Sum – súčet hodnôt premennej.

Percent 0's – percentuálne zastúpenie núl.

Percent 1's – percentuálne zastúpenie jednotiek.

Number of valid cases – počet platných prípadov.

df – stupne voľnosti.

Q – testovacia štatistika.

p - hodnota významnosti. Ak je väčšia ako 0,05, nulovú hypotézu nemožno zamietnuť s 95% spoľahlivosťou. Znamená to, že nie je štatisticky významný rozdiel medzi premennými. Ak je menšia alebo rovná ako 0,05, nulová hypotéza sa s 95% spoľahlivosťou zamietne.

Zamietame nulovú hypotézu, že úlohy sú rovnako obtiažne.

4.7 Evalvácia prípravy dát pre web log mining: Prípadová štúdia

Prípadová štúdia je zameraná na prípravu dát v procese objavovania znalostí, konkrétne na riešenie **problému** evalvácie základných krokov prípravy dát pre web log mining (Munk, Kapusta a Švec, 2009).

Cieľom je zistiť do akej miery je potrebné realizovať prípravu dát pre web log mining s dôrazom na čistenie dát a určiť nevyhnutné kroky pre získanie spoľahlivých dát z logovacieho súboru.

Metodika:

1. Získanie dát z logovacieho súboru webového servera.
2. Príprava dát na štyroch rôznych úrovniach:
    a. Súbor 1: očistené dáta od nepotrebných dát/požiadaviek na obrázky, skripty a štýly - hrubé dáta (raw data),
    b. Súbor 2: očistené dáta od nepotrebných dát a prístupov robotov,
    c. Súbor 3: očistené dáta od nepotrebných dát, prístupov robotov a NAT/proxy zariadení,
    d. Súbor 4: očistené dáta od nepotrebných dát, prístupov robotov a s identifikáciou sedení na základe času.

3. Analýza dát - hľadanie vzorcov správania sa používateľov webu v jednotlivých súboroch.
4. Porozumenie výstupným dátam - vytvorenie dátových matíc z výstupov analýzy.
5. Porovnanie výsledkov analýzy dát spracovaných na rôznej úrovni prípravy dát z hľadiska kvantity a kvality nájdených sekvenčných pravidiel - vzorcov správania sa používateľov pri prehľadávaní webu:
    a. porovnanie podielu nájdených pravidiel v skúmaných súboroch,
    b. porovnanie hodnôt miery podpory (support) a spoľahlivosti (confidence) nájdených pravidiel v skúmaných súboroch.

Použité metódy: popisná štatistika, Kendallov koeficient zhody, neparametrická korelácia, Kendallov koeficient Tau.

4.7.1 Základné techniky prípravy dát

Predpokladom dobre realizovanej analýzy sú kvalitné - spoľahlivé dáta. Dáta logovacieho súboru obsahujú aj nepotrebné, irelevantné, nepresné a neúplné informácie. Pri web log miningu spoľahlivé dáta zabezpečíme kvalitnou prípravou dát z logovacieho súboru.

Nepotrebné údaje sú riadky logovacieho súboru, v ktorých sú zaznamenané požiadavky na obrázky, štýly a skripty alebo iné súbory, ktoré môžu byť vložené do stránky. Táto časť je z celého procesu prípravy dát najjednoduhšia, pretože pozostáva iba z odfiltrovania dát, ktoré nie sú podľa zvolenej šablóny.

Irelevantné údaje sú riadky logovacieho súboru, v ktorých sú zaznamenané prístupy nie používateľov - návštevníkov webu, ale robotov rôznych vyhľadávacích služieb, ktoré prechádzajú celým webom, väčšinou rekurzívne a postupne. Detekcia robotov je možná buď na základe ich identifikácie pomocou poľa User-Agent alebo IP adresy ich porovnaním s databázou www.robotstxt.org. Táto databáza nemusí obsahovať údaje o všetkých vyhľadávacích robotoch avšak tie minoritné predstavujú štatisticky zanedbateľný počet. Inou metódou identifikácie je, či robot pristupoval k súboru robots.txt alebo nie (Lourenco a Belo, 2006). Na základe prístupu k tomuto súboru vieme jednoznačne robota identifikovať aj keď má nesprávne nastavené pole User-Agent.

Nepresnosť dát súvisí s anonymným charakterom dát. Log súbor považujeme za zdroj anonymných dát o používateľovi z toho pohľadu, že nezaznamenávame jeho osobné údaje ani jeho jednoznačnú identifikáciu. Z tohto dôvodu je rekonštrukcia aktivít každého návštevníka náročná. V súčasnej dobe je takmer štandardom, že viacero používateľov zdieľa spoločnú IP adresu, či už sa nachádzajú za určitým NAT (Network Address Translation) zariadením (väčšinou domácnosti) alebo proxy zariadením (väčšie firmy). Autentifikačné mechanizmy môžu identifikáciu používateľa uľahčiť, avšak ich použitie je kvôli ochrane súkromia neželané (Berendt a Spiliopoulou, 2000). Na identifikáciu NAT alebo proxy zariadení používame reverzné doménové záznamy, na základe ktorých vieme rozlíšiť či k danej stránke pristupuje jeden používateľ (alebo malý počet) alebo ich je viacero. V prípade veľkého počtu používateľov z jednej IP adresy ich potrebujeme rozlíšiť na základe identifikácie sedení, ktorej cieľom je rozdeliť jednotlivé prístupy každého používateľa do oddelených relácií (Cooley, Mobasher a Srivastava, 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 (Chen, Park a Yu, 1996). Najjednoduchšou metódou je, ak za sedenie považujeme sériu kliknutí za určitý čas, napr. 30 minút (Berendt a Spiliopoulou, 2000). Reálnu hodnotu pre sedenie môžeme získať na základe empirických dát, v našom prípade pomocou nástroja Google Analytics. Kde na základe hodnoty *avg. time on site*, ktorá reprezentuje priemerný čas používateľa na webovej stránke sme zvolili časové okno (Session Timeout Threshold, STT) dĺžky 10 minút.

V našom experimente sa snažíme zistiť nakoľko sú tieto základné kroky prípravy dát pri použití sekvenčnej analýzy potrebné, t.j. cieľom je vyhodnotiť relevantnosť základných krokov prípravy dát pre sekvenčnú analýzu. Predpokladáme, že jednotlivé úrovne prípravy dát budú mať významný vplyv ako na kvantitu extrahovaných pravidiel, tak aj na ich kvalitu v zmysle ich základných charakteristík kvality.

4.7.2 Porovnanie podielu nájdených pravidiel v skúmaných súboroch

Neočistený súbor (Tabuľka 28) obsahuje takmer 40000 prípadov, ktoré predstavujú prístupy na portál počas jedného týždňa, z ktorých takmer 11% prípadov sú prístupy robotov a viac ako 9% predstavuje prístupy z adries NAT/proxy. Po očistení súborov od robotov a NAT/proxy zariadení sledujeme minimálne rozdiely v počte návštev (costumer's

sequences) a v počte frekventovaných sekvencií. Naopak pri identifikácii sedení sledujeme dvojnásobný nárast návštev/identifikovaných sekvencií a pokles frekventovaných sekvencií viac ako o polovicu.

Tabuľka 28 Počet prístupov a sekvencií v skúmaných súboroch

	Počet prístupov	**Počet identifikovaných sekvencií**	**Počet frekventovaných sekvencií**
Súbor 1: hrubé dáta	39688	4506	90
Súbor 2: očistené od robotov	35374	4454	91
Súbor 3: očistené od robotov a NAT/proxy zariadení	31761	4242	87
Súbor 4: očistené od robotov a s identifikáciou sedení	35374	8875	37

Sledované boli prístupy používateľov na jednotlivé webové časti skúmaného portálu v priebehu jedného týždňa. Výsledkom analýzy sú sekvenčné pravidlá (Tabuľka 29), ktoré sme získali z frekventovaných sekvencií spĺňajúcich minimálnu podporu (v našom prípade *min s* = 0,03). Frekventované sekvencie sme získali z identifikovaných sekvencií, t.j. návštev jednotlivých používateľov portálu v priebehu jedného týždňa.

Tabuľka 29 Nájdené sekvenčné pravidlá v skúmaných súboroch (s - podpora, c - spoľahlivosť)

Sekvenčné pravidlá			Súbor 1		Súbor 2		Súbor 3		Súbor 4	
Predpoklad	**==>**	**Záver**	**s**	**c**	**s**	**c**	**s**	**c**	**s**	**c**
(a178)	==>	(a178)	3,15	38,07	3,19	38,38	3,11	38,48		
(a178)	==>	(a180)	5,22	63	5,28	63,51	5,09	62,97	3,21	59,01
(a180)	==>	(a180), (a180)	4,28	37,92	4,33	38,45	4,13	37,55		
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
(a49)	==>	(a180)	4,86	33,69	4,92	33,8	4,76	34,01		
(a49)	==>	(a180), (a178)			3,01	20,68				
(a49)	==>	(a180), (a180)	3,06	21,23	3,1	21,3				
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
(c7)	==>	(c6)	3,37	19,61	3,39	19,61	3,28	19,41		
(c7)	==>	(c7)	6,5	37,81	6,58	38,05	6,41	37,99	3,63	30,35
(search)	==>	(search)	5,15	52,61	5,21	52,61	5,04	51,82	3,27	50,43
Absolútny počet extrahovaných pravidiel			75		78		72		13	
Relatívny počet extrahovaných pravidel			96,2		100		92,3		16,7	

Medzi výsledkami sekvenčnej analýzy je vysoká zhoda (Tabuľka 29), čo sa týka podielu nájdených pravidiel v prípade prvých troch súborov. Konkrétne najviac pravidiel 78 bolo extrahovaných zo súboru bez robotov (Súbor 2), následne 75 zo súboru hrubé dáta (Súbor 1) a nakoniec 72 pravidiel zo súboru bez robotov a NAT/proxy zariadení (Súbor 3). V súboroch hrubé dáta (Súbor 1) a bez robotov a NAT/proxy zariadení (Súbor 3) predstavuje podiel nájdených pravidiel viac ako 90% z počtu nájdený pravidiel v súbore bez robotov (Súbor 2). Rozdiely v počte nájdených pravidiel medzi prvými troma súbormi sú minimálne.

Predpokladali sme, že očistením dát od robotov a NAT/proxy zariadení sa zásadnejšie zmenia výsledky. V súbore bez robotov (Súbor 2) boli nájdené rovnaké pravidlá ako boli nájdené v súbore hrubé dáta (Súbor 1), rozdiel spočíval iba v troch nových pravidlách, ktoré boli nájdené v súbore bez robotov (Súbor 2). Očistením dát od NAT/proxy zariadení bolo objavených o šesť pravidiel menej ako v súbore bez robotov (Súbor 2), ostatné boli totožné s tými, ktoré boli nájdené v súboroch hrubé dáta (Súbor 1) a bez robotov (Súbor 2).

Naopak veľké rozdiely v počte extrahovaných pravidiel sú medzi súborom s identifikáciou sedení (Súbor 4) a ostatnými, kde podiel nájdených pravidiel v tomto súbore predstavuje menej ako 20% z počtu nájdený pravidiel v súbore bez robotov (Súbor 2). Podiel

nájdených pravidiel je síce najmenší, ale na druhej strane jedine v tomto prípade nebol zaznamenaný výskyt nevysvetliteľných pravidiel (Tabuľka 29), ako napr.:

(a180) ==> (a180), (a180), support = 4,28, confidence = 37,92.

Interpretácia pravidla je nasledujúca: ak používateľ navštívi webovú časť Adresár zamestnancov univerzity s takmer 38% pravdepodobnosťou navštívi webové časti Adresár zamestnancov univerzity, Adresár zamestnancov univerzity.

Identifikáciou sedení sme neumožnili len zahrnutie NAT/proxy zariadení do analýzy, ale sme aj eliminovali problém „jeden pc viac sedení“, čím sa výrazne znížil počet nesúvisiacich sekvencií. Pre internetové kaviarne, knižnice, učebne a pod. je špecifické, že viacero anonymných používateľov používa jeden počítač, tento fakt sa podarilo eliminovať identifikáciou sedení na základe času. Ukazujú to aj nájdené pravidlá, kde jedine vo výsledkoch súboru s identifikáciou sedení (Súbor 4) sa nevyskytovali nevysvetliteľné pravidlá.

4.7.3 Porovnanie hodnôt miery podpory a spoľahlivosti nájdených pravidiel v skúmaných súboroch

Kvalita sekvenčných pravidiel sa posudzuje dvoma ukazovateľmi (Stankovičová, 2009):

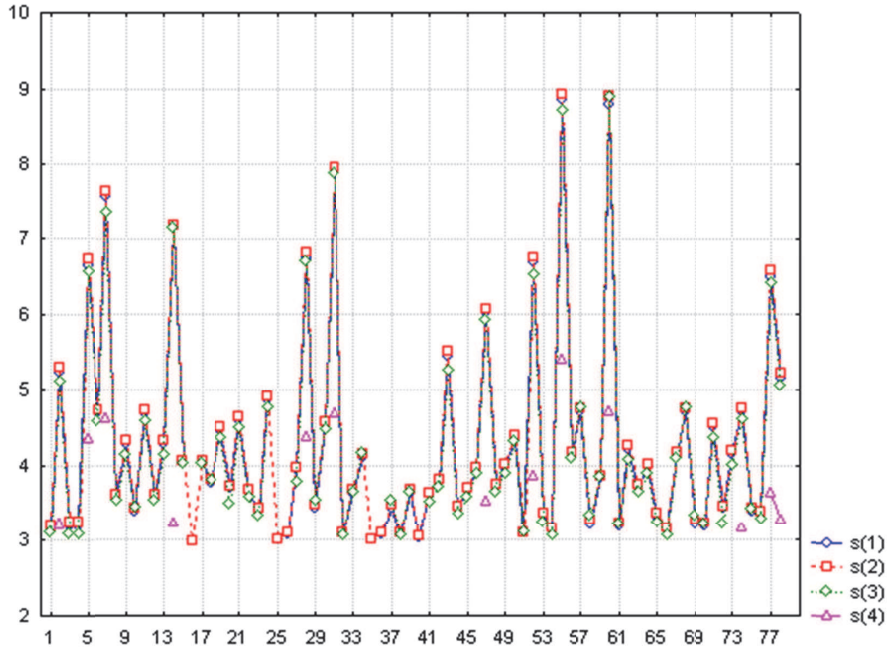
- podpora (support),
- spoľahlivosť (confidence).

Vo výsledkoch sekvenčnej analýzy je vysoká zhoda v kvalite nájdených pravidiel, čo sa týka hodnôt charakteristiky podpory nájdených pravidiel medzi jednotlivými súbormi (Tabuľka 30). Kendallov koeficient zhody predstavuje mieru zhody v podpore nájdených pravidiel medzi jednotlivými súbormi. Hodnota koeficientu je 0,97, pričom 1 znamená dokonalú zhodu a 0 znamená nezhodu.

Tabuľka 30 Kendallov koeficient zhody pre mieru podpory nájdených pravidiel (s - podpora)

	Priemer poradí	**Súčet poradí**	**Priemer**	**Smerodajná odchýlka**
s(1)	2,92	38	6,77	1,3
s(2)	4	52	6,83	1,32
s(3)	2,08	27	6,68	1,33
s(4)	1	13	4	0,73
Kendallov koeficient zhody			0,9716	

Výsledky vizualizuje čiarový graf viacnásobných premenných (Obrázok 40). Graf vizualizuje hodnoty podpory nájdených pravidiel v jednotlivých súboroch. Krivky sa kopírujú, čo nám iba potvrdzuje zistenú zhodu v hodnotách podpory medzi jednotlivými súbormi.



Obrázok 40 Čiarový graf viacnásobných premenných pre mieru podpory nájdených pravidiel

Z korelačnej matici vidíme (Tabuľka 31), že najväčšia miera zhody/závislosti v podpore je medzi pravidlami nájdenými v súbore hrubé dáta (Súbor 1), bez robotov (Súbor 2) a bez robotov a NAT/proxy zariadení (Súbor 3). Naopak menšia zhoda, ale tiež významná ($p < 0,05$) je medzi súborom s identifikáciou sedení (Súbor 4) a ostatnými súbormi.

Tabuľka 31 Kendallove koeficienty Tau pre mieru podpory nájdených pravidiel

	s(1)	**s(2)**	**s(3)**	**s(4)**
s(1)	1	0,998	0,948	0,795
s(2)	0,998	1	0,95	0,795
s(3)	0,948	0,95	1	0,795
s(4)	0,795	0,795	0,795	1

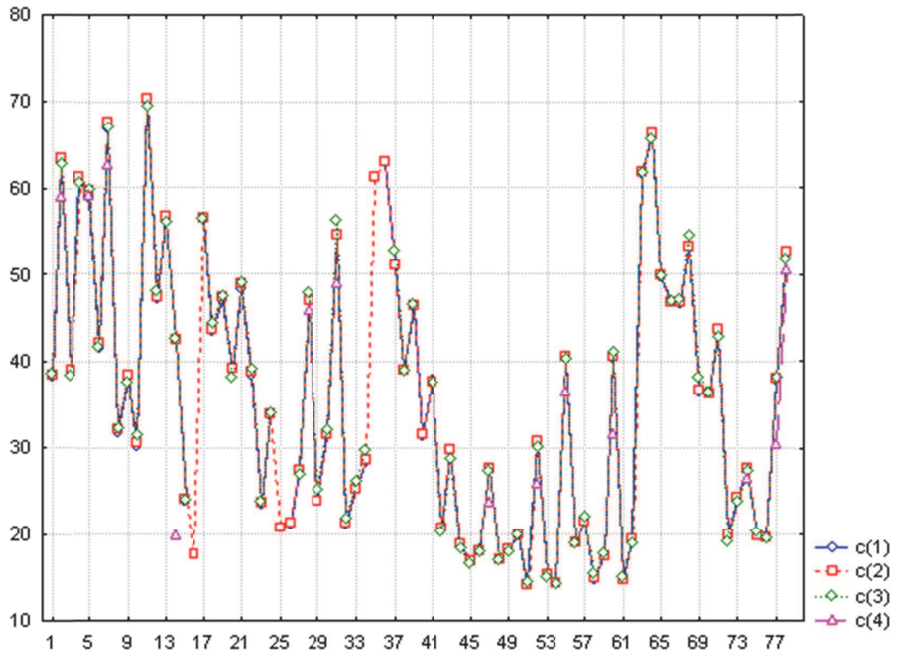
Rovnako sa ukázala pomerne vysoká zhoda v kvalite nájdených pravidiel, čo sa týka hodnôt charakteristiky spoľahlivosti nájdených pravidiel medzi jednotlivými súbormi

(Tabuľka 32). Hodnota Kendalovho koeficientu zhody je 0,61, pričom 1 znamená dokonalú zhodu a 0 znamená nezhodu.

Tabuľka 32 Kendallov koeficient zhody pre mieru spoľahlivosti nájdených pravidiel (c - spoľahlivosť)

	Priemer poradí	Súčet poradí	Priemer	Smerodajná odchýlka
c(1)	2,81	36,5	45,35	13,15
c(2)	3,42	44,5	45,58	13,35
c(3)	2,69	35	45,55	13,37
c(4)	1,08	14	39,97	14,98
Kendallov koeficient zhody				0,6064

Výsledky vizualizuje čiarový graf viacnásobných premenných (Obrázok 41). Graf vizualizuje hodnoty spoľahlivosti nájdených pravidiel v jednotlivých súboroch. Krivky sa kopírujú, čo nám iba potvrdzuje zistenú zhodu v hodnotách spoľahlivosti medzi jednotlivými súbormi.



Obrázok 41 Čiarový graf viacnásobných premenných pre mieru spoľahlivosti nájdených pravidiel

Z korelačnej matici vidíme (Tabuľka 33), že najväčšia miera zhody/závislosti v spoľahlivosti je medzi pravidlami nájdenými v súbore hrubé dáta (Súbor 1), bez robotov (Súbor 2) a bez

robotov a NAT/proxy zariadení (Súbor 3). Naopak menšia, ale tiež významná ($p < 0,05$) je medzi súborom s identifikáciou sedení (Súbor 4) a ostatnými súbormi.

Tabuľka 33 Kendallove koeficienty Tau pre mieru spoľahlivosti nájdených pravidiel

	c(1)	**c(2)**	**c(3)**	**c(4)**
c(1)	1	0,995	0,974	0,744
c(2)	0,995	1	0,973	0,769
c(3)	0,974	0,973	1	0,718
c(4)	0,744	0,769	0,718	1

Očistenie dát od prístupov robotov nemá vplyv na kvalitu extrahovaných pravidiel, v zmysle ich základných charakteristík kvality. Taktiež sa preukázalo, že očistenie dát od prístupov NAT/proxy zariadení nemá významný vplyv na kvalitu extrahovaných pravidiel. Naopak v prípade identifikácii sedení sa zistili najväčšie rozdiely v základných charakteristikách kvality nájdených pravidiel oproti ostatným úrovniam prípravy dát pre web log mining.

4.7.4 Záver

Experiment odhalil viaceré dôležité skutočnosti. Zaujímavosťou je, že najviac pravidiel bolo objavených práve po odstránení záznamov robotov rôznych vyhľadávacích služieb. Ukázalo sa, že rekurzívny charakter prehľadávania portálu robotmi môže skresliť výsledky. Napriek tomu však pri príprave dát pre web log mining nemalo vylúčenie robotov vyhľadávačov na výsledky sekvenčnej analýzy výrazný vplyv. Rovnako na výsledky nemalo výrazný vplyv ani vylúčenie NAT/proxy zariadení. Naopak, výrazný vplyv na spresnenie výsledkov analýzy mala identifikácia sedení používateľov na základe času. Práve identifikácia sedení sa javí ako najdôležitejší faktor celej prípravy dát.

Samozrejme, uvedený experiment môže mať ďalšie modifikácie. Sedenia sa dajú spresniť aj tak, že z jednej IP adresy môžu pristupovať ľudia s rôznym agentom (prehliadačom). Tento faktor taktiež môže spresniť výsledky analýzy. Jednou z metód identifikácie používateľov (skrývajúcich sa za rôzne NAT zariadenia alebo proxy server) je ich rozlíšenie na základe požitého webového prehliadača, t.j. záznamy z rovnakej IP adresy je možné užšie rozdeliť do jednotlivých sedení podľa použitého prehliadača. Týmto prístupom môžeme špecifikovať aj sedenia používateľov z internetových kaviarni, počítačových učební a pod., kde sa za jedným počítačom vystrieda viacero používateľov a predpokladáme, že nie všetci používajú rovnaký webový prehliadač.