Лекция 3.

Сложность информационных графов.
Лемма об измеримости функции сложности на запросе.
Понятие функциональной сложности.
Пример задачи информационного поиска,
для которой не существует оптимального графа.

1 Сложность информационных графов

Из определения функционирования ИГ естественным образом вытекает, что каждому ИГ U можно сопоставить следующую процедуру поиска.

Предполагается, что эта процедура хранит в своей (внешней) памяти структуру ИГ U. Входными данными процедуры является запрос. Выходными данными является множество записей.

Пусть на вход процедуры поступил запрос x. Вводим понятие активного множества вершин и вносим в него в начальный момент корень ИГ U и помечаем его. Далее по очереди просматриваем вершины из активного множества и для каждой из них проделываем следующее:

- если рассматриваемая вершина — лист, то запись, приписанную вершине, включаем в ответ;
- если рассматриваемая вершина переключательная, то вычисляем на запросе x переключатель, соответствующий данной вершине, и если конец ребра, исходящего из рассматриваемой вершины, нагрузка которого равна значению переключателя, непомеченная вершина, то помечаем его и включаем в множество активных вершин;
- если рассматриваемая вершина предикатная, то просматриваем по очереди исходящие из нее ребра и вычисляем значения предикатов, приписанных этим ребрам, на запросе x. Концы ребер, которым соответствуют предикаты со значениями, равными 1, если они непомеченные, помечаем и включаем в множество активных вершин;

- исключаем рассматриваемую вершину из активного множества.

Процедура завершается по исчерпании активного множества.

Заметим, что если ИГ решает задачу I, то множество, полученное на выходе процедуры, будет содержать все те и только те записи библиотеки $\langle U \rangle$, которые удовлетворяют запросу x. То есть полученная процедура решает ЗИП $I = \langle X, V, \rho \rangle$, где $V = \langle U \rangle$, и, значит, является алгоритмом поиска.

Таким образом, ИГ как управляющая система может рассматриваться как модель алгоритма поиска, работающего над данными, организованными в структуру, определяемую структурой ИГ.

В данном разделе мы введем понятия сложности ИГ, которые будут характеризовать такие общепринятые меры сложности алгоритмов поиска как объем памяти, время поиска в худшем случае и время поиска в среднем.

Отметим, что в большинстве работ, посвященных исследованию сложности алгоритмов поиска, под сложностью алгоритма понимается время поиска в худшем случае, и в сравнительно редких случаях исследуется среднее время поиска, хотя для задач поиска, используемых в базах данных, для которых характерны массовость и многократность, исследование среднего времени поиска представляется более актуальным. Некоторое объяснение крена в сторону изучения сложности в худшем случае можно найти в цитате из книги Препарата Ф., Шеймоса М. "Вычислительная геометрия: Введение": "К сожалению, анализ поведения в среднем значительно более сложная вещь, чем анализ худшего случая, по двум причинам: во-первых существенные математические трудности возникают, даже если удачно выбрано исходное распределение; во-вторых, часто с трудом достигается согласие в том, что именно выбранное распределение является реальной моделью изучаемой ситуации. Вот почему преобладающее большинство результатов связано с анализом худших случаев."

Определим понятие сложности ИГ на запросе.

Будем считать, что время вычисления любого переключателя из G примерно одинаково и характеризуется числом a, а время вычисления любого предиката из F — числом b.

Пусть нам дан некий ИГ U и произвольно взятый запрос $x \in X$.

Сложностью ИГ U на запросе x назовем число

$$T(U,x) = a \cdot \sum_{\beta \in \mathcal{P}} \varphi_\beta(x) + b \cdot \sum_{\beta \in \mathcal{R} \setminus \mathcal{P}} \psi_\beta \cdot \varphi_\beta(x).$$

Величина $T(U,x)$ характеризует время работы описанной выше процедуры поиска, сопоставленной ИГ U, поскольку $T(U,x)$ равно количеству переключателей, вычисленных данной процедурой при подаче на ее вход запроса x, умноженное на a, плюс количество вычисленных предикатов, умноженное на b.

Сложность ИГ можно вводить двумя способами. Во-первых, как максимальную сложность на запросе

$$\widehat{T}(U) = \max_{x \in X} T(U,x)$$

(здесь мы берем max, а не sup, так как $T(U,x)$ принимает целые значения, и, значит, sup всегда достигается). Эта величина характеризует время поиска в худшем случае соответствующим ИГ алгоритмом и ее будем называть *В-сложностью ИГ* (верхней сложностью). Эта величина исследуется в большинстве работ, посвященных проблемам сложности задач поиска.

Во-вторых, можно вводить сложность ИГ как среднее значение сложности ИГ на запросе, взятое по множеству всех запросов. С этой целью введем *вероятностное пространство* над множеством запросов X, под которым будем понимать тройку $\langle X, \sigma, \mathbf{P} \rangle$, где σ — некоторая *алгебра подмножеств* множества X, $\mathbf{P}$ — *вероятностная мера* на σ, то есть аддитивная мера, такая, что $\mathbf{P}(X) = 1$.

В связи с тем, что мы ввели вероятностное пространство над множеством запросов, уточним понятие типа. А именно, под *типом* будем понимать тройку $S = \langle X, Y, \rho \rangle$, считая, что множество запросов X рассматривается вместе со своим вероятностным пространством $\langle X, \sigma, \mathbf{P} \rangle$. В тех же случаях, когда мы хотим явно выделить рассматриваемое вероятностное пространство над X, мы будем представлять тип пятеркой $S = \langle X, Y, \rho, \sigma, \mathbf{P} \rangle$.

Скажем, что базовое множество $\mathcal{F} = \langle F, G \rangle$ *измеримое*, если алгебра σ содержит все множества N_f, где $f \in F \cup \widehat{G}$.

Справедлива следующая лемма.

Лемма 1. *Если базовое множество $\mathcal{F} = \langle F, G\rangle$ измеримое, то для любого ИГ U над базовым множеством $\mathcal{F}$ функция $T(U, x)$, как функция от x, является случайной величиной.*

Доказательство. Нам необходимо доказать, что для любого ИГ U над базовым множеством $\mathcal{F}$ и любого действительного числа r множество $\{x \in X : T(U, x) < r\} \in \sigma$.

Покажем, что $(\beta \in \mathcal{R}(U)) \to (N_{\varphi_\beta} \in \sigma)$.

Пусть $\beta \in \mathcal{R}(U)$. Пусть $\mathcal{C}_\beta$ — множество всех ориентированных цепей ИГ U, ведущих из корня в вершину β. Пусть C — некоторая цепь, а c — некоторое ребро. Через $\theta(c)$ обозначим проводимость ребра c.

Нетрудно видеть, что

$$\varphi_\beta = \bigvee_{C \in \mathcal{C}_\beta} \bigwedge_{c \in C} \theta(c).$$

Учитывая, что $N_{f \vee g} = N_f \cup N_g$, $N_{f \wedge g} = N_f \cap N_g$, имеем

$$N_{\varphi_\beta} = \bigcup_{C \in \mathcal{C}_\beta} \bigcap_{c \in C} N_{\theta(c)} \in \sigma.$$

Введем следующее обозначение:

$$\mathcal{M}_r = \{\mathcal{B} \subset \mathcal{R}(U) : |\{\mathcal{B} \cap \mathcal{P}\}| + \sum_{\beta \in \mathcal{B} \setminus \mathcal{P}} \psi_\beta < r\}.$$

Тогда, как нетрудно видеть,

$$\{x \in X : T(U, x) < r\} = \bigcup_{\mathcal{B} \in \mathcal{M}_r} ((\bigcap_{\beta \in \mathcal{B}} N_{\varphi_\beta}) \bigcap (\bigcap_{\beta \in \mathcal{R} \setminus \mathcal{B}} (X \setminus N_{\varphi_\beta}))) \in \sigma.$$

Тем самым лемма доказана. □

Далее всюду будем предполагать, что базовое множество измеримо.

Сложностью ИГ U назовем математическое ожидание величины $T(U, x)$, то есть число

$$T(U) = \mathbf{M}_x\, T(U, x).$$

Если (β, α) — ребро ИГ, то *сложностью этого ребра* назовем число

- $b \cdot \mathbf{P}(N_{\varphi_\beta})$ — если (β, α) — предикатное ребро;

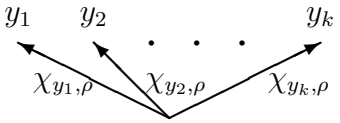

Рис. 1: Информационный граф переборного алгоритма

- $a \cdot \mathbf{P}(N_{\varphi_\beta})/\psi_\beta$ — если это ребро переключательное.

Если β — вершина ИГ, то число $\mathbf{P}(N_{\varphi_\beta})$ назовем *сложностью вершины* β.

Нетрудно показать, что сложность ИГ равна сумме сложностей ребер ИГ. В самом деле

$$\begin{aligned} T(U) &= \mathbf{M}_x\, T(U,x) = \int_X T(U,x)\mathbf{P}(dx) = \\ &= \int_X (b \cdot \sum_{\beta\in\mathcal{R}\setminus\mathcal{P}} \psi_\beta \cdot \varphi_\beta(x) + a \cdot \sum_{\beta\in\mathcal{P}} \varphi_\beta(x))\mathbf{P}(dx) = \\ &= b \cdot \sum_{\beta\in\mathcal{R}\setminus\mathcal{P}} \psi_\beta \int_X \varphi_\beta(x)\mathbf{P}(dx) + a \cdot \sum_{\beta\in\mathcal{P}} \int_X \varphi_\beta(x)\mathbf{P}(dx) = \\ &= b \cdot \sum_{\beta\in\mathcal{R}\setminus\mathcal{P}} \psi_\beta \mathbf{P}(N_{\varphi_\beta}) + a \cdot \sum_{\beta\in\mathcal{P}} \mathbf{P}(N_{\varphi_\beta}). \end{aligned}$$

Далее всюду будем предполагать, что $a = b = 1$.

Пусть нам дан ИГ U.

Объемом $Q(U)$ ИГ U назовем число ребер в ИГ U.

В качестве примера мы можем подсчитать сложность ИГ U, изображенного на рисунке 1. Легко видеть, что $Q(U) = k$ и $T(U) = k$, то есть объем графа минимально возможный, а время максимальное. Это и не удивительно, так как ИГ U соответствует переборному алгоритму поиска.

Пусть нам дана некая ЗИП I. *Сложностью задачи* I *при базовом множестве* $\mathcal{F}$ *и заданном объеме* q назовем число

$$T(I,\mathcal{F},q) = \inf\{T(U) : U \in \mathcal{U}(I,\mathcal{F}) \text{ и } Q(U) \le q\},$$

где $\mathcal{U}(I,\mathcal{F})$ — множество всех ИГ над базовым множеством $\mathcal{F}$, решающих ЗИП I.

Соответственно *B-сложностью задачи I при базовом множестве $\mathcal{F}$ и заданном объеме q* назовем число

$$\widehat{T}(I,\mathcal{F},q)=\min\{\widehat{T}(U):U\in\mathcal{U}(I,\mathcal{F})\text{ и }Q(U)\le q\}$$

(здесь мы берем min, а не inf, так как $\widehat{T}(U)$ принимает целые значения, и, значит, inf всегда достигается).

Число

$$T(I,\mathcal{F})=\inf\{T(U):U\in\mathcal{U}(I,\mathcal{F})\}$$

назовем *сложностью задачи I при базовом множестве $\mathcal{F}$.*

Соответственно *B-сложностью задачи I при базовом множестве $\mathcal{F}$* назовем число

$$\widehat{T}(I,\mathcal{F})=\min\{\widehat{T}(U):U\in\mathcal{U}(I,\mathcal{F})\}.$$

Если k — натуральное число, S — тип задач поиска, то обозначим

$$\mathcal{I}(k,S)=\{I=\langle X,V,\rho\rangle\in S:|V|=k\}.$$

Будем исследовать функции, характеризующие сложность класса ЗИП $\mathcal{I}(k,S)$, такие как функции Шеннона:

$$\widehat{\mathcal{T}}(k,S,\mathcal{F})=\max_{I\in\mathcal{I}(k,S)}\widehat{T}(I,\mathcal{F}),$$

$$\mathcal{T}(k,S,\mathcal{F})=\sup_{I\in\mathcal{I}(k,S)}T(I,\mathcal{F}),$$

(в первом случае мы берем max, а не sup, так как $\widehat{T}(I,\mathcal{F})$ принимает целые значения, и, значит, sup всегда достигается).

Если существует такой ИГ $U\in\mathcal{U}(I,\mathcal{F})$, что $T(U)=T(I,\mathcal{F})$, то ИГ U будем называть *оптимальным* для ЗИП I. Соответственно если $\widehat{T}(U)=\widehat{T}(I,\mathcal{F})$, то ИГ U будем называть *B-оптимальным* для ЗИП I.

Можно привести **пример такой ЗИП и такого базового множества, для которых не существует оптимального ИГ**.

Пусть $X=Y=[0,1]$ и на X задана равномерная вероятностная мера. Пусть отношение поиска есть отношение "$\ge$" для действительных чисел. Пусть библиотека состоит из одной записи $V=\{3/4\}$. Пусть базовое множество имеет вид $\mathcal{F}=\langle F,\emptyset\rangle$, где

$$F=\{f^1\}\cup\{f^2_a:a\in(0,1/2)\},$$

$$f^1(x) = \begin{cases} 0, & \text{если } x \in [0, 1/2] \\ 1, & \text{если } x \in (1/2, 1] \end{cases},$$

$$f_a^2(x) = \begin{cases} 0, & \text{если } x \in (a, 3/4) \\ 1, & \text{если } x \in [0, a] \cup [3/4, 1] \end{cases}.$$

Рассмотрим ЗИП $I = \langle X, V, \geq \rangle$. Поскольку

$$\chi_{3/4,\geq}(x) = \begin{cases} 0, & \text{если } x \in [0, 3/4) \\ 1, & \text{если } x \in [3/4, 1] \end{cases},$$

то $\chi_{3/4,\geq} = f^1 \& f_a^2$, для любого $a \in (0, 1/2)$. Рассмотрим ИГ U_a, состоящий из двух последовательно соединенных ребер, начало первого из которых есть корень ИГ, а конец второго — лист, которому приписана запись $3/4$, первому ребру соответствует предикат f_a^2, а второму — f^1. Нетрудно видеть, что $T(U_a) = 1 + a + 1/4 = 5/4 + a$. Очевидно, что $T(I, \mathcal{F}) = \inf\{T(U_a) : a \in (0, 1/2)\} = 5/4$, но не существует ИГ, чья сложность равна $5/4$.

Упражнения

1. Пусть $X = \{1, 2, ..., N\}$, $S = \langle X, X, =, \mathbf{P}, \sigma \rangle$ — тип поиска идентичных объектов, где $\sigma = 2^X$, $\mathbf{P}$ — равномерная вероятностная мера, то есть для любого $x \in X$ выполняется $\mathbf{P}(x) = 1/N$.

1. Посчитайте сложность, B-сложность и объем информационного графа, полученного при решении упражнения 1 из лекции 8.

2. Посчитайте сложность, B-сложность и объем информационного графа, полученного при решении упражнения 2 из лекции 8.

3. Посчитайте сложность, B-сложность и объем информационного графа, полученного при решении упражнения 3 из лекции 8. Для базового множества и ЗИП, приведенных в упражнении 3 из лекции 8, постройте информационный граф со сложностью, не большей, чем 1.48.

4. Посчитайте сложность, B-сложность и объем информационного графа, полученного при решении упражнения 4 из лекции 8, если $N = 100$. Для какого значения параметра m сложность будет минимальна. Для какого значения параметра m B-сложность будет минимальна. Для какого значения параметра m объем будет минимальным.

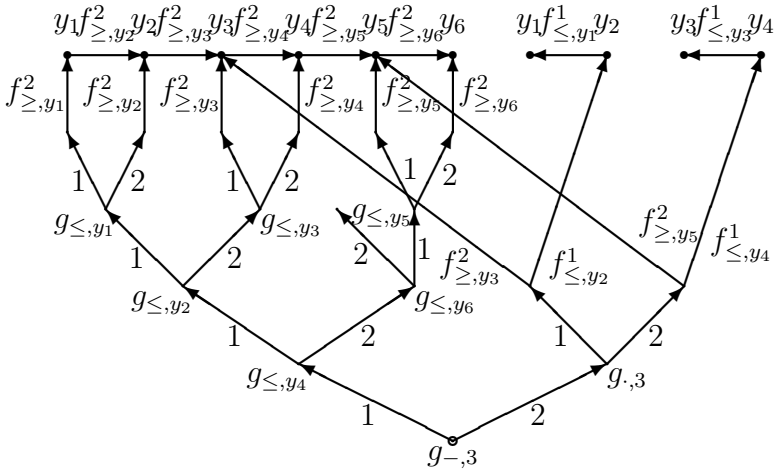

Рис. 2: Решение одномерной задачи интервального поиска

2. Если $X = \{1, 2, ..., N\}$ и на X задана равномерная вероятностная мера, то посчитайте сложность, B-сложность и объем информационного графа, полученного при решении упражнения 5 из лекции 8.

3. Пусть на множестве запросов $X = [0, 1]$ задана равномерная вероятностная мера. Посчитайте сложность, B-сложность и объем информационного графа, полученного при решении упражнения 6 из лекции 8.

4. Пусть на множестве запросов $X_{int} = \{(u, v) : 0 \leq u \leq v \leq 1\}$ задана равномерная вероятностная мера. Посчитайте сложность, B-сложность и объем информационного графа, изображенного на рисунке 2.