

РЕЦЕНЗИЯ

от проф. д-р Радослав Даков Йошинов

на дисертационен труд на тема „Методи за реализация на софтуерни системи за обработка на големи данни (Methods for implementation of data-intensive software systems)“

с автор *Симеон Стоичков Емануилов* –задочен докторант към катедра „Софтуерни технологии“, ФМИ-СУ,

с Научен ръководител: доц. д-р Александър Димов

за придобиване на образователна и научна степен „Доктор“ в професионално направление 4.6 „Информатика и компютърни науки“, докторска програма „Софтуерни технологии“ – Софтуерно инженерство

1. По процедурата

Със заповед № РД 38-283 от 09.06.2025 г. на Ректора на СУ “Климент Охридски” бях назначен в научното жури на дисертационен труд за придобиване на научна степен „доктор“ с автор маг. Симеон Стоичков Емануилов, а на първото заседание на научното жури бях избран за рецензент.

По въпросната процедура официално ми бе изпратена справка за минималните количествени показатели за придобиването на научната степен „Доктор“, съгласно Правилника на СУ “Климент Охридски” за прилагане на ЗРАСРБ и Правилника към него, както и съответните показатели за маг. Симеон Стоичков Емануилов, от които е видно, че той покрива минималните изисквания.

2. Кратки автобиографични данни

Висшето си образование Симеон Стоичков Емануилов започва в Университет по национално и световно стопанство София 2008 - 2012, бакалавър специалност Бизнес информатика,. Завършва Бакалавър 2012 с общ успеваемост много добър.

Магистърската си степен завършва 2012-2014 със специалност Информационни технологии и бизнес разузнаване с отличен успех.

2011 → 2014 година работи в Прософт , като програмист

2018 → 2021 работи в SoftUni Digital, като лектор по анализ, извличане на информация, Търсене, Индексирание, WEB скриптове, Python.

2014 → 2016 работи в NetPeak, като разработчик – скриптове и автоматизация

2016 → 2019 работи във Voxrow, като основен разработчик.

2019 → 2024 работи в Scaleflex последователно, като основен разработчик и ръководител на разработчиците.

3. Актуалност на дисертационния труд

Растежа в обема на данните, известен още като „експлозия на данните“ се дължи на технологичния напредък, разпространението на устройства, генериращи данни, и все по широкото прилагане на приложения с интензивно използване на информация, се превърна в критично предизвикателство в различни сектори. Математическото моделиране, като приложен инструмент в сферата на обработка на данни се появява още в трудовете на Бернули около 1766 г

Необходимостта от централизиран контрол на еинформационните потоци и тяхната обработка може да се счита за основен двигател за развитието на съвременната информатика използваща математическото моделиране като приложен инструмент при реализация на софтуерни системи за обработка на големи данни. Интернет на нещата (IoT) и операциите на предприятията са основните фактори, които допринасят за това нарастване на данните. Наред с това, системите за машинно самообучение (МС) функционират както като значителни потребители, така и като генератори на данни.

Визуализация чрез компютърна графика е съвкупност от методи и технически средства, използвани за създаване и обработка на графични изображения чрез компютър позволяваща творчески да се комбинират и пълноценно да се прилагат модерните средства за графична обработка при решаването на практически задачи от различни приложни области.

Настоящата дисертация представлява принос към областта на разработване на софтуернисистеми за обработка на големи данни , чрез създаване и изучаване на нови математически модели за анализ на данни, включително чрез компютърно моделиране и симулации в рамките на концепцията за Големи данни. Дисертационният труд обхваща следните теми и изследователски области: 1) Архитектурни подходи: Изследване на архитектурните стилове за посрещане на специфичните нужди на системите с интензивно използване на данни. 2) Алгоритми за индексирание и търсене: Разработване на ефикасни структури за индексирание и алгоритми за търсене на високоизмерни данни, с акцент върху търсенето на сходство в милиардни масиви от данни. 3) Техники за интерпретиране на данни: Създаване на методи за подобряване на интерпретируемостта на сложни представяния на данни, като например преобразуване на плътни вектори в лексикални форми, разбираеми за човека. 4) Представяне и съхранение на данни: Изследване на усъвършенствани модели на данни, включително бази данни, ориентирани към колони и векторни представяния, за оптимизиране на съхранението, извличането и обработката на големи по обем данни.

4. Степен на познаване състоянието на проблема и обща характеристика на труда

Обект на изследването са софтуерните системи, предназначени за интензивна обработка на данни, както и ключовите аспекти на тяхното проектиране, реализация и функциониране.

Разработват се математически модели за анализ и адресиране на неотложните предизвикателства, присъщи на системите с интензивно използване на данни, чрез изследване и разработване на решения за повишаване на ефективността, мащабируемостта и интерпретируемостта при обработката на широкомащабни информационни масиви.

Предмет на изследването е разработването, оптимизацията и оценката на иновативни методи и техники, насочени към реализацията, управлението и ефективната експлоатация на софтуерни системи, предназначени за интензивна обработка на данни.

Изследователският фокус е насочен към оптимизирането на техниките за представяне, съхранение, индексирание и извличане на данни. Специално внимание е отделено на колонно-ориентираните бази данни, векторните представяния и алгоритмите за търсене по сходство, приложими към информационни масиви от порядъка на милиарди записи.

Дисертационният труд цели да подобри и интерпретируемостта на комплексните структури от данни. На база обзорен анализ докторантът формулира целта и задачите на дисертационния труд.

Цел на настоящия дисертационен труд е да допринесе за усъвършенстването на методите за реализация на софтуерни системи, работещи с интензивни потоци от данни. Това се постига чрез разработване и оценка на нови подходи за управление и анализ на мащабни информационни колекции, идентифициране на ограниченията в съществуващите практики и предлагане на иновативни решения, насочени към повишаване на мащабируемостта, производителността и интерпретируемостта на тези системи.

За постигане на тази цел са формулирани следните пет задачи, както следва: 1. Анализ и оценка на архитектурни подходи за системи, работещи с интензивни потоци от данни; 2. Проучване и разработване на усъвършенствани техники за индексирание и търсене на високоизмерни данни в информационни масиви; 3. Изследване на методи за подобряване на интерпретируемостта на комплексни представяния в системите с интензивно използване на данни; 4. Изследване и разработване на оптимизирани модели за опериране с големи данни; 5. Провеждане на експериментални изследвания и сравнителен анализ за оценка на ефективността на предложените подходи в практически сценарии.

Дисертационният труд е в размер от 156 страници, включва 21 фигури и 19 таблици, терминологичен речник и списък на съкращенията. Библиографията му е от 152 заглавия. Публикации по темата – 8 заглавия. Дисертационният труд е структуриран в увод, шест основни глави, заключение, раздел с научни и научно-приложни приноси, библиография и приложения (пет на брой), както и декларация за оригиналност на предложението.

В увода е представена актуалността на изследвания проблем, дефинирани са обектът, предметът и целта на дисертационния труд, формулирани са основните задачи и е описана използваната методология.

В първа глава е извършен преглед на системите с интензивно използване на данни и са идентифицирани свързаните с тях предизвикателства.

Във втора глава са разгледани техники за индексирание и клъстеризация, имащи съществено значение за управлението и търсенето в големи информационни масиви.

В трета глава са разгледани проблеми при търсене по сходство в милиарден мащаб. Предложен е нов хибриден алгоритъм за индексирание.

В Четвърта глава се дискутират методи за подобряване на интерпретируемостта на плътни векторни представяния. Представена е нова софтуерна рамка, разработена за тази цел.

В пета глава са анализирани колонно-ориентираните модели на данни и тяхното приложение в контекста на системи с интензивно използване на данни.

В шеста глава са представени резултатите от проведените експериментални изследвания и сравнителни анализи, които валидират ефективността на предложените методи.

В заключението са обобщени основните научни констатации и резултати.

В „Приноси” са представени научните и научно-приложните приноси на дисертацията.

Представен е списък на научните публикации по темата на дисертацията.

Определени са насоки за бъдещи изследвания и развитие.

Не съм забелязал наличие на плагиатство.

5. Съответствие на предложената методика на изследване и поставените цел и задачи на дисертационния труд

Предложената методика на дисертационния труд съответства на основната цел и допринася за усъвършенстването на методите за реализация на софтуерни системи, работещи с интензивни потоци от данни. Това се постига чрез прилагане на комбиниран изследователски подход, включващ теоретичен анализ, проектиране и разработка на софтуерни системи, и емпирична оценка, включваща разработване и оценка на нови подходи за управление и анализ на мащабни информационни колекции, идентифициране на ограниченията в съществуващите практики и предлагане на иновативни решения, насочени към повишаване на мащабируемостта, производителността и интерпретируемостта на тези системи. Методиката включва: Обзор на съществуващата литература; Разработка на системи и библиотеки; Емпирична оценка;

Разработката има интересни обобщения и насоки за бъдещо развитие.

Избраните методи съответстват на основната цел и задачи, поставени за решаване от докторанта.

6. Характеристика на естеството и оценка на достоверността на материала, върху който се градят приносите на дисертационния труд

Методите и моделите, които се създават и използват, отговарят на целевата задача. За някои са получени емпирични числени резултати, а за други са представени концепции за решения и са дадени насоки за бъдещи изследвания.

Не съм забелязал грешки нито в конкретните, нито в концептуалните модели. Намирам също, че предложените стратегии са добре обосновани.

7. Приноси на дисертационния труд

Научно-приложни приноси

1. Анализ, систематизация, методология и оценка на архитектурните подходи за преодоляване на предизвикателствата в системите с интензивно използване на данни. (Глави първа и втора)

2. Разработване на нова методика за хибридно индексване за търсене по сходство в милиарден мащаб. (Глава трета)

3. Създаване на методологията LangVec за подобряване на интерпретацията на високоизмерни векторни представяния (Глава четири).

4. Разработване на колонно-ориентиран модел на данни, оптимизиран за системи с интензивно използване на данни (Глава пета).

Приложни приноси

5. Внедряване и разгръщане на производствена софтуерна система, включваща: а. Библиотеката LangVec Python; б. Производствена система за семантично търсене (Similarix), интегрираща подхода за хибридно индексване; в. Webhook система, ориентирана към колони, която демонстрира прилагането на модела на данните; г. Създаване и оценяване на допълнителни инструменти.

6. Разработване на рамки за оценка и практическо валидиране,

а. Оценка на производителността и интерпретируемостта на LangVec чрез сравнителни тестове на различни по размер набори от данни; б. Проучване на случай за търсене на сходство в милиарден мащаб; в. Анализ на производителността на модела на данни; г. Публикуване на набор от данни с отворен код..

Рецензентът приема така описаните приноси, като препоръчва на автора да се научи по-кратко и систематизирано да излага своите постижения.

8. Степен на личното участие на дисертанта в приносите

За личното участие на докторанта съдя по публикационната дейност на докторанта отразена в публикуваните по дисертацията материали. Докторантката убедително представя постигнатите резултати, с много добра и задълбочена аргументация, както и използва професионално графично оформление на материалите.

Характерът на изследването предполага много добра и широка подготовка в областта на методи и средства за интегриране, обработка и моделиране на

хетерогенни данни. Считаю, че докторантът се е справил успешно, като не поставям под съмнение личното му участие в разработването на дисертационния материал.

9. Преценка на публикациите по дисертационния труд

По темата на дисертационния труд са предоставени 3 публикации в реномирани международни издания. Изнесени са и 8 доклада по темата на научни сесии на ФМИ, както и на научни семинари и международни конференции.

Публикациите в реферирани издания са на английски език и в съавторство, което е приемливо, взимайки предвид мултидисциплинарния характер на изследванията.

Те са публикувани в признати международни издания. Едната е публикувана в рецензирано издание на БАН. Втората статия е публикувана на международната конференция издание на IEEE. Третата е издадена в Az-buki. Има публикации са индексирани в SCOPUS.

Рецензентът отчита че публикационната дейност по темата на дисертацията обхваща периода 2021-2024 година и е направена в реномирани международни научни издания.

Фактът, че всичките публикации са в съавторство, отразява мултидисциплинарността на разглежданата в тях проблематика.

Публикациите отразяват по-съществените резултати, постигнати в дисертационния труд. Докладвани са на достатъчно реномирани научни форуми, което приемам за апробация в научните среди.

10. Съответствие на автореферата с изискванията за изготвянето му и адекватност на отразяване на основните положения и приносите на дисертационния труд

Представеният проект за автореферат е в съответствие с правилника за изготвяне на авторефератите по дисертационните трудове, посочен в сайта на СУ “Климент Охридски” . Отразява постигнатите резултати, както и приносите на автора. Графично е оформен много добре и включва необходимата информация, описваща в резюме дисертационния труд.

11. Мнения, препоръки и бележки

В дисертационния труд се разработва една много сложна, динамично развиваща се и перспективна област – разработването, оптимизацията и оценката на иновативни методи и техники, насочени към реализацията, управлението и ефективната експлоатация на софтуерни системи, предназначени за интензивна обработка на данни.

Представена е софтуерна разработка с инструменти за анализ в реално време с интерактивни визуализации. Анализирани са методи и средства за

индексиране, визуализиране, интегриране, обработка и моделиране на хетерогенни данни, получени от изследователска дейност. Разработена е иновативна софтуерна среда, като е дадено и практическото ѝ приложение.

Това предполага достатъчно задълбочени знания, възможност за интерпретация и формулиране на стратегии за ефективно развитие на областта. Съдържателно и графично материалът е разработен много добре. Този материал представлява интерес за кръг специалисти и ако след преработка се публикува ще има мултиапликативен ефект.

Препоръчам на магистър Симеон Стоичков Емануилов да продължава активната си публикационна дейност в научни списания с импакт фактор.

Препоръчвам по-точна вербализация на неговите постижения - да се научи по-систематизирано да излага своите приноси.

Някои не съществени забележки съм отразил върху копие, което ми бе предоставено.

12. ЗАКЛЮЧЕНИЕ

Съдържанието и приносите на дисертационния труд на магистър Симеон Стоичков Емануилов напълно покрива изискванията на Закона за развитие на академичния състав на Република България, на Правилника за неговото приложение и на Правилника за условията и реда за придобиване на научни степени в СУ "Климент Охридски". Извършена е значителна по обем и съдържание изследователска работа. Има достатъчен брой научно-приложни и приложни приноси. Представени са достатъчен брой публикации по дисертацията публикувани на престижни научни форуми. Безспорно е личното участие на автора в разработката и получените приноси. Това ми дава основание убедено да препоръчам на Уважаемото Научно жури да присъди на **Симеон Стоичков Емануилов** образователна и научната степен „доктор“ в професионално направление 4.6 „Информатика и компютърни науки“, докторска програма „Софтуерни технологии“ – Софтуерно инженерство.

РЕЦЕНЗЕНТ:

/проф. д-р Р. Йошинов/

София, 02.09.2025 г.