VIDEO PRESENTATION SCRIPT

Data Mining: Cải Thiện Apriori

Duration: 5-10 minutes

Nhóm [Tên nhóm]

Năm học 2024 - 2025

Video Recording Requirements

- **Duration**: 5-10 minutes
- **Show presenter**: Yes (hiện hình người báo cáo)
- **Audio quality**: Clear and audible
- **Screen sharing**: Clear and visible

1 [00:00 - 00:45] INTRODUCTION

Visual

Title slide with “CẢI THIỆN APRIORI - Frequent Itemset Mining và Association Rules”

1.1 Speaker Script:

“Xin chào thầy và các bạn. Hôm nay nhóm xin trình bày về đề tài **’Cải thiện Apriori’** trong môn Data Mining. Đề tài tập trung vào việc phân tích Market Basket và tìm các cải tiến cho thuật toán Apriori cổ điển.”

Key Points:

- Greeting and introduction
- Project title clearly stated
- Focus on Market Basket Analysis
- Mention Apriori algorithm improvements

Body Language:

- Smile and maintain eye contact with camera
- Speak clearly and at moderate pace
- Use hand gestures to emphasize key points

2 [00:45 - 01:30] PROBLEM STATEMENT

Visual

Slide showing Market Basket Analysis applications

2.1 Speaker Script:

"Vấn đề chúng ta giải quyết là làm sao để hiểu rõ hành vi mua sắm của khách hàng thông qua việc phân tích dữ liệu giao dịch. Từ đó, có thể phát hiện các sản phẩm thường được mua cùng nhau và đưa ra các gợi ý phù hợp."

"Các ứng dụng thực tế bao gồm: tối ưu hóa sắp đặt sản phẩm, cross-selling opportunities, tạo bundle sản phẩm, và quản lý tồn kho hiệu quả hơn."

Key Points:

- Problem statement: Understanding customer shopping behavior
- Goal: Discover products purchased together
- Real-world applications mentioned

Screen Share:

- Show the Market Basket Analysis slide
- Point to each application as you mention it

3 [01:30 - 02:30] DATASET

Visual

Groceries dataset statistics

3.1 Speaker Script:

"Dataset chúng tôi sử dụng là Groceries dataset, một bộ dữ liệu từ Machine Learning with R. Đây là dữ liệu giao dịch từ một cửa hàng tạp hóa, nơi mỗi giao dịch đại diện cho một giỏ hàng của khách hàng."

"Dataset chứa [số lượng sau khi chạy script.py] giao dịch và [số lượng sau khi chạy script.py] sản phẩm khác nhau. Trung bình mỗi khách hàng mua khoảng 4.4 sản phẩm trong một lần giao dịch."

Key Points:

- Dataset source: Machine Learning with R
- Type: Transactional data from grocery store
- Each transaction = customer basket
- Statistics: [Run script for actual values]

Screen Share:

- Show the Groceries dataset statistics slide
- Highlight key numbers

4 [02:30 - 04:00] APRIORI ALGORITHM

Visual

Apriori algorithm flow and steps

4.1 Speaker Script:

"Apriori là thuật toán cổ điển được đề xuất bởi Agrawal và Srikant vào năm 1994. Nó dựa trên nguyên tắc: tất cả các tập con của một frequent itemset cũng phải là frequent."

"Thuật toán có 4 bước chính:

1. Initialization: Thiết lập minimum support threshold

2. Generate Candidates: Tạo k-itemsets từ (k-1)-itemsets

3. Prune: Loại bỏ items có support thấp hơn ngưỡng

4. Repeat: Tăng k và lặp lại cho đến khi không tìm thấy frequent items nào nữa"

"Chúng tôi cũng sử dụng các chỉ số đánh giá như Support, Confidence, Lift, Leverage, Conviction, và Zhang's Metric để đánh giá chất lượng các association rules."

Key Points:

- Algorithm: Apriori (1994, Agrawal & Srikant)
- Core principle: Subsets of frequent itemsets are also frequent
- 4 main steps explained
- Evaluation metrics mentioned

Screen Share:

- Show the algorithm flow diagram
- Highlight each step as you explain it
- Show the evaluation metrics table

5 [04:00 - 05:30] IMPLEMENTATION & RESULTS

Visual

Pipeline processing and results

5.1 Speaker Script:

"Chúng tôi implement thuật toán sử dụng Python và thư viện mlxtend. Pipeline xử lý bao gồm: Load Data, Encode, Cleaning, Mining, và Rules Generation."

"Kết quả cho thấy Whole milk là sản phẩm phổ biến nhất với khoảng 25.5% support. Các sản phẩm phổ biến tiếp theo bao gồm Other vegetables, Rolls/buns, Soda, và Yogurt."

"Chúng tôi tìm được [số lượng sau khi chạy script.py] association rules với minimum confidence là 0.25. Các rule này cho thấy các mối quan hệ thú vị giữa các sản phẩm."

Key Points:

- Implementation: Python + mlxtend
- Pipeline: 5 steps
- Top product: Whole milk (25.5% support)
- Rules discovered: [Run script for actual value]
- Min confidence: 0.25

Screen Share:

- Show the pipeline diagram
- Display top frequent items table
- Show example association rules

6 [05:30 - 07:30] IMPROVEMENT STRATEGIES

Visual

6 improvement strategies

6.1 Speaker Script:

"Apriori có 4 hạn chế chính: multiple database scans, large candidate sets, high memory usage, và expensive computational cost. Để giải quyết các vấn đề này, chúng tôi đã triển khai 6 chiến lược cải tiến:"

1. **Sampling**: Mine trên sample 30% dữ liệu để giảm chi phí tính toán
2. **DHP (Hash-based)**: Sử dụng hash table để pruning candidates hiệu quả hơn
3. **Transaction Reduction**: Loại bỏ transactions không chứa frequent items
4. **ECLAT (Vertical)**: Sử dụng vertical tid-list format để nhanh hơn
5. **DIC (Dynamic Counting)**: Interleaved counting để giảm database scans
6. **Partitioning**: Chia database thành các partitions để xử lý độc lập

Key Points:

- 4 main limitations of Apriori
- 6 improvement strategies implemented
- Each strategy briefly explained

Screen Share:

- Show the limitations slide
- Display each improvement strategy
- Use diagrams to illustrate concepts

7 [07:30 - 08:30] COMPARISON

Visual

Algorithm comparison table

7.1 Speaker Script:

"So sánh 9 thuật toán cho thấy FP-Growth và FP-Max có hiệu suất tốt nhất. Với Groceries dataset:"

"- FP-Growth nhanh hơn 2-3 lần so với Apriori truyền thống

- FP-Max nhanh hơn 3-5 lần và tối ưu bộ nhớ tốt hơn

- Các cải tiến đề xuất có thể giảm 30-50% execution time"

"Kết quả này cho rằng Apriori vẫn là nền tảng vững chắc, nhưng các thuật toán cải tiến hoặc các phương pháp tiếp cận mới như FP-Growth và FP-Max là lựa chọn tốt hơn cho production environments."

Key Points:

- FP-Growth: 2-3x faster
- FP-Max: 3-5x faster, better memory optimization
- Proposed improvements: 30-50% time reduction
- Recommendation: FP-Growth/FP-Max for production

Screen Share:

- Show the comparison table
- Highlight performance improvements
- Display performance metrics

8 [08:30 - 09:30] CONCLUSION

Visual

Summary and applications

8.1 Speaker Script:

"Tóm lại:"

"- Apriori là nền tảng vững chắc cho frequent itemset mining

- 6 thuật toán cải tiến đã được triển khai và đánh giá

- Các kỹ thuật tối ưu có thể giảm 30-50% execution time

- FP-Growth và FP-Max là lựa chọn tốt nhất cho production environments"

"Ứng dụng thực tế của các thuật toán này rất rộng:"

- **Retail và E-commerce**: Market Basket Analysis
- **Healthcare**: Pattern detection trong bệnh sử
- **Web usage mining**: Phân tích hành vi người dùng
- **Bioinformatics**: Phân tích gene sequences và protein interactions

Key Points:

- Summary of achievements
- Performance improvements
- Real-world applications

Screen Share:

- Show summary slide
- Display application areas

9 [09:30 - 10:00] Q&A

Visual

"Thank you! Questions?" slide

9.1 Speaker Script:

"Cảm ơn thầy và các bạn đã lắng nghe. Nhóm xin nhận câu hỏi."

Key Points:

- Thank the audience
- Invite questions
- Be prepared for:
  - Algorithm complexity questions
  - Implementation details
  - Dataset characteristics
  - Performance metrics

Body Language:

- Smile and appear approachable
- Maintain eye contact
- Listen carefully to questions
- Answer concisely and clearly

10 RECORDING TIPS

10.1 Before Recording:

1. Practice the script 2-3 times to get comfortable
2. Check microphone and camera setup
3. Ensure good lighting (natural light is best)
4. Clean and professional background

10.2 During Recording:

1. **Eye contact**: Look at the camera, not the screen
2. **Speaking pace**: Moderate and clear
3. **Pronunciation**: Clear and articulate
4. **Gestures**: Natural hand movements to emphasize points
5. **Screen sharing**: Make sure the code/slides are clearly visible

10.3 Technical Setup:

- Use a quiet room
- Good internet connection (if streaming)
- Microphone close but not too close
- Camera at eye level
- Screen resolution: 1920x1080 or higher

10.4 Post-Recording:

1. Check audio quality
2. Verify screen clarity
3. Edit out any major mistakes
4. Add captions if possible
5. Keep final video under 10 minutes

11 CHECKLIST

- ☐ Practice script 2-3 times
- ☐ Set up proper lighting
- ☐ Test microphone quality
- ☐ Clean background
- ☐ Check internet connection
- ☐ Prepare slides for screen sharing
- ☐ Have water nearby
- ☐ Test recording software
- ☐ Record test video
- ☐ Check video quality before final recording