

СТАНОВИЩЕ

от доц. д-р инж. Атанас Велков Атанасов,
катедра "Информатика" при ХТМУ

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

с автор маг. Камал Смаил Мохаммед Ал-Барзнжи

за присъждане 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)

Кратки биографични данни и характеристика на научните интереси

Камал Смаил Мохаммед е завършил през 2011г. магистратура по "Компютърни науки" в Университета "Джавахарлал Неру" в Хидерабад, Индия, а в периода 2002-2006г. е следвал и завършил бакалавърска степен по "Компютърни науки" в Университета Дохук в едноименния град в Ирак. От 2012 до 2015г. е ръководител на катедрата по компютърни науки в Университета Рапарин в град Рапарин в Иракски Кюрдистан. От декември 2015г. е докторант в катедра "Информатика" в ХТМУ-София. Докторантурата му е спонсорирана от европейската програма Erasmus Marhaba. Като преподавател в Университета Рапарин той води упражнения и лекционни курсове за магистри и бакалаври в катедра "Компютърни науки" и в катедра "Математика". Научните интереси на Камал Смаил са в областта на базите от данни, обработката на големи данни (Big Data) и машинното обучение (Machine Learning). Камал Смаил Мохаммед е участвал в разработката на системите за информационно осигуряване на студентската, библиотечната, медицинската и стопанската дейности в университета Рапарин.

Преглед на дисертационния труд и анализ на резултатите

Дисертационният труд на Камал Смаил Мохаммед е свързан с извличането на информация и знания от големи масиви от данни (Big Data) с цел подпомагане на различни субекти (предприятия, фирми, организации, здравни и осигурителни фондове и др.) при вземане на решения, които да подобрят бизнеса им и да увеличат тяхната конкурентноспособност. Дисертационният труд е свързан от една страна с алгоритмите за натрупването, обработката и съхранението на големи масиви от данни, а от друга страна с разработката на алгоритми за извличане на знания от тях и формиране на препоръки, които използват методите на изкуствения интелект и машинното обучение (Machine Learning).

В съвременния свят ежеминутно се генерират огромни потоци от данни от най-различни източници (сензори, електронна поща, електронна търговия, социални мрежи и др.). Извличането на полезна информация и знания от такива масиви от данни е от особено значение за фирмите и организациите за подобряване на тяхната дейност, конкурентноспособност, както и за формиране на стратегии за развитие или за формиране на препоръки към клиентите, вземане на решения и др. За осъществяването на това са необходими нови софтуерни системи и алгоритми,

адекватни на скоростта на генериране на такива данни, техния обем, структура и разнообразен характер.

Целта на дисертацията е разработката на цялостна софтуерна платформа за извличане на информация и знания от големи масиви от данни и формиране на препоръки, които да са от полза за работата на фирмите и организациите.

Дисертацията се състои от пет глави.

Първа глава представлява кратък увод в големите данни и формулиране на основните задачи пред дисертационния труд.

Във втора глава е направен литературен обзор на характеристиките на големите данни, извличането на информация и знания от тях и анализ как големите данни могат да помогнат на компаниите и бизнеса за увеличаване на тяхната конкурентноспособност. Анализирани са множество системи за препоръки и предсказване, ориентирани към големи данни. Главата завършва с изводи и формулиране на целта на дисертацията.

В трета глава са дискутирани въпросите и предизвикателствата, свързани с извличането на информация от големите данни, такива като хетерогенност, мащабируемост, скорост, точност, достоверност, поверителност, и инструктивна стойност. В тази глава е анализирана работата на алгоритмите за машинно обучение и са дискутирани възможностите на различни библиотеки за машинно обучение за работа с големи данни.

В четвърта глава са разгледани различни технологии, използвани в системите за препоръки и са разяснени техните силни и слаби страни, както и предизвикателствата, свързани с хибридните системи, създадени за повишаване качеството на препоръките. Дискутирани са различни алгоритми за обучение. Разгледани са метрики за оценка на алгоритмите и качеството на направените препоръки. В главата са представени два нови алгоритъма за съвместно филтриране. Те са базирани на потребител и базирани на продукт и служат за формиране на препоръки на основата на големи количества структурирани данни. Алгоритмите са реализирани на Java и работят в средата на платформата Hadoop и ползват библиотеката Apache Mahout. В главата е направено сравнение между различните техники за определяне на качеството на формираните препоръки с цел подобряване на работата на системите чрез включване на предимствата и отстраняване на недостатъците на всяка една от техниките.

В пета глава е предложена архитектура на софтуерна платформа за облачно базирана хибридна система за препоръки, (Framework for Cloud Based Hybrid Recommender System) и са дискутирани методите и алгоритмите, използвани в системата. Първо са използвани алгоритми за съвместно филтриране, а след това алгоритъм за анализ на мнение, който е свързан с обработка на естествен (натурален) език и текстов анализ за извличане на знания от различни текстови материали и източници. В предложената платформа са комбинирани различни подходи за формиране на две категории от хибридните системи за препоръки. Комбинацията на резултатите от два алгоритъма, по начина предложен в главата, дава по-точни предсказания. Тази комбинация на методи е нова и неизследвана област и може да помогне за повишаване на конкурентноспособността на компании и предприятия. Направена е експериментална оценка на представянето на всички алгоритми (методи) на платформата върху избрани множества от данни.

В дисертацията са цитирани 107 литературни източника, като 94 от тях са от последните 5 години, 12 са от периода от 2008 до 2013 година, а останалите са от преди 2008 година. Всички източници са на латиница.

Характер и оценка на приносите на дисертационния труд

Приемам предложените в дисертационният научно-приложни и приложни

приноси, които бих обобщил:

1. Направен е задълбочен критичен анализ на съществуващите системи и алгоритми за обработка на големи данни и за извличане на информация и знания от тях, които да са от полза за бизнеса на предприятията. Анализирани са възможностите на известните платформи за обработка и извличане на информация от големите данни, както и на алгоритмите за машинно обучение и библиотеките за машинно обучение, ориентирани към работа с големи данни. На основата на анализа са формирани критерии и изисквания към хибридната система за препоръки, обект на дисертационния труд.

2. Разработени са нови алгоритми за съвместно филтриране (Collaborative Filtering) за генериране на препоръки на основата на големи масиви от данни. При дизайна им са използвани най-популярните съвместни моделно-базирани алгоритми, като потребителски- базирания, предметно-базирания, както и моделно-базирания ALS алгоритъм.

3. Проектирани са и са разработени алгоритми за текстов анализ и за генериране на прогнози на основата на големи набори от етикетирани данни, използващи анализа на мненията. Те са базирани на известни алгоритми за класификация, като Naïve Bayes и SVM.

4. Разработени са два нови типа хибридни системи за препоръки, които са 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. .

5. Разработените по-горе нови алгоритми за извличане на знания от големи данни са обединени в софтуерна платформа, наречена облачно базирана хибридна система за препоръки (Cloud Based Hybrid Recommender System) за големи данни. Алгоритмите на предложената платформа са реализирани чрез използването на мащабируемите библиотеки за машинно обучение Apache Mahout и Spark.

6. 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 потвърждават, че разработените хибридни системи дават по-точни препоръки.

7. В разработената софтуерна платформа са решени няколко от най-важните предизвикателства при системите за препоръки, а именно проблемите със студения старт (cold start), проблемът с разреденост (sparsity) на матрицата на рейтингите и проблемът с мащабируемостта (scalability). За по-добри резултати са комбинирани различни подходи. Комбинирането на резултатите на два от алгоритмите по начина, предложен в дисертацията, подобрява точността на крайните резултати. Предложената комбинация е новаторска и не е изследвана до сега. По-точните резултати са предпоставка за повишаване конкурентноспособността на предприятията.

Оценка на съответствието между автореферата и дисертационния труд

Представеният автореферат е от 55 страници и включва увод, пет глави изводи и приноси, както и публикации по дисертацията. По съдържание и структура авторефератът напълно съответства на дисертационния труд и отговаря на изискванията на ППНСЗД на ХТМУ. В него са представени всички значими резултати и приноси в дисертационния труд.

Мнение за публикациите на дисертанта по темата на дисертационния труд

По дисертационния труд са представени пет публикации, които отразяват научни резултати и приноси, свързани с отделните му глави.

Последната публикация от януари 2018 е отпечатана в списание "Engineering and Automation Problems", Moskva, Russia и има импакт фактор (IF 0.133).

Една публикация е отпечатана в журнала на ХТМУ-София "Science, Engineering & Education", том 2. брой 1 от 2017г.

Друга публикация е свързана с доклад на международния симпозиум „Управление на енергийни, индустриални и екологични системи“ през 2017 г. в Баня. Докладът е отпечатан в пълен текст в материалите на конференцията.

Една публикация е приета за печат в списание "Автоматика и Информатика" в брой 3 от 2017г., който ще бъде публикуван през ноември 2018г.

Петата публикация е представена като доклад на 13-тия Международен Конгрес "Machines, Technologies, Materials" през 09.2016 във Варна и е отпечатана в пълен текст във виртуалния журнал "International virtual journal for science, techniques and innovation for the industry MTM", Issue 12/2016.

Представените публикации на докторанта са в съавторство с доц. А. Атанасов, като Камал Смаил е първи съавтор в четири от тях.

Като брой публикациите надвишават количествените параметри, заложи в правилника на ХТМУ.

Лични впечатления за дисертанта

Познавам Камал Ал-Барзнжи от постъпването му като редовен докторант в катедрата. Като ръководител на дисертацията му мога да потвърдя неговия изследователски капацитет при решаване на сложни научни проблеми. Бих отбелязал още, че Камал Ал-Барзнжи е много организиран и дисциплиниран и се справи с изпитите по индивидуалния си план и задачите по дисертацията си за период от две години и два месеца, а не три, както е за всички редовни докторанти. Стипендията му беше за срок от 30 месеца, от които трябва да се изключат времето за признаването на дипломата му за висше образование, както и времето от вътрешната до официалната защита на дисертационния труд. Това донякъде оправдава някои забележки по отношение на оформлението на дисертационния му труд.

Заклучение

Крайната ми оценка относно съдържанието и приносите на дисертационния труд на маг. **Камал Смаил Мохаммед** е **положителна**. Докторантът е извършил значителна по обем и съдържание изследователска работа. Дисертационният труд напълно отговаря на изискванията на Закона за развитие на академичния състав на Република България и на Правилника за придобиване на научни степени и заемане на академични длъжности на ХТМУ-София.

Гореизложеното ми дава основание да препоръчам на Уважаемото Научно жури **да присъди на Камал Смаил Мохаммед Ал-Барзнжи 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. А. Атанасов /

04 юни 2018 г.
София