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. Цель ВКР — изучить структуру публикационной активности в области HPC и создать основу для рекомендательной системы, помогающей выбирать оптимальные условия для размещения научных результатов. В рамках её достижения был проведён количественный анализ публикаций за последние 25 лет. Он подтвердил устойчивое повышение интереса к данной теме.

Во многом рост числа статей связан с развитием методов машинного и глубокого обучения, требующих больших вычислительных мощностей. Эта тенденция хорошо прослеживается на примере публикационной активности в онлайн-библиотеке научных статей arXiv.org. По тегам *high performance computing*, *supercomputer*, *parallel computing*: если в начале XXI века публиковалось 100–200 статей в год, то в ещё не завершённом 2025 году их число превысило десять тысяч. Аккумулированные по годам результаты представлены на рис. 1.

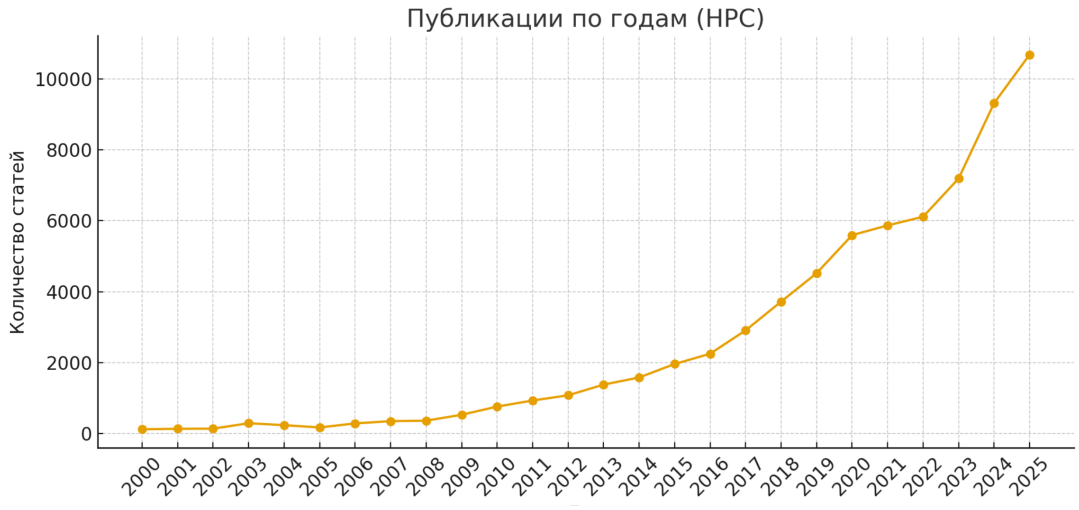

Рис. 1: Количество публикаций в онлайн-библиотеке arXiv.org по запросу *high performance computing*, *supercomputer*, *parallel computing*

Важно понимать, что каждая статья существует не изолированно, а является частью более широкого научного контекста. Однако стремительный рост числа публикаций по тематике высокопроизводительных вычислений затруднил изучение динамики развития области без применения специализированных средств программного анализа

данных. Это усложняет поиск релевантных работ и принятие оптимальных решений при выборе соавторов, исследовательских организаций, площадок для публикации и других параметров метаданных.

В рамках поставленной цели особое внимание уделяется характеристикам, связанным с параметрами публикации научных результатов. Под параметрами в данной работе понимаются такие характеристики, как выбор релевантной научной площадки (журнала или конференции), определение наиболее подходящего тематического направления, а также анализ и предложение авторских коллективов и организаций, ведущих исследования в схожих областях. Разработка такой системы позволит автоматизировать процесс ориентирования в быстро растущем массиве публикаций и упростить исследователю принятие решений о стратегии публикации.

2 Полученные результаты и используемые методы

2.1 Выбор подходящего ресурса

Одним из наиболее известных и богатых по содержанию источников современных научных публикаций по тематике высокопроизводительных вычислений является онлайн-библиотека arXiv.org[1]. 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[2] и OpenAlex[3], 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 (см. приложение 1) был написан программный код, позволяющий получить около 290 тысяч статей, выпущенных за последние 25 лет.

[1] `https://arxiv.org`

[2] `https://www.semanticscholar.org`

[3] `https://openalex.org`

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

При работе с Semantic Scholar возникла 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 до нескольких недель.

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[4]).

2.3 Анализ данных

2.3.1 Число собранных публикаций по годам

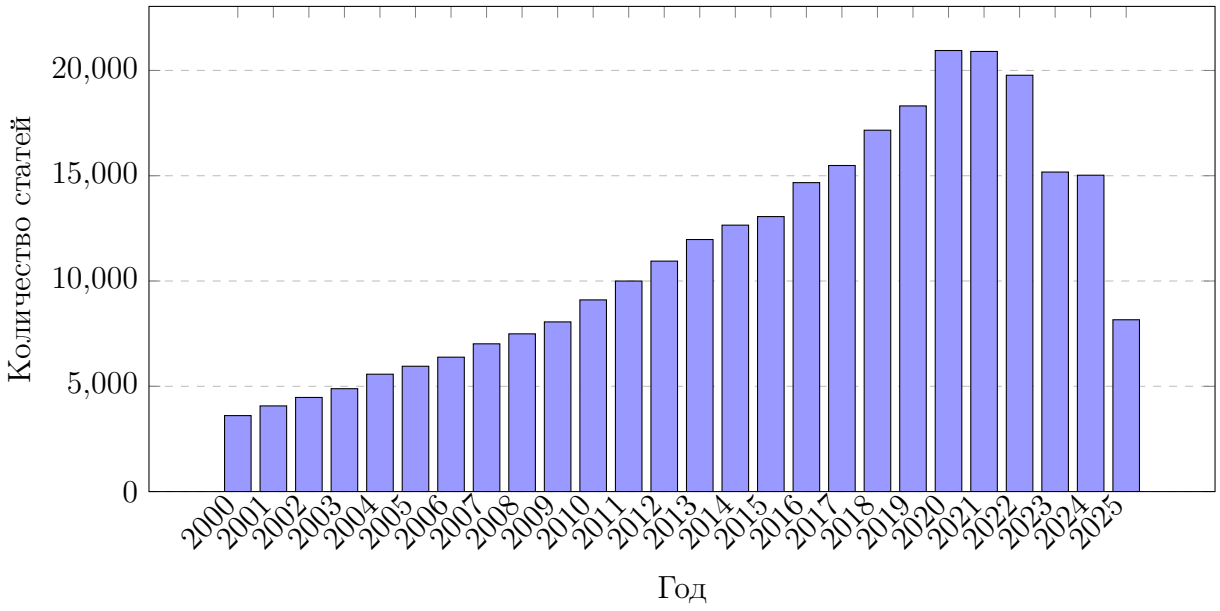

Рис. 2: Распределение числа собранных публикаций по годам

[4]Под тематическим направлением ООН подразумевается отнесение статьи к одной из глобальных Целей устойчивого развития (SDG), используемых Организацией Объединённых Наций для классификации исследований по тематическим областям.

Рост числа публикаций в 2020–2021 годах связан с глобальным всплеском исследований в области машинного обучения и появлением первых трансформерных архитектур, требующих значительных вычислительных ресурсов. Дальнейшее снижение в 2023–2024 годах является естественной нормализацией публикационной активности.

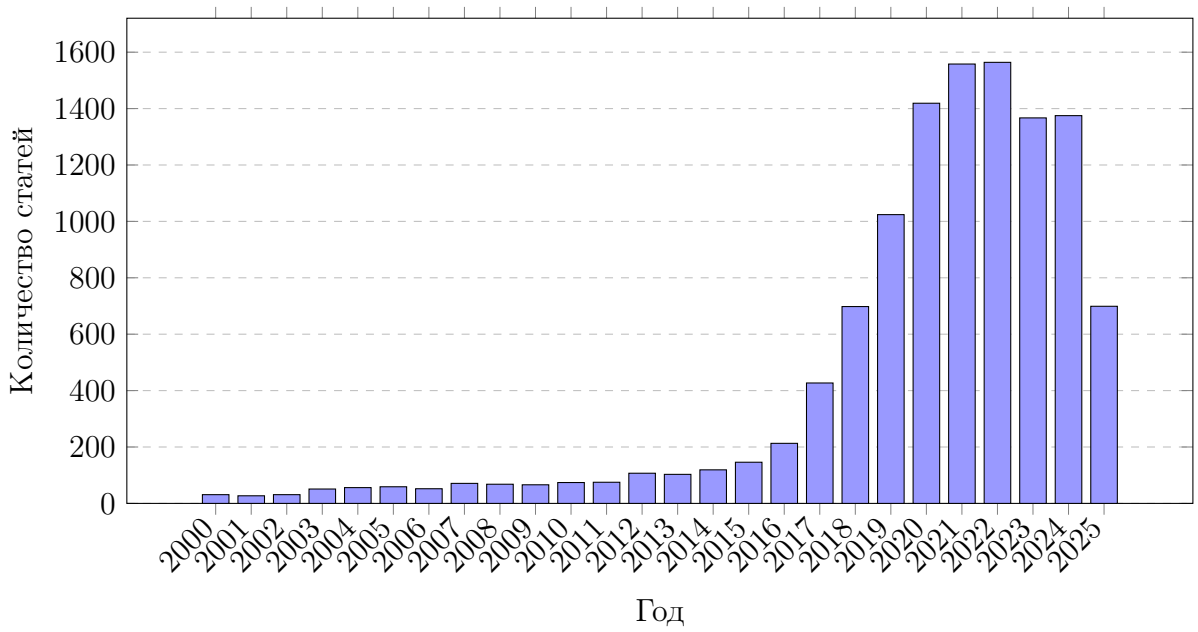

Рис. 3: Количество публикаций по тематике глубокого обучения и трансформеров (барплот)

Анализ числа работ по трансформерам показал значительную корреляцию между их ростом и общим увеличением публикационной активности в рассматриваемый период. Пик публикаций по данному направлению в 2020–2022 годах, а также последующее снижение к средним значениям совпадают с динамикой, наблюдаемой в общем корпусе работ.

2.3.2 Распределение публикаций по типам площадок

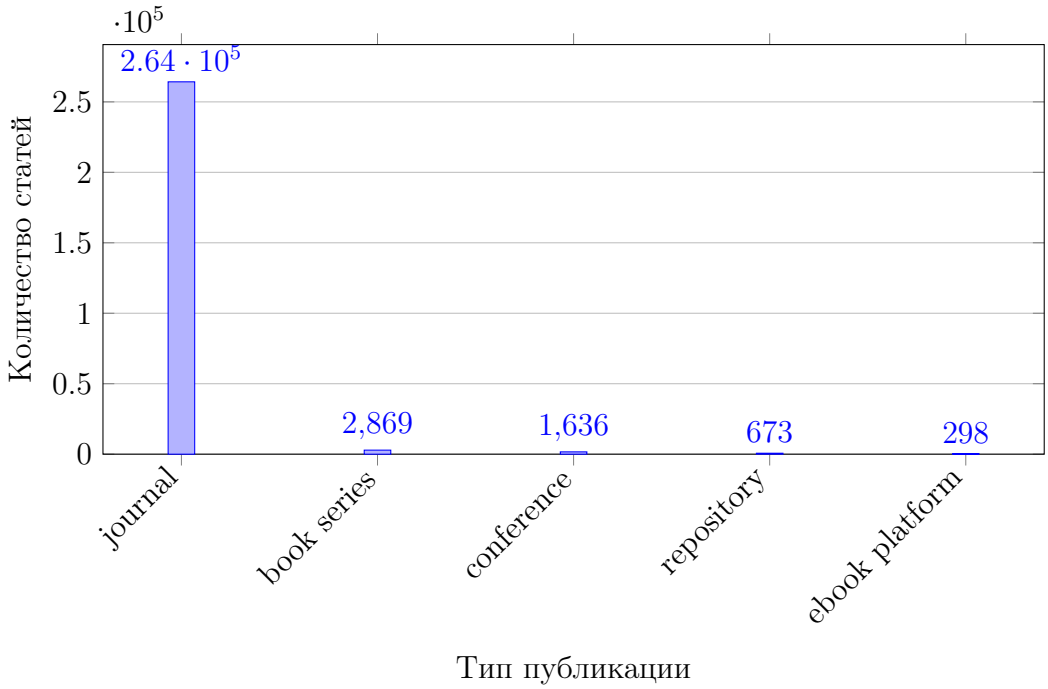

Рис. 4: Распределение публикаций по типам площадок

Большинство публикаций (более 260 тысяч) относится к журнальному формату. Доля конференционных работ и материалов других типов значительно ниже (менее одного процента). Следовательно, основным источником данных для анализа являются журнальные статьи.

2.3.3 Самые крупные площадки по числу публикаций

Площадка	Количество статей
The Astrophysical Journal	16102
Monthly Notices of the Royal Astronomical Society	15077
Astronomy and Astrophysics	12522
Physical Review Letters	9300
Physical Review D	7594
Physical Review B	6794
Journal of High Energy Physics	6486
Physical Review E	4425
The Journal of Chemical Physics	3953
Physics Letters B	3690
Physical Review A	3488
Journal of Cosmology and Astroparticle Physics	2839
The European Physical Journal C	2740
The Astrophysical Journal Letters	2058
Journal of Computational Physics	1869
Physical Review C	1826

Таблица 1: Самые крупные журналы по числу публикаций

Наиболее крупные площадки в выборке представлены журналами по астрофизике, физике высоких энергий и смежным фундаментальным дисциплинам.

Конференция	**Количество статей**
Proceedings of the AAAI Conference on Artificial Intelligence	160
Proceedings of the Genetic and Evolutionary Computation Conference	84
Proceedings of 37th International Cosmic Ray Conference — PoS(ICRC2021)	49
Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining	49
Proceedings of the 30th ACM International Conference on Multimedia	45
Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval	33
Proceedings of the Genetic and Evolutionary Computation Conference Companion	33
Proceedings of the 31st ACM International Conference on Information & Knowledge Management	32
Proceedings of the ACM Web Conference 2022	31
2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)	26
Interspeech 2022	25
2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)	22
IOP Conference Series Materials Science and Engineering	21

Таблица 2: Самые крупные конференции по числу публикаций

Анализ конференционных площадок показал, что наибольшее количество работ приходится на конференции, связанные с искусственным интеллектом, обработкой данных и информационным поиском.

2.3.4 Распределение числа цитирований

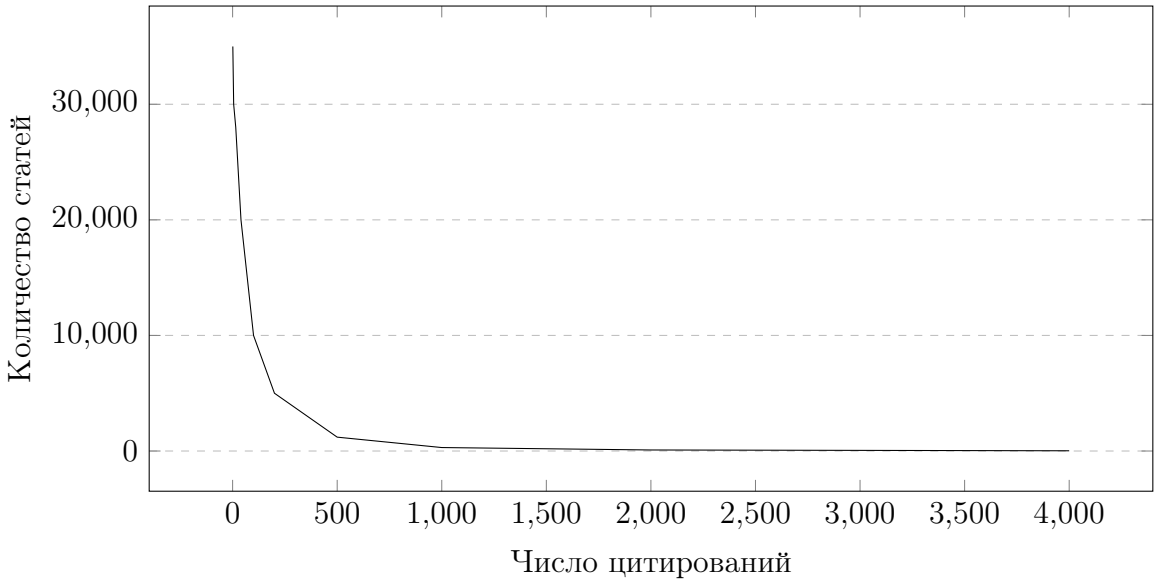

Рис. 5: Распределение числа цитирований

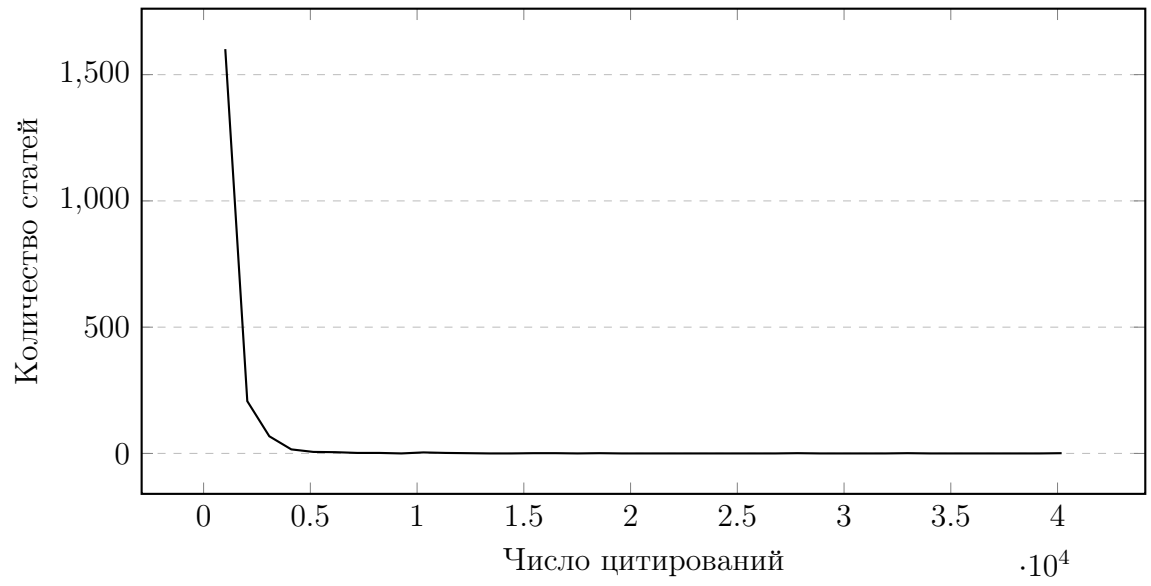

Рис. 6: Распределение числа цитирований среди высокоцитируемых статей (более 500 цитирований)

Большая часть корпуса представлена статьями с числом цитирований до 50. Высокоцитируемые работы встречаются редко и количественно почти не влияют на общую массу

публикаций, но они образуют «длинный хвост» распределения, отражающий наличие небольшого числа публикаций, получивших значительный научный отклик.

2.3.5 Наиболее популярные темы

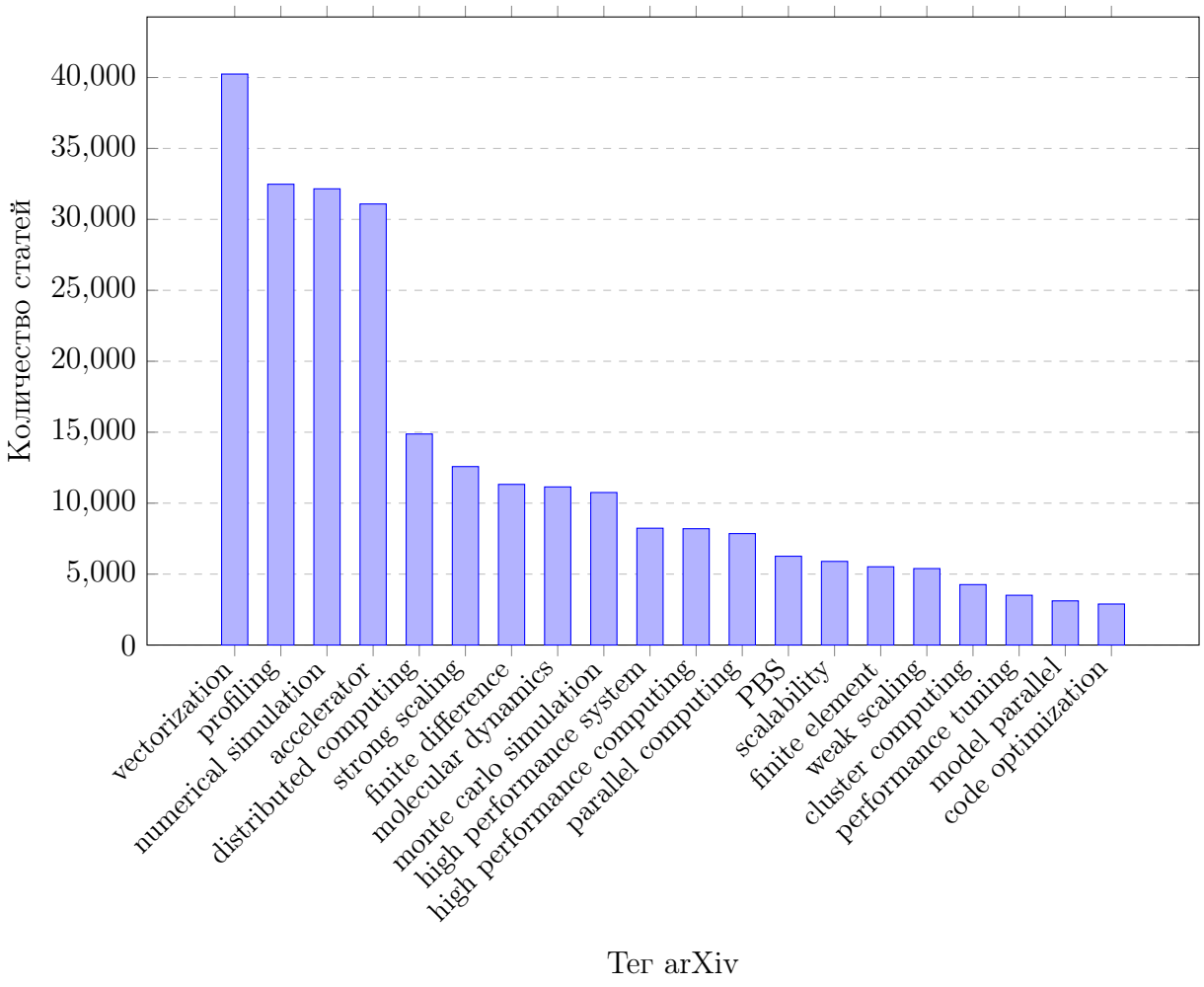

Рис. 7: Топ-20 наиболее популярных тегов

Анализ тегов arXiv показывает, что в выборке доминируют направления, связанные с оптимизацией производительности и численным моделированием. Наиболее распространённые теги — vectorization, profiling, numerical simulation и accelerator, что отражает фокус работ на повышении эффективности вычислений, работе с ускорителями и развитии различных параллельных методов.

2.4 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[5] и SPECTER[6]. Использование моделей SciBERT и SPECTER необходимо для решения основной задачи работы — построения рекомендательной системы, способной анализировать содержимое научных текстов и определять их тематическую близость. Эти модели позволяют получать качественные семантические представления (эмбеддинги) публикаций, что впоследствии используется для выбора релевантных научных площадок, выявления близких работ и анализа структуры исследовательских направлений в области HPC. Для оценки их характеристик были проведены тестовые эксперименты, а также выполнено сравнение с классическим BERT[7], 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 3: 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.

[5] `https://github.com/allenai/scibert`

[6] `https://github.com/allenai/specter`

[7] `https://github.com/google-research/bert`

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

Работа следующего семестра будет направлена на переход от анализа данных к формированию полноценной модели рекомендательной системы. Для этого планируется выполнить следующие шаги.

1. **Выявление значимых параметров и построение зависимостей.** Будет проведён детальный анализ собранной информации с целью определения факторов, влияющих на цитируемость и успешность публикаций. Планируется построение собственного графа цитируемости, что позволит:

   - выявлять центры научного влияния;
   - учитывать связи между научными работами;
   - использовать структурные характеристики (например, PageRank, метрики, аналогичные H-index) в модели рекомендаций.

2. **Интеграция дополнительных метаданных в модельные представления.** Проведённые эксперименты показали, что модели SciBERT и SPECTER уже обладают способностью извлекать качественные семантические представления из текстов научных статей. Для повышения качества тематического и контекстного анализа предполагается включение:

   - информации о площадке публикации (журнал, конференция);
   - года публикации;
   - аффилиаций авторов;
   - ключевых тем и SDG-направлений.

   Это позволит учитывать не только содержание статьи, но и её публикационный контекст и потенциальную видимость.

3. **Комбинация текстовых, графовых и метаданных.** Планируется объединить эмбеддинги SPECTER или SciBERT с характеристиками цитатного графа и метаданными, формируя многомерное представление каждой статьи. Такой подход обеспечит более точное определение тематической близости научных работ.

4. **Разработка прототипа рекомендательной системы.** На основе полученных данных планируется создание модели, способной:

    - предлагать релевантные журналы и конференции для публикации;
    - подбирать потенциальные авторские коллективы и смежные исследовательские направления;
    - прогнозировать потенциальную цитируемость новых статей.

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.

5 Приложение 1

В данном приложении приведён полный список ключевых слов, использованных для тематического поиска статей в области высокопроизводительных вычислений.

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, vectorization, SIMD, AVX, Xeon Phi, BlueGene, Cray, Fugaku, Frontier supercomputer, Summit supercomputer, 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