Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого»

На правах рукописи

Успенский Михаил Борисович

Разработка и исследование методов и моделей обработки диагностической информации для обнаружения и локализации неисправностей в системах хранения данных

Специальность 05.13.01 – Системный анализ, управление и обработка информации

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2020

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Санкт-Петербургский политехнический университет Петра Великого».

Научный руководитель:	кандидат технических наук, доцент **Ицыксон Владимир Михайлович** СПбПУ, директор Высшей школы интеллектуальных систем и суперкомпьютерных технологий
Официальные оппоненты:	
доктор физико-математических наук, Борис Александрович Кулик	Федеральное Государственное Бюджетное Учреждение Науки Институт Проблем Машиноведения Российской Академии Наук, лаборатория интеллектуальных электромеханических систем, ведущий научный сотрудник
кандидат технических наук, доцент, Мелехова Анна Леонидовна	МФТИ, старший преподаватель
Ведущая организация:	**Акционерное общество «Концерн «Научно-производственное объединение «Аврора»**

Защита состоится 26 ноября 2020 года в 16:00 на заседании диссертационного совета У.05.13.01 федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого» (195251, г. Санкт-Петербург, ул. Политехническая, д. 29, III учебный корпус, аудитория 506). С диссертацией можно ознакомиться в библиотеке и на сайте http://www.spbstu.ru. федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого»,

Автореферат разослан «__» _____2020 г.

Ученый секретарь диссертационного совета У.05.13.01

Доктор технических наук, доцент А.Е. Васильев

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Проблема своевременного обнаружения неисправностей в системах хранения данных (СХД) имеет в настоящее время большое значение в связи с резким ростом объема хранимой различными способами информации. По данным совместного доклада экспертов компаний IDC и Seagate, к 2025 году общий объем хранимых данных может превысить 160 зеттабайт. При этом наблюдается устойчивая тенденция к увеличению доли данных, размещенных централизованно, в корпоративных и коммерческих центрах хранения и обработки данных (ЦХОД) с применением СХД различного уровня.

СХД активно используются в банковской сфере и телекоммуникационной сферах, органах государственной власти, в предприятиях оборонно-промышленного и топливно-энергетического комплексов, в сфере образования и науки для хранения персональных данных, финансовых и нормативных документов, проектной документации, обучающих материалов и т.д.

Для повышения сохранности данных и обеспечения их постоянной доступности разработчики СХД применяют все более сложные аппаратные и программные решения, направленные на создание схем избыточности и кэширования на разных уровнях топологии, поэтому современные СХД являются комплексными аппаратно-программными системами, включающими в себя множество взаимосвязанных элементов. Сложность диагностики таких систем постоянно возрастает, так как, по мере увеличения объема хранимых данных, растет сложность применяемых решений. Как следствие, растет важность своевременного и точного обнаружения неисправностей не только в носителях информации (жестких дисках, твердотельных дисках), но и в элементах, не предназначенных непосредственно для хранения данных (контроллерах хранения, элементах сетевой инфраструктуры), а также неисправностей, возникающих в результате межэлементного взаимодействия.

Для обнаружения таких неисправностей необходима разработка новых методов и моделей, обеспечивающих комплексный подход к диагностике на основе анализа расширенного множества параметров СХД. В частности, настоящая работа посвящена решению задач контроля технического состояния, а также поиска мест и определения причин возникающих неисправностей.

Степень разработанности темы исследования.

К СХД применимы актуальные подходы к решению задач диагностирования вычислительных систем, основанные на применении моделей и средств обработки данных мониторинга. Диагностическая модель при этом определяется в соответствии с ГОСТ 20911-89 как формализованное описание объекта, необходимое для решения задач диагностирования.

Известны современные работы: в области диагностики вычислительных систем с использованием моделей С. Чена, Л. Кейроша; в области диагностики на основании анализа данных исследования Дж. Дина, Г. Ванга, Д. Дасгупты, Д. Ли, Ф. Гонсалеса, К. Эйрас-Франко. Диагностика с использованием комбинированного подхода в различных технических объектах рассматривается

в работах А. Слимани, Дж. Луо, Д. Юнга, К. Сундстрема, С. Франка, М. Хини. Диагностическая модель определяется в соответствии с ГОСТ 20911-89 как формализованное описание объекта, необходимое для решения задач диагностирования.

Методы диагностики появления неисправностей носителей информации рассмотрены в работах, выполненных в лабораториях компаний "Майкрософт", "Гугл", "Фейсбук" и др., а также независимыми исследователями: И. Нараянаном, Д. Вангом, Дж. Мюрреем, Ф. Махдисолтани и др.

Одним из наиболее перспективных направлений исследований является диагностика вычислительных систем с использованием журналов программного обеспечения (ПО). Различные аспекты такой диагностики рассмотрены в работах Р. Вааранди, Ф. Киянга, Ж. Мина, С. Бертеро, С. Мессуди и др.

Актуальные технические решения в области диагностики СХД отражены в пакетах программных продуктов компаний "Ай-Би-Эм", "Фуджитсу", "Эйч-Пи", "Делл", "Заббикс" и др.

Анализ состояния предметной области показал необходимость дальнейшего исследования подходов, позволяющих расширить число классов обнаруживаемых неисправностей за счёт возможности обработки разнородной диагностической информации и формализации экспертного знания о функционировании СХД, в том числе путем совместного применения диагностических моделей и методов, основанных на обработке данных мониторинга.

Целью диссертационной работы является расширение множества классов обнаруживаемых неисправностей в системах хранения данных, создание новых методов обработки и анализа диагностической информации и создание инструментальных средств для автоматизации процесса диагностирования.

Научная задача диссертационной работы заключается в разработке моделей, методов и алгоритмов, расширяющих множество возможных классов обнаруживаемых неисправностей в системах хранения данных.

Для достижения поставленных целей в работе решаются следующие задачи:

1) Анализ систем класса СХД как объекта диагностики, определение требований к методам обнаружения неисправностей в СХД и реализующему их ПО. Анализ существующих научных работ и прикладных инструментов в области диагностики вычислительных систем, выполнение сравнительного анализа их характеристик.

2) Разработка метода построения диагностических моделей для систем класса СХД, определяющих отношения между параметрами и возможными состояниями системы и ее элементов, где отношения могут быть заданы и как детерминированные связи между диагностическими сущностями, и как функции машинного обучения.

3) Разработка методов и программных средств преобразования диагностической модели к упрощенному графовому виду для её применения в составе диагностического ПО.

4) Разработка методов и инструментальных средств обнаружения неисправностей в СХД, основанных на анализе текстовой информации, получаемой в процессе мониторинга СХД методами классификации текста с использованием машинного обучения.

5) Экспериментальная проверка разработанных моделей, методов и средств на целевой платформе СХД.

Объектом исследования являются СХД и отношения между диагностическими параметрами, состоянием отдельных элементов СХД и СХД в целом.

Предметом исследования являются модели и методы обнаружения неисправностей в СХД.

Научная новизна положений, выносимых на защиту, заключается в следующем:

- В работе предложен и применен метод построения диагностических моделей систем хранения данных, основанный на использовании онтологической модели и методов обработки и анализа экспертной информации, отличающийся от существующих возможностью задания связей между объектами онтологии путем использования алгоритмов машинного обучения.
- В работе предложен и применен подход к обнаружению неисправностей в системах хранения данных, отличающийся от существующих использованием алгоритма классификации частично структурированных текстовых данных мониторинга программного обеспечения систем хранения данных, основанного на применении методов машинного обучения.
- В работе предложен алгоритм анализа, трансформации и обработки текстовой информации, получаемой в процессе мониторинга программного обеспечения систем хранения данных, отличающийся от известных тем, что позволяет обнаруживать неисправности на основании классификации частично структурированных текстов без детального анализа структуры, формата и порядка поступления сообщений мониторинга.
- В работе предлагается метод обнаружения неисправностей в системах хранения данных, отличающийся от существующих совместным применением онтологической модели и алгоритмов машинного обучения для обработки текстовой информации, получаемой в процессе мониторинга систем хранения данных.

Теоретическая и практическая значимость исследования.

Предложенный в настоящей работе подход развивает научные основы построения диагностических моделей, предназначенных для автоматического и автоматизированного обнаружения неисправностей в СХД, путем применения онтологической модели для описания отношений между состоянием объекта, его элементами и диагностическими параметрами с расширением аппарата методов онтологического моделирования, за счёт добавления возможности описания отношения между понятиями, экземплярами и атрибутами при помощи внешних процедур, реализующих применение алгоритмов машинного обучения.

Практическая значимость исследования заключается в разработке и практической реализации в рамках диагностического ПО метода обнаружения

неисправностей в СХД, использование которого в процессе функционирования СХД позволит повысить надежность хранения данных и обеспечить бесперебойный доступ к ним. Полученное решение может применяться для диагностирования широкого спектра конфигураций СХД путем изменения набора элементов онтологической модели. При этом, если конфигурации СХД отличаются только применением различных количественных характеристик схем избыточности, то какие-либо действия по адаптации онтологической модели не требуются, а полученное решение сохраняет работоспособность в условиях масштабируемости СХД.

Методология и методы диссертационного исследования базируются на междисциплинарном подходе с применением методов диагностики на основании диагностических моделей и методов диагностики, основанных на анализе закономерностей в данных мониторинга, в том числе методами теории графов, теории распознавания образов, онтологического моделирования и семантических сетей, методами обработки естественного языка с использованием средств машинного обучения.

Положения, выносимые на защиту:

1) Метод построения диагностических моделей СХД, отличающихся конфигурацией аппаратных средств, составом программных средств и параметрами использованных схем избыточности, позволяющий обнаруживать большее число типов неисправностей относительно существующих решений за счёт обеспечения возможности совместного использования диагностических параметров разного рода.

2) Метод и алгоритм анализа, трансформации и обработки текстовой информации, получаемой в процессе мониторинга СХД, позволяющий, в отличие от существующих решений, обнаруживать неисправности без детального анализа структуры данных мониторинга, формата и последовательности текстовых сообщений.

3) Комплексный метод обнаружения неисправностей в СХД, основанный на совместном использовании диагностической модели СХД и метода обработки текстовых данных мониторинга СХД с использованием алгоритмов машинного обучения, позволяющий масштабировать онтологическую модель СХД и обеспечивающий увеличение числа обнаруживаемых типов неисправностей относительно существующих средств.

Обоснованность и достоверность научных результатов достигается за счёт использования апробированного математического аппарата, соответствия экспериментальных данных теоретическим предположениям, а также успешного применения разработанных методов и моделей в экспериментальном образце аппаратно-программного комплекса обнаружения неисправностей в СХД.

Реализация результатов работы. Результаты, полученные в настоящем исследовании, использованы для разработки опытного образца программно-аппаратного комплекса предотвращения сбоев в СХД в ходе выполнения работ по разработке программно-аппаратного комплекса прогнозирования неисправностей, выполненных при финансовой поддержке Министерства образования и науки Российской Федерации в рамках Федеральной целевой

программы «Исследования и разработки по приоритетным направлениям развития научно-технического комплекса России на 2014-2020 годы», соглашение о предоставлении субсидии от 03.10.2017 г. № RFMEFI581 17X 0023.

Апробация полученных результатов. Полученные в ходе диссертационного исследования результаты представлены на 7 российских и международных конференциях: Topical Problems of Architecture, Civil Engineering and Environmental Economics, Москва, Россия, 2018; XXIII Международная научно-практическая конференция «Системный анализ в проектировании и управлении» (г. Санкт-Петербург, Россия, 2018 г.); Международная конференция по мягким вычислениям и измерениям (г. Санкт-Петербург, Россия, 2018 г.); International Conference Cyber-Physical Systems and Control (г. Санкт-Петербург, Россия, 2019 г.); 17th IEEE International Symposium on Intelligent Systems and Informatics (г. Суботица, Сербия, 2019 г.); 33rd International Business Information Management Association Conference (IBIMA) (г. Мадрид, Испания, 2019 г.); 2019 International Scientific Conference on Energy, Environmental and Construction Engineering (EECE) (г. Санкт-Петербург, Россия, 2019 г.).

Публикации. Основные результаты по теме диссертационной работы опубликованы в 13 научных работах и приравненных к ним публикациях, в том числе 3 – в журналах из списка рекомендуемых ВАК и 5 – в журналах, индексируемых в базах SCOPUS и Web of Science. По результатам разработки практической части диссертационной работы зарегистрировано 9 программ для ЭВМ.

Личный вклад. Все результаты, представленные в настоящей диссертации, получены автором лично.

Структура и объем диссертационной работы. Диссертация состоит из введения, пяти глав, заключения и 2 приложений. Объем основной части диссертации составляет 150 страницы, полный объем диссертационной работы – 153 страницы, и включает в том числе 27 таблиц, 24 рисунка. Список литературы содержит 154 наименования.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность темы исследования, определяется цель и решаемые задачи, объекты и предмет исследования, формулируются положения, выносимые на защиту, их теоретическая и практическая значимость и научная новизна.

В первой главе приводится описание СХД как объекта диагностики с учётом особенностей реализации целевой платформы. Рассмотрены основные элементы СХД, их взаимосвязь, причины возникновения неисправностей и особенности распространения ошибок в подсистемах СХД.

Проведен сравнительный анализ наиболее актуальных на текущий момент программных и программно-аппаратных средств, предназначенных для обнаружения, предсказания и предотвращения возникновения неисправностей в вычислительных системах, которые применяются или могут быть применены

для диагностики СХД. Предложена их классификация по области применения, по принципу сбора и типу данных значений диагностических параметров, получаемых в результате процесса мониторинга и используемых в процессе диагностики.

Определен и обоснован набор критериев для характеристики основных преимуществ и недостатков проанализированных инструментальных средств, проведено их сравнение в соответствии с данными критериями.

На основании выявленного перечня подходов, наиболее подходящих для обнаружения неисправностей в СХД, проведен анализ современных научных публикаций, направленный на определение наиболее актуальных методов реализации данных подходов, в том числе перспективных алгоритмов обнаружения аномалий в диагностических данных, алгоритмов классификации и кластеризации.

Определена потребность в разработке новых методов и инструментальных средств для обнаружения неисправностей, направленных на более эффективное использование журналов ПО СХД, в том числе как элемента диагностической модели СХД.

Во второй главе представлено описание метода построения диагностической модели, предназначенной для обнаружения неисправностей в СХД. Фактически, задача построения диагностической модели сводится к формализации и определению связей между состоянием объекта, его элементами и диагностическими параметрами, включающими в себя, в том числе, текстовые данные мониторинга. Текстовые данные мониторинга накапливаются в журналах ПО СХД в виде сообщений, последовательно регистрируемых по мере возникновения тех или иных событий в системе.

В рамках исследования предложен подход к созданию онтологической диагностической модели СХД. Под онтологической моделью при этом понимается база знаний на основе онтологии надежности СХД, созданной с применением традиционной методологии, интерпретируемая с использованием новых алгоритмов диагностики и анализа системных журналов для определения технического состояния системы.

Разработанная модель обладает следующими свойствами:

- Онтологическая модель позволяет формально описывать связи между разнородными знаниями о поведении СХД и элементов СХД, получаемыми по результатам экспертных оценок, из анализа статистических данных и исторических данных о функционировании СХД, результатов моделирования и диагностических параметрах разного рода.
- Онтологическая модель позволяет формально описывать иерархическую структуру СХД, их компонентов и подсистем;
- Онтологическая модель удовлетворяет требованию масштабируемости, то есть обеспечивает возможность адаптации для различных конфигураций и архитектур СХД, при этом масштабирование модели происходит с минимальным внесением изменений в реализующий процедуру диагностики программный код.

В диссертационной работе определяются основные понятия (классы) онтологической модели, позволяющие формально описать структуру объекта диагностирования (СХД) и его элементов, возможные состояния объекта диагностирования и его элементов, возможные события в объекте диагностирования и его элементах, а также диагностические параметры.

Понятие **«параметр СХД»** (*P)* предназначено для описания диагностических параметров СХД. Получаемые в процессе мониторинга значения параметров СХД классифицируются в зависимости от описанных в модели возможных значений параметра *{V}* как соответствующие тому или иному событию.

Понятие «**компонент СХД»** (*K*) предназначено для описания базовых структурных элементов СХД. На уровне модели компонент является атомарным и неделимым, то есть максимальная глубина поиска места отказа (неисправности) ограничивается уровнем компонентов.

Понятие **«подсистема СХД»** *(Ss)* предназначено для описания группы элементов СХД, объединенных по какому-либо функциональному признаку. Подсистема СХД может использоваться для группировки понятий не только компонентов, но и подсистем, включая подсистемы такого же уровня.

Понятие «**система»** *(Storage)* соответствует объекту диагностирования (то есть СХД в целом) и предназначено для идентификации группы подсистем верхнего уровня.

Понятие «**состояние**» ***{D}*** предназначено для характеристики здоровья подсистемы, компонента СХД и СХД в целом и может иметь одно из четырех значений: «работоспособное состояние», «предотказное состояние», «уязвимое состояние», «полный отказ».

Понятие **«событие в СХД»** ***{F}*** представляет собой совокупность значений параметров СХД, состояний компонентов или подсистем СХД, характеризующих наступление какой-либо неисправности. Каждое событие при этом классифицируется как одно из состояний ***{D}*** в зависимости от уровня его критичности.

Для определения состояния D_{Ei} в элементе E_i определяется наличие событий в дочерних элементах, начиная с элементов самого низкого уровня вложенности – параметров и далее, вверх по иерархии элементов от компонентов к подсистемам.

Формально онтологическая модель описывается следующим образом:

Параметр $\boldsymbol{P_i}$ СХД:

$$P_i = \{Rp_j, V_j\}_{j=1,..,N} \quad (1)$$

где N – число связей i-ого параметра СХД, $< Rpj, Vj >$ - пара отношение/значение.

Компонент $\boldsymbol{K_i}$ СХД:

$$K_i = <\{P_{ij}\}_{j=1,..,N}; \{Fe_k, De_k, Rep\}_{k=1,..,M} > \quad (2)$$

где N – число параметров, а M – число возможных событий i-ого компонента СХД, $<Fe_k, De_k, Rep>$ - триплет событие/состояние/связь с параметрами, $\{P_{ij}\}$ – набор параметров СХД, относящихся к компоненту.

Подсистема СХД $\boldsymbol{S_i}$:

$$S_i = <\{K_{ij}\}_{j=1,..,N}; \{S_k\}_{k=1,..,M}; \{Fs_t, Ds_t, [Rs_t(K_{ij}), Rs_t(S_k)]\}_{t=1,..,L} > \quad (3)$$

где N – число компонентов, M – вложенных подсистем, а L – событий- i-ой подсистемы СХД, $<Fs, Ds, \{Rs_t, Rs_t\}>$ - триплет <событие/состояние/связь, определяющая компонент или подсистему – источник события>, $\{K_{ij}\}$ – набор компонентов системы.

Тогда система в целом:

$$Storage = <\{S_i\}_{j=1,..,N}; \{Fst_j, Dst_j, Rst_j\}_{j=1,..,M} > \quad (4)$$

где N – число параметров, M – число возможных событий, $<Fst, Dst, Rst>$ - триплет событие/состояние/связь, определяющая подсистему – источник события. Оценка текущего состояния СХД при этом определяется как:

- набор состояний компонентов СХД $\boldsymbol{\{S_c\}}$, подсистем СХД $\boldsymbol{\{S_s\}}$ и СХД в целом $\boldsymbol{\{S\}}$ (из списка: «работоспособное», «предотказное», «неисправность», «полный отказ»);
- набор обнаруживаемых событий в СХД $\boldsymbol{\{F\}}$, каждое из которых соответствует одному из состояний СХД.

Для описания отношений между объектом диагностирования, его элементами, состояниями и диагностическими параметрами применяются два типа связей $\boldsymbol{\{R\}}$ (таблица 1): связи, основанные на онтологических свойствах объектов и свойствах данных, далее именуемые **детерминированными связями,** и связи, для описания которых требуется использование внешней процедуры, далее именуемые **условными связями** (рисунок 1).

Таблица №1: Типы отношений в базе знаний

Имя отношения	Пояснение
manifests_in (if_fatal_manifests_in, if_warning_manifests_in)	Определяет, какому уровню работоспособности компонента соответствует неисправность
shows_in_parameter	Определяет параметры для идентификации неисправности
is_normal (is_normal_when, is_normal_below, is_normal_above)	Определяет диапазон нормальных значений диагностического параметра в простом событии
solves_with/described_by	Для задания ссылки на внешнее правило для события
depends_on (strongly_depends_on, majorly_depends_on и depends_with_ecc_on)	Определяет, каким образом состояние подсистемы СХД зависит от состояний элементов этой подсистемы
consists_of	Описание иерархии компонентов и подсистем конкретной конфигурации СХД в виде дерева
causes (if_failed_causes, if_warning_causes)	Указывает на названия системных явлений, соответствующих состояниям подсистем
interprets_as	Позволяет делать выводы о здоровье СХД как системы в целом

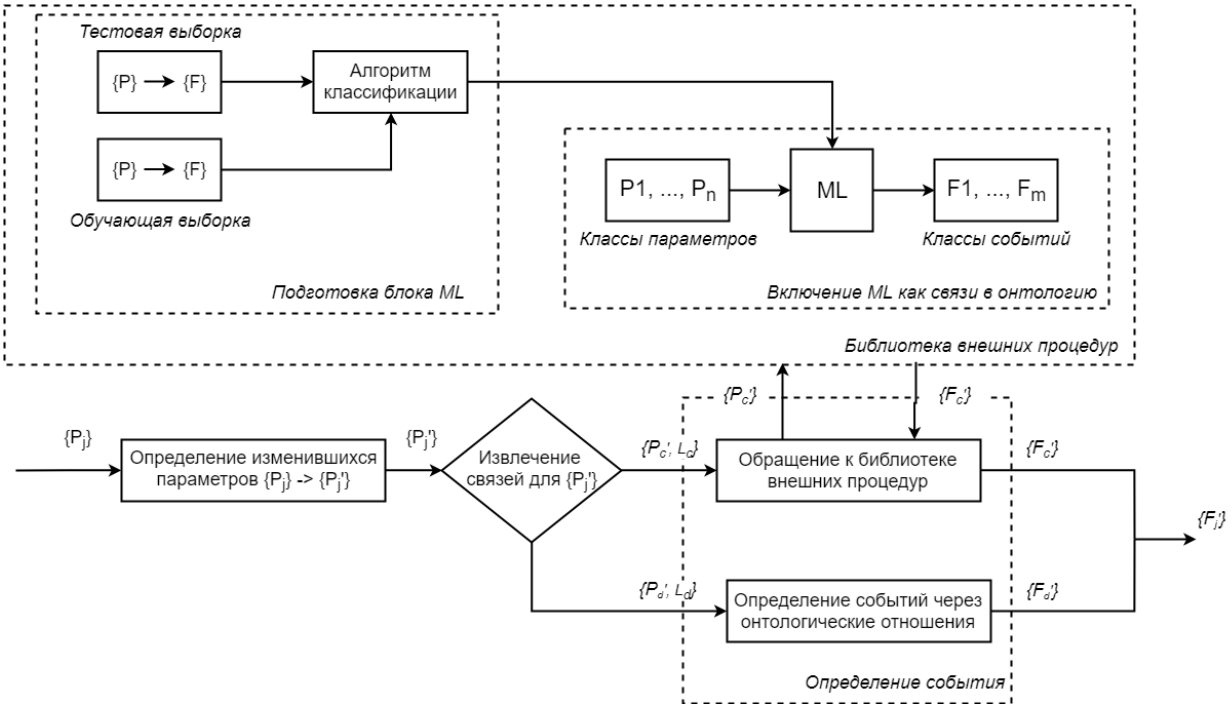

Рисунок 1 – Использование условных связей

В качестве внешней процедуры задается предварительно обученный алгоритм машинного обучения, принимающий на вход вектор значений диагностических параметров и классифицирующий их как одно из связанных понятий. Процедура вызывается в процессе диагностирования при определении события, связанного с указанными диагностическими параметрами.

После извлечения из полного перечня параметров $\{P_i\}$ тех параметров $\{P_i'\}$, значения которых изменились, определяются их возможные связи $\{L\}$ с событиями в компонентах. Далее поиск произошедших событий выполняется параллельно для условных связей и соответствующих им параметров ($\{P_c', Lc\}$) и детерминированных связей и соответствующих параметров ($\{P_d', Ld\}$). При этом параметры $\{P_c'\}$ передаются во внешние процедуры, определяемые в соответствии с $\{Lc\}$.

На рисунке 2 приведен фрагмент онтологии, показывающий, каким образом событие неисправности фиксируется на разных уровнях иерархии топологии СХД.

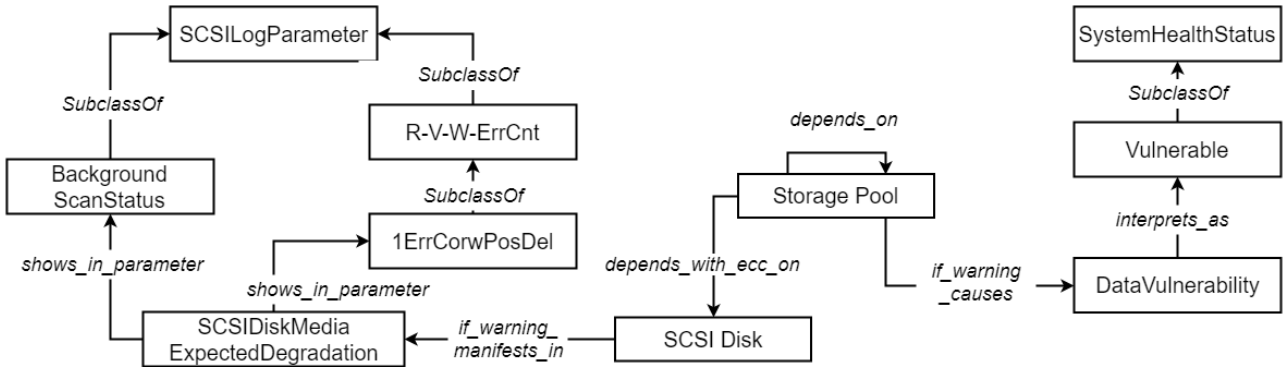

Рисунок 2 –Событие неисправности на разных уровнях иерархии СХД

Таким образом, описанная во второй главе базовая онтологическая модель включает в себя три основных раздела:

- раздел описания топологии СХД;
- раздел описания параметров СХД;
- раздел описания неисправностей.

В заключении второй главы выполнена оценка журналов ПО СХД как возможного источника диагностической информации для онтологической модели, в том числе сформулированы требования к допустимым с точки зрения применимости для разрабатываемых моделей, методов и алгоритмов типам журналов и форматам сообщений.

Третья глава посвящена разработке алгоритма динамического обнаружения неисправностей в СХД, предназначенного для применения в качестве онтологической условной связи и использующего в качестве входных данных блоки сообщений из журналов ПО СХД.

Разработанный алгоритм предполагает комбинированный подход к анализу журналов ПО СХД, рассматривающий каждое сообщение журнала как совокупность структурированного и неструктурированного текста. В процессе работы алгоритма сообщение разбивается на заголовок и текстовую часть в соответствии с набором правил, определяемых в рамках вспомогательного раздела онтологической модели. Для обнаружения возникновения неисправностей используется применение методов классификации текстов с использованием машинного обучения к неструктурированным текстам журналов ПО, составленным путем объединения текстовых частей сообщений в рамках заданного временного интервала.

Алгоритм обнаружения неисправностей состоит из следующих шагов:

1) Разбиение сообщений журналов на заголовочную и текстовую части.
2) Предварительная обработка текстовой части сообщений.
3) Формирование актуальных текстов журналов ПО в результате объединения текстов сообщений, содержащихся в заданном временном интервале.
4) Построение векторного представления текста и определение дополнительных численных параметров для применения методов машинного обучения.
5) Применение предварительно обученного алгоритма классификации к данным, полученным в п.4 для отнесения текстов журналов ПО СХД к заранее определенным классам неисправностей.
6) Локализация источников возникновения неисправностей с использованием онтологических условных связей.

Процедура разбиения сообщений на текстовую часть и заголовок базируется на методе комбинаторного парсинга, причем используется единственный вариант комбинатора, реализующий операцию bind, определяемую как:

```
bind :: Parser a -> (a -> Parser b) -> Parser b
p `bind` f = \inp -> concat [f v inp' | (v,inp') <- p inp]
```

В таком случае разбор строки S набором парсеров L_i, $i\epsilon[0,N]$ производится следующим образом: на вход L_i примитивного парсера подаётся строка, остаток строки подаётся на вход L_{i+1} примитивного парсера, и т.д., после выполнения L_N операция будет считаться завершенной.

Такой метод позволяет отделять заголовочную часть сообщений и использовать содержимое полей заголовка (например, временную метку) в диагностической процедуре.

Предварительная обработка текстовой части сообщения включает в себя следующие действия: токенизация текста, приведение его к единому регистру, удаление всех небуквенных токенов и стоп-слов, стемминг и фильтрация сообщений, где признаки фильтрации задаются для каждого журнала на уровне онтологической модели.

Процедура обнаружения неисправностей в обработанном тексте представляет собой применение алгоритма классификации к вектору X_u, представляющему собой объединение векторного представления текста журнала V_t, и вектора дополнительных признаков X_m. Для повышения эффективности построения векторного представления текста совместно применялись два типа коэффициентов:

- весовой коэффициент $W_{tf\text{-}idf}$, характеризующий важность слова в рамках контекста и представляющий собой отношение частоты появления слова в документе к частоте употребления слова во всех документах из выборки;
- весовой коэффициент W_{err}, характеризующий важность слова для идентификации неисправности, определяемый по положению слова во временном окне.

Таким образом, итоговый вектор V_w, описывающий каждое слово, имеет следующий вид (5):

$$V_w = W_{tf-idf} W_{err}[V]; \quad (5)$$

где *[V]* – векторное представление токена.

Соответственно, вектор V_t, описывающий текущее временное окно, должен быть определен как (6):

$$V_t = \frac{\sum_{i=1}^{n} W_{tf-idf} W_{err}[V_i]}{n}; \quad (6)$$

где n – общее число слов в выборке.

Дополнительные параметры Xm описывают количественные характеристики текста журнала:

- Количество токенов в i-ом журнале;
- Количество сообщений в i-ом журнале;
- Средняя длина сообщения в i-ом журнале;
- Наличие сообщений в i-ом журнале;
- Среднее количество токенов в секунду в i-ом журнале;
- Среднее количество сообщений/секунду в i-ом журнале.

Оценка важности дополнительных параметров выполнялась двумя способами. Первый способ заключается в применения алгоритма Random Forest и расчёта на каждом шаге критерия Джини:

$$G(k) = \sum_{i=0}^{J} P(i) * \big(1 - P(i)\big); \quad (7)$$

где *P(i)*-вероятность классификации i для признака *k.*

Альтернативный способ заключается в расчёте F-критерия на множестве признаков. F-критерий представляет собой отношение межгрупповой дисперсии к внутригрупповой:

$$F = \frac{Var_b}{Var_w} \quad (8)$$

где Var_b – межгрупповая, а Var_w – внутригрупповая дисперсии.

Проверка выполнялась на имеющихся размеченных данных, полученных по результатам тестирования ПО СХД. Общая выборка содержит 5904 пакетов журналов (что составляет порядка 350Гб файлов журналов), размещаемых на контроллерах хранения СХД, в том числе 1574 соответствующих возникавшим когда-либо в процессе эксплуатации СХД неисправностям, с указанием примерного временного интервала возникновения неисправностей. Всего в имеющихся данных был идентифицирован 41 тип различных неисправностей.

Выполненная оценка важности признаков позволила исключить признак «наличие сообщений в i-ом журнале», прочие показали свою важность для классификации.

В четвертой главе приводится описание комбинированного алгоритма, использующего онтологическую модель, и методы машинного обучения и его применения в составе встроенного диагностического ПО.

Диагностическая процедура включает в себя следующие основные шаги, повторяющиеся соответствии с заданным интервалом запуска процедуры обнаружения неисправностей:

1) Сбор данных мониторинга.

2)Для каждого компонента выполняется определение его детерминированного состояния путем подстановки значений параметров в правила определения симптомов (через условные и детерминированные связи)

3) Для каждого компонента выполняется определение состояния связанных подсистем.

4) Для СХД в целом определяется состояние как функция от состояний его подсистем.

На первом этапе процедуры формирования и применения онтологической модели в составе встроенного диагностического ПО целевой СХД (рисунок 3) выполняется подготовка онтологической модели путём наполнения её экспертными данными в редакторе онтологий Stanford Protégé. Для практического применения онтологической модели в составе ПО предложено приведение её к упрощенному графовому виду, предполагающему преобразование традиционного rdf к формату rdf-nquad, и описание модели в формате [узел]-<связь> [узел] [контекст]. Такое преобразование позволяет устранить громоздкие конструкции наследования, применяемые в формате построения онтологии OWL, и свести их к упрощенному виду.

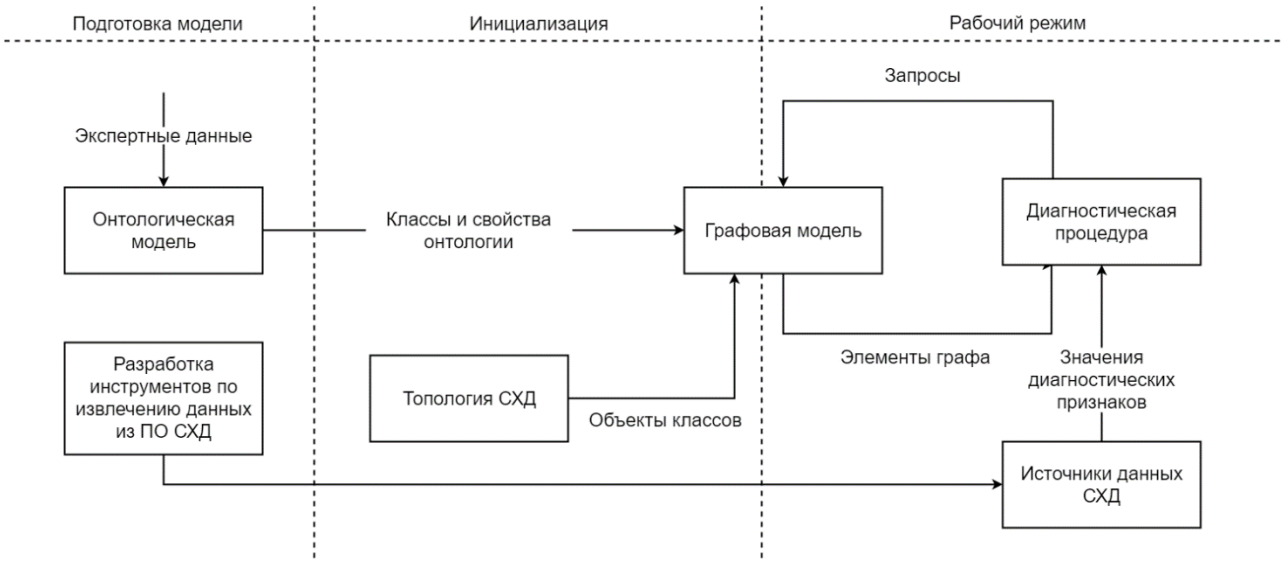

Рисунок 3 – Схема применения моделей

Преобразование к графовому виду выполняется на этапе инициализации, причём структура модели, то есть её классы, объектные свойства и свойства данных, получаются путём преобразования онтологической модели, а экземпляры – путём обращения к встроенному ПО СХД. Сформированная таким образом модель системы помещается в графовую базу данных и используется в рамках диагностической процедуры для получения сведений о взаимосвязи значений диагностических признаков с неисправностями в СХД.

Архитектура реализующего диагностическую процедуру ПО предполагает получение текущего значения параметров СХД, в том числе текстов временных окон журналов ПО СХД из базы данных, заполняемой модулем мониторинга и сбора данных. Экземпляр ПО запускается средствами кластера на одном из контроллеров хранения, входящих в кластер, и определяет состояние системы в целом и её отдельных подсистем, и компонентов.

В пятой главе выполнена экспериментальная проверка работы диагностического ПО на массиве текстов журналов со встроенной моделью и обученными алгоритмами классификации текста. Целью проведения эксперимента являлись:

1) Выбор элементов процедуры классификации, показывающих наилучшие результаты классификации, в том числе собственно алгоритма-классификатора (рассмотрены классические алгоритмы Random Forest, наивный байесовский классификатор, KNN, логистическая регрессия, метод опорных векторов и нейронные сети глубокого обучения LSTM, RNN и LSTM с механизмом внимания), способа векторного представления текста и способа объединения текстов в журналах.

2) Оценка эффективности всего подхода к обнаружению неисправностей в СХД.

Рассматриваемый набор алгоритмов классификации выбран на основании анализа результатов их применения к решению задачи многоклассовой классификации текста, представленных в актуальных научных публикациях (см. Ковасари К. и др., Ю С. и др.) и прикладных исследованиях (Ли С.). В случае, когда применение алгоритма предполагает предварительную настройку гиперпараметров, она выполнялась путем применения комбинации методов

случайного поиска и полного перебора в окрестности предполагаемых оптимальных значений.

Результаты экспериментов показали, что наиболее эффективным является раздельное построение векторного пространства для каждого журнала и объединение их вместе с векторами дополнительных признаков, определенных для каждого журнала, в единое признаковое пространство. Результаты сравнения качества классификации представлены в таблице 2, наибольшую точность при решении задачи классификации текстов журналов ПО СХД показал алгоритм Random Forest.

Таблица 2 – Сравнительная оценка показателей качества классификации

Классификатор	Средняя точность	Средняя полнота	Средняя f-мера	Среднее время обучения модели
Random Forest	0,74	0,71	0,71	183 с
Наивный Байес	0,48	0,27	0,28	205 с
KNN	0,38	0,39	0,37	166 с
Логистическая регрессия	0,28	0,33	0,28	498 с
Метод опорных векторов	0,23	0,26	0,20	430 с
LSTM	0,47	0,44	0,45	~4 часа
GRU	0,24	0,22	0,23	~7 часов
LSTM с механизмом внимания	0,42	0,39	0,40	~ 6 часов

Эксперимент проводился с использованием программных средств имитации неисправностей, так как частота их возникновения в процессе штатного функционирования СХД слишком мала, для того чтобы можно было сделать выводы об эффективности разработанного подхода. По результатам эксперимента получены следующие результаты:

- для неисправностей, определяемых путем применения правил обнаружения, заданных детерминированными онтологическими связями, доля идентифицированных составляет 0,99 на массиве данных в 10000 неисправностей;
- для неисправностей, определяемых путем применения правил, заданных условными онтологическими связями, доля идентифицированных составляет 0,71 на массиве данных в 10000 неисправностей.

ЗАКЛЮЧЕНИЕ

В диссертационной работе решена актуальная научно-техническая задача по расширению множества обнаруживаемых неисправностей в СХД на основе применения предложенного метода диагностирования, базирующегося на совместном использовании онтологической диагностической модели и методов машинного обучения для анализа диагностических параметров.

Получены следующие результаты:

1) Выполнен анализ систем класса СХД как объекта диагностики, определены требования к методам обнаружения неисправностей в СХД и реализующему их ПО. Выполнен анализ существующих научных работ и прикладных инструментов в области диагностики вычислительных систем, сравнительный анализ их характеристик.

2) Разработан метод построения диагностической модели для систем класса СХД, определяющей отношения между параметрами и возможными состояниями СХД и ее элементов, где отношения могут быть заданы и как детерминированные связи между диагностическими сущностями, и как функции машинного обучения.

3) Разработаны методы и программные средства преобразования диагностической модели к упрощенному графовому виду для её применения в составе диагностического ПО.

4) Разработаны методы и инструментальные средства обнаружения неисправностей в СХД, основанные на анализе текстовой информации, получаемой в процессе мониторинга ПО методами классификации текста с использованием машинного обучения.

5) Выполнена экспериментальная проверка разработанных моделей, методов и средств на целевой платформе СХД.

В качестве **рекомендаций и перспектив дальнейшей разработки темы** можно указать исследование моделей векторного представления текстов журналов ПО СХД, основанных на применении нейронных сетей, и определение возможности совместного использования разных алгоритмов для решения задачи классификации.

Полученные результаты соответствуют пунктам 4 «Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации», 5 «Разработка специального математического и алгоритмического обеспечения систем анализа, оптимизации, управления, принятия решений и обработки информации», 11 «Методы и алгоритмы прогнозирования и оценки эффективности, качества и надежности сложных систем.», 12 «Методы получения, анализа и обработки экспертной информации» паспорта специальности 05.13.01 Системный анализ, управление и обработка информации (по отраслям).

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

В журналах, рекомендованных ВАК:

1. Успенский М.Б. Обзор подходов к обнаружению неисправностей в системах хранения данных / Успенский М.Б. // Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление – 2019. - №4 – С.145-158.

2. Успенский М.Б. Применение онтологической модели и алгоритмов классификации текста в задачах обнаружения сбоев систем хранения данных / Успенский М.Б. // Известия Самарского научного центра Российской академии наук – 2020. - №1 - С.107-113.

3. Успенский М.Б. Автоматическое обнаружение сбоев в системах хранения данных с использованием журналов системного программного обеспечения /Успенский М.Б. // Информация и космос – 2020. - №1 – С. 90-96.

В изданиях, индексированных в Web of Science и Scopus

4. Uspenskij, M. B. (2019). Log mining and knowledge-based models in data storage systems diagnostics. //E3S Web of Conferences, Vol.140, №03006.

5. Shirokova S.V., Bolsunovskaiya M.V., Loginova A.V., Uspenskiy M.B. Developing a procedure for conducting a security audit of a software package for predicting storage system failures // MATEC Web Conf, 2018. International Scientific Conference on Energy, Environmental and Construction Engineering (EECE-2018). Volume 245, № 10007.

6. Uspenskij M., Makarov A., Sochnev A., Shirokova S., Petrov V. Development of a software structure for monitoring the working capacity of the data storage system for predicting failures and preventing critical situations // Proceedings of the 33rd International Business Information Management Association Conference, IBIMA 2019: Education Excellence and Innovation Management through Vision 2020, 2019. Pp. 8508-8514.

7. Mamoutova O.V., Shirokova S.V., Uspenskij M.B., Loginova A.V. The ontology-based approach to data storage systems technical diagnostics // E3S Web of Conferences, Vol. 91, № 08018.

8. Mamoutova O.V, Uspenskiy M.B., Sochnev A.V., Smirnov S.V. Bolsunovskaya M.V. Knowledge Based Diagnostic Approach for Enterprise Storage Systems // Proceeding of IEEE 17th International Symposium on Intelligent Systems and Informatics, 2019

В прочих изданиях

9. Макаров А.С., Болсуновская М.В., Широкова С.В., Успенский М.Б., Кузьмичев А.А. Анализ подходов к диагностике систем хранения данных // Мягкие вычисления и измерения: сборник трудов XXI Международной конференции, 2018. Т. 2. С. 61-64

10. Приданова Е.В., Успенский М.Б., Ицыксон В.М. Мониторинг и анализ параметров системы хранения данных для оценки ее состояния // Сборник научных трудов XXIII Международной научно-практической конференции “Системный анализ в проектировании и управлении”, 2019. С. 170-177.

11. Иванов О.И., Михайлов Е.А., Пустоветов В.И., Успенский М.Б. Подсистема подготовки тестовых проектов для контрольно-диагностических комплексов // Вопросы радиоэлектроники. 2013. Т. 1. № 1. С. 99-105.

12. Берлик С.А., Иванов О.И., Успенский М.Б., Пустоветов В.И. Архитектура графической среды аппаратно-программного комплекса КДК // Вопросы радиоэлектроники. 2013. Т. 1. № 1. С. 73-80.

13. Михайлов А.Н., Иванов О.И., Успенский М.Б. Сигнатурный анализатор для многофункционального аппаратно-программного комплекса КДК // Вопросы радиоэлектроники. 2014. Т. 1. № 2. С. 106-112.

Свидетельства о государственной регистрации программ для ЭВМ

1. Успенский М.Б. Программа для сбора параметров системы хранения данных / М.Б. Успенский, В.Д. Петров, А.В. Сочнев, В.И Пустоветов. – Свидетельство о государственной регистрации программы для ЭВМ № 2018660284 от 21.08.2018.
2. Успенский М.Б. Программа имитации функционирования аппаратной компоненты системы хранения данных - носителя информации / М.Б. Успенский, М.Е. Карпов. – Свидетельство о государственной регистрации программы для ЭВМ № 2018665078 от 30.11.2018.
3. Успенский М.Б. Программа имитации функционирования аппаратной компоненты системы хранения данных - контроллера фабрики PCI-Express / М.Б. Успенский, К. Арзыматов. – Свидетельство о государственной регистрации программы для ЭВМ № 2018665160 от 03.12.2018.
4. Успенский М.Б. Программа имитации функционирования аппаратной компоненты системы хранения данных - контроллера хранения данных / М.Б. Успенский, В.С. Белавин. – Свидетельство о государственной регистрации программы для ЭВМ № 2018665676 от 06.12.2018.
5. Успенский М.Б. Программа для сбора и отображения климатических параметров систем хранения данных / М.Б. Успенский, С.В. Смирнов. – Свидетельство о государственной регистрации программы для ЭВМ № 2019614476 от 05.04.2019.
6. Успенский М.Б. Программа для диагностирования системы хранения данных / М.Б. Успенский, М.И. Гущин. – Свидетельство о государственной регистрации программы для ЭВМ № 2019618328 от 27.06.2019.
7. Успенский М.Б. Программа настройки параметров имитационной модели функционирования системы хранения данных / М.Б. Успенский, В.С. Белавин. – Свидетельство о государственной регистрации программы для ЭВМ № 2019618010 от 25.06.2019.
8. Иванов О.И. Программа тестирования и диагностики цифровой радиоэлектронной аппаратуры с использованием многоканального сигнатурного анализа / О.И. Иванов, В.И. Пустоветов, М.Б. Успенский. – Свидетельство о государственной регистрации программы для ЭВМ № 2015661325 от 23.10.2015.
9. Иванов О.И. Программа подготовки комбинированных тестовых проектов / О.И. Иванов, В.И. Пустоветов, М.Б. Успенский. – Свидетельство о государственной регистрации программы для ЭВМ № 2015661326 от 23.10.2015.