На правах рукописи

Гришин Дмитрий Сергеевич

СПОСОБ И УСТРОЙСТВО ДЛЯ МНОЖЕСТВЕННОЙ ПОДБОРКИ ТЕКСТОВЫХ ДАННЫХ В ХРАНИЛИЩАХ НА ОСНОВЕ ПРОДУКЦИОННОГО ПОДХОДА

05.13.05 — Элементы и устройства вычислительной техники и систем управления

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

КУРСК – 2017

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Юго-Западный государственный университет» на кафедре информационных систем и технологий

Научный руководитель: **Титенко Евгений Анатольевич**,
кандидат технических наук, доцент

Официальные оппоненты: **Левин Илья Израилевич**,
доктор технических наук, профессор,
Научно-исследовательский центр супер-ЭВМ и нейрокомпьютеров (г. Таганрог), директор центра

Коробкова Елена Николаевна,
кандидат технических наук, доцент,
Федеральное государственное бюджетное образовательное учреждение высшего образования «Белгородский государственный технологический университет им. В. Г. Шухова», кафедра технической кибернетики, доцент кафедры

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего образования «Орловский государственный университет им. И. С. Тургенева»

Защита состоится 27 декабря 2017 г. в 13:00 ч. в конференц-зале на заседании диссертационного совета Д 212.105.02 при ФГБОУ ВО «Юго-Западный государственный университет» по адресу: 305040, г. Курск, ул. 50 лет Октября, 94.

С диссертацией можно ознакомиться в библиотеке Юго-Западного государственного университета и на сайте www.swsu.ru

Автореферат разослан « » _____ 2017 г.

Ученый секретарь
диссертационного совета
Д 212.105.02 Титенко Евгений Анатольевич

Актуальность.

Важным направлением развития средств вычислительной техники является создание аппаратно-вычислительных средств для поддержки оперативной обработки запросов и доступа к данным в информационных системах, содержащих базы и хранилища текстовых данных. Среди хранимой информации значимая роль принадлежит текстовой и гипертекстовой информации, базовым обработчиком которой являются продукционные системы, учитывающие структурные и лингвистические отношения между символами. Кроме того, продукционная схема вычислений «условие → действие» имеет потенциальные возможности для распараллеливания процессов подборки данных по набору шаблонов и модификации данных.

Особый интерес представляют базы и хранилища с неизменяемыми текстовыми и гипертекстовыми данными — электронные библиотеки, коллекции документов, электронные архивы, справочная информация и т. д. Одномерные формы представления текстовых данных, их размер (сотни мегабайт), многообразие связей, вариативность границ отдельных информационных единиц и др., приводят к экспоненциальным и другим времяемким пропорциям доступа к базам и хранилищам, сложным запросам на доступ к данным. Как итог, запрос на аппаратном уровне декомпозируется последовательностью связанных поисковых процессов по набору шаблонов, уточняемых в процессе поиска.

Как правило, в обрабатывающих серверах БД операции безотступного перебора элементов данных, проверки структурно-лингвистических свойств, установления отношений между шаблонами не имеют соответствующей аппаратной поддержки в виде специализированных устройств ассоциативной обработки текстовых операндов, управления режимами адресаций ячеек оперативной памяти, параллельной обработки безадресных структур, мультиплексирования массивов текстовых данных и др. Итерационный характер организации поиска по набору связанных шаблонов приводит к необходимости учета множественных вариантов в следующем цикле вычислений, что требует динамической буферизации данных непосредственно в устройстве, обеспечивающем параллельную обработку запросов в хранилищах данных.

Несмотря на впечатляющие успехи развития современных микропроцессоров CISC-, RISC-архитектур продукционные операции, реализованные в типовой системе команд, приводят к непродуктивным временным и емкостным затратам на их реализацию вследствии обработки текстов как линейных объектов в одномерном пространстве.

Теоретические и прикладные исследования продукционных систем рассматривались в работах А. А. Маркова, Н. М. Нагорного, Н. А. Шанина, В. И. Городецкого, Дж. Люгера и др. Существенный вклад в изучении вопросов, связанных с базами и хранилищами данных, внесли W. H. Inmon, R. Kimball, E. F. Codd, C. J. Date, R. Wrembel, М. Р. Когаловский, А. А. Барсегянян, М. С. Куприянов, В. В. Степаненко, И. И. Холод и др. Вопросы создания перспективных архитектур и схемотехнических решений для микропроцессоров отражены в работах Е. П. Угрюмова, К. Г. Самофалова, Е. А. Зельдина,

В. А. Потехина, В. В. Корнеева, А. В. Киселева и др. Разработка специализированных устройств, в том числе на программируемых интегральных схемах, и параллельной обработки данных рассматривалась в трудах ученых И. А. Каляева, И. И. Левина, В. В. Сташина, И. С. Еремеева и др.

Вместе с тем вопросы исследования продукционных систем для реализации запросов к базам и хранилищам текстовых данных и аппаратной реализации продукционных операций под обработку неизменяемых текстовых данных нашли только частичное отражение в трудах этих ученых. Сложившееся в современной вычислительной технике проблемная ситуация несоответствия между непродуктивными временными затратами на поисковые операции обработки запросов и структурой микропроцессоров общего назначения отражает основное противоречие современных алгоритмических и аппаратных средств для оперативной реализации продукционных операций под обработку запросов к базам и хранилищам текстовых данных. На разрешение данного противоречия направлено диссертационное исследование.

Научно-техническая задача — разработка программно-аппаратных средств, реализующих строковые операции подборки и модификации текстовых данных в хранилищах.

Объект исследования — способы, вычислительные процессы и технические средства реализации множественной подборки и модификации текстовых данных.

Предмет исследования — способы, алгоритмы и схемотехнические решения реализации базовых продукционных операций для множественной подборки и модификации текстовых данных на основе ассоциативной памяти.

Цель работы — сокращение временных затрат на операции множественной подборки текстовых данных в хранилищах путем разработки модифицированного позиционного представления текста, а также способа и специализированного устройства для параллельного поиска и модификации текстовых данных.

Основные задачи диссертационного исследования:

1. Анализ современных алгоритмических и аппаратных средств, используемых или обеспечивающих множественную подборку текстовых данных в хранилищах. Обоснование направления исследования.
2. Разработка модифицированного позиционного представления текстовых данных для хранилищ, содержащего дополнительную информацию о внутренней организации текста.
3. Разработка способа для множественной подборки текстовых данных в хранилищах на основе разработанного модифицированного позиционного представления текста.
4. Моделирование работы разработанного способа для множественной подборки текстовых данных в хранилищах. Разработка программных средств для моделирования.
5. Разработка структурно-функциональной организации специализированного устройства для множественной подборки текстовых данных в хранилищах, разработка схемотехнических решений основных блоков

и узлов устройства, а также экспериментальная проверка его работоспособности, анализ его временных затрат и аппаратной сложности.

Методы исследования. Для решения поставленных задач использовались методы, основанные на теории проектирования электронно-вычислительных машин, схемотехнике, ассоциативной памяти, математической логике, теории множеств, математической статистики, аппарата продукционных систем, а также теории алгоритмов и прикладного программирования.

Научная новизна и результаты, выносимые на защиту:

1. Модифицированное позиционное представление текста, дополняющее классическое позиционное представление новыми элементами, обеспечивающее тем самым направленность поиска и исключающее ряд неперспективных операций.
2. Способ для множественной подборки данных в модифицированном позиционном представлении текста, позволяющий вычислять сокращенный набор позиций возможных вхождений, разбиением входной подстроки на собственные не пересекающиеся подстроки размером ограниченным константой разработанного представления и получением набора с минимальным количеством элементов среди наборов позиций вхождений этих подстрок.
3. Структурно-функциональная организация специализированного устройства для множественной подборки текстовых данных в хранилищах, отличающаяся в аппаратной реализации операций последовательного и межстрочного сдвига при строковом и матричном представлении обрабатываемых текстовых данных с возможностью временной буферизации граничных элементов строк, выходящих за пределы операционного блока устройства, что позволяет ускоренно реализовывать шаги обработки текстовых данных при различных сочетаниях размеров подстроки и строки-модификатора, в том числе выполнять реверсивные межстрочные сдвиги.

Практическая ценность работы состоит в следующем:

1. Разработанное модифицированное позиционное представление текста для хранилищ текстовых данных позволяет выполнять гибкую настройку соотношения временных затрат подборки к временным затратам его построения, установкой значения максимального размера подстрок, хранящихся в данном представлении, что позволяет расширить область его применения и использовать априорные знания о входных данных для сокращения временных затрат множественной подборки. Если такая информация отсутствует задаваемый пользователем размер подстроки рекомендуется выбирать равным 5, что эмпирически обеспечивает оптимальное соотношение временных затрат подборки данных и предобработки входного текста.
2. Разработанный способ для множественной подборки данных в модифицированном позиционном представлении текста, позволяет вычислять сокращенный набор позиций возможных вхождений. Проведенный анализ временных затрат разработанного способа показал, что данный способ позволяет сократить временные затраты относительно способа подборки на

основе суффиксного дерева на 10 % для искусственно построенных данных и на 17 % для естественных данных.

3. Разработанное ассоциативное устройство позволяет сократить временные затраты модификации данных относительно устройства-прототипа на 14 % и на 80 % относительно устройства-аналога.

Реализация результатов работы. Результаты диссертационной работы внедрены в:

1. Комитет информатизации, государственных и муниципальных услуг Курской области, в части оказания государственных и муниципальных услуг населению и оперативной обработке обращений к базам межведомственных запросов и статистики оказания государственных и муниципальных услуг Курской области.

2. ООО «Инновационные системы управления», (г. Курск). в части реализации способа, алгоритмов и схемотехнических решений основных блоков и узлов устройства для множественной подборки данных, реализованные на ПЛИС 5SGSED8I2F45C2 семейства Stratix.

3. Учебный процесс федерального государственного бюджетного образовательного учреждения высшего образования «Юго-Западный государственный университет» по дисциплине «Интеллектуальные системы и технологии» магистров направления «Информационные системы и технологии» в части реализации средств направленной обработки строковых данных.

Соответствие паспорту специальности. Диссертационная работа соответствует паспорту научной специальности 05.13.05 — «Элементы и устройства вычислительной техники и систем управления» по второму пункту «Теоретический анализ и экспериментальное исследование функционирования элементов и устройств вычислительной техники и систем управления в нормальных и специальных условиях с целью улучшения технико-экономических и эксплуатационных характеристик» в части разработки способа и специализированного ассоциативного устройства для множественной подборки текстовых данных в хранилищах для реализации массовых вычислительных операций и обеспечения/обработки запросов, а также проверки работоспособности разработанного способа и устройства.

Апробация работы. Основные научные результаты работы докладывались и обсуждались на региональной заочной научно-практической конференции «Интеллектуальные информационные системы: тенденции, проблемы, перспективы» (г. Курск, 2013), IX международной научно-практической конференции «Передовые научные разработки» (г. Прага, 2013), XII международной научно-технической конференции «Распознавание — 2015: Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации» (г. Курск, 2015), III региональной заочной научно-практической конференции «Интеллектуальные информационные системы: тенденции, проблемы, перспективы» (г. Курск, 2015), XIII международной научно-технической конференции «Распознавание — 2017: Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации» (г. Курск, 2017), а

также рассматривались на семинарах кафедры программной инженерии и кафедры информационных систем и технологий Юго-Западного государственного университета в 2013–2017 гг.

Публикации по работе. По результатам выполненных разработок и исследований опубликованы 11 работ, в том числе 4 работы в рецензируемых научных журналах и изданиях, получено свидетельство о регистрации программы для электронно-вычислительных машин, а также патент Российской Федерации на изобретение.

Личный вклад. Все выносимые на защиту научные результаты получены соискателем лично. В работах по теме диссертации, опубликованных в соавторстве, личный вклад соискателя сводится к следующему: в [1, 7] разработан способ подборки в позиционном представлении текста, использующий направленный порядок сравнения символов, проведен анализ временных затрат подборки разработанным способом; в [2] разработано модифицированное позиционное представление текста, разработан способ подборки в разработанном представлении, проведен анализ временных затрат подборки разработанным способом; в [3] разработана организация устройства для множественной подборки и модификации данных, обеспечивающее аппаратную поддержку способов подборки и модификации данных; в [4] проведено моделирование работы продукционного способа подборки в модифицированном позиционном представлении текста при различных размерах хранимых подстрок; в [5] модифицирован управляющий цикл работы абстрактной машины поиска данных для поддержки вычислений в хранилищах текстовых данных; в [6] описана организация вычислений в продукционной системе применительно к хранилищам данных; в [8] описана модифицированная продукционная система для текстовых вычислений; в [9] предложен способ и организация матричного устройства параллельной обработки данных, которое позволяет совместить глобальную подборку и локальную модификацию данных; в [10] разработано ассоциативное матричное устройство для множественной подборки данных; в [11] разработано специализированное приложение моделирования работы способов подборки и анализа их временных затрат от параметров входных данных.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, списка литературы и приложения. Основное содержание диссертации изложено на 149 страницах машинописного текста, содержит 48 рисунков, 24 таблицы, список литературы из 107 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИОННОЙ РАБОТЫ

Во введении обоснована актуальность задачи множественной подборки текстовых данных в хранилищах, а также сформулированы цель и задача исследования, научная новизна и основные положения, выносимые на защиту, практическая ценность работы, и другие характеристики работы.

В первой главе проведен анализ современных аппаратных и программных средств, используемых или обеспечивающих множественную подборку текстовых данных в хранилищах.

Результатом выборки или множественной подборки является организованное подмножество данных — система связанных объектов, релевантных запросу, но имеющих различную количественную, качественную, структурную степень близости/ассоциации с поисковыми данными.

Запрос представляет собой процесс отбора из массива индексируемых фрагментов данных таких его элементов, которые удовлетворяют определенным поисковым значениям. При этом большинство современных способов и алгоритмов подборки не ориентированы на параллельную аппаратную обработку входных данных большого размера, таких как данные, находящиеся в хранилищах текстовых данных.

Обзор архитектур аппаратных устройств показал, что они ориентируются на последовательно-конвейерный характер реализации шагов вычислительных процессов множественной подборки и модификации данных на основе полной или сокращенной системы команд. При этом по умолчанию тексты представляются типовыми одномерными структурами данных — строки.

Анализ архитектур и технологий проектирования и производства вычислительных устройств определил программируемые интегральные схемы, как перспективный путь создания специализированных устройств, поскольку программируемые интегральные схемы имеют ряд технологических достоинств (низкая стоимость, надежность и др.), и структурных преимуществ (программируемая структура, большой объем внутренней памяти, многофункциональность и др.) перед микропроцессорами общего назначения.

Оценка основных алгоритмов подборки данных, показал, что по основным требованиям, предъявляемым к множественной подборке текстовых данных в хранилищах (безвозвратный процесс подборки, множественный результат, множественные операнды, метаданные операндов, параллелизм, гибридность представления операндов) наиболее подходящим является алгоритм подборки в суффиксном дереве и алгоритм подборки в позиционном представлении текста. Тем не менее, ни один из них полностью не соответствует всем критериям и специфики подборки в неизменяемых или медленно изменяемых данных. Это обосновывает разработку оригинального способа для множественной подборки текстовых данных в хранилищах.

Сущность предлагаемого в работе подхода заключается в разработке модифицированного позиционного представления текста, содержащего дополнительную информацию о внутренней организации текста, что позволяет осуществлять подборку данных только в выделенных позициях текста, в которых присутствует частичное совпадение входного текста и подстроки, разработке способа множественной подборки в разработанном представлении текстовых данных, и устройства обеспечивающего аппаратную поддержку данного способа и сокращающего временные затраты как подборки, так и модификации текстовых данных в матричном его представлении.

Во второй главе на основе продукционного подхода развивается позиционное преставление текста (ППТ), как представление, позволяющее сократить количество операций множественной подборки данных за счет выполнения подборки только в тех позициях текста, в которых имеется пересечение подстроки и текста как минимум в одной позиции.

Продукционный подход к формированию запросов в базы и хранилища текстовых данных, а также собственно обработке текстовых данных является наиболее распространенным, так как учитывает символьные форматы их представления и базовые операции над ними. Действительно, продукционная обработка символьной информации опирается на правило разбиения текстовых данных на подстроки и последующее манипулирование ими как одномерными объектами т. е. справедливо выражение:

$$\forall i \in [0, |\mathcal{T}| - |\mathcal{P}|], \begin{cases} \mathcal{T} = \langle \mathcal{T}_0, \dots, \mathcal{T}_{i-1}, \mathcal{P}_0, \dots, \mathcal{P}_{|\mathcal{P}|-1}, \mathcal{T}_{i+|\mathcal{P}|}, \dots, \mathcal{T}_{|\mathcal{T}|-1} \rangle \\ \mathcal{T}' = \langle \mathcal{T}_0, \dots, \mathcal{T}_{i-1}, \mathcal{R}_0, \dots, \mathcal{R}_{|\mathcal{R}|-1}, \mathcal{T}_{i+|\mathcal{P}|}, \dots, \mathcal{T}_{|\mathcal{T}|-1} \rangle \end{cases},$$

где i — позиции вхождений входной подстроки в текст;

$\mathcal{T}$ — входной текст размером $|\mathcal{T}|$ символов;

$\mathcal{P}$ — входная подстрока размером $|\mathcal{P}|$ символов;

$\mathcal{R}$ — строка-модификатор размером $|\mathcal{R}|$ символов;

$\mathcal{T}'$ —модифицированный текст.

Вместе с тем повышение эффективности продукционных операций, таких как поиск подстроки и модификация данных возможно за счет введения нелинейности в представление текстовых данных, таких как позиционное представление. Однако основным недостатком подборки данных в ППТ является подборка подстроки полным и многократным, в том числе избыточным, просмотром списков с позициями вхождений каждого символа. Для его устранения в работе предлагается расширить ППТ за счет хранения всех собственных подстрок текста размером до задаваемого пользователем размера $\mathcal{L}$, с соответствующим набором позиций вхождений этих подстрок. Такое представление получило название, модифицированное ППТ (МППТ). Разработанная модификация ППТ, позволяет дополнить классическое позиционное представление новыми элементами, что обеспечивает направленность поиска и исключение ряда неперспективных альтернатив. Пример МППТ «abcabaabca» при $\mathcal{L} = 3$ изображен на рисунке 1.

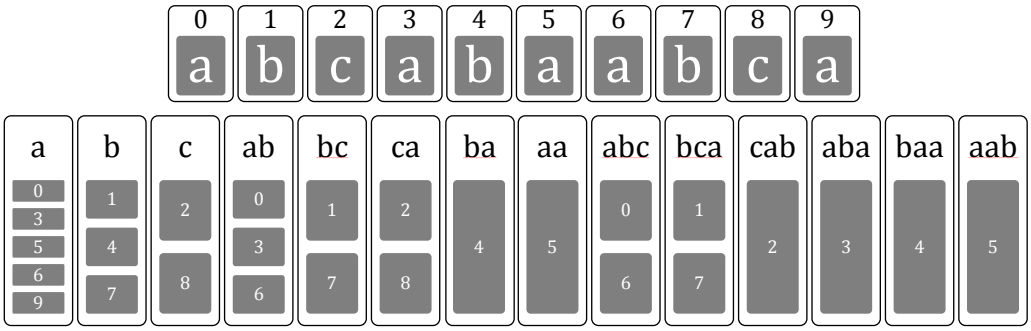

Рисунок 1 – Модифицированное позиционное представление текста «abcabaabca» при $\mathcal{L} = 3$

Такое МППТ является информационной основой для способа множественной подборки данных, в котором можно выделить следующие шаги:

1. Построение модифицированного позиционного представления. Построение МППТ осуществляется последовательным обходом текста, во процессе которого для каждой позиции текста получаются его собственные

подстроки размером равным до $\mathcal{L}$ символов, начинающиеся с текущей позиции, после чего в набор строящегося МППТ для каждой полученной подстроки добавляется позиция ее начала т. е. текущая позиция.

Введем некоторые сокращения, которые используются в данном разделе:

$\mathcal{Z}$ — Модифицированное позиционное представление текста $\mathcal{T}$;

$\mathcal{Z}(\mathcal{s})$ — набор позиций вхождений какой-либо строки $\mathcal{s}$ в текст $\mathcal{T}$;

$\mathcal{Z}_i$ — символ текста $\mathcal{T}$ в позиции i;

$|\mathcal{Z}|$ — размер текста $\mathcal{T}$.

2. Вычисление набора позиций возможных вхождений. Набор позиций возможных вхождений вычисляется на основе набора с минимальным количеством элементов, полученным из МППТ по собственным не пересекающимся подстрокам входной подстроки, размер ℓ которых равен:

$$\ell = \text{Min}(|\mathcal{P}|, \mathcal{L}),$$

где Min — функция получения минимального из параметров.

Решение использовать непересекающиеся подстроки обусловлено тем, что при реализации построения МППТ, для хранения позиций вхождений подстрок используется их хеш-значение, а операция хеширования требует трудоемких вычислений. Набор с минимальным количеством элементов, по которому вычисляется набор позиций возможных вхождений, извлекается из МППТ по подстроке $\mathcal{P}_{\mathfrak{f}\ldots\mathfrak{f}+\ell-1}$, где $\mathfrak{f}$ — позиция ее начала. Такая позиция равна:

$$\mathfrak{f} = Min\left(i \;\middle|\; \begin{cases} i \in \langle 0, \ell, \ldots, \ell \cdot ((|\mathcal{P}| \text{ div } \ell) - 1)\rangle \\ \nexists j \in \langle 0, \ell, \ldots, \ell \cdot ((|\mathcal{P}| \text{ div } \ell) - 1)\rangle \\ |\mathcal{Z}(\mathcal{P}_{i\ldots i+\ell-1})| < \left|\mathcal{Z}\left(\mathcal{P}_{j\ldots j+\ell-1}\right)\right| \end{cases} \right),$$

где div — операция целочисленного деления.

Таким образом набор $\mathcal{B}$ позиций возможных вхождений равен:

$$\mathcal{B} = \left\langle i - \mathfrak{f} \;\middle|\; \begin{cases} i \in \mathcal{Z}\left(\mathcal{P}_{\mathfrak{f}\ldots\mathfrak{f}+\ell-1}\right) \\ i - \mathfrak{f} \in [0, |\mathcal{Z}| - |\mathcal{P}|] \end{cases} \right\rangle.$$

3. Подборка на основе позиций возможных вхождений. Подборка данным способом выполняется на классическом представлении текста и отличается от подборки наивным способом, только тем, что осуществляется только в вычисленных позициях возможных вхождений. Таким образом набор $\mathcal{I}$ позиций вхождений подстроки $\mathcal{P}$ в текст $\mathcal{Z}$ равен:

$$\mathcal{I} = \left\langle i \in \mathcal{B} \;\middle|\; \begin{cases} \forall j \in \langle 0,1, \ldots, |\mathcal{P}| - 1\rangle \\ \mathcal{Z}_{i+j} = \mathcal{P}_j \end{cases} \right\rangle.$$

С алгоритмической точки зрения разработанный способ для множественной подборки данных в МППТ, позволяет вычислить сокращенный набор позиций возможных вхождений что достигается, разбиением входной подстроки на собственные не пересекающиеся подстроки размером ограниченным константой разработанного представления и получением набора позиций возможных вхождений с минимальным количеством элементов среди наборов, полученных по хеш-значению этих подстрок.

В третей главе производится программное моделирование работы разработанного способа для множественной подборки данных, производится анализ временных затрат подборки и предобработки входного текста для

данного способа, а также анализируется зависимость этих временных затрат относительно параметров входных данных. Для моделирования способа подборки в МППТ используется специально разработанное приложение, которое также позволяет моделировать работу следующих способов подборки данных:

– способ подборки Бойера — Мура (БМ);

– способ подборки в модифицированном суффиксном дереве (в МСД);

Способ подборки БМ модифицирован для подборки всех вхождений в текст чтобы соответствовать исследуемой задачи. Суффиксное дерево модифицировано, таким образом, чтобы все позиции, хранящиеся в листьях для всех поддеревьев в суффиксном дереве, хранились в корневой вершине данного поддерева, строится на основе алгоритма Эско Укконена.

В целях определения рационального значения константы $\mathcal{L}$ для МППТ производится моделирование работы способа подборки в МППТ при следующих значениях константы $\mathcal{L}$: 1, 2, 3, 4, 5, 6, 7, 8, 16, 32, 64, 128. Данное моделирование производится для двух типов входных данных, для которых установлены следующие параметры:

1. Искусственно построенные данные.

– размер алфавита: от 2 до 30 (28 итераций);

– размер текста: от 150 до 100000 (100 итераций);

– размер подстроки: от 1 до 128 (127 итераций).

2. Естественные данные.

– размер текста: от 150 до 100000 (100 итераций);

– размер подстроки: от 1 до 128 (127 итераций).

По результатам моделирования работы способа подборки в МППТ построены графики зависимости временных затрат подборки (рисунок 2а) и предобработки входного текста (рисунок 2б) относительно константы $\mathcal{L}$. Для определения рационального значения константы $\mathcal{L}$ графики были усреднены по типу данных и объединены по временным затратам, от минимальных до максимальных (рисунок 2в).

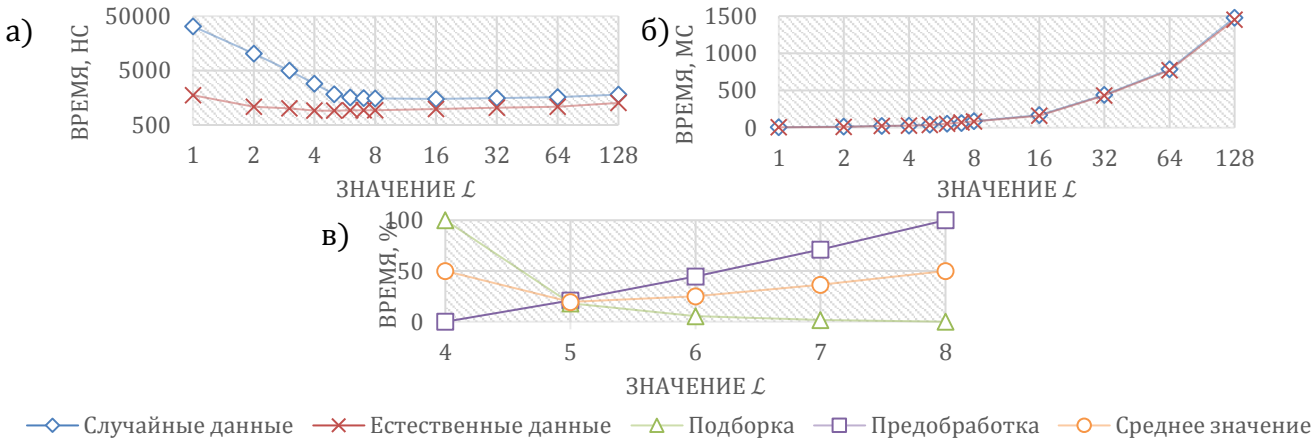

Рисунок 2 – Графики зависимости временных затрат способа подборки в МППТ относительно константы $\mathcal{L}$

МППТ позволяет проводить гибкую настройку соотношения временных затрат подборки к временным затратам его построения, установкой

подходящего значения константы, что позволяет расширить область применения МППТ и использовать предварительные знания о входных данных для оптимизации множественной подборки. Если такая информация отсутствует задаваемый пользователем размер подстроки $\mathcal{L}$ рекомендуется выбирать равным 5, что эмпирически обеспечивает оптимальное соотношение временных затрат подборки данных и предобработки входного текста.

Для сравнения и анализа характеристик разработанного способа подборки в МППТ при $\mathcal{L} = 5$ со способом подборки в МСД, который является одним из самых быстрых способов подборки для решения исследуемой задачи и со способом подборки в БМ, который является одним из самых быстрых среди способов общего назначения производится моделирование работы этих способов для двух типов входных данных, для которых установлены следующие параметры:

1. Искусственно построенные данные.

– размер алфавита: от 2 до 30 (28 итераций);

– размер текста: от 50 до 500000 (100 итераций);

– размер подстроки: от 1 до 20 (19 итераций).

2. Естественные данные.

– размер текста: от 50 до 500000 (100 итераций);

– размер подстроки: от 1 до 20 (300 итераций).

В результате данного моделирования получены средние временные затраты подборки и предобработки входного текста для каждого способа таблица 1.

Таблица 1 – Средние временные затраты

Способ	Время подборки, нс		Время предобработки, мс	
	Искусственные	Естественные	Искусственные	Естественные
БМ	268676	196839	0	0
в МСД	168.2	179.1	700	697
в МППТ	151.4	148.2	121	67

Анализ способов для множественной подборки текстовых данных в хранилищах показал, что разработанный способ подборки позволяет сократить временные затраты относительно способа подборки в МСД на 10 % для искусственно построенных данных и на 17 % для естественных данных.

По результатам моделирования построены графики зависимости временных затрат подборки относительно размера входного текста для искусственно построенных (рисунок 3а) и естественных (рисунок 3б) данных, графики зависимости временных затрат предобработки входного текста от его размера для искусственно построенных (рисунок 3в) и естественных (рисунок 3г) данных. Данные графики демонстрируют, что временные затраты как подборки, так и предобработки для моделируемых способов увеличивается с ростом размера входного текста, а самую меньшую зависимость имеет разработанный способ.

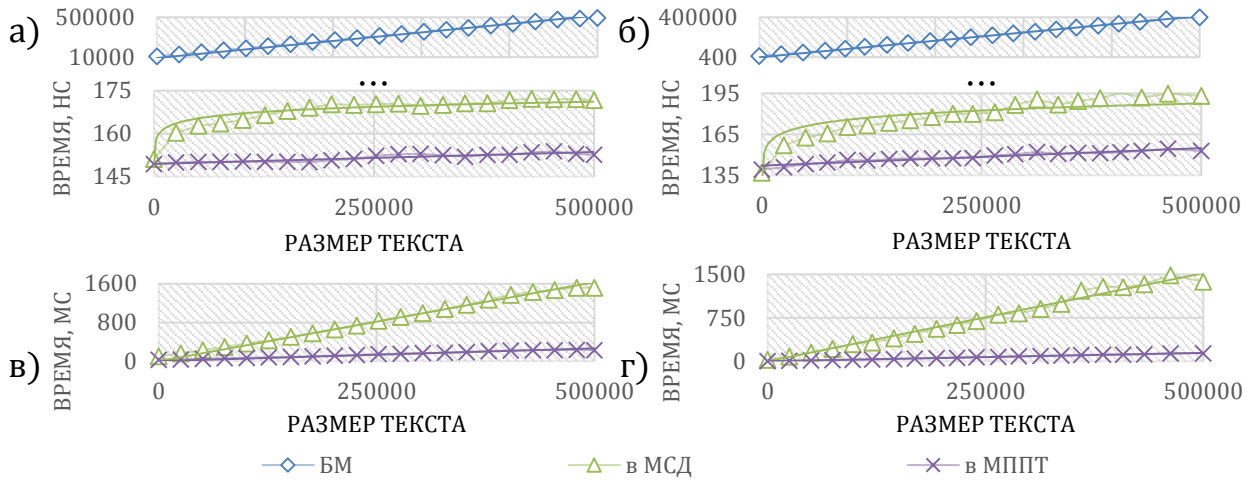

Рисунок 3 – Графики зависимости временных затрат относительно размера входного текста

В результате аппроксимации этих графиков получены уравнения (таблицы 2), подтверждающие такую зависимость.

Таблица 2 – Результаты аппроксимации графиков зависимости временных относительно размера входного текста

Способ	Этап	Уравнение		Коэффициент детерминации	
		Искусственные	Естественные	Искусственные	Естественные
БМ	Подборка	$10^{-6}x$	$8{\cdot}10^{-7}x$	0.9916	0.9973
в МСД	Подборка	$0.0001x^{0.0151}$	$0.0001x^{0.0404}$	0.9107	0.8420
	Предобр.	$0.0033x$	$0.003x$	0.9902	0.9832
в МППТ	Подборка	$8^{-12}x+0.0001$	$3{\cdot}10^{-11}x+0.0001$	0.8432	0.9136
	Предобр.	$0.0005x$	$0.0003x$	0 9422	0.9954

Однако способы подборки в МСД и МППТ используют предобработку входного текста, в отличии от способа подборки БМ, поэтому для оценки эффективности данных способов также построены графики зависимости суммарных временных затрат работы с текстом (суммарные временные затраты подборок + временные затраты предобработки входного текста) относительно количества операций подборки для искусственно построенных (рисунок 4а) и естественных (рисунок 4б) данных. Данные графики алгоритмически подтверждают достижение цели исследования — сокращение временных затрат на операции множественной подборки. Действительно, анализ способов для множественной подборки текстовых данных в хранилищах показал, что количество операций подборки, которое необходимо разработанному способу, чтобы его общие временные затраты работы с текстом стали меньше, чем общие временные затраты работы с текстом способа подборки БМ, для искусственно построенных данных равно 451 (рисунок 4а) в то время как для естественных данных это количество равно 341 (рисунок 4б).

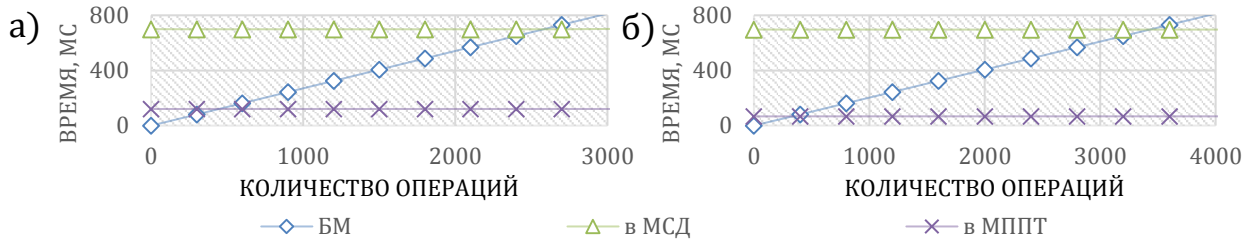

Рисунок 4 – Графики зависимости суммарных временных затрат работы с текстом относительно количества операций подборки

В четвертой главе разрабатывается специализированное ассоциативное устройство для множественной подборки текстовых данных в хранилищах, его структура и схемы функциональных блоков. Проводится моделирование разработанного устройства, устройства замены с однородной операционной частью (прототип) и устройства замены на базе циклического условного сдвигателя (аналог) на ПЛИС в программе Altera Quartus II, анализируется их аппаратная сложность и временные затраты.

Структурно-функциональная организация ассоциативного устройства для множественной подборки данных изображена на рисунке 5. Новизна которой состоит в аппаратной реализации операций последовательного и межстрочного сдвига при одномерном и двумерном циклическом представлении обрабатываемых текстовых данных с возможностью временной буферизации граничных элементов строк, выходящих за пределы операционного блока устройства, что позволяет ускоренно реализовывать шаги обработки текстовых данных при различных сочетаниях размеров подстроки и строки-модификатора, в том числе выполнять реверсивные межстрочные сдвиги.

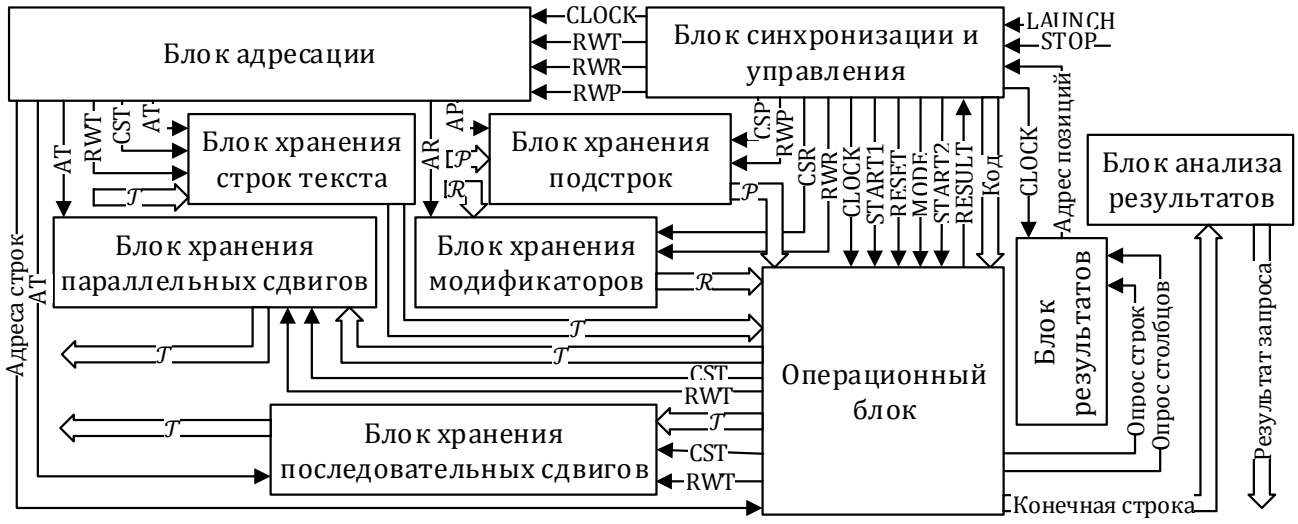

Рисунок 5 – Структурно-функциональная организация разработанного устройства

Разработанное устройство содержит блок синхронизации и управления (БСУ), блок адресации (БА), блок хранения строк текста (БХТ), блок хранения подстрок (БХП), блок хранения строк-модификаторов (БХМ), блок хранения последовательных сдвигов (БХПС), блок хранения параллельных сдвигов

(БХМС), операционный блок (ОБ), блок результатов (БР) и блок анализа результатов (БАР). Основным блоком устройства является операционный блок (рисунок 6), который состоит из ассоциативных запоминающих элементов 1, коммутационных элементов 10, элементов-селекторов 12, маскирующего элемента 36 и преобразователя кода 53.

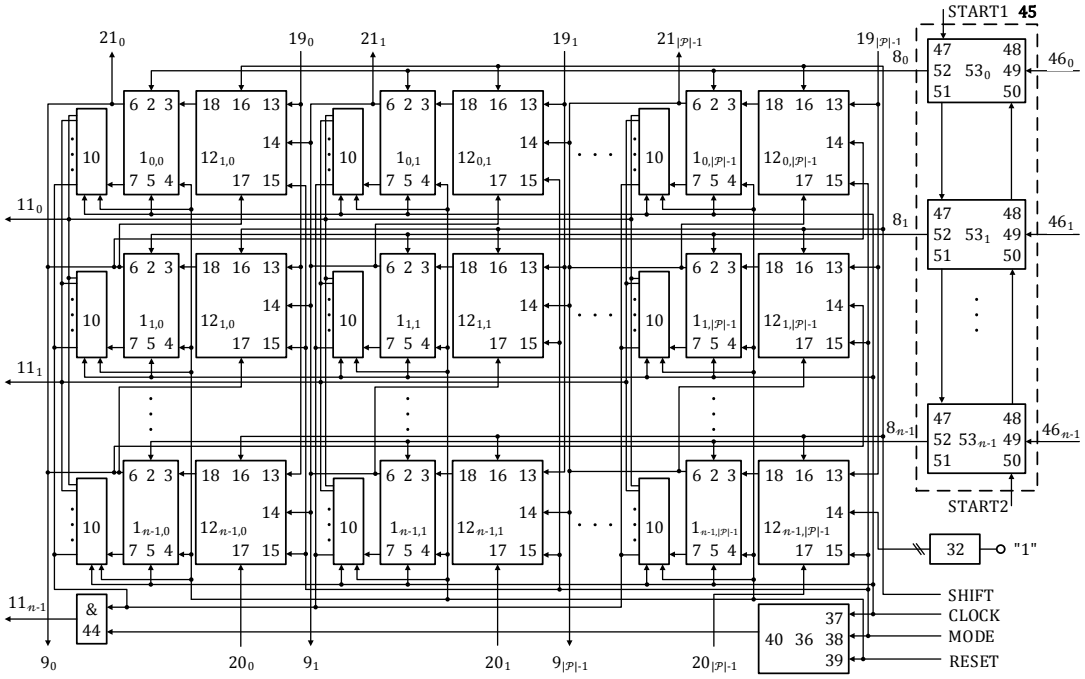

Рисунок 6 – Операционный блок разработанного устройства

Ассоциативный запоминающий элемент 1 (рисунок 7а) реализует функции хранения и сравнения. Межстрочный сдвиг осуществляется при помощи элемента-селектора 12 (рисунок 7б), подключением к выходу 18 входа 17 при подаче сигналов MODE = 0 и SHIFT = 1. Последовательный сдвиг осуществляется подключением к выходу 18 входа 14 при MODE = 1.

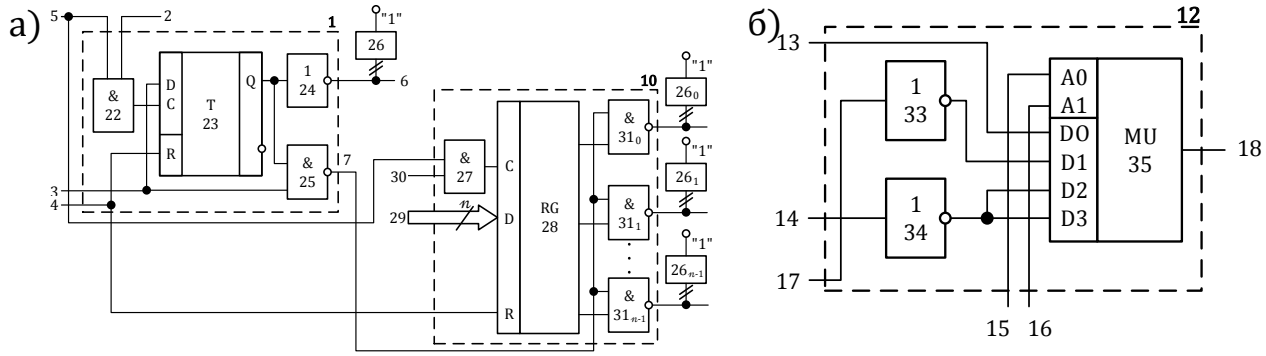

Рисунок 7 – Ассоциативный запоминающий элемент и элемент-селектор

Аппаратная сложность ($\mathcal{k}$) разработанного устройства равна сумме аппаратной сложности всех блоков устройства:

$$\mathcal{k}_{\text{устр.}} = \mathcal{k}_{\text{БСУ}} + \mathcal{k}_{\text{БА}} + \mathcal{k}_{\text{ОБ}} + \mathcal{k}_{\text{БХТ}} + \mathcal{k}_{\text{БХП}} + \mathcal{k}_{\text{БХМ}} + \mathcal{k}_{\text{БХПС}} + \mathcal{k}_{\text{БХМС}} + \mathcal{k}_{\text{БР}} + \mathcal{k}_{\text{БАР}} =$$
$$= Q_{\mathcal{P}} \cdot (36 \cdot |\mathcal{P}| + 111) + Q_{\mathcal{T}} \cdot (36 \cdot |\mathcal{T}| + 111) + Q_{\mathcal{R}} \cdot (72 \cdot |\mathcal{P}| + 111) +$$
$$+ |\mathcal{T}| \cdot (40 \cdot n + 98) + 117 \cdot |\mathcal{P}| + 526 + 52 \cdot n,$$

где $Q_{\mathcal{T}}$ — максимально допустимое количество строк текста в БХТ;

$Q_{\mathcal{P}}$ — максимально допустимое количество подстрок, хранимых в БХП;

$Q_{\mathcal{R}}$ — максимально допустимое количество строк-модификаторов в БХМ.

В целях верификации разработанного способа параллельной замены и устройства, его реализующего, а также анализа его временных затрат проведено моделирование операционного блока разработанного устройства, его прототипа и аналога в системе Quartus II.

В качестве параметров входных данных для моделирования устройств был принят размер текста равный $|\mathcal{T}| = 20$, размер подстроки равный $|\mathcal{P}| = 4$, размер строки-модификатора равный $|\mathcal{R}| = 8$.

Для получения характеристик устройств используется модель ПЛИС 5SGSED8I2F45C2 семейства Stratix V фирмы Altera. Выбранная ПЛИС обладает следующими характеристиками: рабочее напряжение — 0.85 В, количество входов/выходов доступных пользователю — 1158, количество ALMs — 262400, память — 35942400 бит.

Синтезированные имитационные модели устройств успешно проходят проверку на предмет корректности связей, переходных процессов и иных параметров в системе Quartus. Таким образом, они позволяют проанализировать временные затраты работы устройств, в результате которого получены характеристики, представленные в таблице 3.

Таблица 3 – Характеристики операционных блоков устройств

Устройство	Аппаратная сложность, ALMs	Частота, МГц
аналог	23	259
прототип	29	397
разработанное	26	286

Аппаратная сложность операционного блока разработанного устройства равна 26 ALMs, а рабочая частота 286 МГц. Для уточнения результата и приведения сравниваемых данных к одинаковым пропорциям времени необходимо знать временные затраты необходимые устройствам для получения результата работы.

Получение этих временных затрат осуществляется программой ModelSim на трех наборах входных данных. Результаты моделирования работы устройств в ModelSim приведены в таблице 4.

Таблица 4 – Результаты моделирования работы устройств в ModelSim

Время модификации, мс		
устройство-аналог	устройство-прототип	разработанное устройство
1057	230	200
605	187	161
1248	261	221

Таким образом, исходя из полученного среднего времени модификации данных разработанное устройство позволяет сократить временные затраты относительно прототипа на 14 % и на 80 % относительно аналога.

В заключении сформулированы основные результаты и выводы диссертационной работы.

В приложении представлены листинги моделирующих программ.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ РАБОТЫ

В работе решена научно-техническая задача, заключающаяся в разработке программно-аппаратных средств, реализующих строковые операции подборки и модификации текстовых данных в хранилищах.

1. Разработано модифицированное позиционное представление текста, содержащее дополнительную информацию о внутренней организации текста. Такой дополнительной информацией являются позиции вхождения в текст его собственных подстрок, до задаваемого пользователем размера. Данное представление позволяет дополнить классическое позиционное представление новыми элементами, что обеспечивает направленность поиска и исключение ряда неперспективных альтернатив.
2. Разработанное модифицированное позиционное представление текста для хранилищ текстовых данных позволяет выполнять гибкую настройку соотношения временных затрат подборки к временным затратам его построения, установкой подходящего значения максимального размера подстрок, хранящихся в данном представлении, что позволяет расширить область его применения и использовать априорные знания о входных данных для сокращения временных затрат множественной подборки. Если такая информация отсутствует размер подстроки рекомендуется выбирать равным 5, что эмпирически обеспечивает оптимальное соотношение временных затрат подборки данных и предобработки входного текста.
3. Разработан способ для множественной подборки данных в модифицированном позиционном представлении текста, позволяющий вычислять сокращенный набор позиций возможных вхождений. Проведенный анализ временных затрат разработанного способа показал, что разработанный способ подборки позволяет сократить временные затраты относительно способа подборки на основе суффиксного дерева на 10 % для искусственно построенных данных и на 17 % для естественных данных.
4. Анализ способов для множественной подборки текстовых данных в хранилищах показал, что количество операций подборки, которое необходимо разработанному способу, чтобы общие временные затраты работы с текстом разработанного способа подборки данных стали меньше, чем общие временные затраты работы с текстом способа подборки Бойера — Мура, для искусственно построенных данных равно 451 в то время как для естественных данных это количество равно 341.
5. Разработана структурно-функциональная организация специализированного устройства для множественной подборки текстовых данных в хранилищах, новизна которой состоит отличающаяся в аппаратной

реализации операций последовательного и межстрочного сдвига при строковом и матричном представлении обрабатываемых текстовых данных с возможностью временной буферизации граничных элементов строк, выходящих за пределы операционного блока устройства, что позволяет ускоренно реализовывать шаги обработки текстовых данных при различных сочетаниях размеров подстроки и строки-модификатора, в том числе выполнять реверсивные межстрочные сдвиги. Разработаны схемотехнические решения основных блоков и узлов специализированного устройства для множественной подборки текстовых данных в хранилищах, отличающиеся аппаратной поддержкой шагов обработки текстовых данных в строковом и матричном представлении текста.

6. Оценка характеристик разработанного ассоциативного устройства на программируемой логической интегральной схеме Stratix V 5SGSED8I2F45C2 фирмы Altera в среде Quartus II, показала, что аппаратная сложность операционного блока разработанного устройства равна 26 ALMs, а рабочая частота 286 МГц. Анализ временных затрат модификации данных разработанным устройством в программе ModelSim показал, что данное устройство позволяет сократить временные затраты относительно устройства-прототипа на 14 % и на 80 % относительно устройства-аналога.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Рецензируемые научные журналы и издания

1. Гришин, Д. С. Алгоритм поиска подстроки на основе позиционного представления текста для архивно-текстовых данных / Д. С. Гришин, Е. А. Титенко, Н. А. Милостная [и др.] // Известия Юго-Западного государственного университета. — 2015. — № 3(16). — С. 43–48.

2. Гришин, Д. С. Алгоритм построения структуры, представляющей строку в виде хеш-таблицы, состоящей из хешей подстрок данной строки и алгоритм поиска в ней / Д. С. Гришин, Е. А. Титенко // Известия Юго-Западного государственного университета. — 2015. — № 6(63). — С. 62–69.

3. Гришин, Д. С. Ассоциативное матричное устройство для обработки строковых данных в хранилищах текстовой информации / Д. С. Гришин, Е. А. Титенко // Информационные системы и технологии. — 2017. — № 3(101). — С. 72–81.

4. Гришин, Д. С. Способ подборки данных в хранилищах текстовых данных на основе продукционного подхода и моделирование его работы // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 9, № 3 (2017) http://naukovedenie.ru/PDF/77TVN317.pdf (доступ свободный).

Материалы конференций

5. Гришин, Д. С. Модифицированный цикл работы машины вывода / Д. С. Гришин, Е. А. Титенко, В. С. Фастов // Интеллектуальные информационные системы: тенденции, проблемы, перспективы: сборник материалов региональной заочной научно-практической конференции — Курск, 2013. — С. 102–105.

6. Гришин, Д. С. Исчислительная продукционная система и организация параллельных выводов / Д. С. Гришин, Е. А. Титенко, С. Ю. Сазонов // Передовые научные разработки — 2013: сборник трудов IX международной научно-практической конференции. — Прага, 2013. — С. 5–8.

7. Гришин, Д. С. Алгоритм поиска подстроки, использующий позиционное представление текста в качестве дополнительной структуры / Д. С. Гришин, Е. А. Титенко, В. А. Ханис // Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации: материалы XII международной научно-технической конференции «Распознавание — 2015» / ЮЗГУ. — Курск, 2015. — С. 107–110.

8. Гришин, Д. С. Модифицированная продукционная система для параллельных символьных вычислений / Д. С. Гришин, Е. А. Титенко, М. А. Шевченко // ИИС-2015: материалы докладов III региональной заочной научно-практической конференции / ЮЗГУ. — Курск, 2015. — С. 128–130.

9. Гришин, Д. С. Способ и устройство аппаратной поддержки процессов распознавания образов / Д. С. Гришин, Е. А. Титенко, С. Г. Емельянов [и др.] // Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации: материалы XIII международной научно-технической конференции «Распознавание — 2017» / ЮЗГУ. — Курск, 2017. — С. 148–150.

Патенты на изобретение

10. Пат. 2569567 Российская Федерация, МПК G11C 15/00. Способ и ассоциативное матричное устройство для обработки строковых данных / Д. С. Гришин, Е. А. Титенко, А. В. Белокопытов [и др.]; заявитель и патентообладатель Юго-Западный государственный университет. — № 2014110753/08; заявл. 21.03.2014; опубл. 27.11.2015, Бюл. № 33.

Свидетельство о государственной регистрации программы для ЭВМ

11. Гришин, Д. С. Программа тестирования поисковых алгоритмов с применением хэш-данных / Д. С. Гришин, Е. А Титенко, А. В. Гривачев [и др.] // Свидетельство о государственной регистрации программы для ЭВМ № 2016663385, РФ. Зарегистр. в Реестре программ для ЭВМ 07.12.2016 г. Правообладатель: Юго-Западный государственный университет.

Соискатель Гришин Дмитрий Сергеевич

Подписано в печать « » _____ 2017 г .
Формат 60×84 1/16.
Печ. л. 1.0. Тираж 120 экз. Заказ _____.
Юго-Западный государственный университет.
Издательско-полиграфический центр
Юго-Западного государственного университета.
305040, г. Курск, ул. 50 лет Октября, 94.