Алгоритмы и структуры данных

Лекция 1

Сергей Леонидович Бабичев

Содержание лекции

- Свойства алгоритма.
- Сложность алгоритма. O- и Θ- нотации.
- Исполнитель алгоритма.
- Корректность алгоритмов. Инварианты. Индуктивные функции.
- Абстракции. Интерфейс абстракции.

Свойства алгоритма.

Исполнитель

Алгоритм — это последовательность команд для *исполнителя*, обладающая рядом свойств:

- **полезность**, то есть умение решать поставленную задачу.
- **детерминированность**, то есть каждый шаг алгоритма должен быть строго определён во всех возможных ситуациях.
- **конечность**, то есть способность алгоритма завершиться для любого множества входных данных
- **массовость**, то есть применимость алгоритма к разнообразным входным данным.

Алгоритм для своего *исполнения* требует от исполнителя некоторых *ресурсов*. *Программа* есть запись алгоритма на формальном языке.

Алгоритмы вокруг нас

Рецепт приготовления утки по-пекински.

- **Алгоритм** — порядок действий для приготовления.
- **Исполнитель** — повар или мы сами.
- **Полезность** — если получилось вкусно — алгоритм полезен.
- **Детерминированность** — здесь проблемы.
- **Конечность** — тушить «до готовности». Что такое «готовность»?
- **Массовость** — для всех ли ингредиентов возможен хороший результат?
- **Ресурсы** — как ингредиенты, так и оборудование.
- **Программа** — рецепт из книги или с сайта.

Исполнители

Одна задача — несколько алгоритмов — разные используемые ресурсы.
Разные исполнители — разные *элементарные действия* и *элементарные объекты*.
Исполнитель «компьютер»:

- устройство *центральный процессор*
- элементарные действия — сложение, умножение, сравнение, переход ...
- устройство *память* как хранителя элементарных объектов
- элементарные объекты — целые, вещественные числа

Эффективность — способность алгоритма использовать ограниченное количество ресурсов.

Сложность алгоритма. O- и Θ- нотации.

Сложность алгоритма

Что есть сложность алгоритма?

- *комбинационная сложность* — число элементов для реализации алгоритма в виде вычислительного устройства
- *описательная сложность* — длина описания алгоритма на формальном языке
- *вычислительная сложность* — количество элементарных операций, исполняемых алгоритмом для неких входных данных.

Нет циклов — описательная сложность примерно коррелирует с вычислительной.
Есть циклы — интересна асимптотика зависимости времени вычисления от входных данных.

Главный параметр сложности алгоритма

Главный параметр N, наиболее сильно влияющий на скорость исполнения алгоритма. Это может быть:

- размер массива
- количество символов в строке
- количеством битов в записи числа
- если таких параметров несколько — обобщённый параметр, функция от нескольких параметров

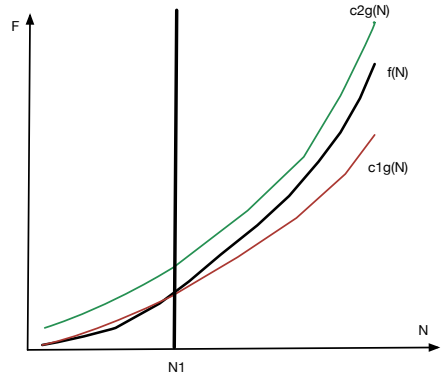
Нотация сложности. Символ Θ

Далее: сложность $\equiv$ вычислительная сложность.

Функция $f(N)$ имеет порядок $\Theta(g(N))$, если существуют постоянные c_1, c_2 и N_1 такие, что для всех $N > N_1$

$$0 \leqslant c_1 g(N) \leqslant f(N) \leqslant c_2 g(N).$$

$\Theta(f(n))$ — класс функций, примерно пропорциональных $f(n)$

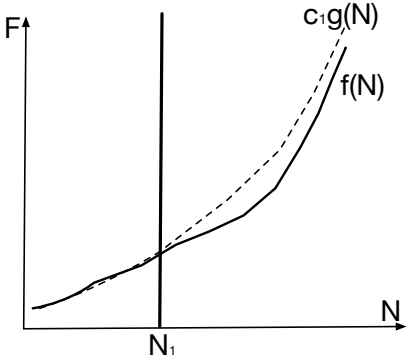


Нотация сложности. Символ O

Функция $f(N)$ имеет порядок $O(g(N))$, если существуют постоянные c_1 и N_1 такие, что для всех $N > N_1$

$$f(N) \leqslant c_1 g(N)$$

$O(f(n))$ — класс функций, ограниченных сверху $cf(n)$.



Приближённое вычисление сложности

- Пусть $F(N)$ — функция сложности алгоритма в зависимости от N
- Тогда если существует такая функция $G(N)$ (асимптотическая функция) и константа C, что

$$\lim_{N \to \infty} \frac{F(N)}{G(N)} = C,$$

то сложность алгоритма $F(N)$ определяется функцией $G(N)$ с коэффициентом амортизации C.

Асимптотика основных зависимостей

Класс сложности определяется по асимптотической зависимости *F(N)*.

- Экспонента с любым коэффициентом превосходит любую степень.
- Степень с любым коэффициентом, большим единицы, превосходит логарифм по любому основанию, большему единицы.
- Логарифм по любому основанию, большему единицы превосходит 1.

- $F(N) = N^3 + 7N^2 - 14N = \Theta(N^3)$
- $F(N) = 1.01^N + N^{10} = \Theta(1.01^N)$
- $F(N) = N^{1.3} + 10 \log_2 N = \Theta(N^{1.3})$

Зависимость времени исполнения от исходных данных

Пусть имеется массив A длиной N элементов.

Алгоритм должен обнаружить самый первый элемент со значением P. Сколько потребуется операций? Просматриваем массив слева направо.

- $K_{min} = 1$
- $K_{max} = N$
- $K_{avg} = \frac{\sum_{i=1}^{N} i}{N} = \frac{N \times (N-1)}{2N} = \frac{N-1}{2}$

  Какой символ здесь подходит, Θ или O?

Зависимость времени исполнения от исходных данных

Пусть имеется массив A длиной N элементов.

Алгоритм должен обнаружить самый первый элемент со значением P. Сколько потребуется операций? Просматриваем массив слева направо.

- $K_{min} = 1$
- $K_{max} = N$
- $K_{avg} = \frac{\sum_{i=1}^{N} i}{N} = \frac{N \times (N-1)}{2N} = \frac{N-1}{2}$

Какой символ здесь подходит, Θ или O?
Для данного алгоритма подходит O-символ: $f(N) = O(N)$.

Зависимость времени исполнения от исходных данных

Пусть имеется массив A длиной N элементов.
Сколько операций потребуется, чтобы сложить все элементы массива?

- $K = N$

  Какой символ здесь подходит, Θ или O?

Зависимость времени исполнения от исходных данных

Пусть имеется массив A длиной N элементов.
Сколько операций потребуется, чтобы сложить все элементы массива?

- $K = N$

  Какой символ здесь подходит, Θ или O?
  Подходит Θ. $F(N) = \Theta(N)$.

Неполиномиальные задачи. Задача о рюкзаке.

- Имеется:
  - N предметов, каждый из которых имеет объём V_i и стоимость C_i, предметы неделимы;
  - рюкзак вместимостью V.
- Требуется:
  - поместить в рюкзак набор предметов максимальной стоимости;
  - суммарный объём выбранных предметов не превышает объёма рюкзака.

Задача о рюкзаке.

- Предметы разрезать на куски нельзя! Разрешим — задача будет иметь простое решение.
- Для решения задачи достаточно перебрать все возможные комбинации из N предметов. Это гарантирует то, что мы не пропустим нужной комбинации.
- Для определения количества комбинаций можно рассуждать так, что K предметов можно выбрать из N предметов C_N^K и так для всех K от 0 до N.

$$F(N) = \sum_{K=0}^{N} C_N^K$$

Задача о рюкзаке.

- Предметы разрезать на куски нельзя! Разрешим — задача будет иметь простое решение.
- Для решения задачи достаточно перебрать все возможные комбинации из N предметов. Это гарантирует то, что мы не пропустим нужной комбинации.
- Для определения количества комбинаций можно рассуждать так, что K предметов можно выбрать из N предметов C_N^K и так для всех K от 0 до N.

$$F(N) = \sum_{K=0}^{N} C_N^K$$

$$F(N) = (1+1)^N = 2^N$$

Одно из решений задачи о рюкзаке

1. Перенумеруем все предметы.
2. Установим максимум стоимости в 0.
3. Составим двоичное число с N разрядами, в котором единица в разряде будет означать, что предмет выбран для укладки в рюкзак (расстановку).
4. Рассмотрим все расстановки, начиная от $000\ldots000$ до $111\ldots111$, для каждой из них подсчитаем значение суммарного объёма.
   1. Если суммарный объём расстановки не превосходит объёма рюкзака, то подсчитывается суммарная стоимость и сравнивается с достигнутым ранее максимумом стоимости.
   2. Если вычисленная суммарная стоимость превосходит максимум, то максимум устанавливается в вычисленную стоимость и запоминается текущая конфигурация.

Свойства алгоритма

Предложенный алгоритм:

1. Детерминированный.
2. Конечный.
3. Массовый.
4. Полезный.

Его сложность $O(2^N)$, так как требуется перебрать все возможные комбинации предметов.

Сложность решения задачи о рюкзаке

- Много ли времени потребуется на решение задачи для $N = 128$?

Сложность решения задачи о рюкзаке

- Много ли времени потребуется на решение задачи для $N = 128$?
- Предположим, на подсчёт одного решения потребуется 10^{-9} секунд, то есть, одна наносекунда.

Сложность решения задачи о рюкзаке

- Много ли времени потребуется на решение задачи для $N = 128$?
- Предположим, на подсчёт одного решения потребуется 10^{-9} секунд, то есть, одна наносекунда.
- Предположим, задачу будет решать триллион компьютеров (10^{12})

Сложность решения задачи о рюкзаке

- Много ли времени потребуется на решение задачи для $N = 128$?
- Предположим, на подсчёт одного решения потребуется 10^{-9} секунд, то есть, одна наносекунда.
- Предположим, задачу будет решать триллион компьютеров (10^{12})
- Тогда общее время решения задачи будет составлять

$$\frac{2^{128} \times 10^{-9}}{10^{12}} \text{секунд}$$

Сложность решения задачи о рюкзаке

- Много ли времени потребуется на решение задачи для $N = 128$?
- Предположим, на подсчёт одного решения потребуется 10^{-9} секунд, то есть, одна наносекунда.
- Предположим, задачу будет решать триллион компьютеров (10^{12})
- Тогда общее время решения задачи будет составлять

$$\frac{2^{128} \times 10^{-9}}{10^{12}} \text{секунд} \approx 10.8 \times 10^{9} \text{лет}$$

NP-задачи

- Задача о рюкзаке относится к классу *NP-сложных*.
- Быстрое (полиномиальное) точное решение таких задач (пока?) не найдено.
- Эта задача к тому же *NP-полная*.
- Если будет найдено решение одной из *NP-полных* задач, то будут решены все задачи из этого класса.
- Сейчас их решают приближённо.

Абстракции.
Интерфейс абстракции.

Понятие абстракции

Появляются *объекты* — появляются *абстракции* — механизм разделения сложных объектов на более простые, без деталировки подробностей разделения.

Функциональная абстракция — разделение функций, *методов*, которые манипулируют с объектами с их реализацией.

Интерфейс абстракции — набор методов, характерных для данной абстракции.

Пример: абстракция последовательности

- **create** — создать объект последовательности. Атрибуты: для чтения или для записи?
- **destroy** — удалить объект.
- **get** — получить очередной элемент последовательности.
- **put** — добавить элемент в последовательность.

Уже прочитанный элемент второй раз не прочитается.
Алгоритмы, рассчитанные на обработку последовательностей, могут иметь сложность по памяти $O(1)$ и по времени $O(N)$.

Пример: абстракция массива

- **create** — создать массив. Статический или динамический?

```
way 1: int a[100];
way 2: int *b = calloc(100, sizeof(int));
way 3: int *c = new int[100];
way 4: vector<int> c(100);
```

- **destroy** — удалить массив. Статический или динамический?

```
way 1: free(b);
way 2: delete [] c;
way 3: // not required
way 4: // not required
```

- **fetch** — обратиться к элементу массива. Основная операция (метод).

```
way 1: int q1 = a[i];
way 2: int q2 = b[i];
way 3: int q3 = c[i];
way 4: int q4 = c[i];
```

Абстракция *стек*

Одна из удобных абстракций — *стек*. Он должен предоставлять нам методы:

- **create** — создать стек. Может быть, потребуется аргумент, определяющий максимальный размер стека.
- **push** — занести элемент в стек. Размер стека увеличивается на единицу. Занесённый элемент становится *вершиной стека*.
- **pop** — извлечь элемент, являющийся вершиной стека и уменьшить размер стека на единицу. Если стек пуст, то значение операции не определено.
- **peek** — получить значение элемента, находящегося на вершине стека, не изменяя стека. Если стек пуст, значение операции не определено.
- **empty** — предикат. Истинен, когда стек пуст.
- **destroy** — уничтожить стек.

Абстракция *множество*

Множество есть совокупность однотипных элементов, на которых определена операция сравнения на равенство.
Обозначение: списком значений внутри фигурных скобок.
Пустое множество: $s = \{\}$.

- **insert** — добавление элемента в множество.

```
{1,2,3}.insert(5) -> {1,2,3,5}
{1,2,3}.insert{2} -> {1,2,3}
```

- **remove** — удалить элемент из множества.

```
{1,2,3}.remove(3) -> {1,2}
{1,2,3}.remove(5) -> ? or {1,2,3}
```

- **in** — определить принадлежность множеству.

```
{1,2,3}.in(2) -> true
{1,2,3}.in(5) -> false
```

- **size** — определить количество элементов в множестве.

Исполнитель алгоритма

Исполнители

Наш основной исполнитель — языки `C` и `C++`.

- Элементарные типы языка отображаются на вычислительную систему: `char`, `int`, `double`.
- Элементарные операции аппаратного исполнителя: операции над элементарными типами и операции передачи управления.
- Типы данных языка — комбинация элементарных типов данных.
- Операции языка — комбинация элементарных операций.

Операции

Пример. Неэлементарная операция языка: цикл `for`.

```
int a[10];
// Initializing a
// ...
int s = 0;
for (int i = 0; i < 10 && a[i] % 10 != 5; i++) {
  s += a[i];
}
```

- Неэлементарный тип: *массив*, его представитель `a`.
- Элементарный тип `int`: его представитель `s`.
- Элементарная операция присваивания (инициализации): `s=0`.
- Неэлементарная операция `for`, состоящая из операций присваивания `i = 0`, двух операций сравнения, и т. д.

Представление типов и сложность

- Целые числа — двоичное представление. `int x; unsigned y;`
- Простые элементарные операции: сложение, вычитание, присваивание, побитовые... `x += y; x &= 0x800;`
- Операции посложнее: сравнение, условное присваивание `x = (y > 0);`
- Сложные элементарные операции: целочисленное умножение. `x *= y;`
- Самые сложные: деление не на степень двойки, нахождение остатка, совершение перехода. `x = y % 10; if (x==6) y = 10;`

Представление типов и сложность

Прямолинейно закодированные

```
if (x > 0) t = 1;
else t = 0;
```

выполняются много медленнее, чем

```
t = x > 0;
```

так как для второго варианта есть готовая машинная команда.

Компиляторы об этом знают и пытаются провести *эквивалентные* преобразования.

Аппаратные исполнители

Популярные архитектуры:

- X86 — изобретена Intel, лицензирована и производится AMD. int, адреса — 32 бита. 32 бита и максимально обрабатываемый аппаратно и быстро целочисленный формат.
- X64 — изобретена AMD, лицензирована и производится Intel. int — 32 бита, но максимально обрабатываемый аппаратно и быстро целочисленный формат — 64 бита.
- ARM схожа с X86, ARM64 — с X64. Телефоны. Планшеты. Серверы (в том числе в TOP5). Начала успешно использоваться для ноутбуков и настольных компьютеров.

Модулярная арифметика

```
unsigned short x = 40000; // x - 16 bits.
x = x + x;
cout << x;
```

Чему равен x?

Модулярная арифметика

```
unsigned short x = 40000; // x - 16 bits.
x = x + x;
cout << x;
```

Чему равен x?

x = 14464.
Почему?

$$x \in [0..2^{16}) = [0..65535).$$

Биты нумеруются от 0 до 15 справа налево.
Все биты результата старше 15-го отсекаются.

Модулярная арифметика

Вся компьютерная арифметика основана на тождествах:

$$\begin{aligned}
(a+b) \pmod m &= (a \pmod m + b \pmod m) \pmod m \\
(a-b) \pmod m &= (a \pmod m - b \pmod m) \pmod m \\
(a \times b) \pmod m &= (a \pmod m \times b \pmod m) \pmod m \\
&\dots
\end{aligned}$$

В качестве m при двоичном представлении выступают числа 2^8, 2^{16}, 2^{32}, 2^{64}

```
unsigned int x,y,z; // 32 bits. [0..4294967295]
...
z = x * y;
```

$$z = (x * y) \pmod{2^{32}}$$

Корректность алгоритмов. Инварианты. Индуктивные функции.

Индуктивное программирование. Индуктивные функции.

Пусть имеется множества M и N. Аргументы функции f — последовательности элементов множества M, значения функции f — элементы множества N.

Если значение функции $f(x_1, x_2, \dots x_n)$ можно восстановить по $f(x_1, x_2, \dots, x_{n-1})$ и элементу x_n, то такая функция называется *индуктивной*.

Индуктивное программирование. Индуктивные функции.

Пусть имеется множества M и N. Аргументы функции f — последовательности элементов множества M, значения функции f — элементы множества N.

Если значение функции $f(x_1, x_2, \dots x_n)$ можно восстановить по $f(x_1, x_2, \dots, x_{n-1})$ и элементу x_n, то такая функция называется *индуктивной*.

Пример: Если мы хотим найти наибольшее значение всех элементов последовательности, то функция $maximum$ — индуктивна, так как

$$maximum(x_1, x_2, \dots, x_n) = \max(maximum(x_1, x_2, \dots, x_{n-1}), x_n)$$

Индуктивные функции и инварианты

- *Предикат* — логическое утверждение, содержащее переменную величину.
- *Инвариант* — предикат, сохраняющий своё значение после исполнения заданных шагов алгоритма.

```
int m = a[0];
for (int i = 1; i < N; i++) {
  if (a[i] > m) {
    m = a[i];
  }
}
```

- Предикат: для любого `i < N` переменная `m` содержит наибольшее значение из элементов `a[0]...a[i]`.
- Инвариант: `m` на каждом шаге равна значению индуктивной функции `maximum`.

Индуктивные функции и инварианты

- Ещё одна индуктивная функция.

```
// Input: array a[n]
// Output: sum of its elements
int sum = 0;
for (int i = 0; i < n; i++) {
  sum += a[i];
}
```

- Предикат: значение переменной `sum` после операции сложения в момент времени `i` есть сумма частичного массива от 0 до `i` включительно.

Доказательство корректности алгоритмов

Инвариант — важнейшее понятие при доказательстве корректности алгоритмов.
Путь доказательства корректности фрагмента алгоритма:

1. выбираем предикат (или группу предикатов), значение которого истинно до начала исполнения фрагмента.
2. исполняем фрагмент, наблюдая за поведением предиката;
3. если после исполнения предикат остался истинным при любых путях прохождения фрагмента, алгоритм корректен относительно значения этого предиката.

Спасибо за внимание.

Следующая лекция — рекурсия