

Analiza ponašanja posjetitelja kataloga weba

Boris Stanić, Marin Vuković, Krešimir Pripužić

Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, Unska 3, Zagreb

Sažetak – Google Analytics, nakon što se uspješno implementira u stranicu, prati ponašanje svih korisnika koji uđu na stranicu i o tome vodi statistiku. Rezultate praćenja je moguće prema potrebi dohvatiti za analizu. U ovom su radu korištena mjerena iz Google Analyticsa za utvrđivanje popularnih kategorija službenog CARNetovog kataloga WWW poslužitelja u Hrvatskoj. Analiza je provedena uz pomoć razvijenog modela koji se temelji na matematičkoj statistici i pseudo slučajnosti.

Uvod

Prema podacima [1] ITU-a (*International Telecommunication Union*) za 2014. godinu u svijetu je bilo 2.9 milijarde registriranih korisnika interneta. Predviđeno je da će se brojka od 3 milijarde korisnika probiti početkom 2015. godine. Broj internet stranica dostupnih korisnicima iznosi otprilike 1.1 milijardu. Internet promet po svakoj pojedinačnoj vezi također raste zahvaljujući sve bržim vezama, koje omogućuju pretraživanje sadržaja koji je memorijski zahtjevniji (na primjer video, audio, slike visoke rezolucije), na pametnim telefonima i sličnim uređajima koji su posebno dizajnirani za obradu i prikaz takvog sadržaja. Nadalje, odstupa se od prakse da se korisnik spaja na mrežu prema potrebi, već je sve više korisnika spojeno cijelo vrijeme, naročito u razvijenijim zemljama.

Vlasnicima internet stranica u interesu je znati kako je sadržaj koji objavljaju prihvaćen od strane korisnika. Najpopularniji servis današnjice za praćenje prometa na stranici je Google Analytics. Prema istraživanju [2] servisa Pingdom koji služi za praćenje dostupnosti i performansi internet stranica iz 2008. godine, 32% od 500 najvećih internet stranica koristi Google Analytics. Istraživanje [3] iz 2015. godine provedeno od strane W3Techsa, stranice specijalizirane za analizu tehnologija zastupljenih u izradi internet stranica, kaže kako 82.2% stranica koristi Google Analytics.

Ideja ovog rada je da na temelju podataka iz Google Analyticsa za stranicu www.hr, koja je službeni CARNetov katalog WWW poslužitelja u Hrvatskoj, izvršiti analizu i ponuditi 3 kategorije koje se mogu istaknuti na početnoj stranici kao preporuka posjetiteljima.

Programsko rješenje je pisano u jeziku Java, a kao ulazni parametar u program potrebno je navesti put do dokumenta s podacima o posjetama. Dokument s podacima potrebno je preuzeti iz servisa Google Analytics u formatu CSV. Nakon pokretanja programa, on će kao rezultat izlistati rangirane kategorije stranice www.hr.

Matematičko modeliranje

Matematički model [4] je matematički opis nekog realnog sistema. Matematički modeli se koriste u širokom spektru područja kao što su inženjerstvo, psihologija, sociologija, statistika idr., odnosno primjenjivi su u svim područjima gdje se vrše mjerena i pokušava naći korelacija između izmjerjenih rezultata. Oni pomažu kada je teško izraziti eksplizitnu pravilnost između nekih podataka, ali je jasno da veza postoji.

Točan postupak za razvijanje matematičkih modela ne postoji. Ako se pokuša izjednačiti dva slučaja kao istovjetna gotovo uvijek se pokaže da postoje parametri koji nisu istovjetni i matematički model koji iznimno dobro funkcioniра za jedan slučaj će pokazati lošije rezultate za drugi slučaj. Tada je najbolje rješenje primijeniti postojeći model na novi slučaj, ali potrebno je unijeti preinake koje će kompenzirati razlike između njih. Razvitak matematičkog modela nikada nije gotov jer je uspješnost i pouzdanost vrlo teško mjeriti. Jedini način za mjerjenje napretka matematičkog modela je usporedba rezultata izmijenjenog modela i modela bez promjena. Odnosno, rezultati koje matematički modeli daju nisu absolutni, već relativni. Naročito je teško razvijati model za slučajeve kada se neki parametri neprestano mijenjaju ili u još gore slučaju kada neki parametri ulaze i izlaze iz sistema.

Programsko rješenje i model razvijeni u ovom radu nije prvi pokušaj da se pokuša na temelju prijašnjeg ponašanja korisnika predvidjeti interes kako starih korisnika tako i korisnika koji po prvi put dolaze na stranicu. Najdalje u području predviđanja afiniteta korisnika su došli internet radiji s ugrađenom podrškom za prepoznavanje ukusa glazbe korisnika [5].

Kada korisnik prvi put pokrene internet radio s ugrađenom takvom podrškom reproducira mu se neka proizvoljna pjesma. Korisnik tada ima izbor ocijeniti pjesmu koju

trenutno sluša i ovisno o njegovoj ocjeni implementirani model uči o njegovom glazbenom ukusu. Ukoliko korisnik trenutačnu glazbu ocijeni lošom ocjenom, model prekida trenutačnu pjesmu i reproducira novu koja po svojim karakteristikama je bitno drugačija od prethodne. Model također pamti da se korisniku glazba s tim karakteristikama ne sviđa. Svakom ocjenom korisnika, model dodatno uči i nakon nekog vremena vrlo uspješno predlaže glazbu korisniku koja se njemu sviđa. Model predlaže glazbu neovisno o izvođaču ili žanru, on parametrizira glazbu po njezinim vokalno-akustičnim parametrima. Proces učenja modela moguće je ubrzati ako se pri prvoj posjeti stranici radija ispuni anketa gdje korisnik može odabrati vrstu glazbe koja se njemu općenito sviđa.

Modeli razvijeni za tu namjenu vrlo su kompleksni. Istovremeno uče o samom korisniku, ali i o glazbi koja se nalazi u bazi podataka. Ukoliko korisnik ocjeni neku pjesmu visokom ocjenom, model za sljedeću pjesmu odabire onu koju su drugi korisnici ocijenili visokom ocjenom, a također su ocijenili visoko i prethodno puštenu pjesmu. Za svaku novu pjesmu dodanu u bazu podataka model ispočetka kreće s učenjem.

Složenost modela uvelike utječe na njegovu pouzdanost i točnost. Složeniji modeli davat će točnije rezultate. Problem kod takvih modela je što često nisu fleksibilni i male promjene u parametrima ili ulazak nekih novih parametara u sistem mogu uvelike poremetiti njihovu točnost. Jednostavniji modeli nisu jako podložni nenadanoj promjeni parametara, ali rezultati koji oni daju često mogu biti konfuzni. Kako zanemaruju određene parametre, ponekad promjena u tim parametrima koje je neki jednostavniji model zanemario može se pokazati vitalnom za točnost rezultata. Kada se odlučuje o složenosti modela, zapravo se odlučuje o razini točnosti i razini robusnosti.

Implementirani model

Model implementiran [6] u programskom rješenju nastao je kombinacijom matematičke statistike i subjektivne procjene parametara. Dio modela koji primjenjuje matematičku statistiku pridonosi dosljednosti modela i daje matematički temelj modelu, dok parametri dobiveni subjektivnom procjenom čine model pogodnim za prilagodbu i daljnja poboljšanja jer promjenom tih parametara može se utjecati na rezultat.

A priori informacije koje model koristi dobivene su iz Google Analyticsa i subjektivnom procjenom. Parametri iz Google Analyticsa su ukupni pregledi i stopa

napuštanja (izražena u postocima), a parametar dobiven subjektivnom procjenom je težinska ocjena važnosti dubine.

Kada model primi datoteku iz Google Analyticsa, za svaku podstranicu www.hr pamti broj ukupnih pregleda u odabranom vremenskom periodu i stopu napuštanja za taj period. Sljedeći korak je izračunavanje koliki broj od ukupnih pregleda otpada na preglede koji su završili napuštanjem. Parametar o takvim posjetama se dobiva množenjem broja ukupnih pregleda sa stopom napuštanja podijeljenom s 100. Također za model je potrebna i informacija o broju pregleda u kojima nije došlo do napuštanja. Taj parametar se dobiva oduzimanjem parametra o posjetama s napuštanjem od ukupnog broja posjeta. Parametri o posjetama kada je nastupilo napuštanje i posjetama kada nije nastupilo su potrebni za izračunavanje pouzdanosti uz pomoć Wilsonove formule.

$$\frac{\hat{p} + \frac{1}{2n} z_{1-\alpha/2}^2 \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{1}{n} z_{1-\alpha/2}^2} \quad (1)$$

Wilsonova formula (1) se pokazala iznimno dobrom za primjenu kako u ovom modelu, tako i u mnogim internetskim stranicama koje rangiraju svoj sadržaj kao što su komentari ili objave (npr. Reddit [7]). Unutar formule se pojavljuju tri parametra. Parametar \hat{p} označava broj posjeta kada nije nastupilo napuštanje, parametar n je ukupni broj pregleda, a parametar $z_{(1-\alpha/2)}$ je kvantil normalne distribucije. U programskom rješenju se koristi s intervalom pouzdanosti od 80%. Interval pouzdanosti u statističkim mjeranjima predstavlja raspon mogućih vrijednosti unutar kojeg se s izvjesnom vjerojatnosti nalazi ta statistička mjera populacije. Što su granice intervala uže, preciznost procjene je veća. Najveća prednost Wilsonove formule je što ima dobra svojstva za mali broj posjeta i za ekstremne vjerojatnosti. Što je veći uzorak to će pouzdanost Wilsonove formule biti bliže stvarnoj vrijednosti. Na primjer ukoliko određena kategorija ima iznimno veliki broj posjeta, ali isto tako i visoku stopu napuštanja njezina ocjena će biti manja od kategorije s puno manjom posjećenošću, ali nižom stopom napuštanja. Upravo iz tog razloga korištenje statističke pouzdanosti je najbolje rješenje za problem, jer na ovaj način će kategorije koje korisnici doista traže dobiti prednosti nad onima za koje zapravo nisu zainteresirani.

Činjenica da Wilsonova formula ocjenu dodjeljuje neovisno o broju posjeta u isto vrijeme može biti i problem jer specijalizirane potkategorije s vrlo malim brojem posjeta

imaju manju stopu napuštanja od onih s velikim brojem posjeta. Rješenje je množenje ocjene pouzdanosti s brojem posjeta. Tako se izbjegava pojava da na najbolje ocijenjene kategorije zapravo otpada vrlo mali broj od ukupnih posjeta.

Prednost [www.hr](#) nad drugim sličnim tražilicama i servisima je detaljna kategorizacija svih stranica. Kada bi se posjetiteljima pokazalo da je [www.hr](#) nadmoćan za traženje internet stranica sa specifičnijim temama, to bi pozitivno utjecalo na percepciju o stranici. Samim tim bi se pozitivno utjecalo i na ukupan broj posjeta stranici. Model zato dodatno daje prednost specifičnim potkategorijama.

Kako se svaka potkategorija može vidjeti u URLu stranice model koristi tu informaciju da dodatno poveća ocjenu specifičnim potkategorijama. Za korekciju ocjene koristi se parametar za težinsku ocjenu važnosti dubine. Potkategorije su navedene slijedno od općenitije do specifičnije, a međusobno su odvojene u URLu znakom „/“. Model za svako pojavljivanje znaka „/“ ocjenu množi s Eulerovim brojem ($e \approx 2.718$). Eulerov broj je određen subjektivnom procjenom i on je trenutna ocjena važnosti dubine. Ocjena važnosti dubine je korigirana više puta i upravo je ona ono što čini model vrlo fleksibilnim. Ukoliko se želi dati veliku prednost iznimno specifičnim kategorijama, ocjenu je potrebno povećati, a ako se želi smanjiti prednost specifičnim kategorijama ocjenu je potrebno smanjiti. Kada ocjena važnosti dubine iznosi Eulerov broj rezultati dobiveni za različita vremenska razdoblja su vrlo logični. Naime, model tada daje visoku ocjenu kategorijama koje su cijelo vrijeme iznimno popularne, ali također za različita razdoblja u najbolje ocijenjenim kategorijama se pojavljuju različite specifične kategorije koje se mogu logično objasniti društveno političkim situacijama u tom razdoblju.

Razvijeni model uz manja testiranja pokazuje dobre rezultate. Za potpuno testiranje potrebno je učiniti ono za što je model i razvijan, odnosno potrebno je kategorije koje model predloži uvrstiti na početnu stranicu [www.hr](#) i pratiti daljnje ponašanje posjetitelja stranice. Rad na modelu nikada nije gotov i potrebno je pažljivo promatrati promjene u ponašanju posjetitelja jednom kada se prijedlozi uvrste na početnu stranicu i po potrebi prilagođavati model. Najveću fleksibilnost nudi ocjena važnosti dubine i njenim ugađanjem se može ponašanje modela usmjeriti u željenom smjeru.

Veliki izazov razvoja modela je što on pokušava objediniti više područja. Model u isto vrijeme pokušava pratiti ponašanje posjetitelja što se može svrstati u sferu psihologije i baviti

se matematičkom statističkom analizom što je sfera prirodnih znanosti. Zbog prirode problema koji se pokušava riješiti drugačiji pristup nije moguć.

Evaluacija i rezultati

Prvi vremenski period u analizi je mjesec svibanj. Svibanj 2015. godine pokazuje da su *news portali* i internet stranice za dopisivanje najtraženije kategorije. Među najzanimljivijim kategorijama se nalaze dvije vezane uz turizam i potkategorija u kojoj se nalaze srednje škole na splitskom području. Svibanj je mjesec kada mnogi ljudi planiraju ljetni godišnji odmor pa povećani interes za turističke teme se vrlo lako može objasniti time. Također, kraj školske godine je početkom lipnja, a upisi u srednje škole nedugo zatim. Interes korisnika za srednjim školama tako se lako objašnjava. Ako usporedimo rezultate iz 2015. godine s onima iz 2014. godine, možemo uz dvije konstantno popularne kategorije (news portali i internet stranice za druženje i razgovor) pronaći čak 3 vezane uz turizam. Još jednu godinu ranije ponovno u vrhu pretraga se nalaze 3 kategorije vezane uz turizam i kategorija s srednjim školama Splita koja se nije pojavila u 2014. Godine 2012., slično kao i 2013. tri kategorije vezane uz turizam i jedna uz srednje škole. Može se uvidjeti da u svibnju svake godine interes za turističke teme i upis u srednje škole se poveća.

Analiza siječnja 2015. godine pokazuje vrlo drukčije interes posjetitelja. Kategorije vezane uz turizam ili srednje škole ne pojavljuju se u rezultatima. Umjesto njih, na samom vrhu rezultata su kategorije vezane uz dopisivanje i druženje. Sociolozi često ukazuju kako je siječanj vrlo depresivan mjesec pa iznimno interes za ovu ionako popularnu kategoriju nije iznenadenje. Kako model prednost daje specifičnim kategorijama, tako se u rezultatima za siječanj pojavljuje vrlo specifična kategorija u kojoj se nalaze akademski slikari. Razlog za povećani interes za takvom jednom kategorijom možda se može objasniti događajem „Noć muzeja“ koji se svake godine održava u siječnju. Ako se vremenski period ograniči na točan datum održavanja tog događaja kategorije vezane uz kulturu su vrlo tražene. Model opet pokazuje pravilnost i za siječnje ranijih godina.

Kada za ulazne datoteke postavimo period ljeta između 2012. i 2014. godine kao preporučene kategorije se uvijek dobiju dvije konstantno popularne. Preostale kategorije su tematski iste, ali nisu iste. Vrlo su popularne kategorije vezane uz turizam, ali konkretno koje potkategorije varira između godina. Takav rezultat može ukazivati da se interes posjetitelja

svake godine mijenja, ali rezultat možemo shvatiti i kao upozorenje da razvijeni model lošije radi za duže vremenske periode.

```
/wwwhr/news/daily/index.hr.html  
/wwwhr/entert/pets/dogs/kennels/index.hr.html  
/wwwhr/entert/penpal/index.hr.html  
/wwwhr/news/index.hr.html  
/wwwhr/tour/travel/auto/rentacar/index.hr.html  
/wwwhr/entert/penpal/sms/index.hr.html  
/wwwhr/business/construction/tvrtke/index.hr.html  
/wwwhr/tour/travel/auto/bus/index.hr.html  
/wwwhr/business/publisher/houses/index.hr.html  
/wwwhr/arts/painting/artists/academic/index.hr.html  
  
/wwwhr/news/daily/index.hr.html  
/wwwhr/entert/penpal/index.hr.html  
/wwwhr/business/construction/tvrtke/index.hr.html  
/wwwhr/news/daily/  
/wwwhr/entert/pets/dogs/kennels/index.hr.html  
/wwwhr/tour/accomm/continental/zagreb/index.hr.html  
/wwwhr/arts/painting/artists/academic/index.hr.html  
/wwwhr/tour/travel/auto/bus/index.hr.html  
/wwwhr/computers/companies/hardware/retail/index.hr.html  
/wwwhr/entert/penpal/sms/index.hr.html  
  
/wwwhr/news/daily/index.hr.html  
/wwwhr/entert/penpal/index.hr.html  
/wwwhr/arts/painting/artists/academic/index.hr.html  
/wwwhr/tour/accomm/continental/zagreb/index.hr.html  
/wwwhr/tour/accomm/islands/dugiotok/index.hr.html  
/wwwhr/education/high/zagreb/index.hr.html  
/wwwhr/entert/pets/dogs/kennels/index.hr.html  
/wwwhr/business/construction/tvrtke/index.hr.html  
/wwwhr/education/high/split/index.hr.html  
/wwwhr/tour/accomm/islands/murter/index.hr.html
```

Slika 1 Primjer izlaza iz modela za ljetne mjeseca 2012., 2013. i 2014. Godine

Takoder testirano je da li veliki svjetski dogadaji utječu na popularnost kategorija. Primjerice, da li se za vrijeme velikih sportskih događanja popularnost kategorije sport poveća. Model nije ni za jedan takav događaj pokazao promjenu u rezultatu. Razlog može biti što kategorija *news portal* je konstantno popularna i posjetioci ukoliko ih zanima nešto vezano uz takav događaj će prije potražiti indirektno na *news portalima* nego pretražujući specijaliziranu kategoriju unutar kataloga.

Zaključak

Uspješno vođenje internet stranice zahtjeva pažljivo praćenje reakcije posjetitelja na sadržaj koji im su nudi. Ručno praćenje je gotovo nemoguće i zato je Google kao pomoć ponudio svoj servis Google Analytics. Analizom podataka Google Analytics servisa može se doći do vrlo korisnih zaključaka o navikama posjetitelja.

Za potrebe analize razvijen je matematički model. Ponašanje modela je prilagođeno specifičnostima kataloga. Detaljna kategorizacija svih web-stranica Hrvatske koju katalog nudi pokušala se istaknuti uvođenjem ocjene važnosti dubine, a statistički izračuni na temelju stope napuštanja daju prednost sadržaju koji posjetitelji uistinu traže.

Provedena evaluacija modela pokazala je da postoje dobri temelji unutar modela, ali da postoji i prostor za napredak. Pravilnosti koje rezultati pokazuju za kratka vremenska razdoblja trebaju se nastojati postići i za duža razdoblja.

Popis literature

- [1] Kende, Michael: "Global Internet Report", Internet society, Ženeva, 2014
- [2] Tech, Blog: "Google Analytics dominates the top 500 websites", s Interneta, <http://royal.pingdom.com/2008/05/28/google-analytics-dominate-the-top-500-websites/>, 9.6.2015.
- [3] Web Technology Surveys: "Usage of traffic analysis tools for websites", s Interneta, http://w3techs.com/technologies/overview/traffic_analysis/all, 9.6.2015.
- [4] Meerscharet,Mark: "Mathematical Modeling, Fourth Edition", Elsevier Inc., Waltham, MA, 2013
- [5] Pandora: "About The Music Genome Project", s Interneta, <http://www.pandora.com/about/mgp>, 9.6.2015
- [6] Barbarosa,Maria: "Basics of Mathematical Modeling", TU Munchen, 2010
- [7] Salihefendic, Amir: "How Reddit ranking algorithm work", s Interneta, <http://amix.dk/blog/post/19588>, 20.5.2015.