

СТАНОВИЩЕ

от доц. д-р инж. Александър Красимиров Ефремов,
катедра "Автоматизация на непрекъснатите производства",
Технически университет – София

по дисертационен труд за придобиване на 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)**

с автор маг. Камал Смаил Мохаммед Ал-Барзнжи

Тема на дисертационния труд: **„Извличане на информация от големи масиви от данни с цел увеличаване конкурентноспособността на предприятията“**

1. Биографични данни на кандидата

Камал Ал-Барзнжи е завършил през 2006 бакалавърска степен по Компютърни науки в Дохукския университет, регион Кюрдистан, Ирак, а през 2011 г. завършва магистратура в Технологичен университет "Джаварлал Неру" – Хайдерабад, Андхра Прадеш, Индия, специалност Компютърни науки и технологии. През декември 2015 г. става редовен докторант към катедра "Информатика", Химикотехнологичен и металургичен университет – София.

През 2009 г. започва преподавателската си кариера в Колеж по инженерство, наука и право към Кояски университет. От 2011 започва да преподава в Рапарински университет. От 2012 до 2015 е ръководител на катедрата по компютърни науки в Университета Рапарин в град Рапарин в Иракски Кюрдистан.

Автор е на 12 публикации, проектирал е и е разработил няколко информационни системи в рамките на Рапаринския университет. Дейностите, свързани с решаването на нетривиални, предизвикателни задачи са намерили отражение и в научните му интереси и разработки, в т.ч. и на представения ми за становище труд. Камал Смаил има интереси в областта на работа с големи масиви от данни, в т.ч. и неструктурирани, машинното обучение (Machine Learning), както и работа с бази от данни.

2. Преглед на дисертационния труд и анализ на резултатите

Трудът засяга много актуални теми, свързани с обработването на големи масиви от данни като фокусът е както върху аналитичните подходи и методи за извличане на закономерности от данни (Data Mining) и препоръчващи системи, така и върху технологичното изпълнение на разработката и по-точно работата с облачни структури и паралелна обработка на данни. В днешно време данните нарастват с по-голяма скорост от тази до момента и тази тенденция се очаква да се запази и в бъдеще. Също така неструктурираните данни се очаква в 2020 да надхвърлят 40 ZB (за сравнение структурираните данни се очаква да нараснат до 5 ZB). В представеният ми за рецензия труд е засегната и актуалната темата за извличане на закономерности от текстови данни.

Трудът на Камал Ал-Барзнжи е насочен към използването на данни с цел вземане на решения, които да подобрят важни бизнес процеси на различни организации, с което да се увеличи тяхната конкурентноспособност. Дисертацията започва с кратко въведение и изясняване на целите и задачите на труда, което е оформено като Първа глава.

Във Втора глава е извършен литературният обзор като тук се засягат теми като Data Mining, бизнес интелигентни системи, разпределени системи и препоръчващи системи в контекста на големите данни. Също така са представени различни архитектури на вече разработени системи и решения подпомагащи различни бизнес процеси. В края на литературния обзор е конкретизирана целта на дисертационния труд и стъпките за постигането ѝ. Целта е създаване на облачно базирана разпределена софтуерна платформа способна да обработва големи обеми от данни и формираща препоръки като решението се търси в посока на хибридните препоръчващи системи. Идентифицирани са три основни характеристики на данните, които се използват в разглежданите системи и които трябва да се отчетат по време на проектирането на системата – това са: обем на данните, скорост постъпване на нови данни и тяхното разнообразие в смисъл на структурирани, неструктурирани и смесени данни.

До голяма степен следващите две глави имат обзорен характер. Трета глава е насочена към средствата, необходими за разработката. От технологична гл.т. са засегнати теми като: Hadoop, разпределени системи, концепцията map-reduce, Spark и NoSQL бази данни. От аналитична гл.т. е извършено запознаване с програмни библиотеки с реализирани методи за обучение без и с учител, с които се решават четирите задачи на Data Mining – за асоциация, клъстеризация, класификация и регресия. Също са описани и реализации на методи за формиране на оптимални препоръки. Представените библиотеки са подходящи за реализация на целта на дисертацията тъй като са съобразени със съвременните възможности на информационните системи.

Четвърта глава засяга темата за препоръчващите системи. Подробно са изследвани традиционните системи – базирани на сътрудничество и на съдържание. В контекста на тези системи са разгледани различни подходи за моделиране и съответните модели. Отделено е внимание и на хибридните препоръчващи системи, както и на различни подходи за съвместяване на работата на различни препоръчващи под-системи. В допълнение са описани концепциите на съвременни подходи за изграждане на препоръчващи системи като: базирани на контекста, базирани на семантиката, P2P, многоезични и др. След обзор на съществуващи препоръчващи системи вниманието се насочва към предизвикателствата, проблемите на системите и начини за справяне с тях. Част от проблемите, които са дискутирани са: липса на исторически данни (cold start), разреденост на данните, скалируемост при увеличаване на размерността на данните, разнообразие на предложенията, преспецифициране на извлечените закономерности от данните, слаба дискриминация на индивидуалните предпочитания (gray sheep) и недостатъчно статистически данни за сходни индивиди (black sheep). Основен момент в препоръчващите системи е мярката за близост, чийто избор може в голяма степен да измени резултата от нейната работа. Затова е отделено внимание на различни измерители на сходство. В главата са представени два алгоритъма за формиране на препоръки, единият от които е базиран на сътрудничество, а другият на съдържание. Главата завършва с предложение за архитектура за формиране на препоръки с използване на голям обем от данни.

В пета глава е описана предложената от Камал Ал-Барзнжи архитектура на облачно базираната софтуерна платформа, с която е реализирана хибридна препоръчваща система. В платформата са реализирани различни подходи за формиране на препоръки, а с комбинирането им се постигат по-точни резултати. Главата съдържа множество експерименти, описани са различни сценарии на проведените експерименти и са обсъдени допълнителни дейности, свързани с подготовката на текстовите данни и с оценката на качеството на извлечените закономерности. Отново са представени различни подходи за моделиране, този път в контекста на елементите на предложената архитектура на системата. По-точно са разгледани техники съставляващи препоръчващите системи и анализа на настроението, който е един от базовите подходи за извличане на информация от текст. Също така е обърнато внимание на готови библиотеки, с които са проведени експериментите. Главата завършва със заключение и намерения за бъдеща работа по подобряване на разработката.

3. Характеристика и оценка на приносите по дисертационния труд

Приемам предложените в дисертационният научно-приложни и приложни приноси, както следва:

1. Нарастващата размерност на данните и смесването на данни от различни системи са причините за реализация разработките по дисертацията подходящи за обработка на големи масиви от данни. Извършен е анализ и структуриране на разпространените платформи, алгоритми за машинно обучение и техни реализации в програмни библиотеки насочени към обработка и извличане на закономерности от големи масиви от данни.
2. Разработени са алгоритми за формиране на препоръки базирани на сътрудничество (Collaborative Filtering), които са ефективни за големи масиви от данни. Алгоритмите се основават на: потребителски-базирания, предметно-базирания и моделно-базирания ALS алгоритъм.
3. Разработени са алгоритми за извличане на информация от текст като обучението е с учител, а също и формиране на препоръки, които се базират на анализа на настроението и са подходящи за големи набори от класифицирани данни. В тази част от работата на докторанта се използват Наивни Бейсови модели (Naïve Bayes) и Метод на опорните вектори (SVM).
4. Разработени са две хибридни препоръчващи системи. Първата система е базирана сътрудничество, а хибридизацията е каскадна. Втората система използва анализ на настроението, а хибридизацията е смесена.
5. Предложена е цялостна архитектура и е реализирана платформата, наречена облачно базирана хибридна препоръчваща система (Cloud Based Hybrid Recommender System), включваща разработените алгоритми за извличане на закономерности от големи данни, както и е добавена функционалност за предварителна подготовка на корпуси от данни.
6. Извършени са експерименти при различни сценарии и с различни варианти за формиране на препоръки с и без исторически данни (cold start) и се анализира качеството на работа на система в отделните случаи. За оценка на качеството на работа на системите е използван коренът от средноквадратичната грешка, а за избягване на преспецифицирането на моделите е използван метод за крос-валидация. От извършения анализ и направените изводи се заключава, че се подобрява точността на препоръките, формирани от разработените системи.

7. В предложената платформа са решени проблемите на препоръчващите системи свързани с: отсъствието на исторически данни, скалируемостта на решението и обработката на разредени масиви от данни, както е типичният случай на данните използвани от този тип системи. За повишаване на точността на препоръките са комбинирани различни подходи.

8. Приносен момент с приложен характер намирам и в реализираната софтуерна платформа за формиране на препоръки чрез използване на библиотеките за машинно обучение Mahout на Apache Hadoop и Spark ML на Apache Spark, като и реализациите на алгоритмите в платформата на езика Java.

4. Оценка на съответствието между автореферата и дисертационния труд

В автореферата са представени всички значими резултати по дисертационния труд. Съдържанието и структура му съответстват на изложението в дисертацията, представени са приносите и в съкратен вид са описани свързаните с тях основни моменти от извършената работа.

5. Мнение за публикациите на дисертанта по темата на дисертационния труд

Петте публикации на докторанта са пряко насочени към целта на труда и обхващат важна част от приносите. Броят им отговаря на изискванията на Правилника за придобиване на научни степени и заемане на академични длъжности на ХТМУ.

6. Критични бележки и коментари

За мен беше удоволствие да се запозная с работата на колегата Камал Ал-Барзнжи. Нямам съществени бележки към представеният ми за становище труд.

7. Лични впечатления за дисертанта

Запознах се с Камал Ал-Барзнжи по време на финализирането на работа му по дисертационния труд. Много ме впечатли неговото вдъхновение, откритост и позитивизъм. Още от личния ми разговор с него добих представа за задълбочените му познания и натрупания опит в работата с неструктурирани данни. Пожелавам на колегата Ал-Барзнжи много професионални успехи в избраната от него бурно развиваща се област.

8. Заключение

Извършената от кандидата работа по дисертацията представя неговите задълбочени познания в избраната област, както и полезните умения, които тепърва ще са търсени от много компании разполагащи с голям обем от данни.

Трудът напълно отговаря на изискванията на Закона за развитие на академичния състав в Република България, Правилника за неговото приложение и Правилника за условията и реда за придобиване на научни степени и заемане на академични длъжности в ХТМУ – София.

С оглед на изложеното до момента давам положителна оценка на дисертационния труд на докторанта Камал Смаил Мохаммед Ал-Барзнжи и предлагам на уважаемото Научно жури да му присъди 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).

07.06.2018,
София

Член на журито:

доц. д-р инж. Александър Ефремов,
ТУ – София