Лекция 4.

Теорема о существовании оптимальных информационных графов. Мощностная нижняя оценка сложности задач информационного поиска. Случай оптимальности переборного алгоритма поиска.

1 Теорема о существовании оптимальных информационных графов

Скажем, что некая вершина ИГ *достижима из корня на запросе* $x \in X$ или просто *достижимая на запросе* $x \in X$, если функция фильтра этой вершины на запросе x равна 1.

Скажем, что некая вершина ИГ *достижима из корня* или просто *достижимая*, если функция фильтра этой вершины не равна тождественному нулю, в противном случае вершину называем *недостижимой*. Скажем, что ребро ИГ *несущественное*, если выполняется хотя бы одно из следующих условий

- ребро исходит из недостижимой вершины,
- ребро является предикатным и входит в корень или в недостижимую вершину,
- ребро является предикатным и не принадлежит ни одной цепи, ведущей из корня в какой либо лист,
- ребро является переключательным, и начало этого ребра не принадлежит ни одной цепи, ведущей из корня в какой либо лист,

- ребро является переключательным, и число, приписанное этому ребру, больше максимально возможного значения переключателя, соответствующего началу этого ребра,

- начало и конец ребра совпадают.

В противном случае ребро называем *существенным*.

Легко заметить, что удаление несущественных ребер из ИГ не изменяет функционирования ИГ и не увеличивает его сложность. Поэтому всегда в дальнейшем мы будем рассматривать ИГ с точностью до несущественных ребер, а точнее будем считать, что все ребра в ИГ — существенные.

Теорема 1 (о существовании оптимальных графов). *Если множество запросов X конечно, то для любой ЗИП $I = \langle X, Y, \rho\rangle$ и любого базового множества $\mathcal{F} = \langle F, G\rangle$ такого, что $\mathcal{U}(I, \mathcal{F}) \neq \emptyset$, существует оптимальный ИГ над базовым множеством $\mathcal{F}$.*

Доказательство. Заметим, что из за конечности множества X множества F и G конечны. В самом деле, если $|X| = m$, то число предикатов, определенных на X не больше, чем 2^m. Следовательно, $|F| \leq 2^m$. Так как область значений любого переключателя есть начальный отрезок натурального ряда, то любой переключатель над X принимает не более m значений, следовательно $|G| < m^m$.

Для произвольного ИГ U обозначим через $N(U)$ подграф графа U, состоящий из ребер, имеющих ненулевую сложность.

Возьмем произвольный ИГ $U_0 \in \mathcal{U}(I, \mathcal{F})$. Обозначим $\mathcal{U}' = \{U \in \mathcal{U}(I, \mathcal{F}) : T(U) \leq T(U_0)\}$, $\mathcal{N}' = \{N(U) : U \in \mathcal{U}'\}$. Очевидно, что

$$T(I, \mathcal{F}) = \inf_{U \in \mathcal{U}(I, \mathcal{F})} T(U) = \inf_{U \in \mathcal{U}'} T(U) = \inf_{U \in \mathcal{N}'} T(U).$$

Для доказательства существования оптимального ИГ для ЗИП I нам достаточно показать конечность множества $\mathcal{N}'$.

Пусть $|F \cup \widehat{G}| = n$. Пусть M — множество, состоящее из тождественно нулевого предиката и всех предикатов, полученных из предикатов множества $F \cup \widehat{G}$ с помощью операций конъюнкции и дизъюнкции. Понятно, что $|M| \leq \min(2^m, 2^{2^n})$. Обозначим $M' = \{f \in M : \mathbf{P}(N_f) > 0\}$. Пусть $\min\limits_{f \in M'} \mathbf{P}(N_f) = r$. По определению $r > 0$. Поскольку для любого ИГ над $\mathcal{F}$ функции фильтров вершин принадлежат множеству M, то сложность

любого предикатного ребра ИГ над $\mathcal{F}$ либо нулевая, либо не меньше чем r, а сложность любого переключательного ребра с ненулевой сложностью не меньше чем r/m. Отсюда в любом ИГ из множества $\mathcal{N}'$ число ребер не больше чем $T(U_0) \cdot m/r$.

Так как из конечности множеств F и G следует конечность числа различных нагрузок ИГ, то, значит, множество $\mathcal{N}'$ конечно.

Что и доказывает теорему. □

2 Мощностная нижняя оценка

В книгу Препарата Ф., Шеймоса М. "Вычислительная геометрия: Введение"на стр. 92 бездоказательно, просто как очевидный факт утверждается, что время поиска по крайней мере не меньше чем время необходимое на перечисление ответа. В нашей модели этот факт находит свое доказательство и носит название мощностной нижней оценки. В связи повсеместностью применимости мощностной нижней оценки часто при оценке времени алгоритма поиска оценивают только разность между временем поиска и временем перечисления ответа.

Пусть нам даны произвольные множества запросов X, записей Y и отношение поиска ρ на $X \times Y$. Причем на множестве запросов задано вероятностное пространство $\langle X, \sigma, \mathbf{P} \rangle$.

Скажем, что базовое множество $\mathcal{F}$ *допустимо для ЗИП* I, если существует ИГ над базовым множеством $\mathcal{F}$, который решает ЗИП I.

Следующий результат, называемый мощностной нижней оценкой, справедлив для любой ЗИП при минимальных ограничениях. Смысл этого результата заключается в том, что время поиска не может быть меньше чем время, необходимое на перечисление ответа.

Справедлива следующая теорема.

Теорема 2 (мощностная нижняя оценка). *Пусть* $I = \langle X, V, \rho \rangle$ — *произвольная ЗИП, такая, что существует такая запись* $y \in V$, *что* $O(y, \rho) \neq \emptyset$, $\mathcal{F}$ — *измеримое базовое множество, допустимое для* I, *тогда*

$$T(I, \mathcal{F}) \geq \max(1, \sum_{y \in V} \mathbf{P}(O(y, \rho))),$$

$$\widehat{T}(I, \mathcal{F}) \geq \max_{x \in X} |\mathcal{J}_I(x)|.$$

Доказательство. Возьмем произвольный ИГ U, решающий задачу I. Такой граф существует, так как $\mathcal{U}(I, \mathcal{F}) \neq \emptyset$.

Возьмем произвольный запрос $x \in X$. Так как ИГ U решает ЗИП I, то ответ на запрос x

$$\mathcal{J}_U(x) = \mathcal{J}_I(x) = \{y \in V : x\rho y\}.$$

Возьмем произвольную запись $y \in \mathcal{J}_U(x)$. Поскольку запись y попала в ответ, то, значит, в ИГ U существует некий лист α, которому приписана запись y и такой, что $\varphi_\alpha(x) = 1$. А так как $\varphi_\alpha(x) = 1$ и никакой лист не совпадает с корнем, то существует цепь, ведущая из корня в лист α, проводимость которой равна 1, и в этой цепи есть ребро, ведущее в α, с проводимостью 1. Это ребро назовем проводящим ребром записи y. Понятно, что разным записям из $\mathcal{J}_U(x)$ соответствуют разные проводящие ребра, так как эти ребра ведут в разные листья. Если проводящее ребро записи предикатное, предикат, приписанный проводящему ребру, обязательно был вычислен перед тем, как мы попали в лист. Если проводящее ребро записи переключательное, то обязательно был вычислен переключатель, приписанный вершине, из которой исходит проводящее ребро. Причем такие переключатели для разных записей из $\mathcal{J}_U(x)$ будут разными, так как только одно из переключательных ребер, исходящих из одной вершины, может иметь проводимость, равную 1. Таким образом, каждой записи из $\mathcal{J}_U(x)$ можно сопоставить переключатель или предикат, вычисляемый непосредственно перед попаданием в соответствующий записи лист. Причем разным записям будут сопоставлены разные переключатели или предикаты. Отсюда следует, что

$$T(U, x) \geq |\mathcal{J}_I(x)|.$$

Следовательно,

$$\begin{aligned}
\widehat{T}(U) &= \max_{x \in X} T(U, x) \geq \max_{x \in X} |\mathcal{J}_I(x)|, \\
T(U) &= \mathbf{M}_x\, T(U, x) \geq \mathbf{M}_x\, |\mathcal{J}_I(x)| = \\
&= \int_X |\mathcal{J}_I(x)| \mathbf{P}(dx) = \int_X |\{y \in V : x\rho y\}| \mathbf{P}(dx) = \\
&= \sum_{y \in V} \int_{O(y,\rho)} \mathbf{P}(dx) = \sum_{y \in V} \mathbf{P}(O(y, \rho)).
\end{aligned}$$

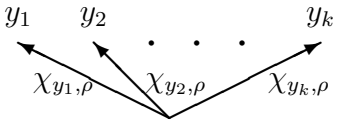

Рис. 1: Информационный граф переборного алгоритма

А так как это неравенство выполняется для любого графа $U \in \mathcal{U}(I, \mathcal{F})$, то

$$\widehat{T}(I, \mathcal{F}) \geq \max_{x \in X} |\mathcal{J}_I(x)|,$$

$$T(I, \mathcal{F}) \geq \sum_{y \in V} \mathbf{P}(O(y, \rho)),$$

Так как в библиотеке V существует такая запись y, что $O(y, \rho) \neq \emptyset$, то в любом ИГ, решающем ЗИП I, существует хотя бы одна цепь, и соответственно хотя бы одно ребро исходит из корня. Следовательно, $T(I, \mathcal{F}) \geq 1$.

Тем самым теорема доказана. □

3 Случай оптимальности перебора

Теорема 1. *Если $I = \langle X, V, \rho \rangle$ — ЗИП и $\mathcal{F} = \langle F, \emptyset \rangle$ — измеримое базовое множество, такие, что для любой записи $y \in V$ выполняется $\chi_{y,\rho} \in F$ и $O(y, \rho) \setminus (\bigcup_{f \in F \setminus \chi_{y,\rho}} N_f) \neq \emptyset$, то*

$$\widehat{T}(I, \mathcal{F}) = T(I, \mathcal{F}) = |V|.$$

Доказательство. Пусть $V = \{y_1, y_2, \ldots, y_k\}$. Каждому $i \in \{\overline{1,k}\}$ сопоставим некоторый запрос $x_i \in O(y, \rho) \setminus (\bigcup_{f \in F \setminus \chi_{y,\rho}} N_f)$. Легко видеть, что для любых таких $i, j \in \{\overline{1,k}\}$, что $i \neq j$, справедливо $x_i \neq x_j$.

Понятно, что $U \in \mathcal{U}(I, \mathcal{F}) \neq \emptyset$, так как ИГ, изображенный на рисунке 1, соответствующий алгоритму перебора, решает ЗИП I. Обозначим этот ИГ через U_0.

Возьмем произвольный ИГ $U \in \mathcal{U}(I, \mathcal{F})$. Так как U решает ЗИП I, то для каждого $i \in \{\overline{1,k}\}$ существуют лист α_i, которому приписана запись y_i, и цепочка ребер C_i, ведущая из корня в лист α_i, которая проводит

запрос x_i. Нетрудно заметить, что для любых таких $i, j \in \{\overline{1,k}\}$, что $i \neq j$, цепочки C_i и C_j не пересекаются по ребрам, поскольку, если бы существовало ребро, принадлежащее и C_i, и C_j, то нагрузка этого ребра принимала бы значение 1 одновременно и на x_i, и на x_j, но по определению x_i и x_j таких функций в множестве F не существует. Отсюда сразу следует, что из корня ИГ U исходит по крайней мере k ребер, и, следовательно, для любого $x \in X$ $T(U,x) \geq k$, откуда $T(U) \geq k$ и $\widehat{T}(U) \geq k$, а в силу произвольности ИГ U $T(I,\mathcal{F}) \geq k$ и $\widehat{T}(I,\mathcal{F}) \geq k$. Но так как для ИГ U_0 справедливо $T(U_0) = \widehat{T}(U_0) = k$, то $T(I,\mathcal{F}) = \widehat{T}(I,\mathcal{F}) = k$, что и требовалось доказать.

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 $\mathbf{P}(x) = 1/N$, $V = \{3, 5, 7, 11, 13, 17, 19\}$. Приведите мощностную нижнюю оценку для ЗИП $I = \langle X, V, =\rangle$.

2. Пусть $S_{dom1} = \langle [0,1], [0,1], \geq, \mathbf{P}, \sigma\rangle$ — тип одномерной задачи о доминировании, где $\mathbf{P}$ — равномерная вероятностная мера на $[0,1]$, $V = \{y_1, y_2, \ldots, y_k\} \subseteq [0,1]$. Приведите мощностную нижнюю оценку для ЗИП $I = \langle [0,1], V, \geq\rangle$.

3. Пусть $S_{int} = \langle X_{int}, Y_{int}, \rho_{int}\rangle$ — тип одномерного интервального поиска и на множестве запросов $X_{int} = \{(u,v) : 0 \leq u \leq v \leq 1\}$ задана равномерная вероятностная мера. $V = \{y_1, y_2, \ldots, y_k\} \subseteq [0,1]$. Приведите мощностную нижнюю оценку для ЗИП $I = \langle X_{int}, V, \rho_{int}\rangle$. Оцените сверху полученную величину.