УДК 004.627

ИССЛЕДОВАНИЕ СЖАТИЯ ПРЕФИКСНОГО ДЕРЕВА В МЕТОДЕ ПРЕФИКСНОЙ ДЕДУПЛИКАЦИИ ДАННЫХ

Богомолов Д.С. (Университет ИТМО)
Научный руководитель – к.т.н. Балакшин П.В.
(Университет ИТМО)

В работе произведена оценка влияния разработанных способов сжатия и восстановления префиксного дерева на характеристики метода префиксной дедупликации данных. Работа содержит описание разработанных решений, а также рекомендации по их реализации и конфигурированию, позволяющие достичь повышения скорости обработки блоков и снижение расходов оперативной памяти.

Введение. Большие объемы данных и высокие скорости их увеличения рождают спрос на системы позволяющие снизить расходы на хранение данных. Дедупликация – это один из подходов, применяемый при построении подобных систем, позволяющий уменьшить избыточность данных за счет их представления ссылками на уникальные блоки данных из хранилища. Метод префиксной дедупликации – представитель малоизученного безхешевого типа дедупликации, использующий префиксное дерево в качестве структуры для поддержания индекса уникальных блоков. В методе каждая ветвь префиксного дерева хранит последовательность сегментов достаточную для идентификации блока, на который лист ветви указывает. Данный подход уже показал свою эффективность, однако исследован не во всех его областях.

Открытым местом для проведения исследований является структура хранения префиксного дерева и его реализация. Для поддержания структуры дерева требуется память под хранение метаданных, что означает сокращение пространства под сегменты и ссылки. В работах посвященных методу, структура префиксного дерева подразумевала хранение сегментов только одного порядкового номера в узлах префиксного дерева. В результате для разрешения коллизии создается ветвь, состоящая из узлов, хранящих один сегмент, начиная с узла, на котором была обнаружена коллизия, до узла, в котором она была разрешена. На практике это означает преобладание в префиксных деревьях наименее эффективных узлов с точки зрения затрат памяти под метаданные на один хранимый сегмент.

Целью работы является исследование влияния сжатия ветвей префиксного дерева, создаваемых при разрешении коллизий, на характеристики метода префиксной дедупликации данных путем хранения сегментов разного порядка в рамках одного узла. Для достижения цели были поставлены и выполнены следующие задачи:

- теоретический анализ влияния сжатия на характеристики метода;
- выработка параметров сжатия, влияющих на эффективность его применения;
- разработка решения по сжатию и восстановлению узлов с учетом возможного конфигурирования на основе существующего прототипа;
- проведение экспериментов над прототипом при различных конфигурациях;
- анализ результатов экспериментов;
- формулировка выводов.

Основная часть. Для анализа влияния были выбраны две ключевые характеристики: расход метаданных и скорость обработки блока данных. При анализе был учтен тот факт, что в случае нахождения коллизии на одном из сегментов сжатого узла, для ее разрешения необходимо восстановить узел. Тем не менее, это не ограничивает возможность хранения других сегментов узла в сжатом виде.

В результате анализа было установлено то, что чем больше размер используемых узлов, тем меньше расход памяти, однако и вероятность восстановления этого узла тоже возрастает с его размером. На вероятность восстановления узла влияет также и диапазон уровней расположенных в нем сегментов: чем дальше от корня дерева, тем ниже вероятность. Вероятность восстановления в свою очередь определяет эффективность сжатия с точки зрения

временных затрат: чем чаще восстанавливаются узлы, тем больше временных затрат на восстановление и тем меньше итоговая выгода, получаемая от сокращения количества создаваемых узлов и увеличение локальности сегментов при обходе дерева. По результатам анализа были выработаны параметры влияющие на эффективность сжатия: количество сегментов, поддерживаемых узлом, уровень начиная с которого разрешено использование узла для сжатия.

На основе полученных параметров, исследований структур данных и существующего прототипа метода было получено комплексное решение по сжатию и восстановлению узлов. Ключевым в понимании этого решения является представление ветви после коллизии в качестве связного списка. В разработанном решении вместо связного списка использовалась его модификация – развёрнутый связный список, которую можно представить как связный список массивов. Спроектированное решение позволило задать размер массивов и уровень его применения, тем самым поддерживая возможность конфигурации в зависимости от выделенных факторов. На эту же конфигурацию полагалось решение по восстановлению узлов, позволяющее разбить сжатый узел на совокупность сжатых узлов меньшего или равного размера и обыкновенных узлов. Над модифицированным прототипом для различных конфигураций системы сжатия и восстановления были проведены эксперименты на наборе данных из резервных копий гетерогенных систем с целью эмпирически доказать целесообразность использования системы. В результате экспериментов удалось подтвердить эффективность разработанных решений: в частности, достичь сокращения расходов оперативной памяти в два раза и сократить скорость обработки на 30 %.

Выводы. Исследование показало эффективность применения сжатия префиксного дерева с целью повышения скорости обработки блоков и снижения расходов оперативной памяти в методе префиксной дедупликации данных. В работе приведены алгоритмы и рекомендации по их реализации и конфигурированию, позволяющие достичь подобного результата. Использование разработанных решений и рекомендаций позволит не только повысить значимые характеристики программных реализаций метода, но и увеличить класс систем, для которых данный метод может быть применен.

Богомолов Д.С. (автор)

Балакшин П.В. (научный руководитель)