UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

Ivana D. Tanasijević

MULTIMEDIJALNE BAZE PODATAKA U UPRAVLJANJU NEMATERIJALNIM KULTURNIM NASLEĐEM

doktorska disertacija

Beograd, 2020.

UNIVERSITY OF BELGRADE
FACULTY OF MATHEMATICS

Ivana D. Tanasijević

MULTIMEDIA DATABASES IN MANAGING THE INTANGIBLE CULTURAL HERITAGE

Doctoral Dissertation

Belgrade, 2020.

Mentor:

Prof. dr Gordana Pavlović-Lažetić, redovni profesor
Univerzitet u Beogradu, Matematički fakultet

Članovi komisije:

Prof. dr Nenad Mitić, redovni profesor
Univerzitet u Beogradu, Matematički fakultet

Prof. dr Gordana Pavlović-Lažetić, redovni profesor
Univerzitet u Beogradu, Matematički fakultet

dr Jelena Graovac, docent
Univerzitet u Beogradu, Matematički fakultet

dr Biljana Sikimić, naučni savetnik
Srpska akademija nauka i umetnosti, Balkanološki institut

Datum odbrane: _____

Porodici

Zahvaljujem se svima koji su pomogli u realizaciji ovog rada. Veliku zahvalnost dugujem svojoj mentorki prof. dr Gordani Pavlović-Lažetić za usmeravanje u izradi ove doktorske disertacije, za brojne stručne savete, kao i za srdačnu saradnju, volju i veliku podršku da se ovaj rad privede kraju.

Zahvaljujem se komisiji na značajnim sugestijama koje su doprinele kvalitetu disertacije. Veoma sam zahvalna prof. dr Nenadu Mitiću na podršci, korisnim savetima i konstruktivnim predlozima kojima je poboljšan kvalitet ovog rada. Zadovoljstvo mi je da se zahvalim doc. dr Jeleni Graovac na stalnoj podršci i spremnosti za saradnju, životnim i stručnim savetima koji su svi zajedno doprineli toku izrade ove disertacije. Posebnu zahvalnost želim da izrazim dr Biljani Sikimić na srdačnoj saradnji i uvođenju u problematiku koja je postala motivacija za sprovedena istraživanja i izradu ove disertacije, kao i na sugestijama koje su doprinele da kvalitet konačnog sadržaja disertacije bude bolji.

Zahvaljujem se profesorima i brojnim kolegama na saradnji, korisnim savetima i podršci. Zahvaljujem se porodici na strpljenju i razumevanju tokom perioda koji sam posvetila radu na ovoj doktorskoj disertaciji. Zahvaljujem se ocu, majci i sestri na stalnoj podršci i ljubavi tokom mog života, sestrićima na neumornoj dečjoj ljubavi tokom celog njihovog života, kao i zetu na podršci i pozitivnom duhu. Zahvaljujem se prijateljima što su verovali u mene i imali strpljenje i razumevanje za moju čestu zauzetost u toku izrade ove disertacije.

Posebno želim da se zahvalim mom Vladi na nesvakidašnjoj nesebičnoj podršci i ljubavi koje su mi bile najlepši izvor snage da ovaj rad dobije svoj konačni oblik.

I.T.

Naslov disertacije: Multimedijalne baze podataka u upravljanju nematerijalnim kulturnim nasleđem

Rezime: Motivacija za izradu ove doktorske disertacije je multimedijalna kolekcija koja je rezultat višegodišnjih terenskih istraživanja istraživača iz Balkanološkog instituta Srpske akademije nauka i umetnosti. Kolekcija se sastoji od materijala u vidu snimljenih intervjua, snimljenih raznih običaja, pridruženih tekstualnih opisa (protokola) i brojnih drugih dokumenata.

Predmet istraživanja ove disertacije je proučavanje mogućnosti i razvoj novih metoda kojima bi se započelo rešavanje problema upravljanja nematerijalnim kulturnim nasleđem Balkana. Podzadaci koji se tom prilikom otvaraju su razvoj adekvatnog dizajna i implementacije multimedijalne baze podataka nematerijalnog kulturnog nasleđa koja bi odgovarala potrebama različitih vrsta korisnika, automatska semantička anotacija protokola uz pomoć metoda obrade prirodnih jezika, kao osnova za polu-automatsku anotaciju multimedijalne kolekcije i uspešnu pretragu po metapodacima koji su u skladu sa CIDOC CRM standardom, istraživanje dodatnih mogućnosti pretrage ove kolekcije u cilju dobijanja novih znanja, kao i razvoj izabranih metoda.

Glavni problem sa dostupnim metodama je u tome što još uvek nema dovoljno razvijene infrastrukture u kontekstu obrade teksta na prirodnom jeziku, organizacije i upravljanja u domenu kulturnog nasleđa na prostoru Balkana i posebno za slučaj srpskog jezika, koja bi se mogla efektivno koristiti za rešavanje postavljenog problema. Stoga, postoji izražena potreba za razvojem metoda kojima bi se došlo do odgovarajućeg rešenja.

Za polu-automatsku anotaciju multimedijalnih materijala korišćena je automatska semantička anotacija protokola koji su pridruženi materijalima. Ona je sprovedena metodama ekstrakcije informacija, prepoznavanja imenovanih entiteta i ekstrakcije tema, tehnikama zasnovanim na pravilima uz pomoć dodatnih resursa poput elektronskih rečnika, tezaurusa i rečnika reči iz specifičnog domena.

Za klasifikaciju tekstualnih protokola u odnosu na tematiku, izvedeno je istraživanje o metodama koje se mogu primeniti za rešavanje problema klasifikacije tekstova na srpskom jeziku, i ponuđena je metoda koja je prilagođena specifičnom domenu koji se obrađuje (nematerijalno kulturno nasleđe), specifičnim problemima koji se rešavaju (klasifikacija protokola u odnosu na tematiku) i srpskom jeziku, kao jednom od morfološki bogatih jezika.

Za rad sa prostornim podacima razvijen je prostorni model koji je pogodan za prikaz rezultata na mapi kao i za postavljanje prostornih upita putem interaktivnog grafičkog prikaza mape lokacija.

Rezultati eksperimenata nad razvijenim metodama pokazuju da korišćenje pristupa zasnovanog na pravilima u kombinaciji sa dodatnim jezičkim resursima i uz ulaganje razumnog truda daje veoma dobre rezultate za zadatak ekstrakcije informacija. Za ekstrakciju imenovanih entiteta dostignuta je F mera 0.87, dok je za ekstrakciju tema dostignuta F mera 0.90, što je u rangu mera iz objavljenih istraživanja nad sličnim problemima i iz sličnih domena.

Rezultati klasifikacije teksta ukazuju da izabrane statističke metode mašinskog učenja u svom osnovnom obliku primenjene na protokole, iako generalno uspešne, daju lošu F meru, 0.44, dok se značajano poboljšanje postiže uz korišćenje semantičkih tehnika, u kom slučaju se dostiže F mera 0.88.

Deo rezultata ove disertacije je sadržan u radovima [267], [266], [94], [265], [264], koji su objavljeni ili prihvaćeni za objavljivanje.

Zaključak koji je izveden na osnovu sprovedenih istraživanja je da je za rešavanje postavljenog problema neophodno angažovanje eksperata iz više oblasti, da su složene potrebe različitih grupa korisnika što usložnjava i zadatak organizacije i upravljanja multimedijalnom kolekcijom, da je domen kulturnog nasleđa veoma bogat semantikom, da kontekst igra veliku ulogu u zadacima ekstrakcije informacija i klasifikacije teksta, kao i da se za ove zadatke razvijene metode obrade prirodnih jezika zanovane na pravilima i na statističkim tehnikama mašinskog učenja pokazuju kao uspešne.

Ključne reči: multimedijalne baze podataka, prostorne baze podataka, obrada teksta na prirodnom jeziku, mašinsko učenje, ekstrakcija informacija, klasifikacija teksta, automatska semantička anotacija, pretraživanje informacija, upravljanje nematerijalnim kulturnim nasleđem

Naučna oblast: računarstvo

Uža naučna oblast: multimedijalne baze podataka, obrada prirodnog jezika

UDK broj: [004.65:004.032.6]+004.85(043.3)

Dissertation title: Multimedia databases in managing the intangible cultural heritage

Abstract: The motivation for writing this doctoral dissertation is a multimedia collection that is the result of many years of field research conducted by researchers from the Institute for Balkan studies of the Serbian Academy of Sciences and Arts. The collection consists of materials in the form of recorded interviews, various recorded customs, associated textual descriptions (protocols) and numerous other documents.

The subject of research of this dissertation is the study of possibilities and the development of new methods that could be used as a starting point in solving the problem of managing the intangible cultural heritage of the Balkans. The subtasks that emerge in this endeavor are the development of adequate design and implementation of a multimedia database of intangible cultural heritage that would meet the needs of different types of users, automatic semantic annotation of protocols using natural language processing methods, as a basis for semi-automatic annotation of the multimedia collection, and successful search by metadata which comply with the CIDOC CRM standard, study of additional search possibilities of this collection in order to gain new knowledge, as well as development of selected methods.

The main problem with the available methods is that there is still not enough developed infrastructure in the context of natural language processing, organization and management in the field of cultural heritage in the Balkans and especially for the Serbian language, which could be effectively used to solve the proposed problem. There is thus a strong need to develop methods to reach an appropriate solution.

For the semi-automatic annotation of multimedia materials, automatic semantic annotation of the protocols associated with the materials was used. It was carried out by methods of information extraction, recognition of named entities and topic extraction, using rule-based techniques with the help of additional resources such as electronic dictionaries, thesauri and vocabularies from a specific domain.

To classify textual protocols in relation to the topic, research was conducted on methods that can be used to solve the problem of classifying texts in the Serbian language, and a method was offered that is adapted to the specific domain being processed (intangible cultural heritage), to the specific problems being solved (classification of protocols in relation to the topic) and to the Serbian language, as one of the morphologically rich languages.

To work with spatial data, a spatial model has been developed that is suitable

for displaying results on a map, as well as for creating spatial queries through an interactive graphical display of a map of locations.

The results of experiments conducted on the developed methods show that the use of a rule-based approach in combination with additional language resources and with putting in a reasonable amount of effort gives very good results for the task of information extraction. An F measure of 0.87 was reached for the extraction of named entities, while an F measure of 0.90 was reached for the extraction of topics, which is in the range of measures from published research from similar problems and domains.

The results of the text classification indicate that the selected statistical methods of machine learning in their basic form when applied to the protocols, although generally successful, give a bad F measure, 0.44, while significant improvement is achieved with the use of semantic techniques, in which case an F measure of 0.88 is reached.

Some of the results presented in this dissertation are contained in the papers [267], [266], [94], [265], [264], which have been published or accepted for publication.

The conclusion drawn from the research is that to solve the given problem it is necessary to engage experts from several fields, that the needs of different groups of users are complex, which complicates the task of organizing and managing the multimedia collection, that the domain of cultural heritage is very rich in semantics, that context plays a major role in the tasks of information extraction and text classification, and finally that for these tasks the developed rule-based methods of natural language processing as well as statistical techniques of machine learning prove to be successful.

Keywords: multimedia databases, spatial databases, natural language processing, machine learning, information extraction, text classification, automatic semantic annotation, information retrieval, managing of intangible cultural heritage

Research area: computer science

Research sub-area: multimedia databases, natural language processing

UDC number: [004.65:004.032.6]+004.85(043.3)

Sadržaj

Glava 1

Uvod

1.1 Kulturno nasleđe i digitalizacija

Kulturno nasleđe je svaki pojam ili objekat koji se odnosi na kulturu nekog naroda ili grupe ljudi. Pod kulturnim nasleđem podrazumeva se materijalno i nematerijalno nasleđe. Materijalno kulturno nasleđe je ono koje se može videti i opipati, dok je nematerijalno kulturno nasleđe ono koje živi u pamćenju ljudi i prenosi se govorom ili praktikovanjem. Konvencija UNESCO iz 2003. godine ([201], [150]) prepoznaje različite domene u kojima se nematerijalno kulturno nasleđe može manifestovati. To su tradicije govora i izražavanja, izvođenje umetničkih dela, društvena dešavanja, rituali, praznični događaji, razvoj i primena zanatskih veština, kao i znanje i njegova primena u vezi sa prirodom i svetom oko nas. Ministarstvo kulture i informacija Republike Srbije oformilo je 2012. godine Nacionalni komitet za nematerijalno kulturno nasleđe u okviru Etnografskog muzeja u Beogradu. Srbija ima svoju stranu ([285]) na UNESCO sajtu svetskog kulturnog nasleđa ([286]). Očuvanje nacionalnog kulturnog nasleđa od izuzetnog je značaja za istovremeno očuvanje i negovanje identiteta svakog pojedinca.

Srpsko kulturno nasleđe je svakako autohtono ali je istovremeno bilo pod uticajem više različitih tradicija ([117], [205], [149], [256]). Kulturna baština Srbije obuhvata materijalno i nematerijalno nasleđe iz različitih istorijskih perioda srpskog naroda ali i svih drugih naroda koji su živeli ili i danas žive na prostorima Srbije.

Istorijat Materijalno kulturno nasleđe Srbije datira još od praistorijskog perioda. Najznačajniji oblici kulture na ovim prostorima predstavljaju arheološka nalazišta Lepenskog Vira i Vinčanske kulture. U rimskom periodu postojale su carske palate

i hramovi koji se i danas mogu videti u Sirmijumu, Viminacijumu, Medijani i Gamzigradu. Od srednjovekovnih spomenika koje su gradili srpski vladari najpoznatiji su brojni manastiri, Visoki Dečani, Žiča, Manasija, Mileševa, Đurđevi Stupovi, Sopoćani, Pećka Patrijaršija, Studenica i Gračanica, koji danas predstavljaju svetsku kulturnu baštinu i zaštićeni su od strane Uneska. Po dolasku osmanske vlasti razvitak srpske kulture stagnira, čak i nazaduje zbog manjka školovanih i pismenih ljudi. U devetnaestom veku mladi umetnici koji su se školovali u inostranstvu obnavljaju kulturu i umetnost i donose savremene stilove svoga doba.

Nematerijalno kulturno nasleđe srpskog naroda (i drugih naroda koji danas žive u Srbiji) seže u daleku prošlost, a danas je još uvek očuvano u načinu proslavljanja pojedinih tradicionalnih praznika kao što su krsna slava, Bogojavljenje, Poklade, Uskrs, Đurđevdan, Badnji dan i Božić. Koreni običaja vezanih za ove praznike potiču još iz doba pre primanja hrišćanstva. Početak srpske književnosti vezuje se za doba Ćirila i Metodija (deveti vek), koji su sastavili prvo slovensko pismo. Miroslavljevo jevanđelje se smatra za najznačajniji spomenik ćiriličnog nasleđa, potiče iz dvanaestog veka, a danas se čuva u Narodnom muzeju u Beogradu. Sistematsko beleženje srpskog folklora: epske i lirske pesme, priče, predanja i male folklorne forme, kao i ogromnog leksičkog blaga, započinje tek početkom devetnaestog veka zahvaljujući delu Vuka Stefanovića Karadžića. Kao nematerijalno kulturno nasleđe Srbije danas su kod Uneska zaštićena tri elementa: slava (2014), kolo (2017) i pevanje uz gusle (2017).

Upotrebom informacionih tehnologija i digitalizacije, u novije vreme, kao rezultat u domenu kulturnog nasleđa je stvaranje digitalnih baza podataka o nematerijalnoj kulturnoj baštini. Kako je postalo tehnički jednostavnije stvaranje digitalnih multimedijalnih dokumenata, to je takvih materijala sve više. Digitalni mediji mogu biti veoma veliki i složeni i vrlo često postaje nemoguće upravljati njima, efikasno pretraživati i dobijati relevantne podatake ili cele dokumente bez uspostavljanja odgovarajuće organizacije i korišćenja pomoćne tehnologije. Pod organizacijom se smatra smeštanje dokumenata u multimedijalnu bazu podataka, uz definisanje odgovarajuće sheme za opis metapodataka i označavanje dokumenata po relevantnim metapodacima.

Uloga informacionih tehnologija Tehnologije koje mogu biti od velike pomoći u radu sa nematerijalnim kulturnim nasleđem su metode pronalaženja informacija u tekstovima na prirodnom jeziku, pretraživanje na osnovu sadržaja i drugih vrsta

dokumenata - slika, audio ili video materijala, posebno prilagođen rad sa prostornim podacima i različite vrste vizuelizacije podataka. Tehnologije istraživanja podataka mogu biti od izuzetne koristi za dobijanje novog znanja koje se može proizvesti na osnovu dostupnih prethodno neistraženih veoma vrednih resursa iz domena kulturnog nasleđa.

Kulturna raznolikost i dugačka tradicija održavanja autentične kulturne baštine iznedrili su specifično nematerijalno nasleđe koje Srbija danas ima. Nedovoljno dobra prezentacija i promocija, loši uslovi za rad i malo interesovanje ljudi, učinili su da se nematerijalno kulturno nasleđe poslednjih godina nađe na društvenoj margini.

Upoznavanje domaće i strane javnosti sa radom, trudom, posvećenošću, ali i preprekama sa kojima se danas susreću ljudi koji neguju nematerijalno kulturno nasleđe, važan je korak ka očuvanju autentične kulturne baštine ([197]).

Detaljniji pregled oblasti nacionalnog kulturnog nasleđa može se naći u [292], [178], [24] i [268].

Vodeći standardi za opis kulturnog nasleđa U cilju jednostavnijeg pristupa, prenosivosti, tumačenja i spajanja informacija o kulturnim dobrima iz više izvora pogodno je imati zajednički model uz pomoć koga se prilikom digitalizacije sva dela kulturnog nasleđa mogu staviti u jedan okvir. To značajno smanjuje vreme potrebno za pravljenje odgovarajućih okvira na pojedinačnim sistemima, kao i za prilagođavanje podataka između različitih sistema. Tokom 2003. godine objavljen je CIDOC Conceptual Reference Model (CRM) ([33]), formalna ontologija za transformaciju i integraciju informacija iz domena kulturnog nasleđa. Nekoliko godina kasnije ovaj model postaje i ISO standard. CRM definiše semantičku shemu podataka i struktura dokumenata putem ontologije. Ovaj standard nije obavezujući u smislu da dokumenti moraju da imaju navedene elemente. Kako dokumenti uglavnom sadrže takve elemente, omogućava se da se oni sistematičnije semantički opišu, a samim tim budu bolje prilagođeni za pridruživanje drugim dokumentima i za razmenu među različitim platformama. CIDOC CRM se bavi opštijim osobinama koje mogu da opisuju dokumente, a implementacije posebnih zadataka treba da vode računa o specifičnim osobinama.

CIDOC CRM definiše koncepte poput klasa (engl. *entity*), relacija (engl. *property*) i pravila nasleđivanja (engl. *inheritance rule*). Primer entiteta su klase `Event` i `Activity`, dok se putem relacije `is_a` opisuje odnos između natklase i potklase (na primer, `Activity is an Event`). Relacije između klasa navode se za označavanje

dodatnih osobina. Jedna od relacija je `to` (na primer, `Monument to a Person`). Više o tome šta je sve obuhvaćeno ovim modelom može se pogledati u kraćem pregledu CIDOC konceptualnog referentnog modela ([57]). Celokupna dokumentacija se nalazi na zvaničnoj strani poslednjeg izdanja CIDOC CRM ([34]).

Dublin Core inicijativa za metapodatke ([58]) je veoma korišćen standard za definisanje metapodataka kojima se opisuju kulturna dobra. Njegov razvoj je započet 1995. godine, i trajao je sve do 2008. godine. Na samom početku standardizovano je petnaest elemenata iz široke upotrebe: autor, učesnik, izdavač, naslov, datum, jezik, format, tematika, opis, identifikator, relacija, izvor, tip fajla, pokrivenost i intelektualna svojina, što je aktuelno i danas.

Preporuke Nacionalnog centra za digitalizaciju (NCD) ([192]) za opisivanje nematerijalnog kulturnog nasleđa Srbije opisane su u radu [202]. U radu [203] dat je aktuelan pregled aktivnosti koje se bave digitalizacijom kulturnog nasleđa Srbije.

1.2 Multimedijalne baze podataka

Multimedijalna baza podataka je kolekcija digitalnih dokumenata koji mogu biti tekst, grafika (crteži i ilustracije), slika, animacija, audio ili video materijali. Mediji se mogu podeliti na statičke, dinamičke i dimenzionalne. Statički mediji su tekst, grafika i slike. Pod dinamičkim medijima podrazumevaju se animacije, audio i video materijali. Dimenzionalni mediji su 3D igre i računarski generisani objekti (poput geometrijskih modela). Multimedijalni podaci mogu da sadrže različite tipove informacija. Grafički podaci se sastoje od fotografija, logoa ili crteža. Audio podaci se sastoje od govora, muzike, ili nekih drugih audio zapisa. Video podaci predstavljaju kombinaciju zvuka i slika. Multimedijalni podaci često nisu potpuno strukturirani, mogu zahtevati više prostora za skladištenje ili netrivijalnu obradu poput primene različitih algoritama kompresije.

Za efikasno rukovanje raznovrsnim i složenim podacima, kao što je multimedijalna baza, neophodan je sistem za upravljanje prilagođen specifičnim zahtevima. Sistem za upravljanje multimedijalnim bazama podataka (Multimedia Database Management System MMDBMS) nudi podršku za definisanje baze, stvaranje, skladištenje, pristup i pretraživanje multimedijalnih podataka, kao i kontrolu multimedijalne baze. Ovakav sistem treba da pruži metode za modelovanje podataka, za njihovo efikasno skladištenje, indeksiranje u skladu sa multimedijalnom prirodom i podršku za složenu pretragu različitih oblika multimedijalnih podataka. MMDBMS zahte-

va, pored standardnih funkcija sistema za upravljanje bazama podataka (Database Management System DBMS), veliki kapacitet za skladištenje dokumenata, multimedijalni interfejs i interaktivnost, mogućnost za postavljanje upita zasnovanih na sadržaju multimedije, izvršavanje indeksiranja i pretraživanja neposredno nad kompresovanim podacima. Multimedijalni sadržaj je često povezan međusobno prostornim i vremenskim karakteristikama, koje se takođe čuvaju u bazi. Ovakvi zahtevi za povezanošću mogu se modelovati relacionim, objektno-relacionim ili objektno-orjentisanim multimedijalnim bazama podataka. Relacionim modelima se modeluju složeni odnosi jednostavnijih tipova podataka, dok je objektna organizacija prikladnija za upravljanje složenim objektima raznovrsne strukture, kojima se predstavlja multimedija.

Primeri aplikacija koje koriste multimedijalne baze mogu biti digitalne biblioteke, video-na-zahtev, muzičke baze podataka, geografski informacioni sistemi, medicinski informacioni sistemi, telemedicina, udaljeno učenje (e-learning), interaktivna televizija, video igre, marketing, elektronske enciklopedije, pravne baze otisaka prstiju i slika, baze podataka danas veoma aktuelnih različitih društvenih mreža za razmenu multimedijalnog sadržaja, kao i inteligentne multimedijalne aplikacije koje omogućavaju integrisanje semantike kao što su audiovizuelna identifikacija osoba, uočavanje neuobičajenih događaja (na primer, kao podrška za bezbedonosne sisteme) i prepoznavanje specifičnog neprimerenog sadržaja.

Istorijat Razvoj multimedijalnih baza podataka tekao je paralelno sa razvojem tradicionalnih baza podataka. Inicijalni razvoj sistema za upravljanje multimedijalnim bazama podataka bio je početkom osamdesetih godina prošlog veka i oslanjao se na mogućnost operativnog sistema za skladištenje velikih binarnih objekata (Binary Large Objects BLOBS) kojima bi se efikasno čuvali ovakvi podaci. Ovi sistemi pretežno su služili kao repozitorijumi. Postojala je mogućnost rada sa različitim tipovima podataka, što je obuhvatalo funkcionalnosti kao što su unošenje, postavljanje upita i pronalaženje podataka.

Jedan od prvih multimedijalnih sistema je ImageAXS ([115]) koji se koristio kao baza sa slikama za ličnu upotrebu. Ovaj sistem postoji i danas sa dodatom podrškom za druge vrste informacija, među kojima je najznačajniji fleksibilan sistem za definisanje metapodataka. Danas se koristi pridružen serverima za muzeje i biblioteke, kao i interaktivnim alatima za veb.

Sistem MediaDB ([133]) predstavlja prelaz između ovog prvobitnog rešenja i

multimedijalnih baza podataka za opštu upotrebu. U to vreme bio je jedinstven po tome što je podržavao klijent / server arhitekturu za multimedijalne aplikacije. Omogućavao je prilagođenu podršku za širi spektar različitih tipova medija. Novina u odnosu na ranije sisteme je i mogućnost definisanja strukturiranih metapodataka koji su potrebni za izvođenje kompleksnih upita nad multimedijalnim podacima. Dodata je podrška za logovanje, konkurentnost, arhiviranje i kontrolu pristupa.

Nakon toga nastaju drugi komercijalni sistemi kojima je zadatak bio da omoguće čuvanje kompleksnih tipova objekata za različite medije ([140]). Objektno-orjentisana paradigma omogućila je način za definisanje novih tipova podataka i operatora za nove vrste medija, kao što su grafika, audio, video, kao i mogućnost međusobnog povezivanja multimedijalnih objekata. Jedan od prvih primera takvog sistema je UniSQL ([49]), koji je objektno-relaciona baza podataka. Ovaj sistem ima podršku za složene tipove objekata kojima se mogu predstaviti različiti mediji. U svakom slučaju, multimedijalni podaci se skladište kao nesemantički veliki objekti koji imaju pridružene ručno anotirane metapodatke. I dalje se nije ulazilo u sadržaj multimedijalnih objekata sa aspekta pretraživanja. Istaknuti primeri objektnih sistema proširenih multimedijalnim funkcionalnostima su Oracle 10g, IBM DB2, IMB Informix ([262]). IBM DB2 Universal Database proširuje objektno-relacioni sistem podrškom za upravljanje grafičkim, audio, video i prostornim objektima ([263]).

Oblast multimedijalnih baza podatka nastavlja da se razvija i u smeru potreba aplikacija sa izraženijim semantičkim sadržajem. Početkom dvehiljaditih godina razvija se standard MPEG-7 (poznat i pod nazivom Multimedia Content Description Interface) ([186]) za opis karakteristika visokog i niskog nivoa multimedijalnog sadržaja. MPEG-7 MDC (Multimedia Data Cartridge) je sistemsko proširenje Oracle DBMS koje omogućava upotrebu multimedijalnog upitnog jezika, pristup medijima, izvršavanje i optimizaciju upita i označavanje resursa oslanjajući se na sheme proistekle iz MPEG-7 standarda. Uključeni su različiti načini za prikupljanje informacija, modelovanje multimedijalnih podataka, označavanje sadržaja i sistem za upravljanje, koji podržava multimedijalne podatke kao objekte prvog reda pogodne za čuvanje i pronalazak na osnovu njihovog semantičkog značaja.

Načini pretrage Pretrage multimedijalnih materijala po sadržaju mogu se vršiti različitim tehnikama prilagođenim za različite medije - tekst, sliku, audio ili video. Aktuelni pregled tehnika se može naći u radu [6]. Primer sistema koji koristi pretragu audio materijala po sadržaju je Shazam ([305]). To je sistem koji na osnovu sekvence

melodije efikasno pronalazi numeru kojoj ona pripada. Sistem Mirage ([180]) izračunava sličnost između audio materijala u kolekciji i automatski generiše plejliste numera koje su slične, pri čemu se sličnost izračunava na osnovu određenih audio analiza sadržaja ([213]). Pretraga po sadržaju video materijala takođe je aktuelno polje istraživanja. Karakteristike po kojima se ovakvi materijali pretražuju mogu biti raznovrsne, od onih nižeg nivoa, koje mogu biti prepoznate statističkim metodama mašinskog učenja, do različitih semantičkih koncepata višeg nivoa ([214], [109]). Sistem iMARS ([196]) kompanije IBM, vrši automatsku anotaciju multimedijalnog materijala semantičkim karakteristikama kao što su scene, događaji i ljudi. Ovaj sistem se sastoji iz dva dela, prvi deo vrši automatsko ekstrahovanje semantičkih karakteristika kojima se označavaju multimedijalni materijali. Drugi deo integriše semantičko pretraživanje sa drugim tehnikama pretrage, kao što su tehnike zasnovane na pretrazi po tekstualnim karakteristikama. Još neki od multimedijalnih sistema u kojima je moguća pretraga po sadržaju su ([52]): MediaMill ([312]), AXESPRO ([172]), CuZero ([315]).

Detaljniji pregled oblasti multimedijalnih baza podataka može se naći u [293], [275] i [70].

1.3 Multimedijalni sistemi za organizaciju kulturnog nasleđa

Veliki broj digitalnih biblioteka sada je dostupan na vebu, što omogućava posebno specijalizovanim istraživačima, ali i široj zajednici, da koriste ove resurse u mnogo dostupnijem i udobnijem maniru.

Neke od najvećih javno dostupnih digitalnih biblioteka kulturnog nasleđa su Europeana ([67]) i Digital Public Library of America (DPLA) ([56]).

Europeana je najveći projekat Evropske Unije u cilju digitalizacije kulturnog nasleđa, uspostavljen 2008. godine. Trenutno se sastoji od oko 57 miliona digitalnih materijala iz preko 3500 institucija, među kojima je i Narodna biblioteka Srbije ([193], [27]). Uopšteno, ona predstavlja platformu za organizaciju i povezivanje velikog broja kolekcija kulturnog nasleđa različitih žitelja Evrope. Ideja je da se razvijenim sistemom višeslojne arhitekture različitim problemima u domenu digitalizacije i organizacije kulturnog nasleđa na tlu Evrope ponudi jednoobrazno rešenje ([38]).

DPLA, uspostavljena 2013. godine, takođe objedinjuje materijale iz različitih institucija i trenutno se sastoji od preko 37 miliona materijala različitih medija (tekst,

slika, audio, video, 3D). Materijali se mogu pretraživati po pridruženim metapodacima koji su u skladu sa Dublin Core standardom.

U domenu kulturnog nasleđa, značajna pažnja posvećuje se sistemima za dinamičko generisanje preporuka i plana obilaska kulturnih znamenitosti prilagođeno korisnikovim interesovanjima. Jedan od takvih sistema primenjen je nad multimedijalnim materijalima iz galerije Uffizi, Italija, ([284]), koji za izračunavanje preporuka određenom korisniku koristi korisnikov definisani zahtev zasnovan na sadržaju, prethodne izbore tog korisnika, ali i prethodne izbore ostalih korisnika ([3]). Ovaj pristup koristi formiranje izbora na osnovu sadržaja (engl. *content-based filtering*) i na osnovu zajednice (engl. *collaborative filtering*). Unapređeni pristup za sužavanje izbora koristi i informacije o kontekstu (engl. *contextual pre-filtering*), kao što su korisnikova lokacija, vremenske prilike i objekti u okolini, nakon čega se primenjuju prethodna dva izvora informacija ([15]).

Multimedijalni sistem za organizaciju nematerijalnog kulturnog nasleđa koji se nalazi u Arhivu etnografije i narodne istorije lombardijske regije ([160]) opisan je u radu [11]. Pretraga po karakteristikama nematerijalnog kulturnog nasleđa je izdvojena u poseban deo koji se naziva IntangibleSearch ([161]). Ovaj sistem omogućava unos novih materijala uz pridruživanje metapodataka. Kao dodatan alat za izbor odgovarajućih vrednosti navodi se MultiWordNet rečnik ([179]). Uz pomoć ovog alata olakšava se ujednačavanje termina prilikom pridruživanja vrednosti za metapodatke. Podržana je pretraga materijala na više načina - putem sistema metapodataka (na primer, tip dokumenta, autor, lokacije), pretrage fraze u tekstu i pretrage na osnovu mape. Podržan je i prikaz na mapi lokacija koje su pridružene materijalima. Kao budući plan razvoja navodi se implementacija pretrage slika po sličnosti i naprednija pretraga teksta uz pomoć MultiWordNet rečnika. Na kraju, navodi se da bi bilo korisno razviti poboljšanje interaktivnosti prilikom prikaza sadržaja.

Razvijene su i veb platforme, poput Omeka ([99]), WissKI ([237]) i Arches ([190]), za organizaciju i deljenje digitalnih multimedijalnih kolekcija. Primer projekta koji koristi Omeka platformu je The rich story of North American square dance ([273]). Ova multimedijalna kolekcija sadrži oko 1800 materijala i može se pretraživati po pridruženim metapodacima koji oslikavaju različite vrste karakteristika, kao na primer datum, lokacija, tip, opis, trajanje, tema, učesnici i potkolekcija.

Sistem za upravljanje multimedijalnim sadržajem MILOS (Multimedia dIgital Library Object Server) ([7]) razvijen je za rad sa heterogenim dokumentima i metapodacima koji su im pridruženi. Razvoj ovog sistema bio je motivisan porastom

potrebe za relevantnim informacijama, porastom cene za čuvanje i pretragu takvih dokumenata, kao i razvojem rešenja kojim bi se omogućilo integrisanje različitih sistema. MILOS je uglavnom orjentisan na organizaciju digitalnih biblioteka i opisuje mogućnost mapiranja sheme metapodataka tako da bude postignuta nezavisnost - jedan dokument može biti opisan pomoću više različitih shema metapodataka. U sistemu MILOS, za skladištenje se koristi XML format, za pretragu se koristi XQuery jezik, navodi se mogućnost pretraživanja slika po sličnosti i pretraživanja teksta full-text pretragom.

Projekat Museum24 ([189]) imao je za cilj da omogući pristup lokalnom kulturnom nasleđu regiona Jamsa u Finskoj, da poveća njegovu vidljivost i učini ga dostupnim široj zajednici, školama i turistima. U okviru projekta napravljen je sistem za organizovanje tekstova i slika.

National Research Institute of Cultural Heritage NRICH ([195]) je organizacija u Koreji koja sprovodi istraživanja i razvojne projekte radi odgovarajućeg prepoznavanja, očuvanja i korišćenja kulturnog nasleđa nacije. NRICH se bavi oblastima arheologije, umetničkog nasleđa, arhitektonskog nasleđa, nematerijalne kulturne baštine i nauke o očuvanju prirode. O nematerijalnoj kulturnoj baštini postoji baza podataka sa spiskom dokumentarnih filmova, knjiga, radova i audio materijala.

Infrastruktura za organizovanje semantičkih podataka heterogenog kulturnog nasleđa Koreje opisana je u radu [135]. Ovo istraživanje se bavi analizom sheme metapodataka posmatranih postojećih baza podataka kulturnog nasleđa Koreje, modelom podataka zasnovanom na ontologiji za kulturno nasleđe Koreje koje bi objedinilo prethodne sheme, definicijom klasa podataka i njihovih osobina, analizom korpusa u cilju razvoja bolje podrške za upotrebu semantičke dimenzije sadržaja.

Sistem za pretraživanje multimedijalnih i geografskih podataka, primenjen na muzeje u Napulju, Italija, opisan je u radu [224]. Predstavljene su mogućnosti integracije između pretraživanja slika po sadržaju i pretraživanja po geografskim informacijama. Nakon pretrage slika po sadržaju izdvajaju se one koje zadovoljavaju geografski kriterijum, a on može biti zasnovan na korisnikovoj lokaciji i lokaciji multimedijalnog materijala iz rezultata pretrage.

Još neke od inicijativa koje su se bavile, ili se još uvek bave, nacionalnim kulturnim nasleđem su Cultural Capital Counts ([46]) (može se vršiti pretraga po državi, kategoriji i oblasti, a ima i mogućnost izbora oblasti putem mape), National Database of Intangible Cultural Heritage of India ([194]) (može se vršiti pretraga po naslovu, tipu i autoru), Asia-Pacific Database on Intangible Cultural Heritage ([12])

(pretraga se vrši na osnovu unete fraze).

U radu [157] napravljen je pregled digitalnih biblioteka kulturnog nasleđa sa sistematizovanim prikazom funkcionalnosti kojima se može vršiti pretraživanje informacija. Jedan od zaključaka koji je izveden u radu je da nijedna od navedenih digitalnih biblioteka nije koristila tehnologije obrade prirodnih jezika kao pomoć u pretraživanju informacija. Naglašeno je da bi integracija ovih tehnologija nemerljivo doprinela kvalitetu rada sa digitalnim bibliotekama kulturnog nasleđa.

Više o multidisciplinarnim istraživanjima koja se bave prethodnim oblastima može se naći u [43], [44] i [104].

Glava 2

Upravljanje tekstualnim podacima

2.1 Uvod

Tekstualni podaci mogu biti raznovrsni. Informacije koje se čuvaju i prenose u tekstualnom obliku su, na primer, različite vrednosti ekstrahovane iz baza podataka, oblici pisanog govora ili misli, simbolička uputstva pisana za razumevanje od strane mašine, mašinski generisani tekstovi u cilju prenošenja znanja, ili, pak, podaci koji se ne mogu uklopiti u konvencionalne sisteme.

Jedan vid čuvanja i organizovanja tekstualnih podataka je u obliku tekstualnog dokumenta.

Istorijat Tekstualni dokument se može sastojati od elemenata sadržaja i elemenata formatiranja. U prvim sistemima za opis dokumenata komande za formatiranje bile su pomešane sa sadržajem (tekstom). Najraniji jezici za obeležavanje teksta, nastali šesdesetih godina prošlog veka, bili su RUNOFF (nastao na Massachusetts Institute of Technology MIT) i FORMAT (razvijen u kompaniji IBM). U njima je prvi put uveden koncept razdvajanja komandi od teksta, a opis se odnosio na određivanje elemenata poput strane, zaglavlja, pasusa, reči, fraza i karaktera ([80]).

Početkom sedamdesetih godina prošlog veka Stanford Artificial Intelligence Laboratory razvija PUB sistem ([269]). U odnosu na prethodne sisteme, dodate su mogućnosti za definisanje kolona, fusnota, odeljaka sa automatskim dodeljivanjem rednih brojeva, oblasti poput naslova ili teksta, i promenljivih poput veličine margine i datuma. PUB je bio sistem između programskog jezika i jezika za opis dokumenta.

Nakon toga, nastaju i drugi jezici za obeležavanje, kao na primer, Generalized Markup Language (GML) ([91]), koji omogućavaju viši nivo apstrakcije, predsta-

vljanje dokumenta preko strukture stabla elemenata sadržaja i naprednije metode za formatiranje. U ovom periodu nastaje TeX sistem ([138]) za koji je Donald Knuth razvio napredne algoritme za formatiranje koji do današnjeg dana nisu prevaziđeni.

Do sredine osamdesetih godina prošlog veka raste raznolikost jezika za obeležavanje, što ima za posledicu složeniju prenosivost. Kao odgovor na taj problem nastaje Standardized Generalized Markup Language (SGML) ([248]) i Open Document Architecture (ODA) ([10]). SGML se pokazao kao uspešan i bio je dobra osnova za razvoj drugih jezika za obeležavanje, kao što je Hypertext Markup Language (HTML) ([19]) koji se intenzivno koristio na vebu. Ispostavlja se da je, sa aspekta obrade, zbog dobre strukturiranosti HTML pogodan za opis stranica teksta.

Extensible Markup Language (XML) ([26]) nastaje kao rešenje koje omogućava da se ostvari bolja semantička struktura koja je potrebna za finije i promišljenije upravljanje aplikacijama koje bi ga koristile. XML je zapravo pojednostavljeni SGML. Suštinska razlika između XML-a i prethodnih reprezentacija je u tome da je XML skoro sasvim deklarativan, odnosno opisuje hijerarhijsku strukturu sadržaja sa eventualnim metapodacima. XML opisuje logičku reprezentaciju dokumenta, ali ne i način na koji treba da bude prikazan. Zbog svoje jednostavnosti i jasne sintakse, iako je napravljen za predstavljanje dokumenata, našao je primenu i u drugim vrstama aplikacija. Postoji dobro organizovan sistem tehnika, jezika, mehanizama obrade i provere grešaka, koji radi sa podacima u XML formatu, kao što su XSLT ([37]), XSD ([274]) i XQuery ([22]).

2.2 Predstavljanje podataka u XML formatu

Jedan od važnijih formata za predstavljanje tekstualnih dokumenata je XML format. XML je jezik za obeležavanje teksta kroz pridruživanje semantičkih informacija određenim njegovim delovima, a koji može biti različitih nivoa složenosti. XML format se, sa druge strane, može koristiti i za organizovanje informacija koje ne potiču izvorno iz nekog dokumenta, već, na primer, mogu biti deo neke veb aplikacije.

Postoje dva tipa XML dokumenata - orjentisan na podatke (engl. *data-centric*) i orjentisan na dokument (engl. *document-centric*), koji se odnose na to da li im je osnovna namena skladištenje podataka i veza između njih ili skladištenje tekstualnih dokumenata. Data-centric dokumenti uglavnom imaju pravilnu strukturu i organizaciju, a sadrže manje jedinice koje su nosioci pojedinačnih podataka. Oni se prete-

žno koriste za prenos podataka iz jednog sistema u drugi i najčešće su napravljeni programski. Document-centric dokumenti, sa druge strane, uglavnom imaju manje pravilnu strukturu i prigodniji su za čitanje korisniku - čoveku. Njih je najčešće napravio čovek za skladištenje već postojećih dokumenata, samo u XML formatu.

Primer jednog teksta sa pridruženim XML etiketama može izgledati ovako:

```
<citat>
  "Za objašnjenje vaseljene potrebno je više preklapajućih teorija, baš
  kao što je više preklapajućih mapa potrebno za prikazivanje Zemlje.",
  iz knjige <naslov>Velika zamisao</naslov>, autora <autor>Stephen
  Hawking</autor> i <autor>Leonard Mlodinow</autor>.
</citat>
```

U ovom primeru navedeno je da je prikazani tekst u stvari citat, sa označenim delovima teksta, naslovom i autorima. Kada se izostave etikete, tekst ostaje nepromenjen i redosled pojavljivanja informacija u tekstu je očuvan.

XML se može koristiti i za organizovanje određenih informacija u strukturiranu formu, na primer:

```
<knjiga>
  <naslov>The grand design<naslov>
  <autor>Hawking, Stephen<autor>
  <autor>Mlodinow, Leonard<autor>
  <godina>2010</godina>
  <izdavac>Random House Digital</izdavac>
</knjiga>
```

Prvi dokument je tipa document-centric, dok je drugi tipa data-centric. U drugom tipu dokumenata redosled pojavljivanja informacija ne mora biti od ključnog značaja, na primer, godine mogu biti navedene na početku ili na kraju bez promene njihove suštine i funkcije.

XML format dozvoljava veliku raznolikost načina na koji se mogu obeležiti podaci. Često je potrebno postaviti određena ograničenja koja treba da budu zadovoljena prilikom označavanja. U tu svrhu se koriste Document Type Definition (DTD) i Extensible Shema Definition (XSD) tehnike za definisanje skupa pravila koje XML dokument treba da zadovoljava da bi se smatrao ispravno formiranim.

DTD

DTD sadrži skup pravila kojima se definiše struktura nekog XML dokumenta. Definisanje strukture podrazumeva definicije elemenata koje dokument može da sadrži, njihovih atributa, tipova podataka, kao i odnosa između elemenata. DTD se može definisati u okviru XML dokumenta ili u posebnom dokumentu.

U sledećem primeru je dat DTD koji se definiše u okviru XML dokumenta:

```
<?xml version="1.0"?>
  <!DOCTYPE poruka [
  <!ELEMENT poruka (primalac,posiljalac,naslov,sadrzaj)>
  <!ELEMENT primalac EMPTY>
  <!ATTLIST primalac ime CDATA #REQUIRED adresa CDATA #REQUIRED>
  <!ELEMENT posiljalac EMPTY>
  <!ATTLIST posiljalac ime CDATA #REQUIRED adresa CDATA #REQUIRED>
  <!ELEMENT naslov (#PCDATA)>
  <!ELEMENT sadrzaj (#PCDATA)>
  ]>

<poruka>
  <primalac ime="Ana" adresa="ana@hello.com"/>
  <posiljalac ime="Marija" adresa="marija@hello.com"/>
  <naslov>Sastanak</naslov>
  <sadrzaj>Vidimo se sutra u 10h.</sadrzaj>
</poruka>
```

Prethodna DTD definicija može se interpretirati na sledeći način:

- `!DOCTYPE poruka` - označava da se kao koreni element mora naći element `poruka`
- `!ELEMENT poruka` - označava da se u okviru elementa `poruka` moraju naći elementi `primalac`, `posiljalac`, `naslov` i `sadrzaj`
- `!ELEMENT primalac EMPTY` - označava da element `primalac` ne treba da ima sadržaj
- `!ATTLIST primalac ime CDATA #REQUIRED adresa CDATA #REQUIRED` - označava da element `primalac` ima dva atributa, ime i adresu, pri čemu su oba atributa obavezna i treba da se sastoje od tekstualnog sadržaja

- `!ELEMENT posiljalac EMPTY` - označava da element `posiljalac` ne treba da ima sadržaj
- `!ATTLIST posiljalac ime CDATA #REQUIRED adresa CDATA #REQUIRED` - označava isto kao i za element `primalac`
- `!ELEMENT naslov (#PCDATA)` - označava da element `naslov` može da sadrži neki tekst
- `!ELEMENT sadrzaj (#PCDATA)` - označava da element `sadrzaj` može da sadrži neki tekst.

DTD se može definisati i van XML dokumenta, u izdvojenom dokumentu, koji se zatim povezuje sa XML dokumentom navođenjem njegove putanje u zaglavlju.

Moguće je hijerarhijski predstaviti odnose između elemenata, sa kardinalnošću i egzistencijom elementa, spojiti informacije iz više elemenata ili u jednom elementu ukazati na drugi element (u stilu primarnog i stranog ključa u relacionim bazama podataka).

XML shema

XSD je alternativni način DTD-u za definisanje strukture XML dokumenta. Omogućava precizniji opis dokumenta u poređenju sa DTD-om, bolje rukovanje imenskim konfliktima, pravljenje sopstvenih tipova podataka i njihovo preciznije navođenje. Pomoću XSD se takođe može proveriti ispravnost XML dokumenta, odnosno da li je dokument u saglasnosti sa navedenom shemom. Može se napraviti paralela između XSD i sheme baze podataka, koja opisuje definicije tabela u bazi.

Primer jedne sheme bi bio:

```
<?xml version = "1.0" encoding = "UTF-8"?>
<xs:schema xmlns:xs = "http://www.w3.org/2001/XMLSchema">
   <xs:element name = "klijent">
      <xs:complexType>
         <xs:sequence>
            <xs:element name="ime" type="xs:string"/>
            <xs:element name="kompanija" type="xs:string"/>
            <xs:element name="telefon" type="xs:integer"/>
         </xs:sequence>
      </xs:complexType>
```

```
    </xs:element>
</xs:schema>
```

U ovom primeru, element `klijent` ima podelemente i on je sadržalac za definicije drugih elemenata - `ime`, `kompanija` i `telefon`.

U shemi se mogu jednostavno koristiti prosti tipovi podataka (`xsd:string`, `xsd:integer`, `xsd:date`). Postoji mogućnost postavljanja ograničenja na proste tipove čime se, praktično, formiraju novi tipovi podataka.

Moguće je definisati globalne tipove, koji se u tom slučaju definišu po jednom van bilo kog elementa i koriste u definicijama elemenata gde god je to potrebno.

Atributi su mehanizam za pridruživanje jednostavnijih informacija u okviru elementa.

2.3 Rad sa podacima u XML formatu

Extensible Stylesheet Language (XSL) ([69]) je familija jezika za rad sa podacima u XML formatu. Sastoji se od XPath (za navigiranje kroz XML dokumente), XSLT (za transformisanje XML dokumenata) i XSL-FO (za formatiranje XML dokumenata). XQuery je upitni jezik koji se koristi za rad sa podacima u XML formatu.

U nastavku su ilustrovane najvažnije karakteristike elemenata od kojih se sastoji XSL, kao i karakteristike jezika XQuery.

XPath

XPath je sintaksa za prolazak kroz XML dokumente. Njegov model podataka ima strukturu stabla. Svaki čvor, bez obzira na tip, ima ime i vrednost. Vrednost svakog čvora je tekst u formi stringa. Čak i ako se čvor sastoji od drugih čvorova, vrednost je string dobijen nadovezivanjem stringova koji su vrednosti čvorova sadržanih u tom čvoru. Čvor može imati roditeljski čvor, decu čvorove, atribute i prostor imena.

XPath 3.0 postaje preporuka W3C standarda 2014. godine. XPath izrazom se opisuju lokacije čvorova u strukturi stabla. On, uopšteno, predstavlja putanju koja se sastoji od koraka razdvojenih kosom crtom (/). Svakim korakom se izdvaja skup čvorova koji su u određenom odnosu sa tekućim čvorom. Ti čvorovi postaju tekući čvorovi za sledeći korak. Skup čvorova koji ostanu nakon primene poslednjeg koraka predstavlja rezultat XPath izraza.

Sintaksa XPath omogućava kretanje kroz hijerarhijsku organizaciju podataka u XML dokumentu takvo da se može napraviti paralela sa kretanjem proz sistem fajlova i direktorijuma u fajl sistemu. Ovde su izdvojena osnovna pravila za pisanje XPath izraza.

Primitive koje se koriste u XPath izrazima su stringovski literali, brojevi, promenljive, pozivi funkcija i izrazi u zagradama.

Primeri formulisanja putanja su:

- `E` - izdvaja se element `E` koji je dete tekućeg čvora
- `/E` - izdvaja se element `E` koji je dete korenog čvora
- `//E` - izdvajaju se svi elementi `E` u dokumentu
- `@A` - izdvaja se atribut `A` tekućeg čvora
- `odeljak[1]/recenica[2]` - izdvaja se drugi element `recenica` koji je dete prvog elementa `odeljak`
- `zaglavlje[starts-with(naslov, 'A')]` - izdvajaju se svi elementi `zaglavlje` koji su deca tekućeg čvora i čiji sadržaj počinje velikim slovom A
- `/korak` - izdvajaju se čvorovi koji se mogu dostići iz korenog čvora primenog navedenog koraka
- `korak` - izdvajaju se čvorovi koji se mogu dostići iz tekućeg čvora primenom navedenog koraka.

Korak može imati naveden i predikat ili listu predikata koje čvor mora da zadovoljava da bi bio izdvojen. Svaki od ovih predikata je XPath za sebe, koji se navodi u uglastim zagradama i izračunava se za svaki čvor ponaosob. Kontekstni čvor je kontekst koji se odnosi na ceo izraz, dok je kontekst predikata čvor koji se testira za dati predikat.

Primer XPath izraza koji pronalazi poslednji paragraf knjige sa zadatim naslovom, gde god da se našla u dokumentu, izgleda ovako:

```
//knjiga/naslov[text()="Kratka istorija čovečanstva"]/paragraf[last()]
```

U okviru specifikacije XPath 3.0 postoji i oko 200 ugrađenih funkcija koje se mogu koristiti. Neke od njih su već korišćene u navedenim primerima. Postoje funkcije za rad sa stringovima, brojevima, logičkim vrednostima, čvorovima, nizovima, i mnoge druge.

XSLT

XSLT je jezik višeg nivoa za definisanje XML transformacija, pri čemu se pojam XML transformacija odnosi na transformisanje jednog XML dokumenta u drugi XML dokument. Na sličan način je koncipiran i SQL jezik koji transformiše tabele u jednu rezultujuću tabelu, s tom razlikom da su ulazni i izlazni dokumenti za XSLT organizovani hijerarhijski ili pomoću strukture stabla, umesto preko tabela kao što je to slučaj sa SQL-om.

Jedna primena XSLT-a je transformisanje podataka u dokument koji sadrži relevantne informacije za prikaz, na primer u HTML, XHTML ([188]) ili Scalable Vector Graphics format (SVG) ([73], [234]).

XSLT koristi šablone kojima se označava šta treba da se uradi sa tekstom koji se nalazi na ulazu. U XSLT jeziku koriste se XML etikete za navođenje radnji koje se žele izvršiti.

Primer jednog šablona bi bio:

```
<xsl:template match="cena">
  <b>$<xsl:value-of select="format-number(., '#0.00')" /></b>
</xsl:template>
```

Navedeni šablon, za svaki element sa nazivom `cena`, sadržaj tog elementa formatira tako što unutar oznake `<b>` dodaje znak `$` i upisuje sadržaj na dve decimale.

Jedna značajna instrukcija je `xsl:apply-templates` pomoću koje se za izabrane elemente primenjuju definisani šabloni. Postoje i brojne druge XSLT instrukcije za definisanje načina na koji se želi raditi sa podacima.

U XSLT postoje operatori `xsl:if`, za uslovno izvršavanje, i `xsl:for-each`, za petlje, a instrukcije se mogu izvršavati u bilo kom redosledu. Teorija na kojoj se zasniva XSLT ima sličnosti sa funkcionalnim programiranjem - svaki deo izlaza je funkcionalno zavisan od određenog dela ulaza. Ove funkcionalne zavisnosti međusobno nisu zavisne.

Tipovi podataka koji se koriste su skupovi čvorova organizovani u strukturu stabla. Kao što SQL vraća skup redova tabele, tako i XSLT izraz vraća skup čvorova početnog stabla.

Za više informacija korisno je pogledati zvaničnu dokumentaciju [128].

XSL-FO

XSL-FO (The Extensible Stylesheet Language Formatting Objects) je jezik za formatiranje stranica, odnosno opisivanje načina za predstavljanje njihovog sadržaja prilikom ispisa.

Primer dokumenta u XSL-FO zapisu je:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<fo:root xmlns:fo="http://www.w3.org/1999/XSL/Format">
  <fo:layout-master-set>
    <fo:simple-page-master master-name="A4"
     page-width="297mm" page-height="210mm"
     margin-top="1cm"   margin-bottom="1cm"
     margin-left="1cm"  margin-right="1cm">
      <fo:region-body   margin="3cm"/>
    </fo:simple-page-master>
  </fo:layout-master-set>
  <fo:page-sequence master-reference="A4">
    <fo:flow flow-name="xsl-region-body">
      <fo:block>Zdravo svete!</fo:block>
    </fo:flow>
  </fo:page-sequence>
</fo:root>
```

U ovom primeru uz pomoć XSL-FO definiše se stranica formata A4 sa navedenim marginama i regionom `body` u kome je ispisan tekst "Zdravo svete!".

XSL-FO sadrži širok spektar mogućnosti za definisanje raznovrsnih elemenata prikaza. Više o tome može se pogledati u knjizi [217].

XQuery

XQuery je funkcionalni jezik za formulisanje upita nad podacima u XML formatu, kao i za formatiranje izlaznih podataka. Iteracije u XQuery jeziku predstavljaju instrukcije paralelnog programiranja koje se nazivaju FLWOR (for, let, where, order by, return). Svaka iteracija u FLWOR izrazu predstavlja zasebnu nit izvršavanja. Rezultati instrukcija u ovim iteracijama u istom FLWOR izrazu ne mogu se koristiti kao ulazi za druge iteracije, niti se mogu prenositi vrednosti promenljivih, što

je i glavna odlika funkcionalnog programiranja. XQuery 1.0 je postao W3C preporuka 2007. godine, XQuery 3.0 postaje preporuka 2014. godine, dok je XQuery 3.1 preporuka od 2017. godine.

Pretraga podataka putem XQuery jezika bazirana je na XPath izrazima uz pomoć kojih se pronalaze određeni čvorovi u XML dokumentu. XQuery 1.0 i XPath 2.0 imaju isti model podataka, šta više XQuery je napravljen tako da je svaki ispravan XPath izraz ujedno ispravan izraz i u XQuery jeziku. Izraz putanje pronalazi čvor tako što ga traži u hijerarhiji stabla. Najkraća putanja je najčešće i najefikasnija s obzirom na to da se za svaki element u putanji pretražuje odgovarajući indeks. Putanja se sastoji od niza koraka razdvojenih kosom crtom (/) ili dvostrukom kosom crtom (//).

Sledećim izrazom:

```
doc('knjige.xml')//zapis/knjiga
```

u fajlu `knjige.xml` pronalaze se svi elementi `knjiga` koji su direktni potomci elemenata `zapis`, a koji su pak potomci korenog čvora, ne obavezno direktni.

FLWOR izraz se sastoji od sledećih pet delova:

- *for* - navodi se element koji se želi pretražiti
- *let* - koristi se da se definiše privremena promenljiva koja će biti aktivna u tekućoj iteraciji
- *where* - navodi se uslov za izdvajanje određenih elemenata od interesa
- *order by* - navodi se redosled u kome treba da budu navedeni rezultati
- *return* - navodi se struktura rezultata.

Deo *return* je obavezan, dok su svi ostali delovi opcioni. Moguće je napraviti nove XML elemente pomoću XQuery jezika koristeći FLWOR izraze.

Ako je dat XML element `knjige`, u fajlu `knjige.xml`, sa sadržajem:

```
<knjige>
  <knjiga>
    <naslov>Autostoperski vodič kroz galaksiju</naslov>
    <autor>Daglas Adams<autor>
    <opis>naučna fantastika</opis>
    <tip>meki povez</tip>
```

```
    <godina_izdanja>1977</godina_izdanja>
    <cena>1000</cena>
  </knjiga>
  <knjiga>
    <naslov>Najveća predstava na Zemlji</naslov>
    <autor>Richard Dawkings<autor>
    <opis>popularna nauka</opis>
    <tip>meki povez</tip>
    <godina_izdanja>2009</godina_izdanja>
    <cena>1500</cena>
  </knjiga>
</knjige>
```

primer FLWOR izraza kojim se reformatira prethodni fajl je:

```
<knjige>
{
  for $knjiga in doc('knjige.xml')/knjige/knjiga
  let $naslov := $knjiga/naslov/text()
  let $cena := $knjiga/cena/text()
  where $knjiga/godina_izdanja/text() lt 2000
  order by $naslov
  return
      <knjiga>
        <naslov>$naslov</naslov>
        <cena>$cena</cena>
      <knjiga>
}
</knjige>
```

FLWOR izrazom se za svaki element `knjiga` iz fajla `knjige.xml`, izdvajaju naslov i cena i smeštaju u pomoćne promenljive, a za one knjige koje su izdate pre 2000. godine, formiraju se novi elementi `knjiga` koji se sastoje samo od elemenata `naslov` i `cena` rezultujućih čvorova. Rezultat se smešta u element `knjige`.

Element koji nastaje izvršavanjem prethodnog izraza je:

```
<knjige>
  <knjiga>
    <naslov>Autostoperski vodič kroz galaksiju</naslov>
```

```
    <cena>1000</cena>
  <knjiga>
</knjige>
```

XQuery podržava standardni skup aritmetičkih operatora, operatora poređenja i logičkih operatora. Ima biblioteku ugrađenih funkcija, a moguće je napisati i sopstvene funkcije. XQuery trenutno podržava širok skup funkcionalnosti, kao što su uslovni izrazi pomoću if-then-else konstrukcije, switch naredbe, kvantifikujući izrazi (some, every), try-catch izrazi, mape i nizovi.

XQuery Update proširenjem može se promeniti sadržaj XML dokumenta. Instrukcije koje su dostupne su insert, delete, replace i rename. Na ovaj način se dodaju novi podaci kao što su elementi, atributi, menjaju ili brišu njihove vrednosti, ili brišu celi elementi.

Više informacija o XQuery jeziku može se naći u [228] i [127].

2.4 Baze podataka za rad sa tekstom

XML dokument se može smatrati bazom podataka u širem smislu. On predstavlja kolekciju podataka, opisuje strukturu i tipove podataka, prenosiv je, može se predstaviti kao stablo ili graf. Ako se poredi sa sistemom za upravljanje bazama podataka, i tu, zajedno sa tehnologijama koje ga okružuju, može ponuditi metod za skladištenje podataka (XML dokument), sheme za definisanje pravila koja podaci treba da ispunjavaju (DTD, XML Shema), upitne jezike (XQuery, XPath), interfejs za pristup iz drugih aplikacija (SAX, DOM) i drugo. Sa druge strane, XML nema mehanizam podrške za transakcije, oporavljanja od grešaka, višekorisnički pristup, okidače i sve ostalo što standardni sistem za upravljanje bazama podataka može da ponudi.

Kada se radi sa velikom količinom podataka, znatno je veća potreba za sistemom koji će moći da upravlja svim aspektima koji su važni u takvom okruženju. Iz tog razloga postoje i XML baze podataka koje se koriste za skladištenje i rad sa velikom količinom informacija u XML formatu. Kako se upotreba XML formata povećava, tako je i potreba za ovakvim bazama podataka veća.

Postoje dve vrste XML baza podataka:

- baze podataka sa podrškom za XML (XML Enabled Database)
- izvorne XML baze podataka (Native XML Database NXD)

Baze podataka sa podrškom za XML su tradicionalne baze podataka, na primer, relacione, koje imaju dodato proširenje za rad sa XML podacima. Rad se zasniva na tabelama sa vrstama i kolonama. Podrška za rad sa XML podacima se sastoji u omogućavanju da se XML dokument smešta na poseban način, kao Character Large Object (CLOB) ili da se sadržaj XML dokumenta mapira u relacionu strukturu (engl. *shredding*). Takvi podaci se prilikom korišćenja vraćaju u prvobitno stanje tako da se sa njima radi kao sa izvornim tekstom u XML formatu. Neke od takvih baza su DB2 ([112]), Informix ([116]), MySQL ([191]), Oracle ([208]), SQL Server ([255]), PostgreSQL ([222]) i MonetDB/SQL ([185]).

NXD su baze podataka posebno napravljene za rad sa XML podacima. Osnovna jedinica za smeštanje podataka su XML dokumenti. One definišu logički model za smeštanje i rad sa XML dokumentima. Ta definicija ne obuhvata ograničenja koja se mogu odnositi na fizički model podataka, koji može biti, na primer, relacioni, hijerarhijski ili objektno-orjentisani. NXD se zasnivaju na kolekcijama i XML dokumentima, umesto na shemama i tabelama. Za formulisanje upita koriste se XPath izrazi. Neke od NXD baza su eXist ([68]), MarkLogic Server ([166]), Xindice ([313]) i MonetDB/XQuery ([184]).

U zavisnosti od tipa XML dokumenta, data-centric ili document-centric, i vrste posla koji se obavlja sa podacima, bira se odgovarajući tip baze. Razlika između ove dve vrste dokumenata u praksi nije uvek jasna. Ne mora da bude pravilo, ali neka vodilja prilikom odabira odgovarajućeg sistema može biti da su za prvu situaciju pogodnije baze podataka sa podrškom za XML, dok su za drugu situaciju pogodnije izvorne XML baze podataka.

NXD mogu biti bazirane na tekstu (engl. *text-based*) ili bazirane na modelu (engl. *model-based*). Text-based su one baze koje podatke čuvaju kao tekstualne fajlove. Zajedničko takvim bazama je da imaju prilagođene indekse koji omogućavaju efikasno pronalaženje podataka bilo gde u dokumentu. Podaci su smešteni sekvencijalno tako da se može dobiti čitav dokument, ili neki njegov deo, jednostavno imajući u vidu da se podaci nalaze na uzastopnim lokacijama. Sa druge strane, ako bi podaci iz dokumenta bili rastavljeni i čuvani u relacionoj bazi podataka, bilo bi potrebno sastaviti ih, kombinovati indekse i potraživati ih iz više zahteva ka memoriji.

Drugi tip NXD je model-based koji, umesto da čuva dokument kao takav, pravi objektni model na osnovu strukture dokumenta i čuva ga u tom obliku. Ovakve NXD mogu jednostavnije da rade sa podacima za koje je potrebno da se pretvore u Document Object Model (DOM) stabla.

Oba modela, text-based i model-based, prilagođena su za rad sa specifičnim oblikom podataka, stoga, kada treba transformisati podatke u drugi oblik, njihova efikasnost očekivano opada.

Većina NXD podržava XML upitne jezike za pretragu, ažuriranje i brisanje, transakcije, zaključavanja, konkurentnost, programski interfejs (API), pristup za izvršavanje upita i dobijanje rezultata putem HTTP protokola.

NXD podržavaju indekse, koji mogu biti indeksi po vrednosti, indeksi po strukturi i full-text indeksi. Prva vrsta indeksa se koristi kada je potrebno da se pronađe podatak na osnovu vrednosti nekog elemenata ili atributa. Druga vrsta indeksa se koristi kada je potrebno pronaći elemente po poziciji. Treća vrsta indeksa indeksira tokene iz teksta i koristi se kada se u tekstu pretražuje konkretna reč, nezavisno da li je u nekom elementu i da li se nalazi u nekoj strukturi.

Više o XML bazama podataka može se naći u [309], [158] i [25].

eXist-db

Ovde su navedene ukratko glavne osobine skladištenja, indeksiranja i izvršavanja upita na arhitekturi eXist-db baze podataka ([68]). eXist-db je sistem otvorenog koda, besplatan i pokriva većinu osobina izvornih XML baza podataka, zbog čega je izabran kao dobar reprezentativni primer NXD baze podataka.

Baza eXist-db je napisana u Java programskom jeziku i može se koristiti kao:

- samostalni serverski proces koji dozvoljava pristup putem HTTP i XML-RPC protokola
- servlet koji radi u okviru veb aplikacije i ima direktan pristup bazi putem XML-RPC, SOAP i WebDAV protokola
- sistem koji je ugrađen u aplikaciju koja ima direktan pristup bazi podataka preko XML:DB API.

Hypertext Transfer Protocol (HTTP) je protokol za komunikaciju između klijenta i servera, a namenjen je za prenos hiperteksta preko veba. XML Remote Procedure Call (XML-RPC) je protokol za udaljene pozive procedura koji koristi HTTP protokol za transfer podataka i XML za enkodiranje poruka. Simple Object Access Protocol (SOAP) je protokol koji se uveliko koristi za XML zasnovane protokole za razvoj veb servisa. WebDAV je HTTP zasnovan protokol koji omogućava distribuirano rukovanje podacima.

Baza eXist-db skladišti XML dokumente u hijerarhijski organizovane kolekcije. Osnovni mehanizam pretrage kolekcija je putem izraza koji koriste XPath sintaksu. Baza omogućava korišćenje indeksa kojima se može ubrzati pretraga. Više je prilagođena radu sa dokument-centric XML fajlovima nego sa data-centric fajlovima. Dokument-centric fajlovi se obično sastoje od veće količine teksta, u kome se nalaze i po neke označene informacije XML etiketama.

XML jezici za postavljanje upita kao što su XPath, Quilt i XQuery koriste izraze u kojima se navode putanje putem kojih se dolazi do konkretnih elemenata podataka od interesa. XML dokument se može predstaviti strukturom stabla, stoga je izraz putanje pogodan jer može opisati putanju do bilo kog čvora u tom stablu. Obilazak čvorova u ovakvoj strukturi može biti vrlo neefikasan za velike dokumente.

Na primer, za putanju:

```
/knjiga//figura/naslov
```

potrebno je naći sve čvorove `knjiga`, sve njihove potčvorove `figura`, ne striktno direktne, i sve njihove direktne potčvorove `naslov`.

Da bi se to izračunalo, potrebno je pretražiti i mnoge čvorove koji ne zadovoljavaju ova svojstva. Stoga, uvode se indeksi kojima se vrši efikasnija pretraga strukture stabla. Za pretragu po vrednosti sa uspehom se koriste indeksi poput heš indeksa ili B stabla sa vrednostima. Kod strukturalnih izdvajanja, kao što je slučaj u prethodnom primeru, drugačiji pristup je poželjan. Potrebno je na brz način identifikovati strukturalne relacije između čvorova kao što su dete-roditelj ili predak-potomak. Dokument se tada pretražuje samo u posebnim slučajevima, kada indeks ne sadrži informacije po kojima se vrši pretraga. Baza eXist-db podržava full-text indekse, indekse zasnovane na n-gramima i indekse opsega, bazirane na biblioteci Lucene ([171]).

eXist-db implementira i proširenje XQuery Update za ažuriranje dokumenata. Ovaj mehanizam se odnosi na trajne objekte, odnosno može se ažurirati dokument koji postoji u bazi, ali se ne može izmeniti privremeni sadržaj koji je nastao izvršavanjem nekog upita.

Više o bazi eXist-db može se naći u radu [175] i knjizi [246].

Glava 3

Upravljanje prostornim podacima

3.1 Uvod

Prostorni podaci su veoma česti u našem okruženju. Prva pomisao na prostorne podatke su geoprostorni podaci - to su oni podaci koji se odnose na informacije o površini planete Zemlje. Slike iz satelita su istaknuti primer ovakvih podataka. Prostorni podaci mogu biti i slivovi reka, mreža puteva, granice među državama, oblasti zelenih površina, lokacije gradova, pozorišta, muzeja, raspored nameštaja u prostoriji ili elemenata na stranici knjige. Ljudsko telo se takođe ponaša kao izvor prostornih podataka. Da bi se izdvojila informacija iz slike poput ljudskog tela, podaci moraju da budu obrađeni poštujući prostorni okvir izvora. Na primer, medicinski trodimenzionalni modeli sadrže prostorne podatke i odnose u referentnom sistemu čiji je okvir ljudsko telo.

Podaci koji naizgled nisu prostorni, ali imaju pridruženu i prostornu dimenziju, takođe se mogu smatrati prostornim podacima. To mogu biti, na primer, temperatura ili meteorološke prilike predstavljene u prostoru. Ukoliko se temperatura izražava kroz vreme, onda su to prostorno-vremenski podaci. Takvi podaci mogu biti kvantitativni (vrednost temperature) ili kvalitativni (sunčano, vedro). Podaci se dele na rasterske i vektorske, u zavisnosti od toga da li su dati skupom kontinualnih ili diskretnih podataka.

Istorijat Osamdesetih godina prošlog veka svaki veliki proizvođač softvera imao je svoj format za čuvanje prostornih podataka. Standardni skup prostornih podataka i operacija koji se koristi u širokoj industriji, pod nazivom OpenGIS, prvi je napravio Open Geospatial Consortium (OGC) ([207]), osnovan krajem devedesetih

godina prošlog veka. Početkom ovog veka napravljena je biblioteka Geospatial Data Abstraction Layer (GDAL) ([85]) koja omogućava prevođenje prostornih podataka u različite formate. Najrasprostranjenije korišćeni formati za vektorske podatke su Geographic Markup Language (GML) ([82]), Keyhole Markup Language (KML) ([131]), CityGML ([36]), GeoJSON ([83]), Well Known Text (WKT) ([308]), ESRI Shapefile (Environmental Systems Research Institute) ([64]), dok su za rasterske podatke na raspolaganju formati kao što su GeoTIFF, ESRI Ascii, SAGA GIS, IDRISI i netCDF ([134]).

Prostorni podaci mogu da sadrže znanje koje bez prostorne dimenzije teško da bi moglo da bude otkriveno ([29]). Jedan istorijski primer iz 1855. godine vezan je za pojavu kolere u oblasti Londona, kada je jedan od epidemiologa mapirao sve lokacija odakle dolaze oboleli i otkrio da se grupišu oko pumpi za vodu ([249]). Nakon što su vlasti isključile pumpe za vodu, bolest je počela da jenjava. Kasnije je naučnim metodama potvrđena povezanost vode sa ovom bolešću. Drugi primer je teorija Gondwanaland da su svi kontinenti nekada bili jedno kopno na osnovu oblika i položaja kontinenata ([261]). Pronalaskom fosilnih ostataka i njihovim proučavanjem kasnije je otkrivena povezanost koja podržava ovu hipotezu. Otkrivanjem implicitnog znanja sadržanog u prostornim podacima bavi se oblast prostorno istraživanje podataka (Spatial Data Mining SDM) ([65], [245]).

3.2 Predstavljanje podataka u GML formatu

GML je jezik za obeležavanje geografskih podataka. To je XML gramatika koju je definisao OGC, napisana u XML shemi za modelovanje, prenos i skladištenje geografskih informacija. Teme koje se prirodno vezuju za GML su transformacija GML podataka, prostorni upiti, geografske analize, razvoj GML mapa, baze podataka zasnovane na GML-u, kao i mnoge aplikacije za mobilne računarske sisteme. GML se odnosi prema prostornim podacima u skladu sa njihovom prirodom, samim tim njegovom pojavom znatno se olakšava rad sa takvom vrstom podataka.

Pogodnost GML-a je, pored ostalih, i u tome što može da se koristi za prenos podataka između različitih aplikacija. On može da bude nosilac raznorodnih informacija. Stručnjaci za biljni svet, na primer, mogu da prikupe podatke o šumama, drveću i drugim biljkama. Odeljenje za životnu sredinu može da prikupi informacije o životinjama i njihovim navikama. Razvojno odeljenje može da ima informacije o stanovništvu i postojećim objektima u izgrađenom okruženju. GML se koristi za in-

tegrisanje svih ovih specifičnih podataka u jedan sistem, koji će sadržati obuhvatniju sliku o posmatranoj oblasti.

On je samo način da se predstavi neki geografski sadržaj, stoga je potrebno da se upotrebi drugi pogodan alat koji može da interpretira njegov sadržaj na način pogodan za vizuelno predstavljanje. Za stvaranje mape, potrebno je transformisati GML elemente u oblik koji je moguće prikazati u nekom grafičkom okruženju ili u veb pretraživaču. Pogodni grafički formati su Scalable Vector Graphics (SVG) ([73]), koji je krajem prošlog veka razvio World Wide Web Consortium (W3C) ([311]), i Extensible 3D (X3D) ([48]) koji je razvio Web3D Consortium ([307]).

GML se zasniva na apstraktnom geografskom modelu. Njime se opisuje svet u terminima geografskih entiteta (engl. *features*). Ovi entiteti su zapravo spisak osobina i geometrija. Osobine mogu predstavljati ime, tip ili vrednost, dok geometrije predstavljaju prostorne podatke poput tačaka, linija i poligona.

Kao jezik, GML se sastoji iz dva dela - sheme koja definiše skrukturu dokumenta i sam GML dokument koji sadrži konkretne podatke. Postojanje sheme omogućava da se definišu generički geografski podaci i da se potom koriste u oznakama koje su prilagođene konkretnom dokumentu. Na primer, ako se definiše reka kao element tipa linije, u dokumentu se svuda može pozivati na reku, umesto na liniju koja na prvi pogled ne daje informaciju o tome šta ta linija predstavlja. Razvojni tim GML-a imao je zamisao da zajednica učestvuje u definisanju raznih shema aplikacije koje bi bile dostupne za korišćenje. Ovo je oblast koja se aktivno razvija, a postoji i lista javno dostupnih shema ([88]).

GML se bazira na XML-u 1.0, koji se bazira na pojmu dokumenta. U GML-u se koristi kolekcija entiteta kao osnova svakog dokumenta. Kolekcija entiteta sastoji se od skupa svih GML objekata sa omotačem (engl. *envelope*) koji prostorno obuhvata sve te objekte, zatim od skupa osobina objekata i liste definicija prostornog referentnog sistema. Ona može sadržati pojedinačne entitete ali i kolekcije entiteta. Na primer, kolekcija entiteta može da bude Evropa, koja sadrži kolekcije entiteta svih država koje se nalaze u njoj. Kolekcije država mogu da sadrže entitete reka, puteva, gradova i raznih objekata. Ovako opisane kolekcije entiteta jednostavno se mogu integrisati, što pojednostavljuje rad u slučaju velikog broja kolekcija.

Entitet predstavlja opis konkretnog objekta, pre nego same geometrije tog objekta. On može da sadrži razne osobine objekta. Na primer, reka je entitet kojim se predstavlja realna pojava sa svim informacijama o njoj, kao što su ime reke, tip reke, geometrija i pritoke. Nije dovoljno izdvojiti samo geometriju, jer u stvarnosti

reka se ne odnosi samo na koordinate preko kojih se prostire.

Entitet može sadržati više geometrija, od kojih se bira ona koja odgovara uglu posmatranja. Na primer, nekada je dovoljno grad predstaviti tačkom, dok je u nekim slučajevima potrebno da se predstavi regionom koji zauzima. Stoga, entitet grad može imati za geometrije i tačku i poligon. Ovakav element geometrije nije tipa multigeometrije, jer se grad ne sastoji od jedne tačke i jednog poligona, već su to razdvojene geometrije od kojih se uzima jedna u zavisnosti od potrebe aplikacije. GML obezbeđuje i definisanje podrazumevane geometrije u slučaju da nije navedeno koja geometrija najviše odgovara.

GML shema

U ovom odeljku opisana su neka opšta pravila, strukture i zavisnosti iz GML shema, pri čemu se posmatra GML verzija 3.2.1. Elementi koji su ovde opisani su deo shema iz prostora imena "gml" ([82]).

Osnovna komponenta za pravljenje GML objekata definisana je na sledeći način:

```
<element name="AbstractObject" abstract="true"/>

<complexType name="AbstractGMLType" abstract="true">
  <sequence>
    <group ref="gml:StandardObjectProperties"/>
  </sequence>
  <attribute ref="gml:id" use="required"/>
</complexType>

<element name="AbstractGML" type="gml:AbstractGMLType" abstract="true"
  substitutionGroup="gml:AbstractObject"/>
```

Apstraktni element `gml:AbstractGML` je bilo koji GML objekat koji ima identitet. Može da se zameni, na primer, elementima tipa GML entiteta ili drugim objektima. On je natklasa svih GML objekata. U kombinaciji sa tipom `gml:AbstractGMLType` predstavlja osnovni šablon koji se koristi u GML shemama.

Shema kojom su opisani tipovi i elementi iz modela entiteta je *features.xsd* i deo je GML-a. Osnovni model entiteta dat je tipom `gml:AbstractFeatureType`:

```
<complexType name="AbstractFeatureType" abstract="true">
  <complexContent>
```

```
    <extension base="gml:AbstractGMLType">
      <sequence>
        <element ref="gml:boundedBy" minOccurs="0"/>
        <element ref="gml:location" minOccurs="0"/>
      </sequence>
    </extension>
  </complexContent>
</complexType>

<element name="AbstractFeature" type="gml:AbstractFeatureType"
  abstract="true" substitutionGroup="gml:AbstractGML"/>
```

Modelu elementa tipa `gml:AbstractFeatureType` dodaju se dve osobine pogodne za geografske entitete u odnosu na model definisan tipom `gml:AbstractGMLType`. Element `gml:boundedBy` sadrži vrednost koja predstavlja omotač koji obuhvata celu instancu konkretnog entiteta i prvenstveno se koristi za brzo pretraživanje entiteta koji se nalaze na određenoj lokaciji.

Sledećim elementom opisuje se minimalni pravougaonik koji sadrži ceo entitet:

```
<element name="boundedBy" type="gml:BoundingShapeType"/>
  <complexType name="BoundingShapeType">
    <sequence>
      <choice>
        <element ref="gml:Envelope"/>
        <element ref="gml:Null"/>
      </choice>
    </sequence>
  </complexType>
</element>
```

Element `gml:Envelope` sadrži par pozicija kojima su određeni gornji levi i donji desni ugao omotača.

Osnovne geometrije U ovom odeljku su opisani osnovni modeli geometrija u GML-u koje se nalaze u XML shemama *geometryBasic0d1d.xsd* i *geometryBasic2d.xsd*. Bilo koji geometrijski element koji ima nasleđenu semantiku `gml:AbstractGeometryType` može se posmatrati kao skup pozicija. Sve pozicije treba da se odnose na neki re-

ferentni koordinatni sistem. Geometrijski model razlikuje tri klase elemenata - geometrijske primitive, agregacije i složene elemente.

Geometrijske primitive su otvorene, odnosno nemaju tačke koje bi se smatrale granicom. Stoga, krive ne sadrže svoje krajnje tačke, površi ne sadrže svoje granične krive i tela ne sadrže svoje granične površi.

```
<complexType name="AbstractGeometryType" abstract="true">
  <complexContent>
    <extension base="gml:AbstractGMLType">
      <attributeGroup ref="gml:SRSReferenceGroup"/>
    </extension>
  </complexContent>
</complexType>

<element name="AbstractGeometry" type="gml:AbstractGeometryType"
  abstract="true" substitutionGroup="gml:AbstractGML"/>
```

Svi geometrijski tipovi se izvode iz ovog apstraktnog nadtipa. Svaki geometrijski element ima `gml:gid`, `gml:name`, `gml:description`. Takođe, svaki element geometrije (element koji je geometrijskog tipa) treba da bude izveden iz grupe `AbstractGML`.

Geometrijske primitive Geometrijske primitive mogu biti jednostavne *0*-dimenzione i *1*-dimenzione, kao što su tačka, kriva i linija (kao specijalni tip krive) i *2*-dimenzione, kao što su površi i pologoni (kao specijalne vrste površi).

Apstraktni tip za sve geometrijske primitive je `gml:AbstractGeometricPrimitiveType`. Geometrijska primitiva je geometrijski objekat koji se ne može razložiti na podobjekte. Sve primitive su orjentisane, što je određeno redosledom njihovih koordinata.

```
<complexType name="AbstractGeometricPrimitiveType" abstract="true">
  <complexContent>
    <extension base="gml:AbstractGeometryType"/>
  </complexContent>
</complexType>
```

Element `AbstractGeometricPrimitive` je nadgrupa za sve predefinisane geometrijske primitive ili one koje korisnik sam definiše.

```
<element name="AbstractGeometricPrimitive"
  type="gml:AbstractGeometricPrimitiveType" abstract="true"
  substitutionGroup="gml:AbstractGeometry"/>
```

Tačka se definiše kao jedan par koordinata i proširenje je prethodnog apstraktog tipa.

```
<complexType name="PointType">
  <complexContent>
    <extension base="gml:AbstractGeometricPrimitiveType">
      <sequence>
        <choice>
          <element ref="gml:pos"/>
          <element ref="gml:coordinates"/>
        </choice>
      </sequence>
    </extension>
  </complexContent>
</complexType>

<element name="Point" type="gml:PointType"
  substitutionGroup="gml:AbstractGeometricPrimitive"/>
```

Apstrakcija krive podržava različite nivoe složenosti. Kriva se može posmatrati kao geometrijska primitiva samo ako je povezana.

```
<complexType name="AbstractCurveType" abstract="true">
  <complexContent>
    <extension base="gml:AbstractGeometricPrimitiveType"/>
  </complexContent>
</complexType>

<element name="AbstractCurve" type="gml:AbstractCurveType" abstract="true"
  substitutionGroup="gml:AbstractGeometricPrimitive"/>
```

Linije su posebna vrsta krivih koje se sastoje iz jednog dela, sa linearnom interpolacijom, odnosno susedne tačke se povezuju dužima. GML podržava više načina da se predstave koordinate linije, a to su preko `gml:pos`, `gml:pointProperty`, `gml:coord` ili `gml:posList` i `gml:coordinates`.

```
<complexType name="LineStringType">
  <complexContent>
    <extension base="gml:AbstractCurveType">
      <sequence>
        <choice>
          <choice minOccurs="2" maxOccurs="unbounded">
            <element ref="gml:pos"/>
            <element ref="gml:pointProperty"/>
            <element ref="gml:pointRep"/>
          </choice>
          <element ref="gml:posList"/>
          <element ref="gml:coordinates"/>
        </choice>
      </sequence>
    </extension>
  </complexContent>
</complexType>

<element name="LineString" type="gml:LineStringType"
  substitutionGroup="gml:AbstractCurve"/>
```

Površ se definiše kao *2*-dimenziona geometrijska primitiva. Kao i kod krivih, definiše se apstrakcija površi kojom se omogućavaju različiti nivoi složenosti. Poligon je posebna vrsta površi koja se sastoji od jednog dela. Granica poligona je u jednoj ravni i koristi se ravanska interpolacija za formiranje unutrašnjosti.

Složene geometrije, kompozicije i agregacije Geometrije se mogu formirati od više geometrijskih primitiva. Geometrijski složeni elementi su zatvoreni skupovi geometrijskih primitiva, odnosno oni sadrže i njihove granice. Unutrašnjosti ovih primitiva su disjunktne. Postoje dva načina objedinjavanja geometrijskih elemenata.

Geometrijske kompozicije su elementi koji predstavljaju geometrijske složene elemente sa jednom osnovnom geometrijom koja je izomorfna nekoj primitivi. Na primer, kriva kao kompozicija je skup kriva čija se geometrija može videti kao kriva. Elementi ovog tipa su `gml:CompositeCurve`, `gml:CompositeSurface` i `gml:CompositeSolid`.

Geometrijske agregacije su proizvoljni skupovi geometrijskih elemenata. Oni nemaju nikakvu dodatnu strukturu, samo postoje da prikupe druge elemente u jedan

skup. Neki od njih su `gml:MultyPoint`, `gml:MultyLineString`, `gml:MultyPolygon` i `gml:MultyGeometry`.

3.3 Rad sa prostornim podacima

Prostorni podaci su svuda oko nas i veoma često su uključeni u brojne dileme, kako u svakodnevnim, tako i u specifičnim situacijama. Neki od načina da se organizuje prostor su: topologija (susednost), mreže (najkraće rastojanje), pravac (severno) i euklidski (razdaljina).

Primeri upita koji uključuju i prostornu dimenziju su:

- Gde je najbliža knjižara?
- Da li na putu do kuće ima apoteke?
- Da li je bilo pacijenata sa sličnim uslovima na slici magnetne rezonance?
- Da li je topologija aminokiselina biosintetičkog gena genoma nađena u još nekoj sekvenci u bazi?
- Ako znamo trend kupovina u sledećoj godini, gde bi bilo najbolje napraviti skladište i prodavnice?
- Kako bi trebalo proširiti mrežu puteva da bi se smanjile gužve u saobraćaju?
- Gde se nalazi osoba kojoj je potrebna pomoć? Koji je najbolji put do nje?

Često se upit formuliše u neodređenim jedinicama, na primer "Ne sećam se koliko je daleko taj restoran. Kada sam u blizini, mogu da se prisetim sa koje je strane ulice i sigurna sam da je pored parka.". U ovom primeru igraju ulogu: razdaljina, određivanje pravca i orjentacije, kao i topološki odnosi (susednost, povezanost, pripadnost).

Model prostornih podataka je skup pravila pomoću kojih se identifikuju i predstavljaju objekti koji se odnose na prostor.

Objektni model se koristi za predstavljanje entiteta koji imaju neki oblik. Na primer, jezero se može predstaviti kao dvodimenzionalni region, reka kao jednodimenziona kriva, gnezdo kao tačka, mreža puteva kao skup jednodimenzionih krivih.

Objektni model je konceptualni i mapiran je u računaru koristeći vektorske strukture podataka. Ove strukture mapiraju regione u poligone, linije u polilinije i tačke u tačke.

Model polja koristi se da predstavi kontinualan ili bezoblični koncept, na primer, temperaturu ili oblake. Polje je funkcija koja preslikava referisani prostorni okvir u domen za taj atribut. Operacije polja mogu da budu lokalne, fokalne i zonske.

Lokalne operacije su one kod kojih vrednost polja na nekoj lokaciji zavisi samo od ulazne vrednosti te lokacije (na primer, preslikavanje lokacije u vrstu drveta {bor, jela, hrast}). Fokalne operacije su one kod kojih vrednost polja na nekoj lokaciji zavisi i od ulaznih vrednosti na lokacijama u bližoj okolini (na primer, izračunavanje nadmorske visine). Gradijent tekuće lokacije zavisi od vrednosti u veoma bliskim lokacijama. Zonske operacije se prirodno vezuju za agregatne operatore (na primer, operacija izračunavanja srednje visine drveća za svaku od vrsta drveta).

Relacije među podacima mogu biti skupovno orjentisane (unija, presek, sadržavanje, članstvo), topološke (dodirivanje, unutar, preklapanje), zasnovane na pravcu (apsolutno - globalni referentni sistem, severno, južno; objektno-relaciono - orjentacija datog objekta, levo, desno; sa tačke gledišta posmatrača - posebno napravljen referentni objekat), metričke (funkcija razdaljine uvodi topologiju, pa je svaki metrički prostor ujedno i topološki), euklidske (posmatra se vektorski prostor sa dve operacije: sabiranje i množenje skalarom).

Operacije koje se definišu prema modelu koji definiše OGC podeljene su u tri kategorije ([107]):

- osnovne operacije - primenjuju se na sve geometrijske tipove (SpatialReference, Envelope, Export, IsEmpty, IsSimple, Boundary)
- operacije koje ispituju topološke relacije između objekata (Equal, Disjoint, Intersect, Touch, Cross, Within, Contains, Overlap)
- opšte operacije za prostorne analize (Distance, Buffer, ConvexHull, Intersection, Union, Difference, SymmDiff).

Istraživanje podataka, kao sistematična pretraga za potencijalno korisnim informacijama ugnježdenim u digitalnim podacima, značajno je kako za akademska istraživanja tako i za industriju. Mnoge korporacije uviđaju da iz informacija koje se skupljaju godinama može da se pronađe nešto korisno što će poboljšati njihov rad. Umesto da predstavljaju problem za skladištenje, podaci su sada izvor dobiti.

U klasičnim analizama podataka, pretpostavlja se da su podaci nezavisno generisani. Kod prostornih podataka je ova pretpostavka sasvim pogrešna. Prostorni podaci su u skladu sa Toblerovim prvim zakonom geografije ([277]): sve zavisi od svega, ali stvari koje su bliske više zavise jedne od drugih, nego one koje su udaljene. Oblast statistike posvećena analizama prostornih podataka naziva se prostorna autokorelacija (Spatial Autocorrelation) ([95]).

Među tehnikama istraživanja prostornih podataka posebno mesto pripada i klasifikaciji. Cilj klasifikacije je da se odredi vrednost nekog atributa relacije na osnovu vrednosti ostalih atributa te relacije. Primer je određivanje pogodne lokacije za život ljudi na osnovu blizine reke, stepena nezagađenosti vazduha, klime u toku godine i drugih relevantnih parametara. U tematskoj klasifikaciji je, na primer, cilj da se kategorizuju pikseli satelitske slike u klase (voda, naselje, šuma). Kod ovakve klasifikacije postoji velika prostorna povezanost, pa pikseli koji su susedni na slici, neretko pripadaju istoj klasi.

Važno je i pronaći male skupove podataka koji se posmatraju kao šum, greška, devijacija ili izuzetak. Ovi skupovi se identifikuju tako što nisu konzistentni sa ostalim skupom podataka. Njihovo pronalaženje može dovesti do otkrivanja neočekivanog znanja ([243]).

3.4 Prostorne baze podataka

Prostorni podaci su, u opštem slučaju, složeniji od neprostornih podataka, stoga oni zahtevaju drugačiji pristup skladištenju, organizaciji i pretrazi.

Uprkos velikom uspehu, ogroman broj današnjih DBMS nije prilagođen da upravlja prostornim podacima, nije ni blizu efikasnog rada sa njima. Relativno jednostavan prostorni upit "Naći sve kupce koji se nalaze u krugu od 50 kilometara od centra kompanije" bio bi "pretežak" za tradicionalnu bazu. Zahtevao bi transformaciju centra kompanije i adrese kupaca u referentni sistem, na primer, geografsku dužinu i širinu, a zatim sekvencijalnu pretragu, bez pomoći višedimenzionih indeksa.

Proučavanje sistema organizacije prostornih baza podataka (Spatial Database Management System SDBMS) vodi ka razvoju modela i algoritama koji efikasno rukuju podacima.

Glavni podstrek za istraživanje SDBMS dolazi od aplikacija poput geografskih informacionih sistema (Geographic Information System GIS), projektovanja uz pomoć računara (Computer Aided Design CAD), informacionih sistema za multime-

diju (Multimedia Information Systems MIS) i skladištenja podataka i proučavanja planete Zemlje od strane NASA.

Primeri komercijalnih proizvoda za rad sa prostornim podacima su proširenja objektno-relacionih servera baza podataka mnogih proizvođača kao što su: Intergraph, Autodesk, Oracle, IBM i Informix. U okviru istraživačkih projekata napravljena su proširenja prostornom funkcionalnošću u sistemima za upravljanje bazama podataka PostgreSQL ([222]), GeO2, Paradise ([55]) i drugim ([244]). Funkcionalnost koja je omogućena ovim prototipima uključuje skupove prostornih podataka kao što su tačka, linija, poligon i skup prostornih operacija poput preseka, zatvorenja i razdaljine. Prostorni tipovi i operacije mogu da budu deo objektno-relacionog upitnog jezika kao što je SQL. Poboljšanje performansi ovih sistema uključuje višedimenzione prostorne indekse i algoritme za prostorne metode, prostorne upite i prostorna spajanja. Upotpunjavanje tradicionalnih baza podataka prostornim podacima zahteva rešavanje netrivijalnih zadataka na različitim nivoima - od prostorne algebre, preko prostornih indeksa i vizuelne obrade upita, do novih pristupa kontroli konkurentnosti, bezbednosti i oporavku.

Primeri baza podataka koje imaju proširenje za rad sa prostornim podacima su GeoMesa ([111], [84]), AllegroGraph ([5], [72]), IBM DB2, IBM Informix, Microsoft SQL Server, MySQL i PostgreSQL.

Detaljniji pregled oblasti prostornih baza podataka može se naći u [154], [243] i [134].

PostgreSQL i PostGIS

PostgreSQL ([222]) je objektno-relaciona baza podataka koja je jedan od prvih i značajnijih sistema koji je uveo specijalizovane operacije za rad sa prostornim podacima, kao što su efikasno skladištenje, prilagođen sistem indeksiranja i upitni jezik koji je u skladu sa prirodom prostornih podataka. Ovaj sistem je otvorenog koda i slobodan, što ga čini prijemčivim za korišćenje u široj zajednici, ali i u profesionalnim okruženjima. Prostorno proširenje PostgreSQL baze naziva se PostGIS ([220]).

Radi boljeg uvida u mogućnosti koje pruža PostGIS, ovde su istaknute osnovne karakteristike prema njegovoj specifikaciji ([221]).

PostGIS definiše tipove kao što su *geometry*, *geography* i *box* tipove. Oni se najjednostavnije mogu opisati na sledeći način:

- *box2d* - tip koji se sastoji od podataka xmin, ymin, xmax, ymax i obično se koristi za 2D omotač geometrije
- *box3d* - tip koji se sastoji od podataka xmin, ymin, zmin, xmax, ymax, zmax, obično se koristi za 3D omotač geometrije
- *geometry* - tip za objekat koji se prostire u ravni i predstavlja objekat u Euklidskom koordinatnom sistemu
- *geography* - tip za objekat koji se prostire u sferi i predstavlja objekat u sfernom koordinatnom sistemu.

Tip *geography* je više prilagođen geografskim podacima, a koordinate se zapisuju longitudama i latitudama. Za razliku od tipa *geometry*, u kome je osnova ravan, kod geografskog tipa osnova je sfera, pa su i izračunavanja složenija jer se i rastojanja računaju u skladu sa time. Posledično, postoji manji broj funkcija koje rade sa tipom *geography* nego sa tipom *geometry*. Tokom vremena se razvijaju sve naprednije metode kojima se omogućava implementacija i složenijih funkcija.

Dodavanje ili brisanje kolone koja se odnosi na geometriju u određenu tabelu može se uraditi uz pomoć funkcije *AddGeometryColumn*, odnosno *DropGeometryColumn*.

Postoje brojne funkcije kojima konvertuje objekat zapisan u drugom formatu (GML, KML, JSON, WKB, WKT), u objekat tipa *geometry*.

Primer stvaranja *geometry* objekta od GML objekta bi bio:

```
SELECT     ST_GeomFromGML(
  '<gml:LineString srsName="EPSG:4269">
    <gml:coordinates>
      44.789081,20.393382 44.820991,20.456553 44.823913,20.495692
    </gml:coordinates>
  </gml:LineString>'
);
```

PostGIS ima brojne funkcije kojima se prave i drugi objekti i izvršava konverzija između objekata kada to ima smisla.

Mogu se izračunati tip objekta, zatvorenje, dimenzija, krajnja tačka, da li je objekat poligon, da li je kolekcija, da li je prazan, da li je zatvoren, da li sadrži prsten, minimalna koordinata i maksimalna koordinata objekta. Mogu se dodavati

ili brisati tačke iz objekta, izračunavati afine transformacije, konvertovati u drugu dimenziju (manju ili veću), spajati objekti, izdvajati samo objekti datog tipa (iz multigeometrije), homogenizovati ili normalizovati kolekcije. Postoje funkcije koje izračunavaju najkraću duž u prostoru koja spaja dva objekta, najbližu tačku i brojne druge funkcije.

Tip *geography* treba koristiti kada dolaze do izražaja sferne osobine podataka, ali po cenu složenosti izračunavanja i manjeg skupa dostupnih funkcija. Sa druge strane, ukoliko je planarna prezentacija dovoljna, efikasnije je koristiti tip *geometry*.

Neki od prostornih operatora definisanih u PostGIS-u su:

- `&&` - vraća TRUE ako prvi objekat preseca drugi objekat
- `&&&` - vraća TRUE ako n-dimenzionalni *bounding-box* prvog objekta preseca n-dimenzionalni *bounding-box* drugog objekta
- `&<` - ispituje da li prvi objekat preseca drugi objekat ili se nalazi sa njegove leve strane
- `&<|` - ispituje da li prvi objekat preseca drugi objekat ili se nalazi ispod njega
- `<<` - da li je *bounding-box* prvog objekta striktno levo od *bounding-box*-a drugog objekta
- `<<|` - da li je *bounding-box* prvog objekta striktno ispod *bounding-box*-a drugog objekta
- `~` - da li *bounding-box* prvog objekta sadrži *bounding-box* drugog objekta
- `<->` - vraća 2D razdaljinu dva objekta
- `<#>` - vraća razdaljinu između *bounding-box*-eva dva objekta

Grupa koja razvija PostGIS planira da podrži i poboljša, u skladu sa standardima OpenGIS i SQL/MM, podršku za niz važnih GIS funkcionalnosti naprednih topoloških operacija kao što su pokrivanja, površine, mreže, poboljšanje korisničkog interfejsa za pregled i izmenu podataka i veb alata za pristup.

Glava 4

Obrada teksta na prirodnom jeziku

4.1 Uvod

Obrada prirodnog jezika (Natural Language Processing NLP) bavi se prepoznavanjem govora, razumevanjem teksta na prirodnom jeziku, stvaranjem teksta na prirodnom jeziku, ekstrakcijom informacija iz teksta, klasifikacijom teksta, sumarizacijom teksta i brojnim drugim zadacima. NLP pripada oblastima lingvistike, računarskih nauka, istraživanja podataka, veštačke inteligencije i smatra se težim zadatkom u oblasti računarskih nauka zbog složenih osobina prirodnog jezika.

Pravila po kojima se informacije stvaraju i prosleđuju nisu jednostavna za razumevanje od strane računara. Ona mogu biti od veoma apstraktnih, kao na primer sarkastične opaske, do vrlo jednostavnih, na primer to da se vlastita imena pišu početnim velikim slovom. Jezik je formiran desetinama hiljada godina unazad i sadrži suptilne detalje i nosioce informacija na različitim nivoima. Čovek, generalno, može uspešno da koristi taj sistem, dok se za mašinu taj zadatak ispostavlja kao težak.

Prirodni jezik se sastoji iz nekoliko slojeva, a to su fonologija, morfologija, leksika, sintaksa, semantika, diskurs i pragmatika ([156]).

Fonologija je deo nauke o jeziku koja se bavi proučavanjem karakteristika i značenja glasova koji se koriste u govoru.

Morfologija je takođe deo nauke o jeziku, a bavi se proučavanjem unutrašnje strukture reči, njihovim vrstama, oblicima i građenjem.

Leksika je skup svih reči jednog jezika i podložna je promeni, u skladu sa živim promenama jezika. Lingvistička disciplina koja se bavi proučavanjem leksike zove se leksikologija.

Sintaksa kao deo nauke o jeziku bavi se strukturom rečenice. Sintaksna analiza

u računarskoj lingvistici obuhvata lematizaciju, morfološku segmentaciju, razdvajanje na reči, označavanje vrsta reči, parsiranje, rastavljanje na rečenice, steming i izdvajanje terminologije.

Semantika se u okviru računarske lingvistike odnosi na proučavanje mogućih značenja koje tekst nosi. To je teži deo obrade i uključuje primenu algoritama za identifikovanje značenja reči i rečenica, kao i načina na koji su rečenice oformljene.

Diskurs se u lingvističkom smislu može opisati kao vezani sled rečenica ili izjava, smisaono povezana celina teksta veća od rečenice. U lingvistici je najveća jedinica proučavanja rečenica, mada se u novije vreme razvija i lingvistika teksta. Analizirajući rečenice prelazi se iz oblasti lingvistike u oblast gde je jezik sredstvo komunikacije i gde diskurs postaje njegova forma izražajnosti.

Pragmatika se bavi svrsishodnom upotrebom jezika u komunikaciji, čije se značenje može interpretirati isključivo poznavajući odgovarajući kontekst. To zahteva šire znanje o svetu i svim stvarima koje nas okružuju, uključujući namere, planove i ciljeve.

Čovek koristi sve slojeve u procesu razumevanja prirodnog jezika. Da bi mašina mogla sa istom efikasnošću da razume jezik, potrebno je da koristi sve izvore znanja koje koristi i čovek.

Istorijat Obrada prirodnih jezika je počela da se razvija četrdesetih godina prošlog veka, a prvi zadaci su bili iz domena mašinskog prevođenja. Alen Turing je 1950. godine formulisao test koji je imao za cilj da ispita "da li mašina može da razmišlja" ([283]), poznat i kao Tjuringov test. Sastojao se od razgovora u kome su učestvovala dva čoveka i jedna mašina. Jedan čovek je bio sa jedne strane i vodio je pisani dijalog sa čovekom i mašinom sa druge strane. Ukoliko prvi čovek ne može da prepozna sa druge strane ko je čovek, a ko mašina, smatra se da je mašina prošla test. Ovo se smatra i početkom intenzivnije računarske obrade teksta.

Godine 1957. Noam Chomsky objavljuje univerzalnu gramatiku za uspostavljanje sintaksičkih struktura koje važe u prirodnom jeziku ([32]). Ideja se bazira na tome da čovek ima urođeni smisao za gramatiku i formiranje jezičkih struktura. Znanje o načinu na koji se formira i koristi jezik već postoji u nama, ali je potrebno ovladavati njime uz vežbu, baš kao i u slučaju hodanja. Tokom narednog perioda postojao je veliki entuzijazam za razvijanje potpune automatizacije mašinskog prevođenja koje bi bilo u rangu sa prevođenjem od strane čoveka. Mislilo se da je to moguće postići u roku od nekoliko godina. Više projekata se bavilo ovom tematikom,

ali sa manjim uspehom od onog koji je bio predviđan nakon prvobitno obećavajućih rezultata. Ispostavilo se da nije dovoljno opisati samo sintaksu, već je potrebno posmatrati i semantički koncept ([156]).

Početni optimizam je uticao na to da se dosta novca uloži u istraživanja u oblasti NLP. Međutim, kako nije bilo zadovoljavajućih rezultata, interesovanje onih koji bi finansirali istraživanja počelo je da opada.

Do osamdesetih godina prošlog veka bilo je različitih sistema i uglavnom su svi koristili pristup zasnovan na formulisanju pravila (Rule-Based RB). U tom periodu počinje da se koristi i mašinsko učenje (Machine Learning ML) za obradu prirodnog jezika, kao neka vrsta revolucije u svetu NLP. Statističke metode, na kojima se zasnivalo mašinsko učenje za obradu prirodnih jezika, vratile su značaj oblasti NLP.

Devedesetih godina prošlog veka uvode se rekurentne neuronske mreže (Recurrent Neural Network RNN) za obradu teksta na prirodnom jeziku koje i danas važe za aktuelnu reč tehnologije u oblasti NLP.

U drugoj deceniji ovog veka u širu upotrebu ulaze duboke neuronske mreže (Deep Neural Network DNN). Predstavljanje reči vektorom i njenim ugrađivanjem u vektorski prostor (engl. *word embedding*) je metoda koja je dala posebno dobre rezultate i aktuelna je oblast istraživanja. Jedan od ciljeva koji se želi postići je razumevanje upita na prirodnom jeziku i njihovo formulisanje u nekom strukturnom jeziku, kao na primer SQL-u. To je znatno složeniji zadatak od pojedinačnih zadataka poput tokenizacije, lematizacije ili određivanja vrsta reči.

Zadaci Neki od zadataka NLP su leksička semantika (pronalaženje značenja reči), mašinsko prevođenje (Machine Translation MT), razumevanje prirodnog jezika (Natural Language Understanding NLU), generisanje prirodnog jezika (Natural Language Generation NLG), prepoznavanje imenovanih entiteta (Named Entity Recognition NER), povezivanje imenovanih entiteta (Named Entity Linking NEL, poznat i kao Named Entity Disambiguation NED), razrešavanje višeznačnosti reči na osnovu konteksta (Word Sense Disambiguation WSD), klasifikacija teksta (Text Classification TC), ekstrakcija informacija (Information Extraction IE), ekstrakcija relacija (Relation Extraction RE), anotacija dokumenata (Document Annotation DA), pretraživanje teksta (Information Retrieval IR), istraživanje teksta (Text Minning TM), analiza emocija (Sentiment Analysis SA), prepoznavanje tema (Topics Detection TD), automatska sumarizacija teksta (Text Summarization TS), razrešavanje koreferenci (Coreference Resolution CR), prepoznavanje govora (Speech Recogniti-

on SR), konverzija teksta u govor (Text to Speech TTS) i odgovaranje na pitanja (Question Answering QA).

4.2 Metode obrade teksta na prirodnom jeziku

Obradi teksta može se pristupiti na više načina. U istraživanjima su najzastupljenije metode zasnovane na pravilima i na mašinskom učenju.

Metode zasnovane na pravilima

Jedan pristup NLP koji koristi semantičke metode zasnovan je na pravilima koja se formulišu na osnovu znanja o jeziku i znanja iz domena koji se obrađuje. Ove metode koriste unapred pripremljene dodatne resurse poput elektronskih rečnika, gramatika, baza podataka sa rečima sličnim po značenju i vezama između reči. Sistemi zasnovani na znanju uglavnom koriste modele konačnih stanja kao što su konačni automati, konačni transduktori i rekurzivne mreže prelaza.

Modeli konačnih stanja koji su zasnovani na pravilima dugo su se koristili u lingvistici zbog svojih dobrih lingvističkih i računarskih osobina. Sa lingvističke strane, njima se mogu opisati najznačajnije osobine jezika. Sa strane računarske obrade, modeli konačnih stanja se najviše koriste zato što su efikasni u smislu vremena i prostora potrebnih za izračunavanja. Pristup zasnovan na gramatikama i pisanju pravila pretpostavlja da je čovek uključen u taj proces. Najveća prednost je što čovek može pratiti izvršavanje upita i može pronaći greške koje pri tome nastaju. Pravila mogu biti napisana tako da budu proširiva da obuhvate nove slučajeve ili da budu izmenljiva u slučaju dobijanja novog znanja.

Najveća mana metoda zasnovanih na pravilima je da zahtevaju obučene eksperte koji bi ih osmislili i ručno napisali. Pravila treba da budu razvijana, ali i isprobavana sve vreme. Tokom vremena mogu postati vrlo komplikovana, teško razumljiva i izmenljiva, a neretko mogu biti i nekonzistentna ([304])

Metode zanovane na mašinskom učenju

Drugi pristup zasnovan je na metodama mašinskog učenja. Mašinskim učenjem se može smatrati učenje kroz iskustvo koje se sprovodi automatski. Metode mogu biti nadgledane, polunadgledane, nenadgledane i zasnovane na učenju sa podrškom.

Nadgledano učenje je najzastupljeniji vid učenja. Svakom ulaznom podatku x pridružuje se željena izlazna vrednost y koju algoritam treba da predvidi. Na osnovu uređenih parova (x, y) treba pronaći optimalnu funkciju koja preslikava ulaz u izlaz. Funkcija koja odgovara realnim podacima nije poznata, tako da se metodama pokušava optimizacija funkcije koja je pretpostavka i koja se naziva hipoteza. Model učenja koji se primenjuje određuje opšti oblik te funkcije, odnosno prostor mogućih hipoteza. Konkretne vrednosti podataka koje se koriste pri obučavanju određuju tačni oblik funkcije, odnosno najbolju hipotezu u prostoru hipoteza. Kod statističkih metoda potrebno je ručno odrediti koji će faktori sa ulaza uticati na izlaz, dok se kod metoda zasnovanim na mrežama ti parametri izračunavaju u samoj mreži. Oni predstavljaju atribute ili svojstva (engl. *features*), tako da se svaki podatak posmatra kao vektor svojih atributa:

$$x = (x_1, x_2, x_3, ..., x_n)$$

Dobar model će imati sposobnost generalizacije, odnosno davaće dobre rezultate i nad podacima koje nije video u procesu treniranja.

Vrednost y koja se traži može biti kontinualna (problem regresije), diskretna (problem klasifikacije, binarne ili višeklasne) i strukturirana (problem strukturne predikcije). Regresija je problem koji se bavi predviđanjem vrednosti nekom narednom ulazu na osnovu prethodnih ulaza i izlaza. Na primer, predviđanje temperature vode ili cene nekretnina. Klasifikacija je problem pridruživanja ulaza nekoj od predefinisanih klasa. Problem strukturne predikcije je predviđanje strukture podatka koja sadrži veze među svojim konstituentima. Primer je mašinsko prevođenje.

Nenadgledano učenje koristi samo ulazne podatke bez znanja o pridruženom željenom izlazu, pa je potrebno pronaći pravilnost u podacima. Tipični zadaci su klasterovanje (deljenje ulaza na podskupove tako da su instance unutar jednog skupa sličnije nego instance iz različitih skupova) i smanjenje dimenzionalnosti (pronalaženje manjeg skupa parametara koji opisuju glavne osobine instanci).

Novija istraživanja sve više koriste nenadgledane ili polu-nadgledane algoritme učenja. Takvi algoritmi mogu da nauče osobine iz podataka koji nisu bili prethodno ručno označeni željenim odgovorima. U opštem slučaju, takav zadatak je znatno teži od nadgledanog učenja i obično daje manje tačne rezultate za istu količinu ulaznih podataka. Danas je na raspolaganju izuzetno velika količina neoznačenih podataka, uključujući, između ostalog, i čitav sadržaj veba, koji često mogu poboljšati inicijalno lošije rezultate.

Učenje sa podrškom se koristi u obučavanju računarskih agenata koji deluju u određenom prostoru akcija. Učenje se vrši na osnovu mogućih akcija i skupa signala podrške. Signal podrške stiže tek na kraju nekog skupa akcija i može biti pozitivan ili negativan. Potrebno je da se utvrdi koja tačno akcija je dovela do takvog rezultata za pridruženi signal podrške, i da se u skladu sa time promeni ponašanje agenta. Ovakav pristup se koristi kod problema za koje je veoma teško ili nemoguće navesti tačnu vrednost izlaza, na primer, kod igara kao što su šah ili go, različitih video igara i drugo.

Najzastupljnije metode mašinskog učenja su statističke i metode zasnovane na "mrežama".

Statističke metode Statističkim metodama u oblasti NLP pokušava se razumeti jezik bez eksplicitnog navođenja kako da se to uradi. To se sprovodi kroz sisteme koji analiziraju ulazni skup podataka u cilju formiranja znanja za izvođenje sopstvenih pravila. Ovaj pristup se zasniva na probabilističkim rezultatima i ne garantuje semantičku tačnost.

Najveća prednost korišćenja statističkih metoda je da mogu da "nauče" neke veze koje važe u podacima bez pisanja uputstava koje zahteva veliku kompetenciju. Za nadgledane metode je potrebno unapred anotirati korpus, ali za to često nije potrebna previše zahtevna obuka. Mašinsko učenje statističkim metodama daje dobre rezultate u oblastima kao što su klasifikacija dokumenata i klasterovanje reči. U tim zadacima, u većini slučajeva, postoji veliki broj ponavljanja iz kojih se statističkim metodama mogu naučiti postojeće zakonitosti. U praksi, mašinsko učenje može znatno ubrzati rešavanje konkretnog zadatka, ali uz pretpostavku da je dostupan veliki korpus anotiranih dokumenata, što sa druge strane i nije tako jednostavno.

Neke od statističkih metoda mašinskog učenja koje se koriste u obradi teksta su ([132], [4]): naivni Bajes (Naive Bayes NB), stabla odlučivanja (Decision Trees DT), potporni vektori (Support Vector Machines SVM), modeli maksimalne entropije (Maximum Entropy Models ME), k najbližih suseda (kNN), skriveni markovljevi modeli (Hidden Markov Models HMM) i uslovna slučajna polja (Conditional Random Fields CRF). Za IE se pretežno koriste HMM i CRF, dok se za TC koriste NB, DT, SVM i kNN.

Metode neuronskih mreža Neuronske mreže su porodica algoritama mašinskog učenja zasnovanih na mreži koja se može sastojati od jednog ili od više slojeva među-

sobno povezanih čvorova. Neuronske mreže koje se sastoje od više slojeva nazivaju se duboke neuronske mreže. Inspirisane su nervnim sistemom ljudi, gde su čvorovi posmatrani kao neuroni, a grane kao sinapse. Svaka grana ima pripadajuću težinu, a mreža definiše računska pravila za prosleđivanje podataka sa ulaznog sloja mreže na izlazni sloj. Uz odgovarajuće definisane funkcije mreže, mogu se obavljati različiti zadaci učenja minimiziranjem grešaka u odnosu na funkciju mreže. Povećanjem korpusa povećava se tačnost rezultata, tako da je kod ovih algoritama postojanje veoma velikog korpusa ključno za dobijanje znatno boljih rezultata.

U oblasti obrade teksta zastupljene su rekurentne, konvolutivne i rekurzivne neuronske mreže ([89]). Rekurentne mreže su pogodne za obradu teksta jer ispoljavaju dobre osobine pri radu sa kontinualnim nizom povezanih pojmova, kakav je tekst. Konvolutivne mreže se uglavnom koriste za obradu slika i videa, ali mogu biti od koristi i prilikom obrade teksta. Njihova prednost je u tome što su pogodne za pronalaženje obrazaca za identifikovanje fraza i idioma u tekstu. Rekurzivne mreže su pogodne za podatke koji se mogu predstaviti u vidu stabala, a rečenice mogu da se predstave u tom obliku tako da se listovima pridružuju reči od kojih se one sastoje.

Reprezentacija teksta u vidu retkog vektora reči iz vreće reči (engl. *bag-of-words*), koristila se u NLP zajednici veoma rano. Reprezentacija u vidu gustog vektora reči koji oslikava i semantičke koncepte i ugrađuje reč u prostor reči, donela je značajno poboljšanje u metodologiji 2013. godine kada je tim koji je predvodio Tomas Mikolov, iz kompanije Google, objavio svoja istraživanja ([177]). Mikolov i njegov tim su uveli metodu koja se može koristiti za učenje visoko kvalitetnih vektora iz skupa koji se broji u milijardama reči, od čega se rečnik različitih reči broji u milionima.

Word2vec ([90]) je grupa modela koji se koriste za pravljenje vektora koji oslikava semantičke veze između reči. Ovi modeli se sastoje od dva sloja neuronske mreže i služe da se rekonstruiše lingvistički kontekst reči. Oni koriste veoma veliki korpus ulaznog teksta za formiranje vektorskog prostora, čija je dimenzija obično nekoliko stotina, tako da se svakoj reči pridružuje vektor iz tog prostora. Vektori se formiraju tako da reči koje nisu slične po kontekstu imaju vektore koji nisu bliski u tom prostoru, i obrnuto. Na osnovu velikog korpusa dokumenata moguće je uhvatiti sličnosti između reči kao što su kralj-kraljica, muškarac-žena, kretanje-kretali, Beograd-Srbija, Rim-Italija. Ako je v(x) vektor reči x, tada je vektor v(Srbija) - v(Beograd) blizak vektoru v(Italija) - v(Rim).

Dva često korišćena osnovna oblika modela iz ove grupe su - kontinualna vreća

reči (Continuous Bag-of-Words Model CBOW) i takozvani kontinualni skip-grami (Continuous Skip-gram Model Skip-gram). Prvi model na osnovu reči u okolini predviđa tekuću reč, dok drugi model na osnovu tekuće reči predviđa skup reči koje je okružuju. Metode se koriste, na primer, za zadatak dopunjavanja nedostajuće reči u rečenici, ili za dovršavanje rečenice. U prethodnim modelima se za analize i predviđanja posmatra lokalni okvir reči. Primer modela koji koristi globalne vektore za predstavljanje šireg konteksta je GloVe ([218]). Ovim modelom se, na primer, bolje rešava problem identifikovanja analogija. Danas su razvijeni i mnogi drugi modeli prilagođeni specifičnijim zadacima.

Neuronskim mrežama se može uhvatiti značenje tek na osnovu veoma velike količine podataka. Prednost neuronskih mreža je u tome da njima potencijalno može da se trenira univerzalno znanje, jer su sposobne za obrađivanje velike količine neanotiranog teksta i za pronalaženje koncepata koji su u njima sadržani.

Uopšteni problem mašinskog učenja je potreba za većim skupom instanci u kojima se iskazuje konkretna zakonitost. Primer je automatsko prevođenje rečenica na prirodnom jeziku. Metode mašinskog učenja zahtevaju znatno veći korpus podataka kako bi programski sistem naučio pravila. Druga mana mašinskog učenja je ta da mala razlika na ulazu može značajno promeniti rezultate, a da čovek ne može da pronađe zbog čega je to tako.

Izbor metode zavisi od nekoliko činilaca: dostupnosti podataka za treniranje, dostupnosti jezičkih resursa, tehnoloških zahteva, zahtevanog nivoa efikasnosti i preciznosti sistema.

4.3 Specifični zadaci obrade teksta na prirodnom jeziku

Ekstrakcija informacija

Ekstrakcija informacija je zadatak automatskog izdvajanja strukturnih informacija iz nestrukturiranog ili polustrukturiranog dokumenta po određenim kriterijumima. Kriterijumi mogu biti definisani na najrazličitije načine, a uvek se odnose na sadržaj dokumenta. Ekstrakcija informacija se ranije uglavnom odnosila na ekstrakciju iz teksta na prirodnom jeziku, ali se u novije vreme može odnositi i na automatsku ekstrakciju sadržaja iz multimedijalnih dokumenata. U slučaju obrade

teksta, jedinica izdvajanja je deo teksta, na primer, određena reč, grupa reči, fraza ili konstrukcija, a može biti i ceo dokument.

Informacije koje se traže su raznovrsne. Moguće je tražiti neki šablon koji postoji u tekstu bez dodeljivanja značenja, ili podatke koji imaju specifično značenje. Primer jednog od zadataka je prepoznavanje imenovanih entiteta kao potklasa zadatka ekstrakcije informacija. On se bavi prepoznavanjem entiteta koji se odnose na kategorije podataka kao što su osobe, lokacije, događaji, datumi ili numerički izrazi. Povezivanje imenovanih entiteta je problem koji se bavi identifikovanjem imenovanih entiteta koji odgovaraju nekom konkretnom značenju. Pronalaženje koreferenci je problem koji se bavi identifikovanjem veza između entiteta takvih da se može zaključiti da se entiteti odnose jedni na druge. Ekstrakcija informacija se bavi i pronalaženjem relacija između entiteta, događaja i njihovih parametara (ko, gde, kada, šta), entiteta koji su relevantni za određeni domen, ili na ekstrakciju određenih struktura.

Informacije mogu biti različitih važnosti, pa je nekada potrebno izdvojiti i težinu ili vrednost. Na primer, kod izdvajanja veština iz biografije ili kod izdvajanja sastojaka iz kulinarskih recepata, neki elementi mogu biti bitniji od drugih.

U opštem slučaju, postoji potreba za nekim informacijama i pretpostavka je da postoji skup tekstova u kojima se nalaze te informacije. Tekstovi su vrlo često nestrukturirani, samim tim i znatno teži za računarsku obradu u smislu formulisanja direktnih upita kojima bi se dobile informacije od značaja. Ovakvih tekstova vrlo često ima u velikom broju, pa je ručno pronalaženje potrebnih informacija u prihvatljivom roku praktično neizvodljivo. Postojanje opisane problematike, uz praktične potrebe za njenim rešavanjem, dovelo je do značajnog razvoja oblasti ekstrakcije informacija.

Istorijat Razvoj oblasti ekstrakcije informacija počeo je sedamdesetih godina prošlog veka i tekao je zajedno sa razvojem obrade prirodnih jezika. Jedan od prvih komercijalnih sistema je JASPER ([8]), koji je služio za obradu vesti o finansijama. Domeni koji su u to vreme imali koristi od razvoja ove oblasti su odbrana od terorizma, mikroelektronika, obrada novinskih članaka o promenama u menadžmentu i obrada satelitskih izveštaja.

Razvojem veba, koji je započeo krajem osamdesetih godina prošlog veka, i njegovim globalnim korišćenjem, sve je više dokumenata sa raznovrsnim sadržajem. Za efikasno pretraživanje takvih dokumenata potrebno je dodeliti im strukturne seman-

tičke podatke koji se zatim mogu jednostavno pretraživati. Najčešća organizacija je da se metodama obrade prirodnih jezika obrađuju dokumenti u cilju traženja željenih informacija, a zatim da se izdvojene informacije čuvaju u nekoj bazi podataka.

Aktuelne oblasti u kojima se koristi ekstrakcija informacija su prevođenje, sistemi za postavljanje pitanja i pronalaženje odgovora, razumevanje teksta na prirodnom jeziku, generisanje teksta na prirodnom jeziku, sumarizacija teksta, digitalni asistent (na primer, Siri ([247]), kompanije Apple), robotika, digitalne biblioteke, obrada pravnih akata, medicinskih zapisa, onlajn vesti, vladinih dokumenata, raznih izveštaja, pravljenje (ažuriranje) kalendara, interakcije na društvenim mrežama, pronalaženje referenci u naučnim radovima i praćenje medija o društvenim pojavama.

Diskusija o metodama za ekstrakciju informacija U radu [31] prikazano je relevantno poređenje aktuelnosti metoda zasnovanih na pravilima sa metodama mašinskog učenja za zadatak ekstrakcije informacija. Prema izvedenim istraživanjima, ispostavlja se da su, u periodu od 2003. godine do 2012. godine, metode zasnovane samo na pravilima zastupljene u tek 3.5% naučnih radova, metode zasnovane na mašinskom učenju zastupljene su u 75% radova, dok je 21% radova koristilo hibridnu metodu. Sa druge strane, veliki proizvođači u svojim alatima za ekstrakciju informacija koristili su pretežno metode zasnovane na pravilima - 67%, zatim metode zasnovane na mašinskom učenju - 17% i metode koje koriste hibridni pristup - u podjednakom odnosu, 17%. Najveće kompanije, kao što su IBM, SAP i Microsoft za ovaj zadatak koriste metode potpuno zasnovane na pravilima.

Objašnjenje manje popularnosti metoda zasnovanih na pravilima u krugu istraživačke zajednice nalazi se u formulacijama poput "prezahtevno je i skupo u kontekstu vremena", kao i "nije previše praktično" ([314]). Sa druge strane, prenebregava se činjenica da je korišćenje mašinskog učenja za zadatak ekstrakcije informacija takođe veoma zahtevno. Da bi se dobili dobri rezultati na praktičnim problemima iz realnog sveta potrebno je definisati problem striktno matematičkim jezikom, razumeti u velikoj meri modele kako bi se izabrali odgovarajući, napraviti izbor karakteristika od značaja koji su u saglasnosti sa izabranim modelima i, na kraju, obezbediti skup označenih podataka koji treba da bude znatno veći od skupa potrebnog za metode zasnovane na pravilima, kako se navodi u ovom radu. Sa stanovišta resursa, kao što su potreban hardver i vreme za izvršavanje, glas ponovo ide u prilog sistemima zasnovanim na pravilima ([30]).

Drugo moguće objašnjenje nalazi se u tome da se ne vidi preterani izazov u istraživanju modela zasnovanih na pravilima, jer ta oblast deluje dovoljno istražena.

Dolazi se do situacije u kojoj je ekstrakcija informacija sve zastupljenija u industriji, pretežno se koriste modeli zasnovani na pravilima sa rešenjima razvijenim samo za te potrebe, dok se, sa druge strane, izvode istraživanja na polju ekstrakcije informacija metodama mašinskog učenja nesrazmerna potrebama tržišta. Rešenje može da se ogleda u saradnji ove dve istraživačke oblasti, definisanju principa, odnosno standardizaciji formulisanja pravila i razvoju metoda mašinskog učenja koje bi učenje uskladili sa novim, prilagođenim, načinom izražavanja pravila.

Istraživanja Neki od programskih sistema koji se koriste za ekstrakciju informacija iz teksta na prirodnim jezicima su GATE ([81]), Unitex ([215]), OpenNLP ([9]), Mallet ([163]), DBpedia Spotlight ([50]), biblioteke pisane za Python programski jezik NLTK ([21]), Stanford NER ([164]), spaCy ([251]) i SrpNER ([145]).

Napravljeno je poređenje sistema Stanford NER, spaCy, NLTK i LingPipe za zadatak ekstrakcije osoba, lokacija i organizacija, na slučaju engleskog jezika ([121]). Stanford NER je baziran na metodi slučajnih polja. spaCy je biblioteka koja je implementirana u programskom jeziku Python, bez objavljenih detalja o metodama koje se koriste. LingPipe je alat koji koristi metode zasnovane na pravilima u kombinaciji sa nadgledanim statističkim metodama mašinskog učenja. NLTK je alat koji koristi metod maksimalne entropije iz familije nadgledanih metoda mašinskog učenja. Za evaluaciju je korišćen WikiGold korpus ([14]). Izračunata su dva tipa mera sličnosti, jedna je potpuno poklapanje i druga delimično poklapanje. Najuspešniji je bio Stanford NER sa ukupnom F merom 0.7075 za potpuno poklapanje i 0.7609 za delimično poklapanje. Testiran je i hibridni metod, koji kombinuje Stanford NER i spaCy, uz dodatak specijalno napravljenog rečnika sa određenim brojem reči iz klasa osoba, lokacija i organizacija. U slučaju konflikta, hibridni metod daje prednost rečniku. Ovaj metod je dostigao ukupnu F meru 0.7976 kada je sličnost računata kao potpuno poklapanje, dok je F mera 0.8742 kada je sličnost računata kao delimično poklapanje.

U radu [159] opisan je alat za označavanje osoba u hrvatskom i slovenačkom jeziku, koji postiže F meru 0.84, što je uporedivo sa Stanford NER sistemom, dok sa određenim varijacijama dostiže F meru 0.92, što je bolje od rezultata koji se dobijaju pomoću Stanford NER ([232]).

Sprovedeno je istraživanje u kome se porede sistemi za ekstrakciju imenovanih

entiteta Stanford NER, spaCy i SrpNER za srpski jezik ([232]). Napravljena su dva skupa tekstova sa po četiri metode za poređenje imenovanih entiteta. Za potrebe istraživanja razvijen je zlatni standard za prepoznavanje osoba u odnosu na koji su testirani prethodni programi za slučaj srpskog jezika. Nakon prilagođavanja oblika ulaznih podataka za svaki od programa i ujednačavanja izlaznih podataka za potrebe poređenja, rezultat je da je SrpNER postigao najbolju preciznost (osim u jednom slučaju), Stanford NER je postigao najbolji odziv, dok je SrpNER postigao najbolju F meru u svim slučajevima, koja se kretala u rangu od 0.802 do 0.877.

Sa druge strane, u radu [162] istraživane su mogućnosti ekstrakcije informacija uz pomoć nadgledanih i polunadgledanih metoda za ekstrakciju ključnih reči metodama neuronskih mreža. Razvijen je troslojni hijerarhijski model koji se sastoji od (1) sloja koji koristi metode ugrađivanja karaktera, reči i izabranih lingvističkih karakteristika, (2) LSTM (Long Short-Term Memory) sloja za označavanje sekvenci i (3) CRF sloja za modelovanje veza između različitih izlaza iz prethodnog sloja. Evaluacija je izvedena nad 500 novinskih članaka, tako da je po jedan paragraf iz svakog članka ručno označen. Trening skup se sastojao od 350 članaka, 50 članaka je služilo za unapređivanje, dok se test skup sastojao od 100 članaka. Za zadatak klasifikacije postignuta je najbolja F mera 52.1 na skupu za razvoj i 46.6 na skupu za testiranje, dok je za zadatak ekstrakcije postignuta najbolja F mera 57.6.

Detaljniji pregled oblasti ekstrakcije informacija može se naći u [41], [233], [66] i [250].

Klasifikacija teksta

Klasifikacija teksta se može opisati kao problem pridruživanja predefinisanih kategorija tekstu koji se obrađuje. Klasifikatori se mogu koristiti da organizuju i kategorizuju tekstove iz najrazličitijih domena. Mogu se organizovati novinski članci prema oblasti, stručni tekstovi prema temama koje obrađuju, medicinski izveštaji prema relevantnosti dijagnoze, različite vrste recenzija prema polaritetu - pozitivno ili negativno, radovi po autorstvu ili iz elektronske pošte izdvojiti neželjena pošta (engl. *spam*).

Klasifikacija može biti binarna (detekcija neželjene pošte, sentiment analiza recenzija) ili višeklasna (od liste predefinisanih klasa bira se jedna, klasifikovanje članaka prema tematici, klasifikovanje filmova prema žanru). Može se odnositi na katalogizaciju (digitalne biblioteke u kojima se podaci razvrstavaju po metapodacima kao što su autor, datum, naslov, kao i po temama za kasnije pretraživanje kataloga),

stvaranje tezaurusa (skupa povezanih termina za kontrolisane rečnike koji sadrže sinonime i šire ili uže termine i nisu hijerarhijski organizovani), taksonomije (organizovanje skupa informacija na različite načine u zavisnosti od svrhe, na primer organizovanje vrsta u biologiji) i ontologije (logičkog okvira kategorija informacija u specifičnom domenu, uključujući sheme i dijagrame za predstavljanje povezanosti).

Koristi od automatske klasifikacije teksta mogu se videti u slučajevima kada postoji veliki korpus nestrukturiranih tekstova za koji bi čoveku bilo potrebno mnogo vremena da ih pročita i obradi. Može se koristiti kao dodatni alat kojim se proverava klasifikacija napravljena od strane čoveka kako bi se potvrdila tačnost ili ispoljile neke osobine koje navode da možda postoji potreba da se revidira neka doneta odluka. Automatska klasifikacija se primenjuje i za analize u realnom vremenu kod kritičnih situacija kada je potrebno dobiti rezultate što je pre moguće.

Istorijat U literaturi se klasifikacija teksta prvi put pojavljuje šesdesetih godina prošlog veka, u radu [167]. Autor je uopštio problem klasifikacije na dva ključna zadatka - (1) pronalaženje relevantnih aspekata posmatranog teksta kojima bi se mogli uočiti odgovarajući indikatori i (2) korišćenje uočenih indikatora za pravilno predviđanje klase kojoj posmatrani tekst pripada.

Do kasnih osamdesetih godina prošlog veka, istraživanja iz oblasti TC uglavnom su se zasnivala na formulisanju pravila za klasifikaciju dokumenata. Uz ubrzani razvoj hardvera i pojavom sve većeg broja domena koji rade sa velikom količinom podataka, devedesetih godina prošlog veka, TC postaje jedna od vodećih podoblasti informacionih sistema. U tom periodu se razvija i oblast mašinskog učenja, sa dobrim rezultatima, tako da ona i u oblasti TC preuzima primat nad tehnologijama zasnovanim na pravilima ([240]).

Diskusija o metodama za klasifikaciju teksta Tekstualni dokumenti mogu sadržati mnoštvo informacija. Neke od njih imaju presudnog uticaja na rezultat klasifikacije, dok neke imaju malog uticaja ili nemaju uopšte. Uobičajeni problem kod klasifikacije teksta je taj što ovih informacija može biti veoma mnogo. Problem pronalaženja informacija koje imaju najviše uticaja na rezultat, kao i onih bez kojih bi klasifikacija bila podjednako uspešna, nije uvek jednostavan.

Kod pristupa zasnovanog na pravilima potrebno je da čovek osmisli i ručno napiše pravila kojima bi se izvršila klasifikacija. U zavisnosti od umešnosti onoga ko razvija

pravila, uglavnom ovakav pristup daje dobre rezultate, ali on zahteva puno analiza i testiranja.

Pristup zasnovan na mašinskom učenju podrazumeva da se napravi numerička reprezentacija dokumenta u obliku vektora. Tekstualni dokument d_j često se predstavlja kao vektor težina termova $\overrightarrow{d_j} = (w_{1j}, w_{2j}, ...w_{|T|j})$, gde je T skup termova koji se pojavljuju bar jednom u bar nekom dokumentu iz skupa za treniranje, dok $0 \leq w_{kj} \leq 1$, okvirno govoreći, označava veličinu koja se odnosi na uticaj terma t_k na značenje dokumenta d_j ([240]). Često korišćeni metod za predstavljanje dokumenta numeričkim vektorom je vreća reči (Bag-of-Words BOW), gde vektor predstavlja frekvencije reči iz predefinisanog rečnika reči.

Informacije kojima se predstavljaju dokumenti nazivaju se karakteristikama ili atributima. Široko je rasprostranjena metoda smanjenja broja atributa kako bi se smanjila dimenzionalnost podataka, uz uslov da nema gubitka informacija. Atribut može predstavljati način da se indeksira dokument (kada dobija vrednost 0 ili 1) ili da se termima dodeli odgovarajuća težina (kada dobija vrednost iz šireg skupa mogućih vrednosti)

U nekim slučajevima broj atributa može se smanjiti pravljenjem restrikcije nad polaznim tekstom po nekom odgovarajućem pravilu, tako da se značajno smanji obim teksta, ali da se ne izostave informacije od važnosti za problem koji se rešava.

Klasifikatori mogu raditi manje efikasno sa tekstovima u celini zbog prevelike dimenzionalnosti u odnosu na današnju raspoloživu tehnologiju. Potrebno je smanjiti dimenzionalnost problema, bez izostavljanja važnih informacija. Neki autori su istraživali i korišćenje fraza umesto reči na leksičkom i statističkom novou. Istraživanja su pokazala da je korišćenje statističkih osobina bolje kada se reči koriste zasebno, dok je korišćenje fraza bolje kada se leksičke i statističke osobine koriste u kombinaciji ([77]).

Istraživanja Smanjenje dimenzionalnosti je zadatak od ključne važnosti za poboljšanje efikasnosti i efektivnosti kod mnogih vrsta obrade informacija, uključujući i klasifikaciju tekstova. U radu [100] izvršeno je poređenje metoda smanjenja dimenzionalnosti. Metode koje su posmatrane su steming, računanje frekvencije dokumenta (DF), računanje frekvencije termova (TFIDF) i latentno semantičko indeksiranje (LSI). Rezultati su pokazali da poslednje tri metode daju znatno bolje rezultate u poređenju sa prve dve metode.

Izvršeno je poređenje metoda klasifikacije nad tekstovima o istraživanju tržišta

([101]), među kojima je bilo pet metoda zasnovanih na rečniku i pet metoda zasnovnih na mašinskom učenju. Metode koje su bile obuhvaćene istraživanjem su SVM, RF, NB, kNN, neuronske mreže i kombinacije lingvističkih metoda (Linguistic Inquiry and Word count LIWC). Metode su testirane na 41 skupu tekstova na različitim jezicima sa vodećih društvenih mreža. Za ovakvu vrstu tekstova ispostavlja se da su najbolje metode RF i NB, dok SVM gotovo ni u jednom slučaju nije imala najbolji rezultat, kao ni metode zasnovane na rečnicima. Istaknuti primer dobrog rezultata je metoda RF za zadatak određivanja polariteta sentimenta u odnosu na tri klase - pozitivno, neutralno, negativno.

Sa druge strane, SVM metoda se pokazuje kao veoma uspešna kada se radi sa zadacima sa većom dimenzionalnošću i uspešnija je u izvođenju generalizacija ([101]). U radu [103] izvršeno je poređenje SVM metode sa NB metodom, ali i njihovim poboljšanjem semantičkim metodama koje se sastoje od korišenja informacija iz Wikitology. Eksperimenti su izvedeni nad tekstovima 20 News Groups. Ovaj skup se sastoji od 20000 dokumenata novinskih članaka razvrstanih na 20 kategorija. Pokazuje se da se semantičkim metodama i SVM metodom poboljšava F mera od 0.865 za baseline do 0.92 za poboljšanje, dok se NB metodom poboljšava F mera 0.681 za baseline do 0.82 za poboljšanje.

Što se tiče metoda zanovanih na neuronskim mrežama za zadatak klasifikacije teksta, do danas je izvršen veliki broj istraživanja. Na primer, u radu [152] su korišćene rekurentne konvolucione neuronske mreže koje su dostigle tačnost 96.49 na skupu 20 News Groups za klasifikovanje novinskih članaka na kategorije politika, religija, računari i rekreacija.

Detaljniji pregled oblasti klasifikacije teksta može se naći u [199], [74], [124], [155] i [223].

Anotacija dokumenata

Anotacija dokumenata je pridruživanje određenih informacija dokumentu. Anotacija, u širem smislu, ne mora podrazumevati elektronske medije. Može se odnositi na anotaciju štampanih medija, na primer, zapisivanje komentara na marginama štampanih materijala. Odnosi se na pridruživanje metapodataka postojećem sadržaju u cilju poboljšanja razumevanja semantike podataka, pretrage ili dobijanja novih informacija. Metapodaci mogu biti umetnuti u sadržaj ili pridruženi dokumentu u nekom drugom obliku. Anotacija se može odnositi na tekstualne dokumente, ali i na dokumente poput slika, audio i video materijala.

Karakteristike koje su identifikovane kao važne za alat koji se bavi anotacijom su ([288]): da li je i na koji način prilagođen korisniku, da li koristi ontologije, koje formate dokumenata podržava, u kom formatu se zapisuju anotacije, kao i da li podržava automatizaciju anotacije i kojim metodama. U navedenom radu dat je paralelni pregled karakteristika izabranih alata za anotaciju dokumenta. Navedeni su i alati koji podržavaju automatizaciju anotacije, i za svaki od alata navedeno je kojim metodama se vrše analize, uključujuči poređenje stringova (sa ili bez korišćenja ontologije), označavanje vrsta reči, prepoznavanje imenovanih entiteta, prepoznavanje šablona, prepoznavanje po sličnosti i bit-mapiranu klasifikaciju.

Istorijat Sama upotrebe dokumenata stvara sa sobom i potrebu za njihovim označavanjem na neki način zbog boljeg uvida u sadržaj ili za potrebe organizacije po nekom zadatom kriterijumu. Stoga, koncept anotacije dokumenata nastaje veoma rano. Sa povećanjem broja računara koji su u upotrebi, osamdesetih godina prošlog veka, broja zadataka koji se izvršavaju elektronski, i posledično broja i raznovrsnosti dokumenata u elektronskoj formi, puno truda je uloženo u razvoj metodologija za povećanje mogućnosti i efikasnosti pretrage dokumenata automatskim dodavanjem novih informacija. Posebno, razvojem semantičkog veba početkom ovog veka, oblast automatske semantičke anotacije počinje ubrzano da se razvija sa osnovnom idejom da se dokumenti predstave semantičkim konceptima kako bi, uporedo razvijane, nove tehnologije mogle što bolje da odgovore sve većim i suptilnijim potrebama korisnika. Težnja za stalnim napredovanjem u ovoj oblasti dovodi do toga da je anotacija dokumenata aktuelno polje istraživanja i danas.

Diskusija o metodama za anotaciju dokumenata Problem anotacije dokumenata rešava se metodama koje mogu biti zasnovane na ručnoj anotaciji ili anotaciji uz pomoć dodatnih tehnologija. U opštem slučaju, anotacija dokumenata je složen i vremenski zahtevan zadatak. Ručna anotacija se može opravdati do nekog nivoa obima i složenosti, nakon čega postaje previše zahtevna i nedovoljno brza u odnosu na očekivanja i realne potrebe. Neke oblasti istraživanja koje se koriste za anotaciju, ili kao pomoć anotaciji, su prepoznavanje govora (Speech Recognition), ekstrakcija ključnih fraza (Key-Phrase Extraction), segmentacija po temi (Topical Segmentation), ručna anotacija (Manual Annotation) i semantička anotacija (Semantic Annotation) ([59]).

Korisno je uvesti automatizaciju, na primer, u vidu saradnje korisnika i sistema u

procesu anotacije dokumenata. Ekstrakcija informacija može se koristiti kao pomoć, bilo kao nenadgledano pridruživanje važnih informacija tekstu, ili kao asistirano pridruživanje u saradnji sa korisnikom - čovekom ([35]). Na primer, korisnik može definisati početna pravila za ekstrakciju, a pri svakoj narednoj anotaciji, mogu se primenjivati metode mašinskog učenja za retreniranje modela na osnovu dosadašnjih potvrđenih anotacija, kako bi se poboljšao skup pravila za ekstrakciju informacija.

Za anotiranje dokumenata koriste se različiti izvori znanja, na primer, kontrolisani rečnici (ili vokabulari, kao konačni skupovi reči), rečnici reči sa opisom značenja (engl. *glossary*, u obliku spiska reči i njihovih značenja), tezaurusi (spisak reči uz dodatnu semantiku, kao što su relacije između reči, sinonimi, antonimi), formalna i neformalna pridruživanja (je pripadnik neke klase, je živo biće, je ženski sat), okviri (ne postoji hijerarhija, već se razvrstavanje vrši po nekoj od pridruženih osobina, na primer, nameštaj napravljen od drveta) i opisi u vidu ograničenja vrednosti ili logičkih pravila. Znanja iz različitih izvora mogu se kombinovati u izgradnji ontologija na kojima bi se zasnivale anotacije dokumenata ([153]).

Primeri kontrolisanih rečnika koji se koriste za anotaciju metapodataka su MeSH (Medical Subject Headings) ([174]), Getty TGN (Thesaurus of Geographic Names) ([272]), i ostali rečnici razvijeni u Getty institutu ([271]). Primer jedne vrste uopštene ontologije je Dublin Core ([58]) koji se koristi kao skup smernica za anotiranje dokumenata. Što je ontologija striktnija u svojim pravilima i smernicama, to je jednostavnija za tumačenje i korišćenje ([39]).

Neke od NLP strategija koje se koriste za automatsku anotaciju dokumenta su Rapid Automatatic Keyword Extraction (RAKE) ([230]), KeyGraph ([204]), TextRank ([176]), Frequency ([173]), Term Frequency-Inverse Document Frequency (TF-IDF) ([282]), KEA ([310]), Latent Dirichlet Allocation (LDA) ([142], [282]), Topic Modeling (TM) ([120]) i kNN ([110], [279]).

Primer poređenja različitih pristupa automatske semantičke anotacije dokumenata, podeljene na četiri potproblema - ekstrakcija koncepta - entiteta (prepoznavanje entiteta korišćenjem baze znanja ([173]), n-grama, RAKE i LDA), aktivacija koncepta - izračunavanje mere pouzdanosti (statističke, hijerarhijske i metode zasnovane na grafovima), izbor anotacije - izbor na osnovu mere pouzdanosti (top-k i kNN) i evaluacija, može se naći u radu [96]. Pokazuje se da od mogućih kombinacija pomenutih strategija, kombinacija u kojoj učestvuju entiteti, metode zasnovane na grafovima i kNN daju rezultate sa najboljom F merom tačnosti nad posmatranim korpusom dokumenata iz oblasti ekonomije, politike i računarstva.

Istraživanja Veliki broj istraživanja se bavi korišćenjem IE i NER za automatsko proizvođenje metapodataka i za semantičku anotaciju. IE metode su bile primenjene na semantičku anotaciju na različitim domenima i jezicima, na primer, anotaciju video novosti na turskom jeziku, biomedicinskom domenu i arheologiji na engleskom jeziku ([76], [148], [297]). U radu [297] predstavljena je semantička anotacija u okviru sistema OPTIMA koja sprovodi NER i RE zadatke koristeći pristup zasnovan na pravilima, ontologijama i rečnicima reči iz domena. Značajan trud je uložen u istraživanje veba i njegove uloge u anotaciji i ekstrakciji entiteta, semantičko indeksiranje i pretragu različitih kolekcija, pristupom zasnovanim na pravilima i na metodama mašinskog učenja. Na primer, u domenu televizijskih i radio novosti, domenu semantičke anotacije samog sadržaja veba, istraživanja stava na osnovu multimedijalnih sadržaja, kao i ekstrakcije informacija iz tekstova mikroblogova ([23], [59], [170]). Još neke od platformi koje koriste automatsku anotaciju pristupima zasnovanim na pravilima i na mašinskom učenju su MultimediaN, AeroDAML, Armadillo, KIM, MnM, MUSE, Ont-O-Mat, SemTag i ONTEA ([151], [225], [238]).

Primer alata za anotiranje koji se zasniva na ontologijama je MnM ([291]), razvijen na Knowledge Media Institute, Open University, u Velikoj Britaniji. MnM omogućava ručnu, polu-automatsku i automatsku anotaciju tekstualnih dokumenata. Za automatsku anotaciju dokumenata koristi metode ekstrakcije informacija, ugrađene kao moduo Amilcare ([242]). Ovaj moduo na početku koristi sistem Annie ([47]), ugrađen u GATE ([47], [81]), za izvršavanje tokenizacije, razdvajanja na rečenice, označavanje vrsta reči, primenu elektronskih rečnika i ekstrakciju imenovanih entiteta. Na osnovu već postojećih anotacija, sistem uči kako da reprodukuje pravila za ekstrakciju informacija, pri čemu skup metapodataka definiše korisnik. Na primer, neka postoji klasa metapodataka "posetilac". Proces anotacije se odvija tako što korisnik u odabranom tekstu ručno obeležava posetioce (na primer, ime i prezime osobe), nakon čega pokreće proces učenja. Amilcare alat na osnovu označenih instanci izvodi pravila za ekstrakciju informacija metodama mašinskog učenja. Korisnik može izabrati drugi tekst i pokrenuti proces označavanja delova teksta klasom metapodataka "posetilac", na osnovu naučenih pravila. Metadapodaci koji se navode u opisu alata su "ima-trajanje", "vreme-početka", "vreme-završetka", "ima-lokaciju", "glavni-agent", "drugi-agenti-koji-učestvuju", "posetilac" i "čovek-ili-organizacija-koja-se-posećuje". MnM sistem je kasnije integrisan sa sistemima Marmot, Badger i Crystal, koji su razvijeni na University of Massachusetts (UMass), a čiji se detalji mogu naći u radu [290].

Melita ([35]) je alat za anotaciju koji koristi analize poput poklapanja stringova, označavanje vrsta reči i prepoznavanje imenovanih entiteta, dok za automatizaciju koristi metode nadgledanog mašinskog učenja.

Schema.org ([98], [236]) je inicijativa koja okuplja zajednicu u izgradnji i održavanju shema za strukturirane podatke na internetu, veb stranicama, u porukama elektronske pošte i drugim domenima. Pokretači inicijative za pravljenje rečnika za Schema.org su Google, Microsoft, Yahoo i Yandex, a neke od zajednica koje učestvuju su iz domena zdravstvene nege ([105]), sporta ([253]), definisanja bibliografskih podataka ([235]) i automotive industrije ([270]).

U radu [206] dat je detaljniji pregled navedenih alata, ali i drugih aktuelnih sistema za semantičku anotaciju.

U okviru projekta DELOS (Network of Excellence on Digital Libraries) objavljen je formalni model za anotaciju digitalnog sadržaja, koji ujedno predstavlja i objedinjenu detaljniju sliku postojećih relevantnih istraživanja [2].

Više o ovoj oblasti može se naći u radovima [137], [226], [102], [170] i [297].

Pretraživanje informacija

Jedan od problema je i kako doći do određenih dokumenata na osnovu korisnikovog upita, čime se bavi oblast pretraživanja informacija. Njen zadatak je pronaći u velikoj kolekciji materijala nestrukturirane forme one materijale koji sadrže informacije koje su opisane kriterijumom pretrage. Najvažniji cilj pretraživanja informacija je da ispravno protumači zahtev za podacima i da ima mehanizam kojim se dobijaju relevantni dokumenti u što boljoj meri kvaliteta.

Problem pretraživanja dokumenata se deli na dva potproblema - indeksiranje i pretraga. Indeksiranjem se prave efikasno pretražive veze između dokumenata i termina koji se pretražuju. Pretraga se odnosi na tumačenje upita i pretraživanje indeksa.

Pretraživanje informacija se koristi na više nivoa složenosti. Najjednostavniji nivo je korišćenje za ličnu upotrebu na ličnom računaru. Kolekcija dokumenata ne mora biti velika, tako da i moć računara takođe ne mora biti velika. Za potrebe institucija ili pretrage u nekom specifičnom domenu, na primer, baza naučnih radova, baza patenata, digitalna biblioteka kulturnog nasleđa ili baza dokumenata neke korporacije, potreban je složeniji sistem. Za ove potrebe uglavnom se koriste jedan ili više servera koji mogu da obrade zahteve. Pretraga dokumenata na vebu je posebno složena jer se dokumenti nalaze na različitim serverima i ima ih veoma mnogo.

Indeksiranje dokumenata u tom slučaju nije ujednačeno i može voditi pogrešnim rezultatima.

Istorijat Oblast pretraživanja informacija nastala je mnogo pre pojave računara i odnosi se na bilo koju vrstu pretraživanja. Automatizovanje procesa pretraživanja informacija započelo je sredinom prošlog veka. Nakon toga se osnivaju grupe koje se bave ovim problemom, da bi se sedamdesetih godina prošlog veka razvile prve metode pretraživanja informacija u veoma velikim sistemima. Najduže i najviše obrađivan problem do sada je pretraživanje informacija u tekstualnim dokumentima (Text Information Retrieval TIR), ali su aktuelne oblasti i pretraživanje slika po sadržaju (Content-based Image Retrieval CBIR), pretraživanje audio materijala (Audio Information Retrieval AIR), vizuelnog materijala (Visual Retrieval VR), video materijala (Video Retrieval VDR), geografskih informacija (Geographic Information Retrieval GIR) i pretraživanje multimedijalnih materijala (Multimedia Information Retrieval MMIR).

Pretraga teksta po sadržaju se može izvesti pretragom celog teksta (engl. *full-text search*) ili pridruživanjem metapodataka koji oslikavaju relevantne aspekte sadržaja, a zatim pretragom po metapodacima. Za ovakvu pretragu po metapodacima potrebno je prethodno izvršiti odgovarajuću anotaciju dokumenta. Dokumentu se mogu pridružiti metapodaci koji nisu deo sadržaja već mogu biti deo šireg konteksta (na primer, tip, autor, izdavač), metapodaci koji se mogu naći u sadržaju (na primer, imenovani entiteti) ili se mogu izračunati na osnovu sadržaja (na primer, pridruživanje tematike).

Diskusija o metodama za pretraživanje informacija Jedna od metoda za rešavanje problema indeksiranja je indeksiranje dokumenata unapred ključnim rečima koje ne moraju biti sadržane u dokumentu. Problem sa ovim pristupom je taj što se kao rezultat dobijaju samo dokumenti kojima je pridružena tačno ta reč koja je navedena u korisničkom upitu.

Druga metoda je indeksiranje dokumenata određenim terminima iz sadržaja po kojima će se vršiti pretraživanje. Informacije se mogu čuvati u matrici sa binarnim sadržajem (sadrži - ne sadrži) tako da su po kolonama navedeni dokumenti, dok su po vrstama navedeni termini koji se traže. Ovakva matrica je uglavnom retka, pa se odgovarajućom transformacijom može prevesti u niz bez gubitka informacija.

Ukoliko se upiti formulišu na fleksibilniji način, na primer tako da se ne koriste

predefinisani termini, tada su potrebne naprednije metode pretrage dokumenata. Složenost metoda za pretragu može rasti ukoliko je potrebno termine zadavati tako da se dva ili više termina nalaze u blizini u tekstu ili da se traže i dokumenti koji sadrže termine koji imaju isto ili slično značenje ([71]).

U metode za rešavanje problema pretrage spada pretraga zasnovana na konceptu, koji je baziran na ontologijama, kada se na osnovu upita traži polje relevantnih pojmova koji udruženi imaju isto značenje kao i polazni upit. Kao rezultat dobijaju se tekstovi u kojima se nalazi koncept koji se podudara sa traženim konceptom. Pretraga zasnovana na sadržaju u tom slučaju omogućava da se na osnovu upita dobiju dokumenti koji sadrže relevantne podatke koji su u vezi sa početnim upitom ([60]).

Istraživanja GoNTogle ([86], [20]) je alat za semantičku anotaciju dokumenata različitih formata (doc, pdf, txt, odt), za njihovo indeksiranje i pretragu na osnovu semantičkih koncepata. Ovaj alat omogućava ručnu i automatsku anotaciju velikog korpusa dokumenata na vebu. Anotacija se zasniva na pridruživanju klasa iz posmatrane ontologije (na primer, klase osoba, lokacija, organizacija). Podržava pretragu po ključnim rečima (dobijaju se dokumenti koji sadrže ključnu reč), pretragu po semantičkim konceptima (rezultat su dokumenti koji su anotirani izabranom klasom iz ontologije) i hibridnu pretragu (kombinacija prethodne dve metode). Sistem podržava i profinjavanje rezultata pretrage, poput, za dokument iz rezultata pretrage: naći relevantne dokumente (to su dokumenti koji su takođe anotirani klasom koja anotira i početni dokument), naći dokumente koji su anotirani direktnom potklasom / natklasom klase kojom je anotiran i početni dokument, naći dokumente koji su anotirani bilo kojom, ne obavezno direktnom, potklasom klase kojom je anotiran i početni dokument. Svim opisanim vrstama pretrage i njihovim profinjenjima se pridružuju mere relevantnosti, a konačna mera relevantnosti je suma svih mera relevantnosti koje su korišćene. Na kraju, za svaki rezultujući dokument, ispisuju se sve metode koje su učestvovale u njegovoj pretrazi, kao i mera relevantnosti.

MUMIS projekat (Multi-Media Indexing and Searching) ([136], [51], [231]) koristi metode poboljšanja pretraživanja podataka u multimedijalnim arhivama. Istaknuta osobina ovog sistema je ta da kombinuje informacije iz više dokumenata u cilju pravljenja kompletnije baze znanja. U okviru ovog projekta istraživane su mogućnosti pretrage multimedijalnih materijala na osnovu njihovih tekstualnih opisa, konkretno pretrage snimaka fudbalskih utakmica na osnovu njihovih izveštaja iz različitih

izvora. Izveštaji mogu da budu manje ili više formalni, da opisuju ko su bili učesnici utakmice, koji igrač je dao gol i kada, koji igrač je dobio žuti ili crveni karton, i kada. U opštem slučaju, tekst opisuje sam tok utakmice. Tekstualni opisi su dobijeni iz više izvora, novina i formalnih izveštaja. Kombinuju se metode ekstrakcije informacija iz tekstualnih opisa zasnovanih na ontologijama iz domena, automatske transkripcije govora i identifikovanja ključnih frejmova u video materijalima. Kao izazovi prilikom spajanja informacija iz ovih izvora navedeni su poravnanje podataka iz različitih izvora koji opisuju isti događaj, postizanje jednoobraznosti podataka iz više izvora u smislu razrešavanja mogućih konflikta u slučaju oprečnih informacija i utvrđivanje pravog redosleda događaja.

Jedno istraživanje koje se bavi pretraživanjem multimedije (slika, audio i video materijala) zasnovanom na konceptima (engl. *concept-based*) u domenu kulturnog nasleđa opisano je u radu [126].

Više o oblasti pretraživanja informacija može se naći u [13], [141] i [165].

4.4 Obrada teksta na srpskom jeziku

Specifičnosti srpskog jezika

Ovde su istaknute najznačajnije karakteristike srpskog jezika u cilju boljeg razumevanja izazova njegove obrade uopšte. Detaljnije specifičnosti se mogu naći u opisu koji je objavljen u okviru meta-net projekta u radu [296].

Prilikom obrade teksta na srpskom jeziku potrebno je imati u vidu njegove sledeće osobine:

- U upotrebi su dva pisma, ćirilično i latinično
- Ima bogat morfološki sistem, kako u smislu izvedenih reči, tako i u smislu različitih oblika za jednu istu reč
- Iako postoji preporučeni redosled reči, on nije i obavezan, tako da je moguće koristiti veliki broj varijacija po pitanju kombinacije reči u rečenici
- Zastupljena je homografija

Pre obrade je korisno ujednačiti tekstove izborom jednog pisma, odnosno prevesti sve ćirilične tekstove u latinicu ili sve latinične u ćirilicu.

Bogata morfologija srpskog jezika podrazumeva da jedna reč može imati više flektivnih oblika. Postoji deset vrsta reči sa velikim brojem potklasa. Promenljive vrste reči u srpskom jeziku su imenice, pridevi, glagoli, zamenice i brojevi, a postoje tri vrste fleksije (deklinacija, konjugacija i poređenje). Unutar sve tri vrste fleksija postoje različite paradigme, uz niz izuzetaka.

Srpski jezik pripada jezicima sa slobodnim redom reči, tačnije, sa slobodnim rasporedom članova koji mogu menjati svoje mesto. Izbor određenog reda zasnovan je na vrlo složenom funkcionalnom sistemu, regulisanom kombinacijom različitih sintaksnih, semantičkih, pragmatičnih i stilskih faktora. Sve permutacije su dozvoljene, ali preferirani redosled je subjekat, predikat, objekat.

Za određeni broj leksema i oblika reči postoje dva različita izgovora, ekavski i ijekavski, etimološki povezani sa starim slovenskim samoglasnikom jat.

Homografija je pojava da postoje reči koje se isto pišu, ali imaju drugačije značenje. U srpskom jeziku je homografija česta, kao posledica bogatog morfološkog sistema. Ona utiče na identifikaciju gramatičkog značenja reči (glagol / pridev / imenica, različiti padeži imenica), ali i samog značenja reči u okviru istog gramatičkog značenja. Na primer, reč "kosi" može imati različita gramatička značenja - kosi travu, kosi toranj, u kosi, ka kosi. U okviru istog gramatičkog značenja može imati drugačije značenje, na primer, kosa kao alat i kosa kao skup dlaka na glavi.

Pored ovoga, srpski jezik ima i druge specifičnosti, zbog kojih je njegova računarska obrada složen zadatak.

Resursi za obradu srpskog jezika

Jezički resursi predstavljaju skupove jezičkih podataka i opisa u mašinski čitljivom formatu i služe za obradu podataka na prirodnom jeziku. Imajući u vidu specifičnosti srpskog jezika za njegovu obradu potreban je bogat skup resursa i alata. Grupa za jezičke resurse i tehnologije na Matematičkom fakultetu u Beogradu u saradnji sa Filološkim fakultetom u Beogradu razvila je brojne resurse i alate koji su neophodni za obradu srpskog jezika, u koje spadaju višejezični paralelni korpusi, elektronski morfološki rečnici, rečnik vlastitih imena, leksičko-semantička mreža Wordnet, višejezična ontologija vlastitih imena Prolex i mnogi drugi. U ovom poglavlju biće reči o nekim od najznačajnijih resursa. Detaljniji pregled resursa za obradu srpskog jezika može se naći u [294] i [63].

Korpusi srpskog jezika Korpus predstavlja kolekciju tekstova na kojoj se mogu vršiti različite analize. Tekstovi se biraju tako da dobro reprezentuju posmatrani prirodni jezik ili neki njegov domen. Korpusi u kojima su praćeni jasni lingvistički kriterijumi su na primer, Dijahroni korpus srpskog jezika formiran na Institutu za eksperimentalnu fonetiku i patologiju govora u Beogradu i Korpus savremenog srpskog jezika razvijan na Matematičkom fakultetu u Beogradu. Korpus savremenog srpskog jezika, u sadašnjem obliku se naziva SrpKor2013, sastoji se od novinskih članaka, administrativnih tekstova, književno umetničkih tekstova, naučnih i naučno popularnih tekstova. Sadrži više od 122 miliona korpusnih reči, anotiran je bibliografski i morfološki. SprKor2013 je besplatan uz prethodnu registraciju ([254]). Lematizirani korpus srpskog jezika SrpLemKor ([139]) podskup je SrpKor2013 korpusa veličine 3.7 miliona reči. Referentni korpus savremenog srpskog jezika SrpKor opisan je u radu ([289]).

Najvažniji višejezični korpusi srpskog jezika su SELFEH (Serbian-English Law Finance Education and Health) ([241]), Englesko-srpski paralelni koprus (SrpEng-Kor) ([61]) i Francusko-srpski paralelni korpus ([75]).

Elektronski morfološki rečnici Elektronski rečnici su rečnici koji se koriste u računarskoj obradi prirodnog jezika. Oni sadrže informacije potrebne za zadatke segmentacije i morfološke analize teksta i od velikog su značaja za njegovu automatsku obradu ([40]). Oni se sastoje od reči i složenica, zajedno sa osnovnim oblicima reči i gramatičkim pravilima za njihove promene. Morfološki rečnici se sastoje od nekoliko delova: DELAS rečnici - sastoje se samo od osnovnih oblika reči, DELAF - sastoje se od svih flektivnih oblika reči, i slično, DELAC - sadrže osnovne oblike složenih reči i DELACF - sadrže izvedene oblike složenih reči.

U slučaju srpskog jezika, postoje elektronski rečnici koje razvija Grupa za jezičke tehnologije na Matematičkom fakultetu Univerziteta u Beogradu. U ovom trenutku postoji javno dostupan morfološki rečnik koji se može preuzeti u sklopu Unitex programa. Prednost korišćenja elektronskog rečnika je u tome što je bogat veoma potebnim informacijama za automatsku obradu teksta.

Na primer, jedan unos za DELAF rečnik je:

```
lekara, lekar.N + Hum + Ek: ms2: ms4: mp2
```

Ovaj unos se sastoji od oblika “lekara” čija je lema “lekar”. “N + HUM + Ek” označava da je to imenica (“N”) o osobi (“HUM”) i u ekavskom (“EK”) izgovoru. Sekvenca “ms2: ms4: mp2” označava sva moguća gramatička značenja ovog oblika

reči. Na primer, "ms2" znači da je imenica u muškom rodu ("m" - mascular), u drugom padežu jednine ("s2" - singular).

U ovim rečnicima postoje podrečnici vlastititih imena i toponima. Imenovani entiteti u srpskom jeziku su analizirani i preko višejezičnih ontologija vlastitih imena (Prolex system) ([169], [144]).

Wordnet Wordnet je informatičko leksičko-semantička mreža, resurs koji nalazi više primena u obradi prirodnog jezika. U njemu su sve reči grupisane u sistem skupova sinonima ili sinseta.

Srpski wordnet ([257]) je leksičko-semantička mreža srpskog jezika. Organizovan je kao sistem čvorova i relacija između čvorova. Svaka reč u sinsetu predstavljena je niskom karaktera ili literalom, za kojom sledi značenje tog konkretnog literala u konkretnom sinsetu. Wordnet je detaljnije opisan u doktorskoj disertaciji [181].

Alati za obradu srpskog jezika

Najviše korišćen lingvistički alat za obradu srpskog jezika je Unitex ([287], [215]). Unitex je skup programa namenjenih za analizu teksta na prirodnom jeziku uz pomoć posebnih dodatnih resursa. Jedan od tih resursa su i elektronski morfološki rečnici koji su namenjeni automatskog obradi teksta. Elektronski rečnici su i glavna prednost Unitex-a, jer u sebi već sadrže brojne semantičke informacije uz pomoć kojih se mogu definisati čitave klase reči po željenoj osobini.

Unitex može izvršiti normalizaciju (izbacivanje praznih karaktera) i tokenizaciju prilagođenu specifičnom jeziku. Nakon što se napravi lista tokena, elektronski rečnici mogu da se primene na tekst ([143]). Zatim, tekst se može obraditi regularnim izrazima ili konačnim transduktorima (Finite State Transducer FST). FST automati mogu se predstaviti kao grafovi sa mogućnošću formiranja izlaznog teksta na osnovu ulaznog teksta prepoznavanjem određenih konstrukcija i umetanjem željenih oznaka.

Pored Unitex-a, široko korišćen alat za obradu teksta na prirodnom jeziku je GATE ([81]), program koji je razvila grupa za obradu prirodnih jezika Univerziteta u Sheffieldu. Izlaz iz programa GATE je tekst sa označenim terminima. Proces označavanja sastoji se od identifikacije znakova i segmentacije, povezivanja sa odgovarajućim vrstama reči (POS) i označavanja određenih pojmova.

Izdvojena istraživanja

Obrada srpskog jezika je polje aktivnog istraživanja. U nastavku su navedena neka od istraživanja koja se bave obradom srpskog jezika, ili obradom prirodnih jezika sa primenama na srpski jezik, a u domenu su ekstrakcije informacija, klasifikacije teksta, anotacije dokumenata ili pretraživanja informacija.

Veći broj istraživanja sprovedeno je u oblasti ekstrakcije informacija, a posebno na polju prepoznavanja i izdvajanja entiteta. U radu [295] predstavljen je pregled resursa i metoda za prepoznavanje imenovanih entiteta na srpskom jeziku. U radu [145] predstavljen je sistem za prepoznavanje i označavanje entiteta koji se oslanja na velike leksičke resurse i konačne transduktore. Sistem prepoznaje nekoliko vrsta imena, vremenskih i numeričkih izraza. Model za poređenje različitih pristupa prepoznavanju imenovanih entiteta u paralelnim tekstovima ("bitekstovima") prikazan je u radu [147]. Automatsko izdvajanje vremenskih izraza iz nestrukturiranih tekstova, kao i izdvajanje ličnih imena, predstavljeni su u [118] i [97].

Opisi metapodataka i sadržaja obrađeni su u doktorskoj disertaciji [280]. Metod dvofaznih transduktora konačnih stanja za ekstrakciju informacija iz polustrukturiranih resursa (na primer, enciklopedija) dat je u [209]. Značajna pažnja se takođe posvećuje ekstrakciji višečlanih izraza i imenovanih entiteta u određenim domenima, na primer, poluautomatska ekstrakcija višečlanih izraza iz korpusa specifičnih za domen poljoprivrede, višečlanih izraza u domenu bibliotekarstva, rudarstvu i geologiji, imenovanih entiteta u kulinarskom domenu, domenu prava ili vremenskih informacija iz meteoroloških tekstova ([211], [260], [146], [200], [210]).

Sprovedena su istraživanja na polju klasifikacije teksta zasnovane na ontologiji, sentiment analize, klasifikacije teksta na osnovu ironije i sarkazma, kategorizacije teksta i detekcije autorstva ([216], [316], [183], [18], [182], [92], [93]).

U radu [17] opisano je istraživanje u kome je ispitan uticaj normalizacije na polju klasifikacije tekstova na osnovu polariteta, uz korišćenje metode ugrađivanja reči u vektorski prostor reči. Sprovedeno je i istraživanje koje ispituje sličnost tekstova na srpskom jeziku uz pomoć statističkih metoda i uz korišćenje ugrađenih reči u vektorski prostor ([79], [16]).

Anotacijom u širem smislu bavi se doktorska disertacija ([289]), čiji je fokus istraživanja anotacija savremenog srpskog jezika.

Istraživanja koja se bave izgradnjom elektronskog korpusa timočkog govora, kao jednog od jezika koji je prema Uneskovoj odluci iz 2010. godine svrstan u ugrožene jezike, opisana su u radovima [298], [301], [300] i [299]. Aktuelna verzija korpusa je

objavljena u [303], dok se na lokaciji [302] može pogledati više o projektu izgradnje modela predstavljanja timočkih govora i pregledati deo ove kolekcije.

Pretraživanje informacija u velikim tekstualnim bazama podataka preindeksiranjem dokumenata pomoću vreće reči i prepoznavanja naziva entiteta opisano je u radovima [258] i [259]. Sistem za nadgledanje veba u kome se koriste konačni automati za pravljenje fraza po kojima se pretražuju i izdvajaju relevantne stranice na vebu prikazan je u radu [212].

4.5 Obrada teksta na prirodnom jeziku u domenu kulturnog nasleđa

Podaci koji se nalaze u kolekcijama o kulturnom nasleđu nose u sebi veliku vrednost koja se može iskoristiti u cilju razumevanja sadržaja kulturne baštine. Ovakve kolekcije su uglavnom multimedijalne, što uključuje i tekstualne zapise kulturnog nasleđa ili tekstualne opise materijala drugih medija. Tekstualni dokumenti su često nestrukturirani, stoga su teže čitljivi u kontekstu mašinske obrade. Potrebno ih je urediti, povezati i obogatiti metapodacima. Ručna obrada je jedna od mogućnosti, ali ujedno i prezahtevna. Stoga je automatizacija procesa obrade više nego dobrodošla. Zbog specifičnosti sadržaja koji nose, efikasno pretraživanje takvih kolekcija treba da bude prilagođeno njegovoj semantici. Ovde su navedena izabrana istraživanja koja se bave takvim podacima iz ugla obrade prirodnih jezika u cilju omogućavanja što boljeg pristupa znanju iz kolekcija sa ovakvim vrednim sadržajem.

CHoral ([106]) je projekat koji istražuje ulogu koju automatska anotacija i tehnologija pretraživanja informacija mogu imati u poboljšanju otkrivanja značajnih reči kojima bi se mogli anotirati dokumenti digitalnih bibioteka.

New South Voices ([198]) pruža pristup više od 700 transkripata intervjua, pripovedaka i razgovora koji dokumentuju život u regionu Charlotte, Severna Karolina, uključujući iskustva i jezik nedavnih imigranata u to područje.

Jedna od radionica o ekstrakciji informacija iz kolekcija kulturne baštine je ENRICH 2013 ([62]). Cilj radionice bio je da se istraži mogućnost pretrage i interakcija sa kolekcijama kulturnog nasleđa. Teme koje su obuhvaćene radionicom su inovativni oblici pretraživanja sadržaja i personalizovano pronalaženje informacija, što odgovara aktivnosti i interesovanjima šire zajednice i doprinosi poboljšanju iskustva u korišćenju ovakvih alata uopšte.

Neke od platformi koje su razvijene za rad sa podacima koji se odnose na kulturno nasleđe su Omeka ([99]), WissKI ([237]) i Arches ([190]). Trenutna situacija je da platforme razvijene za organizaciju kulturnog nasleđa uglavnom nude pretragu koja se odnosi na pretraživanje po pridruženim metapodacima ([227]). Da bi se omogućili složeniji upiti, potrebno je poboljšati kvalitet metapodataka uvođenjem bogatije semantike.

Glava 5

Problem upravljanja multimedijalnim nematerijalnim kulturnim nasleđem

5.1 Uvod

Sistematičan pregled zahteva koje treba da ispunjava dobar multimedijalni sistem kulturnog nasleđa opisan je u radu [45], a biće prikazan u nastavku u kraćim crtama.

Na prvom mestu navedene su korisničke potrebe za određenim funkcionalnostima i podacima. Nije dovoljno napraviti tehnički dobar sistem ako on nije prilagođen potrebama i očekivanjima korisnika. Očekivanja se mogu formulisati u vidu liste funkcionalnosti, nivoa emotivnog zadovoljstva prilikom upotrebe, što bi moglo da se odnosi na intuitivnost i jednostavnost upotrebe, dobru preglednost i upotrebljivost dobijenih informacija, i na kraju prilagođenost korisnikovim navikama, ukusima i očekivanjima u tom segmentu. Navedeni aspekti su isprepleteni i ne mogu se striktno razgraničiti.

Akademska istraživanja često nisu odmah prijemčiva za širu publiku, a čak ni za same istraživače. Potrebno je vreme i iskustvo kako bi se došlo do konačnog oblika prilagođenog krajnjem korisniku. Zadovoljstvo korisnika interakcijom u jednom multimedijalnom sistemu je ono što će najviše uticati na njihovo sveopšte zadovoljstvo sistemom. Dobre funkcionalnosti uz lošu interakciju veoma često odbijaju korisnika od daljeg korišćenja. Veliki trud se ulaže u analizu potreba, i to različitih vrsta korisnika. Različite će biti potrebe za interakcijom, nivoom detaljnosti i načinom prikaza

sadržaja. Na primer, jedne potrebe će imati šira javnost, a neke druge eksperti iz posmatrane oblasti.

Važna kategorija korisnika multimedijalnih sistema kulturnog nasleđa u grupi eksperata su arhivatori. Njihove potrebe u vidu funkcionalnosti su uglavnom i najsloženije u smislu nivoa detaljnosti, nivoa kvaliteta rezultata i količine podataka koji su im potrebni za rad. Arhivatori u domenu kulturnog nasleđa imaju potrebu za veoma specifičnim funkcionalnostima, bogatim skupom metapodataka i različitim dokumentima u multimedijalnoj formi.

Da bi sve ovo bilo omogućeno, potrebno je sadržaj dokumenata i njegovih metapodataka opisati posebnom strukturom, koja ne samo da može jednostavnije da se protumači, već se po potrebi transformiše u drugi odgovarajući oblik. Sledeći zadatak je izbor tehnologija koje će najbolje moći da pokriju sve aspekte prikazanog problema. Takve tehnologije su u ovom trenutku već dovoljno razvijene da mogu da podrže različite zahteve poput pretrage multimedijalnih podataka, dobavljanja velike količine podataka, efikasnog skladištenja i pristupa podacima.

U radu [1] opisano je nekoliko sistema za anotaciju i navedeni su zahtevi koje treba da ispune sistemi za anotaciju i upravljanje digitalnim bibliotekama kako bi bili prihvaćeni od strane korisnika.

Iz perspektive računarskih nauka, zahtevi se mogu podeliti u dve grupe - korisničke i sistemske. U zahteve sa korisničkog nivoa spadaju mogućnosti poput dodeljivanja nove anotacije, izmene ili brisanja postojeće anotacije, okvira privatnosti anotacija (vidljiva samo tom korisniku, grupi korisnika ili svima), dobijanja liste svih anotacija pridruženih digitalnom objektu, pretrage i pregleda postojećih anotacija. Sa sistemskog aspekta neki od važnih zadataka su organizacija korisnika i grupa, stvaranje i čuvanje anotacija, brisanje i izmena anotacija, postavljanje okvira anotacija (privatna, deljena ili javna) i način pretrage anotacija.

Iz perspektive društvenih i humanističkih nauka, sistemi za anotaciju sasvim menjaju mogućnosti i uloge koje korisnici mogu imati u radu sa digitalnom bibliotekom. Korisnici mogu da budu u ulozi autora, evaluatora, mogu da stvaraju kontekst u kome se interpretiraju anotacije i da digitalne biblioteke ne koriste samo kao repozitorijume multimedijalnih materijala, već i kao okruženje za komunikaciju sa ostalima poput učitelja, čitalaca, studenata, izdavača ili obučenih eksperata iz specifičnih domena. Stoga, prilikom dizajniranja sistema potrebno je imati u vidu način na koji će taj sistem koristiti različite grupe korisnika i kako će teći eventualni tokovi komunikacije između njih. Specifikacija ovakvog sistema u većoj meri treba

da zahteva bavljenje modelovanjem potreba korisnika nego tehničkim modelovanjem tokova podataka.

Za uspešan sistem potrebno je da se obe perspektive, i društvenih nauka i računarskih nauka, ujedine u stvaranju dobrog i korisnog sistema digitalnih biblioteka. Spoj eksperata iz ovih oblasti, sa jedne strane istraživača iz domena umetnosti, kulture, sociologije, istorije, lingvistike, i sa druge strane inženjera i istraživača u domenu računarskih nauka, od nemerljivog je značaja za stvaranje inovativnih tehnologija, za kojima danas postoji značajna potreba.

Brojni su izazovi sa kojima se susreću istraživači koji se bave razvojem i primenom tehnika obrade prirodnog jezika u domenu kulturnog nasleđa. Neki od izazova su ti da jezik koji se koristi može biti nestandardan ili arhaičan, da je u upotrebi više vrsta medija iz kojih treba izdvojiti informacije i povezati ih, da postoji slaba podrška u vidu alata i resursa prilagođenih ovom domenu što iziskuje da se sistemi grade "od početka" i postojanje više različitih standarda koji su u upotrebi što zahteva ujednačavanje načina na koji se koriste dokumenti u okviru sistema zbog njihove koherentne upotrebe. Često se istraživanja u oblasti kulturnog nasleđa sprovode na terenu što proizvodi poseban tip materijala u vidu beleški ili protokola. U radu [252] se detaljnije opisuju navedeni problemi, ali i brojni drugi, koji se odnose na pretraživanje informacija u ovom domenu, uz naglašavanje da "domen kulturne baštine predstavlja značajan izazov za jezičku tehnologiju i zahteva razvoj vrlo robusnih i fleksibilnih rešenja".

Nematerijalno kulturno nasleđe je jedan od specifičnih domena koji nije istražen dovoljno na prostoru Balkana u kontekstu obrade prirodnih jezika, što uključuje i srpski jezik. Postojeći alati za obradu prirodnih jezika generalno ne rade tako dobro kada se primene na domen koji je drugačiji od domena nad kojim su oni razvijani i testirani, s obzirom na to da promena domena vodi značajnoj promeni rečnika i sintaksnih struktura koje su u upotrebi ([229], [87]). Zaključci iz različitih istraživanja impliciraju da, u slučaju specifičnih domena, specifični kontekst ima značajnu ulogu u procesu ekstrakcije informacija ([108]). To dalje implicira da kompletna infrastruktura treba da se razvije za obradu srpskog jezika u domenu nematerijalnog kulturnog nasleđe, od ručno razvijanih pravila i specifične ontologije, preko elektronskih rečnika i pravljenja rečnika specifičnih reči, do razvijanja sistema za modelovanje znanja o korišćenju jezičkih konstrukcija u specifičnom domenu poput identifikovanja specifičnih semantičkih struktura.

Detaljniji pregled problema i evaluacije sistema o kulturnom nasleđu može se

naći u radu [78].

5.2 Multimedijalna kolekcija nematerijalnog kulturnog nasleđa Balkana

Istraživanje koje je sprovedeno u ovom doktoratu je motivisano multimedijalnom kolekcijom koja predstavlja rezultat višedecenijskog terenskog istraživanja koje se sprovodi od strane istraživača Balkanološkog instituta Srpske akademije nauka i umetnosti. Istraživači tokom terenskih istraživanja intervjuišu lokalno stanovništvo na različitim lokacijama na području Balkana i šire. Primarni cilj ovih istraživanja je očuvanje informacija o različitim tipovima govora koji se koriste na ovim prostorima, zatim proučavanje njihovih jezičkih karakteristika uz istovremenu dokumentaciju tradicijske kulture, folklora i usmene istorije sagovornika na terenu.

Multimedijalna kolekcija koja je predmet ovog istraživanja sastoji se većinom od audio i video zapisa u obliku intervjua na različitim jezicima i dijalektima koji se koriste na raznim mestima na Balkanu i regionu. Materijali su snimani u Rumuniji, Mađarskoj, Hrvatskoj, Bosni i Hercegovini i Bugarskoj. Na teritoriji Srbije prikupljen je veliki korpus intervjua sa Bunjevcima, Hrvatima, različitim grupama Roma koji govore albanski, romski ili rumunski, Rumunima, Vlasima, Česima, Mađarima i Bugarima. Korpus takođe pokriva celokupnu teritoriju Kosova i Metohije i Sandžaka. Pored audio i video materijala, kolekcija sadrži i mnoštvo fotografija. Intervjui uglavnom imaju pridružene tekstualne opise, takozvane protokole, kao i određeni broj transkripata, od kojih su neki prevedeni na nekoliko jezika. Stotine naučnih radova objavljeno je na osnovu ove kolekcije i uglavnom se tiču istraživanja antropološko-lingvističkih i sociolingvističkih karakteristika na građi ekscerpiranoj iz snimljenih intervjua.

U intervjuima su sagovornici odgovarali na različita pitanja i govorili o raznovrsnim temama. Video materijali mogu biti propratni materijali koji su snimani u toku intervjua dodatno opisujući neke važne aspekte koji su tematika intervuja. Na primer, dok sagovornik opisuje svoj život pokazujući fotografije, istraživač ili neko od prisutnih snima dodatne vizuelne informacije o kojima se govori. Sa druge strane, video materijali mogu sadržati i demonstraciju raznih običaja ili veština, pa je tako neretko snimana proslava seoske slave sa svim propratnim običajima koji se praktikuju. Postoje i snimci izvođenja narodnih igara i individualnih ili kolektivnih tradicionalnih običaja.

Protokoli su tekstualni dokumenti u vidu nestrukturiranog ili polustrukturiranog teksta koji imaju za cilj da opišu određeni audio ili video materijal. U protokolu su opisane osnovne karakteristike snimka na koji se odnosi, kao što su lokacija na kojoj je materijal snimljen, datum, učesnici, etnička i verska pripadnost, jezik na kome se govori i drugo. Pored ovih informacija, u protokolima su navedeni i kratki izvodi iz sadržaja razgovora. Razgovori se odvijaju na različitim jezicima, ali su svi protokoli napisani na srpskom jeziku.

U toku istraživanja multimedijalna kolekcija se sastojala od oko 3000 audio materijala sa oko 2000 sati snimanih intervjua, 200 video materijala, 800 tekstualnih protokola sa oko 600000 reči, brojnih slika i drugih pisanih materijala.

Glava 6

Metode za rešavanje problema upravljanja multimedijalnim nematerijalnim kulturnim nasleđem

Proces rešavanja problema upravljanja multimedijalnim nematerijalnim kulturnim nasleđem se sastoji od više podzadataka. Tok projekta i razvijene metode mogu se sistematizovati na sledeći način:

- Prvi korak koji je sproveden u rešavanju postavljenog problema je analiza potreba korisnika. Zajedno sa korisnicima došlo se do odgovarajućeg početnog dizajna sistema za upravljanje datom kolekcijom koji uključuje grafički korisnički interfejs za ručnu anotaciju materijala, njihovo organizovanje u bazu podataka kao i pretragu i prikaz anotiranih materijala.
- Izvršena je analiza alata i tehnologija potrebnih za tehničko izvođenje sistema i realizovana je njegova izrada.
- Razvijen je prostorni model kojim se geografske karakteristike materijala predstavljaju na mapi.
- Razvijen je model pretrage po prostornim karakteristikama izborom lokacije na interaktivnoj mapi.
- Sledeći korak je bio izgradnja sistema za pomoć u ručnoj anotaciji materijala u vidu automatske semantičke anotacije protokola metapodacima koji se mogu podeliti na imenovane entitete i na teme koje su predmet sadržaja koji se anotira. Automatska anotacija protokola je sprovedena primenom

lingvističko-semantičkih metoda obrade prirodnih jezika u analizi protokola u cilju ekstrakcije metapodataka.

- Razvijene su metode klasifikacije protokola na osnovu postojanja specifične tematike kao predmeta sadržaja metodama mašinskog učenja prilagođene karakteristikama protokola i domena nematerijalnog kulturnog nasleđa.
- Razvijenim metodama ekstrakcije informacija i klasifikacije teksta omogućena je bogatija semantička pretraga baze podataka.
- Razvijena je osnova za izradu vremenskog modela u vidu sistema za ekstrakciju vremenskih entiteta lingvističko-semantičkim metodama obrade prirodnih jezika.
- Postavljena je dobra početna osnova za implementaciju dodatnih metoda pretrage razvojem bogatog sistema semantičkih metapodataka.

Rezultati obrade tekstualnih protokola koriste se kao pomoć pri polu-automatskoj anotaciji i organizaciji dokumenata, pa je ovde prvo prikazana metodologija obrade teksta i njena primena na protokolima. Nakon toga je opisan sistem za anotaciju i organizaciju dokumenata i način na koji su povezani rezultati prethodna dva. Opisane su mogućnosti pretrage materijala u uspostavljenoj organizaciji i implementacija izabranih metoda pretraživanja informacija. Posebno poglavlje posvećeno je prostornom modelu.

6.1 Metode prepoznavanja informacija

Prepoznavanje imenovanih entiteta

Za svrhu prepoznavanja imenovanih entiteta (named entities NE) mogu se koristiti neki od poznatih NER sistema. Iako generalno veoma uspešni, kod obrade specifičnih zahteva ovi već postojeći NER sistemi nisu dovoljni, već je potrebno koristiti dodatne algoritme za obradu.

U proceduri za prepoznavanje imenovanih entiteta koja se primenjuje u ovom radu koriste se:

- elektronski rečnici (na primer, reč “Beograd” se prepoznaje kao lokacija ako se nalazi u rečniku toponima)

- kontekst (na primer, "živi u X", X se prepoznaje kao lokacija, čak i ako X nije nađeno u odgovarajućem rečniku)
- kontekst sa posebnim potkontekstom ("subNE") opšteg NE, koji je određen specifičnim zahtevima domena (na primer, "informator je A" - A se prepoznaje kao potklasa "ispitanik" klase "osoba", ili "razgovor vodila A" - A se prepoznaje kao potklasa "ispitivač" klase "osoba").

Procedura za prepoznavanje imenovanih entiteta može se opisati na sledeći način:

- identifikovati izvore imenovanih entiteta - izvori mogu biti elektronski rečnici, lokalne gramatike (na primer, rečnici ličnih imena, toponima, lokalne gramatike datuma i slično)
- za svaki imenovani entitet, identifikovati kontekstne fraze; za imenovane entitete koji imaju potkategorije identifikovati potkontekst, odnosno kontekstne fraze za potkategorije
- niz uzastopnih reči u tekstu prepoznaje se kao instanca imenovanog entiteta ukoliko se može prepoznati nekim dodatnim izvorom imenovanih entiteta ili ukoliko mu prethodi ili ga sledi kontekstna fraza koja odgovara tom imenovanom entitetu.

Prepoznavanje tema zasnovano na konstrukciji semantičkih struktura

Na početku se analizira tekst i uočavaju se kontekstne fraze koje označavaju izabrane teme. Zatim se identifikuju semantičke strukture koje opisuju uočene kontekstne fraze. Procedura za razvijanje semantičkih struktura sastoji se od identifikovanja specifičnih semantičkih klasa termova koje ih sačinjavaju i odnosa između njih. Semantičke klase termova i njihovi odnosi mogu biti:

- *klasa samostalnih termova* (na primer, "poljoprivreda" može da bude indikator da se radi o tematici "poljoprivreda")
- *klasa specifičnih termova koji se nalaze u specifičnom kontekstu* (na primer, "čuvanje ovaca" često ukazuje na tematiku "poljoprivreda", s obzirom da je specifični term "ovaca" klase "životinja" u specifičnom kontekstu za klasu "životinja" - "čuvanje" u okviru fraze "čuvanje ovaca", koja je indikator tematike "poljoprivreda")

- *klasa opšteg konteksta u kombinaciji sa specifičnim termima* (na primer, "terminologija sejača", može ukazivati na tematiku "poljoprivreda", gde je "terminologija" opšti kontekst, dok je "sejača" specifični term klase "poljoprivreda alati", potklase klase "poljoprivreda").

Procedura za prepoznavanje tema zasnovana na konstrukciji semantičkih struktura sastoji se od sledećih koraka:

- za svaku temu koja se obrađuje, identifikovati u tekstu kontekstne fraze koje označavaju tu temu
- konstruisati semantičke strukture kojima se opisuju identifikovane kontekstne fraze - identifikovati činioce semantičkih struktura i njihove odnose, a zatim i uloge tih činilaca pridruživanjem neke semantičke klase termova
- za svaki od činilaca semantičke strukture napraviti listu osnovnih oblika reči koje opisuju taj činilac
- za svaku reč iz listi, obezbediti dodatne izvore drugih oblika reči - dodatni izvori mogu biti elektronski rečnici, posebno napravljene vreće reči drugih oblika reči ili dopušteno odstupanje reči od početnog oblika po nekom zadatom pravilu
- niz uzastopnih reči u tekstu prepoznaje se kao instanca fraze neke teme ukoliko se može opisati nekom od identifikovanih semantičkih struktura u kojoj su primenjene reči iz odgovarajućih listi ili dodatnih izvora drugih oblika reči.

6.2 Metode ekstrakcije informacija

Ekstrakcija imenovanih entiteta zasnovana na kontekstu

Prvi korak u ekstrakciji imenovanih entiteta je pronaći metapodatke koji su od posebnog značaja za posmatranu klasu dokumenata. To mogu da budu imenovani entiteti kao što su, na primer, osobe, lokacije, datumi i drugi. Ovaj korak se izvodi u saradnji sa ekspertima iz oblasti koja se obrađuje.

Zatim se primenom metodologije za prepoznavanje imenovanih entiteta identifikuju dodatni izvori imenovanih entiteta i kontekstne fraze i razvija biblioteka konačnih transduktora za ekstrakciju imenovanih entiteta.

Pseudoprocedura za ekstrakciju imenovanih entiteta može se opisati sledećim koracima:

- identifikovati listu imenovanih entiteta od značaja
- ustanoviti da li je potrebno neke imenovane entitete deliti u potkategorije i na koji način
- primeniti prethodno opisanu proceduru za prepoznavanje imenovanih entiteta
- izgraditi biblioteku konačnih transduktora koji ekstrahuju imenovani entitet uz pomoć odgovarajućeg izvora (označiti sa "rečnik") ili prepoznavanjem entiteta u kontekstu neke od kontekstnih fraza imenovanog entiteta (označiti sa "kontekst")

Ekstrakcija tema zasnovana na konstrukciji semantičkih struktura

Prvi korak u označavanju tema koje se javljaju u tekstu je identifikovanje sheme po kojoj će se klasifikovati moguće teme. To je problem koji je zavisan od domena i može biti formulisan kao ontologija, tezaurus, liste i drugo. Opciono, iz skupa mogućih tema biraju se teme koje se žele obrađivati. Nakon toga, na osnovu procedure prepoznavanja tema zasnovane na konstrukciji semantičkih struktura, identifikuju se kontekstne fraze koje se odnose na teme i semantičke strukture za opisivanje tih kontekstnih fraza. Napravljene konstrukcije se zatim koriste u razvoju transduktora za ekstrakciju tema iz teksta.

Pseudoprocedura za ekstrakciju tema zasnovana na konstrukciji semantičkih struktura može se opisati sledećim koracima:

- identifikovati shemu po kojoj se klasifikuju moguće teme
- opciono, iz skupa mogućih tema identifikovati teme koje će se obrađivati
- primeniti prethodno opisanu proceduru prepoznavanja tema zasnovanu na konstrukciji semantičkih struktura
- izgraditi biblioteku konačnih transduktora koji ekstrahuju fraze iz teksta koje su prepoznate kao fraze posmatranih tema.

Treniranje modela

Podskupovi za treniranje i testiranje proizvoljno se biraju iz kolekcije svih tekstova, a zatim se ručno obeležavaju imenovani entiteti i teme na osnovu ljudske procene. Primenjuju se pseudoprocedure za ekstrakciju imenovanih entiteta i ekstrakciju fraza koje se odnose na tematiku, čime se dobija biblioteka konačnih transduktora. Ova biblioteka se primenjuje i poboljšava na podskupu za treniranje, obično u nekoliko iteracija, sve dok se ne postigne željena tačnost. Kada je podešavanje modela završeno, konačni transduktori se primenjuju na podskup za testiranje i izračunavaju se mere kvaliteta ekstrakcije. Kako bi se dobili svi dostupni podaci iz tekstova, konačni transduktori se zatim primenjuju na sve raspoložive tekstove i izdvojene informacije se dalje mogu koristiti za zadatke kojima su prvobitno i namenjene.

6.3 Primena metoda ekstrakcije informacija na tekstualne protokole

Protokoli se u opštem slučaju sastoje od nestrukturiranog teksta. U nekim slučajevima informacije se prikazuju u obliku kratkih fraza, a ponekad i kao cele rečenice. Iako je čoveku često jednostavno iz određenog konteksta da odredi ulogu gotovo svakog dela teksta, za računar ovaj zadatak nije lak, posebno zato što ne postoji jedinstvena struktura koju prate svi protokoli. Međutim, ovi protokoli, iako različiti, imaju neke zajedničke karakteristike. Većina njih se može podeliti u dva dela. Prvi deo sadrži metapodatke koji opisuju intervju, dok drugi deo opisuje tok razgovora. Nema svaki protokol obavezno oba dela.

Prvi deo protokola obično sadrži informacije o informatorima koji su intervjuisani, istraživačima koji su imali ulogu ispitivača, mestu i datumu snimanja, jeziku na kome se razgovor vodio, etničkoj pripadnosti i religiji informatora, kao i oznakama protokola.

U drugom delu obično se nalazi tok razgovora. Može se sastojati od popisa kratkih fraza koje predstavljaju teme o kojima se govorilo u intervjuu, ali može se dati i kao kratak ili duži opis u vidu neformalnog transkripta intervjua, ili kao kombinacija prethodnih beleški.

U nastavku je opisana ekstrakcija deskriptivnih metapodataka iz protokola sprovedena po koracima pseudoprocedure za ekstrakciju imenovanih entiteta. Nakon toga, opisana je ekstrakcija metapodataka koji se odnose na teme iz protokola koja

je sprovedena prema opisanoj pseudoproceduri za ekstrakciju tema zasnovanoj na konstrukciji semantičkih struktura.

Ekstrakcija deskriptivnih metapodataka

Identifikovanje informacija od značaja - pronalaženje skupa relevantnih metapodataka Skup metapodataka o razgovorima koji se mogu naći u protokolima i koji se smatraju važnim identifikuje se delom razgovorom sa istraživačima (etnolingvističkim ekspertima) koji su napravili protokole i tako imaju jasnu sliku o tome koje su karakteristike razgovora važne, a delom kroz percepciju podataka sa računarskog stanovišta. Na ovaj način, sledeći imenovani entiteti izabrani su za klase metapodataka:

- oznake protokola koje predstavljaju jedinstvene identifikatore
- osobe koje su učestovovale u razgovoru ili se o njima govori
- različite vrste lokacija
- različite vrste datuma
- jezik na kom se govori
- etničke pripadnosti
- verske pripadnosti

Podela metapodataka na potkategorije Metapodaci, poput osoba i datuma, za potrebe ekstrakcije podeljeni su na potkategorije na sledeći način:

- osobe su podeljene po ulogama na informatore (sagovornike), ispitivače (istraživače) ili ostale (osobe koje nemaju neku od prethodnih uloga)
- datumi su podeljeni po semantici na datume snimanja materijala, na godine rođenja informatora i na ostale datume

Dodatni izvori imenovanih entiteta Za ekstrakciju imenovanih entiteta korišćeni su dodatni izvori poput:

- javno dostupnog elektronskog rečnika ličnih imena i toponima razvijenog od strane Grupe za jezičke tehnologije Matematičkog fakulteta u Beogradu

- sakupljenih ličnih imena i prezimena sa različitih lokacija i javno dostupnih lista ličnih imena na internetu
- sakupljenih naziva toponoma sa različitih javno dostupnih lista toponima na internetu
- sakupljenih i popisanih naziva etniciteta, religija, jezika koji su prisutni neposredno ili posredno na lokalitetu Balkana

Identifikovanje kontekstnih fraza za imenovane entitete i za njihove potkategorije Prva vrsta struktura koje označavaju imenovani entitet sastoje se od jedne ili više reči koje same po sebi mogu predstavljati određenu klasu (bez obzira na kontekst), i takve strukture teksta označene su kao jedna instanca te klase. Na primer, lična imena ili lokacije pronađene u elektronskim rečnicima ili listama sakupljenih ličnih imena ili toponima označavaju se kao takve ne uzimajući u obzir kontekst. Puni datumi i oznake se takođe mogu označiti bez konteksta.

Druga vrsta struktura identifikovana je i obeležena uz pomoć konteksta. Na primer, ako mesto "X" u Srbiji nije prepoznato u rečniku toponima, da bi ono bilo označeno kao lokacija potrebno je da se nađe u kontekstu koji odgovara lokaciji, na primer, "živi u", "iz", "varošica", "mesto", "naseljen u", "oženjen u", "udata u", i slično. Kasnije se taj označeni deo može dekomponovati u kontekst i reč koja je iz određene klase, u ovom slučaju klase lokacija.

Iako se puni datumi mogu označiti bez konteksta, za delimične datume je potreban kontekst da bi bili označeni. Na primer, često se godina osobe pominje bez određenih reči koje bi odredile njenu ulogu, ali se pojavljuje u kontekstu nakon klase "osoba" (na primer, "Ana Ilić, 1950", gde je "1950" godina rođenja), pa je i ta informacija prepoznata kao metapodatak o osobi.

Fraze koje se odnose na religijsku pripadnost mogu biti "pravoslavne veroispovesti", "pravoslavac", "veroispovest: pravoslavna", i slične. Fraze koje se odnose na jezik su "rumunski jezik", "govori srpski". Fraze koje se odnose na etničku pripadnost mogu biti "Srbin", "Bošnjak", "Rumuni", i slične.

Tekstualnoj strukturi se dodeljuje oznaka "rečnik", u značenju da su elektronski rečnik ili odgovarajuća lista učestvovali u njenoj identifikaciji, ili oznaka "kontekst" što znači da je struktura prepoznata pomoću konteksta. Ova metoda je primenjena na sve klase imenovanih entiteta.

Za klase imenovanih entiteta koje su dalje podeljene na potklase potrebne su dodatne metode. Na primer, osobe mogu da budu podeljene prema svojoj ulozi, lokacije i datumi mogu da budu podeljeni prema svojoj semantici. Ovde će biti razmotreno pronalaženje uloga osobe, odnosno razvrstavanje osoba u potklase po ulozi. Uloge koje su razmatrane su informator, ispitivač i ostali.

Da bi se osoba mogla smatrati informatorom, osobi mora prethoditi kontekstna fraza poput "informator [je] [:]", "ispitnik", "sagovornik", "kazivač", "razgovor [je] [vođen] sa", i slično. Može postojati i neki tekst između takve fraze i imena osobe, sve dok taj tekst nije prepoznat kao kontekstna fraza za drugu ulogu. Na primer, u rečenici "razgovor je vođen 19.06.2003. u selu Kalaž sa Anom Ilić, 1950., ovde živi dvadeset godina. Marijom Simić, 1945, doselila se iz drugog mesta, ima dve ćerke. Razgovor je vodila Marija Matić", Ana Ilić i Marija Simić su informatori.

Slično pravilo odnosi se i na pronalaženje ispitivača kojima uglavnom prethodi fraza "intervju vodio", "razgovor [je] vodila", "razgovor [je] snimala", "istraživač [je] [:]", "snimao [je]", i slično. U prethodnom primeru Marija Matić je ispitivač. Nije retkost da se ispitivači navode jedan za drugim bez ponavljanja uloge i eventualno sa nekim tekstom između.

Osobe su klasifikovane kao informatori ili ispitivači čak i ako nisu prisutni u odgovarajućim kontekstima, ali se pojavljuju u odgovarajućim delovima protokola.

Razvoj biblioteke transduktora za ekstrahovanje imenovanih entiteta Za prepoznavanje i ekstrahovanje imenovanih entiteta razvijeni su sledeći transduktori:

- "oznaka"
- "osoba rečnik", "osoba", "osoba kontekst informator", "osoba kontekst ispitivač"
- "datum pun", "datum", "datum kontekst"
- "godina", "godina kontekst"
- "lokacija rečnik", "lokacija", "lokacija kontekst"
- "jezik", "jezik kontekst"
- "etnicitet rečnik"
- "religija rečnik", "religija", "religija kontekst"

Transduktor "oznaka" prepoznaje oznaku na osnovu opisane strukture koju oznaka treba da zadovoljava. Transduktori "osoba rečnik", "lokacija rečnik", "etnicitet rečnik" i "religija rečnik" prepoznaju imenovane entitete na osnovu opisanih dodatnih izvora imenovanih entiteta. Transduktor "datum pun" prepoznaje datum koji je dat punim oblikom datuma. Transduktori "osoba", "datum", "godina" i "lokacija" prepoznaju reči koje su oblika entiteta koji opisuju, ali za koje je potreban dodatni kontekst da bi bile prepoznate kao taj entitet. Transduktori "jezik" i "religija" prepoznaju reči na osnovu dodatnih izvora imenovanih entiteta, ali je potreban kontekst da bi bile prepoznate kao imenovani entitet iz odgovarajuće klase. Transduktorima "osoba kontekst informator", "osoba kontekst ispitivač", "datum kontekst", "godina kontekst", "lokacija kontekst", "jezik kontekst", "religija kontekst" opisani su kontekst u kojima treba da se nađu odgovarajući oblici imenovanih entiteta da bi bili označeni kao takvi.

Kombinovanjem transduktora iz prethodne liste napravljeni su transduktori koji prepoznaju posmatrane imenovane entitete, a koji se načelno mogu opisati na sledeći način:

- entitet informator
  - "osoba kontekst informator" + "osoba rečnik"
  - "osoba kontekst informator" + "osoba"
- entitet ispitivač
  - "osoba kontekst ispitivač" + "osoba rečnik"
  - "osoba kontekst ispitivač" + "osoba"
- entitet ostali
  - "osoba rečnik"
- entitet datum
  - "datum pun"
  - "datum kontekst" + "datum"
  - "datum" + "datum kontekst"
- entitet godina

  - "godina kontekst" + "godina"
  - "godina" + "godina kontekst"
- entitet lokacija
  - "lokacija rečnik"
  - "lokacija kontekst" + "lokacija"
- entitet jezik
  - "jezik kontekst" + "jezik"
  - "jezik" + "jezik kontekst"
- entitet etnicitet
  - "etnicitet rečnik"
- entitet religija
  - "religija rečnik"
  - "religija kontekst" + "religija"
  - "religija" + "religija kontekst"

Dodatno, na neki od sledećih načina:

- "osoba rečnik" + "," + "godina"
- "osoba rečnik" + "(" + "godina" + ")"
- "osoba" + "," + "godina"
- "osoba" + "(" + "godina" + ")"

omogućava se prepoznavanje entiteta informatora i godine i van konteksta za svoje klase entiteta u slučaju da se nađu u neposrednom kontekstu jedno drugome.

Ekstrakcija fraza koje se odnose na tematiku

Pronalaženje skupa relevantnih tema Identifikacija tema o kojima se razgovara je složen zadatak, s obzirom na to da protokoli uključuju veoma raznolike teme, što je zapravo opšti slučaj kod ove vrste intervjua. Stoga, prvo je potrebno izvršiti sistematsku klasifikaciju obuhvaćenih tema.

Tokom razgovora sa učesnicima, istraživači su u velikoj meri koristili već postojeće upitnike iz oblasti etnologije koji se uglavnom zasnivaju na "etnolingvističkom upitniku" [219]. Cilj je bio uključiti što više informatora za razgovor o određenim temama, kako bi se prikupila terminologija koja se više ne koristi, ili retko, da bi se dokumentovali određeni dijalekti za dalje jezičko istraživanje. Informatori su izabrani tako da su imali manje ili više sačuvan dijalekt. S druge strane, mnogo puta je bilo potrebno pronaći (ili prihvatiti) druge teme zavisno od sklonosti ispitanika. Vrlo često su informatori govorili o svom ličnom razmišljanju, događajima ili običajima iz ličnog života. Iako to izlazi iz okvira definisanog upitnikom, ovakav pristup ima važnu ulogu u sticanju poverenja i pokazivanju poštovanja prema informatorima.

Zbog bogatstva etnolingvističkog sadržaja uopšte, postoji vrlo raznolik skup tema. Teme se mogu klasifikovati na više načina. Klasifikacija koja je korišćena u ovom radu zasniva se na dve klasifikacione sheme. Prva je neformalna klasifikacija Biljane Sikimić, jedne od istraživača koji su se bavili prikupljanjem multimedijalnog materijala ([24]). Prema toj klasifikaciji sve teme o kojima se govori u terenskim studijama mogu se svrstati u tri glavne grupe:

- tradicionalna kultura
- usmena istorija
- privatni život.

Ove grupe se dalje mogu podeliti na podteme:

- tradicionalna kultura - kalendar, životni ciklus (rođenje, brak, smrt), običajno pravo (deca, nasledstvo, najam, čast), svakodnevni život (farma, kuća, poslovi, domaće životinje, odeća, hrana, lečenje, zabava, život prošli i sadašnji)
- usmena istorija - pozajmica, kupovina i rad u zadruzi, sezonski radnici, migracija stoke
- privatni život - religioznost, biografske priče.

Druga klasifikacija koju smo konsultovali je tezaurus etnologije i predložen je u [125]. Deo je makrotezaurusa koji pokriva sva područja ljudskog znanja i razvijen je u Narodnoj biblioteci Srbije kao kontrolisani rečnik za bibliografske baze podataka. Etnologija ispituje nacionalnu nematerijalnu i materijalnu baštinu. Obuhvata široko područje znanja, pa prodire u druge mikrotezauruse, poput društvenih odnosa, nacionalne kulture, običajnog prava, religije, sociologije, ekonomije, geologije, terminologije i narodne literature. Jedna od potklasa etnologije je narodna privreda koja se dalje deli na domaću radinost, zanat, rezbarenje kamena, lov, profit od maslina, poljoprivredu, pčelarstvo, ribarstvo, rudarstvo, trgovinu, šumarstvo, vrtlarstvo, stočarstvo. Ova klasifikacija je veoma bogata terminima, u nekim slučajevima postoji jasna hijerarhija dok u nekim slučajevima nije u tom smislu strogo formalna. Za potrebe ovog istraživanja ova klasifikacija je bila veoma korisna, jer je sadržavala sve potrebne koncepte i veze između njih.

U ovom istraživanju, prva klasifikacija se može smatrati osnovom za grubu podelu tema, dok se druga koristi za preciziranje klasifikacione sheme.

Izbor tema koje će se obrađivati Za identifikovanje tema u intervjuima izabrana je oblast narodne privrede iz svakodnevnog života tradicionalne kulture, kao jedna od najzanimljivijih i najpotpunijih tema. Teme iz oblasti narodne privrede ka kojima je bilo usmereno ovo istraživanje su:

- domaća radinost
- lov i ribolov
- pčelarstvo
- poljoprivreda
- rudarstvo
- šumarstvo
- trgovina i
- zanatstvo.

Identifikovanje kontekstnih fraza koje se odnose na izabrane teme Primećeno je da postoji određena pravilnost u korišćenju fraza kojima se opisuje tematika narodne privrede. Uglavnom je to opis neke radnje, događaja ili običaja. Na primer, fraze koje se pojavljuju su "čuvali su stoku", "kako su obrađivali zemlju", "vršidba", "berba kukuruza", i slične.

Konstrukcija semantičkih struktura primenjena na teme iz oblasti narodne privrede Analizom teksta identifikovane su sledeće semantičke klase koje predstavljaju teme koje se obrađuju:

- *samostalni termi* - termi koji sami po sebi predstavljaju određenu temu (npr. "zemljoradnja", "stočarstvo" ili "poljoprivreda")
- *specifični termi u specifičnom kontekstu* - termi koji sami po sebi ne označavaju nužno temu (npr. "životinja", "krompir", "vinograd"), ali u kombinaciji sa specifičnim kontekstom (npr. "čuvanje", "uzgajanje", "orezivanje") predstavljaju specifičnu temu (npr. "čuvanje životinja", "uzgajanje krompira" i "orezivanje vinograda" su fraze iz teme "poljoprivreda")
- *izrazi opšteg konteksta u kombinaciji sa specifičnim termima* - skup izraza (npr. "terminologija za") koji u kombinaciji sa nekim specifičnim termom (npr. "pčele") označavaju određenu temu ("terminologija za pčele", što označava temu "pčelarstvo").

Vodeći se opisanom metodologijom identifikovanja semantičkih struktura kontekstnih fraza koje se mogu naći u tekstu razvijen je sistem po kome se konstruišu semantičke strukture za oblast narodne privrede. Semantičke strukture se sastoje od kombinacije semantičkih klasa termova i od oznake GAP. Ova oznaka predstavlja određeni broj reči koje se mogu naći između termova iz navedenih klasa.

Konstruisane su sledeće semantičke strukture po temama:

- domaća radinost
  - "domaća radinost samostalni termi"
  - "opšti kontekst" + GAP + "domaća radinost termi"
  - "domaća radinost poslovi" + GAP + "domaća radinost termi"
- lov i ribolov

  - "lov i ribolov samostalni termi"
  - "opšti kontekst" + GAP + "lov i ribolov termi"
  - "lov i ribolov poslovi" + GAP + "lov i ribolov termi"
- pčelarstvo
  - "pčelarstvo samostalni termi"
  - "opšti kontekst" + GAP + "pčelarstvo termi"
  - "pčelarstvo poslovi" + GAP + "pčelarstvo termi"
- poljoprivreda
  - "poljoprivreda samostalni termi"
  - "opšti kontekst" + GAP + "poljoprivreda alati"
  - "stočarstvo samostalni termi"
  - "životinje poslovi" + GAP + "životinje termi"
  - "zemljoradnja samostalni termi"
  - "biljke poslovi" + GAP + "biljke termi"
  - "zemljište poslovi" + GAP + "zemljište termi"
- šumarstvo
  - "šumarstvo samostalni termi"
  - "opšti kontekst" + GAP + "šumarstvo termi"
  - "šumarstvo poslovi" + GAP + "šumarstvo termi"
- trgovina
  - "trgovina samostalni termi"
  - "opšti kontekst" + GAP + "trgovina termi"
  - "trgovina poslovi" + GAP + "trgovina termi"
- zanatstvo
  - "zanatstvo samostalni termi"
  - "opšti kontekst" + GAP + "zanatstvo termi"

– "zanatstvo poslovi" + GAP + "zanatstvo termi"

U navedenoj listi, grupe "domaća radinost samostalni termi", "stočarstvo samostalni termi", "zemljoradnja samostalni termi", "poljoprivreda samostalni termi", "šumarstvo samostalni termi", "trgovina samostalni termi", "zanatstvo samostalni termi", "rudarstvo samostalni termi", "lov i ribolov samostalni termi" i "pčelarstvo samostalni termi" predstavljaju klase samostalnih terma. Grupe "domaća radinost termi", "životinje termi", "biljke termi", "zemljište termi", "poljoprivreda alati", "šumarstvo termi", "trgovina termi", "zanatstvo termi", "rudarstvo termi", "lov i ribolov termi", "pčelarstvo termi" predstavljaju klase specifičnih termova. Grupe "domaća radinost poslovi", "životinje poslovi", "biljke poslovi", "zemljište poslovi", "šumarstvo poslovi", "trgovina poslovi", "zanatstvo poslovi", "rudarstvo poslovi", "lov i ribolov poslovi" i "pčelarstvo poslovi" predstavljaju klase termova specifičnog konteksta. Na kraju, grupa "opšti kontekst" predstavlja izraze opšteg konteksta. U slučaju kada su specifični termi u velikoj meri opšti, nije dovoljno kombinovati ih sa opštim kontekstom da bi se dobila specifična tema, tako da se u tom slučaju ta kombinacija izostavlja, kao što je situacija sa klasama "životinje termi", "biljke termi" ili "zemljište termi".

Identifikovanje liste reči semantičkih klasa termova Za predstavljanje reči iz domena korišćene su vreće reči osnovih oblika podeljene na grupe, od kojih su neke date u nastavku sa primerima reči koje se mogu u njima naći:

- "opšti kontekst" - terminologija, bavljenje, raditi
- "domaća radinost samostalni termi" - vunarstvo, rukotvorina, domaća radinost
- "domaća radinost poslovi" - bojenje, pripremanje, prerada
- "domaća radinost termi" - vuna, pređa, ćilim
- "lov i ribolov samostalni termi" - lov i ribolov, pecanje
- "lov i ribolov poslovi" - koristiti, hvatati, čuvati
- "lov i ribolov termi" - udica, divljač
- "pčelarstvo samostalni termi" - pčelarstvo, pčelinjak
- "pčelarstvo poslovi" - vrste, društvo, gajenje

- "pčelarstvo termi" - pčela, košnica, med
- "poljoprivreda samostalni termi" - poljoprivreda
- "poljoprivreda alati" - traktor, vršilica, kombajn
- "stočarstvo samostalni termi" - stošarstvo, govedarstvo, ovčarstvo
- "životinje poslovi" - čuvanje, uzgoj, držanje
- "životinje termi" - kokoške, ovce, marva
- "zemljoradnja samostalni termi" - zemljoradnja, ratarstvo, žetva
- "biljke poslovi" - kopanje, gajenje, branje
- "biljke termi" - kukuruz, pšenica, krompir
- "zemljište poslovi" - imati, steći, obrađivanje
- "zemljište termi" - vinograd, voćnjak, gazdinstvo
- "rudarstvo samostalni termi" - rudarstvo, rudnik, rudar
- "rudarstvo poslovi" - iskopavanje, otkopavanje
- "rudarstvo termi" - ruda, mineral, bakar
- "šumarstvo samostalni termi" - šumarstvo, šumar
- "šumarstvo poslovi" - zaštita, gajenje, podizanje
- "šumarstvo termi" - šuma
- "trgovina samostalni termi" - trgovina, trgovati
- "trgovina poslovi" - nositi, prodavati, prodaja
- "trgovina termi" - pijaca, tezga, trgovac
- "zanatstvo samostalni termi" - zanatstvo, esnaf, zanatlija
- "zanatstvo poslovi" - izučiti, zaposlen, stolarski
- "zanatstvo termi" - zanat, abadžija, drvorezbar

Obezbeđivanje dodatnih izvora drugih oblika reči iz navedenih listi Drugi oblici reči iz prethodnih listi mogu se dobiti iz dostupnih elektronskih rečnika za one reči koje postoje u rečniku.

Za reči koje ne postoje u rečniku, moguće je uraditi razlaganje na koren reči i nastavke, što zahteva ulaganje dodatnih resursa u vidu vremena i stručnosti. Drugo rešenje je dozvoliti određene varijacije reči, kao na primer dozvoliti da poslednja trećina, ili manje ili više, od zadate reči može da odstupa od oblika koji je naveden.

Ukoliko elektronski rečnici sadrže potrebne reči najbolje je koristiti njih. Sa druge strane, za reči koje nisu u rečniku, što može biti česta situacija u slučaju reči iz domena koji je specifičan metoda dopuštanja određene varijacije reči je primenljivija.

Razvoj transduktora za ekstrakciju fraza koje se odnose na tematiku U skladu sa opisanom procedurom za stvaranje semantičkih struktura, za svaku od tema napravljen je po jedan konačni transduktor koji će prepoznavati odgovarajuće kontekstne fraze na osnovu opisanih semantičkih struktura. Na kraju, konačni transduktor za prepoznavanje svih tema iz oblasti narodne privrede predstavlja kompoziciju svih prethodnih transduktora za prepoznavanje tema:

- narodna privreda
  - "domaća radinost"
  - "lov i ribolov"
  - "pčelarstvo"
  - "poljoprivreda"
  - "rudarstvo"
  - "šumarstvo"
  - "trgovina"
  - "zanatstvo"

U fazi treniranja navedeni transduktori su u više iteracija poboljšavani kako bi model što bolje oslikao problem koji se rešava.

6.4 Metode klasifikacije teksta

Sprovedeno istraživanje u ovom delu doktorata imalo je za cilj da primeni neke odabrane pristupe klasifikaciji, da uporedi rezultate, i da ponudi pristup koji

je prilagođen tekstovima koji su bili predmet istraživanja. U radu je razmatrana klasifikacija nadgledanim statističkim metodama mašinskog učenja uz poboljšanje semantičkim metodama. Glavni razlog zbog koga su izabrane ove metode zasniva se na velikoj mogućnosti njihovog prilagođavanja problemu koji se rešava izborom odgovarajućih atributa za predstavljanje dokumenta, uzimajući u obzir obim i karakteristike podataka sa kojima se radi.

Pre predstavljanja specifičnih oblika metoda primenjenih u ovom radu, biće prikazani njihovi osnovni oblici.

Metode za predstavljanje dokumenata

U nastavku su navedene neke od metoda za predstavljanje tekstualnih dokumenata u cilju efikasnijeg izvođenja zadatka klasifikacije teksta.

Predstavljanje dokumenata *n*-gramima Za datu sekvencu tokena $S = (s_1, s_2, \ldots, s_{N+(n-1)})$ iz azbuke tokena Σ, gde su N i n pozitivni celi brojevi, n-gram sekvence S je bilo koja podsekvenca uzastopnih tokena dužine n. Sekvenca $s_i, s_{i+1}, \ldots, s_{i+(n-1)}$ je tada i-ti n-gram sekvence S ([278]).

Pojam n-grama može se definisati na nivou reči, karaktera ili bajtova. Izdvajanje, na primer, n-grama karaktera bi tada izgledalo kao pomeranje okvira dužine n karaktera duž interne prezentacije dokumenta.

Neke od prednosti korišćenja bajt i karakter n-grama u odnosu na n-grame reči u opštem slučaju u obradi prirodnog jezika su:

- nezavisnost od jezika i tematike - nije potrebno preprocesiranje ili procesiranje višeg nivoa koje nije trivijalno, kao na primer izdvajanje reči ili njihovo označavanje
- robusnost - mogu biti tolerantni u izvesnoj meri na slovne greške ili varijacije u zapisima, s obzirom na to da kod korišćenja kraćih n-grama eventualne greške mogu zahvatiti manji broj njih
- slične forme imaju zajedničke n-grame - slične reči (kao što su “metod”, “metoda”, “metodama”) imaju dosta zajedničkog kada se posmatraju kao skupovi n-grama
- efikasnost - sve potrebne informacije čitaju se u jednom prolazu.

Glavna mana upotrebe n-grama u reprezentaciji dokumenta ogleda se u tome da jedan dokument može proizvesti veoma mnogo n-grama. Kod kraćih tekstova ova osobina ne dolazi mnogo do izražaja, čak mogućnost da se dokument predstavi velikim brojem n-grama može biti i prednost.

Predstavljanje dokumenata vrećom reči Ustaljen način izbora atributa je da se termi identifikuju sa rečima. Tada se može raditi sa vrećom reči koja učestvuje u pravljenju reprezentacije tekstualnog dokumenta. Tekstualni dokument se predstavlja vrećom svojih reči uz očuvanje broja njihovog pojavljivanja ili frekvencije pojavljivanja. Vreća reči je specijalan slučaj predstavljanja dokumenta n-gramima reči za $n = 1$.

Prednost ovog modela je u tome što se često samo na osnovu ovih parametara primenom odgovarajuće metode za klasifikaciju može doneti dovoljno dobar zaključak.

Mana metode je što se gubi informacija o redosledu reči. Nekada su važni odnosi zajedničkog pojavljivanja reči, njihova značenja i drugo. Takođe, za morfološki bogate jezike, kao što je srpski jezik, sama vreća reči nije dovoljna, već je uglavnom potrebno koristiti sve oblike reči koje se pominju. Ponekad je korisno izvršiti i zamenu reči osnovnim oblikom ili korenom reči kako bi se smanjio broj različitih reči koje se obrađuju (navedene metode se nazivaju *lematizacija* i *steming*).

Izvođenje restrikcije nad početnim tekstom Kada informacije koje se traže nisu u povoljnom odnosu prema ukupnom broju informacija koje dokument sadrži, posmatranje svih reči koje se pojavljuju može zanemariti reči koje su značajne i dati prednost rečima koje nemaju toliko uticaja na klasifikaciju po posmatranom kriterijumu.

Ukoliko se mogu opisati reči od značaja, može se napraviti restrikcija nad početnim dokumentima tako da se ostave reči od značaja, uz eventualno ostavljanje i određene okoline (konteksta) ukoliko to ima smisla za problem koji se rešava, i da se na dalje dokument predstavlja samo rečima koje su ostale nakon primenjene restrikcije.

Nakon primene restrikcije, dokument se može predstaviti, na primer, nekom od opisanih metoda - n-gramima ili vrećom reči.

Predstavljanje dokumenata značajnim rečima ili značajnim frazama Dokument se može predstaviti i n-gramima značajnih reči ili značajnih fraza.

Predstavljanje dokumenta n-gramima značajnih reči se izvršava tako što se napravi skup svih značajnih reči, od njih se formiraju svi n-grami i najfrekventniji n-grami se dalje koriste kao atributi kojima se predstavlja dokument.

Predstavljanje dokumenta značajnim frazama se može opisati na sličan način s tim da se umesto pravljenja skupa značajnih reči, pravi skup značajnih fraza.

Predstavljanje dokumenata semantičkim entitetima Atributi se mogu odnositi i na drugačije vrste informacija. Na primer, može se opisati prisustvo određenih semantičkih entiteta koji su nosioci informacija. Ovi entiteti mogu biti definisani na različite načine, a u zavisnosti od njihovog značenja u tekstu i to u odnosu na klasifikaciju koja se sprovodi.

Semantički entitet je bilo kakva jedinica koja se može nedvosmisleno definisati, a koja ima neko značenje ili strukturu. Primeri semantičkih entiteta su sledeći:

- autor
- toponim
- ime i prezime
- email adresa
- lista stavki u receptu
- nabrajanje datuma
- url do fotografije
- određena fraza
- određeni struktura fraze
- broj nepraznih linija

Prvo se definišu semantički entiteti za koje se pretpostavlja da bi mogli da imaju uticaja u klasifikaciji. Potom se dokument predstavlja atributima koji označavaju prisustvo / odsustvo, broj ili frekvenciju određenog semantičkog entiteta u tekstu.

Pregled izabranih metoda klasifikacije

Metode mašinskog učenja koje su korišćene za klasifikaciju su metod potpornih vektora (Support Vector Machine SVM), k najbližih suseda (K Nearest Neighborhood kNN) i metod maksimalne entropije (Maximum Entropy MaxEnt), pa će one biti predstavljene u nastavku. Opis metoda je preuzet iz rada [94].

SVM Metoda za klasifikaciju zasnovana na SVM je potvrđena kao uspešna za rešavanje problema klasifikacije na osnovu tematike. To je nadgledana metoda mašinskog učenja koja pronalazi funkcije za ulazno-izlazno mapiranje nad labeliranim podacima iz trening skupa. Zasniva se na pronalaženju hiperravni koja deli prostor tako da se vektor koji pripada posmatranoj klasi nalazi sa jedne strane te hiperravni, dok se vektor koji ne pripada klasi nalazi sa druge strane hiperravni. Iako je početni SVM model bio napravljen za binarnu klasifikaciju, za potrebe ovog istraživanja korišćen je $SVM^{Multiclass}$ predstavljen od strane Joachim ([122], [187]). Razvijen je na osnovu modela koji su predstavili Crammer i Singer ([42]), umesto dekomponovanja problema višeklasne klasifikacije u probleme binarne klasifikacije. U toku faze treniranja, $SVM^{Multiclass}$ pronalazi rešenje za sledeći problem optimizacije:

$$\min_{w,\xi} \frac{\sum_{i=1}^{k} w_i^2}{2} + C\frac{\sum_{i=1}^{n} \xi_i}{n}$$

$$\forall i \in [1..n] \forall y \in [1..k] : [x_i \cdot w_{yi}] \geq [x_i \cdot w_y] + 100\Delta(y_i, y) - \xi_i \tag{6.1}$$

gde $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ je skup za treniranje sa labelama y_i u $[1 : k]$, C je uobičajeni parametar regularizacije greške, $\Delta(y_i, y)$ je funkcija gubitka koja ima vrednosti 0 ako je y_i jednako y, i 1 inače. Za rešavanje ovog optimizacionog problema, $SVM^{multiclass}$ koristi algoritam zasnovan na strukturnom SVM ([281]).

SVM metoda proizvodi dobre rezultate i za manje trening skupove, odnosno nije previše zahtevna u smislu veličine skupa za treniranje.

kNN Metoda koja će biti predstavljena je varijanta kNN algoritma (za $k = 1$) nad n-gramima, a koja je predstavljena od strane Kešelja i njegovih kolega ([130]) za identifikovanje autorstva tekstova. Profil jednog autora je uređeni par $(x_1, f_1), (x_2, f_2), \ldots, (x_L, f_L)$ najfrekventnijih L karakter n-grama x_i i njihovih normalizovanih frekvencija f_i. Autorstvo se određuje na osnovu razlike između dva profila poredeći najfrekventnije n-grame iz oba. Dva identična teksta će imati dva

identična skupa profila od L najfrekventnijih n-grama, a samim tim i meru različitosti jednaku nula. Različiti tekstovi će biti manje ili više slični jedan drugome na osnovu broja najfrekventinijih n-grama koje su im zajednički. Algoritam dat tabelom 6.1 predstavlja detaljan pregled koraka za ovu metodu klasifikacije n-grama zasnovanu na kNN metodi (kNNnGT).

Tabela 6.1: Algoritam metode za klasifikaciju zasnovane na kNN metodi

Treniranje&Validacija_kNNnGT(C,D(Treniranje),D(Validacija),n_min, n_max, L_min, L_max, korak_L)

Ulaz: Skup labela za kategorije C, skup dokumenata za treniranje i validaciju $D(Treniranje), D(Validacija)$, početne i konačne vrednosti (sa korakom) za treniranje parametara klasifikatora: n_min, n_max; L_min, L_max, $korak\_L$.
Izlaz: Tačnost klasifikacije

za svako n od n_min do n_max **radi**
  //Stvaranje skupa "dokumenata kategorije"
  za svako $c \in C$ **radi**
    $doc(c) \leftarrow NadoveziTekstoveSvihDokumenataIzTreningSkupaIzKategorije(D(Trening), c)$
  $D(C) \leftarrow \bigcup_{c \in C} doc(c)$
  //Za svaki dokument kategorije i dokument za validaciju napraviti njegov profil
  za svaki $doc \in D(Validacija) \cup D(C)$ **radi**
    $Ngrami(doc) \leftarrow IzdvojSveNgrame(doc, n)$
    za svako $x \in Ngrami(doc)$ **radi**
      $frekvencije[x] \leftarrow IzracunajNormalizovanuFrekvenciju(x, doc)$
    $Profil(doc) \leftarrow IzlistajNGrameOpadajuePoFrekvenciji(\bigcup_{x \in Ngrami(doc)}(x, frekvencije[x]))$
  za svako L od L_min do L_max sa korakom $korak\_L$ **radi**
    //Skratiti profile kategorija i validacija na dužinu L
    za svaki $doc_t \in D(Validacija)$ **radi**
      $Profil_L(doc_t) \leftarrow Profil(doc_t)|L$
    za svaki $c \in C$ **radi**
      $Profil_L(doc(c)) \leftarrow Profil(doc(c))|L$
    //Izračunati mere različitosti između profila kategorija i validacije
    za svaki $doc_v \in D(Validacija)$ **radi**
      za svaki $doc(c) \in D(C)$ **radi**
        $razlika_{vc} \leftarrow MeraRazlike(Profil_L(doc_v), Profil_L(doc(c)))$
      //Izbor najsličnije kategorije (ili kategorija)
      $c(doc_v) \leftarrow argmin_{c \in C} razlika_{vc}$
    Izračunati tačnosti izvedene kategorizacije
izdvojiti n' i L' koji proizvode najveću tačnost
vrati tačnost za n' i L'

Može se primetiti da mera razlike igra važnu ulogu. Ovde je korišćena mera razlike bazirana na relativnoj udaljenosti koja je predstavljena u [130]:

$$d(\mathcal{P}_1, \mathcal{P}_2) = \sum_{x \in Profil} \left(\frac{2 \cdot (f_1(x) - f_2(x))}{f_1(x) + f_2(x)} \right)^2 \tag{6.2}$$

gde su $f_1(x)$ i $f_2(x)$ frekvencije n-grama x u profilu kategorije $\mathcal{P}_1$ i profilu dokumenta za validaciju $\mathcal{P}_2$, ovim redom.

MaxEnt Model maksimalne entropije je nadgledana metoda mašinskog učenja za predviđanje verovatnoće raspodela labela podataka y maksimiziranjem sledeće funkcije entropije (6.3) koja odgovara podacima za treniranje x, odnosno zadovoljava zadate uslove:

$$H(p) = -\sum_{x,y} p(x)p(y|x)\ logp(y|x) \tag{6.3}$$

Može se pokazati da postoji jedinstveni model p^* sa maksimalnom entropijom u okviru datih ograničenja, u kom slučaju se dobija:

$$p^* = argmax_p\ H(p) \tag{6.4}$$

U ovom istraživanju korišćena je *SharpEntropy* biblioteka, koja je deo SharpNLP, putem C# Apache OpenNLP, a koja se koristi u [183].

Poređenje kvaliteta klasifikacije za različite reprezentacije teksta i metode klasifikacije

U istraživanju koje je opisano u radu [94], u cilju poređenja različitih reprezentacija dokumenata zasnovanih na korišćenju n-grama izvršen je skup eksperimenata koji koriste različite metode mašinskog učenja. Cilj je bio istražiti kako se različite metode ponašaju u rešavanju istog problema, pod istim uslovima. U tu svrhu, za sve metode korišćen je isti ulazni skup podataka, skup za treniranje i testiranje napravljen je na identičan način uz korišćenje 10-tostruke kros validacije i za svaki eksperiment je upotrebljena ista dužina n-grama za sve tipove atributa.

Pod ovim uslovima, mogu se ispitati sledeće osobine: kako se korišćene metode ponašaju kada se menja tip atributa (bajt, karakter ili reč), kako se menjaju rezultati kada se menja glavni parametar atributa - njegova dužina, u slučaju kada se metod i tip atributa ne menjaju, i ako postoji metod, tip atributa ili dužina atributa koji su optimalni za sve skupove tekstova, koji tip atributa je u tom slučaju korišćen.

Izbor atributa se pravi automatski i bez korišćenja dodatnog znanja o domenu koji se obrađuje.

Primena

Sprovedena su istraživanja kojima je ispitivano kako različiti tipovi n-grama (bajt, karakter, reč), njihove različite dužine, i različite metode (SVM, kNN, MaxEnt) utiču na kvalitet klasifikacije različitih skupova tekstova u istom domenu (filmske recenzije) na različitim jezicima (engleski, španski, arapski, francuski, češki, turski i srpski) ([94]). Tekstovi su inicijalno bili podeljeni na dva skupa (pozitivno, negativno) u odnosu na sentiment. U slučaju bajt i karakter n-grama korišćene su

dužine od 2 do 9, dok su u slučaju n-grama reči korišćene dužine od 1 do 3, za sve metode. Upotrebljeni su opcioni parametri za redukciju šuma i nedostajućih vrednosti. Za kNN parametar klasifikacije L (broj najfrekventnijih n-grama) uzimao je vrednost od 1000 do 60000, sa korakom 1000. U MaxEnt metodi, SharpNLP GisModel koristio se u 100 iteracija i vrednost odsecanja je bila 5 (izbacuje se svaki n-gram koji se ne pojavljuje bar 5 puta).

Eksperimentalno je utvrđeno da se kao najuspešnije statističke metode mašinskog učenja pokazuju one koje koriste n-grame bajtova i karaktera. Veći broj drugih metoda sa kojima se vršilo poređenje koristi BOW model, dok metode u ovom radu koriste samo preprocesiranje koje obuhvata uklanjanje interpunkcije, stop reči, ispravljanje slovnih grešaka i ujednačavanje veličine slova.

Iz dobijenih rezultata i poređenja sa metodama iz drugih istraživanja, može se zaključiti da hibridne metode koje kombinuju semantičke metode sa statističkim metodama mašinskog učenja imaju bolje rezultate od metoda koje koriste samo statistički pristup. To dalje navodi na zaključak da bi verovatno i metode predstavljene u ovom istraživanju dale bolje rezultate kada bi bile kombinovane sa nekim semantičkim metodama.

Primena opisane metode poređenja nad navedenim skupovima tekstova omogućila je poređenje rezultata sa drugim objavljenim sprovedenim istraživanjima nad istim skupovima tekstova, čime je postignuta opštost i uporedivost rezultata. Zaključci koji su izvedeni upotrebljeni su, potom, za izbor metode za klasifikaciju tekstualnih protokola na srpskom jeziku.

6.5 Primena metoda klasifikacije teksta na tekstualne protokole

Cilj sprovedenog istraživanja je bio da prikaže mogućnost binarne klasifikacije tekstualnih protokola u odnosu na postojanje tematike iz oblasti narodne privrede.

Tekstualni protokoli su takvi da uglavnom ne preovladava jedna tema tokom celog sadržaja. Može se reći da je češći slučaj da jedan protokol sadrži čitav spisak tema. Sa druge strane, samo jedno pojavljivanje neke fraze može da indicira da se protokol klasifikuje u neku određenu klasu, iako može sadržati i veoma mnogo drugog teksta koji nema dodirnih tačaka sa tom temom. Još jedna karakteristika ovakvih tekstova je da ih nema u velikom broju, što je od ključne važnosti za neke metode

mašinskog učenja. Opisane osobine usložnjavaju problem automatske klasifikacije tekstualnih protokola zasnovane na tematici.

Za potrebe klasifikacije tekstualnih protokola po tematici na raspolaganju su metode kao što su nadgledane statističke metode mašinskog učenja, nenadgledane semantičke metode, metode dubokih ili drugih neuronskih mreža, kao i hibridna metoda koja bi istakla i objedinila prednosti nekih od prethodnih metoda.

Skup tekstualnih protokola sa kojima se radi je specifičan po sadržaju i mali po obimu. Statističke metode mašinskog učenja, poput kNN, MaxExt i SVM, daju dobre rezultate za zadatak klasifikacije tekstova na srpskom jeziku na osnovu skupa za treniranje koji ne mora biti preveliki. Sa druge strane, neki bolji rezultati primenom metoda neuronskih mreža mogu se očekivati tek sa znatno većim skupom tekstova.

Na osnovu prethodnih razmatranja, za potrebe ovog istraživanja izabrano je da se demonstriraju mogućnosti SVM metode za zadatak klasifikacije tekstualnih protokola po tematici, kao metode koja se u eksperimentima pokazala uspešnom u klasifikaciji tekstova na srpskom jeziku. Dobijeni rezultati iz prethodnog istraživanja, čiji je cilj bio poređenje uspešnosti različitih metoda pri korišćenju različitih tipova n-grama, upotrebljeni su za izbor odgovarajućeg tipa n-grama za reprezentaciju tekstualnih protokola. Dodatno, na osnovu sprovedenog poređenja sa drugim aktuelnim istraživanjima i metodama, uz zaključak koji ukazuje na to da bi upotreba semantičkih metoda uz statističke metode potencijalno dala bolje rezultate, testirana je mogućnost poboljšanja statističke metode semantičkim metodama.

Sprovedeni su eksperimenti klasifikacije teksta pri predstavljanju protokola n-gramima bajtova, karaktera i reči, pri čemu su rezultati pri predstavljanju n-gramima karaktera za $n = 5$ bili najbolji, tako da je na dalje razmatran samo taj slučaj. Rezultati su navedeni u poglavlju o rezultatima. Nakon toga, primenjene su i metode za poboljšanje rezultata, a koje su opisane u nastavku.

Klasifikacija teksta pri predstavljanju dokumenata n-gramima karaktera

Primenjena je restrikcija nad polaznim tekstom dokumenta tako što su izostavljene sve reči za koje ne postoji reč iz liste značajnih reči kojoj je slična, uz ostavljanje okoline od m reči, gde je optimalno m utvrđeno eksperimentalno, pri čemu su za značajne reči uzete sve one koje se nalaze u listama reči koje se koriste u semantičkim strukturama za opis fraza tematike narodne privrede.

Funkcija sličnosti koja je upotrebljena zasniva se na poklapanju N početnih karaktera reči D za koju se računa sličnost, sa reči L iz liste značajnih reči, pri čemu se N određuje na sledeći način (dl je dužina reči L):

$$N = \begin{cases} dl, & \text{ako je } dl < 3 \\ ceo\_deo(\frac{2}{3}dl), & \text{inače.} \end{cases} \tag{6.5}$$

Sličnost reči D sa reči L se tada računa na sledeći način:

$$slicna(D, L) = prvih\_n\_karaktera(D, N) == prvih\_n\_karaktera(L, N) \tag{6.6}$$

Može se primetiti da ova funkcija nije simetrična. Na opisani način dobija se da su reči "metodologija", "metoda", "metod", slične reči "metodama". Sa druge strane, ni jedna od ovih reči nije slična sa "metodologija" (osim nje same), ali je njoj slična reč "metodologijama".

Nakon toga, razmatrani su n-grami karaktera za $n = 5$ i to iz skupa napravljenih od sledećih izvora:

- *n-grami iz sopstvenih reči* - dobijaju se iz svih reči dobijenih restrikcija protokola iz skupa za treniranje koji su označeni kao pozitivni
- *n-grami iz značajnih reči* - izračunavaju se na osnovu svih značajnih reči
- *n-grami iz značajnih fraza* - izračunavaju se na osnovu svih značajnih fraza, gde su značajne fraze one koje se mogu dobiti primenom razvijenih semantičkih struktura.

Za svaki od f najfrekventnijih n-grama uvodi se po jedan atribut koji predstavlja pojavljivanje ili odsustvo tog n-grama. Za svaku od tri opisane grupe n-grama izvedeni su pojedinačni eksperimenti i rezultati su prikazani u poglavlju sa rezultatima.

Klasifikacija teksta pri predstavljanju dokumenata semantičkim entitetima

Potrebno je uočiti na koje sve načine se informacije o tematici mogu dobiti iz sadržaja protokola.

Struktura protokola je takva da se informacije o tematici često mogu naći u okviru nekakve liste tema, sa ili bez vodećeg rednog broja teme u razgovoru. Na primer, sledeće linije su indikatori da se radi o tematici narodne privrede:

- 1. vršidba
- 2. sejanje žita
- Tema: pletenje
- Pletenje, pokazivanje radova

Informacija o temi može se naći i van liste u okviru nekog drugog teksta.

Na osnovu prethodnih zapažanja, identifikovane su dve vrste semantičkih entiteta koji mogu pomoći u otkrivanju teme, i oni se mogu opisati na sledeći način:

- *pozicioni semantički entiteti* - entiteti koji označavaju pojavljivanje reči iz domena u nekoj opisanoj strukturi tako da ulogu igra pozicija posmatrane strukture u tekstu
- *kontekstni semantički entiteti* - entiteti koji označavaju pojavljivanje određenih reči iz domena u određenom kontekstu, pri čemu nije važna pozicija pojavljivanja u tekstu.

Za definisanje pozicionih i kontekstnih semantičkih entiteta takođe je upotrebljena metoda prepoznavanja tema zasnovana na konstrukciji semantičkih struktura, koja je opisana u poglavlju 6.1, s tim da je ovde iskorišćen i sledeći korak, odnosno kontekstni semantički entiteti treba da budu oblika i sadržaja semantičkih struktura opisanih u navedenoj metodi, dok pozicioni semantički entiteti mogu da, pored pravila za kontekstne semantičke entitete, uzmu vrednosti i specifičnih terma, uz dodavanje pozicionih ograničenja u svakom slučaju.

Dokument se tada predstavlja atributima koji označavaju prisustvo ili odsustvo semantičkog entiteta na koji se odnose. Odnosno, za svaki semantički entitet definiše se po jedan atribut.

Dodatan metod poboljšanja koji je primenjen je razmatranje različitih funkcija sličnosti prilikom formiranja semantičkih atributa, pri čemu su razmatrane dve mogućnosti:

- *poređenje na podudarnost* - traži se tačno navedeni oblik reči, odnosno sličnost je do na potpuno poklapanje
- *poređenje na sličnost* - traže se i slične reči navedenih reči, pri čemu je sličnost uvedena u potpoglavlju iznad.

U eksperimentima su korišćeni pozicioni atributi i kontekstni atributi i zasebno i udruženi, pri čemu je ispitano ponašanje pri korišćenju obe predstavljene metode za izračunavanje funkcije sličnosti.

Klasifikacija teksta pri predstavljanju dokumenata hibridnom metodom

Hibridna metoda se sastoji od kombinovanja prethodna dva načina za predstavljanje dokumenata. Dokument se predstavlja unijom svojih atributa, koji tada mogu biti:

- *semantički atributi* - atributi koji označavaju pojavljivanje ili odsustvo semantičkih entiteta
- *atributi n-grama* - atributi koji predstavljaju pojavljivanje ili odsustvo odgovarajućeg n-grama

Ovako definisani atributi imaju za cilj da pokriju u što većoj meri i što je moguće preciznije prethodno opisane načine na koje se informacije o temi mogu pojaviti u tekstu. Eksperimentalno je utvrđeno da svaki od ovih skupova atributa daje određeni doprinos u poboljšanju kvaliteta klasifikacije pri njihovom zajedničkom korišćenju, u odnosu na metode gde se oni koriste samostalno, što će biti opisano u poglavlju sa rezultatima.

6.6 Metode semantičke anotacije i organizovanja multimedijalne kolekcije dokumenata u bazu podataka

Početak rada na problemu organizacije multimedijalne kolekcije dokumenata o kulturnom nasleđu Balkana sastojao se od razgovora sa ekspertima Balkanološkog instituta koji su za početak imali za cilj da se dobro formuliše i prenese problematika koju je potrebno rešiti. Tokom više iteracija razgovora bile su izložene specifičnosti same građe kao i problem organizacije i pretrage. Zajedničkim analizama došlo se do skupa funkcionalnosti i početnog oblika interfejsa koji je pogodan za rešavanje ovog problema.

Kao sledeći korak, identifikovani su tipovi dokumenata sa kojima se radi, a to su na prvom mestu tipovi poput audio, video, fotografije i teksta. Zatim su identifikovane postojeće veze među dokumentima, kao i skup veza koji je potrebno očuvati i u novoj organizaciji. Analizirano je šta je to što je potrebno za efektivno i efikasno korišćenje date kolekcije, a što nije bilo optimalno rešeno do tada i što je sputavalo iskorišćenje punog potencijala ovako vredne kolekcije.

Na prvom mestu isticala se potreba za efikasnim i intuitivnim sistemom za pretragu dokumenta po određenim kriterijumima. Da bi se rešio taj zadatak predložena je anotacija dokumenata po parametrima od koristi i organizacija u sistem baze podataka. Za određivanja parametara po kojima je korisno vršiti pretragu, razmatrane su osobine svake od vrsta dokumenata, informacije koje nose i mogući načini za pretraživanje.

U nastavku je navedena skica rešenja koje je realizovano. Ovo rešenje je primenljivo i za druge vrste multimedijalnih kolekcija sa sličnim problemima i potrebama.

Vrste materijala i njima pridruženi atributi

Identifikovane su određene karakteristike (atributi) koje su od značaja za svaku vrstu dokumenata. Ispostavilo se da postoje atributi koji su zajednički za sve vrste materijala, kao i oni koji su primenljivi samo na određenu vrstu materijala.

Audio i video materijali imaju svoj skup atributa (na primer, učesnici, lokacije, datumi, teme), fotografije imaju svoj skup atributa (na primer, lokacija, datum, opis, sadržaj), tekstualni dokumenti takođe svoj (na primer, bibliografski podaci, lokacija, opis). Tekstualni dokumenti koji su propratni dokumenti nekim drugim dokumentima (protokoli i transkripti) se ne označavaju kao zasebni materijali, već se pridružuju materijalu na koji se odnose. Sve to zahtevalo je da se napravi shema kojom će se modelovati osobine za svaki od navedenih tipova materijala.

Osim potrebe za različitim skupovima atributa kod različitih tipova materijala, uočena je potreba da audio i video materijali mogu da imaju različite karakteristike na različitim segmentima u toku njihovog trajanja. Postoje određeni atributi koji su zajednički za ceo materijal, kao što su lokacija snimanja, datum, ispitivač, moguće je i da sagovornik bude jedan tokom trajanja celog materijala. Sa druge strane, teme, opisi, drugi sagovornici mogu da budu promenljivi od segmenta do segmenta. Stoga je predloženo da se napravi mogućnost segmentacije materijala tako da različiti segmenti imaju različite vrednosti za metapodatke.

Kao rešenje napravljene su dve liste atributa:

- osnovna lista atributa, nazvana Globalne informacije i
- lista atributa za segmente, nazvana Segmentne informacije.

Svaki audio ili video materijal ima mogućnost pridruživanja atributa iz osnovne liste atributa. Svakom audio ili video materijalu može se dodati jedan ili više segmenata, čime se dodaje mogućnost pridruživanja informacija iz liste atributa za segmente, tako da svaki segment može imati sopstveni skup pridruženih informacija.

Anotacija nekog materijala ne mora se vršiti odjednom, što je takođe jedna od potreba domena sa kojim se radi. Kako jedan materijal ima mnoštvo raznovrsnih metapodataka veoma često se ne znaju svi podaci odjednom ili je jedna osoba zadužena za jedne metapodatke, a druga za neke druge, na primer, u zavisnosti od svoje ekspertize.

Rešenje je da, osim mogućnosti unosa novog materijala, postoji i mogućnost izmene materijala koji je već anotiran.

Automatizovana anotacija

U osnovnom obliku anotacije sve vrednosti kojima se anotira materijal dodeljuju se ručno (osim prostornih atributa koji se unose ranije zbog pridruženih prostornih podataka). Vrednosti za atribute biraju se iz listi ili upisuju ručno prilikom anotacije. Sve ovo je veoma zahtevan posao, prvenstveno u pogledu truda i vremena da se sve te informacije pronađu, a zatim i da se unesu u sistem.

U pomoć dolazi prikazani sistem za ekstrakciju metapodataka (imenovanih entiteta i tema), kojima se anotira dokument, kao i za klasifikaciju tekstualnih protokola na osnovu teme. Ovaj sistem u velikoj meri doprinosi kvalitetu procesa anotacije, pri čemu su neki od direktnih načina na koji može doprineti sledeći:

- ekstrahovane vrednosti iz protokola mogu se uneti automatski u liste ili polja za unos umesto da se unose ručno
- iz odgovarajućeg protokola mogu se dobiti sve ekstrahovane vrednosti i njima automatski anotirati materijal; čovek svakako treba da proveri valjanost automatske anotacije i da izvrši eventualno potrebne ispravke, zbog čega se konačni proces naziva polu-automatska anotacija
- kada je u fokusu interesovanja neka posebna tematika i kada je potrebno pronaći neoznačene materijale koji se bave tom tematikom, radi pregledanja ili

anotiranja, sistemom za klasifikaciju po tematici mogu se dobiti protokoli koji se bave tom tematikom, a samim tim mogu se pronaći i materijali na koje se odnose dobijeni protokoli.

Indirektnih doprinosa potencijalno ima i znatno više, što može biti zanimljiva tematika za buduća istraživanja.

6.7 Metode pretrage multimedijalne baze podataka

Za rešavanje zadatka pretraživanja informacija potrebno je prvo rešiti zadatak indeksiranja dokumenata, a zatim osmisliti strategije za postavljanje upita za pretraživanje baze koje je adekvatno domenu koji se obrađuje.

U ovom radu, zadatak indeksiranja je sproveden iz pomoć ekstrakcije informacija, klasifikacije teksta i anotacije dokumenata. Informacije koje su ekstrahovane i klase koje su izračunate su osnov za automatsku anotaciju metapodacima, dok je ručna anotacija vid provere i dodeljivanja dodatnih metapodataka koji nisu dodeljeni automatskom anotacijom.

Kulturno nasleđe je prožeto važnim aspektima ljudskih života, a oni se prirodno prostiru kroz prostor i vreme, interakciju sa drugim ljudima, kulturama, običajima i različitim unutrašnjim i spoljnim uticajima. Nematerijalno kulturno nasleđe je veoma bogato semantikom i brojnim isprepletanim implicitinim i eksplicitnim vezama među podacima. Zbog toga je zadatak osmišljavanja adekvatnih upita složen proces i gotovo uvek ima mesta za poboljšanja.

Osim dobijanja konkretnih dokumenata, postoji mogućnost razvoja funkcionalnosti koje bi stvarale novo znanje, na primer, izračunavanje različitih statistika, kvantitativno predstavljanje različitih aspekata od značaja, pronalaženje implicitnih zakonitosti koje važe ili praćenje određenih izabranih karakteristika u odgovarajućem okviru.

Prikaz rezultata bi se mogao sprovesti pomoću nekih od brojnih metoda vizuelizacije podataka. Na prvom mestu je prikaz multimedije na konvencionalan način, putem pregledavanja i preslušavanja video materijala, preslušavanja audio materijala, pregledavanja fotografija i čitanja tekstualnih dokumenata. Sa druge strane, izvedeno znanje bi se moglo predstaviti pomoću različitih grafikona ili dijagrama. Praćenje, na primer, prostornih karakteristika koje su ekstrahovane iz materijala

moglo bi se izvesti njihovim prikazom na geografskoj mapi. Praćenje vremenskih karakteristika moglo bi se vizuelizovati na vremenskoj osi. Navedenim prostornim i vremenskim informacijama mogu se pridružiti i neke druge karakteristike u cilju praćenja njihovog razvoja kroz prostor i vreme, kao na primer, jezičke konstrukcije koje su u upotrebi ili određena praktikovanja poput izvođenja običaja, načina spremanja hrane ili poštovanja određenih socijalnih normi ponašanja.

Metode za obradu protokola koje su razvijene i primenjene u okviru ove disertacije otvaraju brojne mogućnosti za pretragu koje poseduje kolekcija materijala u organizaciji koja je razvijena u ovom istraživanju. Postoje i brojne dodatne metode koje mogu pomoći da se potencijal kolekcije ispolji i upotrebi na najbolji način.

Razvijenim metodama obrade tekstualnih protokola uz kombinaciju sa razvijenim sistemom za organizaciju multimedijalne kolekcije u multimedijalnu bazu podataka, otvorene su mogućnosti za pretragu multimedijalne baze o kulturnom nasleđu Balkana, koja se može sistematizovati podelom na:

- pretragu po tekstualnom kriterijumu na osnovu semantičke anotacije koja je sprovedena
- pretragu po tematici na osnovu sprovedene klasifikacije
- pretragu po prostornom kriterijumu izborom lokacije na interaktivnoj mapi geografskih lokacija
- pretragu po vremenskom kriterijumu izborom vrednosti na vremenskoj osi
- pretragu po kombinaciji prethodno navedenih vrsta kriterijuma.

Ova lista svakako nije konačna, jer je takva i priroda materijala sa kojima se radi - veoma bogata različitim tipovima informacija i semantikom koja se može sagledavati iz različitih aspekata, u različitim nivoima detaljnosti i načinima prikaza, za različite vrste korisnika i drugo. Sistem za organizovanje nematerijalnog kulturnog nasleđa je takav da se uvek može dograđivati novim funkcionalnostima.

Do trenutka pisanja ovog teksta, implementirane su prve tri vrste pretrage, dok su poslednje dve u planu za budući rad na sistemu. Za pretragu po vremenskom kriterijumu napravljena je dobra početna osnova u vidu ekstrakcije vremenskih entiteta. Za dalji rad, potrebno je uraditi normalizaciju vremenskih entiteta i idejno rešiti problem funkcionalnog predstavljanja vremenske ose, kao i tehnički izvesti projekat. Poslednja vrsta pretrage navedena u prethodnoj listi može biti sprovedena

na više načina, kao disjunkcija, konjunkcija ili neka od drugih mogućih kombinacija prethodnih vrsta pretrage. Rešavanje ovog problema bilo bi preporučljivo izvesti u konsultaciji sa korisnicima čije bi povratne informacije o načinu korišćenja sistema za pretragu bile od neizmerne i neophodne pomoći. Nakon toga bilo bi potrebno pristupiti tehničkom izvođenju rešenja do koga bi se došlo prethodnim analizama.

Glava 7

Mapa nematerijalnog kulturnog nasleđa Balkana

Ovo poglavlje je posvećeno modelu koji je razvijen za rad sa prostornim podacima. Pri njegovom razvoju vodilo se računa o efikasnom skladištenju i organizaciji geografskih podataka. Inicijalno skladištenje je sprovedeno korišćenjem GML formata, dok je za organizaciju i rad sa prostornim podacima korišćen PostGIS, prostorno proširenje PostgreSQL baze podataka. Za demonstraciju prikaza mape korišćen je SVG format.

Rad sa prostornim podacima je prikazan na primeru mape Srbije.

7.1 Skladištenje prostornih podataka

Početni oblik podataka u XML i GML formatu

Prostorni podaci se čuvaju u XML fajlu, u okviru GML tagova iz "gml" prostora imena ([82]). Od prostornih entiteta se koriste granica, opština, grad i reka. Ovi entiteti su modelovani tipom `gml:FeaturePropertyType`. Skupovi istovetnih entiteta, poput skupa svih granica, svih opština, gradova ili reka, modeluju se tipom `gml:FeatureArrayPropertyType`. Skup svih nizova entiteta čini jednu kolekciju entiteta, za koju postoji odgovarajući tip `gml:FeatureCollectionType`.

Entiteti su opisani svojim osobinama poput naziva, ispisa naziva, tipa i geometrije. Osim geometrije, sve ostale osobine se čuvaju osnovnim tipom podataka poput stringa ili celog broja. Elementu `geometrija` se pridružuje "gml" tip nekih od osnovnih geometrija. Na primer, geometrija reke je tipa `gml:LineString`, geometri-

ja grada je `gml:Point`, geometrija granice je `gml:Polygon`. U shemi je uz korišćenje već definisanog prostora imena navedeno kog je tipa koja geometrija, tako da je dovoljno u definiciji samo navesti tip iz "gml" prostora imena, čime ona automatski dobija osobine pridruženog tipa. Ovakvom organizacijom je veoma olakšano navođenje kroz dokument i razumevanje strukture dokumenta.

XSD shema Deo sheme koji odgovara fajlu u kome se čuvaju prostorni podaci je sledeća:

```
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
        xmlns:svg="http://www.w3.org/2000/svg"
        xmlns:gml="http://www.opengis.net/gml"
        xmlns:xsi="http://www.opengis.net/gml base/feature.xsd">
    <xs:import namespace="http://www.opengis.net/gml"
        schemaLocation="http://schemas.opengis.net/gml/3.1.1/base/gml.xsd"/>

    <!--Definicije elemenata koji se mogu pojaviti u dokumentu-->
    <xs:element name="mapa_Srbije" type="TipMapeSrbije"/>
    <xs:element name="reke"        type="TipNizaReka"/>
    <xs:element name="reka"        type="TipReka"/>
    <xs:element name="granice"     type="TipNizaGranica"/>
    <xs:element name="granica"     type="TipGranica"/>
    <xs:element name="gradovi"     type="TipNizaGradova"/>
    <xs:element name="grad"        type="TipGrada"/>

    <!--Definicije tipova-->
    <!--Tip za koreni element koji sadrzi sve ostale-->
    <xs:complexType name="TipMapeSrbije">
        <xs:complexContent>
            <xs:extension base="gml:FeatureCollectionType">
                <xs:sequence>
                    <xs:element name="reke"      minOccurs="0"/>
                    <xs:element name="granice"   minOccurs="0"/>
                    <xs:element name="gradovi"   minOccurs="0"/>
                </xs:sequence>
            </xs:extension>
        </xs:complexContent>
```

```
    </xs:complexType>
    ...
    <!--Tipovi za niz gradova i za grad-->
    <xs:complexType name="TipNizaGradova">
        <xs:complexContent>
            <xs:extension base="gml:FeatureArrayPropertyType">
                <xs:sequence>
                    <xs:element name="grad"
                    minOccurs="0" maxOccurs="unbounded"/>
                </xs:sequence>
            </xs:extension>
        </xs:complexContent>
    </xs:complexType>
    <xs:complexType name="TipGrada">
        <xs:complexContent>
            <xs:extension base="gml:FeaturePropertyType">
                <xs:sequence>
                    <xs:element name="geometrija"
                    type="gml:PointType"/>
                    <xs:element name="naziv"
                    type="TipNaziva" minOccurs="0"/>
                </xs:sequence>
                <xs:attribute name="id" type="xs:string"/>
                <xs:attribute name="univerzitet" type="xs:boolean"/>
            </xs:extension>
        </xs:complexType>

</xs:schema>
```

XML dokument Primer dela dokumenta sa prostornim podacima koji odgovara datoj shemi je sledeći:

```
<?xml version="1.0" encoding="UTF-8" ?>
<?xml-stylesheet type="text/xsl" href="moj_gml.xsl"?>
<mapa_Srbije
xmlns:gml="http://www.opengis.net/gml"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
```

```
xmlns:xlink="http://www.w3.org/1999/xlink"
xsi:noNamespaceSchemaLocation="moj_gml.xsd">
<gradovi>
    <grad id="beograd" univerzitet="1">
        <geometrija srsName="EPSG:4326">
            <gml:coord>
                <gml:X>44.820991</gml:X>
                <gml:Y>20.456553</gml:Y>
            </gml:coord>
        </geometrija>
        <naziv id="grad59" x="44.820999" y="20.456559">
        Beograd</naziv>
    </grad>
    ...
</gradovi>
...
</mapa_Srbije>
```

Povezivanje sa bazom PostgreSQL i modulom PostGIS

Pravljenje baze Na početku je potrebno prostorno osposobiti PostgreSQL bazu uz moduo PostGIS, a zatim formirati upite kojima se pravi shema zajedno sa pravljenjem tabela i njihovim popunjavanjem podacima. Upiti za pravljenje tabela sa prostornim atributima su:

```
CREATE DATABASE mapa;
CREATE TABLE mapa.reke (ime varchar(20), utoka varchar(20));
SELECT AddGeometryColumn('mapa', 'reke', 'geometrija', 4326, 'LINESTRING', 2);
CREATE TABLE mapa.granice (ime varchar(20), tip varchar(20));
SELECT AddGeometryColumn('mapa', 'granice', 'geometrija', 4326, 'POLYGON', 2);
CREATE TABLE mapa.gradovi (ime varchar(20), univerzitet varchar(20));
SELECT AddGeometryColumn('mapa', 'gradovi', 'geometrija', 4326, 'POINT', 2);
```

Formiranje upita za unos podataka XSLT pravilima se informacije o geometrijama iz XML fajlova transformišu u format pogodan za čuvanje u PostGIS bazi podataka. Primer XSLT izraza kojim se izvršavaju transformacije je:

```
<?xml version="1.0" encoding="utf-8"?>
<xsl:transform xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
        xmlns:gml="http://www.opengis.net/gml" version="1.0" >
<xsl:output method="text" indent="yes"/>
  <xsl:template match="/">
  <!--Za svaki grad se unose podaci-->
  <xsl:for-each select="//grad">
    INSERT INTO gradovi (ime, geometrija, univerzitet)
      VALUES (
          "<xsl:value-of select="@id"/>",
          ST_GeomFromGML("<gml:Point>
            <gml:coord>
                <gml:X>
                    <xsl:value-of select="./geometrija/gml:coord/gml:X"/>
                </gml:X>
                <gml:Y>
                    <xsl:value-of select="./geometrija/gml:coord/gml:Y"/>
                </gml:Y>
            </gml:coord>
          </gml:Point>"),
          <xsl:value-of select="@univerzitet"/>
      );
  </xsl:for-each>
  ...
  </xsl:template>
</xsl:transform>
```

7.2 Grafički prikaz podataka

Na osnovu postojećih podataka u XML formatu, sledećim XSLT transformacijama se može formirati i SVG format za grafički prikaz podataka:

```
<?xml version="1.0" encoding="utf-8"?>
<xsl:transform xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
xmlns:gml="http://www.opengis.net/gml" version="1.0" >
  <xsl:template match="/">
    <svg width="550" height="600" viewBox="0 0 600 550"
```

```
xmlns="http://www.w3.org/2000/svg"
xmlns:xlink="http://www.w3.org/1999/xlink" >

<defs>
    <symbol id="maliGrad" overflow="visible">
        <circle cx="0" cy="0" r="2"
            style="fill:gray;fill-opacity:0.8;
            stroke:rgb(0,0,0);stroke-width:1"
        />
    </symbol>
    <symbol id="glavniGrad" overflow="visible">
        <circle cx="0" cy="0" r="3"
            style="fill:gray;fill-opacity:0.8;
            stroke:rgb(0,0,0);stroke-width:1"
        />
    </symbol>
</defs>

<g id="granice">
  <!--Za svaku granicu se ispisuje njena geometrija-->
  <xsl:for-each select="//granica">
    <!--Ako je drzavna, linija je vece sirine-->
    <xsl:if test="@tip='drzavna'">
      <polyline fill="none" stroke="#f77777" stroke-width="2"
      points="{./geometrija//gml:posList}"/>
    </xsl:if>
    <!--Ako je pokrajinska, linija je standardne sirine-->
    <xsl:if test="@tip='pokrajinska'">
      <polyline fill="none" stroke="#f77777" stroke-width="0.5"
      points="{./geometrija//gml:posList}"/>
    </xsl:if>
  </xsl:for-each>
</g>
<g id="gradovi">
  <!--Za svaki grad se iscrtava simbol na njegovoj poziciji-->
  <xsl:for-each select="//grad">
    <xsl:choose>
```

```
          <xsl:when test="@id='beograd'">
            <use id="{@id}" xlink:href="#glavniGrad"/>
          </xsl:when>
          <xsl:otherwise>
            <use id="{@id}" x="{./geometrija/gml:coord/gml:X}"
            y="{./geometrija/gml:coord/gml:Y}" xlink:href="#maliGrad"/>
          </xsl:otherwise>
        </xsl:choose>
      </xsl:for-each>
      </g>
      <g id="tekstgrada">
        <!--Za svaki grad se ispisuje njegovo ime-->
        <xsl:for-each select="//grad/naziv">
          <text id="{@id}" fill="#222222" fill-opacity="1" font-family="Arial"
          font-size="6pt" x="{@x}" y="{@y}"><xsl:value-of select="."/></text>
        </xsl:for-each>
      </g>
    </svg>
  </xsl:template>
</xsl:transform>
```

Na slici 7.1 dat je izgled mape posle izvršavanja transformacija u SVG format.

7.3 Formiranje prostornih upita

U ovakvoj organizaciji mogu da se izvršavaju prostorni upiti uz korišćenje prostornih funkcija iz PostGIS modula. Primeri prostornih upita su:

Izračunati najveću razdaljinu dve tačke iz oblasti koja je određena državnom granicom.

```
SELECT ST_AsSVG(ST_LongestLine(c.geometrija, c.geometrija)) AS geometrija,
    ST_Length(ST_LongestLine(c.geometrija, c.geometrija)) AS duzina
FROM granice c
WHERE c.ime='drzavna_granica';
```

Rezultat upita je 556km, a grafički prikaz je dat na slici 7.2.

Prikazati sve gradove u blizini grada Jagodina koji se nalaze u krugu poluprečnika 80 kilometara.

Slika 7.1: Mapa Srbije sa prikazom gradova, državnom i pokrajinskim granicama

Slika 7.2: Prikaz najveće razdaljine dve tačke koje pripadaju državnoj granici

Slika 7.3: Prikaz gradova koji su na razdaljini ne većoj od 80 km od Jagodine

```
WITH krug AS (
          SELECT ST_Buffer(g.geometrija, 80) AS geometrija
          FROM gradovi AS g
          WHERE g.ime='jagodina'
)SELECT ime
 FROM gradovi AS g, krug AS k
 WHERE ST_Intersects(g.geometrija, k.geometrija);
```

Kao rezultat je dobijeno 16 gradova, a njihov grafički prikaz je dat na slici 7.3.

7.4 Mapa prostornih informacija o nematerijalnom kulturnom nasleđu

Za potrebe aplikacije o kulturnom nasleđu Balkana napravljena je mapa uz pomoć prethodno opisanih tehnologija, uz razlike da postoji znatno veći broj gradova i manjih mesta, kao i dodate geometrije i nazivi opština. U prethodnom poglavlju je dat redukovan prikaz geometrija zbog preglednosti.

Mapa kulturnog nasleđa ima tri vrste prikaza:

- Prikazuju se opštine, pri čemu se može izabrati opština klikom na njen grafički prikaz, u kom slučaju se pokreće pretraga baze za materijalima koji su pridruženi datoj opštini
- Prikazuju se mesta, može se izabrati mesto klikom na grafički prikaz mesta, kada se pokreće pretraga baze po izabranom mestu
- Prikazuju se samo mesta koja su pridružena skupu materijala koji je dobijen kao rezultat poslednje izvršenog upita i njima pridružena opština se boji drugom bojom (videti primer sa slike 8.3).

Uz brojne prostorne funkcije PostGIS modula PostgreSQL baze, i u organizaciji podataka koja je sprovedena ovim radom, postoji širok spektar mogućnosti koje se mogu implementirati, a koje su pogodne za materijale sa sadržajem kulturnog nasleđa.

Glava 8

Arhitektura i implementacija sistema

Razvijen je početni sistem za upravljanje nematerijalnim kulturnim nasleđem Balkana koji omogućava rad sa multimedijalnom kolekcijom dokumenata, obradu teksta NLP metodama, anotiranje i organizovanje multimedijalne kolekcije dokumenata u bazu podataka, kao i pretraživanje baze.

Prva generacija sistema, prikazana u radu [267], rešavala je zadatak organizacije kolekcije multimedijalnih materijala u bazu podataka uz ručnu anotaciju materijala.

Druga generacija sistema podrazumeva poboljšanje automatskom semantičkom anotacijom protokola kao pomoći u polu-automatskoj anotaciji multimedijalne kolekcije, kao i poboljšanje prve generacije sistema u različitim njenim delovima.

Automatska semantička anotacija se izvodi uz pomoć metoda obrade prirodnih jezika za zadatke ekstrakcije informacija iz protokola i klasifikacije protokola u odnosu na izabranu tematiku.

Deo razvijenog sistema, koji je zadužen za rad sa multimedijalnom kolekcijom dokumenata, anotiranje multimedijalnog materijala, organizovanje u bazu podataka, i za pretraživanje baze, napisan je u *Java* programskom jeziku. Organizovan je u arhitekturi *klijent / server* i komunicira koristeći *TCP* protokol.

Deo sistema koji se odnosi na implementaciju NLP metoda za klasifikaciju protokola napisan je u jeziku *Python*.

Ceo sistem se sastoji od oko *15000* linija autorskog koda, ne računajući linije koje predstavljaju komentare, i oduhvata oko *100* klasa za omogućavanje funkcionisanja celog sistema.

Dodatni alati koji se koriste u radu sistema su:

- *Unitex/GramLab* program za konstrusanje transduktora i obradu tekstualnih dokumenata ([287])

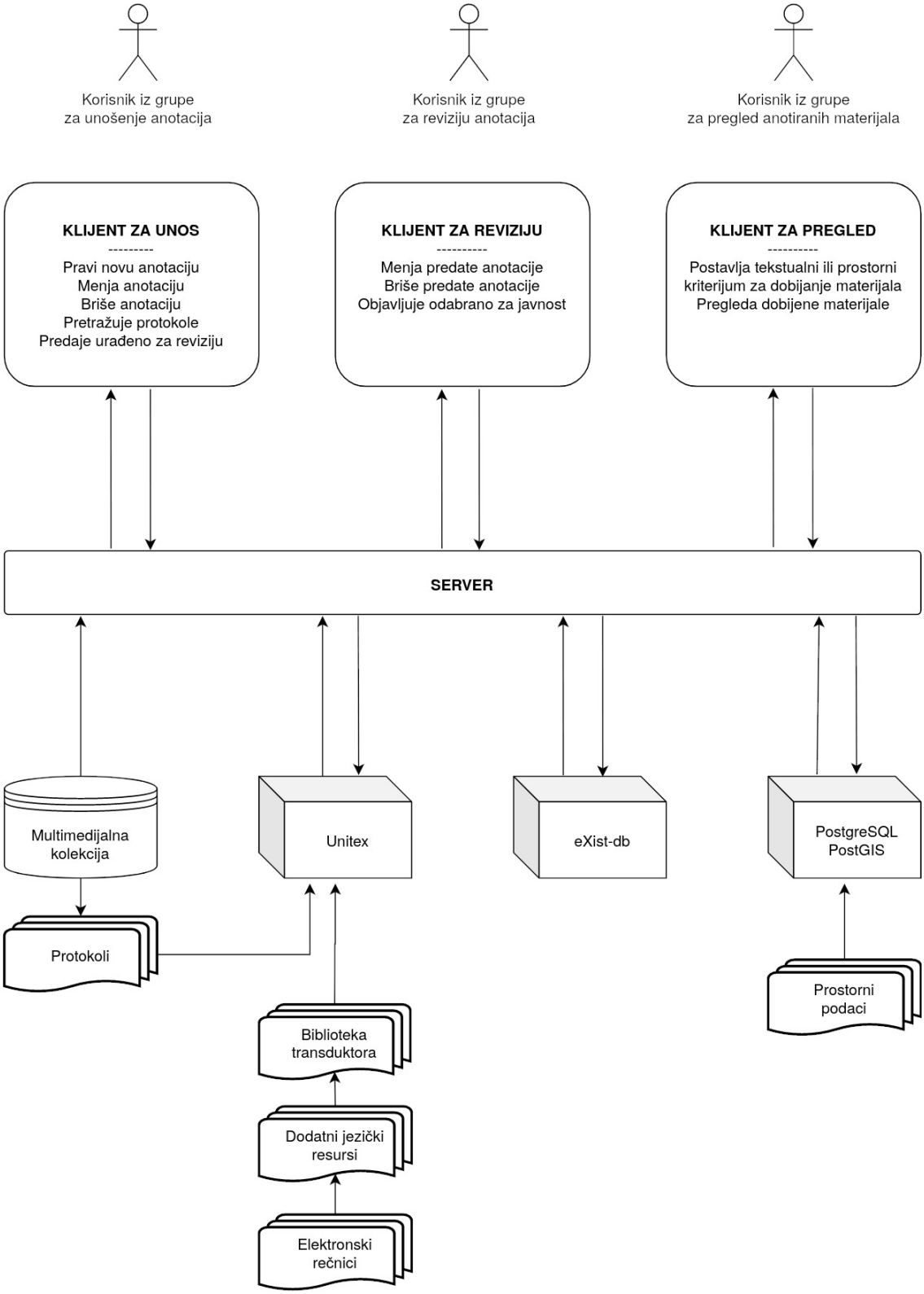
- *scikit-learn* biblioteka u jeziku *Python* za sprovođenje algoritama klasifikacije ([239])
- brojne druge standardne biblioteke u jeziku *Python* za jednostavniji rad sa podacima
- *eXist* baza podataka za rad sa tekstualnim podacima ([68])
- *PostgreSQL* baza podataka ([222]), uz prostorni modul *PostGIS*, za rad sa prostornim podacima ([220])
- *JavaFX* modul za rad sa multimedijom ([119])
- *Tika* biblioteka za prevođenje tekstualnih dokumenata različitih formata u txt format ([276]).

Kompletni sistem ima veliki broj funkcionalnosti, od procesa anotacije, preko različitih vrsta pretrage. Arhitektura celog sistema prikazana je na slici 8.1.

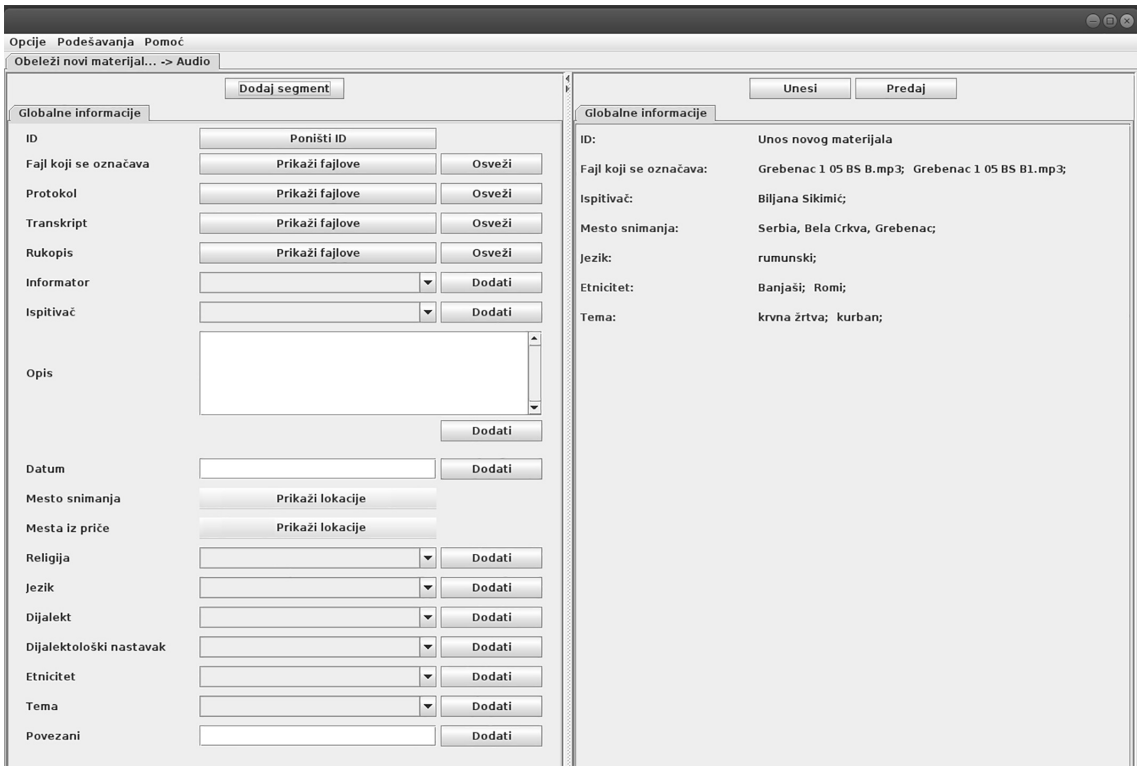
Proces anotacije se izvodi izborom atributa kojima se anotira materijal i unosa novih vrednosti za pojedinačne atribute, ručno ili uz pomoć automatske ekstrakcije informacija iz protokola, izdvajanja grupe materijala uz pomoć sistema za klasifikaciju, izmene ili brisanja već anotiranih materijala, kontrole privatnosti gde korisnik može izabrati da čuva anotaciju za sebe ili da je objavi u okviru grupe korisnika koji se bave revizijom, do revizije unetih anotacija od strane korisnika iz grupe za reviziju i objavljivanja za javnost. Prototip izgleda aplikacije u procesu anotacije može se videti na slici 8.2.

Pretraga i prikaz materijala i pridruženih anotacija je takođe zadatak koji zahteva posebnu pažnju. Omogućena je pretraga multimedije po pridruženim metapodacima tekstualnim formulisanjem kriterijuma ili izborom kriterijuma na mapi lokacija, ali i pretrage tekstualnih protokola putem automatski pridruženih anotacija. Prototip izgleda aplikacije u funkciji pretrage anotiranih materijala i grafičkog prikaza mape Srbije na kome se iscrtavaju opštine i mesta koje odgovaraju dobijenim materijalima dat je na slici 8.3.

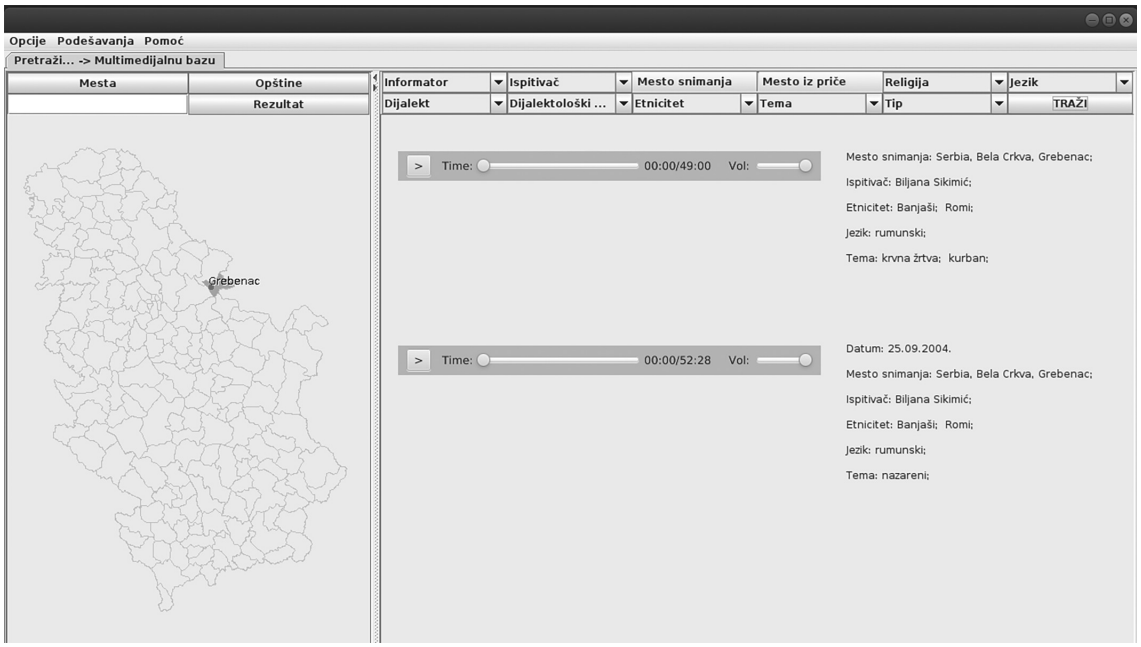
Prikazanim sistemom je realizovan model sa opisanim funkcionalnostima. Budući rad na tehničkoj realizaciji sistema bi mogao da bude u pravcu unapeđenja grafičkog izgleda aplikacije uz pomoć dobijenih povratnih odgovora korisnika koji bi aktivno koristili aplikaciju kao i eksperata iz oblasti grafičkog dizajna. Drugo moguće unapređenje se ogleda u povećanju interaktivnosti i bogatije vizuelizacije informacija koje se mogu dobiti iz baze.



Slika 8.1: Arhitektura sistema



Slika 8.2: Anotacija novog materijala



Slika 8.3: Pretraga baze po izabranom kriterijumu i prikaz materijala

Glava 9

Rezultati eksperimenata i diskusija

U okviru ove disertacije izvedeni su eksperimenti metoda opisanih u poglavlju o metodologiji, a koje metodama obrade prirodnog jezika rešavaju probleme efikasnog automatskog dobijanja informacija iz multimedijalne baze podataka o kulturnom nasleđu. Eksperimenti su izvedeni nad sledećim metodama:

- ekstrakcije informacija primenjene na tekstualne protokole
- poređenja različitih metoda klasifikacije teksta izborom različitih tipova atributa
- klasifikacije tekstova prema pojavljivanju određene tematike primenjene na protokole.

U nastavku su opisani dodatni detalji izvedenih eksperimenata, rezultati i diskusija.

9.1 Analiza skupova tekstova nad kojima su vršeni eksperimenti

Za izvođenje eksperimenata nad metodama ekstrakcije informacija i klasifikacije dokumenata po tematici narodne privrede, korišćeni su tekstovi protokola iz multimedijalne kolekcije o kulturnom nasleđu Balkana. Za izvođenje eksperimenata poređenja kvaliteta različitih metoda klasifikacije pri promeni tipa atributa na osnovu polariteta korišćeni su javno dostupni skupovi filmskih recenzija na različitim jezicima. U nastavku su ukratko opisane osobine navedenih skupova.

Karakteristike skupa tekstualnih protokola

U oblasti kulturnog nasleđa, i posebno za jezike koji nisu široko u upotrebi, kao što je slučaj sa srpskim jezikom, nije jednostavno naći javno dostupne anotirane resurse koji bi se mogli koristiti kao benčmark, čak i ako negde postoje.

U ovom istraživanju korišćen je skup od 800 protokola iz opisane multimedijalne kolekcije nematerijalnog kulturnog nasleđa, koji se sastoji od ukupno oko 600000 reči.

Formiranje jasnih pravila za anotaciju od ključnog je značaja za uspešno sprovedenu ručnu anotaciju korpusa dokumenata ([317]). Za potrebe ručnog anotiranja dela korpusa protokola u ovom istraživanju bilo je moguće napraviti striktna pravila po kojima bi se odlučilo da li protokol sadrži određeni imenovani entitet ili određenu temu koja je među obrađenim temama. Proizveden je zlatni standard tako što su analizirani protokoli i dodeljivane labele na osnovu lične procene autora ([264]). U slučaju imenovanih entiteta nije bilo spornih odluka, odnosno odluka gde bi se procene razlikovale. U slučaju tema, takvih odluka, gde su se razlikovale procene da li protokol sadrži temu narodne privrede ili ne, bilo je na 4% protokola. U procesu evaluacije korišćeni su samo protokoli na kojima su se procene autora poklapale, što je preostalih 96% protokola.

Karakteristike benčmark skupova korišćenih za poređenje metoda klasifikacije pri promeni tipa atributa

Za eksperimente kojima se vršila detekcija polariteta sentimenata u cilju poređenja metoda klasifikacije pri promeni tipa n-grama korišćeno je osam balansiranih, javno dostupnih benčmark skupova filmskih recenzija. Karakteristike benčmark skupova date su u tabeli 9.1.

U većini ovih skupova postojala je i treća, neutralna, kategorija, ali je ona izuzeta iz razmatranja pošto se ovo istraživanje bavilo samo detekcijom polariteta na pozitivno - negativno.

Tabela 9.1: Benčmark skupovi korišćeni u detekciji polariteta nad filmskim recenzijama

Skup podataka	**Jezik**	**Broj pozitivnih recenzija**	**Broj negativnih recenzija**	**Veličina u MB (bez kompresije)**
CornellPD	engleski	1000	1000	7.9
MuchoCine	španski	1351	1274	7.6
OCA	arapski	250	250	2.04
FMR	francuski	1000	1000	1.36
TMR	turski	5331	5331	1.47
CSFD	češki	30897	29716	19.6
SerbMR-2C	srpski	841	841	2.21

9.2 Opis mera koje su korišćene u evaluaciji kvaliteta rezultata

U ovom istraživanju su korišćene mere koje se uobičajeno koriste u oblasti pretraživanja informacija, a to su preciznost (R), odziv (R) i F mera (F). One se izračunavaju uz pomoć broja tačno označenih instanci (True Positives TP), netačno označenih instanci (False Positives FP) i tačnih ali izostavljenih instanci (False Negatives FN), odnosno instanci koje je algoritam označio kao pozitivne, a označene su kao negativne prilikom prethodnog označavanja od strane čoveka ([123]). Mere P, R i F se u skladu sa time izračunavaju na sledeći način:

$$P = \frac{TP}{TP + FP} \tag{9.1}$$

$$R = \frac{TP}{TP + FN} \tag{9.2}$$

$$F = \frac{2PR}{P + R} \tag{9.3}$$

Predstavljene mere mogu biti izračunate na osnovu svih kategorija na dva načina: mikro-merom kada se svi dokumenti posmatraju kao jedan dokument bez obzira na kategoriju, i makro-merom kada se izračunava srednja vrednost mera na svakoj od kategorija. U ovom istraživanju su korišćene makro-mere.

Tabela 9.2: Rezultati označavanja - preciznost, odziv i F mera, računati na skupu za testiranje

Klasa metapodatka	Broj prepoznatih terma uz pomoć rečnika ili obrasca	Broj prepoznatih terma uz pomoć konteksta	Broj prepoznatih terma uz pomoć obe metode	P	R	F
Oznaka	174	0	174	1	0.86	0.92
Informator	0	339	339	0.93	0.84	0.88
Ispitivač	0	181	181	0.98	0.96	0.97
Ostali	356	0	356	1	0.68	0.81
Lokacija	340	742	1082	0.88	0.81	0.84
Datum	191	0	191	1	1	1
Godina	0	80	80	1	0.72	0.84
Etnicitet	646	0	646	0.94	0.79	0.86
Jezik	0	71	71	1	0.76	0.86
Religija	70	7	70	1	0.68	0.81
Imenovani entitet	1777	1420	3197	0.93	0.81	0.87
Tema	196	49	245	0.92	0.89	0.90
Ukupno	1973	1469	3442	0.93	0.82	0.87

9.3 Evaluacija metode ekstrakcije informacija iz tekstualnih protokola

Rezultati

Preciznost, odziv i F mera koje su dobijene na skupu za testiranje za svaku klasu imenovanih entiteta pojedinačno, sumirano za imenovane entitete, sumirano za teme iz oblasti narodne privrede i sumirano za sve klase metapodataka, prikazane su u tabeli 9.2.

Pojedinačne mere za svaku od tema nisu računate jer je broj pojavljanja razdvojen po temama mali, tako da mere nisu reprezentativne. U slučaju imenovanih entiteta F mera je 0.87, dok je u slučaju tema F mera 0.90. Korišćena metodologija rezultira u ukupnoj F meri koja iznosi 0.87.

Da bi osoba bila prepoznata kao informator ili ispitivač potrebno je da se nađe u određenom kontekstu, stoga je prciznost vrlo visoka (0.93), pri čemu je prepoznavanje delimičnog imena računato kao pogrešno označavanje. Na primer, pravilima koja su opisana u konačnim transduktorima nije pokriven slučaj imena koja sadrže dva prezimena ili imaju dodat i nadimak. Prilikom prepoznavanja ispitivača ima manje grešaka u označavanju, odnosno preciznost je viša (0.98), zato što su imena istraživa-

ča po pravilu navođena u manje opisnom obliku, uglavnom navođenjem samo imena i prezimena, dok su imena informatora sadržala i neke dodatne informacije poput nadimka, devojačkog prezimena ili umetnute godine rođenja nakon navođenja imena a pre prezimena. Osobe koje nisu nađene u ovim kontekstima (informator i ispitivač) su prepoznate uz pomoć elektronskog rečnika bez pomoći dodatnog konteksta. Stoga, izračunate mere kvaliteta za klasu "ostali" se mogu uzeti kao aproksimacija mera pokrivenosti ličnih imena dostupnim elektronskim rečnikom.

Lokacije su ekstrahovane uz pomoć rečnika toponima, ali i uz pomoć konteksta, stoga ima više pogrešno označenih i više pogrešno neoznačenih, u poređenju sa klasom "ostali", ali takođe ima i više ekstrahovanih instanci. Najviše grešaka dolazi od konteksta "iz" (na primer, "iz Kulturno umetničkog društva" je pogrešno označeno kao lokacija), dok su konteksti poput "varoš" i "selo" proizvodili tačna označavanja (na primer, "varoš Rača" i "selo Donja Rača").

Svi datumi imaju regularni oblik, tako da je njihova ekstrakcija zasnovana samo na konstruisanju obrazaca bila veoma uspešna. Godine koje se odnose na godine rođenja informatora imaju određeni oblik i ekstrahovane su uz pomoć konteksta, odnosno pronađene su samo one godine koje su navedene striktno nakon imena i prezimena osobe, stoga nema pogrešnih označavanja, ali ima instanci koje nisu označene.

Postoje slučajevi kada određena prezimena imaju isti oblik kao i neki etniciteti, koja bi u tom slučaju pogrešno bila označavana kao etniciteti (na primer, "Bošnjak" i "Crnogorac" su prezimena koja su prepoznata kao etniciteti). Sa druge strane, postoje etniciteti koji nisu pokriveni trenutnim konačnim transduktorima.

Jezici su identifikovani samo na osnovu konteksta tako da nema pogrešno označenih, ali ima slučajeva koji nisu pokriveni (na primer, "na srpskom"), zato što bi obrada takvih slučajeva zahtevala implementaciju drugačije logike da bi se izbegla pogrešna označavanja.

Religije su ekstrahovane uz pomoć konteksta i uz pomoć rečnika tako da bi dopunjavanjem konačnih transduktora i elektronskih rečnika bilo moguće napraviti poboljšanja.

U slučaju zadatka ekstrakcije tema iz tekstova protokola, prepoznavanje samostalnih terma je računato kao prepoznavanje instanci uz pomoć rečnika ili obrasca, dok je prepoznavanje ostalih semantičkih struktura računato kao prepoznavanje uz pomoć konteksta.

Primer dobrog prepoznavanja teme iz netrivijalnog konteksta bi bilo označavanje

Tabela 9.3: Broj ekstrahovanih imenovanih entieta grupisano po klasi imenovanog entiteta

Imenovani entitet	**Broj pojavljivanja**
Oznaka	776
Informator	1334
Ispitivač	719
Ostali	1554
Lokacija	4171
Datum	802
Godina	321
Jezik	287
Etnicitet	2698
Religija	327
Ukupno	12989

fraze "radila je vinograd" klasom "poljoprivreda", zato što se ova fraza odnosi na poljoprivredu.

Primer tačnog neoznačavanja teme uz pomoć konteksta je fraza "prolazila je pored vinograda" koja nije označena kao tema, zato što "vinograd" nije bio u kontekstu specifičnom za temu "poljoprivreda".

Primer pogrešnog označavanja na osnovu konteksta je označavanje samostalnog terma "lov" kao klasa "lov i ribolov", zato što se taj term može nalaziti i u kontekstu "lov mačke i miša" što svakako ne spada u granu narodne privrede. Term "lov" je svrstan u samostalne termove za temu "lov i ribolov" zato što pojavljivanje ove teme nije često, a u protokolima veoma često se koristi samo ova reč kao samostalna da označi priču o tematici "lov i ribolov".

Rezultati metode ekstrakcije informacija primenjene na celu kolekciju protokola

Nakon izračunavanja mera na skupu za testiranje u cilju evaluacije, metoda je primenjena na ceo skup protokola. U tabeli 9.3 prikazan je broj ekstrahovanih imenovanih entiteta posebno za svaku klasu imenovanih entiteta i sumirano za sve imenovane entitete.

U tabeli 9.4 prikazan je broj prepoznatih tema posebno za svaku temu i sumirano za oblast narodne privrede.

Tabela 9.4: Broj prepoznatih fraza iz tematike narodna privreda grupisano po temama

Tema	Broj pojavljivanja
Domaća radinost	160
Lov i ribolov	36
Pčelarstvo	58
Poljoprivreda	640
Rudarstvo	23
Šumarstvo	3
Trgovina	33
Zanatstvo	106
Ukupno	1059

Diskusija

Najvažniji doprinos ovog istraživanja obuhvata metod za ekstrakciju metapodataka kojima se vrši označavanje dokumenata pridruženih multimedijalnoj kolekciji nematerijalne kulturne baštine na prostoru dela Balkana. Pristup koji se koristi zasnovan je na formulisanju pravila i razvijanju biblioteke konačnih transduktora u kombinaciji sa identifikovanjem konteksta koji je specifičan za posmatrani domen, morfološkim elektronskim rečnicima, tezaurusima i dodatnim izvorima domenski specifičnih reči.

Iako su se metode zasnovane na pravilima za zadatak ekstrakcije informacija pokazale korisnim u mnogim domenima i za različite jezike, nije istraživan potpuno njihov uticaj na domenu kulturnog nasleđa, a posebno je oblast Balkana nedovoljno istražena u ovom kontekstu.

Prikazano istraživanje je značajno zato što predstavlja prvi polu-automatski sistem za pomoć pri semantičkoj anotaciji protokola na srpskom jeziku pridruženih materijalima o kulturnom nasleđu Balkana, pretragu materijala na osnovu pridruženih opisnih metapodataka i metapodataka koji se odnose na teme.

Rezultati automatskog indeksiranja dokumenata pokazuju da se ručno pisanim pravilima za ekstrakciju informacija mogu proizvesti visoko kvalitetne semantičke anotacije po izabranoj shemi metapodatka, koja je u skladu sa CIDOC CRM (ISO 21127:2014) standardom, i sa rečničkim resursima prilagođenim domenu nematerijalnog kulturnog nasleđa.

Rezultati evaluacije na osnovu zlatnog standarda su ohrabrujući i u najmanju ruku u rangu rezultata relevantnih istraživanja, iako, zbog specifičnih osobina do-

mena i specifičnih karakteristika sistema nije moguće napraviti direktno poređenje. Za potrebe ovog istraživanja nisu pronađeni radovi koji prikazuju rezultate ekstrakcije informacija u istom domenu nematerijalnog kulturnog nasleđa sa kojima bi se moglo napraviti direktno poređenje rezultata predstavljenih u ovom istraživanju. Što se tiče NER sistema u sličnim istraživanjima, ukupna F mera za NER koja je dobijena u ovom istraživanju (87%) je u rangu sa, na primer, NER sistemima koji se bave arheološkim domenom u kojima F mera iznosi od 68% do 83% za pristupe zasnovane na pravilima i na mašinskom učenju ([28], [297], [317]) ili sistemom za istraživanje stavova u tekstualnom obliku zasnovanom na NER metodama koji ima meru odnosa preciznost-odziv 0.78 ([170]). U terminima ekstrakcije metapodataka koji se odnose na temu, ukupna F mera koja je dobijena u ovom istraživanju (90%) se može porediti sa sličnim inicijativama koje koriste semantičke strukture, na primer, ekstrakcijom podataka iz narativa medicinskog domena u kojima su dobijeni preciznost 93% i odziv 83% ([54]), ili semantičkom anotacijom televizijskih i radio novosti, za koju je objavljena preciznost 100% uz odziv 40.1% ([59]).

Za ekstrakciju informacija koje se odnose na klase metapodataka poput osoba, lokacija, etniciteta i religija, dopunjavanje rečnika domenski specifičnim rečima imalo bi značajnog uticaja na poboljšanje odziva. Poboljšanje bi takođe moglo da se dobije dopunjavanjem pravila za ekstrakciju tako da obuhvate jos slučajeva (čime bi se poboljšao odziv) ali i da opisane slučajeve naprave specifičnijim obuhvatajući i analizu šireg konteksta (čime bi se poboljšala preciznost).

Cilj ovog istraživanja je bio da istraži rezultate koji se mogu dobiti ulaganjem razumnog napora za zadatak ekstrakcije informacija u cilju označavanja metapodataka. Dobijeni rezultati pokazuju da se ovakvim pristupom mogu dobiti sasvim dobri rezultati koji su u skladu da sličnim istraživanjima iz sličnih domena. Pristupom zasnovanim na pravilima, uz pomoć dodatnih izvora rečnika, u slučaju većine metapodataka postoji prostor za unapređenje. Ono što treba imati na umu je da, od nekog trenutka, uloženi trud premašuje dobrobiti koji se mogu dobiti, odnosno za malo poboljšanje potrebno je nesrazmerno više truda.

Prikazani pristup ima i svoja ograničenja i mane. Značajan trud treba da bude uložen od strane eksperata u razvijanju biblioteka konačnih transduktora da bi se dobio neznatno bolji rezultat. Ovde mogu u pomoć doći metode mašinskog učenja, koje se pak zasnivaju na učenju na osnovu označenog velikog skupa podataka. U slučaju kolekcije dokumenata sa kojom se radilo u ovom istraživanju, korpus kao i ostali resursi, nisu bili dovoljni za razvoj i primenu metoda koje su trenutno aktuelne,

kao što su metode dubokog učenja.

Na osnovu sprovedenog istraživanja naučeno je više lekcija. Jedna od stvari koja je naučena je ta da je domen kulturnog nasleđa veoma bogat semantikom, da kontekst igra veliku ulogu u procesu ekstrakcije informacija i da pristup zasnovan na pravilima u kombinaciji sa rečničkim resursima daje veoma dobre rezultate.

Iako predstavljena metodologija objedinjuje domenski specifično znanje i jezičko znanje u biblioteku konačnih transduktora, ona se može prilagoditi za izgradnju sličnih sistema za druge domene i druge jezike.

9.4 Ekstrakcija informacija iz tekstualnih protokola primenom komercijalnog alata IBM SPSS Modelera

IBM SPSS Modeler ([113]) je skup alata koji se koriste za zadatke istraživanja podataka prilagođenih poslovnim korisnicima kao pomoć prilikom donošenja različitih odluka. Modeler omogućava primenu kompletnog procesa istraživanja podataka, počevši od pripreme podataka i infrastrukture za njihovo bolje razumevanje, preko primene različitih metoda za modelovanje preuzetih iz mašinskog učenja, veštačke inteligencije, statistike, metoda jezičkih analiza teksta i tehnologija obrade prirodnih jezika, do brojnih metoda za vizuelizaciju podataka.

Sa IBM SPSS Modelerom mogu se razvijati modeli za različita predviđanja bez prethodnog programiranja. Bogat grafički interfejs omogućava vizuelizaciju celokupnog procesa istraživanja podataka pomoću kombinovanja specifičnih podmodula. Uz pomoć ugrađenih naprednih metoda za analize podataka, mogu se ekstrahovati i sublimirati nova znanja sadržana u postojećim podacima. Predviđene vrednosti se mogu analizirati u cilju razumevanja faktora koji utiču na njihov trenutni oblik, kako bi se eventualno usmerilo ponašanje u skladu sa istraživanjima tržišta tako da bude usaglašeno sa željenim pravcem kretanja poslovanja. Dostupni su algoritmi za modelovanje poput neuronskih mreža, drveta odlučivanja, klasterovanja, sekvenciranja podataka, algoritmi koji su sadržani u Microsoft SQL Server, IBM Db2, Oracle i Netezza bazama podataka i algoritmi koji postoje u Python i Spark bibliotekama. Kompletna dokumentacija za IBM SPSS Modeler može se naći na [114].

IBM SPSS Modeler se može koristiti u Professional i Premium verzijama. Professional verzija se koristi za rad sa strukturiranim podacima, dok Premium verzija

sadrži sve funkcionalnosti koje sadrži i Professional verzija, uz dodatak modula *Text Analytics* koji se koristi za različite napredne jezičke analize nestrukturiranog teksta za brzu i efikasnu obradu uz tehnologije obrade prirodnih jezika i mogućnost korišćenja jezičkih resursa.

IBM SPSS Modeler Text Analytics se isporučuje sa gotovim templejtima sa specijalizovanim rečnicima, bibliotekama i drugim resursima potrebnim za uobičajene probleme kao što su istraživanje zadovoljstva kupaca, istraživanje odnosa sa korisnicima (Customer Relationship Model CRM) ili proučavanje genoma. Sadrži templejte za rad sa tekstovima na engleskom, francuskom, nemačkom, danskom, italijanskom, španskom i portugalskom jeziku. Text Analytics ne sadrži predefinisane templejte za srpski jezik.

Text Analytics omogućava automatsko generisanje koncepata i njihovo kategorisanje na osnovu ugrađenih pravila. Ukoliko su domen ili jezik specifični, što je slučaj sa ekstrakcijom fraza iz tekstova o kulturnom nasleđu i srpskim jezikom, postoji mogućnost da korisnik sam definiše način na koji će se raditi sa tekstom.

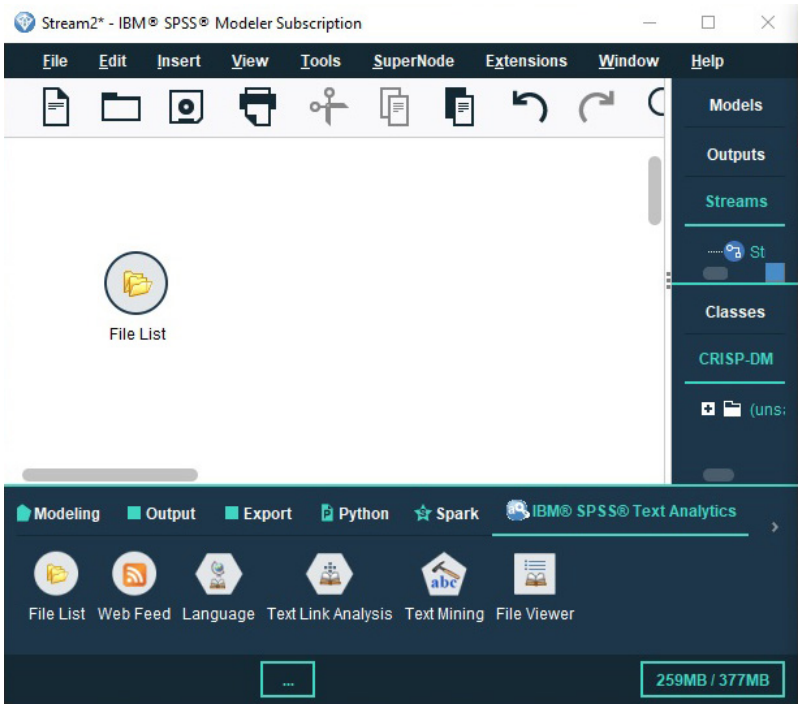
Ovde će biti ukratko opisano kako su iskorišćene funkcionalnosti Text Analytics modula za zadatak ekstrakcije tema iz protokola.

Primer: ekstrakcija tema iz protokola

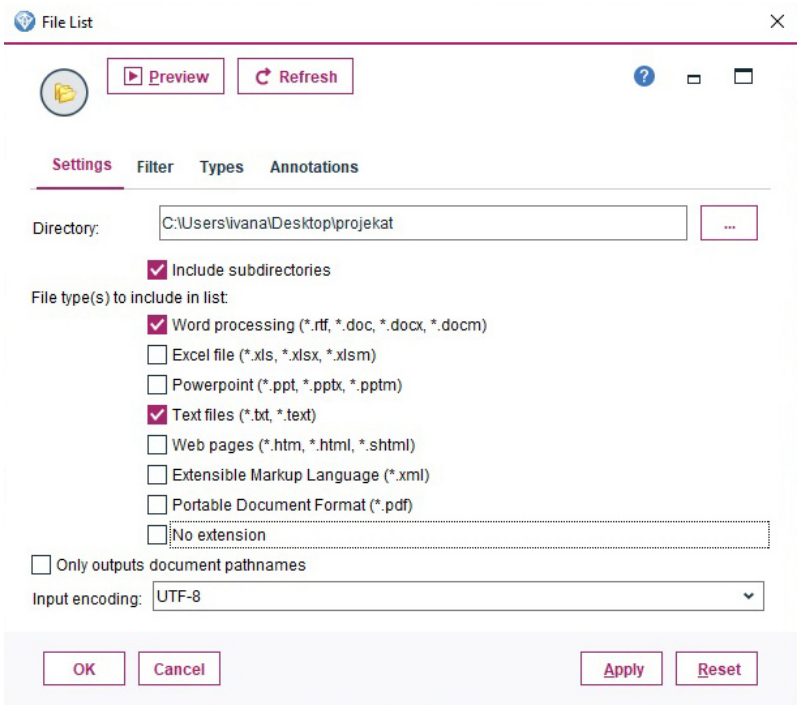
Glavni koraci u procesu ekstrakcije tema su čitanje protokola, formiranje modela za pretraživanje teksta, rad u interaktivnom workbench-u za razvijanje rečnika tipova i koncepata, razvijanje pravila za grupisanje tipova i formiranje kategorija, i čuvanje razvijenog paketa. U nastavku su opisi navedenih koraka.

Čitanje protokola Pošto IBM SPSS Modeler takozvanim čvorovima (engl. *node*) predstavlja aktivnosti i resurse nad kojima se aktivnosti obavljaju, prvo je potrebno napraviti čvor koji odgovara listi fajlova u kojima se nalaze tekstovi koji se obrađuju (u našem slučaju - protokoli). Na slici 9.1 predstavljeno je dodavanje čvora *File List*, a na slici 9.2 postavljanje parametara za taj čvor.

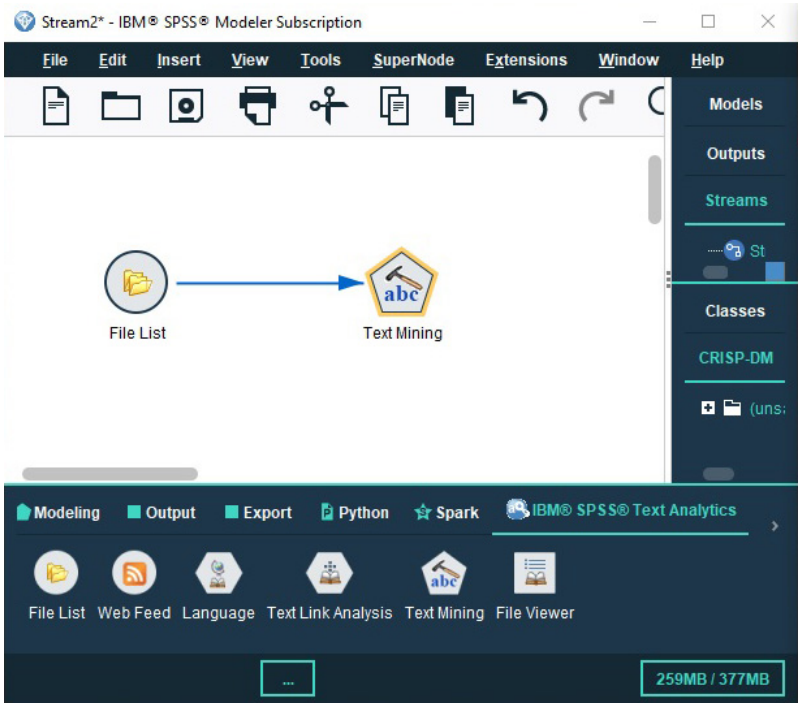
Formiranje modela za pretraživanje teksta Iz IBM SPSS Modeler Text Analytics palete dodati *Text Mining* čvor i spojiti ga sa *File List* čvorom (slika 9.3). U okviru podešavanja *Text Mining* čvora izabrati opciju interaktivnog razvijanja modela.



Slika 9.1: Postavljanje čvora za čitanje protokola.



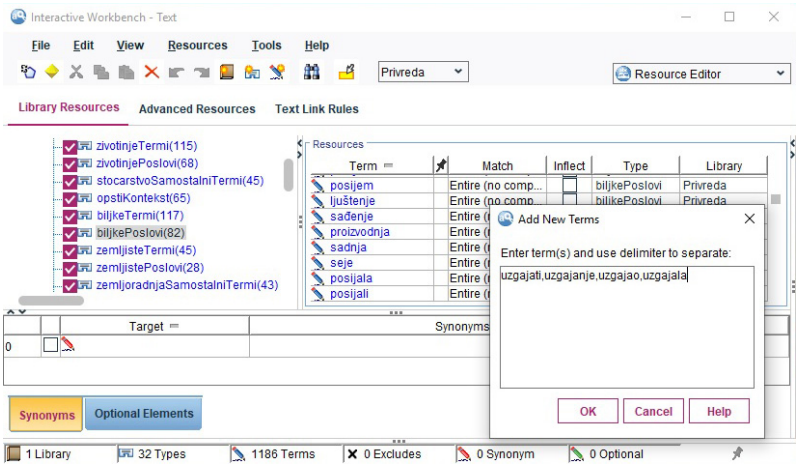
Slika 9.2: Postavljanje parametara za čitanje u *File List* čvoru.



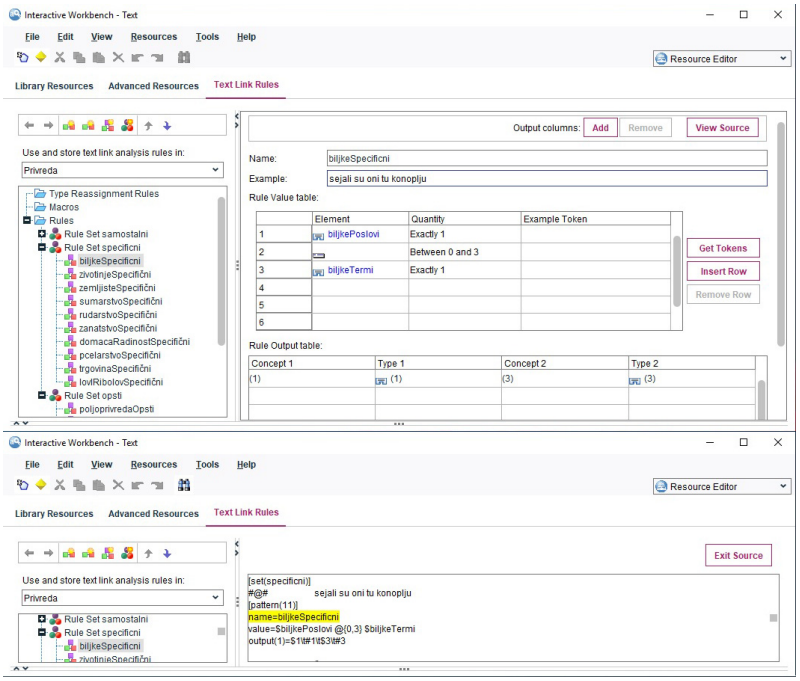
Slika 9.3: Povezivanje *Text Mining* čvora.

Rad u interaktivnom workbench-u U interaktivnom workbench-u postoje četiri vrste pregleda (panela): *Resource Editor*, *Text Link Analysis*, *Clusters* i *Categories and Concepts*. U njima se razvijaju rečnici tipova i koncepata, kao i pravila za grupisanje tipova i formiranje kategorija. Ovde će biti opisani navedeni paneli onako kako su i korišćeni.

Resource Editor Okruženje u kome se mogu praviti jezički resursi, biblioteke *koncepata* (termi koji mogu biti jednosložne ili višesložne reči, na primer, "jabuka" ili "sok od jabuke"), *tipova* (semantičko grupisanje koncepata, na primer, grupe "biljke poslovi" ili "biljke termi"), *sinonima* (reči koje imaju isto značenje u okviru posmatrane problematike, na primer, "terminologija" i "izrazi"), *makroi* (ukoliko je potrebno izvršiti neko drugačije pridruživanje od opisanih, mogu se koristiti makroi, na primer, makroom se mogu grupisati crvene biljke "paradajz", "paprika" i "trešnja") i *pravila* za grupisanje (definisanje obrasca koji treba da zadovoljava neka grupa koncepta, tipova ili makroa da bi bila ekstrahovana, na primer, može se definisati obrazac za prepoznavanje fraza koje se sastoje od koncepta tipa "biljke poslovi" iza koga se nalazi koncept tipa "biljke termi", tako da između mogu da se nađu najviše tri bilo koje reči).



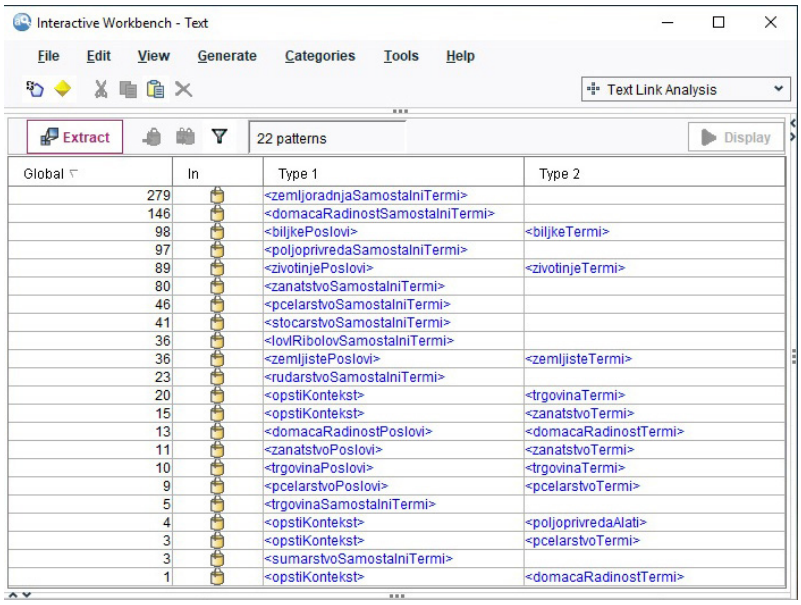
Slika 9.4: Definisanje tipova i pridruživanje termova tipovima.



Slika 9.5: Definisanje pravila grafičkim putem i tekstualno.

Na slici 9.4 je prikazano gde se mogu definisati tipovi i njima pridružiti termi. Na slici 9.5 se nalazi primer formiranja pravila grafičkim putem, ali i prikaz kako se pravila mogu zadavati i tekstualno.

Text Link Analysis Koristi se za ekstrakciju primenom prethodno definisanih elemenata i prikaz rezultata sumirano po pravilima koja su korišćena (slika 9.6). Na



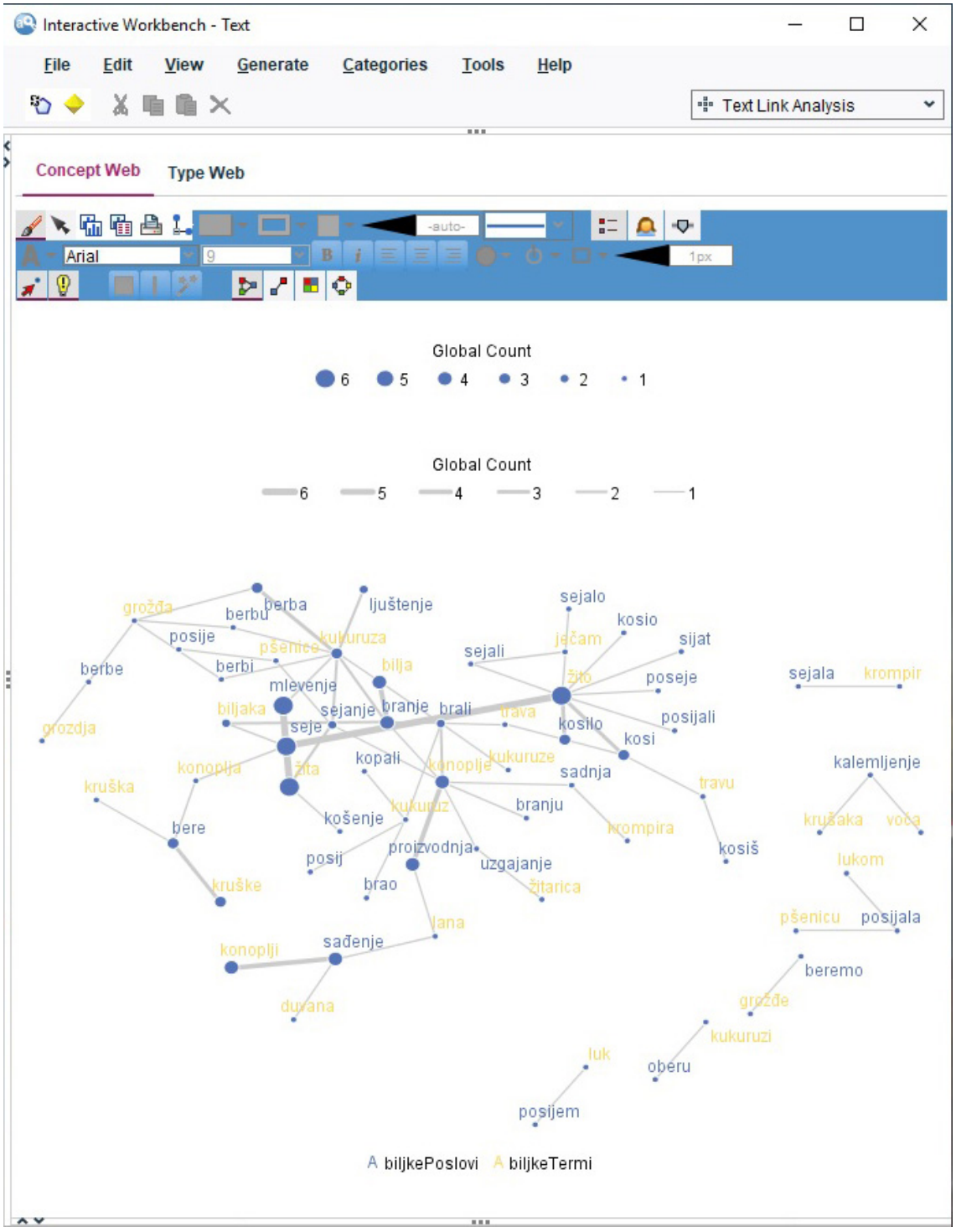
Global	In	Type 1	Type 2
279		<zemljoradnjaSamostalniTermi>	
146		<domacaRadinostSamostalniTermi>	
98		<biljkePoslovi>	<biljkeTermi>
97		<poljoprivredaSamostalniTermi>	
89		<zivotinjePoslovi>	<zivotinjeTermi>
80		<zanatstvoSamostalniTermi>	
46		<pcelarstvoSamostalniTermi>	
41		<stocarstvoSamostalniTermi>	
36		<lovlRibolovSamostalniTermi>	
36		<zemljistePoslovi>	<zemljisteTermi>
23		<rudarstvoSamostalniTermi>	
20		<opstiKontekst>	<trgovinaTermi>
15		<opstiKontekst>	<zanatstvoTermi>
13		<domacaRadinostPoslovi>	<domacaRadinostTermi>
11		<zanatstvoPoslovi>	<zanatstvoTermi>
10		<trgovinaPoslovi>	<trgovinaTermi>
9		<pcelarstvoPoslovi>	<pcelarstvoTermi>
5		<trgovinaSamostalniTermi>	
4		<opstiKontekst>	<poljoprivredaAlati>
3		<opstiKontekst>	<pcelarstvoTermi>
3		<sumarstvoSamostalniTermi>	
1		<opstiKontekst>	<domacaRadinostTermi>

Slika 9.6: Lista pravila sa brojem ekstrahovanih fraza po pravilu. Na listi se nalaze samo pravila na osnovu kojih je pronađena bar jedna fraza.

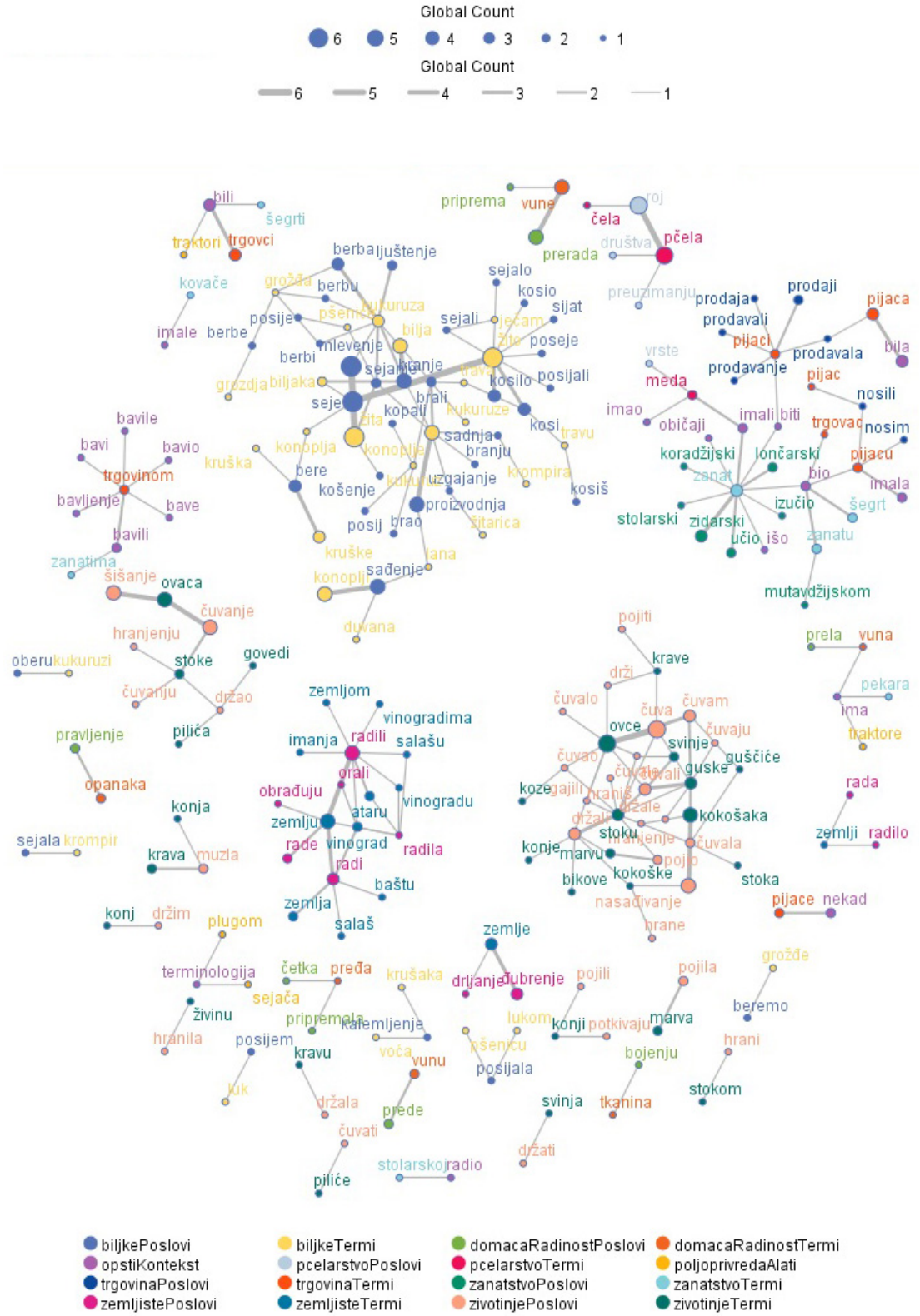
slici 9.7 je prikazana mreža povezanosti koncepata u okviru pravila “biljke poslovi” + GAP + “biljke termi”, gde je GAP sekvenca od nula do tri bilo koje reči. Na slici 9.8 se može videti mreža povezanosti svih ekstrahovanih koncepata u okviru pravila koja zahtevaju spajanje koncepata iz dva tipa, dok je na slici 9.9 mreža povezanosti tipova kroz dokumente.

Categories and Concepts Koristi se za rad sa kategorijama i konceptima. Kategorije služe da se željeni koncepti grupišu po nekom opisanom kriterijumu. U ovom panelu se definišu kategorije, mogu se dodavati ekstrahovani koncepti ili pravila u kategorije, izvršavati ekstrakcija i pregledati rezultat ekstrakcije. U ovom primeru kategorijama su pridružena pravila koja su ranije napravljena. Na slici 9.10 je lista kategorija sa sadržavajućim pravilima. Na slici 9.11 je primer protokola sa kategorijama koje su mu dodeljene. Mreža povezanosti kategorija kroz dokumente je prikazana na slici 9.12.

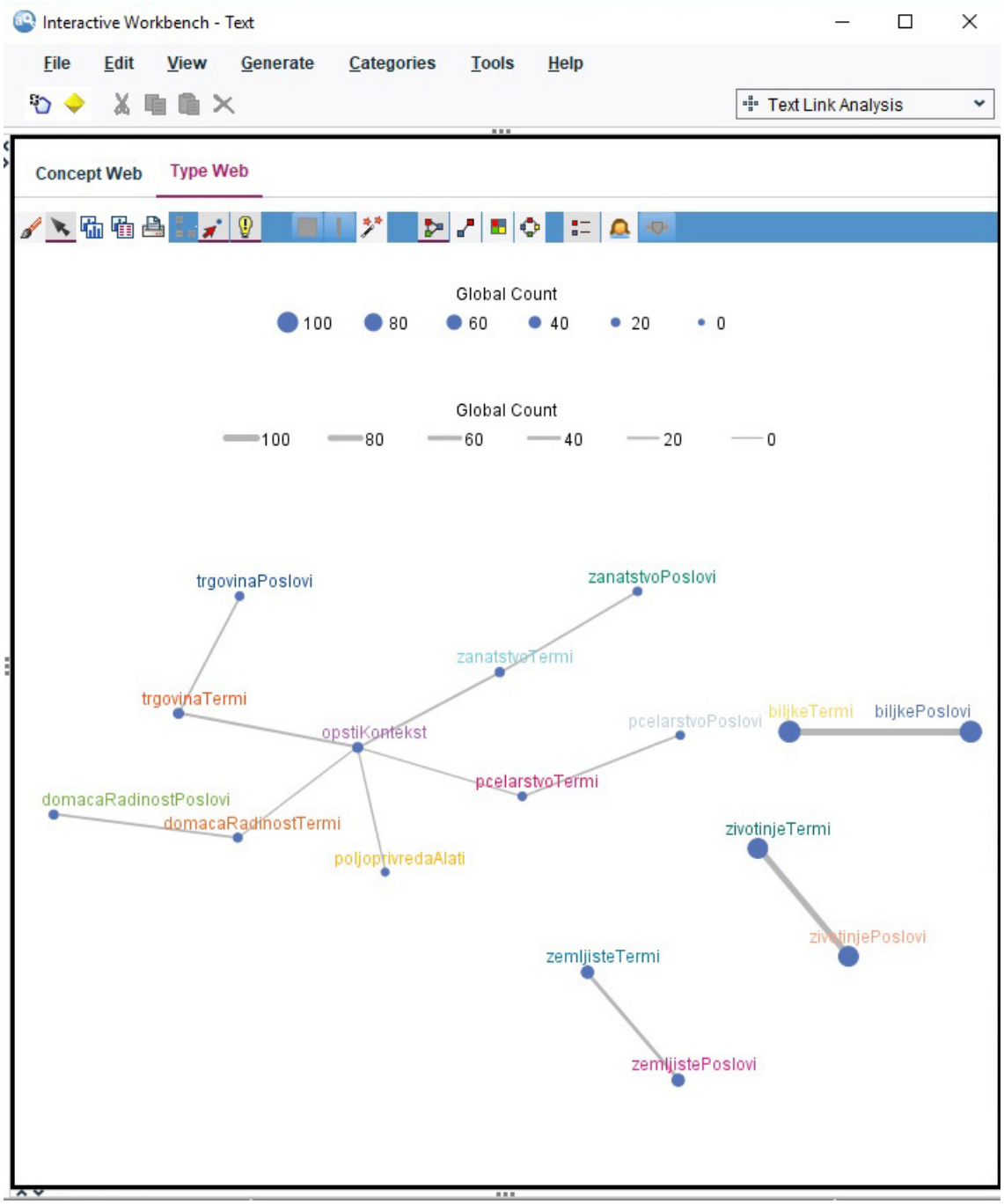
Clusters Omogućava da se otkriju veze između koncepata metodama klasterovanja prema jačini veze između koncepata. Jačina veze je zasnovana na međusobnom zajedničkom pojavljivanju na nivou dokumenta.



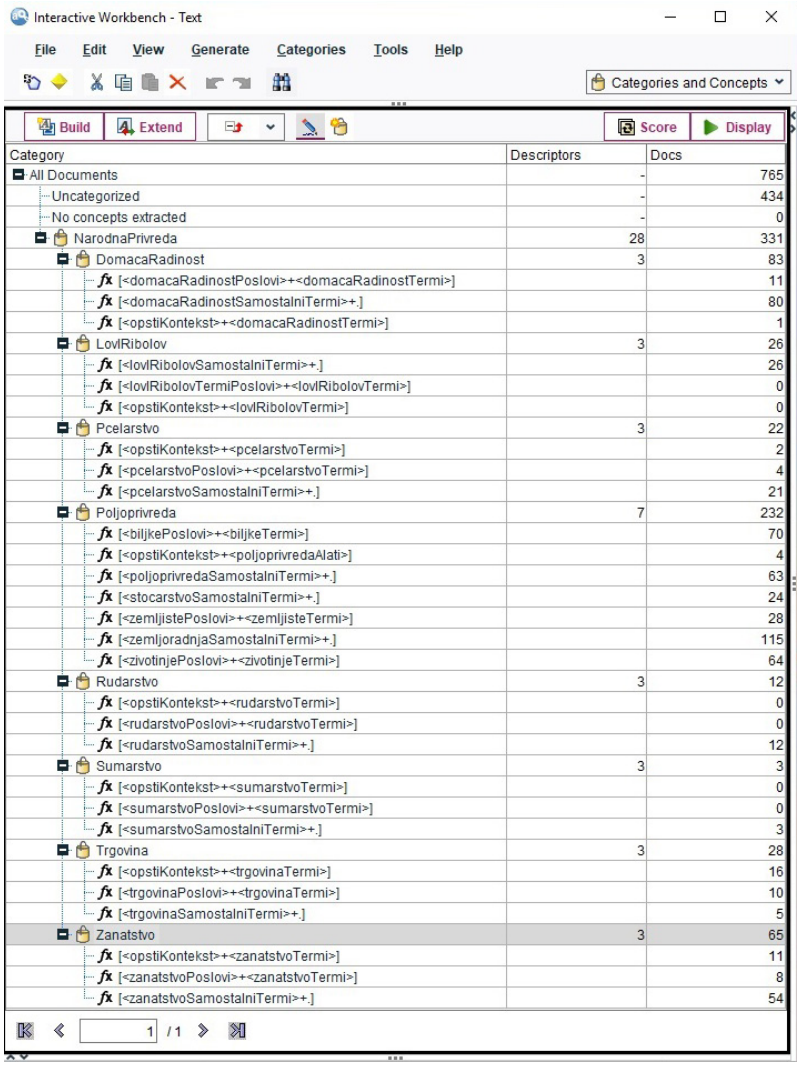
Slika 9.7: Mreža povezanosti koncepata u okviru pravila “biljke poslovi” + GAP + “biljke termi”.



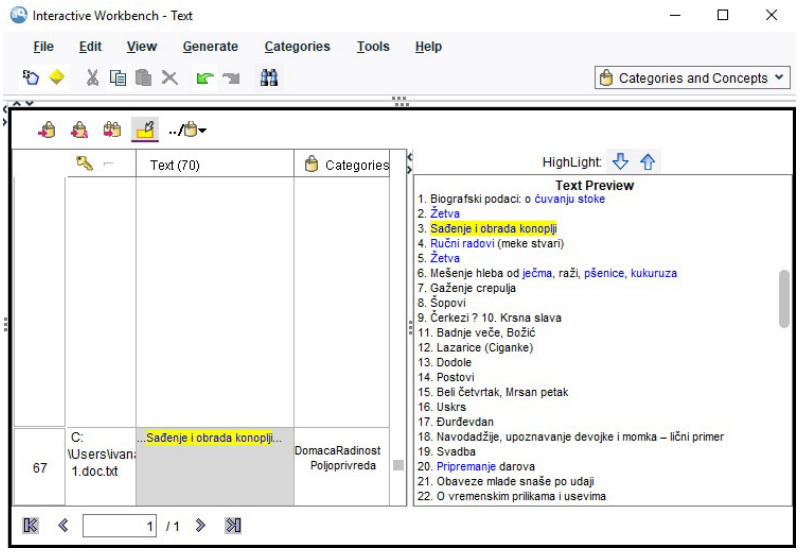
Slika 9.8: Mreža povezanosti svih ekstrahovanih koncepata u okviru pravila koja zahtevaju spajanje koncepata iz dva tipa.



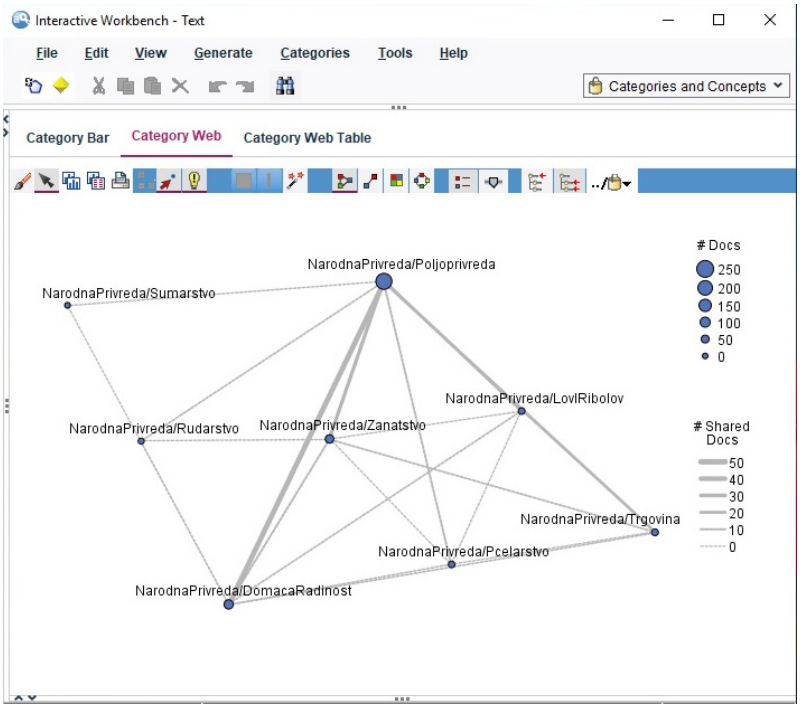
Slika 9.9: Mreža povezanosti tipova.



Slika 9.10: Lista kategorija sa sadržavajućim pravilima, brojem deskriptora po kategoriji, brojem dokumenata u kojima je prepoznata bar jedna fraza odgovarajućim pravilom po svakom pravilu posebno i brojem dokumenata u kojima je prepoznata bar jedna fraza bar jednim pravilom iz odgovarajuće kategorije sumirano po kategorijama.



Slika 9.11: Primer protokola sa kategorijama koje su mu dodeljene, ekstrahovanim frazama na osnovu pravila "biljke poslovi' + GAP + "biljke termi" (žuto) i drugim prepoznatim konceptima (plavo).



Slika 9.12: Mreža povezanosti kategorija kroz dokumente.

Čuvanje razvijenog paketa Paket Text Analysis package (TAP) sadrži definicije jezičkih resursa, rečnike, biblioteke, pravila i definicije kategorija i čuva se izborom odgovarajuće opcije u okviiru *Resource Manager* menija kartice *File*.

Poređenje sa alatom Unitex

IBM SPSS Modeler Text Analytics je ekstrakciju izvršavao primarno vođen definisanim resursima i striktnim podešavanjima tako da su rezultati očekivano u skladu sa rezultatima koji su dobijeni alatom Unitex. U svim kategorijama, osim u kategoriji "poljoprivreda", ekstrahovan je jednak broj instanci (videti tabelu 9.4). U kategoriji "poljoprivreda" alatom IBM SPSS Modeler Text Analytics ekstrahovane su četiri instance više. Reprezentativni primeri dodatno ekstrahovanih instanci su "kosi zito" i "radili su kao radnici u vinogradu". U prvom primeru, "zito" nije u rečniku ali je program ispravno pretpostavio da je ovo dobar spoj. U drugom primeru se između terma "radili", klase "zemljište poslovi", i "vinogradu", klase "zemljište termi", nalaze četiri reči, što nije u skladu sa definisanim pravilom, iako je i ova instanca ispravno ekstrahovana. Jedna mogućnost je, na osnovu načina formiranja koncepata, da program na osnovu svojih predviđanja grupiše koncepte tako da neke dve ili više od ove četiri reči posmatra kao jedan koncept, što je svakako zanimljivo zapažanje koje treba imati u vidu.

Program IBM SPSS Modeler Text Analytics ima mogućnost tolerancije na greške u kucanju i različite varijacije reči. Prethodna razmatranja su bila bez korišćenja ove funkcionalnosti, ali kada se aktivira ova opcija dobija se više instanci koje su uglavnom pogrešne. Na primer, najviše grešaka je dobijeno zato što je na osnovu reči "predenje" prepoznata i reč "predanje", što u srpskom jeziku vodi pogrešnoj ekstrakciji. Drugi primer pogrešne zamene je prepoznavanje reči "prelo" na osnovu reči "prele". Reč "prelo" je često predstavljala društveni događaj, dok "prele" označava aktivnost iz domena domaće radinosti. Primer ispravnog prepoznavanja je prepoznavanje reči "terminoligija" na osnovu reči "terminologija". Pretpostavka je da bi uz bogatiji skup rečnika i drugih jezičkih resursa na srpskom jeziku ova opcija proizvodila bolje rezultate.

Prednost alata Unitex je u mogućnosti korišćenja pogodnosti već razvijenih dostupnih elektronskih rečnika na srpskom jeziku.

9.5 Poređenje kvaliteta klasifikacije za različite reprezentacije teksta i metode klasifikacije

Rezultati

Na benčmark skupovima filmskih recenzija pri detekciji polariteta sentimenata u cilju poređenja različitih reprezentacija dokumenata i metoda klasifikacije, dobijeni su sledeći rezultati.

Pokazuje se da bajt i karakter n-gram modeli daju bolje rezultate od modela koji koriste n-grame reči, pri čemu se može reći da su performanse bajt n-gram modela i karakter n-gram modela vrlo slične. Kod modela koji koristi n-grame reči, najbolje rezultate imao je model sa parametrom $n = 1$, što se zapravo svodi na model vreće reči. U slučaju bajt n-grama, najbolji rezultati su dobijeni za n od 3 do 9, dok su u slučaju karakter n-grama najbolji rezultati dobijeni za n od 5 do 7.

Najbolji rezultati za tri metode - podržavajućih vektora, k najbližih suseda i maksimalne entropije, po tipu n-grama na skupu podataka na srpskom jeziku dati su u tabeli 9.5, dok se uporedni rezultati za druge jezike mogu naći u radu [94].

Tabela 9.5: Rezultati za metode kNN, SVM i MaxEnt primenjene na srpski jezik

		SerbMR-2C		
	n-gram	P	R	F
kNN	bajt	**0.8114**	**0.8114**	**0.8114**
	karakter	0.8060	0.8060	0.8060
	reč	0.6886	0.6886	0.6886
SVM	bajt	**0.8406**	0.8316	**0.8354**
	karakter	0.8240	**0.8400**	0.8308
	reč	0.7847	0.7894	0.7866
MaxEnt	bajt	**0.8783**	**0.8264**	**0.8512**
	karakter	**0.8783**	**0.8264**	**0.8512**
	reč	0.8306	0.7556	0.7909

Napomena: Podebljani brojevi označavaju najbolji rezultat u okviru metode za određeni tip n-grama, dok podvučeni brojevi označavaju najbolji rezultat prema F meri među primenjenim metodama na skupu podataka na srpskom jeziku SerbMR-2C.

Može se zaključiti da se MaxEnt metoda pokazala najboljom posmatrajući meru preciznosti i ukupnu F meru, dok se SVM metoda pokazala sličnom po F meri, ali sa najboljom merom odziva, pa je s obzirom na značaj odziva kada je tematska klasifikacija u relativno maloj kolekciji protokola u pitanju, za klasifikaciju protokola odabrana SVM metoda.

Diskusija

Metode predstavljanja dokumenata zasnovane na n-gramima, opisane u ovom radu, kao i metode klasifikacija teksta, upoređene su sa aktuelnim rezultatima nad svim opisanim benčmark skupovima. Metode klasifikacije teksta sa kojima se vršilo poređenje su bile različitih tipova - statističke, semantičke i hibridne.

Prema dobijenim rezultatima kao najuspešnije pokazuju se statističke metode mašinskog učenja u kombinaciji sa n-gramima bajtova i karaktera. Veći broj drugih metoda sa kojima se vršilo poređenje koristi model vreće reči, dok metode korišćene u ovom radu koriste samo preprocesiranje koje obuhvata uklanjanje interpunkcije, stop reči, ispravljanje slovnih grešaka i ujednačavanje veličine slova. Takođe, metode predstavljene u ovom radu u većini slučajeva proizvode bolje rezultate od semantičkih metoda i od nekih hibridnih metoda. Neke od metoda koristile su "valentne prebacivače" (negacije, augmentative, deminutive) ([129]), neke su koristile metode pretraživanja teksta za ekstrahovanje učestalih podsekvenci reči ili poddrveta zavisnosti iz rečenice ([168]), neke su koristile dodatno znanje iz Wikipedia inkorporirajući ga u semantički moduo u cilju povećanja izražajnosti predstavljanja dokumenta ([306]), dok su neke koristile dodatne specijalizovane rečnike ([53]).

U ovom istraživanju je ispitivano kako reprezentacije dokumenata različitim tipovima n-grama utiču na rezultate prepoznavanja polariteta emocija, odnosno, odgovorom na pitanje da li postoje određeni tipovi n-grama koji se mogu koristiti za predstavljanje dokumenata na različitim jezicima tako da statističke metoda mašinskog učenja budu uspešne u rešavanju postavljenog problema klasifikacije. U eksperimentima su se koristile nadgledane metode mašinskog učenja kNN, SVM i MaxEnt. Isptivane su različite reprezentacije dokumenata sa filmskim recenzijama na jezicima relativno raznolikih paradigmi kao što su engleski, španski, arapski, francuski, turski češki i srpski. U ovom istraživanju nije korišćeno složeno preprocesiranje koje uključuje POS označavanje, parsiranje, i druge jezički zavisne i složene alate za obradu. Uprkos jednostavnosti korišćenih metoda, njima se dobijaju rezultati koji su među najboljima u poređenju sa ostalim nadgledanim statističkim metodama mašinskog učenja.

Pravac daljeg istraživanja može se videti u implementaciji hibridne metode koja bi statističke metode mašinskog učenja dopunila nekom od semantičkih matoda. Takođe, metode klasifikacije po polaritetu emocija bi u daljem radu mogle biti unapređene nekim od metoda neuronskih mreža.

9.6 Evaluacija metode klasifikacije tekstualnih protokola prema tematici

Rezultati

Izvedeni su eksperimenti klasifikacije protokola u odnosu na pojavljivanje tematike narodne privrede. Eksperimentisano je sa različitim načinima reprezentacije protokola. Prvo su ispitane mogućnosti klasifikacije pri reprezentaciji dokumenta karakter n-gramima za $n = 2$ do $n = 9$, pri čemu su najbolji rezultati bili za $n = 5$, tako da je u daljem radu razmatran samo taj slučaj. Za klasifikaciju je korišćena SVM metoda i rezultati su prikazani u tabeli 9.6.

Tabela 9.6: Rezultati metode SVM primenjene na protokole

Tekst	**Atributi**	**P**	**R**	**F**
Originalni tekst	Atributi su najfrekventniji n-grami karaktera za $n = 5$	0.5050	0.3935	0.4435

Kao što se može videti, rezultati nisu obećavajući, tako da je upotrebljeno više metoda za poboljšanje, a koje su opisane u poglavlju o metodologiji.

Prvi metod je restrikcija početnog teksta protokola na reči iz vreće značajnih reči uključujući i njihovu okolinu u tekstu od četiri reči pre i četiri reči posle. Za atribute su razmatrani najfrekventniji n-grami karaktera za $n = 5$ i to iz skupa napravljenih od sledećih izvora:

- *n-grami iz sopstvenih reči*
- *n-grami iz značajnih reči*
- *n-grami iz značajnih fraza.*

Drugi metod poboljšanja je dodavanje semantičkih atributa koji mogu biti:

- *pozicioni atributi*
- *kontekstni atributi.*

Treći metod poboljšanja je razmatranje različitih funkcija sličnosti prilikom formiranja semantičkih atributa, pri čemu su razmatrane dve mogućnosti:

- *poređenje na podudarnost*
- *poređenje na sličnost.*

Pojedinačni rezultati, kao i rezultati hibridnih metoda dati su u tabeli 9.7.

Tabela 9.7: Uporedni rezultati klasifikacije protokola po temi pri korišćenju različitih vrsta atributa

Atributi			**Mere**		
pozicioni atributi	kontekstni atributi	n-gram atributi	P	R	F
-	-	sopstvene reči	0.6667	0.8372	0.7423
		značajne reči	0.8598	0.7132	0.7797
		značajne fraze	0.8175	0.7985	**0.8078**
-	podudarnost		**0.9545**	0.4884	0.6462
		sopstvene reči	0.6750	0.8372	0.7474
		značajne reči	0.8545	0.7287	0.7866
		značajne fraze	0.8125	0.8062	**0.8093**
-	sličnost		0.8585	0.7054	0.7745
		sopstvene reči	0.6646	0.8295	0.7379
		značajne reči	0.8636	0.7364	**0.7950**
		značajne fraze	0.7984	0.7674	0.7826
podudarnost	-		0.8162	0.8605	0.8377
		sopstvene reči	0.6839	0.9225	0.7855
		značajne reči	0.8456	0.8916	**0.8679**
		značajne fraze	0.8056	0.9302	0.8633
podudarnost	podudarnost		0.8162	0.8605	0.8377
		sopstvene reči	0.6821	0.9147	0.7815
		značajne reči	0.8467	0.8992	**0.8722**
		značajne fraze	0.8108	0.9302	0.8664
podudarnost	sličnost		0.8657	0.8992	**0.8821**
		sopstvene reči	0.6784	0.8992	0.7733
		značajne reči	0.8357	0.9070	0.8699
		značajne fraze	0.8054	0.9302	0.8633
sličnost	-		0.8162	0.8605	0.8377
		sopstvene reči	0.6784	0.8992	0.7733
		značajne reči	0.8582	0.8915	**0.8745**
		značajne fraze	0.8054	0.9302	0.8633
sličnost	podudarnost		0.8148	0.8527	0.8333
		sopstvene reči	0.6802	0.9070	0.7774
		značajne reči	0.8582	0.8915	**0.8745**
		značajne fraze	0.8219	0.9302	0.8727
sličnost	sličnost		0.8837	0.8837	**0.8837**
		sopstvene reči	0.6802	0.9070	0.7774
		značajne reči	0.8345	0.8992	0.8657
		značajne fraze	0.8067	**0.9380**	0.8674

Diskusija

Najbolja F mera (0.8837) dobijena je kada su korišćeni semantički atributi uz mogućnosti poređenja do na sličnost bez pomoći atributa n-grama. U ovom slučaju dobijena je druga po redu najbolja preciznost (0.8837), kao i jedan od korektnih odziva (0.8837). Najbolji odziv (0.9380) dobijen je za istu kombinaciju semantičkih atributa, ali uz pomoć i atributa n-grama. Najbolja preciznost (0.9545) dobijena je u eksperimentu gde su se koristili samo atributi konteksta i poredili na podudarnost, ali je u tom slučaju odziv veoma slab, odnosno pronađeno je nešto manje od polovine tačnih instanci.

Može se zaključiti da se SVM metoda mašinskog učenja za klasifikaciju tekstualnih protokola u odnosu na tematiku narodna privreda u kombinaciji sa predstavljanjem dokumenata atributima n-grama u kontekstu F mere značajno popravila korišćenjem semantičkih metoda, od 0.4435 za početnu meru do 0.8837 za poboljšanje. Takođe, može se zaključiti da su kontekstni atributi sličnih mogućnosti kao i atributi n-grama kada se koriste zasebno (F mera se postiže od 0.6462 do 0.8078), ali i da postoji mogućnost određenog poboljšanja u slučajevima njihovog zajedničkog pojavljivanja.

Još se može primetiti da se kao samostalni najbolje pokazuju pozicioni atributi (što je i za očekivanje jer oni eksplicitno nabrajaju teme), s tim da ako je poređenje na podudarnost, potrebno je i dovoljno dopustiti kontrolisanu slobodu kontekstnim atributima uz poređenje na sličnost (dobijena je F mera 0.8821) ili atributima n-grama koji se sastoje od značajnih reči (dobijena je F mera 0.8679). Ako se kod pozicionih atributa dozvoli kontrolisana sloboda (poređenje na sličnost) onda algoritam ima najbolje rezultate ili uz dopuštanje još malo slobode - uz atribute n-grama značajnih reči (0.8745) ili dopuštanjem slobode kontekstnim atributima (0.8837).

Sa druge strane, previše slobode koja se dopušta uz pomoć atributa n-grama koji se dobijaju od svih sopstvenih reči restrikcije vodi ka lošijoj preciznosti (od 0.6646 do 0.6839), tako da je posledično konačna F mera slabija nego kod drugih metoda. Korišćenje n-grama značajnih fraza se pokazalo kao korisno kada nisu korišćeni semantički atributi (poboljšanje od 0.7797 do 0.8078) ili kada su korišćeni kontekstni atributi ali uz poređenje na podudarnost (poboljšanje od 0.7866 do 0.8093).

Glava 10

Zaključak

Značajna količina materijala o kulturnom nasleđu je danas dostupna putem digitalnih biblioteka. Često je potrebna dodatna obuka kako bi se te biblioteke uspešno koristile i kako bi materijali bili interpretirani na odgovarajući način. U dvadesetprvom veku se intenzivno povećava interesovanje za razvojem tehnologija koje bi se bavile istraživanjima u oblasti kulturnog nasleđa. Uspostavljeni su standardi, poput CIDOC CRM i Dublin Core, za označavanje sadržaja o kulturnom nasleđu različitim konceptima. Razvijaju se metode semantičke anotacije dokumenata koje koriste analize teksta, slika, audio i video materijala. U upotrebi su i metode obrade prirodnog jezika za različite analize, poput ekstrakcije informacija, prepoznavanja imenovanih entiteta, semantičke anotacije, ekstrakcije ključnih reči, klasifikacije teksta, pretraživanja teksta i druge.

Glavni doprinosi ove doktorske disertacije su:

- Razvoj novog modela i implementacija multimedijalne baze podataka nematerijalnog kulturnog nasleđa dela Balkana
- Razvoj informatičkog modela dokumenata i prostornog modela geografskih karakteristika sadržaja multimedijalne baze podataka nematerijalnog kulturnog nasleđa
- Razvoj biblioteke konačnih transduktora za ekstrakciju informacija iz tekstualnih dokumenata nematerijalnog kulturnog nasleđa
- Prilagođavanje metoda mašinskog učenja za tematsku klasifikaciju teksta karakteristikama dokumenata u ovoj multimedijalnoj bazi

- Unapređenje razvoja multimedijalne baze podataka kulturnog nasleđa Balkana komponentom polu-automatske anotacije metapodataka uz pomoć automatske semantičke anotacije tekstualnih dokumenata primenom razvijenih metoda ekstrakcije informacija i klasfikacije teksta

- Bogatija semantička pretraga baze podataka primenom razvijenog sistema za anotaciju metapodataka

Za polu-automatsku anotaciju multimedijalnih materijala korišćena je automatska semantička anotacija tekstualnih opisa (protokola) koji su pridruženi materijalima. Automatska semantička anotacija je sprovedena metodama ekstrakcije informacija, prepoznavanja imenovanih entiteta i ekstrakcije tema, metodama zasnovanim na pravilima uz pomoć dodatnih resursa poput elektronskih rečnika, tezaurusa i rečnika reči iz specifičnog domena.

Za klasifikaciju tekstualnih protokola u odnosu na tematiku, izvedeno je istraživanje o metodama koje se mogu primeniti za rešavanje problema klasifikacije tekstova na srpskom jeziku, i ponuđena je metoda koja je prilagođena specifičnom problemu koji se rešava (klasifikacija protokola u odnosu na postojanje teme narodna privreda), domenu koji se obrađuje (nematerijalno kulturno nasleđe) i posmatranom morfološki bogatom jeziku (srpskom jeziku).

Za rad sa prostornim podacima razvijen je interaktivni prostorni model koji je pogodan za prikaz rezultata kao i za postavljanje različitih prostornih upita uz pomoć funkcija prostornog modula PostGIS baze podatka PostgreSQL.

Rezultati eksperimenata izvedenih nad predstavljenim metodama pokazuju da korišćenje pristupa zasnovanog na pravilima za zadatak ekstrakcije informacija iz tekstova na prirodnom jeziku, u kombinaciji sa dodatnim jezičkim resursima i uz ulaganje razumnog truda daje veoma dobre rezultate.

Rezultati eksperimenata klasifikacije teksta ukazuju na zaključak da primenjene semantičke tehnike daju značajan doprinos kvalitetu klasifikacije statističkim metodama mašinskog učenja.

Pokazuje se da kontekst igra veliku ulogu u zadacima ekstrakcije informacija i klasifikacije teksta. Razvijene metode nisu iscrpne u mogućnostima, stoga postoji prostor za dalje unapređenje, što može biti zanimljiva tema za buduća istraživanja.

Domen kulturnog nasleđa je veoma bogat semantikom, tako da bi bilo korisno budući rad usmeriti ka istraživanju načina kako da se raznovrsni podaci analiziraju

i prikažu u cilju boljeg uvida u stvarnost koja je zapisana njima, ali i otkrivanja novih znanja i dodatnih veza između podataka.

Zaključak koji proizlazi iz sprovedenog istraživanja je da je za rešavanje postavljenog problema neophodno angažovanje eksperata iz više oblasti. Potrebno je u saradnji sa korisnicima sistema definisati nove zahteve koji bi bili nadogradnja postojećim početnim zahtevima. Složene su potrebe različitih grupa korisnika što usložnjava i zadatak organizacije i upravljanja multimedijalnom kolekcijom. Za dalji rad na ovom problemu se preporučuje veća uključenost eksperata iz više domena.

Poboljšanje sistema se može videti u bogatijoj vizuelizaciji podataka i interaktivnosti celog sistema sa različitim grupama korisnika, poput studenata, naučnika, eksperata iz različitih oblasti, ali i šire javnosti. Korisno bi bilo razviti dodatne metode pretrage informacija koje su dobijene uspostavljenom bogatom semantičkom shemom metapodataka. Zanimljiva tema za buduća istraživanja bi mogla biti povezivanje više aspekata podataka, poput prostora, vremena i drugih karakteristika koje se prostiru kroz prostor i vreme.

Korisno bi bilo povećati skup tekstova i razviti model koji bi mogao da iskoristi prednosti tehnika neuronskih mreža u obradi prirodnih jezika i za druge, složenije analize teksta.

Bibliografija

[1] Maristella Agosti, Hanne Albrechtsen, Nicola Ferro, Ingo Frommholz, Preben Hansen, Nicola Orio, Emanuele Panizzi, Annelise Mark Pejtersen, and Ulrich Thiel. DiLAS: A digital library annotation service. In *IWAC*, pages 91–101, 2005.

[2] Maristella Agosti and Nicola Ferro. A formal model of annotations of digital content. *ACM Transactions on Information Systems (TOIS)*, 26(1):3–es, 2007.

[3] Massimiliano Albanese, Antonio d'Acierno, Vincenzo Moscato, Fabio Persia, and Antonio Picariello. A multimedia semantic recommender system for cultural heritage applications. In *2011 IEEE fifth international conference on semantic computing*, pages 403–410. IEEE, 2011.

[4] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*, 2017.

[5] *Allegrograph*. Dostupno na adresi `https://allegrograph.com`.

[6] Flora Amato, Luca Greco, Fabio Persia, Silvestro Roberto Poccia, and Aniello De Santo. Content-based multimedia retrieval. In *Data Management in Pervasive Systems*, pages 291–310. Springer, 2015.

[7] Giuseppe Amato, Claudio Gennaro, Fausto Rabitti, and Pasquale Savino. Milos: A multimedia content management system for digital library applications. In *International Conference on Theory and Practice of Digital Libraries*, pages 14–25. Springer, 2004.

[8] Peggy M Andersen, Philip J Hayes, Alison K Huettner, Linda M Schmandt, Irene B Nirenburg, and Steven P Weinstein. Automatic extraction of facts

from press releases to generate news stories. In *Proceedings of the third conference on Applied natural language processing*, pages 170–177. Association for Computational Linguistics, 1992.

[9] *Apache OpenNLP*. Dostupno na adresi `https://opennlp.apache.org/`.

[10] Wolfgang Appelt and Nik Tetteh-Lartey. The formal specification of the iso open document architecture (ODA) standard. *The Computer Journal*, 36(3):269–279, 1993.

[11] Maria Teresa Artese and Isabella Gagliardi. A multimedia system for the management of intangible cultural heritage. *International Journal of Heritage in the Digital Era*, 4(2):149–163, 2015.

[12] *Asia-Pacific Database on Intangible Cultural Heritage*. Dostupno na adresi `http://www.accu.or.jp/ich/en/`.

[13] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[14] Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R Curran. Named entity recognition in wikipedia. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)*, pages 10–18, 2009.

[15] Ilaria Bartolini, Vincenzo Moscato, Ruggero G Pensa, Antonio Penta, Antonio Picariello, Carlo Sansone, and Maria Luisa Sapino. Recommending multimedia visiting paths in cultural heritage applications. *Multimedia Tools and Applications*, 75(7):3813–3842, 2016.

[16] Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. Fine-grained semantic textual similarity for serbian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[17] Vuk Batanović and Boško Nikolić. Sentiment classification of documents in serbian: The effects of morphological normalization and word embeddings. *Telfor Journal*, 9(2):104–109, 2017.

[18] Vuk Batanović, Boško Nikolić, and Milan Milosavljević. Reliable baselines for sentiment analysis in resource-limited languages: The serbian movie review

dataset. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2688–2696, 2016.

[19] Tim Berners-Lee and Dan Connolly. Hypertext markup language-2.0, 1995.

[20] Nikos Bikakis, Giorgos Giannopoulos, Theodore Dalamagas, and Timos Sellis. Integrating keywords and semantics on document annotation and search. In *OTM Confederated International Conferences„ On the Move to Meaningful Internet Systems"*, pages 921–938. Springer, 2010.

[21] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: Analyzing text with the natural language toolkit*. „O'Reilly Media, Inc.", 2009. Dostupno na adresi `https://www.nltk.org/`.

[22] Scott Boag, Don Chamberlin, Mary F Fernández, Daniela Florescu, Jonathan Robie, Jérôme Siméon, and Mugur Stefanescu. *XQuery 1.0: An XML query language*, 2002.

[23] Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A Greenwood, Diana Maynard, and Niraj Aswani. Twitie: An open-source information extraction pipeline for microblog text. In *Proceedings of the international conference recent advances in natural language processing RANLP 2013*, pages 83–90, 2013.

[24] Žarko Bošnjaković and Biljana Sikimić. *Bunjevci: etnodijalektološka istraživanja, 2009.* Nacionalni savet bunjevačke nacionalne manjine, 2013.

[25] Ronald Bourret et al. *XML and Databases*, 1999.

[26] Tim Bray, Jean Paoli, C Michael Sperberg-McQueen, Eve Maler, and Franois Yergeau. Extensible markup language (XML) 1.0, 2000.

[27] Tamara Butigan-Vučaj. Europeana for us: Transforming the western balkan with culture. *BOSNIACA-časopis Nacionalne i univerzitetske biblioteke Bosne i Hercegovine*, 24(24):55–59, 2019.

[28] Kate Byrne and Ewan Klein. Automatic extraction of archaeological events from text. *Proceedings of Computer Applications and Quantitative Methods in Archaeology, Williamsburg, VA*, 2010.

[29] Sanjay Chawla, Shashi Shekhar, Wei Li Wu, and Uygar Ozesmi. *Modeling spatial dependencies for mining geospatial data: An introduction*. Citeseer, 2000.

[30] Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Sriram Raghavan, Frederick R Reiss, and Shivakumar Vaithyanathan. SystemT: An algebraic approach to declarative information extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 128–137. Association for Computational Linguistics, 2010.

[31] Laura Chiticariu, Yunyao Li, and Frederick Reiss. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 827–832, 2013.

[32] Noam Chomsky and David W Lightfoot. *Syntactic structures*. Walter de Gruyter, 2002.

[33] *CIDOC Conceptual Reference Model*. Dostupno na adresi `http://whc.unesco.org`.

[34] *CIDOC Conceptual Reference Model, Last official release*. Dostupno na adresi `http://www.cidoc-crm.org/get-last-official-release`.

[35] Fabio Ciravegna, Alexiei Dingli, Daniela Petrelli, and Yorick Wilks. User-system cooperation in document annotation based on information extraction. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 122–137. Springer, 2002.

[36] *CityGML*. Dostupno na adresi `https://www.ogc.org/standards/citygml`.

[37] James Clark. Xsl transformations (XSLT). *World Wide Web Consortium (W3C)*, 103, 1999. Dostupno na adresi `http://www.w3.org/TR/xslt`.

[38] Cesare Concordia, Stefan Gradmann, and Sjoerd Siebinga. Not just another portal, not just another digital library: A portrait of europeana as an application program interface. *IFLA journal*, 36(1):61–69, 2010.

[39] Oscar Corcho. Ontology based document annotation: Trends and open research problems. *International Journal of Metadata, Semantics and Ontologies*, 1(1):47–57, 2006.

[40] Blandine Courtois and Max Silberztein. Dictionnaires électroniques du français. *Langue française*, 87(1):3–4, 1990.

[41] Jim Cowie and Wendy Lehnert. Information extraction. *Communications of the ACM*, 39(1):80–91, 1996.

[42] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2:265–292, 2002.

[43] Gregory Crane. Cultural heritage digital libraries: Needs and components. In *International Conference on Theory and Practice of Digital Libraries*, pages 626–637. Springer, 2002.

[44] Gregory Crane and Clifford Wulfman. Towards a cultural heritage digital library. In *2003 Joint Conference on Digital Libraries, 2003. Proceedings*, pages 75–86. IEEE, 2003.

[45] Rita Cucchiara, Costantino Grana, Daniele Borghesani, Maristella Agosti, and Andrew D Bagdanov. Multimedia for cultural heritage: Key issues. In *International Workshop on Multimedia for Cultural Heritage*, pages 206–216. Springer, 2011.

[46] *Cultural capital counts.* Dostupno na adresi `http://www.culturalcapitalcounts.eu/index.php/en/`.

[47] Hamish Cunningham. GATE, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254, 2002.

[48] Leonard Daly and Don Brutzman. X3D: Extensible 3D graphics standard [standards in a nutshell]. *IEEE Signal Processing Magazine*, 24(6):130–135, 2007.

[49] Albert D'Andrea and Phil Janus. UniSQL's next-generation object-relational database management system. *ACM Sigmod Record*, 25(3):70–76, 1996.

[50] *DBpedia Spotlight.* Dostupno na adresi `https://www.dbpedia-spotlight.org/`.

[51] Franciska De Jong and Thijs Westerveld. MUMIS: Multimedia indexing and searching. *Content-Based Multimedia Indexing (CBMI 2001)*, pages 423–425, 2001.

[52] Gabriel de Oliveira Barra, Mathias Lux, and Xavier Giro-i Nieto. Large scale content-based video retrieval with LIvRE. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–4. IEEE, 2016.

[53] Rafael del Hoyo, Isabelle Hupont, Francisco J. Lacueva, and David Abadía. Hybrid text affect sensing system for emotional language analysis. In *Proceedings of the international workshop on affective-aware virtual agents and social robots*, page 3. ACM, 2009.

[54] Karl Denecke. Semantic structuring of and information extraction from medical documents using the UMLS. *Methods of Information in Medicine*, 47(05):425–434, 2008.

[55] David J DeWitt, Navin Kabra, Jun Luo, Jignesh M Patel, and Jie-Bing Yu. Client-server paradise. In *VLDB*, volume 94, pages 558–569. Citeseer, 1994.

[56] *Digital Public Library of America*. Dostupno na adresi `http://dp.la`.

[57] Martin Doerr. The CIDOC CRM - an ontological approach to semantic interoperability of metadata. *Ai Magazine - AIM*, 24, 01 2003.

[58] *Doublin Core Metadata Initiative*. Dostupno na adresi `https://dublincore.org/`.

[59] Mike Dowman, Valentin Tablan, Hamish Cunningham, and Borislav Popov. Web-assisted annotation, semantic indexing and search of television and radio news. In *Proceedings of the 14th international conference on World Wide Web*, pages 225–234, 2005.

[60] Ofer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, 29(2):1–34, 2011.

[61] *Englesko-srpski paralelni koprus SrpEngKor*. Dostupno na adresi `http://www.korpus.matf.bg.ac.rs/SrpEngKor/korpus/index1.php`.

[62] *ENRICH.* Dostupno na adresi `http://vvv.cultura-strep.eu/events/enrich-2013`.

[63] Tomaž Erjavec. MULTEXT-east: Morphosyntactic resources for central and eastern european languages. *Language resources and evaluation*, 46(1):131–142, 2012.

[64] *ESRI Shapefile.* Dostupno na adresi `http://downloads.esri.com/support/whitepapers/mo_/shapefile.pdf`.

[65] Martin Ester, Hans-Peter Kriegel, and Jörg Sander. Algorithms and applications for spatial data mining. *Geographic Data Mining and Knowledge Discovery*, 5(6):600, 2001.

[66] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. Open information extraction: The second generation. In *IJCAI*, volume 11, pages 3–10, 2011.

[67] *Europeana.* Dostupno na adresi `http://www.europeana.eu`.

[68] *eXist-db.* Dostupno na adresi `http://exist-db.org`.

[69] *Extensible Stylesheet Language (XSL) Version 1.1.* Dostupno na adresi `https://www.w3.org/TR/xsl11/`.

[70] Christos Faloutsos. *Searching multimedia databases by content*, volume 3. Springer Science & Business Media, 2012.

[71] Christos Faloutsos and Douglas W Oard. A survey of information retrieval and filtering methods. Technical report, University of Maryland, 1998.

[72] Diogo Fernandes and Jorge Bernardino. Graph databases comparison: AllegroGraph, ArangoDB, InfiniteGraph, Neo4J, and OrientDB. In *DATA*, pages 373–380, 2018.

[73] Jon Ferraiolo, Fujisawa Jun, and Dean Jackson. *Scalable vector graphics (SVG) 1.0 specification.* iuniverse Bloomington, 2000.

[74] George Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar):1289–1305, 2003.

[75] *Francusko-srpski paralelni korpus SrpFranKor.* Dostupno na adresi `http://www.korpus.matf.bg.ac.rs/SrpFranKor/korpus/index1.php`.

[76] Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. In *ISMB (supplement of bioinformatics)*, pages 74–82, 2001.

[77] Norbert Fuhr and Chris Buckley. A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems (TOIS)*, 9(3):223–248, 1991.

[78] Norbert Fuhr, Giannis Tsakonas, Trond Aalberg, Maristella Agosti, Preben Hansen, Sarantos Kapidakis, Claus-Peter Klas, László Kovács, Monica Landoni, András Micsik, et al. Evaluation of digital libraries. *International Journal on Digital Libraries*, 8(1):21–38, 2007.

[79] Bojan Furlan, Vuk Batanović, and Boško Nikolić. Semantic similarity of short texts in languages with a deficient natural language processing support. *Decision Support Systems*, 55(3):710–719, 2013.

[80] Richard Furuta, Jeffrey Scofield, and Alan Shaw. Document formatting systems: Survey, concepts, and issues. *ACM Computing Surveys (CSUR)*, 14(3):417–472, 1982.

[81] *General Architecture for Text Engineering GATE.* Dostupno na adresi `http://gate.ac.uk`.

[82] *Geography Markup Language.* Dostupno na adresi `https://www.ogc.org/standards/gml`.

[83] *GeoJSON.* Dostupno na adresi `https://geojson.org/`.

[84] *Geomesa.* Dostupno na adresi `https://www.geomesa.org/`.

[85] *Geospatial Data Abstraction Layer.* Dostupno na adresi `https://gdal.org/`.

[86] Giorgos Giannopoulos, Nikos Bikakis, Theodore Dalamagas, and Timos Sellis. GoNTogle: A tool for semantic annotation and search. In *Extended Semantic Web Conference*, pages 376–380. Springer, 2010.

[87] Daniel Gildea. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 2001.

[88] *GML Application Schemas.* Dostupno na adresi `http://www.ogc.org/standards/gml#schemas`.

[89] Yoav Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.

[90] Yoav Goldberg and Omer Levy. word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[91] CF Goldfarb. Document composition facility: Generalized markup language (GML) users guide. *IBM General Products Division*, 1978.

[92] Jelena Graovac. A variant of n-gram based language-independent text categorization. *Intelligent Data Analysis*, 18(4):677–695, 2014.

[93] Jelena Graovac, Jovana Kovačević, and Gordana Pavlović-Lažetić. Hierarchical vs. flat n-gram-based text categorization: Can we do better? *Computer Science and Information Systems*, 14(1):103–121, 2017.

[94] Jelena Graovac, Miljana Mladenović, and Ivana Tanasijević. NgramSPD: Exploring optimal n-gram model for sentiment polarity detection in different languages. *Intelligent Data Analysis*, 23(2):279–296, 2019.

[95] Daniel A Griffith. *Spatial autocorrelation and spatial filtering: Gaining understanding through theory and scientific visualization.* Springer Science & Business Media, 2003.

[96] Gregor Große-Bölting, Chifumi Nishioka, and Ansgar Scherp. A comparison of different strategies for automated semantic document annotation. In *Proceedings of the 8th International Conference on Knowledge Capture*, pages 1–8, 2015.

[97] Sandra Gucul-Milojević. Personal names in information extraction. *INFOtheca-Journal of Informatics & Librarianship*, 11(1), 2010.

[98] Ramanathan V Guha, Dan Brickley, and Steve Macbeth. Schema.org: Evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51, 2016.

[99] Juliet L Hardesty. Exhibiting library collections online: Omeka in context. *New Library World*, 2014.

[100] Fouzi Harrag, Eyas El-Qawasmah, and Abdul Malik S Al-Salman. Comparing dimension reduction techniques for arabic text classification using bpnn algorithm. In *2010 First International Conference on Integrated Intelligent Computing*, pages 6–11. IEEE, 2010.

[101] Jochen Hartmann, Juliana Huppertz, Christina Schamp, and Mark Heitmann. Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1):20–38, 2019.

[102] Bernhard Haslhofer, Wolfgang Jochum, Ross King, Christian Sadilek, and Karin Schellner. The LEMO annotation framework: Weaving multimedia annotations with the web. *International Journal on Digital Libraries*, 10(1):15–32, 2009.

[103] Sundus Hassan, Muhammad Rafi, and Muhammad Shahid Shaikh. Comparing svm and naïve bayes classifiers for text categorization with wikitology as knowledge enrichment. In *2011 IEEE 14th International Multitopic Conference*, pages 31–34. IEEE, 2011.

[104] Goffredo Haus and Luca A Ludovico. The digital opera house: An architecture for multimedia databases. *Journal of Cultural Heritage*, 7(2):92–97, 2006.

[105] *Healthcare Schema*. Dostupno na adresi `https://www.w3.org/community/schemed/`.

[106] Willemijn Heeren, Laurens van der Werff, Franciska de Jong, Roeland Ordelman, Thijs Verschoor, Arjan van Hessen, and Mies Langelaar. Easy listening: Spoken document retrieval in choral. *Interdisciplinary science reviews*, 34(2-3):236–252, 2009.

[107] John Herring. OpenGIS® implementation standard for geographic information-simple feature access-part 1: Common architecture [corrigendum]. 2011.

[108] Timm Heuss, Bernhard Humm, Christian Henninger, and Thomas Rippl. A comparison of NER tools wrt a domain-specific vocabulary. In *Proceedings of the 10th International Conference on Semantic Systems*, pages 100–107, 2014.

[109] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):797–819, 2011.

[110] Minlie Huang, Aurélie Névéol, and Zhiyong Lu. Recommending MeSH terms for annotating biomedical articles. *Journal of the American Medical Informatics Association*, 18(5):660–667, 2011.

[111] James N Hughes, Andrew Annex, Christopher N Eichelberger, Anthony Fox, Andrew Hulbert, and Michael Ronquest. Geomesa: A distributed architecture for spatio-temporal fusion. In *Geospatial Informatics, Fusion, and Motion Video Analytics V*, volume 9473, page 94730F. International Society for Optics and Photonics, 2015.

[112] *IBM DB2*. Dostupno na adresi `https://www.ibm.com/products/db2-database`.

[113] *IBM SPSS Modeler*. Dostupno na adresi `https://www.ibm.com/products/spss-modeler`.

[114] *IBM SPSS Modeler Documentation*. Dostupno na adresi `https://www.ibm.com/support/knowledgecenter/SS3RA7_sub/modeler_kc_subscription/clementine/knowledge_center/product_landing_subscription.html`.

[115] *ImageAXS*. Dostupno na adresi `https://manualsbrain.com/en/products/digital-arts-sciences-imageaxs-4-dot-1-for-windows/`.

[116] *Informix*. Dostupno na adresi `https://www.ibm.com/products/informix`.

[117] Pavle Ivić. *Istorija srpske kulture*. Dečje novine, 1994.

[118] Jelena B Jaćimović. *Automatsko prepoznavanje i normalizacija vremenskih izraza u nestrukturiranim novinskim i medicinskim tekstovima na srpskom jeziku*. PhD thesis, University of Belgrade, Faculty of Phylology, 2016.

[119] *JavaFX*. Dostupno na adresi `https://openjfx.io`.

[120] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, 2019.

[121] Ridong Jiang, Rafael E Banchs, and Haizhou Li. Evaluating and combining name entity recognition systems. In *Proceedings of the Sixth Named Entity Workshop*, pages 21–27, 2016.

[122] Thorsten Joachims. Making large-scale SVM learning practical. Technical report, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1998.

[123] Thorsten Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers, 2002.

[124] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[125] Sofija Jovanović. *Gradja za tezaurus u oblasti etnologije*, 2003. Dostupno na adresi `www.nb.rs`.

[126] Ridwan Andi Kambau and Zainal Arifin Hasibuan. Concept-based multimedia information retrieval system using ontology search in cultural heritage. In *2017 Second International Conference on Informatics and Computing (ICIC)*, pages 1–6. IEEE, 2017.

[127] Howard Katz, Donald Dean Chamberlin, Denise Draper, Mary Fernandez, Michael Kay, Jonathan Robie, Michael Rys, Jerome Simeon, Jim Tivy, and Philip Wadler. *XQuery from the experts: A guide to the W3C XML query language*. Addison-Wesley Professional, 2004.

[128] Michael Kay. *XSLT 2.0 and XPath 2.0 Programmer's Reference*. John Wiley & Sons, 2011.

[129] Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125, 2006.

[130] Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, pages 255–264, 2003.

[131] *Keyhole Markup Language.* Dostupno na adresi `https://www.ogc.org/standards/kml`.

[132] Wahab Khan, Ali Daud, Jamal A Nasir, and Tehmina Amjad. A survey on the state-of-the-art machine learning models in the context of NLP. *Kuwait journal of Science*, 43(4), 2016.

[133] Setrag Khoshafian and Brad Baker. *Multimedia and imaging databases.* Morgan Kaufmann, 1996.

[134] Milan Kilibarda and Dragutin Protić. *Geovizualizacija i Web kartografija.* Građevinski fakultet, Beograd, 2018.

[135] Sunhyuck Kim, Jaeyeon Ahn, Juhee Suh, Hayun Kim, and Jungwha Kim. Towards a semantic data infrastructure for heterogeneous cultural heritage data-challenges of korean cultural heritage data model (kchdm). In *2015 Digital Heritage*, volume 2, pages 275–282. IEEE, 2015.

[136] Atanas Kiryakov. Ontology and reasoning in MUMIS: Towards the semantic web. Technical report, Technical Report CS–03–03, Department of Computer Science, University of Sheffield, 2003.

[137] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. Semantic annotation, indexing, and retrieval. *Journal of Web Semantics*, 2(1):49–79, 2004.

[138] Donald Ervin Knuth. *TEX: The Program.* Addison-Wesley, 1986.

[139] *Korpus savremenog srpskog jezika SrpLemKor.* Dostupno na adresi `http://www.korpus.matf.bg.ac.rs/SrpLemKor`.

[140] Harald Kosch. *Distributed multimedia database technologies supported by MPEG-7 and MPEG-21.* CRC Press, 2003.

[141] Gerald J Kowalski. *Information retrieval systems: Theory and implementation*, volume 1. Springer, 2007.

[142] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pages 61–68, 2009.

[143] Cvetana Krstev. *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. University of Belgrade, Faculty of Philology, Belgrade, 2008.

[144] Cvetana Krstev and Denis Maurel. A note on the semantic and morphological properties of proper names in the prolex project. *Lingvisticae Investigationes*, 30(1):115–133, 2007.

[145] Cvetana Krstev, Ivan Obradović, Miloš Utvić, and Duško Vitas. A system for named entity recognition based on local grammars. *Journal of Logic and Computation*, 24(2):473–489, 2014.

[146] Cvetana Krstev, Staša Stanković-Vujičić, and Duško Vitas. Approximate measures in the culinary domain: Ontology and lexical resources. In *Proceedings of the 9th Language Technologies Conference IS-LT*, pages 38–43, 2014.

[147] Cvetana Krstev, Andjelka Zečević, Duško Vitas, and T Kyriakopoulou. Nerosetta–an insight into named entity tagging. In *6th Language and Technology Conference*, pages 168–172, 2013.

[148] Dilek Küçük and Adnan Yazıcı. Exploiting information extraction techniques for automatic semantic video indexing with an application to turkish news videos. *Knowledge-Based Systems*, 24(6):844–857, 2011.

[149] *Kulturno nasleđe: putovanje kroz vreme i prostor*. Dostupno na adresi `https://euinfo.rs/kulturno-nasledje-putovanje-kroz-vreme-i-prostor/`.

[150] Richard Kurin. Safeguarding intangible cultural heritage in the 2003 UNESCO convention: A critical appraisal. *Museum international*, 56(1-2):66–77, 2004.

[151] Michal Laclavik, Martin Šeleng, Marek Ciglan, and Ladislav Hluchỳ. Ontea: Platform for pattern based automated semantic annotation. *Computing and informatics*, 28(4):555–579, 2012.

[152] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[153] Ora Lassila and Deborah McGuinness. The role of frame-based representation on the semantic web. *Linköping Electronic Articles in Computer and Information Science*, 6(5):2001, 2001.

[154] Robert Laurini and Derek Thompson. *Fundamentals of spatial information systems*, volume 37. Academic press, 1992.

[155] Ji Young Lee and Franck Dernoncourt. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*, 2016.

[156] Elizabeth D Liddy. Natural language processing. 2001.

[157] Chern Li Liew. Online cultural heritage exhibitions: a survey of information retrieval features. *Program*, 2005.

[158] Ling Liu and M Tamer Özsu. *Encyclopedia of database systems*, volume 6. Springer New York, NY, USA, 2009.

[159] Nikola Ljubešić, Marija Stupar, Tereza Jurić, and Željko Agić. Combining available datasets for building named entity recognition models of croatian and slovene. *Slovenščina 2.0: Empirical, applied and interdisciplinary research*, 2:35–57, 2013.

[160] *Lombardia Digital Archive*. Dostupno na adresi `http://www.aess.itc.cnr.it/ricerca/ricerca_src/home_page.php`.

[161] *Lombardia Digital Archive, Intangible Heritage*. Dostupno na adresi `http://intangiblesearch.eu/home_page.php?db_name=intangible_search&lingua=inglese`.

[162] Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. Scientific information extraction with semi-supervised neural tagging. *arXiv preprint arXiv:1708.06075*, 2017.

[163] *MAchine Learning for LanguagE Toolkit*. Dostupno na adresi `http://mallet.cs.umass.edu/`.

[164] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford CoreNLP natural language

processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations*, pages 55–60, 2014. Dostupno na adresi `https://nlp.stanford.edu/software/CRF-NER.html`.

[165] Cristopher Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to information retrieval*. Cambridge University Press, 2009.

[166] *MarkLogic Server*. Dostupno na adresi `https://www.marklogic.com/product/marklogic-database-overview/`.

[167] Melvin Earl Maron. Automatic indexing: An experimental inquiry. *Journal of the ACM (JACM)*, 8(3):404–417, 1961.

[168] Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. Sentiment classification using word sub-sequences and dependency sub-trees. In *PAKDD*, volume 5, pages 301–311. Springer, 2005.

[169] Denis Maurel. Prolexbase: A multilingual relational lexical database of proper names. In *LREC*, 2008.

[170] Diana Maynard and Jonathon Hare. Entity-based opinion mining from text and multimedia. In *Advances in Social Media Analysis*, pages 65–86. Springer, 2015.

[171] Michael McCandless, Erik Hatcher, and Otis Gospodnetić. *Lucene in action*, volume 2. Manning Greenwich, 2010.

[172] Kevin McGuinness, Noel E O'Connor, Robin Aly, Franciska De Jong, Ken Chatfield, Omkar M Parkhi, Relja Arandjelovic, Andrew Zisserman, Matthijs Douze, and Cordelia Schmid. The AXES PRO video search system. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 307–308, 2013.

[173] Olena Medelyan. *Human-competitive automatic topic indexing*. PhD thesis, The University of Waikato, 2009.

[174] *Medical Subject Headings MeSH*. Dostupno na adresi `https://www.nlm.nih.gov/mesh/introduction.html`.

[175] Wolfgang Meier. eXist: An open source native XML database. In *Net. Object-Days: International Conference on Object-Oriented and Internet-Based Technologies, Concepts, and Applications for a Networked World*, pages 169–183. Springer, 2002.

[176] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[177] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[178] Miloš Milenković. O očuvanju nacionalnog identiteta i kulturne baštine u evropskim integracijama: osnovne zablude i značajnije mogućnosti. *Etnoantropološki problemi*, 8(2):453–470, 2013.

[179] George A Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[180] *Mirage*. Dostupno na adresi `https://hop.at/mirage`.

[181] Jelena D Mitrović. *Elektronski jezički resursi i alati za obradu srpskog jezika i njihovo unapređivanje putem modela grupne raspodele rada*. PhD thesis, Univerzitet u Beogradu-Filološki fakultet, 2018.

[182] Miljana Mladenović, Cvetana Krstev, Jelena Mitrović, and Ranka Stanković. Using lexical resources for irony and sarcasm classification. In *Proceedings of the 8th Balkan Conference in Informatics*, pages 1–8, 2017.

[183] Miljana Mladenović, Jelena Mitrović, Cvetana Krstev, and Duško Vitas. Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems*, 46(3):599–620, 2016.

[184] *MonetDB database system with XQuery front-end*. Dostupno na adresi `https://www.monetdb.org/XQuery`.

[185] *MonetDB/SQL*. Dostupno na adresi `https://www.monetdb.org/`.

[186] *MPEG-7 Standard*. Dostupno na adresi `https://www.iso.org/contents/data/standard/03/77/37778.html`.

[187] SVM Multiclass. *SVM Multiclass*. Dostupno na adresi `http://www.cs.cornell.edu/people/tj/svmlight/svmmulticlass.html`.

[188] Chuck Musciano and Bill Kennedy. *HTML & XHTML: The Definitive Guide: The Definitive Guide*. „ O'Reilly Media, Inc.", 2002.

[189] *Museum24*. Dostupno na adresi `http://www.museo24.fi`.

[190] David Myers, Mario Santana Quintero, Alison Dalgity, and Ioannis Avramides. The arches heritage inventory and management system: A platform for the heritage field. *Journal of Cultural Heritage Management and Sustainable Development*, 2016.

[191] *MySQL*. Dostupno na adresi `https://www.mysql.com`.

[192] *Nacionalni centar za digitalizaciju*. Dostupno na adresi `http://www.ncd.org.rs/ncd_sr/aboutncd.html`.

[193] *Narodna biblioteka Srbije*. Dostupno na adresi `https://www.nb.rs/about_us/icitem.php?id=34445`.

[194] *National Database of Intangible Cultural Heritage of India*. Dostupno na adresi `http://www.sangeetnatak.gov.in/sna/national-inventory.htm`.

[195] *National Research Institute of Cultural Heritage of Korea*. Dostupno na adresi `http://www.nrich.go.kr/english/index.do`.

[196] Apostol Natsev, John R Smith, Jelena Tešić, Lexing Xie, and Rong Yan. IBM multimedia analysis and retrieval system. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 553–554, 2008.

[197] *Nematerijalno kulturno nasledje Srbije*. Dostupno na adresi `http://www.serbia.com/srpski/o-srbiji/kultura/nematerijalno-kulturno-nasledje-srbije/`.

[198] *New South Voices*. Dostupno na adresi `http://nevsouthvoices.uncc.edu`.

[199] Kamal Nigam, John Lafferty, and Andrew McCallum. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, pages 61–67. Stockholom, Sweden, 1999.

[200] Vojkan Nikolić, Branko Markoski, Miodrag Ivković, Kristijan Kuk, and Predrag Đikanović. Information retrieval for unstructured text documents in serbian into the crime domain. In *2015 16th IEEE International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 267–271. IEEE, 2015.

[201] The General Conference of the United Nations Educational. *Text of the Convention for the Safeguarding of the Intangible Cultural Heritage*, 2003. Dostupno na adresi `https://ich.unesco.org/en/convention`.

[202] Zoran Ognjanovć, Tamara Butigan-Vučaj, and Bojan Marinković. NCD recommendation for the national standard for describing digitized heritage in serbia. In *Metadata and Semantics*, pages 45–54. Springer, 2009.

[203] Zoran Ognjanović, Bojan Marinković, Marija Šegan-Radonjić, and Dejan Masliković. Cultural heritage digitization in serbia: Standards, policies, and case studies. *Sustainability*, 11(14):3788, 2019.

[204] Yukio Ohsawa, Nels E Benson, and Masahiko Yachida. KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries-ADL'98-*, pages 12–18. IEEE, 1998.

[205] Radoslav Đokić. *Prožimanja kultura*. Univerzitet umetnosti, 1976.

[206] Pedro Oliveira and Joao Rocha. Semantic annotation tools survey. In *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 301–307. IEEE, 2013.

[207] *Open Geospatial Consortium*. Dostupno na adresi `http://www.ogc.org/`.

[208] *Oracle*. Dostupno na adresi `https://www.oracle.com/index.html`.

[209] Vesna Pajić, Gordana Pavlović-Lažetić, and Miloš Pajić. Information extraction from semi-structured resources: a two-phase finite state transducers approach. In *International Conference on Implementation and Application of Automata*, pages 282–289. Springer, 2011.

[210] Vesna Pajić, Staša Stanković-Vujičić, and Miloš Pajić. Transducers for annotating weather information in meteorological texts in serbian. *INFOtheca–Journal of Informatics & Librarianship*, 13(2):36–51, 2012.

[211] Vesna Pajić, Staša Stanković-Vujičić, Ranka Stanković, and Miloš Pajić. Semi-automatic extraction of multiword terms from domain-specific corpora. *The Electronic Library*, 36(3):550–567, 2018.

[212] Vesna Pajić, Duško Vitas, Gordana Pavlović-Lažetić, and Miloš Pajić. Web-monitoring software system: Finite state machines for monitoring the web. *Computer Science and Information Systems*, 10(1):1–23, 2013.

[213] Elias Pampalk. Audio-based music similarity and retrieval: Combining a spectral similarity model with information extracted from fluctuation patterns. In *Proceedings of the International Symposium on Music Information Retrieval*, 2006.

[214] BV Patel and BB Meshram. Content based video retrieval systems. *arXiv preprint arXiv:1205.1641*, 2012.

[215] Sébastien Paumier. *Unitex-manuel d'utilisation*, 2011.

[216] Gordana Pavlović-Lažetić and Jelena Graovac. Ontology-driven conceptual document classification. In *KDIR*, pages 383–386, 2010.

[217] Dave Pawson. *XSL-FO: Making XML look good in print.* „O'Reilly Media, Inc.", 2002.

[218] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[219] Anna Plotnikova. *Materials for ethnolinguistic investigation of the Balkan-Slavic area.* Institute of Slavic Studies of the Russian Academy of Sciences, 1996.

[220] *PostGIS.* Dostupno na adresi `http://postgis.net`.

[221] *PostGIS Specification.* Dostupno na adresi `https://download.osgeo.org/postgis/docs/postgis-2.5.4.pdf`.

[222] *PostgreSQL.* Dostupno na adresi `https://www.postgresql.org/`.

[223] Joseph D Prusa and Taghi M Khoshgoftaar. Designing a better data representation for deep neural networks and text classification. In *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, pages 411–416. IEEE, 2016.

[224] Erasmo Purificato and Antonio M Rinaldi. Multimedia and geographic data integration for cultural heritage information retrieval. *Multimedia Tools and Applications*, 77(20):27447–27469, 2018.

[225] Lawrence Reeve and Hyoil Han. Survey of semantic annotation platforms. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 1634–1638, 2005.

[226] Dennis Reidsma, Jan Kuper, Thierry Declerck, Horacio Saggion, and Hamish Cunningham. Cross document annotation for multimedia retrieval. In *EACL Workshop on Language Technology and the Semantic Web (NLPXML), Budapest, Hungary*, 2003.

[227] Julian Richards, Stuart Jeffrey, Stewart Waller, Fabio Ciravegna, Sam Chapman, and Ziqi Zhang. The archaeology data service and the archaeotools project: Faceted classification and natural language processing. *Archaeology*, 2:31–56, 2011.

[228] Jonathan Robie. XQuery: A guided tour. *XQuery from the Experts. Addison Wesley*, pages 3–78, 2004.

[229] Douglas Roland and Dan Jurafsky. How verb subcategorization frequencies are affected by corpus choice. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1122–1128, 1998.

[230] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: Applications and theory*, 1:1–20, 2010.

[231] Horacio Saggion, Hamish Cunningham, Kalina Bontcheva, Diana Maynard, Oana Hamza, and Yorik Wilks. Multimedia indexing through multi-source and multi-language information extraction: The MUMIS project. *Data & Knowledge Engineering*, 48(2):247–264, 2004.

[232] Branislava Šandrih, Cvetana Krstev, and Ranka Stanković. Development and evaluation of three named entity recognition systems for serbian-the case of personal names. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1060–1068, 2019.

[233] Sunita Sarawagi. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377, 2008.

[234] *Scalable Vector Graphics.* Dostupno na adresi `https://www.w3.org/Graphics/SVG/`.

[235] *Schema Bib Extend Community.* Dostupno na adresi `https://www.w3.org/community/schemabibex/`.

[236] *Schema.org.* Dostupno na adresi `https://schema.org/`.

[237] Martin Scholz and Guenther Goerz. WissKI: A virtual research environment for cultural heritage. In *ECAI*, volume 242, pages 1017–1018. Citeseer, 2012.

[238] Guus Schreiber, Alia Amin, Lora Aroyo, Mark van Assem, Victor de Boer, Lynda Hardman, Michiel Hildebrand, Borys Omelayenko, Jacco van Osenbruggen, Anna Tordai, et al. Semantic annotation and search of cultural-heritage collections: The MultimediaN e-culture demonstrator. *Journal of Web Semantics*, 6(4):243–249, 2008.

[239] *SciKit.* Dostupno na adresi `https://scikit-learn.org`.

[240] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[241] *Serbian-English Law Finance Education and Health SELFEH.* Dostupno na adresi `http://www.korpus.matf.bg.ac.rs/prezentacija/selfeh.html`.

[242] *Sheffield's Amilcare.* Dostupno na adresi `http://staffwww.dcs.shef.ac.uk/people/F.Ciravegna/Amilcare.html`.

[243] Shashi Shekhar and Sanjay Chawla. *A tour of spatial databases.* Prentice Hall Upper Saddle River, 2003.

[244] Shashi Shekhar, Sanjay Chawla, Sivakumar Ravada, Andrew Fetterer, Xuan Liu, and Chang-tien Lu. Spatial databases-accomplishments and research needs. *IEEE transactions on knowledge and data engineering*, 11(1):45–55, 1999.

[245] Shashi Shekhar, Pusheng Zhang, and Yan Huang. Spatial data mining. In *Data mining and knowledge discovery handbook*, pages 837–854. Springer, 2009.

[246] Erik Siegel and Adam Retter. *eXist: A NoSQL Document Database and Application Platform*. „ O'Reilly Media, Inc.", 2014.

[247] *Siri*. Dostupno na adresi `https://www.apple.com/siri/`.

[248] Joan M Smith. Standard generalized markup language and related standards. *Computer Communications*, 12(2):80–84, 1989.

[249] John Snow. On the mode of transmission of cholera. *Churchill, London*, 1855.

[250] Stephen Soderland. Learning information extraction rules for semi-structured and free text. *Machine learning*, 34(1-3):233–272, 1999.

[251] *spaCy*. Dostupno na adresi `https://spacy.io/`.

[252] Caroline Sporleder. Natural language processing for cultural heritage domains. *Language and Linguistics Compass*, 4(9):750–768, 2010.

[253] *Sport Schema Community Group*. Dostupno na adresi `https://www.w3.org/community/sport-schema/`.

[254] *SprKor2013*. Dostupno na adresi `http://www.korpus.matf.bg.ac.rs`.

[255] *SQL Server*. Dostupno na adresi `https://www.microsoft.com/en-us/sql-server/sql-server-2019`.

[256] *Srpska kultura*. Dostupno na adresi `http://www.serbia.com/srpski/o-srbiji/kultura/`.

[257] *Srpski wordnet SrpWN*. Dostupno na adresi `http://korpus.matf.bg.ac.rs/SrpWN`.

[258] Ranka Stanković, Cvetana Krstev, Ivan Obradović, and Olivera Kitanović. Indexing of textual databases based on lexical resources: A case study for serbian. In *International KEYSTONE Conference on Semantic Keyword-Based Search on Structured Data Sources*, pages 167–181. Springer, 2015.

[259] Ranka Stanković, Cvetana Krstev, Ivan Obradović, and Olivera Kitanović. Improving document retrieval in large domain specific textual databases using lexical resources. In *Transactions on Computational Collective Intelligence XXVI*, pages 162–185. Springer, 2017.

[260] Ranka Stanković, Cvetana Krstev, Duško Vitas, Nikola Vulović, and Olivera Kitanović. Keyword-based search on bilingual digital libraries. In *Semanitic Keyword-based Search on Structured Data Sources*, pages 112–123. Springer, 2016.

[261] Eduard Suess. *Antlitz der Erde*, volume 1. Clarendon Press, 1904.

[262] Sai Sumathi and S Esakkirajan. *Fundamentals of relational database management systems*, volume 47. Springer, 2007.

[263] Mahbubur Rahman Syed. *Multimedia Technologies: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*, volume 3. IGI Global, 2008.

[264] Ivana Tanasijević and Gordana Pavlović-Lažetić. HerCulB: Content-based information extraction and retrieval for cultural heritage of the Balkans. *The electronic library*, 2020. DOI: `https://doi.org/10.1108/EL-03-2020-0052`.

[265] Ivana Tanasijević. Toward automatic tagging of cultural heritage documents. *IPSI Transactions on Advanced Research, TAR*, 15(1), 2019.

[266] Ivana Tanasijević and Gordana Pavlović-Lažetić. Pretraživanje po sadržaju multimedijalne baze nematerijalnog nasleđa. *SKUP 35. godina računarske lingvistike u Srbiji*, pages 87–98, 2014.

[267] Ivana Tanasijević, Biljana Sikimić, and Gordana Pavlović-Lažetić. Multimedia database of the cultural heritage of the Balkans. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2874–2881. Citeseer, 2012.

[268] Aleksandra Terzić, Željko Bjeljac, and Nevena Ćurčić. Common histories, constructed identities: Intangible cultural heritage and the rebranding of serbia. *International Journal of Intangible Heritage*, 10:102–120, 2015.

[269] Larry Tesler. *Pub, the Document Compiler.* Stanford University. Computer Science Department. Artificial Intelligence, 1973.

[270] *The Automotive Ontology Working Group.* Dostupno na adresi `https://www.w3.org/community/gao/`.

[271] *The Getty Research Institute, Vocabularies.* Dostupno na adresi `https://www.getty.edu/research/tools/vocabularies/index.html`.

[272] *The Getty Thesaurus of Geographic Names TGN.* Dostupno na adresi `https://www.getty.edu/research/tools/vocabularies/tgn/index.html`.

[273] *The rich story of North American square dance.* Dostupno na adresi `http://squaredancehistory.org`.

[274] Henry S Thompson, Noah Mendelsohn, D Beech, and M Maloney. W3C XML schema definition language (XSD) 1.1 part 1: Structures. *The World Wide Web Consortium (W3C), W3C Working Draft Dec*, 3, 2009.

[275] Bhavani Thuraisingham. *Managing and mining multimedia databases.* CRC Press, 2001.

[276] *Tika.* Dostupno na adresi `https://tika.apache.org`.

[277] Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

[278] Andrija Tomović, Predrag Janičić, and Vlado Kešelj. n-Gram-based classification and unsupervised hierarchical clustering of genome sequences. *Computer methods and programs in biomedicine*, 81(2):137–153, 2006.

[279] Dolf Trieschnigg, Piotr Pezik, Vivian Lee, Franciska De Jong, Wessel Kraaij, and Dietrich Rebholz-Schuhmann. MeSH up: Effective MeSH text classification for improved document retrieval. *Bioinformatics*, 25(11):1412–1418, 2009.

[280] Aleksandra S Trtovac. *Deskriptori metapodataka i deskriptori sadržaja u pronalaženju informacija u digitalnim bibliotekama.* PhD thesis, University of Belgrade, Faculty of Philology, 2016.

[281] Ioannis Tsochantaridis. Support vector learning for interdependent and structured output spaces. In *Proc. International Conference on Machine Learning (ICML), 2004*, 2004.

[282] Suppawong Tuarob, Line C Pouchard, and C Lee Giles. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 239–248, 2013.

[283] INTELLIGENCE BY AM TURING. Computing machinery and intelligence-AM turing. *Mind*, 59(236):433, 1950.

[284] *Uffizi Gallery, Italy.* Dostupno na adresi `https://www.uffizi.it/en/the-uffizi`.

[285] *UNESCO Country page, Serbia.* Dostupno na adresi `http://whc.unesco.org/en/statesparties/rs`.

[286] *UNESCO World Heritage Centre.* Dostupno na adresi `http://whc.unesco.org`.

[287] *Unitex.* Dostupno na adresi `www-igm.univ-mlv.fr/~unitex`.

[288] Victoria Uren, Philipp Cimiano, José Iria, Siegfried Handschuh, Maria Vargas-Vera, Enrico Motta, and Fabio Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of Web Semantics*, 4(1):14–28, 2006.

[289] Miloš Utvić. *Izgradnja referentnog korpusa savremenog srpskog jezika (The construction of reference corpus of contemporary Serbian).* PhD thesis, University of Belgrade, Faculty of Philology, 2014.

[290] Maria Vargas-Vera, Emanuela Moreale, Arthur Stutt, Enrico Motta, and Fabio Ciravegna. MnM: Semi-automatic ontology population from text. In *Ontologies*, pages 373–402. Springer, 2007.

[291] Maria Vargas-Vera, Enrico Motta, John Domingue, Mattia Lanzoni, Arthur Stutt, and Fabio Ciravegna. MnM: Ontology driven semi-automatic and automatic support for semantic markup. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 379–391. Springer, 2002.

[292] Marilena Vecco. A definition of cultural heritage: From the tangible to the intangible. *Journal of Cultural Heritage*, 11(3):321–324, 2010.

[293] S Vijayarani and A Sakila. Multimedia mining research-an overview. *International Journal of Computer Graphics & Animation*, 5(1):69, 2015.

[294] Duško Vitas, Cvetana Krstev, Ivan Obradović, Ljubomir Popović, and Gordana Pavlović-Lažetič. An overview of resources and basic tools for processing of serbian written texts. In *Proc. of the Workshop on Balkan Language Resources, 1st Balkan Conference in Informatics*. Citeseer, 2003.

[295] Duško Vitas and Gordana Pavlović-Lažetić. Resources and methods for named entity recognition in serbian. *INFOtheca-Journal of Informatics & Librarianship*, 9, 2008.

[296] Duško Vitas, Ljubomir Popović, Cvetana Krstev, Ivan Obradović, Gordana Pavlović-Lažetić, and Mladen Stanojević. *Srpski jezik u digitalnom dobu – The Serbian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. Dostupno na adresi `http://www.meta-net.eu/whitepapers`.

[297] Andreas Vlachidis and Douglas Tudhope. A knowledge-based approach to i nformation e xtraction for semantic interoperability in the archaeology domain. *Journal of the association for information science and technology*, 67(5):1138–1152, 2016.

[298] Teodora Vuković and Maja Milićević. Creation and some ideas for classrom use of an electronic corpus of the dialect of bunjevci. *Minority languages in education and language learning: Challenges and new perspectives*, page 353, 01 2017.

[299] Teodora Vuković, Muheim Nora, Olivier-Andreas Winistörfer, Makarova Anastasia, Šimko Ivan, and Bradjan Sanja. Corpora and processing tools for non-standard contemporary and diachronic balkan slavic. 2019.

[300] Teodora Vuković and Tanja Samardžić. *Areal distribution of the frequency of the post-positive article in the Timok vernacular of Southeast Serbia / Prostorna raspodela frekvencije postpozitivnog člana u timočkom govoru*, pages 181–199. 01 2018.

[301] Teodora Vuković and Biljana Sikimić. *Digitalna zaštita bunjevačkih govora*, 10 2017.

[302] Teodora Vuković, Biljana Sikimić, and Bratislav Vukojičić. *Timočki govori 2015-2016*. Dostupno na adresi `http://balksrv2012.sanu.ac.rs/webdict/timok/index`.

[303] Teodora Vuković. Spoken torlak dialect corpus 1.0 (transcription), 2020. Slovenian language resource repository CLARIN.SI.

[304] Bernhard Waltl, Georg Bonczek, and Florian Matthes. Rule-based information extraction: Advantages, limitations, and perspectives. *Jusletter IT (02 2018)*, 2018.

[305] Avery Wang. An industrial strength audio search algorithm. In *Ismir*, volume 2003, pages 7–13. Washington, DC, 2003.

[306] Pu Wang and Carlotta Domeniconi. Building semantic kernels for text classification using wikipedia. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 713–721. ACM, 2008.

[307] *Web3D Consortium*. Dostupno na adresi `https://www.web3d.org/`.

[308] *Well Known Text*. Dostupno na adresi `https://www.ogc.org/standards/wkt-crs`.

[309] Kevin Williams, Michael Brundage, Patrick Dengler, Jeff Gabriel, Andy Hoskinson, Michael R Kay, Thomas Maxwell, Marcelo Ochoa, Johnny Papa, and Mohan Vanmane. *Professional XML databases*. Wrox press Birmingham, UK, 2000.

[310] Ian H Witten, Gordon W Paynter, Eibe Frank, Carl Gutwin, and Craig G Nevill-Manning. Kea: Practical automated keyphrase extraction. In *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*, pages 129–152. IGI global, 2005.

[311] *World Wide Web Consortium*. Dostupno na adresi `https://www.w3.org`.

[312] Marcel Worring, Cees GM Snoek, Ork de Rooij, Giang P Nguyen, and Arnold WM Smeulders. The mediamill semantic video search engine. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–1213. IEEE, 2007.

[313] *Xindice*. Dostupno na adresi `https://xml.apache.org/xindice/`.

[314] Akane Yakushiji, Yusuke Miyao, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. Automatic construction of predicate-argument structure patterns for biomedical information extraction. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 284–292, 2006.

[315] Eric Zavesky and Shih-Fu Chang. Cuzero: Embracing the frontier of interactive visual search for informed users. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 237–244, 2008.

[316] Anđelka Zečević. N-gram based text classification according to authorship. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 145–149, 2011.

[317] Ziqi Zhang, Sam Chapman, and Fabio Ciravegna. A methodology towards effective and efficient manual document annotation: Addressing annotator discrepancy and annotation quality. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 301–315. Springer, 2010.

Biografija autora

Ivana D. Tanasijević (rođena u Kragujevcu, 24.03.1983. godine) pohađala je osnovnu školu "Milutin i Draginja Todorović" u Kragujevcu, koju je završila 1998. godine kao nosilac Vukove diplome, diplome iz matematike i diplome iz fizike. Srednju školu "Prva kragujevačka gimnazija", specijalizovano odeljenje za talentovane učenike za matematiku i računarstvo po programu Matematičke gimnazije u Beogradu, završila je 2002. godine kao nosilac Vukove diplome. Upisala je Prirodno matematički fakultet Univerziteta u Kragujevcu 2002. godine, smer računarstvo i informatika, nakon čega, 2004. godine, nastavlja studije na Matematičkom fakultetu Univerziteta u Beogradu, takođe na smeru računarstvo i informatika. Matematički fakultet u Beogradu završila je 2008. godine sa prosečnom ocenom 9.62 (od 10.00) i pohvalom fakulteta za postignute uspehe tokom studija. Na istom fakultetu upisala je doktorske studije u okviru kojih je položila izabrane predmete sa prosečnom ocenom 10.00 (od 10.00).

U zvanje saradnik u nastavi na Katedri za računarstvo i informatiku Matematičkog fakulteta Univerziteta u Beogradu prvi put je izabrana 2009. godine. U zvanje asistenta na istoj Katedri prvi put je izabrana 2011. godine. Držala je vežbe iz predmeta "Programiranje 1 (programski jezik C)", "Uvod u arhitekturu računara", "Arhitektura računara (programiranje u asembleru)", "Operativni sistemi (napredno programiranje u Unix sistemu)", "Računarske mreže (mrežno programiranje u Unix sistemu)" i "Projektovanje baza podataka".

Tokom studija bila je korisnik više stipendija, među kojima su stipendije Grada Kragujevca, Vlade Republike Srbije i HipoAlpe Adria banke (posredstvom privrednog društva "Privrednik"). Izabrana je 2008. godine u okviru projekta "Putujemo u Evropu" u cilju upoznavanja drugih kultura i vrednosti, kao jedan od najboljih studenata završnih godina svih privatnih i državnih fakulteta u Srbiji.

Прилог 1.

Изјава о ауторству

Потписани-а Ивана Танасијевић

број индекса 2032/2009

Изјављујем

да је докторска дисертација под насловом

Мултимедијалне базе података у управљању нематеријалним културним наслеђем

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, 15.11.2020.

Ивана Танасијевић

Прилог 2.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Ивана Танасијевић

Број индекса 2032/2009

Студијски програм информатика

Наслов рада Мултимедијалне базе података у управљању нематеријалним културним наслеђем

Ментор проф. др Гордана Павловић-Лажетић

Потписани/а Ивана Танасијевић

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, 15.11.2020.

Ивана Танасијевић

Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Мултимедијалне базе података у управљању нематеријалним културним наслеђем

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство

2. Ауторство - некомерцијално

(3.) Ауторство – некомерцијално – без прераде

4. Ауторство – некомерцијално – делити под истим условима

5. Ауторство – без прераде

6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

Потпис докторанда

У Београду, 15.11.2020.

Ивана Филиповић

1. Ауторство - Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. Ауторство – некомерцијално. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. Ауторство - некомерцијално – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. Ауторство - некомерцијално – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. Ауторство – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. Ауторство - делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.