На правах рукописи

Монгуш Чодураа Михайловна

РАЗРАБОТКА МЕТОДА И СРЕДСТВ ФРАГМЕНТАЦИИ И ДЕФРАГМЕНТАЦИИ ФОРМАЛЬНЫХ КОНТЕКСТОВ

Специальность 05.13.17 — Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Красноярск 2019

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Сибирский федеральный университет», г. Красноярск

Научный руководитель: кандидат физико-математических наук, доцент **Семенова Дарья Владиславовна**

Официальные оппоненты: **Фархадов Маис Паша Оглы**, доктор технических наук, старший научный сотрудник, Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В.А. Трапезникова Российской академии наук, лаборатория автоматизированных систем массового обслуживания и обработки сигналов, заведующий лабораторией

Богаченко Надежда Федоровна, кандидат физико-математических наук, доцент, Федеральное государственное бюджетное образовательное учреждение высшего образования «Омский государственный университет им. Ф.М. Достоевского», кафедра компьютерных технологий и сетей, доцент

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Томский государственный университет»

Защита состоится «21» января 2020 года в 14.00 ч. на заседании диссертационного совета Д 212.099.22, созданного на базе Сибирского федерального университета по адресу: 660074, г. Красноярск, ул. Киренского, 26, ауд. УЛК 112.

С диссертацией можно ознакомиться в библиотеке и на сайте Сибирского федерального университета по адресу http://www.sfu-kras.ru.

Автореферат разослан «___» ноября 2019 г.

Ученый секретарь
диссертационного совета Покидышева Людмила Ивановна

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Во многих задачах интеллектуального анализа данных, в том числе в анализе естественно-языковых текстов, изучаемая предметная область часто описывается в виде объектно-признаковой таблицы, в которой каждый столбец соответствует некоторому признаку, а каждая строка определяет признаковое описание отдельного объекта.

Появление размеченных корпусов естественных языков как элементов информационных систем позволяет получать структурированную информацию и представлять ее в виде объектно-признаковой таблицы. Разметка — главная характеристика корпуса, отличающая корпус от простых электронных коллекций текстов. Она отражает лингвистическую и экстралингвистическую информацию хранимых текстов в корпусе. Чем больше набор признаков, характеризующий каждый текст, тем шире возможности корпуса по поиску текстов для решения различных филологических и лингвистических задач. В рамках корпусов решаются различные прикладные задачи, учитывающие семантическую составляющую анализируемых текстов. Эти задачи в основном сводятся к задачам концептуального моделирования коллекции текстов.

Существует формализованный подход, известный в литературе как анализ формальных понятий (АФП, англ. Formal Concept Analysis), который позволяет построить концептуальную модель предметной области исходя из объектно-признаковой таблицы на основе алгебраической теории решеток Г. Биркгофа. В рамках АФП объектно-признаковая таблица представляется формальным контекстом, отражающим наличие или отсутствие признаков, характерных для исследуемого множества текстов, и моделируется 0,1-матрицей. Основные идеи АФП были сформулированы в начале 80-х годов XX века в работах Р. Вилле и Б. Гантера и развиты в исследованиях российских ученых С. О. Кузнецова, К. А. Найденовой, С. А. Объедкова, Д. И. Игнатова, С. И. Гурова. В АФП каждое формальное понятие определяется с использованием соответствий Галуа и представляет собой пары замкнутых множеств, интерпретируемых как объем и содержание этого понятия. В матричной форме формальному понятию соответствует некоторая максимально полная подматрица 0,1-матрицы, представляющей формальный контекст. С применением методов АФП решаются типовые задачи анализа данных, связанные с классификацией и кластеризацией данных, выявлением зависимостей между данными. В них формальные понятия трактуются как перекрестные ассоциации, кластеры или бикластеры. В рамках АФП решение указанных задач сводится к нахождению всех формальных понятий исходного формального контекста с последующим связыванием их в решетку. Полученная решетка служит концептуальной моделью исследуемой предметной области и основой для решения многих прикладных задач, сводящихся к извлечению знаний из полученной решетки.

При всей привлекательности методов АФП их практическое применение ограничивается высокой трудоемкостью процесса извлечения всех формальных понятий из исходного контекста большой размерности. В задаче нахождения всех формальных понятий требуется найти множество всех формальных понятий для заданного формального контекста. Данная задача относится к комбинаторным перечислительным задачам и является $\#P$-полной. Высокая вычислительная сложность задачи объясняется тем, что в общем случае число формальных понятий экспоненциально зависит от размера исходного формального контекста. Рассматриваемая задача эквивалентна задаче определения всех максимально полных подматриц 0,1-матрицы и может встречаться в различных задачах комбинаторной оптимизации.

Степень разработанности темы исследования. На сегодняшний день для нахождения множества всех формальных понятий и построения решетки известно много алгоритмов и программных средств. Традиционно данные алгоритмы разделяют на две группы: пакетные алгоритмы (Bordat, NextClosure, Close-by-One, Lindig), которые строят решетку из ранее найденных формальных понятий; инкрементные алгоритмы (Nourine, Dowling, Norris), которые достраивают решетку посредством постепенного добавления объектов и пересечения с имеющимися формальными понятиями. Известно, что время выполнения указанных алгоритмов в худшем случае составляет $O(|FC| \cdot |G|^2 \cdot |M|)$, где $|FC|$ — число найденных формальных понятий, $|G|$ — количество объектов, $|M|$ — количество признаков исходного формального контекста. Поскольку величина $|FC|$ может экспоненциально зависеть от $|G|$ и $|M|$, то время выполнения данных алгоритмов также может быть экспоненциальным. Наиболее известными программными системами являются Concept Explorer, ToscanaJ, Galicia, Lattice Minner, OpenFCA, FCART. Многие из них находятся в открытом доступе. Все эти программные средства являются специализированными продуктами, т. е. направлены на решение конкретной задачи анализа данных. Однако, вычислительная сложность задачи нахождения всех формальных понятий формального контекста большой размерности и связывание их в решетку остается открытой проблемой.

В настоящее время актуальны исследования по снижению вычислительной сложности задачи нахождения всех формальных понятий. Первое направление исследований связано с разработкой новых алгоритмов отбора информативных, релевантных формальных понятий при построении решетки. Оно позволяет уменьшить выход рассматриваемой задачи. Второе направление исследований рассматривает уменьшение размерности входа, а значит, повышение производительности существующих алгоритмов нахождения множества всех формальных понятий и родственных с ней задач, путем «неискажающего» разложения формального контекста — декомпозиции исходного контекста

с сохранением всех искомых формальных понятий. Такое направление исследований является более универсальным и рассматривается в настоящей диссертационной работе.

Цель и задачи исследования. Целью диссертационной работы является повышение производительности существующих алгоритмов решения задачи нахождения всех формальных понятий путем декомпозиции формального контекста на фрагменты без потери искомых формальных понятий и разработка на их основе математического и программного обеспечения.

Для достижения цели были поставлены и решены следующие задачи.

1. Разработать и теоретически обосновать метод «неискажающего» разложения формального контекста на фрагменты. Исследовать структуру фрагментов и найти оценку числа фрагментов, получаемых на каждой итерации разложения, определить правила остановки процесса разложения формального контекста на фрагменты без потери формальных понятий.

2. Разработать алгоритмы формирования для заданного формального контекста системы фрагментов, восстановления искомого решения исходя из решений, полученных для подзадач, и реализации возможных запросов на извлечение знаний из решетки формальных понятий.

3. Разработать алгоритмы предобработки формального контекста без потери формальных понятий путем удаления единичных, нулевых и кратных строк и столбцов этого контекста.

4. Создать комплекс программ, реализующий разработанные метод и алгоритмы, для проверки их результативности на случайных формальных контекстах и на реальных данных применительно к корпусу тувинского языка.

Научная новизна результатов, представленных в диссертации.

1. Разработан новый метод декомпозиции формального контекста на фрагменты без потери формальных понятий. В отличие от существующих методов АФП, предложенный метод позволяет уменьшить размерность формального контекста с сохранением всех искомых формальных понятий и тем самым повысить производительность известных алгоритмов нахождения всех формальных понятий формального контекста.

2. Впервые разработан алгоритм реализации предложенного метода «неискажающей» декомпозиции формального контекста. Алгоритм отличается от ранее существующих алгоритмов тем, что разлагает исходный формальный контекст без потери формальных понятий и восстанавливает решение поставленной задачи исходя из решений, полученных для подзадач.

Методы исследования. Для решения поставленных в работе задач использовались современные методы АФП, теории графов и методы объектно-ориентированного программирования.

Теоретическая значимость работы. Предложенный в работе метод «неискажающей» декомпозиции формального контекста может быть использован для развития АФП и комбинаторной оптимизации при решении задач определения всех максимально полных подматриц 0,1-матрицы.

Практическая значимость работы. Применение результатов диссертационной работы при исследовании объектно-признаковых описаний предметных областей позволяет на семантическом уровне решать различные задачи анализа данных, включая классификацию, кластеризацию, обнаружение закономерностей в данных и извлечение знаний из решетки формальных понятий. Исследование естественно-языковых текстов тувинского фольклора в научно-образовательном центре «Тюркология» Тувинского государственного университета с применением предложенных метода и алгоритмов позволяет эффективно решать филологические и лингвистические задачи в рамках корпуса тувинского языка.

Соответствие паспорту специальности. Диссертационная работа соответствует области исследования специальности 05.13.17 — Теоретические основы информатики по п. 5 «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений» (пункты 1, 2 научной новизны).

Положения и результаты, выносимые на защиту.

1. Доказательство корректности метода декомпозиции формального контекста на фрагменты без потери формальных понятий, позволяющего уменьшить размерность формального контекста и тем самым повысить производительность известных алгоритмов нахождения всех формальных понятий формального контекста.

2. Алгоритм реализации предложенного метода «неискажающей» декомпозиции формального контекста и оценки его сложности, а также рекомендации по практическому применению этого алгоритма при анализе данных.

3. Комплекс программ для проверки результативности предложенных метода и алгоритмов на случайных формальных контекстах и на реальных данных применительно к корпусу тувинского языка.

Степень достоверности и апробация результатов работы. Достоверность результатов работы подтверждается строгими математическими доказательствами основных положений, экспериментальной проверкой результатов, численных расчетов на реальных текстовых данных и практической эффективностью программных реализаций.

Основные результаты работы докладывались и обсуждались на III Международной научно-практической конференции молодых ученых, аспирантов и студентов «Актуальные проблемы исследования этноэкологических и этно-

культурных традиций народов Саяно-Алтая» (Кызыл, 2015), Международной конференции студентов, аспирантов и молодых ученых «Молодежь и наука: Проспект Свободный — 2016» (Красноярск, 2016), Международной конференции «Актуальные проблемы прикладной математики и информационных технологий — Аль-Хорезми 2016» (Ташкент, 2016), Международной конференции имени А. Ф. Терпугова «Информационные технологии и математическое моделирование» (Томск, 2016, 2018), VI Международной конференции «Математика, ее приложения и математическое образование» (Улан-Удэ, 2017), Всероссийской научно-практической конференции преподавателей, сотрудников и аспирантов Тувинского государственного университета (Кызыл, 2017), Всероссийской конференции «Компьютерная безопасность и криптография» — SIBECRYPT'19 (Томск, 2019), VII Международной конференции «Знания — Онтологии — Теории» (Новосибирск, 2019), научных семинарах кафедры высшей и прикладной математики Сибирского федерального университета и кафедры информатики и ИКТ Тувинского государственного университета.

Результаты диссертационного исследования переданы в научно-образовательный центр «Тюркология» для использования в научных исследованиях и на кафедру информатики и ИКТ Тувинского государственного университета для внедрения в учебный процесс при подготовке бакалавров по направлению «Фундаментальная информатика и информационные технологии», а также успешно применены для выполнения гранта РФФИ (РГНФ) № 16-34-1-01033 в 2016–2017 гг. Получены свидетельства о государственной регистрации программ для ЭВМ № 2018618907, № 2018615490 от 23.07.2018.

Личное участие автора в получении результатов, изложенных в диссертации. Основные результаты, составляющие новизну диссертационной работы, получены лично автором. Обсуждение метода, алгоритмов, результатов численных экспериментов и подготовка публикаций осуществлялись совместно с научным руководителем и соавторами опубликованных работ.

Публикации. По результатам диссертационных исследований опубликовано 14 печатных работ, из них 4 — в журналах, рекомендованных ВАК [1–4], 8 — в материалах конференций и других изданиях [5–12], получены 2 свидетельства о государственной регистрации программы для ЭВМ [13, 14].

Структура и объем диссертации. Диссертация состоит из введения, трех глав, заключения, списка литературы. Общий объем диссертации составляет 105 страниц; иллюстративный материал представлен 25 рисунками и 20 таблицами; список литературы содержит 127 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы диссертационной работы, определены цель и задачи исследования, указаны научная новизна, практическая и теоретическая значимость выполненных исследований.

В **первой главе** рассматриваются современные модели представления естественно-языковых текстов и их коллекций, приводятся основные положения АФП, используемые в диссертационной работе. Формулируется задача нахождения всех формальных понятий заданного формального контекста.

Приведем основные положения первой главы диссертационной работы. Пусть определены два непустых конечных множества: множество объектов G и множество признаков или свойств M. Пусть также задано непустое отношение инцидентности $I \subseteq G \times M$. Данное отношение содержит информацию о выполнимости свойств из M на объектах из G, т. е. $(g, m) \in I$ означает, что объект g обладает признаком m и наоборот, признак m присущ объекту g. Тройка $K = (G, M, I)$ называется формальным контекстом (или просто контекстом) предметной области. Полагаем, что множества G и M линейно упорядочены (например, лексикографически). В этом случае $K = (G, M, I)$ однозначно задается 0,1-матрицей $T = (t_{ij}) : t_{ij} = 0$ при $(g_i, m_j) \notin I$; $t_{ij} = 1$ при $(g_i, m_j) \in I$ ($i = 1, 2, \ldots, |G|$; $j = 1, 2, \ldots, |M|$). Если $A \subseteq G$ и $B \subseteq M$, то пара операторов, называемых отображениями Галуа,

$$A' = \{m \in M : \forall\, g \in A\ (g, m) \in I\},$$
$$B' = \{g \in G : \forall\, m \in B\ (g, m) \in I\},$$

задает соответствие между частично упорядоченными множествами $(2^G, \subseteq)$ и $(2^M, \subseteq)$. Двойное применение $(\cdot)'$ определяет оператор замыкания $(\cdot)''$ на 2^M в алгебраическом смысле. Если $B = B''$, то B называется замкнутым множеством. Для отображений $(\cdot)'$ характерны свойства антимонотонности и экстенсивности: для любых $B_1, B_2 \subseteq M$, если $B_1 \subseteq B_2$, то $(B_2)' \subseteq (B_1)'$ и $(B_1)'' = ((B_1)')' \subseteq M$. Кроме того, $(B_1 \cup B_2)' = B_1' \cap B_2'$.

Пара множеств (A, B), $A \subseteq G$, $B \subseteq M$, таких, что $A' = B$ и $B' = A$, называется формальным понятием контекста $K = (G, M, I)$ с объемом A и содержанием B. Всякое формальное понятие уникально в заданном контексте, т. е. отличается от других формальных понятий объемом и/или содержанием. Если формальный контекст представлен 0,1-матрицей T, то при $A \neq \varnothing$ и $B \neq \varnothing$ формальному понятию (A, B) отвечает максимальная полная подматрица матрицы T. Строки этой подматрицы соответствуют элементам из A, а столбцы — элементам из B.

Обозначим через FC множество всех формальных понятий формального контекста $K = (G, M, I)$. Пусть $(A_1, B_1), (A_2, B_2) \in FC$. Множество FC частично упорядочено отношением $(A_1, B_1) \sqsubseteq (A_2, B_2)$ тогда и только тогда, когда $A_1 \subseteq A_2$. Отметим, что последнее эквивалентно условию $B_2 \subseteq B_1$. Каждое формальное понятие $(A, B) \in FC$ определяет для исследуемой предметной области совокупность однородных объектов A со своим специфичным набором признаков B. Если $(G, \emptyset) \in FC$, $(\emptyset, M) \in FC$, то формальные понятия $(G, \emptyset)$, $(\emptyset, M)$ называются тривиальными.

Определим на FC операции пересечения $\sqcap$ и объединения $\sqcup$ через одноименные теоретико-множественные операции $\cap$ и $\cup$ следующим образом:

$$(A_1, B_1) \sqcap (A_2, B_2) = (A_1 \cap A_2, (A_1 \cap A_2)'),$$

$$(A_1, B_1) \sqcup (A_2, B_2) = ((B_1 \cap B_2)', B_1 \cap B_2).$$

Тогда $(FC, \sqsubseteq)$ образует полную решетку $L = (FC, \sqcap, \sqcup)$, называемую в АФП решеткой формальных понятий контекста $K = (G, M, I)$.

В АФП задача нахождения всех формальных понятий формулируется следующим образом.

Задан формальный контекст $K = (G, M, I)$.

Требуется найти для $K = (G, M, I)$ множество FC.

Как было доказано С.О. Кузнецовым, данная задача относится к комбинаторным перечислительным задачам и является $\#P$-полной. Высокая вычислительная сложность задачи объясняется тем, что в общем случае число формальных понятий экспоненциально зависит от размера исходного контекста. В настоящее время актуальны исследования по снижению вычислительной сложности данной задачи.

Во **второй главе** содержатся основные результаты диссертационного исследования. В ней решаются задачи 1, 2 и 3 диссертационного исследования. Данные результаты опубликованы в работах [1–4, 7, 11, 12].

Пусть заданы $K = (G, M, I)$ — контекст, FC — множество всех его формальных понятий и T — соответствующая ему 0,1-матрица.

Определение 2.1. Контекст $K_1 = (G_1, M_1, I_1)$ назовем частью контекста $K = (G, M, I)$, если $G_1 \subseteq G$, $M_1 \subseteq M$ и для любых $x \in G_1$, $y \in M_1$ отношение $(x, y) \in I_1$ верно тогда и только тогда, когда $(x, y) \in I$.

Части $K_1 = (G_1, M_1, I_1)$ и $K_2 = (G_2, M_2, I_2)$ контекста $K = (G, M, I)$ будем считать различными, если $G_1 \neq G_2$ и/или $M_1 \neq M_2$.

Определение 2.2. Разложение контекста $K = (G, M, I)$ на конечное множество различных частей назовем «неискажающим» относительно формальных понятий, если оно удовлетворяет следующим условиям: каждая часть содержит, по крайней мере, одно формальное понятие из FC; ни одно формальное понятие из FC не теряется и не возникают новые формальные понятия.

«Неискажающее» (или «безопасное») разложение исходного формального контекста базируется на доказанных предложениях 2.1–2.6 и теореме 2.1 настоящей диссертационной работы, отражающих свойства объектных и признаковых понятий, а также структуру фрагментов этого контекста.

Пусть $g \in G$, $m \in M$ — произвольные элементы контекста $K = (G, M, I)$.

Определение 2.3. Пары множеств (g'', g') и (m', m'') образуют формальные понятия, первое из которых назовем объектным, а второе — признаковым формальным понятием контекста $K = (G, M, I)$.

Обозначим через $O = \{(g'', g') : \forall g \in G\} \subseteq FC$ множество всех объектных формальных понятий и через $S = \{(m', m'') : \forall m \in M\} \subseteq FC$ множество всех признаковых формальных понятий.

Предложение 2.1. *Всякое объектное формальное понятие (g'', g') контекста $K = (G, M, I)$ имеет самое большое по размеру содержание среди других формальных понятий, имеющих в объеме объект $g \in G$, а признаковое формальное понятие (m', m'') обладает самым большим объемом среди других формальных понятий, имеющих в содержании признак $m \in M$.*

Определение 2.4. Пара формальных понятий $(g'', g') \in O$, $(m', m'') \in S$ определяет фрагмент $\omega = (m', g', J)$ как часть контекста $K = (G, M, I)$, если

$$(g'', g') \sqsubseteq (m', m'')\,, \tag{1}$$

что эквивалентно $g'' \subseteq m'$ (или $m'' \subseteq g'$).

Про такой фрагмент будем говорить, что он образован элементами $g \in G$ и $m \in M$. Далее вместо $\omega = (m', g', J)$ будем кратко писать $\omega = (m', g')$ или (m', g').

Предложение 2.2. *Для всякого формального контекста $K = (G, M, I)$ и любых $(g'', g') \in O$, $(m', m'') \in S$ отношение порядка $(g'', g') \sqsubseteq (m', m'')$ выполняется тогда и только тогда, когда $(g, m) \in I$.*

Из предложения 2.2 следует, что число различных фрагментов, порождаемых всевозможными элементами формального контекста $K = (G, M, I)$, не превышает веса 0,1-матрицы T, т. е. величины $\|T\|$ — числа единичных элементов этой матрицы. Очевидно, что $1 \leq \|T\| \leq |G| \cdot |M|$.

Определение 2.5. Будем говорить, что формальное понятие $(A, B) \in FC$ вложено в фрагмент (m', g') формального контекста $K = (G, M, I)$, и писать $(A, B) \preceq (m', g')$, если $A \subseteq m'$, $B \subseteq g'$.

Всякий фрагмент (m', g') не является пустым, поскольку согласно (1) он всегда содержит формальные понятия $(g'', g') \in O$ и $(m', m'') \in S$.

Предложение 2.3. *Всякое нетривиальное формальное понятие (A, B) контекста $K = (G, M, I)$, которое вложено в фрагмент (m', g'), образованный элементами $g \in G$ и $m \in M$, всегда содержит эти элементы и их замыкания, т. е. если $(A, B) \preceq (m', g')$, то неизменно: $g \in A$ и $m \in B$; $g'' \subseteq A$ и $m'' \subseteq B$.*

Согласно предложению 2.3, пару (g'', m'') можно рассматривать в качестве типичного представителя не только фрагмента (m', g'), но и всех формальных понятий контекста $K = (G, M, I)$, вложенных в этот фрагмент. Это правомерно, поскольку подматрица, соответствующая фрагменту (m', g'), во всех строках из g'' и всех столбцах из m'' постоянно имеет единичные элементы. Соответствие между фрагментами и формальными понятиями контекста устанавливает следующая теорема.

Теорема 2.1. *Для всякого формального контекста $K = (G, M, I)$, множества FC всех его формальных понятий и любой пары множеств (A, B), $\varnothing \neq A \subseteq G, \varnothing \neq B \subseteq M$, справедливы высказывания:*

1) *если $(A, B) \in FC$, то всегда в контексте $K = (G, M, I)$ существует фрагмент $\omega = (m', g')$, $g \in G$ и $m \in M$, причем возможно не единственный, в который это формальное понятие вложено;*
2) *если (A, B) — формальное понятие некоторого фрагмента $\omega = (m', g')$ контекста $K = (G, M, I)$, то оно также принадлежит FC.*

Согласно теореме 2.1 разложение контекста $K = (G, M, I)$ на фрагменты является «неискажающим» для любого формального понятия из FC. В теореме 2.1 исключены случаи, когда FC содержит хотя бы одно из тривиальных понятий $(G, \varnothing)$, $(\varnothing, M)$. Поскольку всегда верны $(\varnothing, M) \sqsubseteq (G, \varnothing)$, $(\varnothing, M) \sqsubseteq (G, G')$, $(M', M) \sqsubseteq (G, \varnothing)$, то контекст $K = (G, M, I)$ можно рассматривать как фрагмент (G, M). Следовательно, и даже в этих исключительных случаях каждый фрагмент содержит, по крайней мере, одно формальное понятие из FC, при этом ни одно формальное понятие из FC не теряется.

Из теоремы 2.1 вытекает важное практическое следствие: искомое множество FC может быть восстановлено путем объединения множеств формальных понятий, выявленных в фрагментах контекста $K = (G, M, I)$. Очевидно, что процесс разложения заданного контекста на фрагменты может быть организован итерационно, поскольку каждый выявленный на первой итерации фрагмент можно рассматривать в качестве исходного контекста и вновь подвергать декомпозиции. Оценку числа фрагментов, получаемых на каждой итерации разложения, устанавливает предложение 2.2. Определим теперь правило остановки итерационного процесса разложения. Для этого введем понятие плотности фрагмента. Пусть $|m'| \cdot |g'|$ — размер фрагмента (m', g'), а $\|(m', g')\|$ — число его единичных элементов.

Определение 2.6. Плотностью фрагмента (m', g') назовем величину

$$\sigma(m', g') = \frac{\|(m', g')\|}{|m'| \cdot |g'|}.$$

Верны естественные границы $0 < \sigma(m', g') \leq 1$.

Предложение 2.4. *Если фрагмент* (m', g') *контекста* $K = (G, M, I)$, *образованный элементами* $g \in G$ *и* $m \in M$, *имеет плотность* $\sigma(m', g') = 1$, *то* $g'' = m'$, $m'' = g'$.

Предложение 2.5. *Всякий фрагмент* (m', g') *с плотностью* $\sigma(m', g') = 1$ *содержит ровно одно нетривиальное формальное понятие* (A, B) *контекста* $K = (G, M, I)$, *совпадающее с ним, т. е.* $A = m'$ *и* $B = g'$.

Из предложения 2.5 следует, что фрагмент (m', g') с плотностью 1 вырождается в нетривиальное формальное понятие и не подлежит дальнейшему разложению. Заметим, что время формирования одного фрагмента для формального контекста $K = (G, M, I)$ составляет $O(|G| \cdot |M|)$. В целом, время, необходимое на однократное разложение этого контекста на фрагменты в худшем случае, составляет $O\left(\sigma(G, M) \cdot |G|^2 \cdot |M|^2\right)$, где $\sigma(G, M)$ — плотность исходного формального контекста. Таким образом, чем меньше плотность контекста, тем быстрее осуществляется его разложение на фрагменты. При $\sigma(G, M) = 1$ контекст $K = (G, M, I)$ не подлежит разложению.

Число фрагментов, возникающих на каждой отдельной итерации процесса декомпозиции, в ряде случаев может быть уменьшено за счет удаления вложенных и кратных фрагментов. Рассмотрим для формального контекста $K = (G, M, I)$ множество фрагментов $\Omega = \left\{\omega_1, \omega_2, \ldots, \omega_{\|T\|}\right\}$, где $\omega_i = (m_i', g_i')$, $i = 1, 2, \ldots, \|T\|$.

Определение 2.7. Будем говорить, что фрагмент $\omega_1 = (m_1', g_1')$ вложен в фрагмент $\omega_2 = (m_2', g_2')$, и писать $\omega_1 \preceq \omega_2$, если верны теоретико-множественные включения $m_1' \subseteq m_2',\ g_1' \subseteq g_2'$.

При $m_1' = m_2'$ и $g_1' = g_2'$ фрагменты ω_1 и ω_2 назовем кратными. Будем считать, что фрагменты ω_1 и ω_2 сравнимы между собой, если $\omega_1 \preceq \omega_2$ или $\omega_2 \preceq \omega_1$, иначе несравнимы. Таким образом, множество Ω частично упорядочено относительно введенного выше отношения порядка. Множество типичных представителей частично упорядочено относительно того же отношения только в обратном порядке. Следующее предложение доказывает данный факт.

Предложение 2.6. *Пусть* $h_1 = (g_1'', m_1'')$ — *типичный представитель фрагмента* $\omega_1 = (m_1', g_1')$, *а* $h_2 = (g_2'', m_2'')$ — *типичный представитель фрагмента* $\omega_2 = (m_2', g_2')$. *Тогда, отношение порядка* $h_2 = (g_2'', m_2'') \preceq h_1 = (g_1'', m_1'')$ *выполняется, т. е. верны включения* $m_2'' \subseteq m_1'', g_2'' \subseteq g_1''$, *тогда и только тогда, когда* $\omega_1 \preceq \omega_2$.

С учетом теоремы 2.1 справедливо следующее следствие.

Следствие 2.1. *Для любых* $\omega_1, \omega_2 \in \Omega$ *таких, что* $\omega_1 \preceq \omega_2$, *все формальные понятия фрагмента* ω_1 *также являются формальными понятиями фрагмента* ω_2 *и контекста* $K = (G, M, I)$.

Известно, что в частично упорядоченном множестве всегда можно найти взаимно непересекающиеся цепи. Непустое подмножество $\{\omega_{i1}, \omega_{i2}, \ldots, \omega_{il}\}$ множества Ω является цепью, если все элементы этого подмножества попарно сравнимы между собой и линейно упорядочены $\omega_{i1} \preceq \omega_{i2} \preceq \ldots \preceq \omega_{il}$. Элемент ω_{il} называется максимальным элементом, а величина l — длиной этой цепи. Цепь называется максимальной, если ее объединение с любым, не принадлежащим ей элементом, цепью не является. Две цепи называются взаимно непересекающимися, если они не содержат общих элементов. Число максимальных взаимно непересекающихся цепей и длина самой длинной такой цепи определяются теоремой Дилоурса. Согласно следствию 2.1 максимальный элемент всякой цепи сохраняет все формальные понятия остальных элементов этой цепи. Данные элементы могут быть удалены и тем самым уменьшено число фрагментов, получаемых на каждой отдельной итерации разложения.

Доказательства предложений 2.1–2.6 и теоремы 2.1 базируются на определении отображений Галуа и свойствах замкнутых множеств.

В параграфе 2.2 описан алгоритм FindBoxes, реализующий предложенный в диссертационной работе метод «неискажающего» разложения исходного контекста. Входными данными алгоритма FindBoxes являются формальный контекст $K = (G, M, I)$ и целое положительное число k — число итераций. Результат работы алгоритма: Ω — множество фрагментов и H — множество типичных представителей фрагментов, входящих в Ω.

Алгоритм FindBoxes включает следующие основные процедуры: Boxes, Delete, SearchChains. Процедура Boxes разлагает заданный фрагмент ω, плотность которого отлична от 1, на более мелкие фрагменты и находит для них типичных представителей. Процедура Delete осуществляет удаление кратных фрагментов и фрагментов, совпадающих с исходным фрагментом. Процедура SearchChains выявляет вложенные фрагменты, выполняет построение взаимно непересекающихся цепей частично упорядоченного множества фрагментов Ω_1, и далее находит для этих цепей максимальные элементы. Данная процедура позволяет уменьшать число фрагментов, получаемых на каждой отдельной итерации разложения.

Очевидно, что если число итераций процесса декомпозиции равно k, то разложение можно осуществить за время $O\left(|G|^{2k} \cdot |M|^{2k}\right)$. Если $k = 1$, то алгоритм FindBoxes выполняется за время $O\left(|G|^2 \cdot |M|^2\right)$. Для дополнительного ограничения числа частей следует устанавливать пороговое значение на плотность фрагментов, подлежащих дальнейшему разложению. Это достигается заменой на шаге 10 алгоритма FindBoxes условия $\sigma(\omega) \neq 1$ условием $\sigma(\omega) < \sigma_0$, где $\sigma(\omega)$ — плотность фрагмента, σ_0 — пороговое значение на плотность фрагментов, которые подлежат дальнейшему разложению.

Алгоритм 1. FindBoxes

Вход: исходный контекст $K = (G, M, I)$, k — количество итераций

begin
$\Omega_1 \leftarrow (G, M, I)$ ▷ множество фрагментов, подлежащих дальнейшему разложению
$\Omega_2 \leftarrow \varnothing$ ▷ множество фрагментов, не подлежащих дальнейшему разложению
$H_1 \leftarrow (G'', M'')$ ▷ множество типичных представителей фрагментов, входящих в Ω_1
$H_2 \leftarrow \varnothing$ ▷ множество типичных представителей фрагментов, входящих в Ω_2
while $(k \neq 0 \ \& \ \Omega_1 \neq \varnothing)$ **do**
    $Q \leftarrow \varnothing$
    $R \leftarrow \varnothing$
    for all $\omega \in \Omega_1$ **do**
        if $\sigma(\omega) \neq 1$ **then**
            Boxes(ω, X, Y)
            $Q \leftarrow Q \cup X$
            $R \leftarrow R \cup Y$
        else
            $\Omega_2 \leftarrow \Omega_2 \cup \omega$
            $H_2 \leftarrow H_2 \cup H_1$
        end if
    end for
    $\Omega_1 \leftarrow Q$
    $H_1 \leftarrow R$
    Delete $(\Omega_1 \cup \Omega_2, H_1 \cup H_2)$
    if $\Omega_1 \neq \varnothing$ **then**
        SearchChains(Ω_1, H_1)
    end if
    $k \leftarrow k - 1$
end while
$\Omega \leftarrow \Omega_1 \cup \Omega_2$
$H \leftarrow H_1 \cup H_2$
end

Выход: Ω — множество фрагментов, H — множество типичных представителей фрагментов из Ω

В параграфе 2.3 разработан алгоритм LatticeContext восстановления искомого решения исходя из решений, полученных для подзадач. Вычислительная сложность алгоритма в худшем случае равна

$$O\left(p\left(|G|, |M|\right) \cdot |FC| \cdot |G|^2 \cdot |M|\right),$$

где $p\left(|G|, |M|\right)$ — полином от $|G|$ и $|M|$. Далее в параграфах 2.4 и 2.5 подробно описаны алгоритмы Query1, Query2 реализации запросов на извлечение знаний из решетки формальных понятий и процедуры предобработки исходного формального контекста без потери формальных понятий. Алгоритмы Query1, Query2 предусматривают два вида (X, Y)-запросов, где $X \in G$ и $Y \in M$: построение маршрута, содержащего в каждом узле (X, Y); установление общих и частных понятий для заданного формального понятия (X, Y). Вычислительная сложность этих алгоритмов сопоставима с размером заданной решетки.

В параграфе 2.6 представлен анализ результативности разработанных алгоритмов и сделаны следующие выводы: почти все разработанные в диссертации алгоритмы имеют высокую вычислительную сложность; на практике при удачном задании значений k и σ_0 возможно построение полиномиального числа фрагментов. Проведенные вычислительные эксперименты показали (таблица 1), что применение предложенного метода декомпозиции существенно уменьшает время нахождения всех формальных понятий заданного контекста. Анализировались два случая: случай 1 — в процедуре SearchChains проверка вложенности фрагментов $w_i \preceq w_j$ осуществляется без типичных представителей фрагментов, случай 2 — проверка вложенности фрагментов $w_i \preceq w_j$ выполняется с использованием типичных представителей.

Таблица 1 — Оценка эффективности процесса декомпозиции контекста

	Характеристика исходного контекста				Результаты		
	$\|G\|$	$\|M\|$	$\|\|T\|\|$	$\sigma(G,M)$	N	$\|FC\|$	t, мс
Без разложения на фрагменты	100	20	1000	0,5	–	4962	145125
С разложением на фрагменты (сл. 1)					883	4962	2878
С разложением на фрагменты (сл. 2)					883	4962	**2200**
Без разложения на фрагменты	200	30	2940	0,49	–	10567	794520
С разложением на фрагменты (сл. 1)					2895	10567	97906
С разложением на фрагменты (сл. 2)					2895	10567	**90908**

Из таблицы 1 следует, что:

– значения $|FC|$ в случаях без разложения и с разложением на фрагменты полностью совпадают. Это иллюстрирует справедливость теоремы 2.1, т. е. «неискажаемость» разложения контекста на фрагменты относительно формальных понятий;

– число N фрагментов, образованных при разложении контекста неизменно не превышает величины $\|T\|$, что свидетельствует о правильности предложения 2.2;

– применение предложенного метода декомпозиции дает значительный выигрыш по времени: время выполнения программы FCACorpus при разложении контекста на фрагменты уменьшается в несколько раз;

– проверка вложенности фрагментов по типичным представителям дает незначительный эффект по времени работы алгоритма FindBoxes.

Вычислительные эксперименты также свидетельствуют, что чем выше плотность исходного контекста, тем больше времени требуется на однократное разложение контекста. Для выполнения всего процесса декомпозиции за полиномиальное время рекомендуется k задавать значительно меньше, чем $k \ll |G|/2$, а пороговое значение выбирать из интервала $\sigma(G,M) < \sigma_0 < 1$.

В **третьей главе** описаны программные средства, реализующие разработанные в диссертации метод и алгоритмы в виде отдельных модулей комплекса программ FCACorpus. Цель создания FCACorpus — оценка результативности созданных средств, проведение экспериментальных исследований на электронной коллекции «Тувинские героические сказания». В этой главе представлена четвертая задача диссертационного исследования. Основные результаты этой главы опубликованы в работах [5, 6, 8–10, 13, 14].

В параграфе 3.1 приведено описание разработанного комплекса программ FCACorpus, который реализован на языке программирования C# в интегрированной среде разработки Microsoft Visual Studio Community 2017.

Для его эксплуатации требуется персональный компьютер типа IBM PC Pentium IV с операционной системой Windows XP/Vista/7/8 и оперативной памятью от 512 Мб. Входные данные FCACorpus включают формальный контекст $K = (G, M, I)$, а также другие данные в зависимости от выбранного режима работы этой программы. Допускается ввод контекста из внешнего текстового файла. Результатом работы FCACorpus являются построенная решетка L формальных понятий контекста $K = (G, M, I)$ и результаты запроса, если он был задан во входных данных. Большинство модулей, входящих в FCACorpus, универсальны и не привязаны к каким-либо конкретным базам данных. Для привязки FCACorpus к конкретной предметной области требуется база данных исследуемой предметной области, а также специальный модуль, обеспечивающий информационный интерфейс между базой данных и FCACorpus. Кроме того, необходимы специальные модули, реализующие конкретные прикладные задачи. Привязка FCACorpus к корпусу тувинского языка осуществляется путем разработанных модуля IInterface и базы данных «Тувинские героические сказания».

В параграфе 3.2 рассмотрена организация базы данных «Тувинские героические сказания». База данных «Тувинские героические сказания» — информационная составляющая корпуса тувинского языка, в которой хранятся оцифрованные тексты более 50 произведений и сведения о их сказителях, представленные в виде объектно-признаковой таблицы.

В параграфе 3.3 в рамках корпуса тувинского языка проведены экспериментальные исследования, подтверждающие высокую результативность разработанных в диссертационной работе метода и алгоритмов при решении задачи установления авторского стиля сказителей тувинского героического эпоса. Полученные результаты показали, что разработанные средства могут быть применены не только для распознавания авторского стиля сказителей, но и для других подобных задач анализа текстов в рамках корпуса тувинского языка.

В **заключении** диссертации сформулированы основные результаты и выводы, полученные на основе настоящей диссертационной работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. Разработан и теоретически обоснован метод «неискажающего» разложения формального контекста на фрагменты (теорема 2.1). Исследована структура фрагментов и найдена оценка числа фрагментов, получаемых на каждой итерации разложения, определены правила остановки процесса разложения формального контекста на фрагменты без потери формальных понятий (предложения 2.1–2.6).

2. Разработаны алгоритмы формирования для заданного формального контекста системы фрагментов, восстановления искомого решения исходя из решений, полученных для подзадач, и реализации возможных запросов на извлечение знаний из решетки формальных понятий.

3. Разработаны алгоритмы предобработки формального контекста без потери формальных понятий путем удаления единичных, нулевых и кратных строк и столбцов этого контекста.

4. Создан комплекс программ, реализующий разработанные метод и алгоритмы. Эксперименты показали, что все разработанные в диссертации алгоритмы имеют высокую вычислительную сложность. Однако на практике при удачном задании значений k и σ_0 возможно построение полиномиального числа $|\Omega| = p\left(|G|, |M|\right)$ фрагментов. Эксперименты подтверждают, что увеличение числа итераций приводит к увеличению числа фрагментов, подлежащих дальнейшему разложению, и в свою очередь к увеличению времени выполнения алгоритмов. Поэтому количество итераций разложения k *рекомендуется* задавать значительно меньше, чем $k \ll |G|/2$, а пороговое значение выбирать из интервала $\sigma(G, M) < \sigma_0 < 1$.

Благодарности. Автор выражает искреннюю и глубокую благодарность д.ф.-м.н., профессору Быковой Валентине Владимировне за неоценимую помощь и поддержку на всех этапах выполнения работы.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

В изданиях, рекомендованных ВАК:

1. Bykova V. V., **Mongush Ch. M.** On Algebraic Approach of R. Wille and B. Ganter in the Investigation of Texts // Journal of Siberian Federal University. Mathematics & Physics. — 2017. — № 3 (10). — P. 372–384 (индексируется **Web of Science**, **Scopus**).

2. Быкова В. В., **Монгуш Ч. М.** Алгоритмы концептуального моделирования и классификации текстов в корпусе тувинского языка // Программные продукты и системы. — 2017. — № 3 (30). — С. 487–495.

3. Быкова В. В., **Монгуш Ч. М.** Декомпозиционный подход к исследованию формальных контекстов // Прикладная дискретная математика. — 2019. — № 44. — С. 113–126 (индексируется **Web of Science**, **Scopus**).

4. **Mongush Ch. M.**, Bykova V. V. On decomposition of a binary context without losing formal concepts // Journal of Siberian Federal University. Mathematics & Physics. — 2019. — № 3 (12). — P. 323–330 (индексируется **Web of Science**, **Scopus**).

В материалах конференций и других изданиях:

5. **Монгуш Ч. М.** Методы анализа формальных понятий в исследовании текстов тувинского фольклора // Материалы XV Международной конференции имени А. Ф. Терпугова «Информационные технологии и математическое моделирование». — Томск: Изд-во Том. ун-та, 2016. — Ч. 2. — С. 153–158.

6. Быкова В. В., **Монгуш Ч. М.** Распознавание жанра произведений тувинского фольклора на основе анализа формальных понятий // Труды Международной конференции «Актуальные проблемы прикладной математики и информационных технологий. — Аль-Хорезми 2016». — Бухара: Изд-во Национ. ун-та., 2016. — С. 202–205.

7. Быкова В. В., **Монгуш Ч. М.** Алгебраический подход исследования текстов тувинского фольклора // Материалы VI Международной конференции «Математика, ее приложения и математическое образование». — Улан-Удэ: Изд-во ВСГУТУ, 2017. — С. 277–281.

8. **Монгуш Ч. М.**, Ондар М. В. База данных и средства создания контекстов для представления и анализа тувинского героического эпоса // Программные продукты, системы и алгоритмы. — 2017. — № 3. — С. 1–6.

9. **Монгуш Ч. М.** О классификации произведений тувинского фольклора и распознавании жанра героического эпоса // Материалы XVII Международной конференции имени А. Ф. Терпугова «Информационные технологии и математическое моделирование». — Томск: Изд-во НТЛ, 2018. — С. 257–263.

10. **Монгуш Ч. М.** Программа формирования контекста для электронной коллекции «Тувинские героические сказания» // Инженерный вестник Дона. — 2018. — № 2(49). — С. 119–128.

11. **Монгуш Ч. М.** Алгоритм «безопасной» декомпозиции формального контекста // Прикладная дискретная математика. Приложение (труды Всероссийской конференции «Компьютерная безопасность и криптография»). — 2019. — № 12. — С. 227–232.

12. **Монгуш Ч. М.**, Семенова Д. В. О «неискажающем» разложении бинарного контекста в анализе данных и комбинаторной оптимизации // Материалы VII Международной конференции «Знания – Онтологии – Теории». — Новосибирск: Изд-во Института математики им. С. Л. Соболева СО РАН, НГУ, 2019. — С. 394–395.

Свидетельства о государственной регистрации программы для ЭВМ:

13. **Монгуш Ч. М.** Программа FCACorpus концептуального моделирования тувинских текстов методами анализа формальных понятий. Свидетель-

ство о государственной регистрации программы для ЭВМ № 2018618907. Зарегистрировано в Реестре программ для ЭВМ от 23 июля 2018 г.

14. **Монгуш Ч. М.**, Быкова В. В. Программа формирования контекстов в корпусе тувинского языка. Свидетельство о государственной регистрации программы для ЭВМ № 2018618908. Зарегистрировано в Реестре программ для ЭВМ от 23 июля 2018 г.