Univerzita Pardubice

Fakulta ekonomicko-správní

Ústav systémového inženýrství a informatiky

Modelovanie správania sa používateľov webu v závislosti od času

Dizertačná práca

Autor: Mgr. Ľubomír Benko

Školiteľ: doc. RNDr. Michal Munk, PhD.

Pardubice 2018

University of Pardubice

Faculty of Economics and Administration

Institute of System Engineering and Informatics

Modelling the behavior of web user depending on time

Thesis

Author: Mgr. Ľubomír Benko

Supervisor: doc. RNDr. Michal Munk, PhD.

Pardubice 2018

Prehlásenie

Prehlasujem, že som túto prácu vypracoval samostatne. Všetky literárne pramene a informácie, ktoré som v práci využil, sú uvedené v zozname bibliografických odkazov.

Bol som oboznámený s tým, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorský zákon, hlavne so skutočnosťou, že Univerzita Pardubice má právo na uzatvorenie licenčnej zmluvy o využití tejto práce ako školského diela podľa § 60 odst. 1 autorského zákona, a s tým, že ak dôjde k využitiu práce mnou alebo bude poskytnutá licencia o využití inému subjektu, je Univerzita Pardubice oprávnená odo mňa požadovať primeraný príspevok na úhradu nákladov, ktoré na vytvorenie diela vynaložila, a to podľa okolností až do ich skutočnej výšky.

Súhlasím s prezenčným sprístupnením mojej práce v Univerzitnej knižnici.

V Pardubiciach dňa 27. 9. 2018 Ľubomír Benko

Poďakovanie

Rád by som sa poďakoval môjmu školiteľovi doc. RNDr. Michalovi Munkovi, PhD., za poskytnuté cenné rady a pripomienky, ochotný prístup a odborné vedenie pri spracovaní mojej dizertačnej práci.

Abstrakt

Dizertačná práca sa zaoberá procesom objavovania znalostí na základe používania webu. Špeciálne prípravou dát, ktorá predstavuje časovo najnáročnejšiu fázu celého procesu a modelovaním dát v závislosti od času, ktoré absentuje v oblasti Web Usage Mining-u. Hlavný cieľ práce je zameraný na modelovanie správania sa používateľov webu v závislosti od času. Podstatnú časť práce tvorí optimalizácia prípravy dát v procese Web Usage Mining-u, špeciálne pre úlohu predikcie. V rámci optimalizácie prípravy dát boli vyhodnotené viaceré metódy identifikácie sedení a ako najvhodnejšia bola vybraná metóda Reference Length. Pre potreby optimalizácie potrebných výpočtov pre metódu Reference Length bol vytvorený nový postup odhadu podielu navigačných stránok za pomoci entropie. V ďalšej časti práce sú vytvorené metodiky na zhodnotenie frekventovaných transakcií/sekvencií v čase a predikcie pravdepodobnosti prístupov na webové časti portálu v závislosti od týždňov. Prínos z hľadiska modelovania dát spočíva v detailnom popise modelu a metodike modelovania správania sa používateľov webu v závislosti od času.

Kľúčové slová

Web Usage Mining, Príprava dát, Reference Length, Data Mining, Asociačné pravidlá, Sekvenčné pravidlá, Multinominálny logitový model.

Abstract

The dissertation thesis deals with the process of knowledge discovery based on the use of the web. Mainly with the data preparation that represents the most time-consuming phase of the entire process. Also it deals with the data analysis depending on the time that is absent in the field of Web Usage Mining. The main aim of the thesis is to model the behaviour of the web site users depending on time. The main part of the thesis consists of the optimization of data preparation in the Web Usage Mining process especially of the prediction task. Several method of session identification were evaluated within the data preparation optimization, and the Reference Length method was the most appropriate. The optimization of the necessary calculations of the Reference Length method consisted of creating a novel approach to estimate the ratio of auxiliary pages using entropy. The following part of the thesis introduces developed methods for evaluation of frequent transactions/sequences in time and prediction of the probability

of accesses to the web parts of the web portal depending on the weeks of the year. Benefits in terms of data analysis are based on a detailed description of the model and the methodology of modelling the web site users' behaviour depending on time.

Keywords

Web Usage Mining, Data preparation, Data Mining, Association rule analysis, Sequence rule analysis, Multinomial logit model.

OBSAH

ZOZNAM OBRÁZKOV

ZOZNAM TABULIEK

ZOZNAM SKRATIEK

BIS	Bank for International Settlements
CRISP-DM	CRoss Industry Standard Process for Data Mining
ELF	Extended Log File
KDD	Knowledge Discovery in Databases
KDT	Knowledge Discovery in Texts
PAD	Počítačová analýza dát
USS	User Session Set
STT	Session Timeout Threshold
VLE	Virtual Learning Environment
WALS	Web Access Log Set
WM	Web Mining
WUM	Web Usage Mining

ÚVOD

Každý používateľ za sebou zanecháva stopy na webových sídlach, ktoré sú zaznamenávané v logovacích súboroch. Cieľom zaznamenávania týchto údajov je zväčša analýza návštevnosti, prípadne v prípade elektronických obchodov aj hlbšia analýza správania sa kupujúcich. Hlbšou analýzou logovacích súborov je možné odhaliť správanie sa používateľov webu, prípadne ich návyky. Automatizáciou zberu dát sa tieto dáta začali používať v procese rozhodovania. Hlavným zdrojom týchto dát sú zväčša databázy, elektronické dokumenty a logovacie súbory (Liu, 2011).

Zaznamenané údaje však často majú slabú výpovednú hodnotu, na základe čoho vznikol koncept objavovania znalostí (Knowledge Discovery) (Fayyad et al., 1996). Pod pojmom objavovanie znalostí môžeme rozumieť proces, ktorý zahŕňa výber dát, predspracovanie dát, transformáciu dát, analýzu dát a interpretáciu výsledkov (Fayyad et al., 1996). Pod pojmom znalosti môžeme rozumieť informácie, ktoré sú pre nás užitočné a pomôžu nám napr. v prípade analýzy webu pochopiť správanie sa návštevníkov na webovom portály.

Medzi najznámejšie oblasti objavovania znalostí patrí objavovanie znalostí z databáz (Knowledge Discovery in Databases, KDD), ktoré by mohlo byť definované ako netriviálne získavanie implicitných, predtým neznámych a potencionálne užitočných informácií z dát (Fayyad et al., 1996). Zdrojom dát v tomto prípade sú produkčné databázy a dátové sklady. Analogicky k objavovaniu znalostí z databáz patrí získavanie údajov z textov, pričom sa tento proces nazýva text miningom, resp. objavovanie znalostí z textov (Knowledge Discovery in Texts, KDT) (Hearst, 1999). Podobne ako pri KDD slúžia aj pri KDT k analýze dát štatistické metódy a metódy strojového učenia, pričom najväčšie rozdiely sú v samotnej príprave dát, čiže reprezentácií textu, ktorej sa venovali viacerí autori (Loh et al., 2000; Schmidt et al., 2013; Jin a Srihari, 2007; Jindal a Shweta, 2018; Justicia de la Torre et al., 2018; Tan, 1999; Benko a Munková, 2016; Munk et al., 2016).

V dnešnej dobe je výrazným zdrojom dát hlavne Internet, ktorý predstavuje najdynamickejšie sa rozvíjajúci zdroj informácií. Z nutnosti analýzy aj tohto druhu dát vznikla príbuzná oblasť objavovania znalostí z databáz – objavovanie znalostí z webu (Web Mining, WM) (Liu, 2011). Definícia WM by mohla byť chápaná ako extrakcia

zaujímavých a potencionálne užitočných znalostí a informácií z aktivít súvisiacich s webom (Liu, 2011). Na základe skúmaného druhu dát v procese extrakcie sa WM kategorizuje do troch typov: objavovanie znalostí na základe štruktúry webu (Web Structure Mining), obsah webu (Web Content Mining) a používania webu (Web Log Mining alebo sa zvykne používať aj termín Web Usage Mining) (Liu, 2011).

Najväčšie rozdiely medzi oblasťami objavovania znalostí pri riadení procesu metodikou CRISP-DM sú vo fáze prípravy dát (Data Preparation), pričom práve príprava dát predstavuje časovo najnáročnejšiu fázu v rámci celého procesu objavovania znalostí (Cios et al., 2007; Liu, 2011; Rajenderan, 2012; Zhang et al., 2003). Medzi najnáročnejšie zdroje dát z hľadiska prípravy dát patrí logovací súbor webového servera. Dôvodom je hlavne veľké množstvo zozbieraných nepodstatných údajov a ich nepresnosť, prípadne neúplnosť. V tejto práci budú spomenuté postupy a metodika spracovania tohto druhu vstupných dát a snaha o zefektívnenie práce s nimi.

Záverečná práca pozostáva zo štyroch kapitol: v prvej kapitole je definovaný hlavný cieľ dizertačnej práce a čiastočné ciele potrebné pre jeho dosiahnutie. Druhá kapitola sa zaoberá súčasným stavom riešenej problematiky objavovania znalostí na základe používania webu, pričom je zameraná hlavne na fázu prípravy dát a modelovanie dát. Zároveň sa v druhej kapitole popisuje metodika záverečnej práce rozdelená do jednotlivých častí vychádzajúc z metodiky CRISP-DM. Tretia kapitola predstavuje dosiahnuté výsledky experimentov. Štvrtá kapitola prezentuje vyhodnotenie výsledkov záverečnej práce a ich prínos pre prax.

1 MOTIVÁCIA A CIEĽ PRÁCE

Dizertačná práca je zameraná na proces objavovania znalostí na základe používania webu (Web Usage Mining - WUM). Logovací súbor, ako zdroj dát webového portálu, predstavuje časové dáta vo forme prístupov na webové stránky. Napriek tomu absentuje v oblasti WUM úloha predikcie, kde sa využívajú časové premenné. Výnimku tvoria sekvenčné pravidlá (vzory), v ktorých vystupuje časová premenná iba v obmedzenej forme a slúži len na určenie poradia navštívených stránok v identifikovaných sedeniach. Z tohto dôvodu sme sa rozhodli skúmať správanie sa používateľov webového portálu v závislosti od času a navrhnúť prístupy k riešeniu úlohy predikcie, ktorá v tejto oblasti absentuje.

Zdroj dát (logovací súbor) o používaní webu je bez adekvátneho predspracovania nepoužiteľný pretože obsahuje veľké množstvo nepotrebných, irelevantných, nepresných a neúplných informácií. Z toho dôvodu sme sa zamerali aj na časovo najnáročnejšiu fázu prípravy dát, pričom sme sa snažili zohľadniť špecifiká predikcie. Z týchto dvoch hľadísk vyplýva hlavný cieľ práce. Hlavným cieľom práce je návrh prístupov a metodík k modelovaniu správania sa používateľov webu v závislosti od času. Podstatnú časť práce tvorí optimalizácia prípravy dát v procese WUM, špeciálne pre úlohu predikcie.

Na dosiahnutie hlavného cieľa práce je potrebné riešiť nasledovné úlohy:

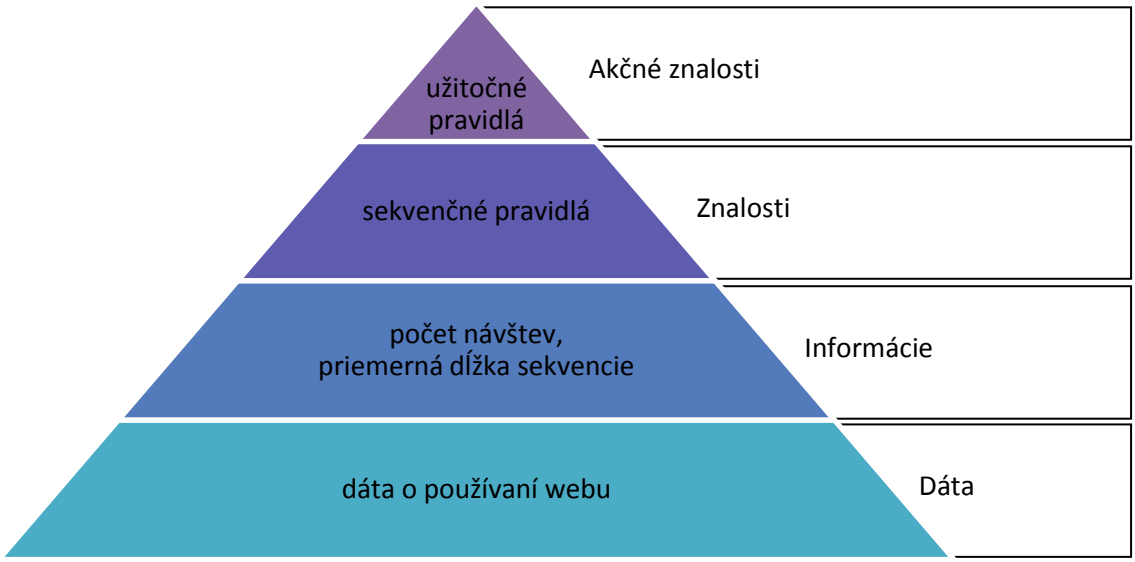
- návrh a implementácia algoritmov identifikácie sedení (časovo/štruktúrovo orientované heuristiky) a dopĺňania ciest;
- optimalizácia algoritmov identifikácie sedení a dopĺňania ciest;
- experimentálne vyhodnotenie relevantnosti optimalizovaných algoritmov predspracovania dát z logovacieho súboru webového servera;
- návrh metodiky predspracovania dát o používaní webu zohľadňujúcej špecifiká predikcie.
- návrh metód predikcie, kde táto úloha objavovania znalostí absentuje v oblasti WUM;
- návrh metodiky na zhodnotenie frekventovaných transakcií/sekvencií v čase;

- návrh metodiky predikcie pravdepodobnosti prístupov na webové časti portálu v závislosti od týždňov, pričom budeme skúmať aj rôzne časové obdobia;
- optimalizácia algoritmov modelovania a evalvácia multinominálneho logitového modelu zohľadňujúcich špecifiká webu.

Predpokladaný prínos záverečnej práce je možné hodnotiť z hľadiska prípravy a modelovania dát, pričom z hľadiska prípravy dát spočíva prínos v návrhu metodiky a odporúčaní pre získanie spoľahlivých dát z logovacieho súboru. Prínos z hľadiska modelovania dát spočíva v detailnom popise modelu a metodike modelovania správania sa používateľov webu v závislosti od času. Súčasťou metodiky bude aj popis možností využitia získaných znalostí.

2 SÚČASNÝ STAV RIEŠENEJ PROBLEMATIKY

Problematika Web Usage Mining a skúmania správania sa používateľov webu je súčasťou viacerých výskumov. WUM v sebe zahŕňa pochopenie správania sa používateľov, keď používajú webové stránky. Podobnú filozofiu je možné použiť aj pre používateľov informačných systémov, ktorých správanie sa v systéme môže odhaliť prípadné chyby alebo prispieť k vylepšeniu systému. Na zaznamenávanie stôp, či už na webových sídlach alebo v informačných systémoch, slúžia logovacie súbory. Skúmanie logovacích súborov odhalí nielen správanie, ale aj návyky používateľov. Keďže sa v logovacích súboroch zaznamenávajú hlavne anonymné údaje, je nutné ich spracovať a pripraviť na analýzu, na čo slúžia metódy predspracovania dát. Predspracovanie dát je dôležitou súčasťou WUM a pre tento účel bolo navrhnutých množstvo techník predspracovania. Cieľom objavovania znalostí na základe používania webu je analýza správania sa používateľov pri prechádzaní webu (Srivastava et al., 2000; Romero et al., 2009).



Obrázok 1 Princíp objavovania znalostí (Zdroj: (Munk a Kapusta, 2014)*)*

Princíp objavovania znalostí (Obrázok 1) je možné priblížiť na objavovaní vzorov správania sa používateľov webu. Podľa Fayyada (Fayyad et al., 1996) môžeme objavovanie znalostí chápať ako proces, ktorý v sebe zahŕňa výber dát, predspracovanie dát, transformáciu dát, analýzu dát a interpretáciu výsledkov. Dáta o používaní webu sa zaznamenávajú do logovacieho súboru webového servera, kde z veľkého objemu dát je možné získať informácie pre lepšie porozumenie dátam. Medzi tieto informácie môžu

patriť napríklad štatistiky počtu prístupov za dané časové obdobie, počet návštev alebo priemerná dĺžka návštev na webe a pod. Výsledkom sekvenčnej analýzy sú sekvenčné pravidlá (sequence rules), ktoré reprezentujú získané znalosti, pričom sa od nájdených pravidiel očakáva nielen jasnosť, ale aj užitočnosť. Len časť objavených znalostí (vzorov správania sa používateľov webu) je možné použiť z hľadiska aplikácie a zvyšné pravidlá sú z hľadiska užitočnosti triviálne, resp. nevysvetliteľné, čiže nepoužiteľné, neprinášajúce žiadne nové znalosti (Berry a Linoff, 2004). Na základe užitočných pravidiel je možné následne identifikovať chyby v navigácií, upravovať odkazy a iné nepresnosti, resp. identifikovať správanie sa používateľov webového portálu.

Prvou fázou v procese objavovania znalostí je porozumenie problematike. Úlohou je pochopiť ciele problému formulovaného z hľadiska modelovania dát. Medzi úlohy objavovania znalostí patrí deskripcia dát a sumarizácia, segmentácia, deskripcia konceptov, klasifikácia, predikcia a analýza závislostí (Liu, 2011). V druhej fáze je cieľom získanie relevantných dát o používaní webu. Zdrojom dát sú dáta o používaní webu, prípadne informačných systémov a pod. Informačné systémy zväčša evidujú údaje o používaní systému vo vlastnej štruktúre, na čo sa používa prevažne databáza. V prípade webových a proxy serverov sú dáta zaznamenávané v spoločnej štandardnej štruktúre v textovom formáte – logovací súbor. Logovací súbor v štandardnej štruktúre – Common Log File (W3C, 1995) zaznamenáva informácie o IP adrese, čase a dátume návštevy a pristupovanom objekte, v prípade rozšírenej podoby (Extended Log File – ELF) dokážeme zaznamenávať aj údaje (Obrázok 2) o odkazovanom objekte a verzii prehliadača používateľa – User Agent (Liu, 2011).

```
84.10.169.115 - - [12/Dec/2016:08:12:22 +0200]
"GET /prijimaciekonanie/moznosti-vs-studia HTTP/1.1" 200 5378
"http://www.ukf.sk/prijimaciekonanie/podmienky-prijatia" "Mozilla/5.0 (Windows NT 5.1)
AppleWebKit/535.1 (KHTML, like Gecko) Chrome/14.0.835.202 Safari/535.1"

...
```

Obrázok 2 Jeden záznam z logovacieho súboru

Ďalším potrebným zdrojom dát v príprave dát je aktuálna mapa webu. Mapa webu obsahuje informácie, či existuje medzi stránkami väzba, čiže či existuje medzi stránkami hypertextový odkaz z jednej stránky na druhú. Najčastejším spôsobom získania mapy webu je pomocou web crawlingu implementovaného v data miningových nástrojoch. Avšak tým, že sú webové portály dynamické a neustále sa menia, tak získať historické

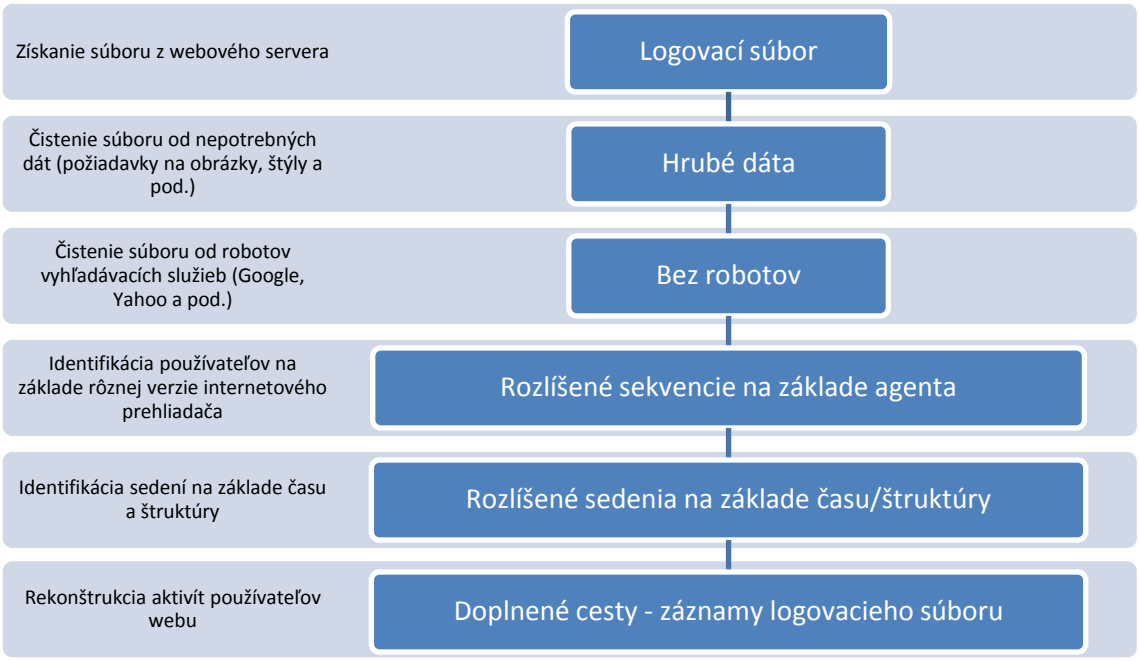
dáta korešpondujúce so skúmaným logovacím súborom, môže byť problematické. Preto alternatívnou metódou je vygenerovanie mapy webu zo samotného logovacieho súboru.

2.1 PRÍPRAVA DÁT

Len z kvalitných dát je možné vykonať dobrú analýzu dát, pričom práve logovacie súbory sú typické tým, že obsahujú značné množstvo nepodstatných údajov, ktoré by mohli analýzu dát pokaziť. V prípade skúmania správania sa používateľov alebo návštevníkov webového portálu je možné použiť nasledovné metódy (Munk et al., 2010b):

- výberové zisťovanie – zisťujú sa odpovede na konkrétne položky dotazníka a návštevník webu si je vedomý predmetu skúmania (Cerna a Poulova, 2008),
- web usage mining – analyzuje sa logovací súbor webového servera, ktorý obsahuje informácie o prístupoch na stránky webového portálu a návštevník o tomto skúmaní nevie, pričom sú jeho údaje do istej miery anonymné (Cooley et al., 1999).

Predpokladom pre prácu s kvalitnými údajmi nie je len ich zber, ale aj príprava pre ďalšie analýzy. Príprava dát sa v angličtine označuje pojmami data preprocessing alebo data preparation. Práve príprava dát patrí medzi časovo aj prostriedkovo najnáročnejšiu fázu v procese objavovania znalostí, hlavne z dôvodu nepodstatných údajov, ktoré sa nachádzajú v logovacích súboroch. Losarwar a Joshi (Losarwar a Joshi, 2012) analyzovali prípravu dát vo WUM a dospeli k záveru, že v oblasti analýzy webu je príprava dát veľmi dôležitá a vyžaduje použitie nástrojov, ktoré sa typicky pre prípravu dát v iných doménach nepoužívajú. V prípade portálov virtuálneho vzdelávacieho prostredia (VLE) prišli autori v článku (Sael et al., 2013) s vlastnou úpravou logovacieho súboru, čím minimalizovali nutnosť prípravy dát a rovno extrahovali všetky potrebné údaje pre analýzu. Napriek tomu, toto riešenie nie je použiteľné v prípade portálov s anonymným prístupom, kde je nutné postupovať klasickým procesom prípravy dát. Názorný prehľad aplikovania techník prípravy dát z logovacieho súboru webového servera poskytuje nasledovná schéma (Obrázok 3).



Obrázok 3 Techniky prípravy dát pre logovací súbor v štruktúre ELF (Zdroj: (Munk a Kapusta, 2014)*)*

2.1.1 Čistenie dát

Čistenie dát od nepotrebných údajov patrí k prvým krokom v príprave dát, pričom je špecifické pre každý webový portál, prípadne dátový zdroj. Cieľom je odstránenie záznamov, resp. odkazov, ktoré nie sú podstatné pre skúmanie správania sa používateľov (Cooley et al., 1999). Medzi takéto odkazy patria hlavne prístupy k obrázkom, flash videám, ikonám kurzora, javascriptom alebo štýlom. Zvyčajný postup identifikácie takýchto záznamov zahŕňa identifikáciu na základe prípony (*.jpg, *.jpeg, *.bmp, *.png, *.gif, *.css, *.js, *.flw, *.swf, *.cur, *.rss, *.ico, *.xml a podobne). Aj pri načítaní len jednej stránky sa všetky tieto požiadavky zapíšu do logovacieho súboru. Okrem požiadavky GET sa tiež do logovacieho súboru zapisujú aj ďalšie požiadavky http protokolu, pričom je potrebné odstrániť aj návratové kódy 4xx/5xx, ktoré identifikujú chybu klienta/servera. Aye vo svojom článku (Aye, 2011) predstavil dva algoritmy pre získavanie dát z databáz a ich následné čistenie v oblasti WUM. Algoritmus pre čistenie dát v sebe zároveň zahŕňal zobrazenie informácie o počte zmazaných údajov a identifikácií počtu unikátnych prístupov na skúmaný webový portál. Srivastava et al. vo svojom článku (Srivastava et al., 2015) predstavili algoritmus určený na čistenie logovacieho súboru od nepotrebných dát, pričom využívajú daný časový interval a taktiež dokážu zoradiť záznamy podľa ich časovej známky. Napriek tomu má predstavený algoritmus problém s veľkým objemom dát a v prípade čistenia väčších logovacích

súborov dochádza k značnému spomaleniu. Spomínaní autori sa nezaoberali čistením logovacích súborov od prístupov robotov vyhľadávacích služieb.

Ďalším krokom čistenia je odstránenie prístupov robotov vyhľadávacích služieb ako napríklad Google, Yahoo, Bing a pod. Keďže roboty pristupujú k webovému portálu sekvenčne, tak nie je vhodné zahrnúť ich aktivitu do skúmania správania sa používateľov. Detekcia robotov prebieha buď na základe ich identifikácie v poli User Agent, alebo na základe IP adresy, ktorú je možné porovnať s databázou robotov (napr. www.robotstxt.org). Autori v článku (Vellingiri a Chenthur Pandian, 2011) sa sústredili na zlepšenie techník na čistenie dát, hlavne na fázu odstraňovania prístupov robotov. Okrem už vyššie spomínaných nepotrebných dát a prístupov robotov odstránili z logovacieho súboru aj všetky prístupy, ktoré mali dĺžku prístupu kratšiu ako dve sekundy. Nithya a Sumathi (Nithya a Sumathi, 2012) predstavili nový prístup odstraňovania prístupov robotov vyhľadávacích služieb, pričom svoje postupy úspešne testovali na anonymnom Microsoft Web Dataset-e a MSNBC.com Anonymous Web Dataset-e. Okrem odstraňovania prístupov robotov sa zameriavali aj na odstránenie lokálneho a globálneho „šumu," resp. znečistenia logovacieho súboru nepotrebnými dátami. Nami navrhnutý algoritmus, ktorý v sebe zahŕňa čistenie dát a zároveň aj odstránenie robotov z logovacieho súboru, prezentujeme v kapitole výsledkov práce, pričom vychádzame z publikovaného článku (Munk et al., 2015).

2.1.2 IDENTIFIKÁCIA POUŽÍVATEĽOV

Keďže sa v logovacom súbore prioritne zaznamenávajú anonymné údaje o používateľoch, tak vzniká problém s jednoznačnou identifikáciou návštevníka webu. Pri analýze nie je potrebné poznať konkrétnu identitu používateľa, ale rozlišovať medzi jednotlivými používateľmi. Avšak predpoklad, že na identifikáciu používateľa stačí IP adresa, je nesprávny, pretože za jednou IP adresou sa môže nachádzať viacero používateľov. Z toho dôvodu je nutné skombinovať viaceré metódy, ako napríklad využitie poľa Cookie (Pabarskaite a Raudys, 2007), prípadne kombinácie IP adresy s poľom User Agent (Srivastava et al., 2000). Viaceré heuristické metódy využívajú hlavne kombináciu IP adresy s poľom User Agent. Ak príde k zmene IP adresy, tak je zrejmé, že je to nový používateľ. Ak je IP adresa rovnaká, tak sa porovnáva pole User Agent, ak príde k zmene, tak je identifikovaný nový používateľ, v opačnom prípade ide o toho istého používateľa (Srivastava et al., 2000). V prípade, že portál vyžaduje

od používateľa registráciu, resp. prihlásenie, tak je identifikácia používateľov zjednodušená, pretože o tom je záznam v logovacom súbore. Podrobnejšej analýze možností identifikácie sedení sa venovali autori v (Pabarskaite a Raudys, 2007; Varnagar et al., 2013), kde jednou z ďalších možností identifikácie sedení bola analýza sekvencie návštev stránok na webovom portály. Bol stanovený predpoklad, že ak na nasledujúcu webovú stránku neexistuje prepojenie zo súčasnej, tak sa musí jednať o nového používateľa.

2.1.3 IDENTIFIKÁCIA SEDENÍ

Používateľ môže navštíviť stránku viackrát, v logovacom súbore sú zaznamenané viacnásobné sedenia (návštevy) pre každého používateľa. Cieľom identifikácie sedení je rozdeliť jednotlivé prístupy každého používateľa do oddelených relácií (Cooley et al., 1999). Sedenie môže byť definované ako postupnosť krokov, ktoré vedú k naplneniu určitej úlohy (Spiliopoulou a Faulstich, 1999) alebo ako postupnosť krokov, ktoré vedú k dosiahnutiu určitého cieľa (Ming-Syan Chen et al., 1998).
Na identifikáciu sedení sa používajú štruktúrovo-orientované heuristiky, časovo-orientované heuristiky(Liu, 2011; Berendt et al., 2003), ako aj kombinácie týchto dvoch prístupov (Munk a Kapusta, 2014).

Identifikácia sedení pomocou štruktúrovo-orientovanej heuristiky

Ak sa v sekvencií sedenia z jednej IP adresy objaví priamy nasledovník stránka, ktorá nie je priamo dosiahnuteľná z predchádzajúcej stránky, tak sa pravdepodobne jedná o iné sedenie používateľa s rovnakou IP adresou (Cooley et al., 1999). Problémom spomínanej heuristiky je možnosť využitia návratového tlačidla v prehliadači, ktoré sa do logovacieho súboru nezapíše, a tak mohol používateľ prejsť na stránku, ktorá s predchádzajúcim záznamom nesúvisela. Rovnaký problém môže predstavovať aj využitie obľúbených záložiek používateľa na prechádzanie medzi stránkami portálu (Berendt a Spiliopoulou, 2000).

Identifikácia sedení pomocou časovo-orientovanej heuristiky

Cooley et al. (Cooley et al., 1999) predstavili časovo orientovanú heuristiku nazvanú h1, ktorá vytvára sedenia považované za sériu kliknutí v priebehu 30 minút. Na druhej strane Spiliopoulou et al. (Spiliopoulou et al., 2003) odporučili identifikovať sedenia na základe odhadu dĺžky času sedenia 10 minút a nazvali to heuristikou h2. Ak máme odhad

dĺžky času sedenia θ, tak sedenie je sekvencia navštívených stránok s časovou známkou, pre ktorú platí:

$$USS = \langle USID, \langle URL_1, DTime_1 \rangle, \dots, \langle URL_k, DTime_k \rangle \rangle, DTime_k - DTime_1 \leq \theta, \quad (1)$$

kde USS (User Session Set) je množina sedení, $DTime_k$ je časová známka posledného záznamu sedenia, $1 \leq k \leq n$ a n je počet záznamov vo $WALS$ (Web Access Log Set - množina webových prístupov zachytená v logovacom súbore a zoradená podľa položky $DTime$) (Cooley et al., 1999). Všetky ďalšie záznamy, ktorých časová známka je väčšia ako $DTime_1 + \theta$ patria do ďalšieho sedenia.

Na určenie konca sedenia a začiatku nového sa často používa aj tzv. časové okno (Session Timeout Threshold, STT). STT je preddefinovaná doba neaktivity, ktorá umožňuje webovým aplikáciám určiť kedy sa začína nové sedenie (Munk a Drlik, 2011b; Kapusta et al., 2014a). Catledge a Pitkow zvolili časové okno (STT) vypočítané na základe priemerného času stráveného na webových stránkach, plus 1,5 násobok smerodajnej odchýlky času stráveného na webových stránkach (Catledge a Pitkow, 1995).

Autori v (Munk et al., 2013b; Kapusta et al., 2013, 2014b) sa prikláňajú k odhadom na základe kvartilového rozpätia, ktoré nie sú ovplyvnené odľahlými hodnotami, napr. $Q_{III} + 1{,}5Q$, kde Q_{III} je horný kvartil (75. percentil) a Q je kvartilové rozpätie (stredných 50 % hodnôt), čiže ak je čas na stránke považovaný za odľahlú hodnotu, začína sa nové sedenie.

Identifikácia sedení pomocou kombinácie štruktúrovo a časovo orientovanej heuristiky

Medzi tieto heuristiky patrí napríklad metóda, ktorá zohľadňuje čas prístupu a odkazujúcu stránku. V logovacích súboroch sa môže zaznamenávať aj stránka, z ktorej používateľ prišiel na aktuálne zobrazenú stránku, tzv. referrer, resp. reference page. To je možné využiť na identifikáciu sedení použitím heuristickej metódy h-ref (Spiliopoulou et al., 2003). Ak máme odhad veľkosti časového okna δ, tak sedenie je sekvencia navštívených stránok s časovou známkou, pre ktorú platí:

$$USS = \langle USID, \langle URL_1, DTime_1, Referrer_1 \rangle, \dots, \langle URL_k, DTime_k, Referrer_k \rangle \rangle, \quad (2)$$

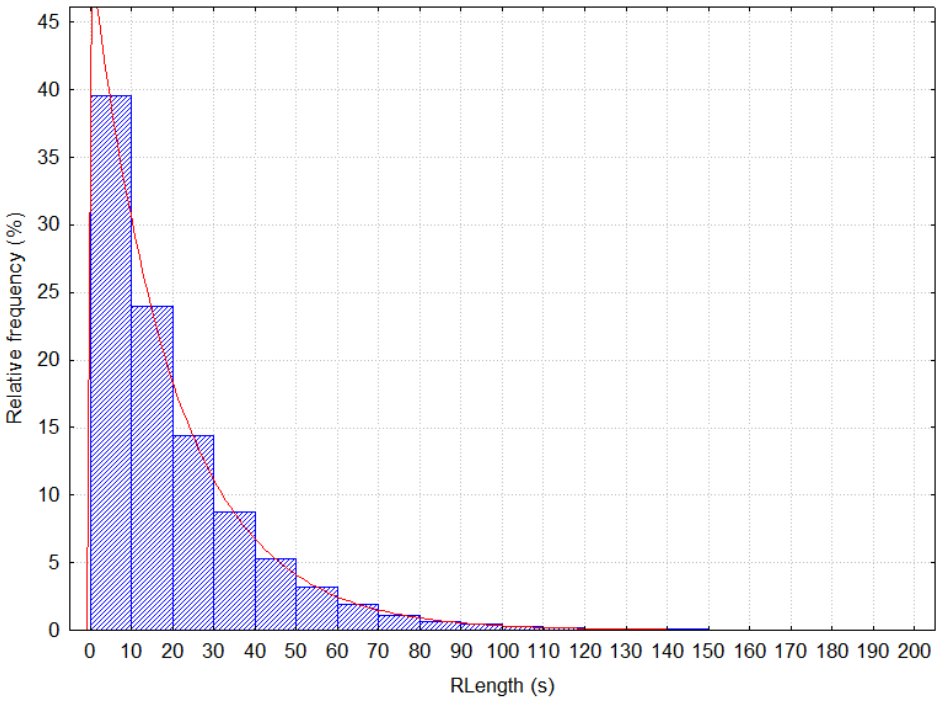
$$Referrer_i = URL_{i-1}, \quad (3)$$

alebo rovnosť neplatí, resp. $Referrer_i$ nie je definovaný, ale:

$$DTime_i - DTime_{i-1} \leq \delta, \tag{4}$$

kde $DTime_k$ je časová známka posledného záznamu sedenia a $1 < i \leq k$.

Metóda Reference Length je založená na predpoklade, že dĺžka času stráveného používateľom na stránke je vo vzťahu s tým, či je stránka klasifikovaná ako obsahová alebo navigačná (Munk a Kapusta, 2014; Kapusta et al., 2012). Na obrázku (Obrázok 4Obrázok 4) je znázornený histogram popisujúci rozdelenie premennej Length, ktorá slúži na reprezentáciu času stráveného na stránke webového portálu. Predpokladá sa, že ľavá strana grafu predstavuje navigačné stránky. Tie slúžia návštevníkom hlavne na rýchly prechod k obsahovým stránkam, ktoré sú ich cieľom. Pravú stranu preto tvoria stránky s obsahom, ktorých dĺžka stráveného času má väčší rozptyl.



Obrázok 4 Rozdelenie premennej RLength (Zdroj: (Munk a Benko, 2018)*)*

Na základe predpokladu exponenciálneho rozdelenia premennej je možné vypočítať hraničný čas C, ktorý slúži na rozlíšenie navigačných stránok od obsahových. Premenná $RLength$ má exponenciálne rozdelenie

$$f(RLength) = \lambda e^{-\lambda RLength}, \tag{5}$$

$$F(RLength) = 1 - e^{-\lambda RLength}, \quad (6)$$

kde $RLength \geq 0$.

Ak je p relatívna početnosť navigačných stránok, tak sa na odhad hraničného času C využije kvantilová funkcia

$$F^{-1}(p, \lambda) = C = \frac{-\ln(1-p)}{\lambda}, \quad (7)$$

pre $0 \leq p < 1$. Maximálne vierohodný odhad parametra λ (priemerná intenzita udalostí) je

$$\hat{\lambda} = \frac{1}{\overline{RLength}}, \quad (8)$$

kde $\overline{RLength}$ je pozorovaný priemer dĺžky návštev.

V okamihu keď je odhadnutý hraničný čas, sedenie môže byť identifikované porovnaním každého času stráveného na stránke s hraničným časom, pričom práve hraničný čas rozdelí stránky na navigačné a obsahové podľa dĺžky času stráveného na konkrétnej stránke (Munk a Benko, 2018; Munk et al., 2015). Následne je sedenie sekvencia navštívených stránok s časovou známkou, pre ktorú platí:

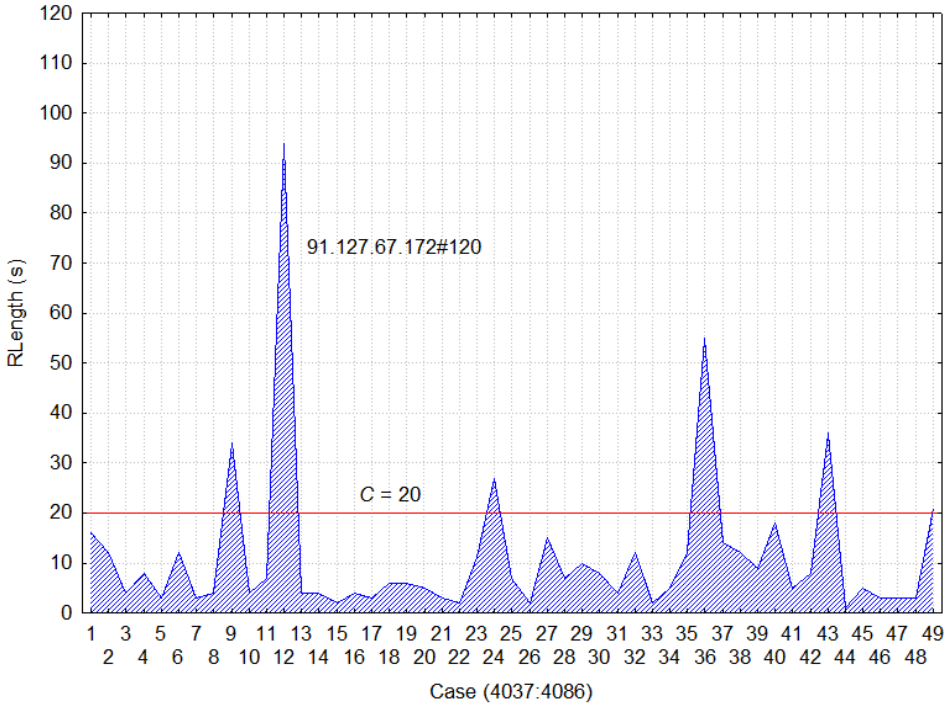
$$\langle USID, \langle URL_1, DTime_1, RLength_1 \rangle, \ldots, \langle URL_k, DTime_k, RLength_k \rangle \rangle, \quad (9)$$

$$RLength_i \leq C, \quad (10)$$

kde $1 \leq i < k$ a pre poslednú stránku sedenia platí:

$$RLength_k > C. \quad (11)$$

Od stránky s vlastnosťou (11) je definované nové sedenie, pričom prvých $k - 1$ stránok je klasifikovaných ako navigačné stránky a posledná k-tá stránka je klasifikovaná ako obsahová.



Obrázok 5 Metóda Reference Length (Zdroj: (Munk a Benko, 2018)*)*

Na obrázku (Obrázok 5) vidíme sekvenciu navštívených stránok z danej IP adresy a agenta, ktorá je usporiadaná podľa času prístupu (os *x*) a čas strávený na stránke (os *y*). Hraničný čas bol 20 sekúnd, kde prvé sedenie je tvorené stránkami 1 až 9, pričom prvých 8 je klasifikovaných ako navigačné a posledná je obsahová. Analogicky sa postupuje pri identifikácií ďalších sedení. Dĺžka času stráveného na stránke je daná rozdielom prístupových časov súčasnej stránky a nasledujúcej, pričom sa nedá vypočítať čas poslednej stránky v sekvencii. Metóda Reference Length predpokladá, že každá posledná stránka je stránka obsahová. Môže sa však stať, že z dôvodu neočakávanej udalosti na strane návštevníka (napr. telefonát) je obsahová stránka klasifikovaná ako navigačná stránka. Rovnako je potrebné vziať do úvahy skutočnosť, že pre každého používateľa môže byť každá stránka inak klasifikovaná, pre jedného to môže byť navigačná stránka, ale pre druhého obsahová a naopak.

Viacerí autori využívajú rôzne metódy identifikácie sedení v oblasti WUM s rôznym úspechom a spoľahlivosťou. Výber správnej metódy často záleží aj od skúmanej domény. Autori v článku (Abdullah et al., 2014) navrhli sekvenčný model a nástroj, ktorý použili na vytvorenie sekvenčného dátového setu. Vo svojom výskume sa zamerali na logovací súbor extrahovaný z VLE, konkrétne z MySQL databázy. Na základe dosiahnutých výsledkov môže pomocou ich nástroja byť vygenerovaný dátový set použitý v rôznych

data miningových nástrojoch bez nutnosti viacerých úprav. Arce et al. (Arce et al., 2014) predstavili prístup Simulated Annealing, ktorý je zameraný na riešenie problému identifikácie sedení, pričom ich cieľom bolo hlavne zvýšiť rýchlosť spracovania veľkého množstva dát. Podarilo sa im zvýšiť rýchlosť spracovania niekoľkonásobne, ale navrhovaný prístup dosahuje spoľahlivé výsledky až po viacnásobnom vykonaní na rovnakom sete dát.

2.1.4 REKONŠTRUKCIA AKTIVÍT POUŽÍVATEĽOV WEBU

Cieľom rekonštrukcie aktivít je určiť, či existujú významné prístupy na web, ktoré nie sú zaznamenané v logovacom súbore. Medzi takéto aktivity môže patriť opätovný návrat na stránku počas toho istého sedenia, ktorú zväčša webový prehliadač načíta z dočasnej (cache) pamäte. Rovnako je vo webových prehliadačoch často využívané tlačidlo späť, ktoré sa taktiež nezaznamenáva v logovacom súbore. Bolo dokázané, že viac ako 50 % prístupov na webe je práve pohyb späť (Tauscher a Greenberg, 1997). Riešením tohto problému je práve rekonštrukcia aktivít, alebo taktiež nazývaná aj dopĺňanie ciest (Cooley et al., 1999; Liu, 2011). Hlavnú rolu v dopĺňaní ciest zohráva hlavne mapa webu webového portálu. Na základe mapy webu je možné zistiť, či medzi stránkami existuje väzba, čiže hypertextový odkaz. Algoritmy pre dopĺňanie ciest zväčšia fungujú na rovnakom princípe, jeden z navrhovaných algoritmov bol predstavený autormi (Li et al., 2008). Navrhovaný algoritmus efektívne doplnil cesty a zvýšil spoľahlivosť skúmaných dát na klasickom logovacom súbore webového portálu.

2.2 MODELOVANIE DÁT

Cieľom aplikácie analytických metód je získanie nových znalostí. Vstupom do analytických nástrojov sú predspracované, prípadne upravené údaje a výstupom sú znalosti. Výber analytickej metódy závisí od cieľa, pre ktorý je model určený. V metodike CRISP-DM (Cross-Industry Standard Process for Data Mining) sa uvádza šesť typov problémov (deskripcia dát a sumarizácia; segmentácia; deskripcia konceptov; klasifikácia; predikcia; analýza závislostí) a k nim odporúčané metódy (Chapman et al., 2000). Medzi najčastejšie z uvedených typov patria klasifikácia a predikcia, kde oba typy si dávajú za cieľ nájsť znalosti použiteľné pre klasifikáciu nových prípadov. Rozdielom medzi klasifikáciou a predikciou je podstatná úloha času v prípade predikcie (Berka, 2003).

Jedna z metód riešenia problémov predikcie a klasifikácie je logistická regresia, ktorá popisuje závislosť kvalitatívnej závislej premennej od jednej alebo viacerých kvantitatívnych nezávislých premenných (Munk et al., 2011a, 2011b). Pri logistickej regresii sa modeluje pravdepodobnosť, že premenná má konkrétnu hodnotu v závislosti na kombinácií hodnôt nezávislých premenných. Výsledky logistickej regresie sú dobre interpretovateľné, ale na druhej strane si vyžadujú dôslednú prípravu dát. Logistická regresia patrí medzi zovšeobecnené lineárne modely a zároveň je podobne ako lineárna regresia založená na štatistickom rozdelení (Stankovičová a Vojtková, 2007). Rozdiel je v tom, že závislá premenná nie je spojitá, ale je diskrétna. Aby bolo možné použiť regresiu, závislá premenná sa transformuje na spojitú hodnotu, ktorá je funkciou pravdepodobnosti výskytu udalosti (Stankovičová a Vojtková, 2007). Nasledujúci popis sa odvoláva prevažne na publikácie (Stankovičová a Vojtková, 2007; Anděl, 2007).

Označme X ako vektor nezávislých premenných. Označme Y ako binárnu závislú premennú, ktorá nadobúda dve možné hodnoty:

- 0 – neúspech, resp. želaná udalosť nenastala;
- 1 – úspech, resp. želaná udalosť nastala.

Označme symbolom p pravdepodobnosť výskytu želanej udalosti ($Y = 1$), čiže modelovanú hodnotu. Potom (12) je podmienená pravdepodobnosť výskytu želanej udalosti za podmienky výskytu vektora nezávislých premenných $\boldsymbol{X}$.

$$p = P(Y = \frac{1}{\boldsymbol{X}}) \tag{12}$$

Podiel $\frac{p}{1-p}$ vyjadruje šancu výskytu želanej udalosti k výskytu neželanej udalosti. Šancu a pravdepodobnosť predstavujú tie isté informácie, len v iných podobách. Vzťah medzi pravdepodobnosťou p a vysvetľujúcimi premennými $\boldsymbol{X}$ nie je lineárny.

Pre finálnu transformáciu závislej premennej na spojitú premennú sa vypočíta logaritmus šancí, tzv. logit, pre ktorý platí: $logit(p) = \ln(\frac{p}{1-p})$. Vzťah medzi logitom a vektorom vysvetľujúcich premenných má už lineárny charakter. Rovnica logistického modelu má tvar (13).

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{13}$$

Vzťah medzi pravdepodobnosťou a vektorom vysvetľujúcich premenných dostaneme spätnou transformáciou a má nelineárny charakter. Je to druh exponenciálnej funkcie (14).

$$p = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k}} = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)}} \qquad (14)$$

Logitovou transformáciou dostávame z nelineárnej závislosti lineárnu závislosť. Kým hodnoty pravdepodobností p_i sú z intervalu $(0; 1)$, hodnoty podielov $\frac{p_i}{1-p_i}$ sú nezáporné hodnoty, hodnoty logitov môžu nadobúdať akékoľvek reálne hodnoty z intervalu $(-\infty; \infty)$.

2.2.1 MODELOVANIA SPRÁVANIA SA POUŽÍVATEĽOV WEBU V ZÁVISLOSTI OD ČASU

Používanie portálov je reprezentované hlavne časovými dátami. Najčastejšie sa časová premenná vyskytuje zväčša len pri extrakcii sekvenčných pravidiel, kde však určuje iba poradie navštívených webových častí. V aplikačnej oblasti absentuje modelovanie správania sa používateľov webu v závislosti od času. Napriek tomu sa snaží viacero autorov sledovať správanie inými spôsobmi. Arbelaitz et al. sa v článku (Arbelaitz et al., 2013) venovali analýze a vytváraniu navigačných profilov návštevníkov turistického webového portálu. Navrhli systém, ktorý s úspešnosťou 60 % dokáže vytvoriť profily, ktoré zodpovedajú reálnym navigačným sekvenciám návštevníkov. Cieľom autorov bolo využiť navigačné profily pre lepšiu personalizáciu webového portálu pre návštevníkov. Využili základné princípy metodiky CRISP-DM na to, aby vytvorili segmenty používateľov so spoločnými záujmami, ale plánujú v budúcnosti rozšíriť navrhovaný program o princíp fuzzy množín. Ďalší autori (Anitha, 2010; Bhawsar et al., 2012) sa tiež zamerali na snahu odhalenia, resp. predikcie ďalšieho kroku návštevníkov webového portálu. Makkar et al. (Makkar et al., 2010) využili na predikciu správania sa používateľov webového portálu Petriho siete, pričom kombinovali informácie získané z logovacieho súboru a štruktúry webového portálu. Carmona et al. (Carmona et al., 2012) sa zamerali na vytvorenie metodiky pre oblasť webových portálov elektronickej komercie, pričom k údajom nepristupovali sekvenčne, ale pomocou nástroja Google Analytics. Následne extrahované dáta spracovali pomocou zhlukovania, asociačnej analýzy a objavovania podskupín. Na základe dosiahnutých výsledkov potom stanovili odporúčania a identifikovali problémové oblasti skúmaného webového portálu pre tím správcov portálu.

Inšpiráciou pri skúmaní správania sa používateľov webu v závislosti od času na báze týždňov nám budú podobné výskumy z iných oblastí. Dabrowska-Zielinska et al. (Dabrowska-Zielinska et al., 2002) sa zamerali na vytvorenie modelu pre skúmanie podmienok na pestovanie plodín v rôznych oblastiach Poľska, pričom na to využili diaľkové skúmanie zeme. Na základe satelitných snímok vypočítali dva indexy pre každý týždeň pre obdobie 14 rokov. Následne sledovali namerané hodnoty v jednotlivých týždňoch počas roka a vyhodnotili najvhodnejšie obdobie pre výsadbu a pestovanie plodín počas roka. Všetky dosiahnuté výsledky podporili meteorologickými pozorovaniami. Raffi et al. (Raffi et al., 2006) sa zamerali na skúmanie vplyvu aplikácie anitretovírusovej liečby do dvanásteho týždňa priebehu choroby a predikcie vývoja liečby v ďalších týždňoch. Na modelovanie výsledkov kritických týždňov 24, 48 a 96 využili predikciu založenú na pozorovaných účinkoch liečby počas prvých dvanástich týždňoch. Vďaka dosiahnutým výsledkom dokážu pacientom už po dvanástich týždňoch liečby odporučiť pokračovanie alebo vysadenie danej liečebnej procedúry. Aj v problematike modelovania povodí využili Verdhen et al. (Verdhen et al., 2014) predikciu v článku, v ktorom sa zamerali na model topenia snehu v Himalájach počas jednotlivých týždňov jarného obdobia. Vo svojom výskume využili údaje z roku 2008, aby vytvorili model predikcie pre spätné roky 2003 a 1983 skúmané na týždennej báze počas jarného obdobia týchto rokov. Okrem dosiahnutých výsledkov v oblasti skúmania zistili aj, že ich vstupné dáta neboli dostatočne spoľahlivé, aby dosiahli efektívny koeficient simulácie pre rok 1983. Odstrániť to chcú v budúcom výskume doplnením ďalších parametrov, ktoré pomôžu zvýšiť kvalitu simulácie.

Na modelovanie správania sa používateľov je možné využiť multinominálny logitový model (Munk a Drlík, 2014), ktorý je špeciálnym prípadom zovšeobecneného lineárneho modelu. Popis modelu sa odvoláva prioritne na príspevky (Munk et al., 2011b, 2011a), ktoré boli zamerané na analýzu používania portálu, resp. virtuálneho vzdelávacieho prostredia. Pri modelovaní správania sa používateľov webu sa pracuje už s predspracovaným logovacím súborom, ktorý je očistený od zbytočných dát a prístupov robotov vyhľadávacích služieb. Následne je nutné vybrať správnu metódu, ktorou sa identifikujú sedenia používateľov, aby bolo možné správne rekonštruovať cestu všetkých používateľov. Pred určovaním modelu si vytvoríme nezávislú premennú čas (*time*), ktorá bude nadobúdať časové hodnoty, reprezentované napríklad hodinami prístupu, týždňami v roku, a pod.

Označíme si π_{ij} ako pravdepodobnosť prístupu na kategóriu j používateľom v hodine i, pričom $j = 1, 2, \dots, J$. Keďže $\sum_{j=1}^{J} \pi_{ij} = 1$, tak máme $J - 1$ parametrov. Nech Y_{ij} je počet prístupov za hodinu i, na kategóriu j, s pozorovanými hodnotami y_{ij}. Potom $n_i = \sum_j y_{ij}$ je počet prístupov v hodine i. Rozdelenie pravdepodobnosti Y_{ij}, v prípade, že n_i je dané, je multinomické:

$$P\left[Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \dots Y_{iJ} = y_{iJ}\right] = \frac{n_i!}{y_{i1}! y_{i2}! \dots y_{iJ}!} \pi_{i1}^{y_{i1}} \pi_{i2}^{y_{i2}} \dots \pi_{iJ}^{y_{iJ}}. \quad (15)$$

Pravdepodobnosti π_{ij} prístupu na kategóriu j s prihliadnutím na hodinu i dostaneme z logitov, ktoré sa modelujú. Logity sú logaritmy:

$$\ln \frac{\pi_{ij}}{\pi_{iJ}}, j = 1, 2, \dots, J - 1; i \epsilon \{0, 1, \dots, 23\}, \quad (16)$$

kde π_{iJ} je pravdepodobnosť poslednej (referenčnej) kategórie. Predpokladáme nasledujúci model:

$$\eta_{ij} = \ln \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j, \quad (17)$$

kde $\boldsymbol{x}_i^T$ je lineárny vektor, α_j je konštanta a $\boldsymbol{\beta}_j$ je vektor regresných koeficientov pre $j = 1, 2, \dots, J - 1$. Vzniká nám takto $J - 1$ rovníc, ktoré zodpovedajú kategóriám $1, 2, \dots, J - 1$. Pravdepodobnosti π_{ij} vypočítame z rovníc:

$$\pi_{iJ} = \frac{1}{1 + \sum_{k=1}^{J-1} e^{\eta_{ik}}}, \quad \pi_{ij} = e^{\eta_{ij}} \pi_{iJ}, \quad j = 1, 2, \dots, J - 1. \quad (18)$$

Odhad maximálnej vierohodnosti parametrov modelu (17) nasleduje po maximalizácií logaritmu multinominálnej funkcie vierohodnosti (15) s pravdepodobnosťami π_{ij} vyjadrenými pomocou parametrov α_j a $\boldsymbol{\beta}_j$.

2.2.2 METODIKA SPRACOVANIA

Predstavenú metodiku spracovaniu uvádzame ku kapitole 3.5, kde bola riešená úloha návrhu metodiky predikcie pravdepodobnosti prístupov na webové časti portálu v závislosti od týždňov, ktorá predstavuje hlavné naplnenie cieľu práce. Pri modelovaní správania sa používateľov webu v závislosti od času sme vychádzali z metodiky CRISP-DM (CRoss Industry Standard Process for Data Mining) (Chapman et al., 2000). Cieľom CRISP-DM je ponúknuť organizáciám porozumenie procesu objavovania znalostí a poskytnúť návod počas plánovania a vykonávania procesu objavovania znalostí. Metodika CRISP-DM pozostáva zo šiestich fáz: porozumenie problematika, porozumenie dátam, príprava dát, modelovanie, vyhodnotenie výsledkov a využitie výsledkov. Analýzu v záverečnej práci budeme vykonávať nad dátami webového servera komerčnej banky, pričom budeme využívať časové premenné na úrovni sekúnd – unixtime, hodín – čas prístupu, týždňov, kvartálov a rokov. Pripravovaný logovací súbor obsahuje záznamy aktivity na vybranom webovom portály počas obdobia viacerých rokov (2009-2015). V rokoch 2009-2010 rezonovala vo svete ekonomická kríza, ktorá mala aj vplyv na správanie sa stakeholderov komerčných bánk (Munk et al., 2013b; Pilkova et al., 2015; Munk et al., 2012). V nasledujúcich rokoch (2011-2015) ekonomická kríza začala odznievať a rovnako predpokladáme aj zmenu v správaní sa stakeholderov komerčných bánk. Zdrojom údajov pre modelovanie správania sa používateľov webu v závislosti od času bol logovací súbor z webového servera významnej domácej komerčnej banky pôsobiacej na Slovensku.

Porozumenie problematike

Cieľom záverečnej práce je modelovať správanie sa používateľov webu v závislosti od času. V našom prípade sa zameriame na rôzne časové obdobia, pričom pomocou modelu je možné predikovať návštevnosť jednotlivých webových častí v závislosti napríklad od hodiny konkrétneho dňa, alebo od týždňa konkrétneho roku a ďalších vysvetľujúcich premenných. Skúmanie správania na základe hodín dňa nám dokáže prezradiť, v ktorých častiach dňa je pravdepodobný prístup návštevníkov na jednotlivé webové kategórie. Môže to byť využité pre správcov a tvorcov webového portálu, aby vedeli naplánovať aktualizáciu alebo úpravu portálu. Skúmanie správania na základe týždňov nám môže prezradiť sezónnosť v navštevovaní jednotlivých webových kategórií a hlavne odhaliť vplyv finančnej krízy na správanie návštevníkov na webovom portály bankovej inštitúcie (hľadanie informácií).

Porozumenie dátam

Zdrojom dát sú dáta webových portálov ukladané v spoločnej štandardnej štruktúre v textovom formáte – logovací súbor alebo vo vlastnej štruktúre najčastejšie organizované v relačnej databáze. V daných zdrojoch sa sledujú nasledovné atribúty:

- IP adresa;
- dátum a čas prístupu;
- URL adresa.

V prípade skúmania portálu, ktorý vyžaduje prihlasovanie používateľov, je možné sledovať aj atribút ID používateľa. Z týchto atribútov sa ďalej vytvárajú premenné, ktoré sú použité priamo na modelovanie, resp. vo fáze prípravy dát. Ďalším podstatným zdrojom dát pre prípravu dát je mapa webu. Mapa webu slúži napríklad vo fáze rekonštrukcie aktivít používateľov webu, kedy sa dopĺňajú cesty medzi stránkami, ktoré v logovacom súbore nie sú zaznamenané, kvôli použitiu tlačidla späť vo webovom prehliadači návštevníka webového portálu.

Príprava dát

Po získaní logovacieho súboru zo skúmaného servera sa prechádza k fáze prípravy dát. Príprava dát patrí medzi časovo aj prostriedkovo najnáročnejšiu fázu v procese objavovania znalostí, hlavne z dôvodu nepodstatných údajov, ktoré sa nachádzajú v logovacích súboroch. Príprava dát pozostáva z nasledujúcich úloh:

1. Čistenie dát – v prípade, ak je zdrojom dát logovací súbor webového servera, tak je nutné očistiť dáta od nepotrebných údajov (požiadavky na obrázky, skripty, štýly a pod.) a rovnako aj od prístupov robotov vyhľadávacích služieb, ktoré pristupujú na webový portál;
2. Identifikácia sedení a dopĺňanie ciest – dopĺňanie ciest je závislé od presnosti identifikácie sedení. V našom prípade bude využitá pre identifikáciu sedení metóda Reference Length;
3. Určenie premenných – logovací súbor obsahuje premenné v tradičnom formáte ELF, preto je nutná transformácia a určenie potrebných premenných pre analýzu správania sa používateľov skúmaného webu. Je vytvorená závislá premenná *category*, ktorej úrovne predstavujú webové časti webového portálu. V prípade,

ak pre webové časti je návštevnosť nízka, je vhodné vytvoriť širšie kategórie na základe ich príslušnosti k obsahu (Munk et al., 2011b). Premenná *category* bude v prípade webového portálu bankovej inštitúcie obsahovať nasledujúce kategórie: *Pricing List, Reputation, Business Conditions, Pillar3 Related, Pillar3 Disclosure Requirements* a *We support*. Taktiež je nutné určiť nezávislé premenné- prediktory, ktoré predstavujú časové premenné vytvorené z časovej známky prístupu. V prípade týždňov sa jedná o premennú *week,* ktorá je vytvorená na základe normy ISO 8601, nadobúdať hodnoty 0-53. Hodnotu 0 dosiahne premenná v prípade, ak sa jedná o týždeň, ktorý začína v predchádzajúcom roku. Ďalšou umelou premennou je premenná *crisis,* ktorá identifikuje obdobie rokov počas finančnej krízy a po finančnej kríze. Následne bola vytvorená umelá premenná *internal* na identifikáciu prístupov zvnútra a zvonka siete organizácie, ktorá nám umožňuje skúmať správanie sa používateľov pristupujúcich zvnútra/zvonka siete organizácie (na základe sád IP adries bola vytvorená premenná *internal*).

Modelovanie

Pri modelovaní sa postupuje nasledovne za predpokladu, že vychádzame z údajov, v ktorých sú zaznamenané individuálne prístupy na webové časti portálu.

1. Určenie modelu – pravdepodobnostné rozdelenie počtov prístupov Y_{ij} v čase i na kategóriu j s pozorovaniami y_{ij}, ak je daný počet prístupov $n_i = \sum_j y_{ij}$ v čase i je multinomické $P\left[Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \dots, Y_{iJ} = y_{iJ}\right] = \frac{n_i!}{y_{i1}!y_{i2}!\dots y_{iJ}!}\pi_{i1}^{y_{i1}}\pi_{i2}^{y_{i2}} \dots \pi_{iJ}^{y_{iJ}}$. Pretože $\sum_{j=1}^{J}\pi_{ij} = 1$ je nutné odhadnúť $J-1$ neznámych pravdepodobností. Odhady sa vypočítajú pomocou metódy maximálnej vierohodnosti, pričom sa v logaritmickej funkcii vierohodnosti $\sum_i \sum_{j=1}^{J} y_{ij} \ln \pi_{ij}$ (19) zavedie logitová transformácia (s voľbou poslednej kategórie za referenčnú) $\eta_{ij} = \ln \frac{\pi_{ij}}{\pi_{iJ}}$ a zároveň sa predpokladá, že logity η_{ij} sú lineárnymi funkciami nezávisle premenných $\eta_{ij} = \alpha_j + \boldsymbol{x}_i^T \boldsymbol{\beta}_j$. Inverznou transformáciou získame $\pi_{iJ} = \frac{1}{1+\sum_{j=1}^{J-1} e^{\eta_{ij}}}, \pi_{ij} = e^{\eta_{ij}}\pi_{ij}, j = 1, 2, \dots, J-1$, resp. $\pi_{iJ} = \frac{1}{1+\sum_{j=1}^{J-1} e^{\alpha_j + x_i^T \boldsymbol{\beta}_j}}, \pi_{ij} = \frac{e^{\alpha_j + x_i^T \boldsymbol{\beta}_j}}{1+\sum_{j=1}^{J-1} e^{\alpha_j + x_i^T \boldsymbol{\beta}_j}}, j = 1, 2, \dots, J-1$ (20). Po dosadení takto vyjadrených pravdepodobností π_{ij} do (19) je logaritmická funkcia

vierohodnosti funkciou s neznámymi parametrami α_j a $\boldsymbol{\beta}_j$ $(j = 1, 2, \dots, J-1)$. Po určení modelu je potrebné identifikovať typ závislosti pre určenie stupňa polynómu a výber prediktorov vrátane umelých premenných.

2. Odhad parametrov modelu $\alpha_j, \boldsymbol{\beta}_j$ maximalizáciou logaritmu multinominálnej funkcie vierohodnosti. Na odhad parametrov pre individuálne údaje bude použitá *STATISTICA Generalized Linear/Nonlinear Models.* Významnosť parametrov bola testovaná pomocou Waldovho testu. Pomocou odhadnutých parametrov je možné vypočítať odhady logitov a z nich pravdepodobnosti výberu jednotlivých kategórií v danom čase.

3. Odhad logitov η_{ij} pre všetky hodnoty nezávislých premenných $\hat{\eta}_{ij} = a_j + \boldsymbol{x}_i^T \boldsymbol{b}_j, j = 1, 2, \dots, J-1$.

4. Odhad pravdepodobností prístupov π_{iJ} v čase i pre referenčnú webovú časť J $\hat{\pi}_{iJ} = \frac{1}{1+\sum_{j=1}^{J-1} e^{\hat{\eta}_{ij}}}$.

5. Odhad pravdepodobností prístupov π_{ij} v čase i pre webovú časť j $\hat{\pi}_{ij} = e^{\hat{\eta}_{ij}} \hat{\pi}_{iJ}, j = 1, 2, \dots, J-1$.

6. Vizualizácia pravdepodobností výberu webovej časti j v čase $i, j = 1, 2, \dots, J$

Vyhodnotenie výsledkov

Po vytvorení modelu je nutné vypočítané výsledky vyhodnotiť, postupovať budeme nasledovne:

1. Určenie empirických početností prístupov y_{ij}.

2. Odhad teoretických početností prístupov $\hat{y}_{ij} = \hat{\pi}_{ij} \sum_j y_{ij}$.

3. Vizualizácia rozdielov empirických a teoretických početností $d_{ij} = y_{ij} - \hat{y}_{ij}$.

4. Identifikácia extrémnych hodnôt d_{ij}, kde $d_{ij} > \bar{d}_j + 2s_j$ reprezentuje podhodnotenú predpoveď a $d_{ij} < \bar{d}_j - 2s_j$ reprezentuje nadhodnotenú predpoveď, kde s_j predstavuje smerodajnú odchýlku a $\bar{d}_j$ priemer rozdielov pre kategóriu j.

5. Výpočet empirických relatívnych početností prístupov $p_{ij} = \frac{y_{ij}}{\sum_j y_{ij}}$.

6. Porovnanie rozdelenia pravdepodobnosti empirických relatívnych početností prístupov a odhadnutých pravdepodobností výberu webovej časti j v čase i $r_{ij} = p_{ij} - \pi_{ij}$, H0: $F(-r) = 1 - F(r)$. Na testovanie nulovej hypotézy, rozdelenie rozdielov párov je symetrické okolo nuly, bude použitý Wilcoxonov párový test.

7. Výpočet empirických logitov $h_{ij} = \ln\left(\frac{p_{ij}}{p_{iJ}}\right), j = 1, 2, \dots, J-1.$

8. Vizualizácia empirických a teoretických logitov pre jednotlivé webové časti, okrem referenčnej.

Využitie výsledkov

Aplikáciou modelu na dáta logovacích súborov, v ktorých je zaznamenané používanie portálu budeme modelovať pravdepodobnosti prístupov na webové časti portálu. Na základe výsledkov odhadu pravdepodobnosti prístupov na webové časti bankového portálu môžeme identifikovať obdobie počas jednotlivých týždňov roka, v ktorých je nutné aktualizovať informácie o treťom pilieri.

3 VÝSLEDKY VÝSKUMU

V tejto kapitole sa zameriame na dosiahnuté výsledky výskumu týkajúce sa optimalizácie prípravy dát v procese WUM a modelovania správania sa používateľov webu. Zameriame sa najprv na experimenty vo fáze prípravy dát, ktoré boli vykonávané nad rôznymi druhmi zdrojov dát. Zdroje dát pochádzali buď z webového portálu s anonymným prístupom alebo webového portálu s povinnou autentifikáciou. V prvom prípade boli zdrojom dát logovacie súbory z univerzitného portálu, prípadne webového portálu z bankovej inštitúcie. V druhom prípade sa jednalo prevažne o portál virtuálneho vzdelávacieho prostredia. Predstavené výsledky jednotlivých výskumov slúžia na splnenie cieľa dizertačnej práce.

3.1 PRÍPRAVA DÁT PORTÁLU S ANONYMNÝM PRÍSTUPOM

Výskumy (Munk a Benko, 2016; Munk, Benko, Gangur, Turcani, 2015) sa zaoberali prioritne fázou prípravy dát, konkrétne však boli zamerané hlavne na fázu identifikácie sedení pomocou metódy Reference Length a vplyvom podielu navigačných stránok na výpočet hraničného času. V prvom rade bol v spomínaných výskumoch použitý logovací súbor univerzitného portálu (Obrázok 6), aj z toho dôvodu bolo nutné očistiť logovací súbor od nepotrebných údajov. Logovací súbor v sebe zahŕňal informácie o IP adrese prístupu, cookie, čase prístupu vo formáte unixového času (unixtime), dátum a čas prístupu, http metódu, URL adresu prístupu, návratový kód, URL adresu predchádzajúcej stránky (referrer) a informáciu o použitom prehliadači (user agent).

```
10.160.0.86;10.160.0.86.1379924192754896;1379931400;23.9.2013
10:16:40;GET;/://0;HTTP/1.1;404;https://www.ukf.sk/struktura-
univerzity;"Mozilla/4.0 (compatible; MSIE 8.0; Windows NT
5.1; Trident/4.0; .NET CLR 2.0.50727)"

192.168.0.32;192.168.0.32.1379924198329561;1379931400;23.9.20
13 10:16:40;GET;/dokumenty/images/udalosti/logo-iab-
slovakia.jpg;HTTP/1.1;200;https://www.ukf.sk/;Mozilla/5.0
(Windows NT 6.0) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/29.0.1547.57 Safari/537.36

10.160.0.86;10.160.0.86.1379924192754896;1379931404;23.9.2013
10:16:44;GET;/struktura-
univerzity/://0;HTTP/1.1;404;https://www.ukf.sk/struktura-
univerzity/pedagogicka-fakulta;"Mozilla/4.0 (compatible; MSIE
8.0; Windows NT 5.1; Trident/4.0; .NET CLR 2.0.50727)"
```

Obrázok 6 Ukážka záznamov logovacieho súboru univerzitného portálu

Na očistenie logovacieho súboru bol vytvorený vlastný algoritmus pomocou programovacieho jazyka Java v prostredí NetBeans. Algoritmus z logovacieho súboru odstraňoval na základe prípony záznamy o obrázkoch (*.jpg, *.jpeg, *.png, *.bmp, *.gif), java skriptoch (*.js), štýloch (*.css), súhrnoch (*.rss), kurzoroch (*.cur), videií (*.flv, *.swf), favikonách (*.ico) a xml súboroch. Zároveň algoritmus obsahoval aj filter obsahujúci HTTP návratové kódy, informujúci o chybách na strane klienta alebo servera. Keďže logovací súbor je nutné očistiť aj od prístupov robotov vyhľadávacích služieb, obsahuje algoritmus dôležitú súčasť, ktorá dokáže v logovacom súbore identifikovať prístupy k súboru „robots.txt.“ Na základe toho bolo možné zaznamenať IP adresy robotov a crawlerov a zaznamenať ich pomocou java triedy HashSet ako tokeny. Následne bolo možné pomocou regulárnych výrazov identifikovať v riadkoch (teda záznamoch) logovacieho súboru neželané tokeny a vytvoriť tak len nový logovací súbor, ktorý neobsahuje nepotrebné dáta.

Z dôvodu skúmania prípravy dát so zameraním na identifikáciu sedení bola v tomto experimente zahrnutá identifikácia používateľov určená na základe poľa IP adresa a User Agent. Identifikácia sedení bola vykonaná pomocou metódy Reference Length, ktorú sme bližšie predstavili v kapitole 2.1.3. Na základe odhadu podielu navigačných stránok v (Kapusta et al., 2012) bol vytvorený algoritmus slúžiaci na identifikáciu sedení. Očistený logovací súbor bol importovaný do databázy, kde bola vytvorená nová premenná, ktorá vychádzala z dátumu a času návštevy webového portálu konkrétneho záznamu. Odhad podielu navigačných stránok je možné vykonať na základe subjektívneho odhadu tvorcu alebo správcu skúmaného webového portálu, alebo je možné vychádzať z mapy webu skúmaného portálu. V Jave bola použitá knižnica Map, ktorá umožnila uchovanie mapy webu a následný odhad podielu navigačných stránok z uloženej mapy webu. Pre výpočet odhadu hraničného času sú nutné vstupné parametre p a λ, kde p je relatívna početnosť navigačných stránok a λ je priemerná intenzita udalostí (bližší popis sa nachádza v kapitole 2.1.3). Experiment bol zároveň zameraný nielen na skúmanie vplyvu výpočtu odhadu podielu navigačných stránok, ale aj na to, akú hrá rolu dopĺňanie ciest.

3.1.1 METODIKA

Experiment v článkoch (Munk et al., 2015; Munk a Benko, 2016) bol realizovaný na základe nasledujúceho postupu, ktorý vychádzal z (Munk et al., 2010a, 2010b):

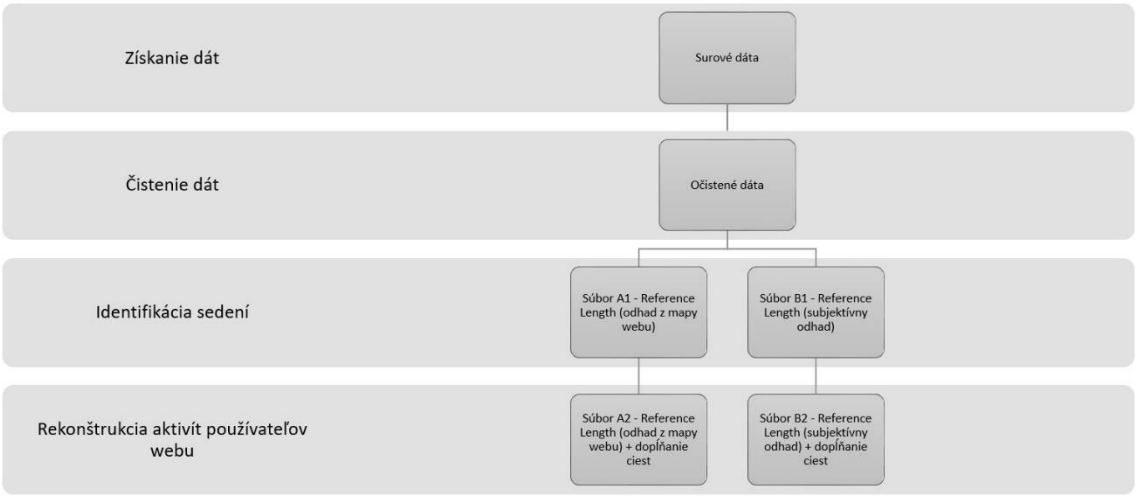
1. Získanie dát – určenie pozorovaných premenných logovacieho súboru z pohľadu získania potrebných údajov (IP adresa, dátum a čas prístupu, URL adresa, a pod.).
2. Vytvorenie dátovej matice – z logovacieho súboru (informácie o prístupoch) a mapy webu (informácie o obsahu webového portálu).
3. Príprava dát na rôznych úrovniach:
    a. identifikácia sedení pomocou Reference Length vypočítané na základe mapy webu,
    b. identifikácia sedení pomocou Reference Length vypočítané na základe mapy webu a doplnené cesty,
    c. identifikácia sedení pomocou Reference Length vypočítané na základe subjektívneho odhadu,
    d. identifikácia sedení pomocou Reference Length vypočítané na základe subjektívneho odhadu a doplnené cesty.
4. Analýza dát – hľadanie vzorcov správania sa používateľov webu v jednotlivých súboroch.
5. Porozumenie výstupným dátam – vytvorenie dátového súboru z výstupov analýz jednotlivých súborov a stanovenie predpokladov.
6. Porovnanie získaných znalostí zo skúmaných súborov predspracovaných na rôznej úrovni prípravy dát z pohľadu kvantity a kvality nájdených pravidiel.

3.1.2 VÝSLEDKY

Boli porovnávané štyri súbory, pripravené na rôznych úrovniach (Obrázok 7). Každý súbor bol očistený od nepotrebných dát rovnakým algoritmom. V prípade identifikácie sedení bol rozdiel práve vo výpočte odhadu podielu navigačných stránok. Porovnávali sme subjektívny odhad a odhad na základe mapy webu, pričom práve subjektívny odhad je najčastejšie využívaný, ale často menej presný. V subjektívnom odhade je navigačná stránka definovaná na základe odporúčania tvorcu alebo správcu webového portálu. Na druhej strane alternatívou by mohol byť výpočet podielu navigačných stránok z mapy webu. V tomto prípade je každá stránka, ktorá obsahuje podstránku, definovaná ako navigačná stránka. V experimentoch (Munk et al., 2015; Munk a Benko, 2016) boli stanovené nasledujúce predpoklady:

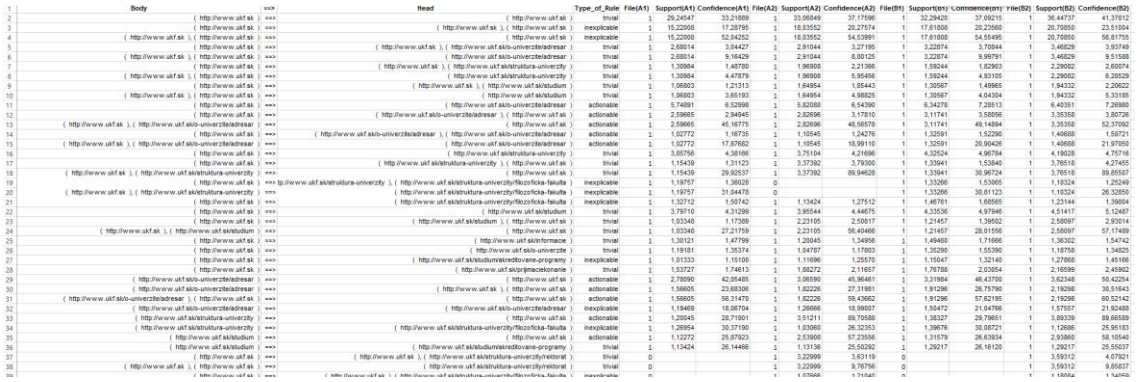
1. Predpokladáme, že identifikácia sedení pomocou Reference Length vypočítané z mapy webu bude mať významný vplyv na kvantitu extrahovaných pravidiel z hľadiska menšieho počtu triviálnych a nevysvetliteľných pravidiel.

2. Predpokladáme, že identifikácia sedení pomocou Reference Length vypočítané z mapy webu bude mať významný vplyv na kvalitu extrahovaných pravidiel z hľadiska ich základných charakteristík kvality.



Obrázok 7 Aplikácia prípravy dát na logovací súbor webového portálu s anonymným prístupom

Súbor A1 obsahoval identifikované sedenia pomocou metódy Reference Length a podiel navigačných stránok bol vypočítaný na základe mapy webu (12,3 %). Súbor A2 obsahoval rovnako ako súbor A1 identifikované sedenia pomocou Reference Length vypočítané na základe mapy webu (12,3 %) a okrem toho boli v tomto súbore doplnené cesty. Súbor B1 obsahoval identifikované sedenia pomocou metódy Reference Length a podiel navigačných stránok bol odhadnutý subjektívne (30 %). V súbore B2 boli identifikované sedenia rovnako ako v súbore B1, pričom boli ešte k tomu doplnené cesty. V ďalšom kroku boli pre každý súbor pomocou sekvenčnej analýzy extrahované sekvenčné pravidlá. Následne boli tieto pravidlá zlúčené do jednej dátovej matice (Obrázok 8), kde sa každé pravidlo nachádzalo práve raz.



Obrázok 8 Ukážka dátovej matice obsahujúcej extrahované sekvenčné pravidlá

Výpočet podielu navigačných stránok má vplyv na počet identifikovaných sedení, kde v súboroch A1 a A2 bolo identifikovaných 51 098 sekvencií a v súboroch B1 a B2 sa objavilo o 20 % menej sekvencií (Tabuľka 1). Pomocou sekvenčnej analýzy (*STATISTICA Sequence, Association, and Link Analysis*) boli z frekventovaných sekvencií extrahované sekvenčné pravidlá s minimálnou podporou 0,01 pre každý súbor (Obrázok 7). Zjednotením súborov do jednej dátovej matice sme získali 78 unikátnych pravidiel. Z toho bolo 35 (45 %) pravidiel identifikovaných v súbore A1, 39 (50 %) pravidiel v súbore B1, 62 (79 %) pravidiel v súbore A2 a 77 (99 %) v súbore B2.

Tabuľka 1 Počet prístupov, sekvencií a pravidiel

	Súbor A1	**Súbor A2**	**Súbor B1**	**Súbor B2**
Počet prístupov	154 681	178 043	154 681	178 806
Počet identifikovaných sekvencií (sedení)	51 098	51 098	40 756	40 756
Počet frekventovaných sekvencií	37	57	43	70
Absolútny počet extrahovaných pravidiel	35	62	39	77
Relatívny počet extrahovaných pravidiel	0,45	0,79	0,50	0,99
Počet užitočných pravidiel	10	10	10	10
Počet triviálnych pravidiel	17	35	19	40
Počet nevysvetliteľných pravidiel	8	17	10	27

Výsledky ukázali (Tabuľka 2Tabuľka 2), že súbory bez dopĺňania ciest (A1, B1) obsahujú skoro podobné pravidlá, až na výnimku štyroch pravidiel (5 %) v prípade súboru so subjektívnym odhadom podielu navigačných stránok (B1). V prípade súborov s dopĺňaním ciest (Tabuľka 3) sa preukázal štatisticky významný rozdiel v 16 nových pravidlách (skoro 21 %) v prípade súboru so subjektívnym odhadom (B2).

Tabuľka 2 Kontingenčné tabuľky pre Súbor A1 x Súbor B1

A1\B1	**0**	**1**	∑
0	39 50,00 %	4 5,13 %	43 55,13 %
1	0 0,00 %	35 44,87 %	35 44,87 %
∑	39 50,00 %	39 50,00 %	78 100,00 %
McNemar (B/C)	*Chi-square* = 2,25000; *df* = 1; *p* = 0,134		

Tabuľka 3 Kontingenčné tabuľky pre Súbor A2 x Súbor B2

A2\B2	**0**	**1**	∑
0	0 0,00 %	16 20,51 %	16 20,51 %
1	1 1,28 %	61 78,21 %	62 79,49 %
∑	1 1,28 %	77 98,72 %	78 100,00 %
McNemar (B/C)	*Chi-square* = 11,52941; *df* = 1; *p* = 0,00069		

Na základe kontingenčných koeficientov (*Kontingenčný koeficient C*, *Cramerovo V*), ktoré reprezentujú stupeň závislosti medzi dvoma nominálnymi premennými, môžeme pozorovať strednú závislosť medzi podielom užitočných, triviálnych a nevysvetliteľných pravidiel a ich výskytom v objavených pravidlách v jednotlivých súboroch bez dopĺňania ciest (A1: 0,40; B1: 0,37). Získané výsledky pre súbory s dopĺňaním ciest boli zaujímavejšie. Stredná závislosť (*Cramerovo V* = 0,32) bola identifikovaná medzi podielom užitočných, triviálnych a nevysvetliteľných pravidiel a ich výskytom v súbore A2 (Tabuľka 4). Hodnota kontingenčného koeficientu pre výskyt pravidiel v súbore B2 bola približne 0,11 (kde 1 predstavuje perfektnú závislosť a 0 žiadnu závislosť), čo znamená, že pre súbor B2 bola len malá závislosť (Tabuľka 5). Taktiež sa ukázalo, že kontingenčný koeficient nie je štatisticky významný (Tabuľka 5). Súbor B2 obsahoval najviac nevysvetliteľných pravidiel, ale podiel užitočných pravidiel bol rovnaký vo všetkých súboroch.

Tabuľka 4 Kontingenčné tabuľky: Výskyt pravidiel x Typ pravidiel: Súbor A2

A2\Type	**Užitočné**	**Triviálne**	**Nevysvetliteľné**
0	0 0,00 %	6 14,63 %	10 37,04 %
1	10 100,00 %	35 85,37 %	17 62,96 %
∑	10 100,00 %	41 100,00 %	27 100,00 %
Pearson **Kon. koeficient C** **Cramerovo V**	*Chi-square* = 7,97115; *df* = 2; *p* = 0,019 0,30450 0,31968		

Tabuľka 5 Kontingenčné tabuľky: Výskyt pravidiel x Typ pravidiel: Súbor B2

B2\Type	Užitočné	Triviálne	Nevysvetliteľné
0	0 0,00 %	1 2,44 %	0 0,00 %
1	10 100,00 %	40 97,56 %	27 100,00 %
∑	10 100,00 %	41 100,00 %	27 100,00 %
Pearson **Kon. koeficient C** **Cramerovo V**	*Chi-square* = 0,91416; *df* = 2; *p* = 0,633 0,10763 0,10826		

Výsledky sekvenčnej analýzy neboli skúmané len na základe kvantity extrahovaných pravidiel, ale aj kvality. Kvalita extrahovaných pravidiel bola posudzovaná na základe dvoch kritérií: podpora a spoľahlivosť (Berry a Linoff, 2004). Štatisticky významné rozdiely boli pozorované medzi súbormi A1, A2, B2 a medzi súbormi B1, B2 ohľadne priemernej podpory extrahovaných pravidiel (Tabuľka 6). V prípade priemernej spoľahlivosti nájdených pravidiel boli identifikované štatisticky významné rozdiely medzi súbormi A1, A2 a medzi súbormi A1, B2 (Tabuľka 7). Vplyv na kvalitu a kvantitu extrahovaných pravidiel sa dokázal až po dopĺňaní ciest. Naopak dopĺňanie ciest je závislé na presnosti identifikácie sedení.

Tabuľka 6 Homogénne skupiny pre podporu extrahovaných pravidiel

Súbor	Podpora	1	2	3
A1	3,425	****		
B1	3,941	****	****	
A2	4,163		****	
B2	4,747			****
Kendallov koeficient konkordancie			0,63692	

Tabuľka 7 Homogénne skupiny pre spoľahlivosť extrahovaných pravidiel

Súbor	Spoľahlivosť	1	2
A1	15,942		****
B1	17,123	****	****
A2	22,320	****	
B2	23,442	****	
Kendallov koeficient konkordancie			0,49568

Na základe dosiahnutých výsledkov bolo možné usúdiť, že výpočet podielu navigačných stránok v metóde Reference Length zohráva významnú rolu až po dopĺňaní ciest. Malo

to vplyv hlavne na zvýšenie počtu triviálnych a nevysvetliteľných pravidiel, pričom extrahovaný počet užitočných pravidiel bol vo všetkých súboroch rovnaký.

V prípade výskumu v článkoch (Munk et al., 2015; Munk a Benko, 2016) boli oba predpoklady, týkajúce sa kvantity a kvality identifikácie sedení pomocou výpočtu z mapy webu, preukázané len čiastočne. Úplne preukázané boli až po dopĺňaní ciest, pričom práve dopĺňanie ciest je závislé od presnosti identifikácie sedení. Metóda Reference Length je dobrým spôsobom identifikácie sedení. Nevýhodou uvedenej metódy je podmienka exponenciálneho rozdelenia premennej *RLength*, ktoré sa musí overiť skôr ako sa začnú identifikovať sedenia. Rôzne spôsoby odhadu podielu navigačných stránok majú vplyv až po dopĺňaní ciest. Výpočet z mapy webu sa ukázal presnejší, než subjektívny odhad, ale nevýhodou mapy webu je prakticky neustála zmena webového portálu a jej prípadná neaktuálnosť. Dostatočné množstvo extrahovaných kvalitných pravidiel umožňuje sofistikovanejšiu analýzu správania sa používateľov na webovom portály.

3.2 PRÍPRAVA DÁT PORTÁLU S POVINNOU AUTENTIFIKÁCIOU

V predchádzajúcej kapitole sme sa zamerali na prípravu dát logovacieho súboru z webového portálu s anonymným prístupom. V praxi sa však často stretávame aj s webovými portálmi, ktoré vyžadujú od návštevníkov registráciu, resp. prihlásenie sa na webovom portály pre ďalšiu činnosť. Články (Benko, 2015; Benko, Reichel a Munk, 2015; Reichel, Kuna, Benko a Munk, 2015; Munk, Drlik, Benko a Reichel, 2017a) boli zamerané práve na skúmanie prípravy dát logovacieho súboru, ktorý pochádzal z webového portálu s povinnou autentifikáciou. V tomto prípade sa jednalo o logovací súbor, ktorý pochádzal z virtuálneho vzdelávacieho prostredia Moodle, konkrétne z kurzu predmetu Počítačová analýza dát (PAD). V kurze sa nachádzalo 69 registrovaných študentov, ktorých správanie bolo skúmané.

Cieľ tohto výskumu blízko súvisel s dvoma predchádzajúcimi výskumami (Munk a Drlik, 2011a, 2011b), ktoré boli zamerané na špecifikáciu potrebných krokov predspracovania dát s úmyslom získania spoľahlivých údajov z logovacích súborov pochádzajúcich z VLE. Prvý experiment (Munk a Drlik, 2011a) bol zameraný na evalváciu vplyvu rôznych premenných (ID používateľa, IP adresa a čas) vhodných pre identifikáciu sedení na objavovanie znalostí reprezentované vzormi správania sa študentov v VLE. Ukázalo sa, že dopĺňanie ciest v kombinácií s týmito troma premennými nemalo významnejší

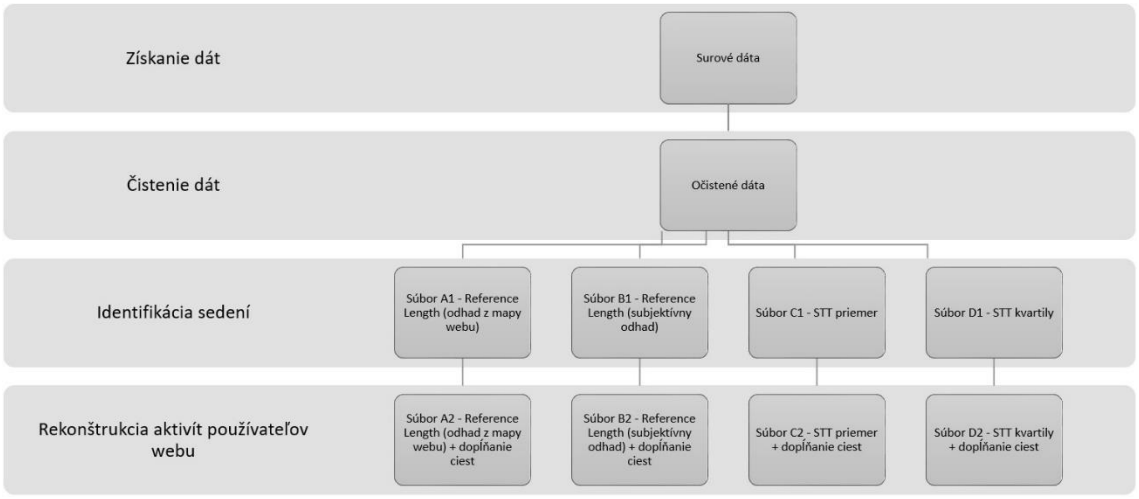
vplyv na kvantitu a ani na kvalitu získaných znalostí. Dopĺňanie ciest malo vplyv len na zvýšenie počtu užitočných pravidiel, avšak nárast nebol štatisticky významný. Naopak identifikácia sedení na základe premennej *čas,* často označovanej aj ako *Session Timeout Threshold* (STT), mala významný vplyv na kvantitu ako aj kvalitu získaných znalostí. Podľa Spiliopoulou et al. (Spiliopoulou et al., 2003) patrí táto technika medzi časovo orientované heuristiky pretože používa hornú hranicu času stráveného na celom portály počas návštevy. Hodnota STT sa používa na určenie kedy sedenie končí a začína nové sedenie. STT je preddefinované obdobie neaktivity, ktoré umožňuje webovým aplikáciám určiť, kedy nastalo nové sedenie (Huynh a Miller, 2009). Správna hodnota STT bola často diskutovaná viacerými autormi v oblasti web mining-u (Catledge a Pitkow, 1995) a taktiež aj edukačného data mining-u (Romero et al., 2014). Druhý experiment (Munk a Drlik, 2011b) bol preto zameraný na evalváciu vplyvu rôznych metód identifikácie sedení s použitím rôznych STT premenných na kvantitu a kvalitu extrahovaných znalostí. V spomínanom experimente bolo dokázané, že hodnota STT má významný vplyv na kvantitu extrahovaných pravidiel. Štatisticky významné rozdiely v priemere premenných *incidence, support* ako aj *confidence* nájdených pravidiel boli dokázané medzi súbormi s rôznym STT odhliadnuc od faktu, či boli súbory upravené dopĺňaním ciest alebo nie. Počet triviálnych a nevysvetliteľných pravidiel bol závislý od hodnoty STT. Identifikácia sedení na základe menšej hodnoty STT malo vplyv na zníženie počtu triviálnych a nevysvetliteľných pravidiel ako aj kvalitu objavených pravidiel na báze základných charakteristík kvality (Munk et al., 2013a). Na druhej strane bolo dokázané, že dopĺňanie ciest nemá vplyv ako na kvantitu tak ani kvalitu extrahovaných pravidiel. Na základe výsledkov týchto dvoch predchádzajúcich experimentov (Munk a Drlik, 2011a, 2011b) boli stanovené ciele pre experiment (Munk et al., 2017a), v ktorom sa porovná viacero rôznych spôsobov metód identifikácie sedení.

Metodika vychádzala z predchádzajúceho výskumu (Munk et al., 2015), avšak proces prípravy dát musel byť prispôsobený logovaciemu súboru získaného z portálu Moodle. Kvôli zmenám v logovacích súboroch systému Moodle (Mazza et al., 2012; Rice, 2011) bolo nutné upraviť logovací súbor do takej miery, aby bolo možné použiť štandardné postupy predspracovania dát. V prvom rade bolo nutné zo súboru extrahovať potrebné premenné, pričom problém bol hlavne s premennou URL, ktorú bolo nutné skrátiť len na identifikačné číslo stránky. Rozdiely v príprave dát oproti logovaciemu súboru z webového portálu s anonymným prístupom boli hlavne vo fáze čistenia a identifikácie

sedení. Logovací súbor VLE neobsahoval žiadne nepotrebné údaje, a preto si fáza čistenia vyžiadala len drobné úpravy vo forme odstránenia záznamov o prístupe správcu portálu a učiteľov. Tieto záznamy boli odstránené, pretože experiment bol zameraný na správanie sa študentov v kurze. Zároveň logovací súbor obsahoval informáciu o používateľovi, preto nebolo nutné identifikovať používateľov.

Dátové matice boli vytvorené z logovacieho súboru, ktorý obsahoval informácie o prístupoch študentov zoradené podľa premenných *ID používateľa, IP adresa* a *čas*. Zároveň bola vytvorená mapa webu elektronického kurzu. Mapa webu bola použitá v spojení s rôznymi technikami rekonštrukcie sedení neskôr v experimente. Mapa webu obsahuje informácie o štruktúre obsahu a navigácie elektronického kurzu. Zohráva veľkú úlohu na spätné dokončenie záznamov cesty, ktorú používateľ vykonal pomocou tlačidla „Späť“ vo svojom prehliadači. Použitie tohto tlačidla nie je automaticky zaznamenávané v logovacích záznamoch VLE. Informácie o existencií prepojení na jednotlivé stránky elektronického kurzu môžu byť extrahované z mapy webu. Mapa webu bola získaná pomocou aplikácie Web Crawling-u implementovanej v použitom *STATISTICA Data Miner*. Mapa webu musela byť upravená, aby korešpondovala s ID stránkami elektronického kurzu. Na základe zoradených záznamov podľa IP adries bolo možné hľadať prepojenia medzi jednotlivými stránkami (Munk a Drlik, 2011b). Premenná *STT* bola vytvorená v logovacom súbore na základe časového okna 100 minút ako ďalší krok vo vytváraní dátových matíc. Hodnota 100 minút bola vybraná s ohľadom na dĺžku priemernej vyučovacej hodiny. Vyučovacia hodina má normálne 90 minút, ale niekedy študenti dokončujú svoje úlohy aj po hodine, počas prestávky, preto bolo pridaných ďalších desať minút k hodnote *STT*. Bolo dokázané, že vybraných ďalších desať minút, bolo po zaokrúhlení v súlade s hodnotou štandardnej odchýlky.

Následne boli identifikované sedenia pomocou rôznych metód identifikácie sedení a doplnené cesty.



Obrázok 9 Aplikácia prípravy dát na logovací súbor VLE

Príprava dát viedla k vytvoreniu štyroch predspracovaných logovacích súborov (Obrázok 9):

- Súbor A1 obsahoval sedenia identifikované metódou Reference Length, pričom odhad podielu navigačných stránok bol vypočítaný z mapy webu (15,32 %).
- Súbor B1 obsahoval sedenia identifikované metódou Reference Length, kde bol využitý subjektívny odhad podielu navigačných stránok (25 %).
- Ďalšou použitou metódou na identifikáciu sedení bola metóda priemernej dĺžky časového okna. Nové sedenie je definované vtedy, keď čas medzi dvoma udalosťami bol väčší ako priemerná dĺžka sedenia. Takto vznikol súbor C1.
- Posledný súbor D1 bol vytvorený aplikáciou metódy, v ktorej sú sedenia identifikované pomocou kvartilov. Bol použitý vzorec $Q_{STT} = Q_{III} + 1{,}5Q$, kde Q_{III} je horný kvartil a Q je kvartilové rozpätie. Nové sedenie bolo definované keď čas medzi dvoma udalosťami v elektronickom kurze bol väčší ako Q_{STT}.

Následne boli po dopĺňaní ciest vytvorené ďalšie štyri súbory (Súbor A2, Súbor B2, Súbor C2, Súbor D2), čiže pre každý súbor z predchádzajúceho použitia metód identifikácie sedení. Sekvenčné pravidlá boli rozdelené do troch skupín: na užitočné, triviálne a nevysvetliteľné pravidlá (Berry a Linoff, 2004). V tomto prípade pozostávali triviálne pravidlá hlavne z pravidiel typu prechod z hlavnej stránky na stránku knihy, kvízu alebo zadania v rámci elektronického kurzu a naopak (napr. *Course => Autotest3*). Nevysvetliteľné pravidlá boli identifikované ako prechod z jednej stránky na tú istú

(napr. *Course => Course*). Užitočné pravidlá predstavovali prechod zo stránky koncového testu na stránku knihy a následne naspäť na stránku koncového testu, pretože takéto správanie nebolo počas koncového testu možné (napr. *Topic3/Autotest2, Topic2/Task2, Activity of lesson => Topic3/Autotest2*). Rozhodnutie, ku ktorému typu sekvenčného pravidla dané pravidlo patrí, bolo vykonané na základe subjektívneho rozhodnutia experta domény. Na základe subjektívneho charakteru rozhodnutia experta, muselo byť jeho rozhodnutie objektívne ohodnotené. Objektivita označuje mieru, do ktorej sú výsledky nezávislé od výskumníka, ako aj od meranej jednotky v zmysle skreslenia merania. V dôsledku toho bolo 89 objavených sekvenčných pravidiel hodnotených ôsmimi ľudskými expertami. Experti boli rozdelený do troch skupín na základe ich úloh v VLE a hlavne v konkrétnom elektronickom kurze: štyria učitelia (T), dvaja vývojári elektronického kurzu (C) a dvaja manažéri VLE (M). Všetci účastníci hodnotili užitočnosť nájdených sekvenčných pravidiel pomocou trojhodnotovej škály. Celkové hodnotenie pre dané pravidlo bolo vypočítané ako vážený priemer jednotlivých hodnotení. Tento prístup zohľadnil nerovnomerné rozdelenie skupín expertov. Použili sme neparametrickú metódu analýzy zohľadňujúc neznáme rozloženie údajov a ordinálnu charakteristiku použitých premenných, napokon bolo použité Kendallovo Tau. Bola testovaná štatistická významnosť vypočítaných koeficientov (Tabuľka 8). Tabuľka (Tabuľka 8) zobrazuje, že medzi hodnoteniami sekvenčných pravidiel jednotlivých expertov je veľká, takmer lineárna závislosť (0,7 – 1).

Tabuľka 8 Kendallove koeficienty Tau medzi hodnoteniami sekvenčných pravidiel jednotlivých expertov

	T1	**T2**	**T3**	**T4**	**C1**	**C2**	**M1**	**M2**
T1	1,000	0,772	0,843	1,000	1,000	0,912	1,000	0,978
T2	0,772	1,000	0,723	0,772	0,772	0,739	0,772	0,757
T3	0,843	0,723	1,000	0,843	0,843	0,734	0,843	0,824
T4	1,000	0,772	0,843	1,000	1,000	0,912	1,000	0,978
C1	1,000	0,772	0,843	1,000	1,000	0,912	1,000	0,978
C2	0,912	0,739	0,734	0,912	0,912	1,000	0,912	0,892
M1	1,000	0,772	0,843	1,000	1,000	0,912	1,000	0,978
M2	0,978	0,757	0,824	0,978	0,978	0,892	0,978	1,000

Koeficienty Kendallovho Tau boli štatisticky významné ($p < 0,05$) medzi jednotlivými skupinami expertov (Tabuľka 9). Inými slovami, môžeme tvrdiť, že vysoké koeficienty

korelácie zabezpečili objektivitu hodnotenia sekvenčných pravidiel. Najväčšia miera zhody bola medzi vývojármi elektronického kurzu (C) a manažérmi VLE (M). Na druhej strane najmenšia miera zhody bola medzi učiteľmi (T) a vývojármi elektronického kurzu. Nulová hypotéza predpokladala, že v hodnotení sekvenčných pravidiel medzi jednotlivými skupinami expertov nebudú štatisticky významné rozdiely. Hypotéza nebola zamietnutá na základe výsledkov Friedmanovho testu (ANOVA Chi kvad. (N = 89, df = 2) = 2,294118; p = 0,31757).

Tabuľka 9 Kendallove koeficienty Tau medzi hodnoteniami sekvenčných pravidiel jednotlivých skupín expertov

	N	**Kendallovo Tau**	**Z**	**p-hodn.**
C & M	89	0,919	12,7536	0,0000
T & M	89	0,795	11,0269	0,0000
T & C	89	0,741	10,2785	0,0000

Hlavný cieľ fázy analýzy dát tohto experimentu spočíval v nájdení vzorov správania sa študentov v jednotlivých súboroch. Sekvenčnou analýzou vykonanou nad všetkými skúmanými súbormi boli extrahované sekvenčné pravidlá. Výsledkom bola sada extrahovaných sekvenčných pravidiel z frekventovaných sekvencií s minimálnou podporou 1 % pre každý súbor.

Pre experiment boli sformulované nasledovné predpoklady (Munk et al., 2017a):

- Predpokladá sa, že identifikácia sedení pomocou metódy Reference Length s odhadom z mapy webu, bude mať významný vplyv na kvantitu extrahovaných pravidiel.
- Predpokladá sa, že identifikácia sedení pomocou metódy Reference Length s odhadom z mapy webu, bude mať významný vplyv na zvýšenie podielu užitočných pravidiel.
- Predpokladá sa, že identifikácia sedení pomocou metódy Reference Length s odhadom z mapy webu, bude mať významný vplyv na kvalitu extrahovaných pravidiel.
- Predpokladá sa, že dopĺňanie ciest bude mať významný vplyv na kvantitu extrahovaných pravidiel.

- Predpokladá sa, že dopĺňanie ciest bude mať významný vplyv na zvýšenie podielu užitočných pravidiel.
- Predpokladá sa, že dopĺňanie ciest bude mať významný vplyv na kvalitu extrahovaných pravidiel.

3.2.1 VÝSLEDKY

Experiment bol v prvom rade zameraný na evalváciu extrahovaných pravidiel zo skúmaných súborov z hľadiska kvantity a následne kvality. Pomocou *STATISTICA Sequence, Association and Link Analysis* modulu boli extrahované sekvenčné pravidlá. Medzi výsledkami sekvenčnej analýzy na základe podielu nájdených pravidiel v prípade súborov bez dopĺňania ciest (A1, B1, C1, D1) a medzi súbormi s dopĺňaním ciest (A2, B2, C2, D2) je vysoká zhoda. Najviac pravidiel bolo extrahovaných zo súborov s dopĺňaním ciest, konkrétne 76 pravidiel bolo extrahovaných zo súboru C2, čo predstavuje viac ako 85 %. Zo súboru A2 bolo extrahovaných 76 pravidiel, reprezentujúcich skoro 81 %. Zo súboru B2 bolo extrahovaných 67 pravidiel (75 %) a zo súboru D2 58 pravidiel (65 %). Vo všeobecnosti bolo viac pravidiel objavených v skúmaných súboroch, v ktorých bolo realizované dopĺňanie ciest. Keď zoberieme do úvahy fakt, že súbory boli vytvorené z toho istého logovacieho súboru, tak je prirodzené, že objavené sekvenčné pravidlá sa čiastočne prekrývajú medzi jednotlivými súbormi.

Na základe výsledkov Q testu je možné povedať, že nulová hypotéza týkajúca sa vplyvu rôznych spôsobov prípravy dát na výskyt pravidiel sa zamieta na hladine významnosti 0,1 % (Tabuľka 10). Kendallov koeficient zhody reprezentuje úroveň zhody v hodnote nájdených pravidiel medzi skúmanými súbormi. Hodnota koeficientu je približne 0,26 pričom 1 znamená perfektnú zhodu a 0 reprezentuje žiadnu zhodu. Nízke hodnoty koeficientu potvrdzujú výsledky Q testu.

Tabuľka 10 Počet prístupov, sekvencií a pravidiel v jednotlivých súboroch

	A1	B1	C1	D1	A2	B2	C2	D2
Počet prístupov	51 841	51 841	51 841	51 841	80 416	79 201	81 163	76 924
Počet identifikovaných sekvencií (sedení)	9342	11052	8256	15019	9342	11052	8256	15019
Počet frekventovaných sekvencií	61	56	66	46	93	86	98	76
Počet nájdených sekvenčných pravidiel	38	34	42	22	72	67	76	58
Percento nájdených sekvenčných pravidiel (Percent 1)	42,7	38,2	47,2	24,7	80,9	75,3	85,4	65,2
Percent 0	57,3	61,8	52,8	75,3	19,1	24,7	14,6	34,8
Cochranov Q test	Q = 158,6498; df = 7; p < 0,000							

Štyri homogénne skupiny (Tabuľka 11) boli identifikované z viacnásobného porovnávania (Tukeyov test), týkali sa priemerného výskytu nájdených pravidiel. Štatisticky významné rozdiely boli preukázané na 5 % hladine významnosti v priemernom výskyte nájdených pravidiel medzi súbormi D1 a C1, D2 a C2 ako aj medzi súbormi bez (X1) a s (X2, kde X = {A, B, C, D}) dopĺňaním ciest. Dopĺňanie ciest má významný vplyv na kvantitu extrahovaných pravidiel. Naopak identifikácia sedení s odhadom z mapy webu nemá vplyv na kvantitu extrahovaných pravidiel v prípade súborov bez a taktiež aj s dopĺňaním ciest.

Tabuľka 11 Homogénne skupiny pre výskyt odvodených pravidiel v skúmaných súboroch

Súbor	Priemerný výskyt	1	2	3	4
D1	0,247				****
B1	0,382	****			****
A1	0,427	****			****
C1	0,472	****			
D2	0,652		****		
B2	0,753		****	****	
A2	0,809		****	****	
C2	0,854			****	
Kendallov koeficient zhody			0,25465		

Výsledky sekvenčnej analýzy môžu byť analyzované aj podrobnejšie, ak zoberieme do úvahy podiel každého typu nájdeného pravidla. Vyžaduje sa, aby extrahované pravidlá boli nielen pochopiteľné, ale aj užitočné. Podobne ako v predchádzajúcom experimente (Munk et al., 2015), sme jednotlivé pravidlá rozdelili na užitočné, triviálne

a nevysvetliteľné (Berry a Linoff, 2004). Jedinou požiadavkou (predpoklad platnosti) použitia Chí-kvadrát testu je dostatočne veľké očakávané početnosti (Hays, 1988). Táto podmienka je porušená, ak sú očakávané početnosti menšie ako 5. V uskutočnených testoch bol predpoklad platnosti Chí-kvadrát testu porušený. Z toho dôvodu boli brané do úvahy nielen výsledky Pearsonovho Chí-kvadrát testu, ale aj hodnoty vypočítaných kontingenčných koeficientov.

Kontingenčné koeficienty (*Kontingenčný koeficient C*, *Cramerovo V*) reprezentujú stupeň závislosti medzi dvoma nominálnymi premennými. Hodnota kontingenčného koeficientu *Cramérovo V* bola približne 0,8 (Tabuľka 12). Medzi podielom užitočných, triviálnych a nevysvetliteľných pravidiel a ich výskytom v sade objavených pravidiel extrahovaných zo súboru A1 bola veľká závislosť. Kontingenčný koeficient bol štatisticky významný. Nulová hypotéza sa zamietala na hladine významnosti 0,1 %, t.j. podiel užitočných, triviálnych a nevysvetliteľných pravidiel závisí na identifikácií sedení pomocou metódy Reference Length s odhadom z mapy webu. V tomto súbore bolo objavených najmenej nevysvetliteľných pravidiel (86 %), pokým 20 užitočných pravidiel extrahovaných zo súboru A1 reprezentuje 95 % všetkých nájdených užitočných pravidiel. Najviac užitočných pravidiel bolo objavených v súbore s identifikovaním sedení pomocou metódy Reference Length s odhadom podielu navigačných stránok z mapy webu (A1).

Tabuľka 12 Kontingenčné tabuľky: Výskyt pravidiel x Typ pravidiel: Súbor A1

A1\Typ	**Užitočné**	**Triviálne**	**Nevysvetliteľné**
0	1 4,76 %	48 88,89 %	2 14,29 %
1	20 95,24 %	6 11,11 %	12 85,71 %
∑	21 100 %	54 100 %	14 100 %
Pearson **Kon. koef. C** **Cramerovo V**	*Chi-square* = 56,30237; *df* = 2; *p* = 0,00000 0,62248 0,79537		

Hodnota kontingenčného koeficientu *Cramerovo V* (Tabuľka 13, 14) bola približne 0,7, kde 1 znamená perfektnú závislosť a 0 znamená žiadnu závislosť. Medzi podielom užitočných, triviálnych a nevysvetliteľných pravidiel a ich výskytom v objavených pravidlách extrahovaných v súbore B1 a C1 bola veľká závislosť; kontingenčný

koeficient bol štatisticky významný. Nulová hypotéza sa zamieta s 99,9 % spoľahlivosťou, t.j. podiel užitočných, triviálnych a nevysvetliteľných pravidiel závisí od identifikácie sedení pomocou metódy Reference Length založenej na odhade z mapy webu (Súbor B1). V súbore C1 boli objavené prevažne triviálne (19 %) a nevysvetliteľné (100 %) pravidlá.

Tabuľka 13 Kontingenčné tabuľky: Výskyt pravidiel x Typ pravidiel: Súbor B1

B1\Typ	**Užitočné**	**Triviálne**	**Nevysvetliteľné**
0	6 28,57 %	48 88,89 %	1 7,14 %
1	15 71,43 %	6 11,11 %	13 92,86 %
∑	21 100 %	54 100 %	14 100 %
Pearson **Kon. koef. C** **Cramerovo V**	*Chi-square* = 44,32209; *df* = 2; *p* = 0,00000 0,57658 0,70569		

Tabuľka 14 Kontingenčné tabuľky: Výskyt pravidiel x Typ pravidiel: Súbor C1

C1\Typ	**Užitočné**	**Triviálne**	**Nevysvetliteľné**
0	3 14,29 %	44 81,48 %	0 0,00 %
1	18 85,71 %	10 18,52 %	14 100,00 %
∑	21 100 %	54 100 %	14 100 %
Pearson **Kon. koef. C** **Cramerovo V**	*Chi-square* = 45,98494; *df* = 2; *p* = 0,000 0,58367 0,71882		

V prípade analýzu súboru D1 bola hodnota kontingenčného koeficientu *Cramerovo V* (Tabuľka 15) približne 0,77. Medzi podielom užitočných, triviálnych a nevysvetliteľných pravidiel a ich výskytom v objavených pravidlách extrahovaných v súbore D1 bola veľká závislosť a kontingenčný koeficient bol štatisticky významný.

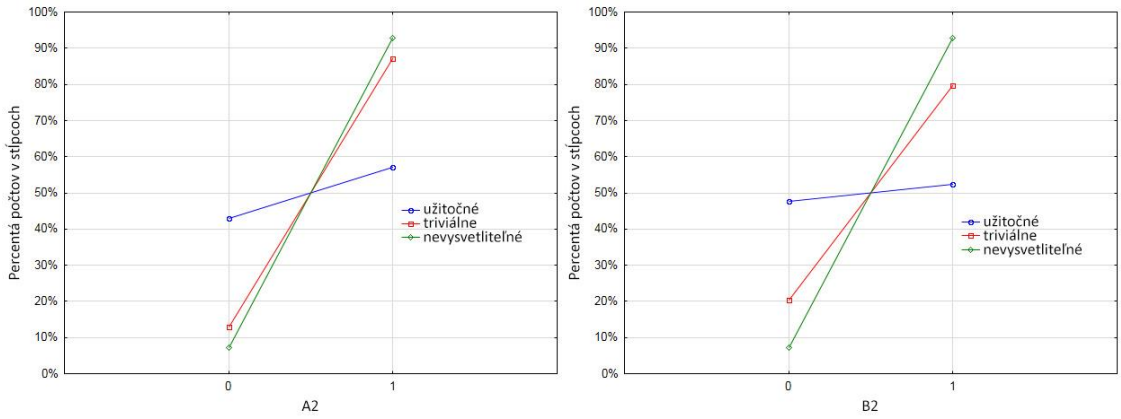
V tomto súbore bolo objavených najmenej užitočných (38 %) a triviálnych (2 %) pravidiel, pričom podiel nevysvetliteľných pravidiel sa príliš nezmenil (93 %).

Tabuľka 15 Kontingenčné tabuľky: Výskyt pravidiel x Typ pravidiel: Súbor D1

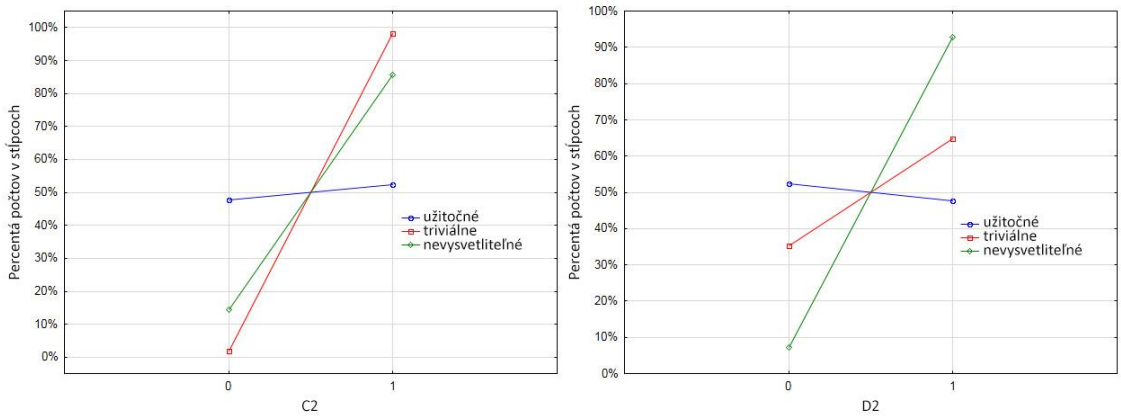
D1\Typ	**Užitočné**	**Triviálne**	**Nevysvetliteľné**
0	13 61,90 %	53 98,15 %	1 7,14 %
1	8 38,10 %	1 1,85 %	13 92,86 %
∑	21 100 %	54 100 %	14 100 %
Pearson **Kon. koef. C** **Cramerovo V**	*Chi-square* = 52,12257; *df* = 2; *p* = 0,000 0,60774 0,76528		

V prípade súborov s dopĺňaním ciest (X2, kde X = {A, B, C, D}) boli hodnoty kontingenčného koeficientu v intervale 0,2-0,5, kde 1 znamená perfektnú závislosť a 0 znamená žiadnu závislosť. Medzi podielom užitočných, triviálnych a nevysvetliteľných pravidiel a ich výskytom v sade objavených pravidiel extrahovaných zo súborov s dopĺňaním ciest bola malá-stredná závislosť.

Grafy (Obrázok 10, 11) vizualizujú frekvencie interakcie – Súbory X2 x Typ pravidiel. V tomto prípade sa krivky navzájom nekopírujú; mali rôzny smer, čo len potvrdzuje výsledky analýzy. Najviac užitočných pravidiel (57 %) bolo nájdených v súbore A2 a najmenej v súbore D2 (48 %), pokým podiel nevysvetliteľných pravidiel sa príliš nezmenil. Pri bližšom pohľade na rozdiely medzi súbormi s a bez dopĺňania ciest (X1 vs. X2, kde X = {A, B, C, D}) s ohľadom na typ pravidiel (Tabuľka 12, 13, 14, 15 a Obrázok 10, 11) sa preukázalo zvyšovanie triviálnych pravidiel a zmenšovanie počtu užitočných pravidiel v súboroch s dopĺňaním ciest (X2, kde X = {A, B, C, D}), pričom podiel nevysvetliteľných pravidiel bol podobný v oboch skupinách súborov (X1, X2, kde X = {A, B, C, D}). Dopĺňanie ciest spôsobilo iba nárast triviálnych pravidiel.



Obrázok 10 Grafy interakcií: Výskyt pravidiel x Typ pravidiel: Súbor A2 a Súbor B2



Obrázok 11 Grafy interakcií: Výskyt pravidiel x Typ pravidiel Súbor C2 a Súbor D2

Kvalita objavených pravidiel bola taktiež hodnotená pomocou dvoch indikátorov: premenných *podpora* a *spoľahlivosť* (Berry a Linoff, 2004). Výsledky sekvenčnej analýzy ukázali rozdiely nielen v kvantite objavených pravidiel, ale taktiež aj v kvalite. Kendallov koeficient zhody reprezentuje stupeň zhody v premennej *podpory* nájdených pravidiel v skúmaných súboroch. Hodnota koeficientu (Tabuľka 16) bola približne 0,71, kde 1 znamená perfektnú zhodu a 0 reprezentuje žiadnu zhodu. Z viacnásobného porovnávania (Tukeyov test) bola identifikovaná len jedna homogénna skupina (Tabuľka 16) pozostávajúca zo skúmaných súborov, týkajúca sa priemernej *podpory* nájdených pravidiel. Medzi skúmanými súbormi neboli preukázané štatisticky významné rozdiely pre premennú *podpory* pre objavené pravidlá.

Tabuľka 16 Homogénne skupiny pre podporu odvodených pravidiel

Súbor	**Priemerná podpora**	**1**
D1	2,631	****
B1	2,785	****
A1	2,819	****
C1	2,821	****
D2	3,417	****
B2	3,521	****
A2	3,557	****
C2	3,562	****
Kendallov koeficient zhody		0,71186

Rozdiely v kvalite boli prezentované pre jednotlivé súbory hodnotami *spoľahlivosti* objavených pravidiel. Kendallov koeficient zhody (Tabuľka 17) bol skoro 0,21, kde 1 znamená perfektnú zhody a 0 reprezentuje žiadnu zhodu. Z viacnásobného porovnávania (Tukeyov test) boli identifikované dve homogénne skupiny (Tabuľka 17), týkajúce sa priemernej *spoľahlivosti* objavených pravidiel. Prvá homogénna skupina pozostáva zo súborov D2, B1, C2, B2, A2, A1, C1 a druhá zo súborov D1, D2, B1, C2, B2, A2. Medzi týmito súbormi neboli identifikované štatisticky významné rozdiely pre premennú *spoľahlivosť* objavených pravidiel. Naopak štatisticky významné rozdiely v hladine významnosti 95 % boli nájdené pre priemernú *spoľahlivosť* objavených pravidiel medzi súbormi D1 a A1, ako aj D1 a C1. Najvyššia hodnota *spoľahlivosti* bola dosiahnutá v prípade súborov s identifikáciou sedení pomocou metódy *STT* založenej na priemere (C1) a pomocou metódy Reference Length založenej na odhade z mapy webu (A1) bez dopĺňania ciest.

Tabuľka 17 Homogénne skupiny pre spoľahlivosť odvodených pravidiel

Súbor	Priemerná spoľahlivosť	1	2
D1	48,724		****
D2	54,124	****	****
B1	54,945	****	****
C2	56,278	****	****
B2	56,254	****	****
A2	56,667	****	****
A1	57,621	****	
C1	58,834	****	
Kendallov koeficient zhody			0,20448

Prvý predpoklad týkajúci sa identifikácie sedení pomocou metódy Reference Length založenej na odhade z mapy webu a jej vplyvu na kvantitu nebol preukázaný. Identifikácia sedení pomocou metódy Reference Length s odhadom podielu navigačných stránok z mapy webu nemá vplyv na kvantitu extrahovaných pravidiel v prípade súborov bez a aj s dopĺňaním ciest. Naopak štvrtý predpoklad týkajúci sa dopĺňania ciest a jeho vplyvu na kvantitu bol preukázaný, dopĺňanie ciest má významný vplyv na kvantitu extrahovaných pravidiel. Štatisticky významné rozdiely v priemernom výskyte nájdených pravidiel boli preukázané medzi súbormi bez a taktiež s dopĺňaním ciest.

V prípade súborov s dopĺňaním ciest, bolo objavených skoro 50 % nových pravidiel. Na druhej strane dopĺňanie ciest spôsobilo len nárast triviálnych pravidiel. Piaty predpoklad týkajúci sa dopĺňania ciest a jeho vplyv na zvyšovanie užitočných pravidiel, nebol preukázaný. Na druhej strane, najmenej nevysvetliteľných pravidiel a najviac užitočných pravidiel bolo nájdených v súbore s identifikáciou sedení metódou Reference Length s odhadom z mapy webu. Druhý predpoklad týkajúci sa identifikácie sedení pomocou metódy Reference Length s odhadom z mapy webu a jej vplyve na zvyšovaní podielu užitočných pravidiel, bol preukázaný.

Tretí predpoklad týkajúci sa identifikácie sedení pomocou metódy Reference Length s odhadom z mapy webu a jej vplyve na kvalitu týkajúcej sa základných metrík kvality bol preukázaný len čiastočne. Štatisticky významný rozdiel bol len v spoľahlivosti medzi objavenými pravidlami v skúmaných súboroch. V prípade súboru s identifikáciou sedení pomocou metódy Reference Length s odhadom z mapy webu bez dopĺňania ciest, bola dosiahnutá najvyššia spoľahlivosť. Na druhej strane, nebol preukázaný šiesty predpoklad

týkajúci sa dopĺňania ciest a jeho vplyvu na kvalitu extrahovaných pravidiel z hľadiska základných metrík kvality.

Bolo preukázané, že identifikácia sedení pomocou metódy Reference Length a mapa webu elektronického kurzu, štatisticky významne zvyšujú počet objavených užitočných sekvenčných pravidiel. Najzaujímavejší prínos týchto výsledkov je, že aplikácia metódy Reference Length bez dopĺňania ciest ako aj prehľadávanie kurzu za účelom získania mapy webu, môže byť automatizované. Tieto závery môžu pomôcť ostatným výskumníkom v oblasti educational data mining vybrať vhodné kroky prípravy edukačných dát a ľahšie sa sústrediť na riešenie špecifických edukačných problémov.

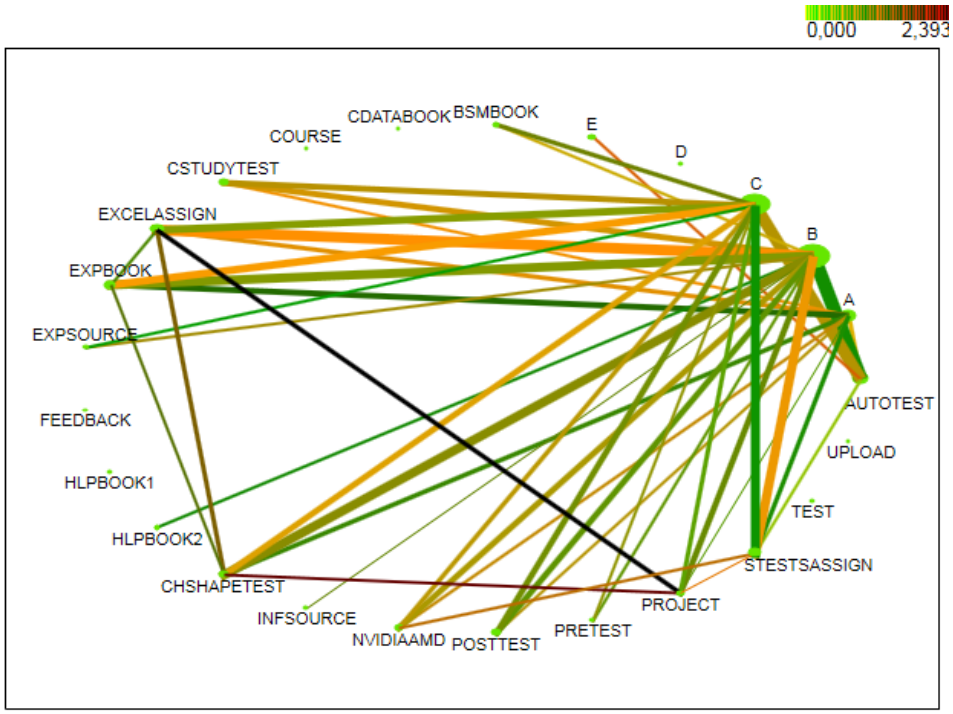
V prípade experimentu (Munk et al., 2017a) autori očakávali, že identifikácia sedení pomocou metódy Reference Length odhadnutej z mapy webu, bude mať významný vplyv na kvantitu extrahovaných pravidiel. Toto nebolo preukázané, keďže sa ukázalo, že identifikácia sedení pomocou metódy Reference Length založenej na mape webu, nemá vplyv na kvantitu extrahovaných pravidiel v prípade súborov bez a s dopĺňaním ciest. Na druhej strane predpoklad týkajúci sa vplyvu metódy založenej na mape webu na zvýšenie počtu užitočných pravidiel bol preukázaný, použitie metódy Reference Length s použitím odhadu podielu navigačných stránok z mapy webu, významne zvýšilo počet objavených užitočných sekvenčných pravidiel.

3.2.2 Analýza správania sa študentov vo VLE

Články (Benko et al., 2015; Reichel et al., 2015) boli zamerané na pozorovanie správania sa študentov v kurze VLE a vplyvu ich správania sa na výsledné hodnotenie z predmetu. Logovací súbor a jeho príprava vychádzala z predchádzajúcej kapitoly 3.2, kde identifikácia sedení bola vykonaná pomocou metódy Reference Length vypočítanej z mapy webu, pričom bola vynechaná fáza dopĺňania ciest, keďže bolo dokázané, že dopĺňanie ciest neprispieva k zvyšovaniu podielu užitočných pravidiel. Cieľom bolo analyzovať tento logovací súbor a hľadať vzory správania sa študentov v kurze VLE a ich následné porovnanie s konečnými hodnoteniami dosiahnutými na konci semestra v predmete PDA. Skúmanými premennými bola hodnotenie a kategória (alebo aktivita), ktorá obsahovala tematicky zlúčené stránky kurzu.

Graf (Obrázok 12) vizualizuje nájdené asociačné pravidlá, pričom veľkosť každého uzlu reprezentuje podporu danej aktivity. Šírka čiary, ktorá spája dve aktivity, popisuje úroveň

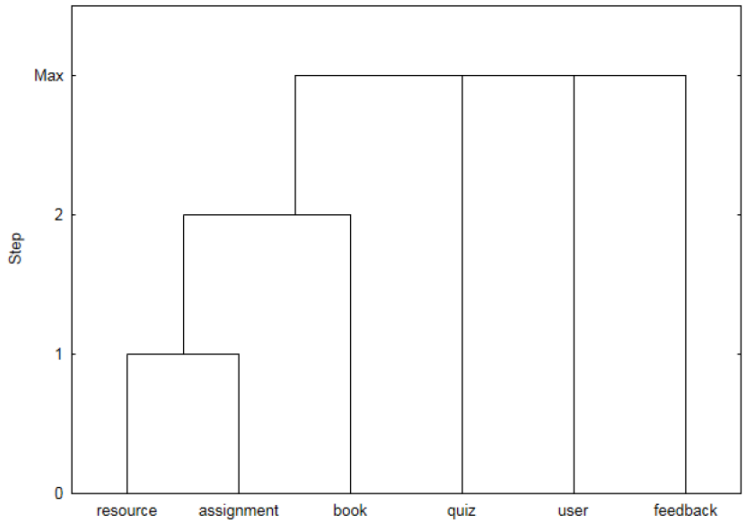
podpory pravidla, respektíve kombinácie dvoch aktivít. Jas čiary reprezentuje *lift* pravidla, čo znamená, či sa dvojica aktivít vyskytovala v jednej transakcii spolu alebo skôr oddelene. Ak je hodnota *lift* vyššia ako 1, tak sa dvojica častejšie vyskytovala spolu v jednej transakcii v súbore navštívených stránok (Reichel a Kuna, 2014). Najviac asociačných pravidiel bolo nájdených pre študentov s najlepším konečným hodnotením, naopak najmenej asociačných pravidiel sa našlo pre študentov s horším hodnotením (D a E). Konkrétne pre hodnotenie D neboli nájdené žiadne pravidlá, ktoré by spĺňali minimálnu podporu vyššiu než 1 %. Najväčší podiel zaujímavosti (*lift* = 1,5) bol identifikovaný pre pravidlo *E => AUTOTEST*. Hodnotenie E a aktivita kurzu *AUTOTEST* boli v transakciách nájdené viac spolu, než oddelene. Taktiež je možné *lift* interpretovať ako všeobecnú závislosť medzi dvoma položkami. Hodnoty väčšie ako 1 naznačujú pozitívnu závislosť a hodnoty menšie ako 1 zase negatívnu závislosť. Ak *lift* pre pravidlo *E => AUTOTEST* je rovný 1,5, tak to znamená, že študent s konečným hodnotením E by používal aktivitu *AUTOTEST* s pravdepodobnosťou 1,5 násobne vyššou ako náhodne vybraný študent. Študenti s hodnotením B a C majú túto pravdepodobnosť rovnú 1. Naopak, pravdepodobnosť návštevy aktivity *AUTOTEST* študentov s hodnotením A je 0,7 násobne nižšia než náhodne vybraný študent.



Obrázok 12 Vizualizácia nájdených pravidiel (Zdroj: (Benko et al., 2015)*)*

Výsledky výskumu (Benko et al., 2015) nepotvrdili veľký stupeň závislosti medzi konečným hodnotením študenta a správaním sa študentov v kurze, ale napriek tomu boli nájdené podstatné rozdiely v správaní študentov. Študenti, ktorí používali VLE na svoje štúdium, boli v konečnom výsledku úspešnejší. Ale výsledky preukázali aj to, že nie všetci úspešní študenti potrebovali absolvovať všetky zadania kurzu. Nadmerné používanie modulu na samo-testovanie môže indikovať, že študent má problémy s preberaným učivom. Moodle takéto informácie zaznamenáva a môžu slúžiť pre učiteľa ako podstatná informácia.

Článok (Reichel et al., 2015) sa zameral na analýzu, ako študenti využívali kurz PDA vytvorený pomocou VLE Moodle. Cieľom bolo porovnať využitie jednotlivých modulov počas semestra. Z dát boli odstránené prístupy na hlavnú stránku kurzu, pretože študenti hlavnú stránku kurzu používajú len na prístup k ostatným aktivitám. Z toho dôvodu mala hlavná stránka kurzu najviac prístupov a tieto údaje by mali nesprávny vplyv na interpretáciu dát. Rovnako boli z experimentu vylúčené aj kategórie, ktorých výskyt nedosiahol minimálne 1 %. Najväčší prístup bol zaznamenaný k aktivite *Quiz*, ktorá obsahovala hlavne samo-testovanie. Pomocou dendrogramu zhlukovej analýzy bolo možné rozdeliť jednotlivé moduly do homogénnych skupín (Obrázok 13). Výsledky zhlukovej analýzy korešpondujú s výsledkami dosiahnutými asociačnou analýzou, ktorá je podrobnejšie popísaná v (Reichel et al., 2015).



Obrázok 13 Dendrogram zobrazujúci homogénne skupiny modulov kurzu (Zdroj: (Reichel et al., 2015)*)*

Skupina modulov, ktorá zahŕňa aktivity *Resource*, *Assignment* a *Book* by sa dal zaradiť do kategórie učiacich sa modulov. Ak sa študenti chceli učiť, tak počas sedení prechádzali

hlavne spomínané moduly. Výrazne využívaný bol aj modul *Quiz*, ktorý slúžil primárne na samo-testovanie študentov. To korešponduje aj s výsledkami predchádzajúceho výskumu (Benko et al., 2015) a ukazuje, že študenti radi využívajú túto funkčnosť na zlepšovanie a overovanie svojich vedomostí. Možným odporúčaním pre tvorcov kurzu a učiteľov by bolo zlepšiť kvalitu týchto modulov.

3.3 Optimalizácia algoritmu identifikácie sedení Reference Length

Články (Munk a Benko, 2017, 2018) sa zaoberajú využitím entropie ako miery neusporiadanosti pri skúmaní podielu navigačných stránok v skúmanom logovacom súbore webového portálu. Články vychádza z predchádzajúceho výskumu (Munk et al., 2015), kde na výpočet podielu navigačných stránok slúžila mapa webu. Ako už bolo spomenuté v predchádzajúcich kapitolách (3.1, 3.2), mapa webu nemusí vždy byť aktuálna, prípadne nemusí byť dostupná a tak nie je možné vykonať výpočet odhadu podielu navigačných stránok pre potreby metódy Reference Length. V takom prípade býva jednou z možností extrahovať mapu webu priamo z logovacieho súboru, čo však môže spôsobiť získanie nekompletnej mapy webu. Alternatívu môže poskytovať využitie entropie ako miery neistoty na odhad podielu navigačných stránok priamo z logovacieho súboru.

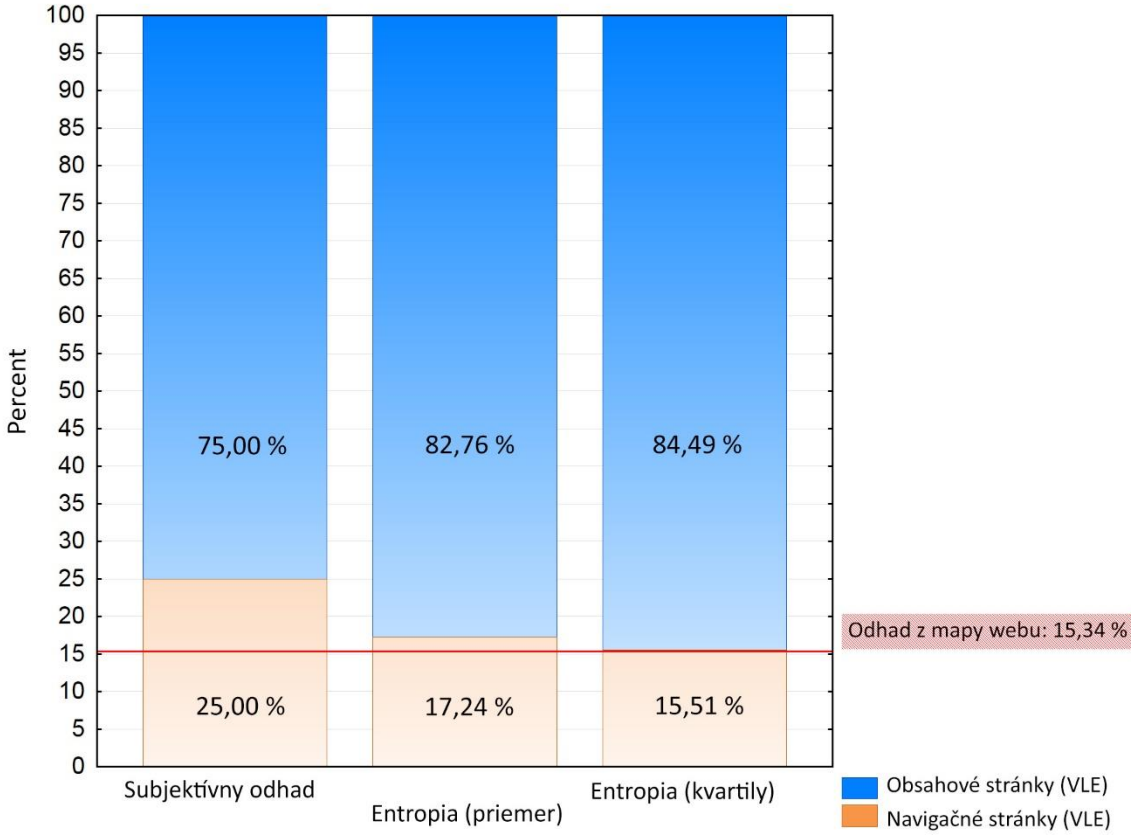
V našom experimente sme využili logovací súbor portálu virtuálneho vzdelávacieho prostredia a logovací súbor portálu s anonymným prístupom. Očakávania experimentu boli nájsť alternatívny spôsob odhadu podielu navigačných stránok v prípade chýbajúcej alebo neúplnej mapy webu. Príprava logovacieho súboru bola vykonaná pomocou štandardných techník predspracovania dát podobne ako v prípade výskumov (Benko et al., 2015; Benko, 2015; Munk et al., 2015). Pripravený logovací súbor bol následne importovaný do databázy, kde sa vykonalo viacero experimentov zahŕňajúcich entropiu, na základe ktorej by bolo možné rozlíšiť navigačné stránky od obsahových stránok na skúmanom webovom portály. Cieľom bolo vytvoriť algoritmus, ktorý by dokázal vypočítať entropiu pre konkrétnu stránku na základe náhodnej premennej *RLength*, ktorá reprezentuje dĺžku času stráveného na každej stránke webového portálu. Pomocou algoritmu bol vypočítaný *Relatívny Priemerný Čas* strávený na stránke používateľom a z tejto premennej bola vypočítaná *Entropia* pre každú stránku a bol vytvorený nový dátový súbor (Tabuľka 18) obsahujúci *Entropiu* pre každý stránku. Následne bola určená

priemerná entropia na celom portály, ktorá slúžila ako hranica, ktorá rozdeľuje navigačné stránky od obsahových. Stránky, ktorých *Entropia* bola väčšia ako *Priemerná Entropia* $Entropia_{Priemer}$ celého portálu, boli klasifikované ako navigačné. Naopak stránky s *Entropiou* menšou ako *Priemerná Entropia* celého portálu, boli klasifikované ako obsahové. Ďalším spôsobom na špecifikáciu časového ohraničenia času stráveného na webovom portály, boli kvartily. Hodnota bola vypočítaná pomocou vzorca: $Entropia_{Kvartily} = Q_{III} + 1.5Q$, kde Q_{III} reprezentuje horný kvartil a Q reprezentuje kvartilové rozpätie. Tento proces bol vykonaný pre každý logovací súbor zvlášť.

Tabuľka 18 Dátová matica logovacích súborov webového portálu rozšírená o entropie

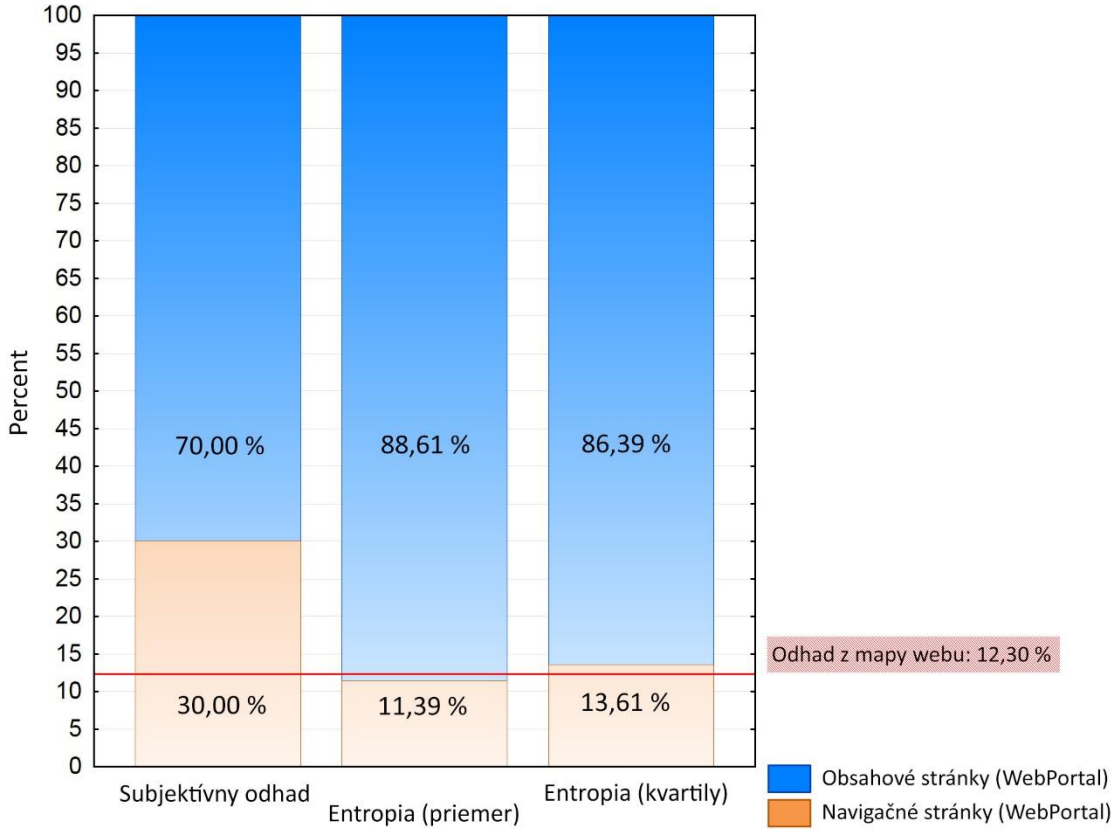
URL	Priemerné RLength	Relatívny priemerný čas	Entropia
/help-desk	39,6471	0,712336	0,241628
/informácie	52,8175	0,948967	0,049708
/konferencie	15,5	0,278487	0,356013
/fakulty	22,1082	0,397216	0,36674
⋮	⋮	⋮	⋮

Skladaný graf (Obrázok 14) vizualizuje dosiahnuté podiely obsahových a navigačných stránok porovnané metódami odhadu podielu navigačných stránok pre portál VLE. V logovacom súbore VLE portálu, bolo identifikovaných 58 stránok. Pomocou algoritmu bolo desať stránok klasifikovaných ako navigačné stránky a zvyšné stránky (48 stránok) bolo klasifikovaných ako obsahové stránky. Preto podiel navigačných stránok VLE portálu bol 17,24 % (Obrázok 14). Pomocou kvartilov bolo deväť stránok klasifikovaných ako navigačné a 49 stránok bolo identifikovaných ako obsahové stránky. Na základe výsledkov predchádzajúcich výskumov, odhad podielu z mapy webu ponúkal najlepšie výsledky pre použitie metódy Reference Length pre identifikáciu sedení. Výsledky (Obrázok 14) pre portál VLE ukázali, že *Entropia Kvartilov* ($Entropia_{Kvartily}$) bolo skoro rovnaká (15,51 %) ako odhad podielu navigačných stránok z mapy webu (15,34 %). *Entropia Priemerov* (17,24 %) bola trochu vyššia, než odhad podielu navigačných stránok z mapy webu (15,34 %), ale stále ponúka presnejšie výsledky, ako subjektívny odhad podielu navigačných stránok (25 %).



Obrázok 14 Podiel navigačných a obsahových stránok na základe rôznych odhadov podielu navigačných stránok pre webový portál VLE

Proces výpočtu entropie sa zopakoval aj pre logovací súbor webového portálu s anonymným prístupom. V tomto prípade bol rozdiel medzi subjektívnym odhadom (30 %) a odhadom z mapy webu (12,30 %) oveľa väčší ako v prípade predchádzajúceho portálu (Obrázok 15). Výsledky predchádzajúcich výskumov ukázali, že odhad podielu navigačných stránok z mapy webu ponúka lepšie výsledky ohľadne kvantity a kvality extrahovaných pravidiel, než subjektívny odhad podielu navigačných stránok. Výsledky odhadu podielu navigačných stránok pomocou entropie ukázali podobné správanie pri webovom portály s anonymným prístupom (Obrázok 15), ako aj pri portály VLE (Obrázok 14). Webový portál s anonymným prístupom obsahoval 1764 stránok. V tomto prípade bolo vhodnejšie použiť *Entropiu Priemerov* (11,39 %). Aj keď aj výsledok odhadu pomocou *Entropie kvartilov* (13,61 %) bol blízko odhadu podielu navigačných stránok pomocou mapy webu (12,30 %). Výber vhodnej metódy pre odhad podielu navigačných stránok pomocou entropie by mohol byť na základe rozhodnutia analytika webového portálu. Odhliadnuc od potreby výberu vhodnej metódy, môžeme povedať, že výsledky oboch metód ponúkajú podobné výsledky ako odhad podielu navigačných stránok pomocou mapy webu.



Obrázok 15 Podiel navigačných a obsahových stránok na základe rôznych odhadov podielu navigačných stránok pre webový portál s anonymným prístupom

V článku (Munk a Benko, 2018) bola predstavená nová metóda odhadu podielu navigačných stránok pomocou entropie (Obrázok 16) ako alternatívny spôsob odhadu podielu navigačných stránok pri identifikácií sedení pomocou metódy Reference Length. Metóda používa entropiu ako nástroj pre odhad podielu navigačných stránok z času stráveného na webových stránkach webového portálu. Keďže metóda pracuje priamo so skúmaným logovacím súborom, výsledky by mali presnejšie korešpondovať so štruktúrou skúmaného webového portálu v čase získania logovacieho súboru. Výhodou to je práve v prípade, ak pracujeme s historickými dátami alebo webovými portálmi, pre ktoré nie je mapa webu k dispozícii alebo nie je úplná. Cieľom výskumu bolo nájsť alternatívny spôsob odhadu podielu navigačných stránok.

Algoritmus Reference Length pomocou Entropie

Odhad hraničného času

1. Výpočet *Relatívneho Priemerného Času* stráveného používateľmi pre každú stránku webového portálu z premennej *Length*

$$Relatívny\ Priemerný\ Čas_j = \frac{\overline{RLegnth_j}}{\overline{RLength}}, j = 1, \dots, J,$$

kde $\overline{RLength_j}$ je priemerný čas strávený na j-tej webovej stránke a $\overline{RLength}$ je celkový priemerný čas

2. Výpočet *Entropie* z *Relatívneho Priemerného Času* pre každú stránku webového portálu
3. Výpočet $Entropia_{Kvartily}$ pre celý webový portál

$$Entropia_{Kvartily} = Q_{III} + 1{,}5Q,$$

kde Q_{III} reprezentuje horný kvartil a Q reprezentuje kvartilové rozpätie

4. Klasifikácia navigačných a obsahových stránok

$$ak\ Entropia_{Kvartily} < Entropia\ potom\ navigačná\ stránka\ inak\ obsahová\ stránka$$

5. Odhad podielu navigačných stránok $p_{Entropia}$

$$p_{Entropia} = \frac{počet\ navigačných\ stránok}{počet\ všetkých\ stránok}$$

6. Výpočet hraničného času C

$$C = \frac{-\ln(1-p_{Entropia})}{\lambda},$$

kde $p_{Entropia}$ je odhad podielu navigačných stránok a odhad parametra λ je $\hat{\lambda} = \frac{1}{\overline{RLength}}$, kde $\overline{RLength}$ je pozorovaný priemer časov strávených na webových stránkach

Identifikácia sedení

1. Pre skúmaný logovací súbor bola vytvorená sekvencia navštívených webových stránok

$$\langle USID, \langle URL_1, DTime_1, RLength_1\rangle, \dots, \langle URL_k, DTime_k, RLength_k\rangle\rangle,$$

kde $USID$ je identifikácia sedení, URL_i je pristupovaná webová stránka s časovou známkou $DTime_i$ a $RLength_i$ je čas strávený používateľom na každej stránke webového portálu s nasledovnou vlastnosťou

$$RLength_i \leq C,$$

kde $1 \leq i < k$ a pre poslednú stránku sedenia platí

$$RLength_k > C$$

2. Od nasledujúcej stránky je definované nové sedenie

Obrázok 16 Algoritmus metódy Reference Length s využitím entropie

3.4 NÁVRH METODIKY NA ZHODNOTENIE FREKVENTOVANÝCH TRANSAKCIÍ/SEKVENCIÍ V ČASE

Výskum prezentovaný v článku (Munk, Pilkova, Benko, Blažeková, 2017b) prispieva k preklenutiu medzery v oblasti dostatočného zverejňovania informácií, čo taktiež prispieva k zvyšovaniu záujmu príslušných stakeholderov o prispievanie k trhovej disciplíne a je relevantné s ich záujmami v rámci Pilieru 3. Článok sa zaoberá analýzou

údajov z webovej stránky zameraných na zverejňovanie informácií komerčných bánk k Pilieru 3 a skúmaním správania sa stakeholderov vo vzťahu k vážnym trhovým turbulenciám. Skúmané dáta pozostávajú z logovacích súborov, ktoré boli predspracované pomocou techník web mining-u a z ktorých boli extrahované frekventované položkové množiny na základe kvartálov a boli vyhodnocované na základe kvantity.

3.4.1 METODIKA

Jedným z čiastočných cieľov bolo vytvorenie metodiky na zhodnotenie frekventovaných transakcií/sekvencií v čase, ktorá bola využitá v článku (Munk et al., 2017b). Pre potreby analýzy správania sa používateľov webu boli zozbierané dáta súvisiace s Pilierom 3 z logovacích súborov bankového webového servera (Munk et al., 2012). Príprava dát bola vykonaná podľa (Kapusta et al., 2013, 2014a), kde bolo potrebné pripraviť logovacie súbory z viacerých serverov, ktoré boli využívané ako vyrovnávače zaťaženia (tzv. load balancer). Metodológia výskumu bola inšpirovaná (Munk et al., 2010a, 2010b, 2015) a aplikovaná na evalváciu frekventovaných položkových množín na základe kvantity. V analýze boli položkové množiny vyhodnocované na základe kvartálov počas rokov 2009-2015. Zdrojom údajov pre našu analýzu boli logovacie súbory z webových serverov významnej domácej komerčnej banky pôsobiacej na Slovensku. Analýza bol vykonaný na vzorke 10 378 751 logovacích prístupov, ktoré boli získané po príprave dát, ktorá zahŕňala čistenie dát, identifikáciu sedení a dopĺňanie ciest. Použitá metodológia bola nasledovná:

1. Získanie logovacích súborov z viacerých serverov.
2. Príprava dát pozostávajúca z viacerých úloh:
    a. Čistenie dát – najdôležitejší krok je očistiť dáta od nepotrebných údajov. Tento krok vedie k získaniu surových dát obsahujúcich len prístupy na webový portál. Čistenie dát v sebe zahŕňa aj odstránenie prístupov robotov vyhľadávacích služieb.
    b. Identifikácia používateľov/sedení – návštevníci boli identifikovaný na základe poľa User Agent a sedenia boli identifikované pomocou metódy Reference Length.

c. Rekonštrukcia aktivít používateľov (dopĺňanie ciest) – zamerané na spätné dopĺňanie záznamov o ceste používateľa webovým portálom pomocou tlačidla „Späť“ vo webovom prehliadači.

3. Analýza dát pozostávajúce z hľadania vzorov správania sa používateľov webu počas kvartálov v skúmanom období. Výsledky boli spracované asociačnou analýzou pomocou *STATISTICA Sequence, Association, & Link Analysis*, ktorá obsahuje implementáciu algoritmu používajúceho a-priori algoritmus spolu s procedúrou stromovej štruktúry, ktorá požaduje len jeden prechod dátami (Statsoft Inc., 2013). Podpora položkovej množiny je daná podielom záznamov v transakcií, ktorá obsahuje položkovú množinu. To znamená, že pre položkovú množinu (A) môže byť podpora (*support*) vypočítaná nasledovne:

$$Podpora(A) = \frac{početnosť\ (A)}{počet\ transakcií\ v\ dátovej\ množine} * 100.$$

Lift pravidiel môže byť vypočítaný podobne. Na základe podpory a spoľahlivosti môžeme určiť Lift pre pravidlo:

$$Lift(ak\ A\ potom\ C) = \frac{spoľahlivosť(ak\ A\ potom\ C)}{podpora(C)},$$

kde $Spoľahlivosť(ak\ A\ potom\ C) = \frac{podpora(ak\ A\ potom\ C)}{podpora(A)} * 100.$

Sústredili sme sa na frekventované položkové množiny extrahované s minimálnou podporou 1 % (Pilkova et al., 2015).

4. Porozumenie výstupným dátam získaných analýzou a definovanie predpokladov.
5. Porovnanie výsledkov analýzy dát z jednotlivých kvartálov skúmaného obdobia.

3.4.2 VÝSLEDKY

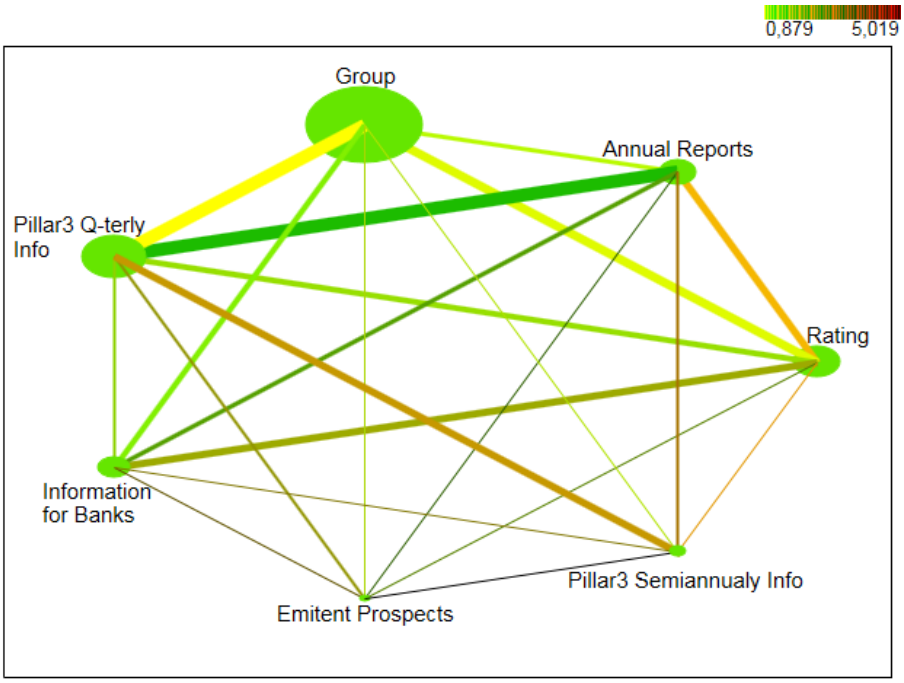
Grafy (Obrázok 16-22) zobrazujú nájdené frekventované položkové množiny, kde veľkosť každého uzlu reprezentuje *podporu* danej webovej časti- 1-položková množina (množina pozostávajúca len z jednej položky). Hrúbka čiary spájajúcej dve webové časti reprezentuje úroveň *podpory* 2-poločkovej množiny alebo kombinácie dvoch webových častí. Jas čiary reprezentuje *lift* páru webových častí.

Počas prvého kvartálu roku 2009 (Obrázok 17) patrila webová časť */Group/* medzi najnavštevovanejšie webové časti s *podporou* skoro 60 %. Webové časti */Pillar3 Q-terly Info/* a */Rating/* sa vyskytovali v identifikovaných sedeniach s pravdepodobnosťou viac ako 30 %, podobne */Annual Reports/* a */Information for Banks/* s pravdepodobnosťou väčšou ako 20 %. Webové časti */Pillar3 Semiannualy Info/* a */Emitent Prospects/*

s pravdepodobnosťou okolo 15 % patrili k menej populárnym. Webové časti */General Shareholder Meeting/* a */Financial Reports/* nesplnili minimálnu *podporu*, resp. pravdepodobnosť výskytu v identifikovaných sekvenciách (transakciách) je menšia ako 1 %.

V prvom kvartály roku 2009 (Obrázok 17), dvojice (*/Group/, /Pillar3 Q-terly Info/*) a (*/Annual Reports/, /Pillar3 Q-terly Info/*) s približne 19 % *podporou* patrili medzi najnavštevovanejšie webové dvojice. Dvojice (*/Rating/, /Group/*), (*/Pillar3 Q-terly Info/, /Pillar3 Semiannualy Info/*), (*/Rating/, /Information for Banks/*), (*/Rating/, /Annual Reports/*), (*/Group/, /Information for Banks/*) a (*/Rating/, /Pillar3 Q-terly Info/*) sa vyskytovali v identifikovaných sekvenciách s pravdepodobnosťou okolo 16 %. Zvyšné dvojice dosiahli pravdepodobnosť v intervale 13-15 %.

Webové časti sú nezávislé (*lift = 1*) v prípade dvojíc (*/Group/, /Pillar3 Q-terly Info/*), (*/Rating/, /Group/*), (*/Annual Reports/, /Group/*) a (*/Group/, /Information for Banks/*). Na rozdiel od ostatných dvojíc, pozitívna korelácia (*lift* > 1) bola identifikovaná a webové časti sa vyskytujú častejšie spolu, než oddelene v identifikovaných sekvenciách. Najvyššia úroveň pozitívnej korelácie (*lift* = 5,02) bola získaná pre pár (*/Emitent Prospects/, /Pillar3 Semiannualy Info/*), tj. ak sa webová časť */Emitent Prospects/* nachádza v sekvencií, tak je 5,02 krát pravdepodobnejšie, že sa tam taktiež nachádza webová časť */Pillar3 Semiannualy Info/*, než v náhodne vybranom sedení. Rovnako to platí aj obrátene, odhliadnuc od orientácie dvojice-pravidla. Podobne vysoká úroveň pozitívnej korelácie (*lift*: 2,5 až 3,5) bola dosiahnutá pre nasledujúce dvojice (*/Information for Banks/, /Emitent Prospects/*), (*/Annual Reports/, /Emitent Prospects/*), (*/Information for Banks/, Pillar3 Semiannualy Info/*), (*/Annual Reports/, /Pillar3 Semiannualy Info/*), (*/Rating/, /Emitent Prospects/*), (*/Pillar3 Q-terly Info/, /Emitent Prospects/*), (*/Rating/, /Pillar3 Semiannualy Info/*), (*/Pillar3 Q-terly Info/, /Pillar3 Semiannualy Info/*), (*/Annual Reports/, /Information for Banks/*) a (*/Rating/, /Information for Banks/*). Zostávajúce dvojice dosiahli *lift* na úrovni rozpätia 1,5 až 2.



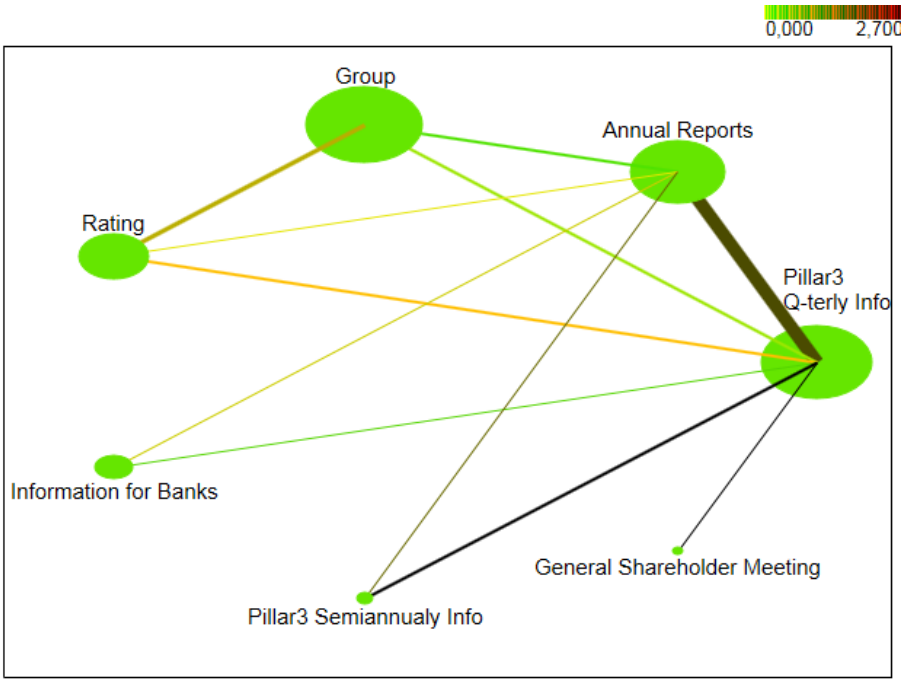
Obrázok 17 Vizualizácia prvého kvartálu roku 2009

Počas prvého kvartálu roku 2010 (Obrázok 18) patrili webové časti */Group/*, */Pillar3 Q-terly Info/* k najnavštevovanejším webovým častiam s *podporou* skoro 40 %. Webové časti */Annual Reports/ a /Rating/* sa vyskytovali s pravdepodobnosťou väčšou ako 20 % a podobne sa vyskytovala s pravdepodobnosťou okolo 10 % webová časť */Information for Banks/*. Menej populárne webové časti */Pillar3 Semiannualy Info/*, */General Shareholder Meeting/*, */Emitent Prospects/* a */Financial Reports/* mali pravdepodobnosť výskytu približne alebo menej ako 1 %.

Počas prvého kvartálu roku 2010 (Obrázok 18) bola najviac navštevovaná dvojica webových častí (*/Pillar3 Q-terly Info/*, */Annual Reports/*) s *podporou* väčšou ako 20 %. Dvojice (*/Group/*, */Rating/*) a (*/Pillar3 Semiannualy Info/*, */Pillar3 Q-terly Info/*) sa vyskytovali v identifikovaných transakciách s *podporou* približne 6 %. Zvyšné dvojice dosiahli pravdepodobnosť menšiu než 3 %.

Vysoká úroveň pozitívnej korelácie (*lift* = 2,7) bola objavená pre dvojice (*/General Shareholder Meeting/*, */Pillar3 Q-terly Info/*) a (*/Pillar3 Semiannualy Info/*, */Pillar3 Q-terly Info/*). Pre dvojice (*/Annual Reports/*, */Pillar3 Q-terly Info/*) a (*/Pillar3 Semiannualy Info/*, */Annual Reports/*) bola tiež objavená pozitívna korelácia (*lift*: 1,5 až 2). V prípade dvojice (*/Ratign/*, */Group/*) sú webové časti považované za nezávislé. Zvyšné webové

časti dosiahli úroveň negatívnej korelácie (*lift*: 0,2 až 0,5) čo znamená, že v identifikovaných sedeniach boli nájdené zväčša oddelene.

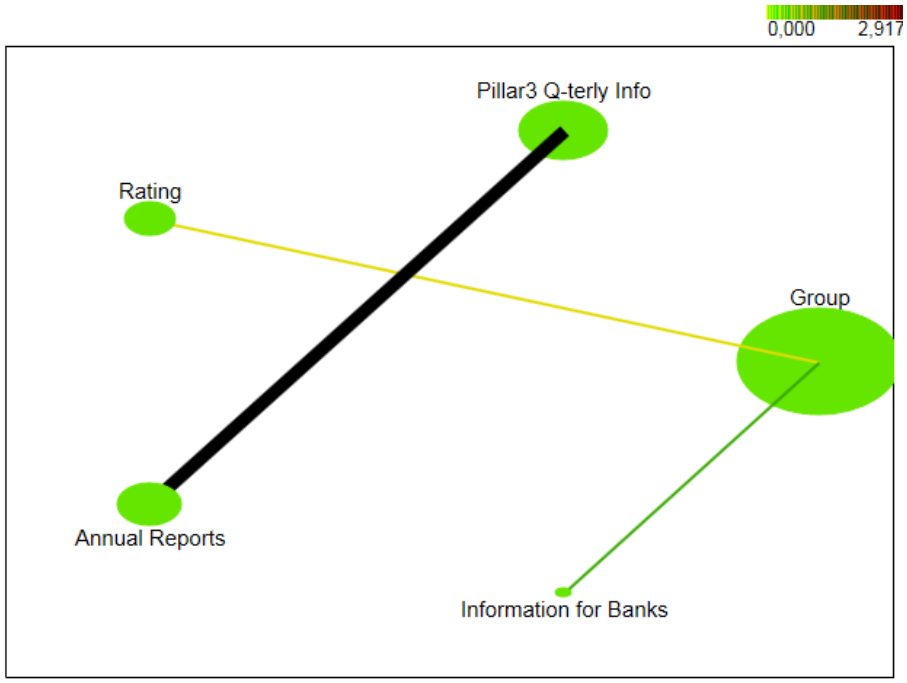


Obrázok 18 Vizualizácia prvého kvartálu roku 2010

V prvom kvartály roku 2011 (Obrázok 19) bola webová časť */Group/* medzi najviac zaujímavé s pravdepodobnosťou 50 %. Webové časti */Pillar3 Q-tearly Info/*, */Annual Report/* a */Ratings/* patrili k menej zaujímavým v porovnaní s webovou častou */Group/* s pravdepodobnosťou približne 20 %. Webová časť */Information for Banks/* mala *podporu* nižšiu než 10 % a webové časti */Pillar3 Semiannualy Info/*, */Emitent Prospects/*, */General Shareholder Meeting/* a */Financial Reports/* mali pravdepodobnosť prístupu nižšiu ako 1 %.

Dvojica (*/Pillar3 Q-terly Info/*, */Annual Reports/*) patrila počas prvého kvartálu roku 2011 (Obrázok 19) s *podporou* 20 % medzi najnavštevovanejšie webové časti. Pravdepodobnosť výskytu identifikovaných dvojíc (*/Group/*, */Rating/*) a (*/Information for Banks/*, */Group/*) dosiahla len približne 5 %. Na zostávajúce dvojice webových častí pristupovali návštevníci s pravdepodobnosťou menšou než 3 %.

Medzi dvojicou webových častí (*/Group/*, */Information for Banks/*) bola objavená nezávislosť (*lift* = 1). Najvyššia úroveň zaujímavosti bola objavená pre dvojicu (*/Pillar3 Q-terly Info/*, */Annual Reports/*) s hodnotu *lift* = 2,9. Na druhej strane dvojica (*/Group/*, */Rating/*) mala negatívnu koreláciu (*lift* = 0,4).

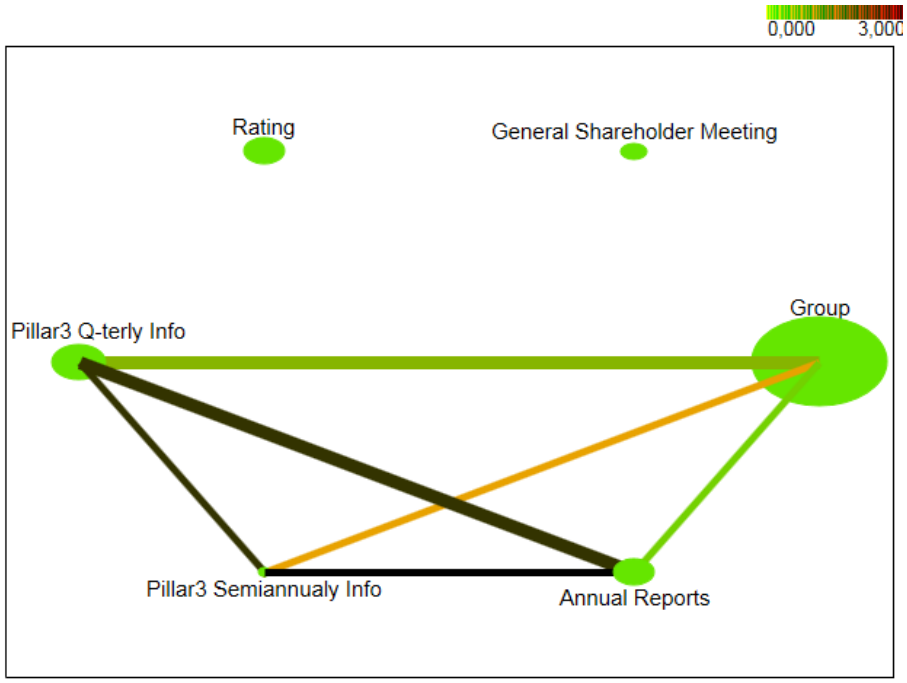


Obrázok 19 Vizualizácia prvého kvartálu roku 2011

Počas prvého kvartálu roku 2012 (Obrázok 20) preukazovali návštevníci portálu podobné správania ako po predchádzajúce roky, kde webová časť */Group/* mala najvyššiu *podporu* 45 %. Webové časti */Pillar3 Q-terly Info/*, */Rating/* a */Annual Reports/* boli pre návštevníkov zaujímavé s pravdepodobnosťou približne 20 %. *Podpora* okolo 10 % bola preukázaná pre webové časti */General Shareholder Meeting/* a */Pillar3 Semiannualy Info/*. Minimálna *podpora* 1 % nebola dosiahnutá pre webové časti */Information for Banks/*, */Emitent Prospects/* a */Financial Reports/*.

Pravdepodobnosť prístupu na dvojice webových častí bola počas prvého kvartálu roku 2012 nižšia (Obrázok 20), pričom dvojice (*/Group/*, */Pillar3 Q-terly Info/*) a (*/Pillar3 Q-terly Info/*, */Annual Reports/*) dosiahli najvyššiu pravdepodobnosť približne 10 %. Ostatné dvojice boli navštevované s *podporou* menšou ako 5 %.

Najvyššia pozitívna korelácia (*lift* = 3) bola nájdená (Obrázok 20) pre pravidlo (*/Annual Reports/*, */Pillar3 Semiannualy Info/*). Podobne vyššiu úroveň (*lift* = 2,4) sme našli pre dvojice (*/Pillar3 Q-terly Info/*, */Annual Reports/*) a (*/Pillar3 Semiannualy Info/*, */Pillar3 Q-terly Info/*). Dvojice webových častí (*/Pillar3 Semiannualy Info/*, */Group/*) a (*/Pillar3 Q-terly Info/*, */Group/*) preukazovali znaky nezávislosti (*lift* = 1). Negatívna korelácia (*lift* = 0,5) bola nájdená pre dvojicu (*/Group/*, */Annual Reports/*).

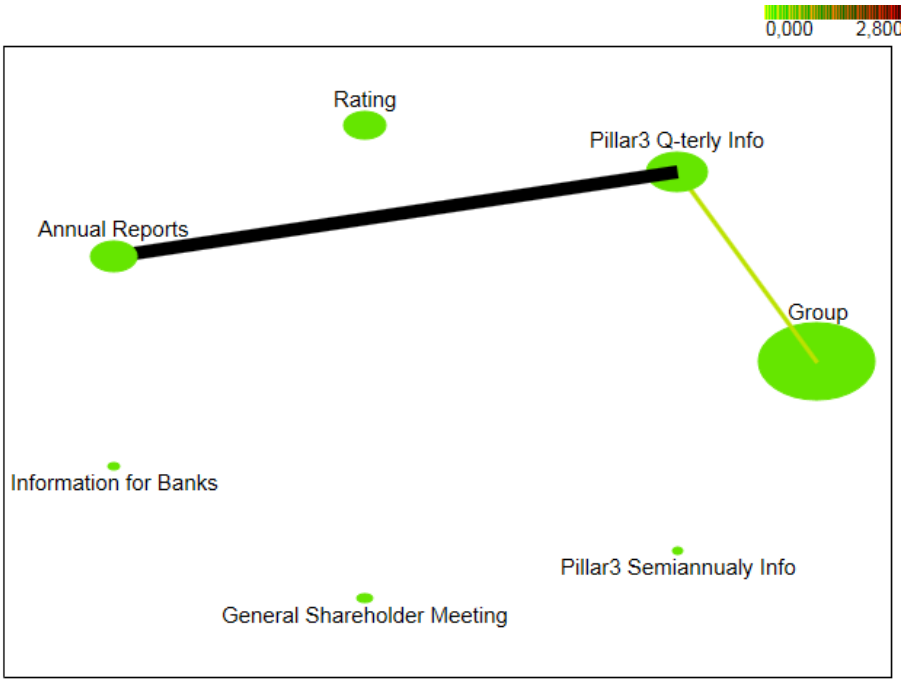


Obrázok 20 Vizualizácia prvého kvartálu roku 2012

Počas prvého kvartálu roku 2013 (Obrázok 21) patrila webová časť */Group/* medzi najviac navštevovanú webovú časť s *podporou* 50 %. Medzi ďalšie populárne webové časti patrili */Pillar3 Q-terly Info/*, */Annual Reports/* a */Rating/* s *podporou* približne 20 %. Webové časti */General Shareholder Meeting/*, */Information for Banks/*, */Pillar3 Semiannualy Info/* s *podporou* menšou ako 5 % boli zaznamenané ako menej zaujímavé. Pravdepodobnosť prístupu pre webové časti */Emitent Prospects/* a */Financial Reports/* nespĺňala požadovanú *podporu* 1 %.

Len dve dvojice webových častí sa vyskytovali v identifikovaných sekvenciách počas prvého kvartálu roku 2013 (Obrázok 21) – (*/Pillar3 Q-terly Info/*, */Annual Reports/*) s *podporou* skoro 15 % a (*/Group/*, */Pillar3 Q-terly Info/*) s *podporou* menšou než 5 %.

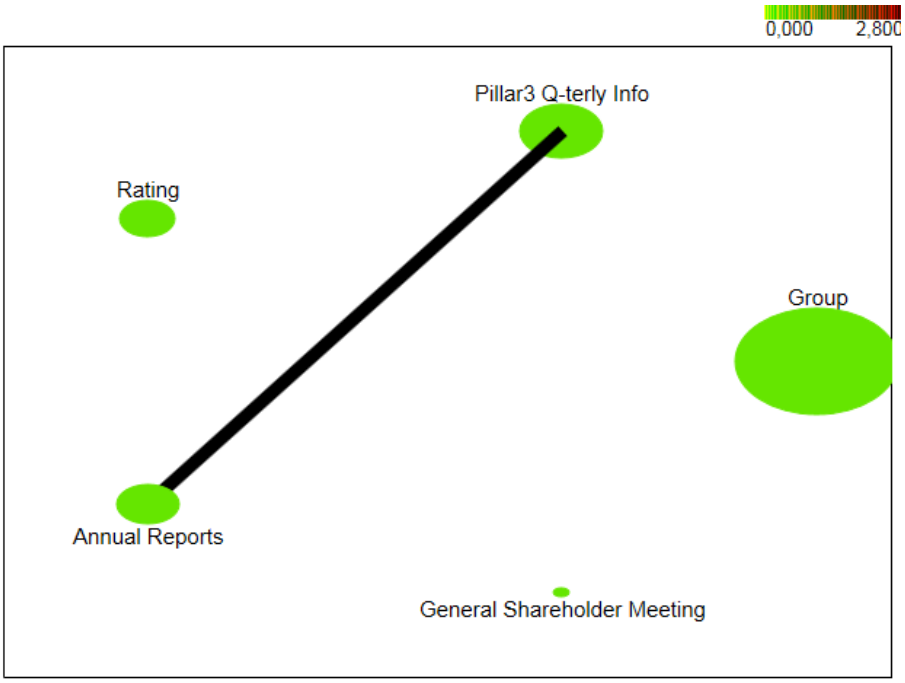
Najvyššia úroveň zaujímavosti (*lift* = 2,8) bola objavená (Obrázok 21) pre pravidlo (*/Annual Reports/*, */Pillar3 Q-terly Info/*). Na druhej strane webové časti (*/Group/*, */Pillar3 Q-terly Info/*) boli nachádzané v identifikovaných sedeniach častejšie oddelené (*lift* = 0,3).



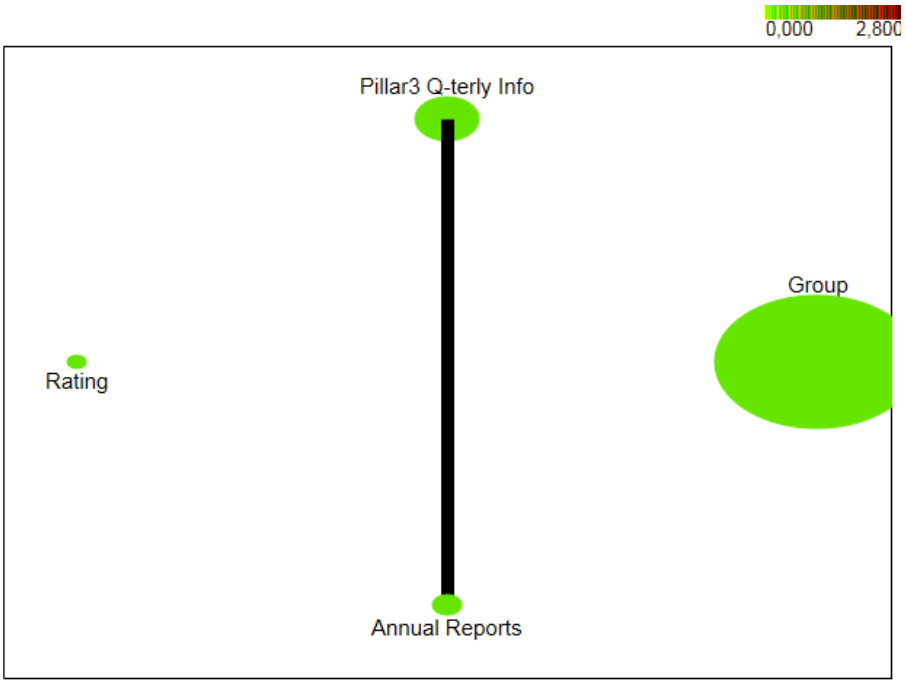
Obrázok 21 Vizualizácia prvého kvartálu roku 2013

Počas prvého kvartálu rokov 2014 (Obrázok 22) a 2015 (Obrázok 23) sa návštevníci webu správali skoro rovnako. V prvom kvartály bola webová časť */Group/* najviac navštevovanejšou webovou časťou s *podporou* 50 %. Webové časti */Pillar3 Q-terly Info/*, */Annual Reports/* a */Rating/* patrili taktiež k často navštevovaným s pravdepodobnosťou okolo 20 %. Jediným rozdielom v správaní bola webová časť */General Shareholder Meeting/*, ktorý v prvom kvartály roku 2014 navštevovali s *podporou* 5 %, ale v prvom kvartály roku 2015 nesplnila webová časť minimálnu *podporu* 1 %. Zvyšné webové časti */Information for Banks/*, */Pillar3 Semiannualy Info/*, */Emitent Prospects/* a */Financial Reports/* nesplnili limit minimálnej *podpory* 1 % počas prvého kvartála roku 2014 a taktiež aj počas prvého kvartálu roku 2015.

Počas prvého kvartálu rokov 2014 (Obrázok 22) a 2015 (Obrázok 23) sa vyskytovala len jedna webová dvojica (*/Pillar3 Q-terly Info/*, */Annual Reports/*) v identifikovaných sedeniach s pravdepodobnosťou skoro 15 %. Okrem toho preukazovala táto dvojica webových častí najvyššiu pozitívnu koreláciu (*lift* = 2,8).



Obrázok 22 Vizualizácia prvého kvartálu roku 2014



Obrázok 23 Vizualizácia prvého kvartálu roku 2015

Na základe Cochranovho Q testu zamietame nulovú hypotézu na hladine významnosti 0,1 % (Q = 65,8594; df = 6; p < 0,0000), ktorá tvrdí, že výskyt frekventovaných položkových množín webových častí nezávisí od času. Najviac frekventovaných položkových množín v prvých kvartáloch (Tabuľka 19) bolo identifikovaných v roku

2009 (približne 64 %), naopak najmenej frekventovaných položkových množín bolo identifikovaných v rokoch 2014 a 2015 (približne 11 % až 14 %).

Z viacnásobného porovnávania (Tabuľka 19) boli identifikované tri homogénne skupiny (15Q1, 14Q1, 11Q1, 13Q1, 12Q1), (13Q1, 12Q1, 10Q1) a (09Q1) na základe priemerného výskytu nájdených frekventovaných položkových množín webových častí. Štatisticky významné rozdiely boli preukázané na hladine významnosti 5 %, čo sa týka priemerného výskytu nájdených frekventovaných položkových množín medzi kvartálom 09Q1 a ostatnými, ako aj medzi kvartálom 10Q1/12Q1/13Q1 a kvartálmi (11Q1, 14Q1, 15Q1).

Tabuľka 19 Homogénne skupiny pre výskyt frekventovaných položkových množín webových častí pre prvý kvartál počas skúmaných rokov

Rok	**Percent**	**1**	**2**	**3**
15Q1	11,36 %	****		
14Q1	13,64 %	****		
11Q1	18,18 %	****		
13Q1	20,45 %	****	****	
12Q1	27,27 %	****	****	
10Q1	40,91 %		****	
09Q1	63,64 %			****

V prípade druhého kvartálu (Q = 27,51515; df = 6; p < 0,000116) sa nulová hypotéza zamieta na hladine významnosti 0,1 % a najviac frekventovaných položkových množín (Tabuľka 20) bolo identifikovaných v roku 2009 (viac ako 45 %), najmenej v rokoch 2014 a 2015 (približne 18 až 20 %).

Z viacnásobných porovnávaní (Tabuľka 20) boli identifikované, na základe priemerného výskytu nájdených frekventovaných položkových množín webových častí, tri homogénne skupiny (15Q2, 14Q2, 12Q2, 13Q2), (14Q2, 12Q2, 13Q2, 11Q2, 10Q2) a (13Q2, 11Q2, 10Q2, 09Q2). Štatisticky významné rozdiely boli preukázané na hladine významnosti 5 %, kde priemerný výskyt nájdených frekventovaných položkových množín medzi kvartálom 09Q2 a (12Q2, 14Q2, 15Q2) ako aj medzi kvartálom 10Q2/11Q2 a kvartálom 15Q2.

Tabuľka 20 Homogénne skupiny pre výskyt frekventovaných položkových množín webových častí pre druhý kvartál počas skúmaných rokov

Rok	Percent	1	2	3
15Q2	18,18 %		****	
14Q2	20,45 %	****	****	
12Q2	22,73 %	****	****	
13Q2	29,55 %	****	****	****
11Q2	38,64 %	****		****
10Q2	38,64 %	****		****
09Q2	45,45 %			****

Pre tretí (Q = 18,54545; df = 6; p < 0,005005) a štvrtý (Q = 32,04706; df = 6; p < 0,000016) kvartál počas skúmaného časového obdobia sa nulová hypotéza zamieta na hladine významnosti 1 %. Najviac frekventovaných položkových množín bolo identifikovaných v roku 2009 (viac ako 38 % pre tretí kvartál (Tabuľka 21) a viac ako 45 % pre štvrtý kvartál (Tabuľka 22)). Na druhej strane najmenej objavených frekventovaných položkových množín bolo v rokoch 2012 a 2015 (približne 16 % v oboch kvartáloch).

Tabuľka 21 Homogénne skupiny pre výskyt frekventovaných položkových množín webových častí pre tretí kvartál počas skúmaných rokov

Rok	Percent	1	2
12Q3	15,91 %	****	
15Q3	15,91 %	****	
11Q3	20,45 %	****	
13Q3	22,73 %	****	****
14Q3	22,73 %	****	****
10Q3	22,73 %	****	****
09Q3	38,64 %		****

Na základe viacnásobných porovnávaní (Tabuľka 21), štatisticky významné rozdiely boli identifikované na hladine významnosti 5 %, v priemernom výskyte nájdených frekventovaných položkových množín medzi kvartálom 09Q3 a kvartálmi (11Q3, 12Q3, 15Q3).

Štatisticky významné rozdiely boli preukázané na hladine významnosti 5 %, priemerný výskyt nájdených frekventovaných položkových množín (Tabuľka 22) medzi kvartálom 09Q4 a ostatnými okrem kvartálu 13Q4.

Tabuľka 22 Homogénne skupiny pre výskyt frekventovaných položkových množín webových častí pre štvrtý kvartál počas skúmaných rokov

Rok	Percent	1	2
12Q4	15,91 %	****	
15Q4	15,91 %	****	
14Q4	18,18 %	****	
10Q4	20,45 %	****	
11Q4	22,73 %	****	
13Q4	29,55 %	****	****
09Q4	45,45 %		****

Udalosť celosvetovej krízy (2009) mala významný dopad na kvantitu extrahovaných frekventovaných položkových množín webových častí.

V prípade kvartálov počas roku 2009 (Q = 8,257732; df = 3; p < 0,040977) sa nulová hypotéza zamieta na hladine významnosti 5 %. Najviac frekventovaných položkových množín v roku 2009 (Tabuľka 23) bolo identifikovaných v prvom kvartály (viac ako 63 %), najmenej v treťom kvartály (približne 39 %).

V roku 2009 boli identifikované dve homogénne skupiny (09Q3, 09Q4, 09Q2) a (09Q4, 09Q2, 09Q1) na základe priemerného výskytu nájdených frekventovaných položkových množín webových častí.

Pre skúmané roky (2010: Q = 12,58065; df = 3; p < 0,005638; 2011: Q = 11,53846; df = 3; p < 0,009144) sa nulová hypotéza zamieta na hladine významnosti 1 %. V roku 2010 (Tabuľka 24) bolo najviac frekventovaných položkových množín nájdených pre prvý kvartál (okolo 41 %), najmenej v štvrtom a treťom kvartály (približne 20 až 21 %). Na druhej strane v roku 2011 (Tabuľka 25) najviac frekventovaných položkových množín bolo identifikovaných v druhom kvartály (okolo 39 %), najmenej v prvom a treťom kvartály (približne 18 až 21 %).

Tri homogénne skupiny (10Q4, 10Q3), (10Q3, 10Q2) a (10Q2, 10Q1) boli nájdené pre rok 2010 (Tabuľka 24) na základe priemerného výskytu nájdených frekventovaných

položkových množín webových častí. Na druhej strane dve homogénne skupiny (11Q1, 11Q3, 11Q4) a (11Q4, 11Q2) boli identifikované pre kvartály v roku 2011 (Tabuľka 25).

Pre zostávajúce roky (2012: $Q = 4{,}153846$; $df = 3$; $p < 0{,}245326$; 2013: $Q = 3{,}255319$; $df = 3$; $p < 0{,}353912$; 2014: $Q = 4{,}565217$; $df = 3$; $p < 0{,}206549$; 2015: $Q = 3.000000$; $df = 3$; $p < 0{,}391627$) neboli nájdené štatisticky významné rozdiely.

Tabuľka 23 Homogénne skupiny pre výskyt frekventovaných položkových množín webových častí pre rok 2009

Kvartál	Percent	1	2
09Q3	38,64 %	****	
09Q4	45,45 %	****	****
09Q2	45,45 %	****	****
09Q1	63,64 %		****

Tabuľka 24 Homogénne skupiny pre výskyt frekventovaných položkových množín webových častí pre rok 2010

Kvartál	Percent	1	2	3
10Q4	20,45 %	****		
10Q3	22,73 %	****	****	
10Q2	38,64 %		****	****
10Q1	40,91 %			****

Tabuľka 25 Homogénne skupiny pre výskyt frekventovaných položkových množín webových častí pre rok 2011

Kvartál	Percent	1	2
11Q1	18,18 %	****	
11Q3	20,45 %	****	
11Q4	22,73 %	****	****
11Q2	38,64 %		****

Prvé kvartály v rokoch 2009 a 2010 počas celosvetovej finančnej krízy mali významný dopad na kvantitu extrahovaných frekventovaných položkových množín webových častí. Výsledky taktiež naznačujú, že banková stakeholderi sa najviac zaujímajú o stav skupiny, čo je v logickým súvisom s Hasan *et al.* (Hasan et al., 2013).

Naše výsledky naznačujú, že ďalšie zmeny vo zverejňovaní informácií komerčnými bankami sú nevyhnutné, ak majú byť mechanizmy trhovej disciplíny efektívne a používané podľa očakávaní a požiadaviek regulátora. Podľa BIS komisie (Bank for International Settlements), by sa mali vyhýbať zverejňovaným informáciám v decembri roku 2015, ktoré nepridávajú výpovednú hodnotu pre pochopenie návštevníkov alebo nekomunikujú užitočné informácie. Okrem toho, informácie, ktoré už viac nemajú zmysel alebo sa netýkajú návštevníkov, by mali byť odstránené. Naše výsledky a závery potvrdzujú tieto nariadenia. Ale kvôli určitým obmedzeniam pri našom výskume, je nutný hlbší výskum v problematike.

V článku (Munk et al., 2017b) sme navrhli metodológiu na evalváciu frekventovaných položkových množín webových častí počas skúmaného časového obdobia. Výsledky výskumu ukazujú, že záujem stakeholderov o zverejňované údaje je nižší po turbulenciách v roku 2009, vyšší v prvom kvartály, taktiež vyšší v spojitosti s výročnými správami (nižší pre samostatné informácie o Pilieri 3). Výsledky výskumu naznačujú, že ďalšie zmeny v zverejňovaní informácií komerčných bánk sú nevyhnutné na dosiahnutie účinného mechanizmu trhovej disciplíny a zmysluplných zverejňovaní podľa očakávania regulátora.

3.5 NÁVRH METODIKY PREDIKCIE PRAVDEPODOBNOSTÍ PRÍSTUPOV NA WEBOVÉ ČASTI PORTÁLU V ZÁVISLOSTI OD ČASU

Modelovali sme pravdepodobnosti prístupu na webové časti skúmaného portálu (*category*) v závislosti od času, kde čas bol reprezentovaný premennou *týždeň* (*week*) (Munk, Pilkova, Benko, Blažeková, 2018). Logovací súbor pochádza z viacerých webových serverov domácej významnej komerčnej banky pôsobiacej na Slovensku. Analýza používania webu bola vykonaná na vzorke 10 378 751 logovaných prístupoch, ktoré boli získané po príprave dát, ktorá zahŕňala čistenie dát, identifikáciu sedení a dopĺňanie ciest. Bolo pozorované spávanie sa používateľov na webovom portály bankovej inštitúcie počas obdobia viacerých rokov. Zisťovali sme vplyv aj ďalších faktorov, či je významné rozlišovať interné a externé prístupy (*internal*), či je možné rozlišovať obdobie rokov počas krízy a po kríze (*crisis*). V prípade umelej premennej *internal* sme identifikovali triviálnu mieru závislosti s premennou *category* (*Kontingenčný koeficient C* = 0,077; *Cramerovo V* = 0,077). Kontingenčné koeficienty

môžu nadobúdať hodnoty od 0 (reprezentuje žiadnu závislosť medzi premennými) po 1 (reprezentuje perfektnú závislosť medzi premennými).

V prípade umelej premennej *crisis* bola identifikovaná malá miera závislosti (*Chi-square* = 81 455,210; *df* = 5; *p* = 0,000; *Kontingenčný koeficient C* = 0,224; *Cramerovo V* = 0,230). Kontingenčný koeficient je štatisticky významný.

Na základe týchto výsledkov sme vytvorili model pre všetky prístupy (nerozlišovali sme externé a interné prístupy) a zahrnuli sme umelú premennú *crisis* do modelu ako ďalší prediktor. Metodika modelovania správania sa používateľov webu v závislosti od času bola popísaná v kapitole 2.2.2.

Napriek tomu, že sme vychádzali z výskumu (Munk et al., 2011a), tak nebolo pri týždňoch jednoznačné, či sa bude jednať o model polynómu druhého, tretieho alebo štvrtého stupňa. Prostredníctvom LR testu je možné porovnať odhady teoretických početností s empirickými početnosťami. Maximum logaritmickej funkcie vierohodnosti je vhodné na porovnanie modelov, pričom čím nadobúda menšiu hodnotu, tým je model vhodnejší. Nevýhodou LR testu je v prípade nie dostatočne veľkých očakávaných početností, kedy dochádza k porušeniu podmienky použitia LR testu. V takom prípade sa na evalváciu modelu používajú alternatívne techniky ako napr. vizualizácia rozdielov empirických a teoretických početností a identifikácia extrémnych hodnôt. V tabuľkách 26-28 sa nachádzajú výsledky LR testov pre všetky tri druhy modely (polynómu druhého, tretieho a štvrtého stupňa). Vo všetkých prípadoch je hodnota LR testu malá, preto je možné všetky modely považovať za vhodné. Hodnota podielu Pearson chí-kvadrátu je približne rovná jednej, čo tiež vypovedá o vhodnosti modelov.

Tabuľka 26 LR test pre model polynómu druhého stupňa

	sv	**Štat.**	**Štat./sv**
Test vierohodnostným pomerom (LR test)	7684410	4529670	0,589462
Pearsonov chí-kvadrát	7684410	7794534	1,014331
Maximum logaritmickej funkcie vierohodnosti		-2264835	

Tabuľka 27 LR test pre model polynómu tretieho stupňa

	sv	**Štat.**	**Štat./sv**
Test vierohodnostným pomerom (LR test)	7684405	4510212	0,586931
Pearsonov chí-kvadrát	7684405	7784416	1,013015
Maximum logaritmickej funkcie vierohodnosti		-2255106	

Tabuľka 28 LR test pre model polynómu štvrtého stupňa

	sv	**Štat.**	**Štat./sv**
Test vierohodnostným pomerom (LR test)	7684400	4503695	0,586083
Pearsonov chí-kvadrát	7684400	7829313	1,018858
Maximum logaritmickej funkcie vierohodnosti		-2251847	

Na základe výsledkov LR testu (Tabuľka 26-28) je najvyššia hodnota v prípade modelu polynómu štvrtého stupňa a rozdiely medzi modelmi s polynómom druhého a tretieho stupňa neboli až také jednoznačné. Ďalšou evalváciou modelu pomocou vizualizácie empirických a teoretických logitov bude možné zvoliť vhodný stupeň polynómu pre skúmaný model.

Na odhad parametrov pre individuálne údaje bola použitá *STATISTICA Generalized Linear/Nonlinear Models.* Významnosť parametrov bola testovaná pomocou Waldovho testu. Boli modelované pravdepodobnosti prístupu na kategórie obsahu portálu v závislosti od času- týždňov prístupu a obdobia krízy. Čas bol reprezentovaný premennou *week* a jej transformáciou v závislosti od stupňa polynómu (*week*^{2}, *week*^{3}, *week*^{4}) a umelou premennou *crisis*, ktorá reprezentovala roky obdobia krízy.

Na základe výsledkov testu všetkých efektov pre model polynómu tretieho stupňa (Tabuľka 29) parametre modelu sú štatisticky významné- vo vytvorenom modeli predstavujú roky krízy a po kríze štatisticky významný znak, ktorý je v modely reprezentovaný umelou premennou *crisis*. Podobne týždne počas roka, reprezentované premennými *week* a jej transformáciou v závislosti od stupňa polynómu, sa ukázali ako štatisticky významné znaky.

Tabuľka 29 Test všetkých efektov pre model polynómu tretieho stupňa

	sv.	Wald. št.	p
Abs. člen	5	52325,51	0,0000
week	5	10800,15	0,0000
week^2	5	18259,02	0,0000
week^3	5	18995,98	0,0000
crisis	5	68622,95	0,0000

Parametre modelu boli odhadnuté pre individuálne údaje v programe *STATISTICA Generalized Linear/Nonlinear Models*. V tabuľke (Tabuľka 30) môžeme vidieť odhad logitov pre model polynómu tretieho stupňa. Významné parametre sú v tabuľke podfarbené, pričom ich významnosť bola testovaná pomocou Waldovho testu. Z tabuľky vyplýva, že logity pre všetky kategórie (okrem kategórie *Pricing List*) sú významne závislé od týždňa prístupu ako aj od jeho transformácií. Na hodnoty týchto logitov významne vplýva obdobie rokov krízy.

Tabuľka 30 Odhad parametrov pre model polynómu tretieho stupňa

	Kategória	Odhad par.	Sm. chyba	Wald. št.	p
week	*Pricing List*	0,0233	0,0025	84,7536	0,0000
week^2	*Pricing List*	0,0001	0,0001	1,3018	0,2539
week^3	*Pricing List*	0,0000	0,0000	0,3505	0,5538
crisis	*Pricing List*	-1,1730	0,0087	18003,3166	0,0000
week	*Reputation*	0,0507	0,0030	291,4300	0,0000
week^2	*Reputation*	-0,0023	0,0001	267,4115	0,0000
week^3	*Reputation*	0,0000	0,0000	305,2583	0,0000
crisis	*Reputation*	-0,9237	0,0101	8362,0226	0,0000
week	*Business Conditions*	-0,0623	0,0028	494,4624	0,0000
week^2	*Business Conditions*	0,0056	0,0001	1814,5684	0,0000
week^3	*Business Conditions*	-0,0001	0,0000	1721,9480	0,0000
crisis	*Business Conditions*	-2,0717	0,0095	47714,1908	0,0000
week	*Pillar3 related*	0,0655	0,0027	574,4090	0,0000
week^2	*Pillar3 related*	-0,0024	0,0001	343,8587	0,0000
week^3	*Pillar3 related*	0,0000	0,0000	301,2951	0,0000
crisis	*Pillar3 related*	-0,8455	0,0093	8189,3534	0,0000
week	*Pillar3 disclosure requirements*	0,1703	0,0030	3182,4843	0,0000
week^2	*Pillar3 disclosure requirements*	-0,0074	0,0001	2640,8711	0,0000
week^3	*Pillar3 disclosure requirements*	0,0001	0,0000	2284,0930	0,0000
crisis	*Pillar3 disclosure requirements*	-0,9935	0,0100	9890,9264	0,0000

Logitový model poskytuje na výstupe odhad pravdepodobností, avšak absolútna veľkosť parametrov modelu ozrejmuje, od ktorých prediktorov najviac závisí skúmaná premenná.

Vysoká absolútna hodnota parametra hovorí o veľkej závislosti, kde kladná hodnota reprezentuje priamo úmernú závislosť a negatívna hodnota nepriamo úmernú závislosť.

Pomocou odhadnutých parametrov je možné vypočítať logity pre každú kategóriu j v čase i. Model polynómu tretieho stupňa pre prístupy počas krízy bol realizovaný podľa vzťahu:

$$\hat{\eta}_{ij} = \alpha_j + \beta_{1j} week_i + \beta_{2j} week_i^2 + \beta_{3j} week_i^3 + \gamma_j crisis_i, j = 1, 2, \dots, J-1, i = 0, 1, 2, \dots, 53. \quad (21)$$

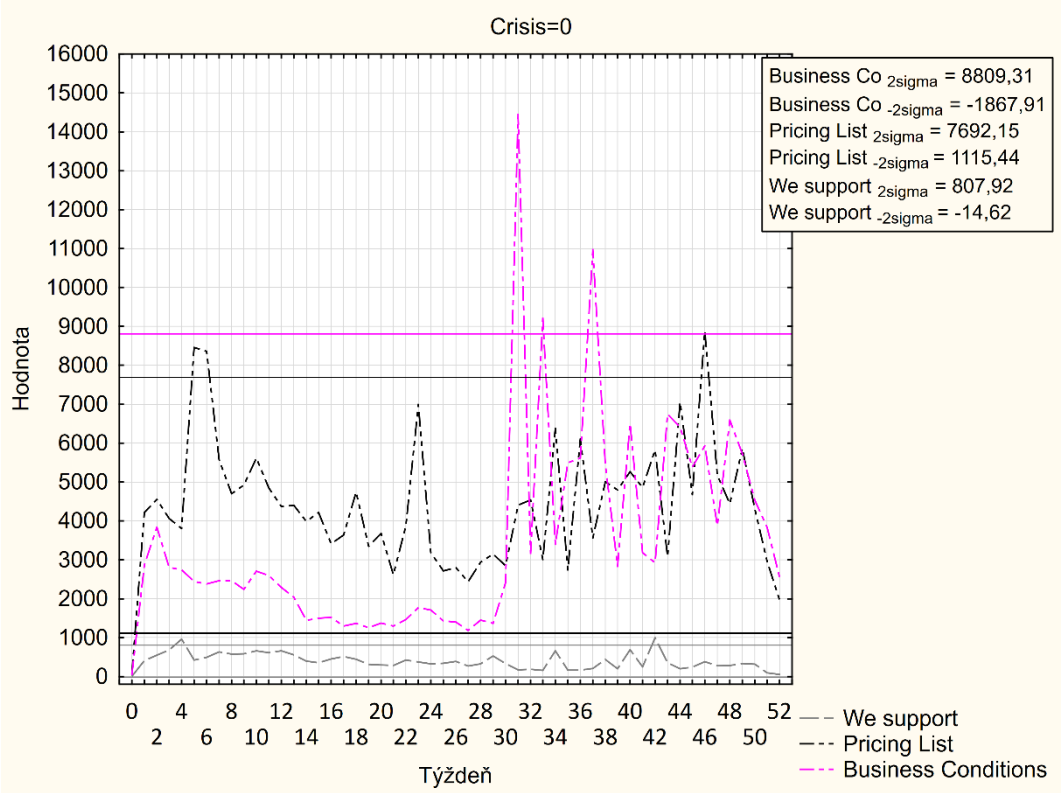
Analogicky boli odhadnuté parametre pre modely polynómu druhého a štvrtého stupňa.

Následne boli vypočítané logity, z ktorých bolo možné odhadnúť pravdepodobnosť pre referenčnú webovú kategóriu. Pri odhade sme vychádzali zo vzorca: $\hat{\pi}_{iJ} = \frac{1}{1+\sum_{j=1}^{J-1} e^{\hat{\eta}_{ij}}}$, kde $\hat{\eta}_{ij}$ sú odhady logitov pre webovú kategóriu j v čase i. Na základe odhadu pravdepodobnosti prístupu na webovú referenčnú kategóriu a odhadnutých logitov, bolo možné odhadnúť pravdepodobnosti prístupov aj pre ostatné webové kategórie: $\hat{\pi}_{ij} = e^{\hat{\eta}_{ij}} \hat{\pi}_{iJ}, j = 1, 2, \dots, J-1$, kde $\hat{\eta}_{ij}$ sú odhady logitov pre webovú kategóriu j v čase i a $\hat{\pi}_{iJ}$ je odhad pravdepodobnosti prístupu na referenčnú webovú kategóriu J v čase i.

Aby sme mohli vyhodnotiť, ktorý model je najvhodnejší, tak sme sa rozhodli porovnať výsledky evalvácie pre jednotlivé modely a až následne potom vizualizovať pravdepodobnosti prístupov na jednotlivé kategórie počas týždňov. Hodnotiť modely budeme na troch úrovniach: na úrovni početností, na úrovni pravdepodobností a na úrovni logitov. Pomocou kontingencie boli určené empirické početnosti prístupov y_{ij} pre webovú kategóriu j v čase i. Vychádzajúc z odhadnutých pravdepodobností prístupov návštevníkov na skúmané webové kategórie, je možné odhadnúť teoretické početnosti prístupov $\hat{y}_{ij} = \hat{\pi}_{ij} \sum_j y_{ij}$, kde $\hat{\pi}_{ij}$ sú odhady pravdepodobností prístupov a y_{ij} sú empirické početnosti prístupov na webovú kategóriu j v čase i. Pre hľadanie problémových častí v početnostiach vizualizujeme rozdiely empirických a teoretických početností $d_{ij} = y_{ij} - \hat{y}_{ij}$, pričom identifikujeme extrémne hodnoty na základe pravidla 2σ.

Po porovnaní rozdielov jednotlivých modelov nebolo možné určiť model, ktorý by bol najvhodnejší, keďže rozdiely početností vyšli skoro pre všetky modely podobne. Jediný značne väčší rozdiel identifikovaný v kategórií *Business Conditions* pre model

polynómu druhého stupňa (Obrázok 24), bol v 31. týždni (polynóm druhého stupňa: 14 445,715; polynóm tretieho stupňa: 14 020,797; polynóm štvrtého stupňa: 13 548,354). Graf (Obrázok 24) vizualizuje rozdiely empirických a teoretických početností prístupov návštevníkov v období rokov po finančnej kríze. Po aplikácií pravidla 2σ boli identifikované extrémne prípady. V tomto prípade sa pre kategóriu *Business Conditions* v 31. týždni jednalo o podhodnotenú predpoveď. Vo všetkých troch modeloch bolo identifikovaných rovnaký počet extrémnych prípadov, ktorých podiel tvoril približne 5,55 % všetkých prípadov.



Obrázok 24 Rozdiely početností modelu polynómu druhého stupňa

Ďalším krokom bolo porovnanie na úrovni pravdepodobností. Vychádzajúc z pozorovaných početností môžeme vypočítať empirické relatívne početnosti prístupov $p_{ij} = \frac{y_{ij}}{\sum_j y_{ij}}$. Ďalej porovnáme rozdelenia pravdepodobností empirických relatívnych početností prístupov a odhadnutých pravdepodobností výberu webovej časti j v čase i: $r_{ij} = p_{ij} - \pi_{ij}$, H0: $F(-r) = 1 - F(r)$. Na testovanie nulovej hypotézy, rozdelenie rozdielov párov je symetrické okolo nuly, bol použitý Wilcoxonov párový test. Zvýraznené hodnoty v tabuľke (Tabuľka 31) nám hovoria o problémových webových kategóriách.

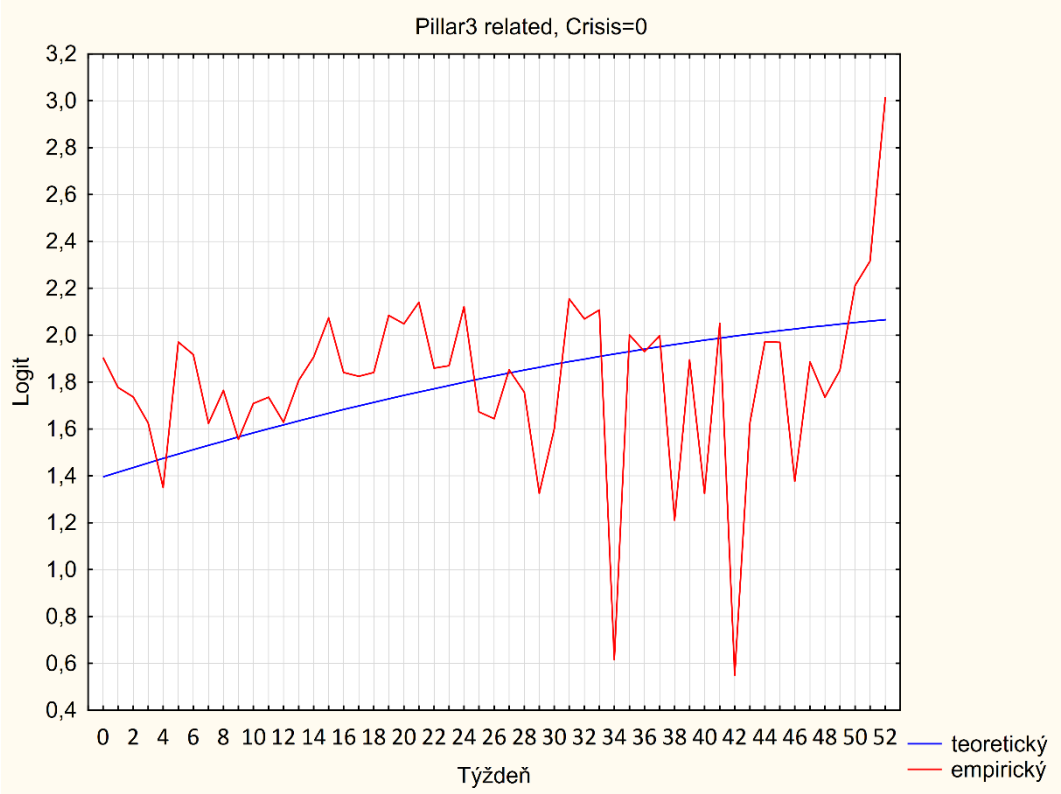
Tabuľka 31 Rozdelenie pravdepodobností pre model polynómu tretieho stupňa

Kategória	Crisis	Počet	T	Z	p-hodn.
We support	1	53	307,0000	3,6164	0,0002
Pricing List	1	53	570,0000	1,2881	0,1977
Reputation	1	53	678,0000	0,3320	0,7399
Business Conditions	1	53	621,0000	0,8366	0,4028
Pillar3 related	1	53	659,0000	0,5002	0,6169
Pillar3 disclosure requirements	1	53	686,0000	0,2612	0,7940
We support	0	53	691,0000	0,2169	0,8283
Pricing List	0	53	696,0000	0,1726	0,8629
Reputation	0	53	597,0000	1,0490	0,2942
Business Conditions	0	53	456,0000	2,2973	0,0216
Pillar3 related	0	53	564,0000	1,3412	0,1799
Pillar3 disclosure requirements	0	53	698,0000	0,1549	0,8769

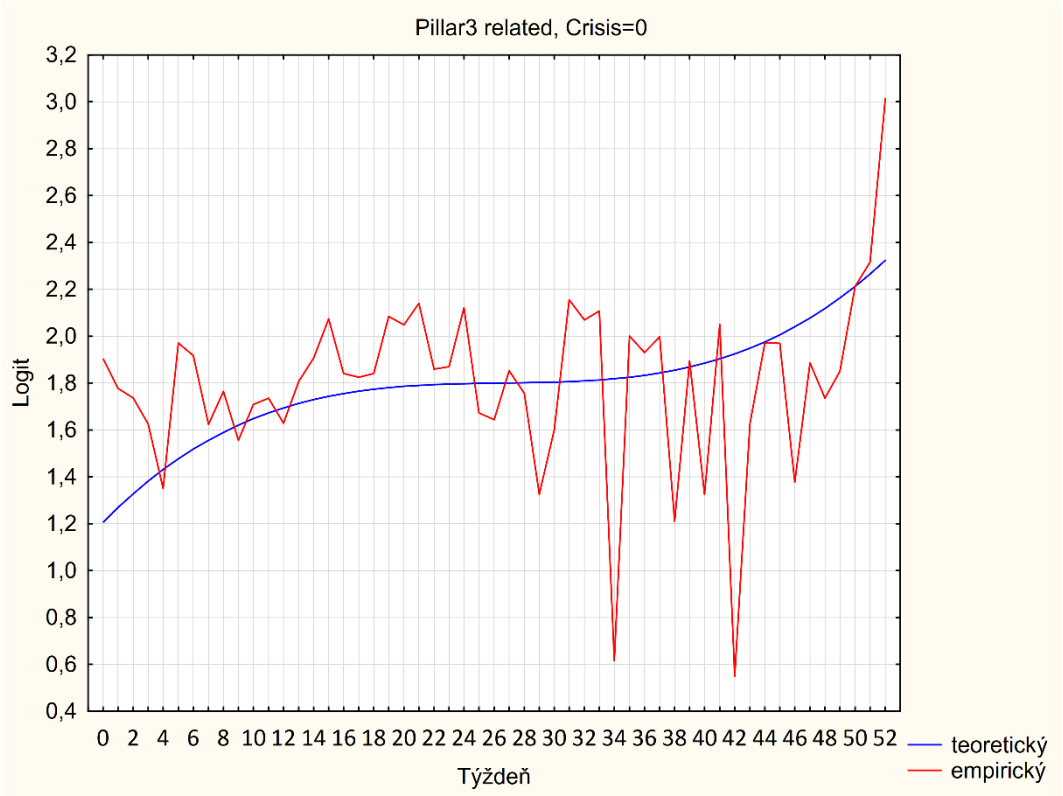
Okrem dvoch problémových kategórií, nebolo ani v tomto prípade porovnanie pravdepodobností jednoznačné, čo sa týka vhodnosti výberu stupňa polynómu. Počas rokov krízy bol objavený problém v kategórii *We support* (polynóm druhého stupňa: p = 0,000320; polynóm tretieho stupňa: p = 0,000299; polynóm štvrtého stupňa: p = 0,000342) a v období po kríze sa jednalo o kategóriu *Business Conditions*, čo odhalilo už aj porovnanie rozdielov početností.

Poslednou možnosťou, ktorá môže pomôcť pri výbere vhodného stupňa polynómu modelu je evalvácia teoretických a empirických logitov. V tomto prípade sledujeme to, či nami odhadnuté teoretické logity fitujú (modelujú) empirické logity vypočítané z empirických relatívnych početností $h_{ij} = \ln\left(\frac{p_{ij}}{p_{iJ}}\right), j = 1, 2, \dots, J-1,$ kde p_{ij} je empirická relatívna početnosť webovej kategórie j v čase i a p_{iJ} je empirická relatívna početnosť referenčnej webovej kategórie J v čase i. Vizualizáciou empirických a teoretických logitov pre jednotlivé webové kategórie (okrem referenčnej) môžeme pozorovať ako teoretické logity modelujú empirické logity. Na základe vizualizácie bolo vidieť, že všetky teoretické logity fitujú empirické logity. Avšak v prípade modelov polynómu druhého stupňa (Obrázok 25) môžeme pozorovať, že teoretické logity predstavujú skôr lineárnu funkciu v prípade niektorých kategórií, než kvadratickú. V prípade teoretických logitov modelov polynómu tretieho stupňa (Obrázok 26) a štvrtého stupňa (Obrázok 27) je vidieť, že lepšie modelujú empirické logity. Avšak model polynómu štvrtého stupňa (Obrázok 27) zachytáva priebeh empirických logitov až príliš podrobne. Model polynómu tretieho stupňa (Obrázok 26) v porovnaní

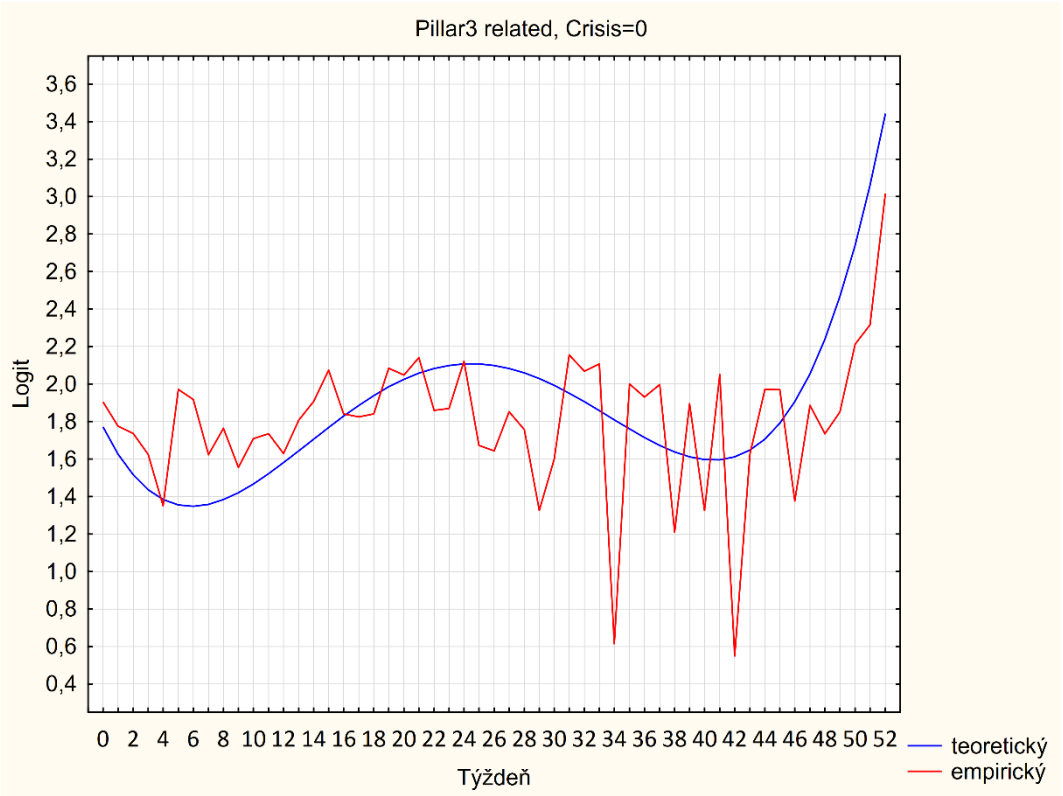
s ním poskytuje podobné fitovanie avšak neprispôsobuje sa až natoľko dátam na úkor trendu. Ako ukážku vhodnosti výberu sme si zvolili kategóriu *Pillar3 related*, ktorú prezentujeme na grafoch (Obrázok 25-27).



Obrázok 25 Vizualizácia logitov pre model polynómu druhého stupňa pre kategóriu Pillar3 related

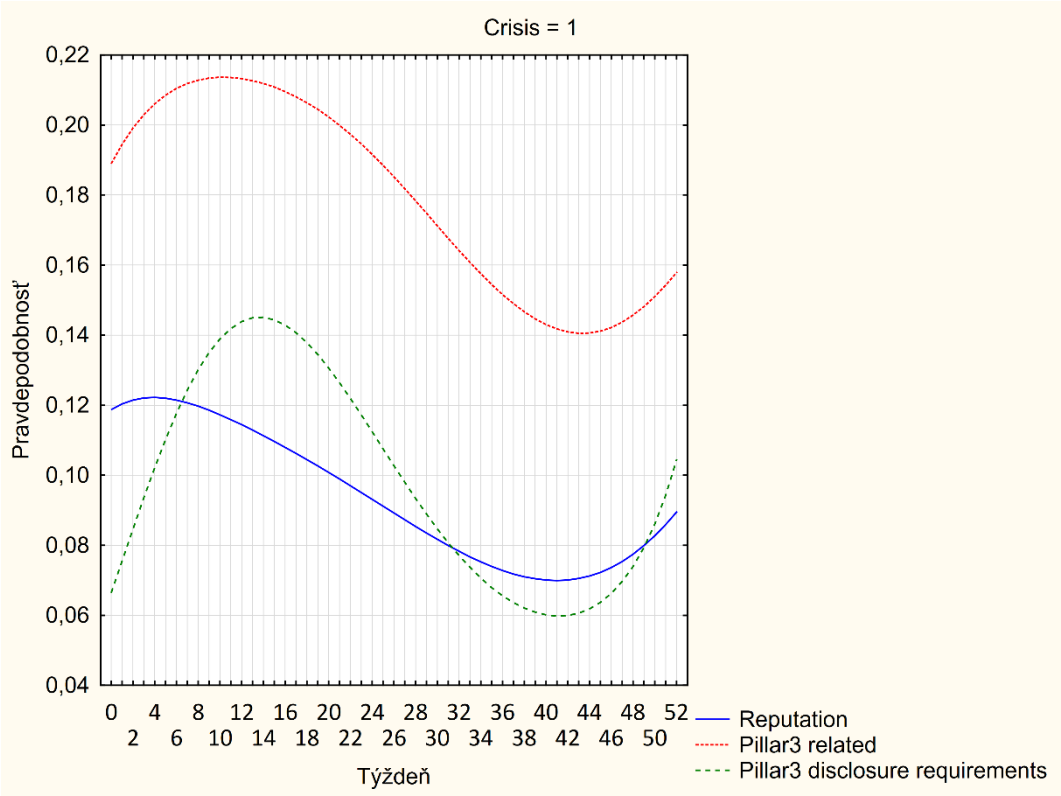


Obrázok 26 Vizualizácia logitov pre model polynómu tretieho stupňa pre kategóriu Pillar3 related

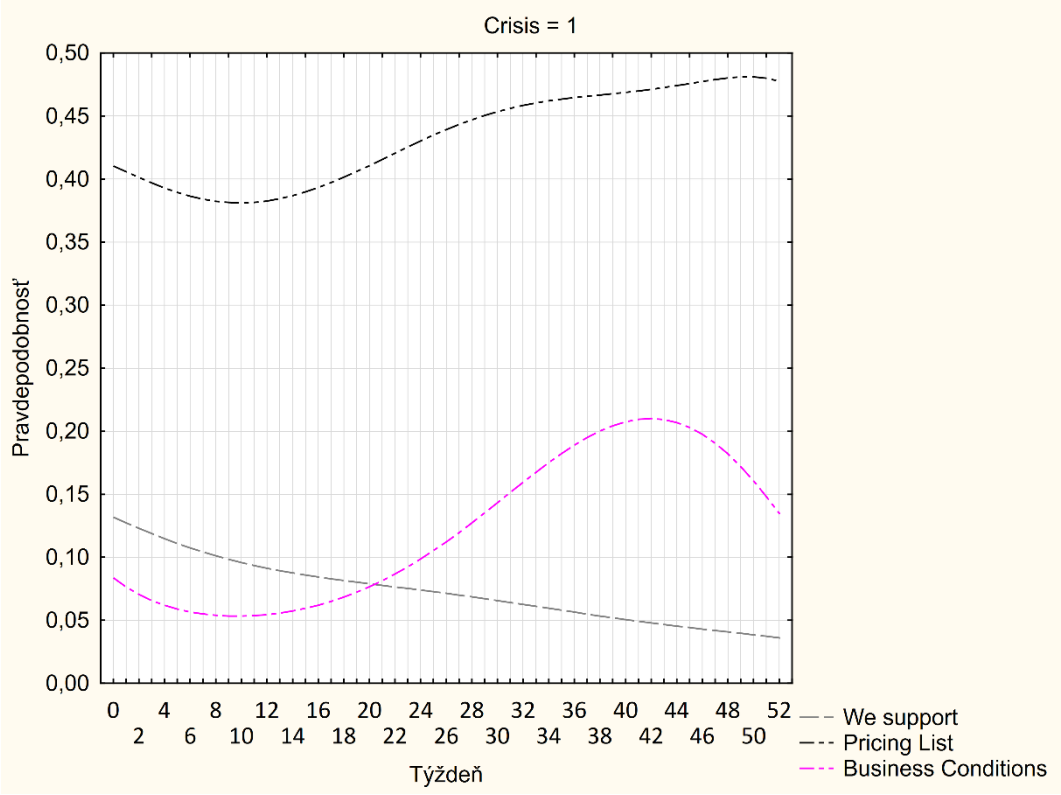


Obrázok 27 Vizualizácia logitov pre model polynómu štvrtého stupňa pre kategóriu Pillar3 related

Grafy (Obrázok 28 a 29) zobrazujú pravdepodobnosti návštevy jednotlivých webových kategórií návštevníkmi počas obdobia rokov finančnej krízy. Najvyšší prístup počas rokov krízy bol odhadnutý pre webovú kategóriu *Pricing List*, pričom najväčšia pravdepodobnosť prístupu bola počas týždňov na konci rokov (50. týždeň dosahuje hodnotu 0,481) a naopak najnižšie odhadnuté hodnoty pre túto kategóriu boli počas týždňov na začiatku roka (10. týždeň dosahuje hodnotu 0,381). Druhou najnavštevovanejšou webovou kategóriou bola skoro polovicu roka kategória *Pillar3 related*, pričom najvyššiu návštevnosť dosahovala počas prvého kvartála roka (10. týždeň dosahuje hodnotu 0,214). V druhej polovici roka začala pravdepodobnosť prístupu na túto webovú kategóriu klesať, pričom opäť výrazne stúpla v posledných štyroch týždňoch roka. Od 33. do 50. týždňa rokov finančnej krízy patrilo *Business Conditions* na základe odhadu pravdepodobností k druhej najnavštevovanejšej webovej kategórií. Najvyššia hodnota 0,210 bola nameraná v 42. týždni. Naopak v prvej polovici roka patrila webová kategória *Business Conditions* podľa odhadu k najmenej navštevovaným webovým kategóriám, pričom najnižšia hodnota 0,053 bola nameraná v 10. týždni. Pravdepodobnosti prístupu na ostatné webové kategórie sú približne podobné v okolí hodnoty 0,10. Zaujímavý je ešte mierny nárast pravdepodobnosti prístupu na webovú časť *Pillar3 disclosure requirements* s dosiahnutým maximom 0,145 v 14. týždni, avšak následne pravdepodobnosť prístupu klesá pod 0,100.

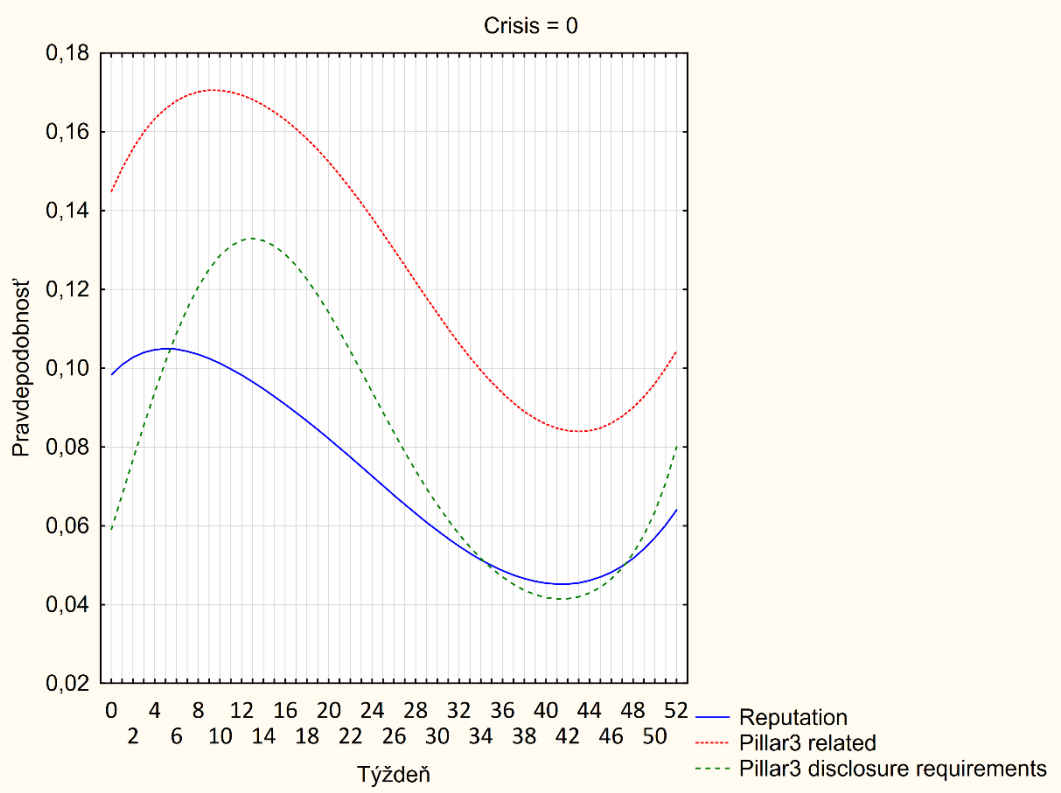


Obrázok 28 Vizualizácia pravdepodobností webových kategórií súvisiacich s trhovou disciplínou v období rokov finančnej krízy

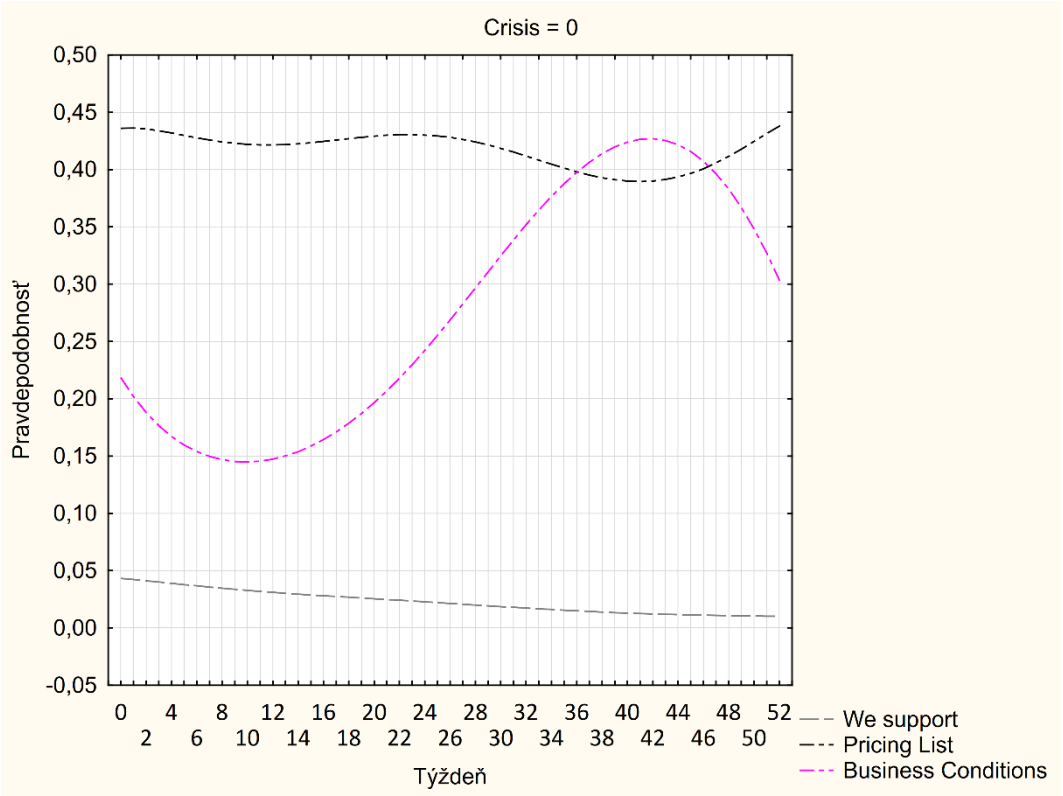


Obrázok 29 Vizualizácia pravdepodobností ostatných webových kategórií v období rokov finančnej krízy

Grafy (Obrázok 30 a 31) zobrazujú pravdepodobnosti návštevy jednotlivých webových kategórií návštevníkmi počas obdobia rokov po finančnej kríze. Najvyšší prístup počas rokov po finančnej kríze bol odhadnutý pre webovú kategóriu *Pricing List*. Okrem obdobia od 36. po 46. týždeň, kedy bol odhadnutý najväčší prístup na webovú kategóriu *Business Conditions*. Najväčší rozdiel v porovnaní s obdobím finančnej krízy je vidieť práve pre kategóriu *Business Conditions*, ktorej pravdepodobnosť prístupu sa skoro zdvojnásobila. Naopak pravdepodobnosť prístupu na webovú kategóriu *We support* klesla po období krízy. V prípade webovej kategórie *Reputation*, *Pillar3 related* a *Pillar3 disclosure requirements* prišlo k poklesu pravdepodobnosti prístupu, avšak správanie počas obdobia krízy a v období po kríze je podobné, čiže väčší záujem o tieto kategórie je na začiatku roka a v priebehu roka klesá, pričom koncom roka opäť začína stúpať.



Obrázok 30 Vizualizácia pravdepodobností webových kategórií súvisiacich s trhovou disciplínou v období rokov po finančnej kríze



Obrázok 31 Vizualizácia pravdepodobností ostatných webových kategórií v období rokov po finančnej kríze

Z podrobnej týždennej analýzy (počas rokov 2009 – 2015) správania sa používateľov na webe zverejňovaných finančných a rizikových informácií komerčnej banky, sme zistili, že výsledky analýzy korešpondujú s výsledkami predchádzajúcej štvrťročnej analýzy (Kapitola 3.4). Najväčší záujem o povinné a doplňujúce Pilier3 informácie mali stakeholderi v prvom štvrťroku, konkrétne môžeme špecifikovať obdobie okolo 10. týždňa, kedy o tieto webové kategórie dosiahnutý najväčší záujem. Na základe toho je možné vyvodiť, že frekvencia požadovaného štvrťročného zverejňovania výsledkov NBS nie je pre účely trhovej disciplíny potrebná. Postačujúce by bolo ročné zverejňovanie týchto informácií, ideálne v týždňoch na začiatku roka.

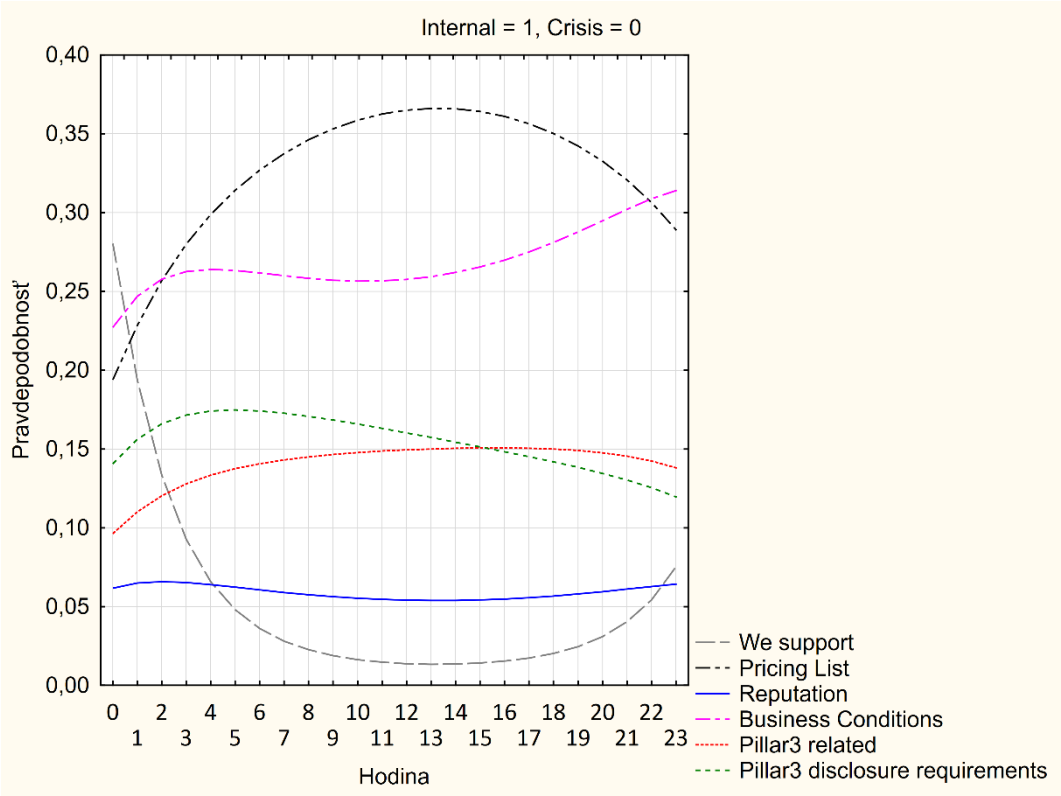
Na základe výsledkov odhadu pravdepodobnosti prístupov na webové časti bankového portálu môžeme identifikovať obdobie počas jednotlivých týždňov roka, v ktorých je nutné aktualizovať informácie o treťom pilieri. Avšak aby bolo možné webové časti aktualizovať, je nutné poznať správanie sa návštevníkov webu nielen počas dlhého časového intervalu, ale aj počas dňa. Ak chceme modelovať správanie sa používateľov webu v závislosti od času (hodín), tak môžeme zopakovať postup ako pri modelovaní v závislosti od týždňov. Časová premenná hodina (*hour*) nadobúda hodnoty 0-23, pričom

na základe výsledkov predchádzajúcich výskumov (Munk et al., 2011a, 2011b) sa ukázalo, že logity sú kvadratickou funkciou času, t.j. bola vytvorená transformácia premennej *hour* na polynóm druhého stupňa (*hour*^{2}). Zisťovali sme aj vplyv ďalších faktorov, či je významné rozlišovať interné a externé prístupy (*internal*), či je možné rozlišovať obdobie rokov počas krízy (2009-2010) a po kríze (2011-2014) (*crisis*). Interné a externé prístupy sme rozlišovali na základe IP adresy prístupu na webovú časť, na základe ktorej sme vedeli identifikovať prístupy zvnútra siete bankovej inštitúcie. V prípade umelej premennej *crisis* bola v prípade prístupov zvnútra siete (*internal = 1*) identifikovaná malá miera závislosti (*Chi-square* = 2551,578; *df* = 5; *p* = 0,000; *Kontingenčný koeficient C = 0,129*; *Cramerovo V* = 0,130). V prípade prístupov zvonka siete (*internal = 0*) bola identifikovaná malá miera závislosti (*Chi-square* = 94 511,58; *df* = 5; *p* = 0,000; *Kontingenčný koeficient C = 0,253*; *Cramerovo V* = 0,261). Na základe týchto výsledkov sme vytvorili model pre interné prístupy a externé prístupy (umelá premenná *internal*) a zároveň sme zahrnuli do oboch modelov obdobie rokov krízy a rokov po kríze (umelá premenná *crisis*) ako ďalší prediktor. Postup bol analogický s postupom modelovania v závislosti od týždňov, pričom sme sa zamerali hlavne na obdobie rokov po finančnej kríze. Model pre prístupy počas krízy bol realizovaný pomocou vzorca:

$$\hat{\eta}_{ij} = \alpha_j + \beta_{1j} hour_i + \beta_{2j} hour_i^2 + \gamma_j crisis_i, j = 1, 2, \dots, J-1, i = 0, 1, 2, \dots, 23. \qquad (22)$$

Graf (Obrázok 32) zobrazuje pravdepodobnosti návštevnosti jednotlivých webových kategórií internými návštevníkmi v období rokov po odznení finančnej krízy počas hodín. Najvyšší prístup bol odhadnutý pre kategóriu *Pricing List*, pričom najväčší záujem o túto kategóriu bol hlavne počas obedných hodín (13. hodina dosahuje hodnotu 0,367). Najnižšie hodnoty prístupu na kategóriu *Pricing List* boli zaznamenané v nočných hodinách (0. hodina dosahuje hodnotu 0,194). Druhou najnavštevovanejšou kategóriou je *Business Conditions,* ktorá má v priebehu prvej polovici dňa približne ustálené percento prístupu (dosahuje hodnotu približne 0,250), avšak môžeme povedať, že rastie s priebehom dňa a k významnejšiemu nárastu dochádza v nočných hodinách (23. hodina dosahuje hodnotu 0,314). V prípade webovej kategórie *Pillar3 disclosure requirements* môžeme vidieť najväčší záujem hlavne v skorých ranných hodinách (5. hodina dosahuje hodnotu 0,175), avšak potom návštevnosť tejto kategórie klesá. Nízku návštevnosť vykazuje webová kategória *Reputation*, pričom by bolo možné tvrdiť, že bola počas celého dňa skoro konštantná (hodnota približne 0,05). Záujem o webovú kategóriu

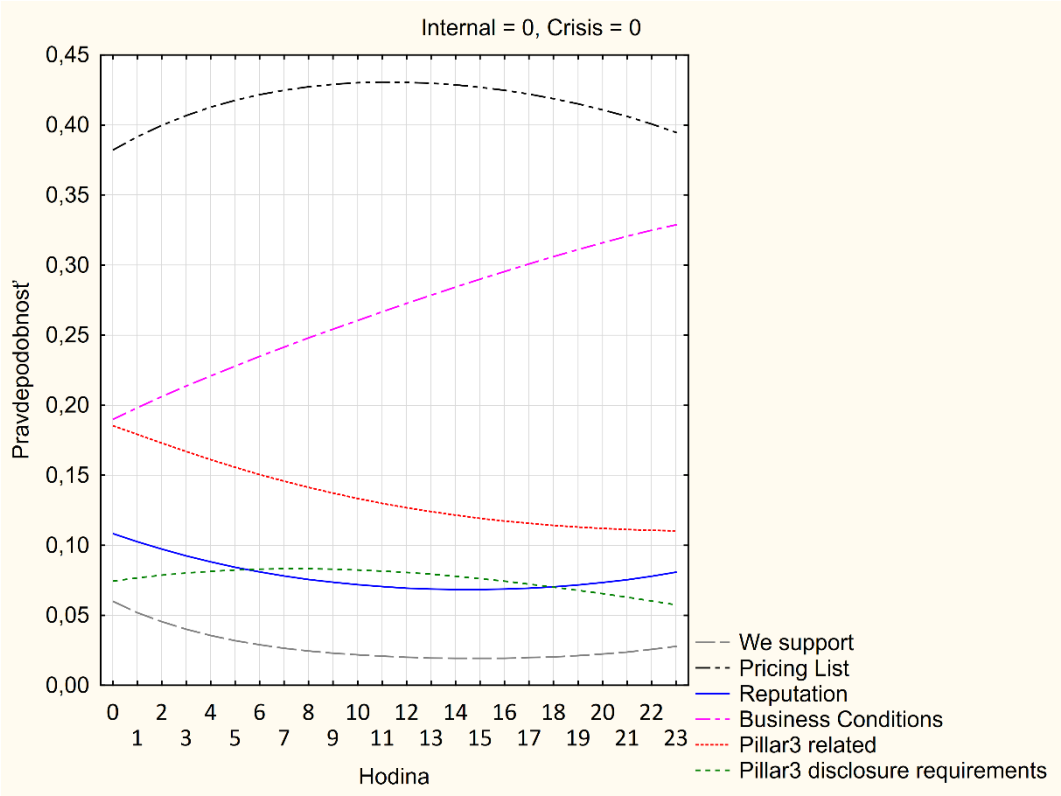
Pillar3 related je prevažne počas dňa konštantný, len v ranných hodinách stúpa zo svojho minima. Zaujímavé správanie môžeme pozorovať pri webovej kategórií *We support*, kde je veľmi výrazný prístup v nočných hodinách a následne dochádza k minimálnemu prístupu na túto kategóriu v priebehu dňa.



Obrázok 32 Vizualizácia pravdepodobností interných prístupov na webové kategórie v období rokov po finančnej kríze v závislosti od času – hodín dňa

Graf (Obrázok 33) zobrazuje externé prístupy na jednotlivé kategórie webového portálu v období rokov po finančnej kríze. Zaujímavé kategórie pre návštevníkov zvonka siete bankovej inštitúcie sú podobné záujmom interných návštevníkov webového portálu. Najnavštevovanejšou je opäť kategória *Pricing List* s najväčším záujmom v obedných hodinách (11. hodina dosahovala hodnotu 0,431) a naopak najmenším v nočných hodinách. Taktiež druhou najnavštevovanejšou kategóriou je *Business Conditions*, kde je najvyšší záujem o kategóriu na konci dňa (23. hodina dosahovala hodnotu 0,329) a najnižšia na začiatku dňa (0. hodina dosahovala hodnotu 0,190). Webové kategórie *Pillar3 related, Pillar3 disclosure requirements* a *Reputation* preukazujú podobné hodnoty a záujem o tieto kategórie má väčšinou počas dňa klesajúci charakter. V prípade webovej kategórie *We support* môžeme povedať, že o ňu návštevníci pristupujúci zvonku

siete nemajú záujem, pretože skoro celý deň je pravdepodobnosť prístupu menšia ako 5 %.



Obrázok 33 Vizualizácia pravdepodobností externých prístupov na webové kategórie v období rokov po finančnej kríze v závislosti od času - hodín dňa

Získané znalosti boli použité ako podklady pre aktualizáciu webového portálu. Bankový portál je povinný zverejňovať a aktualizovať informácie ohľadom Piliera 3 počas roka a preto bolo nutné skúmať aj správanie sa používateľov počas hodín. Pre potreby plánovania zverejňovania informácií by bolo na základe dosiahnutých výsledkov možné odporučiť nočné hodiny (medzi 1-2 hodinou), kedy je prístup na jednotlivé webové kategórie nižší.

4 VYHODNOTENIE VÝSLEDKOV VÝSKUMU

Teoretickým prínosom práce pre oblasť aplikovanej informatiky sú vypracované metodiky predikcie. Vytvorili sme metodiku na zhodnotenie frekventovaných transakcií/sekvencií v čase a bola vytvorená metodika predikcie pravdepodobnosti prístupov na webové časti portálu bankovej inštitúcie v závislosti od týždňov roka, ako aj na základe ďalších časových premenných (hodín dňa). Navrhnuté metodiky môžu byť použité aj pre iný typ webového portálu za predpokladu vhodnej prípravy dát.

V dizertačnej práci sme sa venovali rôznym metódam prípravy dát, pričom sme sa zamerali prevažne na porovnanie a optimalizáciu metód identifikácie sedení a vplyv rekonštrukcie aktivít používateľov pri rôznych metódach identifikácie sedení. Prioritne sme sa zaoberali metódou Reference Length a jej vplyvom na extrahovanie užitočných znalostí z logovacích súborov. Skúmali sme prípravu dát v prípade webového portálu s anonymným prístupom (univerzitný portál) a webového portálu s povinnou autentifikáciou (virtuálne vzdelávacie prostredie Moodle). Metóda identifikácie sedení Reference Length je typická tým, že pre jej použitie potrebujeme vypočítať hraničný čas, ktorý určí koniec sedenia a začiatok nového sedenia. Pre potreby výpočtu hraničného času je nutné poznať podiel navigačných stránok skúmaného webového portálu, ktorý sa zvykol odhadovať subjektívne správcom webového portálu. Našou optimalizáciou metódy Reference Length bol odhad podielu navigačných stránok z mapy webu, čo sa aj preukázalo v našich experimentoch. V nich sme porovnali rôzne metódy identifikácie sedení a aj vplyv dopĺňania ciest na extrahované sekvenčné pravidlá ako po stránke kvantity, tak aj kvality. Na základe dosiahnutých výsledkov môžeme povedať, že identifikácia sedení pomocou metódy Reference Length s odhadom podielu navigačných stránok z mapy webu bez dopĺňania ciest, prispieva k objaveniu kvalitných znalostí. Na druhej strane jednou z nevýhod mapy webu môže byť jej nedostupnosť, prípadne aktuálnosť v súvislosti so skúmaným logovacím súborom. Preto sme v ďalšej fáze optimalizácie algoritmov identifikácie sedení vytvorili alternatívnu metódu výpočtu podielu navigačných stránok priamo z logovacieho súboru pomocou entropie. Takto identifikované sedenia sme následne porovnali s už vypočítanými výsledkami pre oba webové portály (univerzitný portál a virtuálne vzdelávacie prostredie). Výsledky odhadu podielu navigačných stránok pomocou entropie korešpondovali s výsledkami odhadu podielu navigačných stránok z mapy webu. Prínos nového spôsobu výpočtu podielu

navigačných stránok pomocou entropie je v prípade nedostupnosti mapy webu, prípadne aj ako spôsob overenia správnosti odhadu z mapy webu a jej aktuálnosti s ohľadom na skúmaný logovací súbor. Rovnako bol prínosom v prípade nového odhadu podielu navigačných stránok aj prezentovaný algoritmus použitia identifikácie sedení pomocou metódy Reference Length na základe odhadu entropie (Obrázok 16). Týmto boli splnené čiastočné ciele týkajúce sa prípravy dát a optimalizácie algoritmov identifikácie sedení a dopĺňania ciest.

V dizertačnej práci sme predstavili metodiku na zhodnotenie frekventovaných transakcií/sekvencií v čase, ktorú sme aplikovali na skúmanie logovacieho súboru webového portálu bankovej inštitúcie za obdobie rokov počas finančnej krízy a obdobie rokov po finančnej kríze. Boli extrahované frekventované položkové množiny na základe kvartálov a boli vyhodnocované na základe kvantity. Výskum prispieva k preklenutiu medzery v oblasti dostatočného zverejňovania informácií v rámci Pilieru 3, čo taktiež prispieva k zvyšovaniu záujmu príslušných stakeholderov o prispievanie k trhovej disciplíne a je relevantné s ich záujmami v rámci Pilieru 3. Z detailnej štvrťročnej analýzy využívania na webe zverejňovaných finančných a rizikových informácií komerčnej banky sme zistili, že ani obsah ani frekvencia prezentovania informácií nie sú pre nich relevantné. Záujem o tento druh informácií je minimálny, pričom sa o povinné a doplňujúce informácie Pilier 3 zaujímali stakeholderi hlavne v prvom štvrťroku (kedy sú zverejňované celoročné hospodárske výsledky banky). Stakeholderi mali veľmi nízky záujem o informácie týkajúce sa rizika ako povinne bankami zverejňovanej kategórie Pilieru 3. Z toho vyplýva veľmi nízky záujem o riadenie rizika ako jednej z kľúčových oblastí, na ktoré sa ostatné regulácie sústreďovali. Dôvodom môže byť medzi inými aj zložitosť samotnej problematiky manažmentu rizík, ktorá presahuje vedomosti a schopnosti kľúčových stakeholderov to pochopiť. Naše výsledky naznačujú, že ďalšie zmeny vo zverejňovaní informácií komerčnými bankami sú nevyhnutné, ak majú byť mechanizmy trhovej disciplíny efektívne a používané podľa očakávaní a požiadaviek regulátora. Dosiahnuté výsledky a závery potvrdzujú nariadenia zodpovedných inštitúcií v oblasti Pilieru 3.

V poslednom kroku sme skúmali správanie sa používateľov webového portálu bankovej inštitúcie v závislosti od času, konkrétne od týždňov roka a od hodín dňa. Bola vytvorená metodika predikcie pravdepodobnosti prístupov na webové časti v závislosti od času (týždne roka). Na základe predstavenej metodiky sme mohli odhadnúť pravdepodobnosti

prístupov na jednotlivé webové kategórie. Skúmali sme správanie sa návštevníkov webového portálu počas týždňov skúmaných rokov, kedy prebiehala finančná kríza a porovnali sme ho s obdobím rokov po odznení hlavných dopadov finančnej krízy. Tieto znalosti nám pomohli stanoviť odporúčania správcom webového portálu bankovej inštitúcie ohľadom vplyvu zverejňovania a aktualizácií informácií ohľadom Pilieru 3. Výsledky týždennej analýzy korešpondovali s výsledkami štvrťročnej analýzy. Najväčší záujem o povinné a doplňujúce Pilier3 informácie mali stakeholderi na začiatku roka, konkrétne môžeme špecifikovať obdobie okolo 10. týždňa, kedy dosahovali tieto webové kategórie najväčší záujem. Na základe toho je možné stanoviť obdobie zverejňovania týchto informácií na týždne na začiatku roka a znížiť tak požadovanú štvrťročnú frekvenciu, ktorá nie je pre účely trhovej disciplíny potrebná.

ZÁVER

Predmetom skúmania dizertačnej práce bolo modelovanie správania sa používateľov webu v závislosti od času. V prvej fáze sme sa zamerali na prípravu dát v oblasti WUM. Skúmali sme vplyv zvolenej metódy identifikácie sedení na kvantitu a kvalitu extrahovaných znalostí. Vytvorili sme metodiku predspracovania dát (Munk, Benko, Gangur, Turcani, 2015), na základe ktorej sme realizovali sériu experimentov na vyhodnotenie spoľahlivosti jednotlivých metód identifikácie sedení. Experimenty sme realizovali ako nad webovým portálom s anonymným prístupom (Munk, Benko, Gangur, Turcani, 2015; Munk a Benko, 2016), tak nad webovým portálom s povinnou autentifikáciou (Benko, 2015; Benko, Reichel, Kuna, Munk, 2015; Reichel, Kuna, Benko, Munk, 2015; Munk, Drlik, Benko, Reichel, 2017a). Výsledky preukázali vhodnosť použitia metódy Reference Length a jej vplyv na extrahovanie menšieho počtu triviálnych a nevysvetliteľných pravidiel a naopak na väčší počet užitočných pravidiel. Nevýhodou zvolenej metódy je podmienka exponenciálneho rozdelenia premennej *RLength*. Rôzne spôsoby odhadu podielu navigačných stránok majú vplyv až po dopĺňaní ciest. Na druhej strane dopĺňanie ciest spôsobuje nárast triviálnych a nevysvetliteľných pravidiel. Výpočet podielu navigačných stránok z mapy webu sa ukázal presnejší než subjektívny odhad podielu navigačných stránok, ale nevýhodou mapy webu je prakticky neustála zmena webového portálu a jej prípadná neaktuálnosť. Problém s mapou webu sme sa pokúsili odstrániť vytvorením nového postupu odhadu podielu navigačných stránok pomocou entropie vychádzajúc priamo z logovacieho súboru (Munk a Benko, 2017, 2018). Výpočet nášho navrhovaného postupu ukázal, že dosahuje podobné výsledky, ako výpočet z mapy webu a môže predstavovať vhodnú alternatívu v prípade absencie mapy webu. Tento nový prístup bol overený ako na webovom portály s anonymným prístupom, tak na webovom portály s povinnou autentifikáciou.

V ďalšom kroku sme sa zamerali na modelovanie správania sa používateľov webového portálu. Ako zdroj dát nám poslúžili logovacie súbory bankovej inštitúcie za obdobie viacerých rokov (2009-2015). Vytvorili sme metodiku na zhodnotenie frekventovaných transakcií/sekvencií v čase, konkrétne na báze kvartálov roka. Výskum (Munk, Pilkova, Benko, Blažeková, 2017b) prispieva k preklenutiu medzery v oblasti dostatočného zverejňovania informácií v rámci Pilieru 3. Boli extrahované frekventované položkové množiny na základe kvartálov a boli vyhodnocované na základe kvantity. Zamerali sme sa na analýzu údajov z webového portálu zameraného na zverejňovanie informácií

komerčných bánk k Pilieru 3 a skúmaním správania sa stakeholderov vo vzťahu k vážnym trhovým turbulenciám. Výsledky analýzy preukázali nízky záujem stakeholderov o povinne zverejňované informácie Pilieru 3 a ich záujem bol sústredený hlavne na prvý kvartál roka.

Pre potreby analýzy správania sa používateľov webového portálu bola vytvorená metodika predikcie pravdepodobnosti prístupov na webové časti v závislosti od času (týždňov roka) (Munk, Pilkova, Benko, Blažeková, 2018). Zdrojom dát boli opäť logovacie súbory webového portálu bankovej inštitúcie za obdobie viacerých rokov (2009-2015). Na rozdiel od predchádzajúcej analýzy, boli webové časti zlúčené do kategórií a analyzované. Na základe predstavenej metodiky sme mohli odhadnúť pravdepodobnosti prístupov na jednotlivé webové časti. Výsledky týždennej analýzy korešpondovali s výsledkami dosiahnutými predchádzajúcou analýzou, kde sme skúmali frekventované položkové množiny na základe kvartálov. Z pohľadu aplikačnej domény sú získané výsledky využiteľné pre tvorcov regulácií, a to či už na úrovni Európskej Bankovej Autority, Európskej centrálnej banky, Bazilejského výboru a pod.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

ABDULLAH, Zailani, Tutut HERAWAN, Haruna CHIROMA a Mustafa Mat DERIS, 2014. A Sequential Data Preprocessing Tool for Data Mining. In: [online]. B.m.: Springer International Publishing, s. 734–746 [vid. 15. listopad 2016]. Získáno z: doi:10.1007/978-3-319-09150-1_54

ANDĚL, Jiří, 2007. *Statistické metody*. 4. vydanie. Praha: MATFYZPRESS. ISBN 80-7378-003-8.

ANITHA, A, 2010. A New Web Usage Mining Approach for Next Page Access Prediction. *International Journal of Computer Applications*. roč. 8, s. 7–9.

ARBELAITZ, Olatz, Ibai GURRUTXAGA, Aizea LOJO, Javier MUGUERZA, Jesús Maria PÉREZ a Iñigo PERONA, 2013. Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it. *Expert Systems with Applications* [online]. 12., roč. 40, č. 18, s. 7478–7491 [vid. 6. březen 2017]. ISSN 09574174. Získáno z: doi:10.1016/j.eswa.2013.07.040

ARCE, Tomás, Pablo E. ROMÁN, Juan VELÁSQUEZ a Víctor PARADA, 2014. Identifying web sessions with simulated annealing. *Expert Systems with Applications* [online]. 3., roč. 41, č. 4, s. 1593–1600 [vid. 21. říjen 2015]. ISSN 09574174. Získáno z: doi:10.1016/j.eswa.2013.08.056

AYE, Theint Theint, 2011. Web log cleaning for mining of web usage patterns. *2011 3rd International Conference on Computer Research and Development* [online]. roč. 2, s. 490–494. Získáno z: doi:10.1109/ICCRD.2011.5764181

BENKO, Ľubomír, 2015. Vplyv odhadu podielu navigačných stránok na kvantitu extrahovaných vzorov správania sa používateľov webu. In: *Študentská vedecká konferencia 2015*. Nitra: Fakulta prírodných vied, Univerzita Konštantína Filozofa v Nitre, s. 288–294.

BENKO, Ľubomír a Daša MUNKOVÁ, 2016. Application of POS Tagging in Machine Translation Evaluation. In: *DIVAI 2016 : 11th International Scientific Conference on Distance Learning in Applied Informatics, Sturovo, May 2 – 4, 2016*. Sturovo: Wolters Kluwer, ISSN 2464-7489, s. 471–489.

BENKO, Ľubomír, Jaroslav REICHEL a Michal MUNK, 2015. Analysis of student

behavior in virtual learning environment depending on student assessments. In: *ICETA 2015 : 13th International Conference on Emerging eLearning Technologies and Applications, Stary Smokovec, November 26 - 27, 2015*. Stary Smokovec: Danvers : IEEE, s. 33–38.

BERENDT, Bettina, Bamshad MOBASHER, Miki NAKAGAWA a Myra SPILIOPOULOU, 2003. The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis. *Lecture Notes in Computer Science* [online]. roč. 2703, s. 159–179. Získáno z: doi:10.1007/978-3-540-39663-5_10

BERENDT, Bettina a Myra SPILIOPOULOU, 2000. Analysis of navigation behaviour in web sites integrating multiple information systems. *The VLDB Journal* [online]. B.m.: Springer-Verlag, roč. 9, č. 1, s. 56 [vid. 25. únor 2017]. ISSN 10668888. Získáno z: doi:10.1007/s007780050083

BERKA, Petr, 2003. *Dobývání znalostí z databází*. Praha: Academia. ISBN 80-200-1062-9.

BERRY, M.J.A. a G.S. LINOFF, 2004. *Data mining techniques: for marketing, sales, and customer relationship management* [online]. Získáno z: http://portal.acm.org/citation.cfm?id=983642

BHAWSAR, Sawan, Kshitij PATHAK a Vibhor PATIDAR, 2012. New Framework for Web Access Prediction. *International Journal of Computer Technology and Electronics Engineering*. roč. 2, č. 1, s. 48–53.

CARMONA, C.J., S. RAMÍREZ-GALLEGO, F. TORRES, E. BERNAL, M.J. DEL JESUS a S. GARCÍA, 2012. Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. *Expert Systems with Applications* [online]. roč. 39, č. 12, s. 11243–11249. ISSN 09574174. Získáno z: doi:10.1016/j.eswa.2012.03.046

CATLEDGE, Lara D. a James E. PITKOW, 1995. Characterizing browsing strategies in the World-Wide web. *Computer Networks and ISDN Systems* [online]. roč. 27, č. 6, s. 1065–1073. ISSN 01697552. Získáno z: doi:10.1016/0169-7552(95)00043-7

CERNA, Miloslava a Petra POULOVA, 2008. VISIT RATE OF INTERNET PORTALS AND UTILIZATION OF THEIR TOOLS AND SERVICES. *E & M EKONOMIE A MANAGEMENT*. roč. 11, č. 4, s. 132–143. ISSN 1212-3609.

CHAPMAN, Pete, Julian CLINTON, Randy KERBER, Thomas KHABAZA, Thomas

REINARTZ, Colin SHEARER a Rudiger WIRTH, 2000. *CRISP-DM 1.0: Step-by-step Data Mining Guide* [online]. B.m.: SPSS. Získáno z: https://books.google.sk/books?id=po7FtgAACAAJ

CIOS, K.J., W. PEDRYCZ, R.W. SWINIARSKI a L.A. KURGAN, 2007. *Data mining: A knowledge discovery approach* [online]. ISBN 9780387333335. Získáno z: doi:10.1007/978-0-387-36795-8

COOLEY, R, B MOBASHER, J SRIVASTAVA a OTHERS, 1999. Data preparation for mining world wide web browsing patterns. *Knowledge and information systems*. B.m.: Citeseer, roč. 1, č. 1, s. 5–32.

DABROWSKA-ZIELINSKA, K., F. KOGAN, A. CIOLKOSZ, M. GRUSZCZYNSKA a W. KOWALIK, 2002. Modelling of crop growth conditions and crop yield in Poland using AVHRR-based indices. *International Journal of Remote Sensing* [online]. roč. 23, č. 6, s. 1109–1123 [vid. 7. březen 2017]. ISSN 0143-1161. Získáno z: doi:10.1080/01431160110070744

FAYYAD, Usama M., Gregory PIATETSKY-SHAPIRO a Padhraic SMYTH, 1996. From Data Mining to Knowledge Discovery in Databases [online]. roč. 17, č. 3, s. 37–54. ISSN 0738-4602. Získáno z: doi:10.1609/AIMAG.V17I3.1230

HASAN, Iftekhar, Krzysztof JACKOWICZ, Oskar KOWALEWSKI a Łukasz KOZŁOWSKI, 2013. Market discipline during crisis: Evidence from bank depositors in transition countries. *Journal of Banking & Finance* [online]. 12., roč. 37, č. 12, s. 5436–5451 [vid. 1. duben 2016]. ISSN 03784266. Získáno z: doi:10.1016/j.jbankfin.2013.06.007

HAYS, W., 1988. *Statistics*. New York, NY, USA: CBS College Publishing.

HEARST, Marti A., 1999. Untangling text data mining. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* - [online]. Morristown, NJ, USA: Association for Computational Linguistics, s. 3–10 [vid. 4. leden 2017]. ISBN 1558606093. Získáno z: doi:10.3115/1034678.1034679

HUYNH, Toan a James MILLER, 2009. Empirical observations on the session timeout threshold. *Information Processing and Management* [online]. roč. 45, č. 5, s. 513–528. ISSN 03064573. Získáno z: doi:10.1016/j.ipm.2009.04.007

JIN, Wei a Rohini K SRIHARI, 2007. Graph-based Text Representation and Knowledge Discovery. In: *Proceedings of the 2007 ACM Symposium on Applied Computing* [online]. New York, NY, USA: ACM, s. 807–811. SAC '07. ISBN 1-59593-480-4. Získáno z: doi:10.1145/1244002.1244182

JINDAL, R. a SHWETA, 2018. A modified knowledge discovery process in the text documents. *International Journal of Innovative Computing, Information and Control*. roč. 14, č. 3, s. 817–832.

JUSTICIA DE LA TORRE, C., D. SÁNCHEZ, I. BLANCO a M. J. MARTÍN-BAUTISTA, 2018. Text Mining: Techniques, Applications, and Challenges. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* [online]. roč. 26, č. 04, s. 553–582. ISSN 0218-4885. Získáno z: doi:10.1142/S0218488518500265

KAPUSTA, J., M. MUNK, P. SVEC a A. PILKOVA, 2014a. Determining the time window threshold to identify user sessions of stakeholders of a commercial bank portal. In: *Procedia Computer Science* [online]. Získáno z: doi:10.1016/j.procs.2014.05.163

KAPUSTA, Jozef, Michal MUNK a Martin DRLÍK, 2012. Cut-off time calculation for user session identification by reference length. In: *2012 6th International Conference on Application of Information and Communication Technologies, AICT 2012 - Proceedings*.

KAPUSTA, Jozef, Michal MUNK, Peter SVEC a Anna PILKOVA, 2014b. Determining the time window threshold to identify user sessions of stakeholders of a commercial bank portal. *Procedia Computer Science*. roč. 29, s. 1779–1790.

KAPUSTA, Jozef, Anna PILKOVA, Michal MUNK a Peter SVEC, 2013. Data pre-processing for web log mining: Case study of commercial bank website usage analysis. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*. roč. 61, č. 4, s. 973–979.

LI, Yan, Boqin FENG a Qinjiao MAO, 2008. Research on path completion technique in web usage mining. In: *Proceedings - International Symposium on Computer Science and Computational Technology, ISCSCT 2008*. s. 554–559.

LIU, Bing, 2011. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* [online]. ISBN 978-3-642-19459-7. Získáno z: doi:10.1007/978-3-642-19460-3

LOH, Stanley, Leandro Krug WIVES a José Palazzo M DE OLIVEIRA, 2000. Concept-based Knowledge Discovery in Texts Extracted from the Web. *SIGKDD Explor. Newsl.*

[online]. New York, NY, USA: ACM, roč. 2, č. 1, s. 29–39. ISSN 1931-0145. Získáno z: doi:10.1145/360402.360414

LOSARWAR, Vijayashiri a Madhuri JOSHI, 2012. Data Preprocessing in Web Usage Mining. In: *International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore*. s. 1–5.

MAKKAR, Priyanka, Payal GULATI a A SHARMA, 2010. A novel approach for predicting user behavior for improving web performance. *International Journal on Computer Science and Engineering*. roč. 2, č. 4, s. 1233–1236.

MAZZA, Riccardo, Marco BETTONI, Marco FARÉ a Luca MAZZOLA, 2012. MOCLog – Monitoring Online Courses with log data. *1st Moodle Research Conference* [online]. s. 132–139. Získáno z: http://research.moodle.net/mod/data/view.php?d=3&rid=17

MING-SYAN CHEN, Ming-Syan, Jong Soo JONG SOO PARK a P.S. YU, 1998. Efficient data mining for path traversal patterns. *IEEE Transactions on Knowledge and Data Engineering* [online]. B.m.: IEEE Educational Activities Department, roč. 10, č. 2, s. 209–221 [vid. 25. únor 2017]. ISSN 10414347. Získáno z: doi:10.1109/69.683753

MUNK, Michal a Lubomir BENKO, 2018. Using Entropy in Web Usage Data Preprocessing. *Entropy* [online]. roč. 20, č. 1, s. 67. Získáno z: doi:10.3390/e20010067

MUNK, Michal a Ľubomír BENKO, 2016. Improving the Session Identification Using the Ratio of Auxiliary Pages Estimate. *Proceedings of the Mediterranean Conference on Information {&} Communication Technologies 2015: MedICT 2015 Volume 2* [online]. Cham: Springer International Publishing, roč. 381, s. 551–556. Získáno z: doi:10.1007/978-3-319-30298-0_56

MUNK, Michal a Ľubomír BENKO, 2017. Unconventional Usage of Entropy in the Field of Web Usage Data Preprocessing and Machine Translation Evaluation. In: *Applied Physics, System Science and Computers*. Cham: Springer International Publishing.

MUNK, Michal, Ľubomír BENKO, Mikuláš GANGUR a Milan TURČÁNI, 2015. Influence of ratio of auxiliary pages on the pre-processing phase of Web Usage Mining. *E+M Ekonomie a Management* [online]. roč. 18, č. 3, s. 144–159. Získáno z: doi:dx.doi.org/10.15240/tul/001/2015-3-013

MUNK, Michal a Martin DRLIK, 2011a. Impact of different pre-processing tasks on

effective identification of users' behavioral patterns in web-based educational system. In: *Procedia Computer Science*. s. 1640–1649.

MUNK, Michal a Martin DRLIK, 2011b. Influence of Different Session Timeouts Thresholds on Results of Sequence Rule Analysis in Educational Data Mining. In: Hocine CHERIFI, JasniMohamad ZAIN a Eyas EL-QAWASMEH, ed. *Digital Information and Communication Technology and Its Applications SE - 6* [online]. B.m.: Springer Berlin Heidelberg, Communications in Computer and Information Science, s. 60–74. ISBN 978-3-642-21983-2. Získáno z: doi:10.1007/978-3-642-21984-9_6

MUNK, Michal a Martin DRLÍK, 2014. Analysis of stakeholders' behaviour depending on time in virtual learning environment. *Applied Mathematics and Information Sciences*. roč. 8, č. 2, s. 773–785.

MUNK, Michal, Martin DRLIK, Lubomir BENKO a Jaroslav REICHEL, 2017a. Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniques. *IEEE Access* [online]. roč. 5, s. 8989–9004. ISSN 21693536. Získáno z: doi:10.1109/ACCESS.2017.2706302

MUNK, Michal, Martin DRLÍK, Jozef KAPUSTA a Daša MUNKOVÁ, 2013a. Methodology design for data preparation in the process of discovering patterns of web users behaviour. *Applied Mathematics and Information Sciences* [online]. roč. 7, č. 1 L, s. 27–36. ISSN 19350090. Získáno z: http://www.scopus.com/inward/record.url?eid=2-s2.0-84896693902&partnerID=tZOtx3y1

MUNK, Michal, Martin DRLIK a Marta VRÁBELOVÁ, 2011a. Probability Modelling of Accesses to the Course Activities in the Web-Based Educational System. In: [online]. B.m.: Springer Berlin Heidelberg, s. 485–499 [vid. 25. únor 2017]. Získáno z: doi:10.1007/978-3-642-21934-4_39

MUNK, Michal a Jozef KAPUSTA, 2014. *Web Usage Mining: Príprava a modelovanie dát*. Nitra: Univerzita Konštantína Filozofa v Nitre. ISBN 978-80-558-0692-1.

MUNK, Michal, Jozef KAPUSTA a Peter ŠVEC, 2010a. Data preprocessing evaluation for web log mining: Reconstruction of activities of a web visitor. In: *Procedia Computer Science* [online]. s. 2273–2280. Získáno z: doi:10.1016/j.procs.2010.04.255

MUNK, Michal, Jozef KAPUSTA, Peter ŠVEC a Milan TURČÁNI, 2010b. Data Advance Preparation Factors Affecting Results of Sequence Rule Analysis in Web Log

Mining. *E+M Ekonomie a Management*. roč. 13, č. 4, s. 143–160.

MUNK, Michal, Daša MUNKOVÁ a Ľubomír BENKO, 2016. Identification of Relevant and Redundant Automatic Metrics for MT Evaluation. In: *Multi-disciplinary Trends in Artificial Intelligence (MIWAI 2016) Book Series: Lecture Notes in Computer Science* [online]. Cham: Springer International Publishing, s. 141–152 [vid. 4. leden 2017]. Získáno z: doi:10.1007/978-3-319-49397-8_12

MUNK, Michal, Anna PILKOVA, Lubomir BENKO a Petra BLAŽEKOVÁ, 2017b. Pillar 3: market discipline of the key stakeholders in CEE commercial bank and turbulent times. *Journal of Business Economics and Management* [online]. B.m.: Taylor & Francis, roč. 18, č. 5, s. 954–973. Získáno z: doi:10.3846/16111699.2017.1360388

MUNK, Michal, Anna PILKOVA, Ľubomír BENKO a Petra BLAŽEKOVÁ, 2018. Pillar3: Predictive Web Usage Analysis in Banking. *Economic Modelling* (v recenznom konaní).

MUNK, Michal, Anna PILKOVA, Martin DRLIK, Jozef KAPUSTA a Peter SVEC, 2012. Verification of the fulfilment of the purposes of basel ii, pillar 3 through application of the web log mining methods. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*. roč. 60, č. 2, s. 217–222.

MUNK, Michal, Anna PILKOVA, Jozef KAPUSTA, Peter SVEC a Martin DRLIK, 2013b. Pillar 3 and Modelling of Stakeholders' Behaviour at the Commercial Bank Website during the Recent Financial Crisis. *Procedia Computer Science* [online]. roč. 18, s. 1747–1756 [vid. 31. březen 2016]. ISSN 18770509. Získáno z: doi:10.1016/j.procs.2013.05.343

MUNK, Michal, Marta VRÁBELOVÁ a Jozef KAPUSTA, 2011b. Probability modeling of accesses to the web parts of portal. *Procedia Computer Science* [online]. roč. 3, s. 677–683 [vid. 25. únor 2017]. ISSN 18770509. Získáno z: doi:10.1016/j.procs.2010.12.113

NITHYA, P. a P. SUMATHI, 2012. Novel pre-processing technique for web log mining by removing global noise and web robots. In: *2012 NATIONAL CONFERENCE ON COMPUTING AND COMMUNICATION SYSTEMS* [online]. B.m.: IEEE, s. 1–5. ISBN 978-1-4673-1953-9. Získáno z: doi:10.1109/NCCCS.2012.6412976

PABARSKAITE, Zidrina a Aistis RAUDYS, 2007. A process of knowledge discovery from web log data: Systematization and critical review. *Journal of Intelligent Information*

Systems [online]. roč. 28, č. 1, s. 79–104. ISSN 09259902. Získáno z: doi:10.1007/s10844-006-0004-1

PILKOVA, Anna, Michal MUNK, Peter SVEC a Michal MEDO, 2015. Assessment of the Pillar 3 Financial and Risk Information Disclosures Usefulness to the Commercial Banks Users. *Lecture Notes in Artificial Intelligence*. roč. 9227, s. 429–440.

RAFFI, F., C. KATLAMA, M. SAAG, M. WILKINSON, J. CHUNG, L. SMILEY a M. SALGO, 2006. Week-12 Response to Therapy as a Predictor of Week 24, 48, and 96 Outcome in Patients Receiving the HIV Fusion Inhibitor Enfuvirtide in the T-20 versus Optimized Regimen Only (TORO) Trials. *Clinical Infectious Diseases* [online]. B.m.: Oxford University Press, 15.3., roč. 42, č. 6, s. 870–877 [vid. 7. březen 2017]. ISSN 1058-4838. Získáno z: doi:10.1086/500206

RAJENDERAN, Amog, 2012. Data preparation for web mining--a survey. *International Journal of Advanced Computer Research*. B.m.: International Journal of Advanced Computer Research(IJACR), roč. 2, č. 6, s. 1–7.

REICHEL, J. a P. KUNA, 2014. Analysis of students behaviour in virtual environment. In: *2014 IEEE 12th IEEE International Conference on Emerging eLearning Technologies and Applications (ICETA)* [online]. B.m.: IEEE, s. 419–423 [vid. 21. říjen 2015]. ISBN 978-1-4799-7739-0. Získáno z: doi:10.1109/ICETA.2014.7107621

REICHEL, Jaroslav, Peter KUNA, Ľubomír BENKO a Michal MUNK, 2015. Visit rate analysis of course activities : case study. In: *ICETA 2015 : 13th International Conference on Emerging eLearning Technologies and Applications, Stary Smokovec, November 26 - 27, 2015*. Stary Smokovec: Danvers : IEEE, s. 319–324.

RICE, William, 2011. *Moodle 2.0 E-Learning Course Development*. ISBN 9781849515269.

ROMERO, Cristóbal, JoséRaúl ROMERO a Sebastián VENTURA, 2014. A Survey on Pre-Processing Educational Data. In: Alejandro PEÑA-AYALA, ed. *Educational Data Mining SE - 2* [online]. B.m.: Springer International Publishing, Studies in Computational Intelligence, s. 29–64. ISBN 978-3-319-02737-1. Získáno z: doi:10.1007/978-3-319-02738-8_2

ROMERO, Cristóbal, Sebastián VENTURA, Amelia ZAFRA a Paul de BRA, 2009. Applying Web usage mining for personalizing hyperlinks in Web-based adaptive

educational systems. *Computers and Education*. roč. 53, č. 3, s. 828–840.

SAEL, N, A MARZAK a H BEHJA, 2013. Web Usage Mining data preprocessing and multi level analysis on Moodle. In: *Computer Systems and Applications (AICCSA), 2013 ACS International Conference on* [online]. s. 1–7. Získáno z: doi:10.1109/AICCSA.2013.6616427

SCHMIDT, Sebastian, Simon MANSCHITZ, Christoph RENSING a Ralf STEINMETZ, 2013. Extraction of Address Data from Unstructured Text Using Free Knowledge Resources. In: *Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies* [online]. New York, NY, USA: ACM, s. 7:1--7:8. i-Know '13. ISBN 978-1-4503-2300-0. Získáno z: doi:10.1145/2494188.2494193

SPILIOPOULOU, Myra a Lukas C. FAULSTICH, 1999. WUM: A Tool for Web Utilization Analysis. In: [online]. B.m.: Springer Berlin Heidelberg, s. 184–203 [vid. 25. únor 2017]. Získáno z: doi:10.1007/10704656_12

SPILIOPOULOU, Myra, Bamshad MOBASHER, Bettina BERENDT a Miki NAKAGAWA, 2003. A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. *INFORMS Journal on Computing* [online]. B.m.: INFORMS, 5., roč. 15, č. 2, s. 171–190 [vid. 16. říjen 2016]. ISSN 1091-9856. Získáno z: doi:10.1287/ijoc.15.2.171.14445

SRIVASTAVA, Jaideep, Robert COOLEY, Mukund DESHPANDE a Pang-ning TAN, 2000. Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data. *Text* [online]. roč. 1, č. 2, s. 12–23. ISSN 19310145. Získáno z: doi:10.1145/846183.846188

SRIVASTAVA, Mitali, Rakhi GARG a P. K. MISHRA, 2015. Analysis of Data Extraction and Data Cleaning in Web Usage Mining. In: *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015) - ICARCSET '15* [online]. New York, New York, USA: ACM Press, s. 1–6 [vid. 6. březen 2017]. ISBN 9781450334419. Získáno z: doi:10.1145/2743065.2743078

STANKOVIČOVÁ, Iveta a Mária VOJTKOVÁ, 2007. *Viacrozmerné štatistické metódy s aplikáciami*. Bratislava: Iura Edition, spol. s r.o. ISBN 9788080781520.

STATSOFT INC., 2013. *Electronic Statistics Textbook* [online]. B.m.: Tulsa, OK:

StatSoft. Získáno z: http://www.statsoft.com/textbook/

TAN, Ah-Hwee, 1999. Text Mining: The state of the art and the challenges. In: *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases* [online]. ISBN 1940-6029. Získáno z: doi:10.1.1.38.7672

TAUSCHER, Linda a Saul GREENBERG, 1997. Revisitation patterns in World Wide Web navigation. In: *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '97* [online]. New York, New York, USA: ACM Press, s. 399–406 [vid. 25. únor 2017]. ISBN 0897918029. Získáno z: doi:10.1145/258549.258816

VARNAGAR, C R, N N MADHAK, T M KODINARIYA a J N RATHOD, 2013. Web usage mining: A review on process, methods and techniques. In: *2013 International Conference on Information Communication and Embedded Systems, ICICES 2013* [online]. s. 40–46. Získáno z: http://www.scopus.com/inward/record.url?eid=2-s2.0-84879875079&partnerID=40&md5=7dc4b24142841cea2581cd6d2b187b40

VELLINGIRI, J. a S. CHENTHUR PANDIAN, 2011. A novel technique for web log mining with better data cleaning and transaction identification. *Journal of Computer Science* [online]. roč. 7, č. 5, s. 683–689. ISSN 15493636. Získáno z: doi:10.3844/jcssp.2011.683.689

VERDHEN, Anand, Bhagu R. CHAHAR a Om P. SHARMA, 2014. Snowmelt Modelling Approaches in Watershed Models: Computation and Comparison of Efficiencies under Varying Climatic Conditions. *Water Resources Management* [online]. B.m.: Springer Netherlands, 5.9., roč. 28, č. 11, s. 3439–3453 [vid. 7. březen 2017]. ISSN 0920-4741. Získáno z: doi:10.1007/s11269-014-0662-7

W3C, 1995. *Configuration File of W3C httpd* [online] [vid. 23. leden 2017]. Získáno z: https://www.w3.org/Daemon/User/Config/Logging.html

ZHANG, S., C. ZHANG a Q. YANG, 2003. Data preparation for data mining. *Applied Artificial Intelligence* [online]. roč. 17, č. 5–6, s. 375–381. Získáno z: doi:10.1080/713827180