МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

**«МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

**(МОСКОВСКИЙ ПОЛИТЕХ)**

Факультет информационных технологий

Кафедра «Прикладная информатика»

Форма обучения: очная

**ИНДИВИДУАЛЬНОЕ ЗАДАНИЕ**

по дисциплине

«Методы прогнозирования»

Тема: *«Прогнозирование с помощью искусственного интеллекта. Основные направления исследования. Примеры».*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Группа |  | 211-365 |  |  |
| Студент(ы) |  | *А.Д. Заюров*  *И.А. Садаков*  *А.Т. Сыдыкова*  *А.Р. Халитова*  *Н.А. Чугаев*  *К.А. Шатаров* |  |  |
| Преподаватель,  к.п.н., доцент |  | *\_\_\_\_\_\_\_\_\_\_\_\_\_* |  | Н.И. Царькова |
| Оценка работы |  | \_\_\_\_\_\_\_\_\_\_\_\_\_ |  |  |
| Дата |  |  |  | 8 октября 2023 |

МОСКВА 2023

**ОГЛАВЛЕНИЕ**

[ВВЕДЕНИЕ 3](#_Toc147674289)

[ГЛАВА 1. ОСНОВНЫЕ ПОНЯТИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА 4](#_Toc147674290)

[1.1. Определение искусственного интеллекта 4](#_Toc147674291)

[1.2. Виды и технологии искусственного интеллекта 6](#_Toc147674292)

[ГЛАВА 2. ОСНОВНЫЕ НАПРАВЛЕНИЯ ИССЛЕДОВАНИЯ 9](#_Toc147674293)

[2.1. Категории моделей предиктивной аналитики 10](#_Toc147674296)

[2.2. Анатомия рекомендательных систем 11](#_Toc147674297)

[2.2.1. Типы рекомендательных систем 12](#_Toc147674298)

[2.3. Прогнозирование с помощью нейронных сетей 15](#_Toc147674299)

[ГЛАВА 3. ПРИМЕРЫ ПРОГНОЗИРОВАНИЯ С ПОМОЩЬЮ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА 19](#_Toc147674300)

[3.1. Трансформеры 19](#_Toc147674303)

[3.1.1. Позиционные эмбединги 25](#_Toc147674304)

[3.2. Подходы для рекомендательных систем 29](#_Toc147674305)

[3.2.1. Матричная факторизация для рекомендаций 29](#_Toc147674306)

[3.2.2. Матричная факторизация для явной обратной связи 29](#_Toc147674307)

[3.2.3. Контент-ориентированный подход 30](#_Toc147674308)

[3.2.4. Гибридный подход: LightFM 31](#_Toc147674309)

[3.3. Personal Insights 33](#_Toc147674310)

[ЗАКЛЮЧЕНИЕ 35](#_Toc147674311)

[СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ 36](#_Toc147674312)

## ВВЕДЕНИЕ

С развитием научно-технического прогресса увеличивается кол-во собираемой информации и на данный момент не представляется возможным человеку быстро и безошибочно её анализировать для составления прогнозов. Одним из способов решить данную проблему является использование искусственного интеллекта.

Прогнозирование с помощью искусственного интеллекта - это процесс использования алгоритмов и методов искусственного интеллекта для предсказания будущих событий или значений на основе доступных данных. Применение искусственного интеллекта в прогнозировании позволяет автоматизировать и усовершенствовать процесс принятия решений, основанный на анализе больших объемов данных.

Искусственный интеллект включает в себя различные подходы и технологии, такие как машинное обучение, нейронные сети, генетические алгоритмы и другие. Они позволяют извлекать сложные закономерности из данных, обнаруживать скрытые связи и тренды, и основываясь на этом, делать предсказания.

Прогнозирование с помощью искусственного интеллекта может применяться в различных областях, включая финансы, экономику, здравоохранение, логистику, маркетинг и другие. Например, в финансовой сфере искусственный интеллект может использоваться для прогнозирования цен на акции, позволяя инвесторам принимать более осознанные решения.

## ГЛАВА 1. ОСНОВНЫЕ ПОНЯТИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

* 1. Определение искусственного интеллекта

Интеллект (от лат. intellectus – ощущение, восприятие, разумение, понимание, понятие, рассудок), или ум – качество психики, состоящее из способности приспосабливаться к новым ситуациям, способности к обучению и запоминанию на основе опыта, пониманию и применению абстрактных концепций и использованию своих знаний для управления окружающей средой.

Другими словами, интеллект – это общая способность к познанию и решению трудностей, которая объединяет все познавательные способности человека: ощущение, восприятие, память, представление, мышление, воображение.

Существует ряд определений искусственного интеллекта, которые в той или мере дополняют и уточняют друг друга. В начале 1980-х гг. учёные в области теории вычислений Барр и Файгенбаум предложили следующее определение искусственного интеллекта (ИИ).

Искусственный интеллект – это область информатики, которая занимается разработкой интеллектуальных компьютерных систем, то есть систем, обладающих возможностями, которые мы традиционно связываем с человеческим разумом, – понимание языка, обучение, способность рассуждать, решать проблемы и т.д [1].

Дж.Ален считает, что ИИ – это наука о создании машин, решающих задачи, которые могут решать люди. В трактовке М.Т.Джонса ИИ – это процесс создания машин, способных действовать таким образом, что будут восприниматься человеком как разумные.

Психологи (Никерсон, Перкинс и Смит, 1985) перечисляют такие особенности интеллекта:

1. Способность классифицировать паттерны, т.е. разделять неидентичные стимулы на классы.

2. Способность к адаптивному изменению поведения – к обучению.

3. Способность к дедуктивному мышлению, т.е. умение делать выводы из имеющихся посылок.

4. Способность к индуктивному мышлению – к обобщению. Это означает способность выходить за пределы данной информации, т.е. порождать новое знание.

5. Способность разрабатывать и использовать концептуальные модели. Это означает, что у человека имеется некоторая модель мира – представление об его устройстве. Данную модель человек использует для понимания, интерпретации и предсказания событий.

6. Способность к пониманию. Это означает умение видеть отношения в задачах и оценивать их значение для решения задач. Оценка понимания – одна из самых неуловимых проблем в изучении интеллекта.

В качестве рабочего определения примем определение интеллектуальной деятельности, данное В.К.Финном [77-80]:

1. Способность к организации знаний по их существенности и установлению отношения релевантности между имеющимися данными и знаниями.

2. Способность использовать логику в широком смысле этого термина (нахождение следствий, проведение доказательства и т.д.).

3. Наличие рефлексии, т.е. возможность оценивать результаты своей деятельности.

4. Обучение новым знаниям и коррекции ранее приобретенных знаний.

5. Умение порождать гипотезы.

6. Наличие познавательного любопытства.

Более коротко, под интеллектом можно понимать способность решать задачи (в том числе неформализованные), для которых нет стандартных, заранее известных методов решения. При этом важную роль играют приобретение, запоминание и целенаправленное преобразование знаний в процессе обучения на опыте и адаптации к внешним условиям.

Исходя из этого определения, можно трактовать ИИ (artificial intelligence – AI) как свойство автоматических систем брать на себя отдельные функции интеллекта человека, например, выбирать и принимать решения на основе ранее полученного опыта и рационального анализа внешних воздействий.

Существуют и поведенческие (функциональные) определения интеллекта. Таким определением является известный тест Тьюринга [43], суть которого заключается в том, что интеллектуальная система может достаточно долго поддерживать разговор с человеком, оставаясь при этом неузнанной. Такой же точки зрения придерживался и А.Н.Колмогоров. Он полагал, что любая материальная система, с которой можно достаточно долго обсуждать проблемы науки, литературы и искусства, обладает интеллектом.

Основные свойства искусственного интеллекта:

* понимание языка;
* обучение;
* способность мыслить;
* способность действовать.

Искусственный интеллект применим в тех областях, где есть большой объём накопленных данных. Они бывают разных типов. Когда есть несколько типов данных, с ними справиться один аналитик. Искусственный интеллект оказывается эффективным, когда параметров тысячи, причём часть из них неструктурированные.

* 1. Виды и технологии искусственного интеллекта

На сегодняшний день существует четыре основных вида ИИ:

* реактивные машины - системы ИИ, не имеющие памяти и решающие только определенные задачи. Они не способны формировать воспоминания и использовать полученный ранее опыт для выполнения своих функций;
* ограниченная память - системы с памятью, основанной на прошлом опыте. Однако этот опыт не сохраняется и не компилируется в библиотеке информации ИИ;
* теория разума - системы, понимающие человеческие эмоции и намерения, обладающие социальным интеллектом и участвующие в командной работе;
* самосознание - у систем ИИ этого типа формируется представление о себе, благодаря чему они полностью имитируют интеллект человека.

При создании ИИ чаще всего используются технологии:

* машинного обучения. Машинное обучение - это способность компьютера с системой ИИ принимать решения по результатам обработки данных, не придерживаясь четких схем и правил. Иными словами, машина становится способна к поиску закономерностей в сложных для человека задачах с большим числом параметров, к обнаружению точных ответов и верному прогнозированию;
* глубокого обучения - подраздела машинного обучения, позволяющего обнаруживать закономерности в огромным массивах информации (Big Data). Обработкой данных в технологии глубокого обучения занимаются искусственные нейронные сети (ИНС), созданные по аналогии с биологическими нейронными сетями. ИНС предназначаются для моделирования и обработки нелинейных отношений между входными и выходными сигналами. В ИНС применяется алгоритм самообучения, позволяющий им решать определенные задачи с учетом предыдущего опыта и минимальным числом ошибок;
* обработки и генерации естественного языка - составления программного обеспечения для трансформации любых данных в естественный язык, понятный компьютеру и используемый им для ответов человеку.

Также в ИИ применяются технологии компьютерного зрения (для нахождения, распознавания и классификации объектов, извлечения данных из изображений, аналитики полученной информации) и анализа данных для выявления в них закономерностей и прогнозирования событий по его результатам [2].

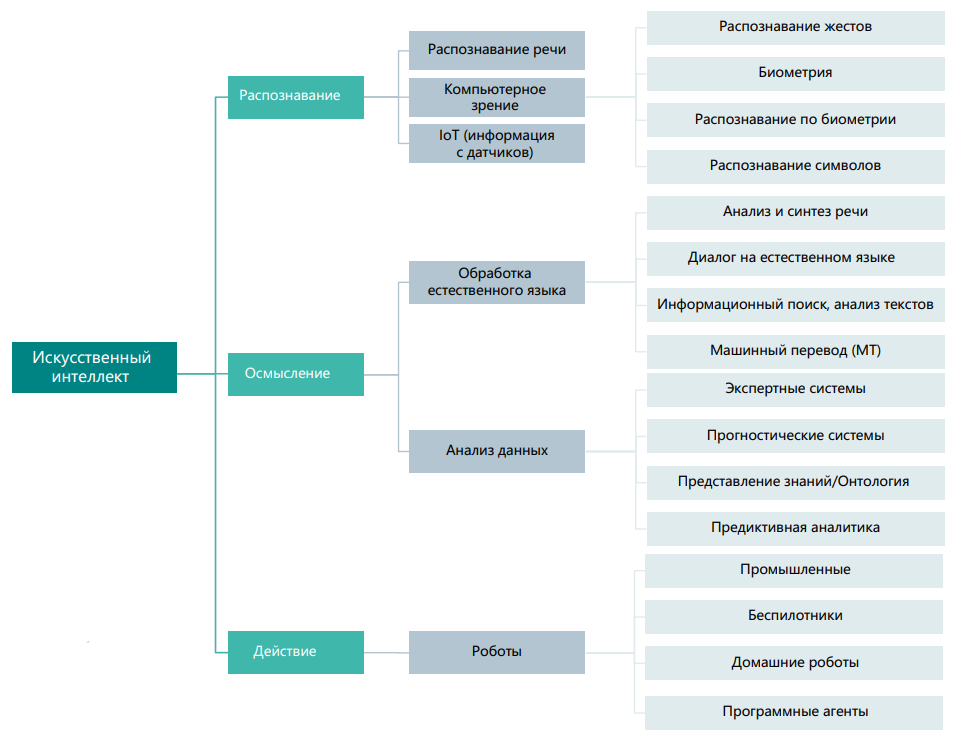


Рисунок 1 – Структура технологий искусственного интеллекта

## ГЛАВА 2. ОСНОВНЫЕ НАПРАВЛЕНИЯ ИССЛЕДОВАНИЯ

Область искусственного интеллекта является крайне неоднородной. В ней существуют различные направления исследований, которые выделяются либо по задаче (или предметной области), требующей интеллектуального анализа, либо по используемому инструментарию, либо по разрабатываемой модели мышления.

К направлениям, выделяемым на основе решаемой задачи, относятся:

* машинный перевод;
* автоматическое реферирование и информационный поиск;
* системы речевого общения;
* игровой интеллект, доказательство теорем и автоматизация научных исследований;
* компьютерное зрение;
* извлечение данных;
* сочинение текстов и музыки и др.

Перечисленные направления характеризуются тем, что значительная часть проводимых в них исследований посвящена не процессам мышления, а предмету интеллектуального анализа.

Направления искусственного интеллекта, выделяемые по развиваемому в них инструментарию, включают:

* искусственные нейронные сети;
* эволюционные вычисления;
* распознавание образов;
* экспертные системы;
* эвристическое программирование;
* мультиагентный подход и т.д.

Отличие данных направлений в том, что в них развивается аппарат решения большого класса задач. Эта группа направлений более неоднородна, чем первая. В ней существуют направления (например, ИНС), которые претендуют на то, чтобы называться отдельным подходом к искусственному интеллекту в целом.

К направлениям третьего типа можно отнести:

* поиск в пространстве решений;
* представление знаний;
* машинное обучение [1].

2. 1. Категории моделей предиктивной аналитики

Основной смысл термина предиктивная аналитика — предиктивное моделирование, то есть оценка имеющихся данных с использованием предиктивной модели, а затем выдача прогноза. Но в целом термин используется для обозначения ряда дисциплин, связанных с аналитикой. Эти дисциплины включают процесс анализа данных и используются в принятии решений. Их можно классифицировать следующим образом:

1. Прогностические модели: связь между результатами работы объекта и его характеристиками моделируется прогностическими моделями. Такая модель оценивает вероятность того, что аналогичный объект с такими же характеристиками покажет такой же результат. Например, такая модель широко применяется в маркетинге, где ожидаются ответы о предпочтениях клиентов. Она моделирует поведение человека, чтобы дать ответы конкретный вопрос. Такая модель обычно считается прямо во время проведения транзакции с конкретным клиентом, рассчитывая его реакцию или, например, риск мошенничества.

2. Описательные модели: Описательная модель выявляет связи между данными для нахождения потенциальных клиентов или групп клиентов. Как и прогностические модели, которые идентифицируют одного клиента, описательные модели определяют множественные отношения между продуктом и клиентами. Вместо ранжирования клиентов по их действиям, она классифицирует клиентов по их предпочтениям. Большое количество отдельных клиентов могут быть объединены в группу, чтобы сделать для нее прогноз в описательной модели.

3. Модели принятия решений: описывают связь между данными, результатом прогноза и решением. Если на основе результатов прогноза необходимо принять решение, которое включает в себя множество параметров, для этого строят модель принятия решения. Такие модели используются, чтобы максимизировать определенный результат или минимизировать конкретные риски. Эти модели встраиваются в бизнес-правила, которые выдают решения для каждой ситуации при любых обстоятельствах.

Прогнозирующая модель строится для предсказания на необходимом уровне детализации. Такая модель генерирует прогностический балл для каждого конкретного объекта (человека, компании, транзакции). Это технология, которая учится на прошлом опыте, чтобы сделать предсказание о будущем поведении. Точность результатов модели зависит от качества данных.

* 1. Анатомия рекомендательных систем

Система рекомендаций (или рекомендательная система) — это класс машинного обучения, который использует данные для прогнозирования, сужения и поиска того, что люди ищут среди экспоненциально растущего числа вариантов.

Система рекомендаций, также известная как рекомендательная система, представляет собой алгоритм искусственного интеллекта, который использует большие объемы данных, чтобы предоставлять пользователям рекомендации по дополнительным продуктам или услугам. Она основывается на различных факторах, таких как предыдущие покупки, история поиска и демографическая информация, чтобы помочь пользователям находить интересные им продукты или услуги, которые они могли бы упустить без такой системы.

Рекомендательные системы учатся понимать предпочтения пользователей и предсказывать их интересы на основе данных о их взаимодействии с продуктами или услугами. Они пользуются популярностью у поставщиков контента и продуктов, потому что могут значительно повысить удовлетворение и продажи, предлагая персонализированные рекомендации, подходящие индивидуальным потребностям пользователей.

* + 1. Типы рекомендательных систем

Существует несколько основных типов рекомендательных систем, вот некоторые из них:

1. Фильтрация на основе контента: Этот тип рекомендательной системы анализирует характеристики и содержание продуктов, которые пользователь уже использовал или оценил, и предлагает ему похожие продукты. Например, если пользователь смотрел фильмы жанра комедии, система рекомендаций на основе контента может предложить ему другие комедийные фильмы.

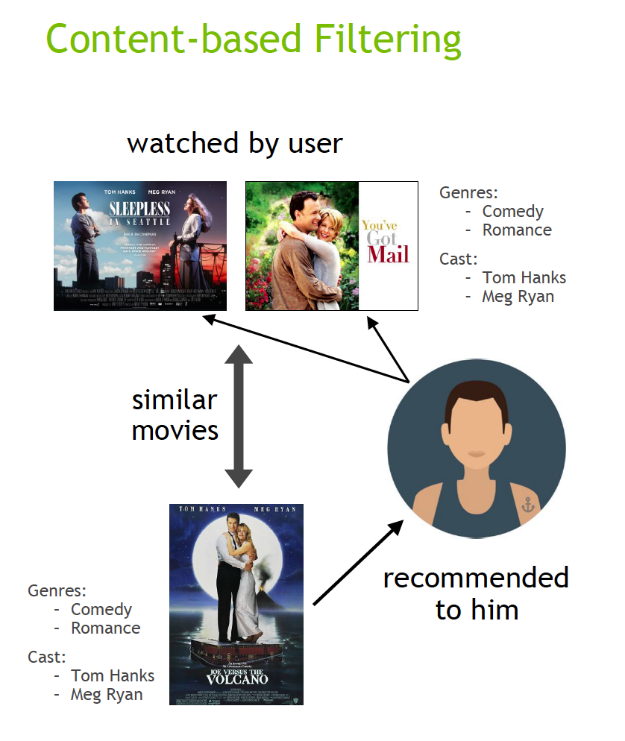


Рисунок 2 – Фильтрация на основе контента

2. Коллаборативная фильтрация: В этом типе системы рекомендаций используется информация о предпочтениях пользователей и их взаимодействии с продуктами. Она анализирует данные о ранее сделанных выборах пользователей и находит схожие пользователей или продукты, чтобы предложить рекомендации. Например, если два пользователя имеют схожие предпочтения в отношении фильмов, система рекомендаций на основе коллаборативной фильтрации может предложить одному из них фильм, который понравился другому пользователю.

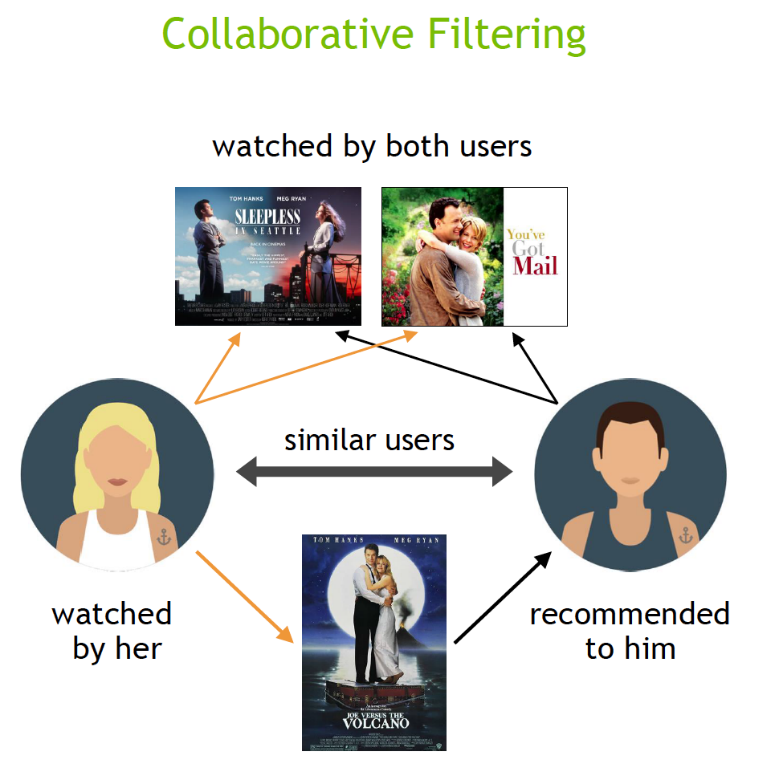


Рисунок 3 – Коллаборативная фильтрация

3. Гибридные системы: Этот тип рекомендательных систем комбинирует различные методы и подходы, такие как фильтрация на основе контента и коллаборативная фильтрация, для предоставления более точных рекомендаций. Гибридные системы позволяют учитывать различные факторы, включая характеристики продуктов, предпочтения пользователей и контекст пользования. Например, рекомендательная система музыкального потока может использовать фильтрацию на основе контента для анализа жанра и стиля песен, а также коллаборативную фильтрацию для анализа предпочтений других пользователей с похожими музыкальными вкусами.

Преимущества рекомендательных систем:

* Персонализация: Рекомендательные системы предлагают персонализированные рекомендации, исходя из предпочтений и интересов каждого конкретного пользователя. Это позволяет создать индивидуальный и уникальный опыт для каждого пользователя, учитывая его предыдущие взаимодействия и предпочтения.
* Улучшение пользовательского опыта: Рекомендательные системы помогают пользователям находить интересные и релевантные им продукты или услуги, которые они могли бы пропустить без таких рекомендаций. Это может значительно улучшить их пользовательский опыт и предоставить им больше удовлетворения.
* Увеличение продаж и доходов: Рекомендательные системы могут привести к увеличению продаж и доходов для поставщиков контента и продуктов. Путем предлагания персонализированных рекомендаций, которые наиболее соответствуют интересам пользователей, рекомендательные системы могут повысить вероятность покупки и увеличить конверсию.
* Уменьшение информационного перегруза: С большим количеством доступной информации может быть сложно найти именно то, что нужно. Рекомендательные системы предлагают достаточно выбора, учитывая предпочтения пользователей, что помогает снизить информационную перегрузку и упростить процесс выбора.
* Расширение ассортимента: Рекомендательные системы могут помочь пользователям обнаруживать новые продукты и услуги, которые они могли бы не заметить самостоятельно. Это может способствовать расширению ассортимента и предложению пользователей более широкого выбора.
  1. Прогнозирование с помощью нейронных сетей

Нейронные сети - это очень мощный и гибкий механизм прогнозирования. При определении того, что нужно прогнозировать, необходимо указывать переменные, которые анализируются и предсказываются. Здесь очень важен использованный уровень детализации. На требуемый уровень детализации влияет множество факторов: доступность и точность данных, стоимость анализа и предпочтения пользователей результатов прогнозирования. В ситуациях, когда наилучший набор переменных неясен, можно попробовать разные альтернативы и выбрать один из вариантов, дающий наилучшие результаты. Обычно так осуществляется выбор при разработке прогнозирующих систем, основанных на анализе исторических данных.

Второй важный этап при построении нейросетевой прогнозирующей системы - это определение следующих трех параметров: периода прогнозирования, горизонта прогнозирования и интервала прогнозирования. Период прогнозирования - это основная единица времени, на которую делается прогноз. Горизонт прогнозирования - это число периодов в будущем, которые покрывает прогноз. То есть, может понадобиться прогноз на 10 дней вперед, с данными на каждый день. В этом случае период - сутки, а горизонт - 10 суток. Наконец, интервал прогнозирования - частота, с которой делается новый прогноз. Часто интервал прогнозирования совпадает с периодом прогнозирования. Выбор периода и горизонта прогнозирования обычно диктуется условиями принятия решений в области, для которой производится прогноз. Выбор этих двух параметров - едва не самое трудное в нейросетевом прогнозировании. Для того чтобы прогнозирование имело смысл, горизонт прогнозирования должен быть не меньше, чем время, необходимое для реализации решения, принятого на основе прогноза. Таким образом, прогнозирование очень сильно зависит от природы принимаемого решения. В некоторых случаях, время, требуемое на реализацию решения, не определено, например, как в случае поставки запасных частей для пополнения запасов ремонтных предприятий. Существуют методы работы в условиях подобной неопределенности, но они повышают вариацию ошибки прогнозирования. Поскольку с увеличением горизонта прогнозирования точность прогноза, обычно, снижается, часто можно улучшить процесс принятия решения, уменьшив время, необходимое на реализацию решения и, следовательно, уменьшив горизонт и ошибку прогнозирования.

Точность прогноза, требуемая для конкретной проблемы, оказывает огромное влияние на прогнозирующую систему. Также огромное влияние на прогноз оказывает обучающая выборка.

Первое, с чем сталкивается пользователь любого нейропакета - это необходимость подготовки данных для нейросети. На практике именно предобработка данных может стать наиболее трудоемким элементом нейросетевого анализа. Причем, знание основных принципов и приемов предобработки данных не менее, а может быть даже более важно, чем знание собственно нейросетевых алгоритмов. Последние, как правило, уже "зашиты" в различных нейроэмуляторах, доступных на рынке. Сам же процесс решения прикладных задач, в том числе и подготовка данных, целиком ложится на плечи пользователя.

Общий алгоритм прогнозирования с помощью нейронной сети состоит из следующих пунктов:

* получение временного ряда с интервалом в выбранную временную итерацию;
* заполнение «пробелов» в истории;
* сглаживание ряда методом скользящих средних (или другим);
* получение ряда относительного изменения прогнозируемой величины;
* формирование таблицы «окон» с глубиной погружения временных интервалов;
* добавление к таблице дополнительных данных (например, изменение величины за предыдущие годы);
* шкалирование;
* определение обучающей и валидационной выборок;
* подбор параметров нейросети;
* обучение нейросети;
* проверка работоспособности нейросети в реальных условиях.

Характерный пример успешного применения нейронных вычислений в финансовой сфере - управление кредитными рисками. Как известно, до выдачи кредита банки проводят сложные статистические расчеты по финансовой надежности заемщика, чтобы оценить вероятность собственных убытков от несвоевременного возврата финансовых средств. Такие расчеты обычно базируются на оценке кредитной истории, динамике развития компании, стабильности ее основных финансовых показателей и многих других факторов. Один широко известный банк США опробовал метод нейронных вычислений и пришел к выводу, что та же задача по уже проделанным расчетам подобного рода решается быстрее и точнее. Например, в одном из случаев оценки 100 тыс. банковских счетов новая система, построенная на базе нейронных вычислений, определила свыше 90% потенциальных неплательщиков [3].

## ГЛАВА 3. ПРИМЕРЫ ПРОГНОЗИРОВАНИЯ С ПОМОЩЬЮ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

2. 1. Трансформеры

Трансформеры (Transformer) - это архитектура нейронных сетей, которая была представлена в статье "Attention is All You Need" в 2017 году и с тех пор стала одной из наиболее важных и широко используемых архитектур в NLP. Она предоставляет мощные инструменты для работы с последовательными данными, такими как тексты, и позволяет моделям понимать зависимости и взаимодействия между словами в тексте.

В основе трансформеров лежит механизм внимания (attention mechanism), который позволяет модели сосредотачиваться на разных частях входных данных с разной степенью важности. Давайте подробнее разберем основные компоненты трансформера:

* Входные данные: Трансформеры обрабатывают текст в виде последовательности векторов слов или токенов. Эти векторы могут быть получены с помощью предварительно обученных эмбеддингов слов (например, Word2Vec или GloVe).
* Многоуровневые энкодеры и декодеры: Трансформер состоит из множества энкодеров и декодеров. Энкодеры преобразуют входные данные во внутреннее представление, а декодеры генерируют выходные данные (например, перевод текста или ответ на вопрос).
* Механизм внимания: Это ключевой компонент трансформера. Механизм внимания позволяет модели определять важность каждого элемента входных данных относительно других элементов. Он создает взвешенную сумму всех элементов с учетом их важности. Это помогает модели улавливать контекст и зависимости между словами.
* Многоголовой механизм внимания: Для повышения производительности и улучшения способности модели к выявлению различных типов зависимостей в тексте трансформер использует многоголовой механизм внимания. Каждая "голова" фокусируется на разных аспектах данных.
* Позиционные эмбеддинги: Трансформеры не учитывают порядок слов в тексте, поэтому для передачи информации о позиции слов в последовательности используются позиционные эмбеддинги.
* Слои нормализации и нейронные сети: Модели трансформера также содержат слои нормализации и нейронные сети, которые выполняют различные операции над данными.

Процесс работы трансформера выглядит следующим образом:

1. Входные данные проходят через несколько слоев энкодера, где применяются механизмы внимания.

2. Энкодеры создают внутреннее представление данных.

3. Выход энкодера передается в декодер (если это задача перевода или генерации текста).

4. Декодер генерирует последовательность выходных данных.

5. Трансформеры могут использоваться для множества задач NLP, таких как машинный перевод, суммаризация текста, чат-боты, анализ тональности и многое другое. Они стали основой для многих современных архитектур и позволили значительно улучшить результаты в области обработки естественного языка.

Общий вид трансформеров представлен на рисунке 4.

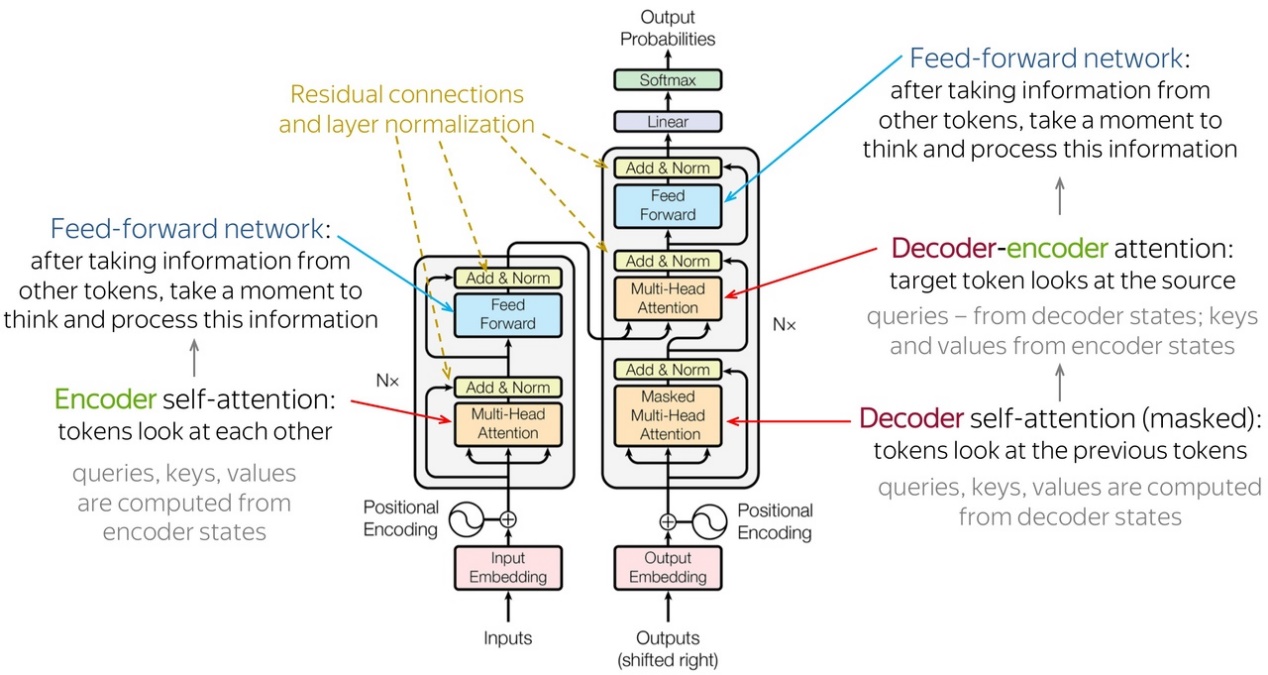


Рисунок 4 – Общий вид трансформеров

Для более детального погружения в трансорфмеры разберём механизм внимания.

Механизм внимания (Attention mechanism) в трансформерах является ключевым элементом, который позволяет моделям обрабатывать последовательности данных (например, текст) и уделять разную важность различным элементам последовательности. Этот механизм имитирует способность человеческого внимания сосредотачиваться на конкретных частях информации в контексте выполнения задачи.

Механизм внимания в трансформерах работает следующим образом:

* Запрос, ключи и значения: При обработке каждого элемента в последовательности создаются три вектора: запрос (query), ключи (keys) и значения (values). Эти векторы вычисляются на основе входных данных и параметров модели. Запрос представляет собой текущий элемент, ключи и значения представляют остальные элементы в последовательности.
* Вычисление весов важности: Для каждой пары "запрос-ключ" вычисляется показатель сходства, который определяет, насколько текущий элемент похож на другие элементы в последовательности. Обычно используется скалярное произведение (dot product) или другая функция сходства.
* Вычисление весов внимания: Показатели сходства нормализуются с помощью функции Softmax, чтобы получить веса внимания. Эти веса показывают, какую важность модель придает каждому элементу в последовательности для текущего элемента.
* Вычисление взвешенной суммы значений: Значения умножаются на соответствующие веса внимания и суммируются, чтобы получить выходной вектор. Это означает, что более важные элементы будут иметь больший вклад в выход.

Пример: Предположим, у нас есть предложение: "Солнце светит ярко". Для слова "светит" механизм внимания может выделить "Солнце" как ключевое слово и присвоить ему больший вес внимания, так как оно важно для понимания значения слова "светит".

Механизм внимания в трансформерах может быть многоголовым, что позволяет модели фокусироваться на разных аспектах данных. Кроме того, он может использоваться как для кодирования информации в энкодерах, так и для декодирования в декодерах (например, при задачах машинного перевода).

Механизм внимания является мощным инструментом для работы с последовательными данными, и он позволяет трансформерам достигать выдающихся результатов в NLP задачах, таких как машинный перевод, суммаризация текста и анализ тональности.

На рисунке 5 представлено, как будет выглядеть механизм для слова it. Как видно, для слова it наиболее важным является tire.

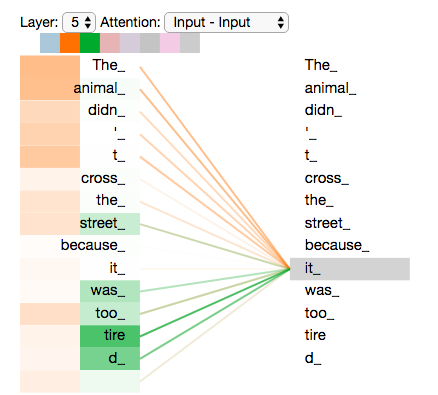


Рисунок 5 – Механизм внимания для слова it

В трансформерах часто используется улучшенная версия этого механзма, а именно Многоголовое внимание.

Многоголовое внимание (Multi-Head Attention) является ключевой компонентой трансформеров в области NLP и обеспечивает модели возможность сфокусироваться на различных аспектах информации входных данных. Это расширение механизма внимания позволяет модели улавливать не только линейные зависимости между словами, но и более сложные и абстрактные отношения.

Основные черты многоголового внимания:

* Параллельность: Вместо одной головы внимания (как в классическом механизме внимания), многоголовое внимание использует несколько голов, каждая из которых обучается находить разные характеристики и зависимости в данных. Эти головы работают параллельно, что позволяет модели обрабатывать информацию более эффективно.
* Проекции: Каждая голова внимания проецирует входные векторы запросов, ключей и значений в новое пространство с весами, которые обучаются в процессе обучения модели. Эти проекции позволяют каждой голове сосредотачиваться на разных аспектах данных.
* Соединение голов: Результаты каждой головы объединяются (например, конкатенируются) и далее проходят через линейное преобразование, чтобы получить окончательный выход многоголового внимания.

Пример использования многоголового внимания: Предположим, у нас есть предложение: "Быстрый бегун победил на соревнованиях". Многоголовое внимание может позволить модели одной голове сконцентрироваться на словах, связанных с бегом и победой, другой голове - на словах, связанных с временем, а третьей - на словах, связанных с соревнованиями. Таким образом, каждая голова будет обрабатывать информацию с разных точек зрения.

Многоголовое внимание помогает моделям трансформера обучаться находить сложные зависимости и абстракции в данных, что делает их эффективными в широком спектре NLP задач, включая машинный перевод, вопросно-ответные системы и анализ текстов.

Пример решения гидры представлен на рисунке 6.

Разберём внимательнее и другие слои Трансформера.

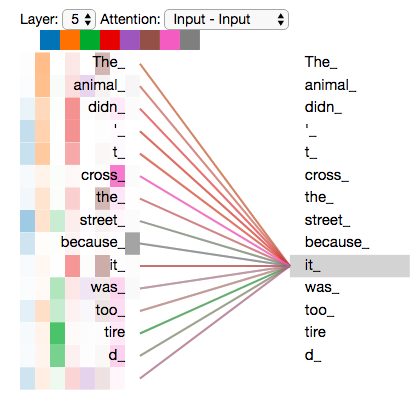


Рисунок 6 – Пример решения гидры

* + 1. Позиционные эмбединги

Позиционные эмбеддинги (Positional Embeddings) - это специальные векторы, которые добавляются к словам или символам в последовательности, чтобы модель могла учитывать позиции слов в тексте. Они играют важную роль в архитектуре трансформеров и позволяют модели учитывать порядок слов, несмотря на отсутствие рекуррентных или свёрточных слоёв.

Рассмотрим, как работают позиционные эмбеддинги:

* Позиционная информация: Важно понимать, что сами по себе слова в трансформерах обрабатываются независимо, и модель не имеет информации о позициях слов в последовательности. Это отличает их от рекуррентных сетей, где порядок слов обрабатывается последовательно.
* Создание позиционных эмбеддингов: Для учёта позиций в последовательности добавляются позиционные эмбеддинги. Обычно это векторы фиксированной длины, которые обозначают позиции слов от начала до конца последовательности. Например, вектор для позиции 1 может выглядеть как **[0.2, 0.4, -0.1, ...]**, а для позиции 2 как **[0.3, -0.1, 0.7, ...].**
* Сложение с эмбеддингами слов: Позиционные эмбеддинги просто складываются с эмбеддингами слов (или символов) в последовательности. Например, если у нас есть слова "я", "люблю", "программирование", и соответствующие им векторы слов, а также позиционные эмбеддинги для каждой позиции в предложении, то эмбеддинги слов складываются с соответствующими позиционными эмбеддингами.
* Сохранение порядка: Это добавление позиционных эмбеддингов позволяет модели учитывать порядок слов в предложении. Теперь модель "знает", что слово на позиции 1 - это первое слово, а слово на позиции 2 - это второе слово, и так далее.
* Смешивание с контекстом: После добавления позиционных эмбеддингов эмбеддинги слов вместе с позиционными информациями передаются в архитектуру трансформера, где модель обрабатывает их с учётом контекста и выполняет задачи NLP, такие как машинный перевод, классификация текстов и многое другое.

Позиционные эмбеддинги позволяют трансформерам сохранять информацию о порядке слов в предложении, что делает их мощными для обработки текстов и решения различных задач в области обработки естественного языка.

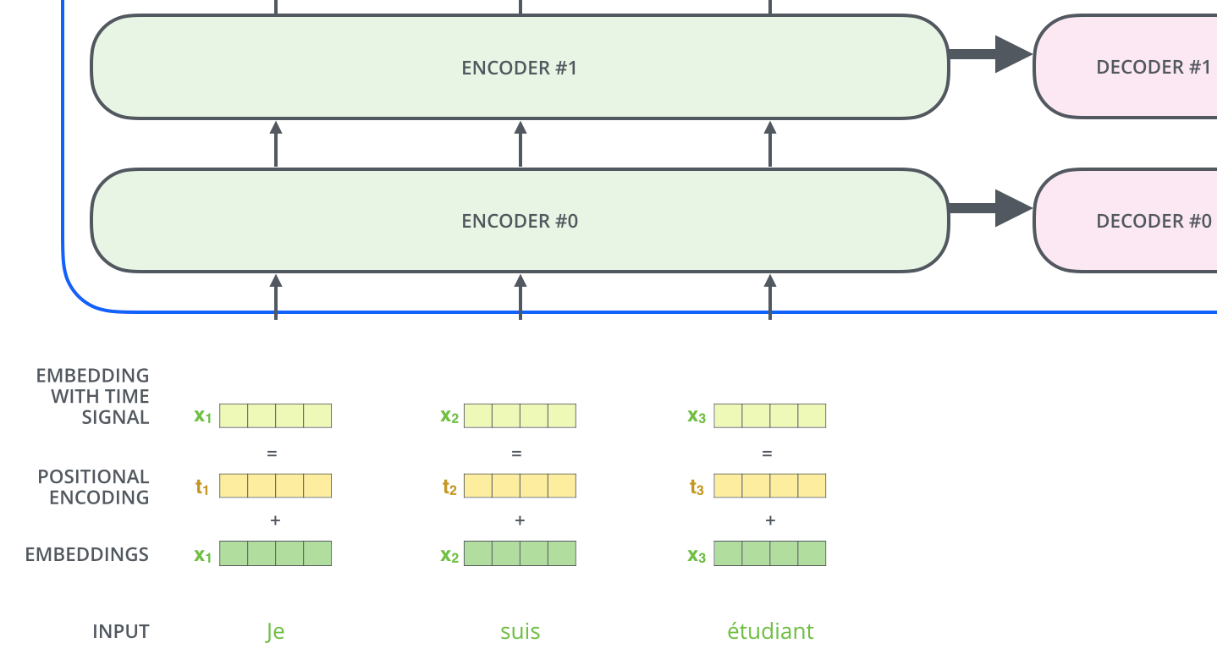


Рисунок 7 – Позиционные эмбеддинги

В заключение, можно подчеркнуть, что трансформеры в машинном обучении представляют собой революционный класс моделей, который существенно изменил подход к обработке естественного языка и другим задачам, связанным с последовательностями данных. Эти модели, впервые представленные в виде архитектуры "Attention Is All You Need", демонстрируют впечатляющую производительность во многих задачах NLP, включая машинный перевод, сентимент-анализ, вопросно-ответные системы и многие другие.

Главной особенностью трансформеров является механизм внимания (attention mechanism), который позволяет моделям учитывать взаимодействие между элементами последовательности и работать с контекстом более эффективно. Этот механизм может быть многоголовым (multi-head), что дополнительно повышает способность модели к анализу сложных зависимостей.

Трансформеры также успешно применяются в обработке изображений, например, в архитектуре Vision Transformer (ViT), что свидетельствует о их универсальности.

Однако следует отметить, что обучение и разработка трансформеров требуют значительных вычислительных ресурсов и данных, что может ограничивать доступ к этой технологии. Тем не менее, с появлением предварительно обученных моделей и фреймворков, таких как Hugging Face Transformers, стало возможным использовать трансформеры в различных приложениях с минимальными усилиями.

Таким образом, трансформеры представляют собой мощный инструмент в области машинного обучения и искусственного интеллекта, и они продолжают привлекать внимание исследователей и инженеров, открывая новые возможности в обработке текста, изображений и других типов данных.

Все state of the art модели на данный момент используют технологию трансформеров. Ниже приведены самые известные

BERT (Bidirectional Encoder Representations from Transformers): BERT - это одна из самых известных моделей для предобучения текстовых данных. Она обучается на больших текстовых корпусах и способна понимать семантику слов и фраз. BERT используется для множества задач, таких как машинный перевод, классификация текстов, вопросно-ответные системы и другие.

GPT (Generative Pre-trained Transformer): GPT - это семейство моделей, начиная с GPT-1 и заканчивая GPT-3. Они специализируются на генерации текста и понимании контекста. GPT-3, например, является одной из самых больших и мощных языковых моделей, способной выполнять разнообразные задачи.

XLNet: XLNet - это модель, разработанная на основе идей BERT и GPT, и она улучшает понимание зависимостей между словами и фразами.

RoBERTa: RoBERTa (A Robustly Optimized BERT Pretraining Approach) - это вариация BERT с оптимизированными гиперпараметрами и обучением на больших корпусах.

* 1. Подходы для рекомендательных систем
     1. Матричная факторизация для рекомендаций

Методы матричной факторизации (MF) лежат в основе многих популярных алгоритмов, включая встраивание слов и тематическое моделирование, и стали доминирующей методологией в рекомендациях на основе совместной фильтрации. MF можно использовать для расчета сходства оценок или взаимодействий пользователей для предоставления рекомендаций. В приведенной ниже простой матрице элементов пользователя Теду и Кэрол нравятся фильмы B и C. Бобу нравится фильм B. Чтобы порекомендовать фильм Бобу, матричная факторизация вычисляет, что пользователям, которым понравился B, понравился и C, поэтому C является возможной рекомендацией для Боба.

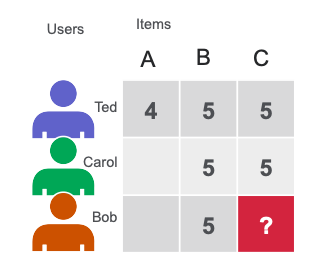


Рисунок 8 – Матричная факторизация

* + 1. Матричная факторизация для явной обратной связи

Самая простая идея – смоделировать взаимодействие пользователя с элементом с помощью линейной модели . Чтобы узнать значения pᵤ и qᵢ, мы можем минимизировать регуляризованную потерю MSE в наборе K пар (u , i), для которых известно rᵤᵢ. Полученный таким образом алгоритм называется вероятностной матричной факторизацией (PMF) .

(1)

(2)

Функцию потерь можно минимизировать двумя разными способами. Первый подход – использовать стохастический градиентный спуск (SGD) . SGD легко реализовать, но с ним могут возникнуть некоторые проблемы, поскольку и pᵤ , и qᵢ неизвестны, и поэтому функция потерь не является выпуклой. Чтобы решить эту проблему, мы можем альтернативно зафиксировать значения pᵤ и qᵢ и получить задачу выпуклой линейной регрессии, которую можно легко решить с помощью обычного метода наименьших квадратов (OLS) . Этот второй метод известен как метод попеременных наименьших квадратов (ALS) и обеспечивает значительное распараллеливание и ускорение.

Алгоритм PMF позже был обобщен алгоритмом разложения по сингулярным значениям (SVD) , который ввел в модель элементы смещения. Более конкретно, bᵤ и bᵢ измеряют наблюдаемые отклонения рейтинга пользователя u и элемента i соответственно, а μ — это общий средний рейтинг. Эти термины часто объясняют большую часть наблюдаемых оценок, поскольку некоторые элементы обычно получают более высокие/худшие оценки, а некоторые пользователи постоянно более/менее щедро выставляют свои оценки.

(3)

* + 1. Контент-ориентированный подход

Методы, основанные на контенте, описывают пользователей и элементы по их известным метаданным. Каждый элемент i представлен набором соответствующих тегов — например, фильмы на платформе IMDb могут быть помечены как «боевик», «комедия» и т. д. Каждый пользователь u представлен профилем пользователя, который может быть создан на основе известной информации о пользователе: например, пол и возраст — или на основе прошлой активности пользователя.

Для обучения модели машинного обучения с помощью этого подхода мы можем использовать модель k-NN. Например, если мы знаем, что пользователь u купил товар i, мы можем порекомендовать вам доступные товары с характеристиками, наиболее похожими на i.

Преимущество этого подхода заключается в том, что метаданные элементов известны заранее, поэтому мы также можем применять их к сценариям холодного запуска, когда на платформу добавляется новый элемент или пользователь, и у нас нет взаимодействия пользователя с элементом для обучения нашей модели. Недостатком является то, что мы не используем полный набор известных взаимодействий между пользователем и элементом (каждый пользователь рассматривается независимо) и что нам необходимо знать метаданные для каждого элемента и пользователя.

* + 1. Гибридный подход: LightFM

Методы совместной фильтрации, основанные на матричной факторизации, часто дают превосходные результаты, но в сценариях холодного запуска, когда данных о взаимодействии новых элементов и пользователей практически нет, они не могут делать хорошие прогнозы, поскольку им не хватает данных для оценки скрытых факторов. Гибридные подходы решают эту проблему, используя известные элементы или метаданные пользователя для улучшения модели матричной факторизации. Библиотека Python LightFM реализует один из самых популярных гибридных алгоритмов.

В LightFM мы предполагаем, что для каждого пользователя u мы собрали набор аннотаций тегов Aᵁ(u) — например , «мужчина» , «возраст <30» … — и аналогично каждый элемент i имеет набор аннотаций Aᴵ(i) — например, «цена > 100 $» , «книга» … Затем мы моделируем каждый тег пользователя с помощью скрытого фактора xᵁₐ ∈ R ᶠ и термина смещения bᵁₐ ∈ R и предполагаем, что векторное представление пользователя pᵤ и связанное с ним смещение bᵤ можно выразить просто как сумму этих членов xᵁₐи bᵁₐ соответственно. Мы применяем тот же подход к тегам предметов, используя скрытые факторы xᴵₐ ∈ Rᶠ и термины смещения bᴵₐ ∈ R. После того, как мы определили pᵤ, qᵢ, bᵤ, bᵢ с помощью этих формул, мы можем использовать ту же линейную модель SVD для описания отношений между этими терминами и rᵤᵢ .

(4)

(5)

(6)

(7)

Обратите внимание, что есть три интересных случая гибридного подхода LightFM:

1. Холодный запуск. Если у нас есть новый элемент i с известными тегами Aᴵ(i) , то мы можем использовать скрытые векторы xᴵₐ (полученные путем подбора нашей модели на предыдущих данных) для вычисления его встраивания qᵢ и, следовательно, оценить для любого пользователя u его рейтинг rᵤᵢ .

2. Нет доступных тегов. Если у нас нет известных метаданных для элементов или пользователей, единственная аннотация, которую мы можем использовать, — это индикаторная функция, то есть отдельная аннотация для каждого пользователя и каждого элемента. Тогда матрицы функций пользователя и элемента становятся матрицами идентификаторов, а LightFM сводится к классическому методу совместной фильтрации, такому как SVD.

3. Контентно-ориентированный против гибридного. Если бы мы использовали только теги пользователей или элементов без аннотаций индикаторов, LightFM был бы почти моделью, основанной на контенте. Поэтому на практике, чтобы улучшить взаимодействие пользователя с элементом, мы также добавляем к известным тегам аннотацию-индикатор, отличающуюся от каждого пользователя и элемента.

* 1. Personal Insights

Пока социальные медиа рассказывают о Социальном рейтинге Китая и о том, как коммунизм уничтожает свободу, в США компания IBM разрабатывает систему, которая анализирует личные данные и переписки населения для определения психического портрета человека и в будущем, кто знает как они будут это использовать.

Изображение выглядит как текст, снимок экрана, графический дизайн, Графика

Автоматически созданное описание

Рисунок 9 – Тотальный контроль

Personality Insights - это инструмент на основе машинного обучения, разработанный IBM, который анализирует текстовые данные, такие как сообщения в социальных сетях, электронные письма и обсуждения на форумах, для определения личностных характеристик человека. Инструмент использует лингвистическую аналитику для определения этих характеристик, которые включают в себя Большую пятерку личностных черт (открытость, добросовестность, экстраверсию, сговорчивость и невротизм), потребности и ценности. Инструмент также может анализировать привычки потребления и временное поведение, если вводимый текст помечен временем.

Personality Insights обучается на основе данных тысяч пользователей, а также данных из их лент в Twitter. Инструмент поддерживает английский, испанский, японский, корейский и арабский языки, но результаты для арабского и корейского языков не столь точны. Точность инструмента измеряется путем сравнения средней абсолютной ошибки (MAE) и средней корреляции между прогнозируемыми баллами и фактическими результатами опроса для различных категорий личностных характеристик.

Основной метод основан на подходе с открытым словарем, разработанном исследователями из Пенсильванского университета. Этот метод был использован для анализа статусов в Facebook 75 000 пользователей-добровольцев. Затем исследователи построили модели, позволяющие предсказать возраст, пол и индивидуальность человека на основе его письменной формы.

Personality Insights использует технологию встраивания слов GloVe с открытым исходным кодом для создания векторных представлений каждого слова входного текста. Затем он передает эти векторы в алгоритм машинного обучения для обучения и тестирования. Инструмент не учитывает демографические данные пользователей, такие как возраст, пол, раса и культура.

Некоторые потенциальные варианты использования Personality Insights включают целевой маркетинг, привлечение клиентов, личные связи, составление резюме и мониторинг психического здоровья. Однако существуют также опасения по поводу потенциальных последствий для прав человека, таких как дискриминация и вопросы неприкосновенности частной жизни. Разработчики должны быть осведомлены об этих проблемах и использовать инструмент ответственно.

## ЗАКЛЮЧЕНИЕ

Прогнозирование с помощью искусственного интеллекта имеет большой потенциал и широкое применение в современном мире. Это инструмент, который может значительно повысить точность и достоверность прогнозов, основанных на больших объемах данных.

Однако, необходимо учитывать, что использование искусственного интеллекта для прогнозирования требует аккуратного подхода. Важно учесть, что полученные прогнозы не являются абсолютно точными и должны быть интерпретированы с учетом контекста и экспертного мнения. Также следует учитывать возможность ошибок и неточностей, связанных с неполными или неправильными данными, настройкой алгоритмов и применением методов.

Тем не менее, применение искусственного интеллекта в прогнозировании уже демонстрирует значительные преимущества. С его помощью можно обнаружить скрытые закономерности, предсказать тренды и риски, что помогает принимать более информированные и осознанные решения в различных сферах деятельности.

Таким образом, прогнозирование с помощью искусственного интеллекта является важным инструментом для принятия решений, основанных на данных. Оно позволяет автоматизировать процесс прогнозирования, повышать точность прогнозов и улучшать качество принимаемых решений.

## СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ

1. Таран Т.А., Зубов Д.А. Искусственный интеллект теория и приложения. Начальное пособие. – Луганск: Издательство ВНУ им. В. Даля, 2006. – 242 с.
2. Альманах Искусственный интеллект. Предсказательная аналитика и системы поддержки принятия решений. Аналитический сборник – Москва, № 5, сентябрь 2020. (URL: <https://cdo2day.ru/wp-content/uploads/2020/11/almanach_ai_5.pdf>)
3. 1 Технологии искусственного интеллекта. [Электронный ресурс]. URL: <https://apr.moscow/content/data/6/11%20%D0%A2%D0%B5%D1%85%D0%BD%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D0%B8%20%D0%B8%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE%20%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82%D0%B0.pdf> (дата обращения 07.10.2023).
4. 2 Технологии искусственного интеллекта и машинного обучения. [Электронный ресурс]. URL: [https://hsbi.hse.ru/articles/tekhnologii-iskusstvennogo-intellekta-i-mashinnogo-obucheniya/#:~:text=%D0%A2%D0%B5%D1%85%D0%BD%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D0%B8%20%D0%B8%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE%20%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82%D0%B0%20%D0%B8%20%D0%BC%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE%20%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D1%8F,-05.03.2021&text=%D0%98%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9%20%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82%20(%D0%98%D0%98)%20%2D%20%D1%8D%D1%82%D0%BE,%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D1%8F%20%D0%BA%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BE%D0%B2%20%D1%81%D0%B0%D0%BC%D0%BE%D1%81%D1%82%D0%BE%D1%8F%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D0%BE%D0%BC%D1%83%20%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D1%8E%20%D0%B7%D0%B0%D0%B4%D0%B0%D1%87.](https://hsbi.hse.ru/articles/tekhnologii-iskusstvennogo-intellekta-i-mashinnogo-obucheniya/#:~:text=%D0%A2%D0%B5%D1%85%D0%BD%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D0%B8%20%D0%B8%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE%20%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82%D0%B0%20%D0%B8%20%D0%BC%D0%B0%D1%88%D0%B8%) (дата обращения 07.10.2023).
5. 3 Прогнозирование с помощью нейронных сетей. [Электронный ресурс]. URL: <https://scienceforum.ru/2018/article/2018005783> (дата обращения 07.10.2023).
6. Recommender Systems — A Complete Guide to Machine Learning Models. [Электронный ресурс]. URL: <https://towardsdatascience.com/recommender-systems-a-complete-guide-to-machine-learning-models> (дата обращения 07.10.2023).
7. What is a Recommendation System | Data Science | NVIDIA Glossary |. [Электронный ресурс]. URL: <https://www.nvidia.com/en-us/glossary/data-science/recommendation-system/> (дата обращения 07.10.2023).
8. Рекомендательные системы: проблемы и методы решения. Часть 1. [Электронный ресурс]. URL: <https://habr.com/ru/companies/prequel/articles/567648/> (дата обращения 07.10.2023).
9. Recommendation System Algorithms: An Overview. [Электронный ресурс]. URL: <https://www.kdnuggets.com/2017/08/recommendation-system-algorithms-overview.html?__cf_chl_tk=57UqnDJKxbXkjGHYcW06ty8CiWeBt1mia7ZtYThgWG4-1696701286-0-gaNycGzNDhA> (дата обращения 07.10.2023).
10. Определение типа личности по тексту (на замену закрытому IBM Watson Personality Insights). [Электронный ресурс]. URL: <https://habr.com/ru/companies/meanotek/articles/689360/> (дата обращения 07.10.2023).