

АЛГОРИТМЫ ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА ИНФОРМАЦИИ В РАСПРЕДЕЛЕННЫХ БАЗАХ ДАННЫХ С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Скалозуб К. А., Ярмош А. Д., Романюк М. В.

Отдел информационных технологий,
центр информатизации и инновационных разработок,
Белорусский государственный университет информатики и радиоэлектроники

Минск, Республика Беларусь

E-mail: {k.skalozub, a.iarmosh, romanuk}@bsuir.by

В статье рассматриваются алгоритмы интеллектуального поиска информации в распределенных базах данных с применением методов машинного обучения. Анализируются подходы к организации поиска, включая использование методов классификации, кластеризации и ранжирования результатов с применением нейросетевых моделей. Показана роль машинного обучения в адаптивной обработке запросов, повышении точности и скорости поиска. Приводятся современные направления развития интеллектуальных алгоритмов в области информационного поиска и их применение в распределенных системах.

ВВЕДЕНИЕ

С ростом объемов данных и числа распределенных хранилищ становится все более актуальной задача эффективного поиска информации в распределенных системах. Традиционные методы информационного поиска (IR, Information Retrieval) часто оказываются недостаточно эффективными при масштабировании: проблемы распределения данных, сетевые задержки, неоднородность источников и высокая нагрузка усложняют задачу.

Интеграция методов машинного обучения (ML) в алгоритмы поиска позволяет повысить качество ранжирования, предсказать релевантность, адаптировать индексацию под нагрузку и динамику данных. В распределенных базах данных (например, распределенных хранилищах, облачных СУБД, мультиагентных схемах) добавляются еще следующие задачи: как обучать модели, когда данные не централизованы, как минимизировать передачу данных, как справляться с задержками, с несинхронностью и т.д.

Цель статьи – рассмотреть основные классы алгоритмов интеллектуального поиска в распределенных базах данных, показать, как методы ML интегрируются в них, выявить трудности и направления для исследований.

1. ОСНОВНЫЕ ПОДХОДЫ АЛГОРИТМОВ ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА С ML

1. Локальные модели и глобальная агрегация. На каждом узле строится локальная модель (например, классификатор, ранжировщик), которая оценивает релевантность документов по запросу. Затем результаты с разных узлов объединяются с учетом дополнительных сигналов (например, доверие узла, статистика качества) и происходит окончательный ранжир [1].
Плюсы: модель обучается локально, не требуется сбор всех данных в одном месте.

Минусы: различие распределений между узлами может ухудшать качество, агрегация сложна.

2. Распределенное обучение (Data/Model Parallelism). Методы распределенного машинного обучения, такие как обучение с разделением данных или с разделением модели, применяются для обучения моделей, используемых для поиска/ранжирования. Особенно актуально, когда признаки или представления документов (векторы, embedding) используются в поиске.
Пример: использование TensorFlow в распределенных системах для обучения нейронных моделей поиска.
3. Обучаемые индексы и итеративное перераспределение данных. Идея: традиционные индексы (B-дерево, хэш-таблицы) заменяются или дополняются ML-моделями, которые прогнозируют положение записи. В распределенном контексте это может сочетаться с динамическим перераспределением данных для балансировки нагрузки и ускорения доступа.
Пример: работа IRLI (Iterative Repartitioning for Learning to Index – метод итеративного перераспределения данных для оптимальной индексации в распределенной среде).
4. Обучение с активным выбором запросов. Метод активного обучения может уменьшить затраты на разметку и улучшить качество определения релевантности документов, особенно в контексте информационного поиска [2].
5. Объяснимые методы поиска (Explainable IR). Поскольку модели поиска могут быть сложными (нейросети, градиентный бустинг и др.), важна интерпретируемость – чтобы понимать, почему система выдала тот или иной результат.

II. СТРУКТУРА ГИБРИДНОЙ АРХИТЕКТУРЫ ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА

1. Партиционирование/шардирование данных. Данные (документы) распределяются по узлам по какому-то ключу (например, тематическому, хеш-функции, географическому признаку). Может быть использовано динамическое перераспределение (repartitioning) в зависимости от нагрузки.
2. Индексация и эмбединг. На каждом узле формируется локальный индекс (например, инвертированный) и представления документов в виде эмбедингов (векторов в пространстве признаков) для более интеллектуального поиска сходства [3].
3. Обучение локальных моделей. На каждом узле обучается модель релевантности (например, градиентный бустинг, нейронная сеть) на локальной выборке, учитывая запросы и клики с этого узла.
4. Обмен метаинформацией/синхронизация. Узлы обмениваются некоторыми сводными статистиками, обобщенными моделями или агрегированными параметрами (например, через алгоритмы Federated Learning). Это может улучшать глобальную модель без передачи всех данных.
5. Маршрутизация запросов. Перед выполнением запроса система решает, на какие узлы его отправлять: полный цикл, часть узлов по тематике, узлы с высоким показателем качества и т.д.
6. Агрегация результатов и коррекция ранжирования. Отбираются наиболее релевантные результаты от каждого узла, затем выполняется агрегация с учетом доверия к узлам, статистики качества, глобальных признаков и возможно применения финального ранжировщика [4].
7. Обновление/дообучение. Периодически (или онлайн) происходит обмен сведениями, дообучение моделей, перераспределение данных и обновление индексов, чтобы система адаптировалась к новым данным и паттернам запросов.

III. ПРОБЛЕМЫ, ОГРАНИЧЕНИЯ И НАПРАВЛЕНИЯ ИССЛЕДОВАНИЙ

- Коммуникационные издержки: как минимизировать передачу данных между узлами, особенно при обмене признаков или моделей;
- неоднородность распределений: как справляться с ситуацией, когда узлы содержат данные разных тематик/доменов;
- конфиденциальность и приватность: важно сохранять приватность при обмене данными и параметрами моделей;
- временная изменчивость данных: как обеспечивать адаптивность (онлайн-обучение, адаптация к дрейфу данных);
- интерпретируемость: пользователям и администраторам важно понимать, почему был выбран тот или иной результат;
- оценка качества: как оценивать качество распределенного поиска, учитывая сеть, задержки, неполные соответствия между узлами;
- совместимость с существующими СУБД: как встроить ML-алгоритмы в существующие распределенные СУБД, не переписывая все «с нуля».

ЗАКЛЮЧЕНИЕ

Алгоритмы интеллектуального поиска информации в распределенных базах данных – тема, сочетающая вызовы из области информационного поиска, распределенных вычислений и машинного обучения. Применение ML-методов позволяет повысить качество поиска, адаптироваться к нагрузке и динамике, но требует продуманных архитектур распределенного обучения, обмена информацией и агрегирования результатов.

Рекомендации:

- Для старта можно использовать гибридную стратегию: локальные модели и обмен сводной информацией;
- при обучении моделей использовать распределенный или федеративный подход, чтобы избежать централизованной передачи данных;
- исследовать алгоритмы перераспределения данных и обучаемых индексов в распределенных средах;
- внедрять элементы объяснимости и оценивать систему не только с точки зрения точности, но и с точки зрения задержек, коммуникаций и адаптивности.

СПИСОК ЛИТЕРАТУРЫ

1. Трофимов, И. Е. Распределенные вычислительные системы для машинного обучения / И. Е. Трофимов // Информационные технологии и вычислительные системы. – 2017. – № 3. – С. 64–65.
2. Аныш, Х. Применение инструментов машинного обучения и интеллектуальный анализ данных в отношении баз данных с небольшим количеством записей / Х. Аныш // Информатика, вычислительная техника и управление. – 2021. – С. 347–359.
3. Raschka, S. Machine Learning with PyTorch and Scikit-Learn / S. Raschka, Y. Liu, V. Mirjalili – Packt Publishing, 2022. – P. 649–660.
4. Макшанов, А. В. Современные технологии интеллектуального анализа данных: учебное пособие для СПО / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь – СПб.: Лань, 2020. – С. 12–19.
5. Львовский, С. М. Набор и верстка в системе LaTeX / С. М. Львовский // Издательство: МЦНМО, 2006. – 448 с.