

МЕХАНИЗМ КЭШИРУЮЩИХ ССЫЛОК В КОНТЕКСТНОМ МОДЕЛИРОВАНИИ

Контекстное моделирование, применяемое в ряде методов сжатия данных, заключается в накоплении статистики о символах входного потока, на основании которой моделируются вероятностные взаимосвязи между ними, что позволяет использовать энтропийные методы сжатия (например, арифметическое кодирование), требующие выполнения оценок вероятностей появления символов для эффективного сжатия данных. Идентичность статистической модели у кодера и декодера гарантирует корректное выполнение адаптивного сжатия. [1]

Рациональным выбором структуры данных для хранения вероятностного распределения символов алфавита для каждого из встреченных контекстов является балансируемое по весу узлов двоичное дерево поиска. Под весом будем понимать «востребованность» узла, т.е. количество раз, когда текущие символы были закодированы или декодированы в данном узле. [2]

Существует два вида древовидных структур, используемых в модели: деревья вероятностей и деревья префиксов. Оба вида деревьев являются BST и в качестве ключа используют символ, чья частота накапливается в счетчике (дерево вероятностей) или символ, являющийся префиксом (первым символом) текущего контекста (дерево префиксов).

Деревья вероятностей хранят счетчики относительных частот символов, встреченных в данном контексте, и манипулируют ими. При этом они являются самобалансирующимися по весу.

Деревья префиксов образуют иерархическую структуру, в которой каждый узел дерева префиксов может содержать указатель на другое дерево префиксов, называемое деревом префиксов контекста более высокого порядка.

Узел дерева префиксов является контекстом, порядок которого соответствует порядку его дерева. Для каждого контекста дерева префиксов более глубокого уровня родительский контекст является суффиксом.

Помимо этого каждый узел дерева префиксов содержит указатель на соответствующее этому контексту дерево вероятностей (дерево вероятностей контекста).

Существенный отрицательный вклад в быстрдействие алгоритма, как показано далее, вносит необходимость навигации по иерархии деревьев контекстной модели.

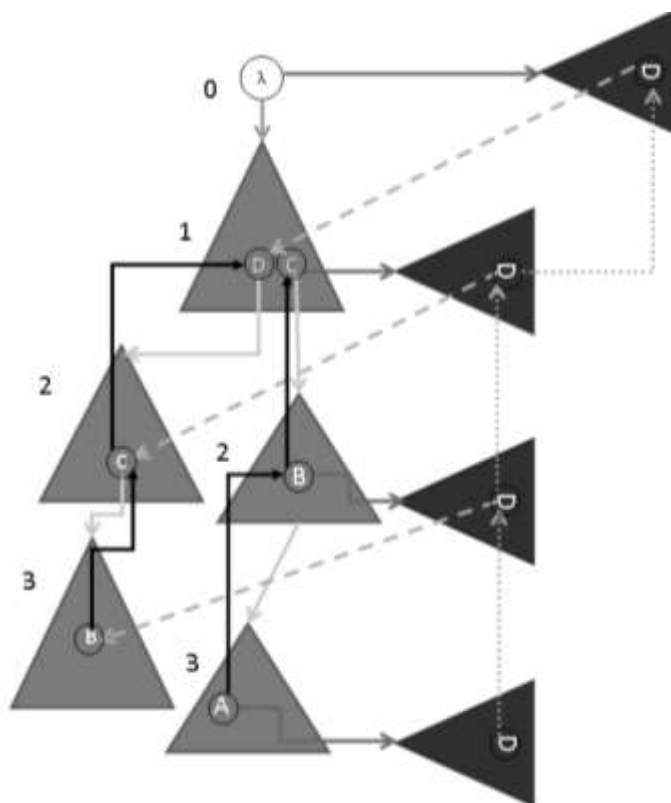


Рис. 1 Иерархическая контекстная модель

На рисунке 1 темными треугольниками обозначены деревья вероятностей, светлыми – деревья префиксов. Светлыми стрелками отмечены указатели на деревья суффиксов контекстов более высокого порядка, темными – указатели из контекстов на деревья вероятностей данных контекстов. Кругами внутри деревьев обозначены их узлы, а кругом с символом λ – вырожденный контекст нулевого порядка. Числами обозначены порядки контекстов.

При поиске узлов обоих видов деревьев осуществляется спуск по этим деревьям.

Наибольший интерес на рисунке 1 представляют собой ссылки черного цвета, а также пунктирные и точечные стрелки, обозначающие ссылки, называемые кэширующими. Рассмотрим их назначение подробнее.

Ссылка черного цвета (в листинге 1 – SUFF) ответственна за переход к суффиксу текущего контекста. Необходимость такой ситуации возникает в том случае, когда дерево вероятностей текущего контекста не включает в себя кодируемого символа, и на выход контекстного моделировщика подается символ ухода. После его кодирования моделировщик должен сделать откат до суффикса текущего контекста. [3] В обычных обстоятельствах мы должны были бы начать обзор с контекста нулевого порядка и, последовательно спускаясь по деревьям префиксов более высоких уровней, достичь суффикса текущего контекста. При этом количество спусков по несбалансированным деревьям префиксов (время поиска элементов имеет порядок $O(n)$) прямо пропорционально порядку текущего контекста, что в среднем составляет около 6 спусков. В случае же использования ссылки SUFF все сводится к одному переходу по ней, требующему константного времени.

Ссылка SUFF создается при добавлении нового узла в дерево префиксов более высокого порядка.

Пунктирная ссылка (в листинге 1 – NEXT) соединяет узел дерева вероятностей с контекстом, в котором необходимо будет осуществлять кодирование следующего символа. Действительно, если мы успешно закодировали символ «D» в контексте «ABC», то следующий символ, какой бы он ни был, мы должны пытаться закодировать в контексте «ABCD». Таким образом, следующий контекст можно всегда предсказать совершенно точно, и ссылка NEXT служит для ускоренного перехода к нему. Без использования ссылки NEXT при кодировании нового символа приходилось бы вручную спускаться по иерархии деревьев суффиксов, что, как уже было показано, чрезвычайно неэффективно.

Ссылка NEXT, как и SUFF, создается при добавлении нового узла в дерево префиксов более высокого порядка.

Точечная ссылка (в листинге 1 – PREV) соединяет узел дерева вероятностей с соответствующим тому же символу узлом дерева вероятностей суффикса текущего контекста. Ее применение обоснованно в той ситуации, когда после успешного кодирования мы хотим перейти к новому контексту, но он еще не создан, так как раньше не встречался. В этом случае нам необходимо его создать, и требуется доступ к дереву префиксов, в которое нужно вставить узел соответствующего контекста. Так, если контекст «ABCD» еще не создан, нам нужно получить доступ к дереву префиксов контекста «BCD», находясь в узле «D» дерева вероятностей контекста «ABC». Как и в предыдущих двух случаях, в наивном алгоритме поиска соответствующего контекста требуется полный спуск по всей иерархии деревьев префиксов до достижения нужного. При использовании же ссылки PREV осуществляется переход к узлу «D» дерева вероятностей суффикса «BC» контекста «ABC», после чего по ссылке NEXT осуществляется переход к контексту «BCD». Контекстом более высокого порядка «ABCD» будет узел дерева префиксов контекста «BCD», поэтому осуществляется однократный спуск по дереву префиксов и создание в нем узла «A». При этом сразу же осуществляется связь узла «D» дерева вероятностей контекста «ABC» с только что созданным контекстом «ABCD» созданием ссылки NEXT и контекста «ABCD» с контекстом «BCD» ссылкой SUFF.

Ссылка PREV создается после добавления узла дерева вероятностей при осуществлении отката до суффикса текущего контекста.

В листинге 1 приведено описание построения контекстной модели при кодировании или декодировании файла в форме псевдокода с использованием кэширующих ссылок. Полуужирным шрифтом выделены действия, требующие спуска по деревьям. Как видно из псевдокода, в цикле осуществляется только спуск по сбалансированному дереву вероятностей. Спуск же по дереву контекстов осуществляется одновременно для каждого кодируемого символа.

Псевдокод намеренно упрощен отсутствием рассмотрения обработки ситуации отката к контексту порядка -1.

Листинг 1 – Псевдокод построения контекстной модели

принять на вход символ S

ЕСЛИ (в дереве вероятностей узла текущего контекста отсутствует узел S)

ТО

добавить в дерево вероятностей узла текущего контекста узел S

запомнить узел S как NODE

закодировать символ ухода в дереве вероятностей узла текущего контекста

сделать откат до суффикса текущего контекста по ссылке SUFF

сделать суффикс текущего контекста текущим контекстом

ЕСЛИ (в дереве вероятностей узла текущего контекста отсутствует узел S)

ТО

ДЕЛАТЬ

добавить в дерево вероятностей узла текущего контекста узел S

создать ссылку PREV из узла NODE к узлу S

запомнить узел S как NODE

закодировать символ ухода в дереве вероятностей узла текущего контекста

сделать откат до суффикса текущего контекста по ссылке SUFF

сделать суффикс текущего контекста текущим контекстом

ПОКА (в дереве вероятностей узла текущего контекста отсутствует узел S)

создать ссылку PREV из узла NODE к узлу S

КОНЕЦ ЕСЛИ

КОНЕЦ ЕСЛИ

закодировать S в дереве вероятностей узла текущего контекста
ЕСЛИ (в узле S определена ссылка NEXT)

ТО

перейти к контексту более высокого порядка по ссылке NEXT
сделать контекст более высокого порядка текущим контекстом
ИНАЧЕ

запомнить узел S как NODE

перейти к узлу S дерева вероятностей суффикса текущего контекста по ссылке PREV

перейти к контексту более высокого порядка по ссылке NEXT
сделать контекст более высокого порядка текущим контекстом
запомнить текущий контекст как СТХТ

добавить в дерево префиксов текущего контекста узел X

сделать префикс X текущего контекста текущим контекстом

создать ссылку SUFF из текущего контекста к контексту СТХТ

создать ссылку NEXT из узла NODE в текущий контекст

КОНЕЦ ЕСЛИ

В данной статье представлены разработанная и структурированная иерархическая контекстная модель, хранящая статистические данные кодируемой информации, а также система кэширующих ссылок, используемая для ускорения доступа к различным контекстам в типичных сценариях кодирования.

Библиографический список

1 Аксенов А.В. Разработка адаптивного алгоритма энтропийного кодирования, Сборник докладов Научной сессии ГУАП, посвященной Всемирному дню космонавтики, СПб, ГУАП, 2010 г.

2. Аксенов А.В. Балансировка деревьев статистической модели энтропийного кодировщика, Сборник докладов Научной сессии ГУАП, посвященной Всемирному дню космонавтики, СПб, ГУАП, 2011 г.

3. <http://marknelson.us/1991/02/01/arithmetic-coding-statistical-modeling-data-compression/> - Mark Nelson. Arithmetical Coding + Statistical Modeling = Data Compression