UDC: 004.832.2:025.4 INFO M: str. 33-40

NOV METOD EKSTRAKCIJE INFORMACIJA BAZIRAN NA TRANSDUKTORIMA
A NEW INFORMATION EXTRACTION METHOD BASED ON TRANSDUCERS

Vesna Pajić, Miloš Pajić, Staša Vujičić Stanković

REZIME: U radu je dat osvrt na oblast ekstrakcije informacije, čije su metode i tehnike nezaobilazne u pretrazi i upravljanju informacijama. Ova oblast u sebi sadrži tehnike drugih oblasti matematike i računarstva, kao što su obrada prirodnih jezika, teorija formalnih jezika, verovatnoća i statistika. Uzimajući u obzir sve specifičnosti zahteva za informacijom i tekstualnih resursa iz kojih se izdvajanje vrši, razvijen je i u radu prikazan nov metod za ekstrakciju informacija nazvan Dvofazni metod baziran na transduktorima. Predstavljena je arhitektura sistema koji implementira ovaj metod kao i primer konkretne primene. Poseban značaj ovaj metod ima u situacijama kada ne postoje već pripremljeni tekstualni korpusi, neophodni za primenu postojećih metoda, posebno onih baziranih na verovatnoći i statistici.

KLJUČNE REČI: ekstrakcija informacija, obrada prirodnih jezika, strukturiranje podataka

ABSTRACT: An overview of the information extraction is given, whose methods and techniques are indispensable in the information search and information management. Information extraction uses and combines techniques and methods from mathematics and computer science, such as natural language processing, formal language theory, probability and statistics. Taking into account all specifics of requests for information and textual resources from which it is extracted, we developed and present a new method for the information extraction called a two-stage method based on the transducers. An architecture of a system that implements this method is presented, along with an example of application. This method has the special significance in situations in which there is a lack of already annotated text corpora, that are necessary for the application of existing methods, especially those based on probability and statistics.

KEY WORDS: information extraction, natural language processing, data structuring

1. UVOD U EKSTRAKCIJU INFORMACIJA

U današnje vreme svedoci smo stvaranja i postojanja velikih kolekcija različitih formata dokumenata i zapisa. Pronalaženje pojedinačnih informacija u takvim kolekcijama je često teško i zahtevno. Sa druge strane, veliki je broj situacija ili problema koji mogu biti uspešno rešeni jedino analizom i upotrebom tih informacija, kao što su analiza komentara korisnika nekog proizvoda ili usluge, analiziranje naučnih studija sa željom da se uspostavi korelacija između nekog medikamenta i uspešnosti terapije, objedinjavanje medicinskih elektronskih kartona pacijenata iz različitih medicinskih institucija u jedinstvenu bazu podataka i drugi.

Svi ovi primeri imaju po nekoliko zajedničkih elemenata: postoji zahtev za određenom informacijom; odgovor na zahtev se uglavnom nalazi unutar nestrukturiranih izvora podataka, kao što su tekst ili slike; nije moguće da čovek obradi te izvore podataka, jer ih ima previše; računari nisu u mogućnosti da direktno postave upit nad podacima, jer podaci nisu strukturirani. Pokušaj rešavanja ovih i sličnih problema doveo je do nastanka čitave jedne oblasti, koja se i danas razvija velikom brzinom i predstavlja podoblast veštačke inteligencije, a u sebi objedinjuje mnoge tehnike drugih naučnih oblasti, kao što su računarska lingvistika, procesiranje prirodnih jezika, verovatnoća i statistika i dr.

Ekstrakcija informacija (engl. *Information extraction*, u daljem tekstu IE) je oblast računarske lingvistike koja obuhvata skup tehnika za pronalaženje informacija u tekstualnim dokumentima, koji su obično nestrukturirani ili polustrukturirani, i predstavljanje tih informacija u strukturiranom obliku. IE se koristi za analiziranje teksta i pronalaženje određenih delova teksta koji sadrže informaciju od interesa, kao i za njenu ekstrakciju iz teksta. S obzirom da je krajnji cilj procesa ekstrakcije informacija omogućavanje dalje obrade podataka, nakon ili tokom izdvajanja potrebno je izvršiti strukturiranje dobijenih informacija. Strukturiranje informacija je proces klasifikovanja informacija u semantičke klase, tj. dodeljivanje značenja dobijenim informacijama.

Strukturiranje informacije se najčešće vrši ili obeležavanjem teksta, tj. umetanjem oznaka (najčešće XML oznaka) koje opisuju značenje delova teksta koji je prepoznat kao informacija, ili kreiranjem i popunjavanjem relacionih baza podataka, ili na neki drugi način. Sam format izlaznih podataka nije od suštinske važnosti za proces ekstrakcije informacije, jer je moguće naknadno konvertovati podatke iz jednog formata u drugi. Ono što jeste važno i što određuje proces ekstrakcije jesu semantičke klase podataka, tj. njihova specifikacija kao i pravila odlučivanja koja informacija pripada kojoj semantičkoj klasi.

2. PRAVILA EKSTRAKCIJE I KONAČNI TRANSDUKTORI

Način na koji se delovi teksta identifikuju kao određene informacije definiše se pravilima ekstrakcije. Postoji veliki broj različitih formata za predstavljanje pravila ekstrakcije. To su *Common Pattern Specification Language* (CSPL) (Appelt i sar., 1993) i formati izvedeni iz njega kao što su JAPE (Cunningam i sar., 2002), obrasci i liste kao u sistemu Rapier (Califf i Mooney, 1999), regularni izrazi kao u sistemu WHISK (Soderland, 1999), SQL izrazi (Jayram i sar., 2006; Reis i sar., 2008) i Datalog izrazi (Shen i sar., 2007). Sve te

reprezentacije imaju dosta zajedničkih osobina i mogu biti jednostavno uopštene. U stvari, suštinski sve navedene reprezentacije mogu biti predstavljene modelima konačnih stanja, na prvom mestu konačnim transduktorima i rekurzivnim mrežama prelaza (Jurafsky i Martin, 2008; Vitas, 2006).

Konačni transduktori (engl. *finite state transducer* ili skraćeno *FST*) su konačne apstraktne mašine koje definišu relacije između dva skupa niski karaktera u smislu da su u mogućnosti da transformišu jednu nisku u drugu. Formalno, FST se definiše kao uređena šestorka $\tau = (\Sigma_1, \Sigma_2, Q, i, F, \Delta)$, pri čemu su:

- Σ_1 i Σ_2 ulazna i izlazna azbuka,
- Q konačni skup stanja,
- $i \in Q$ početno stanje,
- $F \subset Q$ skup završnih stanja,
- $\Delta \subset Q \times \Sigma_1 \times \Sigma_2 \times Q$ relacija tranzicije, čiji se elementi nazivaju *lukovima.*

Konačni transduktori se već dugi niz godina koriste u skoro svim oblastima računarstva, a posebno ulogu imaju i u okviru računarske lingvistike. Glavna osobina transduktora, koja ih izdvaja od ostalih konačnih mašina, jeste da produkuju neki izlaz. Upravo ta osobina i određuje način na koji se konačni transduktori koriste u obradi prirodnih jezika. Takođe, oni mogu biri predstavljeni grafovima, što ih čini veoma udobnim za korišćenje od strane čoveka. U računarskoj lingvistici koriste se za morfološko parsiranje, opisivanje pravopisnih pravila, opisivanja pravila promene reči i sl. Detaljan prikaz teoretske i praktične upotrebe konačnih transduktora u obradi prirodnih jezika može se naći u (Casacuberta i sar. 2005; Friburger i Maurel, 2004; Hobbs i sar. 1997; Jurafsky i Martin, 2000; Kornai, 1999; Pajić, 2010; Pajić, 2011; Pajić i sar. 2011a; Pajić i sar. 2011b; Roche, 1999; Roche i Schabes, 1997).

Konačni transduktori mogu biti veoma kompleksni i teški za kreiranje i modifikovanje, što u praksi dovodi do značajnih problema. Zato se često umesto jednog velikog grafa koristi kolekcija manjih podgrafova. Ovakav pristup ima svoju teoretsku pozadinu u teoriji rekurzivnih mreža prelaza (engl. *Recursive Transition Networks* ili skraćeno *RTN*) (Sastre i Forcada, 2007; Sastre, 2009; Vitas, 2006). U okviru ovog i sličnih istraživanja nije od interesa da li se radi o jednom grafu ili kolekciji grafova i podgrafova, već je važno da je dolaskom u završno stanje izvršena određena transformacija ulazne niske karaktera (prevođenje, umetanje teksta, zamena delova niske i sl.). Zbog toga ćemo u nastavku rada koristiti termin *transduktor* za mašine kojima se vrši transdukcija kad god to bude bilo moguće, misleći pri tom ili na konačni transduktor ili na rekurzivnu mrežu prelaza.

Postoji nekoliko softverskih alata i sistema namenjenih lingvističkim istraživanjima i obradi prirodnih jezika koji su upravo bazirani na transduktorima (Paumier, 2011; Silberztein, 1993). Za obradu teksta i primenu pravila ekstrakcije (primenu transduktora) u okviru ovog istraživanja korišćen je softverski sistem UNITEX (Paumier, 2011).

3. DVOFAZNI METOD BAZIRAN NA TRANSDUKTORIMA

U većini metoda za ekstrakciju informacija lociranje i identifikacija slogova unutar teksta i izdvajanje podataka iz njih se obavlja istovremeno, u okviru jednog istog logičkog procesa. Ovakav pristup često otežava proces ekstrakcije informacija.

Kako bismo prevazišli, ili barem pojednostavli navedeni problemi, razvili smo poseban, nov metod nazvan *Dvofazni metod za ekstrakciju informacija baziran na konačnim transduktorima*. Ovaj metod je namenjen za ekstrakciju određenih entiteta i njihovih osobina (atributa) iz teksta. Pogodan je za korišćenje u situacijama kada se obrađuju tekstovi kao što su veb strane, enciklopedije, neki udžbenici, tj. za sve one resurse kod kojih je moguće na neki način, na osnovu njihove retoričke strukture (poglavlja i odeljci u enciklopediji) ili HTML i XML oznaka (za veb strane), odrediti koji delovi teksta se odnose na koji entitet.

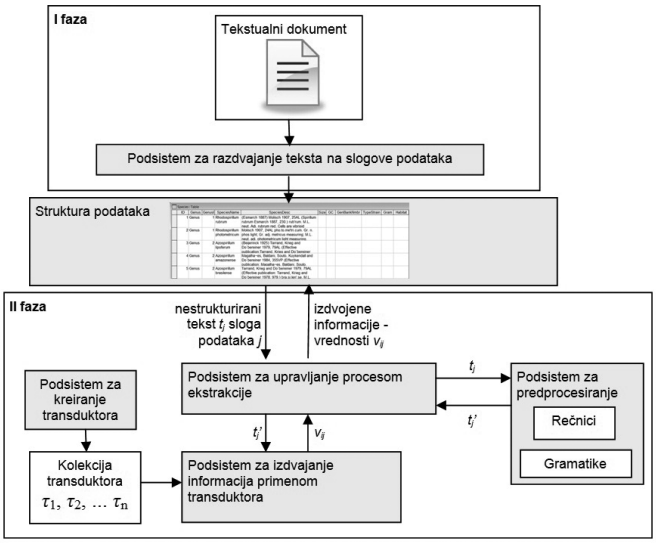
Metod razlikuje dve, logički i vremenski potpuno odvojene faze. Kao ulaz u algoritam prve faze koristi se ceo tekst koji se obrađuje. Algoritam prve faze locira entitete i na osnovu same strukture tekstualnog resursa deli celokupan tekst na manje celine, od kojih svaka odgovara po jednom entitetu (objektu) koji se obrađuje, tj. svaka će predstavljati jedan slog u bazi podataka. Kao izlaz algoritma prve faze kreira se baza podataka sa slogovima koji su identifikovani na neki način (svaki odgovara po jednom entitetu, pa se često upravo prepoznati entitet koristi kao identifikator sloga) i koji sadrže manje delove teksta sa informacijama o objektu. Ove informacije su i dalje u nestrukturiranom obliku, ali su pridružene odgovarajućem entitetu na koji se odnose.

U drugoj fazi cilj je iz delova teksta izvući vrednosti pojedinih atributa i dodeliti im značenje. Za izdvajanje atributa u drugoj fazi koriste se konačni transduktori i rekurzivne mreže prelaza, gde se za svaki atribut koji je potrebno izdvojiti, tj. za svaku osobinu entiteta koja je od interesa u nekom procesu ekstrakcije, kreira po jedan transduktor. Na taj način se u stvari pomoću transduktora zadaju pravila ekstrakcije. Značenje izvučene informacije definisano je samim transduktorom koji je tu informaciju prepoznao, pa se izlaz svakog transduktora ubacuje u odgovarajuće polje baze podataka. Druga faza ne zavisi od strukture samog dokumenta, pa isti transduktori mogu biti primenjeni i na druge tekstualne dokumente koji sadrže odgovarajući tip informacija.

Podela procesa ekstrakcije informacije u dve faze omogućava ne samo bolju efikasnost i veću preciznost, već i korišćenje različitih softverskih alata i sistema za njenu implementaciju. Tako je moguće za implementaciju prve faze metoda korisiti jedan alat (na primer, neki od programskih jezika koji omogućavaju rad sa relacionim bazama), a za implementaciju druge faze neki sasvim drugi alat (na primer, neki od softverskih alata koji su efikasni i jednostavni u radu sa konačnim modelima).

Iako je implementaciju dvofaznog metoda moguće izvršiti na razne načine i različitim postojećim ili novoformiranim

softverskim rešenjima, ipak postoje određeni elementi (podsistemi) koje svaki sistem za implementaciju dvofaznog metoda mora da obezbedi. Arhitektura opšteg sistema za ekstrakciju informacija pomoću dvofaznog metoda baziranog na transduktorima prikazan je na slici 1.



Slika 1.– *Arhitektura sistema za ekstrakciju informacija koji implementira dvofazni metod baziran na transduktorima*

Podsistem za razdvajanje teksta na slogove podataka preuzima tekstualni resurs koji se obrađuje i iz njega izdvaja manje delove teksta t_j, $j=1..m$ koji se odnose na pojedine objekte (entitete). Pri tome, svakom izdvojenom delu teksta se dodeljuje određeno značenje, tj. jasno se utvrđuje na koji entitet se taj deo teksta odnosi. Izdvojeni delovi teksta se čuvaju u nekom formatu pogodnom za dalju obradu, tj. **strukturi za čuvanje podataka**. Dvofazni metod predviđa da to bude relaciona baza podataka, ali je moguće koristiti i neki drugi format.

Pomoću **podsistema za kreiranje transduktora** se kreiraju transduktori (τ_1, τ_2, ... τ_n) namenjeni za izdvajanje konkretnih informacija, tj. vrednosti v_{ij}, $i=1..n$, određenih atributa entiteta iz teksta, za svaki izdvojeni deo teksta t_j, $j=1..m$. Ukoliko postoje već razvijene biblioteke transduktora za ekstrakciju, ovaj podsistem je moguće izostaviti. Da bi u okviru konačnih modela koji se koriste za ekstrakciju informacija bili korišćeni lingvistički resursi ili leksički i morfološki filteri i maske, neophodno je da tekst na koji se primenjuju bude prethodno lingvistički obrađen. **Podsistem za upravljanje procesom ekstrakcije informacija** preuzima jedan po jedan izdvojeni deo teksta t_j iz strukture u kojoj je čuvan i šalje ga prvo **podsistemu za predprocesiranje** na lingvističku obradu. Podsistem za predprocesiranje može da koristi ili da sadrži i različite lingvističke resurse, kao što su elektronski rečnici i gramatike. Lingvistički obrađeni delovi teksta t_j' se dalje prosleđuju podsistemu za izdvajanje informacija primenom transduktora.

Podsistem za izdvajanje informacija primenom transduktora primenjuje niz transduktora (τ_1, τ_2, ... τ_n) na deo teksta t_j' koji mu je prosleđen, pri čemu svaki od transduktora produkuje određeni izlaz. Upravo niske koje predstavljaju izlaze transduktora čine izdvojene informacije (v_{ij}) i bivaju smeštene u polja strukture podataka koja odgovaraju slogu j koji se obrađuje.

U nastavku je opisan razvoj jednog sistema za ekstrakciju informacija i primena dvofaznog metoda unutar njega. Kao podsistem za predprocesiranje, za kreiranje transduktora i za izdvajanje informacija korišćen je UNITEX (Paumier, 2011). Podsistem za razdvajanje teksta na slogove podataka i podsistem za upravljanje procesom ekstrakcije su pisani posebno, koristeći programski jezik Java. Za strukturiranje i čuvanje podataka korišćena je MS Access baza podataka.

4. PRIMENA DVOFAZNOG METODA NA NAUČNU ENCIKLOPEDIJU KAO POLUSTRUKTURIRANI RESURS – KREIRANJE I DOPUNA BAZE PODATAKA O MIKROORGANIZMIMA

U okviru istraživanja sprovedenog u okviru Grupe za bionformatiku[1] Matematičkog fakulteta, Univerziteta u Beogradu, vršena je ekstrakcija podataka iz enciklopedije "*Systematic Bacteriology*" (Garrity, 2005; Garrity i sar., 2005; Krieg i sar., 2010; Vos i sar., 2009) koja sadrži podatke o mikrobima (njihovim fenotipskim i genotipskim karakteristikama) u obliku opisnog teksta, tj. teksta u slobodnoj formi. Pomoću tehnika ekstrakcije podataka kreirana je i popunjena baza podataka sa informacijama o fenotipskim i (u manjoj meri) genotipskim karakteristikama mikroorganizama. Enciklopedija je korišćena samo i isključivo za potrebe naučnog istraživanja. Cilj je bio da se pokaže kako je moguće dobiti strukturirani resurs sa relevantnim podacima, koji je moguće integrisati sa drugim sličnim postojećim resursima i koristiti za dalja istraživanja u oblasti biologije i genetike.

S obzirom da je struktura enciklopedije takva, da je bilo moguće na osnovu retoričke strukture teksta uspostaviti određene veze između entiteta, tj. mikroorganizama u ovom slučaju, i informacija koje se na njih odnose, ovom tekstualnom resursu pristupili smo kao polustrukturiranom, a samu strukturu teksta iskoristili u procesu ekstrakcije. Takođe, podaci dobijeni iz enciklopedije trebalo je da budu što je moguće tačniji, kako bi bilo moguće nad njima sprovoditi dalja istraživanja iz oblasti mikrobiologije i genetike. Zbog toga je odabran pristup baziran na znanju i kreiran je i primenjen dvofazni metod baziran na konačnim transduktorima za ekstrakciju informacija iz teksta.

4.1. *Analiza strukture korišćenog resursa i kreiranje modela baze podataka*

Prva faza dvofaznog metoda jako zavisi od strukture samog resursa iz koga se izvlače podaci, pa je prvi korak u ekstrakciji informacija analiziranje tekstualnog resursa. Podaci koje smo želeli da izdvojimo nalazili su se u četiri toma enciklopedije "*Systematic Bacteriology*", u obliku opisnog, nestrukturiranog teksta na engleskom jeziku. Transformisali smo ovaj tekst iz *.pdf* fromata u *.txt* format, kako bismo obezbedili brži i jed-

[1] http://bioinfo.matf.bg.ac.rs/

nostavniji pristup. Za konverziju teksta korišćen je softverski alat *Abbyy PDF Transformer*[2], i tom prilikom je došlo do gubitka određenih informacija zasnovanih na strukturi dokumenta (neki paragrafi su pogrešno protumačeni, podaci smešteni u tabelama su izgubili strukturu tabele i sl.). Međutim, ovakvo narušavanje strukture nije značajno uticalo na proces ekstrakcije informacije.

Iako su se razlikovali po sadržaju, sva četiri toma enciklopedije su imala veoma sličnu strukturu, koja je iskorišćena u procesu ekstrakcije informacija. Poglavlja enciklopedije su odgovarala taksonomskim kategorijama carstva *Bacteria*. Na primer, prvo poglavlje drugog toma (Garrity, 2005) sadrži informacije o taksonomskoj klasi *Alphaproteobacteria*, sledeće poglavlje je o prvom redu ove klase (*Rhodospiralles*), a zatim slede poglavlja o familijama reda *Rhodospiralles*. Svako poglavlje koje se odnosi na neku familiju praćeno je poglavljima o rodovima te familije. Slika 2 prikazuje izvod iz sadržaja drugog toma enciklopedije.

Slika 2. – *Izvod iz sadržaja enciklopedije "Systematic Bacteriology", Tom 2, deo C*

Opisi vrsta, koji sadrže i podatke koje želimo da ekstrahujemo, dati su na kraju svakog poglavlja o rodu, i to na samom kraju poglavlja. Prethodi im linija teksta "*List of species of the genus* ...". Broj vrsta se razlikuje od roda do roda, ali svaki od opisa vrsta počinje rednim brojem, praćenim imenom vrste i opisom u obliku opisnog teksta (slika 3).

9. **Hyphomicrobium zavarzinii** Hirsch 1989b, 495[VP] (Effective publication: Hirsch 1989, 1903.)
za.wrzin' i.i. M.L. gen. n. *zavarzinii* of Zavarzin, named for G.A. Zavarzin, the Russian microbiologist who isolated these bacteria.

Mother cells drop- or pear-shaped, somewhat slender, with hyphae that rarely branch. Mother cells 0.63 × 1.8 µm (range: 0.5–0.9 × 0.7–2.5 µm). Swarmer cells with 1–3 subpolar flagella. In liquid media under most growth conditions, rosettes are formed, since mother cells produce a polar holdfast. Growth in liquids initially as turbidity and later as a pellicle, with precipitation on the bottom. Colonies on solid media are colorless to light brownish or beige, smooth and shiny, with entire edges.

Chemoorganotrophic, aerobic, oligocarbophilic. Good growth with the following carbon sources: methanol, methylamine·HCl, formate, *n*-butyrate, isovalerate, crotonate, β-hydroxybutyrate, ethanol, *n*-propanol, isobutanol, and glycerol. Growth is stimulated significantly by acetate, *n*-valerate, α-oxoglutarate, galacturonate, formaldehyde, D-glucose, D-mannose, D-melibiose, amygdalin, esculin, chitin, Bacto peptone, DL-lysine, DL-aspartate, and dilute human urine. Nitrogen sources utilized are: NH_4^+, NO_2^-, NO_3^-, and (poorly) Bacto peptone. There is slow growth in the absence of added nitrogen sources (oligonitrophily). Poor growth on sheep blood agar with α-hemolysis. The following antibiotics inhibit growth at 30 µg (per disc): kanamycin, neomycin, and tetracycline. Streptomycin at 10 µg is also inhibitory. There is growth in the presence of 3.5% NaCl. Temperature range: 15–37°C. Optimal pH: 6.5–7.5. Visible light inhibits growth slightly.

Grow anaerobically with nitrate and gas formation (with methanol as the carbon source). With methylamine·HCl and thioglycolate, there is little growth. Catalase and cytochrome oxidase are positive; gelatin liquefaction is negative. Poly-β-hydroxybutyrate is a storage product.

Not pathogenic for mice or guinea pigs.

Genome size: 2.73 × 10^9 Da (strain ZV-580; Kölbel-Boelke et al., 1985).

Habitat: peaty and moist soil near Moscow, Russia.

The mol% G + C of the DNA is: 61.8–64.8 (Bd, T_m, HPLC) (Mandel et al., 1972; Gebers et al., 1986; Urakami and Komagata, 1987b; Urakami et al., 1995b).

Type strain: ATCC 27496, IFAM ZV-622.

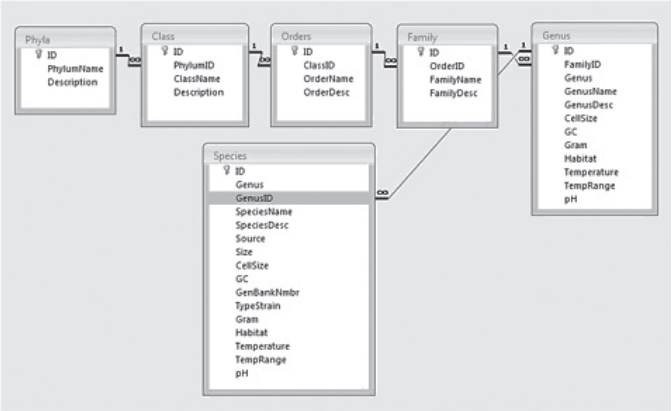
GenBank accession number (16S rRNA): Y14305.

Additional Remarks: Additional strains include IFAM ZV-580, ZV-620, MY-619, MC-625, MC-629, MC-630, and MC-627.

Slika 3. – *Primer opisa vrste. Podvučeni delovi teksta predstavljaju podatke koje želimo da izvučemo iz teksta.*

Pojedinačni atributi o organizmima nalazili su se u slobodnoj tekstualnoj formi, unutar opisa jedne vrste (oni koji su karakteristični za vrstu) ili roda (ukoliko su zajedničkim svim vrstama tog roda). Ti opisi su sadržali različite podatke o organizmima, kao što je njihov oblik, stanište, *Gram* osobina, dominantni soj (*type strain*), veličina genoma, procenat nukleotida guanin i citozin (G+C) u DNK lancu, veličina ćelije, optimalna pH vrednost, *GenBank* (Bilofsky i Christian, 1988) identifikacioni broj, i sl. Upravo su to podaci koje smo želeli da izvučemo iz enciklopedije i smestimo u strukturni oblik, tj. u bazu podataka. Ovi podaci su prikazani podvučeno na slici 3.

Nakon analize strukture, a pre pristupanja prvoj fazi ekstrakcije, kreirana je baza podataka sa tabelama koje odgovaraju taksonomskim kategorijama: *Phyla, Class*, *Order*, *Family*, *Genus*, *GenusIncSed* (za "*Genus Incertae Sedis*", taksonomsku grupu čiji je odnos sa drugim grupama nepoznat ili nedefinisan) i *Species*. Tabele *Phyla*, *Class*, *Order* and *Family* će biti korišćene za čuvanje informacija o taksonomskim kategorijama i njihovim relaicjama. Tabele *Genus, GenusIncSed* i *Species,* sem kolona koje čuvaju relacije među grupama, sadrže i kolone koje će čuvati izvučene podatke o organizmima. Dijagram baze je prikazan na slici 4.



Slika 4. – *Dijagram rezultujuće baze podataka*

4.2.*Prva faza: kreiranje odgovarajuće baze i razbijanje teksta na delove koji odgovaraju pojedinačnim entitetima*

Na osnovu analize strukture enciklopedije, razvijen je algoritam prve faze. On koristi činjenicu da svako poglavlje odgovara jednoj sistematskoj kategoriji (*Class*, *Order*, *Family* i *Genus*). Algoritam čita liniju po liniju teksta sadržaja enciklopedije. Svaka linija ima istu strukturu: prva reč je naziv taksonomske kategorije, praćen rimskim brojem i tačkom, iza koje sledi naziv kategorije (npr. "*Family I. Rhodospirillaceae*"). Algoritam koristi prvu reč u liniji sadržaja kako bi odredio tabelu u koju slog treba da bude ubačen (u navedenom primeru to bi bila tabela "*Family*)". Ime kategorije se koristi kao vrednost polja *FamilyName* sloga koji će biti kreiran (u ovom slučaju kreira se slog za familiju "*Rhodospirillaceae")*. Redosled u kome se različite kategorije pojavljuju u sadržaju iskorišćen je za uspostavljanje veza između kategorija i određivanje pripadnosti, tj. za uspostavljanje veza između tabela. Na primer, struktura tabele *Family* je prikazana u tabeli 1. Polje *OrderID* se koristi za čuvanje veze sa odgovarajućim slogom iz tabele *Order*, tj. sa redom kome neka familija pripada.

[2] http://pdftransformer.abbyy.com/

Tabela 1. – *Struktura tabele Family*

Field Name	Data Type
ID	Integer
OrderID	Integer
FamilyName	Text

Tabela sa opisima vrsta i rodova su popunjavane u sledećem koraku. U okviru I faze popunjavane su samo vrednosti u prva četiri polja (*ID*, *GenusID*, *SpeciesName* i *SpeciesDesc* za tabelu *Species*), dok su ostala polja popunjavana u okviru druge faze. Preciznije, I fazom metoda je izvršeno izdvajanje i strukturiranje većih delova teksta, koji su sadržali informacije o organizmima, a na koje će kasnije biti primenjeni transduktori kako bi se izdvojili konkretni atributi koji opisuju organizme.

Da bi popunio tabelu Species, algoritam prolazi kroz tekst enciklopedije u potrazi za linijom koja počinje sa "*List of species*". Iza ove linije počinju opisi vrsta koje izdvajamo u bazu podataka u okviru prve faze tj. kao rešenje vertikalnog problema. Svaki opis počinje rednim brojem i imenom vrste, na primer

1. Rhodovulum sulfidophilum
(Hansen and Veldkamp 1973) .....

Ova činjenica je iskorišćena kako bi algoritam otkrio slogove za tabelu *Species*, tj. kako bi identifikovao početak i kraj opisa jedne vrste. Na nesreću, tokom konverzije iz .pdf u .txt format neki paragrafi su pogrešno protumačeni, tako da su u tekstualnom fajlu postojale linije koje počinju brojem i tačkom, ali ne predstavljaju početak opisa vrste. Zbog toga je algoritam dizajniran tako da koristi činjenicu da prva reč u imenu vrste predstavlja ime roda kome ona pripada. Samo linije koje zadovoljavaju taj uslov su prepoznate od strane algoritma i tretirane kao početak opisa vrste. Zbog navedene modifikacije, algoritam je imao odličnu efikasnost. Svaki opis vrste koji je postojao u originalnom tekstu je prepoznat i ubačen u bazu podataka. Ovakva vrsta finog podešavanja algoritma je moguća na osnovu analize struktura tesktualnog resursa koji se koristi. Na istraživaču je da odluči do kog nivoa će modifikovati algoritam, kako bi postigao željeni nivo efikasnosti.

Na sličan način je kreirana i tabela *Genus*, pa je nakon završetka prve faze, formirana baza podataka koja je sadržala slogove sa podacima o vrstama i rodovima. Konkretno, za tabelu Species kreirana su i popunjena polja *SpeciesName* (ime vrste), *Genus* (rod kome pripada) i *SpeciesDesc* (tekstualni opis vrste koji sadrži vrednosti pojedinih atributa od interesa). Deo podataka je prikazan na slici 5. Slično, za tabelu *Genus* u popunjena polja *GenusName* (ime roda), *Family* (familija kojoj rod pripada) i *GenusDesc* (tekstualni opis roda koji sadrži vrednosti pojedinih atributa).

Species : Table

ID	Genus	GenusI	SpeciesName	SpeciesDesc	Size	GC	GenBankNmbr	TypeStrain	Gram	Habi
1	Genus	1	Rhodospirillum rubrum	(Esmarch 1887) Molisch 1907, 25AL (Spirillum rubrum Esmarch 1887, 230.) rub'rum. M.L. neut. adj. rubrum red. Cells are vibrioid						
2	Genus	1	Rhodospirillum photometricum	Molisch 1907, 24AL pho.to.me'tri.cum. Gr. n. phos light; Gr. adj. metricus measuring; M.L. neut. adj. photometricum light measuring.						
3	Genus	2	Azospirillum lipoferum	(Beijerinck 1925) Tarrand, Krieg and Do¨bereiner 1979, 79AL (Effective publication:Tarrand, Krieg and Do¨bereiner						
4	Genus	2	Azospirillum amazonense	Magalha~es, Baldani, Souto, Kuykendall and Do¨bereiner 1984, 355VP (Effective publication: Magalha~es, Baldani, Souto,						
5	Genus	2	Azospirillum brasilense	Tarrand, Krieg and Do¨bereiner 1979, 79AL (Effective publication: Tarrand, Krieg and Do¨bereiner 1978, 979.) bra.si.lien'se. M.L.						

Slika 5. – *Izgled tabele Species nakon završene prve faze ekstrakcije*

4.3. TRANSDUKTORI ZA EKSTRAKCIJU INFORMACIJA

Nakon završetka prve faze, a pre početka druge, tekst je izdeljen na manje celine koje su prepoznate kao sadržaoci informacija koje treba da se izvuku iz teksta, pri čemu je za svaku od tih celina identifikovan slog u bazi podataka, tj. mikroorganizam na koji se te informacije odnose. Ti manji delovi teksta se sada nalaze u bazi podataka. U nastavku procesa je potrebno iz njih izvući konkretne podatke.

Pristup ekstrakciji informacija baziran na konačnim transduktorima koristi upravo konačne transduktore za prepoznavanje i izdvajanje pojedinačnih informacija, tj. vrednosti određenih atributa iz teksta. U okviru druge faze navedenog metoda, od strane eksperata kreiraju se konačni transduktori, po jedan za svaki atribut koji je potrebno izvući iz teksta. U okviru pomenutog istraživanja korišćen je softver UNITEX (Paumier, 2011) za kreiranje transduktora, iako je moguće koristiti i druge softvere za tu namenu.

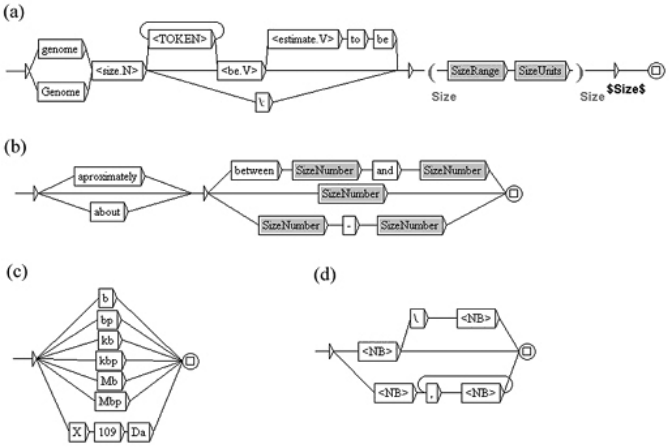
Slika 6 prikazuje nekoliko grafova transduktora korišćenih za ekstrakciju informacije o veličini genoma. Glavni transduktor (Slika 6.(a)) opisuje kontekst u kome se očekuje da se informacija o veličini genoma pojavi unutar teksta, i definiše deo teksta koji će biti ekstrahovan kao podatak. Ovaj kontekst je utvrđen od strane eksperata i baziran je na analizi enciklopedijskog teksta.

Na primer, algoritam koji koristi transduktor prikazan na slici 8 izdvaja veličinu genoma i to tako što prepoznaje izraze kod kojih je reč "*genome*" prva reč. Sledeća reč u izrazu treba da odgovara imenici "*size*" u bilo kom obliku, što je označeno sa <size.N>. Jedino u tom slučaju algoritam nastavlja sa radom, tj. sa prepoznavanjem izraza pomoću transduktora. Informacije o različitim oblicima reči "*size*" se preuzimaju iz elektronskih rečnika. Glavni transduktor koristi i leksičke maske kao što su <be.V> and <estimate.V>, koje prepoznaju bilo koji oblik glagola *to be* ili *to estimate*, kako bi opisao kontekst u kome informacija o veličini genoma može da se pojavi. Specijalan simbol <TOKEN> se odnosi na bilo koji token (bilo koju reč ili karakter koji nije slovo) u tekstu. Samo sekvenca pročitanih reči, koja odgovara putanji definisanoj pomoću transduktora je prepoznata i u tom slučaju se na osnovu tog prepoznatog izraza produkuje i izlaz transduktora.

Izlaz je definisan pomoću zagrada u transduktoru (Slika 6.(a)). Deo teksta koji odgovara delu transduktora između zagrada smešta se u promenljivu, u ovom slučaju nazvanu *Size*. Deo transduktora obeležen karakterom $ definiše šta

će biti izlaz. Transduktor prikazan na slici 6 će kao izlaz vratiti vrednost promenljive *Size*. U nekim drugim slučajevima moguće je da se izlaz definiše kao neki složeniji izraz, a ne samo vrednost promenljive.

Glavni transduktor poziva dva podgrafa, nazvana *SizeRange* i *SizeUnits*. Pozivi podgrafova su obeleženi sivim pravougaonicima, kao na slici 6.(a) i slici 6.(b). Odgovarajući podgrafovi, *SizeRange* i *SizeUnits* su prikazani na slici 6.(b) i slici 6.(c). Graf *SizeRange* opisuje kontekst koji odgovara različitim načinima na koje je moguće izraziti veličinu genoma u enciklopedijskom tekstu, kao što su "*between 2240 and 3787*", "*1256–1276*" ili "*approximately 4061*". Graf *SizeUnits* opisuje sve jedinice korišćene za izražavanje veličine genoma, a koje se pojavljuju u enciklopedijskom tekstu. Graf *SizeNumber* opisuje različite refernce na brojeve koji se pojavljuju u tekstu. Specijalni simbol <NB> prepoznaje bilo koju neprekidnu sekvencu cifara.



Slika 6. Transduktor za ekstrakciju informacija o veličini genoma kreiran pomoću UNITEX softvera (a) Glavni transduktor sadrži pozive podgrafova *SizeRange* i *SizeUnits* i produkuje izlaz $*Size*$; (b) Podgraf *SizeRange* za opisivanje različitih načina za specifikaciju vrednosti veličine genoma; (c) Podgraf *SizeUnits* koji prepoznaje različite jedinice za veličinu genoma; (d) Podgraf *SizeNumber* prepoznaje različite formate brojeva

Sledeće fraze su prepoznate od strane transduktora prikazanog na slici 6. Ekstrahovani podaci, koji se smeštaju u bazu podataka, označeni su podebljanim slovima.

"genome sizes of four G. oxydans strains were estimated to be ***between 2240 and 3787 kb****"*

"genome size of R. prowazekii is ***1,111,523 bp****"*

"genome size of R. africae is ***1.248 kb****"*

"genome size of R. australis is ***1256–1276 kbp****"*

"genome size is ***2.62 X 109 Da****"*

"Genome size: ***2.73 X 109 Da****"*

"genome size is ***1.713 Mbp****"*

"genome size was estimated to be ***approximately 4061 kb****"*

"genome size of all the classical strains examined was ***about 3000 kb****"*

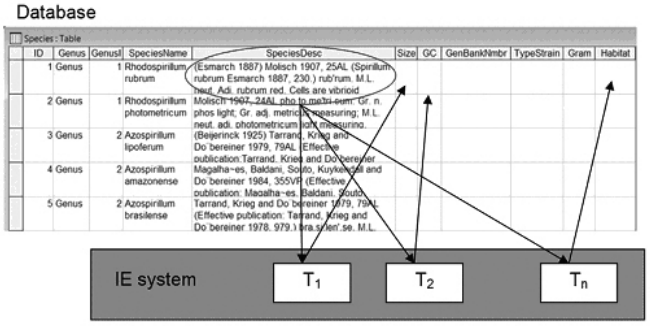
Dakle, transduktor prepoznaje celu frazu, ali izdvaja samo označeni deo teksta. Na sličan način su izdvojeni i drugi atributi jednog organizma.

4.4. *Druga faza: ekstrakcija informacija i njihovo smeštanje u rezultujuću bazu podataka*

U okviru druge faze sistem uzima iz slogova baze podataka delove teksta u nestrukturiranom obliku, u kojima se nalaze informacije o entitetima (vrednosti pojedinih atributa) i analizira ih pomoću grafova transduktora. Za svaki atribut koji je potrebno izvući iz teksta kreira se poseban transduktor. Izlaz koji taj transduktor produkuje nakon primene na tekstualni opis, smešta se u bazu podataka kao vrednost odgovarajućeg atributa za odgovarajući slog. Ovaj postupak je šematski prikazan na slici 7.

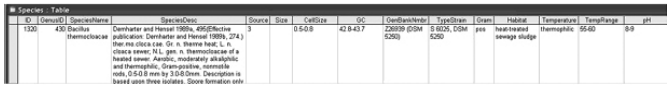
U okviru pripremne obrade tekstualnih opisa izvršena je tokenizacija, normalizacija i primena leksičkih resursa (elektronski rečnik engleskog jezika), kako bi u transduktorima bilo moguće koristiti leksičke i morfološke maske. Kao rezultat, nakon druge faze rezultujuća baza podataka je popunjena podacima, pa je tabela *Species* izgledala kao na slici 8.

Nakon primene modifikovanog dvofaznog metoda na enciklopedijski tekst, svi ekstrahovani podaci su smešteni u relacionu bazu podataka. Pri tom je korišćena Microsfot Access baza za čuvanje i manipulisanje podacima, mada podaci mogu biti veoma lako i jednostavno izvezeni u bilo koji format baze podataka.



Slika 7. – *Ilustracija primena transduktora T_1, T_2, ..., T_n za ekstrakciju informacija*

Za svaki mikroorganizam izvučeni su sledeći podaci iz opisa vrsta: veličina genoma, veličina ćelije, sadržaj G+C, GenBank pristupni broj, *Type strain*, *Gram stain*, stanište, optimalna temperatura, raspon temperature i raspon pH vrednosti. Neki od podataka o krakteristikama vrste su se nalazili u opisima rodova kojima te vrtse pripadaju, pa su isti transduktori primenjeni i na opise u tabelama *Genus* i *GenusIncSed*.



Slika 8. – *Tabela Species nakon završene druge faze procesa ekstrakcije informacije*

Tabela 2. – *Preciznost i odziv transduktora*

Transduktor	Size	CellSize	GC	Gen BankNmbr	Type Strain	Gram	Habitat	Temperature	Temp Range	pH
Preciznost	0.92	0.91	1.00	1.00	0.99	1.00	0.97	1.00	0.84	0.81
Odziv	0.90	0.96	0.96	1.00	0.96	1.00	0.79	0.69	0.66	0.77

4.5. *Preciznost i odziv metoda*

Za potrebe evaluacije uzet je slučajan uzorak od 100 slogova vrsta. Ovi slogovi su pojedinačno analizirani. Upoređeni su njihovi opisi u enciklopediji i vrednosti atributa koje su se nalazile u tim opisima, sa vrednostima atributa koje su izdvojene u bazu podataka, a koji su dobijeni automatski od strane sistema za ekstrakciju informacija. Rezultati ove analize prikazani su u tabeli 2.

Preciznost je bila veoma visoka, što rezultujuću bazu čini veoma pouzdanim resursom za dalja biološka istraživanja. Ovako velika preciznost je posledica činjenice da su transduktori dizajnirani od strane eksperata kako bi izvlačili vrednosti pojedinih atributa, i stoga prepoznaju samo one informacije koje su se nalazile u opisanom kontekstu.

Odziv se razlikovao za različite transduktore, zavisno do kompleksnosti konteksta u kome inforamcija može da se nađe. Na primer, transduktor za Gram osobinu organizama je bio veoma efikasan; ekstrahovao je korektno sva pojavljivanja ovog atributa u tekstu. Neki drugi transduktori nisu bili tako efikasni, kao što je slučaj sa transduktorom za stanište mikroorganizma (*Habitat*). Međutim, daljim podešavanjem, modifikacijom i proširivanjem transduktora, kako bi oni prepoznali i pojavljivanja koja nisu prepoznata u prvom prolazu, proces ekstrakcije informacije bio bi poboljšan i podignut na željeni nivo efikasnosti. Ovo je svakako jedna od važnih prednosti metoda zasnovanih na konačnim transduktorima u odnosu na druge metode ekstrakcije informacije, posebno na one bazirane na probabilističkim modelima.

5. ZAKLJUČAK

Konačni modeli, u računarstvu veoma korišćeni zbog niza osobina koje ih čine pogodnim za modeliranje različitih procesa i algoritama, našli su svoje mesto i u oblasti kakva je ekstrakcija informacija. Tokom godina, njihova upotreba u ovoj oblasti se menjala, od toga da su sistemi za ekstrakciju informacija bili potpuno bazirani na konačnim modelima, pa do toga da se koriste samo u ograničenom skupu koraka unutar procesa ekstrakcije, obično u fazi predprocesiranja tekssta. Pa ipak, njihov značaj je izuzetno velik, jer su nezamenjivi u situacijama kada se traži visoka preciznost procesa ekstrakcije informacije ili kada je potrebno obraditi neki tekst na efikasan način.

Dvofazni metod za ekstrakciju informacija baziran na konačnim transduktorima, koji je prikazan je u ovom radu, predviđa razdvajanje procesa identifikacije slogova podataka u tekstu od procesa izdvajanja konkretnih vrednosti atributa pojedinih entiteta. Na ovaj način omogućena je upotreba različitih softvera i tehnika u jednoj ili u drugoj fazi, čime je moguće poboljšati efikasnost procesa ekstrakcije. Implementacija metoda je demonstrirana projektovanjem konkretnog sistema za ekstrakciju i njegovom primenom na enciklopedijski tekst o mikrorganizmima. Sva pravila ekstrakcije su predstavljena konačnim transduktorima i rekurzivnim mrežama prelaza. Izvršena je i evaluacija sistema i ustanovljena je izuzetno dobra preciznost izdvojenih podataka, kao i mogućnost povećanja odziva daljim podešavanjem transduktora. Kao rezultat primene ovog metoda na enciklopedijski tekst o mikroorganizmima nastala je baza podataka namenjena za dalja istraživanja iz oblasti genetike i bioinformatike. Ovi rezultati objavljeni su u nekoliko radova publikovanih u časopisima ili izlaganih na međunarodnim konferencijama (Pajić, 2011; Pajić i sar. 2011a; Pajić i sar. 2011b) i predstavljaju osnov za dalja istraživanja, jedna u pravcu razvoja dvofaznog metoda, a druga u pravcu upotrebe dobijene baze podataka za istraživanje podataka iz oblasti biologije i bioinformatike.

LITERATURA

[1] Allen J. (1995) *Natural Language Understanding*. Benjamin/ Cummings.

[2] Appelt, D. E., Hobbs, J., Bear, J., Israel, D. i Tyson. M. (1993) FASTUS: A finite-state processor for Information Extraction from real world text. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1172–1178.

[3] Baumgartner, R., Flesca, S. i Gottlob, G. (2001) Visual Web Information Extraction with Lixto. In *Proceedings of the Conference on Very Large Databases (VLDB)*.

[4] Bilofsky, H.S. i Christian, B. (1988) The GenBank® genetic sequence data bank, *Nucl. Acids Res.* 16(5): 1861-1863 doi:10.1093/nar/16.5.1861

[5] Califf, M. E. i Mooney, R. J. (1999) Relational learning of pattern-match rules for Information Extraction. In *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*, pages 328–334, Orlando, FL, July 1999.

[6] Casacuberta, F., Vidal, E. i Picó, D. (2005) Inference of finite-state transducers from regular languages, *Pattern Recognition*, Volume 38, Issue 9, pp.1431-1443

[7] Cunningham, H., Maynard, D. Bontcheva, K. i Tablan, V. (2002) GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, July 2002.

[8] Friburger, N. i Maurel, D. (2004) Finite-state transducer cascades to extract named entities in texts, *Theoretical Computer Science* 313, 93 – 104

[9] Garrity, G. (2005) *Bergey's Manual of Systematic Bacteriology, Volume 2 : The Proteobacteria*, 2005, ISBN 978-0-387-95040-2

[10] Garrity, G., Don, J., Krieg, N.R. i Staley, J.T. (2005) *Bergey's Manual of Systematic Bacteriology, Volume Two: The Proteobacteria (Part C)* , ISBN 978-0-387-24145-6

[11] Grishman, R. i Sundheim, B. (1996) Message Understanding Conference 6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics* (pp 466-471), San Mateo, CA, 1996.
[12] Gross, M. and Perrin, D. (1987) "Electronic Dictionaries and Automata in Computational Linguistics", in *Proceedings of LITP Spring School on Theoretical Computer Science*, Saint-Pierre d'Oleron, France
[13] Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M. i Tyson, M. (1997) FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text, In Roche E. and Y. Schabes, eds., *Finite-State Language Processing*, The MIT Press, Cambridge, MA, pages 383-406.
[14] Jayram, T. S., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S. i Zhu, H. (2006) Avatar information extraction system, *IEEE Data Engineering Bulletin*, vol. 29, pp. 40–48, 2006.
[15] Jurafsky, D. i Martin, J. H. (2008) *Speech and language processing*, 2nd edition, Prentice-Hall Inc.
[16] Kornai, A. (1999) *Extended finite state models of langu*age, Cambridge University Press
[17] Krieg, N.R., Ludwig, W., Whitman, W.B., Hedlund, B.P., Paster, B.J., Staley, J.T., Ward, N., Brown, D. i Parte, A. (2010) *Bergey's Manual of Systematic Bacteriology, Volume 4: The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes*, ISBN 978-0-387-95042-6
[18] Liu, L., Pu, C., i Han, W. (2000) XWRAP: An XML-enabledWrapper Construction System for Web Information Sources. In *Intern. Conference on Data Engineering (ICDE)*, pages 611–621.
[19] Mirhaji, P., Byrne, S., Kunapareddy, N. i Casscells, SW. (2006) "Semantic approach for text understanding of chief complaints data", in *AMIA Annual Symposium Proceedings*, Washington, p.1033
[20] Moens, M.F. (2006) *Information extraction: algorithms and prospects in a retrieval context*, Springer, 2006.
[21] Pajić, V. (2011) Putting Encyclopaedia Knowledge into Structural Form: Finite State Transducers Approach, *Journal of Integrative Bioinformatics, Informationsmanagement in der Biotechnologie e.V.*, Germany, 8(2):164, ISSN 1613-4516.
[22] Pajić, V., Pavlović-Lažetić, G. i Pajić, M. (2011a) Information Extraction from Semi-structured Resources: A Two-Phase Finite State Transducers Approach, *Implementation and Application of Automata: Proceedings of 16th International Conference CIAA*, *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg 282-289, ISBN 3642222552, 9783642222559.
[23] Pajić, V., Pavlović-Lažetić, G., Beljanski, M., Brandt, B. i Pajić, M. (2011b) Towards a Database for Genotype-Phenotype Association Research: Mining Data from Encyclopedia, *International Journal of Data Mining and Bioinformatics*, Inderscience publishers, ISSN (Online): 1748-5681, ISSN (Print): 1748-5673, http://www.inderscience.com/browse/index.php?journalID=189&action=coming.
[24] Paumier, S. (2011) *Unitex 2.1 User Manual*, Universit´e de Marne-la-Vall´ee. http://www-igm.univ-mlv.fr/˜unitex/UnitexManual2.1.pdf
[25] Reiss, F., Raghavan, S., Krishnamurthy, R., Zhu, H. i Vaithyanathan, S. (2008) An algebraic approach to rule-based information extraction, in *ICDE*, 2008.
[26] Roche, E. (1999) Finite state transducers: parsing free and frozen sentences, *Extended finite state models of language*, Cambridge University Press, pp. 108.-120.
[27] Roche, E. i Schabes, Y. (1997) *Finite-state language processing,* The MIT Press
[28] Sastre, J. M. (2009) Efficient Parsing Using Filtered-Popping Recursive Transition Networks, *Lecture Notes in Computer Science*. vol. 5642, pp. 241–244
[29] Sastre, J. M. i Forcada, M. (2007) Efficient parsing using recursive transition networks with output, In Zygmunt Vetulani, editors, *Proceedings of 3rd Language & Technology Conference* (LTC'07) pp. 280–284
[30] Shen, W., Doan, A., Naughton, J. F. i Ramakrishnan, R. (2007) Declarative information extraction using datalog with embedded extraction predicates, in *VLDB*, pp. 1033–1044, 2007.
[31] Silberztein, M. (1993) *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*, Edition Masson, Paris, 1993.
[32] Soderland, S. (1999) Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34(1-3):233–272.
[33] Vitas, D. (2006) *Prevodioci i interpretatori: Uvod u teoriju i metode kompilacije programskih jezika*, Matematički fakultet, Belgrade, Republic of Serbia
[34] Vos, P., Garrity, G., Jones, D., Krieg, N.R., Ludwig, W., Rainey, F.A., Schleifer, K.H. i Whitman W.B. (2009) *Bergey's Manual of Systematic Bacteriology, Volume 3: The Firmicutes*, ISBN 978-0-387-95041-9

Mr Vesna Pajić, Univerzitet u Beogradu, Poljoprivredni fakultet, Beograd
Kontakt: svesna@agrif.bg.ac.rs
Oblasti interesovanja: ekstrakcija informacija, obrada prirodnih jezika, računarska lingvistika, bioinformatika.

Dr Miloš Pajić, Univerzitet u Beogradu, Poljoprivredni fakultet, Beograd
Kontakt: paja@agrif.bg.ac.rs
Oblasti interesovanja: poljoprivredna tehnika, informacioni sistemi u bioinženjeringu, projektovanje tehničkih sistema u poljoprivredi.

M. Sc. Staša Vujičić Stanković, Univerzitet u Beogradu, Matematički fakultet, Beograd
Kontakt: stasa@matf.bg.ac.rs
Oblasti interesovanja: obrada prirodnih jezika, ekstrakcija informacija, baze podataka, teorija formalnih jezika i automata.