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

Акопян Микаэла Тиграновна

—

Отчет о научно-исследовательской работе

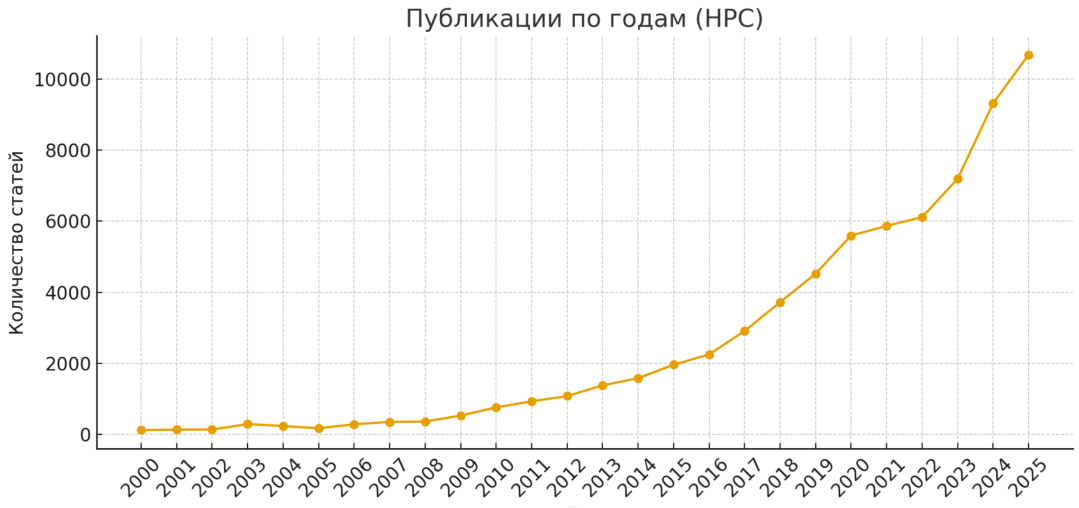
Группа 423
Научный руководитель:
к.т.н.
Григорьева Мария Александровна

Москва, 2025

Содержание

1 Введение

В последние годы область высокопроизводительных вычислений (High Performance Computing, HPC) значительно расширила своё влияние. Развитие современных методов машинного обучения, в частности глубокого обучения, требующего значительных вычислительных ресурсов, привело к повышенному вниманию к эффективности вычислительной техники. Эта тенденция хорошо прослеживается на примере публикационной активности в онлайн-библиотеке научных статей arXiv.org. Рассмотрим статистику по тегам high performance computing, supercomputer, parallel computing: если в начале XXI столетия публиковалось 100-200 статей в год, то в ещё не завершённом 2025 году их число превысило десяти тысяч статей. Аккумулированные по годам результаты представлены на графике ниже.



Важно понимать, что каждая научная статья существует не изолированно, а является частью более широкого научного контекста. Однако недавний резкий рост интереса в данной теме затруднил учёным изучение тенденции без помощи специализированных средств программного анализа данных. Это усложняет поиск релевантных работ и затрудняет принятие оптимальных решений при выборе соавторов, исследовательских организаций, площадок для публикации и других параметров метаданных.

Целью ВКР является анализ данных о публикационной активности в области высокопроизводительных вычислений и разработка рекомендательной системы, способной предлагать подходящие условия для публикации научных результатов.

2 Полученные результаты и используемые методы

2.1 Выбор подходящего ресурса

Одним из наиболее известных и богатых по содержанию источников современных научных публикаций является онлайн-библиотека arXiv.org. Платформа содержит большое количество актуальных научных работ, активно используется исследовательским с ообществом. Однако при всей своей масштабности ресурс обладает существенными ограничениями: в нём нет структурированной информации о цитируемости, месте публикации и ряде других ключевых метаданных.

Существуют альтернативные платформы, Semantic Scholar и OpenAlex, которые предоставляют значительно более полную статистиску публикаций. Но у этих ресурсов есть и свои недостатки: прежде всего, скудная система тематических меток. И OpenAlex, и Semantic Scholar опираются на заранее подготовленные категории, которые описывают содержание работ только лишь в общих чертах.

В отличие от этого, на arXiv каждая статья сопровождается не только категорией, но и набором ключевых слов, указанных авторами. Эти ключевые слова обычно отражают методы, подходы, технические детали и конкретную проблематику исследования. Именно это делает поиск по arXiv значительно более точным и продуктивным.

С учётом описанной неоднозначности было принято решение объединить их сильные стороны: использовать точный тематический поиск по ключевым словам на arXiv.org, а статистику цитируемости, данные о конференциях и другие библиографические характеристики получать из Semantic Scholar и OpenAlex.

2.2 Извлечение информации

У каждого из описанных выше ресурсов существует API с удобными реализациями в виде Python-библиотек. По тематическим ключевым словам `HPC, high performance computing, supercomputer, supercomputing, parallel computing, distributed computing, cluster computing, massively parallel, parallel algorithms, parallel processing, scientific computing, high performance system, compute-intensive, GPU computing, GPU acceleration, GPGPU, CUDA, CUDA kernel, CUDA optimization, NVIDIA GPU, tensor cores, GPU cluster, GPU parallel, multi-GPU, heterogeneous computing, accelerator,`

hardware acceleration, OpenCL, ROCm, AMD GPU, MPI, Message Passing Interface, OpenMP, OpenACC, PGAS, UPC, UPC++, Coarray Fortran, distributed memory, shared memory parallel, hybrid parallelization, CPU cluster, manycore, multicore, vectorizat SIMD, AVX, Xeon Phi, BlueGene, Cray, Fugaku, Frontier supercomputer, Summit supercomp exascale, exascale computing, exascale architecture, BLAS, LAPACK, ScaLAPACK, PETSc, Trilinos, OpenBLAS, cuBLAS, cuDNN, FFT, distributed FFT, parallel linear algebra, distributed training, data parallel, model parallel, pipeline parallel, deep learning at scale, large-scale training, GPU clusters for training, HPC for machine learning, HPC for AI, mixed precision, Horovod, DeepSpeed, Megatron, HPC cloud, cloud supercomputer, cloud GPU cluster, virtual cluster, HPC workload, batch scheduling, Slurm, PBS, job scheduler, performance tuning, auto-tuning, code optimization, profiling, parallel efficiency, scalability, strong scaling, weak scaling, performance portability, Kokkos, RAJA, numerical simulation, finite element, finite difference, lattice Boltzmann, molecular dynamics, computational fluid dynamics, CFD, climate modeling, weather simulation, particle-in-cell, monte carlo simulation, high-speed interconnect, Infiniband, RDMA, parallel filesystem, Lustre filesystem, BeeGFS, distributed storage, fault tolerance, checkpointing, energy-efficient computing, HPC scheduling было получено около 290 тысяч уникальных статей, выпущенных за последние 25 лет.

Каждая из статей имеет уникальный общепризнынный тег DOI, по которому её можно искать в сторонних ресурсах.

При работе с Semantic Scholar возникла серьёзная проблема: авторы проекта отказались предоставлять ключ доступа, без которого невозможна полноценная выгрузка метаданных. Это потребовало перехода на OpenAlex. По объёму и ширине охвата информации OpenAlex не уступает Semantic Scholar, однако накладывает более строгие ограничения на скорость и объём выгрузки, что существенно увеличило время обработки данных.

Полученные данные содержат следующие поля: название, авторы, ключевые слова, дата публикации, DOI, число цитирований, место и тип публикации, теги в OpenAlex (поле, подполе, тематическое направление ООН).

2.3 Запуск SciBERT и SPECTER

В ходе работы был выполнен обзор современных методов анализа и рекомендаций научных публикаций. На основе изученной литературы и практической доступности инструментов были выделены два наиболее перспективных подхода — SciBERT и SPECTER. Для оценки их характеристик были проведены тестовые эксперименты, а также выполнено сравнение с классическим BERT, выступающим в роли базовой модели.
Эксперимент был построен следующим образом: были подготовлены несколько крупных текстовых фрагментов, относящихся к областям биологии, химии и математики. Для каждого текста вычислялись эмбеддинги моделей, после чего проводилось сравнение косинусного сходства между ними и эмбеддингами заранее определённых тематических описаний. Такой подход позволил получить приближённую оценку "уверенности"моделей в принадлежности текста к определённой научной области, а также сравнить качество семантического представления текстов разными архитектурами.

Model	**Biology**	**Chemistry**	**Mathematics**
BERT	0.3628	0.3609	0.3337
SciBERT	0.5107	0.4313	0.3954
SPECTER	0.7047	0.5977	0.5930

Таблица 1: Пример результатов трёх моделей для текста о строении клеток организма

BERT демонстрирует почти равные значения для всех трёх тем, что подтверждает: стандартный BERT не специализируется на научной терминологии и плохо отделяет домены.
SciBERT показывает максимальную близость текста к биологии (0.5107), что соответствует ожиданиям, ведь модель обучена на научных текстах и уверенно различает биологический контекст.
SPECTER даёт наибольшие абсолютные значения сходства по всем темам и наиболее уверенное приближение к биологии (0.7047). Это согласуется с тем, что SPECTER обучен на корпусе научных статей и оптимизирован для понимания научного содержания на уровне документов.

3 Планы на следующий семестр

1. **Выявление подходящей информации для последующего построения зависимости.** Провести анализ собранной информации. Выявить наиболее значимые параметры. Построить график цитируемости.
2. **Проверка имеющихся моделей на собранных данных** Проверка и сравнение BERT, SciBERT, SPECTER. Формирование слабых и сильных сторон методов.
3. **Разработка собственной модели** Разработка модели, способной предлагать подходящие условия для публикации научных результатов.

4 Литература

- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805, 2019.

- Beltagy, I., Lo, K., Cohan, A. *SciBERT: A Pretrained Language Model for Scientific Text*, arXiv:1903.10676, 2019.

- Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D. S. *SPECTER: Document-level Representation Learning using Citation-informed Transformers*, arXiv:2004.07180, 2020.