

# ГИБРИДНАЯ СИСТЕМА СЕМАНТИЧЕСКОГО ПОИСКА ФИЛЬМОВ, НА ОСНОВЕ КОМБИНИРОВАННЫХ ЭМБЕДДИНГОВ

Традиционные системы поиска по ключевым словам не справляются с семантическими запросами. Трансформерные модели (BERT, RoBERTa) кодируют смысл текста в эмбединги, но игнорируют структурированные атрибуты фильмов (жанры, актёров). Предложена гибридная система, объединяющая семантические эмбединги с векторизованными категориальными признаками. Экспериментально показано улучшение Precision@10 на 40% по сравнению с TF-IDF + жанры.

## Введение

Поиск по ключевым словам неэффективен для запросов вроде «добрый мультфильм про рыбок». Трансформеры решают эту проблему, но игнорируют жанры и актёрский состав. Предлагается гибридный подход: комбинированный эмбединг из текста и категориальных признаков.

## I. Архитектура

**Потребление данных:** API TMDb (фильмы, описания, жанры, топ-100 актёров).

**Предобработка:** извлечение сущностей (spaCy), кластеризация DBSCAN для группировки вариантов имён и фильтрации выбросов (не требует задавать число кластеров).

**Генерация гибридных эмбедингов:** конкатенация:

- TF-IDF для актёров и режиссёров;
- MultiLabelBinarizer для жанров;
- RoBERTa для текстовых обзоров ([CLS]-токен).

Итоговая размерность определяется содержанием обучающей выборки.

**Хранение и поиск:** PostgreSQL + pgvector (косинусное сходство), интерфейс на Flask.

## II. Эксперимент

Сравнивались Baseline (TF-IDF + жанры) и Hybrid (RoBERTa + TF-IDF + сущности + жанры). Оценка на 6 фильмах по шкале 0–2.

Тоган Илья Дмитриевич, студент кафедры систем управления БГУИР, ilya.togan@gmail.com.

Научный руководитель: Захарьев Вадим Анатольевич, кандидат технических наук, доцент кафедры систем управления БГУИР, zakharyev@bsuir.by.

| Фильм              | Baseline P@10(2) | Hybrid P@10(2) | Hybrid P@10(≥1) |
|--------------------|------------------|----------------|-----------------|
| Finding Nemo       | 0.70             | 0.80           | 1.00            |
| Minions            | 0.70             | 0.80           | 1.00            |
| Khumba             | 0.60             | 0.70           | 1.00            |
| Pocahontas         | 0.50             | 0.60           | 0.90            |
| American History X | 0.10             | 0.40           | 0.80            |
| American Gangster  | 0.30             | 0.60           | 0.90            |
| Среднее            | 0.48             | 0.67           | 0.95            |

| Метрика                  | Baseline | Hybrid | Улучшение |
|--------------------------|----------|--------|-----------|
| P@10 (релевантность 2)   | 0.48     | 0.67   | +40%      |
| P@10 (≥1)                | 0.83     | 0.95   | +14%      |
| Ср. релевантность топ-10 | 1.32     | 1.62   | +23%      |

**Анализ:** Наибольший прирост для сложных жанров (драма, криминал). Для *American History X* P@10(2) вырос с 0.10 до 0.40, для *American Gangster* – с 0.30 до 0.60.

## III. Выводы

Гибридная система объединяет семантику текста и структурированные метаданные, обеспечивая улучшение Precision@10 на 40% (P@10(≥1)=0.95). DBSCAN эффективно фильтрует шумные сущности.

Дальнейшие направления: более крупные модели (Llama), коллаборативная фильтрация, объяснение рекомендаций.

## Список литературы

1. Devlin, J. BERT / J. Devlin [et al.] // NAACL-HLT. – 2019. – P. 4171-4186.
2. Liu, Y. RoBERTa / Y. Liu [et al.] // arXiv:1907.11692. – 2019.
3. Honnibal, M. spaCy / M. Honnibal, I. Montani. – <https://spacy.io>, 2017.
4. Johnson, J. pgvector / J. Johnson. – <https://github.com/pgvector/pgvector>, 2021.