Хеш-таблицы

Сергей Соболь
ассистент кафедры ДМА ФПМИ БГУ
sobols@bsu.by

Версия от 20 декабря 2015 г.

Содержание

1 Введение

При решении многих задач возникает потребность в структуре данных, реализующей интерфейс *динамического множества* (*set*). Структура данных представляет совокупность попарно различных элементов — ключей. Интерфейс включает три основные операции:

- **Insert**(x) — добавить в множество ключ x;
- **Contains**(x) — проверить, содержится ли в множестве ключ x;
- **Remove**(x) — удалить ключ x из множества.

Естественным способом реализации такого интерфейса является дерево поиска. Если использовать сбалансированные двоичные поисковые деревья (например, AVL-деревья или красно-чёрные деревья), то для множества из N элементов время выполнения каждой из указанных операций будет величиной $O(\log N)$.

Другим способом эффективной реализации интерфейса множества является хеш-таблица. На практике множества, построенные на хеш-таблицах, обычно работают быстрее, чем множества на основе деревьев, и все основные операции выполняются за $O(1)$. Но для конкретных реализаций на конкретных входных данных ситуация может быть прямо противоположной. Трудоёмкость выполнения основных операций с хеш-таблицей может варьироваться от $O(1)$ до $O(N)$.

Мы поговорим про хеш-таблицы и особенно про теоретические аспекты, под ними лежащие. Умения строить хеш-таблицы на практике часто недостаточно для того, чтобы понимать, какие гарантии на время выполнения операций даёт такая структура данных. Оказывается, что хеш-таблицы могут быть быстрее деревьев в некотором доказуемом смысле при выполнении ряда условий.

Отметим, что, кроме множеств, большую практическую роль в программировании играют *ассоциативные массивы*, или *словари*, или *отображения* (*map*). Такие структуры данных хранят пары вида «ключ, значение» и поддерживают операции добавления пары, а также поиска и удаления пары по ключу. Предполагается, что ассоциативный массив не может хранить две пары с одинаковыми ключами. Реализация ассоциативного массива технически немного сложнее, чем множества, но использует те же идеи. Аналогично, ассоциативные массивы обычно строятся либо при помощи поисковых деревьев, либо при помощи хеш-таблиц.

2 Устройство хеш-таблицы

Для простоты будем считать, что ключи являются целыми числами из диапазона от 0 до $n - 1$. Обозначим через K множество возможных ключей:

$$K = \{0, 1, 2, \ldots, n - 1\}.$$

На практике это множество обычно довольно большое. Часто в качестве ключей на практике применяются 32-битные или 64-битные целые числа, т. е. $n = 2^{32} \approx 4{,}2 \cdot 10^9$ или $n = 2^{64} \approx 1{,}8 \cdot 10^{19}$.

2.1 Прямая адресация

Если у нас достаточно памяти для массива, число элементов которого равно числу всех возможных ключей, то для каждого возможного ключа можно отвести ячейку в этом массиве и тем самым иметь возможность добраться до любой записи за время $O(1)$.

Таким образом, для хранения множества используется булев массив T размера n, называемый *таблицей с прямой адресацией*. Элемент $T[i]$ массива содержит истинное значение, если ключ i входит в множество, и ложное значение, если ключ i в множестве отсутствует. Нетрудно видеть, что все три указанные операции легко выполняются за константное время.

Прямая адресация обладает очевидным недостатком: если множество K всевозможных ключей велико, то хранить в памяти массив T размера n непрактично, а то и невозможно. Кроме того, если число реально присутствующих в таблице записей мало по сравнению с n, то много памяти тратится зря. Размер таблицы с прямой адресацией не зависит от того, сколько элементов реально содержится в множестве.

Минимальным адресуемым набором данных в современных компьютерах является один байт, состоящий из восьми битов. Не представляет трудности реализовать таблицу с прямой адресацией так, чтобы каждый бит был использован для хранения одной ячейки. Если n — мощность множества возможных ключей, то для прямой адресации требуется выделить последовательный блок из как минимум n бит памяти. Так, для размеров множества K в 10^9 элементов таблица займёт около 120 МБ памяти. Во многих случаях такой расход памяти неприемлем, особенно когда есть необходимость создавать несколько таблиц. Тем не менее, при сравнительно небольших n метод прямой адресации успешно используется на практике.

2.2 Хеш-функция

Хеш-таблицу можно рассматривать как обобщение обычного массива с прямой адресацией.

Мы задумываем некоторую функцию, называемую *хеш-функцией* (англ. *hash function*), которая отображает множество ключей в некоторое гораздо более узкое множество:

$$h : K \to \{0, 1, \ldots, m-1\}, \tag{1}$$

$$x \mapsto h(x).$$

Величина $h(x)$ называется *хеш-значением* (*hash value*) ключа x.

Далее вместо того, чтобы работать с ключами, мы работаем с хеш-значениями. При этом возникают так называемые *коллизии* (*collisions*). Это ситуации, когда разные ключи получают одинаковые хеш-значения:

$$x \neq y, \quad h(x) = h(y).$$

Хотелось бы выбрать хеш-функцию так, чтобы коллизии были невозможны. Но в общем случае при $m < n$ это неосуществимо: согласно принципу Дирихле, нельзя построить инъективное отображение из большего множества в меньшее.

Пример 1. Приведём примеры некоторых функций. Так, константа $h(x) \equiv 0$ — хеш-функция, для которой любая пара различных ключей будет давать коллизию. Безусловно, такая хеш-функция бесполезна несмотря на то, что она простая и быстро вычисляется. Функция $h(x) = \mathrm{rand}(m)$, всякий раз возвращающая случайное число от 0 до $m-1$ включительно, выбранное равновероятно независимо от x, не может быть использована как хеш-функция, потому что хеш-функция обязана для равных ключей возвращать одинаковые значения. Функция $h(x) = x \bmod m$, возвращающая остаток от деления ключа x на m, является вполне годной для практики хеш-функцией и часто применяется.

Коллизии практически неизбежны, когда требуется осуществлять хеширование случайных подмножеств большого множества K допустимых ключей. Попробуем оценить

вероятность коллизии с точки зрения комбинаторики. Рассуждения аналогичны тем, что используются для объяснения парадокса дней рождения в теории вероятностей. Предположим, что хеш-значения ключей независимы и распределены идеально равномерно от 0 до $m-1$. Пусть мы осуществляем хеширование для N различных ключей ($N \leq m$). Когда мы назначаем всем ключам их хеш-значения, мы по сути строим некоторый вектор длины N, каждый элемент которого принимает одно из m значений. Всего существует m^N таких векторов (размещения с повторениями из m по N). Число векторов, в которых все элементы различны, равно $\frac{m!}{(m-N)!}$ (размещения без повторений из m по N). Разделив вторую величину на первую, мы получим вероятность того, что все элементы вектора различны. Значит, вероятность того, что в векторе найдутся хотя бы два одинаковых элемента, то есть случится коллизия, равна

$$p(m, N) = 1 - \frac{m!}{(m-N)! \times m^N}.$$

Такое выражение неудобно для вычислений, но при помощи формулы Стирлинга для факториала можно вывести приближённые формулы, дающие неплохую точность, когда m велико и N много меньше m, например

$$p(m, N) \approx 1 - e^{\frac{-N^2}{2m}}.$$

Пример 2. Получается, если число m возможных значений хеш-функции равно одному миллиону, мы осуществляем хеширование для $N = 2450$ уникальных ключей, то с вероятностью 95% найдутся такие два ключа, что их хеш-значения будут одинаковыми, то есть будет иметь место коллизия. Интуитивно кажется, что коллизии очень маловероятны, но это не так.

Если бы коллизий не было, хеш-таблицу можно было бы организовать как просто массив, в котором в качестве индексов использовались бы хеш-значения. Но из-за того, что возникают коллизии, структура хеш-таблицы более сложная. Разработано несколько стратегий разрешения коллизий.

2.3 Разрешение коллизий методом цепочек

Образно говоря, хеш-функция раскладывает исходные ключи по *корзинам* (*bins*, *buckets*) или *слотам* (*slots*). Ключ попадает в корзину с номером, равным хеш-значению.

Для хранения элементов с одинаковыми хеш-значениями внутри одной корзины можно использовать связные списки. На верхнем уровне организуется массив размера m — по числу различных значений хеш-функции, каждый элемент которого — это односвязный список, состоящий из ключей, имеющих конкретное хеш-значение. Возникают цепочки ключей, из-за чего метод и получил название *метода цепочек* (англ. *separate chaining*).

Способ разрешения коллизий с помощью цепочек является одним из довольно часто применяемых на практике, хотя он и не единственный.

Обсудим скорость работы такой хеш-таблицы. Посмотрим вначале на выполнение операций вставки. Вначале вычисляется хеш-значение $h(x)$ для ключа x, затем происходит обращение к соответствующему связному списку. Если нет задачи проверять, присутствует элемент x в таблице или нет, то операция вставки может быть реализована за константное время: всегда можно добавить элемент в начало списка (это стоит константу), и не нужно идти по всему связному списку. Однако обычно имеет смысл перед вставкой проверить, есть элемент x в таблице или нет, и добавлять только уникальные элементы. Это удобно в силу ряда причин: можно легко отвечать на запросы о числе элементов в множестве, меньше расход памяти (нередко на практике вставок выполняется много, но среди ключей

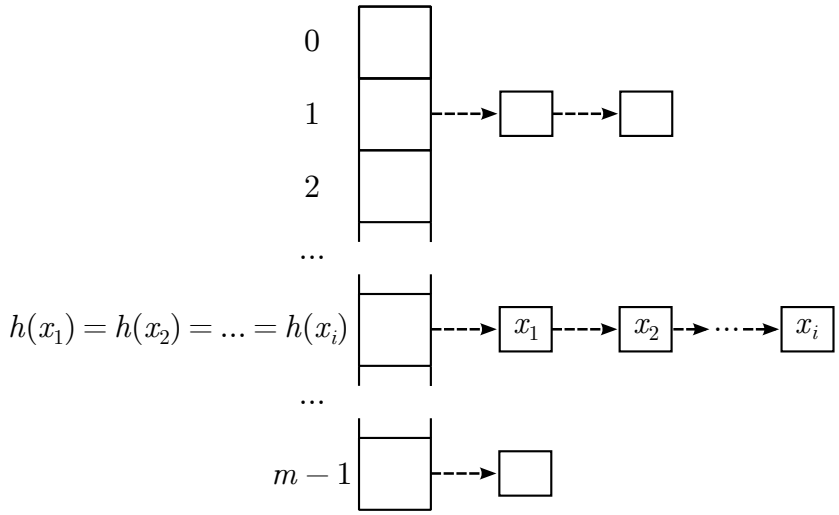

Рис. 1: Разрешение коллизий методом цепочек

мало различных), проще организовать удаление ключа. Поэтому операция вставки вначале выполняет проход по списку, и на это расходуется время, пропорциональное длине соответствующей цепочки.

Удаление ключа x также требует от нас выполнить прохождение списка в поиске элемента x. Отметим следующий факт: в общем случае из односвязного списка удалить элемент из середины сложно. Однако в рассматриваемом случае, несмотря на то, что список односвязный, удалять из него нетрудно, потому что мы движемся слева направо и можем поддерживать указатель на текущий элемент и на предыдущий. При удалении указатель у предыдущего элемента перенаправляется на следующий элемент, а память из-под текущего элемента освобождается.

Про поиск элемента x уже по сути всё сказано.

Получается, что производительность всей конструкции связана с таким параметром, как длина цепочки. Для дальнейшего анализа введём обозначения для длин цепочек. Пусть в каждый момент времени у нас есть набор чисел l_0, ..., l_{m-1} — длины цепочек, для каждого хеш-значения длина своя. Мы будем использовать эти обозначения далее в доказательствах.

Каждая из трёх рассмотренных операций с ключом x требует времени $O(1 + l_i)$, где l_i — длина цепочки, в которую попадает ключ x. Отметим важность слагаемого 1 в асимптотике: даже если цепочка имеет нулевую длину, требуется время на то, чтобы вычислить хеш-значение (мы полагаем, что хеш-функция от ключа вычисляется за константу) и обратиться к соответствующей цепочке.

Пример 3. Используя хеш-функцию $h(x) = x \bmod 10$, сформируем хеш-таблицу размера 10 из последовательности ключей 0, 20, 15, 13, 26, 7, 17, 10. Результат показан на рисунке 2.

2.4 Разрешение коллизий методом открытой адресации

Альтернативный подход называется *открытой адресацией* (англ. *open addressing*). В линейном массиве хранятся непосредственно ключи, а не заголовки связных списков. Когда выполняется операция вставки, ячейки массива проверяются, начиная с того места, в которое указывает хеш-функция, в соответствии с некоторой *последовательностью проб* (англ. *probe sequence*), пока не будет найдено свободное место. Для осуществления поиска ключа массив проходится в той же последовательности, пока либо не будет найден иско-

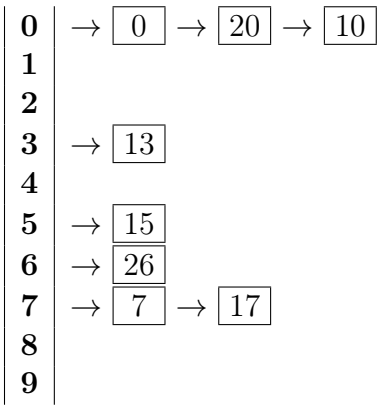

Рис. 2: Пример хеш-таблицы с цепочками

мый ключ, либо не будет обнаружена пустая ячейка (в таком случае утверждается, что искомого ключа нет). О поддержке операции удаления мы поговорим в разделе 2.4.2.

Название «открытая адресация» связано с тем фактом, что положение (адрес) элемента не определяется полностью его хеш-значением. Такой способ также называют «*закрытым хешированием*» (англ. *closed hashing*). Не следует это понятие путать с «открытым хешированием» или «закрытой адресацией»: так иногда называют хеширование с цепочками.

2.4.1 Последовательность проб

Обозначим через $h(x, i)$ номер ячейки в массиве, к которой следует обращаться на i-й попытке при выполнении операций с ключом x. Чтобы формулы были проще, удобно нумеровать попытки с нуля. Последовательность проб для ключа x получается такой:

$$h(x, 0),\ h(x, 1),\ h(x, 2), \ldots$$

Для успешной работы алгоритмов поиска последовательность проб должна быть такой, чтобы в результате m проб все ячейки хеш-таблицы оказались просмотренными ровно по одному разу в каком-то порядке:

$$\forall x \in K \quad \{h(x, i) \mid i = 0, 1, \ldots, m-1\} = \{0, 1, \ldots, m-1\}.$$

Широко используются три вида последовательностей проб: линейная, квадратичная последовательность проб и двойное хеширование.

Линейное пробирование. В этом случае ячейки хеш-таблицы последовательно просматриваются с некоторым фиксированным интервалом c между ячейками:

$$h(x, i) = (h'(x) + c \cdot i) \bmod m, \tag{2}$$

где $h'(x)$ — некоторая хеш-функция.

Для того чтобы все ячейки оказались просмотренными по одному разу, необходимо, чтобы число c было взаимно простым с размером хеш-таблицы m.

Пример 4. Пусть $m = 10$. Формула $h(x, i) = (x + 2i) \bmod 10$ не подходит а качестве последовательности проб. Например, при $x = 5$, подставляя i от 0 до 9, будем получать индексы

$$5, 7, 9, 1, 3, 5, 7, 9, 1, 3,$$

в результате чётные позиции оказываются не посещены. В то же время функция $h(x, i) = (x + 3i) \bmod 10$ работает правильно и возвращает каждый индекс один раз:

$$5, 8, 1, 4, 7, 0, 3, 6, 9, 2.$$

В простейшем случае можно взять единицу в качестве константы c в (2). В таком случае ячейки просматриваются подряд слева направо, за последней ячейкой просматривается первая. Недостаток такой последовательности проб проявляется в том, что на реальных данных часто образуются кластеры из занятых ячеек (длинные последовательности ячеек, идущих подряд). При непрерывном расположении заполненных ячеек увеличивается время добавления нового элемента и других операций. Таким образом, линейная последовательность проб довольно далека от равномерного хеширования.

Квадратичное пробирование. Интервал между ячейками с каждым шагом увеличивается на константу:

$$h(x, i) = (h'(x) + c_1 \cdot i + c_2 \cdot i^2) \bmod m,$$

где числа c_1 и c_2 фиксированы.

Если мы хотим, чтобы при просмотре хеш-таблицы использовались все ячейки, значения c_1 и c_2 нельзя выбирать случайным образом.

На практике такой метод часто работает лучше линейного, разбрасывая ключи более равномерно по массиву. Тенденции к образованию кластеров нет, но аналогичный эффект проявляется в форме образования вторичных кластеров.

Двойное хеширование. Интервал между проверяемыми ячейками фиксирован, как при линейном пробировании, но, в отличие от него, размер интервала вычисляется второй, вспомогательной хеш-функцией, а значит, может быть различным для разных ключей:

$$h(x, i) = (h'(x) + h''(x) \cdot i) \bmod m.$$

Последовательность проб в этой ситуации при работе с ключом x представляет собой арифметическую прогрессию (по модулю m) с первым членом $h'(x)$ и шагом $h''(x)$.

Чтобы последовательность проб покрыла всю таблицу, значение $h''(x)$ должно быть ненулевым и взаимно простым с m.

Для этого можно поступить следующим образом:

- выбрать в качестве m степень двойки, а функцию $h''(x)$ взять такую, чтобы она принимала только нечётные значения;
- выбрать в качестве m простое число и потребовать, чтобы вспомогательная хеш-функция $h''(x)$ принимала значения от 1 до $m - 1$.

2.4.2 Поддержка операции удаления

Удаление элементов в схеме с открытой адресацией несколько затруднено, и все операции немного усложняются. Рассмотрим один из способов. Для ячейки вводится три состояния:

- `EMPTY` — ячейка пуста;
- `KEY` x — ячейка содержит ключ x;
- `DELETED` — ячейка ранее содержала ключ, но он был удалён.

В любой момент времени каждая из m ячеек массива находится в одном из трёх состояний.

Для удобства обычно поддерживают отдельную целочисленную величину — счётчик ячеек, заполненных актуальными ключами (ячеек в состоянии `KEY`). Например, это позволяет быстро отвечать на вопрос об общем числе хранящихся элементов. Иначе для этого всякий раз нужно было бы сканировать весь массив за время $O(m)$.

Итак, изначально все ячейки пусты (состояние `EMPTY`), счётчик выставлен в нуль.

При поиске ключа x необходимо просмотреть все возможные местоположения этого ключа: проверяются ячейки $h(x, 0)$, $h(x, 1)$, ..., пока не найдётся либо ячейка `KEY` x (говорим, что ключ найден), либо ячейка `EMPTY` (говорим, что искомого ключа в таблице нет). Если встречается ячейка `DELETED`, процесс пробирования её игнорирует и идёт дальше.

Перед вставкой ключа x сначала выполняется поиск этого ключа. Если ключ уже есть, вставка не требуется. Иначе проверяется принципиальное наличие ячеек, не занятых ключами (используем счётчик занятых ячеек). Если все m ячеек имеют состояние `KEY`, вставка невозможна: таблица переполнена. Иначе вновь пробуются позиции $h(x, 0)$, $h(x, 1)$, ..., пока не будет найдена либо свободная ячейка `EMPTY`, либо ячейка с удалённым ключом `DELETED` (удалённые ячейки при вставке приравниваются к свободным, а при поиске нет). Ячейка переводится в состояние `KEY` x, счётчик занятых ячеек увеличиваем на единицу.

При удалении ключа x вначале выполняем поиск ключа x. Если такая ячейка найдена, переводим её в состояние `DELETED` и счётчик занятых ячеек уменьшаем на единицу.

Нетрудно видеть, что наличие большого числа `DELETED`-ячеек отрицательно сказывается на времени выполнения операции поиска, а значит и других операций. Чтобы исправить ситуацию, после ряда удалений можно перестраивать хеш-таблицу заново, уничтожая удалённые ячейки.

2.4.3 Плюсы и минусы открытой адресации

Недостаток систем с открытой адресацией состоит в том, что число хранимых ключей не может превышать размер хеш-массива. По факту, даже при использовании хороших хеш-функций, производительность резко падает, когда хеш-таблица оказывается заполненной на 70% и более. Во многих приложениях это приводит к обязательному использованию динамического расширения таблицы, о котором мы поговорим в разделе 2.6.

Схема с открытой адресацией предъявляет более строгие требования к хеш-функции, чтобы механизм работал хорошо. Кроме того, чтобы распределять значения максимально равномерно по корзинам, функция должна минимизировать кластеризацию хеш-значений, которые стоят рядом в последовательности проб. Так, в методе цепочек одна забота — чтобы много ключей не получило одно и то же хеш-значение, и если хеш-значения получились разные, совершенно всё равно, рядом ли они стоят, отличаются ли на единицу и т. д. Если при использовании метода открытой адресации образовались большие кластеры, время выполнения всех операций может стать неприемлемым даже при том, что заполненность таблицы в среднем невысокая и коллизии редки.

Открытая адресация позволяет существенно экономить память, если размер ключа невелик по сравнению с размером указателя. В методе цепочек приходится хранить в массиве указатели на начала списков, а каждый элемент списка хранит, кроме ключа, указатель на следующий элемент, поэтому на все эти указатели расходуется память.

Открытая адресация не требует затрат времени на выделение памяти на каждую новую запись и может быть реализована даже на миниатюрных встраиваемых системах, где полноценный аллокатор недоступен. Также в открытой адресации нет лишней операции обращения по указателю (*indirection*) при доступе к элементу. Открытая адресация обеспечивает лучшую локальность хранения, особенно с линейной функцией проб. Когда размеры ключей небольшие, это даёт лучшую производительность за счёт хорошей рабо-

ты кеша процессора, который ускоряет обращения к оперативной памяти. Однако, когда ключи «тяжёлые» (не целые числа, а составные объекты), они забивают все кеш-линии процессора, к тому же много места в кеше тратится на хранение незанятых ячеек. Как вариант, можно в массиве с открытой адресацией хранить не сами ключи, а указатели на них. Очевидно, часть преимуществ при этом будет утрачена.

Так или иначе, любой подход к реализации хеш-таблицы может работать достаточно быстро на реальных нагрузках. Время, которое занимают операции с хеш-таблицами, обычно составляет малую долю от общего времени работы программы. Расход памяти редко играет решающую роль. Часто выбор между той или иной реализацией хеш-таблицы делается на основании других факторов в зависимости от ситуации.

2.5 Коэффициент заполнения

Критически важным показателем для хеш-таблицы является *коэффициент заполнения* (англ. *load factor*) — отношение числа ключей, которые хранятся в хеш-таблице, к размеру хеш-таблицы:

$$\alpha = \frac{N}{m}.$$

Коэффициент заполнения может быть как меньше, так и больше единицы (не для всех способов разрешения коллизий такое возможно, например, это недопустимо для разрешения коллизий методом открытой адресации).

Пусть для разрешения коллизий используется метод цепочек. На первый взгляд очевидно, что чем больше коэффициент заполнения, тем медленнее работает хеш-таблица. Однако коэффициент заполнения не показывает различия между заполненностью отдельных корзин. Например, пусть есть две хеш-таблицы, в каждой используется 1000 корзин и хранится всего 1000 ключей, поэтому коэффициент заполнения в обоих случаях равен единице. Однако в первой таблице в каждой цепочке по одному ключу, а во второй все ключи лежат в одной длинной цепочке. Очевидно, что вторая хеш-таблица будет работать очень медленно.

Низкий коэффициент заполнения не является абсолютным благом. Если коэффициент близок к нулю, это говорит о том, что большая часть таблицы не используется и память тратится впустую.

2.6 Динамическое изменение размера

Для оптимального использования хеш-таблицы желательно, чтобы её размер был примерно пропорционален числу ключей, которые нужно хранить. На практике редко случается, что число ключей фиксировано и можно заранее выставить хорошее значение параметру m. Если ставить его заведомо большим, то много памяти будет потрачено зря (особенно если нужно организовать много хеш-таблиц с небольшим числом ключей в каждой).

Реализация хеш-таблицы общего назначения обязана поддерживать операцию изменения размера.

Часто используемым приёмом является автоматическое изменение размера, когда коэффициент заполнения превышает некоторый порог $\alpha_{\max}$. Выделяется память под новую, бóльшую таблицу, все элементы из старой таблицы перемещаются в новую, затем память из-под старой хеш-таблицы освобождается. Аналогично, если коэффициент заполненности опускается ниже другого порога $\alpha_{\min}$, элементы перемещаются в хеш-таблицу меньшего размера.

Кстати, чисто технически в языках программирования под хеш-функцией понимается функция, отображающая ключ в произвольное целое число без ограничения на величину (не требуется попадания в промежуток $[0, m)$, как в определении (1)). В этом есть здравый смысл. Хеш-значение — это свойство ключей, и оно не зависит от размера хеш-таблицы. Тот факт, что мы далее берём его по модулю m, — это исключительно особенность его использования внутри конкретной хеш-таблицы конкретного размера. Не нужно предварительно брать остаток. Часто на практике хеш-кодами выступают всевозможные беззнаковые целые числа, и остаток берётся о нужному модулю непосредственно в месте использования.

3 Теория вопроса

Поговорим об оценках на время выполнения операций над хеш-таблицей в случае разрешения коллизий методом цепочек. Метод открытой адресации кажется существенно более сложным для теоретического анализа.

Нетрудно понять, что от выбора хеш-функции многое зависит. Если она будет «плохая» (например константная), то будут иметь место сплошные коллизии (хеш-таблица превратится в связный список). Как определить меру качества хеш-функции и научиться получать «хорошие» хеши?

За годы развития вычислительной техники сформировалась наука о том, как практически строить хеш-функции. Разработаны всевозможные правила на тему того, какие биты ключа x стоит взять, на сколько позиций выполнить поразрядный сдвиг, как применить операцию исключающего или, на что умножить, чтобы получить хеш, который бы выглядел как случайный. Чтобы найти примеры практических хеш-функций, можно взять какую-нибудь стандартную библиотеку языка программирования (мы займёмся этим в разделе 4). Что эти функции гарантируют? Часто ничего. То есть коллизии, конечно, есть. Отметим, что коллизии неистребимы. Любую фиксированную хеш-функцию можно всегда сломать — придумать ситуацию, когда там будут одни сплошные коллизии.

Как же тогда оценивать трудоёмкость? Можно сказать, что всё работает быстро, когда длины цепочек хорошие, но это слабое объяснение, потому что мы знаем, что длины цепочек бывают плохими. Хотелось бы мыслить в терминах худшего случая, а он тут совершенно ужасен. Что делать?

Теоретически есть два направления возможной деятельности. Первое направление заключается в том, чтобы добавить рандомизацию: внести элемент случайности, чтобы не было фиксированного плохого случая. А именно, есть идея хеш-функцию брать не одну какую-то конкретную, которую можно сломать, а, например, организовать программу так, чтобы при каждом запуске хеш-функция была своя. Тогда одним и тем же входом сломать не получится, и можно будет рассуждать о длине цепочки с точки зрения теории вероятностей: говорить о том, какая длина цепочки в среднем, какая у неё дисперсия...

Второе направление состоит в следующем. Пусть, хоть потенциальное множество ключей и велико, в каждый момент времени мы храним ключей меньше, чем число корзин в хеш-таблице. Пусть S — фактическое множество ключей в хеш-таблице, и у этого множества размер N, где $N \leq m$. Тогда понятно, что, вообще говоря, существует хеш-функция, которая не даёт коллизий: если элементов меньше, чем слотов, то их всегда можно отобразить без коллизий. Это хорошая хеш-функция, её называют совершенной. В чём проблема? Она всегда есть, она зависит от набора ключей. Даже если предположить, что набор ключей константный и известен заранее, есть трудность в том, как эффективно задать эту хеш-функцию. Вообще говоря, информации в ней содержится довольно много. Если нам потребуется много памяти, чтобы хеш-функцию задавать, это не очень хорошо. Потом, ещё вопрос, как с этой информацией работать. Возникает проблема примерно того же

сорта, как та, с которой мы пытаемся справиться. Если мы говорим, что функция есть и мы её построим, то как вычислять значения в отдельных точках? В практическом смысле значение функции в точке — небольшое выражение, считающееся за константное время.

Как задавать функцию, которая потенциально существует? Таблицей? Это непрактично, потому что множество K образов для этой функции слишком большое. Если использовать какую-либо ассоциативную структуру, это ровно та же задача, которую мы пытаемся решать (если для вычисления хеш-функции делать запрос в хеш-таблицу, то какую хеш-функцию использовать в этой новой хеш-таблице?). Другое решение: например, можно все значения S отсортировать по возрастанию и в качестве хеш-функции элемента использовать его порядковый номер. Это правильно с точки зрения избавления от коллизий. Но как тогда определять номер элемента во время работы программы? Мы не можем использовать бинарный поиск, потому что он требует логарифмического времени и сведёт на нет смысл построения хеш-таблицы.

Итак, если мы знаем, что $|S| = N$, как нам построить хеш-функцию, которая

- простая, т. е. достаточно константы памяти, чтобы её запомнить,
- быстрая, т. е. требует константного времени на вычисление,
- совершенная, т. е. коллизии отсутствуют.

Оказывается, сделать это можно, и мы такой способ рассмотрим. Здесь также будет использоваться рандомизация, но на промежуточных этапах. В конце функция будет совершенно конкретная, но зависящая от входных данных. Искать мы будем её с помощью рандомизации.

3.1 Идеальное хеширование

Вначале поговорим про рандомизацию. Первое, самое простое, что приходит в голову, — взять в качестве h случайную функцию. Это так называемое *идеальное хеширование* в том смысле, что оно нереализуемо на практике и существует лишь в теории.

Что значит «h — случайная функция»? Функцию можно представить как таблицу, в которой для каждого значения аргумента написано, чему функция равна. Под случайной функцией понимают функцию, у которой каждое значение выбирается случайно равновероятно на множестве $\{0, 1, \ldots, m-1\}$ и все значения выбираются независимо. Мы пытаемся заполнить таблицу числами от 0 до $m-1$, причём числа могут повторяться, количество элементов в этой таблице равно n. Каждая клетка выбирается независимо в соответствии с равномерным распределением.

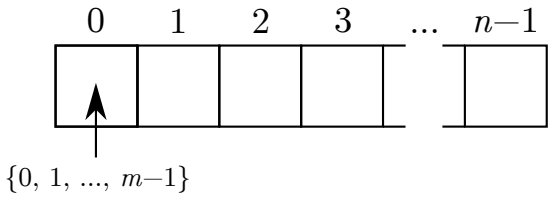

Рис. 3: Идеальное хеширование

3.1.1 Оценка длины цепочки

Какие хорошие свойства даёт идеальная хеш-функция? Например, можно получить хорошую оценку для длины цепочки. А именно, если мы возьмём какой-либо индекс t

среди хеш-значений и посмотрим на длину цепочки l_t, то это, конечно, будет случайная величина. Чему же равно её матожидание $\mathbf{E}\{l_t\}$? Происходит следующее.

Пусть $S = \{x_1, \ldots, x_N\}$. Тогда

$$\mathbf{E}\{l_t\} = \mathbf{E}\left\{\sum_{i=1}^{N} \mathbb{1}_{\{h(x_i)=t\}}\right\},$$

где $\mathbb{1}_{\{h(x_i)=t\}}$ — индикатор того события, что i-й ключ попал в список, соответствующий хеш-значению t, т. е.

$$\mathbb{1}_{\{h(x_i)=t\}} = \begin{cases} 1, & \text{если } h(x_i) = t, \\ 0, & \text{иначе.} \end{cases}$$

Нетрудно видеть, что

$$\mathbf{E}\left\{\mathbb{1}_{\{h(x_i)=t\}}\right\} = 1 \cdot \mathbf{P}\{h(x_i) = t\} + 0 \cdot \mathbf{P}\{h(x_i) \neq t\} = \frac{1}{m}$$

в силу того, что $h(x_i)$ равновероятно принимает некоторое значение из m возможных, одно из которых t, и не зависит от i.

Из линейности математического ожидания следует, что

$$\mathbf{E}\{l_t\} = \sum_{i=1}^{N} \mathbf{E}\left\{\mathbb{1}_{\{h(x_i)=t\}}\right\} = \frac{N}{m} = \alpha,$$

где α — введённый ранее в разделе 2.5 коэффициент заполнения хеш-таблицы, равный отношению количества хранящихся в хеш-таблице элементов к количеству цепочек.

Аналогично можно вычислить дисперсию длины цепочки. Действительно, каждый индикатор является дискретной случайной величиной, имеющей распределение Бернулли с вероятностью успеха $1/m$, поэтому

$$\mathbf{V}\left\{\mathbb{1}_{\{h(x_i)=t\}}\right\} = \frac{1}{m}\left(1 - \frac{1}{m}\right).$$

Поскольку случайные величины независимы, дисперсия суммы равна сумме дисперсий:

$$\mathbf{V}\{l_t\} = \sum_{i=1}^{N} \cdot \mathbf{V}\left\{\mathbb{1}_{\{h(x_i)=t\}}\right\} = \frac{N}{m}\left(1 - \frac{1}{m}\right) < \frac{N}{m} = \alpha.$$

Получается, что длина цепочки имеет константное матожидание и константную дисперсию. В среднем всё ложится равномерно и длина цепочки хорошая.

3.1.2 Сложность идеальной хеш-функции

Однако вернёмся к тому, что в жизни так не бывает и никакой идеальной хеш-функции нет, потому что она слишком сложна. Например, она сложна с точки зрения теории информации, потому что количество информации в этой таблице велико, где-то эту информацию надо хранить и быстро вынимать. Представлять, что хеш-функция — это таблица из n ячеек, непрактично, потому что n обычно очень велико. Если бы можно было создать такую таблицу, не было бы смысла использовать хеширование вместо прямой адресации, когда по ключу прямо вынимается значение.

Построить такую таблицу мы не можем. А что же сделать можно? Таблицу можно строить не сразу, а постепенно. А именно, когда спрашивают значение $h(x)$, можно сгенерировать случайное значение и его выдать. $h(x)$ случайно, но для одного и того же x

хеш-функция должна возвращать одни и те же значения. Поэтому необходимо хранить для уже встретившихся x ранее выданные значения, а это задача, аналогичная той, которую мы решаем (требуется какая-либо ассоциативная таблица).

Идеальных хешей не бывает. Но они обладают небольшим списком свойств, которых, во-первых, хватает для того чтобы производить разные интересные оценки, и, во-вторых, есть способы построения h, которые менее затратны, но для них эти свойства всё ещё тоже выполнены.

Нужно отказаться от сильного свойства полной покоординатной случайности и независимости и заменить на что-то более простое, но так, чтобы, желательно, сохранить оценку матожидания (дисперсию сохранить, увы, не удастся). Кстати, заметим, что для получения оценки матожидания мы мало чем пользовались. Например, полученная оценка верна, даже если хеш-функция на всех аргументах принимает одно и то же случайно выбранное значение. Вообще, в анализе алгоритмов хорошее матожидание мало что значит, важнее оценивать матожидание в совокупности с дисперсией, а ещё лучше оценивать худший случай с высокой вероятностью, т. е. оценку, которая верна «почти всегда», например с точностью до событий, вероятность которых, например, меньше любого полинома. Конечно, матожидание здесь слишком оптимистично.

Что это за свойства, которых часто достаточно и их можно выполнить, не прибегая к большому объёму случайности?

3.2 Универсальное хеширование

Универсальное хеширование (англ. *universal hashing*) — это хеширование, при котором хеш-функция выбирается случайным образом из определённого семейства хеш-функций

$$H = \{h : K \to \{0, 1, \ldots, m-1\}\}.$$

Семейство хеш-функций H называется *универсальным* (англ. *universal*), если для любых двух различных значений ключа x и y вероятность получить коллизию ведёт себя ожидаемым образом, т. е.

$$\mathbf{P}_{h \in H} \{h(x) = h(y)\} = \mathrm{O}\left(\frac{1}{m}\right). \tag{3}$$

Когда константа, скрытая в асимптотике, равна единице, семейство называют *сильно универсальным*. Часто для теоретических рассуждений величина константы особой роли не играет.

Можно пытаться универсальность усилить и ввести понятие k-независимости. Семейство функций H называют *k-независимым*, если для любых различных ключей x_1, x_2, …, x_k и для любых возможных значений хеш-функций t_1, t_2, …, t_k имеет место условие

$$\mathbf{P}_{h \in H} \{h(x_1) = t_1,\ h(x_2) = t_2,\ \ldots,\ h(x_k) = t_k\} = \mathrm{O}\left(\frac{1}{m^k}\right). \tag{4}$$

Формально у нас есть некоторое семейство хеш-функций H с некоторым распределением вероятностей на нём, и $h \in H$ выбирается в соответствии с этим распределением. Важно понимать, что в формулах (3) и (4) вероятности берутся не по значениям ключей (все x_i фиксируются), а по h. Поэтому на множестве всех h есть некоторое распределение (обычно равномерное, но не обязательно). Итак, есть семейство H, и из этого семейства мы h выбираем.

В эту схему, когда хеш-функция выбирается из семейства, отлично вкладывается концепция идеальной хеш-функции. В этом случае H — множество всех функций с равномерным распределением. Свойство универсальности, конечно, всегда справедливо для идеальной хеш-функции. Более того, оно справедливо с жёсткой константой, равной единице.

Поговорим про k-независимость. Формально отметим следующий факт.

Теорема 1. *Из k-независимости вытекает $(k-1)$-независимость.*

Доказательство. Для доказательства можно зафиксировать $x_1, x_2, \ldots, x_{k-1}$ и просуммировать по всем возможным значениям x_k. Там будут непересекающиеся события. Вероятности сложатся, и степень изменится нужным образом. □

Что есть 1-независимость? Из определения следует, что каждое из m значений такая хеш-функция принимает равновероятно, но зависимость между точками возможна. Например, 1-независимостью обладает функция-константа $h(x) = c$, где c выбирается случайно. Это плохой пример хеш-функции, поэтому делаем вывод, что сама по себе 1-независимость хороших свойств не несёт.

Как связать универсальность и k-независимость? Оказывается, k-независимость сильнее, чем универсальность, начиная с некоторого k, причём это k равно двум.

Теорема 2. *2-независимость влечёт за собой универсальность.*

Доказательство. Действительно, 2-независимость означает, что мы понимаем про любую пару заказанных значений, с какой вероятностью она встречается. А в определении универсальности требуется как раз попарное свойство, гласящее, что хеш-значения совпадут.

Как выразить вероятность события

$$\{h(x) = h(y)\}?$$

Нужно разложить событие в комплект условий вида

$$\{h(x) = 0,\ h(y) = 0\} \cup \{h(x) = 1,\ h(y) = 1\} \cup \ldots \cup \{h(x) = m-1,\ h(y) = m-1\}.$$

Получается, что событие распадается на m непересекающихся событий, каждое из которых, согласно определению (4) при $k = 2$, имеет вероятность $\mathrm{O}\left(\frac{1}{m^2}\right)$. После сложения этих m слагаемых мы получаем вероятность $\mathrm{O}(\frac{1}{m})$, что и есть формула (3). □

Интуиция здесь должна быть примерно следующая. k-независимая хеш-функция с точностью до константы, скрытой в «O», неотличима от идеальной, если рассматривать только k элементов. Если в вычислениях начнёт возникать $k+1$ элемент, то уже непонятно, что за свойства у будут хеш-функции. В определении универсальности фигурирует свойство, которое сформулировано в терминах пар элементов. Для идеальных хеш-функций оно верно, поэтому для 2-независимых тоже.

3.2.1 Теоретико-информационные соображения

Оказывается, бывают хеш-функции, для которых свойство универсальности выполнено, при этом они не идеальны.

Главная причина невозможности применения идеальных хеш-функций на практике — то, что H — объект слишком большой, то есть мощность множества H велика. Действительно, чтобы работать с функцией h в алгоритме, нам нужно её зафиксировать и запомнить в памяти. Если мы хотим запомнить элемент множества, нам нужна некоторая память для этого. Чтобы запомнить элемент множества, достаточно $\log_2 |H|$ бит памяти (а в случае равномерного распределения элементов множества и необходимо, в общем же случае следует анализировать энтропию).

В случае идеального хеширования $|H| = m^n$, тогда $\log |H| = \mathrm{O}(n \log m)$. Это очень плохо, потому что полученное выражение линейно по n. Хотелось, чтобы оно было константным или хотя бы субполиномиальным по n.

Пусть размер множества H константный по n и полиномиальный по m: $|H| = \mathrm{poly}(m)$, тогда после логарифмирования получим $\mathrm{O}(\log m)$. А это уже, отметим, константное число машинных слов. Дело в том, что $\log m$ помещается в машинное слово, потому что m — индекс в массиве списков, на котором реализована хеш-таблица. Любой индекс во внутренней памяти должен быть не больше, чем двойка, возведённая в степень длины машинного слова.

3.2.2 Построение универсального семейства для целых чисел

Рассмотрим один способ построения универсального семейства хеш-функций для целочисленных ключей. Метод был предложен Дж. Л. Картером (J. L. Carter) и М. Н. Вегманом (M. N. Wegman) в 1977 г. Фактически семейство будет даже строго универсальным. Способ линейно-алгебраический по содержанию и состоит из двух шагов.

На первом шаге возьмём простое число p, большее или равное n. Например, можно взять ближайшее простое, не меньшее n. По разным свойствам распределений простых чисел мы знаем, что p будет, вообще говоря, недалеко от n. Так, постулат Бертрана гласит, что существует достаточно близкое к n простое число: $n \le p < 2n$.

Пример 5. Для $n = 1000$ можно взять $p = 1009$.

Через $\mathbb{Z}_p$ будем обозначать множество наименьших неотрицательных вычетов по модулю p:

$$\mathbb{Z}_p = \{0, 1, 2, \ldots, p-1\}.$$

На втором шаге мы независимо формируем два случайных вычета по модулю p, один из которых ненулевой:

$$a \in \mathbb{Z}_p \setminus \{0\}, \quad b \in \mathbb{Z}_p,$$

После того как эти два значения выбраны, возникает функция, задаваемая выражением

$$(ax + b) \bmod p. \tag{5}$$

Заметим, что раз $n \le p$, то ключ x, как a и b, является вычетом по модулю p, но старшая часть вычетов запрещена (от n до $p-1$ включительно).

Формула (5) для любого ключа x даёт на выходе вычет по модулю p, который в качестве хеш-значения не годится, потому что p может быть велико. Есть отличный способ сделать его меньше — взять по модулю m:

$$h_{a,b}(x) = [(ax + b) \bmod p] \bmod m. \tag{6}$$

Полученная функция $h_{a,b}(x)$ может служить хеш-функцией для построения хеш-таблиц размера m.

Пример 6. Пусть $m = 10$ и $n = 1000$. Зафиксируем простое число $p = 1009$. Тогда коэффициент a можно выбирать от 1 до 1008 (всего 1008 способов), коэффициент b можно выбирать от 0 до 1008 (всего 1009 способов). Итого можно получить $1008 \times 1009 = 1\,017\,072$ разных функции, например

$$\begin{aligned} h_{3,4}(x) &= ((3x + 4) \bmod 1009) \bmod 10, \\ h_{670,905}(x) &= ((670x + 905) \bmod 1009) \bmod 10, \\ h_{101,0}(x) &= ((101x \bmod 1009) \bmod 10, \\ &\ldots \end{aligned} \tag{7}$$

Заметим, что в общем случае нельзя «для простоты счёта» опустить первое взятие остатка, например $((1010 \bmod 1009) \bmod 10) = 1 \ne 0 = (1010 \bmod 10)$.

Конструкция (6) выглядит странно: часто один вычет по модулю другого вычета не имеет особого смысла, только если один модуль не является делителем другого. Но в данном случае h как раз оказывается хеш-функцией, выбранной равномерно из универсального семейства. Попробуем это понять.

3.2.3 Доказательство универсальности

Теорема 3. *Семейство функций*

$$H = \{h_{a,b} \mid a, b \in \mathbb{Z}_p, a \neq 0\},$$

где $h_{a,b}(x)$ определяется по формуле (6), число p простое, является строго универсальным.

Доказательство. Зафиксируем два различных ключа x и y. Посчитаем вероятность того, что a и b окажутся выбраны так, что хеш-значения у ключей совпадут, т. е.

$$h_{a,b}(x) = h_{a,b}(y),$$

или

$$[(ax + b) \bmod p] \bmod m = [(ay + b) \bmod p] \bmod m. \tag{8}$$

Заметим, что, вообще говоря, хеши могут совпасть на двух стадиях: сначала на стадии арифметики по модулю p, потом на стадии арифметики по модулю m. Понятно, что если хеши совпали на первой стадии, то они совпадут и на второй.

Покажем, что на первой стадии совпадений быть не может.

Лемма 1. *Утверждается, что если $x \neq y$, то $(ax + b) \bmod p \neq (ay + b) \bmod p$.*

Доказательство. Докажем от противного. Пусть

$$ax + b \equiv ay + b \pmod{p}.$$

Вычтем b из обеих частей сравнения и получим

$$ax \equiv ay \pmod{p}.$$

Воспользуемся тем, что число p простое. Вообще, в алгебре удобно работать с простыми числами, потому что целые числа по простому модулю образуют поле. Поскольку a — ненулевой вычет, можно обе части сравнения поделить на a. Получим

$$x \equiv y \pmod{p}$$

Противоречие с тем фактом, что $x \neq y$ и $0 \leq x, y < p$. □

То есть коллизия если и есть, но на второй стадии, когда разные вычеты по модулю p превращаются в одинаковые вычеты по модулю m. Более того, конечно, такие события будут происходить, нужно лишь оценить вероятность этого несчастья.

Для этого нам потребуется понимание того, как себя ведёт выражение $(ax + b) \bmod p$, когда мы перебираем всевозможные a и b, ведь ключ x фиксирован и мы на него не влияем. Очевидно, при этом значение меняется и пробегает всевозможные вычеты (хотя бы потому, что слагаемое b пробегает всевозможные вычеты).

Немного усложним картину. Как ведёт себя пара $((ax+b) \bmod p, (ay+b) \bmod p)$? Определим отображение f:

$$f : (a, b) \to (u, v),$$

которое паре (a, b) ставит в соответствие пары (u, v) по формулам

$$\begin{aligned} u &= (ax + b) \bmod p, \\ v &= (ay + b) \bmod p. \end{aligned} \tag{9}$$

Областью определения функции f являются целые точки квадрата $p \times p$, в котором одна ось называется a, другая ось b, за исключением прямой $a = 0$ (по условию вычет a ненулевой):

$$D_f = \{(a, b) \mid a, b \in \mathbb{Z}_p, a \neq 0\}.$$

Пусть на точку (a, b) мы действуем отображением f. Мы вновь переходим в квадрат $p \times p$, потому что работаем с вычетами, но уже в координатах u и v.

В какие области квадрата $p \times p$ с осями u, v мы можем попасть? На диагональ квадрата попасть мы не можем: в силу леммы 1 при выборе любой допустимой пары a и b значения u и v будут отличаться. Точки множества

$$E_f = \{(u, v) \mid u, v \in \mathbb{Z}_p, u \neq v\}$$

потенциально могут быть образами. Давайте попробуем доказать, что любая точка из этого множества действительно является образом.

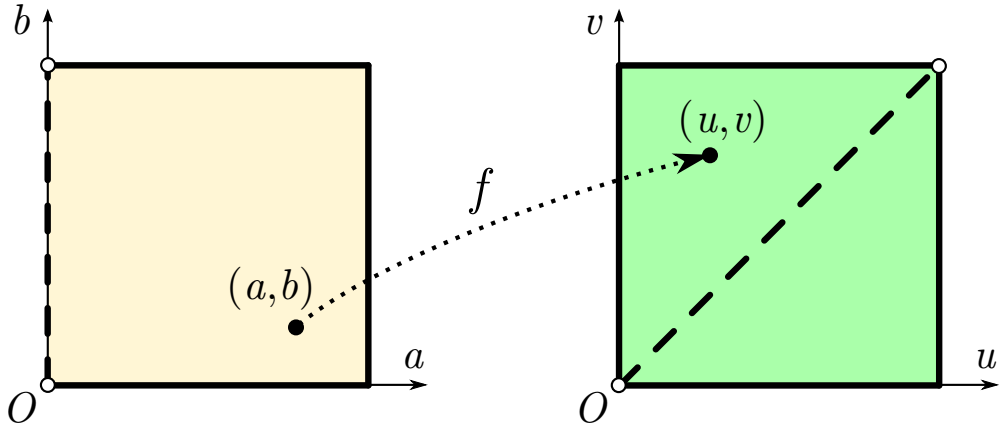

Рис. 4: Действие отображения f

Лемма 2. *Каждая точка множества E_f является образом некоторой точки из D_f при отображении f.*

Доказательство. Посмотрим на происходящее как на систему линейных уравнений над полем $\mathbb{Z}_p$. Заметим, что это система с однозначным решением. А именно, запишем формулы (9) в матричном виде, предполагая, что все операции выполняются по модулю p. Из пространства с координатами (a, b) мы переходим в пространство с координатами (u, v):

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} x & 1 \\ y & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

Матрица линейного отображения f имеет вид

$$A = \begin{pmatrix} x & 1 \\ y & 1 \end{pmatrix}.$$

Её определитель равен $x - y$, и он ненулевой, потому что $x \neq y$. Значит, можно построить обратное отображение f^{-1} пары (u, v) в пару (a, b). Исходя из правила Крамера можно сразу выписать решение системы линейных уравнений (все арифметические операции выполняются в поле $\mathbb{Z}_p$):

$$a = \frac{\begin{vmatrix} u & 1 \\ v & 1 \end{vmatrix}}{\begin{vmatrix} x & 1 \\ y & 1 \end{vmatrix}} = \frac{u - v}{x - y}, \quad b = \frac{\begin{vmatrix} x & u \\ y & v \end{vmatrix}}{\begin{vmatrix} x & 1 \\ y & 1 \end{vmatrix}} = \frac{vx - uy}{x - y}.$$

Важно, что по этим формулам, подставляя точки (u, v) из E_f, мы всегда будем получать точки (a, b) из D_f: мы не попадём на запрещённую сторону квадрата $a = 0$, поскольку диагональ другого квадрата $u = v$ запрещена.

Таким образом, отображение f сюръективно, т. е. отображает D_f *на* E_f. □

Теперь легко понять, что отображение f является на самом деле взаимно-однозначным.

Лемма 3. *Функция f устанавливает биекцию между множеством пар D_f и множеством пар E_f.*

Доказательство. Определим мощности множеств. Так, $|D_f| = p(p-1)$ (это квадрат $p \times p$ без p точек на стороне $a = 0$) и $|E_f| = p(p-1)$ (это квадрат $p \times p$ без p точек на диагонали $u = v$).

По определению функции f из любой точки D_f мы попадаем в некоторую точку E_f. В силу леммы 2 в любую точку множества E_f можно попасть из D_f. С учётом равенства $|D_f| = |E_f|$ делаем вывод, что способ попадания единственный. □

Вернёмся к хеш-функциям. Выпишем выражения (6) для фиксированных x и y:

$$h(x) = \underbrace{[(ax + b) \bmod p]}_{u} \bmod m,$$
$$h(y) = \underbrace{[(ay + b) \bmod p]}_{v} \bmod m.$$

Если мы начнём перебирать пары (a, b) из множества D_f, то пары (u, v) будут пробегать множество E_f, при этом в каждой точке побывают ровно единожды. Из этого следует, что для цели нашей задачи (понять, когда есть коллизия) можно перебирать не a и b по множеству D_f (это неудобно: сложные выражения), а u и v по множеству E_f. Важно, что мы, ничего не забудем, т. е. любую комбинацию значений хеш-функций мы посмотрим, причём посмотрим нужное число раз. Потому что любая комбинация u и v встретится ровно при одной паре a и b.

Тем самым задача упрощается. Нужно понять, какова вероятность следующего события:

$$\{u \bmod m = v \bmod m\}$$

где u, v — пара различных вычетов по модулю p.

Посчитаем эту вероятность самым непосредственным образом по определению: нужно количество случаев, когда равенство достигнуто, поделить на общее количество случаев. Всего для u и v есть $p(p-1)$ вариантов, поэтому знаменатель уже известен. Определим числитель. Подсчитаем, для скольки пар (u, v) по модулю p у них по модулю m тоже совпадают значения.

Лемма 4. *Пусть u — некоторый фиксированный вычет по модулю p. Существует не более $\frac{p-1}{m}$ вычетов v по модулю p таких, что $u \neq v$, но $u \equiv v \pmod m$.*

Доказательство. Какие числа v надо брать, чтобы получать по модулю m такое же значение, как у u? Нетрудно видеть, что все числа, сравнимые с u по модулю m, образуют арифметическую прогрессию с разностью m. Но нас интересуют лишь числа, попадающие в отрезок от 0 до $p-1$. Изобразим это на числовой прямой:

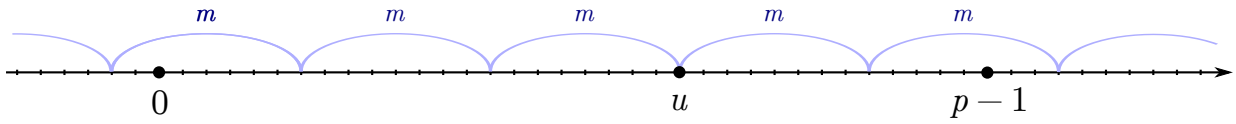

Рис. 5: Варианты выбора вычета v при фиксированном u

На отрезке от 0 до u (левее заданной точки u) подотрезок длины m помещается ровно $\left\lfloor \frac{u-0}{m} \right\rfloor$ раз (округление вниз). Аналогично, можно сделать ровно $\left\lfloor \frac{p-1-u}{m} \right\rfloor$ шагов длины m вправо от точки u, каждый раз получая новое подходящее v.

Всего имеем

$$\left\lfloor \frac{u-0}{m} \right\rfloor + \left\lfloor \frac{p-1-u}{m} \right\rfloor \le \frac{u-0}{m} + \frac{p-1-u}{m} = \frac{p-1}{m}$$

вариантов выбора различных v, что и требовалось доказать. □

Пример 7. Пусть $p = 10$ и $m = 3$. Лемма 4 даёт оценку сверху $\frac{10-1}{3} = 3$ независимо от u. Если $u = 4$, то можно найти два подходящих v: $\{1, 7\}$. Если $u = 6$, то существуют три подходящих v: $\{0, 3, 9\}$, то есть оценка достижима.

Мы вели рассуждения для фиксированного u. А всех u ровно p штук, поэтому полученное значение следует умножить на p.

Подставим и получим

$$\mathbf{P}\left\{u \bmod m = v \bmod m\right\} \le \frac{\frac{p-1}{m} \cdot p}{p(p-1)} = \frac{1}{m}.$$

Таким образом, схема хеширования универсальна, и даже сильно универсальна, потому что имеет константу 1 перед оценкой. □

Пример 8. Пусть $n = 1000$, $m = 10$. Выберем p равным 1009, как в примере 6. Ранее отмечалось, что $|H| = 1\,017\,072$.

Проверим доказанную теорему 3 на практике. Зафиксируем два различных ключа x и y, например $x = 1$ и $y = 2$. Затем промоделируем процесс выбора хеш-функции на компьютере: будем перебирать в двух вложенных циклах a от 1 до 1008 и b от 0 до 1008 включительно, каждый раз будем вычислять хеши $h(x)$ и $h(y)$, сравнивать их и подсчитывать число случаев, когда случается коллизия. Оказывается, что это число равно 100 800 и не зависит от x и y. Значит, вероятность того, что на данной паре ключей при случайном выборе хеш-функции из универсального семейства будет коллизия, равна $\frac{100\,800}{1\,017\,072} \approx 0{,}0991 < 0{,}1 = \frac{1}{m}$.

3.2.4 Универсальное хеширование векторов

Если в качестве ключей выступают не просто целые числа, а объекты других типов, то нужно использовать соответствующие универсальные семейства хеш-функций.

Например, пусть ключ представляет собой вектор фиксированной длины

$$x = (x_1, x_2, \ldots, x_r),$$

составленный из целых чисел, каждое из которых лежит в отрезке от 0 до $n-1$, при этом n небольшое. Размер хеш-таблицы m выберем простым и таким, чтобы выполнялось условие $m \geq n$.

Хеш-функцию будем строить по формуле

$$h_a(x) = \sum_{i=1}^{r} a_i x_i \bmod m, \tag{10}$$

где $a = (a_1, a_2, \ldots, a_r)$ — случайно выбираемый вектор из вычетов по модулю m. Нетрудно видеть, что всего существует m^r таких векторов.

Теорема 4. *Семейство H, составленное из всех функций вида* (10)*, является универсальным семейством хеш-функций.*

Доказательство. См. в книгах [1] и [2, стр. 46]. □

Пример 9. IP-адрес представляет собой 32-битное целое число, которое часто записывают как четвёрку 8-битных целых чисел, например 192.168.1.2. Пусть нужно организовать хранение примерно 250 IP-адресов в хеш-таблице.

Каждый адрес можно рассматривать как четырёхкомпонентный вектор. Тогда получается, что $n = 256$. Размер таблицы m предлагается взять равным 257 (это простое число).

Для выбора хеш-функции из универсального семейства нужно сгенерировать четыре случайных целых числа из отрезка от 0 до 256 включительно, а потом подставить их в формулу (10). Например, если выбрано $a = (87, 23, 125, 4)$, то получается такая хеш-функция:

$$h(x) = (87x_1 + 23x_2 + 125x_3 + 4x_4) \bmod 257.$$

3.2.5 Обобщения

Кроме понятия универсальности, мы вводили также понятие независимости разных сортов. Описанная схема обладает свойством 2-независимости, и из неё, кстати, также вытекает универсальность. Свойство 3-независимости уже не выполнено.

Чтобы построить k-независимую систему систему функций, можно брать полиномы. Оказывается, полинома степени k достаточно, чтобы обеспечить $(k+1)$-независимость. Этот результат интересен для исследований в области теории, хотя вряд ли 3- или 4-независимость влечёт за собой большие преимущества по сравнению с 2-независимостью. Если же брать величину k не константной, а, скажем, порядка $O(\log n)$, то станет невозможно вычислять значение хеш-функции за $O(1)$, как прежде.

3.3 Совершенное хеширование

Напомним постановку задачи. Если множество, которое мы пытаемся отхешировать, по размеру меньше, чем размер хеш-таблицы, то неконструктивно, но факт, существует некоторая хеш-функция, которая не имеет коллизий. Вопрос в том, как эту хеш-функцию компактно задать, т. е. существует ли какое-нибудь не очень большое семейство, из которого такую хеш-функцию всегда можно выбрать, и желательно, чтобы она не просто компактно задавалась (достаточно было константы памяти), но и чтобы она ещё и вычислялась за константное время.

Таких схем сейчас придумано довольно много. Рассмотрим исторически первую схему, называемую FKS — аббревиатура, составленная по инициалам авторов Fredman, Komlós и Szemerédi. Метод был предложен в 1984 г [3]. Метод основан на универсальном хешировании и является рандомизированным.

План действий такой. Во-первых, мы поймём, что если взять какое-нибудь универсальное семейство хеш-функций и хешировать в достаточно большую хеш-таблицу, то можно избежать коллизий. Во-вторых, эту идею мы обобщим на некоторую двухуровневую схему, и у нас получится желаемый результат.

3.3.1 Хеш-таблица без коллизий

Оказывается, что если размер хеш-таблицы достаточно велик, есть хорошая вероятность выбрать такую хеш-функцию, что коллизий вообще не будет.

Предположим, что необходимо организовать хранение ключей из фиксированного множества $S = \{x_1, x_2, \ldots, x_N\}$. Пусть хеш-функция h, которая раскладывает ключи по m цепочкам, выбирается случайно из семейства H. Для простоты рассуждений, чтобы не иметь дела с асимптотической нотацией, будем считать, что семейство хеш-функций H является строго универсальным, т. е. для любых двух ключей x_i и x_j, где $i \neq j$, вероятность того, что хеш-значения у них совпадут, не превышает $1/m$:

$$\mathbf{P}\left\{h(x_i) = h(x_j)\right\} \leq \frac{1}{m}. \tag{11}$$

Сколько коллизий в среднем будет в такой хеш-таблице?

Теорема 5. *Пусть хеш-функция выбирается из строго универсального семейства и распределяет фиксированный набор ключей по m цепочкам. Если в хеш-таблице хранится N ключей, то ожидаемое число коллизий меньше $N^2/2m$.*

Доказательство. Число коллизий является случайной величиной, обозначим его через ξ. Точнее, случайно осуществляется выбор хеш-функции из семейства, а по хеш-функции и набору ключей детерминированно можно подсчитать число коллизий.

Отметим для понимания, что максимальное значение ξ равно C_N^2, или $\frac{N(N-1)}{2}$, и получается в случае, когда все N ключей попадают в одну цепочку, тогда каждая неупорядоченная пара $\{x_i, x_j\}$ даёт одну коллизию. Если $N \leq m$, то минимальное значение ξ равно нулю — нет коллизий.

Вообще, случайная величина ξ имеет дискретное распределение вероятностей и принимает значения из множества $\{0, 1, \ldots, \frac{N(N-1)}{2}\}$, каждое с некоторой вероятностью. На самом деле, нам нет необходимости искать эти вероятности. Для доказательства утверждения теоремы нам только нужно показать, что $\mathbf{E}\left\{\xi\right\} < \frac{N^2}{2m}$.

Рассмотрим пару ключей $\{x_i, x_j\}$, где для определённости $i < j$. Если случается, что $h(x_i) = h(x_j)$, то к счётчику общего числа коллизий прибавляется единица, иначе прибавляется нуль. Отсюда вытекает выражение для случайной величины ξ в виде суммы других случайных величин:

$$\xi = \sum_{1 \leq i < j \leq N} \mathbb{1}_{\{h(x_i)=h(x_j)\}},$$

где

$$\mathbb{1}_{\{h(x_i)=h(x_j)\}} = \begin{cases} 1, & \text{если } h(x_i) = h(x_j), \\ 0, & \text{иначе} \end{cases}$$

— случайная величина, индикатор события «у i-го и j-го ключа совпали хеши».

По определению математического ожидания дискретной случайной величины имеем:

$$\mathbf{E}\left\{\mathbb{1}_{\{h(x_i)=h(x_j)\}}\right\} = 1 \cdot \mathbf{P}\left\{h(x_i) = h(x_j)\right\} + 0 \cdot \mathbf{P}\left\{h(x_i) \neq h(x_j)\right\} =$$

$$= \mathbf{P}\left\{h(x_i) = h(x_j)\right\}.$$

Исходя из определения строгой универсальности (11) при $i \neq j$ получаем:

$$\mathbf{E}\left\{\mathbb{1}_{\{h(x_i)=h(x_j)\}}\right\} \leq \frac{1}{m}.$$

В силу линейности математического ожидания

$$\mathbf{E}\{\xi\} = \mathbf{E}\left\{\sum_{1\leq i<j\leq N} \mathbb{1}_{\{h(x_i)=h(x_j)\}}\right\} = \sum_{1\leq i<j\leq N} \mathbf{E}\left\{\mathbb{1}_{\{h(x_i)=h(x_j)\}}\right\} =$$

$$= \frac{N(N-1)}{2} \cdot \mathbf{E}\left\{\mathbb{1}_{\{h(x_i)=h(x_j)\}}\right\} \leq \frac{N(N-1)}{2} \cdot \frac{1}{m} < \frac{N^2}{2m}.$$

□

Получается следующее. Если выставить размер хеш-таблицы m равным N и выбирать хеш-функцию из универсального семейства, то в среднем будет иметь место не более $N/2$ коллизий. Неплохо, но хотелось бы лучше.

Логично попробовать создать хеш-таблицу большего размера, чтобы прийти к меньшему числу коллизий. Из теоремы 5 следует, что если $m \sim N^2$, то в среднем число коллизий константно. Более того, можно оценить вероятность полного отсутствия коллизий. Покажем, что квадратичный рост обеспечивает отсутствие коллизий с существенной вероятностью.

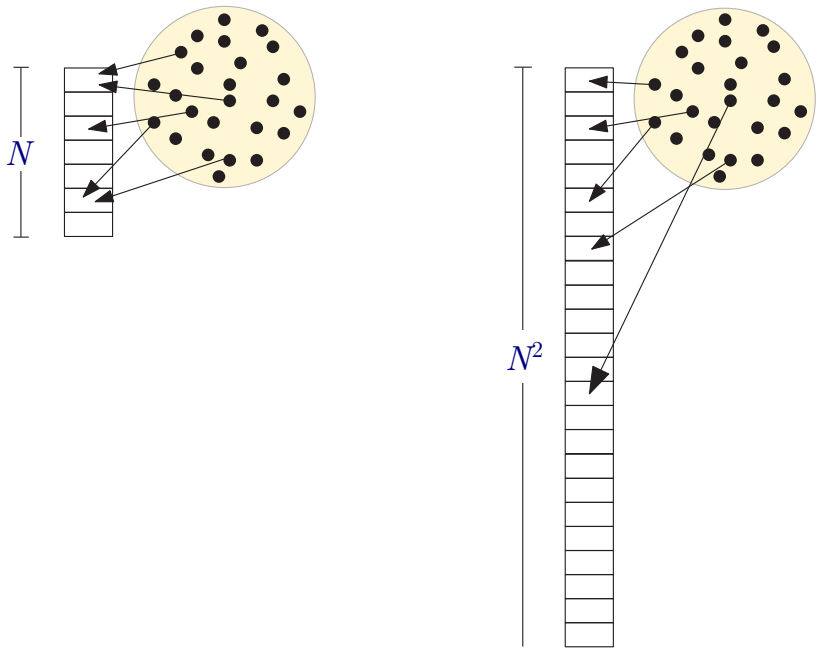

Рис. 6: Хеширование ключей в таблицу размера N и таблицу размера N^2

Теорема 6. *Пусть хеш-функция выбирается из строго универсального семейства. Если размер хеш-таблицы m растёт квадратично с ростом числа хранимых ключей N, т. е. $m = N^2$, то с вероятностью более $1/2$ в хеш-таблице вообще не будет коллизий.*

Доказательство. Пусть ξ — число коллизий. Исходя из теоремы 5, имеем

$$\mathbf{E}\{\xi\} < \frac{N^2}{2N^2} = \frac{1}{2}.$$

Поскольку ξ — неотрицательная случайная величина, для неё справедливо неравенство Маркова:

$$\mathbf{P}\{\xi \geq c\} \leq \frac{\mathbf{E}\{\xi\}}{c}. \tag{12}$$

Интуитивный смысл неравенства Маркова прост: случайная величина превосходит её матожидание в α раз не чаще, чем в $1/\alpha$ случаев.

Положив в (12) константу c равной единице, имеем:

$$\mathbf{P}\{\xi \geq 1\} \leq \frac{\mathbf{E}\{\xi\}}{1} = \mathbf{E}\{\xi\} < \frac{1}{2}. \tag{13}$$

Случайная величина ξ целочисленная, поэтому она либо равна нулю (нет коллизий), либо принимает значение не меньше единицы:

$$\mathbf{P}\{\xi = 0\} + \mathbf{P}\{\xi \geq 1\} = 1. \tag{14}$$

Исходя из (13) и (14),

$$\mathbf{P}\{\xi = 0\} > \frac{1}{2}.$$

□

Таким образом, если размер таблицы квадратичный, то вероятность выбрать такую хеш-функцию из универсального семейства, что коллизий не будет, даже больше, чем вероятность выпадения орла при подбрасывании монеты. Более чем в половине случаев хеш-функции оказываются хорошими.

Предлагается такой рандомизированный алгоритм построения хеш-таблицы без коллизий. Мы берём некоторую хеш-функцию и пробуем. Если коллизий нет, очень хорошо, а если есть, перезапускаем процесс. Пробуем другую хеш-функцию...

Как проверить, даёт ли хеш-функция коллизии? Самым непосредственным образом: будем заполнять хеш-таблицу, и если встретим при добавлении очередного ключа занятую хеш-корзину, то прервём процесс: не повезло, коллизии есть. Необходимо время $\mathrm{O}(m)$, чтобы инициализировать таблицу, и затем $\mathrm{O}(N)$ на проверку каждой новой хеш-функции.

Сколько нужно проб, чтобы наконец найти хеш-функцию, которая коллизий не даёт? Мы имеем дело с распределением Бернулли. Пусть вероятность успеха при каждой пробе равна p. Пусть ζ — количество испытаний случайного эксперимента до наблюдения первого успеха. Известно, что случайная величина ζ имеет геометрическое распределение, и $\mathbf{E}\{\zeta\} = \frac{1}{p}$. Поскольку $p > \frac{1}{2}$ в силу теоремы 6, то $\mathbf{E}\{\zeta\} < 2$. В среднем менее чем со второй попытки мы найдём хорошую хеш-функцию, не дающую коллизий.

Таким образом можно получать совершенные хеш-функции. Но трудность заключается в том, что размер таблицы растёт квадратично с ростом количества элементов, которые мы будем хешировать. Это, конечно, плохо, хотелось бы таблицу линейного размера.

3.3.2 Двухуровневая схема

Построим двухуровневую хеш-таблицу. На первом уровне коллизии возможны. Множество ключей отобразится как-то на цепочки, вполне могут быть цепочки длины больше единицы. Внутри каждой цепочки мы построим ещё одну хеш-таблицу, причём с квадратичным размером, для того чтобы обеспечить отсутствие коллизий внутри каждой цепочки.

Хеш-таблица первого уровня будет иметь линейный размер по сравнению с общим количеством ключей. Конечно, вероятность того, что коллизий там не будет, положительна, но она очень мала, и найти такую хеш-функцию мы не сможем. Легко искать хеш-функции

без коллизий, если мы квадратично расширим хеш-таблицу, но на первом уровне мы этого делать не будем.

Для каждой цепочки мы построим хеш-таблицу второго уровня, уже квадратичного размера.

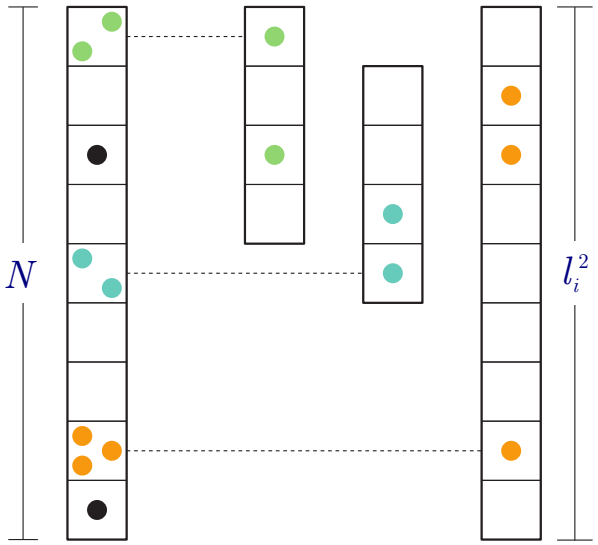

Рис. 7: Двухуровневая совершенная хеш-таблица

Для хеш-таблицы первого уровня $m = N$. На втором уровне строится m хеш-таблиц, и для i-й хеш-таблицы второго уровня $m_i = l_i^2$.

Такой план на самом деле может завершиться успехом, а может завершиться неудачей. Давайте поймём, в чём здесь успех, в чём неудача. Кажется, схема всегда будет работать. Можно всегда взять какую-то хеш-функцию для таблицы первого уровня, получить цепочки и каждую квадратично захешировать. Коллизий суммарно не будет: коллизия на первом уровне разрешится на втором, и время выполнения каждой операции с множеством, построенным таким образом, будет O(1). Возможная неудача заключается в том, что размер структуры данных может оказаться большим. Хеш-таблица первого уровня имеет фиксированный размер N. Рассмотрим суммарный размер всех хеш-таблиц второго уровня:

$$\sum_{i=0}^{m-1} l_i^2.$$

Это случайная величина, которая зависит исключительно от выбора хеш-функции первого уровня. Выбор хеш-функции первого уровня фиксирует нам цепочки. Можно оценить статистические характеристики этой случайной величины.

Теорема 7. *Пусть хеш-функция выбирается из строго универсального семейства хеш-функций и распределяет N ключей по N цепочкам. Тогда математическое ожидание суммы квадратов длин цепочек меньше $2N$.*

Доказательство. Пусть случайная величина ξ обозначает общее число коллизий, как в доказательстве теоремы 5. Попробуем выразить число коллизий через сумму квадратов длин цепочек.

Заметим, что цепочка длины l_i добавляет к общему числу коллизий величину $C_{l_i}^2$ (все пары ключей из одной цепочки дают коллизию по определению, а ключи из разных цепочек гарантированно имеют разные хеши и не имеют коллизий). Тогда общее число кол-

лизий получается суммированием по всем цепочкам:

$$\xi = \sum_{i=0}^{m-1} C_{l_i}^2 = \sum_{i=0}^{m-1} \frac{l_i(l_i-1)}{2}.$$

Перепишем немного по-другому это выражение:

$$\xi = \sum_{i=0}^{m-1} \frac{l_i(l_i-1)}{2} = \sum_{i=0}^{m-1} \frac{l_i^2 - l_i}{2} = \frac{1}{2}\left(\sum_{i=0}^{m-1} l_i^2 - \sum_{i=0}^{m-1} l_i\right) = \frac{1}{2}\left(\sum_{i=0}^{m-1} l_i^2 - N\right)$$

(сумма длин всех цепочек равна общему числу ключей в хеш-таблице). Отсюда выражается сумма квадратов длин:

$$\sum_{i=0}^{m-1} l_i^2 = 2\xi + N.$$

Тогда

$$\mathbf{E}\left\{\sum_{i=0}^{m-1} l_i^2\right\} = \mathbf{E}\left\{2\xi + N\right\} = 2 \cdot \mathbf{E}\left\{\xi\right\} + N < 2 \cdot \frac{N^2}{2N} + N = 2N$$

в силу линейности матожидания и теоремы 5. □

Таким образом, сумма квадратов длин цепочек растёт в среднем линейно. Это удивительный факт, на самом деле.

Матожидание размера двухуровневой структуры данных линейно. Но, вообще говоря, это плохо. Может не повезти, и она будет какого-то большого размера, в худшем случае квадратичного. Что же делать? Решение заключается в том, чтобы снова пробовать. Генерируем одну структуру, и если её размер нас не устраивает, то генерируем новую, и т. д. Неравенство Маркова говорит, что вероятность, с которой размер структуры будет превосходить её матожидание, скажем, в два раза, не больше, чем $\frac{1}{2}$. Размер в среднем линейный, поэтому как минимум в половине случаев он получается линейным.

Таким образом, у нас есть перебор при построении хеш-таблиц второго уровня: мы пробуем, пока все коллизии не уйдут. И у нас есть перебор при построении хеш-таблицы первого уровня: мы пробуем до тех пор, пока размер структуры нас не устроит. Чтобы узнать, какой будет размер структуры, не нужно структуру до конца строить: достаточно посчитать длины цепочек. При построении первого уровня мы не смотрим на коллизии (они там почти наверняка будут), а смотрим на размеры цепочек.

3.3.3 Обобщения совершенного хеширования

Таким образом, мы научились строить статическую таблицу, когда ключи известны заранее. Существуют модификации совершенных хеш-функций, которые являются динамическими.

Есть схемы, которые работают быстрее FKS в полтора-два раза на практике из-за меньшей скрытой константы. Они могут быть интересны, если цель — микрооптимизация программы.

4 Хеш-таблицы на практике

Структуры данных на основе хеш-таблиц реализованы в стандартных библиотеках всех широко используемых языков программирования. В большинстве случаев библиотеки предоставляют как множество, так и ассоциативный массив.

4.1 Требования к ключам

Зачастую при программировании ключами в множествах и ассоциативных массивах выступают не просто целые числа, а какие-либо более сложные объекты, например строки, пары чисел, структуры из нескольких полей и прочие объекты произвольной природы.

Чтобы объект некоторого типа мог выступать ключом в хеш-таблице, необходимо выполнение двух условий:

- должна быть задана хеш-функция, которая ставит в соответствие объекту его хеш-значение (целое число, причём диапазон допустимых значений оговаривается в спецификациях);
- должна быть определена функция, которая для пары объектов отвечает, равны они или нет.

Формально от хеш-функции требуется лишь, чтобы для равных объектов хеш-значения были равными. Функция проверки на равенство должна задавать на множестве объектов отношение эквивалентности, то есть бинарное отношение, для которого выполнены следующие условия:

1. $a = a$ для любого a (свойство рефлексивности);
2. если $a = b$, то $b = a$ (свойство симметричности);
3. если $a = b$ и $b = c$, то $a = c$ (свойство транзитивности).

Для того чтобы объект мог служить ключом в бинарном дереве поиска, требование выдвигается совсем иное: нужно уметь сравнивать любые два элемента. Строго говоря, множество объектов должно быть линейно упорядочено. Напомним, что линейным порядком на множестве называют бинарное отношение $\leq$, которое удовлетворяет следующим условиям:

1. любые два объекта сравнимы между собой, т. е. $a \leq b$ или $b \leq a$ (свойство полноты);
2. если $a \leq b$ и $b \leq a$, то $a = b$ (свойство антисимметричности);
3. если $a \leq b$ и $b \leq c$, то $a \leq c$ (свойство транзитивности).

Такой порядок также называют полным порядком (*total order*). Линейно упорядоченные объекты могут быть размещены на прямой линии один за другим. Первое свойство влечёт за собой свойство рефлексивности, т. е. $a \leq a$. Следовательно, линейный порядок является также частичным порядком. Понятие частично упорядоченного множества слабее (в определении частичного порядка в первом пункте требуется лишь рефлексивность).

От языка программирования зависит способ, при помощи которого можно определять те или иные операции для объектов произвольного типа (подсчёт хеш-функции, проверка на равенство или неравенство).

На программиста возлагается обязанность соблюдать те или иные условия в своём коде. Например, если функция проверки двух объектов на равенство будет возвращать каждый раз случайное значение (истинное или ложное), естественно ожидать, что хеш-таблица будет работать неправильно. Если функция сравнения для множества на основе дерева не обеспечивает линейного упорядочивания, программа будет вести себя непредсказуемо и, возможно, аварийно завершится с ошибкой.

4.2 Объединение хеш-значений

Нередко на практике требуется реализовать хеш-функцию от структуры, содержащей два поля, для каждого из которых по отдельности хеш-функции определены. Например, координаты точек на плоскости хранятся в виде пар целых чисел (x, y), и нужно создать множество точек с использованием хеш-таблицы, а для этого необходимо вычислять хеш-значения от точек. Пусть получены хеш-значения двух координат $h(x)$ и $h(y)$. Как их объединить, чтобы получить хеш от пары? Пусть для простоты верхняя граница возможных значений хеш-функции не фиксирована (где-то дальше в реализации хеш будет взят по нужному модулю m).

Часто на практике программисты для соединения хешей пишут тривиальные функции, например через операцию побитового исключающего или (xor):

```
def combine(hx, hy):
    return hx ^ hy
```

Такой вариант часто работает на практике приемлемо, но не лишён очевидных недостатков. Например, для всех точек с равными координатами x и y хеш-функция будет принимать нулевое значение, и если точек на прямой $y = x$ во входных данных окажется много, производительность будет низкой из-за коллизий. Также очевидно, что разные точки (x, y) и (y, x), симметричные относительно той же прямой, получат одинаковые хеш-значения.

Чтобы подобрать пары, дающие коллизию, было труднее, для объединения хешей используют более сложные функции с обилием «магических» констант и странных операций. Например, в C++-библиотеке `boost` используется примерно такая формула:

```
def combine(hx, hy):
    return hx ^ (hy + 0x9e3779b9 + (hx << 6) + (hx >> 2))
```

Часто берут линейную комбинацию двух хеш-значений с, например, большими взаимно простыми коэффициентами. Как вариант:

```
def combine(hx, hy):
    return hx + 1000000007 * hy
```

Основной смысл таких манипуляций — сделать, чтобы на реально встречающихся в жизни данных коллизии были более редки. Но контрпример при желании можно подобрать. Лучшего универсального решения в этом деле нет.

Отметим, что если научиться объединять хеши двух элементов, то можно последовательно объединить хеши любого числа элементов.

4.3 Проход по содержимому хеш-таблицы

В процессе программирования может возникнуть необходимость выполнить обход всех элементов структуры данных и, например, распечатать их. Например, требуется вывести все ключи, которые содержатся в заданном динамическом множестве.

Функция для итерации по содержимому структуры данных не вводилась в разделе 1, но является полезной, поэтому обычно поддерживается в реализациях хеш-контейнеров, с которыми ведётся работа на практике.

Стоит помнить, что в большинстве реализаций проход по хеш-множествам выполняется в произвольном порядке, не гарантируется какой-либо отсортированности ключей. В случае, если внутренняя реализация хеш-таблицы использует метод цепочек, обычно функция обхода выдаёт сначала все элементы первой корзины (с хеш-значением 0) в порядке их следования в цепочке, затем все элементы второй корзины (с хеш-значением

1), и т. д. Для внешнего наблюдателя может казаться, что ключи выходят в случайном порядке.

Более того, если распечатать элементы хеш-множества, добавить новый ключ, сразу удалить его, вновь распечатать элементы, то порядок может получиться другим. Такое может случиться, если добавление нового ключа привело к перестроению хеш-таблицы с изменением числа m корзин, и элементы были перераспределены по корзинам вновь.

Не стоит нигде в коде закладываться на порядок итерации по хеш-контейнерам: большинство реализаций в разных языках программирования могут гарантировать только, что посещены будут все элементы, не важно в каком порядке.

Наоборот, средства итерации по ключам множества, которое построено на базе бинарного поискового дерева, обычно возвращают ключи в порядке возрастания (выполняется внутренний обход дерева). Порядок фиксирован и каждый раз одинаковый. Часто предсказуемость результата удобна, например, для написания модульных тестов к частям программы.

Таким образом, если порядок итерации важен, возможно, стоит использовать «древесные» структуры данных.

4.4 Хеш-таблицы в C++

Долгое время в языке C++ не было стандартных реализаций структур данных на основе хеш-таблиц. Так, контейнеры `std::set` и `std::map` из STL строятся на основе сбалансированных бинарных поисковых деревьев (во всех популярных реализациях применяются красно-чёрные деревья). Хеш-таблицы существовали в виде нестандартных расширений (например `stdext::hash_set` в Visual Studio) или внешних библиотек (например `boost`). В 2007 г. появились контейнеры `std::tr1::unordered_set`... (черновая версия стандарта).

Наконец, в стандарте C++11 в STL официально были добавлены хеш-таблицы. Стандарт предусматривает четыре контейнера на основе хеш-таблиц, которые отличаются от своих аналогов на основе деревьев наличием префикса `unordered_` в названии. Так, `std::unordered_set` представляет собой динамическое множество, `std::unordered_map` — ассоциативный массив. Существует также два `multi`-контейнера, которые допускают хранение одинаковых ключей.

Стандарт требует, чтобы в построении этих структур данных авторы компиляторов использовали разрешение коллизий методом цепочек. Метод открытой адресации не был стандартизирован из-за внутренних трудностей при удалении элементов. Однако детали реализации хеш-таблиц стандартом не регламентируются.

В качестве хеш-значения в C++ используется число типа `size_t`. Все хеш-контейнеры предоставляют метод `rehash()`, который позволяет установить размер хеш-таблицы (число корзин m). Метод `load_factor()` возвращает текущий коэффициент заполнения.

Рассмотрим более подробно реализацию в компиляторе GCC. Пусть ключи добавляются в `std::unordered_set` по одному. Когда коэффициент заполнения достигает значения 1, происходит перестроение хеш-таблицы: в качестве нового числа корзин берётся первое простое число из заранее составленного списка, не меньшее удвоенного старого числа корзин (таким образом, размер таблицы как минимум удваивается и является простым числом). Длины отдельных цепочек никак не анализируются (появление одной длинной цепочки не повлечёт за собой операцию перестроения).

4.5 Хеш-таблицы в Java

Предполагаем, что речь идёт о языке Java актуальных на момент написания этих строк версий (7 и 8).

Коллекции `HashSet` и `HashMap` реализуются как хеш-таблицы, для разрешения коллизий используется метод цепочек.

Для хеширования целых чисел применяется функция следующего вида:

```
int hash(int h) {
    h ^= (h >>> 20) ^ (h >>> 12);
    return h ^ (h >>> 7) ^ (h >>> 4);
}
```

(здесь операция `>>>` — беззнаковый сдвиг вправо: биты смещаются вправо, число слева дополняется нулями, операция `^` — поразрядное сложение по модулю 2, исключающее или). Затем в классе коллекции результат функции `hash()` берётся по модулю числа корзин, по которым раскладываются элементы. Число корзин, оно же число различных значений хеш-функции m, в Java всегда выбирается как некоторая степень числа 2, чтобы деление на m можно было заменить операцией битового сдвига вправо (современные процессоры выполняют инструкцию деления целых чисел существенно медленнее, чем битовые операции).

В версии Java 8 разработчики озаботились вопросом устойчивости коллекций, использующих хеширование, к коллизиям [5]. В исходном коде библиотеки Java можно найти следующую константу:

```
static final int TREEIFY_THRESHOLD = 8;
```

В случае, если новый ключ попадает в корзину, в которой уже лежат как минимум восемь других ключей, библиотека преобразует связный список для данной корзины в бинарное сбалансированное поисковое дерево. Получается гибридная структура: корзины для тех хеш-значений, где ключей мало, хранятся списками, а корзины, где ключей накопилось много, хранятся в виде деревьев.

4.6 Хеш-таблицы в Python

Рассмотрим реализацию языка программирования CPython версии 2.7.

Встроенный тип `dict` — ассоциативный массив, словарь — очень широко используется в языке. Он реализован в виде хеш-таблицы, где коллизии разрешаются методом открытой адресации. Разработчики предпочли метод открытой адресации методу цепочек ввиду того, что он позволяет значительно сэкономить память на хранении указателей, которые используются в хеш-таблицах с цепочками.

Подробное описание принципов устройства словаря в Python на русском языке можно найти в статье [4].

Интерпретатором CPython поддерживается опция командной строки `-R`, которая активирует на старте случайный выбор начального значения (*seed*), которое затем используется для вычисления хеш-значений от строк и массивов байт.

4.7 Атаки против стандартных хеш-функций

Для целочисленной хеш-функции, которая используется в Java, оказалось нетрудно придумать обратную, поскольку функция выполняет линейное преобразование битов исходного ключа. С использованием этого результата можно строить наборы различных ключей, которые при добавлении в `HashSet` попадают в одну корзину. Пример кода, который генерирует тест против стандартной хеш-таблицы в Java, можно найти в публикации [6].

Более того, на одних и тех же входных данных может наблюдаться деградация производительности сразу нескольких реализаций хеш-таблиц. Так, в рамках соревнования IPSC 2014 была предложена следующая задача [7]. Дано две программы. Первая программа на языке C++, просто читает числа на входе и добавляет их по одному в множество на основе `std::unordered_set`. Вторая программа на языке Java полностью аналогична, выполняет чтение чисел и занесение их в `HashSet`. Требовалось построить такую последовательность из не более чем 50 тысяч 64-битных целых чисел, чтобы время работы обоих программ превысило 10 секунд на сервере организаторов (указывались особенности архитектуры и версии компиляторов обоих языков программирования). Задача была решена многими участниками, т. е. им удалось подобрать такой набор входных данных, что построение обеих хеш-таблиц заняло квадратичное время из-за коллизий.

Существует целый класс атак на серверы, называемый *Hash-flooding DoS* — отказ в обслуживании, вызванный заполнением хеша. Злоумышленники формируют специальные запросы, которые вызывают на сервере большое число хеш-коллизий и оттого медленно обрабатываются. В результате вычислительные ресурсы тратятся впустую, легальные пользователи системы не могут получить доступ к ресурсам либо этот доступ затруднён. Подробнее про этот тип атак можно почитать в презентации [8].

4.8 Криптографические хеш-функции

Для приложений в области криптографии разработаны специальные хеш-функции. Можно считать, что в криптографии множество K возможных ключей бесконечно, и любой блок данных является ключом (в принципе, произвольный массив байт можно рассматривать как двоичную запись некоторого числа). Хеш-функция $h(x)$ называется криптографической, если она удовлетворяет следующим требованиям:

- *необратимость*: для заданного значения хеш-функции c должно быть сложно определить такой ключ x, для которого $h(x) = c$;
- *стойкость к коллизиям первого рода*: для заданного ключа x должно быть вычислительно невозможно подобрать другой ключ y, для которого $h(x) = h(y)$;
- *стойкость к коллизиям второго рода*: должно быть вычислительно невозможно подобрать пару ключей x и y, имеющих одинаковый хеш.

Криптографические хеш-функции обычно не используются в хеш-таблицах, потому что они сравнительно медленно вычисляются и имеют большое множество значений. Зато такие хеш-функции широко применяются в системах контроля версий, системах электронной подписи, во многих системах передачи данных для контроля целостности.

Примерами криптографических хеш-функций являются алгоритмы MD5, SHA-1, SHA-256. Так, метод SHA-1 ставит в соответствие произвольному входному сообщению некоторую 20-байтную величину, т. е. результат вычисления SHA-1 принимает одно из 2^{160} различных значений. Вот пример вычисления SHA-1 от ASCII-строки, где результат записан в шестнадцатеричной системе счисления:

```
SHA-1("The quick brown fox jumps over the lazy dog")
  = 0x2fd4e1c67a2d28fced849ee1bb76e7391b93eb12
```

По состоянию на конец 2015 г. коллизии для SHA-1 неизвестны, т. е. не подобрано пары различных сообщений, у которых хеши были бы одинаковы, хотя работы в этом направлении ведутся. Для алгоритма MD5 коллизии обнаружены, поэтому метод постепенно выходит из широкого использования. Более новые алгоритмы семейства SHA-2 считаются

существенно более стойкими к коллизиям. Тем не менее, следует понимать, что коллизии есть обязательно, потому что нельзя биективно отобразить бесконечное множество в конечное. Вопрос только в том, насколько трудно эти коллизии отыскать.

5 Заключение

Таким образом, рассмотрены подходы к реализации динамических множеств и ассоциативных массивов на основе хеш-таблиц в теории и на практике.

Как уже говорилось, альтернативой хеш-таблицам являются структуры на основе сбалансированных бинарных поисковых деревьев. Возникает вопрос, как теоретический, так и практический, о том, насколько эти способы отличаются, какой из них лучше или хуже. Ответ здесь такой: хеш-таблицы в некотором доказуемом смысле могут быть быстрее.

Интуитивно более-менее понятно, за счёт чего: в случае хеш-таблицы мы работаем с ключами, которые являются просто числами, при этом на этих числах не предполагается никакого порядка, при этом интерфейс не может этот порядок воспроизвести, например нельзя найти все ключи от a до b, просто потому что на ключах порядка как такового нет. Конечно, если ключи — это числа, то порядок возникает, но никакой семантики не несёт. Мы отказываемся от требования упорядоченности, потому можем структуру данных сделать более эффективной.

Так, если к исходному интерфейсу множества (см. раздел 1) добавить операцию типа **LowerBound**(x) — найти первый ключ, больший либо равный x, то это получится уже не просто множество (set), а упорядоченное множество (ordered set). Новая операция осуществляет поиск в окрестности значения x. На первый взгляд, единственный доступный нам приём все эти операции поддерживать — это использовать деревья поиска. Оказывается, нет. Существуют интересные гибриды, находящиеся посередине между деревьями поиска и хеш-таблицами, они в частности реализуют концепцию упорядоченного множества. Это всевозможные деревья Ван Эмде Боасса (Van Emde Boas tree), X-fast-, Y-fast- и Fusion-деревья, у которых в оценках временной сложности появляется двойной логарифм.

Список литературы

[1] Алгоритмы: построение и анализ / Т. Кормен [и др.]. — М.: Вильямс, 2005. — 1296 с.

[2] S. Dasgupta, C. H. Papadimitriou, U. V. Vazirani. Algorithms. New York, USA: McGraw-Hill, 2006. — 336 p.

[3] Fredman, M. L., Komlós, J., and Szemerédi, E. 1984. Storing a Sparse Table with O(1) Worst Case Access Time. J. ACM 31, 3 (Jun. 1984), 538-544.

[4] Реализация словаря в Python 2.7. `http://habrahabr.ru/post/247843/`

[5] JEP 180: Handle Frequent HashMap Collisions with Balanced Trees. `http://openjdk.java.net/jeps/180`

[6] Почему не надо использовать Java.HashSet/HashMap для целых чисел. `http://codeforces.com/blog/entry/4876`

[7] IPSC 2014. Problem H – Hashsets. `http://ipsc.ksp.sk/2014/real/problems/h.html`

[8] J.-Ph. Aumasson, D. J. Bernstein, M. Boßlet. Hash-flooding DoS reloaded: attacks and defenses. `https://131002.net/siphash/siphashdos_appsec12_slides.pdf`