ZAKLJUČNO POROČILO O REZULTATIH OPRAVLJENEGA RAZISKOVALNEGA DELA NA PROJEKTU V OKVIRU CILJNEGA RAZISKOVALNEGA PROGRAMA (CRP) »KONKURENČNOST SLOVENIJE 2006 – 2013«

I. Predstavitev osnovnih podatkov raziskovalnega projekta

1. Naziv težišča v okviru CRP:

KONKURENČNO GOSPODARSTVO IN HITREJŠA RAST

2. Šifra projekta:

V2-0213

3. Naslov projekta:

»Sistemi za statistični semantični splet«

11

3. Naslov projekta

3.1. Naslov projekta v slovenskem jeziku:

»Sistemi za statistični semantični splet«

3.2. Naslov projekta v angleškem jeziku:

System for statistical Semantic Web

4. Ključne besede projekta

4.1. Ključne besede projekta v slovenskem jeziku:

lahke ontologije, čezmodalnost, čezjezičnost, skalabilnost

4.2. Ključne besede projekta v angleškem jeziku:

light weight ontologies, cross-modal, cross-lingual, scalability

5. Naziv nosilne raziskovalne organizacije:

Institut "Jožef Stefan"

5.1. Seznam sodelujočih raziskovalnih organizacij (RO):

6. Sofinancer/sofinancerji:

Ministrstvo za visoko šolstvo, znanost in tehnologijo

7. Šifra ter ime in priimek vodje projekta:

12570	Dunja Mladenič

Datum: 14.10.2008.

Podpis vodje projekta:

doc. dr. Dunja Mladenič

Podpis in žig izvajalca:

prof. dr. Jadran Lenarčič, direktor

Institut "Jožef Stefan" Ljubljana, Slovenija

II. Vsebinska struktura zaključnega poročila o rezultatih raziskovalnega projekta v okviru CRP

1. Cilji projekta:

1.1. Ali so bili cilji projekta doseženi?

- [x] a) v celoti
- [ ] b) delno
- [ ] c) ne

Če b) in c), je potrebna utemeljitev.

1.2. Ali so se cilji projekta med raziskavo spremenili?

- [ ] a) da
- [x] b) ne

Če so se, je potrebna utemeljitev:

2. Vsebinsko poročilo o realizaciji predloženega programa dela[1]:

Delo na projektu je potekalo skladno s predloženim programom in sicer predvsem v smeri izdelave komponente za zajemanje plitvega znanja, komponente za čezmodalnost, komponente za čezjezičnost in nekaj manjših komponent za za polavtomatiko, vizualizacijo in evalvacijo. Pri izdelavi komponent smo sledili že prej definiranim specifikacijam funkcionalnosti komponent.

1. Zajemanje plitvega znanja iz podatkov in predstavitev v obliku »lahke ontologije« (angl. »light weight ontologies«) ter zapis tega znanja v nekaj načinih vključno z standardnimi načini za zapisovanja znanja kot so W3C specifikacije RDF, in OWL, kar bo omogočilo visoko povezljivost oz. interoperabilnost z drugimi sistemi za modeliranje in manipulacijo z znanjem.
2. Čezmodalnost (angl. Cross-Modal) v smislu dela z besedili, grafi oz. omrežji, sliko, kar bomo dosegli z uporabo nekaj pristopov, ki omogočajo kombiniranje različnih predstavitev podatkov v enotne modele. Kombiniranje bomo praktično izvedli tako, da bomo objekte zapisane v raznih modalnostih zapisali v skupnem jeziku nad katerim lahko uporabimo računsko učinkovite metode. Skupni jezik bodo predvidoma redki vektorji značilk oz. atributov (angl. »sparse vectors of features«) nad katerimi lahko uporabimo učinkovite metode iz linearne algebre. Ključen korak za čezmodalnost pa bo način generiranja značilk iz posameznih modalnosti – glede na aplikacije znotraj projekta bomo izbrali modalnosti, med njimi vsekakor besedila.
3. Čezjezičnost (angl. Cross-Lingual) v smislu obravnave besedil zapisanih v različnih jezikih. Šele v zadnjem času so se pojavili korpusi in nekatere rešitve, ki omogočajo delo s takimi korpusi v večjem obsegu. Na IJS razvijamo rešitve, ki slonijo na adaptaciji tradicionalne statistične metode »kanonična korelacijska analiza« (angl. »canonical correlation analysis«). Osnovna ideja je, da dokumente zapisane v enem od podprtih jezikov (za katere imamo poravnan večjezični korpus) lahko zapišemo na jezikovno nevtralen način. Tak zapis omogoča manipulacijo z vsebino dokumentov ne glede na jezik v katerem je dokument zapisan.

Pri razvoju upoštevamo tudi potrebe dveh, v projektu predvidenih, aplikaciji: spremljanje medijske podobe Slovenije v svetovnih pisanih medijih, klasifikacija in povezovanje slovenskih pravnih besedil v evropsko klasifikacijsko shemo EuroVoc.

1. Za namen testiranja in uproabe sistema na predvidenih aplikacijah smo razvili module potebne za zajemanje plitvega znanja iz podatkov (delno opisano v objavi 4). (1) Razvili smo modul za pridobivanje podatkov iz evropske klasifikacijske sheme EvroVoc. EuroVoc je večjezična klasifikacijska shema, ki pokriva zakonodajo. Uporablja jo Evropski parlament, nacionalni parlamenti v Evropi, vladni uradi in evropske organizacije. Zadnja verzija EuroVoc 4.2 obstaja v 21 uradnih jezikih Evropske Unije (Bulgarian, Spanish, Czech, Danish, German, Estonian, Greek, English, French, Italian, Latvian, Lithuanian, Hungarian, Dutch, Polish, Portuguese, Romanian, Slovak, Slovene, Finnish and Swedish). EuroVoc je javno dostopen na spletnih straneh http://europa.eu/eurovoc/ od kod smo izluščili vsebino s pomočjo razvitega modula za zajemanje spletnih podatkov (crawler) posebej prilagojen za strukturo EuroVoc podatkov (razvoj prilagojenega module je bil

[1] Potrebno je napisati vsebinsko raziskovalno poročilo, kjer mora biti na kratko predstavljen program dela z raziskovalno hipotezo in metodološko-teoretičen opis raziskovanja pri njenem preverjanju ali zavračanju vključno s pridobljenimi rezultati projekta.

nepričakovan vendar nujen, ker na žalost, uradno predlagan način pridobivanja podaktov ni deloval). Večjezični del brez strukture smo dobili na spletni straneh v obliki MS Excel datoteke. Izluščeni podatki vsebujejo 13416 vozlišč dveh vrst (6645 descriptors, ostala non-descriptors) in 5 tipov relacij. (2) V sodelovanju z EU projektom 6.O.P. NeOn smo nad našo knjižnico Text-Garden razvili modul za branje in shranjevanje pridobljenih podatkov (lahke ontologije - angl. light weight ontology) EuroVoc2OntoLight.exe, ki sharni ontologijo v binarno datoteko "EuroVoc.OntoLight" in njeno tekstovno predstavitev shrani v tekstovno datoteko "EuroVoc.OntoLight.Txt". (3) Razvili smo modul OntoLight2OntoCfier.exe za prirejanje dokumentov ontologiji (angl. ontology grounding), ki iz ontologije shranjene v formatu ".OntoLight" zgradi model in ga zapiše v datoteko ".OntoCfier" (lahko tudi v tekstovni obliki ".OntoCfier.Txt"). (4) Razširili smo sistem za polavtomatsko gradnjo ontologij OntoGen s komponentno OntoLight, kar omogoča prikaz uporabniških dokumentov (kot so slovenski pravni dokumenti) v kontekstu EuroVoc klasifikacijske sheme.

2. Delo na razvoju komponente za čezmodalnost (angl. Cross-Modal) smo zasnovali na povezovanju podatkov predstavljenih v obliki besedil in v obliki grafov oz. omrežji (opisano v objavah 2 in 3). Nadgradili smo obstoječ pristop za predstavitev grafov z vektorji značilk in razvili dva alternativna pristopa, vsak uporablja drugačno obteževanje značilk pri predstavitvi grafov. Kombiniranje smo izvedli tako, da objekte zapisane v raznih modalnostih (besedilo, graf) zapišemo v skupnem jeziku nad katerim lahko uporabimo računsko učinkovite metode. Za skupni jezik smo izbrali redke vektorje značilk oz. atributov (angl. »sparse vectors of features«) nad katerimi lahko uporabimo učinkovite metode iz linearne algebre. (1) Razvili smo komponento za čezmodalnost v sodelovanju s EU projektom 6.O.P. TAO, kjer je konkreten primer čezmodalnosti sicer nekoliko drugačen - obelava programske kode z upoštevanjem besedila v kodi (imena spremenljivk, komentrji,...) in grafa medsebojnega referenciranja funkcij. (2) Razširili smo sistem za polavtomatsko gradnjo ontologij OntoGen, tako da omogoča delo na podatkih, ki vsebujejo čezmodalne podatke pripravljene z razvito komponentno.

3. Razvili smo metodologijo za avtomatsko generiranje lematizatorja in jo overednotili v poskusih lematizacije Slovenščine (glej objavo 0). Razvili smo komponento za čezjezičnost, ki je zasnovana na uporabi statistične metode KCCA, ki preslika besedila različnih jezikov v skupni, semantični prostor. V sodelovanju s EU projektom 6.O.P. SMART smo izboljšali učinkovitosti uporabljene KCCA metode (glej objavo 5). Zgradili smo model za čezjezičnost na osnovi dokumentov korpusa evropske zakonodaje Acquis Communitarian, ki so prevedeni v različne jezike Evropske Unije. Uporabnost metode smo prikazali na prototipu za čezjezično iskanje informacij (cross-language information retrieval) (glej objavo 7).

4. Skalabilnost, polavtomatske metode in evalvacija. Posebno pozornost posvečamo skalabilnosti sistema, da bo sposoben obravnavati velike količine podatkov, kar bo omogočilo izvedbo predvidenih projektnih aplikacij. Uporabljamo polavtomatske metode, ki kombinirajo »top-down« in »bottom-up« pristope. Rešitev je zasnovana na delih sistema »OntoGen« (http://ontogen.ijs.si), ki implementira polavtomatkse metode – sistem uporablja vrsta uporabnikov po svetu za polavtomatsko modeliranje znanja. V smislu podpore evalvacije rezultatov smo izdelali komponento za avtomatsko evalvacijo ontologij, ki implementira metodo objavljeno v Brank et al., 2006 (glej objavo 6).

5. Za potrebe aplikacije spremljanja medijske podobe Slovenije v svetovnih pisanih medijih smo izdelali komponentno za vizualizacijo agencijskih novic, ki smo jo uporabili na novicah svetovnih medijev (Reuters novice in New York Times novice). Izdelali smo vmesnik za vizualizacijo podatkov, ki implementirta metodo za vizualizacijo besedil skozi čas (kot so n.pr., novice) in ga povezali s komponento za analizo agencijskih novic, ki smo jo prav tako razvili v projektu. Za potrebe aplikacije spremljanja medijske podobe Slovenije smo izdelali komponento za zajemanje novic s spletnega portala znane svetoven medijske hiše New York Times. V dogovoru z New York Times smo tako pridobili njihove novice od leta 1851 do vključno 2007. Aplikacija analize medisjke podobe Slovenije zajema analizo vsebine novic skozi čas, izločanje imenskih entitet iz novic in analizo omrežja sopojavitev imenskih entitet v novicah. Ob tem je potekalo delo na testiranju razvitih komponent in sistema.

6. Za potrebe aplikaciji klasifikacije in povezovanja slovenskih pravnih besedil v evropsko klasifikacijsko shemo EuroVoc smo razvili modul za pridobivanje podatkov iz evropske klasifikacijske sheme EuroVoc. Na osnovi korpusa evropske zakonodaje Acquis Communitarian, ki so označeni s pojmi (descriptors) iz EuroVoc sheme smo zgradili model za avtomatsko klasifikacijo besedil v EuroVoc shemo. Model smo integrirali v obstoječi sistem za polavtomatsko gradnjo ontologij OntoGen (glej točko 1.) in s tem podprli kontekstno odvisno izdelavo pravnih ontologij.

1.01.
0. PLISSON, Joël, LAVRAČ, Nada, MLADENIĆ, Dunja, ERJAVEC, Tomaž. Ripple down rule learning for automated word lemmatisation. AI commun., 2008, vol. 21, no. 1, str. 15-26.

1.08
1. FORTUNA, Blaž, GROBELNIK, Marko, MLADENIĆ, Dunja. Semi-automatic data-driven ontology construction system. In: Zbornik 9. mednarodne multikonference Informacijska družba IS 2006, 9. do 14. oktober 2006, Ljubljana: Institut "Jožef Stefan", 2006, str. 223-226.
(objave delno zajema tudi prispevek tega projekta k specifikacijam, čeprav je večinoma rezultat preteklega dela).

2. GRČAR, Miha, GROBELNIK, Marko, MLADENIĆ, Dunja. Using text mining and link analysis for software mining. In: Proceedings of the Third International Workshop on Mining Comlex Data, MCD 2007, ECML PKDD 2007, The 18th European Conference on Machine Learning and the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, September 21, 2007, Warsaw, Poland., 2007, pp. 1-12.

3. GROBELNIK, Marko, MLADENIĆ, Dunja, FORTUNA, Blaž. From social network to light-weight ontology. In: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07), Hyderabad, India, 06-12 January, 2007 : workshop on Text-Mining & Link-Analysis (TextLink 2007). Hyderabad: International Institute of Information Technology, 2007, 11 str.

4. GROBELNIK, Marko, BRANK, Janez, FORTUNA Blaž, MOZETIČ, Igor. Contextualizing ontologies with ontolight: a pragmatic approach. In: Zbornik 10.

mednarodne multikonference Informacijska družba IS 2007, Ljubljana: Institut "Jožef Stefan", 2007.

5. FORTUNA, Blaž, CRISTIANINI, Nello, SHAWE-TAYLOR, John. A Kernel Canonical Correlation Analysis For Learning The Semantics Of Text. In Kernel methods in bioengineering, communications and image processing, edited by G. Camps-Valls, J. L. Rojo-Álvarez & M. Martínez-Ramón, 2006.

1.16

6. BRANK, Janez, GROBELNIK, Marko, MLADENIĆ, Dunja. Automatic evaluation of ontologies. V: KAO, Anne (ur.), POTEET, Stephen R. (ur.). Natural language processing and text mining. London: Springer, cop. 2007, str. 193-219.

7. FORTUNA, Blaž, RUPNIK, Jan, PAJNTAR, Boštjan, GROBELNIK, Marko, MLADENIĆ, Dunja. Cross-Lingual Search over 22 European Languages. In proceedings of the 10th International Conference on Information Retrieval SIGIR-2008.

2.25
8. MLADENIĆ, Dunja (ed.), GROBELNIK, Marko (ed.). Text mining and link analysis for web and semantic web, KDD 2007 tutorial notes, T4, 2007.

3. Izkoriščanje dobljenih rezultatov:

3.1. Kakšen je potencialni pomen[2] rezultatov vašega raziskovalnega projekta za:

- ☐ a) odkritje novih znanstvenih spoznanj;
- ☒ b) izpopolnitev oziroma razširitev metodološkega instrumentarija;
- ☐ c) razvoj svojega temeljnega raziskovanja;
- ☐ d) razvoj drugih temeljnih znanosti;
- ☒ e) razvoj novih tehnologij in drugih razvojnih raziskav.

3.2. Označite s katerimi družbeno-ekonomskimi cilji (po metodologiji OECD-ja) sovpadajo rezultati vašega raziskovalnega projekta:

☐ a) razvoj kmetijstva, gozdarstva in ribolova - Vključuje RR, ki je v osnovi namenjen razvoju in podpori teh dejavnosti;

☐ b) pospeševanje industrijskega razvoja - vključuje RR, ki v osnovi podpira razvoj industrije, vključno s proizvodnjo, gradbeništvom, prodajo na debelo in drobno, restavracijami in hoteli, bančništvom, zavarovalnicami in drugimi gospodarskimi dejavnostmi;

☐ c) proizvodnja in racionalna izraba energije - vključuje RR-dejavnosti, ki so v funkciji dobave, proizvodnje, hranjenja in distribucije vseh oblik energije. V to skupino je treba vključiti tudi RR vodnih virov in nuklearne energije;

☐ d) razvoj infrastrukture - Ta skupina vključuje dve podskupini:

- transport in telekomunikacije - Vključen je RR, ki je usmerjen v izboljšavo in povečanje varnosti prometnih sistemov, vključno z varnostjo v prometu;
- prostorsko planiranje mest in podeželja - Vključen je RR, ki se nanaša na skupno načrtovanje mest in podeželja, boljše pogoje bivanja in izboljšave v okolju;

☐ e) nadzor in skrb za okolje - Vključuje RR, ki je usmerjen v ohranjevanje fizičnega okolja. Zajema onesnaževanje zraka, voda, zemlje in spodnjih slojev, onesnaženje zaradi hrupa, odlaganja trdnih odpadkov in sevanja. Razdeljen je v dve skupini:

☐ f) zdravstveno varstvo (z izjemo onesnaževanja) - Vključuje RR - programe, ki so usmerjeni v varstvo in izboljšanje človekovega zdravja;

☐ g) družbeni razvoj in storitve - Vključuje RR, ki se nanaša na družbene in kulturne probleme;

☒ h) splošni napredek znanja - Ta skupina zajema RR, ki prispeva k splošnemu napredku znanja in ga ne moremo pripisati določenim ciljem;

☐ i) obramba - Vključuje RR, ki se v osnovi izvaja v vojaške namene, ne glede na njegovo vsebino, ali na možnost posredne civilne uporabe. Vključuje tudi varstvo (obrambo) pred naravnimi nesrečami.

---

[2] Označite lahko več odgovorov.

3.3. Kateri so **neposredni rezultati** vašega raziskovalnega projekta glede na zgoraj označen potencialni pomen in razvojne cilje?

Razvili smo metodologijo za zajemanje plitvega znanja in jo implementirali v programskih modulih: zajemanje znanja iz podatkov, shranjevanje pridobljenih podatkov, avtomatsko prirejanje dokumentov ontologiji, kontekstno gradnjo ontologij.
Razvili smo metodologijo za povezovanju podatkov predstavljenih v obliki besedil in v obliki grafov oz. omrežji in jo implementirali v sistemu za analizo programskih okolji.
Nadgradili smo metodo za čezjezičnost in jo implementorali v progrmaskem modulu, ter pokazali njeno uporabnost na problemu čezjezičnega iskanja informacij.

3.4. Kakšni so lahko **dolgoročni rezultati** vašega raziskovalnega projekta glede na zgoraj označen potencialni pomen in razvojne cilje?

Povezovanje razvitih metod in modulov v kompleksne sisteme za delo s semantičnimi podatki, kot je Evropsk sistem za gradnjo mrežnih ontologij NeOn Toolkit.
Uporaba in prilagoditev razvitih metod in modulov na različnih problemih, kot je gradnja ontologij na osnovi čezmodalnih podatkov in podpora čezjezičnosti.
Nadgradanj izdelanih prototipnih aplikacij za potrebe naročnikov doma in v tujini. Odličan primer je interes, ki ga za našo aplikacijo analize medijskih novic že kaže svetovno znana agencija New York Times.

3.5. Kje obstaja verjetnost, da bodo vaša znanstvena spoznanja deležna zaznavnega odziva?

- ☒ a) v domačih znanstvenih krogih;
- ☒ b) v mednarodnih znanstvenih krogih;
- ☐ c) pri domačih uporabnikih;
- ☒ d) pri mednarodnih uporabnikih.

3.6. Kdo (poleg sofinancerjev) že izraža interes po vaših spoznanjih oziroma rezultatih?

Raziskovalci predvsem iz področja semnatičnih tehnologij, na glavnih mednarodnih kofnerencah iz področja smo imeli predstavitve in demonstracije našega dela vključno s tutorialom na treh najpomembnejših konferencah iz področja semantičnega spleta (svetovna, evropska in azijska) ISWC-2007, ESWC-2008, ASWC-2008.
Mednarodne inštitucije s katerimi sodelujemo na Evropskih projektih 6.O.P. IP NeOn, 6.O.P. STREP TAO, 6.O.P. STREP SWING, 6.O.P. STREP SMART, 6. O.P. STREP IMAGINATION.
Mednarodno uveljavljene korporacije, kot so Microsoft Research, New York Times, Nature, British Telecom, Yahoo.

3.7. Število diplomantov, magistrov in doktorjev, ki so zaključili študij z vključenostjo v raziskovalni projekt?

magisterij Jure Ferlež, diploma Blaž Novak

4. Sodelovanje z tujimi partnerji:

4.1. Navedite število in obliko formalnega raziskovalnega sodelovanja s tujimi raziskovalnimi inštitucijami.

Delo na projektu je potekalo v sodelovanju s partnerji na Evropskih projektih 6.O.P. NeOn, TAO, SWING, SMART, na katerih sodelujemo.

4.2. Kakšni so rezultati tovrstnega sodelovanja?

Sodelovanje je potekalo v več raziskovalnih smeri in v je po potrebi zajemalo tako raziskovalno delo in razvoj metodologij kot tudi razvoj programskih modulov in njihovo povezovanje s obstoječimi sistemi. Podrobnosti so podane v vsebinskem poročilu (točka 2 tega poročila), kjer ob opisih različnih sklopov podamo tudi navezavo na konkretne projekte.

5. Bibliografski rezultati[3] :

Za vodjo projekta in ostale raziskovalce v projektni skupini priložite bibliografske izpise za obdobje zadnjih treh let iz COBISS-a) oz. za medicinske vede iz Inštituta za biomedicinsko informatiko. Na bibliografskih izpisih označite tista dela, ki so nastala v okviru pričujočega projekta.

[3] Bibliografijo raziskovalcev si lahko natisnete sami iz spletne strani:http:/www.izum.si/

6. Druge reference[4] vodje projekta in ostalih raziskovalcev, ki izhajajo iz raziskovalnega projekta:

Prezentacija prototipa za analizo vsebin novic skozi čas in analizo omrežja akterjev v novicah, sestanek na podjetju New York Times, New York, USA, Januar, 2008. Prezentacija nove verzije prototipa in pogovori o sodelovanju na podjetju New York Times, New York, USA, April 2008.

A.01
GROBELNIK, Marko, BRANK, Janez, FORTUNA Blaž, MOZETIČ, Igor. Contextualizing ontologies with ontolight: a pragmatic approach. Informatica (Ljubljana), 2008, vol 32, no. 1.

V članku predlagamo pristop za zajemanje in shranjevanje plitvega zanja, ki omogoča postavitev obstoječe ontologije v kontekst zajetega znanja. Pristop temelji na modelu preprostih oz. lahkih ontologij (angl. light-weight ontologies) in povezovanjem konceptov ontologije z dejanskimi primeri konceptov (angl. ontology grounding). S temi predpostavkami je mogoče učinkovito implementirati osnovne operacije nad ontologijo (klasifikacija, dodajanje vsebine ter določanje preslikav med ontologijami), to pa nam omogoča izkoriščati več velikih ontologij kot vir kontekstualnega znanja o problemskem področju. V članku predstavljamo primer scenarija, kako uporabiti kontekstualne informacije pri polavtomatski izgradnji ontologije iz skupine besedil. Metodo smo implementirali in uporabili na relanih podatkih znanih ontologij EuroVoc, AgroVoc, ASFA, DMoz in Cyc.

A.03
BRANK, Janez, GROBELNIK, Marko, MLADENIĆ, Dunja. Automatic evaluation of ontologies. In: KAO, Anne, POTEET, Stephen R. (eds.). Natural language processing and text mining. London: Springer, cop. 2007, pp. 193-219.

V poglavju v knjigi predlagamo novo metodo za evalvacijo ontologij, podamo pregled metod za evalvacijo ontologij in teoretične okvire v katere umestimo predlagano metodo. Metoda, ki smo jo razvili je predvsem namenjena avtomatski evalvaciji ontologij, ki vključujejo primere. Pristop je zasnovan na principu primerjave z obstoječo ontologijo (angl. golden standard paradigm) v smislu pokritja in razporeditve primerov v ontologijo, medtem ko obstoječi pristopi predvsem uporabljajo primerjavo imen konceptov in relacij (ne pa primere). Metodo smo poimenovali OntoRandIndex, ker temelji na meri Rand index, ki je znana metoda za primerjavo razvrstitev podatkov v skupine. Pomembno je še poudariti, da naš pristop ni odvisen od predstavitve primerov, to kar je pomembno je, da lahko ločimo en primer od drugega. Predlagano metodo smo ovrednotili na relanih podatkih poddrevesa javno dostopne spletne hierarhije DMoz.

B.03 GRČAR, Miha, GROBELNIK, Marko, MLADENIĆ, Dunja. Using text mining and

[4] Navedite tudi druge raziskovalne rezultate iz obdobja financiranja vašega projekta, ki niso zajeti v bibliografske izpise, zlasti pa tiste, ki se nanašajo na prenos znanja in tehnologije.
Navedite tudi podatke o vseh javnih in drugih predstavitvah projekta in njegovih rezultatov vključno s predstavitvami, ki so bile organizirane izključno za naročnika/naročnike projekta.

link analysis for software mining. In: Proceedings of the Third International Workshop on Mining Comlex Data, MCD 2007, ECML PKDD 2007, The 18th European Conference on Machine Learning and the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, September 21, 2007, Warsaw, Poland., 2007, pp. 1-12.

Predlagali smo pristop za statistično analizo programske opreme, ki temelji na povezovanju podatkov predstavljenih v obliki besedil (n.p.r. dokumentacija in komentarji programske kode) in v obliki grafov oz. omrežji (n.p.r. dedovanje med razderi v objektni kodi, klici parametrov). Nadgradili smo obstoječ pristop za predstavitev grafov z vektorji značilk in razvili dva alternativna pristopa, vsak uporablja drugačno obteževanje značilk pri predstavitvi grafov. Kombiniranje smo izvedli tako, da objekte zapisane v raznih modalnostih (besedilo, graf) zapišemo v skupnem jeziku nad katerim lahko uporabimo računsko učinkovite metode. Predlagan pristop smo uporabili na relanih podatkih velikega programskega paketa za analizo naravnega jezika GATE razvitega na Univerzi v Sheffieldu.

B.01 MADENIĆ, Dunja, GROBELNIK, Marko. Text mining and link analysis for web and semantic web, KDD 2007 tutorial notes, 2007.

Imeli smo celodnevno pregledno predavanje na eni izmed najuglednejših mednarodnih konferenc iz področja odkrivanja zakonitosti v podatkih KDD-2007. Predavanje je bilo iz področja analize besedil za potrebe spletnih analiz in semantičnega spleta. Med ostalim smo predstavili tudi nekaj lastnih metod in rešitev, med njimi tudi znanstvene dosežke tega projekta.

B.01. GROBLENIK, Marko, FORTUNA, Blaž, MLADENIČ, Dunja. What Semanitc Webers need to knwo about Machine Learning?, ISWC-2007 tutorial, October 2007.

Imeli smo celodnevno pregledno predavanje na eni izmed najuglednejših mednarodnih kofenrenc iz področja semnatičnega spleta, ISWC-2007. Predavanje je bilo iz področja uporabe metod strojnega učenja za potrebe raziskav in razvoja na področju semantičnega spleta. Med ostalim smo predstavili tudi nekaj lastnih metod in rešitev, med njimi tudi znanstvene in ravojne dosežke tega projekta.

POVZETEK

Delo na projektu je potekalo skladno s predloženim programom in sicer predvsem v smeri izdelave (1) komponente za zajemanje plitvega znanja, (2) komponente za čezmodalnost, (3) komponente za čezjezičnost in (4) nekaj manjših komponent za za polavtomatiko, vizualizacijo in evalvacijo. Pri razvoju upoštevamo tudi potrebe dveh, v projektu predvidenih, aplikaciji: spremljanje medijske podobe Slovenije v svetovnih pisanih medijih, klasifikacija in povezovanje slovenskih pravnih besedil v evropsko klasifikacijsko shemo EuroVoc.

V projektu smo razvili metodologijo za zajemanje plitvega znanja (1) in jo implementirali za potrebe evropske klasifikacijske sheme EvroVoc. Implementirani programski moduli omogočajo: zajemanje znanja iz podatkov, shranjevanje pridobljenih podatkov, avtomatsko prirejanje dokumentov ontologiji, kontekstno gradnjo ontologij. Sistem smo testirali na korpusu besedil evropske zakonodaje. Delo je potekalo v sodelovanju s projektom 6.O.P. NeOn. Delo na razvoju komponente za čezmodalnost (2) smo zasnovali na povezovanju podatkov predstavljenih v obliki besedil in v obliki grafov oz. omrežji. Metodologijo smo implementirali v programski komponenti za analizo čezmodalnih podatkov, ki smo jo povezali z obstoječim sistemom za polavtomatsko gradnjo ontologij OntoGen. Predlagani pristop smo testirali na problemu analize programske knjižnice GATE (University of Sheffield, UK). Delo je potekalo v sodelovanju s projektom 6.O.P. TAO in SWING. Nadgradili smo metodo za čezjezičnost (3) v smislu učinkovitosti in jo implementirali v programski komponenti, ki smo jo povezali z obstoječim sistemom za kontekstno odvisno spletno iskanje SearchPoint. Uporabnost predlaganega pristopa smo pokazali na problemu čezjezičnega iskanja informacij. Delo je potekalo v sodelovanju s projektom 6.O.P. SMART in PASCAL. Posebno pozornost smo posvetili skalabilnosti predlaganih pristopov (4). V smislu podpore evalvacije rezultatov smo v sodelovanju s projektom 6.O.P. SEKT izdelali komponento za avtomatsko evalvacijo ontologij.

Za potrebe aplikacije spremljanja medijske podobe Slovenije v svetovnih pisanih medijih smo izdelali komponentno za vizualizacijo agencijskih novic, ki smo jo uporabili na novicah svetovnih medijev (Reuters novice in New York Times novice). Za potrebe aplikaciji klasifikacije in povezovanja slovenskih pravnih besedil v evropsko klasifikacijsko shemo EuroVoc smo zgradili model za avtomatsko klasifikacijo besedil v EuroVoc shemo. Model smo integrirali v obstoječi sistem za polavtomatsko gradnjo ontologij OntoGen in s tem podprli kontekstno odvisno izdelavo pravnih ontologij. Delo je potekalo v sodelovanju s projektom 6.O.P. NeOn.

ABSTRACT

The project has progressed according to the plan with focus on development of methods and components for (1) knowledge acquisition, (2) cross-modal data analysis, (3) handling cross-lingual data and several support components for (4) supporting scalability, visualization and evaluation. Research and development was guided also by the two prototype applications that we have envisioned: monitoring of Slovenian media image from international news and, support for analysis of Slovenian legal documents using European classification schema EuroVoc.

We have developed a methodology for knowledge acquisition (1) in the form of light-weight ontology. The methodology was implemented for European classification schema EuroVoc, by realizing modules that enable knowledge acquisition, automatic labeling of legal documents using EuroVoc and contextualized ontology generation. The system was tested on European legislation corpus. This work was in collaboration with FP6 project NeOn. Development of methodology for cross-modal data analysis (2) was focused on connecting textual data and graph/network. It was implemented as a component for cross-modal data analysis that we have integrated into an existing system for semi-automatic ontology generation OntoGen. The approach was applied on software mining of a large natural language processing library GATE (University of Sheffield, UK). The work was performed in collaboration with FP6 projects TAO and SWING. Extension of the KCCA method for cross-lingual data analysis (3) in terms of scalability was performed in collaboration with FP6 projects SMART and PASCAL. The developed component was integrated with the existing system for contextual search SearchPoint and applied on the problem of cross-lingual information retrieval. Through all the development special emphases was put on (4) scalability and visualization. In collaboration with FP6 project SEKT we have developed a methodology for automatic evaluation of ontologies.

The prototype application for monitoring Slovenian media image from international news was addressed by a system for visualization of news and actors in the news. It was applied on publicly available Reuters news and on the news archive of New York Times. The prototype application for analysis of legal documents by relating them to European classification schema EuroVoc was focused on enabling contextualized ontology generation, where the context is European legislation. The work was performed in collaboration with FP6 project NeOn.