Московский государственный университет имени М.В.Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра суперкомпьютеров и квантовой информатики

Акопян Микаэла Тиграновна

Исследование возможностей предиктивного анализа публикационной активности

Курсовая работа

Научный руководитель:
к.т.н.
Григорьева Мария Александровна

Москва, 2025

Исследование возможностей предиктивного анализа публикационной активности

Акопян Микаэла Тиграновна

Abstract

Содержание

1 Введение

Физика высоких энергий (ФВЭ) – это раздел физики элементарных частиц, который изучает фундаментальные взаимодействия и структуры материи при экстремально высоких энергиях. Эта область считается крайне важной, так как именно она даёт понимание об пониманию фундаментальных составляющих материи и энергии, и, следовательно, о строении Вселенной. В данный момент ФВЭ является одной из наиболее динамично развивающихся областей современной науки, требющей постоянного анализа огромных объемов обновляющейся информации.

Как и в любой другой научной области, конференции играют ключевую роль в коммуникации учёных, объединяя исследователей со всех краёв света для представления своих достижений и обсуждения результатов. Важным аспектом таких мероприятий являются публикации, которые отражают актуальность тем и уровень интереса научного сообщества.

Информация о всех публикациях сохранятеся на специализированных ресурсах. Одним из таких является InspireHEP — цифровая библиотека в области ФВЭ с открытым доступом. С появлением подобных платформ стало возможным централизованное хранение и доступ к данным о конференциях, публикациях и связанных метаданных. Эти данные открывают новые перспективы для анализа и прогнозирования тенденций в научной деятель- ности, что особенно актуально в условиях увеличивающегося объема информации. С InspireHEP удобнее всего работать, используя существующий API InspireHEP, который позволяет получить информацию об имеющихся на платформе публикациях, авторах, конференциях, цитированиях. Задачей курсовой работы является изучение данных для выявления закономерностей в научной коммуникации, анализа популярности тем и оценки динамики развития областей знаний.

Именно в силу высокой интенсивности научных исследований и большого объёма публикуемых материалов в области ФВЭ возникает необходимость в автоматизированном анализе публикационной активности. Это позволяет систематизировать знания, отслеживать научные тренды и прогнозировать развитие исследовательских направлений. Цифровая библиотека InspireHEP, являясь специализированной платформой, охватывающей большинство публикаций в этой области, предоставляет открытый и структурированный доступ к релевантным данным, что делает её идеальной основой для проведения такого анализа.

2 Постановка задачи

В данной работе рассматривается задача автоматического тематического анализа публикационной активности в области физики высоких энергий на основе открытых данных, полученных с платформы InspireHEP. Основной целью является выделение скрытых тематических направлений научных публикаций и анализ их распределения и динамики.

Входными данными имеющейся задачи являются метаданные о прошедших в 2014-2024 годах научных конференциях, полученные с помощью API платформы InspireHEP. Данные JSON - файлы в частности содержат в себе

- название конференции;
- даты проведения;
- страну и город проведения;
- описание конференции;
- кодовый номер конференции (cnum);
- число публикаций;

Используя кодовый номер конференции, можно получить информацию о всех публикациях, выставленных на данной конференции, а также о связанных с ними метаданных, среди которых

- название публикации;
- авторы;
- краткая аннотация;
- ссылка на полный текст;

В качестве выходных данных требуется получить предсказание динамики популярности различных тематических направлений в области физики высоких энергий, представленных на конференциях в предыдущие годы.

Для достижения поставленной цели необходимо выполнить следующие задачи:

- собрать все необходимые метаданные о прошедших в 2014-2024 годах научных конференциях,
- собрать все необходимые метаданные о выставленных на конференциях публикациях,
- провести тематический анализ публикаций с помощью алгоритмов машинного обучения: выделить темы и провести по ним классификацию документов
- провести анализ динамики тематики публикаций по годам, визуализировать полученные результаты
- присвоить каждой теме её название
- разработать модель предсказания динамики популярности тематики публикаций, оценить её качество на имеющихся данных
- предсказать динамику популярности тематики публикаций на два года вперёд, визуализировать полученные результаты

3 Обзор используемых методов

3.1 Извлючение данных о конферененренциях и публикациях с помощью API InspireHEP

Был изучен API и извлечена информация о прошедших за последние пять лет конференциях и публикациях в них. О конференциях были получены: даты проведения, id в системе, краткие названия, публикации на каждой из них. О каждой публикации: названия, id в системе, авторы, ключевые слова, краткие описания.

Наиболее информативными источниками информации о содержаниях публикаций, безусловно, являются аннотации. Однако в InspireHEP аннотации не являются обязательным атрибутом, и в большом количестве работ отсутствут. Так, только у 30 процентов публикаций есть краткие описания. Аналогично обстоит дело и с ключевыми словами. Поэтому было принято решение работать с названиями статей и извлекать тему из них, так как название является обязательным атрибутом.

3.2 Кластеризация документов по темам

Одним из ключевых этапов анализа публикационной активности является кластеризация документов по темам, позволяющая структурировать корпус текстов и выявить скрытые тематические направления.

Подходящий метод выбирался среди Latent Dirichlet Allocation (LcDA), Non-Negative Matrix Factorization (NMF) и Embedded Topic Model (ETM). LDA – популярный, классический метод, но его ограничение в учёте семантической близости слов стало существенным недостатком при работе с узкоспециализированными текстами. NMF неудобен в интерпретации результатов и требует большое число настроек. После проведённого анализа был выбран ETM, который сочетает вероятностный подход с использованием предобученных эмбеддингов слов. Это позволяет учитывать контекстные связи между словами, что особенно важно для анализа научных публикаций.

3.2.1 Embedded Topic Model (ETM)

ETM — тематическая модель, в которой распределение слов в каждой теме строится не напрямую, а через скалярное произведение вектора темы и векторов слов. Для этого

используются заранее обученные векторные представления слов (эмбеддинги), которые помогают учитывать смысловую близость между слов.
ETM работает следующим образом:

- Каждая тема k представляется вектором $\boldsymbol{\beta}_k \in \mathbb{R}^L$ в эмбеддинговом пространстве размерности L.
- Для каждого документа непосредственно обучается распределение тем $\boldsymbol{\theta}_d = \mathrm{softmax}(\boldsymbol{\eta}_d)$, где $\boldsymbol{\eta}_d$ — вектор логитов, оптимизируемый напрямую градиентным методом.
- Вероятность наблюдения слова w в теме k задаётся через матрицу эмбеддингов $\mathbf{E} = \big[\mathbf{e}_1, \ldots, \mathbf{e}_V\big]^\top \in \mathbb{R}^{V \times L}$:

$$\phi_{k,w} = \big[\mathrm{softmax}\big(\boldsymbol{\beta}_k^\top \mathbf{E}^\top\big)\big]_w.$$

- Каждое слово w_{dn} в документе d порождается как

$$p(w_{dn} = w) = \sum_{k=1}^{K} \theta_{d,k}\, \phi_{k,w}.$$

Все параметры обучаются целиком градиентным методом с помощью оптимизатора Adam.

Для матрицы частот $\mathbf{X} \in \mathbb{N}^{D \times V}$ максимизируется лог-правдоподобие

$$\mathcal{L} = \sum_{d=1}^{D} \sum_{w=1}^{V} X_{dw}\ \log\Big(\big[\boldsymbol{\theta}_d\ \mathrm{softmax}(\boldsymbol{\beta}\mathbf{E}^\top)\big]_w\Big).$$

Алгоритм оптимизации.

1. *Инициализация:* $\boldsymbol{\eta}_d \sim \mathcal{N}(0, 0.01),\ \boldsymbol{\beta}_k \sim \mathcal{N}(0, 0.01)$.
2. *Прямой проход:*

$$\begin{aligned} \boldsymbol{\theta}_d &= \mathrm{softmax}(\boldsymbol{\eta}_d), \\ \boldsymbol{\Phi} &= \mathrm{softmax}(\boldsymbol{\beta}\mathbf{E}^\top), \\ \hat{\mathbf{X}}_d &= D_d\, \boldsymbol{\theta}_d \boldsymbol{\Phi}, \end{aligned}$$

где $D_d = \sum_w X_{dw}$ — длина документа d.

3. *Обратный проход:* вычисление градиента $-\nabla_{\{\eta,\beta\}}\mathcal{L}$ и шаг Adam.

4. *Повторение:* цикл до заданного числа эпох или пока значение $\mathcal{L}$ не стабилизируется.

После сходимости строки $\mathbf{\Phi}$ задают *распределения слов по темам*; ключевые слова темы k — это w с максимальными $\phi_{k,w}$; матрица $\mathbf{\Theta}$ (полученная из $\boldsymbol{\eta}$) характеризует *распределение тем по документам*; компактность параметров ($K \times L$) упрощает расширение модели на большие словари и ускоряет обучение по сравнению с LDA.

3.2.2 Global Vectors for Word Representation (GloVe)

При выборе подходящих предобученных эмбеддингов слов было принято решение остановиться на GloVe, так как они сами по себе легковесные и простые, но в то же время универсальные, достаточно полно покрывают язык и хорошо работают в большинстве задач.

Global Vectors были обучены на глобальной матрице совместных появлений слов в большом корпусе. В отличие от моделей вроде word2vec, которые обучаются на основе предсказания соседей по контексту (локальная информация), GloVe использует глобальную статистику корпуса: матрицу соотношений совместной встречаемости слов. Идея в том, что отношение частот слов содержит семантическую информацию, которую можно эффективно закодировать в векторном пространстве.

Модель минимизирует функцию потерь, которая старается аппроксимировать логарифм количества совместных появлений двух слов через скалярное произведение их векторов. Это позволяет GloVe учитывать как абсолютную, так и относительную частоту слов, благодаря чему модель сохраняет семантические и синтаксические связи между словами. Вектора, полученные с помощью GloVe, хорошо работают в задачах аналогий (king – man + woman = queen) или (Germany - Berlin = France - Paris), классификации, кластеризации и других NLP-задачах.

3.3 Определение названий выявленных кластеров тем

Для определения названий выявленных кластеров были изучены несколько API больших языковых моделей (LLM). Рассматривались ChatGPT, DeepSeek и Gemini. ChatGPT

показал высокую точность и качество генерации названий кластеров, однако он требует для использования API высокой оплаты и использования VPN. DeepSeek, хоть и не нуждался в этом сервисе, плохо справился с поставленной задачей и имел сравнительно невысокие лимиты использования. Лучше всего показал себя Gemini. Он, хоть и нуждался в использовании VPN, имел высокие лимиты использования и позволял генерировать названия кластеров с высокой точностью и качеством.
В результате, для определения названий кластеров тем был выбран Gemini.

3.4 Предсказание на временных рядах

Как возможные варианты предсказания на временных рядах были рассмотрены ARIMA, SARIMA и Prophet.

ARIMA (Autoregressive Integrated Moving Average) — это классическая модель временных рядов, которая использует авторегрессию, интеграцию и скользящее среднее для предсказания будущих значений на основе прошлых данных. Она хорошо работает с линейными временными рядами, но может быть сложной в настройке и требует стационарности данных(свойство временного ряда, при котором его статистические характеристики такие, как дисперисия, ковариация, среднее не меняются со временем.).

SARIMA (Seasonal Autoregressive Integrated Moving Average) — это расширение ARIMA, которое учитывает сезонные компоненты временных рядов. Она добавляет сезонные параметры к модели ARIMA, что полезно, однако, SARIMA может быть сложной в настройке и требует больше вычислительных ресурсов. Кроме того, она обладает такими же недостатками, что и предшественница, и так же, как она требует стационарности данных и ручной настройки параметров.

Prophet — это библиотека от Facebook, которая предназначена для предсказания временных рядов с учетом сезонности и праздников. Она проста в использовании и позволяет быстро получать результаты.

Prophet автоматически обрабатывает пропуски в данных и может работать с нестационарными временными рядами. Она не требует ручной настройки параметров и слабо . Кроме этого, большим преимуществом данной библиотеки являестся поддержка редких и коротких временныв рядов,что актуально для полученных данных о публикациях.

4 Описание практической части

Если в рамках работы писался какой-то код, здесь должно быть его описание: выбранный язык и библиотеки и мотивы выбора, архитектура, схема функционирования, теоретическая сложность алгоритма, характеристики функционирования (скорость/память).

5 Заключение

Здесь надо перечислить все результаты, полученные в ходе работы. Из текста должно быть понятно, в какой мере решена поставленная задача.