VEB-BAZIRAN SISTEM ZA EFIKASNO DOBIJANJE ODGOVORA

Bojan Furlan, Boško Nikolić, Elektrotehnički fakultet, bfurlan@etf.bg.ac.yu, nbosko@etf.bg.ac.yu

Sadržaj – *U radu je predstavljen sistem koji služi za inteligentno pronalaženje odgovora na pitanja korisnika. Odgovor se traži prosleđivanjem pitanja drugim korisnicima sistema, i to osobi koja je najkompetentnija da pruži odgovor. Pretraga se obavlja pomoću posebno realizovanog algoritma veštačke inteligencije. Pri procesiranju pitanja koristi se napredna obrada prirodnog jezika. Takođe, u radu je opisana napredna mogućnost sistema da, sem prosleđivanja pitanja korisnicima sistema, je odgovor moguće tražiti i pozivanjem odgovarajućih aplikacija koje su registrovane kao semantički Web servisi. Sistem ima veliku primenu kako u fazi učenja, tako i kao pomoć pri upoznavanju sa novim pojmovima.*

Ključne reči: veštačka inteligencija, procesiranje prirodnog jezika, semantički Web servisi

1. Uvod

Živimo u svetu velikog tehnološkog razvoja. Proteklih godina postignut je ogroman napredak na polju dostupnosti informacija, pre svega u svetu internet pretraživača. Prvi put u svojoj istoriji, ljudi su u mogućnosti da u izuzetno kratkom vremenu dođu do novih saznanja. Ipak, sva rešenja koja trenutno postoje, poseduju određene slabosti.

Uprkos visokom stepenu razvoja veštačke inteligencije, ljudski mozak je još uvek superiorniji kada su u pitanju moć razumevanja i manipulacije delimično poznatim činjenicama. Na žalost, kada se govori o odgovoru na konkretno pitanje, još uvek ne postoji rešenje zasnovano na veštačkoj inteligenciji koje je u stanju da zameni čoveka - eksperta iz određene oblasti. Ukoliko je potreban logičan odgovor na neko konkretno pitanje, najbolje je da, umesto lutanja po moru informacija koje Internet nesumnjivo pruža, pitati osobu koja će shvatiti pitanje i dati kratak i jasan odgovor. Sa druge strane, daleko je od humanog eksploatacija čoveka za visoko zahtevne intelektualne poslove. Svaki čovek bi trebalo da radi ono u čemu je najbolji i ono što najviše voli.

Stoga, ideja autora je prilično jednostavna: nemojmo pitati mašinu nešto što nije u stanju da odgovori (na trenutnom nivou tehnološkog razvoja), pitajmo je nešto lakše, da pronađe pravu osobu koja je u stanju da pruži korektan odgovor. Na ovaj način čovek je i dalje zadužen za visoko umni deo posla, dok onaj zamorni, koji se ogleda u pronalaženju prave osobe je prepušten mašini. Takođe, sporedni efekat ogleda se u tome da sistem garantuje ispravnost odgovora pronalazeći trenutno najkompetentniju osobu.

Razvoj opisanog sistema zahtevao je pristup zasnovan na tehnikama veštačke inteligencije. Za potrebe ovog projekta osmišljen je poseban algoritam mašinskog učenja koji predstavlja jezgro sistema za inteligentno prosleđivanje pitanja razvijanog na Elektrotehničkom fakultetu [1].

U radu je prvo dat pregled korišćenih tehnologija. Zatim je detaljno opisan primenjeni algoritam i primer njegovog rada. Navedene su mogućnosti proširivanja funkcije i upotrebe sistema. Na kraju je izložen zaključak rada.

2. Pregled korišćenih tehnologija

U ovom odeljku je dat kratak opis tehnologija korišćenih za predstavu baze znanja i implementaciju Web servisa.

2.1 Tehnologije korišćene za implementaciju Web servisa

Windows Communication Foundation (WCF) [2] je skup .NET tehnologija za razvoj distribuiranih softverskih sistema. WCF predstavlja novi tip infrastrukture zasnovan na Web servis arhitekturi koja pruža jednostavnu, sigurnu i pouzdanu komunikaciju. Servisno orijentisan programski model WCF-a baziran je na *.NET Framework* tehnologiji što uprošćava razvoj i održavanje distribuiranih rešenja.

2.2 Tehnologije korišćene za predstavu baze znanja

ANTELOPE (*Advanced Natural Language Object-oriented Processing Environment*) je framework namenjen razvoju aplikacija zasnovanih na procesiranju prirodnog jezika (engl. *Natural Language Processing*) [3]. Primarno korišćene funkcionalnosti ovog programskog okruženja su ekstrahovanje konteksta iz teksta i njegov leksikon zasnovan na WordNet-u, koji je iskorišćen za predstavu baze znanja. Ovaj rečnik takođe u sebi sadrži deo podataka *eXtended WordNet* –a kao i mali deo *SUMO* ontologije [4].

WordNet® predstavlja veliku leksičku bazu podataka Engleskog jezika, razvijenu u okviru Cognitive Science Laboratory, Univerziteta Princeton [5]. Imenice, glagoli, pridevi i prilozi su grupisani u skupove sinonima (engl. *Synset*) opisujući različite koncepte. Ovi skupovi su međusobno spregnuti konceptualno-semantičkim i leksičkim vezama stvarajući mrežu smisleno povezanih reči i koncepata. WordNet sadrži oko 150,000 reči organizovanih u preko 115,000 synset-a što u zbiru čini oko 207,000 reč-smisao parova.

3. Algoritam odlučivanja

Srž ovog sistema predstavlja algoritam veštačke inteligencije koji definiše ponašanje sistema, omogućavajući izbor najkompetentnijeg korisnika za određenu vrstu pitanja. Algoritam je specijalno razvijan za ovaj konkretni problem i zasnovan je na tehnikama mašinskog učenja na pozitivnim primerima.

3.1 Struktura baze znanja

Baza znanja sastoji se od reči grupisanih u skupove sinonima (engl. *Synset*), gde svaki synset sadrži kratku, generalnu definiciju i poseduje različite semantičke veze prema ostalim skupovima sinonima, opisujući različite koncepte. Svi synset-i su opisani svojom lemom, kanoničkom formom lekseme koja se u ovom kontekstu odnosi na skup svih formi koje imaju isto značenje. Na

primer, lema imenice u množini *cars* je *car* ili glagola u prošlom vremenu *went* je *go*.

Bitan deo baze znanja predstavlja semantička mreža koja međusobno povezuje većinu koncepata, predstavljajući vrstu veze tipa *Concept-Concept*. Svaki koncept je jedinstveno određen svojim vezama prema drugima konceptima, što omogućava da bez uticaja na ispravnost rada algoritma mreža koncepata predstavljenih engleskim rečima bude zamenjena mrežom koncepata, npr. predstavljenih srpskim rečima.

Radi ilustracije razmatra se primer koncepta čokolade. Obzirom da čokolada (engl. *Chocolate)*, pripada grupi slatkiša, zatim sadrži mleko, šećer, kakao, itd. i može se opisati kao slatka, ovaj koncept je određenim težinama povezan sa konceptima slatko (engl. *Sweet)*, mleko (engl. *Milk)*, šećer (engl. *Sugar)*, kakao (engl. *Cacao)*, slatkiši (engl. *Sweets)*, itd. Na osnovu navedenog, date koncepte je moguće zameniti konceptima predstavljenim srpskim rečima čokolada, slatko, mleko, šećer, kakao, slatkiši, itd., bez uticaja na ispravnost rada sistema, pošto algoritam principijelno ne zavisi od identifikatora samog koncepta već od veza prema ostalim konceptima kojima je on jedinstveno opisan.

Drugi tip relacija u okviru baze znanja predstavljaju veze između korisnika i koncepta (*User-Concept* vrsta veze). Ova vrsta veze predstavlja osnovu za implementaciju algoritma mašinskog učenja obzirom da je učenje zasnovano na sličnom principu na kojem funkcioniše poimanje novih koncepata u ljudskom umu [6].

Na primer, ako nam neka osoba saopšti da ume da dobro kuva, kako će to biti reprezentovano u našem umu?

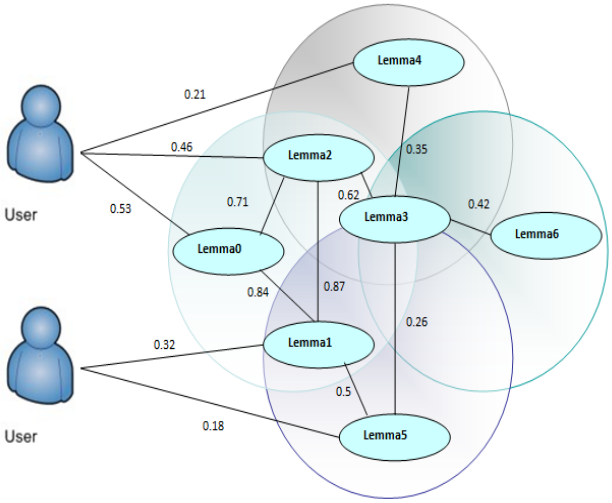
Na jednoj strani će se nalaziti koncept te osobe, dok će na drugoj postojati koncept kuvanja i između ova dva koncepta biće stvorena tanka veza. Zašto će ta veza biti tanka?

Zato što saopštenu činjenicu ljudski um prihvata sa određenom dozom nepoverenja. Kasnije u slučaju da nakon večere koju nam je ta osoba pripremila ne iskusimo stomačne probleme ova veza će postati deblja, dok u suprotnom, ona će postati još tanja što oslikava to da je poverenje u kuvarske sposobnosti našeg domaćina još manje.

Na osnovu toga relacija između korisnika i synset-a predstavlja ekspertizu te osobe za dati koncept, što u široj slici predstavlja ekspertizu u određenoj oblasti znanja, o čemu će biti diskutovano u nastavku rada.

Pojednostavljena ilustracija strukture baze znanja data je na **Slici 1.** Synset-i predstavljaju dijagrame koji su identifikovani svojom lemom. Kao što se može primetiti ovi dijagrami se preklapaju što označava to da određeni skupovi sinonima dele zajedničke reči. Takođe, može se uočiti da neki od koncepata konvergiraju većom verovatnoćom, tj. težinom prema drugim, što znači da skup tih koncepata u široj slici predstavlja određenu oblast znanja.

Težine veza koje povezuju koncepte predstavljaju meru distance, tj. sličnosti, pruženu od strane *Antelope* framework-a. U ovom trenutku postoji samo jedna metrika bazirana na Uzajamnom Informacionom Sadržaju. Druga veoma vredna mera sličnosti zasnovana je na Konceptualnim Vektorima. Na primer, ako reči *samuraj* i *Japan* ne predstavljaju dva slična koncepta sa tačke gledišta Uzajamnog Informacionog Sadržaja, onda iz ugla distance Konceptualnih Vektora su mnogo bliži. Ova metrika, po izjavi autora *Antelope*-a, bi trebalo da bude pridodata u bliskoj budućnosti, dajući ovom sistemu savršeniju meru sličnosti između koncepata.



Slika 1. Ilustracija sturkture baze znanja

3.2 Opis algoritma

Kada korisnik postavi pitanje sistemu procedura njegovog procesiranja izgleda ovako:

1. Sistem ekstrahuje kontekst pitanja dajući svakom uočenom konceptu težinski faktor w_c i formirajući skup *S1*.
2. Zatim sistem dohvata sve koncepte povezane sa konceptima iz skupa *S1* čija je težina veze veća od *s* formirajući skup *S2*.
3. Za svaki koncept iz skupa *S2* sistem izračunava njegovu vrednost po formuli:

$$w_c^j = \max_{i \in lc(j)} w_c^i * w_{lc}^j$$

   * *lc(j)* je podskup svih koncepata iz skupa *S1* koji su povezani sa konceptom *j* iz skupa *S2*.
4. Ako postoje koncepti iz skupa *S2* sa vrednošću većom od *s* oni će biti pridodati skupu *S1* dok će ostali biti uklonjeni iz skupa *S2* i algoritam će nastaviti izvršavanje od koraka 2. U suprotnom algoritam će nastaviti izvršavanje od koraka 5.
5. Sistem sračunava vrednost za svakog korisnika koji poseduje vezu prema nekom od koncepata iz skupa *S1* po formuli:

$$w_u = \sum_{j \in S1} w_c^j * w_{lu}^j$$

6. Sistem odabira prvih N rangiranih korisnika po sračunatoj vrednosti i njima prosleđuje pitanje.

U navedenom algoritmu upotrebljene oznake imaju sledeće značenje:

- ✓ *S1* – skup svih procesiranih koncepata
- ✓ *S2* – skup svih koncepata povezanih sa konceptima iz skupa *S1*
- ✓ w_c – težina datog koncepta

- ✓ w_u – sračunata vrednost za datog korisnika
- ✓ w_{lc} – težina veze tipa *Concept-Concept*
- ✓ w_{lu} – težina veze tipa *User-Concept*
- ✓ s – konstanta koja predstavlja donju granicu u sračunavanju težine određenog koncepta

3.3 Ažuriranje baze znanja

Kada korisnik koji je postavio pitanje dobije odgovor on ima mogućnost da ga oceni. Na ovaj način sistem uči na pozitivnim primerima, menjajući težinu veze između korisnika koji je odgovorio na pitanje i koncepata koji su doveli do odabira datog korisnika. Takođe, korisnik koji je odabran od strane sistema može odbiti da odgovori na određeno pitanje, obzirom da nije kompetentan, na osnovu čega se ažurira težine veza koje su dovele do datog korisnika.
Još jedan vid učenja ogleda se u tome da sistem uvećava težine veza između korisnika koji je postavio pitanje i koncepata ekstrahovanih iz samog pitanja, obzirom da ako neka osoba više puta postavi pitanje vezano za određenu oblast, veoma je verovatno da ta osoba poseduje određeno znanje iz te oblasti, možda ne tako veliko kao neko ko je ekspert, ali dovoljno da to sistem evidentira.

3.4 Održavanje baze znanja konzistentnom

U nastojanju da baza znanja ostane konzistentna sistem ažurira vrednosti brojača za svaki synset, koja reprezentuje neskaliranu maksimalnu težinu veze *User – Concept* za dati skup sinonima. Sračunavanje težine veze se vrši deljenjem neskalirane vrednosti određenog linka sa vrednošću datog brojača.
Obzirom da može doći do prekoračenja, vrednost brojača kada dostigne vrednost blizu maksimalne, deli se sa dva zajedno sa težinama svih veza pridruženih datom brojaču. Ova operacija nema uticaj na težine veza koje koristi algoritam pri svom izvršavanju.

4. Primer rada algoritma

Neka su trenutno ulogovani korisnici vezani određenim težinama za date koncepte i neka deo baze znanja izgleda kao na **Slici 2.** Neka postavljeno pitanje za koje je potrebno pronaći kompetentne korisnike glasi:

Can someone recommend me a good cook book?

Sistem će ekstrahovati kontekst datog pitanja, što u ovom slučaju predstavlja koncept *cooking*, kao i sinonimi *preparation* i *cookery* koji pripadaju istom synset-u, dok ostali ekstrahovani koncepti ovde neće biti dalje razmatrani obzirom da je njihova dodeljena težina zanemarljivo mala. Koncept koji pripada synset-u identifikovanom rečju *cooking* biće pridodat skupu *S1* sa dodeljenom težinom 1. Zatim će sistem dohvatati synset-e koji su povezani sa konceptom *cooking* i sračunavati njihove težine, što je prikazano na slici vrednostima u okviru elipsoide. Na kraju, sistem će sračunati vrednosti trenutno aktivnih korisnika, što prikazuju naznačene vrednosti iznad simbola korisnika.

4.1 Analiza dobijenih vrednosti

U nastavku je data analiza dobijenih rezultata primera sa **Slike 2.**

- ✓ **User1 –** Iako je ovaj korisnik direktno povezan sa konceptom iz postavljenog pitanja, on neće biti visoko rangiran od strane sistema, obzirom da težina veze nije velika. Razlog može biti to što korisnik nije ocenjivan visokom ocenom, od strane drugih korisnika pri odgovorima vezanim za ovu oblast ili jednostavno sistem još uvek nema saznanja o njegovim interesovanjima za navedenu oblast.
- ✓ **User3** - Ovaj korisnik je loše rangiran od strane sistema pošto je povezan samo sa jednim konceptom (*kitchen)* koji poseduje vezu prema konceptu iz postavljenog pitanja (*cooking*). Mala je verovatnoća da će sistem proslediti pitanje ovom korisniku, obzirom da njegova interesovanja možda nisu vezana za kulinarstvo (*cooking*), što je suština postavljenog pitanja, već su npr. vezana za kuhinjske elemente (*kitchen*), jer ovaj korisnik ima afinitete prema stolariji.
- ✓ **User2 -** Ovaj korisnik je najbolje rangiran od strane sistema za dato pitanje. Iako nije direktno povezan sa konceptom iz postavljenog pitanja, jako je povezan sa konceptima koje su veoma bliski ovom konceptu, tj. konvergiraju ka istoj oblasti znanja, pa samim tim, u datom trenutku predstavlja najkompetentniju osobu za dato pitanje.

5. Dalji rad

Obzirom da suštinu sistema predstavlja inteligentno pronalaženje odgovora, ideja autora je da, sem prosleđivanja pitanja korisnicima sistema, je odgovor moguće tražiti i pozivanjem odgovarajućih aplikacija koje su registrovane kao semantički Web servisi. S tim ciljem razvijeno je nekoliko aplikacija, prvenstveno iz oblasti Arhitekture i organizacije računara. Glavni problem sa kojim su se susreli autori predstavlja jezik kojim se upućuju pitanja sistemu, obzirom da je prirodan jezik suviše neregularan i kompleksan. Kako je napredak u razvoju oblasti razumevanja prirodnog jezika ogroman, ali još uvek nedovoljan, dalja istraživanja usmerena su u pravcu iznalaženju sintetičkog regularnog jezika koji bi s jedne strane bio čitljiv i razumljiv za čoveka, a sa druge strane ne suviše kompleksan za procesiranje od strane računara.
Jednu od mogućih smernica predstavlja Esperanto, veštački jezik razvijan više od dve stotine godina, gde je u poslednje vreme učinjen značajan pomak na razvoju ontologije koja opisuje gramatiku i logiku ovog jezika [7].

6. Zaključak

U radu je opisan sistem za efikasno dobijanje odgovora razvijan na Elektrotehničkom fakultetu u Beogradu. Cilj autora je da sistem postane dominantan medijum razmene znanja dopunjujući do sada korišćene vidove komunikacije kao što su veb forumi i mailing liste i pružajući pomoć svojim korisnicima kako u fazi učenja, tako i pri upoznavanju sa novim pojmovima.
Po mišljenju autora, uspeh ovog istraživačkog projekta imaće ogroman značaj za naučnu zajednicu, omogućavajući jednostavnu, brzu i pouzdanu razmenu znanja.

Zahvalnost

Ovom prilikom autori bi hteli da izraze svoju zahvalnost Zahariju Radivojeviću, Milošu Cvetanoviću, Stanku Nikoliću, Pavletu Josipoviću, kao i mnogim drugima koji su svojim savetima i zalaganjem pomogli da se ovaj istraživački rad podigne na viši nivo.

LITERATURA

[1] „aLive!-Sistem za inteligentno prosleđivanje pitanja," Stanko Nikolić, Bojan Furlan, Pavle Josipović, YUINFO 2008, Kopaonik, Srbija.

[2] „Windows Communication Foundation (WCF)", dostupno na Internet veb strani: http://msdn.microsoft.com/en-us/netframework/aa663324.aspx, posećeno 12.3.2008

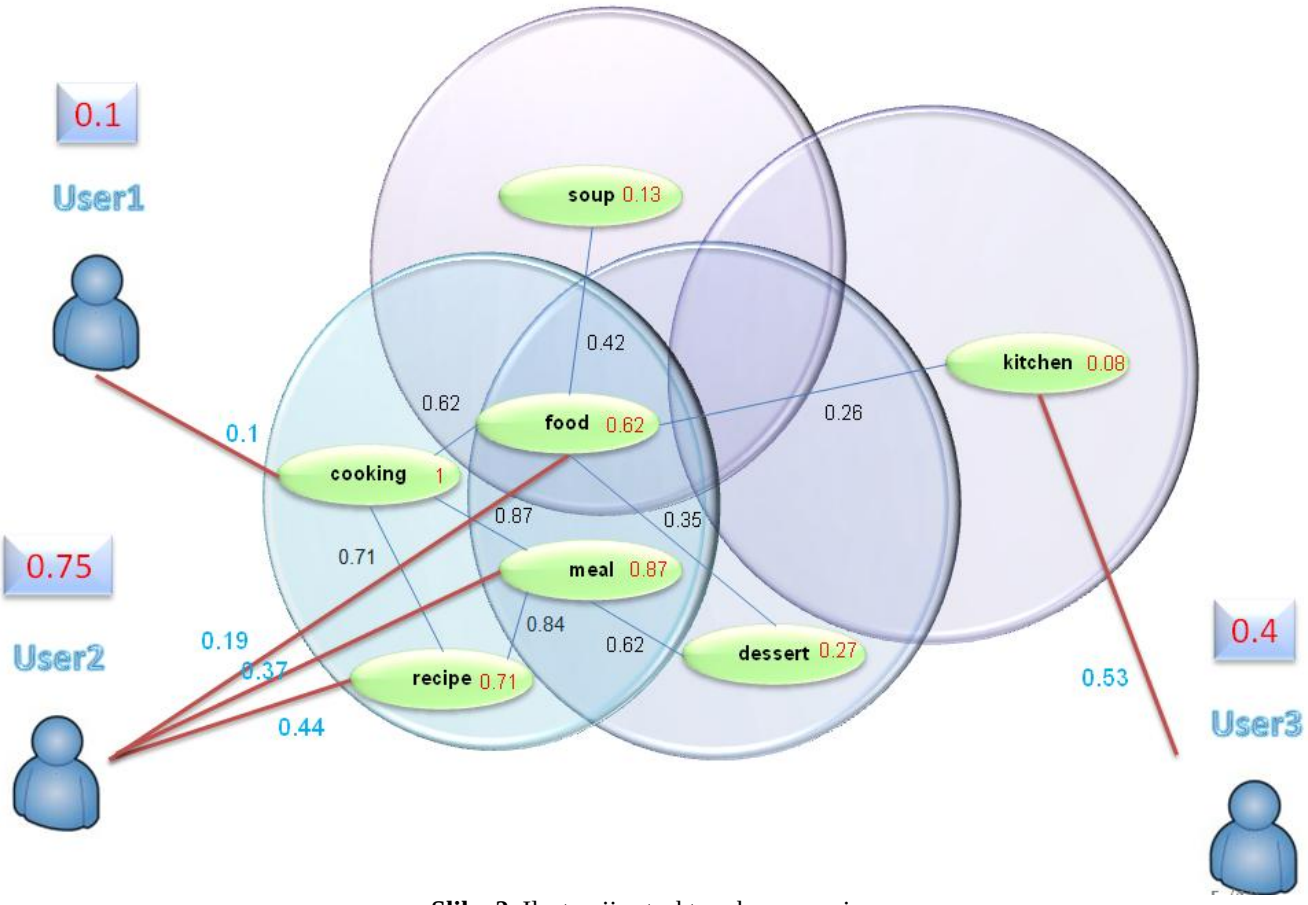
[3] „Antelope", dostupno na Internet veb strani: http://www.proxem.com/, posećeno 12.3.2008

[4] „Suggested Upper Merged Ontology (SUMO)", dostupno na Internet veb strani: http://www.ontologyportal.org/, posećeno 12.3.2008

[5] „Wordnet - a lexical database for the English language", dostupno na Internet veb strani: http://wordnet.princeton.edu/, posećeno 12.3.2008

[6] Stafford,T., Webb, M., "Mind Hacks," **O'Reilly**, November 2004, ISBN : 0-596-00779-5

[7] Romano, M., Severini, L., „Esperanto Ontology", dostupno na Internet veb strani: http://www.epistematica.com/papers/EsperantoOntology.pdf, posećeno 12.3.2008



Slika 2. Ilustracija strukture baze znanja