UNIVERZITET CRNE GORE

Prirodno-matematički fakultet Podgorica

Vijeću Prirodno-matematičkog fakulteta:

Predmet: Izvještaj Komisije o pregledu i ocjeni magistarskog rada *„Primjena genetskog algoritma za optimizaciju parametara algoritma izvlačenja informacija iz administrativnih dokumenata"* kandidata Koste Pavlovića, specijaliste računarskih nauka.

Na osnovu člana 60 Statuta Univerziteta Crne Gore, a u vezi sa članovima 25 i 28 Pravila studiranja na postdiplomskim studijama, na sjednici Vijeća PMF-a od 6. septembra 2018. godine imenovani smo za članove komisije za ocjenu magistarskog rada *„Primjena genetskog algoritma za optimizaciju parametara algoritma izvlačenja informacija iz administrativnih dokumenata"* kandidata Koste Pavlovića, specijaliste računarskih nauka. Kosta Pavlović je predao tekst magistarskog rada 11.09.2018. na uvid javnosti i ocjenu. Nakon uvida u podneseni tekst, a u vezi sa članom 29 Pravila studiranja na postdiplomskim studijama, podnosimo sljedeći

IZVJEŠTAJ

Magistarski rad Koste Pavlovića pod nazivom *„Primjena genetskog algoritma za optimizaciju parametara algoritma izvlačenja informacija iz administrativnih dokumenata"* je napisan na 64 strane, sadrži predgovor, izvod rada, izvod rada na engleskom jeziku, sadržaj i time ispunjava sve zahtjeve propisane članom 27 Pravila studiranja na postdiplomskim studijama. Rad pripada oblasti računarskih nauka, odnosno uže specijalizovanoj oblasti mašinskog učenja, i sastoji se iz šest poglavlja u kojima se detaljno i nadovezano definišu potrebni pojmovi i prikazuju rezultati rada, a pozivajući se na odgovarajuću literaturu.

Cilj rada je da razvije inovativan metod učenja pravila za ekstrakciju informacija. U radu se predlaže metod baziran na kodiranju parametrizovanih pravila za ekstrakciju informacija i učenju tih parametara primjenom genetskog algoritma. Metod je po preformansama uporediv sa drugim rješenjima a u određenim aspektima je superioran u odnosu na njih. Predloženi metod ne samo da smanjuje potrebu za ponovnom implementacijom pravila već i omogućava inkrementalno učenje na novim dokumentima iz iste klase. Kao demonstracija uspješnosti pristupa, implementiran je robustan sistem koji može izvlačiti informacije iz dokumenata koji pripadaju širokom spektru klasa.

Ogromna količina komunikacije između kompanija i dalje se odvija u pisanoj formi. Na taj način se generišu veliki skupovi dokumenata kojima kompanije moraju rukovati u svakom trenutku. Kontinuirani trend ka kancelariji bez papira utiče na razvoj novih tehnika za obradu dokumenata u kompanijama. Tu se posebno ističe proces izvlačenja odnosno ekstrakcije informacija iz administrativnih dokumenata koji se može opisati kao pronalaženje relevantnih strukturiranih informacija iz djelimično strukturiranih slika dokumenata. Na taj način se iz npr. slike računa za komunalne usluge mogu izvući i sačuvati u bazi podataka polja kao što su datum, iznos i broj računa. Poseban problem predstavlja heterogenost dokumenata, pa je proces ekstrakcije relevantnih

strukturiranih informacija iz djelimično strukturiranih dokumenata veoma složen. Ulazni dokumenti su veoma različiti po svojoj strukturi i uključuju pisma, račune, ugovore, faksove i formulare. Ručna obrada ovih podataka je veoma nepraktična i s aspekta organizacije i s aspekta finansija. U cilju prevazilaženja navedenih nedostataka razvijeni su sistemi za automatizovanu obradu velikog broja adminstrativnih dokumenata u što je moguće kraćem vremenu i uz minimalno intervenisanje zaposlenih. U zavisnost od tipa dokumenta, definišu se različiti elementi obrade: klasifikacija dokumenata, obrada višestraničnih dokumenata, razumijevanje sadržaja, usmjeravanje (routing) i izvlačenje informacija.

U dostupnoj literaturi poynata su tri opšta principa izvlačenja informacija: primjena klasifikatora, prepoznavanje prostornih šablona i kodiranje pravila za ekstrakciju. Kreiranje klasifikatora (npr. Support Vector Machines, rekurentne neuronske mreže, drveta odlučivanja...) se vrši na osnovu velikog skupa dokumenata za obučavanje. Formiranje skupa dokumenata za obučavanje jednog ovakvog sistema je dugotrajno i naporno i mora biti izvršeno od strane stručnih osoba, a ovaj proces može izazvati pravne i administrativne probleme u klijentskoj organizaciji u pogledu zaštite podataka. Klasifikatori nijesu pogodni u okruženjima gdje se struktura dokumenata može promijeniti, na primjer, dodavanjem novog polja ili uvođenjem nove klase dokumenata u sistem.

Mnogi potpuno automatizovani komercijalni sistemi ovaj proces zasnivaju na fiksiranim prostornim šablonima koji preslikavaju lokaciju koja je dobijena iz programa za optičko prepoznavanje karaktera u odgovarajuće polje. Ovakav pristup ne daje dobre rezultate kada se raspored polja na dokumentu mijenja.

Kod trećeg pristupa, programer koji kreira sistem za izvlačenje informacija je dužan da na osnovu sopstvenog znanja o dokumentima definiše pravila za ekstrakciju potrebnih informacija. Na ovaj način nije moguće kreirati sisteme kojima je dinamički moguće dodavati nove tipove dokumenata i definisati nova polja za izvlačenje. Takođe, svaka promjena na postojećem tipu dokumenta zahtijeva ponovnu implementaciju pravila.

U ovom radu kreirana je parametrizovana heuristika čije se vrijednosti računaju primjenom niza generičkih pravila. Evolucijom skupa heuristika genetskim algoritmom aproksimirane su optimalne vrijednosti parametara heuristike. Sprovedena su testiranja na različitim klasama dokumenata, sa različitim vrstama polja za ekstrakciju. Prezentovani su dobijeni rezultati, sa posebnim naglaskom na rezultate dobijene na dokumentima sa nepoznatim prostornim šablonima. Algoritam je, u zavisnosti od polja koje se traži, tačne vrijednosti pronalazio u 62% do 100% slučajeva što su performanse uporedive sa drugim pristupima iz literature. Ovo rješenje ponudilo je određena unapređenja u pogledu veličine skupa za obučavanje, kao i mogućnosti dodavanja novih klasa i polja kroz postepeno (inkrementalno) učenje.

U prvom poglavlju predstavljen je problem ekstrakcije informacija iz administrativnih dokumenata. Detaljno su opisani naučni izazovi na koje istraživači mogu naići prilikom izrade sistema koji rješavaju ovaj problem, a opisana je i arhitektura jednog takvog sistema.

U drugom poglavlju dat je pregled postojećih istraživanja i metoda ekstrakcije informacija, a detaljno su opisani pristupi iz nekoliko radova koji su imali najviše uticaja na razvoj algoritma za ekstrakciju informacija u ovom radu.

Treće poglavlje sadrži detaljan opis algoritam za ekstrakciju informacija koji je u suštini heuristička tehnika. Vrijednost predložene heuristike dobija se primjenom niza pravila kojima su dodijeljeni odgovarajući parametri. Što je veća vrijednost parametra veća je i važnost pravila za koje je on vezan. Sve klase dokumenata sa kojima ovaj sistem radi dijele iste vrijednosti parametara, jer svi administrativni dokumenti imaju slične karakteristike: određena polja imaju labele, polja se uglavnom javljaju na sličnim lokacijama, itd. Ove informacije su približno iste važnosti u svim klasama. Iz ovog razloga je u sistem moguće uvesti novu klasu dokumenata bez potrebe za obučavanjem na velikom skupu dokumenata, a i samo inkrementalno obučavanje je veoma jednostavno. Nakon što se izvrši računanje vrijednosti heuristike za svaki od tokena, poljem se proglašava ona grupa tokena za koju je prosječna vrijednost heuristike maksimalna.

Kako bi se dobio sistem sa što boljim performansama potrebno je pronaći optimalne vrijednosti parametara heuristike. Ovaj zadatak je izuzetno težak, pa su u radu ispitani modeli traženja dovoljno dobre aproksimacije ovih vrijednosti. Genetski algoritam se veoma često koristi za generisanje rješenja visokog kvaliteta u zadacima optimizacije, pa je to slučaj i u ovom radu. Opis genetskog algoritma dat je u četvrtom poglavlju. Dodatni razlog za korišćenje genetskog algoritma je činjenica da se u psihologiji smatra da se heuristike uče evolucionim procesima koje genetski algoritam oponaša.

Peto poglavlje sadrži opis izvršenih eksperimenata, kreiranog korpusa dokumenata i dobijenih rezultata. Sprovedena su tri eksperimenta na skupu dokumenata formiranom za potrebe ovog rada kako bi se uporedile performanse tri različite postavke genetskog algoritma. Za procjenu performansi algoritma za ekstrakciju korišćena je recall mjera koja daje najbolji uvid u ponašanje sistema. Prezentovani su rezultati za svaku od klasa dokumenata iz korpusa, kao i za svako polje koje se ekstrahuje sa dokumenata iz tih klasa.

Diskusija o predloženom metodu, zaključci kao i ideje za nastavak istraživanja dati su u šestom poglavlju.

ZAKLJUČAK I PREDLOG

Na osnovu prethodno izloženog, Komisija smatra da je kandidat Kosta Pavlović u potpunosti realizovao postavljene ciljeve. Rad je napisan jasno i pregledno, sa prikazom svih pojmova i tvrđenja koja se koriste. Kandidat je potvrdio da je ovladao složenim matematičkim i računarskim tehnikama i metodama i kroz eksperimente pokazao da predloženi algoritam ima performanse koje su uporedive sa drugim algoritmima ekstrakcije informacija. Rezultati dobijeni u ovom radu predstavljaju dobro koncipirane naučne rezultate i čine dobru osnovu za buduća istraživanja.

Komisija predlaže Vijeću Prirodno-matematičkog fakulteta da rad pod nazivom *„Primjena genetskog algoritma za optimizaciju parametara algoritma izvlačenja informacija iz administrativnih dokumenata“* kandidata Koste Pavlovića prihvati kao magistarski rad i odobri njegovu javnu usmenu odbranu.

U Podgorici, 14.09.2018

Komisija

Prof. dr Predrag Stanišić, član

Prof. dr Savo Tomović, član

Doc. dr Goran Šuković, mentor