

МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ
Н.Г.ЧЕРНЫШЕВСКОГО»**

Кафедра математической теории упругости и биомеханики

**Разработка веб-сервиса рекомендаций на основе
интеллектуального анализа данных**

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 442 группы

направления 09.03.03 «Прикладная информатика (прикладной бакалавриат)»

механико-математического факультета

Харламова Александра Алексеевича

Научный руководитель
доцент, к.ю.н.

дата, подпись

Р. В. Амелин

Зав. кафедрой
д.ф.-м.н., профессор

дата, подпись

Л.Ю. Коссович

Саратов 2019

ВВЕДЕНИЕ

Бакалаврская работа посвящена исследованиям в области рекомендательных систем. Задача рекомендательной системы - предоставить пользователю рекомендации в выборе интересующих его предметов (фильмов, музыки, новостей, товаров и прочего), которые ему наиболее интересны в данный момент времени.

Сегодня проблема рекомендаций товаров и услуг потенциальным потребителям или покупателям в современном мире выходит на первый план. Она состоит в том, чтобы позволить каждому пользователю выбрать среди огромного количества предметов те, которые ему больше подходят или соответствуют его индивидуальным вкусам и предпочтениям.

Для построения системы рекомендации необходим набор данных о людях и их оценках предпочтений рассматриваемого предмета. Этот набор данных должен содержать пользователя оценившего предмет, название рассматриваемого предмета и его оценки. Для более сложных систем рекомендаций используются дополнительные данные, которые могут содержать информацию о тегах назначаемых предмету, кратких характеристических параметров товара или прочая полезная информация.

Целью бакалаврской работы является анализ и разработка веб-сервиса по выработке рекомендаций на примере кинофильмов.

Задачи выполняемой работы являются:

1. Исследование популярных подходов и методов к выработке рекомендаций;
2. Разработка веб-сервиса по рекомендации кинофильмов, на основе изученных методов;
3. Создание дружественного интерфейса для веб-сервиса по рекомендации кинофильмов.

Структура и объем работы. Бакалаврская работа состоит из следующих частей: введение, 3 глав, заключения и списка используемых источников, включающих 21 наименование. Работа изложена на 51 странице, не включая приложения, содержит 23 рисунка.

Основное содержание работы. Во введении описывается актуальность использования рекомендательных систем, а также поставленная цель и задачи данной бакалаврской работы.

Первая глава состоит из 7 подразделов и описывает методы рекомендательных систем. Было выделены следующие 4 вида методов: коллаборативной фильтрации, анализирующие содержимое объектов, основанные на знаниях и основанные на интеллектуальном анализе данных.

Подразделы с 1.2 по 1.6 описывают методы рекомендательных систем подробно. Кратко представим принцип работы каждого из рассмотренных методов.

Коллаборативная фильтрация – это один из методов, использующие известные предпочтения (оценки) одной группы пользователей для прогнозирования неизвестных предпочтений другой группы пользователей [1]. Оценки могут быть как явными (пользователь явно указывает, на сколько баллов он оценивает объект, например от 1 до 5 баллов), так и неявными (например, количество просмотров некоторого фильма). Основная идея данного метода заключается в сравнении между собой интересов различных пользователей или объектов на основе их оценок [2]. Для сравнения данных о пользователях или предметах используются коэффициенты подобия. Используются следующие коэффициенты подобия: евклидово расстояние, коэффициент корреляции Пирсона и коэффициент Танимото.

Оценка по евклидову расстоянию – это один из самых простых методов вычисления подобия. Евклидово расстояние определяет расстояние между двумя точками в многомерном пространстве [5].

Второй способ чуть сложнее в определении степени схожести интересов его вычисляет коэффициент корреляции Пирсона. Коэффициент корреляции - это мера того, насколько хорошо два набора данных ложатся на прямую [6].

Последний третий способ определения меры схожести двух множеств называется коэффициентом Танимото. Данный способ рекомендуется использовать при небольшом количестве вариантов оценок не больше 2-3 (например 1 – понравилось, а 0 – не понравилось) [7].

Когда коллаборативная фильтрация не дает результата, приходится использовать информацию об объектах. Например, был снят новый фильм – продолжение какого-то фильма. Возникает вопрос, а кому его порекомендовать. Возможно, есть смысл предложить новый фильм тем, кто уже посмотрел первую часть фильма. Каждый объект в системе рекомендаций имеет название и другую текстовую информацию, которой можно воспользоваться. На основе информации, которые пользователи предоставляют в своём профиле, можно делать выводы, интересен ли некоторый предмет (фильм) пользователю или нет. Суть данного метода заключается в нахождении предметов соответствующих предпочтениям пользователя из его профиля. В профиле пользователя может содержаться информация о его совершенных покупках, интересующих его товаров или любых других его предпочтений, которая будет полезна для составления индивидуальных рекомендаций для пользователя.

Методы коллаборативной фильтрации и анализе содержимого хорошо работают, тогда когда пользователь часто посещает сервис и отслеживаются его следующие действия: покупки, просмотры, оценки и так далее. Но если это происходит редко (например, покупка дома, автомобиля или компьютера), то тогда эти методы не помогут, потому как система не имеет нужного количества информации о пользователе. Для решения данной задачи используются методы, основанные на знаниях. Их разделяют на использование жестких ограничений и выбор близких объектов. Идея обоих методов том, что пользователь формирует свои требования к предмету рассмотрения, а система пытается

найти. В первом варианте ищутся и рекомендуются только те объекты, которые точно соответствуют всем критериям пользователя. Во втором варианте ищутся объекты с близкими характеристиками предъявленных требований (с применением некоторых мер близости) пользователя системы.

На основе рассмотренных выше методов, есть возможность их совмещения, такая система рекомендаций носит название гибридной. Всего выделяют три популярных вида гибридных систем: монолитной, параллельной и конвейерной организацией.

Системы с монолитной организацией представляют собой один модуль, включающий в себя различные методы обработки имеющихся данных. В большинстве случаев гибридные системы такого рода реализуют коллаборативную фильтрацию, усиленную особенностями содержимого предметов. К примеру, если алгоритм коллаборативной фильтрации рекомендует пользователю несколько книг, то среди них можно выбрать наиболее подходящие с помощью сравнения жанров книг с любимыми жанрами пользователя.

Системы с параллельной организацией принимают исходные данные на каждый модуль системы. После того как все подсистемы выдадут ответ, начинает работать агрегирующий модуль. Существует три основных подхода для выбора наиболее релевантных рекомендаций:

- смешивание всех рекомендаций и выдается результат пользователю по некоторой заданной формуле;
- для каждого объекта подсчитывается сумма рейтингов, полученными параллельными модулями и выводятся объекты с наибольшим значением рейтинга;
- выберется одна подходящая система рекомендаций, реализованная в одном из модулей для каждой конкретной ситуации.

Гибридная система с конвейерной организацией делит всю работу на этапы, которые выполняются последовательно. Результат работы очередного этапа используется в последующем.

Методы, основанные на интеллектуальном анализе данных, соответствуют широко распространенному термину Data Mining, который часто переводится как добыча данных, глубинный анализ данных, извлечение знаний, раскопка знаний в базах данных [3].

Основу методов data mining составляют следующие методы: классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, генетических алгоритмов, эволюционного программирования.

Например, применение методов нейросетевого моделирования открывает новые возможности в части повышения эффективности анализа данных. Искусственная нейронная сеть способна автоматически кластеризовать типы пользователей и динамически выделять новые их типы со временем. Это позволяет при регистрации нового пользователя очень быстро провести анализ его поведения, отнести его к какой-либо группе и начать формировать для него персональные рекомендации на основе предпочтений его группы пользователей. Системе работающей на основе сбора статистики понадобилось бы на это гораздо больше времени, нейронная сеть же начнет выявлять типизацию по первым же шагам пользователя.

Система, построенная на нейронной сети, может быть интегрирована в существующие системы статистического анализа сервиса, при этом обучение нейронной сети может быть основано на уже собранных статистических данных. Нейронная сеть дает хорошие рекомендации на основе предпочтений посетителей, помогает отслеживать результаты работы сервиса и принимать обоснованные решения о повышении эффективности функционала сервиса.

Также в данной главе произведен краткий сравнительный анализ недостатков и достоинств рассмотренных методов.

Вторая глава состоит из 6 подразделов, где описывается процесс разработки веб-сервиса. В первом подразделе перечислены следующие этапы разработки:

- Подготовка данных о пользователях и объектах (фильмов);
- Реализация метода коллаборативной фильтрации;
- Реализация метода, на основе содержимого объекта (фильма);
- Реализация искусственной нейронной сети;
- Обучение искусственной нейронной сети.

На этапе подготовки данных собирается вся необходимая информация для рекомендательного веб-сервиса. Для этого необходимы данные о пользователях и объектах. Сайт проекта MovieLens бесплатно предоставляет различные наборы данных о кинокритиках. Выбран небольшой набор данных, который содержит: 100 000 оценок пользователей, 3 600 тегов и 9 000 фильмов оцененные более 600 пользователями. Так же необходимо собрать больше информации о кинофильмах, для этого необходимо создать поискового робота (паука).

Поисковый робот («веб-паук», «веб-краулер») - программа, являющаяся составной частью поисковой системы и предназначенная для перебора страниц Интернета с целью занесения информации о них в базу данных поисковика [4].

После написания кода для паука направляем его на сайт «Movie DB» (ссылка на сайт: <https://www.themoviedb.org>), процесс сбора информации занимает весьма длительный срок времени, так как нам необходимо извлечь описания для фильмов с более 9 000 названиями.

Для хранения информации страниц с кинофильмами создадим базу данных для хранения текстового индекса. Такой индекс представляет собой список слов, с каждым из которых ассоциировано множество документов, где это слово встречается, и место вхождения слова в документе. На данном этапе анализируется только текст на странице, игнорируя при этом все нетекстовые элементы. Кроме того, индексируются только сами слова, а все знаки

препинания игнорируются. Схема построения индекса будет состоять из пяти таблиц.

Первым методом разрабатывается коллаборативная фильтрация. Для того чтобы предложить фильм критику, нужно просмотреть какие фильмы понравились кинокритикам с похожими вкусами, и выбрать из них те, которые еще не видел. Для ранжирования результатов, вычисляется взвешенная сумма оценок всех критиков. Выдаваться рекомендации будут, путем использования словаря данных о схожести образцов без обращения ко всему набору данных. Результат работы представлен на рисунке 1.



```
Python 3.6.3 Shell
File Edit Shell Debug Options Window Help
5.0 Mamma Mia: Here We Go Again! (2018)
5.0 Dogman (2018)
5.0 Incredibles 2 (2018)
5.0 Sherlock - A Study in Pink (2010)
5.0 Alpha (2018)
5.0 Love (2017)
5.0 Game Night (2018)
5.0 Insidious: The Last Key (2018)
5.0 Die Frauen von Ravensbrück (2005)
5.0 Christina P: Mother Inferior (2017)
5.0 Coco (2017)
5.0 Bonifaces Holiday (1965)
5.0 Black Mirror
5.0 Bliss (2012)
5.0 It (2017)
5.0 Logan Lucky (2017)
5.0 The Godfather Trilogy: 1972-1990 (1992)
5.0 Late Night with Conan O'Brien: The Best of Triumph the Insult Comic Dog (2004)
5.0 Mystère à la Tour Eiffel (2015)
5.0 Munna bhai M.B.B.S. (2003)
Ln: 112 Col: 4
```

Рисунок 1 – Выдача рекомендаций пользователю под номером 87

Вторым методом разрабатывается на основе содержимого объекта. Этот метод эффективен, когда коллаборативная фильтрация не дает результата, тогда приходится использовать информацию об объектах.

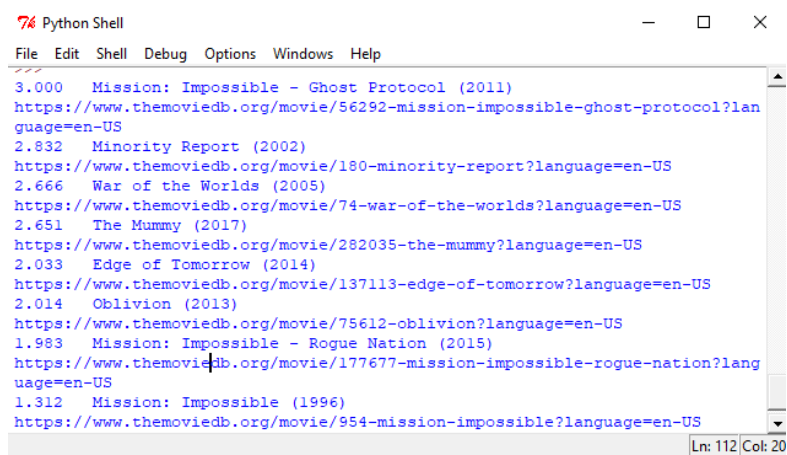
Каждый фильм теперь имеет не только название и оценку, но и краткое описание в виде текста. Если пользователь в системе указал информацию о себе и своих интересах: любимые жанры, киноактеры, режиссеры и тому подобное. То можно сделать выводы, интересен ли конкретный фильм пользователю.

На предыдущих этапах созданный сетевой паук собрал краткую информацию о каждом фильме и занес в базу данных в виде документа, которая будет полезна для построения рекомендаций данным способом. Так как набор документов велик, то предстоит просеять большое количество документов, не относящихся к информации действительно полезной. Для

решения этой проблемы, необходимо присвоить документам ранг относительно данного запроса и вернуть их в порядке убывания рангов. Были рассмотрены следующие способы вычисления у страницы ранга:

- Расчёт ранга по частоте слов;
- Расчёт ранга по расстоянию поисковых слов от начала документа;
- Расчёт расстояния между словами.

Объединив все три метрики поиска документа, получаем следующий результат работы программы, представленный на рисунке 2.



```
Python Shell
File Edit Shell Debug Options Windows Help
3.000 Mission: Impossible - Ghost Protocol (2011)
https://www.themoviedb.org/movie/56292-mission-impossible-ghost-protocol?language=en-US
2.832 Minority Report (2002)
https://www.themoviedb.org/movie/180-minority-report?language=en-US
2.666 War of the Worlds (2005)
https://www.themoviedb.org/movie/74-war-of-the-worlds?language=en-US
2.651 The Mummy (2017)
https://www.themoviedb.org/movie/282035-the-mummy?language=en-US
2.033 Edge of Tomorrow (2014)
https://www.themoviedb.org/movie/137113-edge-of-tomorrow?language=en-US
2.014 Oblivion (2013)
https://www.themoviedb.org/movie/75612-oblivion?language=en-US
1.983 Mission: Impossible - Rogue Nation (2015)
https://www.themoviedb.org/movie/177677-mission-impossible-rogue-nation?language=en-US
1.312 Mission: Impossible (1996)
https://www.themoviedb.org/movie/954-mission-impossible?language=en-US
Ln: 112 Col: 20
```

Рисунок 2 - Поиск «Tom Cruise THRILLER» тремя метриками

Последний третий метод рекомендаций построен на основе искусственной нейронной сети. В данном методе строится искусственная нейронная сеть, которая обучается на основе действий пользователей. В случае поиска нужного фильма, пользователь сообщает насколько ему понравился результат поиска, переходя по одному из результатов и игнорируя все остальное. Обученная нейронная сеть будет использовать результаты поиска прошлых действий пользователя, для упорядочивания результатов.

Для сохранения результатов обучения нейронной сети создается база данных для хранения несколько таблиц, которые будут хранить скрытый уровень нейронной сети и связи между ними.

Обучение нейронной сети достигается предъявлением ей реальных примеров запросов, найденных результатов и выбора пользователя. Для этого необходимо изменять веса связей между узлами, так чтобы нейронная сеть

распознала, как выглядит правильный ответ. Такой алгоритм обучения носит название «обучение с обратным распространением», так как в процессе подстройки весов он продвигается по сети в обратном направлении. В результате обучения нейронной сети получаем результат на рисунке 3.

```
Python Shell
File Edit Shell Debug Options Windows Help
0.999999996276670 Heat (1995)
0.999999990863418 Assassins (1995)
0.999999761385945 Taxi Driver (1976)
0.999999283966775 Get Shorty (1995)
0.994042845338602 Crossing Guard, The (1995)
0.0 Money Train (1995)
0.0 Bad Boys (1995)
0.0 Amateur (1994)
>>>
```

Рисунок 3 – Поиск «crime thriller» обученной нейронной сетью

Третья глава бакалаврской работы состоит из двух подразделов, в которых описан принцип создания прототипа сайта по рекомендации фильмов. В качестве результата предложен интерфейс для пользователя главной и основной страницы веб-сервиса. Пример основной страницы сайта продемонстрирован на рисунке 4.

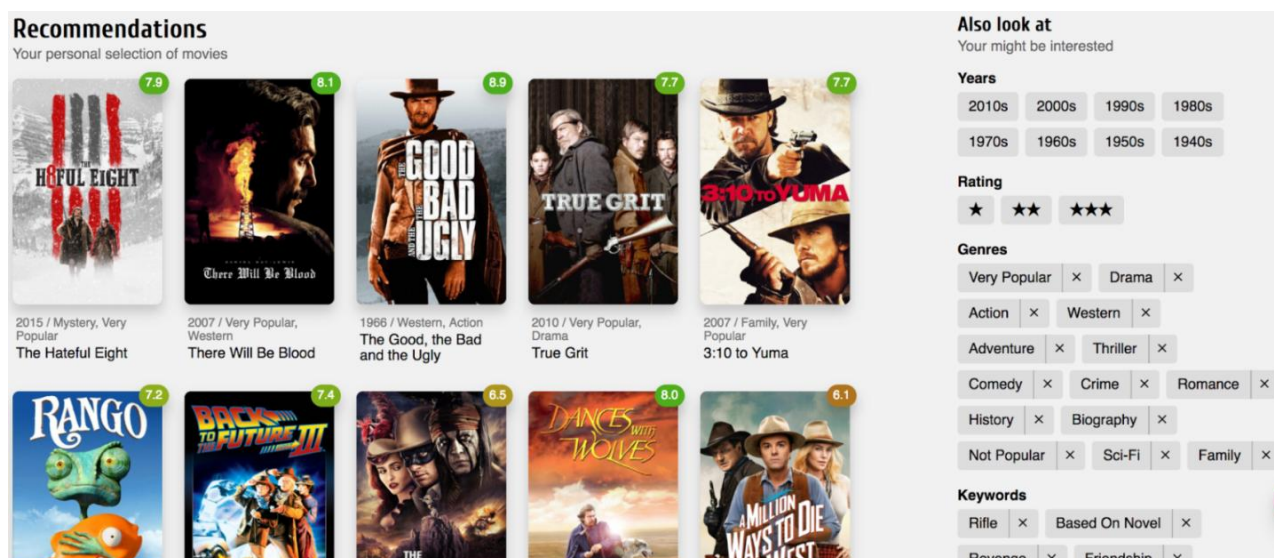


Рисунок 4 – Макет основной страницы по рекомендации фильмов

ЗАКЛЮЧЕНИЕ

В данной работе был произведен анализ различных методов систем рекомендаций, а так же создан веб-сервис по рекомендации кинофильмов. В качестве данных о пользователях были взяты данные с сайта проекта MovieLens, а для сбора более подробной информации о фильмах создали паука и направили его на сайт Movie Database.

В разработанном веб-сервисе использовали следующие методы по выработке рекомендаций: коллаборативной фильтрации, на основе содержимого объекта (фильма) и обученной нейронной сети. Метод коллаборативной фильтрации предлагает рекомендации при достижении достаточного количества оценок предпочтений пользователя, а значит, не даст никаких рекомендаций новым пользователям. Для решения этой проблемы применяем методы систем рекомендаций основанной на содержании объектов. Пользователь заполняет в своем профиле информацию о себе: любимые фильмы, актеры, режиссеры и тому подобное. На основе этих данных ищутся те объекты (фильмы), которые содержат ключевые слова из профиля пользователя. Метод рекомендаций с искусственной нейронной сети позволяют пользователю выбрать интересующие жанры и теги кинофильмов, а затем получить рекомендацию. Все три применяемых метода позволяют максимально качественно вырабатывать рекомендации для пользователей, что способствует укреплению репутации для веб-сервиса.

В третьей части работы разработан прототип дружеского интерфейса для веб-сервиса. С которым пользователю будет максимально комфортно работать с сервисом рекомендаций кинофильмов.

По итогам данной работы была достигнута поставленная цель, после выполнения всех поставленные задач.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 Цылова, Е.Г. Коллаборативная фильтрация как метод учёта интересов пользователей в рекламе продукции / Е.Г. Цылова, Е.Я. Экгауз // Вестник уральского института экономики, управления и права – 2008. –С. 89-94.
- 2 Сегаран, Т. Програмируем коллективный разум / Т. Сегаран. – Пер. с англ. – СПб: Символ-Плюс, 2008. – 368 с.
- 3 Барсегян А. А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP: учеб. Пособие / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – СПб.: БХВ-Петербург, 2007
- 4 Поисковый робот [Электронный ресурс] : [сайт]. - URL: https://ru.wikipedia.org/wiki/Поисковый_робот (дата обращения 25.05.2019) - Загл. с экрана. - Яз. рус.
- 5 Евклидова метрика [Электронный ресурс] : [сайт]. - URL: https://ru.wikipedia.org/wiki/Евклидова_метрика (дата обращения 25.05.2019) - Загл. с экрана. - Яз. рус.
- 6 Корреляция [Электронный ресурс] : [сайт]. - URL: <https://ru.wikipedia.org/wiki/Корреляция> (дата обращения 25.05.2019) - Загл. с экрана. - Яз. рус.
- 7 Степень схожести строк (коэффициент Танимото) [Электронный ресурс] : [сайт]. - URL: <https://grishaev.me/2012/10/05/1> (дата обращения 25.05.2019) - Загл. с экрана. - Яз. рус.
- 8 Искусственная нейронная сеть [Электронный ресурс] : [сайт]. - URL: https://ru.wikipedia.org/wiki/Искусственная_нейронная_сеть (дата обращения 25.05.2019) - Загл. с экрана. - Яз. рус.
- 9 Гомзин, А.Г. Системы рекомендаций: обзор современных подходов / А.Г. Гомзин, А.В. Коршунов // Народное образование. Педагогика : научная статья – 2011. –С. 401-417.
- 10 Барсегян А. А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP: учеб. Пособие / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – СПб.: БХВ-Петербург, 2007