ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

CHƯƠNG 9
Hệ thống khai phá luật kết hợp cho doanh nghiệp

Biên soạn: ThS. Nguyễn Thị Anh Thư

Nội dung

1. Giới thiệu
2. Quy trình triển khai
3. Lựa chọn thuật toán
4. Đánh giá hiệu quả
5. Triển khai hệ thống
6. Giám sát và bảo trì
7. Tổng kết

1. Giới thiệu

Hệ thống khai phá luật kết hợp (Association Rule Mining - ARM) là một công cụ phân tích dữ liệu mạnh mẽ giúp tổ chức hoặc doanh nghiệp khám phá các mối quan hệ ẩn giấu trong dữ liệu.

Được sử dụng trong nhiều lĩnh vực khác nhau.

Luật kết hợp (Association Rule)

Học không giám sát (Unsupervised Learning)

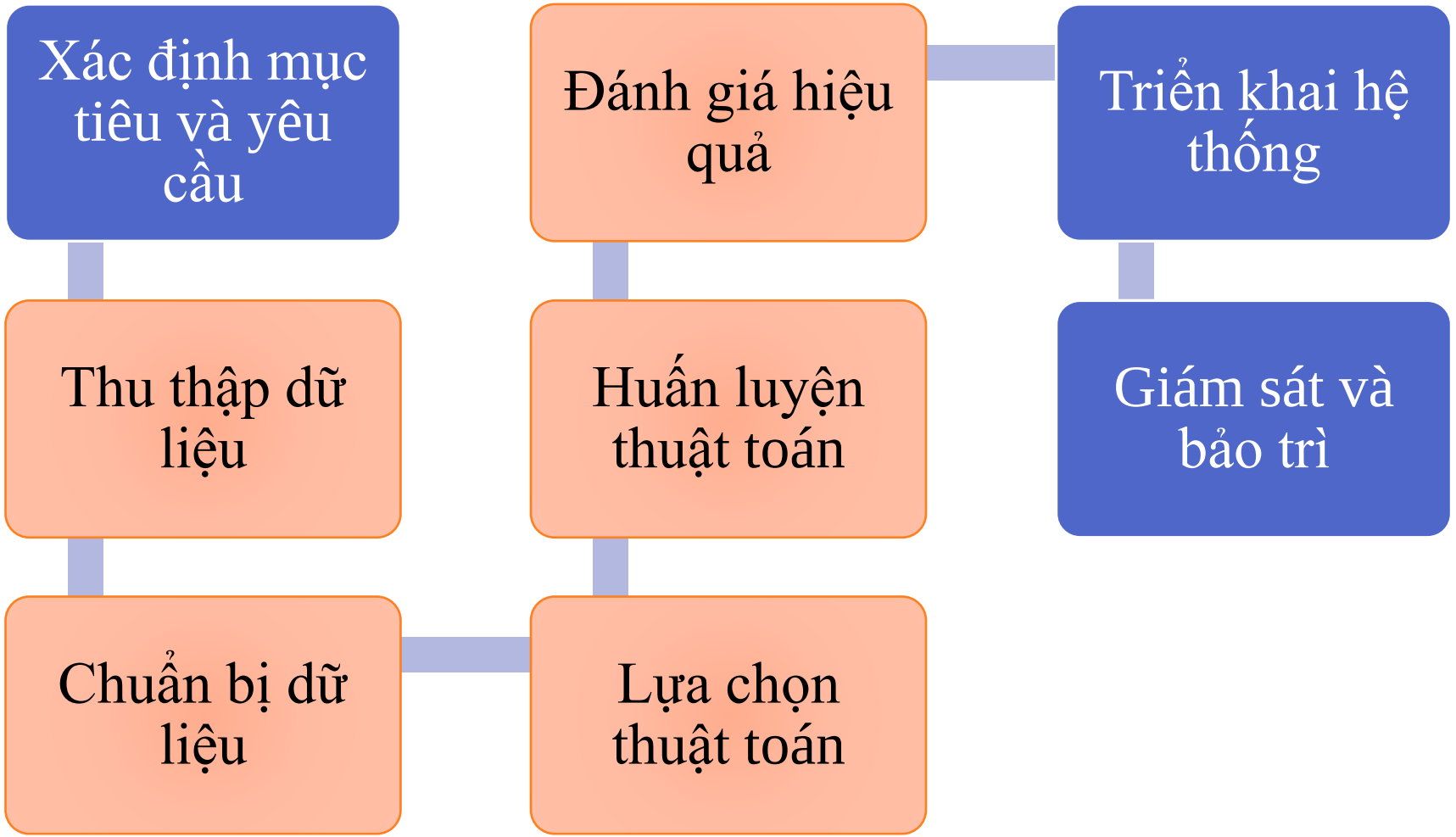
Luật kết hợp (Association Rule)

Tìm kiếm các mối quan hệ tương quan giữa các mục (item) trong tập dữ liệu.

Mục đích:

- Khám phá các quy tắc ẩn trong dữ liệu.
- Xác định các tập mục thường xuyên xuất hiện cùng nhau.
- Tìm kiếm các cơ hội bán hàng chéo (cross-selling) và tiếp thị (marketing).
- Phân tích hành vi khách hàng.

2. Quy trình triển khai



2. Quy trình triển khai

1. Xác định mục tiêu và yêu cầu:

- Xác định rõ ràng mục tiêu của việc khai phá luật kết hợp, ví dụ như tìm kiếm các mẫu quan hệ ẩn trong dữ liệu, dự đoán xu hướng tương lai, v.v.
- Xác định các yêu cầu về hiệu suất, độ chính xác và tính có thể giải thích của hệ thống.

2. Thu thập dữ liệu:

- Thu thập dữ liệu phù hợp với mục tiêu và yêu cầu đã xác định.
- Đảm bảo chất lượng dữ liệu bằng cách loại bỏ các giá trị thiếu, nhiễu và sai sót.

2. Quy trình triển khai

3. Chuẩn bị dữ liệu:

- Chuyển đổi dữ liệu sang định dạng phù hợp với thuật toán khai phá luật kết hợp.
- Chia dữ liệu thành tập huấn luyện, tập kiểm tra và tập xác thực.

4. Lựa chọn thuật toán:

- Lựa chọn thuật toán khai phá luật kết hợp phù hợp với loại dữ liệu và mục tiêu khai phá.
- Một số thuật toán phổ biến bao gồm: Apriori, FP-Growth, CARMA, v.v.

5. Huấn luyện thuật toán:

- Huấn luyện thuật toán khai phá luật kết hợp trên tập dữ liệu huấn luyện.
- Điều chỉnh các tham số của thuật toán để đạt được hiệu quả tốt nhất.

2. Quy trình triển khai

6. Đánh giá hiệu quả:

- Đánh giá hiệu quả của hệ thống khai phá luật kết hợp trên tập kiểm tra.
- Sử dụng các chỉ số hiệu suất như tỷ lệ hỗ trợ, độ tin cậy, độ nâng cao, v.v.

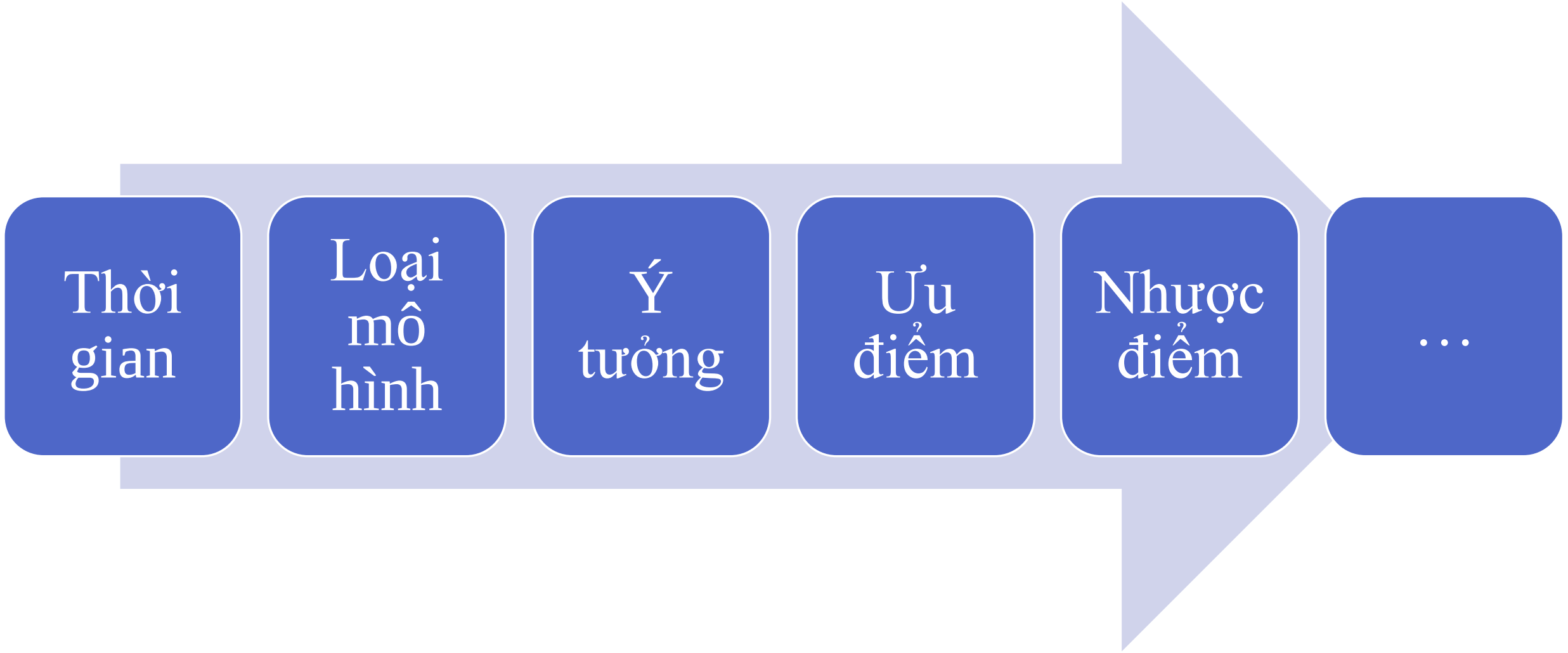
7. Triển khai hệ thống:

- Triển khai hệ thống khai phá luật kết hợp vào môi trường thực tế.

8. Giám sát và bảo trì:

- Giám sát hiệu suất của hệ thống khai phá luật kết hợp theo thời gian.
- Cập nhật hệ thống khi cần thiết để đảm bảo hiệu quả và tính chính xác.

3. Lựa chọn thuật toán



3. Lựa chọn thuật toán

Thuật toán	Ý tưởng	Ưu điểm	Nhược điểm	Thích hợp cho
Apriori	Lặp để tìm tập mục thường xuyên có độ hỗ trợ và độ tin cậy tối thiểu	Đơn giản, dễ hiểu, hiệu quả cho tập dữ liệu nhỏ	Hiệu suất giảm khi kích thước tập dữ liệu tăng, có thể bỏ sót LKH quan trọng	Tập dữ liệu nhỏ
FP-Growth	Cải tiến Apriori, sử dụng cây FP-tree để lưu trữ dữ liệu và tìm kiếm LKH hiệu quả hơn	Hiệu quả hơn Apriori cho tập dữ liệu lớn, có thể tìm ra LKH dài hơn	Phức tạp hơn Apriori, đòi hỏi nhiều bộ nhớ hơn	Tập dữ liệu lớn
CARMA	Sử dụng nén dữ liệu để giảm kích thước tập dữ liệu trước khi khai phá LKH	Hiệu quả cho tập dữ liệu rất lớn, có thể tìm ra LKH hiếm gặp	Phức tạp hơn Apriori và FP-Growth, đòi hỏi nhiều thời gian tính toán hơn	Tập dữ liệu rất lớn
PrefixSpan	Chuyên dụng tìm kiếm LKH có tiền tố chung	Hiệu quả tìm kiếm LKH có tiền tố chung, có thể tìm ra LKH dài hơn	Phức tạp hơn các thuật toán khác, đòi hỏi nhiều thời gian tính toán hơn	Tìm kiếm LKH có tiền tố chung
H-Mine	Sử dụng phương pháp dựa trên đồ thị để tìm kiếm LKH	Có thể tìm ra LKH phức tạp, có thể xử lý tập dữ liệu có nhiễu	Phức tạp hơn các thuật toán khác, đòi hỏi nhiều thời gian tính toán hơn	Tìm kiếm LKH phức tạp, xử lý tập dữ liệu có nhiễu

4. Đánh giá hiệu quả

Đánh giá hiệu quả và mức độ hữu ích của các LKH:

1. Tỷ lệ hỗ trợ (Support):

- Thể hiện tỷ lệ phần trăm các giao dịch chứa tất cả các mục trong LKH.
- **Công thức:** (Số giao dịch chứa LKH / Tổng số giao dịch) * 100%.
- **Ý nghĩa:** Tỷ lệ hỗ trợ cao cho thấy LKH được dự đoán là phổ biến và có ý nghĩa trong dữ liệu.

2. Độ tin cậy (Confidence):

- Biểu thị tỷ lệ phần trăm các giao dịch có chứa các mục trong LKH.
- **Công thức:** (Số giao dịch có chứa LKH / Số giao dịch chứa X) * 100%, với X là một mục bất kỳ trong LKH.
- **Ý nghĩa:** Độ tin cậy cao cho thấy LKH được dự đoán có khả năng xảy ra cao trong thực tế.

4. Đánh giá hiệu quả

Đánh giá hiệu quả và mức độ hữu ích của các LKH:

3. Độ nâng cao (Lift):

- Đánh giá mức độ mà LKH làm tăng khả năng xảy ra các mục cùng nhau so với nếu chúng xảy ra độc lập.
- **Công thức:** (Độ tin cậy / Tỷ lệ hỗ trợ) * 100%.
- **Ý nghĩa:** Độ nâng cao cao cho thấy LKH được dự đoán có mối quan hệ mạnh mẽ giữa các mục trong LKH.

4. Độ dài luật (Rule length):

- Biểu thị số lượng các mục trong LKH.
- **Ý nghĩa:** Độ dài luật ngắn thường dễ hiểu và hữu ích hơn, trong khi độ dài luật dài có thể cung cấp thông tin chi tiết hơn về mối quan hệ giữa các mục.

4. Đánh giá hiệu quả

Đánh giá LKH bằng phương pháp trực quan hóa dữ liệu:

- **Biểu đồ thanh:** Hiển thị tỷ lệ phần trăm các LKH theo độ tin cậy, độ nâng cao, v.v.
- **Biểu đồ Pareto:** Hiển thị 20% LKH có độ tin cậy cao nhất hoặc độ nâng cao cao nhất.
- …

5. Triển khai hệ thống

Lưu trữ LKH một cách hiệu quả trong hệ thống thực tế là chìa khóa để đảm bảo truy cập nhanh chóng, dễ dàng và sử dụng hiệu quả các quy tắc được khai phá.

1. Cấu trúc dữ liệu dạng bảng:

- Sử dụng bảng cơ sở dữ liệu để lưu trữ LKH, với các cột lưu trữ các thuộc tính của LKH như ID quy tắc, tiền đề, kết luận, độ hỗ trợ, độ tin cậy, độ nâng cao, v.v.
- *Ưu điểm*: Đơn giản, dễ dàng truy cập và thao tác bằng các truy vấn SQL.
- *Nhược điểm*: Khó khăn khi lưu trữ LKH phức tạp với nhiều tiền đề và kết luận, hiệu suất truy vấn có thể chậm với lượng lớn LKH.

5. Triển khai hệ thống

Lưu trữ LKH một cách hiệu quả trong hệ thống thực tế là chìa khóa để đảm bảo truy cập nhanh chóng, dễ dàng và sử dụng hiệu quả các quy tắc được khai phá.

2. Cấu trúc dữ liệu đồ thị:

- Biểu diễn LKH dưới dạng đồ thị, với các nút biểu thị các mục trong tiền đề và kết luận, và các cạnh biểu thị mối quan hệ giữa các mục.
- *Ưu điểm*: Hiệu quả cho LKH phức tạp, dễ dàng trực quan hóa và phân tích mối quan hệ giữa các mục.
- *Nhược điểm*: Yêu cầu thuật toán đồ thị để truy cập và thao tác dữ liệu, có thể phức tạp hơn so với cấu trúc bảng.

5. Triển khai hệ thống

Lưu trữ LKH một cách hiệu quả trong hệ thống thực tế là chìa khóa để đảm bảo truy cập nhanh chóng, dễ dàng và sử dụng hiệu quả các quy tắc được khai phá.

3. Cấu trúc dữ liệu dạng cây:

- Biểu diễn LKH dưới dạng cây, với gốc cây là kết luận, và các nhánh con là các tiền đề.
- *Ưu điểm*: Hiệu quả cho LKH có nhiều tiền đề, dễ dàng duyệt và truy cập thông tin theo cấp bậc.
- *Nhược điểm*: Khó khăn khi lưu trữ LKH phức tạp với nhiều kết luận, có thể tốn nhiều bộ nhớ hơn so với cấu trúc bảng.

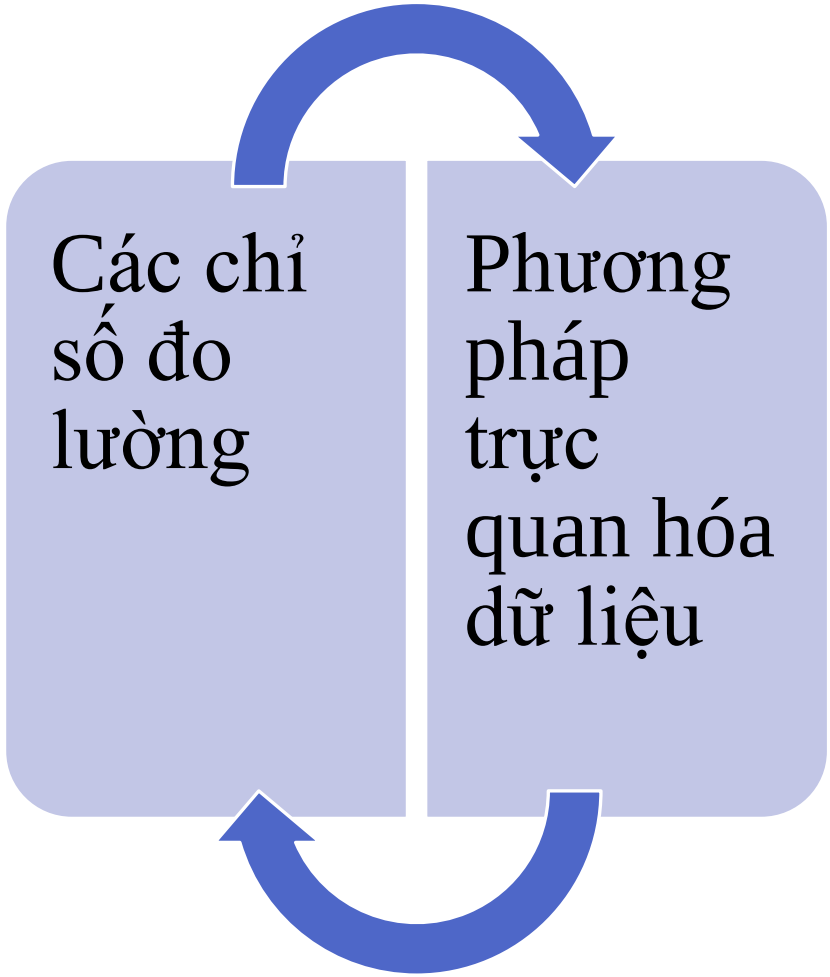
6. Giám sát và bảo trì

Giám sát *hiệu suất hệ thống và chất lượng dữ liệu.*

Cập nhật hệ thống và thuật toán khi cần thiết.

Bảo trì hệ thống để đảm bảo hoạt động ổn định và hiệu quả.

7. Tổng kết



Question & Answer