

## ГИБРИДНАЯ СИСТЕМА СЕМАНТИЧЕСКОГО ПОИСКА ФИЛЬМОВ НА ОСНОВЕ КОМБИНИРОВАННЫХ ЭМБЕДДИНГОВ

*Захарьев В.А. Тоган И.Д.*

*Белорусский государственный университет информатики и радиоэлектроники,  
г. Минск, Республика Беларусь*

*Научный руководитель: Захарьев В.А. – канд. тех. наук, доцент кафедры СУ*

**Аннотация.** Представлен обзор архитектурных решений для гибридных систем семантического поиска, предназначенных для поиска фильмов по текстовым запросам. Рассмотрены основные компоненты таких систем, включая слои генерации эмбедингов (характеристических векторов), хранения и поиска данных. Проведено экспериментальное сравнение базового подхода (TF-IDF + жанры) и предложенной гибридной системы (RoBERTa + TF-IDF + сущности + жанры). Результаты показывают улучшение качества рекомендаций на 40% по метрике Precision@10 для релевантных результатов.

**Ключевые слова:** гибридный поиск, семантический поиск, эмбединги, RoBERTa, pgvector.

**Введение.** Традиционные системы поиска, основанные на сопоставлении ключевых слов, демонстрируют ограниченную эффективность при работе с семантическими запросами. Современные подходы базируются на использовании методов обработки естественного языка (NLP) и глубокого обучения. Трансформерные модели, такие как BERT и RoBERTa, позволяют кодировать смысл текста в многомерные векторы (эмбединги), где расстояние между векторами отражает семантическую близость. Однако использование только текстовых эмбедингов не позволяет учитывать структурированные характеристики фильма. В данной работе предлагается гибридный подход, объединяющий семантические эмбединги описаний фильмов с векторизованными категориальными признаками. Основной вклад работы заключается в разработке архитектуры, демонстрации эффективности гибридного подхода и количественной оценке улучшения качества поиска.

**Основная часть.** В настоящей работе предлагается гибридный подход, объединяющий текстовые эмбединги описаний фильмов на основе модели RoBERTa, векторизованные категориальные признаки (жанры, режиссёры, актёры) и кластерные признаки названий фильмов. Комбинированный вектор формируется путём конкатенации всех компонентов. Хранение и поиск векторов реализованы на базе PostgreSQL с расширением pgvector.

Предложенная архитектура состоит из пяти слоёв: слой потребления данных (загрузка датасетов TMDb и топ-100 актёров); слой предобработки данных (фильтрация фильмов, извлечение именованных сущностей с помощью spaCy, кластеризация названий методом DBSCAN); слой генерации эмбедингов (вычисление гибридных векторов); слой хранения векторов (PostgreSQL с расширением pgvector); слой поиска и визуализации (веб-интерфейс на Flask).

Гибридный эмбединг формируется конкатенацией трёх групп признаков: категориальные признаки (сценаристы, режиссёры, кластер названия, извлечённые сущности) преобразуются с помощью TF-IDF с параметрами min\_df=2 и max\_df=0.2; жанры кодируются с помощью MultiLabelBinarizer; текстовый обзор обрабатывается моделью RoBERTa-base с усреднением выходных эмбедингов токенов.

Для хранения эмбедингов используется PostgreSQL с расширением pgvector. Поиск ближайших соседей реализован через оператор косинусного расстояния <=>. Веб-интерфейс на Flask поддерживает поиск по названию фильма. При отсутствии точного совпадения пользователю выводятся подсказки из 20 случайных названий.

Для оценки качества предложенного подхода проведено сравнение двух

конфигураций системы: базовый подход (Baseline) – используются только TF-IDF векторы для категориальных признаков и жанры, без RoBERTa и извлечения сущностей; основная система (Hybrid) – полная версия с RoBERTa, сущностями spaCy и TF-IDF.

Оценка проводилась на 6 фильмах, представляющих различные жанры: анимационные комедии (Finding Nemo, Minions, Khumba), историческая анимация (Pocahontas) и драмы/криминал (American History X, American Gangster). Для каждого фильма оценивались топ-10 рекомендаций по шкале от 0 до 2: 2 балла – фильм очень похож по жанру, тематике, сюжету; 1 балл – частичное пересечение (общий жанр, но разная стилистика); 0 баллов – фильм не похож.

Основные метрики оценки: Precision@10 (хорошие) – доля фильмов с оценкой 2 среди первых 10 рекомендаций; Precision@10 (хорошие+средние) – доля фильмов с оценкой 1 или 2; Средняя релевантность (0–2) – средняя арифметическая оценка по первым 10 рекомендациям.

Результаты для базового подхода представлены в таблице 1.

Таблица 1 – Результаты базового подхода (TF-IDF + жанры)

| P@10 (2) | P@10 (>=1) | Ср. рел. |
|----------|------------|----------|
| 0.70     | 1.00       | 1.70     |
| 0.70     | 1.00       | 1.70     |
| 0.60     | 0.90       | 1.50     |
| 0.50     | 0.80       | 1.30     |
| 0.10     | 0.70       | 0.80     |
| 0.30     | 0.60       | 0.90     |
| 0.48     | 0.83       | 1.32     |

Результаты для гибридной системы представлены в таблице 2.

Таблица 2 – Результаты гибридной системы (RoBERTa + TF-IDF + сущности + жанры)

| P@10 (2) | P@10 (>=1) | Ср. рел. |
|----------|------------|----------|
| 0.80     | 1.00       | 1.80     |
| 0.80     | 1.00       | 1.80     |
| 0.70     | 1.00       | 1.70     |
| 0.70     | 0.90       | 1.60     |
| 0.40     | 0.90       | 1.30     |
| 0.60     | 0.90       | 1.50     |
| 0.67     | 0.95       | 1.62     |

Сравнение средних значений метрик представлено в таблице 3.

Таблица 3 – Сравнение базового и гибридного подходов

| Метрика                | Базовый подход | Гибридная система | Улучшение |
|------------------------|----------------|-------------------|-----------|
| P@10 (хорошие)         | 0.48           | 0.67              | 40,00%    |
| P@10 (хорошие+средние) | 0.83           | 0.95              | 14,00%    |
| Средняя релевантность  | 1.32           | 1.62              | 23,00%    |

Наибольший прирост качества наблюдается для сложных жанров (драма, криминал). Для фильма American History X показатель P@10 (хорошие) увеличился с 0.10 до 0.40, а для American Gangster – с 0.30 до 0.60. Анимационные фильмы изначально демонстрировали высокие результаты, однако гибридная система также обеспечила стабильное улучшение.

**Заключение.** Разработанная гибридная система позволяет учитывать как семантику текстовых описаний, так и структурированные метаданные фильмов. Экспериментальная оценка на 6 фильмах различных жанров показала улучшение качества рекомендаций на 40% по метрике Precision@10 для релевантных результатов по сравнению с базовым подходом, использующим только TF-IDF и жанры. Наибольший эффект достигнут для сложных жанров (драма, криминал), где добавление семантических эмбедингов и именованных сущностей

значительно повысило релевантность рекомендаций. Дальнейшие направления развития включают использование более крупных языковых моделей, добавление коллаборативной фильтрации и разработку механизмов объяснения рекомендаций.

### ***Список литературы***

1. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // *Proceedings of NAACL-HLT*. – 2019. – P. 4171-4186.
2. Liu, Y. RoBERTa: A Robustly Optimized BERT Pretraining Approach / Y. Liu [et al.] // *arXiv preprint arXiv:1907.11692*. – 2019.
3. Neumann, M. ScaNN: Efficient Vector Similarity Search / M. Neumann, P. R. King // *Proceedings of VLDB*. – 2020.
4. Honnibal, M. spaCy: Industrial-strength Natural Language Processing / M. Honnibal, I. Montani // <https://spacy.io>. – 2017.
5. Johnson, J. pgvector: Open-source vector similarity search for PostgreSQL / J. Johnson // <https://github.com/pgvector/pgvector>. – 2021.

UDC 004.032.26

## **HYBRID SEMANTIC SEARCH SYSTEM FOR MOVIES BASED ON COMBINED EMBEDDINGS**

*Zakharayev V.A. Togan I.D.*

*Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus*

*Alekseev V.F. – Cand. of Sci., associate professor, associate professor of the department of ICSD*

**Annotation.** An overview of architectural solutions for hybrid semantic search systems designed for movie search based on text queries is presented. The main components of such systems are considered, including the layers of embedding generation, data storage and retrieval. An experimental comparison of the baseline approach (TF-IDF + genres) and the proposed hybrid system (RoBERTa + TF-IDF + entities + genres) is carried out. The results show a 40% improvement in recommendation quality according to the Precision@10 metric for relevant results.

**Keywords:** hybrid search, semantic search, embeddings, RoBERTa, pgvector.