

Вероятностные структуры данных - фильтр Блума

Вергунова А.Е., студентка,

Кубанский государственный технологический университет,

г. Краснодар, Россия

Аннотация: Представлены методы для решения задач хранения и поиска информации на примере фильтра Блума. Приведено описание внутренней структуры фильтра Блума и принципы его работы. Такая структура данных используется в обработке больших объемов данных для эффективного определения принадлежности элемента к множеству, когда задача должна решаться в сжатом размере памяти и быстро. Предложены способы улучшения фильтра, и обсуждаются стратегии для кластеризации и распараллеливание системы, системы с возможностью удаления.

Ключевые слова: большие данные, вероятностные структуры данных, хеш-функция, хранение, фильтр Блума, эффективность, минимизация хранения.

Фильтр Блума – это вероятностная структура данных [1], позволяющая хранить некое множество элементов, а также быстро отвечать на запрос о том, есть ли данный элемент во множестве или нет. При этом существует возможность получить не правильный результат. Если система работы с данными толерантна к незначительным накладным расходам на ошибки, то с помощью фильтра Блума можно значительно повысить ее производительность. Поэтому фильтр Блума нашел применение во многих сферах - в больших данных [2, 3], в дедупликации [4], в сетевых технологиях, включая информационную безопасность [5], файловых системах [2], индексировании и т.п. Фильтр Блума неприменим там, где предъявляются строгие требования к ответам системы, например, базам данных, содержащим данные пользовательской аутентификации (логины, пароли и т.п.).

...

Полный текст во вложении