Парфенов Денис Васильевич

Российский технологический университет, Институт кибернетики
promasterden@yandex.ru

1 Эффективное нахождение пересечения N неупорядоченных дискретных множеств

Тема рекомендована студентам 2–4 курсов и магистрантам и допускает варианты:

1.1 Сравнительный обзор методов отыскания пересечения неупорядоченных множеств

1.2 Методы отыскания пересечения неупорядоченных множеств с применением хэш-функций

1.3 Вычислительно эффективные методы нахождения пересечения N множеств

1.4 Проблема выбора хэш-функций для отыскания пересечения неупорядоченных множеств

Данная задача характеризуется тем, что:

- имеет яркую практическую направленность (в частности, важна в современных системах хранения и обработки больших данных (big data) и искусственного интеллекта),
- её постановка вполне очевидна,
- оптимальный в общем случае с позиций минимизации вычислительной сложности алгоритм решения до сих пор неизвестен,
- известно несколько относительно несложных и хорошо понятных субоптимальных алгоритмов,
- задача характеризуется значительным потенциалом для дальнейших исследований,
- допускается применение численного эксперимента.

Ссылки на библиографию и программные реализации современных методов отыскиваются лучше в англоязычных источниках по ключевым словосочетаниям: unordered set intersection, hash functions, hash-based dictionaries. Старые подходы, связанные с сортировкой, предлагается игнорировать.

2 Методы разрешения хэш-коллизий

Тема рассчитана на студентов 2–4 курсов и магистрантов.

Работа с хэш-функциями сегодня служит доминирующим методом быстрого доступа к данным. Непреодолимый изъян даже лучших конструкций хэшей - возможность так называемого хэш-промаха, т.е. ситуации соответствия многим различным элементам данных одного и того же

значения хэш-функции. В результате различение таких данных требует дополнительных усилий. Степень проявления проблемы зависит как от грамотного выбора метода хэширования с учётом свойств данных, так и от метода обработки таких промахов. Предлагается проанализировать литературу и написать обоснованный сопоставительный обзор современных методов разрешения хэш-коллизий. Ознакомление с вопросом рекомендуется начать с
`https://en.wikipedia.org/wiki/Hash_table#Collision_resolution`,
далее изучить теорию по соответствующему разделу классической книги Д.Э. Кнута "Искусство программирования для ЭВМ" (лучше воспользоваться более новым англоязычным изданием, D.E. Knuth, The Art of Computer Programming), затем освоить современные журнальные публикации, легко отыскиваемые по соответствующим ключевым выражениям.

3 Алгоритмическая поддержка хранилища интервальных данных

Тема рассчитана на студентов 3–4 курсов и магистрантов.

Стандартный арсенал средств алгоритмической организации операций с хранилищем дискретных данных (таких как список, упорядоченное и неупорядоченное множество и т.д.) поддерживает такие атомарные операции, как поиск элемента, вставка нового элемента, удаление элемента. Известны эффективные алгоритмы осуществления этих операций.

Открытой исследовательской задачей является синтез эффективных алгоритмов для аналогичных операций с хранилищем интервальных данных. Каждый элемент данных представляет собой интервал значений некоторой *ключевой* вещественной величины. Хранилище представляет собой объединение таких интервалов: дублирование информации в хранилище и пересекающиеся интервалы в нём исключены. Алгоритмы работы должны обеспечивать вычислительно эффективное выяснение вхождения проверяемого интервала в хранилище (полного или частичного, посредством операции пересечения), удаление проверяемого интервала из хранилища (также с обработкой ситуаций его отсутствия, полного или частичного наличия в хранилище), вставки интервала. Данная проблематика важна для разработки систем хранения данных нового поколения.

4 Модификации метода эмпирических мод для анализа временных рядов

Тема рассчитана на студентов 2–4 курсов и магистрантов.

Метод эмпирических мод (N. Huang, 1995, 1998) прочно вошёл в инструментарий прикладного анализа временных рядов, в том числе в экономике. Основные причины:

- устойчивость к нестационарности временных рядов,
- способность автоматически адаптироваться под их структуру на разных масштабах частотно-временной шкалы,
- адекватное моделирование существенно нелинейных процессов,
- высокая агрегирующая способность,
- малая вычислительная сложность.

Тем не менее, оригинальный вариант, предложенный Н. Хуангом, имеет несколько характерных недостатков, среди которых — неоднозначность разделения компонент, артефакты при изменении масштаба одной из компонент составного процесса и другие. Это стимулирует дальнейшие исследования и появление новых модификаций метода. Их рассмотрению и сопоставлению и предлагается посвятить работу. Рекомендуемая стартовая точка:
`https://ru.wikipedia.org/wiki/Empirical_Mode_Decomposition`,
затем англоязычная версия этой статьи, которая выведет на словосочетания для поиска нового материала.

5 Сравнение методов минимаксного приближения функций

Тема рассчитана на студентов 3–4 курсов и магистрантов.

Наиболее известным методом минимаксного (в смысле Чебышёва) приближения функций на заданном множестве узлов является алгоритм Е.Я. Ремеза (Remez algorithm). С него и стоит начать, разобравшись в таких аспектах, как область применимости, точность и объём вычислений, вычислительные проблемы. Далее логично ознакомиться с его модификациями, которые ослабляют, но не снимают принципиальных ограничений. Основная цель работы – сравнение с более новыми подходами, в частности, Osborne – Watson и развиваемыми А.Б. Богатырёвым в задаче синтеза фильтров.

6 Численные методы отыскания коэффициентов модели временных рядов ARIMA

Тема рассчитана на студентов 3–4 курсов и магистрантов.

ARIMA является относительно старым, хорошо изученным и зарекомендовавшим себя инструментом предсказания временных рядов, до сих пор интенсивно применяемым в прикладной статистике и экономике. Имеется огромная библиография по данному вопросу. Однако, очень часто за рамками описания остаются конкретные численные методы нахождения порядков модели и её авторегрессионных и остаточных коэффициентов, а они не единственны. С этим связана путаница, когда разные (корректные!) реализации одного теоретического подхода дают сильно различающиеся результаты. Цель предлагаемой работы - ознакомление с двумя-тремя наиважнейшими и их численное сравнение на искусственных (сгенерированных пользователем) и реальных наборах данных (data sets, traces), которые легко добываются в соответствующих сетевых ресурсах.