

СТАНОВИЩЕ

по процедура за защита на дисертационен труд на тема:
„Methods for implementation of data-intensive software systems“
(*„Методи за реализация на софтуерни системи за обработка на големи данни“*)
за придобиване на
образователна и научна степен „доктор“

от

кандидат: **Симеон Стоичков Емануилов,**

Област на висше образование: **4. Природни науки, математика и информатика**

Професионално направление: **4.6. Информатика и компютърни науки**

Докторска програма: **„Софтуерни технологии“ – Софтуерно инженерство,**

катедра: **„Софтуерни технологии“,**

Факултет по математика и информатика (ФМИ),

Софийски университет „Св. Климент Охридски“ (СУ),

Становището е изготвено от: **доц. д-р Елисавета Василева Гурова, Факултет по математика и информатика, Софийски университет „Св. Кл. Охридски“**

в качеството ми на член на научното жури, съгласно Заповед № РД-38-283/ 09.06.2025 г. на Ректора на Софийския университет.

1. Обща характеристика на дисертационния труд и представените материали

Дисертационният труд е в обем 157 страници, съдържа увод, 6 глави, заключение и 5 приложения. В увода авторът е формулирал целта на дисертацията: *„да се постигне напредък в прилагането на софтуерни системи с интензивно използване на данни чрез разработване и оценяване на нови методи за управление и анализ на големи колекции от данни. Допълнително, да се идентифицират ограниченията в настоящите подходи и да се предложат нови решения, които подобряват мащабируемостта, производителността и интерпретируемостта.“*

Първите две глави са насочени към проучване и анализ на научната литература и идентифициране на тенденциите за дизайн на системи за обработка на големи обеми от данни, както и техники за индексване и клъстеризация, важни за управлението и търсенето в големи масиви от данни. На тази основа в трета глава е предложен

хибриден алгоритъм за индексирание, а в четвърта глава - техниката LangVec, предназначена за подобряване на интерпретацията на плътни вектори. В пета глава се разглеждат моделите на данни, ориентирани към колони, а в шеста глава са дадени експерименти и анализи на предложените методи.

Текстът на дисертационния труд е написан изцяло на английски език. В библиографията са включени 152 източника. Използвани са стандартни научни методи, свързани с проучване на научната литература, разработване на нови софтуерни инструменти, тестване и оценка с оглед на поставените цели.

Представени са 3 научни статии по дисертационния труд, автореферат и резюме на дисертацията, както и административни документи по процедурата.

2. Данни и лични впечатления за кандидата

Кандидатът Симеон Емануилов е докторант в катедра „Софтуерни технологии“ и редовно е представял резултатите от изследванията си пред катедрата, както и по време на Пролетната научна сесия на ФМИ.

3. Съдържателен анализ на научните и научноприложните постижения на кандидата, съдържащи се в представения дисертационен труд и публикациите към него, включени по процедурата

Представеният дисертационен труд е насочен към решаване на проблеми и предизвикателства от реалната практика и по-конкретно за подобряване на ефективността на софтуерни системи с интензивно използване на големи данни и разработване на нови методи за управление и анализ на големи данни.

Първоначално в дисертацията е представено състоянието на научните изследвания и тенденциите за дизайн на системи за обработка на големи обеми от данни. Особено внимание е обърнато на архитектурните стилове при системите с интензивно използване на данни и са анализирани техните силни и слаби страни от гледна точка на способностите за обработка на данните и гарантиране на качествените изисквания. Накрая са изложени предизвикателства и тенденции в областта на системи за обработка на големи обеми от данни като основа за следващи изследвания. Анализът на научната литература е задълбочен във втора глава, където са разгледани и анализирани техники за индексирание и клъстеризация, важни за управлението и търсенето в големи масиви от данни. Задълбочено са проучени предизвикателствата на широкомащабни системи за търсене по сходство, както и

съществуващите подходи за индексирание. От значение за дисертацията са идентифицираните ограничения в съществуващите подходи за ефективна обработка на големи масиви от данни.

Проучването на научната литература е в основата на 3 оригинални подхода, които са и *основните научно-приложни приноси на кандидата*, изложени в следващите глави: хибриден алгоритъм за индексирание, метод за подобряване на интерпретацията на плътни вектори, и модели на данни, ориентирани към колони.

- В трета глава е представен нов подход: хибриден алгоритъм за индексирание, който осигурява практически и икономически ефективно решение за търсене на сходство със сложни възможности за филтриране, независимо от идентифицираните ограничения. Предложеният алгоритъм е тестван в реални условия чрез специално проектиран експеримент за търсене на сходство в милиарден мащаб, както и е сравнен със съществуващи алгоритми за индексирание.

- Друго значително предизвикателство, свързано с интерпретирането на високомерни векторни е в основата на четвърта глава, където е предложен метод, предназначен за подобряване на интерпретируемостта на плътни вектори чрез съпоставянето им с лексикални представяния, които могат да се четат от човека. Представена е методологията LangVec и ключовите ѝ компоненти, вкл. дефиниране на лексикон, изчисляване на проценти и картографиране на вектори в думи. Обърнато е внимание на практическото приложение на LangVec и използването му в различни сценарии. Изтъкнати са предимствата на LangVec в приложения като семантично търсене, дедупликация на данни и клъстеризация.

- В пета глава изложението адресира нуждата от ефективни механизми за съхранение и извличане на данни за системите с интензивно използване на данни. Кандидатът разглежда модел на данни, ориентиран към колони, като изяснява основните принципи, съображения за проектиране, техники за оптимизация и характеристики на производителност. Представен е специфичен дизайн на колонно-ориентиран модел, използващ webhook система като илюстративен пример. За да се оцени ефективността на подхода е направен сравнителен анализ с традиционна система за бази данни, ориентирана към редовете.

Последната шеста глава представя експерименти и анализи на предложените методи.

4. Аprobация на резултатите

Представени са 3 оригинални публикации по дисертацията в съавторство с научния ръководител, две в списания, индексирани и реферирани в WoS (Q4):

1. Simeon Emanuilov, Aleksandar Dimov, Billion-scale Similarity Search Using a Hybrid Indexing Approach with Advanced Filtering, Cybernetics and Information Technologies - Bulgarian Academy of Sciences, 2024, ISSN (print):1311-9702, ISSN (online):1314-4081
2. Simeon Emanuilov, Aleksandar Dimov, Lexical Representation of Dense Numerical Vectors: Introducing LangVec, Mathematics and Informatics - Az-buki, 2024, ISSN (print):1310-2230, ISSN (online):1314-8532, doi:10.53656/math2024-3-1-lex

и една в Scopus:

3. Simeon Emanuilov, Aleksandar Dimov, Column-oriented data model for data-intensive systems, Proceedings in the International Scientific Conference "Computer Science' 2022", Publisher: IEEE Xplore Digital Library, 2022, ISSN (online):978-1-6654-9777-0, doi:10.1109/COMSCI55378.2022.9912610

Налице са 3 цитирания на първата публикация от други автори (https://scholar.google.com/scholar?oi=bibs&hl=en&cites=1761747379818227492&as_sdt=5).

Научните трудове отговарят на минималните национални изисквания за придобиване на образователна и научна степен „доктор“ в професионално направление 4.6. Информатика и компютърни науки.

5. Качества на автореферата

Авторефератът е с обем 53 страници и представя коректно резултатите и съдържанието на дисертационния труд.

6. Критични бележки и препоръки

Нямам съществени критични забележки по дисертационния труд.

Като пропуск в автореферата може да се отбележи анализа в глава 1 на архитектурните стилове при системите с интензивно използване на данни.

7. Заключение

След като се запознах с представените в процедурата дисертационен труд и придружаващите го научни трудове и въз основа на направения анализ на тяхната

значимост и съдържащи се в тях научно-приложни приноси, **потвърждавам**, че представените материали, качеството и оригиналността на представените в дисертационния труд и научните публикации резултати, отговарят на изискванията на ЗРАСРБ, Правилника за приложението му и съответния Правилник на СУ „Св. Климент Охридски“ за придобиване от кандидата на образователната и научна степен „доктор“ в научната област „Природни науки, математика и информатика“, професионално направление: „Информатика и компютърни науки“. Кандидатът удовлетворява минималните национални изисквания в професионалното направление и не е установено плагиатство в представените по конкурса научни трудове.

Въз основа на гореизложеното, **препоръчвам** на научното жури да присъди на **Симеон Стоичков Емануилов** образователна и научна степен „доктор“ в научна област: 4. „Природни науки, математика и информатика“, професионално направление: 4.6 „Информатика и компютърни науки“, докторска програма: „Софтуерни технологии“ – Софтуерно инженерство.

08.09.2025 г.

Изготвил становището:

доц. д-р Елисавета Гурова