

СТАНОВИЩЕ

по процедура за защита на дисертационен труд на тема:
„Methods for implementation of data-intensive software systems“
за придобиване на образователна и научна степен „доктор“

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, Факултет по
математика и информатика (ФМИ), Софийски университет „Св. Климент
Охридски“ (СУ)**

Становището е изготвено от: доц. д-р Вася Красимилова Атанасова от Института по биофизика и биомедицинско инженерство – БАН, секция „Биоинформатика и математическо моделиране“, в качеството ми на член на Научното жури, съгласно Заповед № РД-38-283 / 09.06.2025 г. на Ректора на Софийския университет.

Представеният дисертационен труд се състои от въведение, шест глави, заключение, библиография от 152 цитирани източника, списъци с публикациите и презентациите по дисертационния труд, и пет приложения. Като структура и съдържание трудът е прилежно организиран и демонстрира методичния и структуриран подход на докторанта към изследванията и софтуерните му разработки. Темата на труда е с безспорна актуалност и значимост в светлината на наблюдавания експоненциален растеж на приложенията за обработка и търсене по семантика и сходство в информационни масиви от порядъка на милиарди записи, а поставените цели и задачи са методологически издържани, и адекватно реализирани и документирани.

Дисертационният труд е оформен консистентно, по детайлно структуриран начин, с възприет експозиционен, но на места шаблонен, стил. В областта на информатиката и с оглед избраната тематика, не е необичайно, но се забелязва в някои раздели свръхупотреба на номерирани и неномерирани списъци и генерични текстове за преход между тях. В този смисъл, в интерес на кандидата би било в бъдещите си научни трудове да изгради свой индивидуален стил и да намери собствения си научен глас, с който да изразява приносите си в аналитична перспектива и с необходимата критичност и ангажираност.

Приемам претенцията на докторанта за приносите на дисертационния му труд, систематизирани като научно-приложни (четири) и приложни (два) в раздела „Приноси“ и в Таблица 18 със съответствията между преносите, разделите от дисертационния труд и публикациите / линковете, послужили за академично свидетелство или валидация на резултатите.

1. Извършен е задълбочен анализ и е разработена методология за оценка на архитектурни стилове като споделени данни, *sharding* и кеширане и др., които са систематизирани съпоставени по отношение на скалируемостта и производителността им с цел вземането на ефективни решения при проектиране на системи с интензивно използване на данни.

2. Разработена е нова методика за хибридно индексване за търсене по сходство в милиарден мащаб. Предложената методика комбинира възможността за ефективното извличане на данни от пространства с висока размерност с постигнатата по-ниска изчислителна „цена“ благодарение на: а) проектирания метод за интеграция на плътни вектори с филтриращи атрибути, б) усъвършенстваната плоска индексна структура за многомерно филтриране, и в) стратегията за управление на паметта, позволяваща обработка на масиви по-големи от наличната оперативна памет.
3. Реализирана е методологията LangVec за усъвършенстване на интерпретируемостта на вектори с висока размерност. Подходът включва дефиниране на лексикон, изчисляване на персентилно разпределение и разработената на тази основа нова техника за установяване на съответствия (mapping) за трансформация на векторните величини в лексикални представяния.
4. Създаден е колонно-ориентиран модел на данни, специално подобрен за нуждите на системи, обработващи данни в милиарден мащаб. Ползите от него включват повишена производителност при изпълнение на заявките, подобрена компресия на данните, по-голяма скалируемост. Този принос включва алгоритъм за автоматично изграждане на колонната структура и набор от оптимизационни техники, които ускоряват постъпването (ingestion) на данни и се изразява в увеличена ефективност на съхранение и скорост на достъп до данните и повишени възможности за обработката им в условия на високи аналитични натоварвания и милиони записи.
5. Приложен принос е разработената и внедрена на производствена софтуерна система, включваща библиотеката *LangVec Python*, системата за семантично търсене *Similarix*, колонно-ориентирана webhook система и високоскоростния инструмент за сваляне на файлове *gotake*.
6. Дефинирането на система за оценка и практическа валидация на разработените компоненти на системата също е валиден приложен принос на дисертационния труд.

Прави впечатление описаният в Глава 1 възприет от автора систематичен подход към изготвянето на литературния обзор и списъка от 152 цитирани литературни източника (тук отбелязвам разминаването с бройката, посочена на стр. 10, 151).

От тях четири са трудове в съавторство на докторанта, но това са и единствените цитирани работи на български учени. Би било добре литературният обзор да демонстрира повече приемственост с други предхождащи изследвания, правени в България (само за пример, в обзора на методи за редуциране на размерността в т. 2.6 е уместно споменаването на разработвания в България от десетина години метод интеркритериален анализ).

За отбелязване е, че около 10% от общия брой на цитираните източници (14 на брой) са дадени със статут на препринти в депозитната база ArXiv – а с такъв статут са трудове, незадължително преминали през процедури по рецензиране. Дори в динамично развиваща се област като информатиката този процент може да се интерпретира като твърде висок в контекста на подлежащ на защита дисертационен труд. По-обстойна проверка показва, че 8 от тези препринти – дори и към момента на подаване на документите за процедурата – вече са били реализирани в публикации с библиографски данни (1 статия и 7 доклада в

сборници от конференции) и е било коректно да се цитират като такива, 1 препринт е бил изнесен доклад, но неприет за публикуване в сборника от конференцията, поради получени отрицателни рецензии, и 5 действително са със статут на препринти без информация да са предлагани или публикувани в рецензирани издания. Такова редуциране на цитиранията-препринти е в полза за дисертационния труд, но е добре докторантът да има изградено ясно разбиране за разликата, както и да актуализира списъка с литературата си, в случай че в бъдеще смята отново да реферира към тези трудове.

[30]	arXiv: 2407.09107 от 2024 г.	Публикуван доклад на 2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA), 10-13.09.2024, Padova, Italy; DOI: 10.1109/ETFA61755.2024.10711136
[62]	arXiv: 2403.15927 от 2024 г.	Публикуван доклад на 2025 23rd International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 26-29.05.2025, Linköping, Sweden; DOI: 10.23919/WiOpt66569.2025.11123337
[72]	arXiv: 1802.03426 от 2018 г.	-
[73]	arXiv: 1403.2877 от 2014 г.	-
[79]	arXiv: 1705.07874 от 2017 г.	Публикуван доклад на Advances in Neural Information Processing Systems 30 (NIPS 2017), 04-09.12.2017, Long Beach, US https://neurips.cc/Conferences/2017/Schedule?showEvent=9253
[87]	arXiv: 2109.03022 от 2021 г.	-
[105]	arXiv: 2105.09613 от 2021 г.	-
[108]	arXiv: 1408.2927 от 2014 г.	-
[115]	arXiv: 2401.06251 от 2024 г.	Статия в сп. Information Fusion, Volume 122, October 2025, 103152 (Достъпна онлайн от 03.04.2025) https://doi.org/10.1016/j.inffus.2025.103152
[128]	arXiv: 1301.3781 от 2013 г.	Изнесен доклад на International Conference on Learning Representations, 2-5.05.2013, Scottsdale, Arizona (невключен в сборника с публикувани доклади) https://iclr.cc/archive/2013/program-details/program.html https://openreview.net/forum?id=idpCdOWtqXd60
[130]	arXiv: 1810.04805 от 2018 г.	Публикуван доклад на 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), 02-06.07.2019, Minneapolis, Minnesota https://aclanthology.org/N19-1423
[133]	arXiv: 1906.04341 от 2019 г.	Публикуван доклад на 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 01.08.2019, Florence, Italy https://aclanthology.org/W19-4828/
[134]	arXiv: 1805.01070 от 2018 г.	Публикуван доклад на 56th Annual Meeting of the Association for Computational Linguistics, 15-20.07.2018, Melbourne, Australia https://aclanthology.org/P18-1198/

[135]	arXiv: 1905.09418 от 2019 г.	Публикуван доклад на 57th Annual Meeting of the Association for Computational Linguistics, 28.07 – 02.08.2019, Florence, Italy https://aclanthology.org/P19-1580/
-------	--------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Без претенции за изчерпателност, ще отбележа, че прегледът на цитираните източници показва още, че се срещат и случаи на:

- непълно изброяване на съавторите, например в цитирания като препринт [130] (публикуван доклад, дори отличен с награда „Best Long Paper”), освен посоченият първи автор J. Devlin, съавтори са и още трима души, в т.ч. българската колега д-р Кристина Тутанова. Идентични са случаите и с препринт [79] (един пропуснат съавтор + статут на публикуван доклад от конференция), [128] (трима пропуснати съавтори), [133] (трима пропуснати съавтори + статут на публикуван доклад от конференция), [148] (един пропуснат съавтор).
- непълни / неточни заглавия – например от заглавието на препринт [134] (отново труд със статут на публикуван доклад на конференция), е пропусната малка част; а в цитирания източник [70]: посоченото заглавие съответства на една от главите от книгата („Principal Component Analysis for Special Types of Data“), а не заглавието на самата книга („Principal Component Analysis“), което липсва заедно със страниците.
- цитиране на „самиздат“ ръкопис вместо съответстващата му глава от публикувана впоследствие книга, какъвто е случаят с качения през февруари 2023 г. ръкопис [100] на личния сайт на автора, вместо респективната глава 8 от монография в Springer, издадена през 2024 г. (DOI: [10.1007/978-3-031-51462-3](https://doi.org/10.1007/978-3-031-51462-3)).
- неточни / непълни библиографски данни – например „IEEE” вместо “IEEE”; липсващи дати и локации на конференции в повечето цитирания на доклади; липсващи конкретни реферирани страници от цитираните книги, и др.

Като цяло, библиографията не е изпълнена съгласно АРА стандарта, въпреки твърдението за това в дисертационния труд и в авторефератите. Препоръката ми е, при успешна защита, преди предаване на документите по процедурата към НАЦИД, където ще бъдат качено копие и на дисертационния труд, колкото се може повече от тези пропуски в библиографията да бъдат отстранени, което е от интерес и на докторанта, и на цитираните автори.

Към документите по процедурата и в списъка на стр. 136 са представени три публикации по дисертационния труд:

1. *Billion-scale Similarity Search Using a Hybrid Indexing Approach with Advanced Filtering* – статия в сп. Cybernetics and Information Technologies с IF(2024) = 1.1, JIF квантил Q4; и SJR(2024) = 0.358, Scopus квантил Q2.
2. *Lexical Representation of Dense Numerical Vectors: Introducing LangVec* – статия в сп. Mathematics and Informatics с IF(2024) = 0.3, JIF квантил Q4.
3. *Column-oriented data model for data-intensive systems* – доклад на конференцията 10th International Scientific Conference on Computer Science ‘2022 – в сборник, индексирани в Scopus, без SJR / квантил.

Установява се леко разминаване, тъй като на стр. 133 от труда сред академичните свидетелства е вписана още една, четвърта статия от 2022 г., която по-долу ще бъде коментирана във връзка с цитиранията по нея.

Ще отбележа във връзка със статията “*Billion-scale...*” факта, че към момента на изготвяне на това становище, тя не се появява индексирана в *Scopus*, а само в *Web of Science*, въпреки че е публикувана преди повече от 8 месеца, а както предходните три, така и следващите два броя на списанието са коректно индексирани. Това обаче се отнася и за всички останали 10 статии от същия брой 24(4) на списанието, така че го отдавам на временен технически проблем със *Scopus*, независещ от докторанта.

И в трите публикации докторантът е първи автор и те са в съавторство само с научния ръководител, което следва да подчертае високия дял на принос и ангажираност към тези три труда. Наукометричните данни на публикациите по дисертационния труд показват съответствие с Правилника за приложение на ЗРАСРБ в СУ „Св. Климент Охридски“, като покриват и надхвърлят изискуемите точки по показател „Г“ за придобиване на ОНС „доктор“.

Също много добър атестат за работата на докторанта е и наличието (по данни от *Google Scholar* и *Web of Science*) на пет независими цитирания от чужди учени на две статии с участието на докторанта:

- статия, озаглавена “*Architectural Approaches to Overcome Challenges in the Development*” (DOI: 10.54941/ahfe1002521) с 2 цитирания: от май 2024 г. и януари 2025 г. – и двете в списания, индексирани в *Web of Science* с квартали Q3 и импакт фактори съответно 1.7 и 0.9.
- статията по дисертационния труд “*Billion-scale...*” е вече с 3 цитирания от 2025 г., представляващи:
 - статия (DOI: 10.48084/etasr.11575) в индексираното в *Scopus* сп. “*Engineering, Technology and Applied Science Research*” (SJR (2024) = 0.332, Q2) от август 2025 г.
 - препринт (arXiv: 2508.18617) в ArXiv от август 2025 г.
 - препринт (DOI: 10.13140/RG.2.2.10178.90569) в ResearchGate от май 2025 г.

И в двата препринта се посочва на кои конференции/списания са подадени и, съдейки по профила им и високата наукометрия на част от съавторите, е вероятно скоро да променят статута си от препринти на официални публикации. В този смисъл тези пет цитирания имат висока добавена стойност, тъй като представляват независимо признание и раннокариерен сигнал за видимостта, полезността и влиянието на приносите на докторанта върху научната област и общността.

Двата варианта на автореферата (на български и английски език) коректно отговарят в обем и обхват на дисертационния труд и го представят в съкратен, но въпреки това четивен и достъпен вид. Минимални разминавания между заглавията на раздели в дисертационния труд и автореферата на английски са установени в:

- заглавието на Глава 6 (“*Case study*” в труда, “*Experimental studies*” в автореферата).
- заглавието на подраздел 3.4 (“*Limitations and trade-offs*” в дисертационния труд, “*Limitations*” в автореферата на английски).

Въпреки че не е задължително изискване, утвърдена практика е номерацията на подраздели, фигури и таблици в авторефератите да следват оригиналната номерация от

дисертационните трудове. Тази практика в случая не е последвана, от което леко се затруднява проследимостта на съответствията между двата материала. Впечатляващо обаче е усилието всичките 152 източника от библиографията в дисертацията да намерят място в авторефератите – и като списък, и като цитирания в текста, благодарение на което този тип номерация е запазена консистентна.

Представеният дисертационен труд би бил в още по-голяма степен ценен и полезен, ако беше изготвен на български език, поради по-голямата му достъпност и възможността да се въведе и утвърди терминология от областта, която липсва, или се среща във вариации. От предоставения български вариант на автореферата личи способността на докторанта в голяма степен да се справи с такава задача, поради което мнението ми е, че подходящо подбрани части от труда, преведени на български, биха послужили като полезен учебен ресурс за студентите по информатика, така че едно подобно допълнително усилие в бъдеще би било препоръчително.

Така формулираните критични бележки по стила и съдържанието, както и изброените пропуски и неточности в библиографията не намаляват научните достойнства на представения дисертационния труд, а имат за цел да дадат на докторанта полезна обратна връзка относно бъдещата му научна и публикационна дейност.

Заклучение

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 и приложни приноси, **потвърждавам**, че представеният дисертационен труд и научните публикации към него, както и качеството и оригиналността на представените в тях резултати и постижения, 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“/научна степен „доктор на науките“ в научна област 4. „Природни науки, математика и информатика“, професионално направление 4.6. „Информатика и компютърни науки“. В частност, 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 уважаемото Научно 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.

Дата: 9 септември 2025 г.

Изготвил становището:

(доц. д-р Вася Атанасова, ИБФБМИ-БАН)