

ИЗВЕШТАЈ О ОЦЕНИ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

I ПОДАЦИ О КОМИСИЈИ
<p>1. Датум и орган који је именовao комисију</p> <p>7. април 2016, Наставно-научно веће Економског факултета у Суботици, Универзитет у Новом Саду, на 7. седници</p> <p>2. Састав комисије са назнаком имена и презимена сваког члана, звања, назива уже научне области за коју је изабран у звање, датума избора у звање и назив факултета, установе у којој је члан комисије запослен:</p> <p>1. Проф. др Ласло Шереш, ванредни професор, Информациони системи и инжењеринг, 14. јул 2011, Универзитет у Новом Саду, Економски факултет у Суботици</p> <p>2. Проф. др Зита Бошњак, редовни професор, Информациони системи и инжењеринг, 14. децембар 2006, Универзитет у Новом Саду, Економски факултет у Суботици</p> <p>3. Проф. др Владан Девеџић, редовни професор, Софтверско инжењерство, јун 2005, Универзитет у Београду, Факултет организационих наука.</p>
II ПОДАЦИ О КАНДИДАТУ
<p>1. Име, име једног родитеља, презиме: Маја, Ненад, Димитријевић</p> <p>2. Датум рођења, општина, држава: 25.12.1974. године, Нови Сад, Република Србија</p> <p>3. Назив факултета, назив студијског програма дипломских академских студија – мастер и стечени стручни назив:</p> <p>Природно-математички факултет, Информатика, дипломирани информатичар</p> <p>4. Година уписа на докторске студије и назив студијског програма докторских студија – нема</p> <p>5. Назив факултета, назив магистарске тезе, научна област и датум одбране:</p> <p>University of British Columbia, Department of Computer Science, Vancouver, Canada, <i>„Data mining for Co-occurring Motion Trajectories“</i>, рачунарске науке, мај, 2002.</p> <p>6. Научна област из које је стечено академско звање магистра наука:</p> <p>информатичке науке</p>
III НАСЛОВ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:
„УНАПРЕЂЕЊЕ ПОСТУПАКА ЗА ОТКРИВАЊЕ АСОЦИЈАТИВНИХ ПРАВИЛА О КОРИШЋЕЊУ ВЕБ САЈТОВА“

IV ПРЕГЛЕД ДОКТОРСKE ДИСЕРТАЦИЈЕ:

Докторска дисертација кандидаткиње Маје Димитријевић под називом „Унапређење поступака за откривање асоцијативних правила о коришћењу веб сајтова“ садржи осам поглавља, која су изложена на укупно 148 страница, А4 формата. Докторска дисертација обухвата 17 слика, 14 графика и 23 табеле. При изради докторске дисертације кандидаткиња је консултовала 142 референце. У наставку је приказан кратак садржај дисертације:

Увод (од 4. до 8. странице)

1. Асоцијативна правила (од 9. до 14. странице)

- 1.1. Дефиниције асоцијативних правила
- 1.2. Улога асоцијативних правила у процесу откривања знања
- 1.3. Алгоритми за откривање асоцијативних правила
- 1.4. Домени примене асоцијативних правила

2. Web mining (од 15. до 22. странице)

- 2.1. Преглед web mining метода
- 2.2. Структура веб лог података
- 2.3. Примена асоцијативних правила у web mining-у

3. Упоредна анализа АП алгоритама (од 23. до 31. странице)

- 3.1. Преглед најзначајнијих алгоритама за откривање асоцијативних правила
- 3.2. Утицај избора параметара на квалитет откривеног знања

4. Мере интересантности АП алгоритама (од 32. до 59. странице)

- 4.1. Преглед математичких функција као мера интересантности асоцијативних правила
- 4.2. Утицај мера интересантности на квалитет откривеног знања
- 4.3. Мере интересантности у web mining-у

5. Софтверски систем за анализу веб лог података (од 60. до 94. странице)

- 5.1. Елементи објектно-оријентисаног дизајна
- 5.2. Архитектура система
- 5.3. Пред-процесирање веб лог података
- 5.4. Уграђени алгоритми за проналажење асоцијативних правила
- 5.5. Избор параметара и мера интересантности
- 5.6. Имплементација и уградња нових мера интересантности
- 5.7. Приказ корисничког интерфејса
- 5.8. Тестирање система на експерименталним подацима

6. Резултати истраживања (од 95. до 128. странице)

- 6.1. Утицај мера интересантности АП при анализи веб лог података
- 6.2. Имплементација и проширење функционалности софтверског система за проналажење асоцијативних правила додавањем нових мера интересантности
- 6.3. Анализа квалитета знања добијеног проширеним софтверским системом

7. Закључак (од 129. до 135. странице)

Литература

V ВРЕДНОВАЊЕ ПОЈЕДИНИХ ДЕЛОВА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

У уводном поглављу дефинисани су предмет, проблем и циљеви истраживања. Дат је кратак преглед основних појмова везаних за примену асоцијативних правила, као и мотивација за унапређење квалитета откривених асоцијативних правила о коришћењу веб сајтова.

У првом поглављу дате су дефиниције асоцијативних правила и описана њихова улога у процесу откривања знања у великим базама података. Дат је преглед алгоритама за откривање асоцијативних правила и разноврсних домена њихове примене.

У другом поглављу дат је преглед *web mining* метода, са акцентом на методе за примену асоцијативних правила на откривање знања у подацима о коришћењу веб сајтова.

У трећем поглављу дата је упоредна анализа најзначајнијих алгоритама за откривање асоцијативних правила у трансакционим базама података, као и метода за њихову евалуацију.

У четвртном поглављу дат је преглед математичких функција као мера интересантности асоцијативних правила. Потом је дефинисана „локална Z-score“ функција, као мера статистичке очекиваности асоцијативног правила у односу на неко општије асоцијативно правило. Дефинисан је услов за елиминацију асоцијативних правила која су статистички очекивана у односу на општија асоцијативна правила, која такође постоје у скупу свих правила. Приказана је двојака примена претходно дефинисаног општег услова елиминације. У првом случају он се примењује за елиминацију асоцијативних правила статистички очекиваних у односу на општија асоцијативна правила мање дужине. У другом случају примењује се за елиминацију асоцијативних правила која су статистички очекивана у присуству концептне хијерархије веб страница. Предложено је коришћење концептне хијерархије веб страница генерисане на основу асоцијативних правила са максималном вредности confidence мере интересантности. Оваква асоцијативна правила се јављају у подацима о коришћењу веб сајтова, а последица су повезаности веб страница хиперлинк структуром. У оквиру четвртог поглавља предложене су дефиниције „локалних мера интересантности“ асоцијативних правила која нису елиминисана као статистички очекивана у односу на неко општије асоцијативно правило, применљиве на асоцијативна правила чија лева и десна страна садрже потконцепт заједничког натконцепта у оквиру концептне хијерархије. Показано је да примена локалних мера интересантности умањује интересантност једног дела асоцијативних правила, приближавајући је њиховој стварној интересантности са становишта аналитичара података.

У петом поглављу приказан је софтверски систем за откривање асоцијативних правила о коришћењу веб сајтова развијен у оквиру ове докторске дисертације. Имплементирани су методе за претпроцесирање веб лог података, алгоритми за генерисање асоцијативних правила, методе за елиминацију статистички очекиваних правила предложене у четвртном поглављу дисертације, као и методе за рангирање откривених асоцијативних правила коришћењем више различитих мера интересантности. Приказани софтверски систем је дизајниран користећи објектно-оријентисану методологију, имајући у виду модуларност и проширивост. Систем је имплементиран у програмском језику C#, као самостална Windows desktop апликација. Дат је приказ објектно-оријентисаног дизајна овог софтверског система, као и корисничког интерфејса.

Шесто поглавље приказује резултате експерименталног истраживања, где је имплементирани софтверски систем примењен на два стварна скупа података. У првој фази експерименталног истраживања показује се да број откривених асоцијативних правила о коришћењу веб сајтова расте експоненцијално са смањењем минималног support параметра и порастом броја фреквентних веб страница. Потом се примењују једноставније методе елиминације статистички очекиваних асоцијативних правила – елиминација кластер правила и статистички очекиваних правила у скупу свих веб сесија и испитује њихова ефикасност на смањење величине скупа откривених асоцијативних правила. На скупу асоцијативних правила претходно пречишћеном применом једноставнијих метода елиминације тестира се ефикасност метода елиминације статистички очекиваних асоцијативних правила предложених у оквиру ове дисертације. Спроведени експерименти показују да се елиминацијом асоцијативних правила која су статистички очекивана у односу на краћа и општија асоцијативна правила (предложеног у четвртном поглављу), величина

скупа откривених правила смањује у интервалу од 73% до 93%, што значајно олакшава сналажење аналитичарима података при одабиру стварно корисних правила. Потом се у следећој фази експеримената, применом методе елиминације асоцијативних правила која су статистички очекивана у присуству концептне хијерархије веб страница (предложеног у четвртном поглављу), постиже додатно смањење величине скупа откривених асоцијативних правила у интервалу од 14% до 34%. У последњој фази експерименталног истраживања приказаног у шестом поглављу извршено је рангирање асоцијативних правила која преостају у скупу откривених правила после његовог пречишћавања, применом више статистичких мера интересантности, укључујући „локалне“ мере интересантности предложене у четвртном поглављу. Резултати рангирања различитим мерама интересантности поређени су применом Spearman-овог коефицијента корелације ранга.

У последњем поглављу изведени су закључци на основу резултата истраживања, указано је на постигнуте циљеве истраживања, научни допринос дисертације, као и на будуће правце истраживања.

VI СПИСАК НАУЧНИХ И СТРУЧНИХ РАДОВА КОЈИ СУ ОБЈАВЉЕНИ ИЛИ ПРИХВАЋЕНИ ЗА ОБЈАВЉИВАЊЕ НА ОСНОВУ РЕЗУЛТАТА ИСТРАЖИВАЊА У ОКВИРУ РАДА НА ДОКТОРСКОЈ ДИСЕРТАЦИЈИ

Радови објављени у међународним часописима

Dimitrijević M., Bošnjak Z., *Pruning statistically insignificant association rules in the presence of high-confidence rules in web usage data*, *Procedia Computer Science Journal*, 2014, Elsevier, Vol 35, 2014, pp. 271–280. (indexed – Kobson – Web of Science)

Dimitrijević M., Subić N., Bošnjak Z., *Improving the Interestingness of Web Usage Association Rules Containing Common Web Site Menu Items*, *The Online Journal of Applied Knowledge Management*, Vol 2, 2014, pp. 82-92. (received best paper award)

Dimitrijević M., Krunić T., *Privacy preserving association rule mining applied to web usage data*, *Monitoring and Expertise in Safety Engineering*, Vol 4/3, 2014, pp. 1-7.

Dimitrijević M., Krunić T., *Association Rules for Improving Website Effectiveness: Case Analysis*, *The Online Journal of Applied Knowledge Management*, Vol 1, 2013, pp. 56-63.

Dimitrijević, M., Bošnjak Z., *Web Usage Mining System*, *Interdisciplinary Journal of Information, Knowledge, and Management*, Vol 6, 2011, pp. 137-150. (indexed – Scopus)

Dimitrijević, M., Bošnjak Z., *Discovering Interesting Association Rules in the Web Log Usage Data*, *Interdisciplinary Journal of Information, Knowledge, and Management*, Vol 5, 2010, pp. 191-207. (indexed – Scopus)

Радови објављени у зборницима међународних научних скупова

Dimitrijević, M., Subić, N., Gemović, B. *Objektno-orijentisani softver za pripremu podataka o korišćenju veb sajtova u procesu otkrivanja znanja*, 2. Међународно саветовање “Управљање знанјем и информатика”, *Zbornik radova*, str. 87-92, 2016, Kopaonik, Srbija.

Dimitrijević, M., Krunić, T., Subić, N., *Prečišćavanje skupa asocijativnih pravila o korišćenju veb sajtova primenom lokalne Z-score mere interesantnosti*, 1. Међународно саветовање “Управљање знанјем и информатика”, *Zbornik radova*, str. 101-107, 2015, Kopaonik, Srbija.

VII ЗАКЉУЧЦИ ОДНОСНО РЕЗУЛТАТИ ИСТРАЖИВАЊА

Један од основних фактора који негативно утичу на употребљивост алгоритама за откривање асоцијативних правила је тенденција генерисања превеликог броја правила у којима се аналитичари података тешко сналазе при одабиру стварно корисних правила. Овај проблем је погоршан у случају асоцијативних правила о коришћењу веб сајтова услед снажне корелације између веб страница, која је последица њихове повезаности хиперлинк структуром веб сајтова. Као последица тога генерише се превелики број асоцијативних правила која су заправо статистички очекивана и самим тим

неинтересантна аналитичарима података.

Основни циљ овог истраживања био је:

1. *Развој теоријског оквира и унапређење метода за проналажење и вредновање асоцијативних правила у веб сервер лог подацима.*

Овај циљ био је разложен на потциљеве:

- 1.1 *Упоредна анализа примене различитих мера интересантности правила кроз аспект њихове употребљивости за анализу веб сервер лог података*

Извршена је упоредна анализа различитих статистичких мера интересантности асоцијативних правила о коришћењу веб сајтова. Експериментално истраживање је извршено користећи два стварна скупа веб сервер лог података. Резултати рангирања асоцијативних правила путем различитих статистичких мера интересантности су поређени коришћењем Spearman-ovog коефицијента корелације ранга.

- 1.2 *Формулисање смерница при одабиру мера интересантности асоцијативних правила.*

У оквиру овог потциља, ради повећања квалитета откривених асоцијативних правила, предложена је метода за елиминацију асоцијативних правила која су статистички очекивана у односу на општија асоцијативна правила која такође постоје у скупу откривених правила. Показано је да се на овај начин елиминишу асоцијативна правила чија је вредност статистичких мера интересантности несразмерно велика у односу на њихову стварну интересантност са становишта аналитичара података. Предложени метод за елиминацију статистички очекиваних асоцијативних правила базиран је на Z-score статистичкој мери. Показано је да се овај метод може применити двојачко. У првом случају, елиминишу се асоцијативна правила која су статистички очекивана у односу на општија асоцијативна правила мање дужине. У другом случају, елиминишу се општија асоцијативна правила дефинисана у односу на концептну хијерархију веб страница. Показано да је предложени метод елиминације статистички очекиваних асоцијативних правила општији и статистички оправданији у односу на методе коришћене у претходним истраживањима.

Коришћена је концептна хијерархија генерисана на основу асоцијативних правила са максималним вредностима confidence мере интересантности. Тако дефинисана концептна хијерархија погодна је за анализу веб сервер лог података, и јавља се као последица повезаности веб страница хиперлинк структуром. Ограничење коришћења овакве концептне хијерархије је што она не обухвата све везе потконцепт/натконцепт, које би могао дефинисати експерт, или који би евентуално могли бити аутоматски екстраховани семантичком анализом веб сајта. Предности њеног коришћења су што за њено генерисање није потребно експертско знање, нити се захтева додатно рачунско процесирање, јер се она заснива на већ откривеним асоцијативним правилима.

Такође је предложена модификација стандардних статистичких мера интересантности, дефинисана у односу на исту концептну хијерархију, којом се повећава квалитет рангирања асоцијативних правила која не могу бити елиминисана као статистички очекивана. Теоретски и експериментално је показано да примена предложених модификованих мера интересантности умањује интересантност асоцијативних правила чија лева и десна страна садрже потконцепт заједничког натконцепта у оквиру концептне хијерархије, приближујући је њиховој стварној интересантности са становишта аналитичара података.

Практични циљ истраживања био је:

2. *Развој софтверског система за анализу веб лог података са могућношћу одабира постојећих и генерисања нових мера интересантности асоцијативних правила.*

Овај циљ био је разложен на потциљеве:

- 2.1 *Анализа функционалности одређеног софтверског система за data mining и предлог његовог проширења*

Анализиран је популарни Weka систем за data mining – софтвер отвореног типа који обухвата широк опсег data mining метода. Дат је пример употребе Weka система за анализу веб лог података у циљу

откривања асоцијативних правила о коришћењу веб сајта на стварном скупу података и изложена ограничења коришћења Weka система.

Предложено је коришћење софтверског система који би био специјализован за откривање асоцијативних правила о коришћењу веб сајтова, и који би интегрисао све фазе процеса откривања асоцијативних правила. Предложени софтверски систем обухвата припрему веб лог података, примену алгоритма за откривање асоцијативних правила, имплементацију предложених метода за елиминисање неинтересантних асоцијативних правила, као и рангирање преосталих асоцијативних правила применом стандардних и модификованих мера интересантности.

2.2 Проширење функционалности постојећег софтверског система у циљу побољшања квалитета откривеног знања

Имплементиран је софтверски систем за откривање асоцијативних правила о коришћењу веб сајтова као самостална, објектно-оријентисана софтверска апликација. Рад имплементираних софтверског система је експериментално испитан на два стварна скупа веб лог података која садрже 485.143, односно 403.915 корисничких захтева за релевантним веб страницама, распоређених у 203.380, односно 66.746 корисничких веб сесија.

У првој фази експерименталног истраживања потврђено је да број откривених асоцијативних правила о коришћењу веб сајтова применом стандардних алгоритама за њихово генерисање расте експоненцијално са бројем фреквентних веб страница. Потом је експериментално показано да и после примене основних метода елиминације (елиминација такозваних кластер правила и статистички очекиваних правила у скупу свих веб сесија), преостаје превелики број асоцијативних правила која имају неоправдано повишене вредности стандардних статистичких мера интересантности.

Експерименти којима се тестира ефикасност предложених метода елиминације асоцијативних правила која су статистички очекивана у односу на краћа и општија асоцијативна правила, показали су да се применом ове методе величина скупа откривених правила значајно смањује. Потом је експериментално показано да се применом предложених метода елиминације асоцијативних правила која су статистички очекивана у присуству концептне хијерархије веб страница, постиже додатно смањење величине скупа откривених асоцијативних правила.

Обзиром да примењене методе елиминације не доводе до губитка потенцијално интересантних асоцијативних правила, учинак смањења величине скупа откривених асоцијативних правила применом ових метода је директно пропорционалан повећању квалитета и употребљивости скупа откривених асоцијативних правила.

У првој фази експерименталног истраживања тестирано је рангирање асоцијативних правила која преостају у скупу откривених правила после његовог пречишћавања, применом више статистичких мера интересантности, укључујући предложене модификоване локалне мере интересантности. Потврђено је да примена локалних мера интересантности умањује интересантност асоцијативних правила чија лева и десна страна садрже потконцепт заједничког натконцепта у оквиру концептне хијерархије. На овај начин се вредност статистичких мера интересантности откривених асоцијативних правила приближава њиховој стварној интересантности, додатно олакшавајући аналитичарима података одабир стварно корисних правила.

Елиминација неинтересантних, односно статистички очекиваних асоцијативних правила једна је од кључних фаза поступака за откривање асоцијативних правила, којом се знатно повећава употребљивост скупа откривених асоцијативних правила. Методе за елиминацију неинтересантних асоцијативних правила, као и за њихово рангирање коришћењем локалних мера интересантности предложене у овој дисертацији су математички јасно дефинисане, и прилагођене примени у домену откривања знања у веб лог подацима. Софтверски систем развијен у оквиру овог истраживања интегрисао све фазе процеса откривања асоцијативних правила о коришћењу веб сајтова. Његовом применом експериментално је потврђена ефикасност предложених метода у смислу повећања квалитета знања садржаног у скупу откривених асоцијативних правила у веб лог подацима.

Иако су методе предложене у оквиру овог истраживања примењене на откривање знања у веб лог подацима, оне би се могле уопштити и применити и на откривање асоцијативних правила у другим

доменима, у случајевима када је могуће креирати концептну хијерархију објеката у датој бази знања. Креирање овакве концептне хијерархије било би могуће коришћењем експертског знања, или аутоматски, што може представљати правац будућег истраживања.

VIII ОЦЕНА НАЧИНА ПРИКАЗА И ТУМАЧЕЊА РЕЗУЛТАТА ИСТРАЖИВАЊА

Докторска дисертација кандидаткиње Маје Димитријевић у потпуности је испунила циљеве дефинисане у пријави теме. Емпиријско истраживање је у складу са насловом докторске дисертације. Кандидаткиња је дала одговор на постављена истраживачка питања и реализовала дефинисане циљеве истраживања. Резултати су приказани јасно и прегледно, уз навођење претходних истраживачких резултата из ове области. Резултати експеримената су приказани табеларно и графички, и изведени су јасни и прецизни закључци.

Комисија доноси позитивну оцену начина приказа и тумачења резултата истраживања.

IX КОНАЧНА ОЦЕНА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

1. Да ли је дисертација написана у складу са образложењем наведеним у пријави теме

Докторска дисертација је у потпуности **написана у складу са образложењима** које је кандидаткиња навела у пријави теме.

2. Да ли дисертација садржи све битне елементе

Да, докторска дисертација **садржи све битне елементе**. Кратак увод и прва три поглавља дају преглед досадашњих истраживања релевантних за тематику дисертације. Четврто поглавље представља оригиналан теоретски допринос дисертације. Пето поглавље описује практични допринос истраживања, шесто садржи експерименталне резултате, док су у закључку сумирани резултати истраживања. Списак литературе на крају дисертације се састоји од 142 референце које су у тексту цитиране на исправан начин.

3. По чему је дисертација оригиналан допринос науци

Дисертација садржи оригиналан научни допринос из области откривања знања у подацима. Унапређене су методе за пречишћавање скупа откривених асоцијативних правила о коришћењу веб сајтова, као и методе за њихово квалитетније рангирање статистичким мерама интересантности. Предложени критеријуми за елиминацију статистички очекиваних асоцијативних правила и за њихово рангирање су математички прецизно дефинисани. Имплементиран је софтверски систем за откривање асоцијативних правила о коришћењу веб сајтова, који обухвата све кораке процеса откривања знања. Коришћењем имплементираних софтверских система је на стварним скуповима података о коришћењу веб сајтова експериментално испитана ефикасност унапређених метода за откривање, пречишћавање и рангирање асоцијативних правила.

4. Недостаци дисертације и њихов утицај на резултат истраживања

Докторска дисертација нема уочених недостатака који би могли утицати на постизање, интерпретацију или презентовање дефинисаних циљева истраживања.

X ПРЕДЛОГ:
На основу укупне оцене дисертације, комисија предлаже:
На основу напред дефинисаних ставова и у складу са анализом постигнутих резултата и доприноса докторске дисертације, Комисија предлаже Наставно-научном већу Економског факулета у Суботици да прихвати позитивну оцену докторске дисертације „Унапређење поступака за откривање асоцијативних правила о коришћењу веб сајтова“ и да кандидаткињи Маји Димитријевић одобри јавну одбрану докторске дисертације пред усвојеном Комисијом.

У Суботици, 8. априла 2016. године

ЧЛАНОВИ КОМИСИЈЕ:

Проф. др Ласло Шереш, ванрени професор
Економски факултет у Суботици,
Универзитет у Новом Саду

Проф. др Зита Бошњак, редовни професор
Економски факултет у Суботици,
Универзитет у Новом Саду

Проф. др Владан Девеџић, редовни професор
Факултет организационих наука,
Универзитет у Београду

НАПОМЕНА: Члан комисије који не жели да потпише извештај јер се не слаже са мишљењем већине чланова комисије, дужан је да унесе у извештај образложење односно разлоге због којих не жели да потпише извештај.