

ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертационную работу ЦЫМБЛЕРА Михаила Леонидовича «Интеллектуальный анализ данных в СУБД», представленную на соискание ученой степени доктора физико-математических наук по специальности 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Как известно, интеллектуальный анализ данных направлен на извлечение из них знаний, доступных для понимания и необходимых для принятия решений в различных сферах деятельности. Стремительное увеличение объема информации, наблюдаемое в настоящее время, требует создания новых инструментов для ее хранения и обработки. При этом эффективная навигация и извлечение полезной информации приводит к значительным временным затратам. Диссертационная работа М.Л. Цымблера как раз и посвящена разработке одного из таких инструментов – масштабируемых методов и параллельных алгоритмов интеллектуального анализа данных на платформе современных вычислительных кластеров с вычислительными узлами на базе многоядерных ускорителей. Одним из лейтмотивом проведенного исследования является приближение аналитических алгоритмов вплотную к анализируемым данным. При этом функции интеллектуального анализа данных распараллеливаются для платформы многоядерных ускорителей и встраиваются в реляционную СУБД.

В настоящее время в сфере систем управления данными доминируют реляционные СУБД, несмотря на популярность инструментов класса NoSQL. Выполнение анализа данных внутри СУБД дает возможность избежать существенных накладных расходов на экспорт–импорт данных. Реляционная СУБД распараллеливается для платформы вычислительных кластеров посредством

модификаций исходного кода свободной последовательной СУБД, что является существенно более экономичным решением, чем разработка параллельной СУБД «с нуля».

Учитывая теоретическую и практическую значимость решения проблемы разработки эффективных методов и алгоритмов хранения, обработки и интеллектуального анализа Больших данных, считаю, что выбранная соискателем тема исследования и предлагаемые им методы и подходы, несомненно, являются **актуальными**.

Целью проведенного диссертационного исследования была разработка комплекса параллельных алгоритмов, составляющих программную платформу интеллектуального анализа данных средствами СУБД с открытым кодом. Для достижения этой цели были решены **следующие задачи**:

- разработать методы и алгоритмы для распараллеливания СУБД с открытым кодом и апробировать их на выбранной СУБД;
- разработать методы и алгоритмы, обеспечивающие проведение интеллектуального анализа данных в рамках параллельной СУБД на вычислительных кластерах с многоядерными ускорителями;
- разработать комплекс параллельных алгоритмов для кластеризации данных, поиска шаблонов и анализа временных рядов в параллельной реляционной СУБД;
- провести вычислительные эксперименты, результаты которых позволили бы оценить эффективность предложенных методов и алгоритмов и их преимущества перед известными мировыми аналогами.

Все перечисленные задачи были успешно решены, получен ряд новых интересных научных результатов, **основными** из которых, по моему мнению, являются следующие:

1. Разработан, исследован и успешно апробирован комплекс параллельных алгоритмов интеллектуального анализа данных средствами СУБД на кластерных вычислительных системах с многоядерными ускорителями, в том числе, алгоритмы: поиска похожих подпоследовательностей временного ряда для кластеров с многоядерными ускорителями; поиска диссонансов временного ряда для многоядерных ускорителей; поиска частых наборов для многоядерных ускорителей; нечеткой кластеризации данных и кластеризации графа для параллельной СУБД на основе фрагментного параллелизма; кластеризации данных для многоядерных ускорителей.

2. Предложен подход к внедрению параллельных алгоритмов интеллектуального анализа данных в реляционные СУБД.

3. Предложен метод внедрения фрагментного параллелизма в последовательные СУБД с открытым исходным кодом.

Все эти результаты являются **новыми**. В целом же **научная новизна результатов диссертации** характеризуется следующим.

Разработан ряд новых параллельных алгоритмов для интеллектуального анализа данных на многоядерном ускорителе: поиска частых наборов, кластеризация данных, поиска похожих и аномальных подпоследовательностей во временных рядах.

Предложен новый метод инкапсуляции параллелизма в последовательную СУБД с открытым исходным кодом, подразумевающий сравнительно небольшие по объему изменения в исходном коде.

Разработаны новые параллельные алгоритмы интеллектуального анализа данных для параллельной СУБД.

Предложен оригинальный метод выполнения интеллектуального анализа данных внутри реляционной СУБД, инкапсулирующий параллельные алгоритмы анализа данных на многоядерных ускорителях.

Теоретическая значимость исследования, проведенного соискателем, состоит в том, что разработанные методы и подходы являются существенным развитием теории СУБД и позволяют внедрить в последовательные реляционные СУБД параллельные обработку и интеллектуальный анализ данных. При этом разработанные параллельные алгоритмы решения ряда задач интеллектуального анализа данных для многоядерных ускорителей обеспечивают ускорение, близкое к линейному.

Практическая ценность диссертационного исследования заключается в том, что предложенные методы и подходы уже реализованы на практике: выполнено распараллеливание свободной СУБД PostgreSQL и в полученный прототип параллельной СУБД PargreSQL внедрены разработанные параллельные алгоритмы интеллектуального анализа данных.

Обоснованность и достоверность результатов диссертации подтверждаются вычислительными экспериментами с реальными и синтетическими данными, которые проведены согласно общепринятым стандартам, и тем, что выполнено сравнение полученных результатов с существующими подходами и результатами.

Кратко охарактеризуем структуру и содержание диссертации.

Диссертация состоит из введения, пяти глав, заключения и библиографии. Объем диссертации составляет 260 страниц, объем библиографии – 274 наименования.

Во введении обоснована актуальность темы и охарактеризована степень ее разработанности; сформулированы цели и задачи исследования; раскрыты новизна, теоретическая и практическая значимость полученных результатов; методологическая основа диссертационного исследования.

В главе 1 рассмотрены типовые задачи анализа данных и дан обзор методов и подходов к интеграции интеллектуального анализа данных в реляционные СУБД.

Во второй главе рассмотрены задачи кластеризации данных сверхбольшого объема, которые не могут быть размещены в оперативной памяти, в реляционных СУБД и параллельного поиска шаблонов на многоядерных ускорителях. Предложены алгоритмы dbParGraph кластеризации графа и pgFCM нечеткой кластеризации данных для параллельной СУБД на основе фрагментного параллелизма. Разработаны параллельные алгоритмы PDIC и DDCapriori поиска частых наборов для многоядерных ускорителей (архитектур Intel MIC и IBM Cell соответственно). Представлены результаты вычислительных экспериментов, характеризующих эффективность разработанных алгоритмов.

Глава 3 посвящена методам интеллектуального анализа временных рядов на платформе современных многопроцессорных многоядерных вычислительных систем. Предложены новые параллельные алгоритмы PBM поиска похожих подпоследовательностей во временном ряде для кластерных систем с узлами на базе многоядерных ускорителей и MDD поиска диссонансов во временном ряде для многоядерных ускорителей. Представлены результаты вычислительных экспериментов, исследующих эффективность предложенных алгоритмов.

В четвертой главе представлен подход к интеграции интеллектуального анализа данных в реляционную СУБД. Предложенный подход предполагает внедрение параллельных алгоритмов интеллектуального анализа данных в реляционную СУБД с открытым кодом на основе функций, определяемых пользователем. В качестве целевой площадки реализации подхода использована СУБД PostgreSQL. Описана системная архитектура и методы реализации

подхода. Предложены параллельные алгоритмы PBlockwise вычисления матрицы евклидовых расстояний и PРАМ кластеризации данных для ускорителей Intel MIC, внедряемые в СУБД. Представлены результаты вычислительных экспериментов, исследующих эффективность предложенного подхода.

В пятой главе предложены методы, позволяющие внедрить фрагментный параллелизм в свободную реляционную СУБД посредством модификации ее открытых исходных кодов. Описаны архитектура и принципы реализации параллельной СУБД PargreSQL на платформе вычислительного кластера, получаемой посредством распараллеливания свободной СУБД PostgreSQL. Представлены результаты вычислительных экспериментов, характеризующих эффективность предложенных методов.

В заключении в краткой форме изложены итоги выполненного диссертационного исследования.

Отмечу, что как сама диссертация, так и ее автореферат написаны и оформлены очень хорошо и практически не имеют грамматических и стилистических погрешностей. Первая глава выглядит почти как готовое учебное пособие. Нестандартно выглядит и заключение, в котором, в частности, указаны отличия настоящей работы от ранее выполненных родственных работ других авторов, а также сформулированы направления дальнейших исследований.

Публикации и апробация результатов. Соискатель имеет 12 статей, опубликованных в журналах из Перечня ВАК, и 9 статей в изданиях, проиндексированных в библиографических базах данных Scopus и Web of Science. Все эти публикации с достаточной полнотой отражают основные результаты диссертации. Кроме того, зарегистрированы у программы для ЭВМ.

Результаты исследования прошли также значительную апробацию на международных и всероссийских научных конференциях. Автореферат диссертации достаточно полно и правильно отражает ее содержание.

ЗАМЕЧАНИЯ ПО ДИССЕРТАЦИИ

1. В диссертации недостаточно четко отражена актуальность именно реализованных методов. В частности, отсутствует анализ методов нейронных сетей, дисперсионного анализа, в том числе, метода главных компонент. Для временных рядов отсутствует упоминание методов их прогнозирования. Кроме того, целесообразно было отразить проблемы, общие для всех выбранных и проанализированных методов.

2. Представленные обзор и сопоставление с имеющимися аналогами, а также проведенное тестирование недостаточно полны. В частности, отсутствует описание недостатков использования для анализа данных, кроме их высокой стоимости, коммерческих систем Oracle, DB2 и Terradata. Приведенное в диссертации сопоставление с некоммерческими проектами не охватывает такие известные системы, как SciDB и Postgres XL. Проведенные тесты реализованной диссертантом распределенной системы не охватывают анализа на нагрузочную способность, в частности, не проанализирована возможность возникновения дедлоков.

3. Почему в проведенном исследовании сразу была выбрана СУБД PostgreSQL? В диссертации не рассматриваются существующие в настоящий момент на рынке открытые (или частично открытые) реляционные параллельные СУБД, в том числе, на основе PostgreSQL, такие, как Greenplum, CockroachDB, а также Postgres-XL и SciDB (уже упомянутые выше). Известно, что некоторые из названных СУБД предлагаются и применяются для хранения и анализа данных в коммерческих предприятиях. Представляется полезным (может быть, в будущем) провести сравнения используемых в них технологий параллелизма с рассматриваемой в работе технологией фрагментарного параллелизма, а также оценить возможность интеграции разработанных модулей интеллектуального анализа.

4. В диссертации практически во всех ее главах проведено сравнение разработанных методов с некоторыми из тех методов, которые были названы ранее в главе 1 без какого-либо их сравнения между собой. Только при анализе полученных результатов говорится, что сравнение проводилось с лучшими методами. Естественно возникает вопрос, почему не было выполнено сразу сравнение известных методов между собой или сразу не были названы лучшие, чтобы потом проводить сравнение именно с ними.

Далее, снова про различные сравнения. На с. 35 автореферата сказано, что при сравнении с лучшими аналогами алгоритм MDD показал быстроедействие, опережающее конкурентов в 2–3 раза. Аналогично на с. 38 автореферата сказано, что при сравнении с лучшими аналогами алгоритм PBlockwise показал быстроедействие, опережающее конкурентов. Хотелось бы, чтобы были названы эти аналоги и эти конкуренты.

Еще: говорится об экспериментах с алгоритмом PBlockwise на суперкомпьютере ИВМиМГ СО РАН, но, к сожалению, никакие сведения об этом суперкомпьютере не приведены, поэтому естественно возникает вопрос, насколько при этом проведенное сравнение оправдано.

Наконец, в диссертации сказано, что с учетом указанных накладных расходов на экспорт–импорт информации разработанный алгоритм dbParagraph опережает аналоги по производительности. А как учитывать и оценивать накладные расходы, ничего не сказано.

5. Разработанные многочисленные алгоритмы – один из основных результатов проведенного исследования, который выносится на защиту. Все эти алгоритмы представлены на очень высоком уровне абстракции (фактически описаны «на пальцах»). Такое описание достаточно для понимания применяемых принципов и подходов, но затруднительно для чтения. Целесообразно было вынести эти фрагменты программ в отдельное приложение.

6. В работе описаны алгоритмы кластеризации данных, где каждый из полученных фрагментов таблицы физически хранится на различных вычислительных узлах кластера. Возникает вопрос: предусмотрено ли динамическое обновление такой структуры хранения (для таблиц с быстро изменяющимися данными) с целью поддержки оптимального фрагментирования?

7. Во введении, в разделе «Степень разработанности темы» (с. 8 и далее), названы многие научные группы и их руководители без каких-либо ссылок на их публикации (хотя бы основные), что затрудняет чтение и понимание акцентов, сделанных диссертантом.

8. В автореферате имеется 30 пронумерованных формул, практически на все из них в автореферате нет ссылок. Эта нумерация не совпадает с нумерацией, использованной в диссертации. Зачем она применена в автореферате, не ясно.

Высказанные замечания по диссертационной работе не снижают значимости полученных результатов и оценки общего высокого научного уровня работы.

ЗАКЛЮЧЕНИЕ

Диссертационная работа представляет собой самостоятельно выполненную и законченную научно-квалификационную работу, в которой автором предложен новый метод инкапсуляции параллелизма в свободную СУБД и разработан комплекс новых параллельных алгоритмов интеллектуального анализа данных в СУБД на вычислительных кластерах с многоядерными ускорителями. Полученные соискателем результаты могут быть квалифицированы как научное достижение высокого уровня, диссертационная работа в полной мере соответствует требованиям Положения о порядке присуждения ученых

степеней, включая п. 9, а ее автор, Михаил Леонидович Цымблер, достоин присуждения ему ученой степени доктора физико-математических наук по специальности 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Официальный оппонент, профессор Подпись А.М. Елизаров удостоверяю
кафедры «Программная инженерия»

Высшей школы информационных

технологий и интеллектуальных систем Проректор по научной деятельности

Казанского (Приволжского) федерального университета, доктор ности Казанского (Приволжского)

физико-математических наук федерального университета,

(01.02.05 – Механика жидкости и профессор

газа), профессор, Заслуженный деятель науки Республики Татарстан

Александр Михайлович Елизаров

«9» января 2020 г.



Данис Карлович Нурғалиев

«9» января 2020 г.

Казанский (Приволжский) федеральный университет, 420008, г. Казань, ул. Кремлевская, 35.

Тел. 8-9376151553,

e-mail: amelizarov@gmail.com

СВЕДЕНИЯ ОБ ОФИЦИАЛЬНОМ ОППОНЕНТЕ

по диссертации Цымблера М.Л. «Интеллектуальный анализ данных в СУБД», представленной на соискание ученой степени доктора физико-математических наук по специальности 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Фамилия, имя, отчество	Елизаров Александр Михайлович
Ученая степень (с указанием номера и шифра специальности)	Доктор физико-математических наук, 01.02.05 – Механика жидкости, газа и плазмы
Ученое звание	Профессор
Организация основного места работы	Федеральное государственное автономное образовательное учреждение высшего обра- зования «Казанский (Приволжский) феде- ральный университет», Высшая школа ин- формационных технологий и интеллектуаль- ных систем (ВШ ИТИС КФУ)
Занимаемая должность	Профессор кафедры программной инжене- рии
Почтовый адрес	420008, Казань, ул. Кремлевская, дом 35
Телефон	+7 (843) 233-72-46
Адрес электронной почты	amelizarov@gmail.com

Список основных публикаций по теме диссертации в рецензируемых научных изданиях

1. Биряльцев Е.В., Галимов М.Р., Демидов Д.Е., Елизаров А.М. Платформенный подход к выполнению исследовательских и проектных работ с использованием высокопроизводительных вычислений// Программные системы: теория и приложения. 2019. Т. 10, № 2 (41). С. 121–153.

2. Eugeny Biryaltsev, Marat Galimov, Denis Demidov, Aleksandr Elizarov. The platform approach to research and development using high-performance computing// Program Systems: Theory and Applications, 2019, V. 10, No 2(41). P. 93–119.

3. Biryal'tsev E.V., Galimov M.R., Elizarov A.M. Work flow-based Internet Platform for Mass Supercomputing// Lobachevskii Journal of Mathematics. 2018. V. 39, No 5. P. 647-654.

4. Биряльцев Е.В., Галимов М.Р., Елизаров А.М. Программная платформа массового суперкомпьютинга // Доклады Академии наук. 2017. Т. 473. № 3. С. 277-281.

5. Беляева А.А., Биряльцев Е.В., Галимов М.Р., Демидов Д.Е., Елизаров А.М., Жибрик О.Н. Кластерная архитектура программно-технических средств организации высокопроизводительных систем для нефтегазовой промышленности // Программные системы: теория и приложения. 2017. Т. 8. № 1 (32). С. 151-171.

6. Елизаров А.М., Липачёв Е.К., Невзорова О.А., Соловьев В.Д. Методы и средства семантического структурирования электронных математических документов // Доклады Академии наук. 2016. Т. 457. № 6. С. 642.

7. Елизаров А.М., Жижченко А.Б., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К. Онтологии математического знания и рекомендательная система для коллекций физико-математических документов // Доклады Академии наук. 2016. Т. 467. № 4. С. 392.

8. Биряльцев Е.В., Богданов П.Б., Галимов М.Р., Демидов Д.Е., Елизаров А.М. Программно-техническая платформа высокопроизводительных вычислений для нефтегазовой промышленности // Программные системы: теория и приложения. 2016. Т. 7. № 1 (28). С. 15-27.

9. Elizarov A.M., Lipachev E.K., Haidarov S.M. Automated processing service system of large collections of scientific documents. 18th International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016), Ershovo, Moscow Region, Russia, October 11-14, 2016. CEUR Workshop Proceedings, vol. 1752. pp. 58-64.

10. Biryal'tsev E., Elizarov A., Zhil'tsov N., Lipachev E., Nevzorova O., Solov'ev V. Methods for analyzing semantic data of electronic collections in mathematics. Automatic Documentation and Mathematical Linguistics. 2014. vol. 48, no. 2. pp. 81-85.

11. Биряльцев Е.В., Елизаров А.М., Жильцов Н.Г., Липачёв Е.К., Невзорова О.А., Соловьев В.Д. Методы анализа семантических данных математических электронных коллекций // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2014. № 4. С. 12-17.

/ Елизаров А.М. /

Подпись А.М. Елизарова заверяю

Директор ВШ ИТИС КФУ



/ Абрамский М.М. /

9 января 2020 года