

СТАНОВИЩЕ

по дисертационен труд на тема
„Извличане на информация от големи масиви от данни с цел увеличаване
конкурентоспособността на предприятията”

за присъждане на образователната и научна степен „доктор”
по професионално направление 4.6. Информатика и компютърни науки
с автор маг. Камал Смаил Мохаммед Ал-Барзнжи

от доц. д-р инж. Георги Томов Ружеков, Факултет Автоматика,
Технически университет - София

1. Кратки биографични данни и характеристика на научните интереси

Камал Смаил Ал-Барзнжи притежава магистърска и бакалавърска степени по "Компютърни науки" които е придобил през 2011 и 2006 г. в университети в Индия и Ирак. След дипломирането си той работи като преподавател и ръководител на катедра по "Компютърни науки" и зам-ръководител на катедрата по "Математика и компютри" в университета Рапарин в Ирак. След спечелване на стипендия по програмата Erasmus е зачислен като редовен докторант в катедра "Информатика", ХТМУ София. Научните му интереси са области, като математиката, информатиката, обработката на големи данни (Big Data) и на извличането на знания от тях на базата на методите на изкуствения интелект.

2. Преглед на дисертационния труд и анализ на резултатите

Дисертацията третира важни и актуални проблеми от областта на едно бързо развиващо се научно направление, каквото е извличането на знания и информация от големи обеми данни. Използването на тези знания за формиране на препоръки и прогнози може да, „добави стойност” в дейността на фирмите или предприятията като по този начин ги позиционират по добре спрямо конкуренцията. В съвременния дигитален свят всеки ден се произвеждат огромни количества данни (Big Data) с най-различен формат (структуриран или неструктуриран). Съхраняването, обработката и извличането на знания от тях е огромно предизвикателство и множество университети, научни институти и софтуерни фирми и са концентрирали усилията си за решаване на проблемите в тази област. Разработката на алгоритми за препоръки на основата на големите данни е във фокуса на дисертационния труд и е опит за решаването на някои от изброените проблеми.

Дисертацията се състои от пет глави.

Първа глава е въведение в проблемите и предизвикателствата при обработката на големите данни и кратко описание на целите на дисертационния труд.

Втора глава представлява литературен обзор върху големите данни, техните характеристики, класификация и процеси свързани с обработката им, анализът и интерпретацията и извличането на информация от тях. Описани са основните източници и потребители на големите данни, и как големите данни (добавят стойност) подобряват дейностите на предприятията, чрез вземане на информирани решения. В главата са анализирани водещите разпределени системи и платформите за съхранение, анализ и извличане на информация от големи данни, като Hadoop, HDFS, MapReduce, Hive, HBase, Pig, Mahout, Sqoop, Oozie, Cassandra, както и базираните на тях комерсиални такива. Разгледани са редица системи за препоръки, като са анализирани техните предимства и недостатъци.

Главата завършва с формулиране на основната цел на дисертацията – разработка на облачно-базирана хибридна система за препоръки, използваща базирани на памет и на предмети съвместни филтриращи алгоритми за предсказване, както и на хибридни алгоритми, използващи методите на анализа на мнения (Sentiment Analysis).

В тази глава, както и в останалата част на дисертацията, са цитирани 107 литературни източника, като над 90 са от последни 5 години, а останалите са от последните 10 години. Всички източници са на английски език.

В трета глава е направен анализ на техниките на извличане на знания от големите данни и на техниките на машинното обучение. Анализирани са възможностите на екосистемата Hadoop и нейните основни подсистеми HDFS и MapReduce за обработка на големи данни. Анализирани са също стандартните-релационни бази данни и NoSQL базите от данни от гледна точка на съхранението на структурирани и неструктурирани сурови и обработени данни. Разгледани са особеностите на облачно-базираните системи и изчисления с тях. Особено внимание е отделено на алгоритмите и библиотеките за машинно обучение Apache Mahout и Spark, които са в основата на разработваните в дисертационни труд алгоритми и системи за препоръки. По същество анализите и респективно приносите в тази глава имат научно-приложен характер.

В четвърта глава е направен детайлен анализ на съществуващите традиционни и модерни системи и техники за препоръки. Детайлно са разгледани предимствата и недостатъци на съвместното филтриране (базирано на памет и на модел и техните особености), филтрирането, базирано на съдържание, както и различните техники за хибридното филтриране. Представени са предизвикателствата и проблемите пред изброените техники на филтрирани, като студен старт, разреденост на матрицата на рейтингите, мащабируемост и др. Разгледани са възможностите на множество системи за препоръки и са коментирани техните положителни и отрицателни страни. Анализирани са метриките за подобие/близост, известни от системите с изкуствен интелект, които могат да се използват за откриване на близост между потребители или търсени от тях предмети. На основата на този анализ са подбрани две най-подходящи метрики, които са използвани в разработените в главата алгоритми за съвместно филтриране. Формулирани са и метрики за оценка на разработените алгоритми (RMSE, Precision, Recall, и F1-Score). Като краен резултат в тази глава е предложена архитектура за система за препоръки комбинираща техниките на два алгоритъма - потребителски и предметно базирани, която дава по-добри резултати от системите, базирани само на един от тези алгоритми. Алгоритмите са тествани с използването на две различни метрики за близост на основата на известен набор от данни (movie lens data set) с известни резултати от препоръки за конкретен потребител. Новите комбинирани алгоритми дават по-добри резултати, като оценката е на основата на формулираните в главата 4 метрики. Като научно-приложен принос тук може да се посочат тези алгоритми и ползваните от тях метрики, а като приложен принос разработката на свързания с тях софтуер.

Петта глава обединява разработените в предходната глава алгоритми в цялостна облачно-базирана хибридна система за препоръки за големи данни. Първо се използва нов алгоритъм за съвместно филтриране на големи данни за формиране на препоръки. За тази цел се използват главно релационни данни. След това се използва нов алгоритъм за анализ на мнение за формиране на препоръки от големи данни. За тази цел се използват текстови данни. Първият алгоритъм може да работи с исторически данни на дадено предприятие, които са под формата на структурирани таблици. Вторият алгоритъм може да използва данни от социални

медии, свързани с предприятието и неговите продукти. Накрая, гореописаната платформа комбинира резултатите от алгоритмите, и прилага две категории хибридни препоръчители. Комбинацията на резултатите от два алгоритъма е по-полезна за бизнес интелигентността и анализа. Тази комбинация е нова, неизследвана територия и е обект на изследване в настоящата дисертация. Алгоритмите за съвместно филтриране изграждат модел на базата на поведението на потребителя в миналото (закупени или избрани продукти и/или числен рейтинг, даден на тези продукти), а също така подобни решения взети от други потребители. След това моделът се използва, за да предскаже продуктите (или рейтинга) на продуктите, от които потребителят може да се интересува. Понеже методите за съвместно филтриране имат проблеми със студения старт, към мащабируемостта и разредеността е приложена каскадна хибридизация, при която един препоръчител подобрява препоръките. Като първо се използва алгоритъм за съвместно филтриране базиран на памет (предметно базиран) и второ алгоритъм за съвместно филтриране базиран на модел (алтернативна регресия на най-малките квадрати /ALS/). Направен е сравнителен анализ на алгоритмите поотделно, както и на хибридният, като е доказано, че последният дава по-добри препоръки.

В главата са разработени и два класификационни алгоритъма за анализ на мнение, с цел формиране на препоръки за големи масиви от данни, базирана на смесена хибридизация. Те използват методите на изкуствения интелект за анализ и извличане на позитивни или отрицателни мнения от естествен текст (Naïve Bayes и линеен SVM). Хибридните алгоритми са тествани като са ползвани две множества от данни (от Amazon и Yelp) с етикети, като всяко едно от тях е разделено на пет части по произволен начин и са ползвани 5 комбинации от по 80% данни за обучение и 20% данни за тестване (предсказване), като най-накрая резултатите са осреднени.

Предложената архитектура на хибридната система за препоръки, както и на хибридните алгоритми за съвместно филтриране и анали на мнение са с безспорни научно-приложни приноси, а тяхната софтуерна имплементация с приложен характер.

3. Оценка на съответствието между автореферата и дисертационния труд

Авторефератът съответства по съдържание и структура на дисертационния труд и е в обем от 55 страници. Той включва: пет глави, от които първата е въведение; в останалите четири глави в синтезиран вид са представени значимите научно-приложни и приложни постижения в дисертационния труд, като всяка глава завършва с изводи; приноси и списък с публикациите на докторанта.

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

Камал Ал-Барзнжи е представил пет публикации по дисертационния си труд. Първата публикация със заглавие "Mapreduce Solution For Handling Large Data Efficiently" е представена като доклад на 13-тия международен конгрес "Machines, Technologies, Materials" във Варна през 2016г. и е отпечатана в пълен текст в "International virtual journal for science, techniques and innovation for the industry MTM", брой 12 /2016. Втората със заглавие "Review Of Big Data And Big Data Mining For Adding Big Value To Enterprises" е отпечатана в брой 2 от 2017г. на списание "Science, Engineering & Education" на XTMU_София. Третата публикация "A Framework for Cloud based Hybrid Recommender System (FCHRS) for Big Data Mining" е представена като доклад на XXV-тия Международен симпозиум управление на енергийни, индустриални и екологични системи" през 2017г. в Банкя и е отпечатана в пълен текст в материалите на симпозиума. Четвъртата публикация "Особености и предизвикателства при

извличането на големи данни" е приета за печат в списание Автоматика и Информатика, брой 3/2017, който ще бъде отпечатан тази година. Последната публикация "Comparison of memory based filtering techniques for generating recommendations on large data" е с импакт фактор и е отпечатана в Journal "Engineering and Automation Problems", брой 1.от 2018г., Moskva, Russia.

Всичките пет публикации на Камал Ал-Барзнжи са в съавторство с научния му ръководител, като в четири от тях той е първи съавтор. Броят на публикациите отговаря на изискванията на Правилника за придобиване на научни степени и заемане на академични длъжности на ХТМУ. Публикациите му отразяват основни научни резултати и постижения от дисертационния му труд .

5. Критични бележки и препоръки

5.1. Глава 4 представлява едновременно анализ на съществуващи системи и заедно с това са разработени алгоритми, тествани са с различни метрики. Смятам, че би било по-добре частта за синтез да е в следващата глава.

5.2. Една от основните характеристики на алгоритмите, с които се обработват Големите данни е скоростта. Направена ли е някаква оценка на бързодействието на алгоритмите и сравнение с други известни?

5.3. Алгоритмите са тествани с етикирани данни. Възможно ли е тези алгоритми да се тестват с реални данни и да се сравни получения резултат с други системи?

5.4. Възможно ли е тези алгоритми да бъдат разпаралелени?

5.5. Може ли да се направи оценка на необходимите компютърни ресурси в зависимост от количеството на данните и може ли да се сравни с други алгоритми?

6. Лични впечатления за дисертанта

Не познавам лично Камал Ал-Барзнжи, но от прочита и анализа на дисертационния му труд и от биографичните и данни оставам с впечатление, че той е изграден научен работник с отлични познания както в областта на информатиката, анализа на големи данни изкуствения интелект и разпределените изчисления.

7. Заключение

Постиганията на докторанта показват, че той притежава отлична научна подготовка в областта на информационните технологии, големите данни и машинното обучение. Представеният дисертационен труд напълно отговаря на изискванията в Закона за развитие на академичния състав в Република България, Правилника на неговото прилагане и Правилника за придобиване на научни степени и заемане на академични длъжности на ХТМУ. Въз основа на всичко това давам положителна оценка на дисертационния труд на Камал Смаил Мохаммед Ал-Барзнжи и препоръчвам на уважаемото научно жури да му бъде присъдена **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).

Член на научното жури:.....

/доц. д-р инж. Г. Ружеков /

7 юни 2018 г.

София