МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ М.В. ЛОМОНОСОВА
МЕХАНИКО-МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ

На правах рукописи
УДК 519.71

Кучеренко Наталья Сергеевна

СЛОЖНОСТЬ ПОИСКА В СЛУЧАЙНЫХ БАЗАХ ДАННЫХ

01.01.09 — дискретная математика и математическая кибернетика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

МОСКВА — 2010

Работа выполнена на кафедре Математической теории интеллектуальных систем Механико-математического факультета Московского государственного университета имени М. В. Ломоносова.

Научный руководитель:	доктор физико-математических наук, профессор Гасанов Эльяр Эльдарович
Официальные оппоненты:	доктор физико-математических наук, профессор Зубков Андрей Михайлович
	кандидат физико-математических наук, доцент Применко Эдуард Андреевич
Ведущая организация:	Российский Государственный Гуманитарный Университет (РГГУ)

Защита диссертации состоится 26 ноября 2010 г. в 16 ч. 45 мин. на заседании диссертационного совета Д.501.001.84 при Московском государственном университете имени М. В. Ломоносова по адресу: Российская Федерация, 119991, Москва, ГСП-1, Ленинские горы, д. 1, МГУ, Механико-математический факультет, аудитория 14-08.

С диссертацией можно ознакомиться в библиотеке Механико-математического факультета МГУ М. В. Ломоносова (Главное здание, 14 этаж).

Автореферат разослан 26 октября 2010 г.

Ученый секретарь
диссертационного совета
Д.501.001.84 при МГУ,
доктор физико-математических наук,
профессор А. О. Иванов

Общая характеристика работы

Актуальность темы

Одним из актуальных направлений дискретной математики и математической кибернетики является направление хранения и поиска информации в базах данных. Развитие этого направления невозможно без тщательного анализа накопленного опыта и построения математической модели баз данных. Исследование математической модели помогает решать задачи, позволяющие повысить эффективность поиска в базах данных.

Под базой данных в работе понимается способ хранения и представления информации и соответствующие этим способам алгоритмы поиска информации. Самой распространенной задачей поиска, которая встречается в любой базе данных, является задача поиска по ключу[1]. Суть ее состоит в том, что любой объект базы данных имеет свой уникальный ключ. Это может быть порядковый номер, уникальное имя, или уникальное значение некоторого поля, например, номер паспорта. Задача состоит в том, чтобы по заданному в запросе ключу найти в базе данных объект с этим ключом.

Более формально задачу поиска по ключу можно описать следующим образом[2]. Имеется некоторое множество объектов Y, на котором введен линейный порядок. Данные — это конечное подмножество V, $V \subset Y$. Множество V далее будет называться также библиотекой. Множество запросов X совпадает с множеством Y. Требуется по произвольному запросу из множества X найти в библиотеке V объект равный запросу, если такого объекта нет — указать в какой промежуток между данными попал запрос. Полагается, что для решения этой задачи можно сравнивать любые два объекта из множеств X и Y.

Рассматривается случай, когда множества X и Y представляют собой интервал $(0, 1)$ и база данных — статическая, то есть библиотека V фиксирована. Предполагается, что к статической базе данных происходит многократное обращение с запросами на поиск по ключу, поэтому при ее проектировании

[1] Кнут Д. Э. *Искусство программирования.* Издательский дом "Вильямс", Москва, 2000, Т. 3.

[2] Гасанов Э. Э., Кудрявцев В. Б. *Теория хранения и поиска информации.* Физматлит, Москва, 2002.

внимание акцентируется на организации данных и алгоритме поиска, минимизирующих среднее время поиска.

В данной работе задача поиска по ключу исследуется с позиции информационно-графовой модели данных[2]. В этой модели структура данных задается управляющей системой — информационным графом (ИГ), который представляет собой ориентированный граф, ребра и вершины которого нагружены элементами данных и функциями. Алгоритм поиска — это "волновой" процесс на графе, который управляется нагрузочными функциями. Под сложностью информационного графа понимается среднее время поиска. Информационный граф, на котором достигается минимум сложности, называется оптимальным.

В пятидесятые годы двадцатого века возникла идея представлять алгоритмы для задачи поиска по ключу с помощью деревьев. В 1959 году Э. Н. Гильберт и Э. Ф. Мур показали, что можно построить оптимальное дерево поиска за $O(n^3)$ шагов (n — мощность библиотеки), и привели оценки сложности такого дерева поиска[3]. Оптимальным называлось дерево поиска, имеющее минимальную сложность среди всех деревьев. В 1971 году Д. Э. Кнут показал, что построение оптимального дерева поиска можно улучшить до порядка $O(n^2)$ шагов[4]. Дальнейшие упрощения методов построения были произведены в 1977 А. М. Гарсия и М. Л. Вочем[5]. Их метод позволяет построить оптимальное дерево поиска за $O(n \cdot \log_2 n)$ шагов.

В общем случае оптимальный информационный граф не является древовидным, поэтому возникает вопрос, есть ли среди множества древовидных информационных графов оптимальный и применимы ли имеющиеся результаты к построению оптимального ИГ. В случае утвердительного ответа, возникает следующий вопрос о сложности оптимального информационного графа.

Любую задачу поиска по ключу можно решить с помощью информационного графа, реализующего бинарный поиск, со сложностью равной лога-

[3] Gilbert E. N, Moore E. F., Variable-length binary encodings. *Bell System Tech. J.,* — 1959. **38**, — 933–968.

[4] Knuth D. E., Optimum binary search trees. *Acta Informatica,* — 1971. **1**, — 14–25.

[5] Garsia A. M., Wachs M. L., A new algorithm for minimum cost binary trees. *SICOMP,* — 1977. **4,** — 622– 642.

рифму от мощности библиотеки V. Сложность же оптимального ИГ в зависимости от конкретной задачи поиска может быть как логарифмом, так и константой. Поэтому возникает вопрос о поведении средней сложности оптимального информационного графа для случайных библиотек.

Цель работы

Целью работы является исследование структуры оптимального информационного графа для решения задачи поиска по ключу и изучение поведения его средней сложности для случайных библиотек.

Основные методы исследования

В работе используются методы теории графов, теории вероятностей, математического анализа, комбинаторики, алгебры.

Научная новизна

Результаты работы являются новыми и состоят в следующем.

1. Показано, что для любой задачи поиска по ключу существует оптимальный информационный граф с древовидной структурой, в котором количество всех используемых операций сравнения равняется мощности библиотеки. Описан условный алгоритм построения такого ИГ. Приведены универсальные оценки сложности оптимального ИГ.

2. Рассмотрено поведение сложности оптимального ИГ для двух «равномерных» библиотек. В первом случае библиотека представляет собой равномерную сетку на интервале $(0, 1)$, во втором случае библиотека — случайный равномерно распределенный вектор, и запросы также распределены равномерно. В первом случае установлено, что сложность оптимального ИГ имеет асимптотику двоичного логарифма от мощности библиотеки n. Во втором случае показано, что средняя сложность оптимального ИГ также имеет асимптотику $\log_2 n$, $n \to \infty$, и получена нижняя оценка средней сложности оптимального ИГ, которая отличается от $\log_2 n$ на константу.

3. Исследовано поведение средней сложности оптимального ИГ в случае, когда распределение данных и запросов может быть отлично от равномерного. Распределение данных задается функцией плотности g, а распределение запросов — функцией плотности f. При слабых ограничениях на f и g доказана нижняя оценка средней сложности оптимального ИГ. Получены условия для f и g, при которых средняя сложность оптимального ИГ имеет порядок логарифма от мощности библиотеки. Уточнены эти условия до получения асимптотики такой сложности.

4. Рассмотрены случаи, когда средняя сложность оптимального информационного графа является ограниченной функцией при увеличении мощности библиотеки. Показано, что для любого отрезка вида $[b, b+2]$, $b \in \mathbb{R}$, $b > 1$, можно построить такие функции плотности распределения запросов f и данных g, для которых средняя сложность оптимального ИГ не выходит за пределы отрезка при увеличении мощности библиотеки.

5. Рассмотрены случаи, когда средняя сложность оптимального ИГ является неограниченно возрастающей функцией по порядку меньшей, чем логарифм. Описано семейство S возможных асимптотик и семейство S^* возможных порядков функций роста, которые являются неограниченно возрастающими, но по порядку меньшими, чем логарифм от мощности библиотеки.

Теоретическая и практическая ценность

Работа носит теоретический характер и может быть полезна специалистам по синтезу и сложности управляющих систем. Результаты работы могут быть использованы при проектировании баз данных.

Апробация работы

Результаты диссертации неоднократно докладывались на семинарах механико-математического факультета МГУ им. М. В. Ломоносова: на семинаре «Вопросы сложности алгоритмов поиска» под руководством профессора Э. Э. Гасанова (2006–2009 гг.), на семинаре «Теория автоматов» под руководством академика В. Б. Кудрявцева (2007–2009 гг.).

Они докладывались также на следующих конференциях: IX международная конференция «Интеллектуальные системы и компьютерные науки» (Москва, 2006 г.), IX международный семинар «Дискретная математика и ее приложения», посвященный 75-летию со дня рождения академика О. Б. Лупанова (Москва, 2007 г.), международная конференция студентов, аспирантов и молодых ученых «Ломоносов» (Москва, 2007 г., 2008 г., 2009 г.), международная конференция «Современные проблемы математики, механики и их приложений», посвященная 70-летию ректора МГУ академика В. А. Садовничего (Москва, 2009 г.), X Международный семинар «Дискретная математика и ее приложения» (Москва, 2010 г.), научная конференция «Ломоносовские чтения» (Москва, 2007 г., 2008 г., 2010 г.).

Публикации

Основные результаты диссертации опубликованы в 6 работах автора, список которых приведен в конце автореферата [1–6].

Структура и объем диссертации

Диссертация состоит из введения и трех глав. Объем диссертации 179 страниц. Список литературы содержит 23 наименования.

Краткое содержание работы

Во введении приведен краткий исторический обзор по тематике работы, изложены цели и методы исследования, а так же структура диссертации. Затем для каждой главы содержательно описываются полученные в ней результаты.

В **первой главе** приводится две формализации задачи поиска по ключу — расширенная задача поиска идентичных объектов (РЗПИО), когда при отсутствии объекта в базе данных указывается позиция запроса относительно данных, и задача поиска идентичных объектов (ЗПИО), когда позиция запроса не указывается. Операция сравнения формализуется как переклю-

чатель. Алгоритм поиска, который использует только операции сравнения, формализуется как информационный граф с переключательной нагрузкой (ИГПН). Также в главе дается определение сложности ИГПН и определение оптимального ИГПН. Показывается, что сложности оптимальных информационных графов с переключательной нагрузкой для обоих способов формализации равны. Описывается алгоритм преобразования ИГПН для решения ЗПИО в ИГПН для решения РЗПИО, который не меняет сложность информационного графа с переключательной нагрузкой. Эти результаты позволяют исследовать сложность задачи поиска по ключу в более «простой» формализации задачи поиска идентичных объектов.

Также в этой главе исследуется структура оптимального ИГПН, доказываются универсальные оценки для его сложности и описывается алгоритм его построения. Приводится определение ИГПН бинарного поиска, доказываются универсальные оценки для его сложности и описывается алгоритм его построения. По разделам это распределено следующим образом.

В **разделе 1.1** 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 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 является случайной величиной, принимающей значения из интервала $(0,1)$. Распределение запросов задается функцией плотности f. 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 четверка $((0,1), V', \rho_\in, f(x))$, 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 $x \in y$. Запрос x является случайной величиной, принимающей значения из интервала $(0,1)$. Распределение запросов задается функцией плотности f. Введем обозначения $y_0 = 0$, $y_{n+1} = 1$. Отношение $\rho_\in$ задает функцию ответов $J'(x)$, определенную на множестве запросов $(0,1)$, следующим образом

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

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

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

Также в разделе дается определение информационного графа. Приведем содержательное описание этого определения. Информационный граф — это ориентированный граф, 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).

Множество ИГПН, которые решают задачу поиска I, обозначим через S_I. 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. *Сложностью ИГПН U из множества S_I* называется математическое ожидание величины $T(U, x)$, которое можно записать в виде

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

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.*

В **разделе 1.2** приводятся результаты главы и необходимые для их формулировки понятия. Вводится понятие древовидного ИГПН и определяется множество D_I, состоящее из древовидных ИГПН, решающих задачу поиска I и удовлетворяющих жестким требованиям к структуре.

В **разделе 1.3** приводится доказательство результата о том, что для любой задачи поиска идентичных объектов I можно построить оптимальный ИГ принадлежащий множеству D_I.

Теорема 1. *Для любой задачи поиска идентичных объектов I существует оптимальный информационный граф с переключательной нагрузкой, который содержится во множестве D_I.*

Для расширенной задачи поиска идентичных объектов в разделе доказывается аналогичный результат. Однако он не выводится заново, а получается из теоремы 1 с помощью *алгоритма преобразования* ИГПН из множества D_I, где I — задача поиска идентичных объектов, в ИГПН из множества $D_{I'}$, где РЗПИО I' — расширение I, и теоремы 3 о том, что сложности ЗПИО I и РЗПИО I' равны.

Теорема 2. *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' из множества $D_{I'}$, где РЗПИО I' — расширение задачи поиска идентичных объектов I. При этом сложность ИГПН U равна сложности получившегося ИГПН U'*

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

Теорема 3. *Сложность любой задачи поиска идентичных объектов I и ее расширения I' равны*

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

В силу теоремы 3 сложность задачи поиска по ключу не зависит от выбора одного из двух способов формализации. Поэтому далее в работе исследуется более «простая» формализация в виде ЗПИО, и при необходимости даются пояснения как эти результаты перенести для РЗПИО.

В **разделе 1.4** приводится алгоритм построения оптимального ИГПН, основанного на результатах, полученных в 1959 году Э. Н. Гильбертом и Э. Ф. Муром[3]. Под алгоритмом построения понимается условный алгоритм с операциями сложения и нахождения минимального для пар вещественных чисел. Сложность условного алгоритма — количество таких операций. Доказывается корректность алгоритма построения и показывается, что его сложность по порядку не больше куба от мощности библиотеки.

Теорема 4. *Существует условный алгоритм, который для любой задачи поиска идентичных объектов*

$$I = ((0,1), V, \rho_=, f),\ V = \{y_{i_1}, y_{i_2}, \ldots, y_{i_n}\},$$

строит оптимальный ИГПН из множества D_I. При этом сложность построения по порядку не больше n^3, где $n = |V|$.

Известно, что задачу поиска идентичных объектов можно решить с помощью бинарного поиска. В **разделе 1.5** дается определение информационного

графа бинарного поиска и показывается, что для любой задачи поиска идентичных объектов I сложность ИГ бинарного поиска не больше $]\log_2(n+1)[$ и не меньше $[\log_2(n+1)]$, где n — мощность библиотеки задачи поиска. Алгоритм построения такого ИГ требует линейного от мощности библиотеки числа операций сложения, деления и взятия целой части над вещественными числами.

Теорема 5. *Существует условный алгоритм, который для любой задачи поиска идентичных объектов*

$$I = ((0,1), V, \rho_=, f),\ V = \{y_{i_1}, y_{i_2}, \ldots, y_{i_n}\},$$

строит информационный граф бинарного поиска U_B из множества D_I. При этом сложность построения по порядку не больше n, где $n = |V|$. Для построенного ИГПН U_B верны следующие оценки

$$[\log_2(n+1)] \leqslant T_I(U_B) \leqslant]\log_2(n+1)[.$$

В **разделе 1.6** доказывается, что сложность любой задачи поиска не меньше единицы и не больше верхней оценки сложности ИГ бинарного поиска $]\log_2(n+1)[$. При этом строится пример задачи поиска, для которой нижняя оценка достигается, и строится задача поиска со сложностью не меньше чем $\log_2(n+1)$, где n — мощность библиотеки задачи поиска.

Теорема 6. *Для сложности любой задачи поиска идентичных объектов $I = ((0,1),\ V,\ \rho_=,\ f)$ верно*

$$1 \leqslant T(I) \leqslant]\log_2(n+1)[,$$

где $n = |V|$. При этом для любого $n \in \mathbb{N}$ существуют задачи поиска идентичных объектов $I'_n = ((0,1),\ V',\ \rho_=,\ f')$, $|V'| = n$, и $I''_n = ((0,1),\ V'',\ \rho_=,\ f'')$, $|V''| = n$, такие, что

$$T(I'_n) = 1, \quad T(I''_n) \geqslant \log_2(n+1).$$

В силу теоремы 6 сложность оптимального ИГПН может быть как логарифмом от мощности библиотеки, так и константой в зависимости от конкретной задачи поиска идентичных объектов. В следующих главах автором

исследуется поведение средней сложности оптимального информационного графа с переключательной нагрузкой на классах задач.

Вторая глава посвящена классам задач поиска, для которых средняя сложность оптимального ИГПН имеет порядок логарифма от мощности библиотеки. В **разделе 2.1** приводятся основные понятия и результаты главы.

В **разделе 2.2** рассмотрены два «равномерных» класса задач поиска идентичных объектов. Библиотека первого класса задач представляет собой равномерную сетку на интервале $(0,1)$, библиотека второго — вариационный ряд равномерно распределенных случайных величин.

Первый класс задач поиска обозначим через I_n^f

$$I_n^f = ((0,1), V, \rho_=, f),\ V = \left(\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}\right),\ n \in \mathbb{N}.$$

Автором получен следующий результат.

Теорема 7. *Для любой функции плотности распределения* $f(x)$

$$T(I_n^f) \sim \log_2 n \quad (n \to \infty).$$

При этом если функция плотности ограничена константой c, $c > 0$, *то*

$$T(I_n^f) > \log_2(n+1) - \log_2\log_2(n+1) - 1 - c.$$

Библиотека V второго рассматриваемого в работе «равномерного» класса задач поиска представляет собой вектор

$$V = (y_{(1)}, y_{(2)}, \ldots, y_{(n)}),$$

где $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ — вариационный ряд, составленный из независимых равномерно распределенных на интервале $(0,1)$ случайных величин $y_1, y_2, \ldots, y_n$.

Обозначим через $T_n(V)$ случайную величину, которая при реализации $\widehat{V}$ вектора V равна сложности ЗПИО $((0,1),\ \widehat{V},\ \rho_=,\ \chi(0,1))$, где функция плотности распределения запросов $\chi(0,1)$ является функцией плотности равномерного распределения на интервале $(0,1)$.

Автором показано, что для любого $n, n \in \mathbb{N}$, математическое ожидание сложности оптимального ИГПН для равномерно распределенных запросов и

для случайной библиотеки V, являющейся вариационным рядом равномерно распределенных случайных величин, незначительно отличается от сложности логарифмического поиска, а именно верны следующие оценки

Теорема 8. *Пусть* $\mathbf{M}_V(T_n(V))$ — *математическое ожидание* $T_n(V)$, *тогда для любого* $n \in \mathbb{N}$

$$]\log_2(n+1)[\geqslant \mathbf{M}_V(T_n(V)) \geqslant \log_2(n+1) - \frac{1-\gamma_{n+1}}{\ln 2},$$

где $\gamma_n = \sum_{i=1}^{n} \frac{1}{i} - \ln n$.

Последовательность γ_n при $n \to +\infty$ — сходящаяся. Предел этой последовательности называется постоянной Эйлера γ, $\gamma = 0,57\ldots$

Разделы 2.3 и 2.4 посвящены классам задач поиска, для которых распределение данных и запросов может быть отлично от равномерного. Случайная библиотека V задается с помощью функции плотности g как вектор

$$V = (y_{(1)}, y_{(2)}, \ldots, y_{(n)}),$$

где $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ — вариационный ряд, составленный из независимых случайных величин $y_1, \ldots, y_n$ с функцией плотности распределения $g(x)$. Обозначим через $T_n^{(f,g)}(V)$ случайную величину, которая при реализации $\widehat{V}$ случайной библиотеки V, заданной с помощью функции плотности g, равна сложности ЗПИО $((0,1),\ \widehat{V},\ \rho_{=},\ f)$, где f — функция плотности распределения запросов.

В **разделе 2.3** изучается при каких условия на функции плотности f и g математическое ожидание величины $T_n^{(f,g)}(V)$ имеет порядок логарифма от мощности библиотеки.

Автором показано, как при известных ограничениях на функции плотности получить нижнюю оценку средней сложности оптимального ИГПН на классе задач.

Теорема 9. *Пусть для функций плотности* f *и* g *существует* $s, s \in \mathbb{N}$, *не пересекающихся интервалов* $(a_i, b_i) \subset (0,1), i = 1, \ldots, s$, *таких что*

$$\forall i,\ i = 1, \ldots, s,\ \exists e_1^i, e_2^i, k_1^i, k_2^i \in \mathbb{R}^+ :$$

$$\forall x \in (a_i, b_i) \quad e_1^i \geqslant f(x) \geqslant e_2^i > 0, \quad k_1^i \geqslant g(x) \geqslant k_2^i > 0.$$

Тогда при $n \to \infty$

$$]\log_2(n+1)[\geqslant \mathbf{M}_V(T_n^{(f,g)}(V)) \geqslant \sum_{i=1}^{s} u_i \cdot \log_2(1 + r_i n) - \sum_{i=1}^{s} v_i + O\left(\frac{1}{n^{1/3}}\right),$$

где $u_i = \frac{e_2^i \cdot r_i}{k_1^i}$, $v_i = \frac{e_1^i \cdot r_i}{k_2^i}$, $r_i = \int_{a_i}^{b_i} g(x)\,dx, i = 1, \ldots, s$.

Следствием из этой теоремы является достаточное условие на функции плотности, при котором средняя сложность оптимального ИГПН имеет порядок логарифма от мощности библиотеки.

Следствие 9.1. *Если существует невырожденный интервал* (a, b), *на котором функции плотности* f *и* g *одновременно ограничены и отделены от нуля, то*

$$\mathbf{M}_V(T_n^{(f,g)}(V)) \asymp \log_2 n \quad (n \to \infty),$$

где $n = |V|$.

В **разделе 2.4** исследуются классы задач поиска, для которых автором получена асимптотика средней сложности оптимального ИГПН.

Обозначим носитель функции f через $supp(f)$, $supp(f) = \{x \in (0,1) : f(x) \neq 0\}$. Назовем функцию плотности f *функцией с конечно-интервальным носителем,* если $supp(f)$ имеет вид

$$supp(f) = \sqcup_{i=1}^{s}(a_i, b_i) \sqcup K, \quad K \subseteq \{a_1, \ldots, a_s, b_1, \ldots, b_s\},$$

где $a_i < b_i$, $b_i < a_{i+1}$, $\forall i = 1 \ldots (s-1)$, $a_s < b_s$.

Автором доказана следующая теорема

Теорема 10. *Пусть функции плотности* f *и* g *имеют конечно-интервальный носитель, ограничены и отделены от нуля на множестве* $supp(f)$ *и* $supp(g)$ *соответственно, и существует интервал, на котором функции отделены от нуля одновременно. Пусть также функции* f *и* g *интегрируемы по Риману. Тогда*

$$\mathbf{M}_V(T_n^{(f,g)}(V)) \sim \log_2 n \cdot \int_B f(x)\,dx \quad (n \to \infty),$$

где $B = supp(g)$.

В **третьей главе** исследуются классы задач поиска идентичных объектов, для которых математическое ожидание сложности оптимального информационного графа с переключательной нагрузкой имеет порядок, отличный от логарифма.

В **разделе 3.1** приводятся основные понятия и результаты главы.

Раздел 3.2 посвящен классам задач поиска идентичных объектов, для которых математическое ожидание сложности оптимального ИГПН при увеличении мощности случайной библиотеки n является ограниченной функцией. В работе автором показано существование таких функций плотности f и g, что математическое ожидание сложности оптимального ИГПН $T_n^{(f,g)}(V)$ является ограниченной функцией при $n \to \infty$. При этом верна следующая теорема

Теорема 11. *Для любого числа* $n' \in \mathbb{N}$ *существуют функции плотности* f' *и* g' *такие, что*

$$\mathbf{M}_V(T_{n'}^{(f',g')}(V)) = 1.$$

Для любого вещественного числа b, $b > 1$, *существуют такие функции плотности* f'' *и* g'', *что*

$$\exists n_0 \in \mathbb{N}: \quad \forall n > n_0 \quad b + 2 > \mathbf{M}_V(T_n^{(f'',g'')}(V)) \geqslant b.$$

Раздел 3.3 посвящен классам задач поиска идентичных объектов для которых математическое ожидание сложности оптимального ИГПН является неограниченно возрастающей функцией по порядку меньшей, чем логарифм. Автором исследуются возможные асимптотики и порядки функций роста средней сложности поиска для таких классов задач.

Положительная, возрастающая функция $r(x)$ называется *сохраняющей асимптотику,* если выполнено условие

$$\forall c \in \mathbb{R} \quad r(x + c) \sim r(x) \quad (x \to \infty).$$

Семейство S возможных асимптотик промежуточных функций роста состоит из функций вида $r(\log_2 \log_2(n))$, где неограниченно возрастающая, положительная и дифференцируемая функция $r(x)$, определенная на интервале

$(x_0, +\infty), x_0 \geqslant 0$, сохраняет асимптотику и имеет в качестве производной монотонную, положительную и непрерывную функцию $r'(x)$, удовлетворяющую условию:

$$\exists \alpha > 0, \alpha \in \mathbb{R} : \overline{\lim_{x \to +\infty}} \frac{r'(x)}{x^\alpha} < 1.$$

Все функции из S являются неограниченно возрастающими и имеют порядок меньше чем $\log_2 n$ при $n \to \infty$.

Автором показано, что для функций семейства S верна следующая теорема.

Теорема 12. *Для любой функции $r(\log_2 \log_2(n))$ из семейства S существуют функции плотности f и g такие, что для математического ожидания сложности оптимального ИГПН верно*

$$\mathbf{M}_V(T_n^{(f,g)}(V)) \sim r(\log_2 \log_2(n)) \quad (n \to \infty).$$

В качестве примера показано, что множество

$$\{c \cdot (\underbrace{\log_2 \ldots \log_2}_{i+1} n)^\alpha | i \in \mathbb{N}, \alpha \in \mathbb{R}_+, c \in \mathbb{R}_+\},$$

где $\mathbb{R}_+$ — множество положительных вещественных чисел, является подсемейством для S.

Следствие 12.1. *Для любого натурального i, для любых положительных и вещественных α и c существуют функции плотности f и g такие, что*

$$\mathbf{M}_V(T_n^{(f,g)}(V)) \sim c \cdot (\underbrace{\log_2 \ldots \log_2}_{i+1} n)^\alpha \quad (n \to \infty).$$

Положительная, возрастающая функция $r(x)$ называется *сохраняющей порядок,* если

$$\forall c \in \mathbb{R}, c \neq 0, \quad r(c \cdot x) \asymp r(x) \quad (x \to \infty).$$

Семейство S^* возможных порядков промежуточных функций роста состоит из функций вида $r(\log_2(n))$, где неограниченно возрастающая, положительная и дифференцируемая функция $r(x)$, определенная на интервале $(x_0, +\infty)$,

$x_0 \geqslant 0$, сохраняет порядок и имеет в качестве производной монотонную, положительную и непрерывную функцию $r'(x)$, удовлетворяющую условию:

$$\exists \alpha \in \mathbb{R}, 0 < \alpha < 1 : \overline{\lim_{x \to +\infty}} \frac{r'(x)}{x^{\alpha-1}} \leqslant 1.$$

Все функции из S^* являются неограниченно возрастающими и имеют порядок меньше чем $\log_2 n$ при $n \to \infty$. В отличие от класса S, в классе S^* есть функции по порядку большие чем любая функция из S, например $(\log_2 n)^\alpha$, $0 < \alpha < 1$. Автором показано, что для функций семейства S^* верна следующая теорема

Теорема 13. *Для любой функции $r(\log_2(n))$ из семейства S^* существуют функции плотности f и g, такие что для математического ожидания сложности оптимального ИГПН $\mathbf{M}_V(T_n^{(f,g)}(V))$ верно*

$$\mathbf{M}_V(T_n^{(f,g)}(V)) \asymp r(\log_2(n)) \quad (n \to \infty).$$

В качестве примера показано, что множество

$$\{(\log_2 n)^\alpha | 0 < \alpha < 1, \alpha \in \mathbb{R}\},$$

является подсемейством для S^*.

Следствие 13.1. *Для любого вещественного α, $0 < \alpha < 1$, существуют функции плотности f и g, такие что*

$$\mathbf{M}_V(T_n^{(f,g)}(V)) \asymp (\log_2 n)^\alpha \quad (n \to \infty).$$

Благодарности

Автор выражает благодарность своему научному руководителю профессору Гасанову Эльяру Эльдаровичу за постановку задачи и постоянное внимание к работе и академику Кудрявцеву Валерию Борисовичу за ценные советы и замечания.

Автор также благодарен всему коллективу кафедры Математической теории интеллектуальных систем за поддержку и внимание.

Публикации автора по теме диссертации

[1] Кучеренко Н. С. Сложность поиска идентичных объектов в случайных базах данных. *Интеллектуальные системы* (2007) **11,** № 1–4, с. 525–550.

[2] Кучеренко Н. С. О промежуточных функциях роста сложности поиска для случайных баз данных. *Интеллектуальные системы* (2009) **13,** № 1–4. с. 361–395.

[3] Кучеренко Н. С. О сложности поиска идентичных объектов для случайных баз данных. В кн.: *Материалы IX Международной конференции "Интеллектуальные системы и компьютерные науки".* Механико-математический факультет МГУ, Москва, 2006, Т. 1, с. 171.

[4] Кучеренко Н. С. Оценки сложности поиска идентичных объектов для случайных баз данных. В кн.: *Материалы IX Международного семинара "Дискретная математика и ее приложения", посвященного 75-летию со дня рождения академика О. Б. Лупанова.* Механико-математический факультет МГУ, Москва, 2007, с. 329–331.

[5] Кучеренко Н. С. О средней сложности поиска идентичных объектов для случайных баз данных. В кн.: *Материалы международной конференции "Современные проблемы математики, механики и их приложений", посвященная 70-летию ректора МГУ академика В. А. Садовничего.* Механико-математический факультет МГУ, Москва, 2009, с. 362.

[6] Кучеренко Н. С. Асимптотика промежуточных функций роста сложности поиска для случайных баз данных. В кн.: *Материалы X Международного семинара "Дискретная математика и ее приложения".* Механико-математический факультет МГУ, Москва, 2010. с. 373–375.