Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko

Denis Vodišek

PROFILIRANJE IN SLEDENJE UPORABNIKOM NA SPLETU

Diplomsko delo

Maribor, junij 2017

PROFILIRANJE IN SLEDENJE UPORABNIKOM NA SPLETU

Diplomsko delo

Študent:	Denis Vodišek
Študijski program:	Univerzitetni študijski program informatika in tehnologije komuniciranja
Smer:	Informacijski sistemi
Mentor FERI:	doc. dr. Marko Hölbl, univ. dipl. inž. rač. in inf.

Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko
Smetanova ulica 17
2000 Maribor, Slovenija

FERI

Številka: E1079427
Datum in kraj: 11. 04. 2017, Maribor

Na osnovi 330. člena Statuta Univerze v Mariboru (Statut UM – UPB 11, Ur. l. RS, št. 44/2015) izdajam

SKLEP O ZAKLJUČNEM DELU

1. **Denisu Vodišku,** študentu študijskega programa prve stopnje UN INFORMATIKA IN TEHNOLOGIJE KOMUNICIRANJA, smer Informacijski sistemi, se dovoljuje izdelati zaključno delo.

2. Tema zaključnega dela je pretežno s področja Inštituta za informatiko.

3. **MENTOR:** **doc. dr. Marko Hölbl**

4. **Naslov zaključnega dela:**
PROFILIRANJE IN SLEDENJE UPORABNIKOM NA SPLETU

5. **Naslov zaključnega dela v angleškem jeziku:**
ONLINE USER TRACKING AND PROFILING

6. Rok za izdelavo in oddajo zaključnega dela je 30. 09. 2017. Zaključno delo je potrebno izdelati skladno z "Navodili za izdelavo zaključnega dela" in ga v treh izvodih (dva trdo vezana izvoda in en v spiralo vezan izvod) oddati v pristojnem referatu članice. Hkrati se odda tudi izjava mentor-ja/-ice (in morebitnega somentor-ja/-ice) o ustreznosti zaključnega dela.

Pravni pouk: Zoper ta sklep je možna pritožba na Senat članice v roku 10 delovnih dni od dneva prejema sklepa.

Dekan:
red. prof. dr. Borut Žalik

Obvestiti:
- kandidata,
- mentor-ja/-ico,
- odložiti v arhiv.

ZAHVALA

Zahvaljujem se mentorju doc. dr. Marku Hölblu za pomoč in usmerjanje pri opravljanju diplomskega dela.

Posebna zahvala gre tudi moji družini in prijateljem, ki so me podpirali tekom celotnega študija.

PROFILIRANJE IN SLEDENJE UPORABNIKOM NA SPLETU

Ključne besede: Spletna analitika, sledenje, profiliranje, Google Analytics, splet, zbiranje podatkov, R, podatkovno rudarjenje

UDK: 004.774.6(043.2)

Povzetek

Na današnjem spletu je skorajda težko ostati neviden. Podjetja, mediji in posamezniki se čedalje bolj zavedajo, da lahko s poznavanjem svojih strank oziroma uporabnikov ugotovijo njihove želje, interese in omogočijo odlično uporabniško izkušnjo, saj je ravno to tisto, kar odloča o nadaljnjem ravnanju uporabnika na strani. V diplomskem delu smo vzpostavil storitev za sledenje uporabnikom z brezplačno storitvijo Google Analytics. Pridobljene podatke smo z orodjem za podatkovno rudarjenje R združili v skupine in ustvarili uporabniške profile. Zajeli smo glavne metodologije zbiranja podatkov, piškotke in orodje, ki jih bomo uporabili za spletno analitiko. Prav tako smo obravnavali tudi zasebnost uporabnika na spletu in načine sledenja ter profiliranja uporabnikov. V diplomskem delu smo tudi implementirali storitve na že obstoječo spletno stran in zbirali podatke s po meri ustvarjenimi razsežnosti in meritvami. Podatke smo obdelali v programu R in jih vizualno predstavili. Ugotovili smo, da lahko s spletno analitiko natančno odkrijemo, kaj uporabniki počnejo na strani, kaj jih zanima in na podlagi teh informacij ustvarili uporabniške profile. Tako lahko z ustvarjenimi profili uporabnikom iste skupine priporočamo podobne rezultate.

ONLINE USER TRACKING AND PROFILING

Key words: Web analytics, profiling, tracking, Google Analytics, web, collecting data, R, data mining

UDK: 004.774.6(043.2)

Abstract

It is hard to stay invisible on today's web. Companies, the media and individuals are aware that by knowing their clients and users, they can identify their wishes, interests and provide them excellent user experience, because this is exactly what determines the user's further behavior on the web site. In this diploma work, we set up a user tracking service by using Google Analytics. With the gathered data, we create user profiles using data mining tool R. We describe the main methodologies of collecting data, cookies and the tool, which we will use for web analytics. We also describe the privacy on the web and the ways of tracking and profiling users. In the diploma, we implement the service on the web site and collect the data with custom dimensions and metrics. In the end, we processed data in R tool and even visually present them. We also write down our findings. We find out that with web analytics we can identify users' behavior on the web site, what are they interested in and then we created user profiles based on the collected data. With the profiles created, we can recommend similar results to users of the same group.

KAZALO

KAZALO SLIK

KAZALO TABEL

KAZALO PROGRAMSKE KODE

Seznam uporabljenih kratic

GA – Google Analytics

GATC – Google Analytics Tracking Code

HTTP – HyperText Transfer Protocol

IP – Internet Protocol

HTML - Hypertext Markup Language

API - Application programming interface

AI – Analitična Inteligenca

URL – Uniform Resource Locator

WEKA – Waikato Environment for Knowledge Analysis

CSV – Comma-Seperated Values

XML – eXtensible Markup Language

1. UVOD

Uporabnik ob vsakem obisku spletne strani pusti svoje sledi. Sledi enega uporabnika, kaj je na strani počel, koliko dolgo je bil, kaj ga je zanimalo, ne pomenijo veliko, a če zbiramo podatke od več uporabnikov, ustvarimo različne profile in s tem identificiramo interese in stvari, ki so za uporabnika zanimivi. Pridobljene podatke lahko uporabi sistem oziroma storitev ali spletni analitiki, da bolje razumejo uporabnika in posledično dajejo možnost, da izboljšajo zadovoljstvo uporabnika.

V tej diplomski nalogi bomo poiskali in integrirali spletno storitev za spletno analitiko, zbiral podatke o uporabnikih na spletni strani, uporabnike bomo tudi profilirali in predstavili v grafični obliki.

Najprej bomo opisali kaj spletna statistika je, kako se zbirajo podatki, katera orodja za analizo podatkov poznamo, nato bomo opisali kakšni so načini sledenja uporabnikov in metode profiliranja. Na koncu sledi praktični del v katerem bomo zbirali podatke z izbrano storitvijo jih obdelali v orodju za podatkovno rudarjenje in jih predstavili v grafični obliki.

1.1. OPREDELITEV PROBLEMA

Pogoj za uspešno poslovanje na spletu je neprestano izboljševanje in optimiziranje spletne strategije, navigacije strani in vsebine strani. Slaba spletna stran lahko zelo škoduje podjetju. Ugotoviti moramo kaj nam povzroča težave – ali so to slabe marketinške kampanje, so slabe ocene izdelkov/produktov, ali zna stran uporabniku podati tisto kar si želi. Spletna statistika nam zagotavlja, da pridobimo vse potrebne informacije in zmožnost merjenja učinkovitosti sprememb [1].

Kot pomožna veja spletni statistiki, se je za potrebe natančnih informacij, kaj uporabniki klikajo, kakšno je njihovo obnašanje na spletni strani, razvilo sledenje uporabnikov in posledično tudi njihovo profiliranje.

Spletne tehnologije sledenja uporabljamo za zbiranje, shranjevanje in povezovanje uporabnikove zgodovine in navad brskanja. Glavna motivacija za spletno sledenje so

oglaševalne agencije, organi pregona in obveščevalne agencije, testi uporabnosti in pridobitev uporabnikovih navad [2].

Prav v slednjem smo odkrili problem, katerega bomo odpravili v diplomski nalogi. Namreč v podjetju v katerem opravljam študentsko delo, smo končali projekt za turistično podjetje, kjer smo ustvarili internetni sistem za rezervacije. Ne vemo, kdo so naši uporabniki, kaj si ogledujejo in kakšne so njihove želje.

1.2. CILJI

Cilj diplomske naloge je poiskati in integrirati ustrezno spletno storitev za sledenje in profiliranje uporabnikov, ki bo zmožna pridobiti potrebne oz. uporabne podatke o uporabnikih. V ospredje bomo postavili integracijo storitve v spletno stran, pridobitev podatkov, koliko so ti podatki uporabni in kaj lahko razberemo iz njih ter analizo o pridobljenih podatkih. Poleg prej naštetih zastavljenih ciljev, bomo obravnavali tudi kaj spletna statistika je, kako se podatki zbirajo, kaj so piškotki, načini in metode profiliranja ter sledenja uporabnikom.

Cilji, ki jih želimo doseči:

- Raziskati delovanje storitev za sledenje in profiliranje uporabnikov;
- Poiskati ustrezno rešitev in jo integrirati ter prilagoditi za naše potrebe;
- Raziskati uporabnost pridobljenih podatkov;
- Gručenje podatkov (ustvarjanje profilov uporabnikov);
- Priprava podatkov za priporočilni sistem.

1.3. METODOLOGIJA

Pri sledenju uporabnikom bomo uporabili dokumentacijo, ki je dosegljiva na spletu. Pregledali bomo obstoječe odprto-kodne rešitve in jih prilagodili našim potrebam.

Okolje, v katerem bomo pridobili potrebne informacije bomo izbrali po preučitvi obstoječih rešitev in dokumentacije, podatke pa bomo pridobili na dejanski spletni strani.

Seznam metod, katere načrtujemo uporabiti:

- Deskriptivna metoda (študija literature);
- Integracija storitve (integracija v spletno stran);
- Empirična metoda (obdelava in analiza podatkov);
- Razprava (razpravljanje o pridobljenih rezultatih, podajanje mnenja).

1.4. PREDPOSTAVKE IN OMEJITVE DIPLOMSKEGA DELA

Predpostavili bomo, da diplomska naloga ne bo zajemala kompletno izdelavo priporočilnega sistema, temveč samo pridobitev podatkov o uporabnikih, katerim bomo sledili na spletni strani, ki bodo primerni in uporabni za priporočilni sistem. Seveda bo število pridobljenih podatkov odvisnih od števila obiskovalcev na spletni strani.

Raziskava in uporaba storitve za sledenje uporabnikov se bo omejila na raziskavo obstoječih odprtokodnih rešitev. Cilj ni načrtovanje in implementacija lastne storitve, temveč uporaba in integracija najbolj ustrezne storitve za naše potrebe.

2. ZBIRANJE PODATKOV O UPRABNIKIH NA SPLETU

2.1. SPLETNA ANALITIKA

Spletna analitika je tehnologija in metoda za zbiranje, merjenje, analiziranje in poročanje podatkov iz spletnih strani in spletnih aplikacij [3].

Raste neprestano že od samega razvoja »World Wide Web«. Razvila se je od preproste http (Hypertext Transfer Protocol) funkcije za beleženje dnevnika prometa do celovite rešitve za sledenje, analiziranje uporabniških podatkov in njihovo poročanje. Področje spletne analitike se močno razvija z velikim številom novih orodji, platform in delovnih mest [4].

Spletna analitika je postala kritična komponenta mnogih poslovnih odločitev. Z nenehno rastjo števila raznovrstnih transakcij preko spletnih vmesnikov, je sposobnost razumevanja in pregleda aktivnosti spletne strani nujno potrebna. Moderne spletne strani vsebujejo bogate vsebine, ki prinašajo uporabne informacije. Spletne strani se pojavljajo od enostavnega statičnega HTML-ja do zelo prefinjenih dinamičnih vsebin. Ob gostovanju takšnih vsebin se soočamo z različnimi izzivi, kot so zagotavljanje virov in zaznavanje anomalij prometa uporabnikov. Še večji izziv je identificiranje poslovnih vpogledov, kateri temeljijo na vzorcih spletnega dostopa. Najbolj uporaben vir za pridobivanje vpogledov so dnevniki spletnih dostopov in orodja za spletno analitiko. Z pridobitvijo teh podatkov lahko ugotovimo kaj se je zgodilo, kaj pričakovati in kako dobro stvari delujejo [5].

2.2. METODOLOGIJE ZBIRANJA PODATKOV

Ključ za uspešno uporabo zbranih informacij je, da vemo, kako ravnat z njimi—kaj nam lahko doprinesejo, njihove omejitve. To zahteva popolno razumevanje metodologij zbiranja podatkov. V bistvu, obstajajo dve pogosti tehniki: Oznake na strani in strežniške dnevniške datoteke [1].

JavaScript oznake

Prva najpogostejša tehnika zbiranja podatkov deluje na strani odjemalca. Tako imenovane oznake na strani zbirajo podatke glede na brskalnik odjemalca in pošiljajo pridobljene informacije na oddaljen strežnik. Informacijo o obiskovalcih zajame JavaScript koda, ki je po navadi postavljena pred HTML značko `<body>`, tako da je prisotna na vsaki naloženi strani znotraj domene. Oseba, ki uporablja orodje za analitiko pregleda poročila spletnega mesta z oddaljenega strežnika [1].

Proces pridobivanja podatkov je naslednji:

1. Odjemalec pošlje zahtevo za spletno stran;
2. Strežnik sprejme zahtevo;
3. Strežnik pošlje odgovor, torej spletno stran s pripeto kodo za zbiranje podatkov;

Ko se stran naloži, se izvede JavaScript koda, katera zajame obiskano stran, podatke o seji obiskovalca in piškotke. Nato pridobljene podatke pošlje na podatkovni strežnik.

Dnevniške datoteke

Beleženje strežniškega dnevnika je tradicionalna metoda in najstarejša metoda zbiranja podatkov. Razvita je bila, da bi zajela napake, katere so povzročali strežniki in čez čas se je izkazalo, da se te podatke lahko analizira in ustvari dobičke in tako se je beleženje dnevnikov iz tehničnega vidika usmerilo na marketing.[6]

Dnevniška datoteka se generira s pomočjo strežnika, kateri beleži strežniške aktivnosti in HTTP zahtevke v tekstovnem formatu. Obstaja več različnih formatov dnevniških datotek. Najbolj pogosto beleženi podatki so strežniški IP, datum/čas, HTTP zahteva, status odgovora in velikost odgovora [7].

Proces zbiranja podatkov je naslednji:

1. Odjemalec napiše URL v brskalnik;
2. Zahteva se pošlje na strežnik;

3. Strežnik sprejme zahtevo in ustvari vnos v dnevnik;
4. Strežnik nato pošlje stran nazaj odjemalcu.

Seveda imata obe tehniki svoje omejitve. Spodnja tabela prikazuje razlike med tehnikama. Znan mit je ta, da je postavljanje oznak na spletno stran bolj superiorna od ostalih tehnik, ampak je to odvisno, kaj želimo izvedeti. Če kombiniramo obe tehnike, prednosti ene zabrišejo slabosti druge. Uporabo obeh tehnik lahko z drugimi besedami označimo za hibridno metodo [1].

Tabela 1: Primerjava metodologij

Metodologija	Prednosti	Slabosti
Oznake na strani	• ”Real-time” zbiranje in procesiranje podatkov • Ponudnik analitike poskrbi za hrambo podatkov in arhiviranje • Sledenje dogodkov (JavaScript, Flash, Ajax) • Zagotavlja natančnejše sledenje sej	• Požarni zid lahko zavrne oznake na strani • Robotki ignorirajo oznake • Ob napaki lahko izgubimo vse podatke
Dnevniške datoteke	• Brez težav z požarnim zidom • Sledi pajkom in robotkom	• Brez sledenja dogodkov • Robotki večajo števec obiskovalcev • Netočnost predpomnilnika

Ostale metode zbiranja podatkov

Kljub temu, da so JavaScript oznake in dnevniške datoteke najbolj razširjene metode za zbiranje podatkov na spletu, nista edini. Poznamo še merilno kodo (ang. web beacon) in vohljanje paketov (ang. packet sniffing).

Merilna koda je tehnologija meri uspešnost oglasne pasice in njene klike. Čeprav ni pogosto uporabljena metoda, se jo še vedno da zaslediti na spletu. Prednost merilne kode je sledenje obnašanja uporabnika preko več spletnih strani. Od tod lahko dobimo odgovore, kako je pasica uspešna na različnih straneh. Ker isti strežnik zbira podatke, piškotke in sledi uporabnikom, je oglaševalcu poenostavljeno spremljanje pridobljenih podatkov [8].

Metoda vohljanja paketov je zelo napredna, če govorimo o tehnološkem znanju. Njena glavna prednost je ta, da ne potrebuje oznak na spletnih straneh ampak vse informacije pridobiva z strojno opremo (»vohljačem«).

Proces poteka tako:

1. Odjemalec zahteva spletno stran;
2. Zahtevo sprejme strojna oprema in zbira podatke o zahtevi;
3. Vohljač nato pošlje zahtevo do strežnika;
4. Strežnik pošlje odgovor a ga vohljač prestreže, shrani informacije in pošlje naprej do odjemalca.

2.3. PIŠKOTKI

Piškotki imajo v današnji družbi bolj negativen prizvok kot pozitiven. Veliko je paranoje in ne razumevanje na tem področju, še posebno v povezavi z zasebnostjo. Kot navaja definicija je piškotek: »Sporočilo poslano spletnemu brskalniku s strani strežnika. Brskalnik hrani sporočilo v tekstovni datoteki. To sporočilo se nato pošlje na strežnik ob vsaki zahtevi spletne strani na strežniku.« Vsakič, ko brskalnik/odjemalec zahteva spletno stran s strežnika, bo le-ta vedno

preveril, če piškotki od odjemalca že obstajajo. V istem trenutku ga bo tudi shranil oziroma posodobil v brskalnik obiskovalca [6].

Sejni piškotki

Ti piškotki so prehodni. Obstajajo le tako dolgo, kot je obiskovalec v interakciji s spletno stranjo. Tipično obstajajo zato, da hranijo podatke katere uporablja strežnik (npr. hranjenje izdelkov v košarici, dokler ste na spletni strani) in zato, da lahko spletna analitika razume obnašanje med obiskom spletne strani. Ti piškotki izginejo po zaključku seje [6].

Trajni piškotki

Piškotki se nastavijo, ko obiskovalec obišče spletno stran in ostanejo tudi po zaključku seje. Obstajajo dokler ne poteče njihov rok trajanja. Vsebuje unikatni ID, s katerim sledi spletnim brskalnikom, kadar obiščejo spletno stran [6].

Super piškotek

Super piškotki (ang. supercookie) je piškotek, kateri izhaja iz najvišjega nivoja domene (npr. .com ali .si) ali javne pripone (npr. co.uk). Običajno piškotki izvirajo iz specifičnega imena domene (npr. primer.si). Pri super piškotkih se pojavlja vprašanje varnosti, zato so pogosto blokirani pri brskalnikih. Če ga brskalnik ni blokiral, lahko napadalec nastavi piškotek in morebiti prekine ali prestavi legitimne zahteve na drugo spletno mesto. Na primer piškotek z izvorom .si domene, bi lahko zlonamerno vplival na zahtevek na stran primer.si, čeprav piškotek ne izvira iz primer.si. Na ta način lahko izvedemo lažen vpis ali spremenimo uporabnikove informacije [9].

Zombi piškotki

Zombi piškotek je piškotek kateri se ponovno ustvari po izbrisu. Piškotki se poustvarijo s pomočjo varnostnih kopij, katere so shranjene izven brskalnikovih datotek za shrambo piškotkov. Lahko so shranjeni na spletu ali direktno na uporabnikovem računalniku[9].

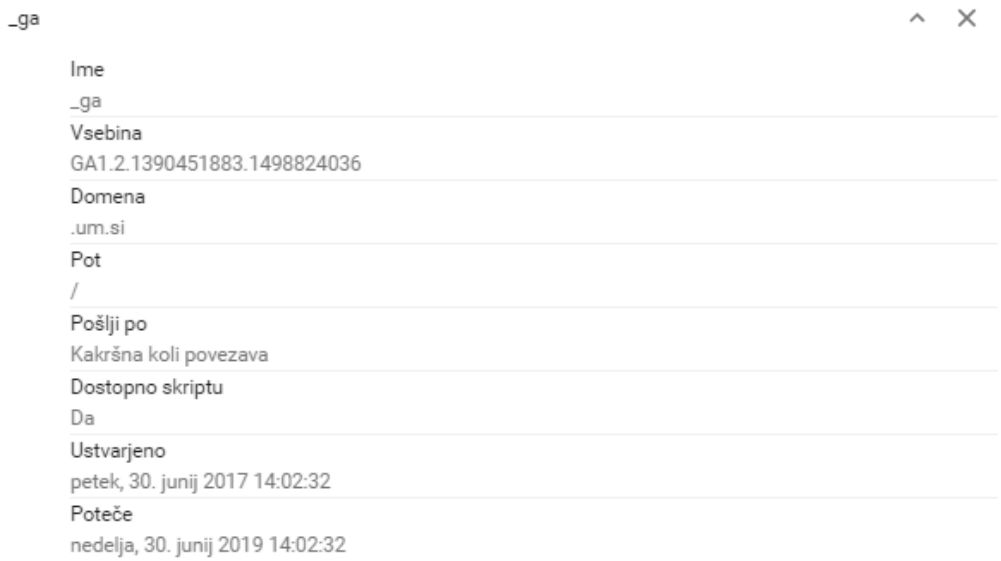
2.3.1. DELOVANJE PIŠKOTKOV

Spletno bazirane aplikacije pogosto uporabljajo piškotke za obdržanje stanja v HTTP protokolu. Kot del odgovora strežnik pošlje še arbitrarno informacijo, tako imenovani piškotek, v `Set-Cookie` glavi odgovora. Ta informacija je lahko karkoli: uporabniški identifikator, bazni ključ, skratka kar koli strežnik potrebuje, da lahko nadaljuje, kjer je nazadnje končal. Pod normalnimi pogoji, odjemalec vrne informacijo v glavi piškotka. Strežnik izbira ali bo vključil nov piškotek v odgovor, kateri bo nadomestil starega. Tako obstaja nenapisana »pogodba« med strežnikom in odjemalcem: strežnik se zanaša na odjemalca, da shrani strežnikovo stanje in da ga vrne ob naslednjem obisku [10].

Piškotki sami po sebi niso škodljivi, ne nastanejo zaradi vdora v računalnik ali zaradi zlonamernih programov. Je zgolj košček informacije, katero si strežnik in odjemalec ves čas pošiljata med sabo. Piškotki so tekstovne datoteke, ki ne vzamejo veliko prostora (do 4KB), shranjeni so lokalno, tako da se jih lahko po želji izbriše [10].

Omejevanje odjemalcu, da vrne piškotek na strežnik od katerega ga je prejel je zelo omejujoče. Organizacije imajo pogosto več strežnikov in le-ti potrebujejo dostop do iste informacije, da lahko zagotovijo brezhibno izkušnjo in delovanje. Ko strežnik pošlje odjemalcu piškotek, strežnik določi na katere druge strežnike lahko odjemalec pošlje piškotek v naslednjih zahtevah [10].

Vse vrednosti katere vidimo spodaj na sliki so s strani brskalnika in so uporabljene da sledijo obnašanju obiskovalcev na strani. Vsak piškotek vsebuje unikaten ID, ki identificira brskalnik. Piškotek identificira tudi vir, torej kdo je nastavil piškotek in nastavi prirejene spremenljivke, ki pomagajo dodatnemu sledenju. Poznavanje uporabnika lahko veliko doprinese k uporabniški izkušnji v naslednjih obiskih na strani, saj mu lahko prilagodimo stran in napišemo »dobrodošel nazaj« ter tako ustvarimo bolj individualno izkušnjo.



Slika 1: Primer piškotka

2.3.2. IMPLEMENTACIJA PIŠKOTKOV

Piškotki, kot smo že omenili v prejšnjih poglavjih se nastavijo s `Set-Cookie` v HTTP zahtevku. Zahtevek pove brskalniku naj shrani piškotek in ga pošlje nazaj ob zahtevi. Na primer, kot vidimo v spodnji sliki brskalnik pošlje svoje prvo zahtevo za spletno stran www.primer.org [11]. Nato v razdelku odgovor strežnik v glavi določi 2 piškotka. To stori s pomočjo določila `Set-Cookie`.

```
#zahteva

GET /index.html HTTP/1.1
HOST: www.primer.org
...

#odgovor

HTTP/1.0 200 OK
Content-type: text/html
Set-Cookie: theme=light
Set-Cookie: sessionToken=abc123; Expires=Wed, 09 Jun 2021 10:18:14 GMT
...
```

Programska koda 1: Primer zahteve in odgovora

Strežnikov HTTP odgovor vsebuje vsebino obiskane strani in hkrati pove brskalniku naj nastavi dva piškotka. Prvi piškotek `theme` velja za sejnega piškotka in nima roka trajanja. Drugi `sessionsToken` velja za trajni piškotek, ker vsebuje atribut rok trajanja [11].

Nato brskalnik pošlje zahtevo za obiskanje druge strani na isti domeni. Zahteva v opisu vsebuje `Cookie HTTP`, kateri vsebuje piškotka, ki sta se nastavila ob obisku strani.

Tako strežnik ve, da je ta zahteva v relaciji s prejšnjo. Strežnik bo odgovoril z zahtevano stranjo in morebiti vključil še več piškotkov v odgovor, z namenom, da bi dodal, spremenil ali izbrisal piškotke [11].

2.3.3. ATRIBUTI PIŠKOTKOV

Piškotki lahko imajo poleg imena in vrednosti še več drugih atributov. Brskalniki ne vključujejo atributov piškotkov v svoji zahtevi, ampak se na podlagi njih odločijo kdaj piškotek zbrisati, blokirati ali poslati na strežnik.

Domena (ang. domain) in pot (ang. path) sta atributa, ki definirata področje piškotka. Pravzaprav povesta h kateri spletni strani piškotek spada. Zaradi varnostnih razlogov lahko piškotke nastavi samo trenutna domena in njene poddomene. Na primer, stran www.primer.org ne mora nastaviti piškotka, kateri ima domeno strani www.test.si, ker to bi potem dovoljevalo strani primer.org kontrolo nad piškotki strani test.si [12].

Če domena in pot piškotka nista nastavljena s strani strežnika, se kot privzeto nastavita na trenutno domeno in pot. A vendar je velika razlika, če atribute nastavimo oziroma jih ne. V prejšnjem primeru, če atributov ne bi nastavili, bi bil piškotek poslan samo na test.si. V nasprotnem primeru, torej če bi nastavili domeno in pot, potem bi to veljalo še za vse poddomene test.si [12].

```
HTTP/1.0 200 OK
Set-Cookie: LSID=DQAAAK…Eaem_vYg; Path=/primer; Expires=Wed, 13 Jan 2021 22:23:01 GMT; Secure; HttpOnly
Set-Cookie: HSID=AYQEVn…DKrdst; Domain=.test.si; Path=/; Expires=Wed, 13 Jan 2021 22:23:01 GMT; HttpOnly
Set-Cookie: SSID=Ap4P…GTEq; Domain=test.si; Path=/; Expires=Wed, 13 Jan 2021 22:23:01 GMT; Secure; HttpOnly
```

Programska koda 2: Primer domene in poti piškotka

Prvi piškotek, `LSID`, nima domenskega atributa, pot pa ima nastavljeno na `/primer`. Ta primer pove brskalniku, da naj uporabi piškotek le takrat kadar bomo zahtevali strani znotraj določene

domene. Ostala dva piškotka bosta uporabljena kadar bo brskalnik zahteval katero koli poddomeno in pot na določeni domeni [12].

Rok trajanja (ang. expires) in maksimalna starost (ang. max-age) piškotka sta atributa, katera definirata čas piškotku. Atribut rok trajanja definira točno določen datum in čas, kdaj naj brskalnik izbriše piškotek. Alternativa roku trajanja je atribut maksimalna starost. Atribut nastavi časovni interval, koliko časa bo še lahko aktiven [12].

Poznamo tudi »Secure« in »HttpOnly« atributa. Prvi omenjeni zagotavlja kriptirano komunikacijo piškotkov in usmerjal brskalnike k uporabi varnih oziroma kriptiranih povezav. Če strežnik nastavi piškot z »Secure« atributom, a ga nato pošlje preko ne varne povezave, je lahko informacija prestrežena na poti z napadom s posrednikom (ang. man in the middle attack). »HttpOnly« atribut diktira brskalnikom naj piškotka ne pošiljajo preko drugih kanalov, kot le preko HTTP (in HTTPs). Kar pomeni, da do piškotka ni mogoče dostopati preko skript [12].

2.4. ZASEBNOST UPORABNIKOV NA SPLETU

Spletna iskanja lahko odkrijejo občutljive informacije o uporabnikovem zasebnem življenju spletnemu brskalniku in internetnim prisluškovalcem. Uporabniki raziskujejo po spletu, da bi pridobili nove informacije oziroma da bi našli določene spletne strani. Vendar poleg iskanja puščajo tudi sledi svojih interesov in namenov. Ta informacija je lahko uporabljena s strani iskalnikov in spletnim prisluškovalcem, da bi zgradili profile uporabnikov in oblikovali občutljive osebne podatke o njih[13].

Skrbi glede zasebnosti

Ni dvoma, da so internetni uporabniki zaskrbljeni glede spletne zasebnosti. Kot je ugotovil Hoffman in drugi so ugotovili, da je več kot 90% internetnih uporabnikov zavrnilo posredovanje osebnih informacij ali pa so si jih izmislili zaradi nezaupanja spletu [14]. Razvoj in hitro širjenje spletnih socialnih omrežij sta sprožila nov val skrbi o zasebnosti na spletu. Socialna omrežja so odlične platforme za hitro širitev in vzdrževanje osebnih ali profesionalnih vezi, a za to se zahteva veliko osebnih podatkov, kar povzroča večja tveganja za kršitve zasebnosti [15].

Uporabniki še posebej izpostavljajo naslednje skrbi [16]:

- Skrita sledenja ob obisku strani;
- E-mail naslovi in ostale osebne informacije bodo zajeta in uporabljena za oglaševalske namene brez njihovega dovoljenja;
- Osebne informacije bodo prodane brez dovoljenja;
- Kraja kreditne kartice.

Težave z zasebnostjo

Zgodovina brskalnika je nedvomno povezana z osebnimi podatki. Strani lahko odkrijejo uporabnikovo lokacijo, interese, nakupe, status zaposlitve, spolno usmeritev in še več. Preučevanje uporabnikovih sledi, ki jih pusti na strani, lahko doprinese številne ugotovitve o njemu. Dodatna težava se pojavi tudi, kadar se na stran vključi vsebina z zunanjim delovanjem (ang. third party). Glavni namen je seveda pridobivanje podatkov za oglaševanje. Možno je, da lastnik spletnega mesta proda uporabnikovo identiteto. Največkrat lahko to vidimo v obliki kviza. Do razkritje podatkov lahko pride na več načinov. Eden izmed njih je, da lastnik spletnega mesta nenamerno pošilja identifikacijsko informacijo preko URL. Druga možnost je, da zunanja aplikacija izvaja križno izvajanje skripte (»XSS«), preko katere lahko ugotovi identiteto uporabnika [17].

2.5. ORODJA ZA ZBIRANJE PODATKOV

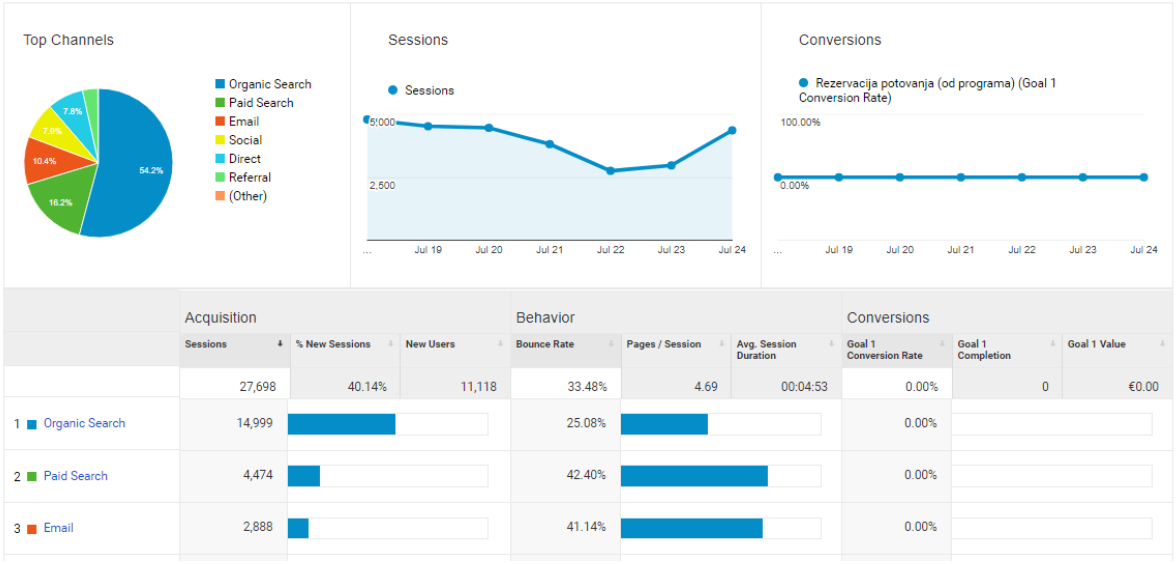
Področje spletne analitike se je v zadnjih letih močno razvilo. Obstaja veliko najrazličnejših orodij za zbiranje podatkov, nekatera imajo več funkcionalnosti, nekatera manj, mi bomo opisali Google Analytics. Google Analytics (v nadaljevanju tudi GA) je najbolj priljubljena rešitev za spletno analitiko in katero bomo tudi uporabili v okviru diplomske naloge

2.5.1. GOOGLE ANALYTICS

Google Analytics je brezplačna rešitev, ki zagotavlja statistične in osnovne analitične funkcionalnosti za marketinške namene in SEO optimizacijo. Storitev je dosegljiva vsem, kateri imajo Googlov račun.

Poročilo o celotni oglaševalski akciji

GA omogoča sledenje in primerjanje vaših obiskovalcev – od brezplačnega organskega iskanja, plačljivih oglasov, e-mail novic in od vseh ostalih iskanj in mediji, ki preusmerjajo obiskovalca na spletno stran [1].



Slika 2: Poročilo o prihodih na spletno stran

Poročilo o e-poslovanju

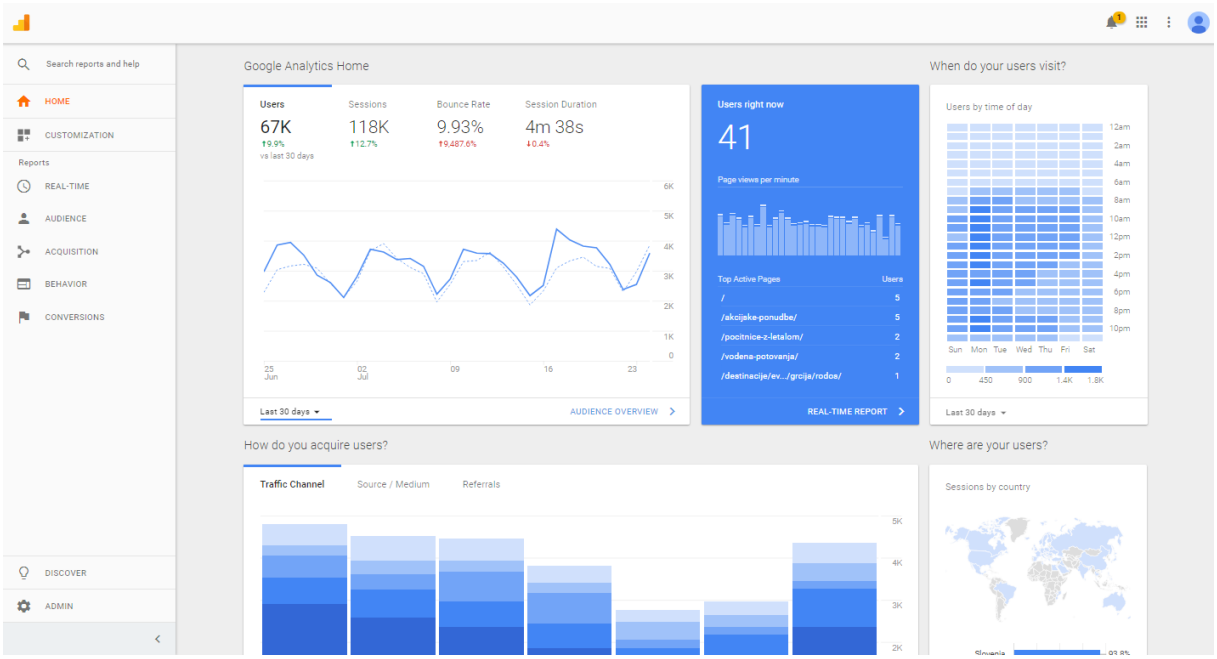
Sledimo lahko transakcijam oglaševalskih akcij in ključnim besedam, pridobimo metrike zvestobe in latence, identificiramo naše vire prihodkov [1].

Vizualizacija tokov

Tokovi (ang. funnels) so poti obiskovalcev preden dosežemo dosego cilja. Očiten cilj tokov na primer je proces nakupa pri e-poslovanju. Z vizualizacijo obiskovalčevih tokov lahko odkrijemo pri katerih straneh izgubimo obisk in kam se obiskovalci nato odpravijo [1].

Prilagojena nadzorna plošča

Nadzorna plošča je izbor skrajšanih poročil iz glavnih delov GA. Na nadzorni plošči lahko postavimo in organiziramo ključne podatke. Dodamo lahko do 12 poročil, katere lahko spreminjamo in jim določimo vrstni red. Nadzorna plošča je omejena na uporabnika, torej vsak uporabnik lahko ima drugačno nadzorno ploščo [1].



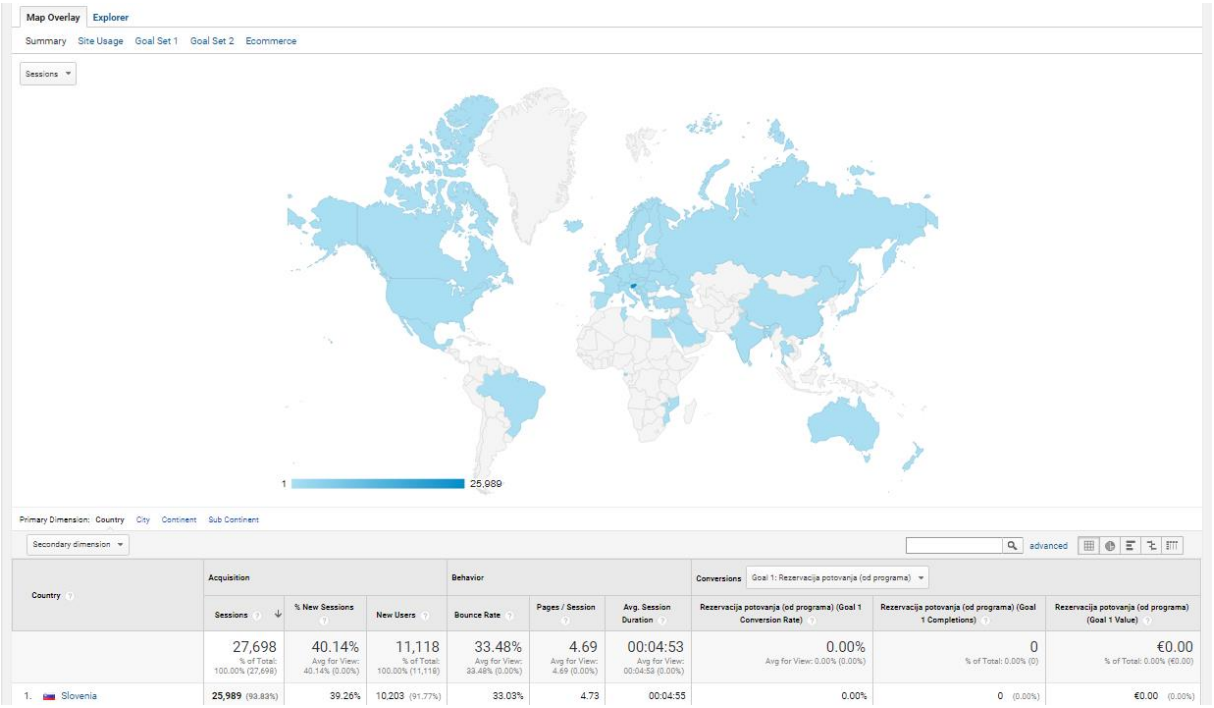
Slika 3: Nadzorna plošča

Poročilo strani

Poročilo strani je grafični prikaz priljubljenosti vaših povezav na strani. Je preprost prikaz, katere strani privabljajo obiskovalce in katere ne [1].

Poročilo v obliki zemljevida

Podobno kot poročilo strani je tudi poročilo v obliki zemljevida grafični prikaz podatkov, katero prikazuje lokacijo obiskovalcev, ki so obiskali spletno stran. Glede na IP naslov se pridobi lokacija in se prikaže na svetovni, regionalni ter državni ravni [1].



Slika 4: Poročilo v obliki zemljevida

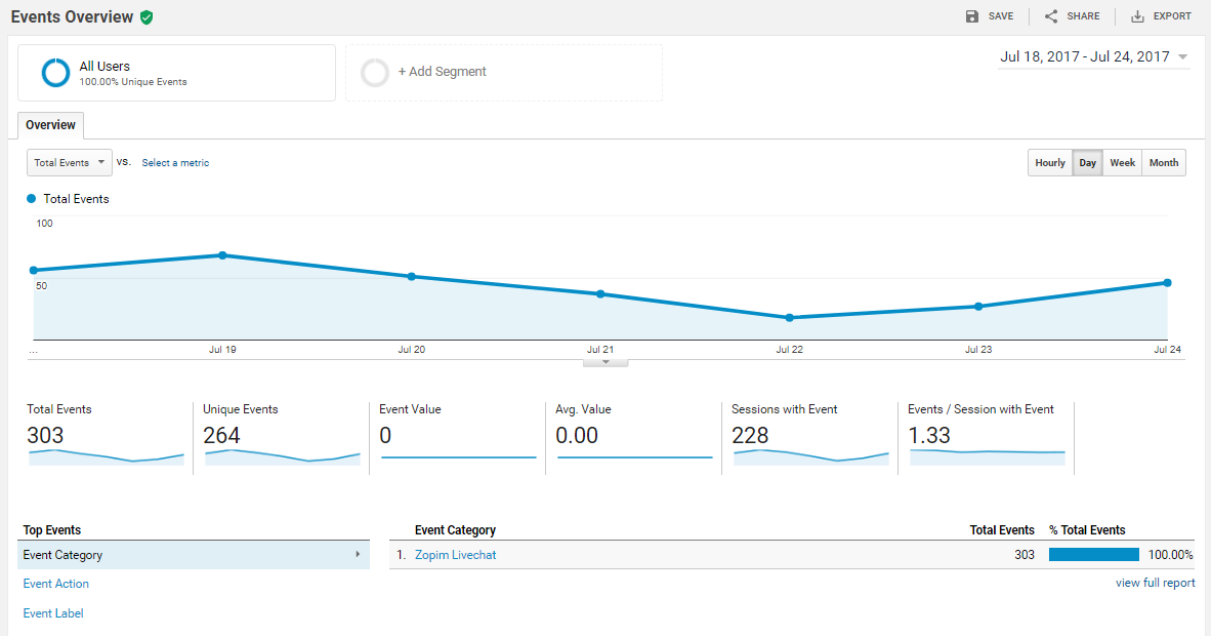
Izvoz podatkov in nastavljanje urnikov

Podatki so lahko ročno izvoženi v najrazličnejših formatih, tako v CSV formatu, PDF, TSV in tudi celo v odprtokodnem XML. Lahko se nastavi, da se poročilo pošlje avtomatsko na različne e-mail naslove [1].

Sledenje dogodkom

Dogodki so definirani kot akcije na strani, katere ne generirajo novih ogledov na strani. Dogodkom sledimo na primer, če spletna stran vsebuje različne gradnike, vstavljene vsebine iz drugih virov in želimo videti kakšna je njihova interakcija z uporabniki. Pravzaprav kakršnem koli dogajanju na spletni strani lahko sledimo, tukaj govorimo o kliku na predvajanje, stop, izbiri

spustnega menuja, prenosu datoteke itd. Sledenje dogodkom generira poročilo ločeno od ostalih poročil. Lahko je tudi grupirano v kategorije [1].



Slika 5: Sledenje dogodkom

Gibalni grafikoni

Grafikoni dodajo sofisticirano večrazsežnostno analitiko. Izberemo lahko metrike za x-os, y-os, velikost in barvo mehurčka ter nato spremljamo kako so te metrike v interakciji med sabo [1].

API in razvijalska platforma

GA API omogoča razvijalcem razširitev GA na nove in kreativne načine. Razvijalci lahko integrirajo podatke v obstoječe produkte oziroma ustvarijo svoje aplikacije za katere niti ni potrebna prisotnost Googla [1].

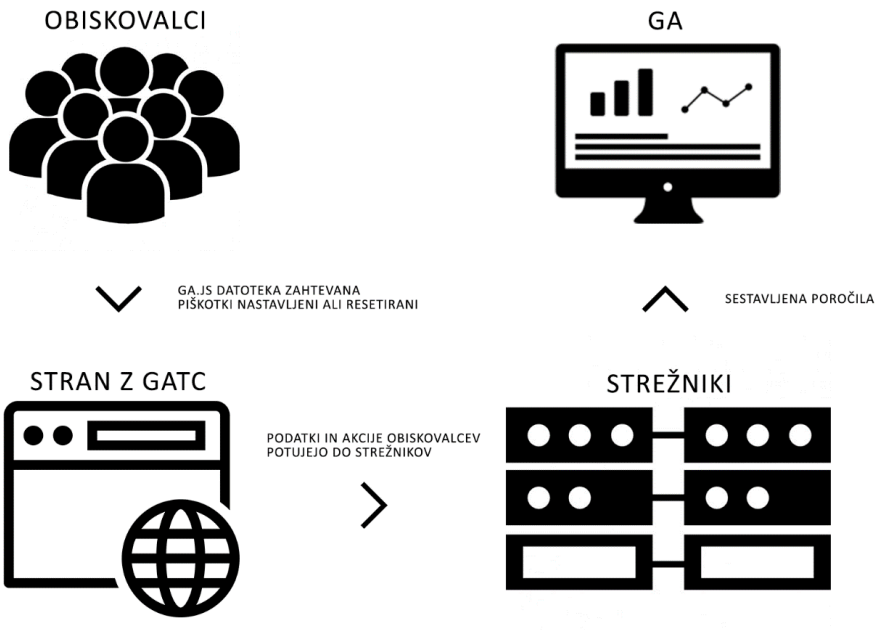
Analitična Inteligenca

Analitična inteligenca zagotavlja avtomatska opozorila za pomembnejše spremembe v podatkovnih vzorcih z naše spletne strani. Namesto, da moramo ves čas spremljati poročila in prečesavati podatke nam to delo naredi analitična inteligenca[1].

Kako GA deluje? Kot smo že omenili v prejšnjih poglavjih, je GA storitev, ki uporablja oznako na strani. S to metodo so vse zbirke podatkov, procesiranje podatkov, posodabljanje storitve in vzdrževanje upravljane s strani Googla.

Spodaj je opisan potek procesa in toka podatkov [1].

1. Ko obiskovalec pride na stran in če imamo GATC (»Google Analytics Tracking Code«), potem se avtomatično pošlje zahteva za datoteko http://www.google-analytics.com/ga.js. To je GA glavna datoteka, katera je prenesena enkrat na uporabnikovo sejo. Naslednje zahteve bodo pridobljene iz brskalnikovega predpomnilnika;
2. Za vsak ogled posamezne strani, GATC pošlje to informacijo na Googlovo zbirko podatkov. Celoten prenos do oddaljene zbirke podatkov traja zgolj delček sekunde;
3. Vsako uro, GA procesira zbrane podatke in posodobi GA poročila.



Slika 6: Delovanje Google Analytics

3. IMPLEMENTACIJA STORITVE IN PROFILIRANJE UPORABNIKOV NA SPLETU

3.1. SLEDENJE SPLETNIM UPORABNIKOM Z GOOGLE ANALYTICS

Seveda, če želimo začeti, se je potrebno najprej prijaviti oziroma ustvariti račun. Tega postopka ne bomo opisovali in bomo preskočili na dodajanje oznak na naše strani. Najpomembnejši del GA je oznaka (»GATC«), katera se prilepi na stran. Brez nje se podatki o uporabnikih ne bodo zbirali. GATC je delček kode JavaScript, ki je prilepljen na strani. Koda je skrita in deluje kot posrednik za zbiranje informacij o obiskovalcih in pošiljanje zbranih podatkov na strežnik.

Sledenje spletnim mestom

To je koda za sledenje storitve Universal Analytics za to znamko.
Če želite izkoristiti vse prednosti storitve Universal Analytics za to znamko, kopirajte in prilepite to kodo na vsako spletno stran, ki ji želite slediti.

```
<script>
  (function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[r]=i[r]||function(){
  (i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new Date();a=s.createElement(o),
  m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;m.parentNode.insertBefore(a,m)
  })(window,document,'script','https://www.google-analytics.com/analytics.js','ga');

  ga('create', 'UA-[redacted]', 'auto');
  ga('send', 'pageview');

</script>
```

Programska koda 3:Koda za sledenje spletnim mestom

Tako, ko to prilepimo v glavo ali nogo HTML datoteke, se bo koda izvršila na vsaki strani in tako se bodo beležile informacije, katero stran je uporabnik obiskal, koliko časa je bil na spletnem mestu itd..

Ko imamo GATC na strani, nato samo čakamo, da nam Google sestavi poročilo o naših pridobljenih informacijah. Pri storitvi ni potrebno namestiti podatkovnih baz ali imeti gostovanje, saj za vse to poskrbi Google sam.

3.2. PROFILIRANJE SPLETNIH UPORABNIKOV

Profiliranje uporabnika lahko definirano kot proces identificiranja podatkov o uporabnikovih interesih. Informacijo lahko sistemi uporabijo, da bi bolje spoznali uporabnika, ga razumeli in s tem povečali zadovoljstvo ob prihodnjih obiskih na strani [18].

Zajemanje informacij o uporabnikih in njihov interesih je glavna funkcionalnost ustvarjanja profilov. V zadnjem času je bilo narejenih veliko raziskav na področju priporočilnih sistemov in različnih tehnik profiliranja. Profiliranje se je v grobem razvilo s pomočjo napredka v podatkovnem rudarjenjem in z razvojem strojnega učenja. Starejši sistemi so delovali tako, da so pridobivali podatke direktno od uporabnika, ampak ta metoda ni bila uspešna, saj niso želeli izpolnjevati dodatna vnosna polja in podati informacijo direktno. Dandanes je profiliranje bolj osredotočeno na pridobivanje podatkov, ki temeljijo na dejanjih uporabnika [18].

Eksplicitna tehnika profiliranja

Pri tem pristopu je obnašanje uporabnika predvideno z že pridobljenimi podatki. Ti podatki so največkrat pridobljeni z izpolnjevanjem anket ali kakšnih drugih obrazcev. Tehniki se tudi drugače reče statično profiliranje. Težava pri tej tehniki je, da se uporabnik izogiba izpolnjevanju obrazcev, tako da natančnost profiliranja ni visoka.

Implicitna tehnika profiliranja

Danny Poo [19] razloži tudi dinamični pristop profiliranja, pri katerem sledimo vedenju uporabnika brez njegove vednosti. Torej uporabimo piškotke oziroma katero drugo tehniko sledenja in z njo zbiramo potrebne podatke, ki nam bodo koristile pri poznavanju uporabnika v prihodnosti.

Hibridna tehnika profiliranja

Kot nam že ime pove je ta tehnika kombinacija zgornjih dveh. Ta pristop zagotavlja učinkovitejše profiliranje[18].

3.3. GRUČNI ALGORITMI

3.3.1. K MEANS

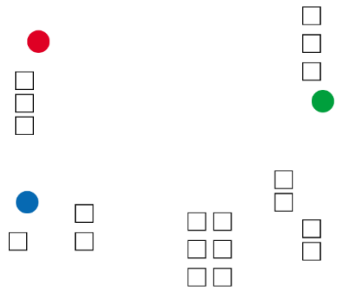
K-means je primer nenadzorovanega učenja, kateri je uporaben, ko imamo podatke pri katerih kategorije oziroma skupine niso definirane. Cilj algoritma je, da najde skupine (toliko, kot jih želimo imeti) v danem podatkovnem naboru. Algoritem iterativno dodaja podatke k ustvarjenim skupinam. Rezultati K-means gručenja so [20]:

1. Centroidi K skupin, ki se lahko uporabljajo za označevanje novih podatkov;
2. Podatki so razvrščeni po skupinah.

V poslovnem svetu se algoritem uporablja, da bi našli skupine, katere niso bile eksplicitno določene. Ta podatek, bi jim pomagal pri nadaljnjih odločitvah, še posebej pri segmentaciji strank, definiranju uporabnikov glede na interese, ločevanje med uporabniki in boti.

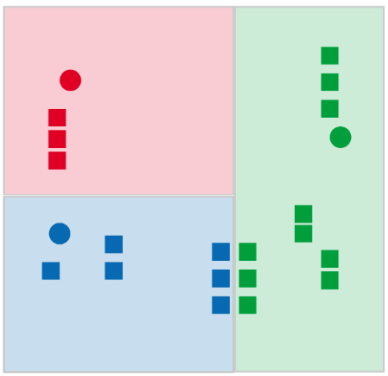
Cilj algoritma je minimiziranje objektne funkcije, ki je dejansko kvadratna funkcija napake. Algoritem pa deluje tako:

1. K točke (v tem primeru je k=3) predstavljajo začetne skupine centroidov;



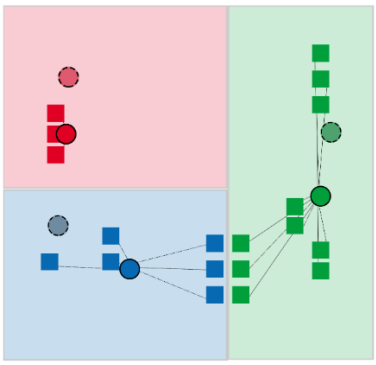
Slika 7: Ustvarjanje začetnih skupin

2. Skupine se ustvarjajo z združevanjem najbližjih centroidov;



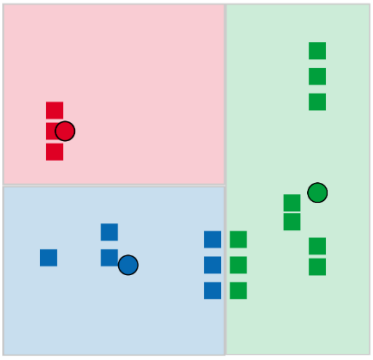
Slika 8: Združevanje skupin

3. Centroid vsake skupine postane novo povprečje;



Slika 9: Novo ustvarjeni centroidi

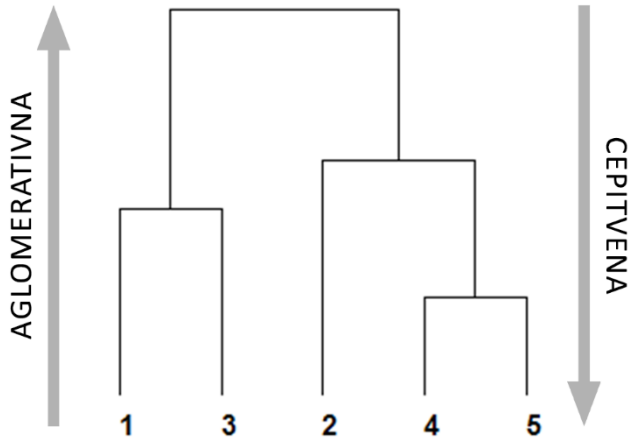
4. Koraka 2 in 3 se ponavljata dokler ni dosežena konvergenca[21].



Slika 10: Končne skupine

3.3.2. HIERARHIČNO GRUČENJE

Hierarhično gručenje vsebuje skupine, katere imajo vnaprej določeno sortiranje od zgoraj navzdol. Na primer tudi datoteke in mape na trdem disku so organizirane hierarhično[22]. Hierarhično gručenje zgradi hierarhijo gruč oziroma drevo gruč, drevesu pravimo tudi dendrogram ali drevo združevanja. Obstajata dve metodi gručenja. Ena gradi hierarhijo od spodaj navzgor, temu pravimo aglomerativna metoda gručenja. Začne se z eno točko in se nato rekurzivno dodajata 2 oziroma več gruč. Konča se, ko dosežemo k število gruč. Druga metoda pa se imenuje metoda cepitve. Ta metoda je obratna prvi, torej gradi hierarhijo od zgoraj navzdol. Metoda se začne z veliko gručo, katera se rekurzivno deli na manjše gruče. Konča se ko dosežemo k število gruč [23].



Slika 11: Dendogram

Preden se opravi gručenje, je potrebno določiti matriko razdalj. Nato se med procesom najbližje gruče združujejo v nove in na koncu staro matriko nadomesti nova. Obstajajo tudi različne metode kako je izračunana razdalja med vsako gručo. Navedli bomo samo metode, brez natančnejših opisov[22].

Minimalna, maksimalna in povprečna metoda

Minimalna metoda je razdalja med dvema gručama je definirana kot najkrajša razdalja med dvema točkama v vsaki gruči. Maksimalna metoda je ravno obratna kot minimalna, torej je

definirana kot najdaljša razdalja med dvema točkama v vsaki gruči. Povprečna metoda je razdalja definirana kot povprečna razdalja med vsako točko v eni gruči in vsemi točkami v drugi gruči[22].

3.3.3. DBSCAN

DBSCAN je algoritem, ki skuša najti gruče glede na njihovo gostoto točk . Ključna ideja algoritma je, da združuje točke katere so tesno skupaj (točke z veliko sosedi). DBSCAN je eden izmed najbolj uporabljenih algoritmov. DBSCAN loči točke na tri tipe [23]:

- Jedrne točke: To so točke katere so znotraj gruče;
- Mejne točke: Mejne točke so točke, katere niso temeljne, ampak so sosede temeljnim točkam;
- Točke šuma: So katere koli točke, katere niso temeljne ali mejne.

Iskanje gruč pri DBSCAN algoritmu poteka v slednjem vrstnem redu: Potrebno je vnesti dva parametra, to sta velikost okolice in minimalno število točk v okolici. Začne se z poljubno izhodiščno točko, katera še ni bila obiskana. Ko se točka izbere, se preveri ali ima okoli sebe dovolj sosednjih točk, če ima, potem se začne proces gručenja, v nasprotnem primeru označimo točko kot šum.

Če se ugotovi, da je točka del gruče, potem isto velja tudi za njeno sosesko. Algoritem ves čas sprejema nove točke in če so si točke med seboj dovolj blizu spadajo pod isto gručo[23].

3.4. ORODJE ZA PODATKOVNO RUDARJENJE R

R je brezplačno programsko okolje za statistično obdelavo in grafični prikaz podatkov. Deluje na UNIX platformah na Windows in MacOS sistemih. Je GNU projekt in ima svoj programski jezik imenovan R. R zagotavlja širok spekter statističnih (linearno in nelinearno modeliranje, klasični statistični testi, klasifikacija, gručenje, …) in grafičnih tehnik, hkrati pa je zelo razširljiv [24].

R in njegovo okolje

R je skupek programske opreme za manipulacijo nad podatki, izračune in grafične prikaze. Vključuje [24]:s

- Učinkovito obdelavo in hranjenje podatkov;
- Zbirko operatorjev za izračune polj in matrik;
- Veliko zbirko orodij za analizo podatkov in njihovem prikazu;
- Dobro razvit, enostaven in učinkovit programski jezik, ki vključuje pogoje, zanke in uporabniško določene rekurzivne funkcije.

Izraz omenjen v naslovu - okolje naj bi ga opredelil, kot popolnoma načrtovan skladen sistem. R je zasnovan okrog lastnega programskega jezika in uporabnikom omogoča dodajanje dodatnih funkcionalnosti z definiranjem novih funkcij. Napredni uporabniki lahko uporabljajo programski jezik C in tako lahko direktno manipulirajo z objekti R-ja [24].

R je tudi razširljiv. Razširi se ga preko paketov, ki jih najdemo na CRAN spletni strani. R ima svoj lasten format dokumentacije imenovan LaTeX. Ta se uporablja za dostavo celovitih dokumentacij [24].

Slabosti in prednosti programa

Prednosti so [25]:

- Brezplačna uporaba;
- Uporablja veliko knjižnic za različne statistične obdelave;
- Velika skupnost in podpora;
- Delovanje na različnih operacijskih sistemih.

Slabosti so [25]:

- Ni uporabniško prijazen. Potrebno napisati veliko kode, da pridobimo rezultat

- Izračuni v pomnilniku samem, dovoljujejo procesiranje velikosti podatkov toliko kot je RAM pomnilnik.

4. PREGLED PRIDOBLJENIH PODATKOV IN GRADNJA UPORABNIŠKIH PROFILOV

Kot že omenjeno v prejšnjih poglavjih smo podatke zbirali s pomočjo Google Analytics. Podatki so se zbirali en mesec. V tem času smo pridobili približno 67.000 uporabnikov. Te podatke bomo izvozili in jih uporabili pri profiliranju. V diplomski nalogi smo se odločili, da bomo delali z metrikami uporabnikov, s katerimi lahko ugotovimo v katerem cenovnem razredu uporabniki iščejo ponudbe.

Začeli smo tako, da smo povezali program R in storitev Google Analytics med seboj. Ker je program R zelo agilen, omogoča razširitve. Obstaja tudi razširitev za Google Analytics, s katero lahko direktno dostopamo do podatkov iz spletne strani, kateri so shranjeni v Googlovi bazi in tako pišemo svoja povpraševanja in sestavljamo svoja poročila.

Na začetku smo si na računalnik namestili program R in program R-studio, kateri deluje popolnoma isto kot program R vendar ima bolj uporabniku prijazen vmesnik. Ko smo si namestili potrebna programa, smo namestili paket Google Analytics.

Ko se namestitev razširitve konča, lahko začnemo knjižnico uporabljati. Kot vidimo na spodnji sliki, moramo najprej nastaviti knjižnico, s katero bomo delali. Nato se s funkcijo `ga_auth()` prijavimo v naš GA račun in izberemo profil v katerem želimo izvajati povpraševanja. S temi koraki smo povezali R in Google Analytics ter omogočili delovanje razširitve.

Iz Google Analytics smo vzeli naslednje potrebne podatke za obdelavo: userId*, country, pagepath, stOseb*, date in cena*.

* spremenljivka je bila ustvarjena po meri

```
library(googleAnalyticsR)
ga_auth()
start_date <- "2017-06-01"
end_date <- "2017-07-02"
id <- [redacted]
ga_data <- google_analytics(id = id,
                            start = start_date,
                            end = end_date,
                            metrics = "ga:itemQuantity",
                            dimensions = "ga:pagepath, ga:dimension1, ga:dimension2,
                            ga:dimension3, ga:date, ga:country",
                            max = 67000)
2017-07-15 13:57:06> Request to profileId: [redacted]
2017-07-15 13:57:06> Fetched: itemQuantity pagepath dimension1 dimension2 dimension3 date country.
[67000] total results out of a possible [67267], Start-Index: 1
```

Programska koda 4: Povpraševalni stavek

Da izvedemo preprosto povpraševanje in prikažemo rezultate moramo določiti ID profila, s katerega želimo pridobivati podatke, nastaviti moramo časovno obdobje metrike in dimenzije. Ostali parametri, kot so na primer segmenti, največje možno število rezultatov, sortiranje niso obvezni. Primer povpraševalnega stavka lahko najdemo na zgornji sliki.

Iz stolpca, kjer so zapisani URL-ji, smo programsko izrezali niz, tako, da smo dobili samo unikatno številko iskanih regij. Na primer iz povezave `/rezultati/?rangeType=variable&rg=2788516&persons=2`, smo pridobili unikatno številko 2788516. Tako smo si zagotovili lažjo obdelavo podatkov.

Vse podatke smo preoblikovali v numerično obliko, tako da jih lahko gručni algoritem obdela. Najprej smo se lotili združevanja uporabnikov in regij. Ustvarili smo novo tabelo poimenovano uporabniki in uporabili slednji ukaz.

```
# dodaj id regij k uporabniku
uporabniki <- merge(x=uporabniki, y=regije, by.x='userId', by.y='userId', sort=TRUE, all=TRUE)

uporabniki$regije <- as.numeric(uporabniki$regije)
```

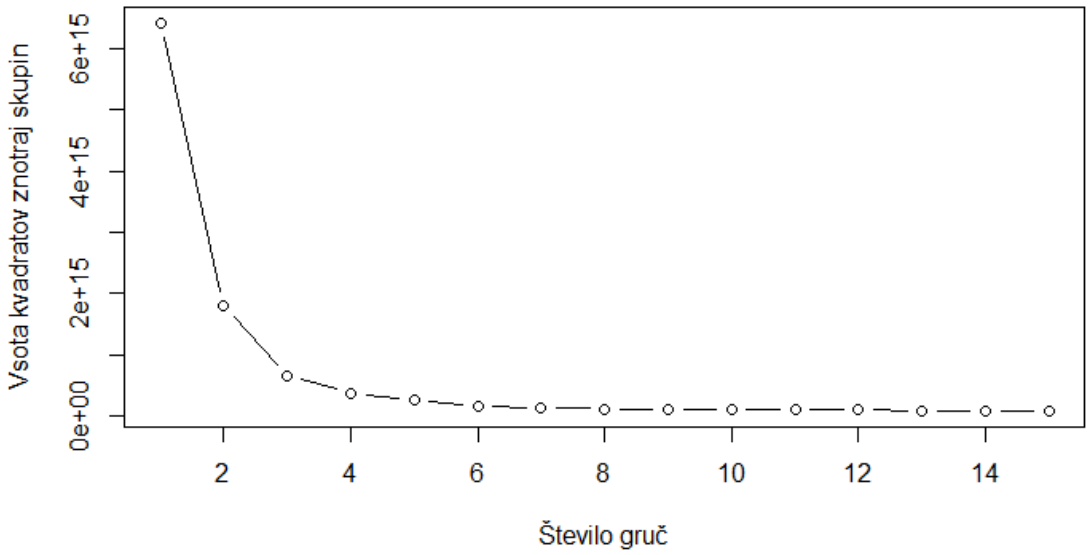
Programska koda 5: Združevanje tabel

Nato smo izračunali še frekvenco, koliko krat je bila regija iskana, ter ponovili ukaz iz zgornje slike in združili podatke v eno tabelo. Postopke smo uporabili še za združitev države in uporabnika, ter izračunali povprečje koliko so produkti, ki jih je uporabnik iskal, stali. Zadnji

podatek nam bo veliko pripomogel, saj bomo lahko z njim ugotovili, kateri produkti mu bolje ustrezajo glede na ceno.

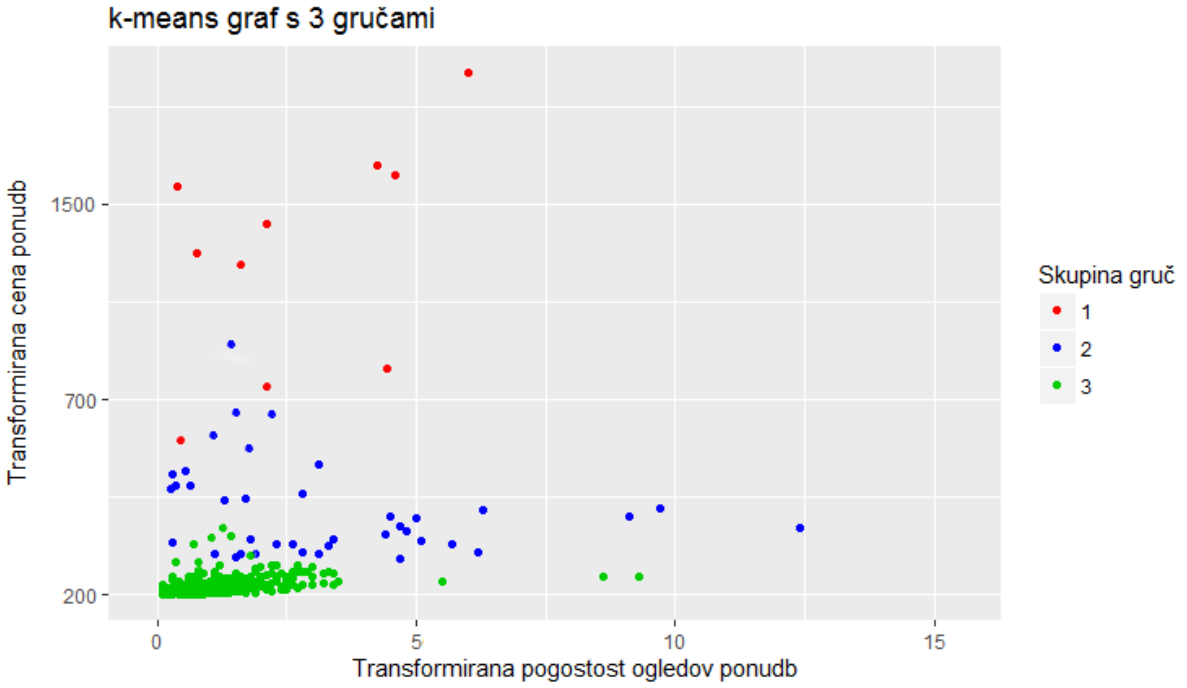
K-means gručenje najbolje deluje s sorazmerno porazdeljenimi vhodnimi spremenljivkami. Standardizacija vhodnih spremenljivk je zelo pomembna, v nasprotnem primeru se pojavljajo večja odstopanja in tako večji vpliv na končne rezultate.

Ko smo vse podatke obdelali in ustvari končno tabelo za profiliranje smo izvedli meritve, ki nam bodo pomagale pri gručenju. Izračunali smo vsoto kvadratov znotraj skupin. Vsota kvadratov znotraj skupin predstavlja kot nekakšno merilo enakosti. Na vsakem koraku se najbližje skupine združijo v gručo. Razdaljo med skupino in gručo izračunamo s pomočjo Evklidske enačbe, tako da merimo razdaljo med skupino in središčem gruče. Izračun vsote kvadratov znotraj skupine se ponavlja tako dolgo, dokler ne doseže največje število ponovitev. Naši rezultati na spodnji sliki sicer ne kažejo najboljših rezultatov. Kot vidimo že pri dveh gručah začnemo izgubljati vrednosti.

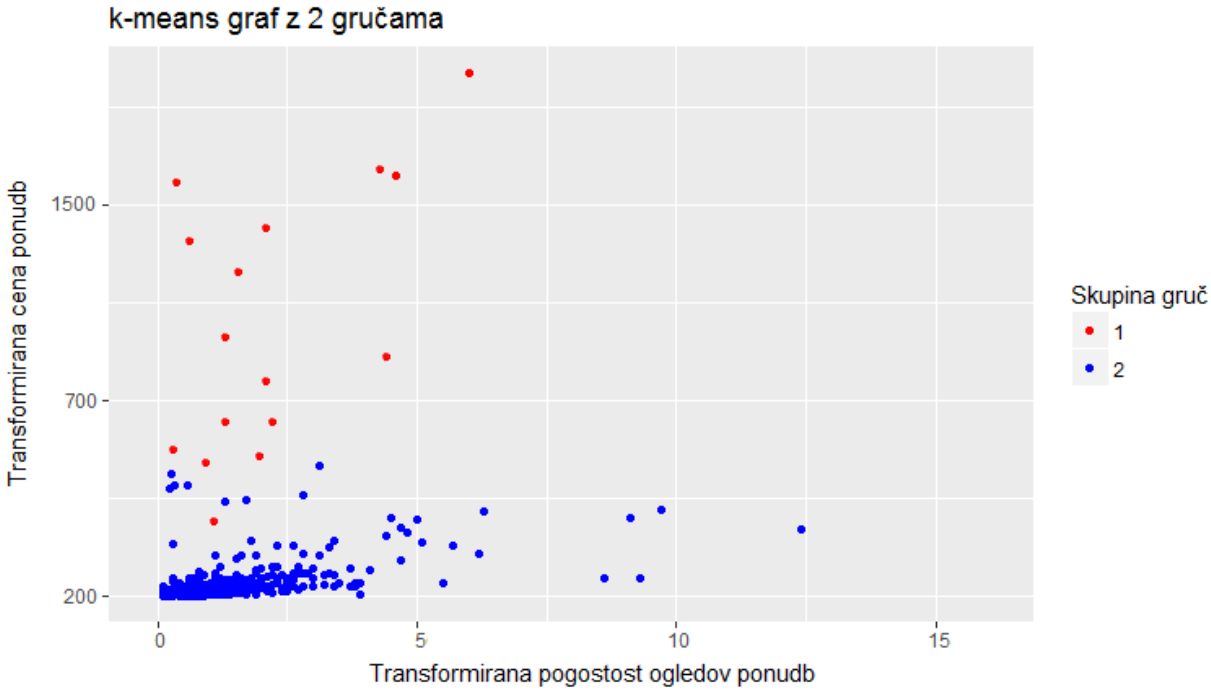


Slika 12: Vsota kvadratov znotraj skupin

Ker nam je zgornji test vsot kvadratov znotraj skupin pokazal, kater gruče nam lahko prinesejo najboljše podatke, smo se odločili, da bomo uporabili gručni algoritem z dvema in tremi gručami. Tako bomo imeli še vizualno predstavo o pridobljenih podatkih in lažje razumevanje.



Slika 13: Vizualizacija gručenja s 3 gručami



Slika 14: Vizualizacija gručenja z 2 gručama

Zadnji dve sliki sta vizualna prikaza gručenja. Razlika med gručenjem s 3 in 2 gručami ni velika, vidimo lahko, da je modra barva zavzela uporabnike, kateri so bili prej obarvano z zeleno,

opazimo lahko tudi več rdečih točk v spodnjem delu grafa, kar nakazuje, da se je centroid gruče prestavil nižje. Različne barve točk predstavljajo različne skupine gruč. Če so točke iste barve, pomeni ta spadajo pod isto gručo in imajo skupne lastnosti.

Sodeč po obeh grafih lahko sklepamo, da uporabniki obarvani z modro in zeleno barvo po večni iščejo ponudbe, ki so v istem cenovnem razredu. Na grafu z 3 gručami lahko ugotovimo, da zelene točke predstavljajo uporabniki, kateri iščejo ponudbe od 200€ pa vse do 400€, te ponudbe pa si ogledajo tudi do 4-krat. Težje je profilirati uporabnike modre gruče, kajti njihov razpon iskanih ponudb se razteza vse od 300€ do 700€. Pri tej gruči lahko tudi opazimo, da višja kot je cena ponudb majn pogosto so si jih ogledali. Rdeči uporabniki iščejo po ponudbah višjega cenovnega razreda in hkrati manj pogosto kot ostali uporabniki. Njihovi iskalni parametri cen se gibljejo od 600€ in vse do 1500€. Zanimivo je, da si ogledajo zelo malo ponudb-samo do 2.5 ponudbe na sejo. Torej pri teh uporabnikih bi lahko zelo napredovali, če bi jim priporočili prave ponudbe za njih.

4.1. UGOTOVITVE

Kar smo pri implementaciji, obdelavi in prikazu rezultatov ugotovili je to, da nam lahko spletna analitika bistveno pripomore pri ugotavljanju kaj uporabniki želijo in kako jim lahko pri tem pomagamo. Spletna analitika nam omogoča neomejeno število podatkov, s katerimi se lahko poigramo in ugotovimo stvari, katere nas zanimajo.

S pridobljenimi rezultati oziroma podatki bi lahko z veliko uporabnih informacij in na podlagi teh oblikovali uporabniške profile in jim lažje priporočali ponudbe. Uporabniki iste gruče, bi dobivali priporočila ponudb, katere uporabniki iščejo znotraj te gruče. Tako, ne bi uporabnik, ki išče potovanje v Afriko, dobil priporočila naj si ogleda skandinavske dežele.

Za povratne informacije, bi lahko tudi postavljali vprašanja uporabnikom ali jim je priporočena ponudba olajšala iskanje na strani ali se jim je zdela neuporabna.

Na podlagi podatkov, bi lahko ugotovili še več različnih informacij. Ugotovili bi lahko na primer koliko Slovencev išče ponudbe v določenem cenovnem razredu ali pa koliko ljudi je obiskalo ponudbe istih destinacij.

Če bi želeli še globje raziskati naše uporabnike, bi lahko šli še dlje in ustvarili sledenje dogodkom uporabnikov in tako ugotovili, koliko ljudi je rezerviralo ponudbe in koliko ponudb je pred tem že pregledal. Tako, bi lahko natančno ugotovili, kaj določeni kupci iščejo in kaj jih privlači, da rezervirajo določeno ponudbo.

5. ZAKLJUČEK

Najpomembnejše vprašanje, ki si ga lahko zastavimo je: kako lahko prepričamo uporabnike, da si ogledajo naše storitve, kupijo izdelek, naročijo novice? Odgovor se skriva v podatkih. Potrebno je pogledati v podatke in razumeti, kaj se dogaja na spletni strani. Odkriti moramo, kaj si uporabniki želijo in personalizirati stran, da jim čim bolje služi.

Diplomsko delo smo si razdelili na dva dela. Na teoretični in praktični del. V teoretičnem delu smo predelali metodologije zbiranja podatkov, opisali kaj so JavaScript oznake, kako se beležijo dnevniške datoteke in ostale tehnologije zbiranja podatkov. Nato smo opisali kaj spletna statistika je in kaj nam omogoča. Povedali smo kaj so piškotki in opisali vrste na katere jih delimo, zajeli smo tudi zasebnost na spletu, kje so uporabniki najbolj skeptični in opisali storitev za spletno analitiko. V praktičnem delu smo najprej vzpostavili storitev za spletno analitiko Google Analytics in tako vzpostavili sledenje uporabnikom, podatke smo pridobili z integracijo storitve s programom za podatkovno rudarjenje R. Podatke smo preoblikovali, da so bili ustrezni za gručni algoritem. Ko smo imeli končne podatke smo izvedli še testiranje, na koliko gruč je najbolj smiselno razdeliti uporabnike in rezultate vizualno prikazali. Ugotovili, da uporabniki po večini gledajo ponudbe v istem cenovnem razredu. Ker trenutno na spletni strani ni vzpostavljenega priporočilnega sistema, nam ti podatki pridejo zelo prav in bi se lahko iz njih razvil priporočilni sistem.

V diplomskem delu smo dosegli vse cilje, katere smo si zadali. Torej raziskali smo delovanje storitev za sledenje uporabnikom, poiskali smo ustrezno rešitev ter jo integrirali v program za podatkovno rudarjenje. Raziskali smo uporabnost pridobljenih podatkov in ustvarili uporabniške profile.

Soočili smo se tudi z nekaj ovirami. Kar smo spoznali za zelo motečo so omejitve storitve GA. GA ni odprto-kodna rešitev, tako, da ni možno pridobiti vseh podatkov. Na primer GA meri tudi razne demografske podatke uporabnikov, a le teh ni možno pridobiti glede na uporabniški ID, prav tako isto velja za lokacijske podatke. Prav tako se je težava pojavila v programu R, kateri lahko vrne maksimalno 10 000 vrstic. Tako smo morali izvesti več povpraševalnih stavkov, in združevati tabele, da smo dobili končno z vsemi rezultati.

Za nadaljnje delo, glede na to, da podjetje na spletni strani nima vzpostavljenega priporočilnega sistema, bi jim ga lahko vzpostavili, s tem izboljšali uporabniško izkušnjo in povečali prihodke podjetja. Z uporabljenim programom R, se lahko programsko povezujemo na spletno storitev in tako bi lahko sproti posodabljali podatke za priporočanje.

6. VIRI

[1] B. Clifton, *Advanced web metrics with Google Analytics*. Wiley, 2012.

[2] N. Schmucker, "Web Tracking," *SNET2 Semin. Pap.*, 2011.

[3] Waa, "Web Analytics Definitions [Report]," *Web Anal. Assoc.*, pp. 1–32, 2007.

[4] J. Lovett, "US Web Analytics Forecast, 2008 To 2014." Forrester, Cambridge, p. 9, 2009.

[5] A. Ganapathi and S. Zhang, "Web Analytics and the Art of Data Summarization."

[6] A. Kaushik, *Web analytics: an hour a day*. 2007.

[7] Z. Guangzhi and S. Peltsverger, "Web Analytics Overview," in *Encyclopedia of Information Science and Technology*, Third., IGI Global, Editors: Mehdi Khosrow-Pour, 2015.

[8] D. Waisberg and A. Kaushik, "Web Analytics 2.0: empowering customer centricity," *Orig. Search Engine Mark. …*, vol. 2, no. 1, p. 7, 2009.

[9] "Cookie - HTTP | MDN," 2017. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Cookie. [Accessed: 11-Jul-2017].

[10] D. M. Kristol, "HTTP Cookies: Standards, Privacy, and Politics," May 2001.

[11] "Set-Cookie - HTTP | MDN," 2017. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Set-Cookie. [Accessed: 11-Jul-2017].

[12] A. Barth, "HTTP State Management Mechanism," RFC Editor, 2011.

[13] A. Gervais, R. Shokri, A. Singla, S. Capkun, and V. Lenders, "Quantifying Web-Search Privacy," *Proc. 2014 ACM SIGSAC Conf. Comput. Commun. Secur.*, pp. 966–977, 2014.

[14] D. L. Hoffman, T. P. Novak, and M. Peralta, "Building Consumer Trust Online," *Commun. ACM*, vol. 42, no. 4, pp. 80–85, Apr. 1999.

[15] S. Trepte and L. Reinecke, *Privacy online : perspectives on privacy and self-disclosure in the social web*. Springer-Verlag, 2011.

[16] W. Chung and J. Paynter, "Privacy issues on the Internet," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2002, vol. 2002–Janua.

[17] J. R. Mayer and J. C. Mitchell, “Third-Party Web Tracking: Policy and Technology,” in *2012 IEEE Symposium on Security and Privacy*, 2012, pp. 413–427.

[18] S. Kanoje, S. Girase, and D. Mukhopadhyay, “User Profiling Trends, Techniques and Applications,” *Int. J. Adv. Found. Res. Comput.*, vol. 1, no. 1, 2014.

[19] D. Poo, B. Chng, and J.-M. Goh, “A Hybrid Approach for User Profiling.”

[20] A. Trevino, “Introduction to K-means Clustering,” 2016. [Online]. Available: https://www.datascience.com/blog/introduction-to-k-means-clustering-algorithm-learn-data-science-tutorials. [Accessed: 27-Jul-2017].

[21] “Clustering - K-means.” [Online]. Available: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html. [Accessed: 28-Jul-2017].

[22] S. Sayad, “Hierarchical Clustering.” [Online]. Available: http://www.saedsayad.com/clustering_hierarchical.htm. [Accessed: 28-Jul-2017].

[23] B. Chaudhari and M. Parikh, “A Comparative Study of clustering algorithms Using weka tools,” *Int. J. Appl. or Innov. Eng. Manag.*, vol. 1, no. 2, 2012.

[24] CRAN, “R: What is R?” 2016.

[25] M. Brys, *Using Google Analytics with R*. 2016.