AUTOMATIZIRANO OZNAČAVANJE NEDOSTUPNIH I IZMIJENJENIH SJEDIŠTA KATALOGA WWW.HR

Krešimir Pripužić, Marin Vuković, Gordan Gledec, FER, Zagreb

Sažetak

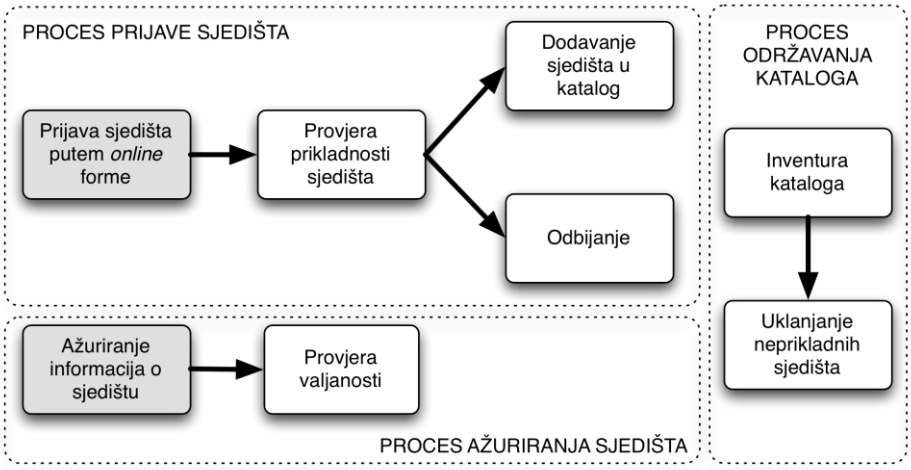
Rad se bavi problemima koji se javljaju tijekom održavanju kataloga www.hr zbog promjena sjedišta weba nastalih nakon njihova uključivanja u katalog. Pri tome je cilj ukloniti iz kataloga sva ona sjedišta koja više nisu dostupna ili su promijenila svoj sadržaj. U radu se predlaže potpuno automatiziran sustav za označavanje ovakvih sjedišta. Obzirom na velik broj sjedišta u katalogu, predloženi sustav značajno olakšava i ubrzava proces inventure kataloga.

1. Katalog www.hr

Katalog www.hr rezultat je projekta "WWW.HR - početna stranica Hrvatske" koji se pod pokroviteljstvom Hrvatske akademske i istraživačke mreže CARNet, radi na Zavodu za telekomunikacije Fakulteta elektrotehnike i računarstva (FER) Sveučilišta u Zagrebu. Projekt je započeo na Zavodu za telekomunikacije u veljači 1994. Godine, a 1996. godine WWW.HR postaje projekt CARNeta [1]. Sjedišta u katalogu razvrstana su prema kategorijama, od kojih se svaka dijeli u više potkategorija. Najniže potkategorije u hijerarhiji tako sadrže isključivo sjedišta profilirana određenoj djelatnosti, usluzi ili slično. Uz redovite poslove održavanja sjedišta, svake se godine uvode sadržajne novosti i tehnološka poboljšanja.

2. Životni ciklus sjedišta u katalogu

Životni ciklus sjedišta u katalogu prikazan je na slici 1 a sastoji se od tri procesa. Prvi proces je prijava sjedišta u katalog koja se provodi na zahtjev privatne ili fizičke osobe. Takav zahtjev osoba pokreće ispunjavanjem *online* forme za prijavu sjedišta dostupne u okviru kataloga www.hr. Tijekom prijave provjerava se dostupnost prijavljenog sjedišta te se zatim popunjavaju informacije o sjedištu kao što su njegov naslov, metapodaci, ključne riječi i slično. Osoba koja prijavljuje sjedište podatke o njemu dobija na uvid te ih može korigirati ili proširiti prema potrebi. Nakon zaprimanja prijave administrator kataloga provjerava odgovaraju li unesene informacije o sjedištu stvarnom stanju, je li sadržaj sjedišta prikladan za objavu u katalogu, je li sjedište zaista hrvatsko te odgovara li predložena katagorija sjedišta stvarnoj kategoriji. Ukoliko su navedeni uvjeti ispunjeni sjedište se uvrštava u katalog.



Slika 1 - Životni ciklus sjedišta u katalogu www.hr

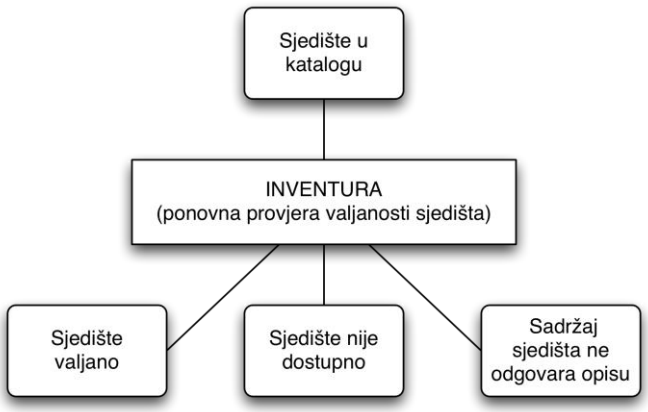
Kako se sjedišta s vremenom mijenjaju, korisnicima je omogućeno ažuriranje informacija o sjedištu. Ono je obuhvaćeno prikazanim procesom ažuriranja sjedišta, a također se pokreće zahtjevom korisnika.

Proces održavanja kataloga svodi se na inventuru svih sjedišta u katalogu. Cilj ovog procesa je čišćenje kataloga od sjedišta koja su tijekom prijave doista bila prikladna za objavu u katalogu, no s vremenom su ukinuta ili im je sadržaj izmijenjen.

Obzirom da je u 2011. godini katalog sadržavao više od 25 000 sjedišta, proces njegove ručne inventure bi bio izuzetno dugotrajan ako se ona ne bi provodila automatizirano. U ovom radu je predstavljen sustav za automatsko označavanje onih sjedišta u katalogu koja su u međuvremenu postala neprikladna za objavu. Pojam neprikladnosti sjedišta podrazumijeva više različitih aspekata koji su detaljno obrađeni u nastavku.

3. Označavanje sjedišta u katalogu

Prema slici 2, rezultat inventure kataloga jesu oznake svih sjedišta u katalogu koje indiciraju da je sjedište je ili nije valjano. Uz najpovoljniji ishod – valjano sjedište – sjedište može biti nevaljano zbog toga što više uopće nije dostupno ili zbog toga što informacije koje su o njemu pohranjene u katalogu ne odgovaraju stvarnom stanju.



Slika 2 - Moguće oznake sjedišta u katalogu

3.1. Nedostupno sjedište

Nedostupno sjedište je najjednostavnije automatizirano identificirati pokušajem dohvaćanja njegova sadržaja. Ukoliko je sjedište dostupno, poslužitelj na kojem se sjedište nalazi dati će pozitivan odgovor na upućeni HTTP zahtjev (200 OK). Međutim, ako sjedište više ne postoji, tada će odgovor biti negativan, odnosno ukazat će na određenu grešku prilikom povezivanja sa sjedištem. Greške mogu biti u rasponu 3xx, čime se indicira određena promjena ili preseljenje sjedišta, 4xx, što ukazuje na grešku u klijentu ili 5xx, što ukazuje na pogrešku poslužitelja na kojem se nalazi ispitivano sjedište weba [2].

Nedostupno sjedište je jednostavno identificirati i ovako označeno sjedište u katalogu se može automatski izbaciti, jer ne postoji mogućnost pogreške u označavanju, kao što je to slučaj sa oznakama opisanim u nastavku.

3.2. Sadržaj sjedišta ne odgovara opisu

Automatizirano utvrđivanje sadržaja sjedišta podrazumijeva usporedbu pohranjenih informacija u katalogu sa stvarnim informacijama na sjedištu te je stoga znatno kompleksnije od prvog slučaja. Ipak, najčešći praktični slučaj ovakve oznake jest kada se umjesto ukinute stranice na automatizirani upit javlja davatelj smještaja i održavanja web stranica s informacijom o nepostojećem sjedištu i, najčešće, vlastitim promotivnim porukama. U tom je slučaju jednostavnom analizom sadržaja sjedišta moguće utvrditi o čemu se radi.

Znatno su rjeđi slučajevi kada, primjerice, tvrtka čije se sjedište automatizirano provjerava promijeni djelatnost. Tada se u pravilu mijenjaju opis sjedišta, ključne riječi i kategorija. Takva promjena se uočava izdvajanjem ključnih riječi iz sadržaja i oznaka sjedišta te usporedbom izdvojenih riječi s opisom pohranjenim u okviru kataloga www.hr. Koncept izdvajanja ključnih riječi i usporedbe sličan je konceptu korištenom kod detektiranja lažnih poruka opisanim u radu [3].

Međutim, takva usporedba nam može dati samo indiciju da je promijenjen sadržaj sjedišta. U tom smislu, proces izdvajanja ključnih riječi i usporedbe nikada ne može biti apsolutno točan, pa se ovako označeno sjedište nikako ne smije automatski izbaciti iz kataloga. Prema tome, o ovako označenim sjedištima potrebno je obavijestiti administratora kataloga koji mora provjeriti je li uistinu došlo do promjena.

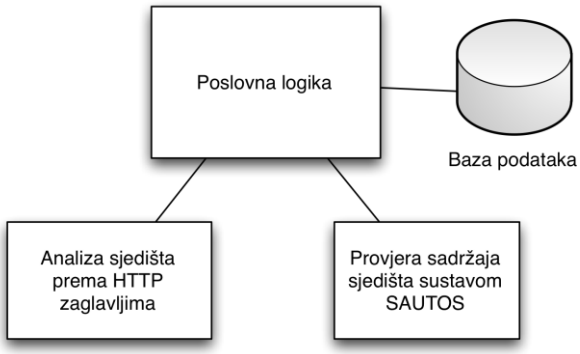
4. Arhitektura rješenja za automatizirano označavanje sjedišta kataloga

Arhitektura rješenja za automatizirano označavanje sjedišta kataloga www.hr prikazana je na slici 3. Rješenje se izvodi neovisno o samom katalogu i pokreće se s zasebnog računala. Unos u sustav je lista adresa (URL) sjedišta koja treba označiti. Naime, zbog mogućeg opterećenja računala s kojeg se provjeravaju sjedišta potrebno je ograničiti broj sjedišta za provjeru u jednom prolazu. U praksi se ovakva provjera može provoditi odjednom za kategoriju ili podkategoriju, ovisno o broju sjedišta koje sadržavaju.

Poslovna logika upravlja slijedom provjera a sjedišta se, u trenutnoj fazi, provjeravaju slijedno iako je ovakav proces jednostavno raspodijeliti na više računala.

Prvo se provjeravaju sjedišta prema HTTP zaglavljima na način objašnjen u poglavlju 3.1. Uočena nedostupna sjedišta se automatski suspendiraju iz kataloga, dok se premještena i slična sjedišta označavaju za daljnju analizu administratora. Ako je sjedište označeno u ovom smislu proces označavanja se prekida.

Drugi korak je provjera sadržaja sjedišta na ključne riječi, prema poglavlju 3.2. Provjera sadržaja se provodi predloženim sustavom SAUTOS koji je opisan u nastavku. U ovom se dijelu prvenstveno orijentira na moguće poruke davatelja smještaja ili slične stranice. Ukoliko se pokaže da ispitivano sjedište sadrži više kritičnih ključnih riječi, proces označavanja se prekida. Međutim, ako nisu uočene kritične ključne riječi, proces se nastavlja usporedbom pronađenih ključnih riječi s ključnim riječima pohranjenim u katalogu. Ako postoji indicija da sjedište nije ispravno zbog bilo kojeg od navedenih razloga, sjedište se prosljeđuje administratoru na konačnu analizu. Uz sjedište se u izvještaju za administratora navodi oznaka i razlog takve oznake.



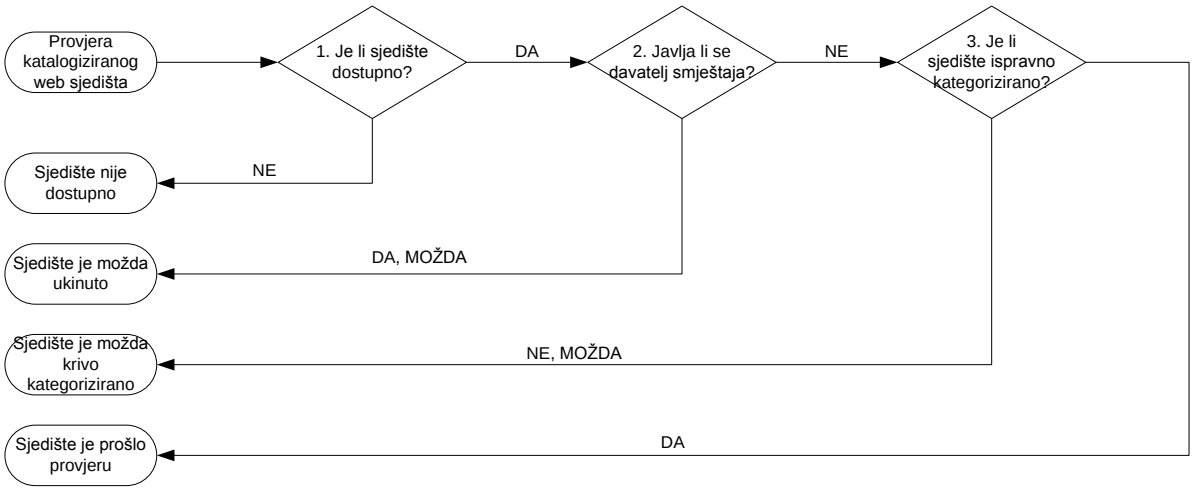
Slika 3 - Arhitektura rješenja za označavanje sjedišta kataloga www.hr

5. Postupak provjere sadržaja sjedišta

Sustav za automatizirano utvrđivanje točnosti opisa web sjedišta kataloga www.hr (SAUTOS) periodički provjerava sva sjedišta pohranjena u katalogu. Kao što je prikazano na Slici 4, provjera pojedinog sjedišta weba se sastoji od tri koraka: provjere dostupnosti, provjere smještaja i provjere ispravne kategoriziranosti.

U prvom koraku se provjerava je li sjedište uopće dostupno korisnicima, prema konceptima opisanim u poglavlju 3.1. Drugi korak uključuje provjeru javljanja davatelja smještaja i održavanja web stranica s informacijama o nepostojećem sjedištu i vlastitim promotivnim porukama. Rezultat ovog procesa jest vjerojatnost da se umjesto web sjedišta javio davatelj njegova smještaja. Ukoliko SAUTOS zaključi da se javilo web sjedište, a ne njegov davatelj smještaja, daljnja provjera se nastavlja. Inače se administrator kataloga obavještava o tome da bi web sjedište moglo biti ugašeno s procijenjenom vjerojatnošću ovog događaja. Ručnom provjerom sumnjivog web sjedišta administrator utvrđuje je li sjedište doista ugašeno. Ovom provjerom se analiziraju metapodaci o sjedištu i ključne riječi njegove naslovnice uz pomoć umjetne neuronske mreže s nadziranim učenjem koja je objašnjena u nastavku.

U trećem koraku se provjerava ispravnost kategorizacije web sjedišta. Ukoliko SAUTOS zaključi da je sjedište ispravno kategorizirano (i dostupno), administrator kataloga neće biti obaviješten o ovom sjedištu. U suprotnom se administrator sustava obavještava o tome da postoji mogućnost krive kategoriziranosti sjedišta s procijenjenom vjerojatnošću ovog događaja. Nakon zaprimanja jedne takve poruke, administrator ručnom provjerom sumnjivog web sjedišta utvrđuje je li sjedište ispravno kategorizirano ili ne. Ova provjera uključuje izdvajanje ključnih riječi i metapodataka o sjedištu iz njegove naslovnice te njihovu usporedbu s opisom sjedišta u katalogu. Sustav SAUTOS vrši ovu provjeru uz pomoć umjetne neuronske mreže s nadziranim učenjem koja je objašnjena u nastavku.



Slika 4: Dijagram toka rada sustava SAUTOS

5.1. Provjera javljanja davatelja web smještaja uz pomoć umjetne neuronske mreže

Pretprocesiranje informacijskog sadržaja na ulazu u umjetnu neuronsku mrežu predstavlja najvažniji korak u radu sustava koji se temelji na neuronskim mrežama [2] zato što neuronske mreže nisu u stanju ispravno raditi s krivo pretprocesiranim podacima. SAUTOS pretprocesira sadržaje web sjedišta uz pomoć metoda za obradu tekstualnih podataka koje se koriste u području pretraživanja informacija [6]. Iz sadržaja naslovnice sjedišta izdvajaju se metapodaci i tekst te se svode na normalni oblik, kao što je objašnjeno u [3]. Zatim se od tog normiranog sadržaja web sjedišta stvara vektor težina normiranih ključnih riječi koji se predaje na ulaz neuronske mreže. Pri tome se izbacuju one ključne riječi koje nemaju značajnu ulogu za otkrivanje javljanja davatelja web smještaja. Ovime se smanjuju dimenzije ulaznih vektora i smanjuje prostor rješenja čime se značajno ubrzavaju procesi izvođenja i učenja neuronske mreže.

Inicijalni postupak treniranja (učenja) neuronske mreže se provodi s poznatim stranicama davatelja web smještaja i reprezentativnim podskupom katalogiziranih web sjedišta. Prilikom rada sustava, skup poznatih stranica davatelja web sadržaja dodatno se proširuje stranicama naknadno otkrivenih davatelja web smještaja. Ovime se omogućava visoka razina efikasnosti otkrivanja stranica davatelja web smještaja.

5.2. Provjera ispravne kategoriziranosti web sjedišta uz pomoć umjetne neuronske mreže

Slično kao u slučaju provjere javljanja davatelja web smještaja, pri provjeri ispravne kategoriziranosti dostupnog web sjedišta, SAUTOS pretprocesira njegov sadržaj na način da se iz sadržaja naslovnice sjedišta izdvajaju i svode na normalni oblik metapodaci i tekst. Zatim se od tog normiranog sadržaja web sjedišta stvara vektor težina normiranih ključnih riječi koji se predaje na ulaz neuronske mreže. Osim toga, na ulaz neuronske mreže se istovremeno također predaje i kategorija te vektor težina normiranog kratkog opisa i ključnih riječi koje čine opis web sjedišta u katalogu. Pri tome se opet izbacuju one ključne riječi koje nemaju značajnu ulogu za otkrivanje ispravne kategoriziranosti web sjedišta da bi se ubrzali procesi izvođenja i učenja neuronske mreže.

Inicijalni postupak treniranja ove neuronske mreže provodi se s reprezentativnim podskupom ispravno i neispravno katalogiziranih web sjedišta. Prilikom rada sustava, skup neispravno katalogiziranih web sjedišta se proširuje naknadno otkrivenim neispravno kategoriziranim sjedištima. Stalnim ažuriranjem poznatog skupa neispravno kategoriziranih web sjedišta i periodičkim treniranjem neuronske mreže, postiže se vrlo visoka razina efikasnosti otkrivanja ispravne kategoriziranosti web sjedišta.

6. Zaključak

U radu su opisani koncepti na kojima se temelji automatizirano označavanje nevaljalih sjedišta u katalogu www.hr. Obzirom na velik broj sjedišta uvrštenih u katalog, nužno je osmisliti barem dijelomično automatizirano rješenje kako bi se smanjio broj sumnjivih sjedišta koje administrator mora osobno pregledati. Ključno je da se sjedišta u najmanjoj mogućoj mjeri automatizirano izbacuju iz kataloga, jer se mora osigurati da se niti u kojem slučaju ispravno sjedište greškom ne izbaci iz kataloga. Predloženi sustav za automatizirano označavanje u mogućnosti je značajno smanjiti početnu brojku od više od 25000 sjedišta koje administrator mora osobno pregledati. Time se omogućuje brža inventura kataloga koja se iz tog razloga može provoditi češće, čime se osigurava da se ukupan broj nevaljalih sjedišta u katalogu u svakom trenutku svede na minimum.

6. Literatura

[1] Sjedište www.hr, pristupano u srpnju 2011.

[2] Fielding, R. and Gettys, J. and Mogul, J. and Frystyk, H. and Masinter, L. and Leach, P. and Berners-Lee, T.: Hypertext Transfer Protocol -- HTTP/1.1, 1999, RFC Editor, United States.

[3] M. Vuković, K. Pripužić, H. Belani: An Intelligent Automatic Hoax Detection System // Lecture Notes in Computer Science/Knowledge-Based and Intelligent Information and Engineering Systems 5711/2009. 1 (2009) ; 318-325

[4] Google Labs. Google Safe Browsing API. http://code.google. com/apis/safebrowsing/, pristupano u srpnju 2011.

[5] Bishop, C. M. Neural Networks for Pattern Recognition. New York: Oxford University Press. 2004.

[6] Manning, C. D., Raghavan, P i Schütze, H. Introduction to Information Retrieval, Cambridge: Cambridge University Press. 2008.