

Поиск представителя в задаче о метрической близости

А. П. Пивоваров

Работа посвящена описанию алгоритма поиска представителя в задаче о метрической близости. Пусть $R \in (0, 2)$, $X = [0, 1]^2$ — множество запросов, $Y = [0, 1]^2$ — множество записей, V — конечное подмножество Y , ρ — отношение поиска на $X \times Y$, такое что $x \rho y \Leftrightarrow \max(|x_1 - y_1|, |x_2 - y_2|) \leq R/2$, где $x = (x_1, x_2) \in X$ и $y = (y_1, y_2) \in Y$. Поиск представителя в задаче о метрической близости состоит том, чтобы для любого $x \in X$ находить и выдавать в качестве ответа любую запись $y \in V$, такую что $x \rho y$, либо указывать, что таких записей нет. Для поиска представителя в задаче о метрической близости разработан алгоритм, объем памяти которого линеен по числу записей в библиотеке $k = |V|$, время поиска в среднем константа, а в худшем случае порядка $\log k$.

Введение

Для оценки алгоритмов поиска в данной работе рассматривается информационно-графовая модель формализации понятия алгоритма поиска.

В данной работе исследуется следующая задача. Дано конечное множество точек из квадрата $[0, 1]^2$ (это множество называют библиотекой). Запрос на поиск представляет собой некоторую точку квадрата $x \in [0, 1]^2$. Надо перечислить все точки библиотеки, отстоящие от x не более чем на некоторую заранее заданную величину по каждой из двух координат. Это так называемая задача о метрической близости в ее стандартной постановке.

В такой постановке задача о метрической близости может быть решена за константное в среднем время (без перечисления ответа) при затратах памяти порядка $k^{1+\varepsilon}$ (k — число записей в библиотеке), где константа времени растет обратно пропорционально ε (алгоритм с такими характеристиками описан в [1]).

Поиск представителя представляет собой модификацию задачи информационного поиска, состоящую в том, что требуется найти не все записи из библиотеки, удовлетворяющие запросу, а любую одну из таких записей, если они есть, либо выдать в качестве ответа пустое множество, если таких записей в библиотеке нет. Фактически из всего ответа задачи в ее стандартной постановке достаточно выдать только одну запись (ее как раз и называют представителем).

В данной работе рассматривается поиск представителя в задаче о метрической близости. Основное утверждение состоит в том, что существует алгоритм поиска представителя в задаче о метрической близости, время работы которого в среднем константа, а затраты памяти имеют порядок k , где k — число записей в библиотеке.

Автор выражает глубокую благодарность профессору Э. Э. Гасанову за постановку задачи и помощь в работе.

1. Основные понятия и формулировка результатов

Для описания алгоритма поиска воспользуемся информационно-графовой моделью поиска, в которой алгоритм описывается с помощью структуры, называемой *информационным графом* (ИГ). Следуя [2], дадим определение понятию ИГ.

В формальном определении понятия ИГ используются 4 множества:

- множество запросов X ;
- множество записей Y ;
- множество F *одноместных предикатов*, заданных на множестве X ;

- множество G *одноместных переключателей*, заданных на множестве X (*переключатели* — это функции, область значений которых является начальным отрезком натурального ряда).

Пару $\mathcal{F} = \langle F, G \rangle$ будем называть *базовым множеством*.

Определение понятия ИГ разбивается на два шага. На первом шаге раскрывается структурная (схемная) часть этого понятия, на втором — функциональная.

Определение ИГ с точки зрения его структуры.

Пусть нам дана ориентированная многополюсная сеть.

Выделим в ней один полюс и назовем его *корнем*, а остальные полюса назовем *листьями*.

Выделим в сети некоторые вершины и назовем их *точками переключения* (полюса могут быть точками переключения).

Если β — вершина сети, то через ψ_β обозначим *полустепень исхода* вершины β .

Каждой точке переключения β сопоставим некий символ из G . Это соответствие назовем *нагрузкой точек переключения*.

Для каждой точки переключения β ребрам, из нее исходящим, поставим во взаимно однозначное соответствие числа из множества $\{1, \psi_\beta\}$. Эти ребра назовем *переключательными*, а это соответствие — *нагрузкой переключательных ребер*.

Ребра, не являющиеся переключательными, назовем *предикатными*.

Каждому предикатному ребру сети сопоставим некоторый символ из множества F . Это соответствие назовем *нагрузкой предикатных ребер*.

Сопоставим каждому листу сети некоторую запись из множества Y . Это соответствие назовем *нагрузкой листьев*.

Полученную нагруженную сеть назовем *информационным графом* над базовым множеством $\mathcal{F} = \langle F, G \rangle$.

Определение функционирования ИГ.

Скажем, что предикатное ребро проводит запрос $x \in X$, если предикат, приписанный этому ребру, принимает значение 1 на запросе x ;

переключательное ребро, которому приписан номер n , проводит запрос $x \in X$, если переключатель, приписанный началу этого ребра, принимает значение n на запросе x ; ориентированная цепочка ребер проводит запрос $x \in X$, если каждое ребро цепочки проводит запрос x ; запрос $x \in X$ проходит в вершину β ИГ, если существует ориентированная цепочка, ведущая из корня в вершину β , которая проводит запрос x ; запись y , приписанная листу α , попадает в ответ ИГ на запрос $x \in X$, если запрос x проходит в лист α . *Ответом ИГ U на запрос x* назовем множество записей, попавших в ответ ИГ на запрос x , и обозначим его $\mathcal{J}_U(x)$. Эту функцию $\mathcal{J}_U(x) : X \rightarrow 2^Y$ будем считать результатом функционирования ИГ U и называть *функцией ответа ИГ U* .

Объемом $Q(U)$ ИГ U назовем число ребер в ИГ U . Эта величина характеризует объем памяти, требуемый для реализации алгоритма поиска, задаваемого ИГ U .

Сложностью ИГ U на запросе x назовем число $T(U, x)$, равное сумме числа переключателей и предикатов, вычисленных в процессе обработки запроса x . Эту величину также называют временем работы U на запросе x . *Верхней сложностью* (или временем работы в худшем случае) называют величину $\hat{T}(U) = \max_{x \in X} T(U, x)$. В случае, когда на X введено вероятностное пространство, а переключатели и предикаты являются измеримыми функциями, в [2] показано, что $T(U, x)$ является случайной величиной относительно x и можно рассматривать $T(U) = \mathbf{M}_x T(U, x)$ — *сложность ИГ U в среднем* (или время работы U в среднем).

Понятие ИГ полностью определено.

Постановка задачи.

Пусть $S = \langle X, Y, \rho \rangle$ — тип задач информационного поиска. $I = \langle X, V, \rho \rangle$ — конкретная ЗИП. $\mathcal{J}_I(x) : X \rightarrow 2^Y$ — функция ответа на запрос: $\mathcal{J}_I(x) = \{y \in V \mid x \rho y\}$. Будем говорить, что функция $\mathcal{J}(x) : X \rightarrow 2^Y$ является *функцией предъявления представителя* для задачи I , если для любого запроса $x \in X$ выполнены следующие условия:

$$\begin{aligned} \mathcal{J}(x) &\subseteq \mathcal{J}_I(x), \\ \mathcal{J}_I(x) \neq \emptyset &\Rightarrow \mathcal{J}(x) \neq \emptyset. \end{aligned}$$

Будем говорить, что информационный граф U над некоторым базовым множеством \mathcal{F} решает задачу поиска представителя для ЗИП I , если его функция ответа $\mathcal{J}_U(x)$ является функцией предъявления представителя для задачи I .

Рассмотрим тип задач поиска, описывающий задачу о метрической близости: $S = \langle [0, 1]^2, [0, 1]^2, \rho_{\square}^R \rangle$, где $x \rho_{\square}^R y \Leftrightarrow \rho_{\square}(x, y) = \max(|x_1 - y_1|, |x_2 - y_2|) \leq R/2$, $R \in (0, 2)$.

Будем рассматривать базовое множество $\mathcal{F} = \langle F, G \rangle$, где

$$\begin{aligned} F &= F_0 \cup F_1, \\ F_0 &= \{f^{id}(x) \equiv 1\}, \\ F_1 &= \{\chi_{y, \rho_{\square}^R}(x) \mid y \in Y\}, \\ G &= G_1 \cup G_2, \\ G_1 &= \{g_m(x) = m(i(x_1) - 1) + i(x_2) \mid m \in \mathbb{N}\}, \text{ где } i(x) = \max(1, \lceil mx \rceil), \\ G_2 &= \left\{ g_{\leq, a}^i(x) = \begin{cases} 1, & \text{при } x_i \leq a, \\ 2, & \text{при } x_i > a \end{cases}, \right. \\ &\quad \left. g_{\geq, a}^i(x) = \begin{cases} 1, & \text{при } x_i \geq a, \\ 2, & \text{при } x_i < a \end{cases}, \text{ где } i \in \{1, 2\}, a \in \mathbb{R} \right\}. \end{aligned}$$

Теорема 1. Пусть дана задача $I = \langle [0, 1]^2, V, \rho_{\square}^R \rangle$. Если вероятностная мера на X задается ограниченной функцией плотности вероятности $p(x) \leq c$, то для любого $m \in \mathbb{N}$: $1/m < R/2$ существует ИГ U над базовым множеством \mathcal{F} , решающий задачу поиска представителя для I , удовлетворяющий условиям:

$$\begin{aligned} T(U) &\leq 1 + 4c \frac{(mR - 1)(2 + R)^2}{(mR - 2)^2} (2 + \lceil \log_2 |V| \rceil), \\ \hat{T}(U) &\leq 9 + 4 \lceil \log_2 |V| \rceil, \\ Q(U) &\leq 4k + m^2 \left(1 + 60 \frac{(mR - 1)(2 + R)^2}{(mR - 2)^2} \right). \end{aligned}$$

Следствие 1. Если последовательность R_k удовлетворяет условию $\frac{1}{R_k} = o\left(\frac{\sqrt{k}}{\log_2 k}\right)$, то для последовательности задач $I_k = \langle [0, 1]^2, V, \rho_{\square}^{R_k} \rangle$,

где $|V| = k$, существует последовательность ИГ U_k , решающих соответствующие задачи поиска представителя, такие, что выполнено:

$$\begin{aligned} Q(U_k) &\lesssim 4k, \\ T(U_k) &\sim 1, \\ \hat{T}(U_k) &\leq 9 + 4 \lceil \log_2 k \rceil. \end{aligned}$$

2. Вспомогательные утверждения

Лемма 1. Если фигура $\Phi = \bigcup_{i=1}^n \Pi_i$ является конечным объединением прямоугольников Π_i , стороны которых параллельны координатным осям, и таких, что каждая из сторон любого прямоугольника не меньше h , где $h \in \mathbb{R}, h > 0$, периметр Φ равен L , а площадь S , то выполнено неравенство $Lh \leq 4S$.

Доказательство. Границу Φ можно разбить на конечное число отрезков. $\partial\Phi = \bigcup_{j=1}^p l_j$, где каждый отрезок l_j либо вертикальный, либо горизонтальный и $\sum_{j=1}^p |l_j| = L$. Для каждого отрезка границы l_j построим прямоугольник T_j следующим образом: одной из сторон T_j возьмем l_j , ширина T_j равна h , и расположен T_j по ту же сторону от l_j , что и Φ . Тогда так как $\Phi = \bigcup_{i=1}^n \Pi_i$, то каждая точка отрезка l_j является точкой границы одного из прямоугольников Π_i , а раз так, то и соответствующая этой точке полоса прямоугольника T_j тоже принадлежит Π_i . Получаем $T_j \subseteq \bigcup_{i=1}^n \Pi_i = \Phi$. При этом каждая точка $x \in \Phi$ может принадлежать не более чем четырем разным прямоугольникам T_j . Получаем $\sum_{j=1}^p S(T_j) \leq 4S$. С другой стороны, $\sum_{j=1}^p S(T_j) = \sum_{j=1}^p h|l_j| = h \sum_{j=1}^p |l_j| = Lh$, откуда $Lh = \sum_{j=1}^p S(T_j) \leq 4S$. Лемма доказана.

$P \subseteq \mathbb{Z}^2$ будем называть *фигурой* в \mathbb{Z}^2 . Если $|P| < \infty$, то фигуру P будем называть конечной. Для любой точки $(i, j) \in \mathbb{Z}^2$ введем $N_{i,j} = \{(i-1, j), (i+1, j), (i, j-1), (i, j+1)\}$ — множество соседей точки (i, j) . Путем от точки $a \in \mathbb{Z}^2$ до точки $b \in \mathbb{Z}^2$ будем называть конечную последовательность точек $a = x_0, \dots, x_N = b$, такую, что x_{k-1} и x_k — соседние для всех $k \in \{1, \dots, N\}$. Фигуру P будем называть *связной*, если для всех $a, b \in P, a \neq b$ существует путь от a до b , каждая точка которого принадлежит P . Формально фигура, не содержащая точек или содержащая одну точку является связной. Если P — фигура в \mathbb{Z}^2 , то введем функции $t_P(i, j), q_P(i, j)$ следующим образом:

$$\begin{aligned} t_P(i, j) &= |N_{i,j} \cap P|, \\ q_P(i, j) &= |N_{i,j} \setminus P|. \end{aligned}$$

Очевидно, что для всех $(i, j) \in \mathbb{Z}^2$ справедливо $t_P(i, j) + q_P(i, j) = 4$. $t_P(i, j)$ это количество соседей (i, j) , лежащих в множестве P , аналогично $q_P(i, j)$ это количество соседей (i, j) , лежащих вне множества P . Определим для фигуры P периметр $l(P)$ и площадь $S(P)$:

$$\begin{aligned} l(P) &= \sum_{(i,j) \in P} q_P(i, j), \\ S(P) &= |P|. \end{aligned}$$

Такое определение периметра и площади соответствует отображению \mathbb{Z}^2 на \mathbb{R}^2 , при котором точке сопоставляется квадрат размера 1. Обозначим $(i, j)^* = \{(i', j') \in \mathbb{Z}^2: |i - i'| \leq 1 \text{ и } |j - j'| \leq 1\}$. Аналогично, для фигуры P введем $P^* = \bigcup_{(i,j) \in P} (i, j)^*$. Очевидно, для любой фигуры P справедливо $P \subseteq P^*$.

Лемма 2. Для любой конечной связной фигуры P справедливо $|P^*| \leq |P| + l(P) + 4$.

Доказательство. Будем доказывать утверждение леммы индукцией по $|P|$.

$|P| = 0$. Тогда $P = \emptyset$. Тогда $|P^*| = |P| = l(P) = 0$ и утверждение выполнено.

$|P| = 1$. Тогда P состоит из одной точки a , $|P^*| = 9$, $l(P) = 4$ и в неравенстве, составляющем утверждение леммы достигается равенство.

Пусть утверждение леммы доказано для любого $P \subseteq \mathbb{Z}^2$, такого что $|P| \leq n$. Докажем утверждение для P : $|P| = n + 1$.

Так как $|P| > 1$, и P — связно, то для всех $(i, j) \in P$ выполнено $t_P(i, j) \geq 1$.

Если существует $a = (i, j) \in P$, такое что $t_P(i, j) = 1$, то рассмотрим $P' = P \setminus \{a\}$. Пусть $l' = l(P')$, а $l = l(P)$ соответственно. Для P' справедливы соотношения:

$$\begin{aligned} |P'| &= n, \\ l' &= l - 2, \\ |P'^*| &\leq |P^*| \leq |P'^*| + 3. \end{aligned}$$

Кроме того, P' — связно и по предположению индукции имеем $|P'^*| \leq |P'| + l' + 4$. Итого получаем цепочку неравенств

$$|P^*| \leq |P'^*| + 3 \leq |P'| + l' + 7 = n + l - 2 + 7 = n + 1 + l + 4 = |P| + l + 4,$$

и утверждение леммы для рассматриваемого случая доказано.

Если же для всех $a = (i, j) \in P$ выполнено $t_P(i, j) \geq 2$, то рассмотрим $i_0 = \max\{i \mid \exists j \in \mathbb{Z}: (i, j) \in P\}$ и из множества $\{a \in P \mid a = (i_0, j)\}$ выберем элемент $a_0 = (i_0, j_0)$ с наибольшей второй координатой. Тогда точки $(i_0 + 1, j_0)$ и $(i_0, j_0 + 1)$ уже не принадлежат P . Поэтому $t_P(i_0, j_0) \leq 2$. Но мы имеем для всех точек P обратное неравенство $t_P(i, j) \geq 2$. Следовательно, $t_P(i_0, j_0) = 2$ и точки $(i_0 - 1, j_0)$ и $(i_0, j_0 - 1)$ принадлежат P .

Рассмотрим $P' = P \setminus \{a_0\}$. P' уже не обязательно является связной. Предположим, P' связно. Тогда для нее справедливы следующие утверждения:

$$\begin{aligned} |P'| &= n, \\ l(P') &= l(P) = l, \\ |P'^*| &\leq |P^*| \leq |P'^*| + 1. \end{aligned}$$

И кроме того по предположению индукции $|P'^*| \leq |P'| + l + 4$. Получаем:

$$|P^*| \leq |P'^*| + 1 \leq |P'| + l + 5 \leq n + l + 5 = n + 1 + l + 4 = |P| + l + 4.$$

Остается рассмотреть случай, когда P' уже не является связной. Тогда $P' = P_1 \cup P_2$, где P_1 и P_2 — связные. Для этих фигур выполнены следующие утверждения:

$$\begin{aligned} |P_1| + |P_2| &= n, \\ |P_1^* \cap P_2^*| &\geq 4, \\ |P'^*| &\leq |P^*| \leq |P'^*| + 1, \\ |P_1^*| + |P_2^*| &= |P_1^* \cup P_2^*| + |P_1^* \cap P_2^*| \geq |P'^*| + 4, \\ l(P_1) + l(P_2) &= l(P), \\ |P_i^*| &\leq |P_i| + l(P_i) + 4, \text{ при } i = 1, 2, \\ |P^*| &\leq |P'^*| + 1 \leq |P_1^*| + |P_2^*| - 3 \leq |P_1| + |P_2| + l(P_1) + l(P_2) + 5 = \\ &= n + l(P) + 5 = |P| + l(P) + 4. \end{aligned}$$

Тем самым, лемма полностью доказана.

3. Доказательство основного результата

В данном разделе приводится доказательство теоремы 1.

Построим искомый ИГ U для заданной библиотеки V , и величин R и m , таких что $0 < R < 2$ и $1/m < R/2$.

Суть предлагаемого алгоритма будет состоять в следующем: множество всех запросов $[0, 1]^2$ разбивается на m^2 маленьких квадратов и для каждого такого квадрата строится некоторым способом решение задачи (для большинства квадратов решение будет тривиальным). Обработка конкретного запроса проходит в два этапа: сначала определяется квадрат, в который попал запрос, и затем используется алгоритм решения, соответствующий этому квадрату.

Возьмем переключательную вершину с переключателем $g_m(x)$ и объявим ее корнем. С точки зрения логики алгоритма это будет значить, что мы определяем квадрат равномерной сетки, в который

попал запрос. Будем называть эти квадраты клетками. Концу каждого ребра, исходящего из корня взаимно-однозначно сопоставим соответствующую клетку. Если запрос x прошел в некоторую вершину α , то x обязательно принадлежит клетке, сопоставленной вершине α . Посмотрим, как может выглядеть запрос, принадлежащий некоторой определенной клетке.

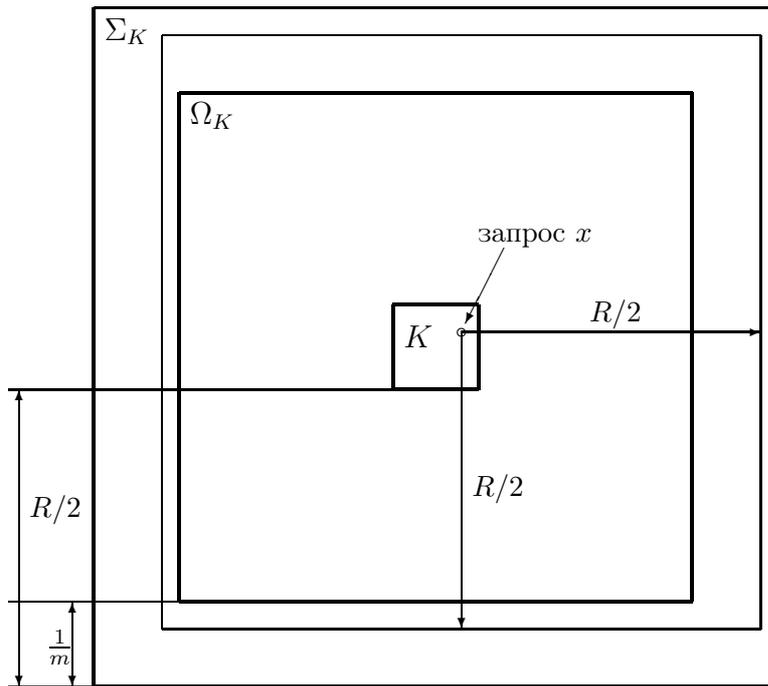


Рис. 1. Квадраты Ω_K и Σ_K .

Как показано на рисунке 1, для данной клетки K мы можем выделить два вложенных квадрата Ω_K и Σ_K с центрами, совпадающими с центром клетки и сторонами $R - 1/m$ и $R + 1/m$ соответственно. При этом можно утверждать, что квадрат, соответствующий запросу, лежащему в клетке K , полностью содержит в себе Ω_K и полностью содержится в Σ_K . Рассмотрим возможные варианты:

- 1) $\Omega_K \cap V \neq \emptyset$. Тогда существует $y \in \Omega_K \cap V$. Эта запись удо-

влетворяет любому запросу, лежащему в клетке K . Припишем вершине ИГ, соответствующей клетке K , запись y .

- 2) $\Sigma_K \cap V = \emptyset$. Тогда для любого запроса, лежащего в клетке K ответом будет пустое множество записей. Вершину ИГ, соответствующую такой клетке K мы уже больше трогать не будем.
- 3) $\Sigma_K \cap V \neq \emptyset$, но $\Omega_K \cap V = \emptyset$. То есть в маленьком квадрате записей библиотеки нет, а в большом — есть. Такие клетки назовем *нетривиальными*, а записи из множества $\Sigma_K \cap V$ будем называть соответствующими нетривиальной клетке K . Если запрос x попал в нетривиальную клетку K , то $\mathcal{J}_I(x) \subseteq \Sigma_K \cap V$, то есть только записи, соответствующие клетке K могут удовлетворять запросу x .

Для каждой нетривиальной клетки K нам придется выпустить из соответствующей ей вершины граф U_K , решающий задачу поиска представителя из множества записей, соответствующих K .

Однако рассматривать все записи принадлежащие рамке $\Sigma_K \setminus \Omega_K$ не обязательно. Указанную рамку можно разбить на восемь частей: четыре прямоугольные боковые части, включающие все свои стороны, кроме прилежащих к квадрату Ω_K ; и четыре угловые части, где каждая такая часть представляет собой квадрат такого же размера, что и клетка K . При этом, если $K = \left[\frac{i-1}{m}; \frac{i}{m} \right] \times \left[\frac{j-1}{m}; \frac{j}{m} \right]$, то указанные части имеют вид:

$$\begin{aligned} \Pi_{\leftarrow}^K &= \left[\frac{i-1}{m} - R/2; \frac{i}{m} - R/2 \right) \times \left[\frac{j}{m} - R/2; \frac{j-1}{m} + R/2 \right], \\ \Pi_{\rightarrow}^K &= \left(\frac{i-1}{m} + R/2; \frac{i}{m} + R/2 \right] \times \left[\frac{j}{m} - R/2; \frac{j-1}{m} + R/2 \right], \\ \Pi_{\downarrow}^K &= \left[\frac{i}{m} - R/2; \frac{i-1}{m} + R/2 \right) \times \left[\frac{j-1}{m} - R/2; \frac{j}{m} - R/2 \right), \\ \Pi_{\uparrow}^K &= \left[\frac{i}{m} - R/2; \frac{i-1}{m} + R/2 \right) \times \left(\frac{j-1}{m} + R/2; \frac{j}{m} + R/2 \right], \\ \Pi_{\swarrow}^K &= \left[\frac{i-1}{m} - R/2; \frac{i}{m} - R/2 \right) \times \left[\frac{j-1}{m} - R/2; \frac{j}{m} - R/2 \right), \end{aligned}$$

$$\begin{aligned} \Pi_{\swarrow}^K &= \left[\frac{i-1}{m} - R/2; \frac{i}{m} - R/2 \right) \times \left(\frac{j-1}{m} + R/2; \frac{j}{m} + R/2 \right], \\ \Pi_{\nearrow}^K &= \left(\frac{i-1}{m} + R/2; \frac{i}{m} + R/2 \right] \times \left(\frac{j-1}{m} + R/2; \frac{j}{m} + R/2 \right), \\ \Pi_{\searrow}^K &= \left(\frac{i-1}{m} + R/2; \frac{i}{m} + R/2 \right] \times \left[\frac{j-1}{m} - R/2; \frac{j}{m} - R/2 \right). \end{aligned}$$

Возьмем для примера левый прямоугольник рамки Π_{\swarrow}^K . Рассмотрим любой запрос, лежащий в клетке K . Для такого запроса все записи библиотеки, лежащие в Π_{\swarrow}^K удовлетворяют запросу по второй координате, но не обязательно по первой. При этом попадание в ответ на запрос какой-либо записи из Π_{\swarrow}^K влечет за собой попадание в ответ и самой правой записи из Π_{\swarrow}^K . Поэтому при рассмотрении нетривиальной клетки K из множества $V \cap \Pi_{\swarrow}^K$ достаточно оставить только одну — самую правую запись. Важно отметить, что отброшенные точки не рассматриваются только при построении графа U_K . Вполне возможно, что для других нетривиальных клеток эти точки уже не будут отбрасываться. Аналогично из $V \cap \Pi_{\nearrow}^K$ оставляется самая левая запись, из $V \cap \Pi_{\uparrow}^K$ нижняя, а из $V \cap \Pi_{\downarrow}^K$ — верхняя.

Несколько похожая ситуация и с $V \cap \Pi_{\searrow}^K$. Только для этих множеств рассматривать придется не одну запись, а целый слой записей — $\left\{ y = (y_1, y_2) \in V \cap \Pi_{\searrow}^K \mid \nexists y' = (y'_1, y'_2) \in V \cap \Pi_{\searrow}^K, y'_1 \geq y_1 \text{ и } y'_2 \geq y_2 \right\}$. Аналогично рассматриваются множества $V \cap \Pi_{\swarrow}^K$, $V \cap \Pi_{\nearrow}^K$ и $V \cap \Pi_{\downarrow}^K$.

Оставшиеся после таких действий записи, соответствующие нетривиальной клетке K обозначим соответственно $V_{\swarrow}^K, V_{\nearrow}^K, V_{\uparrow}^K, V_{\downarrow}^K, V_{\searrow}^K, V_{\swarrow}^K, V_{\nearrow}^K, V_{\searrow}^K$. Из первых четырех множеств каждое либо пусто, либо содержит одну запись. Каждое из $V_{\nearrow}^K, V_{\searrow}^K$ либо пусто, либо содержит набор попарно несравнимых точек. Каждое из $V_{\swarrow}^K, V_{\searrow}^K$ либо пусто, либо содержит набор попарно строго сравнимых точек.

Рисунок 2 показывает, как могут выглядеть отбрасываемые и рассматриваемые записи.

Как было только что показано, для каждой нетривиальной клетки нам будет нужно в худшем случае хранить 4 записи из боковых

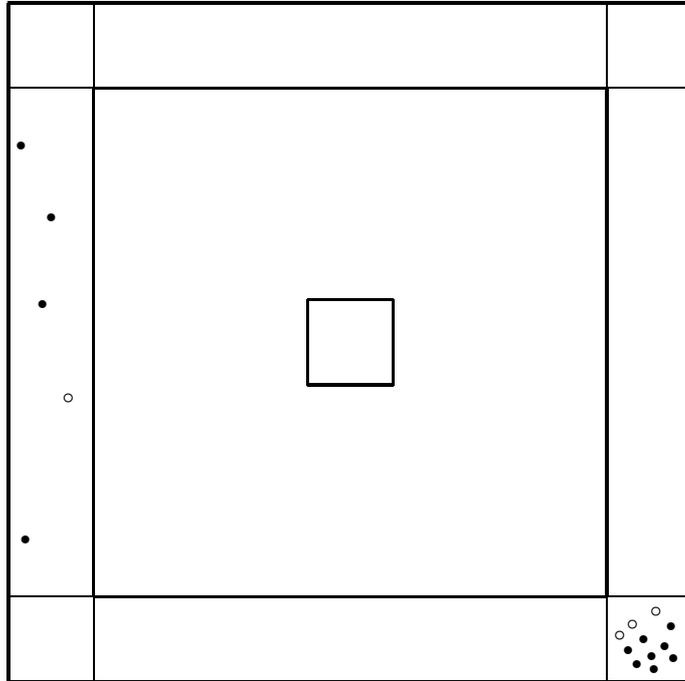


Рис. 2. Отбрасываемые точки сплошные.

прямоугольников рамки $\Sigma_K \setminus \Omega_K$ и 4 множества для каждого угла рамки: $V_{\swarrow}^K, V_{\nwarrow}^K, V_{\searrow}^K, V_{\swarrow}^K$. Никакая запись библиотеки не может принадлежать одновременно $V_{\swarrow}^{K'}$ и $V_{\swarrow}^{K''}$, если K' и K'' — две разные клетки, так как если $K' \neq K''$, то $\Pi_{\swarrow}^{K'} \cap \Pi_{\swarrow}^{K''} = \emptyset$.

Если считать, что на клетки разбит не только квадрат $[0; 1]^2$, но и вся плоскость, то для любой клетки K множество клеток $\{K' \mid \Pi_{\swarrow}^{K'} \cap \Pi_{\swarrow}^K \neq \emptyset\}$ состоит ровно из четырех клеток, расположенных квадратом 2×2 . Обозначим их K'_1, K'_2, K'_3, K'_4 начиная с нижнего левого по часовой стрелке. При этом записи из V_{\swarrow}^K не могут попасть и в $V_{\swarrow}^{K'_1}$ и в $V_{\swarrow}^{K'_3}$ одновременно. Кроме того в каждый из $V_{\swarrow}^{K'_i}$ может попасть только одна запись из V_{\swarrow}^K , так как в каждом $V_{\swarrow}^{K'_i}$ записи попарно строго сравнимы, а в V_{\swarrow}^K попарно несравнимы. Таким обра-

зом не более трех точек V_{\swarrow}^K попадают в какие-либо $V_{\nwarrow}^{K'}$. Аналогично обстоит дело с V_{\searrow}^K .

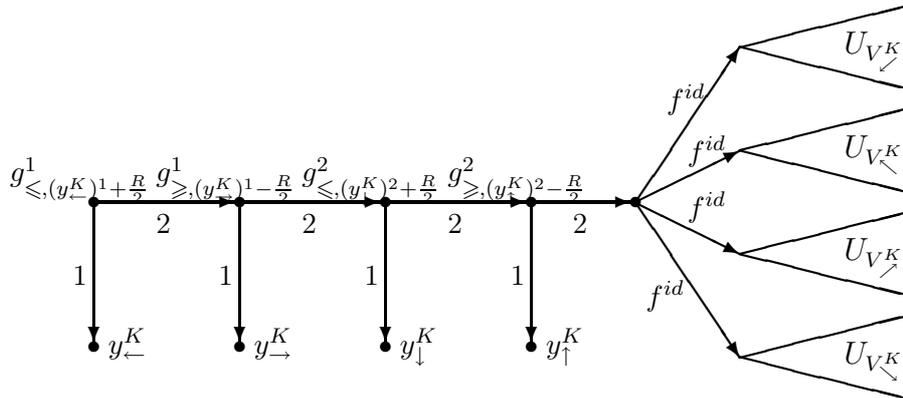


Рис. 3. Общий вид графов U_K .

Построим граф U_K способом, указанным на рисунке 3. При этом, если, например, $V_{\swarrow}^K = \emptyset$, то соответствующий граф $U_{V_{\swarrow}^K}$ строить не нужно. Точно также обстоит дело и с y_{\leftarrow}^K — если в левом прямоугольнике рамки нет точек, то соответствующий элемент U_K опускается. Графы $U_{V_{\swarrow}^K}$, $U_{V_{\nwarrow}^K}$, $U_{V_{\searrow}^K}$, $U_{V_{\swarrow}^K}$ представляют собой бинарные деревья поиска [2], у которых переключательными являются внутренние вершины, у которых потомки — внутренние вершины. Сложность таких графов оценивается следующим образом:

$$\hat{T}(U_{V_{\swarrow}^K}) \leq 1 + \lceil \log_2 |V_{\swarrow}^K| \rceil,$$

$$Q(U_{V_{\swarrow}^K}) \leq 2|V_{\swarrow}^K| - 1.$$

Аналогичные оценки сложности для графов $U_{V_{\nwarrow}^K}$, $U_{V_{\searrow}^K}$, $U_{V_{\swarrow}^K}$.

Отсюда для U_K получаем следующие оценки:

$$\hat{T}(U_K) \leq 8 + 4 \lceil \log_2 k \rceil,$$

$$Q(U_K) \leq 8 + 2(|V_{\swarrow}^K| + |V_{\nwarrow}^K| + |V_{\searrow}^K| + |V_{\swarrow}^K|).$$

Таким образом сразу получаем оценки сложности U :

$$\hat{T}(U) \leq 9 + 4 \lceil \log_2 k \rceil, \tag{1}$$

$$T(U) \leq 1 + \frac{4c}{m^2} (\lceil \log_2 k \rceil + 2)N, \tag{2}$$

где N равно числу нетривиальных клеток.

Обозначим множество нетривиальных клеток \mathcal{K} . Введем множества V_1, V_2 :

$$V_1 = \bigcup_{K \in \mathcal{K}} (V_{\nearrow}^K \cup V_{\searrow}^K), \quad V_2 = \bigcup_{K \in \mathcal{K}} (V_{\swarrow}^K \cup V_{\nwarrow}^K).$$

Как уже говорилось выше, для каждой нетривиальной клетки K не более трех элементов V_{\nearrow}^K принадлежат каким-либо $V_{\swarrow}^{K'}$ и не более трех элементов принадлежат каким-либо $V_{\nwarrow}^{K'}$. То есть не более шести элементов из V_{\nearrow}^K лежат в V_2 . Аналогично, не более шести элементов V_{\searrow}^K лежат в V_2 . Всего получаем не более $12N$ элементов V_1 лежит в V_2 , иными словами $|V_1 \cap V_2| \leq 12N$. Откуда получаем:

$$\begin{aligned} \sum_{K \in \mathcal{K}} Q(U_K) &\leq 8N + 2(2|V_1| + 2|V_2|) \leq 8N + 4(k + 12N) \leq 4k + 60N, \\ Q(U) &= m^2 + \sum_{K \in \mathcal{K}} Q(U_K) \leq 4k + m^2 + 60N. \end{aligned} \tag{3}$$

Оценим число нетривиальных клеток N .

Будем называть *окном* прямоугольник W , состоящий из клеток. При этом считается, что на клетки разбивается не только квадрат $[0, 1]^2$, но и все \mathbb{R}^2 . Множество всех клеток будем обозначать \mathbb{K} .

Если K — клетка, то обозначим K^* множество, состоящее из самой клетки K и восьми ее соседей (по вертикали, горизонтали и диагонали). Аналогично, если $W \subseteq \mathbb{K}$, то $W^* = \bigcup_{K \in W} K^*$.

Если W, W_1, \dots, W_n — наборы клеток, и $W = \bigcup_{i=1}^n W_i$, то легко убедиться, что $W^* = \bigcup_{i=1}^n W_i^*$.

Если $\Pi \subseteq [0, 1]^2$ — прямоугольник, то введем Π^+, Π^- следующим образом:

$$\begin{aligned}\Pi^+ &= \{K \in \mathbb{K} \mid K \cap \Pi \neq \emptyset\}, \\ \Pi^- &= \{K \in \mathbb{K} \mid K \subseteq \Pi\}.\end{aligned}$$

Π^+ — окно для любого прямоугольника Π , а Π^- — всегда либо окно, либо пустое множество. Кроме того, всегда выполнено соотношение $\Pi^- \subseteq \Pi \subseteq \Pi^+$. При этом, если Π^- — окно, то $\Pi^+ \subseteq (\Pi^-)^*$.

Лемма 3. Число нетривиальных клеток $N \leq m^2 \frac{(mR-1)(2+R)^2}{(mR-2)^2}$.

Доказательство. Пусть множество записей из библиотеки $V = \{y_1, \dots, y_k\} \subseteq Y = [0, 1]^2$. Рассмотрим расширение нашей задачи — заменим множество запросов $X = [0, 1]^2$ на $X = \mathbb{R}^2$. Количество нетривиальных клеток при этом может только возрасти. Рассмотрим тени записей (уже в расширенной задаче) $\Pi_y := \{x \in \mathbb{R}^2 : \rho_{\square}(x, y) \leq R/2\}$. Теперь для всех $y \in V$ рассмотрим Π_y^+ и Π_y^- . Так как $R > 2/m$, то они оба являются окнами и $\Pi_y^- \subseteq \Pi_y \subseteq \Pi_y^+$. При этом любая из ширин Π_y^- принадлежит промежутку $(R - 2/m, R]$ и каждая из ширин Π_y^+ принадлежит промежутку $(R, R + 2/m]$. Смысл введенных множеств следующий: любая клетка из Π_y^- является тривиальной (запись y подходит любому запросу, попавшему в такую клетку), аналогично ни одной клетке, не попавшей в Π_y^+ не принадлежит запрос, которому бы удовлетворяла запись y . Введем множества Y_V^+, Y_V^- :

$$\begin{aligned}Y_V^+ &:= \bigcup_{y \in V} \Pi_y^+, \\ Y_V^- &:= \bigcup_{y \in V} \Pi_y^-\end{aligned}$$

Множество Y_V^- — является объединением всех тривиальных клеток первого типа. Множество $\mathbb{K} \setminus Y_V^+$ — объединение тривиальных клеток второго типа. Нетривиальные же клетки, составляют множество $Y_V^+ \setminus Y_V^-$. Разделим Y_V^- на связные части (соседними считаются клетки, по вертикали и горизонтали, но не по диагонали). При этом получится

некоторое разбиение $V = \bigsqcup_{i=1}^s V_i$. $Y_{V_i}^+ := \bigcup_{y \in V_i} \Pi_y^+$, $Y_{V_i}^- := \bigcup_{y \in V_i} \Pi_y^-$. Так как $\Pi_y^+ \subseteq (\Pi_y^-)^*$, то $Y_{V_i}^+ \subseteq (Y_{V_i}^-)^*$. То есть $Y_{V_i}^+ \setminus Y_{V_i}^- \subseteq (Y_{V_i}^-)^* \setminus Y_{V_i}^-$. Получаем цепочку включений:

$$Y_V^+ \setminus Y_V^- = \bigcup_{i=1}^s Y_{V_i}^+ \setminus \bigcup_{i=1}^s Y_{V_i}^- \subseteq \bigcup_{i=1}^s (Y_{V_i}^+ \setminus Y_{V_i}^-) \subseteq \bigcup_{i=1}^s ((Y_{V_i}^-)^* \setminus Y_{V_i}^-).$$

Применяя лемму 2 и учитывая, что периметр из утверждения леммы будет измерен в единицах $\frac{1}{m}$, получаем:

$$|(Y_{V_i}^-)^* \setminus Y_{V_i}^-| \leq 4 + L(Y_{V_i}^-)m.$$

По лемме 1 справедлива оценка:

$$L(Y_{V_i}^-) \leq 4 \frac{S(Y_{V_i}^-)}{R - 2/m}.$$

Таким образом получаем:

$$|(Y_{V_i}^-)^* \setminus Y_{V_i}^-| \leq 4 + L(Y_{V_i}^-)m \leq 4 + 4m \frac{S(Y_{V_i}^-)}{R - 2/m}.$$

Поэтому для числа N' нетривиальных клеток в расширенной задаче имеем:

$$\begin{aligned} N' = |Y_V^+ \setminus Y_V^-| &\leq \sum_{i=1}^s |(Y_{V_i}^-)^* \setminus Y_{V_i}^-| \leq \\ &\leq \sum_{i=1}^s \left(4 + 4m \frac{S(Y_{V_i}^-)}{R - 2/m} \right) \leq 4s + 4m \frac{(1 + R/2)^2}{R - 2/m}. \end{aligned}$$

Но $\sum_{i=1}^s S(Y_{V_i}^-) \leq (1 + R/2)^2$ и для всех i выполнено $S(Y_{V_i}^-) \geq (R - 2/m)^2$. Поэтому $s \leq \frac{(1+R/2)^2}{(R-2/m)^2}$. Итого $N' \leq 4 \frac{(1+R/2)^2}{(R-2/m)^2} + 4m \frac{(1+R/2)^2}{R-2/m} = m^2 \frac{(mR-1)(2+R)^2}{(mR-2)^2}$. Лемма доказана.

Подставляя в формулы (1),(2),(3) оценку для числа нетривиальных клеток, получим, что характеристики ИГ U удовлетворяют оценкам, сформулированным в утверждении теоремы 1. Тем самым, теорема 1 полностью доказана.

Доказательство следствия 1. $\frac{1}{R_k} = o\left(\frac{\sqrt{k}}{\log_2 k}\right)$ по определению означает, что $\frac{\log_2 k}{R_k \sqrt{k}} = \alpha_k = o(1)$. Возьмем $m_k = \lceil \sqrt{k \alpha_k} \rceil$. Тогда выполнено $m_k = o\left(\sqrt{k}\right)$, и в то же время $\frac{m_k R_k}{\log_2 k} \geq \frac{\sqrt{k \alpha_k} R_k}{\log_2 k} = \frac{\sqrt{k \alpha_k}}{\log_2 k} \frac{\log_2 k}{\sqrt{k \alpha_k}} = \frac{1}{\sqrt{\alpha_k}} \rightarrow \infty$. Откуда, в частности, имеем $m_k R_k \rightarrow \infty$. Теперь рассмотрим интересующую нас последовательность:

$$\begin{aligned} 1 \leq T(U_k) &\leq 1 + 4c \frac{(m_k R_k - 1)(2 + R_k)^2}{(m_k R_k - 2)^2} (2 + \lceil \log_2 k \rceil) \leq \\ &\leq 1 + 64c \frac{m_k R_k - 1}{(m_k R_k - 2)^2} (3 + \log_2 k) \sim 1 + 64c \frac{1}{m_k R_k} \log_2 k \rightarrow 1. \end{aligned}$$

Таким образом, $T(U_k) \sim 1$. Аналогично получаем $Q(U_k) \lesssim 4k$. Оценка $\hat{T}(U_k) \leq 9 + 4 \lceil \log_2 k \rceil$ следует непосредственно из утверждения теоремы 1. Утверждения следствия полностью доказано.

Список литературы

- [1] Gasanov E. E. On Functional Complexity of Two-dimensional Manhattan Metrics Closeness Problem // Emerging Database Research In East Europe. Proceedings of the pre-conference workshop of VLDB. 2003. P. 51–56.
- [2] Гасанов Э. Э., Кудрявцев В. Б. Теория хранения и поиска информации. М.: ФИЗМАТЛИТ, 2002.