

НАСТАВНО-НАУЧНОМ ВЕЋУ

Предмет: Подобност теме и кандидата **Браниславе Цвијетић**, магистра техничких наука област аутоматика и информатика, за израду докторске дисертације „**Рестаурирање структуре података коришћењем техника машинског учења**” (на енглеском „*Restoration of data structures using machine learning techniques*”)

Одлуком 5045/16-1 бр. од 04.03.2022. године, именовани смо за чланове Комисије за оцену подобности теме и кандидата Браниславе Цвијетић за израду докторске дисертације и научне заснованости теме „Рестаурирање структуре података коришћењем техника машинског учења” (на енглеском „*Restoration of data structures using machine learning techniques*”)

На основу материјала приложеног уз Захтев кандидата, Комисија подноси следећи

ИЗВЕШТАЈ

1. Подаци о кандидату

1.1. Биографски подаци

Бранислава Цвијетић је рођена 26.01.1978. године у Сарајеву, Босна и Херцеговина. Електротехнички факултет Универзитета у Источном Сарајеву, уписала је 1996. године, а 2002. године је дипломирала на одсеку аутоматика и електроника на тему „Пројектовање и имплементација ПИД контролера” код проф. др Милице Б. Наумовић.

Постдипломске студије, одсек аутоматика и информатика, уписала је 2003. године на Електротехничком факултету Универзитета у Источном Сарајеву. Магистрирала је 2013. године код проф. др Слободанке Ђорђевић – Кајан на тему „Пројектовање и имплементација сигурности код система база података”.

Докторске студије уписала је 2016. године на модулу рачунарска техника и информатика Електротехничког факултета Београд и положила све испите са просечном оценом 10.00. Упоредно са редовим радним задацима, дефинисаним радним местом на којем је ангажована, бави се истраживањем проблема у области вештачке интелигенције, односно примене машинског учења у решавању проблема у домену рада.

Од 2003. године до 2006. године радила је на радном месту виши стручни сарадник – програмер у Републичком заводу за статистику Републике Српске. Од 2006. до 2007. радила је на радном месту стручни сарадник за базе података у Централној изборној комисији БиХ. Од 2007. до 2010. године радила је у Државној агенцији за истраге и заштиту. Од 2010. године запослена је у Агенцији за статистику БиХ на радном месту стручни саветник за пројектовање и програмирање.

У периоду од 15.01. - 15.06.2018. била је на стручном усавршавању у Заводу за статистику Републике Грчке у оквиру Програма обуке за статистичаре у Еуростату и државама чланицама Европске уније – Вишекорисничког програма статистичке сарадње. Стручно усавршавање је било на тему „Рад на експерименталним пројектима везаним за Велике податке”.

1.2. Стечено научноистраживачко искуство

Кандидат је докторске академске студије на изборном подручју Електротехника и рачунарство на ЕТФ-у уписао 2016/2017. школске године. Од тренутка уписа до тренутка подношења Захтева за одобрење докторске дисертације положио је све испите (са просечном оценом 10,0) и обавио све обавезе предвиђене докторским академским студијама на изборном подручју Електротехника и рачунарство. Списак положених испита и обављених обавеза је дат у следећој табели:

бр.	Назив испитија / обавезе	Оцена	ЕСПБ	Датум
1	Вештачка интелигенција и експертски системи	10	9	02.07.2018.
2	Софтверски дефинисане мреже	10	9	15.01.2018.
3	Студијски истраживачки рад II	-	15	30.09.2021.
4	Научно стручни рад	-	3	30.09.2021.
5	Студијски истраживачки рад I	-	15	30.09.2020

Бранислава Цвијетић се бави развојем метода за рестаурирање структура података коришћењем техника машинског учења. Циљ истраживања је да се предложи алгоритам и развије аутоматизовано софтверско решење за рестаурирање структура унутар података код којих је дошло до нарушавања структуре, а које је модулarno са становишта броја функција за рачунање сличности између структура необележеног скупа података. Такође, решење треба да омогући проширивање са новим алгоритмима машинског учења, могућношћу промене одређених параметара у погледу дефинисања оквира података за истраживање, додавања правила за одређивање врсте података и сл. Садржај генерисаних ентитета даље се може машински обрадити у релационим базама података и као такав послужити у процесу стицања знања и доношења пословних одлука, што представља основни значај предложеног истраживања. Додатни значај огледа се у чињеници да ће алгоритам предвидети могућност сажимања великог скупа података и омогућити примену алгоритама машинског учења уз уштеду времена и рачунарских ресурса, а решење неће бити ограничено доменом примене. За потребе рада биће креирани скупови података са нарушеном структуром на основу јавно доступних података. Бранислава Цвијетић је предложила решење за рестаурирање структура података коришћењем техника машинског учења.

Бранислава Цвијетић до сада је објавила један рад у међународном часопису (ван СЦИ листе) као први аутор и пет радова на међународној конференцији (пет као први аутор).

Дана 29.03.2022. године одржана је јавна усмена одбрана теме докторске дисертације (докторски испит) пред Комисијом у саставу: др Милош Цветановић, ванредни професор са Електротехничког факултета; др Зоран Шеварац, ванредни професор са Факултета организационих наука; др Горан Квашев, ванредни професор са Електротехничког

факултета. Оцена Комисије је да је кандидат Бранислава Цвијетић задовољила све критеријуме и одбранио предложену тему докторске дисертације.

У наставку је списак радова који показују резултате досадашњег рада и афинитета који су у складу са предложеном дисертацијом кандидата Браниславе Цвијетић.

Радови у међународним часописима (ван СЦИ листе):

B. Cvijetic, Z. Radivojevic, Application of Machine Learning in the Process of Classification of Advertised Jobs, International Journal of Electrical Engineering and Computing, Vol. 4, No. 2, pp. 93 - 100, Dec, 2020

Радови на домаћим конференцијама:

B. Cvijetiћ, Z. Radivojeviћ, Primjena mašinskog uћenja u procesu klasifikacije oглаћених radnih mјesta, XIV међународни симпозијум INFOTЕH-JAHORINA 2020, pp. 146 - 151, Jahorina, Istoћно Sarajevo, Republika Srpska, Bosna i Hercegovina, Mar, 2020

B. Cvijetiћ, Z. Radivojeviћ, Primjena Weka softvera i konvolucijskih neuronskih mreжа u procesu klasifikovanja digitalnih fotografija, XVIII међународни симпозијум INFOTЕH-JAHORINA 2019, pp. 224 - 229, Jahorina, Istoћно Sarajevo, Republika Srpska, Bosna i Hercegovina, Mar, 2019

B. Cvijetiћ, Otvaranje javnih podataka u Bosni i Hercegovini, XIV међународни симпозијум INFOTЕH-JAHORINA 2015, pp. 670 - 674, Jahorina, Istoћно Sarajevo, Republika Srpska, Bosna i Hercegovina, Mar, 2015

B. Cvijetiћ, D. Avram, M. Bratiћ, Upotreba kriptografije za sigurnost baza podataka, VII међународни симпозијум INFOTЕH-JAHORINA 2008, Jahorina, Istoћно Sarajevo, Republika Srpska, Bosna i Hercegovina, Mar, 2008

M. Naumoviћ, D. Avram, B. Cvijetiћ, Neuronske mreже u sistemima upravljanja kroz primere, IV међународни симпозијум INFOTЕH-JAHORINA 2005, Jahorina, Istoћно Sarajevo, Republika Srpska, Bosna i Hercegovina, Mar, 2005

1.3. Оцена подобности кандидата за рад на предложеној теми

Приликом писања овог Извештаја, Комисија је размотрила све претходно наведене чињенице и резултате у досадашњем раду Браниславе Цвијетић. Са посебном пажњом су размотрена искуства у истраживачком раду и узимајући све наведено, Комисија је стекла увид у њену способност за научно истраживачки рад и рад на предложеној теми докторске дисертације, која је, такође, препозната и вреднована и од стране научне заједнице.

2. Предмет, циљ и значај истраживања

Са напретком информационо-комуникационих технологија, количина похрањених и електронски доступних података, значајно се повећала током последњих година. Електронски подаци се могу налазити у различитим облицима, почевши од неструктурираних података који се чувају у текстуалним документима, полуструктурираним XML подацима, па до структурираних података који се чувају у релационим базама података. За приступ подацима могу се користити различити интерфејси, као на пример: веб претраживачи, структурирани упитни језик (енгл. *Structured Query Language*) и десктоп апликације. У пракси постоји потреба да се из таквих података дође до одређених знања

(енгл. *Knowledge Extraction*) и на основу тих знања да се донесу одлуке, без обзира о којој сфери живота је реч [1].

У складу са наведеним, подаци се могу поделити у три категорије: неструктуриране (енгл. *Unstructured data*), структуриране (енгл. *Structured data*) и полуструктуриране (енгл. *Semi-structured data*) податке. Неструктурирани подаци не морају нужно имати неки дефинисан формат или следити неко правило по питању садржаја података. У неструктуриране податке спадају текстуални подаци, видео, звук и фотографија. Насупрот томе, структурирани подаци су организовани у семантичке ентитете. Слични ентитети су груписани у виду релација, а ентитети исте групе имају исти опис, тј. исте атрибуте. Описи за све ентитете у некој групи чине схему. Атрибути обично имају предефинисан формат и дужину, а обично сви следе неки предефинисани редослед. Пример структурираних података укључује различите релационе податке који су смештени у релационим базама података. Трећа врста података су полуструктурирани подаци. Они имају одређене структуралне информације, али нису у потпуности структурирани. Насупрот структурираним подацима, код полуструктурираних података, ентитети исте групе могу да имају различите атрибуте, а редослед атрибута није нарочито важан, док се за означавање самих атрибута користе одговарајући маркери или лабеле.

Извлачење информација (енгл. *Information extraction - IE*) из полуструктурираних и неструктурираних података је област која се стално истражује пошто постоји потреба да се из таквих података извуку табеле [2 - 7], да се раздвоје релационе од нерелационих табела [8], као и да се извуку релационе табеле из хетерогених извора података [9 - 12].

Предмет истраживања у склопу ове докторске дисертације представља развој проширеног система за рестаурирање структура унутар скупа података за који се поуздано зна да поседују структуру, али која је из било ког разлога нарушена. Овакав скуп података у свом саставу може садржавати различите верзије исправних структура једног или више ентитета, а које су мењане током времена (нпр. замена у редоследу атрибута, броју атрибута и сл.). Такође може садржавати и нарушене структуре ентитета које су настале при експорту структурираних података услед коришћења различитих делимитера за раздвајање атрибута ентитета, као и новог реда. Осим наведених особина којима се описује овакав скуп података, проблем настаје и због саме хетерогености података. Подаци са великом флукутацијом и различитим типом података називају се хетерогени подаци. Хетерогени скупови структурираних података могу садржавати информације о датумима или временским интервалима, експлицитне нумеричке податке (нпр. година), кодиране нумеричке податка (нпр. 0 или 1 за присуство или одсуство), категоријске податке (нпр. земље) и текстуална поља (нпр. кратки опис). Такође, код оваквих скупова података постоји проблем који се огледа у чињеници да за неки атрибут могу недостајати вредности, тј. вредност таквих атрибута могу да остану непопуњене или попуњене неком подразумеваном вредношћу. Ови скупови података могу спадати у категорију великих података (енгл. *Big data*), тешко обрадиви са доступним рачунарским ресурсима, а најчешће не садрже мета податке. За скуп података, са претходно наведеним карактеристикама, може се рећи да је неструктуриран, али да се поуздано зна да је некад имао структуру.

Циљ истраживања, у склопу ове дисертације, је да се предложи алгоритам и развије аутоматизовано софтверско решење за рестаурирање структура унутар података код којих је дошло до нарушавања структуре, а које је модуларно са становишта броја функција за рачунање сличности између структура необележеног скупа података. Такође, решење треба да омогући проширивање са новим алгоритмима машинског учења, могућношћу промене одређених параметара у погледу дефинисања оквира података за истраживање, додавања правила за одређивање врсте података и сл. Наиме, софтверско решење треба да буде аутоматизовано на начин да на улазу прихвати скуп података за који је познато да је раније био структуриран, те да није познат делимитер који је коришћен за раздвајање атрибута, а као излаз да генерише ентитете. Садржај генерисаних ентитета даље се може машински обрадити у релационим базама података и као такав послужити у процесу стицања знања и

доношења пословних одлука, што представља основни значај предложеног истраживања. Додатни значај огледа се у чињеници да ће алгоритам предвидети могућност сажимања великог скупа података и омогућити примену алгоритама машинског учења уз уштеду времена и рачунарских ресурса [13]. Такође, значај овог истраживања је и у томе што решење неће бити ограничено доменом примене, а рад доменских стручњака ће бити минималан. Поред свега наведеног, значајан резултат истраживања је тај да ће, а за потребе верификације предложеног алгоритама и омогућавања даљих истраживања, бити креирани скупови података са нарушеном структуром коришћењем јавно доступних података.

Списак референци везаних за тезу:

- [1] O. Rusu, I. Halcu, O. Grigoriu, G. Neculoiu, V. Sandulescu, M. Marinescu, and V. Marinescu, "Converting unstructured and semi-structured data into knowledge," Proceedings - RoEduNet IEEE International Conference, pp. 1 - 4, 2013
- [2] J. C. Roldán, P. Jiménez, and R. Corchuelo, "On extracting data from tables that are encoded using HTML," Knowledge-Based Systems, 2019
- [3] E. Koci, M. Thiele, O. Romero, and W. Lehner, "Table identification and reconstruction in spreadsheets," In International Conference on Advanced Information Systems Engineering, pp. 527 - 541, Springer, 2017
- [4] T. Döhmen, H. Mühleisen, and P. Boncz, "Multi-hypothesis CSV parsing," In Proceedings of the 29th International Conference on Scientific and Statistical Database Management, SSDBM '17, ACM, pp. 16:1 - 16:12, 2017
- [5] C. Christodoulakis, E. B. Munson, M. Gabel, A.D. Brown, and R. J. Miller, "Pytheas: Pattern-based Table Discovery in CSV Files," Proceedings of the VLDB Endowment, Vol. 13, No. 11, pp. 2075 - 2089, 2020
- [6] L. Jiang, G. Vitagliano and F. Naumann, "Structure Detection in Verbose CSV Files," 24th International Conference on Extending Database Technology (EDBT), pp. 23 - 26, 2021
- [7] H. Dong, S. Liu, S. Han, Z. Fu, and D. Zhang, "TableSense: Spreadsheet Table Detection with Convolutional Neural Networks," AAAI, Vol. 33, no. 01, pp. 69 - 76, 2019
- [8] M. J. Cafarella, A.Y. Halevy, Y. Zhang, D. ZheWang, and Eugene Wu, "Uncovering the Relational Web," Proceedings of the 11th International Workshop on Web and Databases (WebDB 2008), Vancouver, Canada, 2008
- [9] Z. Chen and M. Cafarella, "Automatic web spreadsheet data extraction," In International Workshop on Semantic Search over the Web, pp. 1 - 8, 2013
- [10] J. Eberius, C. Werner, M. Thiele, K. Braunschweig, L. Dannecker, and W. Lehner, "Deexcelerator: A framework for extracting relational data from partially structured documents," In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 2477 - 2480, 2013
- [11] A. O. Shigarov and A. A. Mikhailov, "Rule-based spreadsheet data transformation from arbitrary to relational tables," Information Systems (IS) 71, pp. 123 - 136, 2017
- [12] G. Xiang, X. Ji, J. Zhou, Y. Yin, and X. Shen, "A flexible method for converting non-relational text data to relational data for data acquisition of equipment operation," Procedia Manufacturing, Vol. 37, 2019
- [13] A. S. Shirkhorshidi, S. Aghabozorgi, T.Y. Wah, and T. Herawan, "Big Data Clustering: A Review," Computational Science and Its Applications – ICCSA 2014, Vol. 8583, 2014

3. Полазне хипотезе

Полазне хипотезе у овом истраживању су:

- Могуће је предложити алгоритам, заснован на коришћењу техника машинског учења, којим се рестаурира структура код података за које се поуздано зна да поседују структуру, али која је из било ког разлога нарушена.
- Предложени алгоритам моћи ће да даје боље резултате у односу на постојећа решења која су примењива при решавању проблема рестаурирања структуре података са нарушеном структуром.
- Могуће је проширити предложени алгоритам скупом правила и функција за откривање скаларних типова атрибута унутар идентификоване структуре или више структура.
- Могуће је проширити предложени алгоритам скупом правила и функција за откривање векторских типова података.

4. Научне методе истраживања

У овом истраживању биће примењене дескриптивне, експерименталне и комперативне методе.

У првом делу истраживања биће извршен детаљан преглед постојећих решења из отворене литературе, а која се бави проблемом детекције структура из података са нарушеном структуром.

Након тога ће бити анализирани постојећи репозиторијуми података у циљу проналажења адекватних тестних података, односно података који садрже податке који имају један или више референцијалних ограничења, као и векторске типове података фиксне и променљиве дужине.

Потом ће бити анализирана међузависност између података унутар структура, коришћењем постојећих статистичких метода и алгоритама машинског учења. У складу са добијеним резултатима претходних анализа, биће предложен алгоритам за детекцију структура у подацима са нарушеном структуром.

У наредном делу истраживања биће имплементиран прототип алгоритма за откривање структура у подацима са нарушеном структуром, а који се заснива на алгоритмима машинског учења.

У експерименталном делу, над идентификованим улазним подацима, биће извршена анализа резултата примене алгоритама машинског учења, као и детаљно поређење резултата у погледу перформанси и комплексности алгоритама предложених решења. Наиме, на основу резултата добијених током експеримента, применом компаративних метода, биће извршен одабир техника и алгоритама машинског учења.

5. Очекивани научни допринос

Основни допринос раду јесте развој алгоритама за рестаурирање структура код података са нарушеном структуром, односно откривање ентитета и односа међу атрибутима ентитета.

Такође се очекује да ће истраживање дати и следеће научне доприносе:

- Евалуација постојећих алгоритама машинског учења који се могу применити у домену проблема откривања структуре података са нарушеном структуром, као и предлога њихове класификације са аспекта могућности примене у фазама развоја софтверског решења.
- Предложен нови алгоритма који користи технике машинског учења.
- Креирани нови репозиторијуми података са нарушеном структуром на основу постојећих јавно доступних скупова података у отвореној литератури и индустрији. Скупови података ће садржати модификоване податке, тј. податке где је вештачки направљено

нарушавање структуре података (на пример, замена атрибута, уклањање или додавање атрибута и сл.).

- Модуларно софтверско решење развијено сходно предложеном алгоритму за откривање структуре унутар података са нарушеном структуром и прошириво са новим модулима за откривање зависности између података, као и њиховог груписања.
- Дизајнирани експерименти погодни за евалуацију перформанси предложеног алгоритма и развијеног софтверског решења.

Сви добијени резултати биће дати на увид јавности кроз публикације на научно-стручним скуповима и часописима.

6. План истраживања и структура рада

Планирано је да се истраживање спроведе у седам (7) фаза.

- Прва фаза обухвата анализу доступне релевантне научно-стручне литературе, као и анализе јавно доступних скупова података који ће се користити у истраживању. У првој фази биће креирани тестни скупови података за потребе процене добијеног решења.
- Друга фаза обухвата анализу постојећих алгоритама и решења за откривање структура унутар података са нарушеном структуром.
- Трећа фаза обухвата предлог алгоритма, функција и метода за препознавање структура података на основу типова података, њиховог распореда, учесталости појављивања одређеног распореда података, као и развоја софтверског решења заснованог на предложеном алгоритму. У трећој фази биће креирано софтверско решење.
- Четврта фаза обухвата тестирање и евалуацију предложеног алгоритма за препознавање структура ентитета. У овој фази се врши и процена да ли софтверско решење исправно груписало ентитете и припадајуће атрибуте на нивоу ентитета тамо где је дошло до нарушавања структуре због непознатог делимитера.
- Пета фаза обухвата тестирање и евалуације предложеног алгоритма за препознавање стварних вредности атрибута унутар препознатих структура и њиховог даљег груписања уколико се ради о погрешно идентификованим стварним вредностима атрибута скаларног или векторског типа података.
- У шестој фази обавља се интегрално тестирање и провера рада предложеног алгоритма и развијеног софтверског решења над креираним скуповима података. Наиме, врши се провера да ли и колико успешно, развијено софтверско решење може препознати ентитете и њихове припадајуће атрибуте.
- Седма фаза се односи на анализу добијених резултата и доношења одговарајућих закључака.

Уколико резултати и закључци из седме фазе буду указивали да предложени алгоритам не даје очекиване резултате, приступиће се понављању одговарајућих фаза и извршити кориговање корака рада сходно новим претпоставкама.

7. Закључак и предлог

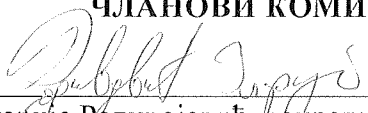
Приликом писања извештаја, Комисија је размотрила релевантне чињенице о подобности теме и кандидата. Узимајући у обзир резултате досадашњег рада кандидата **Браниславе Цвијетић**, Комисија је закључила да кандидат јесте способан за научно истраживачки рад. Детаљном анализом предложене теме под називом „**Рестаурирање структуре података коришћењем техника машинског учења**” (на енглеском „*Restoration of data structures using machine learning techniques*”), Комисија је у потпуности сагласна да оваква тема задовољава све услове подобности за докторску дисертацију.

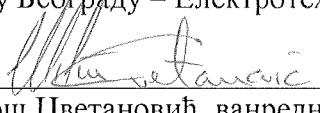
Комисија констатује да предложена тема припада ужој научној области **Рачунарска техника и информатика** и да је Универзитет у Београду - Електротехнички факултет (ЕТФ) матичан за ову научну област. За ментора докторске дисертације комисија предлаже **др Захарија Радивојевића, ванредног професора** са ЕТФ-а, и прилаже уз овај Извештај и списак радова др Захарија Радивојевића, који га квалификују за ментора докторске дисертације.

Имајући у виду све претходно наведено, Комисија има част и задовољство да Наставно-научном већу ЕТФ-а предложи да прихвати тему докторске дисертације „Рестаурирање структуре података коришћењем техника машинског учења” (на енглеском „*Restoration of data structures using machine learning techniques*”) и да кандидату Бранислави Цвијетић омогући приступ њеној изради.

У Београду, 01.04.2022.

ЧЛАНОВИ КОМИСИЈЕ


др Захарије Радивојевић, ванредни професор
Универзитет у Београду – Електротехнички факултет


др Милош Цветановић, ванредни професор
Универзитет у Београду – Електротехнички факултет


др Зоран Шеварац, ванредни професор
Универзитет у Београду – Факултет организационих наука


др Горан Квашчев, ванредни професор
Универзитет у Београду – Електротехнички факултет