

ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертационную работу Тищенко Владимира Александровича
«Методы построения многоуровневого классификатора по лексикографическому признаку применительно к ключевому уровню массива ООСУБД НИКА»,
представленную на соискание учёной степени кандидата технических наук по специальности 05.13.01 - «Системный анализ, управление и обработка информации (информационно-вычислительное обеспечение)»

1. Актуальность темы

Составление онтологий предметной области является универсальным современным подходом к классификации документов. Этот подход дополнил более точным определением «эталонного» документа эффективные методы классификации с использованием машинного обучения. Как некоторый вид онтологий можно выделить алфавитные классификаторы, которые наряду с полем ввода являются естественным способом поиска ключевых слов в массиве, особенно в случае мобильных устройств. Актуальной является проблема автоматизированного построения многоуровневых алфавитных классификаторов при больших объёмах ключевых слов. В рамках этой проблемы возникает задача построения классификатора, оптимального с точки зрения числа переходов в классификаторе при поиске ключа.

Следовательно, диссертационная работа Тищенко В.А. актуальна, имеет важно теоретическое и практическое значение, а проводимые исследования актуальны, как в настоящее время, так и на обозримую перспективу.

2. Научная новизна проведенных исследований и полученных результатов

На основании выполненных исследований были получены теоретические положения, совокупность которых можно квалифицировать как решение актуальной научной задачи. Автором были получены новые научные результаты:

- представлены в аналитическом виде 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;

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

- разработан алгоритм построения оптимального классификатора по лексикографическому признаку с использованием префиксного дерева сочетаний при минимизации функционала общего числа операций в дереве, позволяющий определить максимальное число ключей в классе оптимального классификатора;

- предложен метод построения оптимального алфавитного классификатора для организации интерактивного интерфейса с ключевым массивом, в котором решается задача минимизации функционала общего числа операций в классификаторе.

3. Обоснованность и достоверность результатов

Основные результаты диссертационной работы подтверждены корректными математическими выводами, экспериментальной проверкой и согласованы с уже имеющимися известными данными.

4. Практическая значимость результатов диссертационной работы

Практическая значимость результатов диссертации заключается в том, что полученные автором результаты научных исследований реализованы в гипертекстовой системе для ООСУБД НИКА, в данный момент внедрённой в информационно-поисковых системах. Проектирование таких систем производилось с учётом разработанной модели архитектуры гипертекстовой системы, которая может быть использована как фиксированный набор интерфейсов и программных модулей, обеспечивающий возможность построения системы подсказок в виде автозаполнения для полей ввода, использующихся в поисковых системах. Практический интерес представляет дальнейшая апробация и развитие отдельных элементов этой модели в контексте сложных информационно-справочных систем.

Приоритет практических решений подтвержден патентом на изобретение и свидетельством о государственной регистрации программы для ЭВМ. Практическая значимость результатов диссертации подтверждается соответствующими документами о внедрении.

5. Оценка содержания диссертационной работы, завершённость и соответствие автореферату

Диссертационную работу можно охарактеризовать глубиной исследования поставленной задачи. Работа носит законченный характер, выполнена на современном уровне и оставляет хорошее впечатление.

Диссертация содержит 207 страниц и состоит из введения, пяти глав, заключения, библиографического списка литературы из 128 источников и четырёх приложений.

Введение содержит краткое описание поставленной научной задачи, обоснование актуальности темы работы, цель, объект, предмет и методы

исследования. Представлены основные положения, выносимые на защиту и их новизна, теоретическая и практическая значимость работы.

В *главе 1* рассматриваются основные проблемы методов классификации и ставится задача построения оптимального алфавитного классификатора для организации интерактивного интерфейса ключевого массива. В конце главы даётся обоснование классификации уникальными алфавитными ключами.

Глава 2 посвящена методам построения алфавитного классификатора. В начале главы вводится понятие префиксного дерева и дан обзор работ по цифровому поиску. В качестве основной структуры классификатора берётся префиксное дерево в комбинации с линейным списком ключей, рассмотренное Сассенгатом. Это дерево, сжатое по поддеревьям, составляет основу классификатора. Последний подраздел главы посвящён характеристикам классификатора: средней длине ключа префикса класса, среднему и максимальному числу ключей в классе, а также регрессионной зависимости между ними.

В *главе 3* вводится функционал общего числа операций в классификаторе в качестве целевой функции для решения задачи построения оптимального классификатора. Минимум функционала даёт значения характеристик для оптимального классификатора, а из регрессионной зависимости, полученной в главе 2, определяется средняя длина ключа оптимального классификатора.

Глава 4. Модель БД НИКА может быть описана в рамках некоторого формализма, который разработали Абитебул и Хулл для изучения проблем реструктуризации объектно-ориентированных баз данных. БД представляет собой сеть объектов с выделенными иерархиями. С другой стороны, динамически создаваемые документы, содержащие объекты БД, составляют гипертекст. Посредством модели гипертекста в виде сетей Петри показано, что модель ООСУБД НИКА естественным образом вписывается в гипертекстовую систему. Правомерность описания объектов БД НИКА в виде гипертекста следует из двойственности структуры объектов БД и структуры html/xml-документов, содержащих эти объекты. Формальное описание предметной области формируется в виде связанных друг с другом иерархий объектов, что является естественным при представлении сложно-структурированной информации. Объекты БД инкапсулируют методы отображения этих объектов в гипертекстовые документы и составляют ядро системы. Для более тонкой настройки отображения объектов используется надстройка над ядром в виде стилевых таблиц. В гипертекстовой системе NKWSystem предусмотрены

различные методы и атрибуты для построения различных классификаторов - однобуквенных двухбуквенных, многоуровневых, диапазонных и др.

В главе 5 описано практическое применение многоуровневого алфавитного классификатора на основе префиксного дерева. Приведён реальный пример построения оптимального классификатора. Для построения оптимального классификатора используется сжатое префиксное дерево, пример которого представляет собой оптимальный классификатор для поля ФИО.

Заключение содержит краткую сводку результатов, полученных в работе, примеры внедрений, а также перспективы развития методов построения классификаторов.

Автореферат диссертации достаточно полно и правильно отражает ее содержание и основные результаты, выносимые на защиту.

Материал диссертации и автореферата изложены ясным, четким и лаконичным языком. В тексте диссертации и автореферата содержится достаточное количество иллюстративного материала.

6. Соответствие специальности

Содержание диссертации соответствует паспорту специальности 05.13.01 - «Системный анализ, управление и обработка информации (информационно-вычислительное обеспечение)» - формуле специальности и области исследования.

7. Публикации и апробация основных результатов работы

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 государственной регистрации программы, реализующей гипертекстовую систему для ООСУБД НИКА.

Научные публикации соответствуют содержанию диссертации.

Основные результаты работы прошли апробацию на 10 профильных и международных конференциях и научно-исследовательских семинарах.

Результаты были получены в процессе выполнения работ по нескольким научным грантам РФФИ, непосредственно относящихся к теме диссертации.

8. Основные результаты диссертационной работы

Основные результаты теоретических и экспериментальных исследований, приведенных в диссертационной работе в соответствии с поставленной научной задачей, могут быть сформулированы, следующим образом:

- разработан метод модельных распределений для анализа неравномерности распределения ключей массива по n -граммным префиксам на основе префиксного дерева сочетаний, для модельных распределений получено аналитическое выражение средней длины ключа;

- разработан метод построения классификатора по лексикографическому признаку на основе регрессионной зависимости средней длины префикса алфавитного классификатора от максимального числа ключей на любой префикс с помощью ортогональных полиномов Чебышева;

- разработан метод построения оптимального классификатора по лексикографическому признаку с использованием префиксного дерева сочетаний на основе оптимизации функционала общего числа операций в классификаторе, позволяющий формализовать интерактивный способ доступа к ключевому массиву;

- получены функции плотности распределений длины префикса класса и числа ключей в классе для алфавитного классификатора путём кумулянтного разложения в ряд Эджворта, необходимые для выделения различных типов индексов по полям БД;

- на основе полученных результатов разработана гипертекстовая система на основе ООСУБД НИКА NKWSystem.

9. Замечания

По работе можно сделать следующие замечания:

- 1) основной результат диссертации опирается во многом на публикации конца 50-60-х годов, таких авторов как Фредкин, Брианде, Моррисон, Сассенгат, а современные публикации используются только для сравнения полученных результатов; ссылки на них в основном сделаны в обзоре префиксного дерева сочетаний (в подразделе 2.1), например, при сравнении с деревом LC-trie;

- 2) в работе до конца не формализовано математически понятие «сжатое по поддеревьям» префиксное дерево и не определяется то, как оно соотносится с деревом LC-trie;

- 3) с математической точки зрения является некорректным делать обобщение на основе частных примеров без формулировки теорем и их доказательств - в частности, в работе рассмотрен только отдельный пример нахождения минимума функционала общего числа операций в классификаторе на заданных диапазонах и его существование обобщается на любой классификатор.

10. Заключение о соответствии работы требованиям ВАК

Несмотря на имеющиеся замечания, диссертация Тищенко А.В. является законченной научно-квалификационной работой, содержащей решение актуальной научной задачи, выполнена на высоком научно-техническом уровне. Диссертация написана автором самостоятельно. Результаты работы являются достоверными, обладают научной новизной и практически значимы.

Считаю, что диссертационная работа «Методы построения многоуровневого классификатора по лексикографическому признаку применительно к ключевому уровню массива ООСУБД НИКА» полностью соответствует критериям ВАК РФ, предъявляемым к кандидатским диссертациям, а её автор - Тищенко Владимир Александрович, заслуживает присуждения ему учёной степени кандидата технических наук по специальности 05.13.01 - «Системный анализ, управление и обработка информации (информационно-вычислительное обеспечение)».

Официальный оппонент профессор
кафедры «Вычислительные системы,
сети и информационная
безопасность», доктор технических
наук по специальности 05.12.13 -
Системы, сети и устройства
телекоммуникаций, доцент.
Федеральное государственное
автономное образовательное
учреждение высшего образования
«Российский университет
транспорта».

Адрес: 127994, ГСП-4, г. Москва, ул.
Образцова, д. 9, стр. 9
Телефон: +7 (495)681-13-40
Электронная почта: info@rut-miit.ru
<https://rut-miit.ru/>

Сафонова Сафонова Ирина Евгеньевна

28.01.2021 г.

Подпись руки *Сафоновой И.Е.*
Заверяю _____
Начальник Отраслевого центра подготовки
научно - педагогических кадров
высшей квалификации *С.Н. Коржин*

