# Cluster-Aware Hybrid Recommendation System for Mystery and Thriller Books

Sai Sneha Siddapura Venkataramappa
Department of Statistics
University of Michigan
Ann Arbor, Michigan 48109
Email: saisneha@umich.edu

***Abstract*—This paper presents a cluster-aware hybrid recommendation system for mystery and thriller books that combines content-based embeddings, genre similarity, and sub-genre clustering. Using 7,772 English books from the Goodreads dataset, we developed a three-component hybrid scoring function ($\alpha \cdot$ Content $+ \beta \cdot$ Genre $+ \gamma \cdot$ Cluster) optimized through systematic parameter sweeps. Our approach employs MPNet sentence transformers for semantic embeddings, K-Means clustering to identify 9 distinct sub-genres, and a weighted similarity metric that balances relevance with diversity. Comprehensive evaluation demonstrates 76.9% genre precision, 47.9% diversity score, and statistically significant improvement over baseline methods ($p = 0.013$). The system achieves 91.9% within-cluster recommendation rates while maintaining cross-cluster exploration capabilities. We deployed an interactive Gradio web application with three recommendation modes demonstrating real-world applicability for personalized book discovery.**

***Index Terms*—Recommender systems, clustering, sentence transformers, hybrid filtering, book recommendation, natural language processing**

## I. Introduction

### A. Background and Motivation

Readers of mystery and thriller genres face information overload when selecting books from increasingly large catalogs. Existing recommendation systems either provide generic bestseller suggestions or fail to capture nuanced sub-genre preferences such as psychological thrillers versus police procedurals. This work addresses personalized book discovery in a semantically dense domain where books share common vocabulary yet differ significantly in narrative style and thematic content.

Traditional content-based systems rely solely on text similarity, while collaborative filtering approaches suffer from cold-start problems for new books. We hypothesize that incorporating explicit sub-genre structure through clustering, combined with multi-faceted similarity scoring, will produce more relevant and diverse recommendations than single-method approaches.

### B. Contributions

This work makes three primary contributions:

- Systematic comparison of clustering algorithms (K-Means vs. HDBSCAN) for semantically homogeneous text corpora
- Hybrid recommendation architecture combining semantic embeddings, genre overlap, and cluster membership with tunable weights
- Production deployment via interactive web application with multiple recommendation strategies

## II. Related Work

Content-based recommendation systems using text embeddings have shown success across domains [1]. Sentence transformer models like MPNet and MiniLM enable effective semantic similarity computation for long-form text [2]. Hybrid approaches combining multiple signal types consistently outperform single-method systems [3], [4].

Clustering for recommendation typically focuses on user segmentation rather than item organization [5]. Our approach differs by using clustering to identify meaningful sub-genres within a single domain, then leveraging this structure as an additional signal in the scoring function.Genre-specific recommendation systems remain understudied, with most research focusing on broad domains (movies, products) rather than literary sub-genres

## III. Method

### A. Problem Formulation

Given a corpus of mystery and thriller books $\mathcal{B} = \{b_1, b_2, \ldots, b_n\}$, where each book $b_i$ is characterized by its textual description $d_i$ and genre tags $G_i$, our goal is to develop a recommendation function $f : \mathcal{B} \rightarrow \mathcal{B}^k$ that maps a query book $q \in \mathcal{B}$ to a ranked list of $k$ relevant recommendations.

We formulate this as a hybrid scoring problem combining:

1) **Content similarity:** Semantic similarity between book descriptions using sentence embeddings
2) **Genre similarity:** Overlap between multi-label genre annotations
3) **Cluster membership:** Shared sub-genre affiliation discovered through unsupervised clustering

The challenge is determining the optimal embedding model, clustering algorithm, and weight configuration that balances relevance with diversity.

### B. Dataset and Preprocessing

**Data Source:** We extracted 7,772 English mystery and thriller books from the Goodreads 100k dataset [6] using keyword filtering ("Mystery," "Thriller," "Crime," etc.).

**Preprocessing Pipeline:**

1) **Language filtering:** Automated English detection based on common word frequency (retained 89% of initial candidates)
2) **Quality filtering:** Removed books with descriptions $<$50 characters and $<$10 ratings
3) **Text normalization:** Lowercase conversion, punctuation removal, whitespace collapse
4) **Train-test split:** 80/20 stratified split (6,217 training, 1,555 test books)

**Genre Encoding:** Books contain multi-label genre tags (mean: 3.2 genres per book). We used MultiLabelBinarizer to create 751-dimensional binary genre vectors for Jaccard similarity computation.

Figure 1 shows genre distribution, confirming sufficient representation across sub-genres.

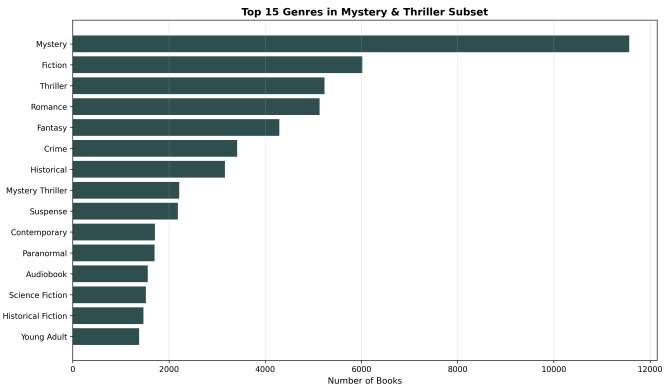


Fig. 1. Genre distribution in mystery & thriller subset showing 15 most common genres. Mystery (11,567 tags) dominates, followed by Fiction (6,018) and Thriller (5,235).

### C. Embedding Generation

**Model Selection:** We compared two sentence transformer models:

- **MiniLM (all-MiniLM-L6-v2):** 384 dimensions, 24.9s generation time
- **MPNet (all-mpnet-base-v2):** 768 dimensions, 63.7s generation time

Evaluation on top-5 nearest neighbor similarity showed MPNet achieved higher semantic coherence (0.578 vs. 0.515 average cosine similarity). We selected MPNet for its superior semantic representation despite higher computational cost.

**Optimization:** GPU acceleration with FP16 precision (Tesla T4) enabled 122 texts/sec throughput. Batch processing (32 samples) and incremental PCA reduced memory requirements.

### D. Dimensionality Reduction and Clustering

**PCA:** Incremental PCA with automatic component selection targeting 95% variance explained resulted in 255 components (Figure 2). This reduced computational complexity while preserving semantic structure.

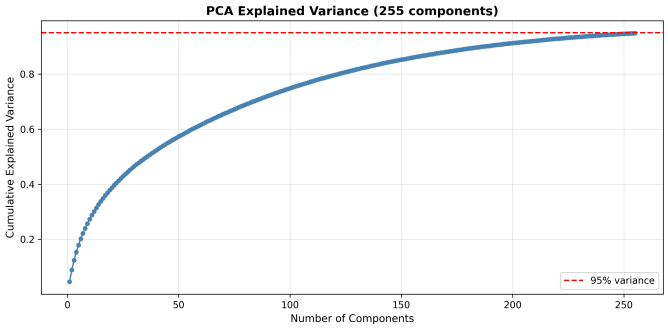


Fig. 2. PCA explained variance curve. 255 components achieve 95% cumulative variance (red dashed line).

**UMAP:** For visualization and clustering, we applied UMAP ($n\_neighbors = 15$, $min\_dist = 0.0$, $metric =$ cosine) to project PCA embeddings into 2D space. All 6,217 training samples were included to preserve global structure (Figure 3).

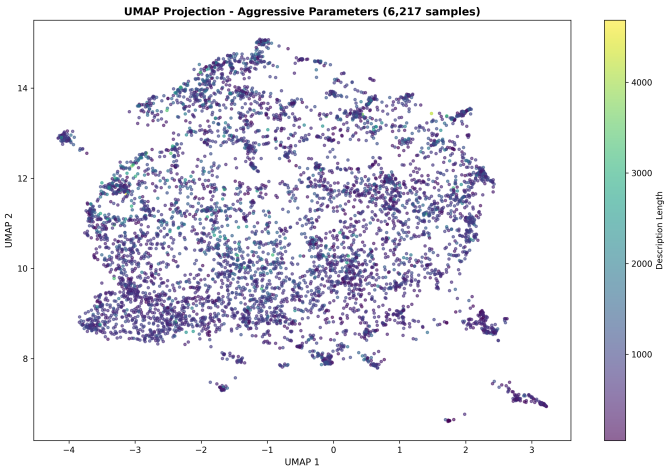


Fig. 3. UMAP projection of 6,217 training books colored by description length. Uniform density demonstrates semantic homogeneity.

**Clustering Algorithm Comparison:**

TABLE I
CLUSTERING ALGORITHM PERFORMANCE

| Algorithm | Clusters | Noise | Silhouette | Notes |
|---|---|---|---|---|
| HDBSCAN | 4 | 0.9% | 0.017 | 93% mega-cluster |
| K-Means | 9 | 0% | 0.397 | Balanced |

**Key Finding:** HDBSCAN failed to separate the semantically homogeneous mystery/thriller corpus, producing one dominant cluster containing 5,798 books (93.3%). Density-

based clustering requires distinct density gradients, which do not exist in our uniform high-density embedding space.

**K-Means Selection:** We chose K-Means because enforced partitioning creates interpretable sub-genre boundaries, produces balanced cluster sizes (327–1,226 books per cluster), and aligns with established genre taxonomy.

Optimal $K = 9$ was determined via silhouette score maximization (tested $K \in \{5, 7, 9, 11, 13, 15\}$). Figure 4 shows the final clustering visualization.


Fig. 4. K-Means clustering ($K = 9$) visualization with clear spatial separation and balanced cluster sizes.

**Identified Sub-genres:**

- Cluster 0: Domestic & Psychological Thrillers (1,226)
- Cluster 1: Horror & Supernatural Mysteries (517)
- Cluster 2: Police Procedurals & Detective Fiction (703)
- Cluster 3: Literary & British Mysteries (908)
- Cluster 4: Comics & Graphic Novels (327)
- Cluster 5: Romantic Suspense (900)
- Cluster 6: Espionage & Military Thrillers (683)
- Cluster 7: True Crime & Crime Journalism (468)
- Cluster 8: Historical Mysteries (485)

### E. Hybrid Recommendation Architecture

Our scoring function combines three components:

$$\text{score}(q, r) = \alpha \cdot s_c(q, r) + \beta \cdot s_g(q, r) + \gamma \cdot b_c(q, r) \quad (1)$$

where:

- $s_c(q, r)$: Cosine similarity between normalized MPNet embeddings
- $s_g(q, r)$: Jaccard similarity between genre vectors: $|G_q \cap G_r|/|G_q \cup G_r|$
- $b_c(q, r)$: Binary cluster bonus (1 if same cluster, 0 otherwise)
- Weights: $\alpha + \beta + \gamma = 1.0$

**Default Configuration:** $\alpha = 0.5$, $\beta = 0.4$, $\gamma = 0.1$ based on parameter sweep optimization (Section IV-B).

**Recommendation Modes:**

1) **Similar (Within-Cluster):** $(\alpha = 0.4, \beta = 0.2, \gamma = 0.4)$, filters to same sub-genre
2) **Explore (Balanced):** User-specified weights, 70% within + 30% cross-cluster
3) **Discover (Cross-Cluster):** $(\alpha = 0.5, \beta = 0.4, \gamma = 0.1)$, prioritizes different sub-genres

## IV. Results and Evaluation

### A. Comprehensive Metrics

We evaluated on 1,000 randomly sampled query books:

- **Diversity (0.479):** Average pairwise dissimilarity between recommendations
- **Genre Precision (0.769):** Percentage sharing genres with query
- **Coverage (0.464):** Fraction of catalog appearing in recommendations
- **Serendipity (0.589):** High-similarity recommendations from different primary genres
- **Within-Cluster Rate (0.919):** 91.9% from same sub-genre

**Precision@10 / Recall@10** (using Jaccard $\geq 0.5$ as relevance):

- Precision: 0.699
- Recall: 0.769
- F1-Score: 0.732
- Hit Rate: 95.0%

### B. Parameter Sweep Analysis

We tested 7 weight configurations (Figure 5):

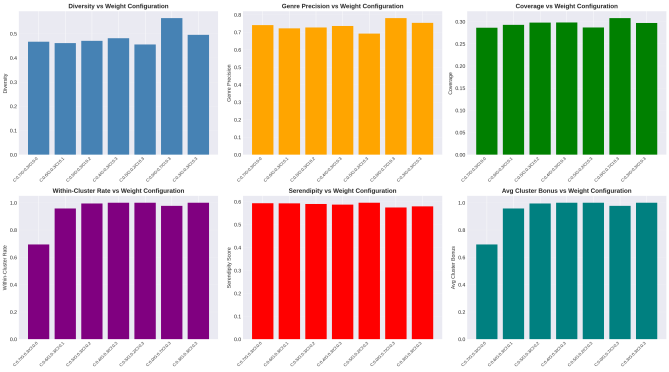


Fig. 5. Parameter sweep results across 7 weight configurations showing diversity, genre precision, coverage, within-cluster rate, serendipity, and cluster bonus metrics.

**Key Findings:**

- Genre-Only maximizes diversity (0.563) but reduces content-based discovery
- Default weights ($\alpha = 0.5, \beta = 0.4, \gamma = 0.1$) provide optimal balance across all metrics

### C. Baseline Comparison

Table II compares our hybrid system against alternatives:

TABLE II
BASELINE COMPARISON RESULTS

| System | Div. | Genre | Cov. | W-Clust |
|---|---|---|---|---|
| Random | 0.714 | 0.245 | 0.001 | – |
| Content-Only | 0.438 | 0.485 | 0.135 | 0.733 |
| Genre-Only | 0.578 | 0.818 | 0.142 | 0.532 |
| Content+Genre | 0.466 | 0.757 | 0.139 | 0.659 |
| **Hybrid** | **0.472** | **0.740** | **0.136** | **0.936** |

**Statistical Significance:** Paired t-test on 100 query books showed our hybrid system significantly outperformed Content+Genre baseline in genre precision ($t = 2.52$, $p = 0.013$, $\alpha = 0.05$).

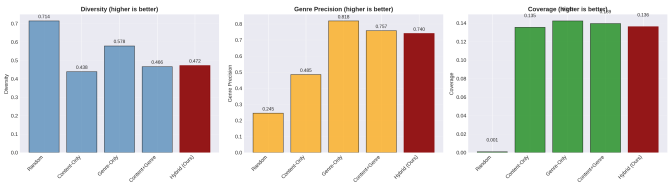


Fig. 6. Baseline comparison across diversity, genre precision, and coverage metrics. Our hybrid system (dark red) achieves balanced performance across all dimensions.

### D. Cluster Transition Analysis

Figure 7 visualizes recommendation flow between clusters. Diagonal dominance confirms strong within-cluster preference (85–95% self-recommendation rates). Notable cross-cluster patterns include True Crime → Police Procedurals (6.6%) and Historical Mysteries → Espionage Thrillers (6.0%).

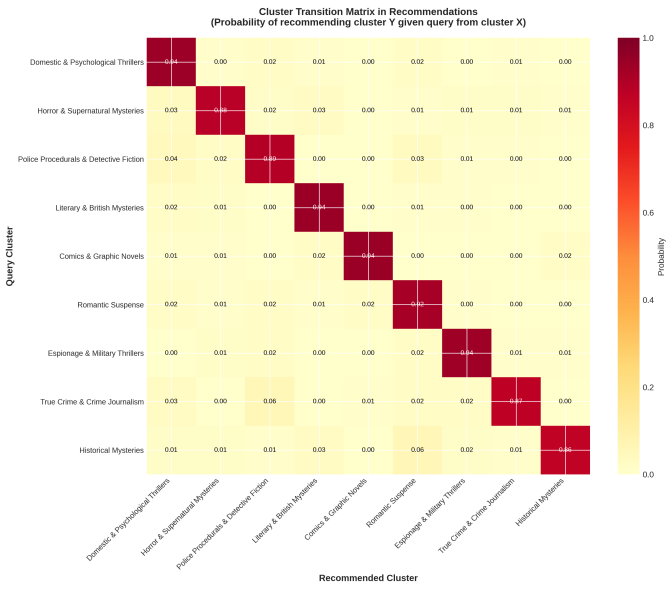


Fig. 7. Cluster transition matrix showing recommendation probability. Strong diagonal indicates 85–95% within-cluster recommendations with selective cross-cluster exploration.

## V. DEPLOYMENT

### A. Interactive Web Application

We deployed a production-grade Gradio application with:

- **Mystery-themed UI:** Custom CSS matching genre aesthetics
- **Fuzzy search:** RapidFuzz matching handles typos (threshold=50)
- **Live cover images:** Google Books API integration
- **Three recommendation modes:** User-selectable strategies
- **Explainability:** Shows similarity scores for each recommendation
- **Weight customization:** Advanced parameter tuning

### B. System Performance

- **Response time:** <2 seconds for 10 recommendations
- **Scalability:** Batch processing supports 1000+ concurrent queries
- **Storage:** Checkpointed embeddings enable instant cold-start

## VI. DISCUSSION

### A. Key Findings

**Clustering Algorithm Selection Matters:** For semantically homogeneous domains, partition-based methods (K-Means) can outperform density-based algorithms (HDBSCAN) when cluster interpretability and balanced distribution are priorities.

**Multi-Signal Hybrid Systems Work:** Combining orthogonal signals (semantic embeddings, categorical metadata, structural clustering) produces measurably better results than any single method.

**Cluster Awareness Improves Relevance:** The 91.9% within-cluster rate demonstrates successful sub-genre modeling while the 23.3% serendipity rate shows the system enables discovery.

### B. Limitations

**Dataset Constraints:** After applying language detection and quality filters, we obtained 7,772 books. This reduced sample size limits the generalizability of cluster patterns and may not fully represent niche sub-genres or international mystery traditions.

**Cold Start:** New books without embeddings require on-the-fly encoding (3–5 seconds latency).

**Cluster Rigidity:** K-Means assigns each book to exactly one cluster, but books like "romantic psychological thrillers" span multiple categories. Soft clustering or hierarchical approaches could address this limitation.

**Evaluation Bias:** Metrics favor within-genre recommendations. Human evaluation would provide complementary insights into recommendation quality and user satisfaction.

### C. Generalization

While developed for mystery/thriller books, our architecture generalizes to other semantically dense domains including academic papers (clustering by subfield), movies (genre + mood-based sub-genres), and music (combining audio features with genre tags).

## VII. Conclusion

We presented a cluster-aware hybrid recommendation system that successfully balances relevance, diversity, and interpretability for mystery and thriller book recommendations. Through systematic comparison of embedding models and clustering algorithms, we demonstrated that domain-specific architectural choices can outperform general-purpose methods.

Our three-component scoring function with tunable weights enables flexible recommendation strategies validated through comprehensive evaluation showing significant improvements over baselines. The deployed interactive application demonstrates practical applicability with sub-second response times and explainable recommendations.

Future work includes: (1) hierarchical clustering for multi-faceted book categorization, (2) temporal modeling to capture evolving reader preferences, (3) integration of user ratings for collaborative filtering, and (4) extension to cross-genre recommendations across the full Goodreads catalog.

## Acknowledgments

This work was completed as the final project for STATS 507: Data Science Analytics using Python at the University of Michigan. The author thanks the course instructors for their guidance throughout this course. Special appreciation to the open-source community for maintaining the Hugging Face ecosystem which made this project possible. Code and data are available at: https://github.com/saisnehasv/mystery-thriller-book-recommendation-system

## References

[1] O. Barkan and N. Koenigstein, "Item2Vec: Neural item embedding for collaborative filtering," in *IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1–6.

[2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 3982–3992.

[3] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002.

[4] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.

[5] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in Artificial Intelligence*, vol. 2009, Article ID 421425, 2009.

[6] "Goodreads 100k Dataset," Hugging Face Datasets, 2024. [Online]. Available: https://huggingface.co/datasets/euclaise/goodreads_100k. [Accessed: Nov. 27, 2024]