УНИВЕРЗИТЕТ „ГОЦЕ ДЕЛЧЕВ“ - ШТИП

ФАКУЛТЕТ ЗА ИНФОРМАТИКА

ШТИП

МАРЈАН ВЕЛКОСКИ

АНАЛИЗА НА ВЕБ ЛОГОВИ СО ПРИМЕНА НА ТЕХНИКИ ЗА ПОДАТОЧНО РУДАРЕЊЕ

- МАГИСТЕРСКИ ТРУД -

Штип, декември 2013 година

Комисија за оценка и одбрана

Ментор: **Цвета Мартиновска-Банде**
Вон. проф. д-р, Факултет за информатика
Универзитет „Гоце Делчев“ - Штип

Член: **Сашо Коцевски**
Доц. д-р, Факултет за информатика
Универзитет „Гоце Делчев“ - Штип

Член: **Татјана Атанасова-Пачемска**
Вон. проф. д-р, Факултет за информатика
Универзитет „Гоце Делчев“ - Штип

Членови на комисија за оценка и одбрана

Претседател: **Сашо Коцевски**
Доц. д-р, Факултет за информатика
Универзитет „Гоце Делчев“ - Штип

Член: **Цвета Мартиновска-Банде, ментор**
Вон. проф. д-р, Факултет за информатика
Универзитет „Гоце Делчев“ - Штип

Член: **Татјана Атанасова-Пачемска**
Вон. проф. д-р, Факултет за информатика
Универзитет „Гоце Делчев“ - Штип

Научно поле: Информатика
Научна област: Информациони системи и мрежи
Дата на одбрана:_____
Дата на промоција:_____

БЛАГОДАРНОСТ И ПОСВЕТА

Посебна благодарност до моите родители Љубомир и Ана, кои секогаш ми даваа безрезервна поддршка за да чекорам понатаму ...

Овој магистерски труд го посветувам на мојата сопруга Славица и децата Јана и Марио. Деновите се малку за да се опише големината на нивното значење ... Љубовта од сопругата и семејството е најголемиот благослов во животот.

Публикувани трудови

Marjan Velkoski and Cveta Martinovska-Bande, Analyzing Web Server Access Log Files Using Data Mining Techniques, International Conference on Applied Internet and Information Technologies, 2013

October 25, 2013, Zrenjanin, Serbia, http://www.tfzr.uns.ac.rs/aiit/

НАСЛОВ НА ТРУДОТ

АНАЛИЗА НА ВЕБ ЛОГОВИ СО ПРИМЕНА НА ТЕХНИКИ ЗА ПОДАТОЧНО РУДАРЕЊЕ

Апстракт

Денес WWW(World Wide Web) сервисот не се смета само како мрежа за собирање податоци, купување продукти и добивање услуги, туку како и социјална средина за интеракција и споделување на информациите. Бројот на веб-страниците продолжува да расте и станува се повеќе тешко за корисниците да ги најдат и извлечат корисните информациите. Како решение на овој проблем во последната декада е податочното рударење.

Целта на оваа теза е со помош на техники за податочно рударење да се утврди дали постојат какви било шеми, врски и законитост во податоците за пристап на веб-страниците за веб-локацијата на Секретаријатот за европски прашања. Конкретно, оваа теза истражува дали има какви било разлики во шемите за пристап помеѓу: (1) Посетители од Македонија и посетители надвор од Македонија, (2) Посетители од Секретаријатот за европски прашања и посетители надвор од Секретаријатот за европски прашања, (3) Посетители од Секретаријатот за европски прашања и посетители надвор од Секретаријатот за европски прашања, во Република Македонија.

Во текот на истражувањето беа користени различни техники на податочно рударење како што се класификациски правила, асоцијативни правила, групирање(кластерирање) и атрибутивна селекција и истите ќе бидат користени со четири различни групи на функции. Целокупниот процес на откривање на шемата беше поделен на три главни чекори: (1) Идентификација на трансакција и издвојување на групи на карактеристики (2) Пронаоѓање на шеми за пристап и (3) Анализа на откриени шеми за нивните интереси.

Откривањето на шемите како и нивното значење во текот на работењето на оваа теза сугерира дека техниките за податочно рударење со соодветните групи на функции може да произведе многу интересни шеми.

Клучни зборови: податочно рударење, веб рударење, веб дневници за пристап, кориснички пристапи.

TITLE

ANALYZING WEB SERVER ACCESS LOG FILES USING DATA MINING TECHNIQUES

Abstract

Nowadays web is not only considered as a network for acquiring data, buying products and obtaining services but as a social environment for interaction and information sharing. The number of Web sites continues to grow and becomes increasingly difficult for users to find and extract useful information. As a solution to this problem in the last decade is the data link mining.

The aim of this thesis is using techniques for data mining to determine whether the data on web page access contain any patterns, links and rules concerning the web site of the Secretariat for European Affairs. More specifically, this thesis carries out research as to whether there are any differences between the access patterns between: (1) Visitors from Macedonia and visitors outside of Macedonia; (2) Visitors from the Secretariat for European Affairs and visitors outside of the Secretariat for European Affairs; (3) Visitors from the Secretariat for European Affairs and visitors outside of the Secretariat for European Affairs but within Macedonia.

In the research several techniques of data mining, such as classification rules, association rules, clustering and attributive selection were used with four different groups of function. The overall process of discovering a pattern was divided into three principal steps: (1) Transaction identification and separation of groups of characteristics; (2) Discovering access patters; (3) Analysis of the patterns of interest found.

The discovery of patterns and of their significance during the research that is the subject of this thesis suggests that the data mining techniques, along with the appropriate function groups may result in a wide variety of access patterns.

Key words: data mining, web mining, web access logs, web usage mining.

СОДРЖИНА

ВОВЕД

Континуираниот процес на развојот и дигитализацијата на општеството, автоматското запишување на податоците, текстовите и другите содржини во дигитален формат, поврзувањето на компјутерите и развојот на WWW сервисот придонесуваат за постојан пораст на базите на податоци. Истовремено, со растењето на базите на податоци се појавува и потребата за нивна анализа и визуализација, сè со цел да се дојде до корисни податоци, информации и знаења. Информатичката технологија дава можност да се управува со овој огромен обем на податоци и дава можност да се пронајдат вистински податоци и информации кои претставуваат база за донесување на успешни работни одлуки. Без разлика за која област и подрачје се работи, стандардниот пристап на анализа на податоци се базира врз работа на аналитичарот кој ги обработува податоците со примена на компјутерски програми или без нив. Во рамките на развојот на информатичките системи, како што е спомнато и погоре количината на податоците многу брзо расте, па оттука каква било обработка на податоците без употреба на компјутер и современи алгоритми и техники на обработка е потполно неефикасна и практично невозможна. Оваа современа анализа е предуслов за донесување квалитетни одлуки во модерното работење, а исто така и за квалитетна научно-истражувачка работа. Резултатот од анализата на податоци е откривање на некое ново корисно знаење кое е скриено во внатрешноста на овој длабок океан од податоци.

Откривањето знаење подразбира целокупен процес на откривање на нови корисни знаења од податоците, додека податочното рударење е техника која е само еден чекор во тој процес. *Податочното рударење* е методологија за извлекување на знаење од податоци и се чини дека претставува единствено решение за овој растечки проблем. Со зголемената популарност на World Wide Web сервисот автоматски се собираат огромни количини на податоци, како што се корисничките адреси или URL барањата од веб-серверите до кои се пристапува во таканаречени датотеки за пристап (access log files). Откривањето на односите и шемите на однесување кои постојат во овие датотеки може да обезбеди значајни и корисни информации за подобрување

на ефикасноста во целокупното работење, преструктуирање на веб-локацијата, како и на клиентите кои се фокусирани во електронската трговија.

Веб-локацијата на Секретаријатот за европски прашања содржи информации за процесот и подготовката на Република Македонија за членство во Европската унија. Таа содржи многу веб-страници, кои обезбедуваат информации за претпристапна помош, обуки, бази на податоци, регистрација на проекти и информации за процесот на преведување.

При посетата сите посетители зад себе оставаат траги од нивната посета и од однесувањата во облик на дневници кои се чуваат во веб-серверите. Овие дневници како што беше спомнато, всушност, се документи кои автоматски се креираат и се нарекуваат веб-дневници за пристап (web access log). *Рударење на кориснички пристапи (Web Usage Mining)* е процес на примена на техники за податочно рударење за откривање на користените податоци(шеми) од веб-дневниците кои ги идентификуваат однесувањата на веб-корисниците [14]. Целиот овој процес вклучува неколку чекори: предпроцесирање(*Preprocessing),* откривање на шема*(Pattern Discovery)* и анализа на шема(*Pattern Analysis)*.

Во зависност од потребите може да се изберат различни техники за податочно рударење. Едни од најчесто користените техники се асоцијативни правила, класификација, кластерирање и избор на атрибути. Секои од овие техники податоците ги анализираат на различни начини и со одредени предности и недостатоци, кои треба да се земат предвид пред нивната употреба.

Целта на овој магистерски труд е со помош на техниките за податочно рударење да се откријат некои интересни шеми за посетителите на веб-страницата за интеграција на Македонија во Европска унија на Секретаријатот за европски прашања. Акцентот на оваа теза е ставен на откривањето дали има било какви разлики во шемите за пристап помеѓу:

1. Посетителите од Македонија и посетителите надвор од Македонија.
2. Посетители од Секретаријатот за европски прашања и посетителите надвор од Секретаријатот за европски прашања.
3. Посетителите од Секретаријатот за европски прашања и посетители

надвор од Секретаријатот за европски прашања, во Република Македонија.

Во оваа магистерска работа се користат следните алгоритми за податочно рударење на веб-дневници:

- алгоритмите за класификација 1R и J48.
- Алгоритам за асоцијативни правила Apriori
- Алгоритам за кластерирање - EM
- Алгоритам за селекција на атрибути, корелација базирана на евалуатор на изборот на карактеристики за подгрупа.

Податоците кои се користени при анализата се добиени од пристапните дневници од 01 до 31 декември, 2012 год.

1. ТЕХНИКИ И АЛАТКИ ЗА ПОДАТОЧНО РУДАРЕЊЕ

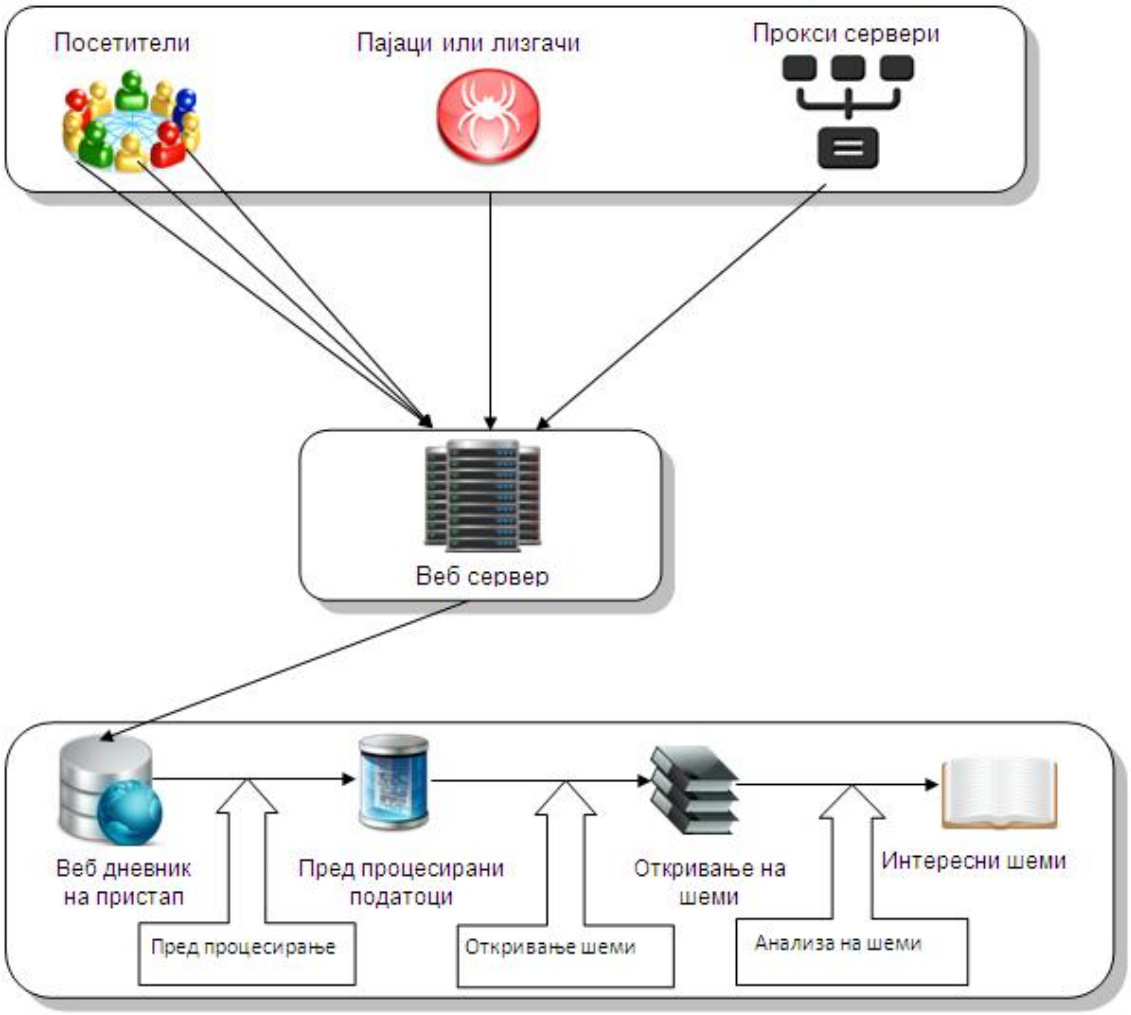
Во ова поглавје подетално ќе се дискутира за целокупниот процес на податочно рударење на податоци. Во оддел 1.1 се дискутира за фазите и техниките на податочното рударење и софтверската алатка за рударење на податоци кои се користат во магистерскиот труд. За видовите на веб-рударење и примената на техниките за рударење врз веб-податоците се дискутира во оддел 2.1, одделот 2.2 ги опфаќа деталите за веб-дневниците од серверот и чекорите за претпроцесирање од примената на веб-рударењето. Одделот 3.1 ги опишува сите истражувања и препораки поврзани со рударење на податоци од веб-дневници од страна на истражувачи кои ја обработуваат оваа проблематика.

1.1 Податочно рударење

Податочното рударење е нова методологија и станува сè популарна поради нејзината способност на автоматски или полуавтоматски начин да наоѓања корисни информации, разбирливи шеми на податоци, знаење и трендови од огромната база на податоци, односно онаму каде што е многу тешко и напорно да се применат традиционалните техники на знаење и човечкото сфаќање. Податочното рударење користи софистицирани математички алгоритми за сегментот на податоци и оценка на веројатноста на идните настани. Податочното рударење е, исто така, познато како откривање на знаење од базите на податоци (Knowledge Discovery in Databases – KDD). Главни својства на податочно рударење се: автоматско откривање на шаблони, предвидување на можни резултати, креирање на вистинити резултати и се фокусира на големи збирки на податоци и бази на податоци. Податочното рударење може да одговори на прашања кои не можат да се решат преку едноставни пребарувања и техники на извештаи.

Слика 1.1 претставува краток опис за главната задача на процесот на рударење на кориснички пристапи. Првиот чекор, предпроцесирање е задача во која главно влегува чистење на податоците, идентификација на корисникот, идентификација на сесија и завршување на патот на посетителот. Ова не е ни

малку лесна задача бидејќи има кеширање на страници и пристапи од страна на веб-роботите. Вториот чекор, откривање на шема, е процес во кој се инволвираат алгоритми и техники на податочно рударење на претпроцесираните податоци со цел да се откријат интересни шеми и третиот чекор, анализа на шема, се однесува на анализа на откриени шеми од вториот чекор за процена на нивните интереси.



Слика 1.1 Чекори на процесот рударење на кориснички пристапи
Figure 1.1 Steps of WebUsage Mining Process

Целокупниот процес на податочно рударење може да се подели во 4 фази и тоа: *собирање на податоци, предпроцесирање(подготовка) на податоци, откривање на шеми и анализа на откриените шеми.*

Собирање на податоци е фаза во која се врши собирање на податоци и идентификување на карактеристиките за кои се смета дека ќе бидат корисни за податочно рударење. Бидејќи податоците се собираат од разни извори тие се

разновидни и истите треба да се подготват, односно да се нормализираат. Тој процес на нормализација се врши со трансформација од една структура во друга и на тој начин тие стануваат погодни за користење. Откако се откриени карактеристиките на шемите од претходниот процес и истите се конечни, се врши форматирање на истите во облик погоден за податочно рударење во зависност од алатката која ќе се користи.

1.1.1 Техники на рударење и откривање на шема

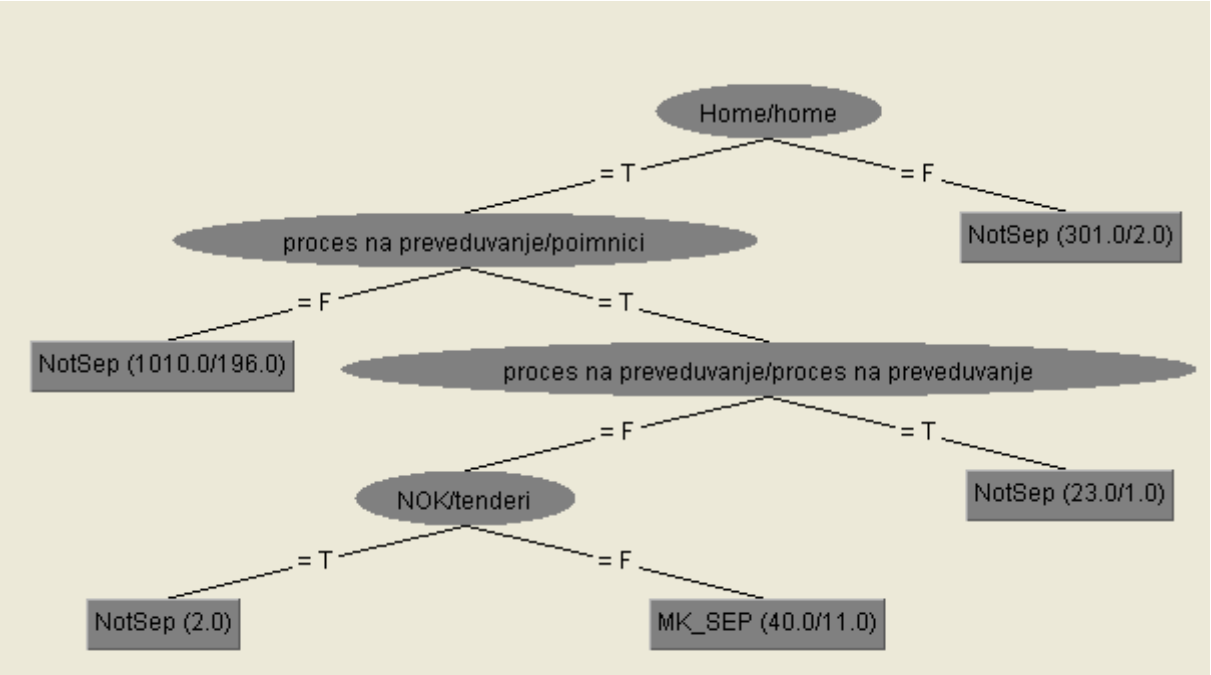
За евалуација на експериментите на овој труд се користи алатката WEKA која е кратенка од Waiakato Enviroment for Knowledge Analysis, односно (Waiakato) околина за анализа на знаењето [28]. За упешно да се примени оваа алатка потребно е податоците да бидат подготвени и соодветно форматирани за да може да се откријат интересни шеми. WEKA е колекција од алгоритми за машинско учење за решавање проблеми од податочно рударење. WEKA е софтвер со отворен код издаден според Општа јавна лиценца/GNU General Public License. WEKA содржи алатки за податочно рударење и тоа: претпроцесирање, класификација, групирање, селекција на атрибути и визуелизација. Техниките за податочно рударење кои се користат за овој труд се опишани подолу.

1. **Класификација:**

   Класификација(classification) е техника на распределување на објекти врз основа на пронаоѓање сличност на објектот со однапред предефинирани класи или категории. Сличноста на двата објекта се одредува со анализа на нивните карактеристики кои најдобро ги опишуваат особините на дадената класа.

   Постојат повеќе техники на класификација и една од моќните и најпопуларните т.е *дрво на одлука(*Decision Trees). *Дрво на одлука* е структура што претставува класификациски алгоритам во форма на стеблеста структура во кој се разликуваат два типа на јазли поврзани со гранки. Во *дрво на одлука*, јазлите претставуваат јазол на одлука(“decision node”) и ја тестираат вредноста на атрибутот, а листовите ги претставуваат крајни јазли (“Leaf node”) кои претставуваат предефинирани класи на кои припаѓаат атрибутите. Слика 1.2 покажува *дрво на одлука* од 5 листови и 4 јазли на одлука. Еден дел на ова *дрво*

на одлука може да се толкува како: ако посетители ја посетат страницата „Home/home“, а исто така ја посетат страницата „Proces na preveduvanje/poimnici“, тогаш тие се посетители кои се занимаваат со превод или посетители кои потенцијално ќе се занимаваат со превод.



Слика 1.2 Пример за Дрво на Одлука
Figure 1.2 An Example of a Decision Tree

Класификацијата е индуктивна операција која го сместува секој атрибут(случај, инстанца) од множеството податоци во една од специфичните однапред одредени(познати) класи, врз основа на карактеристиките на атрибутот. Во групата на податоци атрибутот што претставува класа се нарекува променлива класа.

При процесот на класификација најпрво се спроведува фаза на обука, а потоа се врши класификација на атрибутите. Фазата на обука се спроведува над множеството податоци за обука кое е одбрано. Сите податоци кои припаѓаат во тоа множество имаат позната класификација, односно припаѓаат во некоја класа. Оваа фаза претставува техника на надгледувано учење. Еден од можните проблеми кој може да се појави е разликата во точно класифицираните случаи на податоци за обуки и тестирање и е познат како overfitting(пренаучување). Со цел да се избегне

пренаучување, потребно е да се користи дополнителна техника cross-validation (вкрстено потврдување). Техниката на cross-validation (вкрстено потврдување), која се базира на „повторно земање мостри“(\re-sampling") [31], се користи за проценка на стапката на грешка(error rate). Во k-fold вкрстеното потврдување, податоците се поделени во k подгрупи со еднаква големина. Моделот е обучен k пати, со различни подгрупи кои се испуштени од групата за обука секој пат. Точноста се мери само за испуштената подгрупа и стапката на грешка се пресметува како просекот од стапките на грешка од к извршувања.

За потребите на експерименти на оваа теза ќе се користат алогритми за класификација „1R“ и „J48“.Овие алгоритми и нивните параметри се опишани подолу.

(а) **1R:** Алгоритамот генерира дрво на одлучување со едно ниво изразено во форма на серија на правила кои сите тестираат еден одреден атрибут. Но, исто така е забележано дека правилата кои работат само со еден атрибут работат зачудувачки добро [11]. Идејата е следна: правиме правила кои тестираат еден атрибут и соодветна гранка и секоја гранка одговара на различна вредност на атрибутот. Се оценува стапката на грешка за секое правило на атрибут и се избира најдоброто, односно се избира оној кој дава најголема точност. Грешката претставува мерка за примерите кои не припаѓаат на побројните класи во гранките на стеблото. Како што е споменато во [11], 1R често ги класифицира случаите доста точно и наоѓа добри правила за карактеризирање на структурата на податоците.

Во Табела 1.1 се покажани „временски“ групи на податоци кои се однесуваат на условите по кои може да се игра некоја неодредена игра [11]. Атрибути во оваа група на податоци се: време, температура, влажност и ветровитост. Сите овие атрибути имаат номинални вредности. „Play“ е променлива класа која одговара на состојбата за играње и чија вредност може да биде „да“ или „не“.

Слика 1.3 покажува делумен резулат од програмот WEKA со 1R алгоритамот со користење на „временската“ група на податоци. Како што може да се забележи првата клаузула на правилото може да се толкува како: ако *времето* е „сончево“ тогаш *Play* = „не“. Втората и третата клаузула на правилото може да се толкува како: ако времето е „дождливо“ или „облачно“ тогаш *Play* = „да“.

Табела 1.1 Пример за време
Table 1.1 A Sample of the Weather Data

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes

```
===          1R model (full training set) ===

outlook

        sunny      -> no
        overcast   -> yes
        rainy      -> yes
```

Слика 1.3 *Делумен излез од WEKA 1R програмата со користење на „временската“ група на податоци.*
Figure 1.3 Partial Output Produced by the WEKA 1R Program Using the Weather Data

(б) **J48:** J48 е класификатор кој е настанат врз основа на популарниот алгоритам за изградба на стебло на одлуки C4.5 и претставува негово подобрување како што е споменато во [11]. C4.5 е метод кој најмногу се користи за машинско учење [11]. J48 алгоритмот работи со номинални и нумерички атрибути. Овие алгоритми главно се состојат од две концептуални фази: растење и чистење (кроење). Фазата чистење, односно кроење на дрвото се однесува на поедноставување на дрвото на одлука и зголемување на точноста на предвидување. Кроење на дрвото се постигнува со отстранување на едно или повеќе поддрва и нивна замена со листови. Поддрво во

дадениот јазол се чисти со отстранување на неговите гранки и замена со лист кој се обележува како најчеста класа во поддрвото кое се заменува. Со замената на поддрвото со лист, алгоритамот очекува понизок процент на грешка при предвидувањето и зголемување на квалитетот на класификациското дрво. Со ова се овозможува побрза класификација и подобрување на точноста на конкретното класифицирање на податоци за тестирање [13]. Постојат две основни стратегии за смалување:

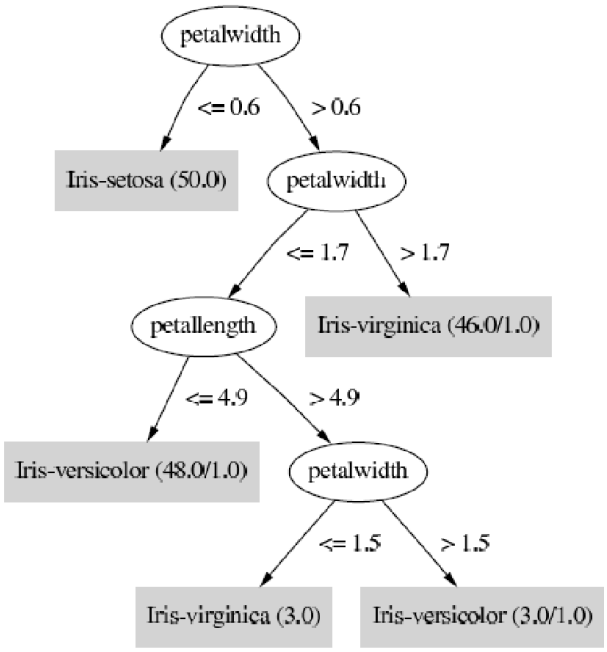
1. Постсмалување или смалување наназад(Post-pruning). Оваа метода најпрво го составува целото стебло, а потоа дополнително го смалува. Оваа метода во пракса најчесто се користи. Во пралтиката постојат два различни методи:
   - Замена на подстебло (Subtree replacement)
   - Подигнување на подстебло (Subtree raising)
2. Априори смалување или смалување нанапред *(Pre-pruning).* Оваа метода во текот на градењето на стеблото бара одлука за стопирање на процесот на градење потстебла, а тоа е многу корисно бидејќи се заштедува време.

Најчесто вообичаената вредност на *доверба* во J48 се зема да е 0,25. Доколку ова вредност се намали тоа значи повеќе кроење на дрвото. Вториот важен параметар е *минимален број на предмети* чија задача е да се елиминираат тестовите за кои бројот на случаи е помал од вредноста на групата за овој параметар.

Табела 1.2 покажува примерок на групата на ирис податоци [11]. Како што може да се види од оваа табела, четирите атрибути во групата на податоци се *должина на ливчето, ширина на ливчето, должина на цветното ливче* и *ширина на цветното ливче*. Сите атрибути имаат нумерички вредности. *„Класа“* е променлива класа која одговара на видовите на растенија.

Табела 1.2 Примерок на Ирис податоци
Table 1.2 A Sample of the Iris Data

SepalLength	SepalWidth	PetalLength	PetalWidth	Class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica
7.1	3.0	5.9	2.1	Iris-virginica
6.3	2.9	5.6	1.8	Iris-virginica



Слика 1.4 Дрво на одлуки изработено од WEKA J48 програмата со користење на Ирис податоци
Figure 1.4 Decision Tree Produced by the WEKA J48 Program Using the Iris Data

2. **Правила на асоцијација**

Асоцијативните правила(association rules) ја наоѓаат корелацијата помеѓу предметите во големите збирки на податоци и фреквенцијата на појавување парови од одредени предмети. Нека R1 е правило(релација) каде што:

R1=„Елементот А се појавува заедно со елементот Б во 10 % случаи“
Десет проценти (десет насто) е мерка на фреквенција на појавување на парот елементи А и Б, односно ако се појавил елементот А, тогаш се појавил и елементот Б и претставува *прецизност(confidence). Прецизност* е бројот на случаи во кои парот елементи А и Б се појавуваат по правило, изразени како процент од вкупниот број на случаи во кои се појавил елементот А. *Поддршка(support)* е мерка за тоа колку често елементите А и Б се појавуваат заедно во однос на сите случаи.

Рударењето на податоци со асоцијативни правила се состои од пребарување на групи од објекти кои се појавуваат заедно во одреден контекст. Бидејќи може да постојат многу корелации помеѓу објектите, прецизноста и поддршката даваат голем придонес и помош во одлуката на аналитичарите кои правила ќе ги задржат а кои ќе ги отстранат.

Асоцијативните (здружените) правила навистина не се разликуваат од правилата на класификација, освен дека тие може да го предвидат секој атрибут, не само на класата, и тоа им дава слобода да се предвидат комбинации на атрибути. Исто така, поврзаните правила не се наменети да се користат заедно како група, како правила за класификација. Различни поврзани правила изразуваат различни законитости коишто придонесуваат за базата на податоци, а тие обично предвидуваат различни работи. Бидејќи може да се изведат многу различни поврзани правила може од мала база на податоци, интересот е ограничен на оние кои се однесуваат на разумно голем број на случаи и имаат разумно висока точност на случаи на кои тие се однесуваат. Опфатот на поврзани правила е бројот на случаи за кои тие се предвидени точно - ова често се нарекува *поддршка*. Неговата *точност*, често наречен *доверба* е бројот на случаи што се предвидени правилно, изразени како процент од сите случаи на кои се однесува. За еден елемент велиме дека е чест ако неговата фреквенција е поголема од вредноста на *довербата.* Алгоритамот априори применува итеративен пристап во кои к-групи на елементи се користат за да се испитаат (к+1)-групи на податоци. За да се подобри ефикасноста на генерирање чести групи на елементи се користи важна особина наречена априори во која се наведува дека доколку некој елемент не се појавува често тогаш сите негови супергрупи исто така

не се честа појава и тие може да се отфрлат. Параметрите *доверба* и *поддршка* помагаат да се добијат правилата за асоцијација за бараната заинтересираност и истите треба да се обезбедени од страна на корисникот. Алгоритмот априори работи само со номинални атрибути.

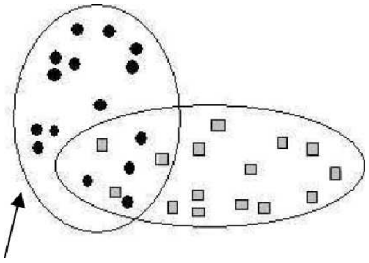
Правила на асоцијација кои се генерирани од WEKA-Априори програмата со користење на „временската“ група на податоци се прикажани на Слика 1.5. Генерирани се вкупно четири правила. Првото правило може да се толкува како: ако *влажноста* е „нормална“ и *ветер* „нема“ тогаш *Play* = „да“(играј ја играта). Довербата на ова правило е 1.

```
Minimum support: 0.4
Minimum confidence: 0.9

Best rules found:

 1. humidity=normal windy=false ==> play=yes conf:(1)
 2. temperature=cool ==> humidity=normal conf:(1)
 3. outlook=overcast ==> play=yes conf:(1)
 4. temperature=cool play=yes ==> humidity=normal conf:(1)
```

Слика 1.5 Делумен излез на WEKA-Априори програмата со користење на „временски“ податоци.
Figure 1.5 Partial Output of the WEKA Apriori Program Using the Weather Data

3. **Групирање (кластерирање):**

Групирање(clusters) е дескриптивна техника која ги групира податочните елементи кои имаат слични карактеристики или особини заедно и ги одвојува различните елементи по групи. Овде групите не се познати однапред. Еден пример за групирањето на веб-посетителите е да се определат групи на посетители кои покажуваат слично однесување во прелистувањето и врз основа на тоа откритие може да се обезбедат соодветни информации за посетителите. По формирањето на групите, непознатите случаи може да се распределат во една или повеќе соодветни групи.



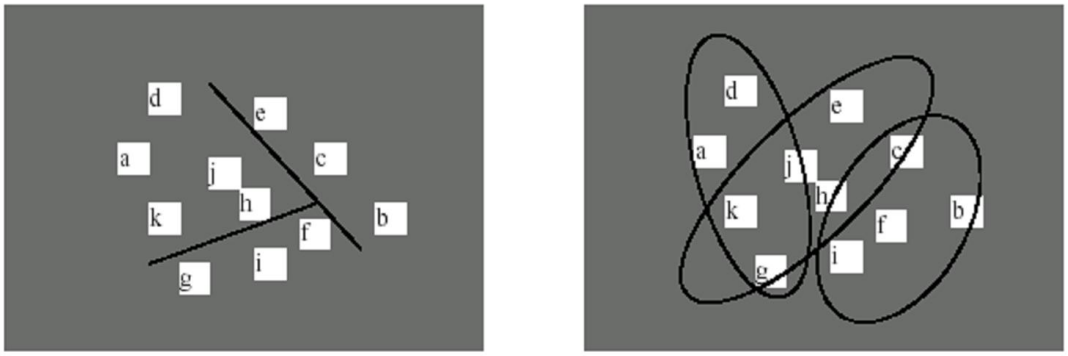
„Процес на преведување“ група „Поимници“ група

Слика 1.6 *Пример на групирање*
Figure 1.6 An Example of Clustering

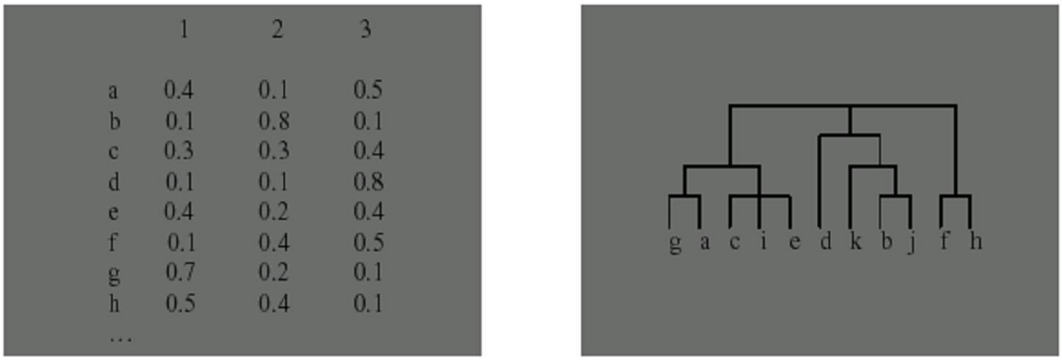
Сликата 1.6 покажува дека се формирани две групи. Од сликата може да се забележи дека, „Поимници“ е група од посетители кои имаат тенденција да прелистуваат информации за процесот на преведување. Се забележува дека некои од посетителите им припаѓаат и на двете групи и оттаму овие групи се преклопуваат.

Кластерирање е статистичка операција на групирање објекти во ограничен број на групи-кластери чиј број не е однапред познат. Кластерите се комбинација од објекти кои имаат слични карактеристики и основа на кластерирањето е диструбуција на објектите во групи. Бидејќи не постои однапред знаење за класите, овој процес е дескриптивен и претставува техника на ненагледно учење. Алгоритмите за кластерирање тежнееат да ги поделат групите или класите на објекти во подгрупи во кои сличноста помеѓу објектите во една подгрупа е максимална, а сличноста меѓу групите е минимизирана [13]. Кластерите формирани на крајот на процесот може да бидат веројатни, а тоа значи дека објектот припаѓа на одреден кластер со одредена веројатност, како што е опишано во [11]. Некои алгоритми на групирање овозможуваат некој пример да спаѓа во повеќе групи, додека други поврзуваат примери со групите преку функциите на веројатност. За секој пример постои бројчен износ кој одговара на веројатноста тој пример да припаѓа на одредена група. Други, пак, алгоритми градат хиерархиска структура на групи, така што на најголемо ниво просторот на примерот поделен на два подгрупи кои поединечно се делат на две подгрупи и така

натаму. После групирањето често се прават стебла на одлуки или збир на правила кои го доделуваат секој пример на групата на која и припаѓа.



Слика 1.7 Првата слика покажува како примерите се групираат во групи. Втората слика ја прикажува можноста некој пример да припаѓа во повеќе групи
Figure 1.7 The first picture shows how the examples are grouped in groups. Another photo shows one such possibility to belong to several groups



Слика 1.8 Првата слика го прикажува поврзувањето на примерите со групите преку функции на веројатност. Втората слика покажува добивање на хиерархиска структура на групи
Figure 1.8 The first picture shows the connecting groups through examples of probability functions. Another photo shows obtaining hierarchical structure of groups

Методите на кластерирање можат да се поделат на неколку методи: партиционирање, хиерархиски и методи засновани на густина, мрежа и модели.

Алгоритамот EM/OM-максимални очекувања (eng. Expectation maximization) е претставник на метод заснован на модели. Предложен е со цел да се зголеми веројатноста на членство во кластерот [9]. Во првиот чекор, алгоритмот EM пресметува веројатност на членство во кластерот, а во вториот чекор ги максимизира овие веројатности, како што е спомнато од [11].

Комплексноста на EM алгоритамот е линеарна со бројот на објекти кои се анализираат, бројот на влезни карактеристики и бројот на итерации.

На Слика 1.9 е прикажан излез на кластер формиран од страна на EM алгоритам со користење на „временски“ податоци. Алгоритамот EM генерира излез со пресметување *износ* за секој номинален атрибут. *Износот* е бројот на секоја посебна вредност на атрибут нормализиран во веројатност [11]. Атрибутот време може да се толкува како: случаи кои припаѓаат на овој кластер кои имаат 6% (од вкупно 17%) веројатност, нивната вредност на атрибутот *изглед (време)* е *сончево*. Слично и случаите кои припаѓаат на овој кластер кои имаат 5% (од вкупно 17%) веројатност, нивната вредност на атрибутот *изглед (време)* е *облачно*. На сличен начин, исто така, може да се толкуваат и излезите за од другите атрибути. На слика 1.9 е прикажан еден кластер на случаи во кои е прикажана веројатноста за секоја вредност на секој атрибут.

Attribute: outlook		(sunny)	(overcast)	(rainy)	
Discrete Estimator.	Counts	= 6	5	6	(Total = 17)
Attribute: temperature		(hot)	(mild)	(cool)	
Discrete Estimator.	Counts	= 5	7	5	(Total = 17)
Attribute: humidity		(high)	(normal)	(low)	
Discrete Estimator.	Counts	= 1	8	8	(Total = 17)
Attribute: windy		(false)	(true)		
Discrete Estimator.	Counts	= 9	7		(Total = 16)
Attribute: play		(no)	(yes)		
Discrete Estimator.	Counts	= 6	10		(Total = 16)

Слика 1.9 Забележан делумен излез на WEKA EM програмата со користење на „временски“ податоци

Figure 1.9 Annotated Partial Output of the WEKA EM Program Using the Weather Data

4. **Селекција на атрибут(Attribute selection):**

Селекција на атрибути*(Attribute Selection)* е процес на селекција на најрелевантни и најдискриминирачки атрибути од групата на податоци. По идентификацијата на релевантните атрибути, техниките за податочно рударење може да се применат на овие идентификувани атрибути се со цел да се подобрат перформансите и резултатите. Затоа, процесот на селекција на атрибути пожелно е да се користи пред примена на други техники за податочно рударење, како што е техниката дрво на одлука [11].

Селекција на атрибут претставува процес на отстранување на непотребни атрибути кои се сметаат за ирелевантни за задачата на податочно рударење, а воедно се идентификуваат повеќе релевантни и дискриминирачки атрибути. Како и да е, присуството на атрибути кои не се корисни за класификација може да се мешаат со релевантните атрибути, а со тоа да ги деградира перформансите на класификација [11]. Селекцијата на атрибут ја подобрува ефикасноста на алгоритмите за податочно рударење со бришење на несоодветни атрибути [11]. За извршување на овој процес потребно е да избереме проценител на атрибути и метод на пребарување подгрупи. За оценување подгрупи на атрибути кои се во корелација меѓу себе се избира *CfsSubsetEval* проценител на атрибути [11]. *CfsSubsetEval* проценителот ги избира атрибутите кои се во корелација со класната променлива и имаат помалку интеркорелација помеѓу нив, како што е спомнато во [11]. Како што е спомнато погоре во процесот на селекција на атрибут, покрај проценителот на атрибути се бара метод за пребарување. Негова задача е да пребарува низ можните подгрупи на атрибути кои најдобро ја предвидуваат класата. *Best-first* е метод на пребарување што чува список на подгрупи кои се оценети по редослед на нивните перформанси, како што е опишано во [11]. Оттука, кога изведувањето ќе почне да соодветствува за подгрупа, пребарувањето може да се наврати на претходната подгрупа.

На слика 1.10 е прикажан примерок на излез изработен од CfsSubsetEval проценителот на атрибути со методот на пребарување Best-first со користење на ирис податоци. Може да се види дека, *должината на*

цветниот лист и *ширината на цветниот лист* се двата атрибути кои се избрани како најрелевантни и дискриминирачки.

```
Search Method:
        Best first.
        Total number of subsets evaluated: 12

Attribute Subset Evaluator (supervised, Class (nominal): 5 class):
        CFS Subset Evaluator

Selected attributes: 3,4 : 2
                     petallength
                     petalwidth
```

Слика 1.10 Делумен излез од WEKA за селекција на атрибут со користење на „временската“ група на податоци
Figure 1.10 Partial Output Produced by the WEKA for Attribute Selection Using the IrisData

1.1.2 Анализа на шема

Во процесот на рударење на користените веб-податоци како последна фаза е анализата на откриената шема. Мотивацијата позади овој процес на анализа на шемите е проценката на заинтересираноста на откриените шеми и филтрирање на неинтересните шеми на податоци. Точната методологија за анализа зависи од апликацијата за која податочното рударење се спроведува.

Техниките за визуелизација, како што се графички урнеци или распределба на бои за различни вредности, често можат успешно да ги потенцираат целокупните шеми на податоците.

Информациите за содржината и структурата на веб-страницата може да се искористат за пречистување на страници кои содржат одреден тип на содржина.

1.1.3 Алатка за податочно рударење WEKA (Waiakato Enviroment for Knowledge Analysis)

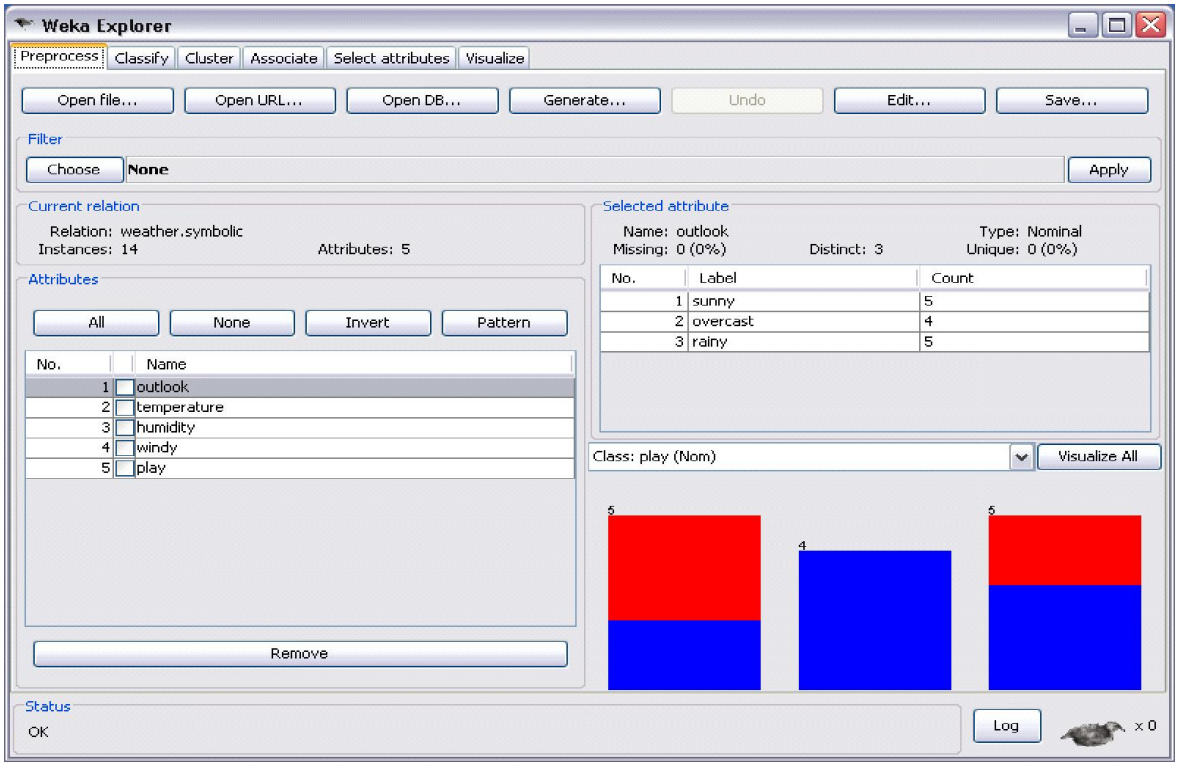
WEKA GUI Chooser прозорот го прикажува графичкиот интерфејс кој содржи три опции:

- Simple CLI, командно линиски интерфејс;
- Explorer, графички интерфејс кој овозможува анализа на податоци со WEKA;

- Experimenter, графички интерфејс кој овозможува спроведување на експерименти и поврзување на статистички тестови помеѓу повеќе шеми за учење.

Explorer графичкиот интерфејс содржи повеќе подменија: Preprocess, Classify, Cluster, Associate, Selecţ attributes, Visualize.

Во подменито Preprocess прво е потребно да се отвори датотека со податоци во arff или csv формат. Прикажани се Base Relation Ii Working Relation кои се идентични под услов да не се користи постапка на филтрирање Apply Filter.



Слика 1.11 Приказ на WEKA Explorer графички интерфејс и неговото подмени Process

Figure 1.11 Showing the WEKA Explorer graphical interface and its sub Process

Во подменито Classify потребно е да се одбере класификатор и тест опција како и априбутот на кој се однесува предикцијата. По почнувањето на класификацијата во Classify output прозорот се прикажани резултатите од предикцијата. Test options box разликува четири различни типови на тестирања:

- *Use training set,* класификаторот се евалуира според тоа како добро врши предикција на класите примери на кои се тренира. Бројот на погрешни класификации на податоци за тренирање не е добар показател на перформансите на нашиот предиктивен модел со идни податоци.
- *Supplied test set,* класификаротот се евалуира според тоа како добро ја врши предикцијата на класата според примерите кои се вчитуваат од датотеката.
- *Cross-validation,* класификаторот се евалуира со помош на вкрстено-потврдување користејќи различен број на предавања кои можат да се специфицираат. Прво целиот податок се поделува на к подгрупи податоци со еднаква далечина. Сите тие групи служат како податоци за тренирање на модел. Тоа се вика k-fold cross-validation. Поделбата не 10 делови се покажува како најдобра.
- *Perscentage split,* класификатор се евалуира според тоа колку добро врши предвидување на податоците кои се задржуваат за тестирање. Количината на тие податоци во вкупниот збир на податоци се одредува со запишување на тие податоци во поле.

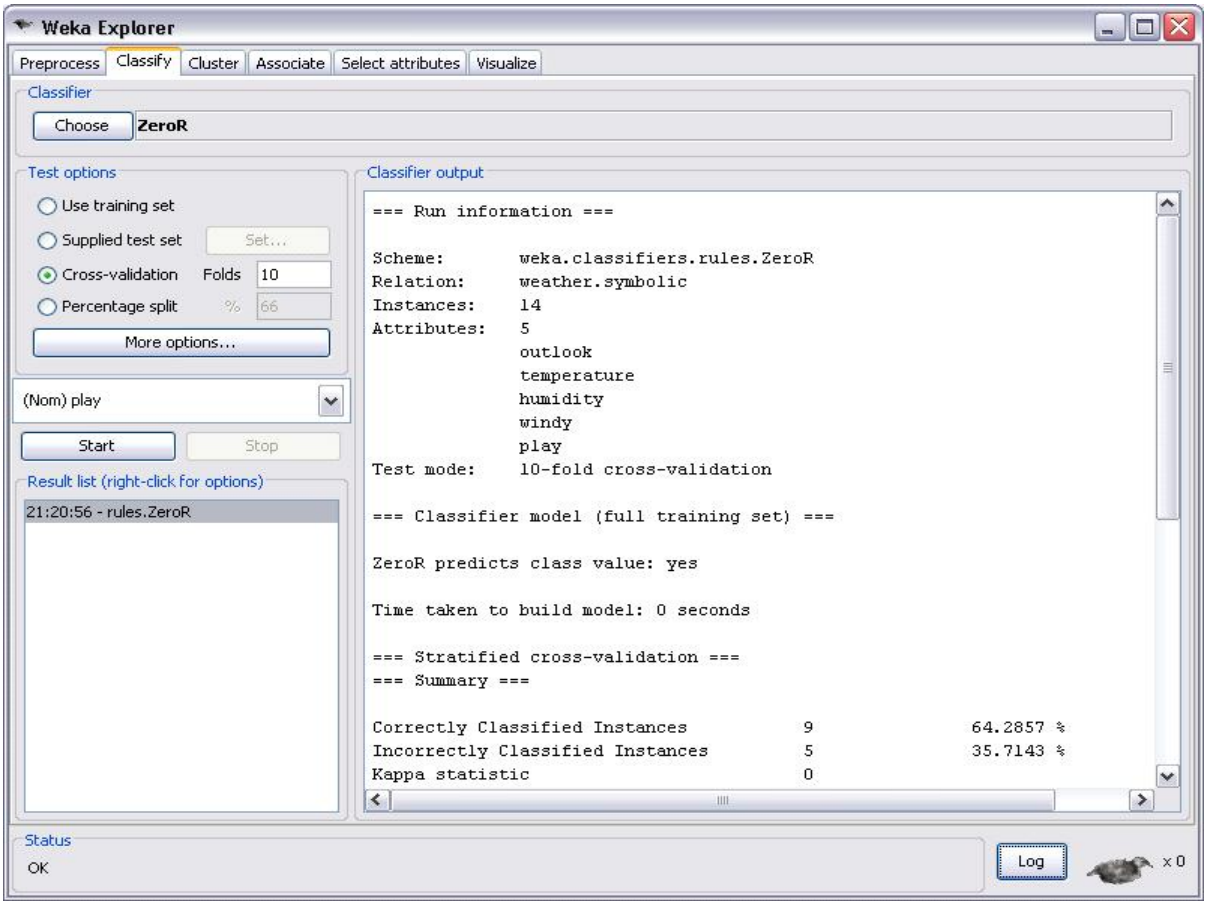
Прикажаните излезни резултати се поделени според секциите:

- *Run information,* што претставува информација за одбраниот класификатор, датотеки со податоци, за релациите, бројот на примери, бројот и имињата на атрибутите и избраните опции на тестот.
- *Classifier model (full training set),* претставува текстуална презентација на класификацискиот модел кој е произведен на целиот сет податоци.
- *Summary,* претствува листа на статистички параметри кои предочуваат колку точно класификаторот успеа да изврши предикција.
- *Detaled Accurasy By Class,* претставува резултат на претпоставената точност на податоците спрема вредностите на класата.
- *Confusion matrix,* прикажува колку примери се доделени на секоја класа. Елементите на матрицата на контигентност/случајност се претставуваат: во редови - број на тестови по класи (class number) и во колони - број на тестови кои се класифицирани како секоја од класите (classified as). Матрицата овозможува увид во тоа колку е добра класификацијата.

Идеалниот класификатор содржи нули секаде освен на дијагоналата на матрицата на контигентност.

Резултатите можат да се подготват или визуализираат така што со десниот клик на маусот да се одбере една од извршените класификации внатре во Result list (листата на резултати).

- *Save result buffer,* опција која овозможува зачувување на резултатите на класификацијата.
- *Visualize classifer errors,* опција која донесува визуализациски прозор која го исцртува резултатот на класификацијата. Исправно класифицираните примери се исцртуваат со крстови, а неисправно класифицираните примери се исцртуваат со квадратчиња.
- *Visualize tree,* опција која графички ја претставува структурата на класификацискиот модел доколку е можна визуализацијата; тоа е можно само кај некои класификатори.
- *Vizalization margin curve,* опција која дава цртеж со означени предикциски маргини. Маргините означуваат разлика помеѓу веројатноста на предикцијата за останатите класи. На пр. Бостинг алгоритамот постигнува подобри резултати на тесниот сет на податоци ако се зголеми маргината на податоците за тренирање.
- *Visualize thereshold curve,* опција која дава цртеж кој илустрира мерење помеѓу предикцијата на некои класи наспрема вредноста на одбраниот праг. На пример, со веќе дефинираната вредност на прагот од 0.5 предикациска веројатност на позитивните мора да биде поголема од 0.5 за примерите кои се предвидени како позитивни. Овој цртеж не може да се употреби во анализа на ROC кривата (TP наспроти FP).



Слика 1.12 Приказ WEKA Explorer графичкото учење и неговото подмени Classify со класификаторот ZeroR и резултати
Figure 1.12 Display WEKA Explorer graphical learning and its submenu Classify the classifier ZeroR and results

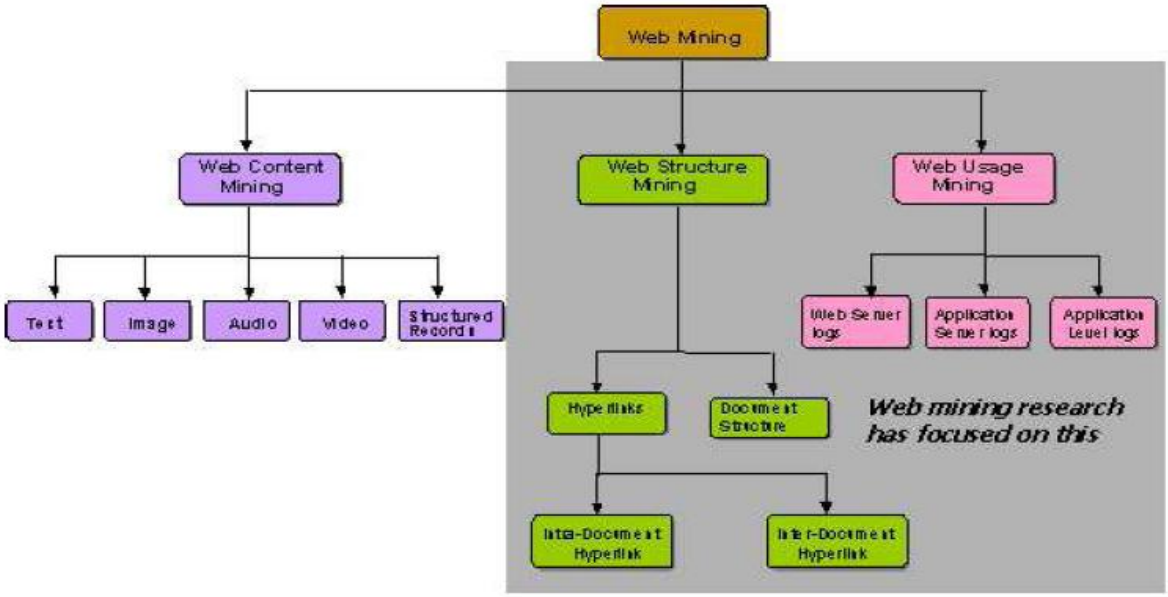
2. Анализа на веб-дневници

2.1 Веб рударење

Доаѓањето на World Wide Web сервисот, предизвика драматичен пораст во користењето на интернетот. World Wide Web е преносен медиум каде што е сместен широк спектар на информации коишто континуирано стануваат огромен извор на информации.

Како последица на ова, се наметнува потребата аналитичарите да го насочат вниманието кон извлекување на корисни информации и знаења со употреба на техниките за податочно рударење. *Веб-рударење (Web mining)* претставува примена на техники за податочно рударење за извлекување на знаења од веб-податоци вклучувајќи веб-документи, хиперлинкови помеѓу документите, користење дневници на веб-локации и слично [12]. Веб-рударењето се користи за разбирање на однесувањето на корисниците, оценување на ефикасноста на веб-страницата и за одредување успешност на некоја компанија.

Врз основа на примарните видови податоци кои се користат во процесот на податочно рударење, веб-рударењето може да се категоризира во три вида: *рударење на структура, рударење на содржина* и *рударење* на корисничките податоци во зависност од тоа кој дел од веб-локацијата се истражува [24].



Слика 2.1 Веб-рударење
Figure 2.1 Web mining

2.1.1 Структурно рударење

Целта на веб структурното рударење е да се открие корисно знаење од хиперлинковите (или линкови на кусо) кои претставуваат структура на веб-локацијата односно може да се направи класификација на веб-страниците врз основа на нивната организација. Структурното рударење може да се користи за да се категоризираат веб-страниците и тоа:

1. *Хиперлинкови*: хиперврска е структурна единица која поврзува една локација во веб-страница со друга различна локација, или во рамките на истата веб-страна или различна веб-страница. Линкот кој поврзува различни делови од истата страница се нарекува *интрадокумент хиперлинк* и хиперврска која поврзува две различни страници се нарекува *интердокумент хиперврска;*

2. Структура на документот: Покрај тоа, содржините во рамките на веб-страница можат исто така, да бидат организирани во дрво-структурирана форма врз основа на различни HTML и XML тагови во рамките на страницата.

Можностите за употреба на оваа категорија на веб-рударењето е да се генерираат информации како што е сличноста помеѓу различни веб-страници [12].

Структурата на една веб-страница може да се прикаже на типичен веб-графикон кој се состои од веб-страници како јазли и хиперлинкови како рабови кои ги поврзуваат поврзаните страни.

2.1.2 Содржинско рударење

Содржинското рударење е процес на извлекување корисни информации од содржината на веб-документите со примена на алгоритмите за податочно рударење. Содржината на веб-документите се *вистинските* податоци за кои веб-страницата е дизајнирана за да ги пренесе на корисниците [12]. Страниците може да се од најразлични типови на податоци: неструктуриран текст, графика, звук, видео и полуструктуриран хипертекст, па оттука прозлегува дека постојат различни категории на содржинско рударење. Тоа е поврзано со податочното рударење, бидејќи многу техники за податочно рударење може да се применат во содржинско рударење, но истовремено и различно бидејќи веб-податоците се главно полуструктурирани и/или неструктурирани, додека рударењето на податоци првенствено се занимава со

структурирани податоци. Тоа, исто така, е поврзано со рударење на текст бидејќи голем дел од веб содржини се текстови, но исто така и различно од текстуалното рударење бидејќи вебот е полуструктуриран, додека текстуалното рударење се фокусира на неструктурирани текстови [24]. Може да се изработи концептуална шема [15] која може да ја опише семантиката на управување со голем број на неструктурирани веб-податоци.

2.1.3 Рударење на корисничките пристапи

Третиот вид на веб-рударење е рударење на кориснички пристапи. Овој тип на рударење овозможува колекција на информации од веб-пристапот за веб-страниците. Дизајнирањето на веб-страница е тежок и комплексен проблем. Дневниците од пристапите на корисниците овозможуваат да се набљудуваат корисниците, односно нивната интерактивност во сајтот и да се направат подобрувања на структурата на сајтот. Доколку успешно се изврши анализа на навиките на веб-посетителите можно е да се добијат индиции за тековните трендови на пазарот и да се предвидат идните трендови на потенцијалните клиенти. Доколку со анализата се утврди дека посетителите долго остануваат, тоа е доволен индикатор за потребата од преструктуирање на веб-страницата за да им помогне на посетителите брзо да дојдат до саканата информација. Со корисничкото рударење врз основа на информациите што поточно преферираат посетителите може да се понудат интересни содржини. За постигнување на оваа цел се предлага да се користат *адаптивни сајтови* кои користат информации за пристапните шеми на корисникот за подобрување на нивната организација и презентација [17].

Адаптивните сајтови ги набљудуваат корисничките активности, потешкотиите на корисникот и учат за типовите на корисници, вообичаениот пристап и проблемите кои доаѓаат со веб-локацијата. Во [5] со користење правило за асоцијација и групирање врз основа на користените URL адреси се воспоставува техниката за следење на вообичаен профил и истата е предложена во [5]. Техниката за веб-персонализација предложена во [4] се базира на откривање на правило за асоцијација од употребата на податоци. Во [18] е развиена методологија за оценување на квалитетот на веб-страницата врз основа на откривање и споредба на шеми за навигација на клиентите и

оние кои не се клиенти е предложена техника на динамичка адаптација на веб-локацијата.

2.2 Веб дневници за пристап и рударење на кориснички пристапи

Основен извор на податоци за рударење кориснички пристапи се дневниците од веб-серверите кои во себе содржат веб-дневници за пристап (web access logs). Секој пристап до серверот од страна на корисникот кој одговара на секое HTTP барање, автоматски се генерира еднократно влез(запис) во веб дневникот за пристап.

. Овие записи може да се чуваат во NCSA, Common Log Format (CLF) или Extended Log Format (ELF) и тоа се три најчести формати дефинирани од страна на W3C (World Wide Web Consortium). Веб-дневникот за пристап во W3C Extended Log Format формат ги содржи следниве елементи:

date time c-ip cs-username cs-method cs-uri-stem cs-uri-query sc-status sc-bytes cs-version cs(User-Agent) cs(Referer)

Веб-дневник за пристап во ELF формат има информации од c-ip е адресата или името на домаќинот(машина) на посетителот од каде што е направено барањето(request), cs-username/userID се користи за автентификација на посетителот, date, time датумот и времето на барање на страницата. Cs-method и cs-uri-stem се однесува на методот и url-то. Методот е средство за барање на страница. Тоа може да се GET, PUT, POST или HEAD. cs-uri-stem е URL-то на страната, која се бара. Cs-version е протоколот и е средство за комуникацијата која се користи, на пример HTTP/1.0. cs-status е Статус е завршниот код. На пример, 200 е код за успех. Полето cs-bytes покажува големина на бајти префрлени како резултат на барање на страница. Extended Log Format, во прилог на овие информации го зачувува cs(Referer) упатувачот кој е страницата од која е дојдено ова барање и cs(User-Agent) агент е веб прелистувач кој се користи. Слика 2.2 покажува веб-дневник за пристап во Extended Log Format.

date	time	cs-method	cs-uri-stem	cs-uri-query	cs-username	c-ip	cs-version	cs(User-Agent)	cs(Referer)	sc-status	sc-bytes
01.01.2013	00:01:06	GET	/content/informacii/sobranie_II_20	-	-	157.55.33.113	HTTP/1.1	Mozilla/5.0+(compatible;+bingbot/2.0;++http://www.bing.com/	-	200	856514
01.01.2013	00:03:30	GET	/cdp/govdc1.sei.gov.mk_SEI+CA(1).c	-	-	95.156.10.60	HTTP/1.1	securityd+(unknown+version)+CFNetwork/609+Darwin/13.0.0	-	404	1795
01.01.2013	00:03:34	GET	/cdp/govdc1.sei.gov.mk_SEI+CA(1).c	-	-	95.156.10.60	HTTP/1.1	securityd+(unknown+version)+CFNetwork/609+Darwin/13.0.0	-	404	1795
01.01.2013	00:03:50	GET	/content/Dokumenti/MK/Nacionaln	-	-	220.181.108.1	HTTP/1.1	Mozilla/5.0+(compatible;+Baiduspider/2.0;++http://www.baidu	-	200	5594172
01.01.2013	00:04:29	GET	/cdp/govdc1.sei.gov.mk_SEI+CA(1).c	-	-	95.156.10.60	HTTP/1.1	securityd+(unknown+version)+CFNetwork/609+Darwin/13.0.0	-	404	1795
01.01.2013	00:04:55	GET	/content/Poimnici/poimnik_opst.xls	-	-	93.182.139.13	HTTP/1.1	Opera/9.80+(Windows+NT+5.1;+U;+en)+Presto/2.10.289+Versio	http://www.sea.gov.mk/	200	1830669
01.01.2013	00:05:50	GET	/Default.aspx	ControlID=Nc	-	66.249.73.70	HTTP/1.1	DoCoMo/2.0+N905i(c100;TB;W24H16)+(compatible;+Googlebot	-	200	34496
01.01.2013	00:08:02	GET	/content/Dokumenti/MK/ANNEX+1-	-	-	109.202.102.1	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/535.1+	http://www.sea.gov.mk/	200	758509
01.01.2013	00:10:49	GET	/content/Poimnici/poimnik_prasalr	start=8744	-	66.249.73.70	HTTP/1.1	Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.	-	200	9952
01.01.2013	00:13:01	GET	/cdp/govdc1.sei.gov.mk_SEI+CA(1).c	-	-	95.156.10.60	HTTP/1.1	securityd+(unknown+version)+CFNetwork/609+Darwin/13.0.0	-	404	1795

Слика 2.2 Веб дневник за пристап во Extended Log Format
Figure 2.2 Web Access Login Extended Log Format

Целокупниот процес на рударење кориснички податоци како што е прикажано на слика 1.1 на страница 13 може да се подели во три чекори. *Предпроцесирање* вклучува отстранување на ирелеватни податоци од дневниците. *Шемата за откритивање* вклучува примена на разни техники и алгоритми за претпроцесирање на податоци. *Анализа на шема* е индетификација на корисните шеми од откриените шеми.

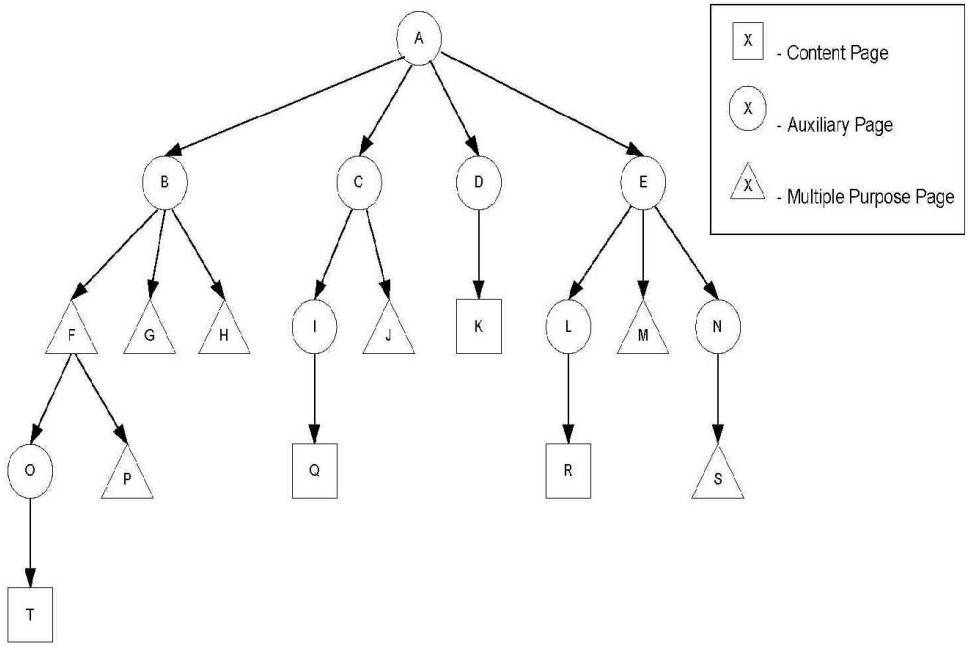
За да се изврши податочно рударење на кориснички пристапи потребни се техники за подготовка на податоци. Иако овие техники се презентирани во [9, 23]. Сепак, не постои многу литература која го опишува претпроцесирањето во детали за рударење на веб корисничките пристапи од веб-дневниците [6].

Постојат повеќе проблеми при рударење на кориснички пристапи. Еден од проблемите за добивање на јасна слика на пристап на веб страницата е предизвикан од страна на веб-прелистувачи и прокси сервери. Од страна на корисникот при посета на веб-страница се користат веб-прелистувачи кои ги зачувуваат страници кои биле посетени. Во случај да се прелистува иста страница која била претходно посетена, веб-прелистувачот ја прикажува страницата наместо да испраќа друго барање до веб-серверот и ова претставува веќе проблем бидејќи не може да се добие јасна слика. За да се подобрат перформансите на серверот и мрежната инфраструктура се користат прокси сервери кои ги кешираат често посетените страници локално. Проблемите предизвикани од страна на веб-прелистувачите и прокси серверите може да се реши со употреба на колачиња и *далечински агенти* [23].

Како што е спомнато во [9] огромниот број на записи кои ги прават веб-роботите во дневниците за пристап може да бидат избришани со помош на хеуристика. Друг начин на идентификација на веб-роботи е преку името на домаќинот. Со изготвување на список на веб-роботи, записите можат да бидат отстранети од веб-дневниците за пристап. Како што е споменато во оддел 3.1, веб-роботите пристапуваат до „robots.txt“ датотеката за веб-дозволи и со изработка на листа на IP адреси кои пристапиле до „robots.txt“ датотеката може да се направи бришење. Да напомнеме дека со ова не е комплетно отстранет овој проблем бидејќи, се појавуваат нови роботи и се појавиваат алатките за скенирање на страници и тие пак остануваат во веб-дневникот за пристап .

Постојат неколку техники за подготовка на податоци за идентификување на посетителите и нивните трансакции. Оваа задача е многу комплицирана поради постоењето на локалните кеширања, корпорациските firewall-и, како и прокси сервери. Имено, ако IP адресата е иста, ако дневникот на агентот покажува промена во софтверот на интернет-пребарувачот или оперативниот систем, разумна претпоставка е дека секој различен тип на агент за IP адреса претставува друг корисник. На пример, да ја земеме веб-локацијата прикажана на слика 2.3 и примерокот на информации собрани од пристап, агент, и дневници прикажани во Слика 2.4. Сите записи на дневниците во веб-дневникот имаат иста IP адреса и непознато(незапишано) корисничко име. Сепак, петтиот, шестиот, осмиот и десеттиот запис се пристапи со користење на различни агенти од другите, што укажува дека најавите претставуваат најмалку две кориснички сесии. Следниот хеуристички метод за идентификација на корисникот е да се користи дневник за пристап, односно упатувач кој претставува страница од каде е дојдено ова барање и топологија на локацијата за изградба на пребарувачки патеки за секој корисник. Ако една побарана страница не е директно поврзана со хиперлинк од која било од страниците посетени од страна на корисникот, повторно, хеуристиката претпоставува дека постои друг корисник со истата IP адреса. Гледајќи ја слика 2.4 примерок за повторен дневник, на третиот запис, страница L, не е директно поврзана од страниците A или B. Исто така, на седмиот запис, страница P е достапна од страната L, но не од кој било друг претходен дневник во дневникот. Ова би требало да укаже на тоа дека постои и трет корисник со

истата IP адреса. Затоа, по чекорот на идентификација на корисникот со примерокот на дневник, три уникатни корисници се идентификувани со прелистување на патеките на **A-B-F-0-G-A-D**, **A-B-C-J** и **L-R**, соодветно. Важно е да се напомне дека ова се само хеуристики за идентификување на корисници. Двајца корисници со иста IP адреса, кои го користат истиот прелистувач на истиот тип на машина може лесно да се збунат како еден корисник, доколку тие се во потрага на истиот сет од страници. Спротивно на тоа, еден корисник со користење на два различни пребарувачи, или кои типови во URL директно, без користење на структурата на локациите може да биде бележан како повеќе корисници.



Слика. 2.3 Пример веб-локација— Стрелки меѓу страниците кои претставуваат хипертекст врски

Figure. 2.3 Sample Web Site - Arrows between the pages represent hypertext links

#	IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referred	Agent
1	123.456.78.9	-	[25/Apr/1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
2	123.456.78.9	-	[25/Apr/1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
3	123.456.78.9	-	[25/Apr/1998:03:05:39 -0500]	"GET L.html HTTP/1.0"	200	4130	-	Mozilla/3.04 (Win95, I)
4	123.456.78.9	-	[25/Apr/1998:03:06:02 -0500]	"GET F.html HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
5	123.456.78.9	-	[25/Apr/1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
6	123.456.78.9	-	[25/Apr/1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
7	123.456.78.9	-	[25/Apr/1998:03:07:55 -0500]	"GET R.html HTTP/1.0"	200	8140	L.html	Mozilla/3.04 (Win95, I)
8	123.456.78.9	-	[25/Apr/1998:03:09:50 -0500]	"GET C.html HTTP/1.0"	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
9	123.456.78.9	-	[25/Apr/1998:03:10:02 -0500]	"GET O.html HTTP/1.0"	200	2270	F.html	Mozilla/3.04 (Win95, I)
10	123.456.78.9	-	[25/Apr/1998:03:10:45 -0500]	"GET J.html HTTP/1.0"	200	9430	C.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
11	123.456.78.9	-	[25/Apr/1998:03:12:23 -0500]	"GET G.html HTTP/1.0"	200	7220	B.html	Mozilla/3.04 (Win95, I)
12	123.456.78.9	-	[25/Apr/1998:05:05:22 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
13	123.456.78.9	-	[25/Apr/1998:05:06:03 -0500]	"GET D.html HTTP/1.0"	200	1680	A.html	Mozilla/3.04 (Win95, I)

Слика. 2.4 Пример-Информации од веб-дневниците за пристап
Figure. 2.4 Sample-Information from web access logs

Многу често се случува, веб-посетителите во неколку наврати поради зголемен интерес или било која друга причина да посетуваат веб-страница повеќе од еднаш. Исто така, многу е веројатно од еден ист компјутер и од иста страница еден корисник да ја завршува неговата или нејзината посета и друг корисник да ја почнува посетата.

Целта на идентификација на сесијата е да се поделат пристапените страници на секој корисник во индивидуални сесии. Наједноставниот метод за постигнување на ова е преку истек на време, каде што ако времето помеѓу барања на страница надминува одреден лимит, се претпоставува дека корисникот почнува нова сесија. Многу комерцијални производи користат 30 минути како стандарден истек на време и воспоставен истек од 25,5 минути врз основа на емпириски податоци. Користејќи опсег од 30 минути, патеката за корисникот 1 од најавениот дневник е поделен на две одделни сесии, бидејќи во последните две референци се повеќе од еден час подоцна од првите пет. Чекорот на идентификација на сесијата резултира во четири сесии на корисникот кои се состојат од **A-B-F-0-G, A-D, A-B-C-J, и L-R.**

Утврден е друг проблем во сигурното идентификување на сесии на уникатни корисници за тоа дали постојат важни пристапи кои не се евидентирани во логот за пристап. Овој проблем е наведен како *завршување на патека (Path completition)*. Методи слични на оние што се користат за

идентификација на корисникот може да се користат за завршување на патеката. Ако некое барање за страница е направено да не е директно поврзано со последната страница на бараниот корисник, преку записот упатувач може да се провери за да се види од која страница доаѓа барањето. Ако страницата не е во последната историја на пребарување на корисникот, претпоставката е дека корисникот го обележал копчето „назад“ кое е на располагање на повеќето пребарувачи, повикувајќи се на кеширани верзии на страници додека новата страница била побарана. Ако рефererираниот дневник(упатувачот) не е јасен, топологијата на локацијата може да се користи за да се постигне истиот ефект. Ако повеќе од една страница во историјата на корисникот содржи линк до бараната страница, се претпоставува дека страницата која е најблиску до претходно бараната страница е изворот на новото барање. Референците на страницата која недостасува кои се имплицирани преку овој метод се додаваат на датотеката од сесијата на корисникот. Тогаш е потребен алгоритам за да се процени времето на секоја додадена референца на страница. Едноставен метод на подигање на временскиот печат е да се претпостави дека секоја посета на страница која веќе е видена, ефективно ќе се третира како помошна страница. Просечната должина на упатување за помошни страници за локацијата може да се користи за проценка на времето на пристап до исчезнати страници. Гледајќи ги сликите 2.3 и 2.4, страница G не е директно достапна од страницата 0. Дневникот упатувач за оваа страница G бара листи на страница како бараната страница. Ова укажува дека корисникот 1 прави “backtracked” до страница B употребувајќи го копчето за враќање назад пред да се бара страница G. Затоа, страниците F и B треба да се додадат во сесија на датотеката за корисникот 1. Повторно, иако е можно дека корисникот го знае URL за страница G и го внесува тоа директно, ова е малку веројатно, и не треба да се појавува доволно често за да влијаат на алгоритмите за рударење. Чекорот на завршување на патека резултира во патеките на корисникот на **A-B-F-0-F-B-G**, **A-D**, **A-B-A-C-J**, и **L-R**. Резултатите од секој од претпроцесирачките чекори се сумирани во табела 2.1.

Табела 2.1 Резиме на резултатите од примероците на дневникот претпроцесирање
Table 2.1 Summary of Sample Log Preprocessing Results

Задача	**Резултат**
Исчистен лог	• A-B-L-F-A-B-R- C-0-J-G-A-D
Идентификација на корисникот	• A-B-F-0-G-A-D • A-B-C-J • L-R
Идентификација на сесијата	• A-B-F-0-G • A-D • A-B-C-J • L-R
Завршување на патека	• A-B-F-0-F-B-G • A-D • A-B-A-C-J • L-R

Секоја сесија на корисник во датотеката на сесија на корисникот може да се смета на два начина; или како една трансакција на референци на многу страници, или како збир на многу трансакции кој се состои од референца на една страница. Целта на идентификација на трансакцијата е да се создадат значајни кластери на референци за секој корисник. Постојат три поделени процеси за идентификација на трансакција. Првите два, референтна должина *(Reference Length)* и максимална пренасочена референца *(Maximal forward reference)*, прават обид да се идентификуваат семантички значајните трансакции, а третиот не е базиран на модел на пребарување туку со временски прозорец(рок).

За разлика од сесиите, при определување на трансакции предвид се зема топологијата на веб-локацијата односно се прави разлика на трансакции врз основа на типот на страницата која се појавува во нив, а тие се делат на

помошно-содржински *(Auxiliary-Content Transactions)* и исклучително-содржински *(Content-only transactions).* Сесиите кои се исклучително-содржински се позначајни во однос на помошно-содржинските трансакции, бидејќи асоцијативните правила се откриваат само помеѓу страниците кои содржат податоци важни за корисникот.

Пристапот на идентификација на трансакција со референтна должина *(Reference Length)* се базира на претпоставката дека износот на времето што корисникот го поминува на страницата се поврзува со тоа дали страницата треба да се класифицира како помошно-содржинска или исклучително-содржински страница за тој корисник.

Пристапот на идентификација на трансакција со максимална пренасочена референца (*Maximal forward reference)* е начин на создавање на трансакција во која една трансакција претставуваат сите страници сè до моментот додека корисникот не врати еден чекор назад.

Едноставен начин да се идентификуваат поединечни трансакции е да се користи дозволеното време. Ако временскиот период помеѓу две последователни посети на страница е повеќе од фиксен временски период, тогаш тоа треба да се смета како крај на трансакцијата и почетокот на друга. Триесет минути често се користи како истек на временски период, како што е споменато во [9, 23].

3. ПОДГОТОВКА НА ПОДАТОЦИ

Во овој дел се опишуваат извршените претпроцесирачки задачи и спроведените експерименти. Во дел 3.1 се дискутира за техниките на претпроцесирање и структурата на експериментите односно организацијата на експериментите и задачата за откривање на модели за секој од експериментите, додека делот 3.2 опфаќа се што дискутиравме во 2.2 во врска со задачите за подготовка на податоците. Откако ќе се опишат извршените претпроцесирачки задачи и спроведените експерименти следуваат детали за спроведените експерименти.

3.1 Техники на претпроцесирање

Фазата на претпроцесирање е најпредизвикувачка фаза во овој магистерски труд пред сè поради необработените и нечисти податоци кои можат да бидат од најразличен вид и времето кое е потребно да се исчистат сите ирелевантни податоци кои се вклучени. Оваа фаза одзема највеќе време(60% од времето) за изработка на овој труд и се состои од следните задачи: отстранување на некорисни (со шум) податоци, идентификација на посетители и нивни трансакции и издвојување на карактеристики. Предпроцесирањето може да се подели во три различни чекори:

1. *Отстранување на пристап на веб-роботи и филтрирање на слики и податоци со шум:*

   Веб-роботите (пајаци или лизгачи) се алатки кои вршат скенирање на сите веб страници од веб-локацијата и притоа генерира огромен број на барања до веб-серверот. Сите овие барања се запишани во веб дневниците за пристап. Згора на тоа, овие барања се надвор од анализата, бидејќи цел на интерес е однесувањето на корисниците. Сите веб-роботи вообичаено пристапуваат до „robots.txt“ датотеката која се креира од страна на администратор на веб-локацијата за дозвола за пристап. Исто така, некои роботи може да се идентификуваат со набљудување на Host името на IP адресите и корисничките агентите познати како роботи.

Општо, веб-страниците покрај текст содржат и други типови на податоци, како што се слика, звук или видео-датотеки. При секоке барање испратено до веб-серверот автоматски евидентира записи за содржината на страница која била побарана во моментот ново исто време евидентира и за која било слика, звук или видео-датотеки, кои биле испратени. Бидејќи овие типови на податоци не се цел на анализа на дневникот на пристап, истите нема да се земаат предвид. Треба да се земе предвид дека доколку и овие типови на податоци се анализираат истите може да дадат интересни показатели за структурата на веб-локацијата, перформансите на сообраќајот и корисничката мотивација [20].

При посета на веб-локацијата понекогаш се случува серверот да е недостапен или да паѓа кога корисникот сака да се најави (идентификува) или пак бара некои страници кои повеќе не постојат. При вакви случувања веб-серверот прави записи во дневникот за пристап со соодветен код. Во зависност од целта на анализата, овие записи по потреба може да бидат земени предвид или отфрлени.

2. *Извлекување на трансакции:* После процесот на отстранување некорисни (со шум) податоци, следен чекор е секвенците од записите во дневникот за пристап мора да бидат групирани во логички единици и тие претставуваат веб-трансакции. За трансакцијата се смета еден запис во дневникот или збир на записи во одреден временски период од страна на еден посетител од иста машина.

   При процесот на извлекување на трансакции за секој корисник се јавуваат и одредени проблеми [23]. Употребата на повеќе IP адреси од еден корисник, поставувањето на лимит на сесија што е максималниот временски период за прелистување на веб во една посета, прокси сервери и структурата на хиперлинк создаваат проблеми при идентификување на посетителите.

   Еден од начините кои може да се искористи за идентификување на трансакции е *упатната должина (reference length)* [23]. Веб-локацијата содржи страници кои се од различен тип и тоа како навигациски и

содржински. Посетителите поминуваат повеќе време на *навигациските* страници отколку на страниците за *содржина и ова е концептот на м*оделот на упатната должина. Друг начин за извлекување трансакции е со користење на *максимална напредна референца* (maximal forward references) [23] и таа претставува страница на која се пристапува пред страницата да биде повторно посетена. На пример, во една посета, ако некој посетител ги посети страниците I-J-K-I-L-J тогаш максималните напредни референци се К и L.За идентификување на пристап на корисник може да се искористи хеуристички метод *упатен дневник (referrer log)* [23]. Упатен дневник претставува датотека во која се чува запис на документ кога се бара заедно со запис на *упатната* страница преку која доаѓа барањето. На пример, доколку посетителот во моментот ја разгледува страницата „/Претпристапна поддршка“ и бара на страницата „/Претпристапна поддршка/ ИПА“, тогаш запис во упатниот дневник „/претпристапна подршка“ може да биде упатна страница за да се дојде до страницата ИПА. Употребата на овој хеуристички метод во комбинација со структурата на веб-локацијата може да помогне да се идентификуваат несекојдневните посетители.

3. *Извлекување на карактеристики и форматирање:* Во овој процес кој воедно е и последниот чекор од претпроцесирањето се врши форматирање на податоците во согласност со алгоритмите за податочно рударење кои се применуваат за да се извлечат карактеристиките од трансакциите. Ова значи идентификување на релевантните атрибути кои се сметаат дека ќе бидат корисни и обемноста на податоците ќе биде намалена.

   Вообичаено е да посетителите при посета на една веб-страница да се задржуваат на поинтересни страници за нив во однос на помалку интересните. Оваа карактеристиката е наречена „интерес“ е извлечена од [7] за да се моделираат корисничките параметри во различните пристапени страни на веб-страницата. При дефинирање на оваа карактеристика се земаат предвид факторите големината на страниците

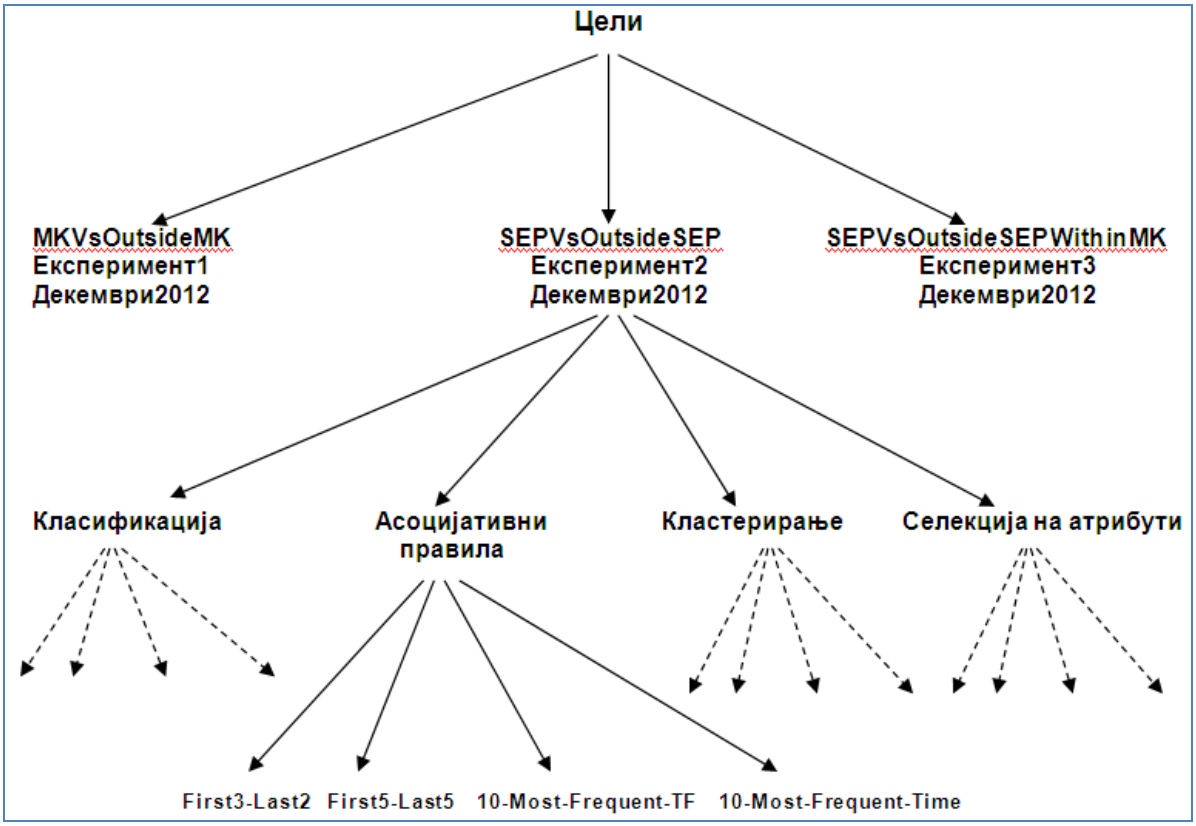
и брзината на мрежата бидејќи овие влијаат на тоа колку долго ќе се задржат посетителите на страниците.

На крај, за да може успешно да се применат алатките за податочно рударење потребно е векторите со карактеристика на фиксна должина да се конвертираат во соодветен формат кој ќе биде применлив за податочно рударење.

Во овој дел се опишува организацијата на спроведените експерименти и задачите за откривање на модели за секој експеримент. Слика 3.1 дава хиерархиски преглед на задачите за откривање на модели во експериментите на оваа теза и е наменета да ја покаже организацијата на експериментите.

Како што спомнавме претходно, овој магистерски труд има три цели. За спроведување на сите експерименти се користеше датотека на кориснички пристапи од декември 2012 година. Имено, сите експеримети кои се разработуваа беа спроведени врз една датотека за постигнување една цел и вкупно беа спроведени 3 експерименти. На пример, од слика 3.1, експериментот број 2 е спроведен за целта SEPVsNotSEP со помош на датотеката декември 2012.

Исто така, од сликата 3.1, може да се забележи дека за секој експеримент се користат 4 техники на податочно рударење и тоа: класификација, асоцијативни правила, групирање(кластерирање) и селекција на атрибути. Од претпроцесираните податоци се извлекоа четири различни групи на карактеристики First3-Last2, First5-Last5, 10-Most-Frequent-TF и 10-Most-Frequent-Time. Со користење на овие групи на карактеристики за секоја од споменатите техники беше извршена задача на откривање модели. - Како пример, може да се изврши задача за откривање модели за да се најдат правилата за асоцијација со помош на 10-Most-Frequent-TF групата на карактеристики и истите подетално се опишани во делот 3.2.3.



Слика 3.1 Хиерархиски приказ на извршените задачи за откривање модели
Figure 3.1 A Hierarchical View of the Pattern Discovery Task Performed

3.2 Обработка на податоци

3.2.1 Податоци од веб-дневникот

Деталите за податоците од веб-дневникот користени за експериментите се прикажани на табела 3.1. Како што е спомнато претходно, сите записи од пристапот и пребарувањето на веб-локацијата од страна на корисниците се чуваат во веб-дневникот кој е генериран од страна на Microsoft IIS 6.0 во формат Extended Log Format (ELF) слика 3.2 бидејќи веб-локацијата на СЕП е развиена и поставена на Microsoft Windows2003 платформа. За подготовка на податоците се користеа perl скрипти кои беа овозможени од таканаречената алатка WUMprep[29] која е дел од Open Source-Project HypKnowSys и истата може да се користи како самостојна или во конјункција со друга алатка за податочно рударење. Бидејќи алатката WUMprep работи исклучително само со дневници во NCSA формат неопходно беше да се направи конверзија од ELF во NCSA формат со

алатката RCONV[22]. На слика 3.3 е прикажан крајниот резултат добиен од употребата на RCONV конверторот.

```
#Software: Microsoft Internet Information Services 6.0
#Version: 1.0
#Date: 2012-11-05 07:57:01
#Fields: date time cs-method cs-uri-stem cs-uri-query cs-username c-ip cs-
version cs(User-Agent) cs(Referer) sc-status sc-bytes
2012-12-21 20:53:32 GET
/Content/Publications/Documents/Dogovor+od+Lisabon(1).pdf - - 77.29.28.212
HTTP/1.1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/537.11+
2012-12-21 20:57:17 GET /default.aspx ContentID=47 - 173.199.114.115 HTTP/1.1
"Mozilla/5.0+(compatible;+AhrefsBot/4.0;++http://ahrefs.com/robot/)" - 200
27256
2012-12-21 20:58:09 GET
/Content/Publications/Documents/Dogovor+od+Lisabon(1).pdf - - 89.205.15.152
HTTP/1.1
Mozilla/5.0+(Windows+NT+6.1)+AppleWebKit/537.11+(KHTML,+like+Gecko)+Chrome/23
.0.1271.97+Safari/537.11
http://www.pfk.uklo.edu.mk/index1.php?page=rezultati&grad= 200 131331
```

Слика 3.2 Веб-дневник за пристап во Extended Log Format
Figure 3.2 Web Access Login Extended Log Format

```
#Fields: Host Rfc932 Username Date:time timezone Request statuscode Bytes
Referrer User_agent
173.199.114.115.ahrefs.com - - [21/Dec/2012:20:57:17 +0100] "GET
/default.aspx?ContentID=47 HTTP/1.1" 200 27256 "-"
""Mozilla/5.0+(compatible;+AhrefsBot/4.0;++http://ahrefs.com/robot/)""
89.205.15.152.robi.com.mk - - [21/Dec/2012:20:58:09 +0100] "GET
/Content/Publications/Documents/Dogovor+od+Lisabon(1).pdf HTTP/1.1" 200
131331 "http://www.pfk.uklo.edu.mk/index1.php?page=rezultati&grad="
"Mozilla/5.0+(Windows+NT+6.1)+AppleWebKit/537.11+(KHTML,+like+Gecko)+Chrome/2
3.0.1271.97+Safari/537.11"
```

Слика 3.3 Веб-дневник за пристап во NCSA
Figure 3.3 Web Access Login in NCSA

Како што беше спомнато и претходно записите направени од страна на веб-роботите(лизгачите) не се од интерес за овој труд и истите не беа користени за експериментите. Веб-роботите обично ја отвораат датотеката „robots.txt“ претставува дозвола за пристап до веб-страниците од веб-локацијата. Затоа, овие записи беа отстранети со подготовка на список на веб-роботите кои ја отвориле датотеката „robots.txt“. За оваа активност се користеше скриптата RemoveRobots.pl. Меѓутоа, оваа техника не ги отстранува сите записи на веб-роботите, како што е споменато во Оддел 2.2.

Записите на IP адреси во дневникот кои веќе не постоеле исто така не беа од корист за експериментите. Оттука, и тие записи беа отстранети.

Записите од слики и видеа, записите од сервери-посредници, записите на погрешни барања освен со код 2/x/x и дупли записи, како што е опишано во Оддел 3.1, не беа корисни за експериментите на оваа теза и оттука, истите беа отстранети со помош на скриптата LogFilter.pl.

Табела 3.1 *Детали за логовите за веб-пристап*
Table3.1 *Details of Web Access Log Files*

Датотека за веб-пристап	Број на внесови	Период
access2012	452563	01/12/2012 - 31/12/2012

3.2.2 Идентификација на трансакции

Во 3.1 е дискутирано за идентификација на трансакции и за начините на кои може да се идентификуваат. Еден од начините за идентификација на трансакции е временската рамка “Time window” . Нека еден IP-ден се состои од сите записи на страниците кои се пристапени од една IP-адреса во текот на еден ден. IP-деновите беа извлечени врз основа на IP-адресите, веднаш откако нерелевантните записи беа отстранети од веб-дневникот на пристап. Трансакциите беа идентификувани од IP-деновите според времетраењето меѓу две последователни посети кои се споредува со одреден временски интервал. Доколку времетраењето меѓу отворањето на две страници „A“ и „B“ во еден IP-ден е подолго од 30 минути, тогаш „A“ се смета за последна страна која е пристапена во една трансакција и „B“ како прва страница отворена во друга нова трансакција. Периодот од 30 минути се смета за соодветен за да се разграничат две трансакции, како што е дискутирано во оддел 3.1. Можно е да се случи еден посетител да отвори страница во 23:15 и да заврши во 00:10, и тогаш во овој случај велиме дека има две трансакции наместо една. Процесот на идентификацијата на трансакции беше направен со користење на скриптата Sessionize.pl.

Трансакциите кои имаат најмалку 5 посетени страници се претпоставуваше дека ќе бидат корисни за ова податочно рударење. Трансакциите кои содржеа помалку од 5 посетени страници не беа користени. Табела 3.2 го покажува бројот на трансакции користени за експериментите од секоја датотека.

Табела 3.2 Број на трансакции користени од дневниците за веб-пристап
Table 3.2 Number of Transaction Used from Web Access Log Files

Веб пристап до датотека	Број на трансакции
access2012	1377

3.2.3 Групи на карактеристики

Претходно спомнавме во дел 3.1 дека последна фаза од претпроцесирањето е извлекување на карактеристики и форматирање. Откако ќе се изврши идентификација на трансакциите, следниот чекор е извлекувањето карактеристики од трансакции. За експериментите на овој труд, за процесот на откривање модели, беа извлечени четири групи карактеристики:

1. **First3-Last2:** Една инстанца на оваа група на карактеристики, се состои од првите 3 и последните 2 страници посетени од еден посетител во трансакцијата. Доколку бројот на посетени страни во една трансакција се помалку од 5 страници во тој случај велиме, на еден примерок ќе му недостасува една или повеќе атрибутски вредности. Оваа група на карактеристики беше избрана врз основа на претпоставката дека првите 3 и последните 2 посетени страници ќе содржат критични информации за интересот на корисниците и нивното однесување. Ова ќе помогне за разликување, односно за категоризирање на класите посетители.

   Пример за инстанца претставена со оваа група е покажан на слика 3.2. Примероците се претставени во формат „одделени со запирка“. Името на атрибутот во оваа група го покажува редоследот на посетените страници. На пример, атрибутот „F1“ одговара на првата страница посетена во една трансакција, „F2“ одговара на втората посетена страница. На сличен начин,

атрибутот „L2“ одговара на втората последна страница посетена во една трансакција и „L1“ одговара на последната посетена страница во една трансакција. Последниот атрибут беше користен во улога на варијабла на класата и овој атрибут ги претставува различните класи на кои припаѓаат примероците.

```
khost,F1,F2,F3,L2,L1,Country

pslux.ec.europa.eu,Home/Novosti,Home/Novosti,Home/home,Home/Novosti,
Home/Novosti,NMK

ws.ukim.edu.mk,Home/home,za nas/organizacija,NOK/tenderi,za nas/organizacija,
Home/home,mk

ghisa2.como.polimi.it,Home/home,Home/Novosti,za nas/organizacija,za nas/proektni
edinici,NOK/tenderi,NMK
192.168.0.1,Home/home,proces na preveduvanje/poimnici,Home/home,proces na
preveduvanje/poimnici,Home/home,MK_SEP
```

Слика 3.2 Пример за примероци со групата First3-Last2
Figure 3.2 A Sample of Instances Using the First3-Last2 Feature Set

2. **First5-Last5:** Оваа група е иста како групата First3-Last2, освен што во оваа група, примерокот се состои од првите 5 и последните 5 посетени страници од посетителот во една трансакција. Можно е да се посетени помалку од десет страници во една трансакција. Во тој случај велиме дека на примерокот ќе му недостасува една или повеќе вредности. Оваа групата повеќе се користи за рударење на повеќе информации во споредба со групата First3-Last2.

   На слика 3.3 е прикажан пример за инстанца. Примероците се претставени во формат „одделени со запирка“. Името на атрибутот во оваа група го покажува редоследот на посетените страници. На пример, атрибутот „FF1“ одговара на првата страница посетена во една трансакција, „FF2“ одговара на втората посетена страница. На сличен начин, атрибутот „LL2“ одговара на втората последна страница посетена во една трансакција и „LL1“ одговара на последната посетена страница во една трансакција. Последниот атрибут беше користен во улога на варијабла на класата и овој атрибут ги претставува различните класи на кои припаѓаат примероците.

Khost,FF1,FF2,FF3,FF4,FF5,LL5,LL4,LL3,LL2,LL1,Country
192.168.0.1,Home/home,Home/NPAA,Home/IzvestaiEU,Home/home,Kon pregovori/NPAA,Home/home,Home/NPAA,dokumenti/bazi na podatoci,Home/ZPVRM,
Home/home,SEP

62.220.196.2,Home/IzvestaiEU,Home/home,Home/ZPVRM,dokumenti/register na dokumenti,Home/AP,Home/linkovi,Home/home,Home/NPAA,Home/kontakti,Home/sitemap,OutsideSEP

Слика 3.3 Пример за примероци со групата First5-Last5
Figure 3.3 A Sample of Instances Using the First5-Last5 Feature Set

3. **10-Most-Frequent-TF:** За оваа група на карактеристики потребно е да се најдат најчесто посетуваните страници и да се направи анализа од веб- дневниците за пристап. Беа избрани 10 најчесто посетени страници како атрибути на оваа група. Атрибутската вредност во еден примерок е „T“ ако таа страница била посетена во трансакцијата и „F“ доколку е поинаку. На слика 3.4 е прикажан пример за примерок на групата 10-Most-Frequent-TF даден на слика 3.4.

Khost,Home/home,proces na preveduvanje/poimnici,Home/Novosti,proces na preveduvanje/proces na preveduvanje,prepristapna poddrska/IPA,za nas/organizacija,NOK/tenderi,dokumenti/register na dokumenti, za nas/proektni edinici,Home/IzvestaiEU,Country

ctel-31-11-127-106.cabletel.com.mk,T,T,F,T,F,F,F,F,F,F,mk

intranet.un.org.mk,T,F,F,F,T,F,F,F,F,F,mk

host242-147-static.206-37-b.business.telecomitalia.it,T,F,F,F,F,F,F,F,
F,F,NMK

Слика 3.4 Пример за примероци од групата 6-Most-Frequent-TF
Figure 3.4 A Sample of Instances Using the 6-Most-Frequent-TF Feature Set

4. Се сметаше дека ќе биде можно да се категоризираат посетителите кои прелистуваат однесувања врз основа на тоа дали посетиле често посетувана страница или не. Оттука, оваа група беше избрана за експериментите 10-Most-Frequent-Time**:** Оваа група е иста како претходната 10-Most-Frequent-TF дури ги има и истите атрибути. Во оваа група, атрибутска вредност е времето кое посетителот го поминал на конкретната често посетувана страница т.е. времето поминато во секунди

наместо „T“ или „F“, како што се користи во групата 10-Most-Frequent-TF. Времетраењето се пресметува ако се земе временската разлика меѓу отворањето на две последователни страници. Атрибутската вредност „0“ укажува дека посетителот не ја посетил оваа конкретна страна во таа трансакција. На слика слика 3.5 прикажан е пример за примероци од групата 10-Most-Frequent-Time. Можно е една од најпосетуваните страни да е посетена последна во една трансакција.

Во тој случај, времетраењето на посетата за оваа страница не може да се пресмета. Затоа, во овој случај, на соодветниот атрибут ќе му недостига вредност.

Оваа група е базирана врз претпоставката дека времето поминато на често посетувани страници може да биде фактор за разграничување и категоризирање на посетители.

```
Khost,Home/home,proces na preveduvanje/poimnici,Home/Novosti,proces na preveduvanje/proces na preveduvanje,prepristapna poddrska/IPA,za nas/organizacija,NOK/tenderi,dokumenti/register na dokumenti,za nas/proektni edinici,Home/IzvestaiEU,Country
mail.mtsp.gov.mk,0,0,0,0,0,20,0,0,8,0,OutsideSEP
77.29.135.225,384,0,0,105,0,0,0,0,29,0,OutsideSEP
192.168.0.1,0,0,25,18,0,0,0,0,0,0,SEP
192.168.0.19,0,0,0,4,0,11,6,0,0,0,SEP
```

Слика 3.5 Пример за примероци од групата 6-Most-Frequent-TF
Figure 3.5 A Sample of Instances Using the 10-Most-Frequent-Time Feature Set

3.2.4 Форматирање

Откако се најдени карактеристичните групи потребно е врз нив да се применат алатките за податочно рударење. За да се користат овие алатки во процесот на откривање шеми (модели) опишан во Оддел 2.1.3, потребно е примероците од карактеристичните групи да се претворат во соодветен формат.

За експериментите се користеше алатката за податочно рударење WEKA, опишана во Оддел 2.1.3. За потребите на WEKA потребно е влезните податоци да се претставуваат во соодветен формат кој е познат

за неа, односно во ARFF или CSV формат. На слика 3.6 е прикажан пример за примероци со соодветен формат користејќи ја групата 10-Most-Frequent-Time.

```
Khost,Home/home,proces na preveduvanje/poimnici,Home/Novosti,proces na preveduvanje/proces na preveduvanje,prepristapna poddrska/IPA,za nas/organizacija,NOK/tenderi,dokumenti/register na dokumenti,za nas/proektni edinici,Home/IzvestaiEU,Country

pslux.ec.europa.eu,95,0,0,0,0,0,0,10,0,0,NMK
mail.dzr.gov.mk,0,0,0,0,0,52,0,0,26,0,mk
mail.mtsp.gov.mk,0,0,0,0,791,0,0,477,0,0,mk
gw.ujp.gov.mk,440,0,0,0,0,0,0,0,0,65,mk
ptr.abcom.al,70,0,0,0,0,0,23,0,0,0,NMK
89.205.34.34.robi.com.mk,0,0,206,0,0,0,28,0,0,0,mk
ctel-31-11-96-6.cabletel.com.mk,0,0,0,0,0,0,0,0,0,609,mk
ctel-92-53-27-32.cabletel.com.mk,103,0,0,0,0,0,50,0,0,0,mk
host242-147-static.206-37-b.business.telecomitalia.it,0,0,0,0,0,0,0,0,0,0,NMK
mail.trinity-systems.com.mk,0,0,0,0,191,0,82,0,0,0,mk
D57D9DE3.static.ziggozakelijk.nl,1381,0,16,0,0,0,17,0,0,0,NMK
ws.ukim.edu.mk,0,0,28,0,20,9,8,0,45,0,mk
146.255.84.126,0,0,0,0,0,0,0,0,0,0,mk
192.168.0.1,691,381,0,0,0,0,0,0,0,0,mk
80.77.151.254.neotel.mk,0,0,0,0,0,0,96,0,0,5,mk
85.234.194.68.static.edpnet.net,8,0,1633,0,0,0,0,0,0,0,NMK
62.162.46.67,0,0,0,0,0,0,0,0,0,0,mk
77.28.7.244,0,0,35,55,0,0,0,0,0,0,mk
77.28.98.160,447,0,0,0,0,36,0,0,0,1319,mk
77.28.98.160,0,0,0,0,0,19,0,0,0,0,0,mk
77.28.98.70,0,0,0,0,0,0,0,0,0,0,0,mk
77.29.199.169,50,0,0,0,0,34,0,0,0,0,0,mk
77.29.199.82,0,0,0,191,0,0,0,0,0,0,mk
```

```
77.29.51.220,0,494,0,0,0,0,0,0,0,0,mk
77.29.51.220,0,2126,0,81,0,0,0,0,0,0,mk
77.29.51.220,0,572,0,0,0,0,0,0,0,0,mk
77.29.52.252,0,0,0,0,8,0,0,0,0,0,mk
77.29.55.159,0,2307,0,0,0,0,0,0,0,0,mk
80.77.152.66,0,0,75,0,0,0,0,0,0,0,mk
88.85.115.20,0,0,281,0,0,0,0,0,0,0,mk
91.226.20.127,1616,0,115,0,18,0,0,0,0,0,mk
91.226.20.127,4,0,155,0,0,0,0,0,0,0,mk
95.180.199.74,0,0,0,0,557,0,32,0,0,0,mk
```

Слика 3.6 Примерок за податоци со формат CSV, користејќи ја групата 10-Most-Frequent-Time

Figure 3.6 A Sample of Data in CSV Format Using the 10-Most-Frequent-Time Feature Set

4. РЕЗУЛТАТИ ОД ЕКСПЕРИМЕНТИТЕ

Сите експерименти кои се набројани и опишани за секои од нив се спроведе испитување, додека за се што беше интересно откриено при испитувањето се направи анализа на интересните шеми. Сите резултати кои се добиени за секој спроведен експеримент се објаснети и прикажани со своите детали. Бидејќи при испитувањето се добиваат голем број на секакви резултати (значајни и незначајни), во детали се дискутирани и опишани само резулатите од првиот експеримент, додека за другите експерименти се дискутираат само значајните излезни резултати. За сите експерименти постојат табели во кои се прикажани значењето на добиените резулати.

4.1 Експеримент 1: MKVsOutsideMK2012

Целта на првиот експеримент кој беше спроведен е за да споредат шемите на пристап меѓу посетители во Македонија и посетители надвор од Македонија со помош на датотеката „access2012“, како што е дадено во Табела 3.1. Се претпоставуваше дека ќе има разлика во шемите на користење меѓу посетителите на веб-страницата на Секретаријатот за европски прашања во и надвор од Македонија. Овој експеримент беше наменет да одреди дали тоа е навистина така и какви би биле разликите.

За процесот на откривање шеми се користеа следните техники за податочно рударење класификација, групирање, асоцијативни правила и селекција на атрибути. Вредностите на променливата на класата за секој примерок од овој експеримент беше „MK“ или „NotMK“. Оваа вредност беше одредена од host name или IP адесата на примерокот. На пример, ако host name завршува со поднизата „.mk“, тогаш тој посетител се смета дека доаѓа од Македонија, и оттука вредноста се одредува како „MK“. Ако host name не ја содржи поднизата „.mk“ на крајот, посетителот се смета дека доаѓа надвор од Македонија и вредноста на променливата на класата се одредува како „NotMK“.

Извршената експерименталната работа и резултатите добиени од експериментот се сумирани и прикажани во табела 4.1. Последната колона(Significant Patterns Discovered) од табелата ги покажува резултатите кои

се добиени од експериментите дали се значајни или не. Сите добиени резултати се дискутираат во детали соодветно.

4.1.1 Класификација

За да може добиените резулати да се сметаат за значајни потребно е точната класификација да надмине некој одреден праг. Овој праг не може да се очекува со висока точност затоа што постојат грешки кои се познати, но има и непознати грешки кои се вклучени во процесот на препроцесирачка работа како што беше дискутирано во делот 3.2, и ова е избрано со точност од 70%. Точност од 50% ако има проблем со 2 класи може да се постигне со погодување. Точност од 70% е значително подобрување на погодување. Проблемот на пренаучување(overfitting) беше решен со повторно изведување на алгоритамот со различни параметри како што е опишано во литературата за техники на класификација во дел 2.1.1

Табела 4.1 Експеримент 1: MKVsOutsideMK – Збир на резултати
Table 4.1 Experiment1: MKVsOutsideMK - Summary of results

Техники на податочно рударење	Веб дневник за пристап	Користени множества на податоци	Откриени значајни шеми
Класификација	access2012	First3-Last2	ДА
		First5-Last5	ДА
		10-Most-Frequent-TF	ДА
		10-Most-Frequent-Time	ДА
Асоцијативни правила	access2012	First3-Last2	ДА
		First5-Last5	Да
		10-Most-Frequent-TF	НЕ
Кластерирање	access2012	First3-Last2	ДА
		First5-Last5	ДА
		10-Most-Frequent-TF	НЕ
		10-Most-Frequent-Time	НЕ
Селекција на атрибути	access2012	First5-Last5	ДА
		First3-Last2	ДА
		10-Most-Frequent-TF	ДА
		10-Most-Frequent-Time	ДА

1. **First3-Last2**

Со примена на OneR класификаторот врз множеството податоци First3-Last2 се доби дека точноста на класификацијата достигнува 91.028% и атрибутот F3 (трета посетена страница) е атрибут со најголем коефициент на

информациска добивка. Со ова точно се класифицирани 487/535 случаи(инстанци), односно точно се класифицирани 91.028%, а неточно 8.972% случаи. Од матрица на контингентност/случајност може да се утврди дека лажно позитивни се 48, а лажно негативни нема. Оваа група на карактеристики дава значајна шема на употреба.

Имено, ако посетителите од Македонија најчесто ја посетиле изворната страница на СЕП тогаш посетуваат и други делови кои даваат информации за: извештаи и новости од ЕУ, претпристапна поддршка, кон процесите за преговори, документи и регистарот на документи, процес на преведување, за структурата, организацијата и работата на Секретаријатот за европски прашања, контакт со институцијата, за заменикот на претседателот на Владата на Република Македонија задолжен за европски прашања (ЗПВРМ) и неговата биографија, комуникација со јавноста, обука, стипендии, набавки, огласи, конкурси и тендери. Посетителите надвор од Македонија кои ја посетиле изворната страница, тогаш ја посетиле Билтен Европа и страницата за претпристапна поддршка, односно Twining.

```
=== Classifier model (full training set) ===
F3:
        Home/Bilten_evropa          -> NMK
        dokumenti/sloboden pristap          -> mk
        Home/Publikacii      -> mk
        za nas/sistem na upravuvanje       -> mk
        Home/biografija      -> mk
        Home/kontakti        -> mk
        prepristapna poddrska/IPA -> mk
        Kon pregovori/hronologija  -> mk
        za nas/proektni edinici       -> mk
        obuka/opsto  -> mk
        Home/home  -> mk
        Home/sitemap         -> mk
        Home/IzvestaiEU     -> mk
        Home/NPAA -> mk
        Kon pregovori/proces na pristapuvanje   -> mk
        obuka/stipendii      -> mk
        Home/HLAD -> mk
        Home/Novosti         -> mk
        za nas/organizacija  -> mk
        NOK/tenderi  -> mk
        proces na preveduvanje/poimnici -> mk
        Home/EU_novosti   -> mk
```

```
        proces na preveduvanje/proces na preveduvanje        -> mk
        Kon pregovori/NPAA          -> mk
        dokumenti/bazi na podatoci          -> mk
        dokumenti/register na dokumenti  -> mk
        komunikacija so javnosta/komunikaciski proekti-> mk
        Home/ZPVRM          -> mk
        Kon pregovori/pregovori     -> mk
        prepristapna poddrska/Druga stranska pomos  -> mk
        prepristapna poddrska/ORIO          -> mk
        Kon pregovori/SSA  -> mk
        prepristapna poddrska/Twinning   -> NMK
        Home/GI       -> mk
        Home/AP       -> mk
        Home/linkovi -> mk
        Home/prasalnik       -> mk
        prepristapna poddrska/TAIEX       -> mk
        dokumenti/registracija na proekti  -> mk
        prepristapna poddrska/Programi na unijata        -> mk
        komunikacija so javnosta/ispituvanje na javno mislenje          -> mk
        Home/MK_INFO      -> mk

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances         487               91.028  %
Incorrectly Classified Instances        48               8.972  %
Kappa statistic                          0.1308
Mean absolute error                      0.0897
Root mean squared error                  0.2995
Relative absolute error                 50.7731 %
Root relative squared error            101.1153 %
Total Number of Instances              535

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.077        0          1       0.077     0.143      0.538          NMK
                 1            0.923      0.91    1         0.953      0.538          mk
Weighted Avg.    0.91      0.833      0.918     0.91      0.874      0.538

=== Confusion Matrix ===

   a   b   <-- classified as
   4  48 |   a = NMK
   0 483 |   b = mk
```

Слика 4.1 Пример за излезни резултати од WEKA OneR Program за групата First3-Last2

Figure 4.1 A Sample Output of the WEKA 1R Program for group First3-Last2

2. First5-Last5

Со примена на класификаторот врз множеството податоци First5-Last5 се доби дека точноста на класификација достигнува 93.333% и атрибутот F33 (трета посетена страна) е атрибут со најголем коефициент на информациска добивка. Со ова точно се класифицирани 112/120 случаи (инстанци), односно точно се класифицирани 93.333%, а неточно 6.667% случаи. Од матрицата на контингентност може да се утврди дека лажно позитивни се 1, а лажно негативни 7. Оваа група на карактеристики дава, исто така значајна шема на употреба и добиените резултати се скоро исти, но со мали незначителни разлики со групата First3-Last2.

Посетителите надвор од Македонија кои ја посетиле изворната страница, тогаш ја посетиле страниците за ХЛАД и страниците за „кон преговори“ односно процесот на пристапување.

```
Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         112               93.3333 %
Incorrectly Classified Instances         8                6.6667 %
Kappa statistic                          0.4
Mean absolute error                      0.0667
Root mean squared error                  0.2582
Relative absolute error                 42.069  %
Root relative squared error             93.3914 %
Total Number of Instances              120

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.991    0.7       0.94      0.991     0.965      0.645    mk
                 0.3      0.009     0.75      0.3       0.429      0.645    NMK
Weighted Avg.    0.933    0.642     0.924     0.933     0.92       0.645

=== Confusion Matrix ===

   a   b   <-- classified as
 109   1 |   a = mk
   7   3 |   b = NMK
```

Слика 4.1 Пример за излезни резултати од WEKA OneR Program за групата First5-Last5

Figure 4.1 A Sample Output of the WEKA 1R Program for group First5-Last5

3. 10-Most-Frequent-TF

Користејќи ја оваа група на функции при откривањето на шема дава две интересни шеми на употреба. Почетната страница на веб-локацијата на СЕП е избрана како корен од повеќето страници од алгоритмите 1R и J48. Примери на резултати на алгоритмите 1R и J48 соодветно се прикажани во слика 4.3 и слика 4.4. Резултатите може да се толкуваат како што е објаснето во оддел 1.1.1. Резултатот на алгоритамот 1R е прикажан во слика 4.3 може да се толкува како: ако посетителите ја посетуваат почетната страница, тогаш тие се од Македонија и ако не ја посетуваат, почетната страница тогаш повторно се од Македонија. Ова се должи на фактот дека вработените во СЕП доволно ја познаваат веб -страницата и пристапот до одредени делови и пристапот го прават директно.

Резултатите добиени преку J48 алгоритам како што е прикажано на слика 4.4, исто така, се во согласност со резултатите произведени од страна на алгоритмот 1R. Дрвото на одлучување произведено од страна на J48 алгоритам, исто така откри уште една шема. Како што може да се види од дрвото на одлучување на слика 4.5, ако посетителите ја посетуваат изворната страница, и исто така ја посетуваат страницата поврзана со процесот на преведување, тогаш тие се најчесто од Македонија.

```
Home/home:
        T       -> mk
        F       -> mk
(1195/1376 instances correct)


Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1195               86.8459 %
Incorrectly Classified Instances       181               13.1541 %
```

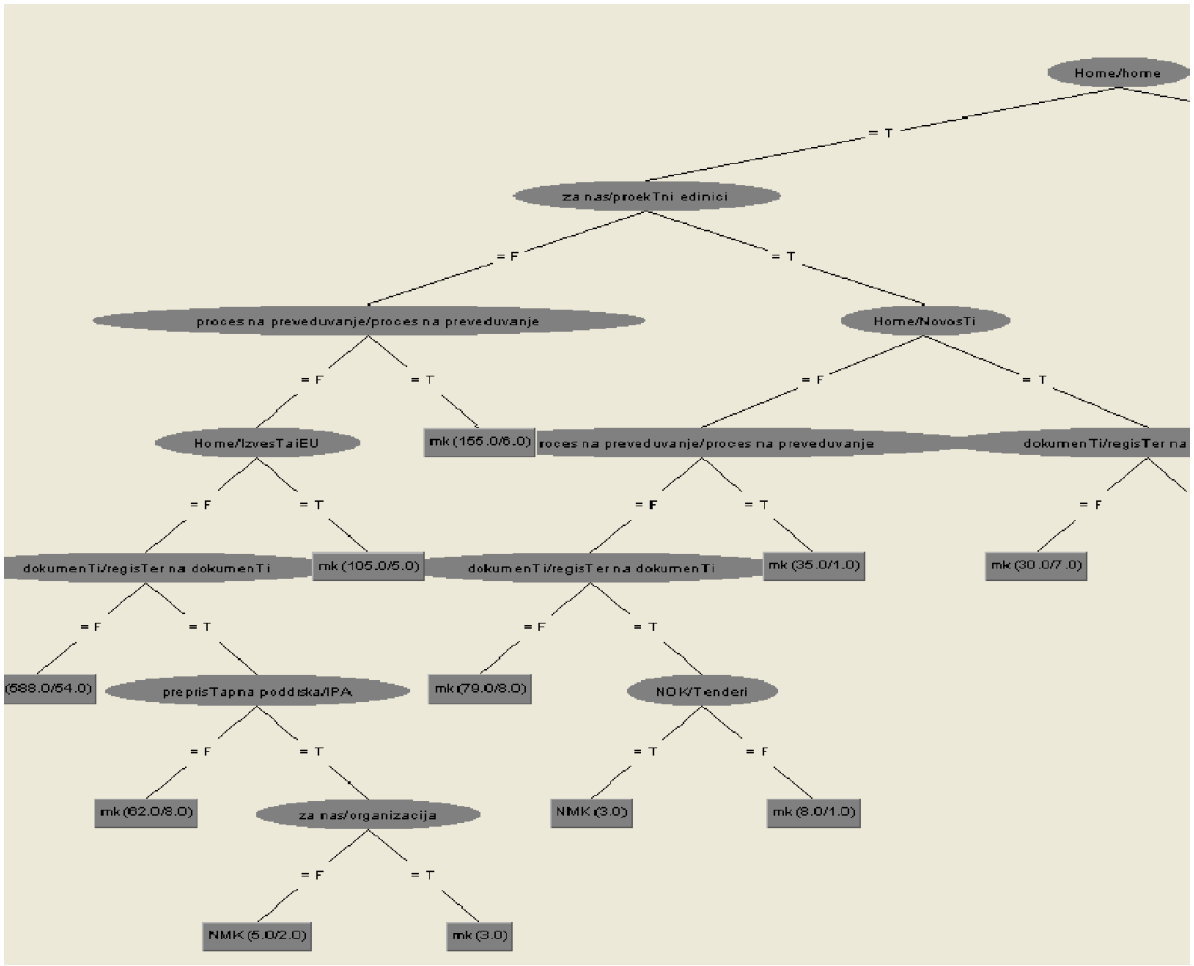
Слика 4.3 Пример за излезни резултати од WEKA OneR Program за експериментот 1

Figure 4.3 A Sample Output of the WEKA 1R Program in Experiment1

```
J48 unpruned tree
------------------

Home/home = T
|   za nas/proekTni edinici = F
|   |   proces na preveduvanje/proces na preveduvanje = F
|   |   |   Home/IzvesTaiEU = F
|   |   |   |   dokumenTi/regisTer na dokumenTi = F: mk (588.0/54.0)
|   |   |   |   dokumenTi/regisTer na dokumenTi = T
|   |   |   |   |   preprisTapna poddrska/IPA = F: mk (62.0/8.0)
|   |   |   |   |   preprisTapna poddrska/IPA = T
|   |   |   |   |   |   za nas/organizacija = F: NMK (5.0/2.0)
|   |   |   |   |   |   za nas/organizacija = T: mk (3.0)
|   |   |   Home/IzvesTaiEU = T: mk (105.0/5.0)
|   |   proces na preveduvanje/proces na preveduvanje = T: mk (155.0/6.0)
|   za nas/proekTni edinici = T
|   |   Home/NovosTi = F
|   |   |   proces na preveduvanje/proces na preveduvanje = F
|   |   |   |   dokumenTi/regisTer na dokumenTi = F: mk (79.0/8.0)
|   |   |   |   dokumenTi/regisTer na dokumenTi = T
|   |   |   |   |   NOK/Tenderi = T: NMK (3.0)
|   |   |   |   |   NOK/Tenderi = F: mk (8.0/1.0)
|   |   |   proces na preveduvanje/proces na preveduvanje = T: mk (35.0/1.0)
|   |   Home/NovosTi = T
|   |   |   dokumenTi/regisTer na dokumenTi = F: mk (30.0/7.0)
|   |   |   dokumenTi/regisTer na dokumenTi = T: NMK (2.0)
Home/home = F
|   proces na preveduvanje/poimnici = F: mk (203.0/43.0)
|   proces na preveduvanje/poimnici = T
|   |   za nas/organizacija = F
|   |   |   Home/IzvesTaiEU = F
|   |   |   |   proces na preveduvanje/proces na preveduvanje = F
|   |   |   |   |   preprisTapna poddrska/IPA = F: mk (76.0/32.0)
|   |   |   |   |   preprisTapna poddrska/IPA = T: NMK (4.0/1.0)
|   |   |   |   proces na preveduvanje/proces na preveduvanje = T: mk (14.0/1.0)
|   |   |   Home/IzvesTaiEU = T: NMK (2.0)
|   |   za nas/organizacija = T: NMK (2.0)

Number of Leaves  :     18

Size of the tree :      35


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1189               86.4099 %
Incorrectly Classified Instances       187               13.5901 %
```

Слика 4.4 Пример за излезни резулатати од WEKA J48 Program за експериментот 1

Figure 4.4 A Sample Output of the WEKA J48 Program in experiment1

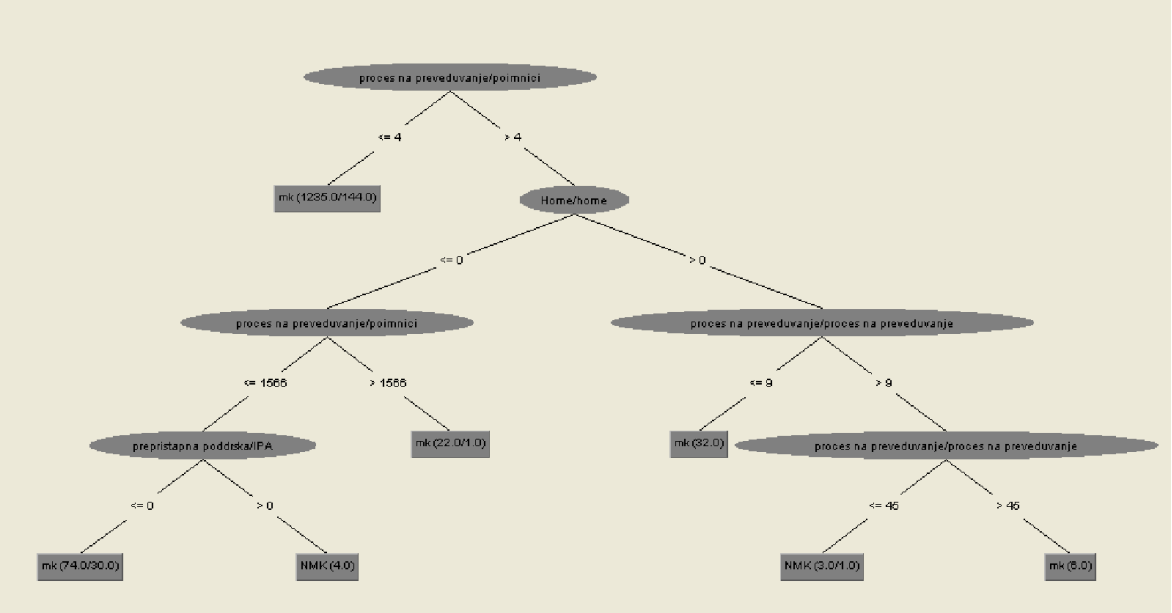


Слика 4.5 Дрво на одлука WEKA J48 програмот користејќи 10-Most-Frequent-TF податоци за експериментот 1

Figure 4.5 Decision Tree of the WEKA J48 Program Using the 10-Most-Frequent-TF Feature Set in Experiment1

4. 10-Most-Frequent-Time

Примената на двата алгоритми 1R и J48 врз оваа група на карактеристики и добиените резултати ја избираат почетната страница како најзначаен атрибут. Овој резултат е во согласност со резултатите добиени во оддел 4.1.1 користејќи група на функции 10-Most-Frequent-TF. Како што може да се види од слика 4.6 дрвото на одлучување покажува дека ако посетителите ја посетуваат изворната страница, и исто така ја посетуваат страницата поврзана со процесот на преведување, тогаш тие се најчесто од Македонија.



Слика 4.6 Дрво на одлука WEKA J48 програмот користејќи 10-Most-Frequent-TF податоци за експериментот 1

Figure 4.6 Decision Tree of the WEKA J48 Program Using the 10-Most-Frequent-Time Feature Set in Experiment1

Овој резултат е во согласност со резултатот добиен со користење на групата функции 10-Most-Frequent-TF.

Исто така, од слика 4.6 делумно може да се види дека ако посетителите не ја посетуваат изворната страница, а ја посетуваат страницата за „процес на преведување“, тогаш тие се од Македонија. Овој резултат веројатно доминира од страна на вработените на СЕП кои ја знаат страница за „процес на преведување“, и оттука тие ја посетуваат оваа страница директно.

4.1.2 Асоцијативни правила

Асоцијативните правила генерирани од алгоритмот Априори се анализирани за откривање на нивните предности.

1. First3-Last2

   Алгоритмот Априори открива 7 правила. Тие се прикажани на слика 4.7. Асоцијативните правила може да се толкуваат како што е објаснето во Оддел 1.1.1. Првото правило може да се толкува како: ако првата и претпоследната посетена страница е изворната страница тогаш посетителот е од Македонија. Второто правило може да се толкува како:

ако претпоследната посетена страница е изворната страница тогаш посетителот е од Македонија. Третото правило може да се толкува како: ако третата посетена страница е изворната и посетителот е од Македонија, тогаш прва посетена страница е изворната страница. Четвртото правило: ако првата и третата посетена страница е изворната тогаш тие посетители се од Македонија. Петтото правило: ако трета посетена страница е изворната тогаш исто така првата посетена страница била изворната. Шесто правило: ако третата посетена страница е изворната тогаш тие посетители се од Македонија, и последното правило: ако првата страница е изворна тогаш посетителите се од Македонија.

Сите овие правила укажуваат на тоа дека ако посетителите се од Македонија, тогаш тие ја посетуваат изворната страница. Овој резултат е во согласност со класификацијата на резултатот забележан во оддел 4.1.1.

2. First5-Last5

   Алгоритамот Априори најде десет интересни правила кои, исто така се слични како оние што се користат во групата на функција First3-Last2.

3. 10-Most-Frequent-TF

   Добиените резулатати за оваа група на карактеристики не беа интересни поради фактот дека бројот на посетени страници беше многу помал од бројот на страниците кои не беа посетени во трансакцијата. Оттука, алгоритамот Априори генерира правила во кои сите атрибутни вредности имале „F“ вредност. Еден резултат на добиените асоцијативни правила е прикажан на слика 4.8.

   Оттука, беше одлучено да се разгледаат само примерите во кои најмалку една вредност на атрибутот е „T". Сепак, не постои значителна разлика во генерираните правила. Исто така, се разгледуваат само оние случаи во кои 2 или повеќе атрибути имаа „T" вредност. Но, нема многу случаи кои го задоволуваат овој услов. Значи, беше констатирано дека со алгоритамот Априори со оваа група на функција веројатно нема да произлезат интересни асоцијативни правила. Затоа, оваа група на

функција не се користи за асоцијативно наоѓање во понатамошните експерименти.

```
Apriori
=======

Minimum support: 0.1 (54 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 4

Size of set of large itemsets L(2): 5

Size of set of large itemsets L(3): 2

Best rules found:

 1. F1=Home/home L2=Home/home 93 ==> Country=mk 88    conf:(0.95)
 2. L2=Home/home 107 ==> Country=mk 101    conf:(0.94)
 3. F3=Home/home Country=mk 106 ==> F1=Home/home 99    conf:(0.93)
 4. F1=Home/home F3=Home/home 106 ==> Country=mk 99    conf:(0.93)
 5. F3=Home/home 114 ==> F1=Home/home 106    conf:(0.93)
 6. F3=Home/home 114 ==> Country=mk 106    conf:(0.93)
 7. F1=Home/home 404 ==> Country=mk 367    conf:(0.91)
```

Слика 4.7 Делумен излез на WEKA Apriori програмот користејќи First3-Last2 податоци за експериментот 1

Figure 4.7 Partial Output of the WEKA Apriori Program Using the First3-Last2 Feature Set in Experiment1

```
1. Home/home=T CounTry=mk 977 ==> proces na preveduvanje/poimnici=F 918
conf:(0.94)
2. Home/home=T 1075 ==> proces na preveduvanje/poimnici=F 1010    conf:(0.94)
3. za nas/organizacija=F NOK/Tenderi=F 965 ==> za nas/proekTni edinici=F 902
conf:(0.93)
4. preprisTapna poddrska/IPA=F za nas/organizacija=F 978 ==> za nas/proekTni
edinici=F 909    conf:(0.93)
5. za nas/organizacija=F CounTry=mk 995 ==> za nas/proekTni edinici=F 921
conf:(0.93)
6. za nas/organizacija=F 1148 ==> za nas/proekTni edinici=F 1061    conf:(0.92)
```

7. za nas/organizacija=F dokumenTi/regisTer na dokumenTi=F 982 ==> za nas/proekTni edinici=F 905 conf:(0.92)

8. za nas/organizacija=F Home/IzvesTaiEU=F 996 ==> za nas/proekTni edinici=F 914 conf:(0.92)

9. proces na preveduvanje/poimnici=F za nas/organizacija=F 989 ==> za nas/proekTni edinici=F 903 conf:(0.91)

10. Home/home=T proces na preveduvanje/poimnici=F 1010 ==> CounTry=mk 918 conf:(0.91)

Слика 4.8 Делумен излез на WEKA Apriori програмот користејќи 10-Most-Frequent-TF податоци за експериментот 1
Figure 4.8 Partial Output of the WEKA Apriori Program Using the 10-Most-Frequent-TF Feature Set in Experiment1

4. **10-Most-Frequent-Time**

Алгоритмот Априори не работи со нумерички атрибути. Според тоа, групата на функции не се користи за рударење на асоцијативни правила.

4.1.3 Кластерирање

Кластерите кои беа формирани од страна на EM алгоритмот беа анализирани за да се открие дали постојат групи на посетители кои покажуваат слично однесување при посета на веб-локацијата на СЕП.

1. First3-Last2

Излезот на EM алгоритмот покажа два интересни кластери. Едниот кластер е означен со Cluster0 и се однесува на посетители надвор од Македонија и такви се 21% од вкупниот број посетители и Cluster1 кој се однесува на посетители од Македонија кои се 79% од вкупниот број посетители. Резултатот од EM алгоритмот може да се толкува како што е објаснето во оддел 1.1.1. Едниот кластер Cluster1 кој беше формиран од посетители од Македонија кои освен почетната страница пребаруваат страници поврзани со организацијата и работата на СЕП, новости и набавки, огласи и конкурси во СЕП. Одреден излезен резултат на овој кластер е прикажан на слика 4.9. Како што може да се види од оваа слика, овој кластер на посетители има тенденција да ги посети трите страници “Home/home/" и "Home/Novosti", “za nas/organizacija” и “NOK/tenderi”. Излезниот резултат, исто така, покажува дека овие 3 се само страници

што овој кластер на посетители најчесто ги посетуваат. Оттука, може да се извлече дека некои посетители од Македонија ја посетуваат веб-страницата на СЕП за пристап до информации за СЕП, новости, набавки, конкурси и огласи во СЕП.

```
EM
==
Number of clusters: 2
                                        Cluster
Attribute                                0       1
                                       (0.21)  (0.79)
=============================================================
F1
    Home/home                          24.1236 381.8764
F2
 za nas/organizacija                   2.2247  49.7753
 Home/Novosti                          1.0803  43.9197
F3
  Home/home                            6.5413 109.4587
  za nas/organizacija                  8.7708  32.2292
L2
  Home/home                            6.6491 102.3509
  NOK/tenderi                          3.1326  21.8674
  Home/Novosti                         4.0535  20.9465
L1
  Home/home                            2.4074  52.5926
  Home/Novosti                         2.7317  35.2683

Clustered Instances
0      113 ( 21%)
1      422 ( 79%)
Log likelihood: -14.16034
Class attribute: Country
Classes to Clusters:
   0   1  <-- assigned to cluster
  15  37 | NMK
  98 385 | mk
Cluster 0 <-- NMK
Cluster 1 <-- mk
Incorrectly clustered instances :        135.0    25.2336 %
```

Слика 4.9 Коментари од делумен излез од Cluster1 со WEKA EM Program Using користејќи 10-Most-Frequent-TF податоци за експериментот 1
Figure 4.9 Annotated Partial Output of Cluster-1 by the WEKA EM Program Using the 10-Most-Frequent-TF Feature Set in Experiment1

Другиот интересен формиран кластер Cluster0 беше од посетители надвор од Македонија кои освен почетната страница пребаруваат информации за процесот на преведување поимници. Друг забележителен делумен излез на овој кластер е прикажан на слика 4.10. Излезниот

резултат се толкува на истиот начин како излезниот резултат на слика 4.9. Како што може да се види од оваа слика, овој кластер на посетители има тенденција да ги посети трите страници „Home/home/", „proces na preveduvanje/poimnici" и „Prepristapna poddrska/IPA и Druga stranska pomos". Излезниот резултат, исто така, покажува дека овие 3 се само страници што овој кластер на посетители најчесто ги посетуваат. Оттука, може да се извлече дека некои посетители надвор од Македонија ја посетуваат веб-страницата на СЕП за пристап до процесот на преведување-поимници и претпристапна поддршка.

Овој резултат е во согласност со резултатот добиен со користење на групата функции 10-Most-Frequent-TF во оддел 4.1.1.

```
EM
==
Number of clusters: 2
                                          Cluster
Attribute                                   0       1
                                          (0.21)  (0.79)
============================================================


F1
  Home/home                              24.1236 381.8764
  proces na preveduvanje/poimnici         20.9813   1.0187
F2
proces na preveduvanje/poimnici          14.0638   3.9362
Home/home                                33.2271  20.7729
F3
proces na preveduvanje/poimnici          17.9757   1.0243
prepristapna poddrska/IPA                11.6955  11.3045
L2
proces na preveduvanje/poimnici           4.4278   7.5722
prepristapna poddrska/Druga stranska pomos  13.5134   2.4866
L1
```

proces na preveduvanje/poimnici	21.4128	1.5872
prepristapna poddrska/IPA	11.1993	20.8007

Слика 4.10 Коментари од делумен излез од Cluster0 со WEKA EM Program користејќи 10-Most-Frequent-TF податоци за експериментот 1
Figure 4.10 Annotated Partial Output of Cluster0 by the WEKA EM Program Using the 10-Most-Frequent-TF Feature Set in Experiment1

2. First5-Last5

   Формираните кластери обезбедуваат интересни информации за шемата на пристап. Оваа група на функција се користи за техниката на кластери при што се добива дека корисниците од Македонија се однесуваат скоро исто како групата на функција First3-Last2 со дополнителен интерес за препристапна поддршка, кон преговори, регистарот на документи, а корисниците кои не се од Македонија се однесуваат скоро исто како групата на функција First3-Last2.

3. 10-Most-Frequent-TF

   Излезните резултати на EM алгоритмот не покажаа интересни кластери. Имено, тој изгенерира кластери за најчесто посетените страници за кои пресметковната вредност за „F“ е поголема отколку нивната пресметковна вредност за „T“. Оттука, оваа група на функција не е пронајдена како погодна за употреба на кластери и не е користена за понатамошни експерименти.

4. 10-Most-Frequent-Time

   Излезните резулати од WEKA EM програмата која ја користи оваа група на функција не произведе некој интересен кластер. Поради големиот процент на некоректно кластерирани инстанци од 46,875%, оваа група на функција не е погодна за употреба со техника на кластерирање и дополнително откривање на шема со користење на техника на кластерирање со оваа група на функција, не е извршена.

4.1.4 Селекција на атрибут

Атрибутите избрани од страна на процесот на селекција на атрибут користејќи „cfssetEval“ проценителот на атрибут заедно со истражувачкиот метод „BestFirst“ беа анализирани за нивните предности.

1. First3-Last2

   Резултатите произведени со користење на оваа група ги покажува „F1“, „F3“, „L2“ како најзначајни атрибути.

   „F1“, „F3“, „L2“ се избрани атрибути и како што е објаснето во делот 3.2.3, „F1“ соодветствува на првата посетена страница во трансакцијата, "F3" соодвествува на третата страница во трансакцијата и "L2" соодвествува на претпоследната посетена страница во трансакцијата. Овој резултат покажува дека првата, третата и претпоследната страница на посетителите се најрелевантни и дискриминирачки.

```
=== Run information ===
Evaluator:    weka.attributeSelection.CfsSubsetEval
Search:       weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     WEKA32_MK_NotMK-weka.filters.unsupervised.attribute.Remove-R1-3,10-13
Instances:    535
Attributes:   6
              F1
              F2
              F3
              L2
              L1
              Country
Evaluation mode:    evaluate on all training data

=== Attribute Selection on all input data ===
Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 15
        Merit of best subset found:    0.042
Attribute Subset Evaluator (supervised, Class (nominal): 6 Country):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,3,4 : 3
                     F1
                     F3
                     L2
```

Слика 4.11 Излезен резулат со WEKA AttributeSelection Program користејќи First3-Last2 податоци за експериментот 1

Figure 4.11 Output by the WEKA AttributeSelection Program Using the First3-Last2 Feature Set in Experiment1

2. First5-Last5

   Селекцијата на атрибут со користење на оваа група на функција избрана како „FF1“ е најзначаен атрибут. За втори, трети, четврти и петти

најзначајни атрибути од вкупно десет беа соодветно избрани „FF2“, „FF3“, „LL5“ и „LL4“. „FF1“ е прва посетена страница во една трансакција, „FF2“ и „FF3“ се соодветно — втора и трета посетена страна во трансакцијата. „LL5“ и „LL4“ се шеста и седма страница посетена во трансакцијата.

3. 10-Most-Frequent-TF

   Селекцијата на атрибут со користење на оваа група на функција е насочена кон почетната страница Home/home како најзначаен атрибут и атрибутите process na preveduvanje/poimnici и process na preveduvanje/process na preveduvanje. Овој резултат е во согласност со резултатите добиени во Оддел 4.1.1 користејќи група на функции 10-Most-Frequent-TF.

```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval
Search:       weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     WEKATOP10F_MK_NotMK-weka.filters.unsupervised.attribute.Remove-R1-6
Instances:    1376
Attributes:   11
              Home/home
              proces na preveduvanje/poimnici
              Home/Novosti
              proces na preveduvanje/proces na preveduvanje
              prepristapna poddrska/IPA
              za nas/organizacija
              NOK/tenderi
              dokumenti/register na dokumenti
              za nas/proektni edinici
              Home/IzvestaiEU
              Country
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
              Best first.
              Start set: no attributes
              Search direction: forward
              Stale search after 5 node expansions
              Total number of subsets evaluated: 46
              Merit of best subset found:    0.052

Attribute Subset Evaluator (supervised, Class (nominal): 11 Country):
```

```
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,2,4 : 3
                Home/home
                proces na preveduvanje/poimnici
                proces na preveduvanje/proces na preveduvanje
```

Слика 4.12 Излезен резулат со WEKA AttributeSelection Program користејќи 10-Most-Frequent-TF.податоци за експериментот 1
Figure 4.12 Output by the WEKA AttributeSelection Program Using the 10-Most-Frequent-TF Feature Set in Experiment1

4. 10-Most-Frequent-Time

   Излезниот резултат со користење на оваа група на функција е насочен во страница „process na preveduvanje/poimnici" како најзначаен атрибут. Изворот беше единствениот избран атрибут. Овој резултат е истиот како оној со користење на групата функции 10-Most-Frequent-TF.

4.2 Експеримент 2: SEPVsOutsideSEP

Вториот експеримент беше направен за да се споредат шемите на пристап меѓу посетители од Секретаријатот за европски прашања и надвор од него со помош на датотеката „access2012" дадена во табела 3.1. Овој експеримент беше осмислен врз основа на претпоставката дека посетителите од СЕП може да остварат пристап до веб-локацијата различно од оние надвор од СЕП. Целта на експериментот беше да се пронајдат евентуалните разлики.

Техниките за податочно рударење користени за процесот на откривање шеми беше ист како и за експериментот 1. Вредноста на променливата на класата за секој примерок од експериментот беше „SEP" или „NotSEP". Оваа вредност беше одредена од host name на примерокот или IP адресата. На пример, ако host name ја содржеше поднизата „.sep." или „.sea." или IP адресата 192.168.0.1-254, тогаш тој посетител се смета дека е од СЕП, и оттука вредноста се одредува како „SEP". Ако host name не ја содржи поднизата „.sep." или „.sea.", посетителот се смета дека доаѓа надвор од СЕП и вредноста на варијаблата на класата се одредува како „NotSEP".

Табела 4.2 ги сумира експерименталната работа и резултатите добиени од експериментот. Само значајните добиени резултати се дискутираат детално во нивните сегменти.

Табела 4.2 Експеримент 2: SEPVsOutsideSEP – Збир на резултати
Table 4.2 Experiment 2: SEPVsOutsideSEP - Summary of results

Техники на податочно рударење	Веб дневник за пристап	Користени множества на податоци	Откриени значајни шеми
Класификација	access2012	First3-Last2	ДА
		First5-Last5	ДА
		10-Most-Frequent-TF	ДА
		10-Most-Frequent-Time	ДА
Асоцијативни правила	access2012	First3-Last2	ДА
		First5-Last5	НЕ
		10-Most-Frequent-TF	НЕ
Кластерирање	access2012	First3-Last2	ДА
		First5-Last5	ДА
		10-Most-Frequent-TF	НЕ
		10-Most-Frequent-Time	НЕ
Селекција на трибути	access2012	First3-Last2	ДА
		First5-Last5	ДА
		10-Most-Frequent-TF	ДА
		10-Most-Frequent-Time	ДА

4.2.1 Класификација

1. First3-Last2

Со примена на OneR класификаторот врз множеството податоци First3-Last2 се доби дека точноста на класификација достигнува 82.0561% и атрибутот L2 (претпоследна посетена страна) е атрибут со најголем коефициент на информациска добивка. Со ова точно се =класифицирани 445/535 случаи(инстанци), односно точно се класифицирани 82.0561%, а неточно 17.9439% случаи. Оваа група на карактеристики дава значајна шема на употреба.

Овој резултат добиен со користење на оваа група е, ако посетителите не се од СЕП, тогаш најчесто ја посетуваат изворната страница на СЕП, но исто така посетуваат и други страници кои даваат информации за: извештаи и новости од ЕУ, предпристапна поддршка,

кон процесите за преговори, документи и регистарот на документи, процесот на преведување, за структурата, организацијата и работата на Секретаријатот, контакт со институцијата, за заменикот на претседателот на Владата на Република Македонија задолжен за европски прашања(ЗПВРМ) и неговата биографија, комуникација со јавноста, обука, стипендии, набавки, огласи, конкурси и тендери. Ако посетителите се од СЕП и ја посетиле изворната страница, тогаш го посетиле и поздравното обраќање, односно страницата на заменикот на претседателот на Владата на Република Македонија задолжен за европски прашања. Ова најверојатно се должи поради интересот на вработените во СЕП за перспективите, визијата и идните предизвици на заменикот на претседателот на Владата на Република Македонија задолжен за европски прашања како координатор на работата на органите на државната управа и на другите органи и институции за подготовка на Република Македонија за членство во Европската унија.

2. First5-Last5

Со примена на класификаторот врз множеството податоци First5-Last5 се доби дека точноста на класификација достигнува 79.1667% и атрибутот FF3 (трета посетена страница) е атрибут со најголем коефициент на информациска добивка. Со ова точно се класифицирани 108/120 случаи(инстанци), односно точно се класифицирани 79.1667%, а неточно 20.8333% случаи. Оваа група на карактеристики дава исто така значајна шема на употреба и добиените резултати се скоро исти, но со мали незначителни разлики со групата First3-Last2, додека пак, ако посетителите се од СЕП и ја посетиле изворната страница, тогаш исто така ги посетиле страниците за документи, односно регистарот за документи, предпристапна поддршка, односно ИПА(инструмент за претпристапна помош) и страницата за комуникација со јавноста односно испитување на јавното мислење.

3. 10-Most-Frequent-TF

Овој резултат добиен со користење на оваа група на функции е во согласност со резултатите добиени во Оддел 4.2.1 користејќи група на функции 10-Most-Frequent-TF. Излезните резултати на алгоритамите 1R и

J48 покажуваат дека изворниот атрибут(Home/home) е избран од двата како најзначаен атрибут. Дрвото на одлучување генерирано од страна на алгоритмот J48 посочи дека ако посетителите ја посетиле изворната страница и тие, исто така ја посетиле страницата за процес на преведување, односно поимници, тогаш тие се надвор од СЕП.

4. 10-Most-Frequent-Time

Резултатите произведени од страна на двата алгоритми 1R и J48 ја избираат изворната страница „Home/home“ како најзначаен атрибут. Овој резултат е во согласност со резултатите добиени во оддел 4.2.1 користејќи група на функции 10-Most-Frequent-TF. Од дрвото на одлучување се покажува дека ако посетителите ја посетуваат изворната страница и новости, исто така ја посетуваат страницата поврзана со процесот на преведување, тогаш тие се надвор од СЕП. Овој резултат е во согласност со резултатот добиен со користење на групата функции 10-Most-Frequent-TF.

Исто така, од дрвото на одлука се добива следното: ако посетителите ја посетуваат изворната страница, и ја посетуваат страницата за „предпристапна подршка/ИПА“, тогаш тие се од СЕП.

4.2.2 Асоцијативни правила

1. First3-Last2

Алгоритмот Априори открива 2 правила. Асоцијативните правила може да се толкуваат како што е објаснето во оддел 1.1.1. Првото правило може да се толкува како: ако посетителот е од СЕП тогаш првата посетена страна F1 е изворната страница „Home/home“. Второто правило може да се толкува како: ако третата посетена страница е изворната страница, тогаш на посетителот, прва, односно почетна страница која ја посетил е изворната “Home/home“.

2. First5-Last5

Алгоритамот Априори најде само едно единствено правило и тоа може да се толкува како: ако посетителот е од СЕП тогаш на посетителот прва односно почетна страница која ја посетил е изворната „Home/home“.

3. 10-Most-Frequent-TF

Произведените резултати не беа интересни поради фактот дека бројот на посетени страници беше многу помал од бројот на страниците кои не беа посетени во трансакцијата. Оттука, алгоритамот Априори генерира правила во кои сите атрибутни вредности имале „F” вредност.

Оттука, беше одлучено да се разгледаат само примерите во кои најмалку една вредност на атрибутот е „T". Сепак, не постои значителна разлика во генерираните правила. Исто така, се разгледуваат само оние случаи во кои 2 или повеќе атрибути имаа „T“ вредност. Но, нема многу случаи кои го задоволуваат овој услов. Значи, беше констатирано дека со алгоритмот Априори со оваа група на функција не е веројатно да се произведуваат интересни асоцијативни правила. Затоа, оваа група на функција не се користи за придружно наоѓање во понатамошните експерименти.

4. 10-Most-Frequent-Time

Алгоритмот Априори не работи со нумерички атрибути. Според тоа, групата на функции не се користи за рударење на асоцијативни правила.

4.2.3 Кластерирање

Кластерите формирани од страна на EM алгоритмот беа анализирани за да се открие дали постојат групи на посетители кои покажуваат слично однесување при посета на веб-локацијата на СЕП.

1. First3-Last2

Излезот на EM алгоритмот покажа два интересни кластери. Едниот кластер е означен со Cluster0 и се однесува на посетители од СЕП и такви се 21% од вкупниот број посетители и Cluster1 кој се однесува на посетители надвор од СЕП кои се 79% од вкупниот број посетители. Едниот кластер кој беше формиран од посетители од СЕП, кои освен почетната страница пребаруваат страници поврзани со процесот на преведување, односно поимници во СЕП. Од излезните резултати на овој кластер може да се види дека посетителите имаат тенденција да ги посетат четирите страници „Proces na preveduvanje/poimnici/“ и „Proces na preveduvanje/process na preveduvanje", “prepristapna podrska/druga stranska poomos” и „prepristapna podrska/IPA”. Излезниот резултат, исто така,

покажува дека овие 4 се само страници што овој кластер на посетители најчесто ги посетуваат. Оттука, може да се извлече дека некои посетители од СЕП ја посетуваат веб-страницата на СЕП за пристап до информации за предпристапна помош и процесот на превод во СЕП.

Другиот интересен формиран кластер беше од посетители надвор од СЕП, кои освен почетната страна пребаруваат информации за СЕП и новости. Друг забележителен делумен излез на овој кластер е дека овој кластер на посетители има тенденција да ги посети четирите страници „Home/home/“, „proces na preveduvanje/process na preveduvanje“, „za nas/organizacija“ и „Dokumenti/registar na dokumenti“. Излезниот резултат, исто така, покажува дека овие 4 се само страници што овој кластер на посетители најчесто ги посетуваат. Оттука, може да се извлече дека некои посетители надвор од СЕП ја посетуваат веб-страницата на СЕП за пристап до процесот на преведување, новости, за организацијата на СЕП и за регистарот на документи.

2. First5-Last5

   Формираните кластери обезбедуваат интересни информации за шемата на пристап. Оваа група на функција се користи за техниката на кластерирање при што се добива дека корисниците кои не се од СЕП се однесуваат скоро исто како групата на функција First3-Last2 со дополнителен интерес за претпристапна поддршка, кон преговори, регистарот на документи, бази на податоци, регистрација на документи и проектни единици. Корисниците кои се од СЕП се однесуваат како корисници кои ги интересира само процесот на превод, односно поимници и тоа без да ја посетат изворната страница.

3. 10-Most-Frequent-TF

   Излезот на EM алгоритмот не покажа интересни кластери. Имено, тој изгенерира кластери за најчесто посетените страници за кои пресметковната вредност за „F“ е поголема отколку нивната пресметковна вредност за „T“. Оттука, оваа група на функција не е погодна за употреба на кластери и не е користена за понатамошни експерименти.

4. 10-Most-Frequent-Time

   WEKA EM програма која ја користи оваа група на функција не произведе некој интересен кластер. Поради големиот процент на некоректно кластерирани инстанци од 42,587%, оваа група не е погодна за употреба со техниката на кластерирање и дополнително откривање на шема со користење на техниката на кластерирање за оваа група , не е извршена.

4.2.4 Селекција на атрибут

Атрибутите избрани од страна на процесот на селекција на атрибут користејќи „cfssetEval“ проценителот на атрибут заедно со истражувачкиот метод „BestFirst“ беа анализирани за нивните предности.

1. First3-Last2

   Резултатите произведени со користење на оваа група ги покажува „F1“, „F3“, „L1“ како најзначајни атрибути.

   „F1“, „F3“, „L1“ се избрани атрибути и како што е објаснето во делот 3.2.3, „F1“ соодветствува на првата посетена страница во трансакцијата, „F3“ соодветствува на третата страница во трансакцијата и „L1“ соодветствува на последната посетена страница во трансакцијата. Овој резултат покажува дека првата, третата и последната страница на посетителите се најрелевантни и дискриминирачки.

2. First5-Last5

   Селекцијата на атрибут со користење на оваа група на функција избрана како „FF3“ е најзначаен атрибут. За втор најзначаен атрибут од вкупно десет беше соодветно избран „LL3“. „FF1“ е прва посетена страница во една трансакција, „LL3“ е осмата страница посетена во трансакцијата.

3. 10-Most-Frequent-TF

   Селекцијата на атрибут со користење на оваа група на функција е насочена во изворната страница „Home/home“ како најзначаен атрибут и атрибутот “NOK/Tenderi“. Овој резултат е во согласност со резултатите добиени во Оддел 4.2.1 користејќи група на функции 10-Most-Frequent-TF.

4. 10-Most-Frequent-Time

   Излезниот резултат со користење на оваа група на функција е насочен во изворната страница “Home/home”, “Home/Novosti”, “NOK/Tenderi”,

“dokumenti/register na dokumenti” и “za nas/proektni edinici” како најзначајни атрибути.

4.3 Експеримент 3: SEPVsOutsideSEPWithinMK

Третиот експеримент беше направен за да се споредат моделите на пристап меѓу посетители од СЕП и надвор од него, но од Македонија, со помош на датотеката „access2012“, како што е покажано во табела 3.1. Овој експеримент се засноваше на претпоставката дека може да има разлика меѓу шемите на пристап до веб-локацијата на СЕП меѓу посетителите од СЕП и оние надвор од СЕП, но од Македонија.

Исто како и во експериментот 1 за процесот на откривање модели беа користени истите техники за податочно рударење. Вредноста на променливата на класата за секој примерок од експериментот беше „SEP“ или „OutsideSEP“. Оваа вредност беше одредена од host name на примерокот. На пример, ако host name ја содржеше поднизата „.sep.“, „.sea.“ или IP адресата 192.168.0.1-254, тогаш тој посетител се сметаше дека е од СЕП. Ако host name не ја содржеше поднизата „.sep.“ или „.sea.“, а завршуваше со поднизата „.mk“, посетителот се сметаше дека доаѓа надвор од СЕП, но од Македонија. Оттука, вредноста на променливата на класата се одредуваше како „OutsideSEP“. Примероците во кои host name не содржеше „.sep.“, „.sea.“ или не завршуваше на „.mk“ не се земаа предвид.

Бројот на користени трансакции за овој експеримент е прикажан во табела 3.3.

Табела 3.3 *Број на трансакции за експериментот 3*

Експеримент	Број на трансакции
Experiment3: SEPsOutsideSEPWithinMK	1196

Табела 4.3 ги сумира експерименталната работа и резултатите добиени од експериментот. Значајните добиени резултати се дискутираат детално во нивните сегменти.

Табела 4.3 Експеримент 3: SEPVsOutsideSEPWithinMK – Збир на резултати

Table 4.3 Experiment 3: SEPVsOutsideSEPWithinMK - Summary of results

Техники на податочно рударење	Веб дневник за пристап	Користени множества на податоци	Откриени значајни шеми
Класификација	access2012	First3-Last2	ДА
		First5-Last5	ДА
		10-Most-Frequent-TF	ДА
		10-Most-Frequent-Time	ДА
Асоцијативни правила	access2012	First3-Last2	ДА
		First5-Last5	ДА
		10-Most-Frequent-TF	НЕ
Кластерирање	access2012	First3-Last2	НЕ
		First5-Last5	ДА
		10-Most-Frequent-TF	НЕ
		10-Most-Frequent-Time	НЕ
Селекција на атрибути	access2012	First3-Last2	ДА
		First5-Last5	ДА
		10-Most-Frequent-TF	ДА
		10-Most-Frequent-Time	ДА

4.3.1 Класификација

1. First3-Last2

Со примена на OneR класификаторот врз множеството податоци First3-Last2 се доби дека точноста на класификација достигнува 77.2257% и атрибутот L2 (претпоследна посетена страна) е атрибут со најголем коефициент на информациска добивка. Со ова точно се класифицирани 393/483 случаи(инстанци), односно точно се класифицирани 77.2257%, а неточно 22.7743% случаи. Оваа група на карактеристики дава значајна шема на употреба.

Имено, ако посетителите од Македонија најчесто ја посетиле изворната страница на СЕП тогаш посетуваат и други делови кои даваат информации за: извештаи и новости од ЕУ, предпристапна поддршка, кон процесите за преговори, документи(бази на податоци, регистрација на проекти и регистерот на документи), процес на преведување, за структурата, организацијата и работата на Секретаријатот за европски прашања, контакт

со институцијата, за биографија, говори и интервјуа на заменикот на претседателот на Владата на Република Македонија задолжен за европски прашања (ЗПВРМ), комуникација со јавноста, обука, стипендии, набавки, огласи, конкурси и тендери. Ако посетителите се внатре од СЕП и оние кои ја посетиле изворната страница, тогаш најверојатно ја посетиле страницата за поздравен говор на ЗПВРМ.

2. First5-Last5

Со примена на класификаторот врз множеството податоци First5-Last5 се доби дека точноста на класификација достигнува 77.2727% и атрибутот FF3 (трета посетена страница) е атрибут со најголем коефициент на информациска добивка. Со ова точно се класифицирани 98/110 случаи(инстанци), односно точно се класифицирани 77.2727%, а неточно 22.7273 % случаи. Оваа група на карактеристики дава, исто така значајна шема на употреба и добиените резултати се многу, скоро исти, со мали незначителни разлики со групата First3-Last2.

Посетителите внатре од СЕП кои ја посетиле изворната страница, тогаш ги посетиле и страниците за документи, односно регистерот на документи, претпристапна поддршка, односно ИПА и комуникација со јавноста односно испитување на јавното мислење.

3. 10-Most-Frequent-TF

Откривањето на шема користејќи ја оваа група на функции дава две интересни шеми на употреба. Изворната страница на веб-локацијатана СЕП е избрана како страница со најголема информациска добивка од повеќето страници од алгоритмите 1R и J48. Резултатите може да се толкуваат како што е објаснето во оддел 2.1.1. Резултатот на алгоритамот 1R може да се толкува како: ако посетителите ја посетуваат изворната страница, тогаш тие се од Македонија и ако не ја посетуваат, тогаш повторно се од Македонија. Ова се должи на фактот дека посетителите доволно ја познаваат веб страницата и пристапот до одредени делови и пристапот го прават директно.

Резултатите добиени преку J48 алгоритмот, исто така, се во согласност со резултатите произведени од страна на алгоритмот 1R. Дрвото на одлучување генерирано од страна на J48 алгоритам, исто така откри уште две шеми. За првата шема од дрвото на одлучување може да се заклучи,

ако посетителите ја посетуваат изворната страница, и исто така ја посетуваат страницата поврзана со процесот на преведување, тогаш тие се најчесто од Македонија. Втората добиена шема е: ако посетителите ја посетуваат изворната страница, и исто така не ја посетуваат страницата за превод, тогаш тие најчесто се од Македонија.

4. 10-Most-Frequent-Time

Резултатите произведени од страна на двата алгоритми 1R и J48 ја избираат изворната страница како најзначаен атрибут. Овој резултат е во согласност со резултатите добиени во оддел 4.3.1 користејќи група на функции 10-Most-Frequent-TF. Дрвото на одлучување покажува дека ако посетителите ја посетуваат изворната страница и се надвор од СЕП, тогаш тие временски помалку од 1.5 минути се задржуваат пократко во однос на посетителите од СЕП.

Исто така, делумно од дрвото на одлучување може да се види дека, ако посетителите кои ја посетиле страницата за новости тогаш тие посетители се од СЕП и се задржуваат подолго време во однос на посетителите кои се надвор од СЕП.

4.3.2 Асоцијативни(здружени) правила

Асоцијативните правила генерирани од алгоритмот Априори се анализирани за откривање на нивните предности.

1. First3-Last2

   Алгоритмот Априори открива 2 правила. Асоцијативните правила може да се толкуваат како што е објаснето во Оддел 1.1.3. Првото правило може да се толкува како: ако третата посетена страница е изворната страница тогаш на посетителот, прва посетена страна му била изворната. Сите овие правила укажуваат на тоа дека ако посетителите се од СЕП, тогаш тие најчесто ја посетуваат изворната страница. Овој резултат е во согласност со класификацијата на резултатот забележан во Оддел 4.3.1.

2. First5-Last5

   Алгоритамот Априори најде десет интересни правила кои исто така се слични како оние што се користат во групата на функција First3-Last2.

3. 10-Most-Frequent-TF

Произведените резултати не беа интересни поради фактот дека бројот на посетени страници беше многу помал од бројот на страниците кои не беа посетени во трансакцијата. Оттука, алгоритамот Априори генерира правила во кои сите атрибутни вредности имале „F"вредност.

Оттука, беше одлучено да се разгледаат само примерите во кои најмалку една вредност на атрибутот е „T". Сепак, не постои значителна разлика во генерираните правила. Исто така, се разгледуваат само оние случаи во кои 2 или повеќе атрибути имаа „T" вредност. Но, нема многу случаи кои го задоволуваат овој услов. Значи, беше констатирано дека со алгоритмот Априори со оваа група на функција не е веројатно да се произведуваат интересни асоцијативни правила. Затоа, оваа група на функција не се користи за придружно наоѓање во понатамошните експерименти.

4. **10-Most-Frequent-Time**

Алгоритмот Априори не работи со нумерички атрибути. Според тоа, групата на функции не се користи за рударење на асоцијативни правила.

4.3.3 Кластерање

Кластерите формирани од страна на EM алгоритмот беа анализирани за да се открие дали постојат групи на посетители кои покажуваат слично однесување при посета на веб-локацијата на СЕП.

1. First3-Last2

WEKA EM програмата која ја користи оваа група на функција не произведе некој интересен кластер. Поради големиот процент на некоректно кластерирани инстанци од 37.6812%, оваа група на функција не е пронајдена како погодна за употреба со техниката на кластерирање и дополнително откривање на шема со користење на техниката на кластерирање со оваа група на функција, не е извршена.

2. First5-Last5

Излезот на EM алгоритмот покажа два интересни кластери. Едниот кластер е означен со Cluster0 и се однесува на посетители надвор од СЕП, но во Македонија и такви се 79% од вкупниот број посетители и Cluster1 кој се однесува на посетители од СЕП кои се 21% од вкупниот

број посетители. Резултатот од ЕМ алгоритмот може да се толкува како што е објаснето во оддел 1.1.1. Едниот кластер кој беше формиран од посетители надвор од СЕП Cluster0, но во рамките на Македонија, кои освен почетната страница пребаруваат страници поврзани со системот на управување, организацијата и работата на СЕП, контакти, новости, набавки, огласи и конкурси, претпристапна поддршка, регистрација на документи(проекти), база на податоци и извештај за ЕУ. Од излезениот резултат на овој кластер може да се види дека, овој кластер на посетители има тенденција да ги посети страниците „Home/home/“, „Home/Novosti, izvestaiEU“, „za nas/organizacija,sistem na upravuvanje” и „NOK/tenderi“ и „process na preveduvanje/process na preveduvanje“. Излезниот резултат, исто така, покажува дека овие се страниците што овој кластер на посетители најчесто ги посетуваат. Оттука, може да се извлече дека некои посетители од Македонија ја посетуваат веб - страницата на СЕП за пристап до информации за СЕП, контакти, новости, како и набавки, конкурси и огласи, регистрација на проекти, процес на преведување и претпристапна поддршка во СЕП.

Другиот интересен формиран кластер Cluster1 беше од посетители внатре од СЕП, кои освен почетната страница пребаруваат информации за процесот на преведување, за проектни единици, публикации односно прирачници за ЕУ за систем на управување. Излезниот резултат се толкува на следниов начин: овој кластер на посетители има тенденција да ги посети страниците „Home/home/“, „proces na preveduvanje/poimnici“, „process na preveduvanje“, „za nas/proektni edinici, sistem za upravuvanje” и „Home/Publikacii“. Излезниот резултат, исто така, покажува дека овие страници што овој кластер на посетители најчесто ги посетуваат се најинтересни за нив. Оттука, може да се извлече дека некои посетители од СЕП ја посетуваат веб-страницата на СЕП за пристап до процесот на преведување-поимници, за проектните единици, публикациите и системот за управување. Овој резултат се должи поради запознавање на вработените со обврските и работата на СЕП како стручна служба на

Владата на Република Македонија задолжена за координација на процесот на приближување на Македонија кон ЕУ.

3. 10-Most-Frequent-TF

   Излезот на EM алгоритмот не покажа интересни кластери. Имено, тој изгенерира кластери за најчесто посетените страници за кои пресметковната вредност за „F" е поголема отклолку нивната пресметковна вредност за „T". Оттука, оваа група на функција не е погодна за употреба на кластери и не е користена за понатамошни експерименти.

4. 10-Most-Frequent-Time

   WEKA EM програмата која ја користи оваа група на функција не произведе некој интересен кластер. Поради големиот процент на некоректно кластерирани инстанци од 42.4268%, оваа група на функција не е погодна за употреба со техниката на кластерирање и дополнително откривање на шема со користење на техниката на кластерирање со оваа група на функција не е извршена.

4.3.4 Селекција на атрибут

Атрибутите избрани од страна на процесот на селекција на атрибут користејќи „cfssetEval", проценителот на атрибутот заедно со истражувачкиот метод „BestFirst" беа анализирани за нивните предности.

1. First3-Last2

   Резултатите произведени со користење на оваа група ги покажува „F1", „F3", „L1" како најзначајни атрибути.

   „F1", „F3", „L1" се избрани атрибути и како што е објаснето во делот 3.2.3, „F1" соодветствува на првата посетена страница во трансакцијата, „F3" соодветствува на третата страница во трансакцијата и „L1" соодветствува на последната посетена страна во трансакцијата. Овој резултат покажува дека првата, третата и последната страница на посетителите се најрелевантни и дискриминирачки.

2. First5-Last5

   Селекцијата на атрибут со користење на оваа група на функција избрана како „FF3" е најзначаен атрибут. За втор и трет најзначаен атрибут од вкупно десет беа соодветно избрани „FF5" и „LL3". „FF3" е трета посетена

страница во една трансакција, „FF5” е соодветно петта посетена страница во трансакцијата и „LL3“ е осма посетена страница во трансакцијата.

3. 10-Most-Frequent-TF

   Селекцијата на атрибут со користење на оваа група на функција е насочена во изворната страница Home/home како најзначаен атрибут и атрибутите Home/Novosti, Nok/tenderi и dokumenti/register na dokumenti. Овој резултат е во согласност со резултатите добиени во Оддел 4.3.1 користејќи група на функции 10-Most-Frequent-TF.

4. 10-Most-Frequent-Time

   Излезниот резултат со користење на оваа група на функција е насочен во изворната страница Home/home како најзначаен атрибут и атрибутот za nas/proektni edinici.

5. АЛАТКИ ЗА АНАЛИЗА НА ВЕБ-ДНЕВНИЦИ

Како што е спомнато погоре, развојот и дигитализацијата на општеството, автоматското запишување на податоците, текстовите и другите содржини во дигитален формат истовремено придонесуваат за огромен развој на WWW сервисот, односно појава на многу веб-сервери. Денеска, овие веб-сервери стануваат се популарни и се пристапуваат од страна на милиони веб-корисници 24 часа. При посетата сите овие посетители зад себе оставаат траги од нивната посета и однесувања во облик на веб дневници кои се чуваат во веб-серверите. Истовремено, со нараснувањето на овие веб-дневници се појавува потребата за анализа и визуелизација на истите. Со анализа на веб-дневниците снимени во овие сервери, како и на прокси серверите, можно е да се дојде до знаења за однесувањето и структурата на самите веб-корисници. За анализа на веб-дневниците се користат софтвери кои се наречени софтвери за анализа на веб-дневници (Web log analysis software).

Софтверите за анализа на веб-дневници се еден вид на веб аналитички софтвер кој врши парсирање на серверскиот дневник од веб-серверот и врз основа на вредностите кои се содржат во дневникот, произлегуваат показатели за тоа кога, како и од страна на кого се посетени веб-страниците на серверот. Обично, извештаите кои се генерирани од дневник датотеките се појавуваат веднаш, но исто така, дневник датотеките алтернативно може да се парсираат во база на податоци и да се генерираат извештаи во зависност од потребата. Сите овие софтвери делумно се раликуваат меѓусебно, но исто така, имаат и некои заеднички особини како што се:

- Број на посети и број на уникатни посетители
- Времетраење на посетите и последни посети
- Најавени корисници, и последно најавени посети
- Денови во неделата и најгуст сообраќај
- Домени / земји на посетители
- Листа на хостови(домаќини)
- Вкупен број на прегледани страници
- Најмногу видени, влезни и излезни страници
- Типови на датотеки

- Користени оперативни системи
- Користени пребарувачи
- Роботи
- HTTP насочник(од каде е пристапено до страницата)
- Клучни фрази и зборови користени да се најде анализираната веб-локација
- HTTP грешки
- Некои од софтверите, исто така, даваат извештај за тоа кој е на веб-локацијата, следење на посетителот после кликање, време на посетата и навигациска страна

5.1 Апликации за анализа на веб-дневници и каректиристики

Како што е спомнато погоре постојат различни типови на апликации за анализа на веб-дневници кои повеќе или помалку се слични помеѓу себе во своите функционалности и карактеристики. Сите тие се разликуваат според типовите на лиценци за користење и тоа: комерцијални и бесплатни, а според начинот на хостирање на софтверот се делат на Self-hosted software и Hosted/Software as a service.

5.1.1 Self-hosted software

Self-hosted software е модел на самостојна инсталација, развивање и одржување на софтверот и истиот може да биде free/libre/open-source software (FLOSS), односно слободен софтвер/код и Proprietary software, односно комерцијален софтвер.

Слободниот софтвер/код (FLOOS) е софтвер кој е слободен за користење, а исто така и кодот, односно на корисниците им дава можност и права да го користат, копираат, проучуваат, менуваат, развиваат и подобруваат дизајнот преку достапноста на изворниот код.

Во табелата 5.1 е прикажан табеларен преглед на софтвери (FLOOS) за анализа на веб-дневници.

Табела 5.1 Ова е табела за споредба на слободни софтвери за веб анализа под слободна софтверска лиценца (FSL)
Table 5.1 This is a comparison table of web analytics software released under a free software license.

Name Име	Програмски јазик	Подржани бази на податоци	Метод на следење	Последна стабилна издадена верзија	Лиценца
Analog	C	Logfile-based	Web log files	6.0	GNU GPL
AWStats	Perl	Logfile-based	Web log files	7.1	GNU GPL
CrawlTrack	PHP	MySQL	PHP pagetag	3.3.2	GNU GPL
Open Web Analytics	PHP	MySQL	JavaScript or PHP pagetag	1.5.2	GNU GPL
Piwik	PHP	MySQL	JavaScript or PHP pagetag or Web log files	1.12	GNU GPL
W3Perl	Perl	Logfile-based	Web log files	3.17	GNU GPL
Webalizer	C	Logfile-based	Web log files	2.23-05	GNU GPL

Комерцијалниот софтвер (Proprietary software or closed source) е компјутерски софтвер лиценциран под ексклузивно законско право на носителот на авторските права, со намера дека лиценцата е дадена со право да се користи софтверот само под одредени услови и ограничен од други намени, како што се промена, споделување, проучување, и редистрибуирање. Обично изворниот код на комерцијален софтвер не е ставен односно не е достапен на располагање.

Во табелата 5.2 е прикажан табеларен преглед на комерцијални софтвери(Proprietary software) за анализа на веб-дневници.

Табела 5.2 Ова е табела за споредба на комерцијални софтвери за веб анализа
Table 5.2 This is a comparison table of web analytics proprietary software

Име	Компанија	Платформа	Подржани бази на податоци	Метод на следење	Последна издадена верзија	Цена во долари
Angelfish	Actual Metrics	Linux/Windows	Proprietary	Page tags or log analysis	1.1	$1,295
Mint	Mint	PHP	MySQL	Cookies via JavaScript	2.17	$30/Site
Sawmill	Flowerfire Inc	Windows/Linux/BSD /POSIX	MS SQL/MySQL/Oracle Database/PostgreSQL/ Proprietary	Cookies via JavaScript & Logs	8.5	mixed, from $99/profile
Splunk	Splunk Inc.	Windows/Linux/BSD /Solaris	Proprietary	Web log files	4.3	Negotiable, 500MB per day free
Urchin	Google	Windows/Linux/BSD	MySQL, PostgreSQL	Cookies & Logs	7.0	Sale has been discontinued
Tealeaf cx*	Tealeaf	Windows/Linux	MS SQL/Proprietary	Network traffic monitor	8.4	See web-site
Unica NetInsight	IBM	Windows/Linux/Solaris	MS SQL/DB2/Oracle Database/Netezza	Web log files & Cookies	8.6 (as of 2012-05-15)	Various pricing options

5.1.2 Hosted/Software as a service

Софтвер како услуга (engl. Software as a Service) е модел на дистрибуција на софтвер во кој произведувачот на апликативното решение ја изработува апликацијата, управува со самата апликација и со околината која ја одржува (hosting), а на корисниците им е достапна по мрежа. Главната разлика на SaaS моделот во однос на традиционалниот модел е дека софтверот не се купува, туку се плаќа

услугата на неговото користење. Предности на оваа услуга се: нема потреба за набавка и одржување на инфраструктурата за инсталација и одржување на новиот софтвер, заштеда на време за имплементација, достапност, сигурност и надградба на целиот систем.

Во табела 5.3 е прикажан табеларен преглед на софтверот (SaaS) за анализа на веб-дневници.

Табела 5.3 Ова е табела за споредба на хостирани софтвери за веб анализа софтвер како сервис

Table 5.3. This is a comparison table of hosted web analytics software as a service.

Име	Компанија	Метод на следење	Последна стабилна издадена верзија	Цена во долари
Analyzer	AT Internet	Cookies via JavaScript	N/A	Negotiable
Bango Mobile Web Analytics	Bango plc	Mobile ID and cookies	4.0	од $49/month
ClickTale	ClickTale	Cookies via Javascript	N/A	Free / Negotiable
Google Analytics	Google	Cookies via JavaScript	N/A	Free (Standard), $150,000 Annual (Premium)
Insight	Omniture (Adobe Systems)	Cookies via JavaScript	N/A	Negotiable
Mapmyuser.com	Mapmyuser, LLC	Cookies via JavaScript	N/A	Free
Quantcast	Quantcast Corporation	Cookies via JavaScript	N/A	Free

Име	Компанија	Метод на следење	Последна стабилна издадена верзија	Цена во долари
SiteCatalyst	Omniture (Adobe Systems)	Cookies via JavaScript	15	Negotiable
StatCounter	StatCounter	Cookies via JavaScript	N/A	Free - $5/month ... $119/month
Webtrekk Q3	Webtrekk	Cookies via JavaScript	N/A	From $202/month
Webtrends	Webtrends	Cookies via JavaScript	N/A	N/A
Woopra	iFusion Labs LLC	Cookies via JavaScript	1.2	Free - $499.95+/month
Yahoo! Web Analytics	Yahoo!	Cookies via JavaScript	Not available anymore	Free

5.2 Веб-анализатор - Deep Log Analyzer

Deep Log Analyzer е софтверско решение за веб-анализа кое е напредно и прифатливо за мали и средни веб-локации. Со ова решение може да се анализира однесувањето на посетителите на веб-локацијата и да се добие комплетна статистика за користењето на веб-локацијата во неколку лесни чекори. Со овој веб анализатор се знае точно од каде клиентите доаѓаат, како се движат низ страниците на веб-локацијата, и каде истата ја напуштаат. Овие сеопфатни информации може да помогнат за да се привлечат повеќе посетители и истите да се претворат во задоволни посетители.

Со Deep Log Analyzer може да се видат извештаи за сите пристапени ресурси од веб-локацијата, активноста на посетителите и навигацијата, веб-локации преку кои е дојдено до анализираната веб-локација, како пребарувале

за да пристапат до веб-локацијата, лизгачи кои го пребаруваат сајтот, кои прелистувачи и оперативни системи ги користи посетителот, грешки на веб-серверот и друго. Исто така, овој анализатор го прави лесен за примена тоа што може да следи статистички промени со текот на времето и да направи споредба на извештаите за различни временски периоди.

Овој веб-анализатор е конфигурабилен, односно овозможува да се креираат специфични извештаи или стандардни извештаи кои ќе ги задоволат сопствените специфичните потреби. Исто така, Deep Log Analyzer-от статистичките податоци за веб-локацијата ги чува во стандардната MS Access база на податоци, а со тоа пак се овозможува лесен пристап до базата од други апликации.

Главни карактеристики:

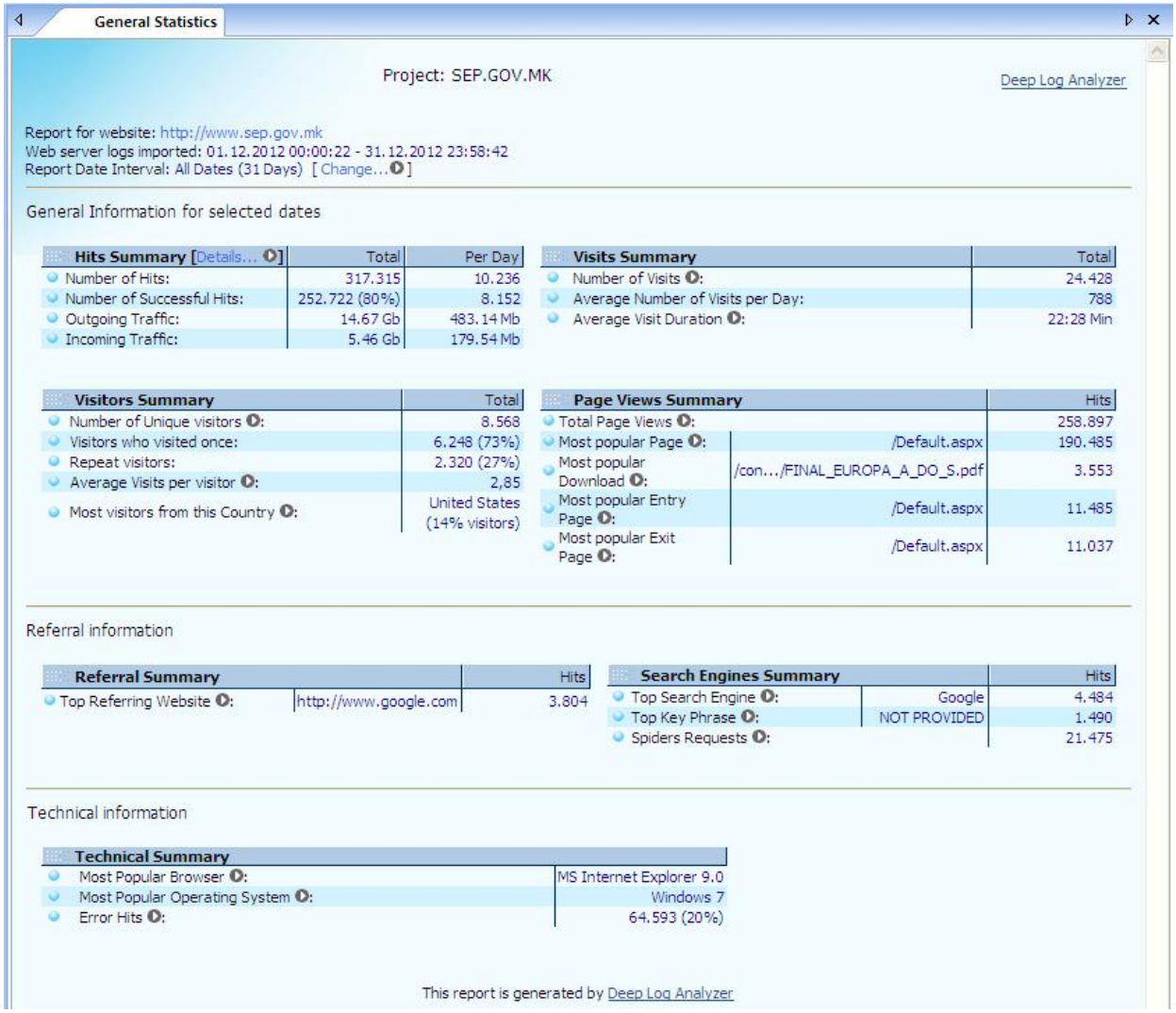
- Напредни веб-статистики и веб-аналитички извештаи презентирани интерактивно
- Анализа на дневник датотеки од сите популарни веб-сервери, дневници превземени преку FTP сервисот, архивирани дневници без да се распакуваат
- Одбирање опсег на датум за статистика и генерирање глобален извештај . Споредба на извештаите за различни интервали
- Преку 40 стандардни извештаи и креирање на единствени сопствени специфични извештаи за сопствената локација во HTML формат или Excel формат
- Deep Log Analyzer користи MS Access MDB база на податоци за складирање на информации, кои се добиени од веб-дневник датотеки. Оваа опција овозможува да се напишат свои прашања(queries), ако е потребно. Програмата не мора повторно да процесира стари денвник датотеки за следниот циклус на анализа.

Може да се зададат автоматски задачи преку скрипти за периодично извршување.

5.2.1 Резултати од експериментот за анализа на дневник датотеката со Deep Log Analyzer

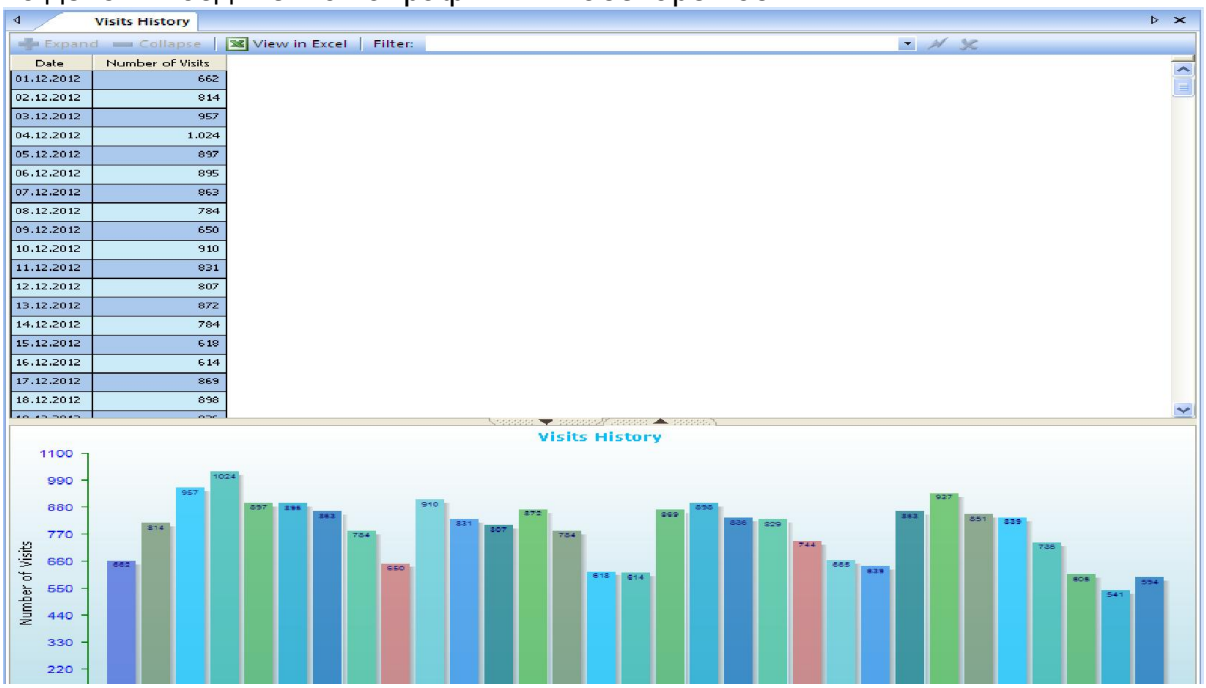
Експерименталното испитување беше спроведено за дневник датотеката од табела 3.1 за периодот од 1.12.2012 до 31.12.2012 година и делумно излезните резултати се прикажани на следните слики:

На слика 5.1 е даден целокупниот преглед на извештај за веб-локацијата на СЕП генериран од веб-дневникот за пристап за периодот од 01.12 до 21.12.2013 година. Од извештајот може да се добијат информации за вкупниот влезен и излезен сообраќај, вкупен број на посети, број на посети по денови и просечното времетраење на посетата, кои се влезно и излезни страни на веб-локацијата, преку кои веб-локации најчесто се доаѓа до веб-локацијата на СЕП и сл.



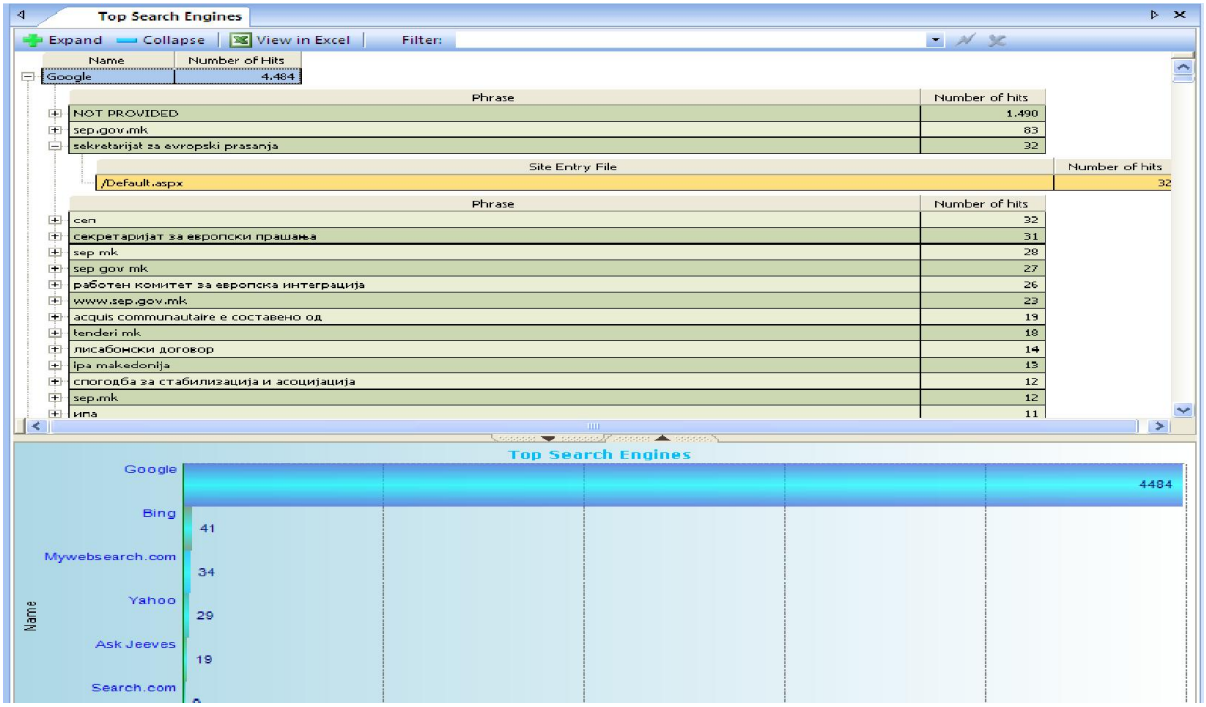
Слика 5.1 Преглед на извештајот за веб-локација
Picture 5.1 Web site overview report

На слика 5.2 е прикажана историја на посетата на веб-локацијата на СЕП по денови поединечно во графички и табеларен облик.



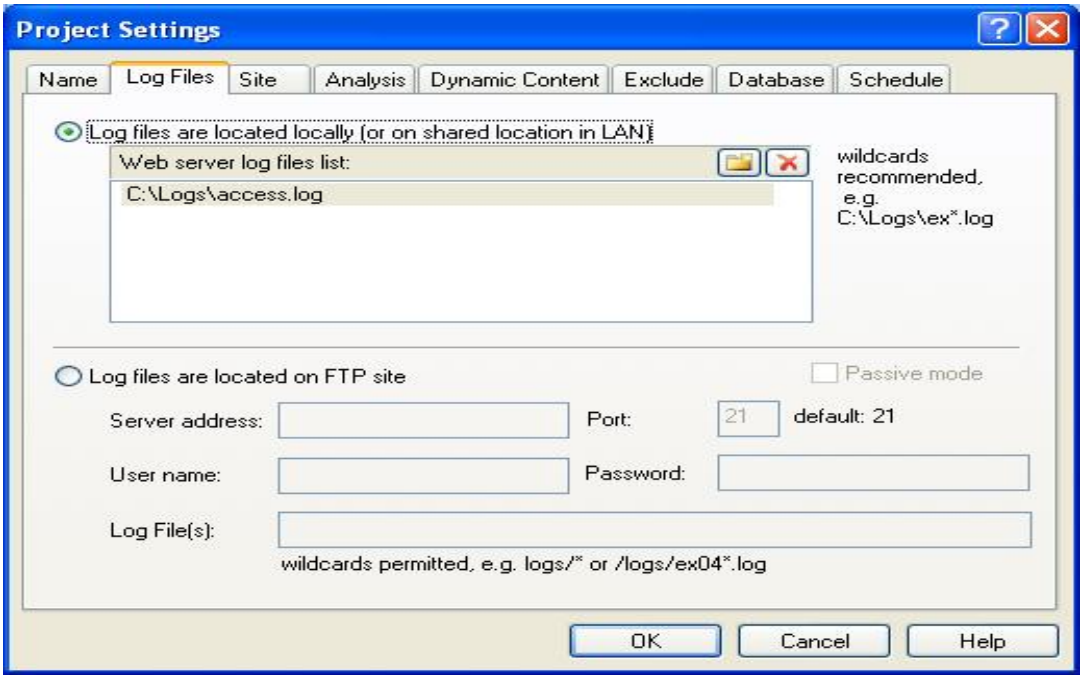
Слика 5.2 Извештај за историја на посети прикажани со графичка шема
Picture 5.2 Visits History report with graphical chart

На слика 5.3 е прикажан извештај за првите топ 5 прелистувачи со три нивоа на хиерахија. Како што може да се види google.com е најповеќе користен прелистувач за пребарување на содржини.



Слика 5.3 Извештај за топ прелистувачи со 3 нивоа на хиерархија
Picture 5.3 Top Search Engines report with 3 levels of hierarchy

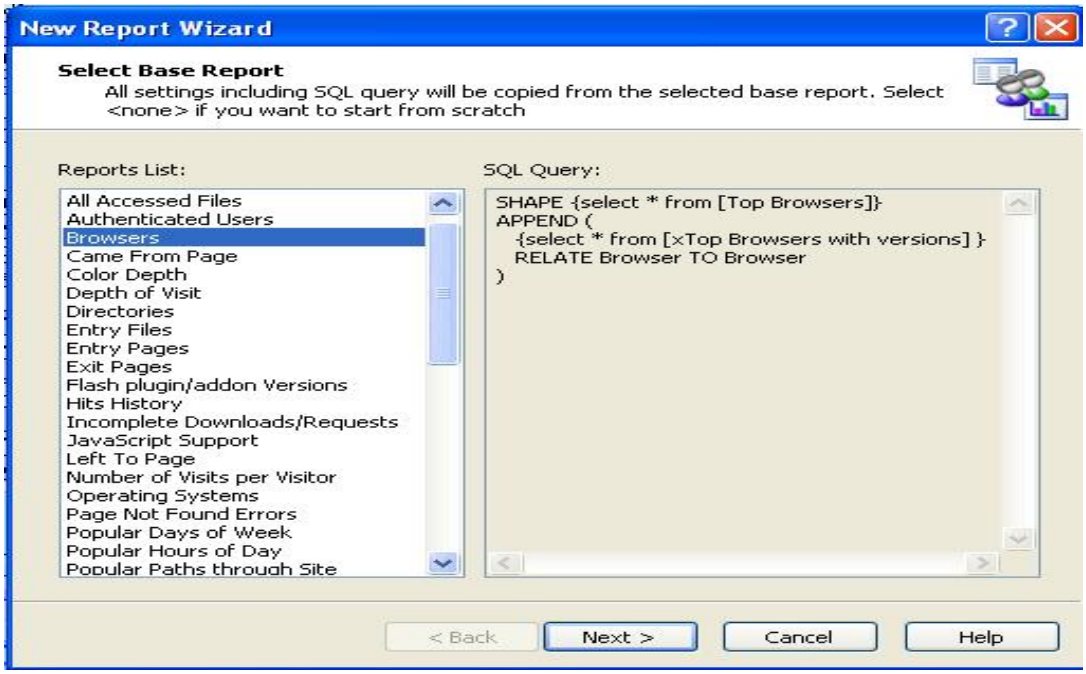
Исто така во зависност од локацијата каде се наоѓа веб-дневникот за пристап, може програмата да се прилагоди и подеси според сопствените потреби слика 5.4.



Слика 5. 4 Подесување на проектот. Патека до локацијата на дневник датотеката

Picture 5. 4 Project settings window. Path to the location of the log file.

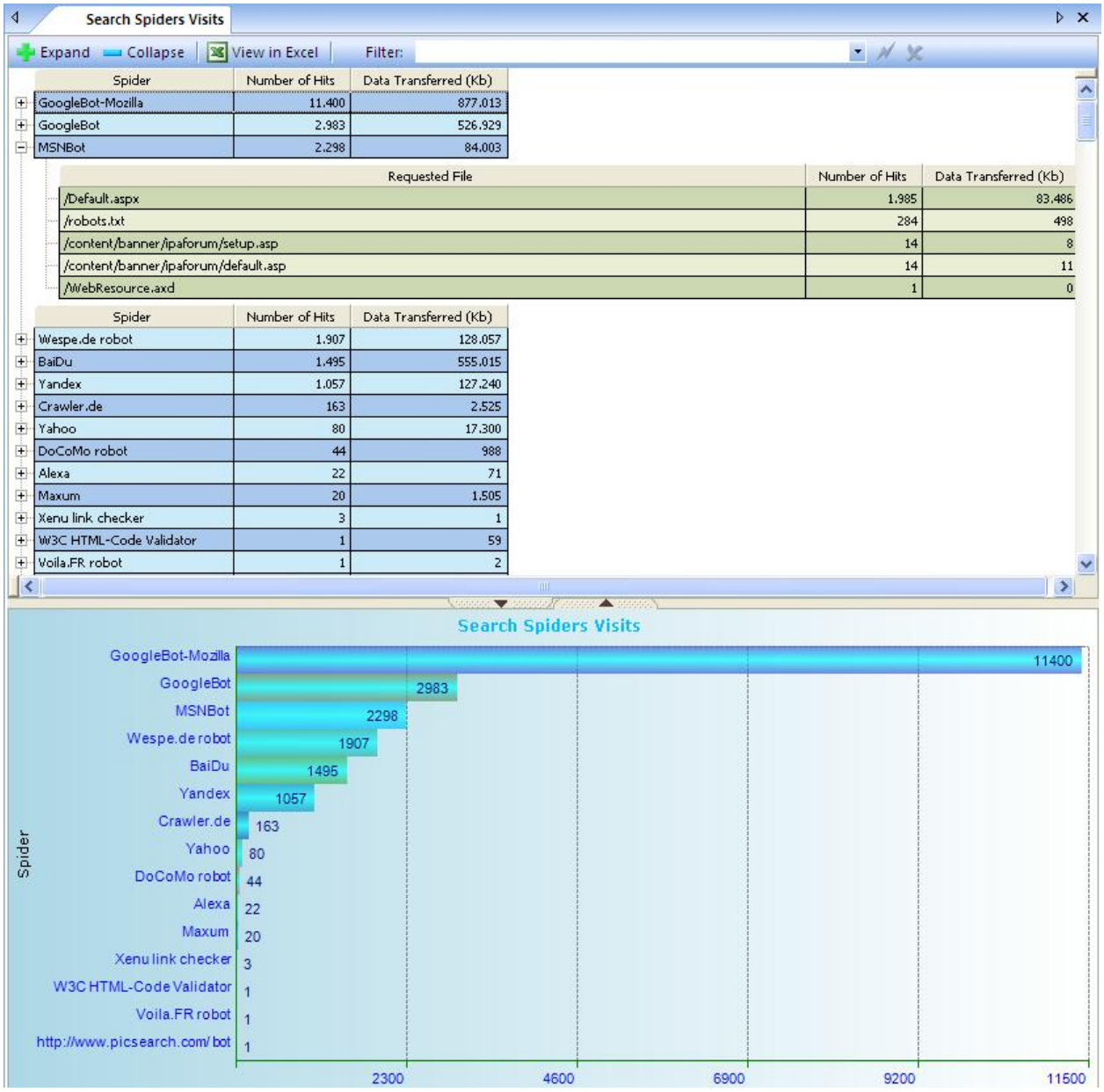
На слика 5.5 е прикажан волшебник за креирање на извештај согласно сопствените потреби и желби.



Слика 5.5 Волшебник за креирање сопствен извештај

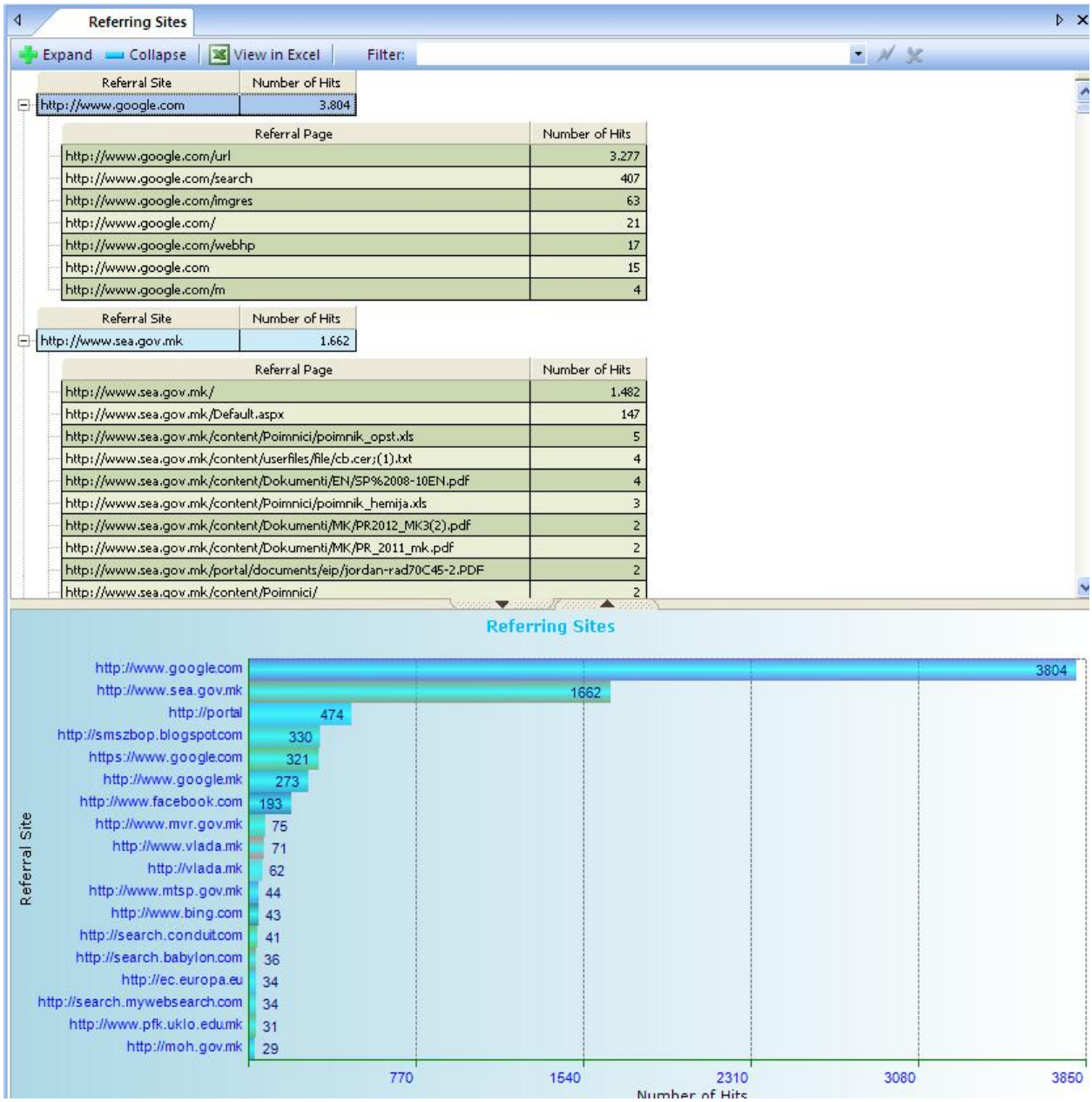
Picture 5.5 New custom report wizard

Исто така овој веб-анализатор овозможува да се видат сите интернет-роботи кои систематски ја пребаруваат веб-локацијата се со цел да изврши индексирање на веб-локацијата слика 5.6.



Слика 5.6 Извештај за “spiders” или “webcrawlers” кои пристапиле до веб-локацијата рангирани по број на хитови
Figure 5.6 Report "spiders" or "webcrawlers" who acceded to the Web Site ranked by number of hits

На слика 5.7 е прикажан список на страници преку кои е дојдено до веб-локацијата на СЕП како и вкупниот вкупниот број(хитови) на датотеки кои се превземени од серверот кога страницата е вчитана.



Слика 5.7 Страници кои ги носат посетителите до веб сајтот кои се рангирани според број на хитови од упатувачите

Figure 5.7 Web pages that bring visitors to the website who are ranked by number of hits from referrer

6. ПРИМЕНА НА ТЕХНИКИТЕ ЗА ПОДАТОЧНО РУДАРЕЊЕ ОД БЕЗБЕДНОСНИ ПРИЧИНИ

6.1 Вовед

Кибернитичкиот простор покрај многубројните примени во целокупниот општествен и приватен живот, исто така се користи интензивно и за електронска трговија и бизнис. Наназад веќе неколку години банките и други финансиски организации спроведуваат трансакции преку интернет со користење на различни географски локации на компјутерски системи. Бизнисите кои прифаќаат трансакции преку интернет може да се стекнат со конкурентска предност во однос на класичните трансакции. Интернетот претставува уникатен збир на безбедносни прашања, поради неговата отвореност и присутност, односно неговата безбедност е клучното прашање за користење на електронската трговија и бизнис. Клиентите доставуваат информации преку веб/интернет само ако се уверени дека нивните приватни информации,(броеви на кредитни картички) се безбедени. Затоа, денешните сервиси базирани на веб/интернет мора да содржат решенија кои обезбедуваат безбедност како примарна компонента во нивниот дизајн и развој.

Веб-сервисите генерално се однесуваат на веб-базирани апликации кои ја овозможуваат таа услуга достапна за претпријатијата да направат трансакции на интернет и за корисниците кои сакаат да ги споделат документите и информациите едни со други преку интернет. Стандард кој прави, односно дава можност да се опише оваа комуникација на некој структуриран начин е Web Services Definition Language (WSDL). WSDL е во XML формат за опишување на веб-сервисите односно нивната функцијата и локацијата на сервисот. WSDL практично претставува договор помеѓу клиентот и серверот на чијашто основа се дефинира што точно овозможува сервисот и на кој начин се врши размената на податоци помеѓу нив.

Но оваа отвореност и интеграција има цена. Без соодветна безбедна заштита и ефикасно управување со безбедноста, овие карактеристики може да се злоупотребат за напад на достапноста и интегритетот на информациските системи и нивната мрежна поврзаност. Постојат неколку типични начини на напаѓачот со кои може да се добие незаконски пристап до системот за

информации, или да се направи тој да биде недостапен за легитимните корисници. За откривање на вакви напади потребно е идентификација на профили или потписи за низа дејствија на напаѓачот. Со користење на техниките на податочно рударење може да се откријат напади како профили на напади.

Податочното рударење се третира како техника за интелигентна и автоматска помош на луѓето при анализа на големи количини на податоци за да се идентификуваат валидни, несекојдневни, и потенцијално корисни шеми на податоци. Оваа техника дава ветување во помагање на организациите дека ќе најдат скриени шеми на однесувања во нивните податоци кои понатаму може да се искористат за да се предвиди однесувањето на потрошувачите, така што тие можат подобро да ги планираат нивните производи и процеси.

Клучот за спречување на безбедносни напади е да се разбере и да се идентификуваат слабостите и да преземе соодветна акција. Напад е спроведувањето на закана со користење на систем од слабости. Ранливост е слабост во безбедносниот систем со кој кои би можеле да се изврши напад. За спречување на сето ова, контролата е заштитна мерка - една акција, уред, постапка, или техника - која ги намалува слабостите.

Во табела 6.1 се првите 10 безбедносни слабости како што е соопштено периодично од страна на Sans Институтот. Причините за постоењето на овие слабости се: лабав софтвер за дизајн и развој, систем-администраторите се премногу зафатени за да инсталираат безбедносни закрпи навремено, и несоодветни политики и процедури. Друг фактор е присутноста на интернет, каде пропустите се објавуваат брзо и насекаде.

Мотивот на напаѓачите може да биде од чисто хакирање од финансиска корист. Напаѓачите се или високо технички оспособени, или тие понекогаш провалуваат во мрежата преку обиди и грешки. Незадоволни вработени имаат повеќе права за пристап до компјутерските мрежи во компанијата во споредба со напаѓачите од надвор.

Табела 6.1 Топ 10 безбедносни слабости
Table 6.1 Top 10 security vulnerabilities

Ранливости на Windows системите	**Ранливости на Unix системите**
1. Internet Information Services (IIS)	1. Remote Procedure Calls (RPC)
2. Microsoft Data Access Components (MDAC) -- Remote Data Services	2. Apache Web Server
3. Microsoft SQL Server	3. Secure Shell (SSH)
4. NETBIOS -- Unprotected Windows Networking Shares	4. Simple Network Management Protocol (SNMP)
5. Anonymous Logon -- Null Sessions	5. File Transfer Protocol (FTP)
6. LAN Manager Authentication -- Weak LM Hashing	6. R-Services -- Trust Relationships
7. General Windows Authentication -- Accounts with No Passwords or Weak Passwords	7. Line Printer Daemon (LPD)
8. Internet Explorer	8. Sendmail
9. Remote Registry Access	9. BIND/DNS
10. Windows Scripting Host	10. General Unix Authentication -- Accounts with No Passwords or Weak Passwords

6.2 Типични безбедносни напади

Безбедносниот напад се случува кога напаѓачот користи една или повеќе безбедносни слабости. Да се подобри безбедноста, треба да се намалат безбедносните пропусти. Во овој дел се презентираме некои типични безбедносни напади, кои укажуваат на слабостите кои се злоупотребуваат за извршување на напади.

6.2.1 Denial-of-Service Attack

Denial-of-Service (DoS) напад е организиран така што пречи или потполно го застанува нормалното функционирање на веб-локациите, серверите или мрежните ресурси. Постојат различни начини со кои хакерите го постигнуваат ова. Една од највообичаените начини е едноставно праќање преголем број на барања кон серверот. Нападот е успешен, ако легитимни корисници повеќе не може да пристапат до ресурсите и услугите кои се понудени од страна на серверот кои ќе бидат недостапни.

Нападите може да бидат насочени кон оперативниот систем или на мрежата. Напаѓачот може да испрати специјално направени пакети кон софтверските сервиси кои работат на серверот кој ќе биде жртва. Тој ќе биде успешен ако мрежата не е во можност да се направи разлика меѓу легитимниот сообраќај и малициозниот или лажниот сообраќај.

6.2.2 SQL Injection

Нападот со подметнување на SQL наредба (eng. SQL injection) е метода која ја злоупотребува сигурносната ранливост кај базите на податоци на веб-апликацијата. Тоа е најрашириениот и најопасниот напад кој се случува на апликацискиот слој TCP/IP. Спомнатата ранливост напаѓачот ја искористува со преименување на SQL наредбата со помош на некои знаци и потоа веб-апликацијата ја испраќа до базата на податоци така што се откриваат чуствителни податоци или се извршуваат недозволени команди над нив. Со помош на подметнување на SQL наредба напаѓачот може да преземе потполна контрола на податоците во базата.

6.2.3 Cross-Site Scripting (XSS)

Cross-Site Scripting (XSS) е еден од најчестите облици на хакерски напади на интернет. XSS нападите претежно се насочени кон корисниците, а поретко кон скриптите на серверите. Тие се користат со манипулација на скриптите на страната на веб-серверот и потоа тие секогаш се извршуваат кога ќе се исчита страницата и се извршуваат на начин на кој сака злонамерникот.

Главни карактеристики на XSS нападот

- XSS напади тесе извршуваат на ранливи веб-апликации
- Во XSS нападите жртва е корисникот, а не апликацијата
- Во XSS нападите злонамерната содржина се испорачува на корисниците со помош на JavaScript-а

XSS ранливоста настанува кога веб-апликацијата ги зема податоците од корисникот и потоа динамички ги вклучува во веб-скриптата без претходна проверка.

6.2.4 HTTP GET attack

Друга можност за искористување на ранливоста на веб-апликациите исто така е и преку HTTP протоколот. За многу веб-апликации, потребно е клиентот да биде во можност да испрати информација до серверот. HTML 2.0 и подоцнежните верзии поддржуваат во рамките на HTML документот да им се овозможи на податоци да бидат испратени до веб-серверите. Еден од атрибутите е метод(Method) што укажува на тоа како податоците се доставуваат до веб-серверот. Валидни вредности за атрибутот Method се GET и POST. Во METHOD=GET, внесените вредности од страна на корисникот се спојуваат со URL-то, одвоени со посебен карактер (обично?); полињата се одделени со &; простор е претставен со +.
На пример, следниов URL:
http://www.gadgets.com?customer=John+Doe&address= 101+Main+Street&cardno=1234567890&cc=visacard укажува на тоа дека името на корисникот и адресата со бројот на кредитната картичка се испраќаат на веб-серверот во www.gadgets.com. Корисникот (напаѓачот) може да биде во можност да ја користи оваа функција да добие пристап до комерцијални информации доколку соодветните безбедносни механизми не се соодветно подесени.

6.3 Техники за податачно рударење за откривање на упад

Во овој дел ќе се разгледуваат веб-дневниците од серверот, вовед во напад на потписи и презентација како безбедносниот напад на потписи се користи во комбинација со податочното рударење за да се открие безбедносен упад. Поконкретно, прво се опишува релевантноста и важноста на различните дневник датотеки кои се на располагање, па потоа се дефинираат специфичните шеми на податоци во дневник датотеките за напад како напад

на потпис и потоа ќе се користат техники на податочно рударење за пребарување. Ефикасноста и брзината на целокупниот процес дури и може да доведе до предвидување на напад.

6.3.1 Дневници(Logs)

Секоја посета на веб-локацијата од страна на корисникот создава запис на она што се случува за време на таа сесија во дневникот на серверот. Ако веб-локацијата е многу посетена може да се генерираат илјадници записи во дневникот. Веб-дневник датотеката содржи информации за IP адресата на компјутерот од посетителот на веб-страницата, датум и време, име и големината на датотеката која се пребарува. Дневниците може се разликуваат според видот на серверот и форматот на датотеката.

Пример за неколку типични дневници и што тие содржат:

• Access Log - секоја трансакција помеѓу серверот и пребарувачите (датум, време, името на доменот или IP-адреса, големината на трансакцијата).

• Referrer Log - води евиденција за патот на посетителот до локацијата (почетна URL од која дошол посетителот).

• Agent Log – води евиденција за типот и верзијата на прелистувачот.

6.3.2 Податочно рударење на логови(Mining Logs)

За да се добијат корисни информации податоците кои се достапни од веб-дневниците треба треба да се “рударат”. Податочното рударење на податоците нуди ветување во откривањето на скриени шеми на податоци кои може да се искористат за да се предвиди однесувањето на (злонамерени) корисници. Користењето на техниките за податочно рударење во Intrusion Detection е релативно нов концепт. Главната идеја во изградбата на Intrusion Detection е да се користат ревизорски програми со чија помош ќе се извлечат обемен сет на функции кои ги опишуваат мрежните конекции или host сесиите, и потоа се применуваат техники на податочни рударења со правила на учење со кои ќе се запишува однесувањето при упади и нормални активности. Моделите за откривање на нови упади се вклучени во Intrusion Detection системите (IDS) преку meta-learning (или ко-оперативец) процесот на учење. Силата на овој приод е во класификацијата, мета-учење, и асоцијативните (здружените) правила.

6.3.3 Attack signatures

Attack signatures претставува уникатно уредување (комбинација) на информации кои може да се користат за идентификација на обид на напад. Во комбинација со податочно рударење не само што може да се детектираат, туку може однапред да се предвидат нападите. Attack signatures го концизира начинот на кој напаѓачот ќе се движи низ ресурсите и акции кои ќе ги превземе. На пример, во denial-of-service напади, напаѓачот може да испрати голем број на речиси истовремени барања за TCP конекции од една или повеќе IP адреси без одговарање на серверот дека потврдил прием.

Постојат најразлични видови на потпис - напади и секогаш треба да се обрне внимание да се направи сеопфатна анализа за откривање на овие напади. Нецелосните напад - потписи резултираат со лажно-позитивни или лажно-негативни детекции. Друг важен момент како што е спомнато и погоре е употребата на податочно рударење за да се откријат овие напади. Мора да се има предвид дека количината на дневници собрани од веб-сервисите може да резултираат во мулти-тера бајти на бази податоци. Со такви големи количини на податоци и доколку примената на податочно рударење не биде ефикасна тогаш цела оваа постапка станува исклучително скапа.

6.3.4 Безбедносни заштитни мерки (Security Safeguards)

Основната цел на примената на заштитните мерки е намалување на ризикот на безбедноста на едно прифатливо, односно пожелно ниво. Тие можат да бидат проактивни за да се спречат безбедносни или реактивни инциденти, да се заштитат информациите кога веќе инцидентот е откриен. Во секој случај, тие мора да бидат ефективни, да не се заобиколат и со минимално влијание врз работењето. Примери на заштитни мерки се: избегнување (водејќи евиденција за безбедносните инциденти кои се случуваат, на пример, отстранување на слабости), ограничувањето на пристапот (на пример, преку намалување на бројот на влезни точки од каде нападите може да потекнуваат), пренос (префрлување одговорност на ризик за некој друг, на пример, преку осигурување или надворешна подршка) и за ублажување (минимизирање на влијанието врз инцидентот, на пример, преку намалување на својот делокруг или подобрување на откривањето на нападот).

Една од главните заштитни мерки е да се открие и да се намалат односно да се отстранат слабостите. Главните причини за постоечките слабости се лабавиот софтвер (дизајн и развој) или проблемите поврзани со систем-администрација. Постоењето на грешки во софтверот најчесто се должи на:

•не се научени основните постулати за безбедноста при програмирање,

•софтвер процесите не се универзални, како и

•постоење на наследен(преземен) код.

Проблемите поврзани со систем-администрација се должи на несоодветните политики и процедури или систем-администраторите се премногу зафатени со многу машини за администрирање, премногу платформи и апликации за поддршка и премногу надградби и закрпи за примена.

6.4 AQUNETIX

Како што е спомнато во делот 6.1.1 безбедноста на веб и мрежните сервиси е од круцијално значење и треба да биде приоритет во секоја е-област. Постојат голем број на програми(алатки) кои комуницираат со веб-апликациите со цел автоматски да се идентификуваат потенцијалните технички, безбедносни, архитектонски/логички и други слабости на веб-апликациите. Во табела 6.2 е прикажана листа на алатки за скенирање безбедноста на веб-страниците.

Табела 6.2 Листа на алатки за скенирање на веб-апликации
Table 6.2 List of web application security scanners

Комерцијални софтвери	**Software-as-a-Service Providers**	**Отворен / Слободен код**
Acunetix WVS by Acunetix	AppScan OnDemand by IBM	Arachni by Tasos Laskos
AppScan by IBM	ClickToSecure by Cenzic	Grabber by Romain Gaucher
Burp Suite Professional by PortSwigger	QualysGuard Web Application Scanning by Qualys	Grendel-Scan by David Byrne and Eric Duprey
Hailstorm by Cenzic	Sentinel by WhiteHat	Paros by Chinotec
N-Stalker by N-Stalker	Veracode Web Application Security by Veracode	Andiparos
Nessus by Tenable Network Security	VUPEN Web Application Security Scanner by VUPEN	Zed Attack Proxy
NetSparker by Mavituna		Powerfuzzer by Marcin

Security NeXpose by Rapid7 NTOSpider by NTObjectives ParosPro by MileSCAN Technologies Retina Web Security Scanner by eEye Digital Security WebApp360 by nCircle WebInspect by HP WebKing by Parasoft Websecurify by GNUCITIZEN	Security WebInspect by HP WebScanService by Elanize KG	Kozlowski SecurityQA Toolbar by iSEC Partners Skipfish by Michal Zalewski W3AF by Andres Riancho Wapiti by Nicolas Surribas Watcher by Casaba Security WATOBO by siberas Websecurify by GNUCITIZEN Zero Day Scan

Acunetix WVS е комерцијален софтвер кој врши тестирање на веб-апликации и нивната ранливост. Бидејќи може да се испроба бесплатно и истиот е оценет како еден од најдобрите алатки беше одбран за тестирање на веб-локацијата на Секретаријатот за европски прашања.

Acunetix Web Vulnerability Scanner вклучува многу иновативни функционалности, но ќе се напоменат само неколку најважни:

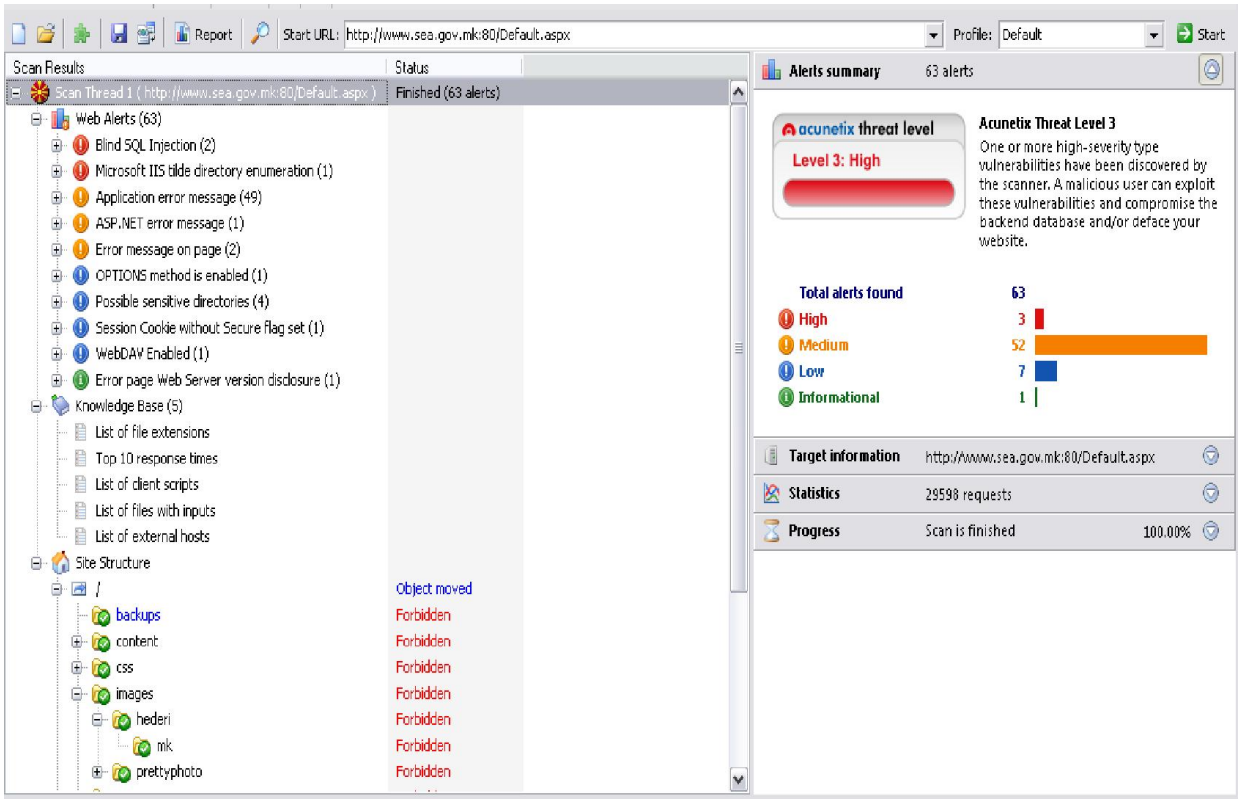
- *Скенирање AJAX и Web 2.0 технологија на* ранливост - Врвниот CSA (Client Script Analyzer) Engine овозможува темелно скенирање на најактуелните и најкомплексните AJAX / Web 2.0 веб-апликации и пронаоѓање на веб-ранливост;
- *Длабинско тестирање на SQL Injection, Cross Site Scripting (XSS) и останати ранливости со користење на AcuSensor технологија* – важно е да се напомене дека за скенирање на веб-ранливоста не е важен само бројот на напади кој скенерот може да ги детектира туку комплексноста и деталноста со кој скенерот ги лансира SQL Injection, Cross Site Scripting и останатите напади. AcuSensor технологијата врз основа на мал број предупредувања за ранливост покажува каде се

наоѓаат пропустите внатре во кодот;

- *Тестирање на зоните заштитени со лозинка и веб-формуларите со користење на автоматски пополнувач на формуларот* – со користење на алатка за снимање макро (Login Sequence Recorder) може да се сними логин секвенца или процесот на пополнување формулар;
- *Port Scanning i Network Alerts* – скенирање на порти во однос на веб-серверот на кој е хостирана веб-локацијата и автоматски ги идентификува мрежните сервиси кои користат отворени порти, лансирајќи серија на безбедносни тестови на тие сервиси;
- Анализа на локацијата во однос на Google Hacking Database - (GHDB) е база на прашалници кои се користени од страна на хакери со цел да се идентификуваат осетливи информации(логон форми, информации за мрежна безбедност и итн.) или други важни точки кои може да се злоупотребат пред хакерот да ги примени;
- Скенирање и анализа на веб-локации кои содржат flash содржини, SOAP и AJAX.

6.4.1 Резултати од експериментот за испитување сигурност на веб-локација со ACUNETIX скенерот за веб ранливост

Слабости кои беа откриени за време на скенирањето на веб-страницата се прикажани во реално време на слика 6.1. Исто така е прикажана и „Мапа на структура на веб-локацијата" со листата на датотеки и папки кои беа откриени.



Слика 6.1 Резултат од скенирањето и информациски прозорец
Picture 6.1 Scan Result and Information window

Слабостите кои беа откриени се категоризирани според 4 нивоа:

Ниво 3 висок ризик (High Risk Alert Level 3) – Слабости кои се категоризираат како најопасни и ја поставуваат веб-локацијата на максимален ризик за хакирање и кражба на податоци.

Ниво 2 среден ризик (Medium Risk Alert Level 2) – Пропусти предизвикани од конфигурацијата на серверот и недостатоци на кодот од локацијата, со кои се олеснува нарушување и упад на серверот.

Ниво 1 низок ризик (Low Risk Alert Level 1) – Слабости кои произлегуваат од недостатокот на енкрипција на сообраќајот на податоци, или разоткривање на патеката до директориумот.

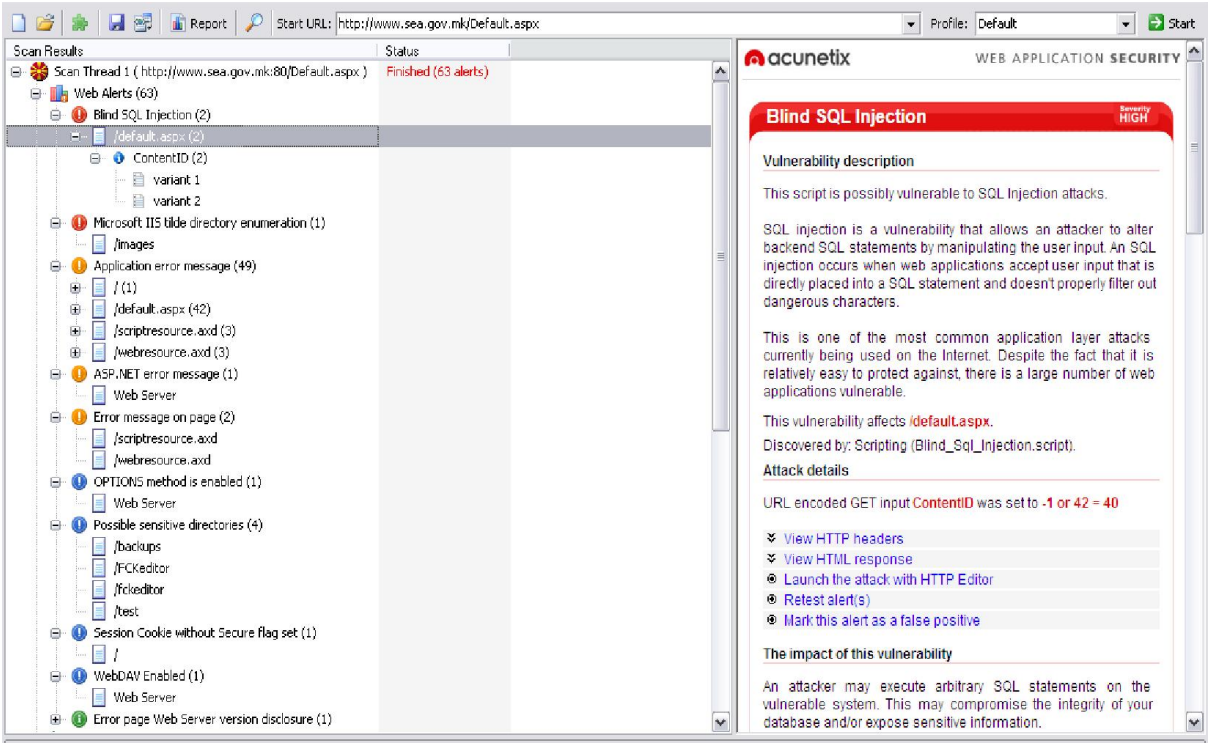
Информативен сигнал (Informational Alert) – Локации кои се подложни на откривање информации преку Google хакирање или откривање е-пошта.

Од слика 6.1 може да се забележи дека се откриени вкупно 63 слабости на веб-локацијата на СЕП и тоа:

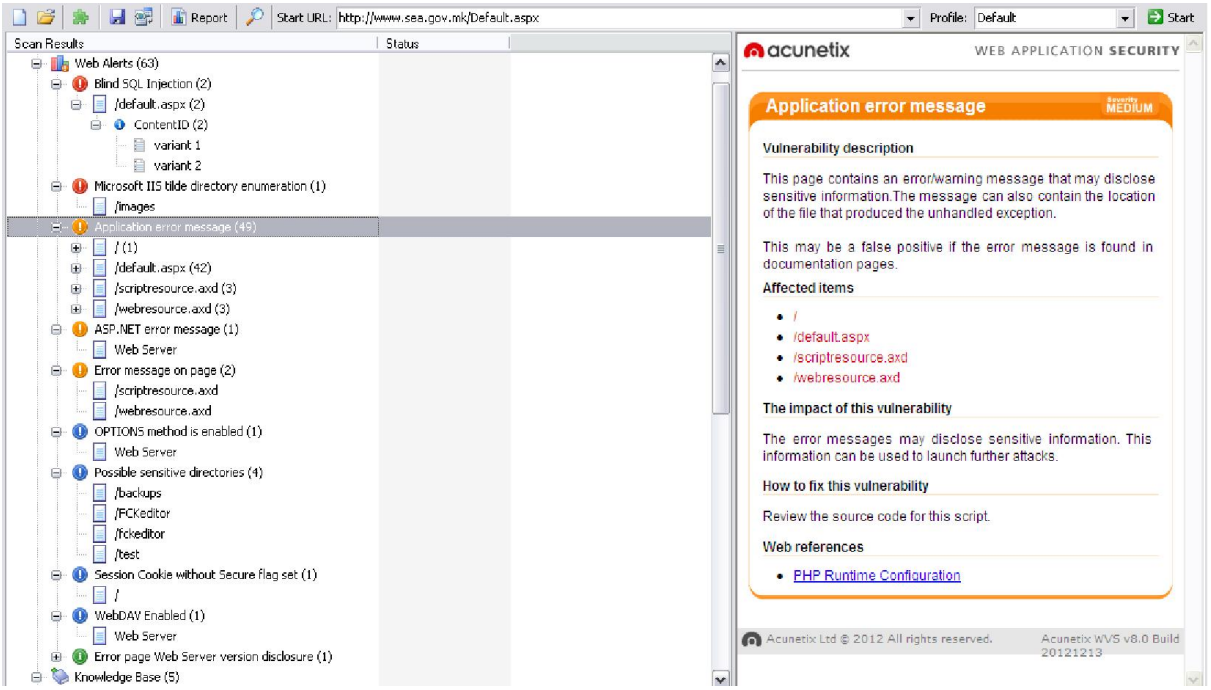
- *Ниво 3 висок ризик (High Risk Alert Level 3)* – најдени се вкупно 3

слабости слика 6.1 од кои две слабости за можен SQL Injection напад како што е објаснато во 3.2 и една слабост за можен напад преку кој може да се пристапи до важни датотеки и папки кои нормално не се видливи.

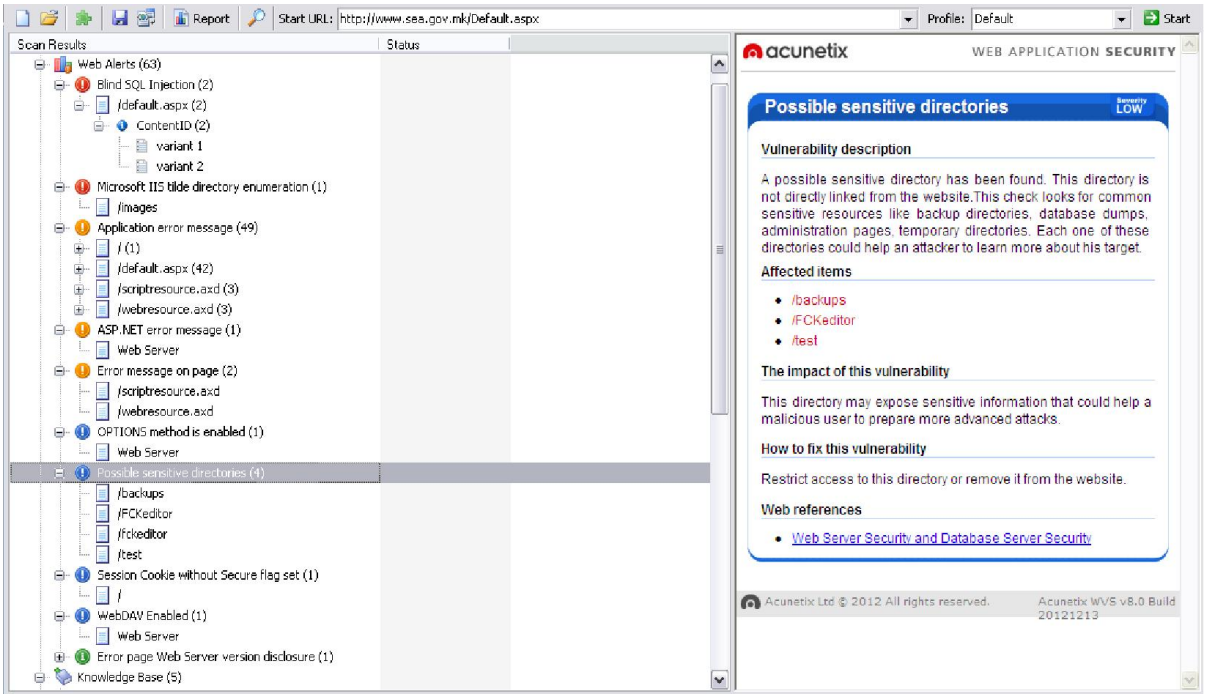
- *Ниво 2 среден ризик (Medium Risk Alert Level 2)* – најдени се вкупно 52 пораки за грешки, слика 6.1 и се однесуваат на кодот на апликацијата и со соодветни препораки каде да се изврши ревизија на изворните кодовите.
- *Ниво 1 низок ризик (Low Risk Alert Level 1* - најдени се вкупно 7 слабости, слика 6.1 од кои една се однесува на HTTP методот OPTIONS преку кој може хакерите да подготват понапреден напад и препораката е да се оневозможи користење на OPTIONS, четири се однесуваат на можноста да се откријат чуствителни папки со препорака да се има исклучително рестриктивен пристап до нив или да се тргнат од веб локацијата и уште две кои се однесуваат на сесиски колачиња со препорака на користење на безбедносни знамиња за сесиите да бидат прифатени само преку SSL канал и користење на WebDAV(далечинки корисници менуваат содржина) екстензијата на HTTP протоколот да биде правилно конфигурирана или да се исклучи.
- *Информативен сигнал (Informational Alert)* – најдена е една слабост при пребарување на страницата која не постои односно страницата со грешка враќа информација за верзијата на веб серверот и листата на модули кои се достапни.



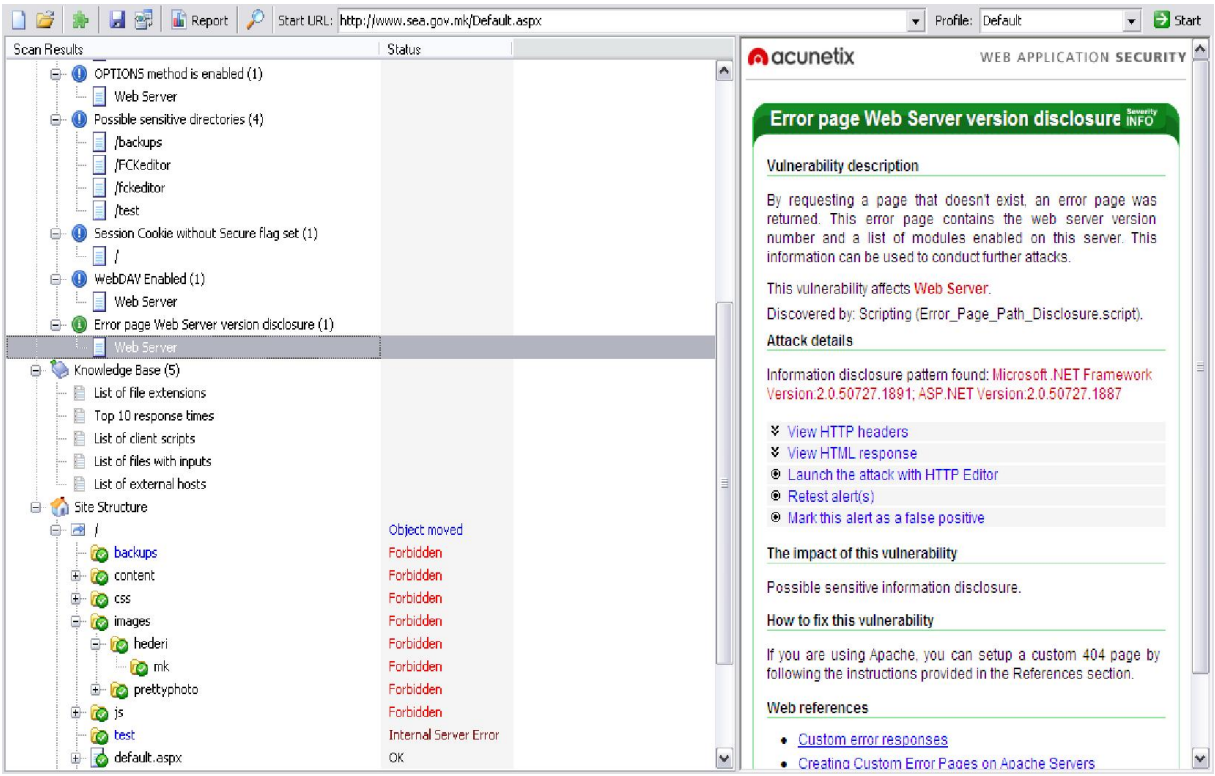
Слика 6.2 Резултат од скенирањето со високо ниво на ризик
Picture 6.2 The result of the scan with a high level of risk



Слика 6.3 Резултат од скенирањето со средно ниво на ризик
Picture 6.3 The result of the scan with a medium level of risk



Слика 6.4 Резултат од скенирањето со ниско ниво на ризик
Picture 6.4 The result of the scan with a low level of risk



Слика 6.5 Резултат од скенирањето со информативен карактер
Picture 6.5 The result of scanning for information

7. Заклучоци и идна работа

7.1 Заклучоци

Целта на експериментот 1 е да се споредат пристапните шеми помеѓу посетителите од Македонија и посетителите надвор од Македонија. Посетителите од Македонија најмногу ја посетуваат изворната страница, но исто така и директно посетуваат страници, а посетителите надвор од Македонија најчесто директно ги посетуваат специфичните страници. Ова е веројатно затоа што тие ги користат пребарувачите. Неколку резултати покажуваат дека посетители надвор од Македонија претежно пристапуваат до страници поврзани со претпристапна поддршка (twining) и друга странска помош, процесот на пристапување на Македонија кон ЕУ, преведување и координација на процесот на подготовка на националната верзија на правото на ЕУ. Некои резултати покажуваат дека интересот на посетителите од Македонија се страници поврзани со организацијата и работата на СЕП, новостите, огласите и конкурсите во СЕП. Затоа, администраторот на интернет- страницата треба да овозможи информациите во овие страници секогаш да се точни и тековни.

Откривањето на шеми според кои пристапуваат корисниците може да послужи за преуредување на веб-локацијата.

Целта на експериментот 2 е да се споредат пристапните шеми помеѓу посетители кои се од СЕП и посетители надвор од СЕП. Резултатите покажуваат дека посетителите од СЕП најмногу ја посетуваат изворната страница, но исто така, и директно посетуваат страници и ова се должи на фактот дека вработените во СЕП доволно добро ја познаваат веб-страницата. Посетителите надвор од СЕП најчесто ја посетуваат изворната страница. Дополнителна откриена шема е дека посетителите надвор од СЕП имаат тенденција да дојдат до информации за институцијата, новостите, процесот на преведување, документи, односно за регистарот на документи, регистрација на проекти, претпристапна поддршка, процесот на преведување, кариера и можности за вработување. Тоа е веројатно затоа што посетителите имаат тенденција да погледнат информации за можностите за вработување,

искористување на странската претпристапна помош и регистрација на проекти за искористување финансиска неповратна помош.

Целта на експериментот 3 е да се споредат пристапните шеми помеѓу посетителите од СЕП и посетителите надвор од СЕП, но од Македонија. Некои од откриените шеми се во согласност со оние откриени во експериментите 1 и 2. Откриената шема покажува дека посетителите надвор од СЕП обично поминуваат помалку време на страниците, во споредба со посетителите од рамките на СЕП. Може да се заклучи дека оваа разлика во поминатото време може да се должи на фактот дека посетителите од СЕП ја користат интернет-страницата како извор на податоци за нивните тековни обврски и ги следат сите новости, измени и други информации. Исто така, посетителите надвор од СЕП од Македонија, би можеле да го одложат пребарувањето поради одредени пречки во мрежата.

Беше забележано дека групите на функции First3-Last2 и First5-Last5 се прилагодени и даваат корисни шеми за асоцијативни правила, додека групите на функции 10-Most-Frequent-TF и 10-Most-Frequent-Time се погодни за техниката на класификација.

Збирно, може да потврдиме дека беа откриени три главни шеми: (1) Посетителите од Македонија најмногу ја посетуваат изворната страница, но исто така и директно посетуваат страници, а посетителите надвор од Македонија најчесто директно ги посетуваат специфичните страници. Како и да е, некои (2) од посетителите надвор од СЕП, но од Македонија имаат тенденција да дојдат до информации за институцијата, новостите, процесот на преведување, документи, односно за регистарот на документи, регистрација на проекти, претпристапна поддршка, процесот на преведување, кариера и можности за вработување. Тоа е веројатно затоа што посетителите имаат тенденција да погледнат информации за можностите за вработување, искористување на странската претпристапна помош и регистрација на проекти за искористување финансиска неповратна помош и (3) посетителите надвор од СЕП обично поминуваат помалку време на страниците, во споредба со посетителите од рамките на СЕП. Може да се заклучи дека оваа разлика во поминатото време може да се должи на фактот дека посетителите од СЕП ја користат интернет-страницата како извор на податоци за нивните тековни

обврски и ги следат сите новости, измени и други информации поврзани со процесот на интеграција на Република Македонија во Европската унија.

Со примена на алатките за веб-анализа и испитување на веб-ранливоста се забележаа доста значајни информации. Со алатката ACUNETIX WVS беше забележано дека се откриени вкупно 63 слабости на веб-локацијата на СЕП и тоа: ниво на висок ризик најдени се вкупно 3 слабости од кои две слабости за можен SQL Injection и една слабост за можен напад преку кој може да се пристапи до важни датотеки и папки кои, нормално, не се видливи, ниво на среден ризик најдени се вкупно 52 пораки за грешки и се однесуваат на кодот на апликацијата, ниво на низок ризик најдени се вкупно 7 слабости од кои една се однесува на HTTP методот OPTIONS преку кој може хакерите да подготват понапреден напад, четири се однесуваат на можноста да се откријат чуствителни папки и уште две кои се однесуваат на сесиски колачиња и на ниво на информативен сигнал најдена е една слабост при пребарување на страницата која не постои, односно страницата со грешка враќа информација за верзијата на веб-серверот и листата на модули кои се достапни.

Исто така, со примена на алатката за веб-анализа Deep Log Analyzer се дојде до комплетна статистика за користењето на веб-локацијата и до информации за пристапени ресурси од веб-локацијата, активноста на посетителите и навигацијата, веб-локации преку кои е дојдено до анализираната локација, како пребарувале за да пристапат до веб-локацијата, лизгачи кои ја пребаруваат локацијата, кои прелистувачи и оперативни системи ги користи посетителот, грешки на веб-серверот и друго.

7.2 Идна работа

Оваа магистерска работа ги покрива основите на рударењето на веб-податоци со примена на техники за податочно рударење, но има уште многу што може да се направи со овие податоци. Иако оваа истражувачка работа е процес во кој се одзема многу време, но со поцелосна и поопфатна анализа може да се дојде до значајни знаења и корист кои ќе бидат стекнати во текот на овој процес. За таа цел, треба да се извршат дополнителни работни активности во сите експерименти. Во дополнителна работа може да се вклучат анализи на трансакции со помалку од 5 посетени страници бидејќи реално постојат такви посети и во нив може да се откријат интересни шеми на пристап.

Поради ограниченото време и хардверските ресурси во оваа магистерска работа земан е краток временски интервал од дневникот на веб пристапните дневници за анализа и сметам дека во иднина дека доколку се изврши анализа или споредба на различни или подолги временски периоди ќе се добијат дополнителни интересни знаења и шеми. Во претпроцесирачката фаза од податочното рударење многу е важно да се изврши успешна идентификација на посетителот, односно дали посетителот е човек или веб робот (лизгач). Во горе наведените експерименти се сметаше дека ако суфиксот „.mk"се наоѓа во host name на машината на посетителот, тогаш тој е од Македонија. Но, исто така има случаи кога не може да се најде host-името иако се знае од која IP адреса доаѓа посетителот и затоа треба да се применат точни и софистицирани техники за идентификување на локации и посетители кои ќе овозможат поточни информации, а со тоа крајните резулати ќе бидат подобрени. Друга активност која сигурно ќе придонесе точни информации за посетата на веб-локацијата е имплементација на знаење за структурата на сајтот, колачиња и завршна патека. Знаењето за целокупната структура на сајтот при анализа на пристапните дневници ќе придонесе за пополнување на страниците кои фалат при посетата од корисникот, а со тоа ќе се обезбедат поточни информации за посетата. Употребата на колачиња ќе овозможи управување на посетителите и серверите и посигурна идентификација на посетителот. Исто така, посебен акцент треба да се стави на веб-роботите(лизгачи) кои не пристапуваат до датотеката „robots.txt" и истите прават записи во веб пристапните датотеки. Доколку не се идентификуваат успешно, овие роботи ќе имаме неточни податоци до одреден степен, па затоа потребно е да се користат информации за веб-прелистувачи со кои подобро ќе се идентификуваат истите и крајните резулати ќе се подобрат.

Имајќи го предвид фактот дека сигурноста на веб-локацијата и нејзината ранливост се многу важни како идна работа во согласност со алатката ACUNETIX WVS и посочените ранливости вкупно на број 63, треба да се преземат дополнителни активности. Најпрво потребно е да се отстранат сите можности за SQL напади, да се направи ревизија(прекодирање) на посочените ранливи кодови во согласност со стандардите и препораките за безбедност, да

се направи порестриктивен пристап до почуствителните папки и да се користат исклучиво SSL канали при прифаќање безбедносни знамиња за сесиите.

Добиените сеопфатни информации и статистички извештаи од алатката Deep Log Analyzer може да помогнат за преземање на активности за идна работа за оптимизација и изработка на адаптивни страници за да се привлечат повеќе посетители и истите да се претворат во задоволни посетители.

КОРИСТЕНА ЛИТЕРАТУРАТА

1. Acunetix Web Vulnerability Scanner, http://www.acunetix.com/vulnerability-scanner/

2. Akshay Upadhyay, Balram Purswani. Web Usage Mining has Pattern Discovery. International Journal of Scientific and Research Publications, Volume 3, Issue 2, February 2013

3. Ankit R Kharwar, Viral Kapadia, A Complete PreProcessing Method For Web Usage Mining. Ganpat university journal of engineering & technology, volume 1, issue 1, March 2011

4. B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalization based on association rule discovery from web usage data. In Proc. of the 3rd ACM Workshop on Web Information and Data Management (WIDM01),Atlanta, 2001.

5. B. Mobasher, R. Cooley, and J. Srivastava. Creating adaptive web sites through usage-based clustering of urls. In IEEE Knowledge and Data Engineering Workshop (KDEX'99), 1999.

6. C. R. Anderson. A machine learning approach to web personalization. PHD Thesis, University of Washington, 2002.

7. Feng Tao and Fionn Murtagh. Towards knowledge discovery from www log data. In Proc. of the International Conference on Information Technology: Coding and Com-puting, 2000.

8. Free and open-source software, http://en.wikipedia.org/wiki/FLOSS

9. H. Ishikawa, M. Ohta, S. Yokoyama, J. Nakayama, and K. Katayama. On the efefectiveness of web usage mining for page recommendation and restructuring. Web, Web-Services, and Database Systems.

10. I.H. Witten and E. Frank. Data mining - Practical machine learning tools and techiniques. Second edition, Morgan Kauffmann Publishers, 2005.

11. I.H. Witten and E. Frank. Data mining - Practical machine learning tools and techiniques using JAVA implementations. Third edition, Morgan Kauffmann Publishers, 2009.

12. Jaideep Srivastava, Prasanna Desikan, Vipin Kumar. Chapter 3 Web Mining Accomplishments & Future Directions, www.ieee.org.ar/downloads/Srivastava-tut-paper.pdf

13. J. Han and M. Kamber. Data mining: Concepts and techniques. Morgan Kauffmann Publishers, 2001.

14. J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations, 1(2):12–23, 2000.

15. K.W. Tan, H. Han, and R. Elmasri. Web data cleansing and preparation for ontology extraction using WordNet. In First International Conference on Web Information Systems Engineering (WISE'00), volume 2.

16. List of web analytics software, http://en.wikipedia.org/wiki/List_of_web_analytics_software

17. M. Perkowitz and O. Etzioni. Adaptive sites: Automatically learning from user access patterns. In Proc. of the Sixth International WWW Conference, Santa Clara, CA., 1997.

18. M. Spiliopoulou, C. Pohle, and L.C. Faulstich. Improving the effectiveness of a website with web usage mining. In Advances in Web Usage Analysis and User Profiling, Berlin, Springer, pp. 14162, 2000.

19. M. Spiliopoulou. Web usage mining for web site evaluation. Communications of the ACM, 43(8), 2001.

20. O. R. Zaiane, M. Xin, and J. Han. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In Advances in Digital Libraries, pages 19–29, 1998.

21. Proprietary software, http://en.wikipedia.org/wiki/Proprietary_software

22. Rconvlog (Rebex Internet Log Converter), http://www.rebex.net/rconvlog/default.aspx

23. R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems V1(1).

24. R. Kosala and H. Blockeel. Web mining research: A survey. ACM SIGKDD, 2(1):1–15, 2000.

25. Software as a Service, Storage as a Service, http://www.webopedia.com/TERM/S/SaaS.html

26. Web Analytics Software - Deep Log Analyzer, http://www.deep-software.com/

27. Web log analysis software, http://en.wikipedia.org/wiki/Web_log_analysis_software

28. WEKA: www.cs.waikato.ac.nz/ml/weka

29. WUMprep (Web mining pre-processing), http://sourceforge.net/projects/hypknowsys/

30. Xindong Wu and Vipin Kumar, The Top Ten Algorithms in Data Mining, pages 1–17, 2009.

31. X.O. Kaynak, Model Selection with Cross-Validations and Bootstraps, LNCS 2714, pp. 573–580, 2003.Springer-Verlag Berlin Heidelberg 2003.

32. Marjan Velkoski and Cveta Martinovska Bande, Analyzing Web Server Access Log Files Using Data Mining Techniques International Conference on Applied Internet and Information Technologies, 2013

Марјан Велкоски

Анализа на веб логови со примена на техники за податочно рударење

Универзитет „Гоце Делчев“ - Штип