Pronalaženje adresa i geokodiranje sjedišta weba uvrštenih u katalog WWW.HR

Ivan Šemanjski, Marin Vuković
{ivan.semanjski, marin.vukovic}@fer.hr
Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva, Unska 3, Zagreb

Sažetak *– Rad se bavi izdvajanjem adresa sa sjedišta weba korištenjem metoda zasnovanih na pravilima. U radu su predstavljeni određeni problemi koji se javljaju pri izdvajanju adresa sa stranica hrvatskih web sjedišta te predložena rješenja. Izdvojene se adrese geokodiraju korištenjem javno dostupnih servisa. Predložene metode su evaluirane na različitim skupovima stvarnih podataka iz kataloga WWW.HR.*

1. Uvod

Katalog WWW.HR sadrži više od 25000 sjedišta weba iz Hrvatske ili vezanih uz Hrvatsku. Razvoj mobilnih tehnologija i lokacijski temeljenih usluga nameće zahtjeve za geokodiranjem različitih sadržaja dostupnih na Internetu kako bi se takvi sadržaji mogli uključiti u lokacijske usluge. Kod lokacijskih usluga u kontekstu kataloga WWW.HR prvenstveno se fokusira na geokodiranje sjedišta obuhvaćenih katalogom. Geokodiranje sjedišta kataloga može omogućiti pretraživanje kataloga ovisno o lokaciji korisnika, što je posebno pogodno za mobilne aplikacije koje se izvode na pametnim telefonima svjesnim vlastite lokacije. U tom smislu, cilj geokodiranja sjedišta jest mobilna aplikacija WWW.HR koja će omogućavati pregled i lokacijsko pretraživanje kataloga kako bi se krajnjim korisnicima omogućila prilagođena informacija u svojstvu žutih stranica.

Postupak geokodiranja sjedišta kataloga WWW.HR podrazumijeva dva temeljna koraka: pronalaženje lokalnih adresa u tijelu analiziranog sjedišta te geokodiranje pronađene adrese pomoću javno dostupnih servisa.

Pronalaženje lokalnih adresa provodi se analizom teksta (engl. *text mining*). Međutim, najveći dio informacija na webu nalazi se u obliku nestrukturiranog teksta (najčešće pisan prirodnim jezikom). Neki izvještaji [2] spominju da je čak više od 90% podataka u "digitalnom svemiru" nestrukturirano pa je problem pronalaženja kvalitetnih i potrebnih

informacija još više otežan. U početku razvoja područja izdvajanja informacija nestrukturirani tekst bio je promatran kao "vreća riječi" (engl. *bag of words*) [3].

Jednom pronađena lokalna adresa geokodira se javno dostupnim servisima Yahoo, Bing i Google. Postupak geokodiranja ujedno se može promatrati i kao validacija izdvojene adrese, u smislu da se krivo pronađena adresa neće moći geokodirati korištenjem bilo kojeg od navedenih servisa.

2. Slična istraživanja

Istraživanja na području izdvajanja adresa sa sjedišta weba promatrana su kroz sustave zasnovane na pravilima. Takvi se sustavi najčešće koriste registrima geografskih pojmova. Autori [4] koriste se segmentacijom DOM stabla sjedišta weba na smislene blokove zadržavajući informacije o strukturi bloka. Počinju od najmanjih blokova koje najčešće čine tekstualni čvorovi pa ih onda proširuju susjednim čvorovima ovisno o njihovoj vizualnoj sličnosti i susjedstvu. Autori zatim, pomalo naivno, definiraju da je indikator za postojanje adrese segment koji sadrži tekst poput "adresa za slanje" ili "adresa ureda". Rezultati ovog istraživanja, rađenog na samo 44 web stranice, pokazuju preciznost od 0,89. Autori u radu [5] za razliku od većine prethodnih istraživanja zadržavaju strukturu pretraživane web stranice. U radu koriste nekoliko ručno izrađenih pravila te malu bazom geografskih naziva i ključnih riječi. Pravila su izrađena na temelju 10 najčešćih zapisa adresa na australskim web stranicama. Rezultati istraživanja daju odziv od 0,73 te F-mjeru od 0,83. Autori [6] u svom radu također koriste sustav zasnovan na pravilima i to njih 6 od početnih 18 pravila jer su se kroz praksu pokazala kao najpouzdanija pravila za pronalazak adresa. Eksperiment je proveden na bazi od 4 milijuna brazilskih web stranica, od čega je na 603.798 njih pronađeno jedna ili više adresa što čini 14,77% stranica ukupne kolekcije.

McCurley [7], koji je postavio temelje ovakvom pretraživanju, istražuje različite izvore za geografski kontekst sjedišta weba. Izvori uključuju informacije o poslužitelju, kontekst iz sadržaja, adrese i poštanske brojeve, telefonske brojeve, nazive geografskih elemenata (jezera, škole, crkve, parkovi, itd.), kontekst izveden iz poveznica i drugih izvora (npr. žute stranice). U prototipu sustava koristi se baza poštanskih brojeva SAD-a gdje je svaki poštanski broj povezan se geokodiranom lokacijom koja označava centar regije koja sadrži taj poštanski broj. Sličan, ali prošireni princip primjenjuje se i u ovom radu, što je objašnjeno u nastavku.

3. Pronalaženje adresa

Sustav razvijen u ovom radu temelji se na sustavu zasnovanom na pravilima pa se proces prepoznavanja adresa temelji na definiranim pravilima pretraživanja. U tu se svrhu koriste regularni izrazi. Budući da adrese u hrvatskom adresnom sustavu nemaju točno definiranu strukturu niti prepoznatljivi početak zapisa moraju se definirati određena pravila za njihovo uspješno izdvajanje.

Inicijalnim pregledom nekoliko stotina sjedišta weba nasumično odabranih iz kataloga utvrđeno je da zapisi adresa najčešće sadrži sljedeće entitete:

– *Ulica* – ulica, naselje, trg, aleju, šetalište, itd.

– *Broj* – arapski ili rimski brojevi i njihove kombinacije te skraćenica *bb*

– *PB* – peteroznamenkasti poštanski broj

– *Naselje* – naziv naselja

Pravila zapravo predstavljaju određene znakove ograničenja kao graničnike pri pretraživanju *Ulice* i *Broja*. Naime, nakon što su pronađeni poštanski broj *PB* i naziv naselja *Naselje* potrebno je ispred ili iza para tih dvaju entiteta probati pronaći entitete *Ulica* i *Broj*.

Početni pristup postupku izdvajanja adrese bio je postupak čitanja redak po redak. Kada bi se u određenom retku pronašao par (*PB*, *Naselje*) izdvojili bi se svi znakovi od početka pročitanog reda do elementa *PB* (obrazac odgovara strukturi *Ulica*+*Broj*+*PB*+*Naselje*), odnosno, prema strukturi *PB*+*Naselje*+*Ulica*+*Broj*, izdvojili bi se svi znakovi od elementa *Naselje* do kraja retka. Pregledavanjem izvornog kôda nasumično odabranih HTML stranica zaključeno je da je takav pristup neuspješan, budući da u recima u kojima se navodi adresa može biti i drugog teksta jer blok s adresom nije posebno izdvojen.

U sljedećem su pretraživanju primijenjena zapažanja o smještaju adrese u različitim strukturama HTML dokumenta poput blok ili linijskih elemenata. Primijećeno je da je to poprilično čest slučaj, ali konačni rezultati nisu bili dovoljno precizni jer su bile izdvojene i neke riječi koje nisu pripadale nazivu ulice. Međutim, u tom se pretraživanju mogao primijetiti određeni obrazac koji se pojavljivao u većini rezultata, a to je da su *Ulica* i *Broj* odvojeni od ostatka teksta određenim znakovima ograničenja: {, : ; - < > = |}.

Izlomljene su zagrade (< i >) dio oznaka HTML-a, dok su ostali znakovi umetnuti kao dio teksta.

Prikupljena saznanja o najčešćim formatima zapisa adresa, smještaju adresa u strukturi HTML-a, pokazatelju postojanja adrese te prethodno definirani znakovi ograničenja bili su temeljna podloga pri definiranju obrazaca, odnosno regularnih izraza prikazanih u tablici 1.

Tablica 1: Regularni izrazi za izdvajanje adrese

Element	Regularni izraz
Ulica	[*slovo razmak* . - ( ) 0-9]+
Broj	[0-9 / *rimski_broj bez_broja*]+
PB-prefiks	(HR *razmak** -? *razmak**)
PB	([0-9]{5}) | ([0-9]{2} [0-9]{3})
Naselje	[*slovo razmak* . -]+
Graničnik	[, : ; | – - < >]+

Adresa za format: ***Ulica + Broj + PB + Naselje***	{*Graničnik*} {*Ulica*} {*Broj*} {*Graničnik*} {*PB-prefiks*}?{*PB*}
Adresa za format: ***PB + Naselje + Ulica + Broj***	{*PB*} {*Naselje*} {*Graničnik*} {*Ulica*} {*Broj*} {*Graničnik*}
Adresa za format: ***Ulica + Broj + Naselje + PB***	{*Graničnik*} {*Ulica*} {*Broj*} {*Graničnik*} {*Naselje*} {*Graničnik*}?{*PB-prefiks*}?{*PB*}

4. Geokodiranje adresa

Geokodiranje izdvojenih adresa provodi se pomoću javno dostupnih servisa za geokodiranje. Takvi su servisi većinom prilično napredni i osim što podržavaju geokodiranje hrvatskih adresa, omogućuju i prikaz geokodiranih lokacija na digitalnim kartama putem web preglednika. Bitno je naglasiti da su u radu razmatrane samo besplatne usluge. Geokodiranje hrvatskih adresa pomoću besplatnih javnih usluga nije sveobuhvatno i dovoljno precizno jer su podržani samo gradovi i veća naselja. Zbog toga rezultat geokodiranja ponekad može biti nepovoljan, iako su adrese pravilno izdvojene. Usluge koje su se koristile za geokodiranje su: *Yahoo! PlaceFinder*, *Bing Maps Location API* i *Google Maps Geocoding API*.

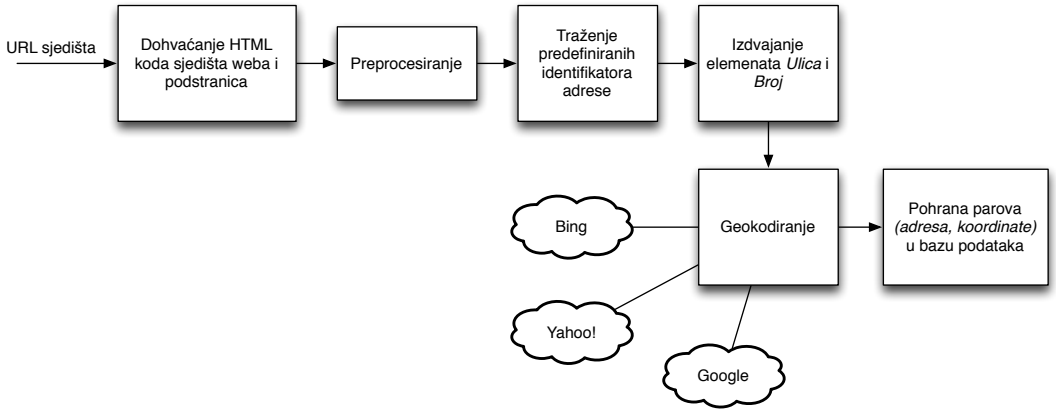
Geokodiranja adresa se provodilo slanjem posebno definiranog upita nekoj od usluga, a rezultat geokodiranja je zatim primljen u obliku odgovora formiranog u strukturi JSON. Iz tog tekstualnog formata su se dalje izdvajale koordinate i podaci o preciznosti i uspješnosti geokodiranja. Formati upita svojstveni su svakoj usluzi, a jedinstveno je da kao parametre primaju entitete *Ulica*, *Broj*, *PB* i *Naselje*.

5. Sustav za pronalaženje i geokodiranje adresa

Predloženi sustav prikazan je na slici 1. Ulazni parametar sustava jest URL web sjedišta kojeg treba procesirati, URL_i. Izlaz sustava predstavljen je parom $(address, coords)_i$ koji predstavljaju adresu pronađenu na ulaznom URL parametru te geokodirani zapis iste. Prema tome, sustav se može opisati kao preslikavanje:

$$URL_i \rightarrow (address, coords)[URL_i]$$

Prema slici 1, sustav se izvodi na sljedeći način. Sustav dohvaća sjedište identificirano URLom, i ako je isto dostupno, iz sadržaja početne stranice se izdvajaju URL-ovi podstranica. Izdvojeni URL-ovi podstranica stavljaju se u listu čekanja za obradu. Nakon toga slijedi proces obrade dohvaćenog sadržaja. Prvo se procesira početna stranica sjedišta weba, a ako proces izdvajanja ne uspije, onda se dohvaćaju i procesiraju izdvojene podstranice.



Slika 1 - Sustav za pronalaženje i geokodiranje adresa

Prvi korak u obradi dohvaćenog sadržaja je preprocesiranje u kojem se vrši zadatak čišćenja HTML-a. Naime, u tome se koraku uklanjaju određeni elementi koje nisu bitni za proces izdvajanja adresa: element *head* i njegov sadržaj, komentari te elementi *script* kojima se dohvaćaju *JavaScript* datoteke. Nakon pročišćavanja sadržaja od navedenih elemenata, drugi je korak uklanjanje atributa svih preostalih elemenata. Atributi nisu bitni za pretraživanje adresa, a njihovo uklanjanje smanjuje veličinu konačnog stringa za pretraživanje. Zadnji korak faze preprocesiranja je zamjena svih otvarajućih i zatvarajućih oznaka posebnom oznakom `<$$>` te uklanjanje svih bjelina između navedenih oznaka i stapanje susjednih oznaka u jednu oznaku kako bi se dobio dugačak niz znakova u jednoj liniji. Specijalna oznaka `<$$>` koristi se kako bi se izradili što jednostavniji obrasci za izdvajanje adresa.

Nakon faze preprocesiranja slijedi faza pretraživanja. Prvo se traži indikator postojanja adrese, odnosno par (*PB*, *Naselje*) pomoću obrazaca koji odgovaraju tim elementima. Ako indikator nije pronađen, sustav provjerava ima li još stranica u listi za čekanje i u slučaju pozitivnog ishoda procesiranje kreće od početka s pretraživanjem nove stranice. Ako je indikator pronađen, prelazi se na sljedeću fazu, fazu izdvajanja entiteta *Ulica* i *Broj*. U toj se fazi koriste pripadajući regularni izrazi. Dohvaćeni niz znakova prolazi kroz fazu filtriranja. U slučaju neispravne adrese, odnosno neprolazne ocjene filtra, vrši se daljnje traženje i izdvajanje elemenata *Ulica* i *Broj* koji stoje uz pronađeni poštanski broj i naselje. Ako je dohvaćeni niz znakova uspješno prošao filtar, prosljeđuje se metodi za geokodiranje.

Metoda za geokodiranje vrši pretvorbu ulaznih parametara *Ulica*, *Broj*, *PB*, *Naselje* u jedan URL za upit na javnu uslugu za geokodiranje. Sustav geokodiranja sastoji se od tri razine, od kojih svaka predstavlja jednu uslugu ovisno o njezinom prioritetu. Ako usluga većeg prioriteta ne uspije geokodirati dohvaćenu adresu, prelazi se na sljedeću razinu. Najveći prioritet ima usluga *Yahoo! PlaceFinder*, zatim *Bing Maps*, a najmanji prioritet ima usluga *Google Maps*. Odgovor na postavljeni upit je u nekom strukturiranom formatu iz kojeg se dohvaćaju odgovarajući parametri (geografska širina i dužina) te provjerava uspješnost i preciznost odgovora. Dohvaćene se geografske koordinate zatim povezuju s identifikatorom u katalogu hrvatskih poslužitelja WWW.HR i pohranjuju u odgovarajuću bazu podataka.

6. Evaluacija sustava

Izvor podataka nad kojima se provodi eksperimentalna evaluacija sustava jest katalog hrvatskih sjedišta weba WWW.HR. Ta se baza podataka, u trenutku pisanja, sastoji od nešto više od 25.000 zapisa podataka o sjedištima weba različitih kategorija poput gospodarstva, zabave, turizma, kulture, društva, itd. Evaluacija izrađenog programskog rješenje rađena je na različitom broju sjedišta weba iz kategorije gospodarstva, budući da su u tu kategoriju svrstane većinom pravne osobe kod kojih postoji velika vjerojatnost da imaju pohranjenu adresu na svom sjedištu.

Evaluacija sustava provedena je na skupu od 250 sjedišta weba. Skup ulaznih podataka sadržavao je samo sjedišta koja su na početnoj ili nekoj od podstranica sadržavala zapis adrese. Ovakav skup je odabran kako bi se najpreciznije uočili nedostatci sustava u smislu lažno negativnih rezultata (*false negative*), odnosno kako bi se utvrdilo raspoznaje li sustav precizno adrese koje sigurno postoje. Obzirom na strog pristup raspoznavanju adresa

sa traženjem poštanskog broja mjesta, smatramo da je mogućnost lažnih pozitivnih rezultata (*false positive*) znatno manja od mogućnosti lažno negativnih rezultata na koje je evaluacija izložena u ovom radu usmjerena. Rezultati analize skupa koji sadrži sjedišta sa i bez navedene adrese izloženi su u radu [8].

Dobiveni rezultati pokazali su da je broj sjedišta na kojima su izdvojene adrese u skupu od 250 web sjedišta bio 238, što čini 95% ukupnog broja sjedišta weba. U skupu od 238 pronađenih zapisa validacija ispravnosti svake adrese provedena je putem geokodiranja, a neuspješno geokodirane adrese su dodatno ručno provjerene pomoću internetskih tražilica. Nakon tako provedenog istraživanja zaključeno je da je od 213 pronađenih zapisa samo 1 zapis krivo identificiran (*false positive*) što znači da je preciznost ovog sustava visokih 0,995. Za ovako visoku preciznost odgovorno je ograničenje sustava koje nalaže postojanje poštanskog broja u adresi.

Budući da su sva sjedišta sadržavala adrese jer su unaprijed odabrana sjedišta koja ih sadrže izračunat je i odziv sustava kao omjer broja pronađenih ispravnih adresa te ukupnog broja ispravnih adresa u skupu ulaznih podataka. Budući da je ispravno izdvojeno 237 adresa na skupu od ukupno 250 ispravnih adresa, odziv sustava je 0,948. Odziv bi mogao biti i veći jer, od 12 sjedišta na kojima adrese nisu uspješno izdvojene, pet sjedišta imalo je poveznice na kontaktne informacije generirane pomoću *JavaScripta*. Skriptni jezik *JavaScript* izvršava se u pregledniku, a njegove se akcije temelje na interakciji korisnika. Budući da se pretraživanje i izdvajanje provodilo strojno, bez interakcije korisnika, podstranice koje su sadržavale zapise adresa nisu bile dohvaćene i obrađene.

Kako se validacija izdvajanja adresa oslanja na javno dostupne servise za gaokodiranje izdvojenih adresa, u nastavku su prikazani podaci evaluacije tih servisa s hrvatskim adresama. Evaluacija servisa za geokodiranje provedena je na skupu od 237 sjedišta weba od kojih je svako sadržavalo po jednu adresu, izdvojenu u procesu evaluacije sustava. Sve su adrese bile s područja Republike Hrvatske, a postojanje istih je ručno provjereno. Uspješnost svake usluge je ispitivana zasebno koristeći navedeni skup sjedišta. Ispitivanje je provedeno na način da se pojedinoj usluzi svaka adresa slala zasebno pomoću posebno definiranog upita, a zatim je provjeren primljeni odgovor. Dobiveni rezultati prikazani su sljedećom tablicom (Tablica Tablica 2).

Tablica 2: Evaluacija usluga za geokodiranje

Naziv usluge	Broj geokodiranih adresa

Yahoo! PlaceFinder	197 (83%)
Bing Maps Location API	213 (90%)
Google Maps Geocoding API	172 (73%)

Rezultati potvrđuju sumnju da se validacija ispravnosti izdvojenih adresa trenutno ne može osloniti isključivo na rezultate geokodiranja pomoću javnih servisa, budući da niti jedna usluga nema uspješnost 100%. Međutim, uspješnost od 90% geokodiranih adresa usluge *Bing Maps* prilično je impresivna. Rezultat *Yahooa* je zadovoljavajući, dok je *Google* sa svojih 73% prilično razočarao ako se uzme u obzir da je ta usluga najpopularnija.

7. Zaključak

U radu je predstavljen sustav za izdvajanje i geokodiranje lokalnih adresa sa sjedišta weba. Evaluacija je provedena na skupu od 250 sjedišta kataloga WWW.HR koji je ujedno i motiv za razvoj ovog sustava. Evaluacija je, uz ograničenje postojanja poštanskog broja, pokazala vrlo visoku uspješnost izdvajanja adresa. Na temelju rezultata može se zaključiti da je predloženi sustav u mogućnosti potpuno zadovoljiti zahtjeve te populirati bazu podataka adresa sjedišta uvrštenih u katalog WWW.HR.

Nedostatak predloženog rješenja jest u činjenici da se oslanja na postojanje poštanskog broja mjesta. Iako većina sjedišta doista sadrži potpun oblik adrese, buduća istraživanja će se usmjeriti na mogućnost izdvajanja adrese bez oslanjanja na poštanski broj kako bi se proširila funkcionalnost sutava. Prema sličnim istraživanjima, u tom se slučaju očekuje znatno veći broj lažno pozitivnih rezultata, dok su u izloženom radu veći problem predstavljali lažno negativni rezultati.

Cilj predloženog sustava jest povezati jednu fizičku adresu tvrtke i sjedište u katalogu, pa je opisani sustav usmjeren na pronalaženje jedne adrese sjedišta. Međutim, sjedište može sadržavati i više adresa (npr. poslovnice), pa će se algoritam prilagoditi kako bi mogao pronaći i više adresa na jednom sjedištu. Međutim, u tom je slučaju potrebno izvesti dodatne zahvate na logici kataloga WWW.HR kako bi se jedna adresa web sjedišta mogla logički povezati sa više fizičkih adresa.

Uz navedeno, razmotrit će se za koje kateogrije uopće ima smisla izdvajati lokalne adrese (primjerice, za osobne stranice, blogove i slično nije potrebno) te će se predloženi sustav evaluirati na čitavom skupu odabranih kategorija iz kataloga.

8. Popis literature

1. Rowe, B., Wood, D., Link, A., Simoni, D. *Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program*, Durham, SAD: RTI Internatinal, 2010
2. Gantz, J., Reinsel, D. *Extracting Value from Chaos*, IDC, 2011.
3. Mooney, R. J., Bunescu, R. Mining Knowledge from Text Using Information Extraction, *ACM SIGKDD Explorations Newsletter - Natural language processing and text mining*, 7,1 (2005), 3-10
4. Can, L., Qian, Z., Meng, X., Lin, W. Postal Address Detection from Web Documents. WIRI 2005: 40-45.
5. Asadi, S., Yang, G., Zhou, X., et al. Pattern-based Extraction of Addresses from Web Page Content, *In Proceedings of 10th Asia-Pacific Web Conference - APWeb 2008*, (2008), 407-418
6. Borges, K., Laender, A., Medeiros, C., Davis Jr., C. Discovering Geographic Locations in Web Pages Using Urban Addresses. *In Proceedings of the 4th ACM workshop on Geographical information retrieval*. ACM: New York, NY, SAD (2007), 31-36
7. McCurley, K. S. Geospatial mapping and navigation of the web. *In Proceedings of the Tenth International Conference on World Wide Web*, ACM Press (2001), 221-229
8. Ivan Šemanjski: "Izdvajanje i geokodiranje adresa sa sjedišta weba", diplomski rad, br.303, Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, 2012.