt được: mô hình IPRDB, hệ quản trị IPRDB-SQLite và các thử nghiệm mẫu.
* Đánh giá hiệu quả của mô hình và hệ thống trong việc xử lý thông tin không chắc chắn.
* Hạn chế của nghiên cứu: ví dụ như phạm vi thử nghiệm, hiệu năng, hoặc những giả định lý thuyết.
* Đề xuất hướng nghiên cứu tiếp theo: mở rộng sang các hệ quản trị khác, tối ưu hóa truy vấn, ứng dụng vào dữ liệu thực tế, v.v..

## **10. Danh mục tài liệu tham khảo**

[1] E. F. Codd, “A relational model of data for large shared data banks,” *Commun. 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 Trans. Database Syst.*, vol. 21, no. 3, pp. 339–369, 1996.  
[4] T. Imielinski and W.J.R. Lipski, “Incomplete information in relational databases,” *J. ACM*, vol. 31, no. 4, pp. 761–791, 1984.  
[5] D. Barbara, H. Garcia-Molina, and D. Porter, “The management of probabilistic data,” *IEEE Trans. Knowl. Data Eng.*, 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,” *Proc. 3rd Int. Conf. Artificial Intelligence and Smart Energy*, Coimbatore, India, pp. 763–769, 2023.  
[7] N. Fuhr and T. Rölleke, “A probabilistic relational algebra for the integration of information retrieval and database systems,” *ACM Trans. Inf. Syst.*, vol. 15, no. 1, pp. 32–66, 1997.  
[8] S. Zhang and C. Zhang, “A probabilistic data model and its semantics,” *J. Res. Pract. Inf. Technol.*, 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 Trans. Knowl. Data Eng.*, 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,” *J. Artif. Intell.*, vol. 295, no. 11, pp. 103474–103513, 2021.  
[12] H. Debbi, “Explaining query answers in probabilistic databases,” *Int. J. Interact. Multimedia Artif. Intell.*, 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 Trans. Database Syst.*, vol. 22, no. 3, pp. 419–469, 1997.  
[14] H. Nguyen, “Extending relational database model for uncertain information,” *J. Comput. Sci. Cybern.*, vol. 35, no. 4, pp. 355–372, 2019.  
[15] W. Zhao, A. Dekhtyar, and J. Goldsmith, “Databases for interval probabilities,” *Int. J. Intell. Syst.*, vol. 19, no. 9, pp. 789–815, 2004.  
[16] R. Ross and V.S. Subrahmanian, “Aggregate operators in probabilistic databases,” *J. ACM*, vol. 52, no. 1, pp. 54–101, 2005.  
[17] D. Dey and S. Sarkar, “Generalized normal forms for probabilistic relational data,” *IEEE Trans. Knowl. Data Eng.*, 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,” *Ann. Math. Artif. Intell.*, vol. 33, pp. 205–252, 2001.  
[19] S.K. Lee, “An extended relational database model for uncertain and imprecise information,” *Proc. 18th Int. Conf. Very Large Data Bases*, Vancouver, Canada, pp. 211–220, 1992.  
[20] H. Nguyen, “A probabilistic relational database model and algebra,” *J. Comput. Sci. Cybern.*, vol. 31, no. 4, pp. 305–321, 2015.