Oponentský posudek disertační práce

Autor práce: **Mgr. Ľubomír Benko**

Název práce: **Modelovanie správania sa používateľov webu v závislosti od času**

Cíle a podcíle práce jsou přehledně definovány v úvodu celé práce i v úvodu dílčích výzkumů, které naplňují definované podcíle. Jako hlavní cíl si autor vytýčil analýzu chování uživatelů webů v závislosti na čase. Autor provedl celou řadu výzkumu, které navázali na předchozí práce v dané oblasti a jejichž výsledky umožnili naplnění hlavního cíle. Klíčovým se ukazuje zejména rozsáhlý a detailní výzkum zabývající se identifikací sezení uživatele webu dle záznamu v logovacích souborech webového serveru. Autor navrhl několik nových funkčních postupů, které zlepšují dosavadní algoritmy, řešící danou problematiku.

Obsah a struktura práce

Disertační práce je členěna do 4 kapitol. V rešeršní části autor se zabývá zejména popisem přípravy dat, jejich čistěním, identifikací uživatelů a identifikací sezení. V dalším se poté zabývá rekonstrukcí aktivit uživatele a modelováním chování uživatele v závislosti na čase pomocí modelu logistické regrese.

V další části autor ukazuje způsob přípravy dat na serveru s anonymním přístupem a stejně tak na serveru školního LMS s autentifikací uživatele. Velká pozornosti je věnována výzkumu a optimalizaci algoritmu Reference Length pro identifikaci sezení uživatele. V tomto algoritmu autor využívá entropie webových stránek k odhadu podílu navigačních stránek.

Zbývající dvě oblasti výzkumu se zaměřují na analýzu chování uživatelů serveru bankovní instituce s ohledem na hodnocení frekventovaných transakcí pro jednotlivé části webu v čase a analýzu pravděpodobnosti přístupu na vybrané části webu s ohledem na čas.

Úroveň zpracování práce, přínos práce

Práce je psána odborným jazykem, který odpovídá problematice práce. Někdy se v textu vyskytují drobné nepřesnosti (str. 28 vztah 12 neodpovídá podmíněné pravděpodobnosti – chyba lomítka), popř. nečitelný obsah obrázku (str. 40 obrázek č. 8). U provedených statistických testů je možné uvést důvod volby daného testu v případě použití více možných testů (např. Friedmanova testu, Turkey testu). Stejně tak je možné detailněji odůvodnit volbu modelu logistické regrese oproti jiným metodám.

Přínosem práce je zejména návrh a ověření metodiky přípravy dat zejména v identifikaci sezení a optimalizaci již navržených algoritmů pomocí nové metody odhadu podílu

navigačních stránek pomocí entropie. Stejně tak novými jsou metodika modelování chování uživatelů webu a hodnocení frekventovaných transakcí v závislosti na čase a také metodika pro určení pravděpodobnosti přístupu uživatelů na vybrané části webu v závislosti na čase pomocí modelu logistické regrese. Tyto postupy umožňují praktické využití v komerční oblasti a efektivnější řízení aktualizace komerčních webů dle chování uživatelů webu.

Formální náležitosti práce

Po formální stránce je práce zpracována na odpovídající úrovni. Některé obrázky nejsou příliš čitelné. Ostatní obrázky i tabulky jsou vhodně zapracovány do textu práce. Práce je vhodně členěna do kapitol s ohledem na prováděné výzkumy i jejich posloupnost směřující k naplnění hlavního cíle.

Způsob naplnění cílů práce

Cíle i dílčí cíle práce jsou na začátku práce jasně a přehledně definovány. Dále jsou znovu uvedeny v každé příslušné části zabývající se výzkumem daného tématu. Kromě metodiky, uvedené v obecné části, je pro každý výzkum uvedena podrobná metodika spolu s výsledky. Všechny vytčené cíle práce tak byly naplněny a umožnily splnění hlavního cíle celé práce. V závěru práce je možné kromě shrnutí uvést také případná omezení vytvořených modelů a z nich vyplývající doporučení pro další výzkum v dané oblasti.

Celkově lze konstatovat, že předložená práce splňuje svým obsahem i zpracováním požadavky kladené na disertační práci. Práci hodnotím kladně a doporučuji ji k obhajobě.

Otázky

1. Proč autor používá Friedmanův test (str. 49) místo Kruskall-Walis testu, popř. ANOVA? Jsou použité výběry závislé?
2. Proč pro vytvoření homogenních skupin v mnohonásobném porovnávání byl z dostupných testů použit právě Turkey test? Jsou použité výběry vyvážené?
3. Zvažoval autor i další metody pro predikci chování uživatelů s ohledem na čas?
4. Není vhodnější v případě zkoumání chování studentů v kurzu PDA použít vhodně strukturované očištěné logovací soubory LMS Moodle, které požadované informace ve vhodném formátu již poskytují?
5. Lze identifikovat některá omezení použitých modelů a postupů?

23. 11. 2018

doc. RNDr. Mikuláš Gangur, Ph.D.

Posudok oponenta

Oponent: prof. RNDr. Jozef Komorník,DrSc.

Doktorand: Mgr. Ľubomír Benko

Téma dizertačnej práce: *Modelovanie správania sa používateľov webu v závislosti od času*

Práca je zameraná na získavanie znalostí z používania webu (web usage mining). Téma práce je aktuálna, vzhľadom na absenciu úlohy predikcie v oblasti modelovania správania sa používateľov webu. Autor sa v práci primárne zaoberá návrhom metodík modelovania správania sa používateľov webu v závislosti od času. V riešenej problematike sa doktorand dobre zorientoval o čom svedčí viac ako 80 použitých zdrojov a prehľadne spracovaný súčasný stav problematiky. Hlavným cieľom práce bol návrh prístupov a metodík k modelovaniu správania sa používateľov webu v závislosti od času. Hlavný cieľ práce ako aj z neho vyplývajúce úlohy sú definované jasne a presne, pričom autor ich všetky splnil.

Autor práce sa upriamil nielen na modelovanie pravdepodobností prístupov používateľov (stakeholderov) portálu skúmanej komerčnej banky vo vzťahu k Pillar3 informáciám, ale aj na kvantitatívne zhodnotenie ich vzorov správania sa, a to netradične, z časového hľadiska. Obidva prístupy sa vhodne dopĺňajú a prinášajú detailný pohľad na skúmané obdobie rokov 2009-2015. Pozitívne hodnotím, že sa autor neobmedzil iba na modelovanie pravdepodobností prístupov v závislosti od času, resp. na zhodnocovanie vzorov správania sa používateľov webu v čase. Dopracovaním sa k rovnakým záverom overil správnosť obidvoch prístupov.

Kladne hodnotím práve zhodnocovanie vzorov správania sa používateľov webu v čase, pričom sa autor z časti inšpiroval metaanalýzou, v ktorej zhodnocoval frekventované množiny webových častí súvisiacich s Pillar3 informáciami. V názve kapitoly 3.4 sú uvedené nielen transakcie, ale aj sekvencie, pričom sa však autor obmedzil iba na asociačné pravidlá. Zaujímavým by bolo zamerať sa nielen na frekventované transakcie, ale aj na samotné pravidlá a s nimi súvisiace charakteristiky.

Skúšali ste zhodnocovať extrahované vzory správania sa z hľadiska ich kvality?

Časovo najnáročnejšou fázou v procese získavania znalostí je predspracovanie dát. Autor práce neopomína túto dôležitú časť procesu, bez ktorej by nebolo možné získať správne výsledky. Nezaoberá sa len zhodnotením jednotlivých krokov predspracovania dát, ale prináša aj nové prístupy, napr. k identifikácii sedení prostredníctvom entropie. Predspracovanie dát pozostávalo z čistenia dát, identifikácie sedení, dopĺňania ciest a vytvárania premenných. Práve vytváranie premenných, či už závislej premennej alebo samotných prediktorov bolo nevyhnutné pre modelovanie pravdepodobností prístupov v závislosti od času.

Výsledky práce boli riešené v zhode so stanovenými úlohami. Závery práce sumarizujú výsledky a prínos práce.

Výsledky práce je možné využiť vo fázach predspracovania dát, modelovania dát a evalvácie výsledkov v procese získavania znalostí z používania webu.

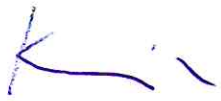
Zároveň oceňujem fakt, že autor práce bol aktívnym spoluriešiteľom projektov Vedeckej grantovej agentúry (VEGA 1/0392/13, VEGA 1/0776/18), riešených na FPV UKF v Nitre a FM UK v Bratislave, prostredníctvom ktorých si osvojoval metódy a postupy získavania znalostí aplikovaných v doméne bankovníctva.

Práca spĺňa požiadavky kladené na tento typ záverečnej práce a prácu odporúčam v predloženej podobe obhajovať a po jej úspešnej obhajobe navrhujem, aby Mgr. Ľubomírovi Benkovi bol udelený akademický titul

Philosophiae Doctor (PhD.)

v študijnom programe Aplikovaná informatika.

V Bratislave, 13.11.2018 Podpis:



doc. Ing. Pavel Petr, PhD.,
Ústav systémového inženýrství a informatiky
Fakulta ekonomicko-správní, Univerzita Pardubice

Oponentní posudek

Disertant: Mgr. Ľubomír Benko
Název disertační práce: Modelovanie správania sa používateľov webu v závislosti od času
Školitel: doc. RNDr. Michal Munk, PhD.

Předložená disertační práce je zpracovaná na 109 stranách a je rozdělena do čtyř kapitol. V první kapitole je definován cíl práce a úlohy, které je třeba řešit na dosažení tohoto cíle. Druhá kapitola je věnována analýze současného stavu řešené problematiky, což je problematika získávání znalostí z používání webu. Výsledky výzkumu a závěry jsou uvedeny v kapitole tři a čtyři.

V hodnocení předložené disertační práce se budu věnovat následujícím hlediskům:

- aktuálnosti zvoleného tématu;
- splnění sledovaného cíle;
- zvoleným metodám zpracování;
- výsledky disertační práce s uvedením nových poznatků.

Aktuálnost zvoleného tématu.
Zvolené téma disertační práce je nesporně aktuální. Problematika identifikace uživatelů a identifikace sezení uživatelů je velmi zajímavým tématem. Výsledky lze využít při optimalizaci struktury webu i získávání zajímavých a užitečných znalostí o chování uživatelů webu.

Splnění sledovaného cíle.
Cílem práce je návrh přístupů a metodik k modelování chování uživatelů webu v závislosti na čase. Lze konstatovat, že tento cíl, i všechny dílčí cíle práce byly splněny.
Autor se v jednotlivých částech práce velice podrobně zabýval problematikou přípravy dat i vlastním modelováním chování uživatelů. Problém identifikace uživatelů a sezení analyzoval na příkladu dat z portálu s anonymním přístupem i s povinnou autentifikací.

Zvolené metody zpracování.
Použité metody považuji za odpovídající. Kladně hodnotím i prezentování metodiky pro každou část výzkumu s uvedením výsledků.

Výsledky disertační práce s uvedením nových poznatků.
Výsledky disertační práce jsou v závěru práce vhodně sumarizovány. Lze konstatovat, že autor celkem přehledně ukázal na obtížnost problematiky identifikace uživatelů webu a jednotlivých sezení. Přínos práce lze vidět v návrhu metodiky procesu přípravy dat a jejím ověření. Dále i v optimalizaci popisovaných algoritmů s využitím entropie pro odhad podílu navigačních stránek webového portálu.

Připomínky k práci a otázky.
Celkově lze práci po formální stránce hodnotit jako zdařilou. Mám zde však pár připomínek.
Autor od začátku práce hovoří o struktuře logovacích souboru a poukazuje na to, že obsahuje nepodstatná a nepotřebná data, nepřesnosti, neúplnost. Bohužel až na str. 20 je uvedeno, co si představuje pod pojmem nepotřebná data. Na str. 24 v textu je odkaz na obrázek rozdělení proměnné Length, ale na obrázku je již RLength. Uvítal bych u matematických výrazů preciznější popis symbolů a důslednější odkaz na literaturu.
Autor se v práci odkazuje na metodiku CRISP-DM, ale příliš ji nedodržuje (například v části porozumění problémů absentuje definování cíle a kritéria jeho naplnění).
Je na škodu věci, že v textu je poměrně značné množství nepřesností a stylistických prohřešků.
Uvítal bych, kdyby na podporu verbálního tvrzení při přípravě dat a popisování modelů a výsledků autor více prezentoval ukázky vstupních souborů a jednotlivých výsledků. Velice často zde hovoří například o pravidlech, ale není zde uvedena ani jedna ukázka výstupních pravidel.

Otázky:
1. Objasněte způsob výpočtu entropie pro každou stránku webu.
2. Upřesněte váš cíl a kritéria pro hodnocení jeho naplnění v souladu s metodikou CRISP-DM.

Závěr.
Disertační práce vyhovuje z hlediska rozsahu, zpracování a přínosů požadavkům kladeným na disertační práci.
Disertační práci hodnotím kladně a doporučuji ji předložit k obhajobě a v případě úspěšné obhajoby navrhuji Mgr. Ľubomírovi Benkovi udělit hodnost Ph.D. v studijním programu Aplikovaná informatika.

V Pardubicích dne 25. 11. 2018