Задача поиска по ключу с определением позиции

Н. С. Кучеренко

В работе изучаются два способа формализации задачи поиска по ключу — задача поиска идентичных объектов (ЗПИО) и расширенная ЗПИО. Показано, что сложность поиска для этих способов формализации одинакова. Описан алгоритм преобразования, который алгоритм поиска для ЗПИО преобразует в алгоритм поиска для расширенной ЗПИО без изменения сложности поиска.

Ключевые слова: базы данных, поиск по ключу, алгоритм поиска.

Введение

Теория хранения и поиска информации является важным разделом теории интеллектуальных систем. Одним из ключевых объектов этой теории является информационный граф (ИГ) — управляющая система, которая позволяет рассматривать имеющиеся модели данных и задачи, связанные с ними, с более общих позиций [3].

В работе рассматривается задача поиска по ключу [4]. Предполагается, что каждому объекту базы данных сопоставлен уникальный ключ. На множестве всех возможных ключей введен линейный порядок, и ключи объектов базы данных упорядочены. К базе данных приходит запрос в виде значения ключа для искомого объекта. Цель поиска — выяснить есть ли в базе данных объект, ключ которого равен запросу.

Изучаются два способа формализации задачи поиска по ключу. Первый вариант формализации — это задача поиска идентичных объектов (ЗПИО) для интервала $(0, 1)$. Алгоритм поиска для решения этой задачи при отсутствии объекта, ключ которого равен за-

просу, отвечает, что объекта в базе данных нет. Второй способ формализации — это расширенная задача поиска идентичных объектов (РЗПИО) для интервала $(0,1)$. При этом способе формализации в случае отсутствия объекта в базе данных алгоритм поиска выдает промежуток между ключами, в который попал запрос.

Независимо от способа формализации задачи алгоритмы поиска представляются с помощью информационных графов. Предполагается, что при поиске можно использовать только операции сравнения. Сложность информационного графа вводится как среднее количество операций сравнения вычисленных при обработке запроса. Поскольку для решения РЗПИО требуется выяснить больше информации о базе данных, возникает вопрос, на сколько увеличивается сложность оптимального по вычислительным возможностям ИГ для РЗПИО по сравнению с «более простой» ЗПИО.

Ранее в работе [1] автором было показано, что оптимальный информационный граф для задачи поиска идентичных объектов I принадлежит множеству D_I — множеству древовидных ИГ с жестко ограниченной структурой.

В данной работе построен алгоритм преобразования древовидного ИГ из множества D_I в древовидный информационный граф, который решает расширенную задачу поиска идентичных объектов. При этом сложность информационного графа при преобразовании не изменяется.

Также автором показано, что сложности оптимальных информационных графов для задачи поиска идентичных объектов и расширенной задачи поиска идентичных объектов равны. Из этого следует, что при изучении сложности РЗПИО можно переходить к «более простой» ЗПИО. Для задачи поиска идентичных объектов I известны алгоритмы построения оптимального ИГ из множества D_I и информационного графа бинарного поиска из множества D_I. После применения в конце этих алгоритмов построения алгоритма преобразования получаются алгоритмы построения оптимального ИГ и информационного графа бинарного поиска для расширенной задачи поиска идентичных объектов.

Автор выражает благодарность своему научному руководителю профессору Э. Э. Гасанову за постановку задачи и внимание к работе и академику В. Б. Кудрявцеву за ценные советы и замечания.

1. Основные понятия и результаты

В данной работе алгоритмы поиска, использующие только операции сравнения, исследуются с позиции информационно-графовой модели данных [3]. В информационно-графовой модели структура данных задается ориентированным графом, ребра и вершины которого нагружены элементами данных и функциями. Алгоритм поиска — это «волновой» процесс на графе, который управляется нагрузочными функциями. Нагрузочные функции разделены на два класса: предикаты и переключатели. Поскольку для операций сравнения выбрано представление в виде переключателей, в работе используется информационный граф, множество нагрузочных функций которого состоит только из класса переключателей. Информационный граф такого вида называется *информационным графом с переключательной нагрузкой* (ИГПН).

Для задачи поиска по ключу приводится два способа ее формализации — задача поиска идентичных объектов на интервале $(0,1)$ и расширенная задача поиска идентичных объектов на интервале $(0,1)$.

Задача поиска идентичных объектов на интервале $(0,1)$ формализуется [3] как четверка $((0,1),\ V,\ \rho_{=},\ f(x))$, где $(0,1)$ — множество запросов. Конечный набор V точек интервала $(0,1)$, $V = (y_1, y_2, \ldots, y_n)$, элементы которого упорядочены по возрастанию и не равны между собой, называется библиотекой, а элементы библиотеки — записями. Отношение поиска $\rho_{=}$ — это отношение равенства на множестве $(0,1)\times(0,1)$. Предполагается, что запрос $x \in (0,1)$ — случайная величина, принимающая значения из интервала $(0,1)$, с функцией плотности $f(x)$. Отношение $\rho_{=}$ задает функцию ответов $J(x)$, определенную на множестве запросов $(0,1)$, следующим образом

$$J(x) = \begin{cases} \{y_i\}, & \text{если } \exists i \in [1..n] : y_i = x, \\ \varnothing, & \text{иначе.} \end{cases}$$

Расширенная задача поиска идентичных объектов (РЗПИО) на интервале $(0,1)$ — это задача поиска I', которая представляет собой четверку $((0,1),\ V',\ \rho_{\in},\ f(x))$, где $(0,1)$ — множество запросов. Обозначим через Y множество, состоящее из всех подинтервалов и точек интервала $(0,1)$. Конечный набор V' состоит из таких элементов множества Y, что они представляют собой разбиение интервала $(0,1)$ следующего вида

$$V' = ((0, y_1), \{y_1\}, (y_1, y_2), \{y_2\}, \ldots, \{y_n\}, (y_n, 1)), \quad y_1 < y_2 < \cdots < y_n.$$

Набор V называется библиотекой, а элементы библиотеки — записями. Отношение поиска $\rho_\in$ — это отношение на множестве $(0,1) \times Y$. Запрос x, $x \in (0,1)$, находится в отношении $\rho_\in$ с элементом y множества Y тогда и только тогда, когда $x \in y$. Предполагается, что запрос $x \in (0,1)$ — случайная величина, принимающая значение из интервала $(0,1)$, с функцией плотности $f(x)$. Отношение $\rho_\in$ задает функцию ответов $J'(x)$, определенную на множестве запросов $(0,1)$, следующим образом

$$J'(x) = \begin{cases} \{y_i\}, & \text{если } \exists i \in [1..n] : y_i = x, \\ (y_j, y_{j+1}), & \text{если } \exists j \in [0..n] : y_j < x < y_{j+1}, \end{cases}$$

где $y_0 = 0$, $y_{n+1} = 1$.

Расширением ЗПИО $I = ((0,1), V, \rho_=, f(x))$, где $V = (y_1, y_2, \ldots, y_n)$, назовем РЗПИО I' $((0,1)$, V', $\rho_\in$, $f(x))$, где

$$V' = ((0, y_1), \{y_1\}, (y_1, y_2), \{y_2\}, \ldots, \{y_n\}, (y_n, 1)), \quad y_1 < y_2 < \cdots < y_n.$$

Для реализации функции ответов задачи поиска используются операции сравнения вещественных чисел из интервала $(0,1)$, которые заданы в виде функции $p(y,x)$

$$\forall y, x \in (0,1) \quad p(y,x) = \begin{cases} 1, & \text{если } x < y, \\ 2, & \text{если } x = y, \\ 3, & \text{если } x > y. \end{cases}$$

Функция $p_a(x) = p(a,x)$, $a \in (0,1)$, называется переключателем. Определим базовое множество переключателей $G = \{p_a(x) \mid a \in (0,1)\}$. Над базовым множеством G строится информационный граф с переключательной нагрузкой, который хранит в себе схему вызова переключателей из G.

Информационный граф с переключательной нагрузкой — это связный ориентированный конечный граф без кратных ребер и петель, вершины и ребра которого размечены следующим образом. Выделены два подмножества вершин P и L, так что $P \cap L = \varnothing$.

Вершины множества P называются *переключательными*, а вершины множества L — *листьями*. Ребра, исходящие из переключательных вершин, называются также *переключательными*. Среди переключательных вершин выделена одна вершина v_0, которая называется *корневой*. Переключательным вершинам сопоставлены переключатели из G. Переключатель $p_a(x)$, сопоставленный переключательной вершине v, называется ее нагрузкой и обозначается $[v]_P = p_a(x)$. Листьям сопоставляются записи библиотеки. Это сопоставление обозначается $[v]_L = y$, $y \in V$, $v \in L$; y называется нагрузкой листа v. Переключательным ребрам сопоставляются элементы множества $\{\Lambda, 1, 2, 3\}$, при этом из переключательной вершины не выходит двух ребер с одинаковыми номерами. Если ребру (v_{i_1}, v_{i_2}) сопоставлен номер $i \in \{1, 2, 3\}$, то номер i называется нагрузкой ребра (v_{i_1}, v_{i_2}) и обозначается $[(v_{i_1}, v_{i_2})] = i$.

Определим *функционирование* информационного графа с переключательной нагрузкой. Переключательное ребро (v_{i_1}, v_{i_2}), которому приписан номер i, *проводит запрос* $x \in (0, 1)$, если переключатель, приписанный началу этого ребра, принимает значение i на запросе x, $[(v_{i_1}, v_{i_2})] = [v_{i_1}]_P(x)$. Если $[(v_{i_1}, v_{i_2})] \neq [v_{i_1}]_P(x)$, то переключательное ребро (v_{i_1}, v_{i_2}), *не проводит запрос* $x \in X$. Полагаем, что переключательное ребро, которому был сопоставлен символ Λ, не проводит запросы. Также не проводят запросы все остальные ребра, отличные от переключательных. Ориентированный маршрут проводит запрос $x \in X$, если каждое его ребро проводит запрос x. Запрос $x \in (0, 1)$ *проходит в вершину* v *или достигает вершину* v, если существует ориентированный маршрут из корневой вершины в вершину v, который проводит запрос x. Запись y, приписанная листу v', называется *результатом функционирования информационного графа с переключательной нагрузкой на запросе* x, если запрос x проходит в лист v'. *Результат функционирования информационного графа с переключательной нагрузкой* U — это функция $R_U(x)$, которая каждому запросу x сопоставляет множество всех записей, которые являются результатом функционирования ИГПН на запросе x.

Информационный граф с переключательной нагрузкой U *решает* задачу поиска I, если результат функционирования ИГПН U совпадает с функцией ответов задачи поиска на любом запросе x из интервала $(0, 1)$. Множество ИГПН, которые решают задачу поиска I, обозначим через S_I.

Если в информационном графе с переключательной нагрузкой U удалить «не функциональные» элементы, а именно, не переключательные ребра и ребра, которым сопоставлен символ Λ, а затем оставить только одну компоненту связности, которой принадлежит корневая вершина, то результат функционирования $R_U(x)$ не изменится. Поэтому далее рассматриваются ИГПН с точностью до этих операций удаления.

Рассмотрим информационный граф с переключательной нагрузкой из множества S_I, где I некоторая задача поиска. *Сложностью ИГПН U на запросе x* называется число переключательных вершин, которые достигаются запросом x при функционировании U. Сложность можно представить в виде

$$T(U,x) = \sum_{v \in P_U} \varphi_v(x),$$

где P_U — множество переключательных вершин, а $\varphi_v(x)$ — предикат, принимающий значение единица, если x проходит в вершину v, и нуль в противном случае. Предикат $\varphi_v(x)$ называется *функцией фильтра вершины v*. Величина $T(U,x)$ как функция от x является случайной величиной [3]. *Сложностью ИГПН U из множества S_I* называется математическое ожидание величины $T(U,x)$, которое можно записать в виде

$$T_I(U) = \mathbf{M}(T(U,x)) = \int_0^1 T(U,x) f(x)\, dx,$$

где $f(x)$ — функция плотности распределения запросов в задаче поиска I.

Сложностью задачи поиска I называется величина

$$T(I) = \inf_{U \in S_I} T_I(U).$$

Информационный граф с переключательной нагрузкой, на котором достигается инфимум, называется *оптимальным информационным графом с переключательной нагрузкой для задачи поиска I*.

Информационный граф с переключательной нагрузкой — это особым образом размеченный ориентированный конечный граф без кратных ребер и петель. Информационный граф без разметки назовем *основанием* информационного графа.

Фундаментом ориентированного графа называется неориентированный граф, получаемый из ориентированного после удаления ориентации на всех ребрах [5]. Ориентированный граф, у которого выделена корневая вершина, назовем *деревом, ориентированным от корня*, если выполнены следующие условия.

1) Фундаментом ориентированного графа является дерево.
2) В корневую вершину не входит ни одно ребро.
3) В любую вершину графа из корневой вершины ведет ориентированный маршрут.

Информационный граф с переключательной нагрузкой назовем *древовидным*, если его основание является деревом, ориентированным от корневой вершины информационного графа.

Точками разбиения ЗПИО $I=((0,1),V,\rho_{=},f)$, $V=(y_1,y_2,\ldots,y_n)$, назовем набор точек $\widehat{V}$ который равен библиотеке, $\widehat{V}=V$. *Точками разбиения* РЗПИО $((0,1),V',\rho_{\in},f(x))$, где

$$V'=((0,y_1),\{y_1\},(y_1,y_2),\{y_2\},\ldots,\{y_n\},(y_n,1)),\quad y_1<y_2<\cdots<y_n,$$

назовем набор $\widehat{V}'=(y_1,\,y_2,\ldots,\,y_n)$. Заметим, что наборы точек разбиения ЗПИО и ее расширения равны.

Для задачи поиска I со множеством точек разбиения $\widehat{V}=(y_1,\,y_2,\ldots,\,y_n)$ введем множество D_I. Информационный граф с переключательной нагрузкой U принадлежит множеству D_I тогда и только тогда, когда выполнены следующие условия.

1) ИГПН U является древовидным и решает задачу поиска I.
2) ИГПН U содержит n внутренних вершин, которые являются переключательными и которым взаимно однозначно сопоставлены n переключателей

$$\{p_{y_1},p_{y_2},\ldots,p_{y_n}\}.$$

3) Из каждой переключательной вершины ИГПН U исходит три ребра, которым взаимно однозначно сопоставлены числа из множества $\{1,2,3\}$.

Для информационного графа с переключательной нагрузкой из множества D_I введем понятие *номера ориентированного маршрута*. Номер ориентированного маршрута — это набор $(a_1, a_2, \ldots, a_s)$, где s — количество ребер в ориентированном маршруте, a_i — нагрузка i-го переключательного ребра (нумерация ребер начинается от корневой вершины). Поскольку по определению в ИГПН из множества D_I все ребра являются переключательными с нагрузкой из множества $\{1, 2, 3\}$, то определение номера корректно. На множестве номеров введем отношение порядка. Для сравнения двух номеров $a = (a_1, a_2, \ldots, a_s)$, $s > 0$ и $b = (b_1, b_2, \ldots, b_t)$, $t > 0$, в конец более короткого номера допишем нули так, чтобы длины номеров сравнялись. Обозначим преобразованную пару номеров

$$a' = (a'_1, a'_2, \ldots, a'_{\max(s,t)}), \quad b' = (b'_1, b'_2, \ldots, b'_{\max(s,t)}).$$

Номера a и b равны, если они равны как наборы, то есть $s = t$ и равны соответствующие элементы

$$a_1 = b_1, a_2 = b_2, \ldots, a_s = b_t.$$

Номер a больше b, если существует такое j, $1 \leqslant j \leqslant \max(s,t)$, что

$$a'_1 = b'_1, \ldots, a'_{j-1} = b'_{j-1}, a'_j > b'_j.$$

Опишем *алгоритм преобразования* информационного графа с переключательной нагрузкой U из множества D_I, где I — задача поиска идентичный объектов $I = ((0,1), V, \rho_=, f)$, $V = (y_1, y_2, \ldots, y_n)$, в информационный граф U' из множества $D_{I'}$, где I' — расширенная задача поиска идентичных объектов $((0,1), V', \rho_\in, f(x))$,

$$V' = ((0, y_1), \{y_1\}, (y_1, y_2), \{y_2\}, \ldots, \{y_n\}, (y_n, 1)) .$$

1) Рассмотрим все висячие вершины ИГПН U, которые не являются листовыми. Обозначим это множество через R.
2) Каждой вершине α из множества R сопоставим номер ориентированного маршрута, который ведет из корневой вершины в рассматриваемую вершину α.
3) Упорядочим вершины множества R в соответствии с их номерами. Обозначим их $(\alpha_0, \alpha_1, \ldots, \alpha_{|R|})$, где α_0 — вершина с наименьшим номером.

4) Введем обозначения $y_0 = 0$, $y_n = 1$. Сопоставим вершинам множества R интервалы следующим образом

$$\forall j \leqslant n \quad \alpha_j = (y_j, y_{j+1}).$$

5) Обозначим вершины множества R, которым сопоставлены интервалы, листовыми.
6) Обозначим полученный ИГПН через U'.
7) Алгоритм закончен.

Автором показано, что алгоритм преобразования работает правильно и не меняет сложность информационного графа с переключательной нагрузкой.

Теорема 1. *Для любой задачи поиска идентичных объектов I и для любого информационного графа с переключательной нагрузкой U из множества D_I, верно, что после применения алгоритма преобразования к ИГПН U получается ИГПН U', который принадлежит множеству $D_{I'}$, где РЗПИО I' — расширение задачи поиска идентичных объектов I. При этом сложность ИГПН U равна сложности получившегося ИГПН U'*

$$T_I(U) = T_{I'}(U').$$

В статье [1] автором было показано, что для любой ЗПИО I во множестве D_I содержится оптимальный информационный граф с переключательной нагрузкой для задачи I. Используя этот результат и теорему 1 в работе доказано, что задача поиска идентичных объектов и ее расширение имеют одинаковую сложность.

Теорема 2. *Сложность любой задачи поиска идентичных объектов I и ее расширения I' равны*

$$T(I) = T(I').$$

В силу этой теоремы при исследовании сложности РЗПИО $I' = ((0,1), V', \rho_{\in}, f(x))$, где

$$V' = ((0, y_1), \{y_1\}, (y_1, y_2), \{y_2\}, \ldots, \{y_n\}, (y_n, 1)), \quad y_1 < y_2 < \cdots < y_n,$$

можно переходить к исследованию сложности ЗПИО $I = ((0,1),\ V,\ \rho_=,\ f)$, где $V = (y_1,\ y_2, \ldots,\ y_n)$. При этом, если описан алгоритм построения оптимального ИГПН из множества D_I, то после добавления в его конец алгоритма преобразования получается алгоритм построения оптимального ИГПН для РЗПИО I' из множества $D_{I'}$. Аналогичным образом может быть получен алгоритм построения информационного графа бинарного поиска из множества $D_{I'}$.

2. Доказательство теоремы 1

Для доказательства теоремы понадобится понятие проводимости ориентированного маршрута, проводимости вершины и кода вершины. Поскольку при определении информационного графа с переключательной нагрузкой мы договорились не рассматривать нефункциональные элементы (ребра, которые не являются переключательными, и ребра которым сопоставлен символ Λ), то все ребра любого ориентированного маршрута являются переключательными с нагрузкой из множества $\{1, 2, 3\}$.

Длиной ориентированного маршрута назовем число его ребер. *Проводимостью ориентированного маршрута* M назовем множество всех запросов x, $x \in (0,1)$, которые ориентированный маршрут M проводит. Проводимость ориентированного маршрута M обозначим через $X(M)$. Для проводимости ориентированного маршрута верна следующая лемма.

Лемма 1. *Проводимостью любого ориентированного маршрута в любом информационном графе с переключательной нагрузкой является подмножество интервала* $(0,1)$, *которое может быть только либо пустым множеством, либо точкой, либо интервалом.*

Доказательство. Рассмотрим произвольный информационный граф с переключательной нагрузкой и произвольный ориентированный маршрут M в нем

$$M = \left(v_{i_0}, \left(v_{i_0}, v_{i_1}\right), v_{i_1}, \ldots, \left(v_{i_{l-1}}, v_{i_l}\right), v_{i_l}\right).$$

Поскольку все ребра являются переключательными, то вершины $v_{i_0}, v_{i_1}, \ldots, v_{i_{l-1}}$ являются переключательными.

По определению переключательное ребро $(v_{i_j}, v_{i_{j+1}})$, которому приписан номер m_j, проводит запрос $x \in (0,1)$, если переключатель $[v_{i_j}]_P = p_{a_j}(x)$, приписанный началу этого ребра, принимает значение m_j на запросе x, то есть, если $p_{a_j}(x) = m_j$. Условие $p_{a_j}(x) = m_j$ задает одно из трех неравенств

$$\begin{aligned} x < a_j, &\quad \text{если } m_j = 1, \\ x = a_j, &\quad \text{если } m_j = 2, \\ x > a_j, &\quad \text{если } m_j = 3, \end{aligned}$$

при выполнении которого запрос проходит по ребру $(v_{i_j}, v_{i_{j+1}})$. Ориентированный маршрут проводит запрос $x \in (0,1)$, тогда и только тогда, когда каждое его ребро проводит запрос x. Из этого следует, что запрос проходит по ориентированному маршруту M тогда и только тогда, когда он удовлетворяет системе, составленной из неравенств, заданных каждым ребром маршрута M. Множество запросов из интервала $(0,1)$, которые удовлетворяют такой системе может быть пустым множеством, либо точкой, либо подинтервалом интервала $(0,1)$.

Доказательство леммы 1 закончено.

Проводимостью вершины α назовем множество N_α, которое состоит из всех запросов x, $x \in (0,1)$, которые проходят в вершину α. Проводимость корневой вершины положим равной интервалу $(0,1)$. Заметим, что если информационный граф с переключательной нагрузкой — древовидный, то проводимость вершины, отличной от корневой, равна проводимости ориентированного маршрута, ведущего из корня в эту вершину.

Номер, который сопоставлен ориентированному маршруту, ведущему из корня в вершину α, назовем *кодом* вершины α.

Лемма 2. *В любом информационном графе с переключательной нагрузкой* U *из множества* D_I, $I = ((0,1), V, \rho_=, f)$, $V = (y_1, y_2, \ldots, y_n)$, *содержатся* $2n+1$ *висячие вершины. При упорядочении висячих вершин в соответствии с возрастанием их кода*

$$\alpha_1, \alpha_2, \alpha_3, \alpha_4, \ldots, \alpha_{2n}, \alpha_{2n+1},$$

их проводимости соответственно равны

$$(0, y_1), \{y_1\}, (y_1, y_2), \{y_2\}, \ldots, \{y_n\}, (y_n, 1).$$

При этом только висячие вершины с проводимостями вида $\{y_i\}$ являются листовыми, остальные висячие вершины не являются ни листовыми, ни переключательными.

Доказательство. По определению множества D_I в информационном графе с переключательной нагрузкой U используются только переключатели вида

$$\{p_{y1}, p_{y2}, \ldots, p_{y_n}\}.$$

Поэтому условие для проводимости любого ориентированного маршрута является системой из равенств и неравенств вида

$$x < y_i, x = y_i, x > y_i,$$

где $y_i \in V$. Решением этой системы может быть только либо пустое множество, либо множество вида $\{y_i\}$ или (y_j, y_{j+l}), где y_i, y_j и y_{j+l} либо записи библиотеки, либо точки 0 или 1.

Для любой записи библиотеки y_i в ИГПН U должен существовать лист с нагрузкой $\{y_i\}$ и записью $\{y_i\}$. Это следует из того, что запрос, равный y_i, должен достигать листовой вершины с записью $\{y_i\}$, и при этом другие запросы этого листа достигнуть не могут, то есть проводимость этого листа $\{y_i\}$.

Покажем, что проводимость у висячей вершины ИГПН U может быть либо точкой $\{y_i\}$, либо интервалом вида (y_j, y_{j+1}), который не содержит записей библиотеки.

Допустим противное, что существует висячая вершина, у которой проводимость имеет вид (y_j, y_{j+l}), $l > 0$. Тогда запрос x', равный записи библиотеки y_k, $y_j < y_k < y_{j+l}$, проходит по ориентированному маршруту от корня в эту вершину. Поскольку ИГПН решает задачу поиска I, то запрос x' должен достигать листовой вершины с нагрузкой $\{y_k\}$. Поскольку все ребра ориентированного маршрута переключательные, то листовой может быть только висячая вершина. Но рассматриваемая висячая вершина не может быть листовой, так как в нее проходят запросы, не равные записям библиотеки. Поэтому проводимость висячей вершины не может быть интервалом, содержащим записи библиотеки.

Предположим, что существует висячая вершина, у которой проводимость равна пустому множеству. Рассмотрим ориентированный маршрут, ведущий в нее из корневой вершины. Проводимости вершин этого ориентированного маршрута образуют цепочку вложенных

друг в друга множеств. Поскольку проводимость корневой вершины равна интервалу $(0,1)$, то в ориентированном маршруте существует такое ребро (α,β), что $N_\alpha \neq \varnothing$, $N_\beta = \varnothing$. Чтобы такая ситуация имела место, переключатель g_{y_i} вершины α должен удовлетворять условию

$$y_i \notin N_\alpha.$$

По определению множества D_I переключатель g_{y_i} можно сопоставить только одной вершине ИГПН U. Значит равенство $x = y_i$ может быть только в системе условий для проводимости ориентированного маршрута, который проходит через вершину α и выходит из нее по ребру с нагрузкой 2. Однако система любого такого ориентированного маршрута не совместна, а это значит, что в ИГПН U нет вершины с проводимостью $\{y_i\}$. Это противоречит тому, что ИГПН U решает ЗПИО I, так как в этом случае есть лист записью $\{y_i\}$ и проводимостью $\{y_i\}$.

Покажем, что для любого интервала (y_j, y_{j+1}), $j = 0, 1, \ldots, n$, $y_0 = 0, y_{n+1} = 1$, существует висячая вершина с такой проводимостью. Рассмотрим произвольный интервал (y_j, y_{j+1}). Рассмотрим ориентированный маршрут, по которому проходит запрос x' такой, что $x' \in (y_j, y_{j+1})$. Этот ориентированный маршрут не может заканчиваться во внутренней вершине ИГПН U, поскольку каждая внутренняя вершина является переключательной, и из нее выходит три ребра с нагрузками соответственно 1,2 и 3, при этом одно из этих ребер обязательно проводит запрос x'. Ориентированный маршрут, по которому проходит запрос x', ведет из корневой вершины в висячую вершину, обозначим ее γ. В проводимости вершины γ содержится x'. Поскольку проводимость висячей вершины либо состоит только из одной точки, равной записи библиотеки, либо является интервалом вида (y_l, y_{l+1}), $l = 0, 1, \ldots, n$, $y_0 = 0$, $y_{n+1} = 1$, то $N_\gamma = (y_j, y_{j+1})$.

Таким образом, мы показали, что в информационном графе с переключательной нагрузкой U есть $2n + 1$ висячих вершин, проводимости которых образуют множество

$$\{(0, y_1), \{y_1\}, (y_1, y_2), \{y_2\}, \ldots, \{y_n\}, (y_n, 1)\}.$$

Мы показали, что висячих вершин с проводимостью не из этого множества быть не может. Так же не может быть двух висячих вершин с одинаковой проводимостью, так как ИГПН U древовидный, и

каждый запрос может пройти только в одну висячую вершину. Висячая вершина с проводимостью, состоящей из одной точки, является листовой, а висячая вершина с проводимостью, равной интервалу, не является ни листовой, ни переключательной.

Для доказательства леммы осталось показать, что для двух висячих вершин α и β, таких что код No_α вершины α меньше кода No_β вершины β, верно

$$\forall a \in N_\alpha \; \forall b \in N_\beta : \; a < b.$$

Рассмотрим ориентированные маршруты из корневой вершины в вершины α и β. У этих ориентированных маршрутов общее начало. Поскольку код первой вершины меньше второй, значит из последней вершины общего начала запросы по первому маршруту выходят по ребру с меньшим номером, чем запросы, которые проходят по второму маршруту. Значит любой запрос из N_α меньше любого запроса из N_β.

Доказательстве леммы 2 закончено.

При применении алгоритма разметки к информационному графу с переключательной нагрузкой U $(n+1)$ висячие вершины, не являющиеся ни листовыми, ни переключательными, добавятся во множество листовых. Этим висячим вершинам сопоставится нагрузка, которая соответствует их проводимости. То есть, получится информационный граф U' с $2n+1$ листьями, проводимости которых равны

$$(0, y_1), \{y_1\}, (y_1, y_2), \{y_2\}, \ldots, \{y_n\}, (y_n, 1).$$

Нагрузка каждой листовой вершины равна ее проводимости, поэтому результат функционирования ИГПН U' совпадает с функцией ответов J' расширенной задачи поиска идентичных объектов $I' = (V',\, \rho_\in,\, f(x))$, где

$$V' = ((0, y_1), \{y_1\}, (y_1, y_2), \{y_2\}, \ldots, \{y_n\}, (y_n, 1)), \quad y_1 < y_2 < \cdots < y_n,$$

$$J'(x) = \begin{cases} \{y_i\}, & \text{если } \exists i \in [1..n] : y_i = x, \\ (y_j, y_{j+1}), & \text{если } \exists j \in [0..n] : y_j < x < y_{j+1}, \end{cases}$$

$y_0 = 0$, $y_{n+1} = 1$.

Таким образом, U' решает РЗПИО I' и принадлежит множеству $D_{I'}$. При применении алгоритма количество переключательных вершин и их нагрузка не изменились, поэтому

$$T_I(U) = T_{I'}(U').$$

Доказательство теоремы 1 закончено.

3. Доказательство теоремы 2

Рассмотрим произвольную ЗПИО $I = ((0,1),\ V,\ \rho_=,\ f)$, где $V = (y_1,\ y_2, \ldots,\ y_n)$, и ее расширение $I' = ((0,1),\ V',\ \rho_\in,\ f(x))$, где

$$V' = ((0,y_1), \{y_1\}, (y_1,y_2), \{y_2\}, \ldots, \{y_n\}, (y_n,1)), \quad y_1 < y_2 < \cdots < y_n.$$

Рассмотрим произвольный ИГПН U' из множества $S_{I'}$. Поскольку ИГПН U' решает I', то в нем содержатся как минимум $2n+1$ листовые вершины, проводимости и нагрузки которых соответственно равны

$$(0,y_1), \{y_1\}, (y_1,y_2), \{y_2\}, \ldots, \{y_n\}, (y_n,1).$$

Из множества листовых вершин удалим все листья, которым сопоставлена нагрузка в виде интервала. Получившийся ИГПН U будет содержать только листовые вершины с проводимостями и записями, имеющими вид $\{y_i\}$, $y_i \in V$. ИГПН U решает задачу поиска I, поскольку все запросы, равные записям библиотеки V, достигают соответствующих им листьев, а другие запросы в эти листья не проходят. Сложность ИГПН U равна сложности ИГПН U', поскольку переключательные вершины при перестройке не затрагивались. Из этого следует, что

$$\inf_{U' \in S_{I'}} T_{I'}(U') \geqslant \inf_{U \in S_I} T_I(U).$$

В статье [1] было показано, что для любой ЗПИО I во множестве D_I содержится оптимальный информационный граф с переключательной нагрузкой для задачи I, поэтому верно следующее тождество

$$\inf_{U \in S_I} T(U) = \inf_{U \in D_I} T(U).$$

В силу теоремы 1 любой ИГПН из множества D_I после алгоритма преобразования становится ИГПН из множества $D_{I'}$, при этом его сложность не изменяется. Поэтому верно следующее неравенство

$$\inf_{U \in D_I} T(U) \geqslant \inf_{U \in D_{I'}} T(U).$$

Поскольку множество $D_{I'}$ является подмножеством множества $S_{I'}$ верно

$$\inf_{U \in D_{I'}} T(U) \geqslant \inf_{U \in S_{I'}} T(U).$$

Получаем, что

$$\inf_{U \in S_I} T(U) = \inf_{U \in S_{I'}} T(U).$$

Из этого следует, что $T(I) = T(I')$, так как

$$T(I) = \inf_{U \in S_I} T(U) = \inf_{U \in S_{I'}} T(U) = T(I').$$

Доказательство теоремы 2 закончено.

Список литературы

[1] Кучеренко Н. С. Сложность поиска идентичных объектов в случайных базах данных // Интеллектуальные системы. 2007. Т. 11, вып. 1–4. С. 525–550.

[2] Кучеренко Н. С. О промежуточных функциях роста сложности поиска для случайных баз данных // Интеллектуальные системы. 2009. Т. 13, вып. 1–4. С. 361–395.

[3] Гасанов Э. Э., Кудрявцев В. Б. Теория хранения и поиска информации. М.: Физматлит, 2002.

[4] Кнут Д. Э. Искусство программирования. Т. 3. М.: Изд. дом «Вильямс», 2000.

[5] Татт У. Теория графов. М.: Мир, 1988.