Наставно-научном већу

Математичког факултета

Универзитета у Београду

Одлуком Наставно-научног већа Математичког факултета Универзитета у Београду донетом на 363. седници одржаној 28.06.2019. године именовани смо за чланове комисије за преглед и оцену докторске дисертације „Мултимедијалне базе података у управљању нематеријалним културним наслеђем" кандидата Иване Танасијевић, дипломираног математичара. После прегледања поднетог рукописа подносимо следећи

ИЗВЕШТАЈ

БИОГРАФСКИ ПОДАЦИ

Ивана (Данило) Танасијевић рођена је 24.03.1983. у Крагујевцу, где је завршила основну школу и гимназију. Основне студије студијског програма Рачунарство и информатика на Математиком факултету у Београду завршила је 2008. године са просечном оценом 9,61. Докторске студије студијског програма Информатика уписала је 2009. године на Математичком факултету Универзитета у Београду и положила сви испите предвиђене планом и програмом са просечном оценом 10,00.

Ивана Танасијевић је од 2009. године запослена на Математичком факултету Универзитета у Београду као асистент. До сада је изводила вежбе из низа предмета на основним и мастер студијама:

- Пројектовање база података
- Увод у оперативне системе и рачунарске мреже
- Теорија оперативних система
- Увод у архитектуру рачунара
- Архитектура рачунара
- Програмирање 1 (програмски језик Ц)
- Програмирање 2 (програмски језик Ц)

Област научног интересовања је рачунарска обрада природног језика и културног наслеђа и базе података. Учесник је неколико домаћих и међународних истраживачких пројеката из ових области. Од 2010. године учествовала је на пројекту "Инфраструктура за електронски подржано учење у Србији" Министарства просвете, науке и технолошког развоја Србија, као и на билатералном пројекту Балканолошког института САНУ и Института за славистику Хумболт универзитета у Берлину "Towards a Social Construction Grammar: A New Approach to Construction Theory and Cross-Cultural Narrative Analysis" и FP7 пројекта "Central and South - East European Resources" (CESAR).

У току докторских студија учествовала је на више летњих школа и радионица у земљи и иностранству ("Third workshop: Coprus annotation", Балканолошки институт, САНУ, Београд, 2016, "Second Workshop", Институт за славистику, Хумболт универзитет, Берлин, 2015, "Workshop: Methodology of Transcription, Corpus Planning and Annotation", Балканолошки институт, САНУ, Београд, 2015, Летња школа "Технологија обраде природних језика у циљу подршке индустријским потребама", Универзитет Babeș-Bolyai, Клуж-Напока, Румунија, 2011. године).

ОБЈАВЉЕНИ НАУЧНИ РАДОВИ И САОПШТЕЊА СА НАУЧНИХ СКУПОВА

[1] Јелена Граовац, Миљана Младеновић, Ивана Танасијевић, "NgramSPD: Exploring Optimal N-gram Model for Sentiment Polarity Detection in Different Languages", Intelligent Data Analysis 23(2), 2019 (IF 0.691)

[2] Ивана Танасијевић, Гордана Павловић-Лажетић, "HerCulB: Content-based Information Extraction and Retrieval for Cultural Heritage of the Balkans", прихваћен за објављивање у The Electronic Library, Emerald Publishing, DOI (10.1108/EL-03-2020-0052), ISSN: 0264-0473 (IF 0.954)

[3] Ивана Танасијевић, "Toward automatic tagging of cultural heritage documents", Special issue - "ICT Research at the University of Belgrade and at its Foreign Guests", IPSI BgD Transactions on Advanced Research (TAR), Volume 15, Number 1, ISSN 1820 - 4511, 2019

[4] Ивана Танасијевић, Биљана Сикимић, Гордана Павловић-Лажетић, "Multimedia Database of the Cultural Heritage of the Balkans", Language Resourses and Evaluation Conference (LREC), European Language Resources Association (ELRA), Istanbul, 2012, ISBN 978-2-9517408-7-7, pp 2874-2881

[5] Ивана Танасијевић, Гордана Павловић-Лажетић, "Cultural Heritage Information Retrieval by Metadata", in Natural Language Processing for Serbian - Resources and Applications, ISBN 978-86-7589-088-1, pp 87-98, 2014

Саопштења на домаћим научним скуповима

[6] Ивана Танасијевић, Биљана Сикимић, Гордана Павловић-Лажетић, "Мултимедијална база нематеријалног културног наслеђа Балкана", Предавање по позиву на скупу "Информатика 2013 - Нови трендови у развоју информационих система", Друштво за информатику Србије, стр. 16-20, 2013

[7] Ивана Танасијевић, Биљана Сикимић, Сташа Вујучић-Станковић, "Дигитализација и организовање нематеријалног културног наслеђа Балкана", конференција "Нове технологије и стандарди, дигитализација националне баштине" (НЦД), Београд, 2011

[8] Ивана Танасијевић, "Дигитална туристичка мапа Београда", конференција "Нове технологије и стандарди, дигитализација националне баштине" (НЦД), Београд, 2011

ПРЕДМЕТ ДИСЕРТАЦИЈЕ

Предмет истраживања ове докторске дисертације је развој нових метода за решавање проблема управљања нематеријалим културним наслеђем.

Истраживање проведено у оквиру ове дисертације мотивисано је мултимедијалном колекцијом која је, у тренутку преузимања грађе, представљала резултат петнаестогодишњег теренског истраживања које су извели истраживачи Балканолошког института САНУ. Истраживачи су интервјуисали локално становништво на различитим локацијама у подручју Балкана. Примарни циљ њиховог истраживања био је очување информација о различитим типовима говора који се користе на овим просторима као и проучавање њихових језичких карактеристика. Колекција се састоји од аудио и видео материјала, фотографија, рукописа и текстуалних протокола. За већину интервјуа написани су текстуални протоколи у виду слободног, неструктуираног или полуструктуираног текста који имају за циљ да опишу одређени аудио или видео материјал.

Задаци који се решавају у оквиру ове дисертације су развој адекватног дизајна и имплементације мултимедијалне базе података нематеријалног културног наслеђа која би

одговарала потребама различитих корисника, аутоматска семантичка анотација протокола методама обраде природног језика као основа за полу-аутоматску анотацију мултимедијалне колекције и успешну тематску претрагу и претрагу по метаподацима који су у складу са CIDOC CRM стандардом. Докторска дисертација се бави и повезивањем садржаја базе са геолокацијским информацијама на мапи као и истраживањем додатних могућности претраге ове колекције у циљу добијања нових знања.

Дизајн и имплементација мултимедијалне базе уз подршку пуне текстуалне и просторне функционалности спроводи се у оквиру PostGIS и eXist база података као методолошки погодних алата за рад са просторним и XML обележеним текстуалним подацима.

За полу-аутоматску анотацију мултимедијалних материјала коришћена је аутоматска семантичка анотација протокола који су придружени материјалима. Она је спроведена над информатичким моделом документа методама екстракције информација, препознавања именованих ентитета и екстракције тема, техникама заснованим на правилима уз помоћ додатних ресурса попут електронских речника, тезауруса и речника речи из специфичног домена. За класификацију текстуалних протокола у односу на тематику, изведено је истраживање о методама које се могу применити за решавање проблема класификације текстова на српском језику, и понуђена је метода (модификована CBM метода са комбинованим семантичким и позиционим својствима и различитим опцијама за представљање текста) која је прилагођена специфичном домену који се обрађује (нематеријално културно наслеђе), специфичним проблемима који се решавају (класификација протокола у односу на тематику) и српском језику, као једном од морфолошки богатих језика.

За рад са просторним подацима развијен је просторни модел који је погодан за приказ резултата на мапи као и за постављање просторних упита путем интерактивног графичког приказа мапе локација.

Резултати експеримената над развијеним методама показују да коришћење приступа заснованог на правилима у комбинацији са додатним језичким ресурсима даје веома добре резултате за задатак екстракције информација, као и да се методе засноване на статистичким техникама машинског учења показују успешним у решавању задатка тематске класификације и у овом специфичном контексту. Добијени резултати имају посебну димензију с обзиром на значај који очување националног кутурног наслеђа има за очување и неговање идентитета сваког појединца.

ПРИКАЗ ДИСЕРТАЦИЈЕ

Рукопис има 178 страна и обухвата 9 поглавља и закључак као и списак коришћене литературе од 317 библоиографских јединица, 8 табела и 18 слика. Структура рукописа је следећа.

У уводном поглављу уведени су концепти којима се дисертација бави – културно наслеђе, мултимедијалне базе података и мултимедијални системи за организацију културног наслеђа. Приказане су специфичности српског културног наслеђа, улога информационих технологија у очувању и коришћењу као и водећи стандарди за опис културног наслеђа, начини претраге мултимедијалне базе, јавно доступне дигиталне библиотеке културног наслеђа, пројекти, веб платформе и системи за дељење дигиталних мултимедијалних

колекција културног наслеђа.

Главе 2-4 посвећене су инфраструктури за информатичку обраду текстуалних и просторних података. У глави 2 представљено је управљање текстуалним подацима у XML формату као и типови база података за рад са текстом – базе података са подршком за XML (*XML Enabled Databases*) и изворне XML базе података (*Native XML Databases*), и посебно XML база података eXist-db. Глава 3 посвећена је управљању просторним подацима – представљању података у GML формату, раду са просторним подацима и просторним базама података, посебно PostgreSQL и PostGIS - објектно-релационој бази података која је једна од првих и значајнијих система са специјализованим операцијама за рад са просторним подацима и њеном просторном проширењу PostGIS. Глава 4 представља методолошку основу истраживања ове дисертације: представљене су методе обраде текста на природном језику - методе засноване на правилима и методе засноване на машинском учењу, специфични задаци обраде текста на природном језику – екстракција информација, класификација текста, анотација докумената, претраживање информација, као и специфичности обраде текста на српском језику и припадни ресурси и алати. Глава се завршава кратким уводом у специфичности обраде текста на природном језику у домену културног наслеђа.

Главе 5-9 представљају главни део дисертације. У њима се излаже проблем управљања мултимедијалним нематеријалним културним наслеђем и методе за његово решавање, архитектура и имплементација система за управљање нематеријалним кутурним наслеђем Балкана као и резултати примене развијених метода у решавању постављених задатака. Глава 5 посвећена је различитим аспектима управљања нематеријалним културним наслеђем – од захтева који се постављају пред одговарајуће системе, преко изазова у обради текстова из домена културног наслеђа, до специфичности културног наслеђа простора Балкана. У кратким цртама се описује мултимедијална колекција нематеријалног културног наслеђа Балкана која је мотивисала истраживања ове дисертације.

Глава 6 представља главни резултат дисертације - целовиту методологију решавања специфичног проблема управљања нематеријалним културним наслеђем – методе препознавања и екстракције информација (именованих ентитета – методе засноване на контексту и тема – методе засноване на конструкцији семантичких структура) и класификације текста са применама на текстуалне протоколе, методе семантичке анотације и организовања мултимедијалне колекције у базу података као и методе претраге мултимедијалне базе. Представљен је и развој библиотеке трансдуктора за екстракцију именованих ентитета и тематских фраза. Приказане су методе представљања документа н-грамима карактера, значајним речима или фразама, врећом речи, семанстичким ентитетима – и њиховим комбинацијама, и методе класификације к најближих суседа (kNN), потпорних вектора (SVM) и максималне ентропије (MaxEnt).

У глави 7 описане су методе примењене у изградњи мапе нематеријалног културног наслеђа Балкана – складиштење просторних података у XML и GML формату, повезивање са базом PostgreSQL и модулом PostGIS, графички приказ података и формирање просторних упита.

Глава 8 приказује архитектуру система и елементе имплементације. Део развијеног система написан је у Java програмском језику, организован је у архитектуру клијент/сервер и комуницира користећи TCP протокол. Део система који се односи на имплементацију NLP метода за класификацију протокола написан је у језику Python. Цео систем се састоји од око 15000 линија ауторског кода и обухвата око 100 класа које обезбеђују фунционалност

система, уз коришћење бројних других софтверских алатки за обраду текста, базе података, библиотека алгоритама машинског учења, за рад са мултимедијом и сл.

Глава 9 садржи резултате експеримената екстракције информације примењене на тектуалним протоколима, поређења различитих метода класификације текста избором различитих типова атрибута и класификације текстова према појављивању одређене тематике. За извођење експеримената над методама екстракције информација и тематске класификације коришћени су текстови протокола из мултимедијалне колекције о културном наслеђу Балкана. Посебна тачка Главе 9 посвећена је екстракцији информација из текстуалних протокола применом комерцијалног алата IBM SPSS Modeler. Један од његових модула, Text Analytics, користи се за различите напредне језичке анализе неструктурираног текста за брзу и ефикасну обраду уз технологије обраде природних језика и могућност коришћења језичких ресурса. Приказана је функционалност овог модула за задатак екстракције тема из протокола и установљено је да овај комерцијални алат има пун потенцијал за извођење овог задатка уз обезбеђене неопходне језичке ресурсе. За извођење експеримената поређења квалитета различитих метода класификације при промени атрибута на основу поларитета коришћени су јавно доступни скупови филмских рецензија на различитим језицима. Коришћене су уобичајене мере успешности које се користе у области претраживања информација – прецизност (P), одзив (R) и F мера. Методе екстракције информација из текстуалних протокола резултују укупном F мером (именовани ентитети и теме) од 0.87 што је у рангу (или превазилази) са актуелним резултатима за екстракцију информација уопште. Да би се одабрала најбоља метода за тематску класификацију текстуалних протокола, извршено је поређење метода представљања докумената и класификације на скупу јавно доступних скупова филмских рецензија. Методом потпорних вектора и представљањем документа протокола различитим комбинацијама атрибута н-грама и семантичких атрибута (контекстних и позиционих) добијени су најбољи резултати тематске класификације протокола.

У глави 10 дат је закључак, наведени су научни доприноси дисертације и правци даљих истраживања.

НАУЧНИ ДОПРИНОС ДИСЕРТАЦИЈЕ

- Развој новог модела и имплементација мултимедијалне базе података нематеријалног културног наслеђа дела Балкана

- Развој информатичког модела докумената и просторног модела географских карактеристика садржаја мултимедијалне базе података нематеријалног културног наслеђа

- Аутоматска семантичка анотација садржаја мултимедијалне колекције нематеријалног културног наслеђа, применом развијених метода екстракције информација и класификације текста, као основа за успешну претрагу по метаподацима

- Развој метода претраге садржаја мултимедијалне базе података нематеријалног културног наслеђа по просторним карактеристикама избором локације на интерактивној мапи

ЗАКЉУЧАК

У рукопису „Мултимедијалне базе података у управљању нематеријалним културним наслеђем" кандидат Ивана Танасијевић је показала широко и систематично познавање области обраде природних језика, мултимедијалних база података и машинског учења. Такође, овладала је значајним доменом примене ових знања – доменом управљања нематеријалним културним наслеђем који представља значајан изазов за информатичке технологије уопште и посебно језичке технологије. Развила је методу аутоматске семантичке анотације садржаја и извршила прилагођавање методе потпорних вектора тематској класификацији докумената мултимедијалне колекције нематеријалног културног наслеђа. Иако представљена методологија обједињује доменски специфично знање и језичко знање у библиотеку коначних трансдуктора и колекцију изабраних типова атрибута, она се може прилагодити за изградњу сличних система за друге домене и друге језике. Развијена је и основа за израду временског модела у виду система за екстракцију временских ентитета лингвистичко-семантичким методама обраде природних језика и постављена добра основа за имплементацију додатних метода претраге развојем богатог система семантичких метаподатака.

Кандидат Ивана Танасијевић је кроз истраживачки рад приказан у овој дисертацији дала значајан допринос области примене информационих технологија у решавању проблема управљања мултимедијалним базама података специфичне намене, и посебно, у решавању проблема управљања доменом нематријалног културног наслеђа који је од изузетног значаја за очување и неговање идентитета сваког појединца. Развијене методе имају адекватну примену и у низу других значајних домена. Стога предлажемо Наставно-научном већу Математичког факултета Универзитета у Београду, да рукопис "Мултимедијалне базе података у управљању нематеријалним културним наслеђем" кандидата Иване Танасијевић, прихвати као докторску дисертацију и одреди комисију за јавну одбрану.

У Београду, 10.10.2020. године

Чланови комисије:

др Гордана Павловић-Лажетић , редовни професор,
Математички факултет, Универзитет у Београду

др Ненад Митић, редовни професор,
Математички факултет, Универзитет у Београду

др Јелена Граовац, доцент,
Математички факултет, Универзитет у Београду

др Биљана Сикимић, научни саветник
Балканолошки институт САНУ