АВТОМАТИЗАЦИЯ ОБРАБОТКИ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ СТРУКТУРЫ МАТЕРИАЛОВ

Шиманский Н. А., Баглов А. В., Хорошко Л. С.
IDA Technologies
Кафедра физики твердого тела и нанотехнологий, физический факультет,
Белорусский государственный университет
Минск, Республика Беларусь
E-mail: nikita.shymanski@gmail.com, baglov@bsu.by, khoroshko@bsu.by

В данной работе рассмотрена возможность использования современных методов автоматизации и работы с большимиданными, а также использования современных облачных технологий для решения прикладных задач материаловедения на примере одного из самых распространенных методов анализа структуры кристаллическских материалов – дифракции рентгеновских лучей. Предложено программное решение, реализующее возможность эффективного масштабирования задач и глобализации базы данных. Приводятся примеры оптимизации решения задачи с применением разработанного решения.

Введение

Современная интеграция и глобализация международных исследовательских проектов делает актуальным создание общих баз данных, содержащих большое количество обработанных и верифицированных экспериментальных данных в виде, удобном для дальнейшего использования и обращения с ними. В настоящее время подобные базы данных, касающиеся структурных свойств материалов, могут быть как открытыми, так и предоставляемыми ограниченному кругу лиц в соответствии с приобретенной лицензией [1, 2], однако, конечная обработка результата поиска и сопоставление с собственным результатом осуществляется исследователем непосредственно, поскольку даже автоматизированные пакеты предполагают, во-первых, ручное сужение круга поиска, во-вторых, тщательное визуальное сопоставление и выбор окончательного решения.

С применением средств автоматизации, нейронных сетей и методов машинного обучения становится возможным не только предсказание строения и свойств наноматериалов, но и решение широкого ряда материаловедческих задач, а также снижение временных и трудовых затрат на эти процессы, что рассмотрено в данной работе на примере кейса рентгеновского дифракционного анализа материалов

I. Актуальность и задачи исследования

Рассмотрим частный случай решения распространенной задачи исследовательского материаловедения на примере обработки результатов одного из базовых методов исследований свойств и структуры кристаллических материалов – дифракции рентгеновских лучей (*XRD*, от англ. «*X-Ray Diffraction*»), детальный анализ результатов которого может дать исчерпывающую информацию о строении материала, однако сопряжён с большим количеством специфических операций обработки [3-5]. Такие факторы как уровень шума на дифрактограммах, базовая линия шума, наличие гало, разрывы данных и т.д. значительно увеличивают время на оценку и интерпретацию результатов. Для сложных дифрактограмм обработка и анализ результатов, особенно при отсутствии элементарной автоматизации, может превышать по времени проведение самого эксперимента *XRD*.

Также имеет место проблема оперативного предоставления доступа к результатам исследований заинтересованным исследователям, что на данный момент чаще всего реализуется созданием общих облачных хранилищ, пересылкой и-мейлов и т.п., но тем не менее, не отменяет многократной самостоятельной обработки результатов каждым исследователем в ручном режиме.

Современные объемы обрабатываемых данных, а также скорость их распространения все чаще требуют подходов *Big Data* (таких как *Data Warehouses / Data Lakes*), поскольку традиционные базы данных перестают демонстрировать устойчивую возможность обработки и анализа данных по мере роста их объема в хранилищах. Классические реляционные (а также *NoSQL*) базы демонстрируют высокие скорости записи и чтения данных, а также их чрезвычайную консистентность, но испытывают особенные трудности работы с данными в тех случаях, когда хранилище данных возрастает до определенных пределов (терабайты) и должно быть распределено. В таких случаях операции параллельного чтения/записи могут значительно замедляться, что в конечном итоге приводит к перегрузке и полному отказу сервисов. Облачная платформа способна сохранять, каталогизировать, классифицировать и исследовать «на лету» практически неограниченные объемы дифрактограмм и иных спектроскопических данных (в том числе необработанных/неизвестных), полученных в ходе экспериментов.

II. Описание реализации

Платформа создаваемого решения представляет собой использование совокупности облачных сервисов глобального провайдера *Amazon Web Services (AWS)*, что подразумевает высокую доступность для пользователей во всех регионах планеты. Выбранная конфигурация в виде совокупности таких сервисов как *AWS Cognito, AWS DocumentDB, AWS S3 (Data Lake), AWS Redshift / Spectrum, AWS SageMaker, etc.* гарантирует безотказный доступ к сервисам, позволяет горизонтально масштабировать нагрузки практически без ограничений, а также хранить и обрабатывать пользовательские данные, исчисляемые петабайтами.

Пользовательский интерфейс приложения позволяет проводить автоматизированную обработку дифрактограмм: удалить шум на графике, выровнять базовую линию шума, провести автоматический поиск пиков дифрактограммы и их позиций, а также провести аппроксимацию всех найденных или заданных пиков, и др.

III. Верификация результатов исследования

Методы математического моделирования, реализованные в приложении, позволяют обеспечить достаточно высокую точность в оценке распределения и позиционирования пиков *XRD*. Точность может быть дополнительно повышена увеличением количества итераций (циклов) при аппроксимации пика за счет дополнительного вычислительного ресурса.

По мере накопления материала в облачной базе данных достоверность прогнозов искусственного интеллекта растет, а в качестве стартового набора могут быть использованы как реальные дифрактограммы низкодефектных кристаллов, так и эталонные, сгенерированные специализированными научными программами (например, *Vesta* и др.). Полностью автоматическая оценка дифрактограмм, с точки зрения временных затрат на анализ полученных данных, значительно превосходит традиционную, а в рамках разрабатываемого проекта реализуется постоянное пополнение пользовательской базы данных и обучаемость моделей.

На примере постановки практической работы в учебной группе оценка затраченного времени для исследования типовой дифрактограммы показывает, что в среднем для полного анализа дифрактограммы в ручном режиме требуется от 2 часов, с использованием демо-версии лицензируемого платного пакета Match! от 20 мин, то платформе требуется в среднем 12–18 секунд для проведения автоматического анализа такой же дифрактограммы, при этом будет произведено устранение шумов и корректировка базовой линии.

IV. Выводы

Достоверность автоматизированного анализа с использованием нейросетей все еще является дискуссионным предметом в научном сообществе, но ее уровень, как и точность предсказаний, непрерывно растет по мере накопления банков данных *Data Warehouse / Date Lake* и совершенствования технологий в области *BigData* и *Machine Learning*, что открывает дальнейшие перспективы для дальнейшего углубления интеграции науки и информационных технологий. Методы обучения нейронных моделей при этом показывают перспективные результаты в области исследования и предсказания свойств наноматериалов [6]. Используемые вычислительные методологии не налагают строгих ограничений на источник и методики анализа материалов, которые могут быть автоматически обработаны, в частности, существует перспектива подобной автоматизации обработки результатов, полученных с помощью различных материаловедческих методов, включая Рамановское рассеяние (комбинационное рассеяние света), электронная Оже-спектроскопия, Резерфордовское обратное рассеяние, люминесцентная спектроскопия и др.

1. Crystallography Open Database [Электронный ресурс]. – Режим доступа: `http://www.crystallography.net/cod/`. – Дата доступа: 16.10.2023.
2. Cambridge Structural Database System [Электронный ресурс]. – Режим доступа: https://software.chem.ucla.edu/CSD/. – Дата доступа: 16.10.2023.
3. Автоматизация обработки результатов исследования структуры и свойств наноматериалов / Н.А.Шиманский, А.В.Баглов, Л.С.Хорошко // BIG DATA и анализ высокого уровня = BIG DATA and Advanced Analytics : сб. науч. ст. IX Междунар. науч.-практ. конф. В 2 ч. Ч. 1 (Республика Беларусь, Минск, 17–18 мая 2023 года) / редкол. : В.А. Богуш [и др.]. – Минск : БГУИР, 2023. –296-300.
4. Облачные решения Big Data&Machine Learning в сфере автоматизации материаловедческих исследований / Л.С. Хорошко, А.В. Баглов // Система «Наука–Технологии–Инновации»: методология, опыт, перспективы: материалы Международной научно-практической конференции, Минск, 28-29 сентября 2023 г. / Под ред. В.В. Гончарова. –Мн.: Центр системного анализа и стратегических исследований НАН Беларуси, 2023. – *в печати.*
5. Трушин, В.Н. Рентгеновский фазовый анализ поликристаллических материалов / В.Н. Трушин, П.В. Андреев, М.А. Фаддеев // Электронное учебно-методическое пособие. – Нижний Новгород: Нижегородский госуниверситет, 2012. – 89 с.
6. Использование методов глубокого обучения для решения научно-технических задач в области материаловедения / Л.С. Хорошко, А.В. Баглов // Система «Наука–Технологии–Инновации»: методология, опыт, перспективы: материалы Международной научно-практической конференции, Минск, 24-25 сентября 2020 г. / Под ред. В.В. Гончарова. –Мн.: Центр системного анализа и стратегических исследований НАН Беларуси, 2020. – С. 578–582.