На правах рукописи

Тищенко Владимир Александрович

Методы построения многоуровневого классификатора по лексикографическому признаку применительно к ключевому уровню массива ООСУБД НИКА

Специальность 05.13.01 – Системный анализ, управление и обработка информации (информационно-вычислительное обеспечение)

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Москва – 2020

Работа выполнена в отделе №94 «Организация банков данных» отделения «Математическое обеспечение вычислительной техники» Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук»

Научный руководитель:	доктор технических наук **Соловьёв Александр Владимирович**
Официальные оппоненты:	**Портнов Евгений Михайлович** доктор технических наук, федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Московский институт электронной техники», профессор по кафедре информатики и программного обеспечения вычислительных систем **Сафонова Ирина Евгеньевна** доктор технических наук, федеральное государственное автономное образовательное учреждение высшего образования «Российский университет транспорта (МИИТ)», профессор кафедры «Вычислительные системы и сети» Института управления и информационных технологий
Ведущая организация:	Федеральное государственное автономное образовательное бюджетное учреждение высшего образования «Южный федеральный университет»

Защита состоится 15 марта 2021 г. в 12 часов 30 минут на заседании диссертационного совета. Д 002.073.04 на базе Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН) по адресу: 117312, г. Москва, проспект 60-летия Октября, 9 (конференц-зал, 1-й этаж).

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН по адресу: г. Москва, ул. Вавилова, д. 40 и на официальном сайте ФИЦ ИУ РАН http://www.frccsc.ru.

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, просьба высылать по адресу: 119333, г. Москва, ул. Вавилова, д. 44, кор. 2, учёному секретарю диссертационного совета Д 002.073.04.

Автореферат разослан 14 января 2021 г.
Телефон для справок: +7 (499) 135-51-64.

Ученый секретарь
диссертационного совета
Д 002.073.04,
доктор технических наук,
профессор

В.Н. Крутько

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Проблема быстрого интерактивного поиска данных в массиве в форме классификатора по лексикографическому признаку формулируется в следующем виде. Существует информационная потребность в оптимальном классификаторе по причине увеличения объёмов данных и повышение сложности структуры данных. Сформированные в виде баз данных они составляют основу информационных систем в интернет, поэтому при работе с ними, особенно с использованием сенсорных устройств без клавиатуры, для пользователя критично количество “кликов” при поиске ключа. С другой стороны, проблема построения классификатора не решена, несмотря на многочисленные попытки в виде однобуквенных, двухбуквенных алфавитных классификаторов или диапазонных классификаторов и т.п. Такая противоречивая ситуация и определяет необходимость решения важной научно-технической проблемы построения оптимального классификатора как альтернативы к запросной системе. При изменении БД необходимо динамически перестраивать классификаторы, являющиеся способом доступа к ключевым массивам. В более простых случаях, когда, например, строится классификатор для БД товаров потребления достаточно разделить товары на различные классы. В случае фактографических БД со сложной структурой, в которых есть более сотни полей и по этим полям построены индексы с большим количеством ключей для объектов одного класса, например по фамилиям, для организации доступа к таким ключевым массивам необходимо построение классификаторов по лексикографическому признаку.

Что касается проблемы организации доступа к базе данных (БД), то обычная практика такова, что БД хранения рассматривается отдельно от системы доступа. При этом обычно исследуются запросы, их оптимизация, среднее время, минимальное время, минимизация этих временных характеристик. Остаётся не изученной проблема интерактивного доступа и проблема оптимизации такого доступа. Распространённой формой интерактивного доступа к базам данных служит поисковая форма, содержащая запросные поля. Существующий способ поиска в ключевом массиве в виде поля ввода необходимо дополнить **классификатором по лексикографическому признаку**, который предоставляет визуальный просмотр класса ключей на выбранный префикс и даёт возможность обнаружить наличие или отсутствие префиксов искомых ключей, что ускоряет поиск. Такая же проблема интерактивного доступа существует и для ключевых массивов больших объёмов структурно сложных БД, опубликованных в интернет. Широкое распространение мобильных устройств, не оснащённых клавиатурой, ставит проблему удобного интерфейса с пользователем. Классификатор предоставляет такой интерфейс в виде иерархического списка существующих префиксов-подсказок к ключам массива.

Стандартным интерфейсом для просмотра реляционных баз данных, в том числе в интернет, является табличная форма представления данных с выбором столбцов (например, Oracle, MS Access и др.). В случае поисковых систем в интернет (google, yahoo, yandex и др. — проиндексированы миллиарды страниц, индексы содержат петабайты данных) стандартным способом отображения результатов поиска является список, разделённый на части. Навигация в списке с большим количеством текстовых ссылок на документы при поиске требуемой информации может быть упрощена посредством классификатора. Для более удобного способа навигации по данным, представленным в виде таблицы, списка текстовых ключей или записей с текстовыми ключами служит классификатор, предоставляющий способ обзора массива текстовых ключей и позволяющий найти по префиксам искомые ключи, упорядоченные в алфавитном порядке. В случае библиографических баз данных встречаются такие классификаторы, но чаще всего однобуквенные, что является недостаточным при больших объёмах массивов, разделённых на группы ключей. Если брать группы по 20 ключей, то для

оптимальной навигации по массиву объёмом в несколько тысяч ключей уже недостаточно однобуквенного указателя. Под оптимальной навигацией понимается минимальное число переходов пользователя в классификаторе по ссылкам при поиске по ключу. Примером многоуровневого классификатора может служить БД "Жертв политического террора в СССР" (3,1 млн. имён — 2019г.). Здесь применяется трёхуровневый алфавитный классификатор. Однако он не оптимален в том смысле, что на втором уровне здесь происходит деление на трёхбуквенные диапазоны ключей, которые не позволяют определить наличие или отсутствие префиксов внутри диапазонов, а для этого необходимо просматривать третий уровень классификатора. Третий уровень представляет собой список фамилий, на каждую из которых присутствует различное число ключей. Такой классификатор не оптимизирован по общему числу операций, а префиксы выбраны "по смыслу" и им можно пользоваться до тех пор, пока не будут добавлены новые ключи в массив, т.к. он становится не удобным в использовании из-за большого количества одинаковых фамилий. В результате увеличения числа записей в новой версии БД используется только форма ввода, т.к. построенный классификатор стал непригодным из-за большего объёма данных. Вышеизложенное описание проблемы и приведённые примеры показывают, что для БД больших объёмов является **актуальным** построение оптимального классификатора, который строится автоматически по исходному массиву при минимизации общего числа операций в классификаторе. **Актуальность темы** можно охарактеризовать следующими положениями:

- Электронная публикация структурно сложных БД больших объёмов.
- Изменение БД требует динамической перестройки классификатора.
- Широкое распространение мобильных устройств без клавиатуры.
- Список не даёт обзора ключей на предмет существования или отсутствия определённых классов ключей на заданные префиксы.
- Системы с многоуровневыми классификаторами по лексикографическому признаку существуют и востребованы, но не оптимизированы по числу переходов пользователя.
- Классификаторы по различным категориям в виде онтологий применяются в информационных системах и являются естественными элементами пользовательского интерфейса.

Также можно сделать вывод о том, что данная область **малоизучена**, т.к. не существует на данный момент стандартных средств построения такого классификатора, библиографические БД ограничиваются однобуквенным указателем, существующие системы не используют специально разработанных методов построения классификаторов. В диссертации предлагается решение проблемы в виде метода построения оптимального классификатора, который является **новаторским**. Метод имеет существенные достоинства по **сравнению** с другими методами интерактивного доступа.

- Список, разделённый на части, не даёт возможности обзора существующих ключей массива и быстрой навигации.
- Классификатор дополняет онтологии, используемые для формализации областей знаний.
- Приведённый пример системы, который является единичным, с использованием нескольких уровней классификатора не оптимизирован по числу нажатий пользователя. Возможная причина узкого распространения таких систем в отсутствии метода автоматического построения оптимального классификатора.
- Оптимизированный классификатор построен на основе методов формализации задачи интерактивного доступа в виде оптимального классификатора, построенного посредством минимизации функционала

общего числа операций (на примере СУБД НИКА). Такой пользовательский интерфейс даёт наиболее быстрый способ перехода к искомым ключам в виде “префиксов”-подсказок.

Концептуально классификатор представляет собой “схему” ключевого уровня массива и является элементом базы знаний в данной предметной области, является средством визуализации ключевого уровня массива, а также альтернативным способом быстрого перехода по ключу в массиве по отношению к полю ввода. Существует множество подходов для разделения объектов на классы. Основу проблемы классификации по лексикографическому признаку составляет префиксное дерево, представляющее собой пространство состояний классификатора. Это же дерево (или древовидная структура trie) лежит в основе лучевого поиска. Лучевой поиск — это технология быстрого многопутевого принятия решения по индексу, в котором содержатся буквы, имеющиеся в ключах исходного массива. Особенность классификатора состоит в том, что структура trie лучевого поиска, размещаемая в памяти компьютера, берётся за основу организации интерактивного интерфейса в форме многоуровневого иерархического классификатора.

В первой половине 60-х годов Эдвард Сассенгат составил комбинированную стратегию цифрового поиска, сочетающую бинарный и последовательный поиск. Лучевая память в виде древовидной структуры trie впервые была создана Брианде. Моррисон предложил сжатую структуру trie в виде дерева PATRICIA без однопутевых ветвей. Такие структуры могут применяться для организации кэша. В настоящее время trie также активно применяется, например, в 2015 г. вышла статья[1] нидерландских специалистов о построении словаря RDF-данных с использованием упомянутых структур.

Проблема доступа к данным в виде некоторой разновидности структуры trie, размещённой в памяти, достаточно хорошо изучена. При этом остаётся не рассмотренной проблема, связанная с организацией интерактивного доступа к данным. Организация многоуровневого классификатора по лексикографическому признаку на основе префиксного дерева как пользовательского интерфейса ключевого массива позволит сделать некоторый шаг в этом направлении. Применение комбинированной стратегии лучевого поиска Сассенгата и сжатого дерева PATRICIA Моррисона составляют основу для построения классификатора.

Относительно **степени разработанности** метода необходимо отметить, что метод лучевого поиска Сассенгата неоднократно использовался и модифицировался в различных исследованиях вплоть до настоящего времени. Применение префиксного дерева в качестве интерфейса с пользователем является новым способом использования этой структуры. Разработанные методы адаптируют метод Сассенгата для организации пользовательского интерфейса. Известно, что среднее время доступа к таким структурам для n строк составляет $O(\log n)$, а пространственная сложность порядка $O(n)$. В 2005г. Резник на основании более поздних исследований выделил класс структур trie со значительно более быстрым временем доступа порядка $O(\log \log n)$ и изучил его асимптотические свойства. К таким структурам относятся LC-trie (сжатые по уровню структуры trie). Оптимальный классификатор является сжатым по поддеревьям деревом trie и играет роль интерфейса с пользователем, поэтому в нём существенна временная, а не пространственная сложность. В разделе 2.1 диссертации получено, что временные показатели классификатора лучше обычного дерева trie, но хуже

[1] Hamid R. Bazoobandi , Steven Rooij , Jacopo Urbani , Annette Teije , Frank Harmelen , Henri Bal, A Compact In-Memory Dictionary for RDF Data, Proceedings of the 12th European Semantic Web Conference on The Semantic Web. Latest Advances and New Domains, May 31-June 04, 2015

сжатого по уровням дерева LC-trie. Дерево LC-trie не применимо в виде интерфейса с пользователем, т.к. должно быть заполнено всеми буквами алфавита на заданное число уровней. Применение структуры, используемой в лучевом поиске, для организации интерактивного доступа к ключевому массиву имеет свои особенности. Для построения оптимального классификатора необходимо минимизировать общее число операций в алфавитном классификаторе при поиске по ключу. При этом необходимо получить оптимальное число переходов при поиске ключа в классификаторе и число ключей в списке на выбранный префикс.

Ещё одним важным аспектом построения оптимального классификатора является применение **алгоритмов минимаксного размещения букв** или их сочетаний по каналам обслуживания при параллельном подсчёте числа операций. Самые быстрые алгоритмы дают удвоенное оптимальное решение за линейное время. Улучшить этот класс алгоритмов нельзя, т.к. было доказано, что при коэффициенте аппроксимации меньшим, чем 2 получается NP-трудная задача[2]. В виду небольшого числа однобуквенных и двухбуквенных префиксов и числа каналов обслуживания возможно применение аппроксимации $1+\varepsilon$ или даже алгоритмов точного решения задачи.

Основная **цель** диссертационной работы заключается в разработке методов построения многоуровневого классификатора по лексикографическому признаку, адаптирующего структуру лучевого поиска Сассенгата для организации интерактивного доступа к ключевому массиву для повышения эффективности, надёжности и качества гипертекстовой системы.

Для её достижения требуется решить следующие **задачи**:

1. На основе структуры лучевого поиска Сассенгата необходимо проанализировать неравномерность распределения ключей массива по n-граммным префиксам посредством средней длины префикса классификатора с использованием модельных неравномерных распределений на примере индексных текстовых полей в ООСУБД НИКА.
2. Для исследования вида случайных распределений величин длины префикса класса и числа ключей в классе выделить характерные из семейства распределений, зафиксировав максимальное число ключей в классе.
3. Построить регрессионную модель зависимости средней длины префикса класса от максимального числа ключей в классе, позволяющую сопоставить классификатору среднюю длину префикса.
4. Выбрать оптимальный классификатор посредством минимизации функционала общего числа операций в классификаторе на заданных диапазонах максимального числа ключей в классе, разделённого на равные группы, и числа ключей в группе. Оптимальному классификатору сопоставить среднюю длину префикса из регрессионной зависимости.
5. Разработать программу для расчёта оптимального значения функционала общего числа операций в классификаторе, соответствующего оптимальному классификатору, с использованием нескольких параллельных каналов обработки буквенных префиксов с близкими средними частотами на основе подходящих минимаксных алгоритмов размещения объектов.

Основные **положения, выносимые на защиту**:

1. Метод модельных распределений для анализа неравномерности распределения ключей массива по n-граммным префиксам на основе префиксного дерева сочетаний.
2. Функции плотности распределения для случайных величин длины префикса класса и числа ключей в классе при фиксированном максимальном

---

[2] *Gonzalez T.* Clustering to minimize the maximum intercluster distance // Theoretical Computer Science. — 1985. — Vol. 38. — P. 293–306.

значении числа ключей в классе в виде асимптотического разложения, основанного на нормальном распределении. Распределение длины префикса получилось мультимодальным, а распределение числа ключей в классе унимодально при небольшом максимальном значении числа ключей.

3. Метод построения классификатора на основе регрессионной зависимости средней длины префикса алфавитного классификатора от максимального числа ключей на любой префикс методом ортогональных полиномов Чебышева.

4. Метод построения оптимального классификатора на основе математической модели и алгоритма выбора оптимального классификатора с использованием префиксного дерева сочетаний посредством минимизации функционала общего числа операций в классификаторе.

5. Программные модули, реализующие подходы и алгоритмы, представленные в диссертации.

Научная новизна, выносимых на защиту результатов состоит в следующем:

– в рамках диссертационной работы впервые формулируется и решается задача адаптации структуры лучевого поиска Сассенгата для организации интерактивного доступа к ключевому массиву в глобальных гипертекстовых системах, требующая новых моделей и подходов;

– впервые представлены в аналитическом виде характерные случайные распределения длины префикса класса и числа ключей в классе, выбранные из семейства случайных распределений, для фиксированного максимального числа ключей в классе с использованием разложения в ряд Эджворта;

– впервые предложена модель регрессии на ортогональных полиномах для зависимости средней длины ключа от максимального числа ключей в классе для определения средней длины ключа оптимального классификатора;

– впервые предложен алгоритм построения оптимального классификатора по лексикографическому признаку на основе префиксного дерева сочетаний при минимизации функционала общего числа операций в дереве, в результате которого также определяется максимальное число ключей в классе оптимального классификатора.

Методы исследования, используемые в диссертационной работе, включают в себя системный анализ, методы математической статистики, регрессионный анализ, методы оптимизации, методы классификации и кластеризации.

Объектом исследования являются классификаторы по лексикографическому признаку на основе префиксных деревьев.

Предметом исследования являются методы построения классификаторов по лексикографическому признаку и их оптимизация для организации интерактивного доступа к ключевому массиву.

Теоретическая и практическая значимость работы. Диссертационная работа имеет как теоретическую, так и практическую значимость.

Теоретическая значимость работы заключается в первую очередь в постановке исследуемой задачи, предложенном оптимальном классификаторе по лексикографическому признаку, а также в разработанных моделях и алгоритмах для описания свойств оптимального классификатора на основе префиксного дерева сочетаний в виде функционала общего числа операций и задачи построения регрессионной зависимости средней длины префикса классификатора от максимального числа ключей в классе. Полученные результаты могут быть использованы для дальнейшего развития науки в данной области.

Практическая значимость диссертационной работы подтверждается тем, что её результаты внедрены (см. справки о внедрении) в информационно-справочную систему на основе электронной публикации материалов ежегодника

“Системные исследования” за более чем 25-летний период издания. Другим применением является автоматизированная система управления электронными публикациями баз данных «Философия и методология науки в журнале ”Вопросы философии”». Система включает в себя коллекцию из полутора тысяч статей за более чем 50-летний период издания. Наконец, третье применение — это информационно-поисковая система по репрессированным за годы советской власти, содержащая на данный момент более 36 тыс. биографических справок о пострадавших. Система содержит индексы по текстовым полям, представленные в виде многоуровневых классификаторов по лексикографическому признаку с использованием полученных результатов.

Соответствие паспорту специальности. Содержание диссертации соответствует п.2 «Формализация и постановка задач системного анализа, оптимизации, управления, принятия решений и обработки информации» в части создания методов и алгоритмов построения классификаторов по лексикографическому признаку и их оптимизации, а также соответствующих алгоритмов; п.4 «Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации» в части создания программного и мат. обеспечения гипертекстовой системы СУБД НИКА; п.12 «Визуализация, трансформация и анализ информации на основе компьютерных методов обработки информации» паспорта специальности 05.13.01 — Системный анализ, управление и обработка информации (информационно-вычислительное обеспечение) в части адаптации структуры цифрового поиска для организации интерактивного доступа к ключевому массиву.

Степень достоверности и апробация результатов. Основные положения и результаты диссертационной работы были представлены на следующих международных научных конференциях:

– XVIII Международной конференции “Advances in Science and Technology”, Москва, 2019;

– конференции "XXII Ежегодная международная конференция ПСТГУ", секция факультета ИПМ 2011;

– конференции "XXIII Ежегодная международная конференция ПСТГУ", секция факультета ИПМ, 2012;

– конференции "XXIV Ежегодная международная конференция ПСТГУ", секция факультета ИПМ, 2013;

– конференции "XXV Ежегодная международная конференция ПСТГУ", секция факультета ИПМ, 2014;

– конференции "Информационные системы в науке - 95", Москва, 1995;

– конференции "The 3-rd international workshop on "Advances in databases and information systems", ACM SIGMOD, 1996;

– III Международной конференции "Развитие и применение открытых систем", 1996.

Помимо научных конференций результаты диссертационной работы были обсуждены на научно-исследовательском семинаре в ОИВТА РАН.

Публикации. Основные научные результаты диссертации изложены в 20 публикациях, в том числе: 10 статей в изданиях, рекомендованных ВАК РФ, 10 публикаций в изданиях, входящих в российские базы цитирования, патент на изобретение, а также свидетельство на программу для электронных вычислительных машин (ЭВМ). ПСТГУ была выдана грамота и медаль "За труды" за результаты применения программы для электронных вычислительных машин (ЭВМ).

Структура и объём диссертации. Диссертация содержит введение, пять глав, заключение, список сокращений, библиографию, список публикаций автора по

теме диссертации, а также четыре приложения. Число страниц в работе — 207, рисунков — 44, таблиц — 23. Число наименований в библиографии составляет 128.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** даётся обоснование актуальности проблемы связанной с организацией интерактивного доступа к ключевому массиву сложноструктурированной БД на основе оптимального классификатора. Формулируются задачи, которые необходимо решить для достижения цели построения многоуровневого классификатора. Кратко сформулировано основное содержание глав, а также приведены основные библиографические ссылки, в том числе и на публикации автора.

В **первой главе** приводится обзор, посвящённый классификации, в котором рассматриваются основные проблемы методов классификации, и даётся постановка задачи исследования, связанного с построением оптимального классификатора. Показывается, что решение этой задачи вносит научный вклад в интерактивные методы доступа к БД. Подчёркивается значимость предлагаемого метода с точки зрения глобальных гипертекстовых систем на основе баз данных. В конце вводной главы даётся обоснование классификации посредством уникальных алфавитных ключей.

Во **второй главе** проводится обзор публикаций по структуре лучевого поиска, называемой префиксным деревом сочетаний (ПДС) или деревом PATRICIA Моррисона (см.рис.1а), проанализирована сложность операций в данной структуре. Вводятся понятие ПДС и понятие классификатора с использованием ПДС, рассмотрены различные виды многоуровневого классификатора, а также модельные распределения ключей по буквенным сочетаниям, разработаны характеристики классификатора в виде случайных величин: длины префикса в ПДС и числа ключей в классе на данный префикс, построена регрессионная зависимость этих величин.

Классификатор представляет собой сжатое по поддеревьям ПДС (отличие только в том, что любой префикс классификатора — это путь от корня ПДС) и использует структуру лучевого поиска, предложенную Сассенгатом. PATRICIA — сжатое префиксное дерево trie без однопутевых ветвей. Структура trie была предложена Брианде и Фредкиным. Сассенгат использует метод, комбинирующий несколько первых уровней дерева с прерыванием ветвления на некотором уровне и списки ключей на соответствующие префиксы (см. пример на рис.1b).

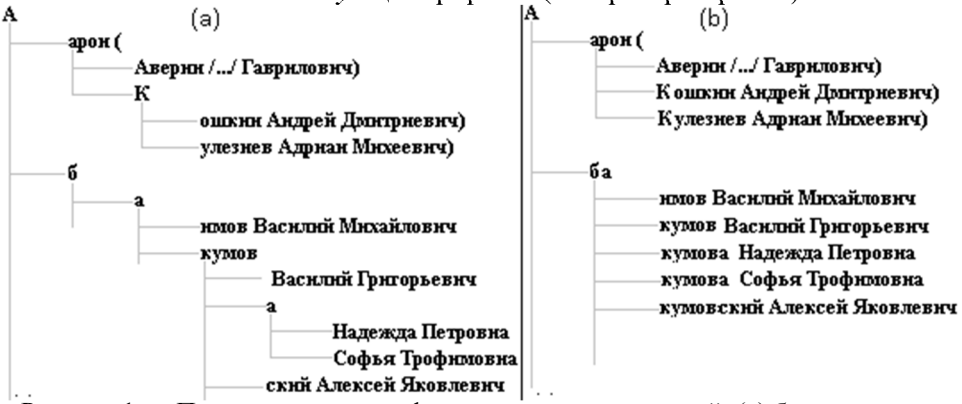

Рисунок 1 — Пример сжатого префиксного дерева сочетаний: (a) без прерывания (b) с прерыванием ветвления

Пусть N — число ключей в массиве, a — мощность алфавита. В случае равномерного распределения вершин по буквенным префиксам длина префикса k в ПДС связана строгой зависимостью с числом ключей n в классе $k(n)=\log_a(N/n)$. В случае неравномерного распределения эта зависимость является регрессионной.

Пусть n — заданное максимальное число ключей в классе. Средняя длина ключа в классификаторе определяется по формуле $M\kappa(n) = \sum_{i=1}^{n} k_{si} n_{si} / N$. Здесь s_i — i-ое буквенное сочетание, все сочетания занумерованы в порядке левого обхода классификатора; n_{si} — число вершин в группе с ключом s_i: $n_{si} \le n$. Длина ключа для неравномерного случая всегда будет не меньше, чем длина ключа для равномерного случая: $k(1) \le M\kappa(n)$.

Для исследования неравномерности распределений ключей по префиксам автором диссертации предложен **метод модельных распределений.** В методе рассматриваются модельные неравномерные распределением ключей по префиксам заданной длины. Для большей длины распределение считается равномерным. Тогда мат. ожидание длины ключа классификатора выражается формулой $M[\kappa_{\mu_m}] = m + \sum_{i=1}^{a^m} P_i \log_a P_i + \log_a (N/n)$. Здесь m — длина префикса, на которую ключи распределены неравномерно. Такая модель хорошо показывает отличие неравномерного случая от равномерного. Предлагаемая модель объясняет отличие мат. ожидания случайной величины длины ключа $M[\kappa_{\mu m}]$ при неравномерном распределении от длины ключа k(n) в равномерном случае посредством энтропии $H(P_1,P_2,\ldots) = -\Sigma P_i \log_a P_i$. Данный метод позволяет увидеть, что в случае реальных приложений буквенные ключи распределены по префиксам неравномерно на всю глубину префиксного дерева, т.к. модельное мат. ожидание наиболее близко к мат. ожиданию для неравномерного случая, когда величина m близка к средней длине ключа массива. Отсюда можно сделать вывод, что различные префиксы на любом уровне классификатора могут содержать различное случайное число ключей.

Для построения классификатора автором диссертации введены его характеристики в виде случайных величин длины ключа класса в ПДС и числа ключей в классе и аппроксимированы их **функции плотности распределения.** В случае ПДС для индекса по полю ФИО объёмом 34 657 ключей были получены функции плотности для указанных случайных величин посредством разложения в ряд Эджворта. Для случайной длины ключа класса a это распределение является мультимодальным, а для случайного числа ключей в классе n_c — унимодально или не имеет характерного максимума (при заданном максимальном значении числа ключей в классе n).

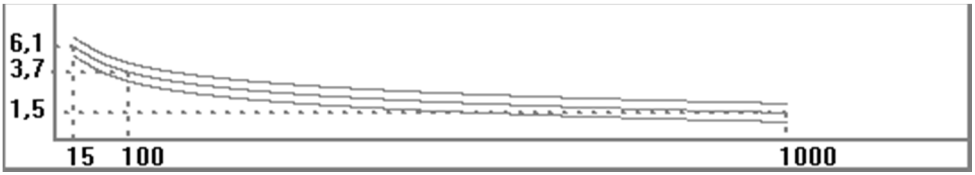

Рисунок 2 — Регрессионная зависимость $k_r(n)$

Для практической реализации классификатора автором диссертации разработан **метод на основе полиномиальной регрессионной модели** зависимости рассматриваемых величин**.** На первом шаге рассчитывается регрессионная зависимости $k_r(n)$, с использованием полиномиальной регрессионной модели $k_r(n) = \ln\left(N / \left(\sum_{j=0}^{p} a_j n^j \right) \right) / \ln a$. На втором шаге посредством соответствующей спецификации, реализованной в гипертекстовой системе СУБД НИКА, по значению

$\varkappa$, определённому из зависимости $\varkappa(n)$ по данному n, строится классификатор. В работе строилась зависимость от максимального числа ключей в классе $k_r(n) = 5{,}867 - \log_{30}(0{,}0034n^3 - 0{,}240n^2 + 6{,}798n - 57{,}561)$ при $14 < n \le 1000$. На рис.2. показана полученная зависимость в указанном диапазоне, а также зависимости соответствующие 95%-ым доверительным интервалам (при возрастании n доверительная область сужается).

Зависимость позволяет по заданному максимальному числу ключей в классе n получить среднюю длину ключа классификатора $k_г$. Ограничением метода является область определения n, на которой была построена зависимость. Зависимость $k_г(n)$ является актуальной для получения характеристик оптимального классификатора. Зависимость $k_г(n)$ задаёт класс классификаторов с различными параметрами $k_г$ и n, но не позволяет выбрать классификатор с оптимальным числом переходов, который рассматривается в третьей главе.

В **третьей главе** описан разработанный автором **метод построения оптимального классификатора по лексикографическому признаку**, в рамках которого рассматривается модель для оптимизации функционала общего числа операций. В отличии от метода, описанного во второй главе, данный метод позволяет построить классификатор с оптимальным числом переходов. Комбинированная структура префиксного дерева и списка ключей задаёт целый класс возможных классификаторов, среди которых необходимо выбрать оптимальный с точки зрения наиболее быстрого перехода на искомый ключ. При этом для каждого префикса определяется длина, при которой следует прерывать ветвление и переходить на линейный список, а также число уровней, на которые разбивается многоуровневый классификатор. Для этого решается задача минимизации функционала общего числа операций $S_{оп}^*$. **Постановка задачи** имеет вид:

$$S_{оп}^* = \min_{n \in D(n), n_g \in D(n_g)} S_{оп}(n, n_g)$$

По оптимальному значению $S_{оп}^*$ определяются соответствующие аргументы

$$(n^*, n_g^*) = \arg \min_{n \in D(n), n_g \in D(n_g)} S_{оп}(n, n_g)$$

n^* — оптимальное максимальное число ключей в классе и n_g^* — оптимальное число ключей в группе. Из регрессионной зависимости, описанной в главе 2, по значению n^* определяется оптимальная средняя длина ключа $k^*=k(n^*)$ в оптимальном классификаторе.

При равномерном распределении ключей по n-граммным префиксам функционал общего числа операций имеет следующий вид.

$$S_{оп}(n, n_g) = \sum_{h=1}^{h_m} n^{h-1} \left(\frac{m(m-1)}{2} n_g + \frac{n_g(n_g+1)}{2}(m-1) + \frac{r(r-1)}{2} \right) \qquad (1)$$

Здесь h=1,…, h_m — номер уровня в классификаторе; $h_m=[k_m/\Delta k]+1$ — число уровней в классификаторе; Δk — длина n-грамм, добавляемых на каждом уровне; $n = a^{\Delta k}$ — число ключей в классе; $n_g \le n$ — число ключей в группе; $N = a^{km}$ — число ключей в массиве, k_m — длина ключа массива; $r = \{n / n_g\}$ — число ключей в последней группе (при r=0 берётся $r=n_g$); $m = [n / n_g] + l(r)$, $l(r>0)=1$, $l(0)=0$ — число групп в классе.

В неравномерном случае процесс построения классификатора осуществляется с помощью ПДС, в котором записаны все частоты встречаемости префиксов. При построении классификатора каждый класс состоит не более чем из n ключей. Каждый класс разбивается на группы ключей по n вершин. В последней группе класса может быть менее чем n вершин.

В диссертационной работе при расчётах использовался функционал общего числа операций для одноуровневого классификатора следующего вида:

$$S_{on}\left(n,n_g\right)=\sum_{i=1}^{k_m}\sum_{j=1}^{n(k)}\left(S_g\left(k,i,n_g\right)+S_{gk}\left(k,i,n_g\right)+S_{gr}\left(k,i,\right)\right)+S_k\left(n_g\right) \quad (2)$$

Здесь первая сумма означает суммирование по всем длинам ключей от 1 до k_m, вторая сумма означает суммирование по всем ключам длины k от 1 до $n(k)$, где $n(k)$ — общее число ключей длины k.

Первое слагаемое $S_g(k,i,n)$ — суммарное число операций прохода по группам ключей для класса с i-ым ключом длины k, где $m(k,i)$ — число групп класса по n ключей, кроме последней:

$$S_g\left(k,i,n_g\right)=\sum_{j=0}^{m(k,i)-1}jn_g=\frac{m(k,i)\left(m(k,i)-1\right)}{2}n_g \quad (3)$$

Второе слагаемое $S_{gk}(k,i,n_g)$ — суммарное число операций прохода по ключам групп для класса с i-ым ключом длины k, кроме последней группы, в которой может быть менее чем n ключей:

$$S_{gk}\left(k,i,n_g\right)=\sum_{w=1}^{n_g}w\left(m(k,i)-1\right)=\frac{n_g\left(n_g+1\right)}{2}\left(m(k,i)-1\right) \quad (4)$$

Третье слагаемое $S_{gr}(k,i)$ — суммарное число операций прохода по ключам последней группы числом $r(k,i)\le n$ для класса с i-ым ключом длины k:

$$S_{gr}\left(k,i\right)=\frac{r(k,i)\left(r(k,i)+1\right)}{2} \quad (5)$$

Последнее слагаемое $S_k(n_g)$ — общее число операций для алфавитных ключей классификатора, составляющих один класс:

$$S_k\left(n_g\right)=\frac{m_a\left(m_a-1\right)}{2}n_g+\frac{n_g\left(n_g+1\right)}{2}\left(m_a-1\right)+\frac{r_a\left(r_a+1\right)}{2} \quad (6)$$

Для многоуровневого классификатора функционал общего числа операций имеет вид:

$$S_{\text{оп}}\left(n,n_g\right)=\sum_{h=1}^{h_m}\sum_{k=1}^{k_m}\sum_{i=1}^{n(h,k)}\left(S_g\left(h,k,i,n_g\right)+S_{gk}\left(h,k,i,n_g\right)+S_{gr}\left(h,k,i\right)\right) \quad (7)$$

Внешняя сумма соответствует сумме по всем уровням многоуровневого классификатора от 1 до $h_m=[\bar{k}/\Delta k]+1$, включая ключевой уровень массива. Здесь $\bar{k}$ обозначает среднюю длину ключа на последнем уровне классификатора, а Δk — среднее число букв, которое добавляется к ключу на каждом уровне классификатора. Функционал (7) можно использовать при больших объёмах массива, например, более 1 млрд. ключей. При меньших объёмах следует использовать функционал (2) и при необходимости надстраивать расчётный уровень классификатора посредством однобуквенного и двухбуквенного уровня.

Метод построения оптимального классификатора задаёт вид функционала общего числа операций $S_{on}(n,n_g)$ в виде формулы (2) или в более общем случае — в

виде (7) и определяет, что для оптимального классификатора $S^*_{on} = \min_{n \in D(n), n_g \in D(n_g)} S_{on}(n, n_g)$.

В конце **третьей главы** рассматривается разработанный автором **алгоритм расчёта оптимального классификатора по лексикографическому признаку**, реализующий метод, предложенный в данной главе, на основе функционала (2). Для ключевого массива строится префиксное дерево (ПДС) и каждому префиксу сопоставляется его частота встречаемости. В соответствии с методом, комбинирующим ПДС и список, на последнем уровне классификатора располагается список ключей на выбранный префикс. Алгоритм состоит в подсчёте $S_{on}(n,n_g)$ для различных n, n_g по формуле (2) и выборе минимального значения $S_{on}(n,n_g)$ на основе полного перебора значений для натуральных величин n и n_g на заданных диапазонах, например, $1 \le n \le 1000$, $10 \le n_g \le 100$.

Для проведения расчётов, связанных с ПДС, оптимальным способом следует распараллелить по нескольким каналам обслуживания расчёт оптимального значения функционала общего числа операций $S_{on}{}^* = S_{on}(n^*, n_g{}^*)$ по формуле (2) для различных грамм и возможно биграмм. Общее оптимальное значение функционала получается в виде суммы значений по каналам:

$$S^*_{on} = \sum_{j=1}^{c} S^*_{on}(j) \qquad (8)$$

Величина c обозначает число каналов обслуживания. Нетрудно видеть, что для распределения грамм или биграмм по каналам обслуживания с наиболее близкими общими средними частотами необходимо решить задачу, связанную с минимаксным размещением объектов.

Существует точное решение этой задачи, но за субэкспоненциальное время $n^{O(\sqrt{P})}$. Здесь P — число кластеров, по которым размещаются объекты. Приближённое решение существует в виде аппроксимации $1+\varepsilon$, например, за время $O(P^n)$ для малых P. Оба класса алгоритмов являются NP-трудными. Более того, Гонзалез, в частности, доказал, что алгоритмы с точным решением и с аппроксимирующим коэффициентом меньшим двух являются NP-трудными. Поэтому непрактично искать решения посредством алгоритмов из указанных классов. Гонзалез предложил решение методом выделения отдалённых точек, который гарантирует удвоенное оптимальное решение за линейное время.

Алгоритм расчёта оптимального значения функционала $S_{on}{}^*$ содержит следующие шаги:

1. Методом выделения отдалённых точек (или путём последовательного разделения алфавита) выделяется соответствующее число групп грамм или биграмм по числу каналов обслуживания с близкими средними частотами.
2. По каждому каналу параллельно производится расчёт функционала $S_{on}{}^*(j)$ по формуле (2).
3. По формуле (8) вычисляется общее значение $S_{on}{}^*$ для всего алфавита.

Таблица 1 — Проблема распределения буквенных сочетаний по каналам обслуживания

N	n_1 / t_1	n_2 / t_2	n_3 / t_3	n_4 / t_4
34 657	8 013	8 046	9 892	8 706
	905	920	1 107	889
103 823	22 945	25 875	27 077	27 926
	3 042	3 370	3 448	3 541
923 915	203 295	230 040	241 750	248 829
	36 942	43 181	45 677	46 223

4 216 674	931 042	1 042 443	1 100 402	1 142 787
	261 021	282 925	301 379	302 345
11 655 046	2 568 183	2 875 210	3 036 400	3 175 253
	843 701	843 701	957 721	1 007 096

Алгоритм расчёта оптимального классификатора был применён к ключевым массивам различных объёмов. Для параллельных расчётов использовались 4 канала обслуживания с частотой каждого ядра процессора 2,5ГГц (Core 2 Quad Q9300, DDR2 8GB, 120GB SSD). В данном случае получилось, что применение деления подряд на 4 части алфавита по наиболее близким частотам по процессам даёт более близкие средние частоты, чем применение метода выделения отдалённых точек.

В таблице 1 приводятся частоты n_i по четырём каналам и временные интервалы t_i (в миллисекундах) расчёта функционала для соответствующей последовательности букв s_i при различных объёмах N ключевого массива. Буквенные последовательности по каналам s_1=АБВГ, s_2=ДЕЖЗИЙК, s_3=ЛМНОПР, s_4=С–Я.

В таблице 2 приведены различные значения для минимума функционала $S_{on}(n^*,20)/N$ при различных объёмах массива N. Здесь также приводятся значения для времён создания ПДС T_c и обработки ПДС при расчёте общего числа операций T_o. Кроме того, даются значения для соответствующих размеров файлов ПДС V и оптимальных значений максимального числа вершин в классе n^*.

Таблица 2 — Временные характеристики ПДС и оптимальные значения функционала при n_g=20

N	T_c, с	T_o, с	V, МБ	n^*	$S^*(n^*,20)/N$
34 657	5,4<1м	3,4<1м	9,6	176	15
103 823	18,0<1м	6,4<1м	23,5	333	18
923 915	211,1<4м	68,9<2м	215,9	619	31
4 216 674	1 155,1<20м	513,9<9м	954,1	1 078	45
11 655 046	8 187,1<137м	1 021,2<18м	2 892,8	1 259	58

В таблице 3 представлены длины префиксов при равномерном распределении ключей по префиксам K_p, средние длины префиксов при неравномерном распределении ключей по префиксам $K_н(1)$, соответствующее среднее число уровней для каждого объёма массива $K_o(n^*)$, средние длины префиксов $K_н(n^*)$ для оптимального значения максимального числа ключей в классе n^* и соответствующее среднее число ключей в классе n_{cp}. Таблица составлена на основе соответствующих экспериментальных данных, на основе которых строятся регрессионные зависимости средней длины префикса от максимального числа ключей в классе $K_н(n^*)$. Здесь n_g — число ключей в группе бралось равным 20.

Таблица 3 — Среднее число уровней и средние длины ключей оптимальных классификаторов

N	K_p	$K_н(1)$	$K_o(n^*)$	$K_н(n^*)$	n^*	$n_{ср}$
34 657	1,553	10,969	3,068	3,106	176	74,158
103 823	1,688	10,803	3,268	3,294	333	47,402
923 915	2,149	13,789	4,789	4,866	619	272,233
4 216 674	2,432	15,693	5,591	5,683	1 078	465,787
11 655 046	2,685	17,244	6,320	6,458	1 259	526,695

На рисунке 3 приводятся зависимости оптимального по числу операций максимального числа ключей в классе от числа ключей в группе $n^*(n_g)$ и числа операций, оптимизированного по n^*, от числа ключей в группе $S_{on}(n^*,n_g)$. Характер зависимости $S_{on}(n,n_g^*)$ при различных N такой же как показано на рисунке 4, на котором приводится зависимость числа операций в классификаторе, построенном для индекса по полю ФИО, содержащего N=34 657 ключей, для числа ключей в группе $n_g^*=20$.

Вид приведённой зависимости является типичным в смысле минимума функционала $S_{on}(n,n_g^*)$. Точное значение в точке минимума равно n^*=176 (или n^*=177) и $S_{оп}^*=S_{оп}(n^*,n_g^*)$=505 080. Фактически минимум $S_{оп}^*$ достигается при n=15. При этом величина $n^*(n_g)$ имеет сильные колебания.

Поиск оптимального значения функционала общего числа операций в алфавитном классификаторе S_{on}^* организован в виде полного перебора значений n и n_g на заданных диапазонах.

В результате по оптимальному значению максимального числа ключей в классе n^*, определённого по минимальному значению функционала общего числа операций $S_{on}(n,n_g)$, из регрессионной зависимости определяется оптимальная длина префикса классификатора $k(n^*)$. Таким образом, оптимальный классификатор для индекса по фамилиям объёмом 34 657 в БД по репрессированным будет содержать максимальное число ключей в классе n^*=176 при числе ключей в группе n_g^*=20 и средней длине ключа класса k^*=3,106. Для других объёмов массивов N значения $S_{оп}^*$ (см. формулу (2)), n^* и $k^*=k(n^*)=K_н(n^*)$ приведены в таблицах 2 и 3. В результате применения метода к ключевому массиву строится соответствующий оптимальный классификатор. Минимизация функционала общего числа операций S_{on} (см. формулу (2) или (7)) позволяет определить параметры оптимального классификатора n_g^* и n^*, по которому из регрессионной зависимости $k(n)$ определяется третий параметр k.

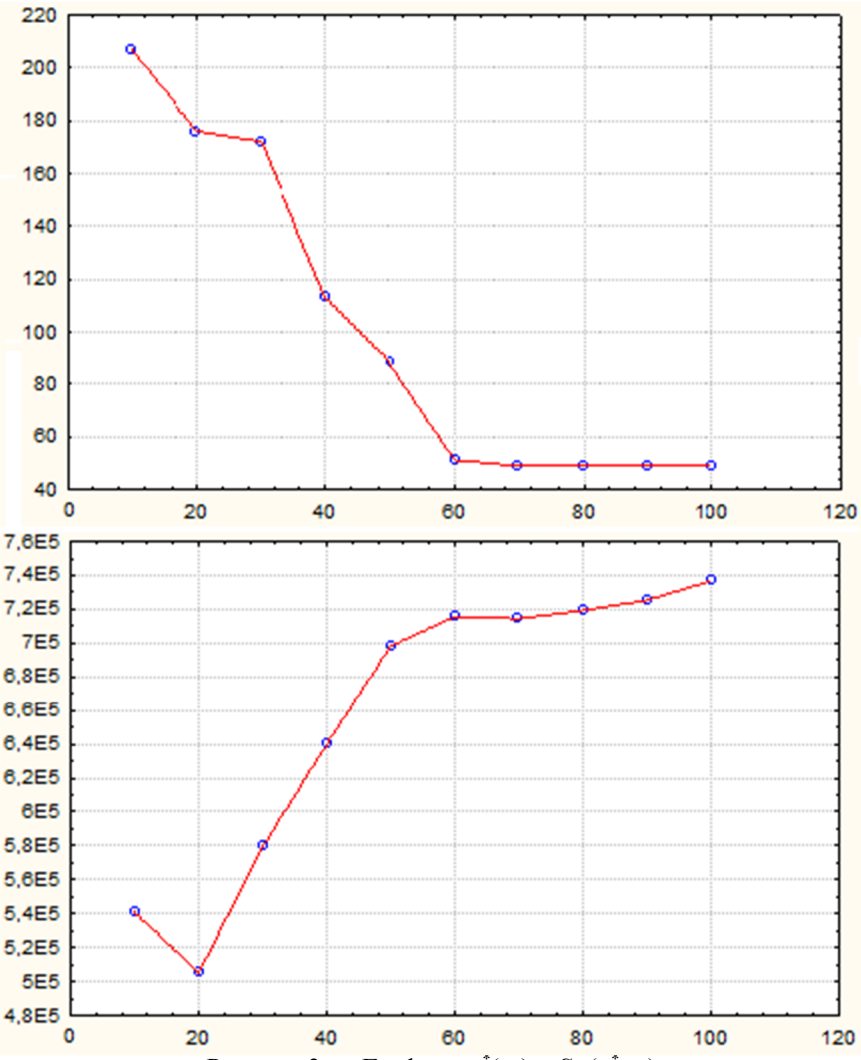

Рисунок 3 — Графики $n^*(n_g)$ и $S_{on}(n^*,n_g)$

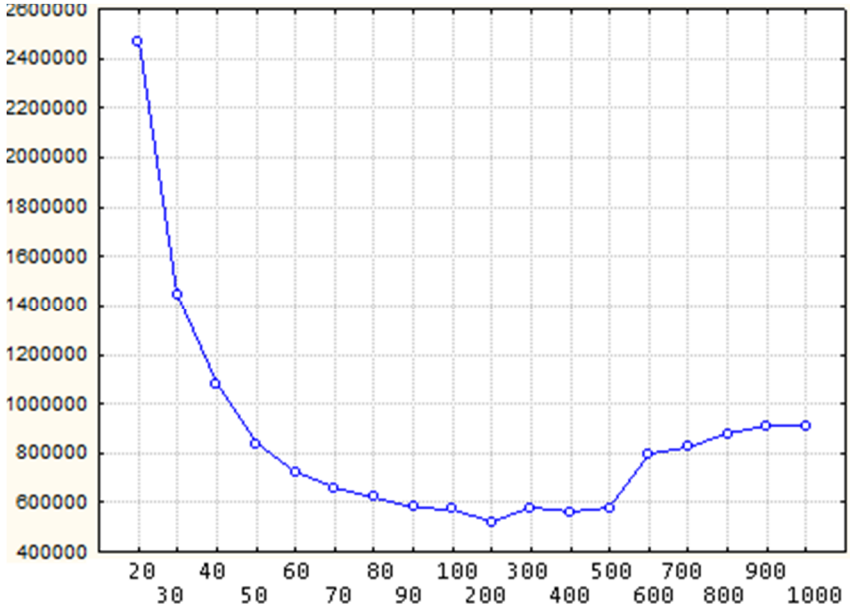

Рисунок 4 — Схематичный вид зависимости числа операций от максимального числа вершин в классе $S_{on}(n,n_g^*)$ при n_g*=20.

Четвёртая глава содержит теорию, методы и средства построения гипертекстовой системы на основе ООСУБД НИКА, средствами которой реализуется классификатор по лексикографическому признаку. Первый пункт данной главы содержит формальное описание модели СУБД НИКА[3] и модели гипертекста в виде сетей Петри; второй — формальное описание предметной области в виде схемы базы данных и определение классификатора по лексикографическому признаку в виде схемы БД НИКА; третий — описание спецификаций отображения вершин БД в виде гипертекстовых документов, соответствующих видам представления данных по Емельянову; четвёртый — описание возможностей уточнения спецификаций посредством расширяемого языка стилей.

ООСУБД НИКА имеет модель данных в виде сети с выделенными иерархиями вершин, задаваемую схемой описания данных. Схема данных определяет классификатор, который разбивает все вершины БД на классы эквивалентностей. Любая координата в схеме БД задаёт класс эквивалентных вершин или объектов заданного типа. Тип объекта БД определяется в смысле Абитебула.

Абитебул определяет понятие “релятивизма данных”, когда системы могут эффективно преобразовывать данные между различными их представлениями.

Идентификация текущей точки в БД осуществляется с помощью задания полной координаты вершины в БД.

Модель СУБД НИКА легко представима в виде модели гипертекста Для построения гипертекстовой системы объекты БД отображаются в виде гипертекстовых документов в соответствии с методами отображения, которые

[3] Годунов А.Н., Емельянов Н.Е., Косьмынин А.Н., Солдатов В.А. СУБД НИКА // Системы управления базами данных и знаний. М.: "Финансы и статистика", 1991. С.208-249.

инкапсулируют объекты БД. Методы или спецификации отображения задаются в вершинах схемы БД. По умолчанию объекты БД отображаются по уровням иерархии — нетерминальные вершины в виде гипертекстовых ссылок на подчинённые, терминальные вершины в виде пары имя и значение. Спецификации отображения вершин составляют ядро гипертекстовой системы. Такое отображение является правомерным в виду двойственности структуры БД и структуры гипертекстовых документов.

Основные спецификации соответствуют видам отображения сложно структурированных данных на двумерной плоскости. Профессор Емельянов Н.Е. в своей работе "Теоретический анализ документного интерфейса"[4] исследует релятивизм данных и определяет четыре вида представления данных: в виде последовательности (одномерное представление), таблицы (двумерное представление), иерархии (n-мерное представление) или смешанного представления (комбинация первых трёх видов). Дополнительно к основным спецификациям определяются спецификации для отображения терминальных вершин различными способами. Можно сказать, что спецификации отображения задают свои классы эквивалентностей вершин, представленных в виде гипертекстовых документов, как и схема БД.

Расширяемый язык таблиц стилей XSL позволяет задавать произвольное форматирование XML-документов. В этом смысле язык XSL можно рассматривать как надстройку над ядром гипертекстовой системы.

Исходный массив ключей представляется в виде ПДС. На рисунке 5 приведена схема описания данных для многоуровневого классификатора, основанного на ПДС.

Здесь "Буквы" — массив, элементы которого состоят из текстового ключа массива "Альфа", номера уровня классификатора "Уровень", циклической ссылки на шаблон "Буквы" так, что ПДС может содержать любое число уровней, и циклической ссылки на значение "=>".

Оптимальный классификатор определяется в смысле функционала $S_{on}(n,n_g)$, описанного в третьей главе. Для построения классификатора берутся оптимальные значения максимального числа ключей в классе n^* и числа ключей в группе n_g^*. Поскольку класс разбивается на группы, то $n_g^* \le n^*$. Каждый префикс исходного массива может быть получен из последовательности значений ключевого поля "Альфа" вложенных массивов "Буквы". Классификатор строится, начиная с последнего уровня. Исходный массив на основе ПДС делится на классы с числом ключей, не превышающим n^*. Префиксы классов исходного массива составляют последний уровень классификатора. Полученный уровень классификатора разбивается на классы префиксов также как исходный массив. Итерационный процесс прекращается, когда число префиксов на текущем уровне классификатора будет не более n^*. Для каждого массива "Буквы" определён счётчик всех терминальных элементов в поддереве ПДС, который соответствует частоте встречаемости $\nu(s_i)$ данного префикса s_i. Префикс s_i добавляется на последний уровень классификатора, если $\nu(s_i) \le n^*$ и для всех префиксов префикса s_i их частота превышает n^*. При добавлении префикса на произвольный уровень классификатора частота префикса корректируется, отсчитывая от предыдущего уровня префиксов, как от исходного массива.

[4] Емельянов Н.Е. Теоретический анализ документного интерфейса, М., препринт ВНИИСИ, 1987, 40 с.

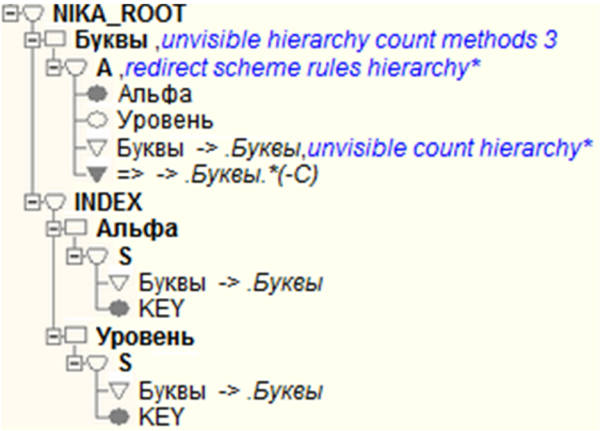

Рисунок 5 — описание данных для многоуровневого классификатора

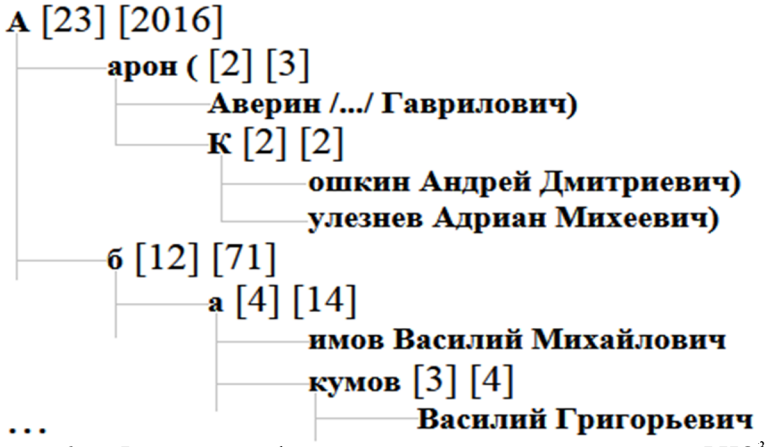

Рисунок 6 — Фрагмент префиксного дерева сочетаний для поля ФИО[34 657]

В **пятой главе** описаны практические примеры применения многоуровневого алфавитного классификатора на основе ПДС с помощью гипертекстовой системы СУБД НИКА в глобальных сетях. Данные отображаются в виде HTML/XML-документов. При реализации классификатора автором был использован опыт, накопленный профессором Емельяновым с сотрудниками при разработке языка OOML, подобного SGML, для СУБД НИКА. OOML служит прототипом для SGML. OOML был реализован в системе МАГИС, которая является документным интерфейсом для СУБД НИКА. В этом смысле реализация гипертекстовой системы для СУБД НИКА с использованием HTML/XML подобна реализации языка OOML, но только с добавлением гипертекстовых ссылок. На рисунке 6 рассматривается пример фрагмента ПДС для индекса по полю ФИО[34 657] в БД по репрессированным. В верхнем индексе указано общее количество имён в индексе. На рисунке 6 в квадратных скобках указаны соответствующие частоты встречаемости префиксов. Первая величина — частота встречаемости относительно подчинённого уровня, вторая величина — частота встречаемости относительно поддерева.

В результате расчёта оптимального классификатора для поля ФИО[34 657] получается минимум функционала общего числа операций $S_{оп}^{*}$=505 080 при максимальном числе ключей в классе n^{*}=176, числе ключей в группе n_g^{*}=20 и

средней длине ключа класса κ^*=3,106. Эти параметры задаются в качестве входных к спецификации, реализующей алфавитный классификатор в гипертекстовой системе СУБД НИКА. В результате на основе ПДС формируется двухуровневый классификатор для поля ФИО[34 657]. Со второго уровня классификатора, фрагмент которого показан на рисунке 7, осуществляется переход на ключевой уровень исходного массива на искомый префикс при выборе соответствующей гипертекстовой ссылки. Вторая цифра в квадратных скобках — это частота встречаемости данного префикса, которая не может превышать n^*=176. Фрагмент классификатора на рисунке 7 состоит из первой группы на префикс "А" с числом префиксов n_g^*=20. Общее число префиксов классификатора составляет 1519.

Аарон (*Аарон (Аверин /.../ Гаврилович)---Аарон (Кулезнев Адриан Михеевич)* [2] [3]
Аб *Абаимов Василий Михайлович---Абызова Мария Васильевна* [12] [71]
Ав *Аввакум (Боровков Григорий Антонович)---Автухов Василий Григорьевич* [9] [97]
Аг *Агаев Михаил Евграфович---Агупова Ольга Осиповна* [8] [84]
Ад *Ададурова Александра Николаевна---Адушкина Александра Ивановна* [6] [26]
Аж *Ажгеревич Надежда Григорьевна---Ажмяков Илья Терентьевич* [2] [4]
Аз *Азанов Иван Абрамович---Азрапкин Прокопий Петрович* [5] [24]
Аикин Андрей Васильевич *Аикин Андрей Васильевич---Аикин Андрей Васильевич*
Ай *Айвазов Иван Георгиевич---Аймаков Николай Семенович* [2] [2]
Ак *Академов Всеволод Николаевич---Акутин Василий Егорович* [6] [93]
Ала *Алабин Владимир Александрович---Алашев (Алашеев?) Михаил Иванович* [6] [10]
Алдо *Алдокимова Антонина Федоровна---Алдошкин Трофим Епифанович* [2] [2]
Алебастров Николай Михайлович *Алебастров Николай Михайлович---Алебастров Николай Михайлович*
Алевтина (*Алевтина (Васильнина Софья Михайловна)---Алевтина (Овчинникова Мария Александровна)* [3] [3]
Алеев *Алеев Василий Павлович---Алеев Михаил Иванович* [2] [2]
Алейников *Алейников Петр Евлампиевич---Алейникова Олимпиада Назаровна* [2] [3]
Алекина Екатерина *Алекина Екатерина---Алекина Екатерина*
Алекса *Алексагин Никита Тимофеевич---Алексапольский Василий Алименьевич (Алинич?)* [3] [134]
Алексе *Алексеев (Аскольдов-Алексеев) Сергей Александрович---Алексенцева Домна Егоровна* [3] [90]
Алекси *Алексий---Алексия (Тимашева-Беринг) Александра Григорьевна* [3] [46]

Рисунок 7 — Фрагмент классификатора для поля ФИО[34 657]

Второй уровень классификатора для поля ФИО[34 657] надстраивается ещё одним уровнем, состоящим из букв алфавита с гипертекстовыми ссылками на соответствующие префиксы второго уровня (поскольку число сочетаний на втором уровне на каждую букву, кроме "К", не превышает n^*). С расчётного уровня классификатора на рисунке 7 при выборе префикса осуществляется переход на группы ключей массива на данный префикс.

При больших объёмах массива число уровней классификатора может быть большим и в общем случае расчёт числа операций производится по формуле (7).

Построенный оптимальный классификатор позволяет находить ключ в среднем за 8 переходов по ссылкам: 4 — по префиксным ссылкам при выборе класса ключей (1 переход с однобуквенного уровня классификатора на основной и 3 при переходе на искомый префикс) и 4 — по ссылкам при переходе по группам на уровне ключевого массива, в худшем случае — за 16 переходов (на последнюю группу на префикс "Кра"), в лучшем случае — за 1 переход, например, на первую группу для букв "Щ", "Э", "Ю", в которых все ключи вмещается в один класс. Такой классификатор оптимизирует доступ к массиву ключей, отображаемого в виде групп, заданного размера. Очевидно, что такой классификатор качественно

превосходит последовательный список, разбитый на группы, с возможностью перехода на следующую группу. В отличии от равномерного классификатора с диапазонами ключей такой классификатор даёт возможность определения отсутствия ключей на заданный префикс на префиксных уровнях, не требуя перехода к самим ключам массива на данный префикс. Наконец, среди всех возможных классификаторов с различными параметрами такой классификатор оптимизирован по общему числу операций. В отношении к поиску посредством поля ввода является альтернативным методом и удобен для использования на мобильных устройствах, не оснащённых клавиатурой.

Алфавитные классификаторы служат одним из основных элементов в справочно-информационной системе по репрессированным и доступны для использования в свободном режиме по адресу http://martyrs.pstbi.ru.

В **заключении** приведены основные результаты диссертации.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ДИССЕРТАЦИИ

Решением поставленных задач является оптимальный классификатор по лексикографическому признаку в смысле минимального в среднем числа переходов к искомому ключу применительно к ключевым массивам сложноструктурированной БД. Для построения оптимального классификатора были решены следующие задачи: проанализирована неравномерность распределения ключей по префиксам методом модельных распределений, исследован вид случайных величин длины префикса класса и числа ключей в классе посредством соответствующих функций плотности, исследована зависимость средней длины ключа префикса класса от максимального числа ключей в классе на основе регрессионной модели, среди возможных классификаторов выбран оптимальный посредством минимизации общего числа операций в классификаторе и реализована программа, выполняющая указанные задачи для заданного ключевого массива. Главным результатом диссертации является методы построения классификаторов по лексикографическому признаку, среди которых выделяется метод построения оптимального классификатора. Остальные методы позволяют исследовать различные характеристики классификаторов и выделить класс алфавитных классификаторов, к которому применяется метод оптимизации функционала общего числа операций. В результате получены следующие основные результаты.

1. Разработан метод модельных распределений для анализа неравномерности распределения ключей массива по n-граммным префиксам на основе префиксного дерева сочетаний. Для модельных распределений получена средняя длина ключа в виде формулы с использованием энтропии, которая показала, что при практическом применении ключи распределены по буквенным сочетаниям неравномерно на всю длину ключа.
2. Получены функции плотности распределений длины префикса класса $f(\kappa)$ и числа ключей в классе $f(n)$ для алфавитного классификатора путём кумулянтного разложения в ряд Эджворта. Аналитические выражения функции плотности распределения интересны с точки зрения выделения различных типов индексов по полям БД. Например, для индекса ФИО$^{32\ 127}$ при одном ключе в классе это распределение имеет 4 моды в точках 10, 5, 17, 29 в порядке убывания частотных вероятностей. Первые 3 моды соответствуют в среднем имени, фамилии и отчеству.
3. Разработан метод построения классификатора по лексикографическому признаку на основе регрессионной зависимости $\kappa(n)$ средней длины префикса алфавитного классификатора κ от максимального числа ключей на любой префикс n методом ортогональных полиномов Чебышева. Зависимость $\kappa(n)$ позволяет определить среднюю длину ключа классификатора при заданном максимально числе ключей в классе и используется также и в случае оптимального классификатора.

4. Разработан метод построения оптимального классификатора по лексикографическому признаку с использованием префиксного дерева сочетаний на основе оптимизации функционала общего числа операций в классификаторе $S_{on}^*=S_{on}(n^*,n_g^*)$, что позволяет формализовать интерактивный способ доступа к ключевому массиву. С помощью приведённого в разделе 3.5 алгоритмы рассчитывается величина S_{on}^* для заданного ключевого массива. Из регрессионной зависимости получается параметр оптимального классификатора средней длины префикса $\kappa^*=\kappa(n^*)$. Полученные характеристики применяются как параметры спецификаций, отображающих классификатор.
5. Разработана с использованием полученных результатов гипертекстовая система на основе ООСУБД НИКА NKWSystem, которая функционально состоит из ядра спецификаций отображения объектов БД, и надстройки над ядром в виде расширяемого языка стилей, для тонкой настройки отображения объектов БД. Одним из применений указанного функционала является отображение объекта типа массив в виде оптимального классификатора по лексикографическому признаку с использованием спецификаций отображения префиксного дерева сочетаний с оптимальными параметрами.

Важно также привести вывод из раздела 2.1 (Понятие префиксного дерева сочетаний), что классификатор по лексикографическому признаку — сжатое по поддеревьям префиксное дерево trie играет роль интерфейса с пользователем, поэтому в нём существенна временная, а не пространственная сложность. Временные показатели классификатора получились лучше приблизительно в 1,6 раза обычного дерева trie, но хуже приблизительно в 1,7 раза сжатого по уровням дерева *LC-trie*, которое не применимо в виде интерфейса с пользователем.

Внедрение результатов работы. В качестве практического применения перечисленных результатов можно рассматривать различные внедрения автором диссертации гипертекстовой системы ООСУБД НИКА. В 1996 году она была внедрена в виде информационно-поисковой системы по репрессированным на базе сайта кафедры информатики ПСТГУ, в 1997 году внедрена в виде системы “СУБД про СУБД” на выделенном ресурсе ИСА РАН и зарегистрирована в Государственном регистре баз данных, в 1998 году внедрена в виде информационной системы на основе Ежегодника “Системные исследования” в рамках проектов РФФИ (Емельянов,Садовский,1997), в 2009 году внедрена в виде информационного ресурса редакции журнала “Вопросы философии” для организации полнотекстового доступа (Чернозуб,Емельянов,2012). Основу информационной системы по репрессированным составляет электронная публикация БД “За Христа пострадавшие”, которая доступна по адресу http://martyrs.pstbi.ru (Емельянов,Тищенко,1999,Тищенко,2019). Для приведённых внедрений к диссертации прилагаются *справки о внедрении*. Также получено свидетельство о государственной регистрации программы *“Гипертекстовая система для ООСУБД НИКА”* (Тищенко и др., 2019). Важным результатом является получение *патента на изобретение* (Арлазаров,Тищенко,2019).

Практическое внедрение результатов позволило проверить теоретические результаты на примере построенных оптимальных классификаторов для конкретных индексов БД, рассчитав оптимальные значения функционала общего числа операций, а также построить более оптимальный алгоритм его расчёта на основе параллельных вычислений.

Перспективы развития классификатора по лексикографическому признаку формулируются в виде следующих направлений развития.

- Выделение типов различных классификаторов на основе функций плотности распределения длины префикса класса, соответствующих

различным полям БД, по которым строятся индексы, например, “ФИО“, “Адрес“, “Должность“.

- Применение префиксного дерева для автозаполнения поля ввода формы.
- Развитие пользовательского интерфейса классификатора в направлении, связанном с одновременным использованием различных панелей n-грамм для навигации по префиксам и ключам массива.
- Применение минимаксных методов размещения объектов для составления равночастотных в среднем буквенных последовательностей для параллельной обработки

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ АВТОРОМ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, входящих в Перечень ВАК РФ

1. Тищенко, В.А. Выбор оптимального алфавитного классификатора при минимизации общего числа операций / В.А. Тищенко // Труды Института системного анализа Российской академии наук (ИСА РАН). Под ред. чл.-корр. РАН Арлазарова В.Л. - М.: Поли Принт Сервис, 2018. -Т. 68 - №1. - С. 54-57.

2. Тищенко, В.А. Методология построения многоуровневого индекса ключевого массива по лексикографическому признаку на основе метода регрессионного анализа на примере СУБД НИКА / Н.Е. Емельянов, В.А. Тищенко // Труды Института системного анализа Российской академии наук (ИСА РАН) РАН. Обработка информационных и графических ресурсов. Под ред. чл.-корр. РАН Арлазарова В.Л. М.: URSS. 2010. -Т. 58 - С. 6-17.

3. Тищенко, В.А. Проблемы построения многоуровневого алфавитного классификатора (на примере ключевого уровня массива СУБД НИКА) / А.В. Соловьёв, В.А. Тищенко // Труды Института системного анализа Российской академии наук (ИСА РАН). Под ред. чл.-корр. РАН Арлазарова В.Л. - М.: Поли Принт Сервис, 2018. -Т. 68 - №1. - С. 63-73.

4. Тищенко, В.А. Представление гипертекста в СУБД НИКА / Н.Е. Емельянов, В.А. Тищенко // Труды Института системного анализа Российской академии наук (ИСА РАН). Под ред. чл.-корр. РАН Арлазарова В.Л. и д.т.н. проф. Емельянова Н.Е. - М., 2009. -Т. 45 - С. 17-36.

5. Тищенко, В.А. О создании информационной системы «Философия и методология науки в журнале „Вопросы философии“ / С.П. Чернозуб, В.Н. Садовский, Н.Е. Емельянов, В.В. Келле, В.И. Тищенко, В.А. Тищенко, И.Б. Чернышева, Е.А. Богомолова, Т.В. Никонова, Д.И. Сергеев, Н.С. Смирнова // Системные исследования. Методологические проблемы: Ежегодник 2011--2012. Вып.36/2011—2012. Под ред. чл.-корр. Попкова Ю.С., д.филос.н Садовского В.Н., к.филос.н. Тищенко В.И. М.: URSS, 2012. С.239-247.

6. Тищенко, В.А. Построение web-сервера для периодических изданий на материале ежегодника “Системные исследования” / Н.Е. Емельянов, В.Н. Садовский, В.А. Тищенко, И.Б. Чернышева // Системные исследования. Методологические проблемы. Ежегодник 1997, изд-во "Эдиториал УРСС”, 1997, с. 313-323 (http://sr.isa.ac.ru/sr-97/emelyan.html).

7. Тищенко, В.А. Применение автозаполнения для перехода по ключевым словам на искомые значения в массиве СУБД НИКА / В.А. Тищенко // Материалы XXIII Ежегодной богословской конференции ПСТГУ – М., 2013г. - т.1 - С. 325-328.

8. Тищенко, В.А. Принципы построения web-сервера на основе объектно-ориентированной базы данных. / Н.Е. Емельянов, В.А. Тищенко // Информационные технологии и вычислительные системы. — 1997, N 4. С.90-99. — Отпечатано в ВЦ РАН. Москва.

9. Тищенко, В.А. Методы отображения объектов для построения web-сервера объектно-ориентированной базы данных / Н.Е. Емельянов, В.А. Тищенко // Развитие безбумажных технологий в организационных системах / Сборник трудов

ИСА РАН / Под ред. д.т.н. проф. Арлазарова В.Л. и д.т.н. проф. Емельянова Н.Е. М.: URSS. 1999. С. 96-109.

10. Тищенко, В.А. Применение языка XSL для отображения БД НИКА / В.А. Тищенко // Организационное управление и искусственный интеллект . Сборник трудов ИСА РАН / Под ред. д.т.н. проф. Арлазарова В.Л. и д.т.н. проф. Емельянова Н.Е. М.: URSS. 2003. С. 149-175.

Публикации в трудах профильных конференций

11. Тищенко В.А. Организация интерактивного доступа к ключевому массиву на основе классификатора по лексикографическому признаку // Материалы XVIII Международной научно-практической конференции “Advances in Science and Technology”, 31 января 2019, с.108-111.

12. Tishchenko, V.A. “Web server on the basis of NIKA DBMS” / N.E. Emelyanov, I.V. Muhanov, V.A. Tishchenko // proceedings of the third international workshop on “Advances in databases and information systems”, ACM SIGMOD, Moscow, sep. 10-13, 1996, Vol.2, p.58-59.

13. Тищенко, В.А. “Использование баз данных в составе BBS” / Н.Е. Емельянов, В.А. Тищенко // материалы конференции “Информационные системы в науке - 95”, Москва, 1995.

14. Тищенко, В.А. Развитие теории, методов и средств индексации, поиска и отображения объектов в сложных структурах и документах / Н.Е. Емельянов, А.С. Богданов, А.П. Романов, А.В. Соловьев, В.А. Тищенко, И.Б. Чернышова // Отчет о НИР № 96-01-01840 (Российский фонд фундаментальных исследований).

15. Тищенко, В.А. Инструментальные средства построения информационных систем / Н.Е. Емельянов, А.С. Богданов, И.В. Муханов, А.В. Соловьев, В.А. Тищенко, С.А. Хабарова, И.В. Щелкачёва // Отчет о НИР № 96-07-89394-в (Российский фонд фундаментальных исследований).

16. Тищенко, В.А. Информационная система по истории христианства в России в XX веке / В.Н. Воробьев, Н.В. Сомин, Л.С. Аристова, Н.Е. Емельянов, А.В. Мазырин, А.В. Соловьев, Н.С. Соловьева, В.А. Тищенко, И.В. Щелкачева // Отчет о НИР № 97-07-90055 (Российский фонд фундаментальных исследований).

17. Тищенко, В. А. СУБД НИКА и гипертекстовые информационные системы в INTERNET. / Н.Е. Емельянов, И.В. Муханов, В. А. Тищенко // Телематика-96. — 1996.

18. Тищенко, В.А. WWW-сервер на основе СУБД НИКА / Н.Е. Емельянов, И.В. Муханов, В.А. Тищенко // Издание Международного Центра Научно-технической информации, 3-я Международная конференция «Развитие и применение открытых систем» — 1996.

Патенты, свидетельства о регистрации программ для ЭВМ

19. Устройство отыскания информации по ключевым словам : Патент на изобретение № 2679967 C1 Российская Федерация / В.Л. Арлазаров, В.А. Тищенко ; патентообладатель Общество с ограниченной ответственностью «Смарт Энджинс Рус». — № 2018109096; заявл. 14.03.2018 ; опубл. 14.02.2019. Бюл. № 5.

20. Программа для электронной публикации баз данных «Гипертекстовая система для ООСУБД НИКА» : свидетельство о государственной регистрации программы для ЭВМ № 2019612352 / В.А. Тищенко и др. — № 2019611090 ; заявл. 07.02.2019 ; опубл. 18.02.2019.

Тищенко Владимир Александрович
«Методы построения многоуровневого классификатора по лексикографическому признаку применительно к ключевому уровню массива ООСУБД НИКА»
Подписано в печать: 22.09.2020 Объём: 1,5 усл. п.л. Тираж: 100 экз.
Отпечатано в типографии «Реглет»
115184, г. Москва, Пятницкий переулок, д. 3
8 (495) 971-17-64; reglet.ru