**Đề tài luận văn:**

“Khung phân loại lai ghép cho bài toán đề xuất mục mới: Kết hợp phân loại nhanh và tìm kiếm tương đồng”

*Tiếng Anh: A Hybrid Taxonomy-based Framework for Cold-start Item Recommendation: Combining Fast Classification and Similarity-based Matching*

Trong khoảng thời gian từ 15/04/2025 đến 30/06/2025. Kế hoạch được phân chia thành 7 giai đoạn chính.

**Tổng quan kế hoạch**

1. **Giai đoạn 1 (15/04-30/04):** Nghiên cứu tài liệu và chuẩn bị
2. **Giai đoạn 2 (01/05-15/05):** Thu thập và phân tích dữ liệu
3. **Giai đoạn 3 (16/05-31/05):** Xây dựng taxonomy và biểu diễn đặc trưng
4. **Giai đoạn 4 (01/06-10/06):** Phát triển mô hình phân loại nhanh
5. **Giai đoạn 5 (11/06-18/06):** Phát triển kỹ thuật tìm kiếm tương đồng
6. **Giai đoạn 6 (19/06-25/06):** Kết hợp phân loại và tìm kiếm tương đồng
7. **Giai đoạn 7 (26/06-30/06):** Đánh giá và so sánh

Mỗi giai đoạn được chia thành các công việc chi tiết với thời gian bắt đầu, kết thúc, người thực hiện, và đầu ra cụ thể. Kế hoạch này đảm bảo:

## Bảng phân công công việc chi tiết

* **Đề tài:** Khung phân loại lai ghép cho bài toán đề xuất mục mới: Kết hợp phân loại nhanh và tìm kiếm tương đồng  
  **Thời gian thực hiện:** 15/04/2025 - 30/06/2025  
  **Thành viên:** Hà và Dương

### **Giai đoạn 1: Nghiên cứu tài liệu và chuẩn bị (15/04/2025 - 30/04/2025)**

| **STT** | **Công việc chi tiết** | **Người thực hiện** | **Thời gian bắt đầu** | **Thời gian kết thúc** | **Đầu ra** |
| --- | --- | --- | --- | --- | --- |
| 1.1 | Tổng hợp và phân loại tài liệu về vấn đề cold-start recommendation từ các nguồn học thuật (IEEE, ACM, Springer) | Hà | 15/04/2025 | 19/04/2025 | Danh sách 30-40 bài báo được phân loại theo phương pháp, năm xuất bản |
| 1.2 | Tổng hợp tài liệu về phương pháp phân loại (taxonomy) và embedding trong hệ thống đề xuất | Dương | 15/04/2025 | 19/04/2025 | Danh sách 10-20 bài báo liên quan đến taxonomy và embedding |
| 1.3 | Đánh giá các phương pháp state-of-the-art hiện có để giải quyết vấn đề cold-start | Hà | 20/04/2025 | 23/04/2025 | Bảng so sánh chi tiết các phương pháp, ưu/nhược điểm |
| 1.4 | Nghiên cứu và tổng hợp về các kỹ thuật biểu diễn taxonomy và phương pháp xây dựng taxonomy tự động | Dương | 20/04/2025 | 23/04/2025 | Báo cáo tổng hợp về các kỹ thuật biểu diễn taxonomy |
| 1.5 | Tìm hiểu chi tiết về các kỹ thuật tìm kiếm tương đồng hiệu quả (ANN, KNN, HNSW) và thư viện tương ứng | Hà | 24/04/2025 | 26/04/2025 | Báo cáo so sánh các thuật toán tìm kiếm gần đúng |
| 1.6 | Đánh giá và lựa chọn bộ dữ liệu phù hợp cho nghiên cứu (MovieLens, Amazon, Yelp) | Dương | 24/04/2025 | 26/04/2025 | Báo cáo đánh giá 3-5 bộ dữ liệu kèm đặc tính |
| 1.7 | Thiết kế sơ đồ khái niệm (conceptual framework) của mô hình lai ghép đề xuất | Hà & Dương | 27/04/2025 | 30/04/2025 | Sơ đồ khái niệm và đề cương chi tiết của phương pháp |

### **Giai đoạn 2: Thu thập và phân tích dữ liệu (01/05/2025 - 15/05/2025)**

| **STT** | **Công việc chi tiết** | **Người thực hiện** | **Thời gian bắt đầu** | **Thời gian kết thúc** | **Đầu ra** |
| --- | --- | --- | --- | --- | --- |
| 2.1 | Tải và chuẩn bị bộ dữ liệu chính (MovieLens 20M, Amazon Review, Yelp) | Dương | 01/05/2025 | 03/05/2025 | Script tải và cơ sở dữ liệu thô |
| 2.2 | Phân tích thống kê dữ liệu (phân bố đánh giá, mật độ tương tác, số lượng user/item) | Hà | 01/05/2025 | 03/05/2025 | Báo cáo phân tích thống kê, biểu đồ phân bố |
| 2.3 | Tiền xử lý dữ liệu văn bản (mô tả sản phẩm, reviews): tokenization, stemming, loại bỏ stopwords | Dương | 04/05/2025 | 07/05/2025 | Dataset văn bản đã chuẩn hóa |
| 2.4 | Chuẩn hóa metadata có cấu trúc (thể loại, thuộc tính, tags) | Hà | 04/05/2025 | 07/05/2025 | Metadata đã chuẩn hóa dạng bảng/JSON |
| 2.5 | Tách dữ liệu thành các tập train (70%), validation (15%), test (15%) | Dương | 08/05/2025 | 09/05/2025 | Tập dữ liệu đã phân chia |
| 2.6 | Tạo tập mô phỏng cold-start bằng cách tách một phần item làm "mục mới" | Hà | 08/05/2025 | 09/05/2025 | Tập dữ liệu cold-start simulation |
| 2.7 | Trích xuất đặc trưng cơ bản từ dữ liệu văn bản (TF-IDF, Bag-of-Words) | Dương | 10/05/2025 | 12/05/2025 | Ma trận đặc trưng văn bản |
| 2.8 | Phân tích mối tương quan giữa đặc trưng nội dung và hành vi tương tác | Hà | 10/05/2025 | 12/05/2025 | Báo cáo phân tích tương quan |
| 2.9 | Kiểm tra tính đầy đủ và chất lượng dữ liệu, xử lý dữ liệu thiếu/nhiễu | Hà & Dương | 13/05/2025 | 15/05/2025 | Bộ dữ liệu hoàn chỉnh sẵn sàng cho mô hình |

### **Giai đoạn 3: Xây dựng taxonomy và biểu diễn đặc trưng (16/05/2025 - 31/05/2025)**

| **STT** | **Công việc chi tiết** | **Người thực hiện** | **Thời gian bắt đầu** | **Thời gian kết thúc** | **Đầu ra** |
| --- | --- | --- | --- | --- | --- |
| 3.1 | Xây dựng taxonomy tự động bằng phân cụm phân cấp (Hierarchical Clustering) | Dương | 16/05/2025 | 19/05/2025 | Cấu trúc phân loại ban đầu dạng cây/đồ thị |
| 3.2 | Cải thiện taxonomy bằng phương pháp Topic Modeling (LDA, NMF) | Hà | 16/05/2025 | 19/05/2025 | Các chủ đề và phân cụm được tăng cường |
| 3.3 | Tích hợp thông tin ngữ nghĩa vào taxonomy dựa trên mô hình Word2Vec/GloVe | Dương | 20/05/2025 | 22/05/2025 | Taxonomy được làm giàu với quan hệ ngữ nghĩa |
| 3.4 | Biểu diễn items bằng vector nhúng từ mô tả văn bản (FastText, USE) | Hà | 20/05/2025 | 22/05/2025 | Ma trận embedding vectors cho items |
| 3.5 | Phát triển phương pháp kết hợp các nguồn đặc trưng khác nhau (văn bản, metadata, taxonomy) | Dương | 23/05/2025 | 26/05/2025 | Vector đặc trưng tổng hợp cho mỗi item |
| 3.6 | Đánh giá chất lượng taxonomy dựa trên độ nhất quán nội bộ và độ tách biệt giữa các nhóm | Hà | 23/05/2025 | 26/05/2025 | Báo cáo đánh giá chất lượng taxonomy |
| 3.7 | Thực hiện giảm chiều dữ liệu (PCA, t-SNE) và trực quan hóa không gian đặc trưng | Dương | 27/05/2025 | 29/05/2025 | Vector đặc trưng chiều thấp và biểu đồ trực quan |
| 3.8 | Đánh giá và tinh chỉnh các tham số biểu diễn đặc trưng | Hà & Dương | 30/05/2025 | 31/05/2025 | Báo cáo tối ưu hóa tham số và biểu diễn cuối cùng |

### **Giai đoạn 4: Phát triển mô hình phân loại nhanh (01/06/2025 - 10/06/2025)**

| **STT** | **Công việc chi tiết** | **Người thực hiện** | **Thời gian bắt đầu** | **Thời gian kết thúc** | **Đầu ra** |
| --- | --- | --- | --- | --- | --- |
| 4.1 | Thiết kế kiến trúc mô hình phân loại dựa trên taxonomy | Dương | 01/06/2025 | 02/06/2025 | Thiết kế kiến trúc mô hình phân loại |
| 4.2 | Huấn luyện các mô hình phân loại (Random Forest, XGBoost, LightGBM) | Hà | 01/06/2025 | 02/06/2025 | Các mô hình phân loại ban đầu |
| 4.3 | Tối ưu siêu tham số cho mô hình phân loại bằng Grid Search/Random Search | Dương | 03/06/2025 | 05/06/2025 | Bộ tham số tối ưu cho mỗi mô hình |
| 4.4 | Đánh giá hiệu năng các mô hình phân loại trên tập validation | Hà | 03/06/2025 | 05/06/2025 | Báo cáo đánh giá (accuracy, precision, recall, F1) |
| 4.5 | Tối ưu hóa tốc độ phân loại bằng kỹ thuật pruning và feature selection | Dương | 06/06/2025 | 07/06/2025 | Mô hình tối ưu về tốc độ và hiệu suất |
| 4.6 | Phân tích lỗi và cải tiến mô hình phân loại | Hà | 06/06/2025 | 07/06/2025 | Báo cáo phân tích lỗi và kế hoạch cải tiến |
| 4.7 | Đánh giá độ chính xác phân loại theo các cấp độ khác nhau của taxonomy | Hà & Dương | 08/06/2025 | 10/06/2025 | Báo cáo hiệu suất phân loại theo cấp độ phân cấp |

### **Giai đoạn 5: Phát triển kỹ thuật tìm kiếm tương đồng (11/06/2025 - 18/06/2025)**

| **STT** | **Công việc chi tiết** | **Người thực hiện** | **Thời gian bắt đầu** | **Thời gian kết thúc** | **Đầu ra** |
| --- | --- | --- | --- | --- | --- |
| 5.1 | Triển khai thuật toán KNN cơ bản dựa trên vector đặc trưng | Dương | 11/06/2025 | 12/06/2025 | Mô hình KNN cơ bản |
| 5.2 | Tích hợp các thư viện ANN hiệu suất cao (FAISS, Annoy, NMSLIB) | Hà | 11/06/2025 | 12/06/2025 | Mô hình ANN tương ứng với từng thư viện |
| 5.3 | Tối ưu cấu trúc chỉ mục cho tìm kiếm tương đồng hiệu quả | Dương | 13/06/2025 | 14/06/2025 | Cấu trúc chỉ mục tối ưu (HNSW, IVF) |
| 5.4 | Đánh giá hiệu suất tìm kiếm (thời gian truy vấn, độ chính xác) | Hà | 13/06/2025 | 14/06/2025 | Báo cáo benchmark hiệu suất các phương pháp |
| 5.5 | Phát triển phương pháp tìm kiếm tương đồng dựa trên cấu trúc taxonomy | Dương | 15/06/2025 | 16/06/2025 | Mô hình tìm kiếm kết hợp taxonomy |
| 5.6 | So sánh các phương pháp tìm kiếm tương đồng khác nhau | Hà | 15/06/2025 | 16/06/2025 | Bảng so sánh chi tiết các phương pháp |
| 5.7 | Đánh giá tính khả thi của tìm kiếm tương đồng trên tập dữ liệu lớn | Hà & Dương | 17/06/2025 | 18/06/2025 | Báo cáo khả năng mở rộng của phương pháp |

### **Giai đoạn 6: Kết hợp phân loại và tìm kiếm tương đồng (19/06/2025 - 25/06/2025)**

| **STT** | **Công việc chi tiết** | **Người thực hiện** | **Thời gian bắt đầu** | **Thời gian kết thúc** | **Đầu ra** |
| --- | --- | --- | --- | --- | --- |
| 6.1 | Thiết kế kiến trúc framework kết hợp phân loại và tìm kiếm tương đồng | Hà & Dương | 19/06/2025 | 20/06/2025 | Sơ đồ kiến trúc framework tích hợp |
| 6.2 | Triển khai cơ chế kết hợp kết quả từ phân loại và tìm kiếm tương đồng | Dương | 21/06/2025 | 22/06/2025 | Mô-đun kết hợp hoàn chỉnh |
| 6.3 | Phát triển thuật toán re-ranking dựa trên kết quả từ hai phương pháp | Hà | 21/06/2025 | 22/06/2025 | Thuật toán re-ranking với đa tiêu chí |
| 6.4 | Tối ưu tham số trọng số giữa kết quả phân loại và tìm kiếm tương đồng | Dương | 23/06/2025 | 24/06/2025 | Bộ tham số tối ưu cho việc kết hợp |
| 6.5 | Xây dựng pipeline hoàn chỉnh cho quy trình đề xuất mục mới | Hà | 23/06/2025 | 24/06/2025 | Pipeline tích hợp các thành phần |
| 6.6 | Kiểm thử toàn diện hệ thống trên các kịch bản khác nhau | Hà & Dương | 25/06/2025 | 25/06/2025 | Báo cáo kiểm thử và đánh giá ban đầu |

### **Giai đoạn 7: Đánh giá và so sánh (26/06/2025 - 30/06/2025)**

| **STT** | **Công việc chi tiết** | **Người thực hiện** | **Thời gian bắt đầu** | **Thời gian kết thúc** | **Đầu ra** |
| --- | --- | --- | --- | --- | --- |
| 7.1 | Thiết lập và triển khai các baseline để so sánh (Popularity-based, Content-based) | Dương | 26/06/2025 | 27/06/2025 | Kết quả các mô hình baseline |
| 7.2 | Đánh giá chi tiết framework đề xuất trên tập dữ liệu cold-start | Hà | 26/06/2025 | 27/06/2025 | Kết quả đánh giá (Precision@k, Recall@k, NDCG@k) |
| 7.3 | So sánh với các phương pháp state-of-the-art hiện đại | Dương | 28/06/2025 | 28/06/2025 | Bảng so sánh toàn diện với các phương pháp khác |
| 7.4 | Phân tích chi phí tính toán và hiệu suất của framework | Hà | 28/06/2025 | 28/06/2025 | Báo cáo sử dụng tài nguyên và thời gian phản hồi |
| 7.5 | Phân tích tính mở rộng và khả năng áp dụng trên quy mô lớn | Dương | 29/06/2025 | 29/06/2025 | Báo cáo đánh giá khả năng mở rộng |
| 7.6 | Tổng hợp kết quả và rút ra kết luận cho nghiên cứu | Hà & Dương | 30/06/2025 | 30/06/2025 | Báo cáo tổng kết kết quả và đóng góp chính |
| 7.7 | Xác định hướng phát triển và cải tiến tiếp theo | Hà & Dương | 30/06/2025 | 30/06/2025 | Đề xuất hướng nghiên cứu tương lai |