

РЕЦЕНЗИЯ

върху дисертационния труд на
инж. Камал Смаил Мохаммед Ал Барнзжи

със заглавие:

**„ИЗВЛИЧАНЕ НА ИНФОРМАЦИЯ ОТ ГОЛЕМИ МАСИВИ ОТ ДАННИ
С ЦЕЛ УВЕЛИЧАВАНЕ НА КОНКУРЕНТНОСПОСОБНОСТТА НА
ПРЕДПРИЯТИЯТА“**

за придобиване на образователната и научна степен „ДОКТОР“
по научна специалност „4.6. Информатика и компютърни науки
(Информатика)“

ОТ

проф. д-р Идилия Александрова Бачкова

ХТМУ, кат. „Автоматизация на производството“
бул. „Кл. Охридски“ 8, 1756 София
Тел.0888755232, e-mail: idilia@uctm.edu

1. КРАТКИ БИОГРАФИЧНИ ДАННИ И ХАРАКТЕРИСТИКА НА НАУЧНИТЕ ИНТЕРЕСИ НА КАНДИДАТА

Камал Смаил Мохаммед Ал-Барнзжи е роден на 1.01.1983 г. в Ирак. Има завършени бакалавърска и магистърска образователни степени по "Компютърни науки", като последната е придобита през 2011 г. от Технологичния Университет "Джавахарлал Неру" в Хидерабад, Индия. Бакалавърската степен е от Университета Дохук в едноименния град в Ирак и е придобита през 2006 г. От 2011 г. работи като преподавател и изследовател в Университета Рапарин в град Рапарин, Иракски Кюрдистан, като води упражнения и лекционни курсове за магистри и бакалаври към катедрите "Компютърни науки" и "Математика". В периода 2012 - 2015 г. е ръководител на катедра „Компютърни науки“ в Университета Рапарин. От декември 2015 г. е докторант в ХТМУ-София, катедра "Информатика". Докторантурата е по европейската програма Erasmus Marhaba. Научните интереси на Камал Смаил са в следните области: бази данни, анализ на големи данни (Big Data), машинно обучение (Machine Learning). Камал Смаил Мохаммед е участвал в разработката на голям брой проекти, свързани предимно с разработката на бази данни за различни области.

2. ПРЕГЛЕД НА ДИСЕРТАЦИОННИЯ ТРУД И АНАЛИЗ НА РЕЗУЛТАТИТЕ

Представният ми за рецензия дисертационен труд е написан на 155 страници и включва 5 глави, приноси, списък на цитираната и използвана литература, и списък с авторските публикации по темата на дисертацията. Представени са и списъци на фигури (58 бр.), таблици (13 бр.) и съкращения.

Дисертацията е в областта на големите масиви от данни (Big Data) и извличането на закономерности (информация и знание) от тях, една от основните технологии, на която се опира Четвъртата индустриална революция (Индустрия 4.0) и интернет на нещата. Темата на дисертационния труд е изключително актуална и значима, поради натрупаните, вследствие на развитието и използването на информационните технологии, огромни количества данни, които могат да бъдат използвани за подобряване на мениджмънта и управлението на редица разнообразни процеси и дейности в съвременния свят. Основната идея на дисертацията, е свързана със създаването на разширима, детайлна рамка за разработка и прилагане на различни алгоритми в областта на големите масиви от данни, с цел повишаване на добавената стойност, а оттам и конкурентноспособността на предприятията. Акцентът е поставен върху алгоритми, които могат да се възползват от паралелната мощ на съвременните изчислителни устройства и системи. Темата на дисертацията изисква знания, както от областта на натрупване, обработка и съхранение на големи масиви от данни и софтуерно инженерство, така също и добро познаване на методите на изкуствения интелект и машинното обучение за извличане на информация и знание.

Първа глава представлява кратко въведение в големите масиви от данни, в което също са дефинирани целта и задачите на дисертационния труд. Целта е формулирана като изграждане на детайлна рамка на хибридна система за препоръки в облак за извличане на закономерности от големи масиви от данни. Формулирани са 10 основни задачи, свързани с постигане на целта, стартирайки от анализа на съвременното състояние на изследванията в областта; предлагане на съвременна рамка и алгоритми за разработката на хибридна система за препоръки на базата на големи масиви от данни; разработката и имплементацията на нови алгоритми за извличане на бизнес интелигентност от големи масиви от данни, както и такива за съвместно филтриране за целите на генериране на препоръки върху големи масиви от данни и анализ на текст с оглед на предсказване (прогнози) върху големи данни; интеграция на всички модули и тестване на системата за различни множества от данни; експериментална оценка на всички алгоритми, използвани в рамката; заключения и препоръка за бъдещо усъвършенстване на системата.

Във **втора глава**, в обем от 24 страници е направен литературен обзор на големите масиви от данни (характеристики, класификация, процеси) и на използваните технологии, свързани с извличането на информация и знания от тях, както и анализ на ползата от обработка на големи данни за бизнеса и повишаване на неговата конкурентноспособност. Анализирани са някои системи за препоръки и предсказване, ориентирани към големи данни. Главата завършва с изводи и формулиране на целта на дисертацията, като са очертани 5 подцели.

В **Трета глава**, която е в обем от 31 страници, на базата на дефиниране на основните предизвикателства пред извличането на информация от големи данни са анализирани различни методи и рамки за извличане на закономерности (информация) от големи данни, като: „Hadoop Framework“, „Apache Mahout“, „Apache Spark“, „NoSQL databases“ и „Cloud Computing“. В тази глава също са анализирани алгоритмите за машинно обучение и са дискутирани възможностите на различни библиотеки за машинно обучение за работа с големи данни, като „Apache Mahout“, „Apache Spark“. И тази глава завършва с изводи.

Четвърта глава е в обем от 48 страници и в нея е представена една съществена част от разработката по дисертационния труд. Представена е същността на системите за препоръки, класификацията на тези системи и са анализирани различни технологии, използвани за тяхната реализация, като метода на съвместното филтриране, технологиите за филтриране на базата на съдържание, хибридно филтриране, препоръки базирани на знания и др. - като представители на традиционните техники, и такива на модерните, съвременни методи за изготвяне на препоръки, като базираните на контекст или семантично базираните методи и др. Дискутирани са различни примери на машини за препоръки, както и съществуващите рамки за тяхната разработка. Разгледани са различни мерки за сходство и за оценка на алгоритмите и

качеството на направените препоръки. Предложени и разработени са 2 метода за съвместно филтриране на базата на памет (базирани на потребители и базирани на елементи) за генериране на препоръки за голям брой структурирани данни, използвайки Apache Mahout на Hadoop Framework. Получените резултати са сравнени по отношение на качеството на генерираните препоръки с оглед на получаване на по-добри резултати чрез включване на предимствата и отстраняване на недостатъците на всеки един от методите. Главата завършва с изводи.

В **пета глава** (обем от 38 страници) е предложена архитектурата на облачно базирана хибридна система за препоръки, (Framework for Cloud Based Hybrid Recommender System) като са представени методите и алгоритмите, използвани от системата. На първо място това са алгоритмите за съвместно филтриране, базирани на памет и на модели, а след това алгоритъмът за анализ на мнение, който е свързан с обработка на естествен език и текстови анализ за извличане на знания от различни текстови материали и източници. Системата комбинира резултатите от съвместното изпълнение на различните методи и алгоритми.

В дисертацията са цитирани 107 литературни източника. Всички литературни източници са на английски език, като близо 90% от тях са публикувани през последните 5 години.

3. ОЦЕНКА НА СЪОТВЕТСТВИЕТО МЕЖДУ АВТОРЕФЕРАТА И ДИСЕРТАЦИОННИЯ ТРУД

Авторефератът е в обем от 55 страници и е изготвен в съответствие с изискванията, като адекватно отразява основните положения и приносите на дисертационния труд, съдържа целите и задачите на дисертационния труд, представя по подходящ начин предложените подходи и методи, използваните средства и получените резултати. Известно разминаване има при отразяване на приносите, които в автореферата са 8, а в дисертацията – 6, но смисълът е запазен и не са включвани нови приноси. Недостатък на автореферата е и липсата на номерация на разделите, което затруднява проследяването на съответствието с дисертационния труд.

4. ХАРАКТЕРИСТИКА И ОЦЕНКА НА ПРИНОСИТЕ В ДИСЕРТАЦИОННИЯ ТРУД

- Направените от докторанта изводи в края на всяка глава на дисертацията отразяват правилно постигнатото, но някои от резултатите не са достатъчно добре прецизирани.
- Съгласна съм с направената в края на дисертацията самооценка на приносите, с изключение на последния дефиниран принос в автореферата, който по-скоро представлява някакъв извод или обобщение.
- Приносите в дисертационния труд оценявам като научно-приложни и приложни:
 - Към научно-приложните приноси спадат:

- Анализирани са същността, предизвикателствата и проблемите, свързани с извличането на информация от големи масиви данни, както и методите за обработка и съхранение на големи данни;
 - Анализирани са някои от основните програмни средства за обработка и извличане на информация от големи данни;
 - Анализирани са алгоритмите за машинно обучение и библиотеките за машинно обучение, ориентирани към работа с големи данни;
 - Анализирани са мерки за подобие (similarity measures) и показатели за оценка (evaluation metrics), използвани за целите на системата за препоръки и оценяване на вземаните решения;
 - Предложена е софтуерна рамка на облачно базирана хибридна система за препоръки, (Framework for Cloud Based Hybrid Recommender System) за извличане на информация от големи данни;
 - Разработени са алгоритми за съвместно филтриране за генериране на препоръки от големи масиви с данни, на основата на най-популярните техники за съвместно филтриране, а именно: базирани на потребител и базирани на продукт – представители на методите, базирани на памет и ALS от тези, базирани на модел;
 - Разработени са алгоритми за анализ на мнение за текстов анализ и предсказване на тенденции от големи етикирани множества от данни. За целта са използвани и адаптирани два от най-известните алгоритми за класификация: Naïve Bayes и Support Vector Machine;
 - Разработени са два типа Хибридни Системи за Препоръки, които са нови и различни от съществуващите хибридни системи. В първия тип са използвани само алгоритми за съвместно филтриране, а системата за препоръки е базирана на каскадна хибридизация. Във втория тип са използвани алгоритми за анализ на мнение, а системата за препоръки е базирана на смесена хибридизация.
- Към приложните приноси отнасям следните приноси:
- Предложената Облачно базираната хибридна система за препоръки е реализирана на платформата Hadoop, използвайки Apache Mahout, инструмент за машинно обучение, за осигуряване на мащабируема система за ефективна обработка на големи масиви от данни.
 - Алгоритмите на предложената платформа са реализирани чрез използването на мащабируемите библиотеки за машинно обучение Apache Mahout и Apache Spark;
 - Проведени са експерименти с цел оценяване представянето на всички алгоритми и модели на предложената платформа. За оценка на работата на алгоритмите са използвани средноквадратична грешка (RMSE) и известни

критерии (крос-валидация и др.) за оценка на различните модели и системи за препоръки. Резултатите от оценката, базирани на известни комплекти от данни (с известни етиктирани данни) потвърждават, че разработените хибридни системи дават по-точни препоръки.

5. МНЕНИЕ ЗА ПУБЛИКАЦИИТЕ НА ДИСЕРТАНТА ПО ТЕМАТА НА ДИСЕРТАЦИОННИЯ ТРУД

В списъка с авторски публикации са представни 5 труда, от които 1 е публикуван в руското списание „Engineering and Automation Problems” с IF 0.133, 2 са публикувани в български специализирани списания „Автоматика и информатика” и “International virtual journal for science, techniques and innovation for the industry MTM”, 1 - в университетското списание „Science, Engineering and Education”, както и 1 доклад на международна конференция, проведена в България. Всичките 5 публикации са в съавторство с ръководителя на дисертанта, като в 4 от публикациите Камал Ал-Барнзжи е на първо място. Липсва самостоятелна публикация. Считам, че в публикациите са отразени основните постижения на дисертационния труд, като по този начин той е получил необходимата публичност сред специалистите в това професионално направление.

6. КРИТИЧНИ БЕЛЕЖКИ И КОМЕНТАРИ

Нямам критични бележки, с които да оспорвам основните научно-приложни и приложни приноси на докторанта. Част от критичните бележки и коментари по дисертационния труд, отправени в рамките на вътрешната защита на докторанта, са взети под внимание в окончателната версия на дисертационния труд. Акцентирам върху някои неотчетени и/или по-късно констатирани критични забележки:

- Към дисертационния труд не е приложена декларация за оригиналност съгласно чл.11 (2) от ППЗРАСРБ на ХТМУ.
- Обемът на обзора, представен във Втора глава и от части в Трета, Четвърта и Пета глави е твърде голям, като са забелязани и следните тематични повторения:
 - Раздел 2.7. към Втора глава съвпада тематично с раздел 2.4 на същата глава;
 - Раздел 3.2. към Трета глава е излишен или трябва да бъде отнесен към Втора глава;
 - Раздел 3.5.2.1. и 3.4.2 съвпадат тематично и разглеждат библиотеката Apache Mahout. Следва да са в един раздел;
 - Раздел 3.5.2.2. и 3.4.3 съвпадат тематично и разглеждат Apache Spark. Следва да са в един раздел;
 - Раздел 5.2 към Пета глава е излишен и съвпада тематично със съдържанието на Четвърта Глава.

- Забелязват се известни пропуски в структурирането и оформянето на дисертационния труд, както следва:
 - Излишно въвеждане на подраздели с номера 4.3.1.2.1., 4.4.1., 5.2.1., тъй като са единствени. Раздел 4.4.1 трябва да се представи като раздел 4.5, тъй като съдържанието е тематично различно от това в раздел 4.4.;
 - Не са прецизирани заглавията на Трета, Четвърта и Пета глави, както и на някои раздели, като например подраздел 2.3.3., чието заглавие съвпада с това на раздел 2.3.;
 - Лошо качество на използваните в дисертационния труд формули;
 - Считам, че заглавие в дисертационен труд под формата на въпрос, като например “Why NoSQL?” е неподходящо и трябва да се избягва;
 - Използването на декомпозиция в подраздели на 5-то и 6-то ниво затруднява проследяването на причинно-следствената връзка и при препоръчителния обем за един дисертационен труд трябва да се избягва, като например подраздели 3.4.4.1.x, 5.3.1.4.x, 4.3.1.1.1.x или 4.3.1.1.2.x и др.
 - На някои места в дисертационния труд е налице неправилно цитиране. Литературният източник е цитиран преди края на последното изречение, а трябва позицията му да е след края на параграфа, като например стр.51 – литературен източник [32] или стр.52 – литературен източник [61].
- Целите и задачите на дисертацията са отразени на две места в дисертационния труд – в раздели 1.1. и 1.2. на Първа глава (Въведението) и в раздел 2.10 на Втора глава, като има известно разминаване в дефинираните цели и задачи (подцели).
- Термините „рамка“ и „платформа“ са използвани на много места като синоними, което според мен е грешно, имайки предвид стандартизираните дефиниции на двата термина. Кое е правилното?
- Твърдението в раздел 3.4.4, че съществуват 3 типа Бази Данни е невярно или не пълно. Защо?

5. ЛИЧНИ ВПЕЧАТЛЕНИЯ ЗА ДИСЕРТАНТА

Познавам докторанта Камал Ал-Барнзжи от неговите участия с доклади на 13-тия Международен Конгрес „Машины, технологии, материали“ (MTM'16) и XXV Международен симпозиум „Управление на енергийни, индустриални и екологични системи“ проведени през 2016 и 2017 години, респективно, където представя компетентно своите изследвания. От участието му на тези научни форуми проличава неговата добра теоретична и базова подготовка, както и задълбочените му практически умения, които му помогнаха да постигне в срок целите на дисертационния труд и да представи една завършена и успешна разработка.

6. ЗАКЛЮЧЕНИЕ

Независимо от направените по-горе критични бележки, оценката ми за дисертационния труд, автореферата, научните публикации и приноси на маг. Камал Смаил Мохаммед Ал-Барнзжи е ПОЛОЖИТЕЛНА. Предвид на актуалността на проблема, степента на новост, качеството на разработка, и постигнатите в дисертационния труд научно-приложни и приложни приноси, считам, че дисертационният труд отговаря напълно на изискванията на Закона за развитие на Академичния състав на Р. България и Правилника за условията за придобиване на научни степени и заемане на академични длъжности в ХТМУ-София. Това ми дава основание, да препоръчам на членовете на почитаемото Научно жури да приемат положително дисертационния труд на маг. Камал Смаил Мохаммед Ал-Барнзжи и да му присъдят образователната и научна степен „доктор“ в област на висшето образование - 4. Природни науки, математика и информатика, професионално направление - 4.6. Информатика и компютърни науки (Информатика).

София
юни, 2018 г.

Рецензент:
/проф. д-р Идилия Бачкова/