SUPERVIZOR+

Marko Bajec[1], Štefan Furlan[2], Aleš Kumer[1], Lovro Šubelj[1], Slavko Žitnik[2], Dejan Lavbič[1]
{marko.bajec, ales.kumer, lovro.subelj, dejan.lavbic}@fri.uni-lj.si
{stefan.furlan, slavko.zitnik}@optilab.net

Abstract

SUPERVIZOR+

Supervizor+ is a simple web application that helps to identify and visualise patterns of relationships among business and physical entities in Slovenia. The application is based on automatic information retrieval from various data sources publicly available on the web, such as AJPES, UJP, ZZS, Supervizor, Wikipedia, Facebook, Google+, LinkedIn, Finance, Dnevnik, Delo, etc. Widely known as web content mining, automatic information retrieval from web is not an easy task and represents a complex challenge for practitioners and researchers. In this paper we shortly represent the idea Supervisor+ is based on as well as the technologies behind its implementation. While Supervisor+ clearly demonstrates the ability of state of the art techniques in web content mining, its ambition is higher; i.e. to become a community tool where users will not only be able to browse and mine the acquired data but also to contribute their personal knowledge to the system.

Ključne besede

Ekstrakcija podatkov, spletno rudarjenje, integracija podatkov, socialna omrežja

Key words

Information Retrieval, Web Mining, Data Integration, Social Networks

1. AVTOMATSKA EKSTRAKCIJA PODATKOV S SPLETA

V *Laboratoriju za podatkovne tehnologije* Fakultete za računalništvo in informatiko v Ljubljani se ukvarjamo z *zajemom*, *predstavitvijo*, *obdelavo*, *analizo* in *vizualizacijo* podatkov. Eno od področij, v katerega vlagamo veliko raziskovalnega napora, je izdelava metod in pristopov za avtomatsko ekstrakcijo podatkov, ki so javno ali prosto dostopni. Najbolj obsežen takšen vir je splet, kjer je na voljo nepregledna množica podatkov, zajetih v tekstih, objavljenih na spletni straneh, v blogih, v raznih socialnih omrežjih itn. Dodatno so na spletu dostopne ogromne količine slik in videoposnetkov, ki prav tako predstavljajo vir prosto dostopnih podatkov. Večina teh podatkov je nestrukturiranih – na primer zapisi v prostem tekstu brez definirane strukture. Kar nekaj pa je tudi virov, ki so bodisi strukturirani skladno z definirano obliko, na primer spletna enciklopedija Wikipedia (podmnožica teh podatkov je na voljo tudi v strukturirani obliki v DBpediji) ali morda javni registri. Takšni so npr. javno dostopni podatki o poslovnih subjektih, ki jih ureja in v delno strukturirani obliki objavlja Agencija Republike Slovenije za javnopravne evidence in storitve, AJPES ter drugi (npr. ZZZS, PIRS, GZS,…). Za iskanje po spletu so nam na voljo spletni iskalniki kot je na primer Google. Slednji so danes že zelo izpopolnjeni in nam omogočajo, da na osnovi ključnih besed najdemo najbolj relevantne zadetke, vključno s slikovnim gradivom in avdio-video posnetki. Zadetke potem ročno pregledamo in tako izluščimo podatke, ki nas zanimajo. Vseh zadetkov seveda ne moremo pregledati, saj je splet neobvladljivo velik, ročno pregledovanje pa zamudno. Zato se navadno zadovoljimo s pregledom nekaj najbolj relevantnih zadetkov.

Avtomatska ekstrakcija podatkov s spleta (ang. *web content mining*) bi pomenila velik korak naprej [1]. Sama ideja še zdaleč ni nova in če bi razvoj sledil viziji Tima Berners-Leeja, enega izmed pionirjev svetovnega spleta, bi danes imeli opravka s t.i. *semantičnim spletom*, kjer bi bili podatki strukturirano zapisani na način, razumljiv tudi računalnikom in drugim napravam. Z razmahom interneta je Timovo napredno razmišljanje zamenjala preprostost dodajanja spletnih vsebin, kar je botrovalo izjemni rasti spleta, sama ideja o semantičnem spletu, ki ga pogosto imenujemo tudi »splet prihodnosti«, pa se zdi bolj kot ne utopična. Vprašanje pa je, ali si lahko kljub razsežnosti spleta in nestrukturiranosti njegove vsebine kakorkoli pomagamo z računalnikom in avtomatsko obdelavo spletnih vsebin. V tem prispevku predstavljamo primer računalniške rešitve, ki izkorišča koncept avtomatske ekstrakcije podatkov s spleta in s pomočjo nekaterih naprednejših tehnologij omogoča iskanje povezav med poslovnimi subjekti in njihovimi lastniki oziroma zastopniki. Rešitev smo poimenovali Supervizor+.

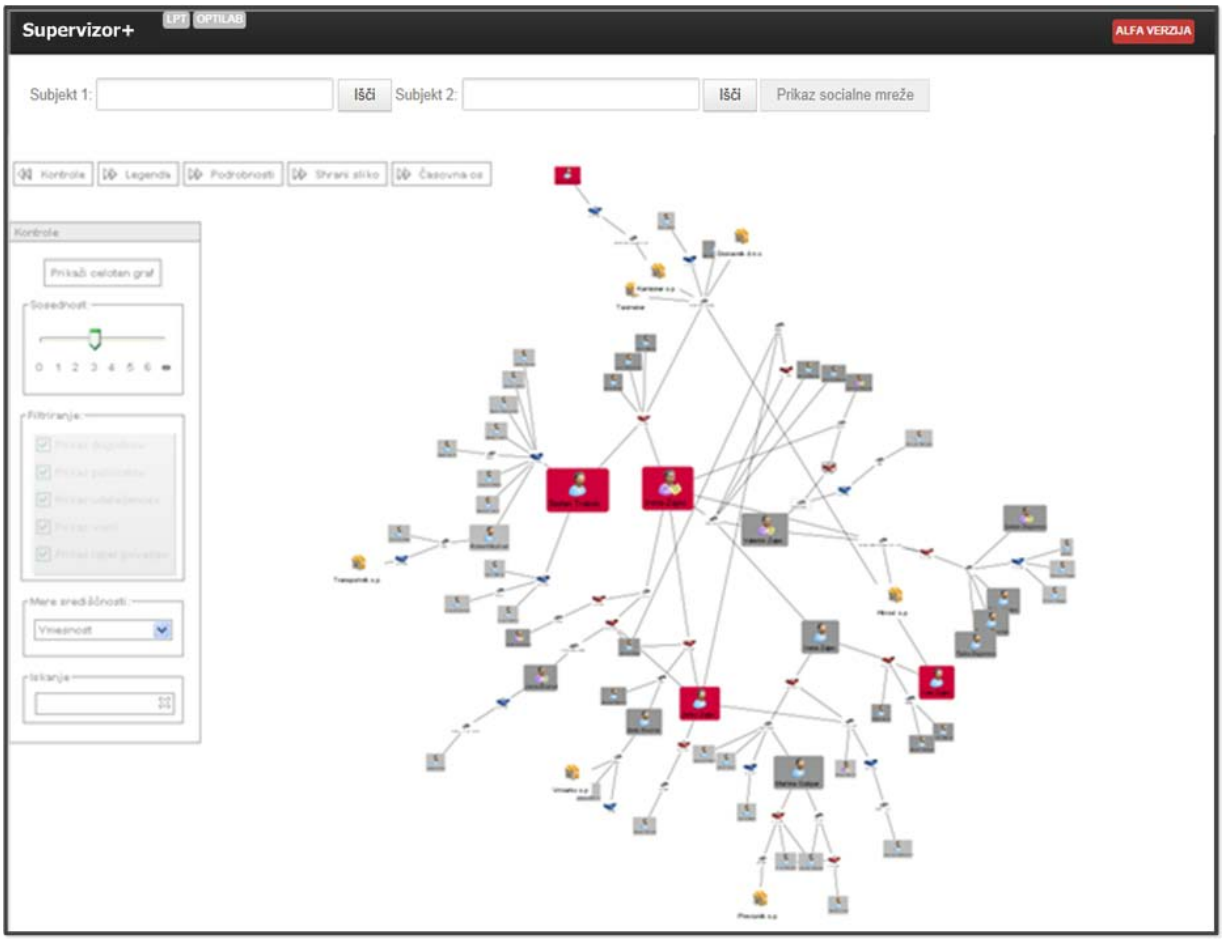
2. SUPERVIZOR+

Supervizor+ je spletna računalniška rešitev za iskanje povezav med poslovnimi subjekti in osebami v Sloveniji. Povezave so vizualizirane s pomočjo grafa, kjer vozliča predstavljajo pravne ali fizične osebe (pravni subjekti se delijo naprej na društva, stranke, gospodarske in negospodarske subjekte ipd.), povezave med vozlišči pa predstavljajo eno izmed naslednjih pomenskih odvisnosti:

1. Poslovne odvisnosti:
    a. subjekt je lastnik ali solastnik drugega subjekta,
    b. fizični subjekt je zaposlen pri poslovnem subjektu[1],
    c. gospodarski subjekt je prejemnik sredstev negospodarskega subjekta.
2. Politične odvisnosti
    a. fizični subjekt je pripadnik politične stranke,
    b. fizični subjekt je simpatizer politične stranke.
3. Družbene odvisnosti:
    a. fizični subjekt je sorodnik drugega fizičnega subjekta,
    b. fizični subjekt je prijatelj drugega fizičnega subjekta,
    c. subjekt je registriran na istem naslovu kot drugi subjekt.

S pomočjo povezav, ki so vzpostavljene med subjekti, lahko učinkovito pregledujemo različne vrste skupnosti (večje število tesno povezanih subjektov) ter druge neobičajne vzorce ([2], [3]). Sistem omogoča tudi iskanje povezav med poljubnima dvema subjektoma. Če takšna povezava ali več povezav obstaja, sistem vizualizira del grafa, ki vsebuje omenjene povezave. Primer ekrana Supervizorja+ prikazuje slika spodaj.

[1] Zanimiva spletna stran, ki ponuja podatke o zaposlitvah oseb, njihovih funkcijah, odgovornostih, pooblastilih itn. je **Jigsaw**, www.jigsaw.com. Katalog podatkov, ki nastaja se v okviru Jigsaw izjemno hitro raste.



Slika 1: Primer ekrana aplikacije Supervizor+

3. KAKO DELUJE SUPERVIZOR+

3.1 Spletni viri podatkov

Večina podatkov, ki jih Supervizor+ uporablja, je javno dostopnih prek spleta. Ključni viri so naslednji:

- Strukturirani poslovni viri (AJPES, UJP, ZZZS, Supervizor, …)
- Strukturirani spletni viri (Wikipedia ipd.)
- Socialna in profesionalna omrežja (Facebook, Google+, LinkedIn itd.)
- Spletni časopisi (Finance, Dnevnik, Delo itd.)
- Drugo (prvih nekaj zadetkov, pridobljenih z uporabo meta-iskalnika).

Zaradi učinkovitosti Supervizor+ podatkov ne bere neposredno s spleta, temveč se le-ti predhodno zajamejo v interno podatkovno shrambo. Podatki iz Supervizorja, AJPES-a in drugih poslovnih virov so zajeti s pomočjo spletnega robota, ki v izbranem viru za vsak poslovni subjekt (uporabljen je AJPES-ov šifrant poslovnih subjektov) poišče spletno stran, prebere in analizira podatke ter jih zapiše v podatkovno shrambo. Za pridobivanje podatkov iz socialnih in profesionalnih omrežij se uporabljajo vmesniki, ki jih posamezna omrežja

ponujajo. Poseben izziv predstavlja avtomatski zajem podatkov iz spletnih časopisov ter drugih spletnih virov. Slednje poteka v treh korakih: (1) iskanje zadetkov s pomočjo ključnih besed (pri spletnih časopisih se uporabljajo interni iskalniki, pri drugih spletnih virih pa meta-iskalnik, ki združuje zadetke več splošnih iskalnikov), (2) ekstrakcija teksta (slike, pasice ipd. se ignorirajo) ter (3) avtomatska ekstrakcija podatkov iz teksta.

3.2 Zainteresirana javnost kot vir podatkov ter ocenjevanje zaupanja

Poseben vir podatkov Supervizorju+ predstavlja zainteresirana javnost. Aplikacija je namreč zasnovana po vzoru skupinskega oz. družbenega razvoja vsebin (ang. *Collaborative content development*) in poljubnemu registriranemu uporabniku omogoča dvosmerno interakcijo; torej ne le, da po podatkih Supervizorja+ brska, temveč tudi, da v shrambo prispeva dodatne podatke, ki jih sam pozna. Supervizor+ od uporabnika zahteva, da vsak dodaten podatek opredeli z virom, dodanim podatkom pa dodeli poseben status. Za tako dodane podatke in povezave lahko drugi uporabniki glasujejo in tako vplivajo na stopnjo zaupanja posameznega podatka. Stopnja zaupanja je dodeljena tudi posameznemu viru ter se modelira v odvisnosti od vira samega in načina ekstrakcije podatkov. Pri iskanju po podatkih Supervizorja+ lahko opredelimo najnižjo stopnjo zaupanja virov in podatkov, ki naj jih Supervizor+ še upošteva pri iskanju povezav, ter na ta način vključujemo ali izključujemo posamezne vire podatkov.

3.3 Iskanje povezav in posebnih vzorcev povezav

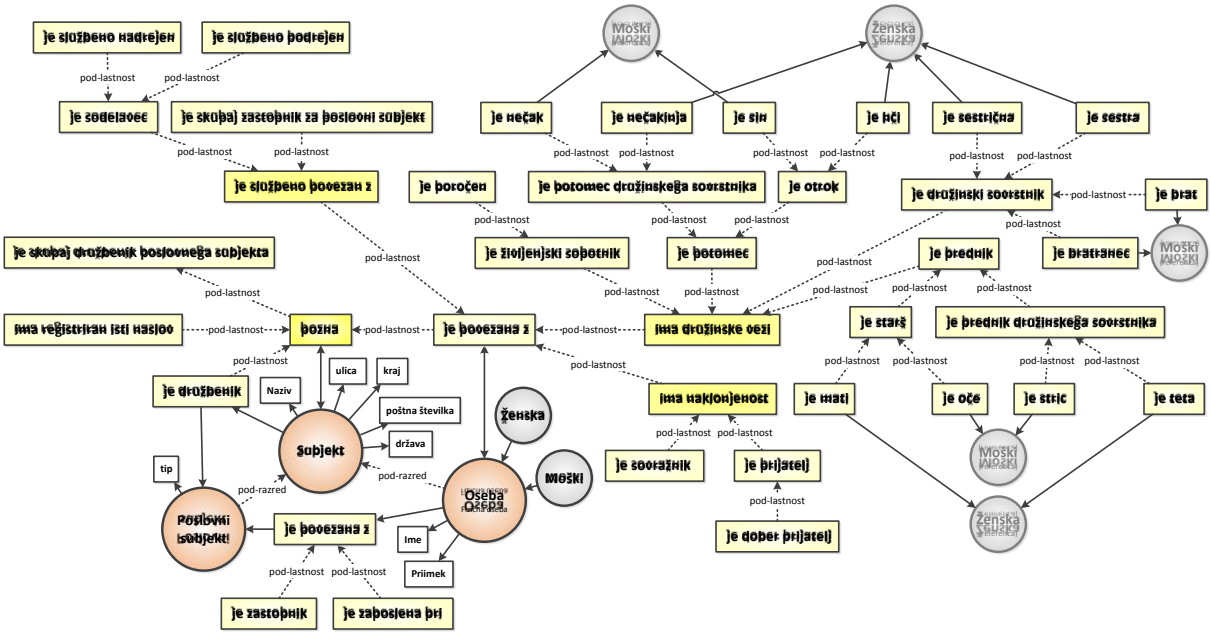
Supervizor+ omogoča učinkovito iskanje po zajetih podatkih. Podatke lahko vizualiziramo, prikazujemo različne podrobnosti, skrivamo posamezne elemente, iščemo povezave med subjekti itn. Poleg naštetega omogoča odkrivanje posebnih vzorcev, ki jih v naprej opredelimo, ter identifikacijo skupnosti, to je množice subjektov, ki so nenavadno tesno povezani med seboj. Zaradi procesorske zahtevnosti iskanje komun in vzorcev ne poteka v realnem času, temveč v ozadju.

4. UPORABLJENE TEHNOLOGIJE

4.1 Zapis podatkov in poizvedovanje

Supervizor+ je realiziran s pomočjo odprtokodnega ogrodja Jena [4]. Jena predstavlja razvojno okolje za razvoj semantičnih aplikacij s pomočjo sintaks in jezikov RDF, RDFS, OWL in vključuje podatkovno shrambo TDB, poizvedovalni jezik SPARQL ter mehanizem za sklepanje. Pri obsežnih ontologijah, kjer je večina podatkov v obliki primerkov (ABox komponenta), se zaradi performančnih omejitev mehanizem sklepanja v praksi ne uporablja. V primeru Supervizor+ se mehanizem sklepanja uporablja zgolj na shematski ravni (TBox komponenta), medtem ko izpeljane trojčke primerkov persistenčno shranjujemo neposredno v podatkovni shrambi. Ontologija je posledično izjemno velika in vključuj več kot 5 milijonov trojčkov.

Vsi podatki, ki jih Supervizor+ pridobi s spleta in nadalje uporablja, so shranjeni v RDF obliki [5] (ang. *Resource Description Framework*), ki jo predpisuje W3C (ang. World-Wide-Web Consortium) za konceptualno opisovanje in modeliranje spletnih virov. Ontologija, ki predstavlja model podatkovne shrambe, je prikazana na sliki spodaj.



Slika 2: Ontologija Supervizorja+

Za poizvedovanje po podatkih, ki so zapisni v shrambi, Supervizor+ uporablja poizvedovalni jezik SPARQL. Poizvedovanje je realizirano s strežniško komponento Joseki, ki prejema ukaze s pomočjo metode REST, izvede poizvedbe ter vrne rezultate posebni komponenti za vizualizacijo podatkov [4].

4.2 Vizualizacija podatkov

Za potrebe vizualizacije podatkov in povezav Supervizor+ uporablja spletno komponento WebNet podjetja Optilab d.o.o., ki je sodelovalo pri razvoju aplikacije. Komponenta WebNet podatke vizualizira v obliki dinamičnega grafa, ki ga uporabnik lahko poljubno dodatno prilagaja. Sami podatki se v komponento prenašajo kot vhodni parameter v sintaksi GraphML [6], ki temelji na XML jeziku. Sintaksa je obsežna, vendar enostavna za uporabo. Komponenta WebNet je vpeta v spletno aplikacijo in kot taka omogoča:

- Vizualno manipuliranje z grafom;
- Prikaz podrobnosti o posameznem vozlišču, povezavi;
- Prikaz/skrivanje (filtriranje) vozlišč ali povezav določene vrste;
- Opazovanje nastajanja grafa skozi čas s pomočjo drsnika;
- Izbiro največje razdalje med vozlišči, ki se še prikazujejo;
- Prikaz odkritih posebnih vzorcev v omrežju;
- Prikaz odkritih komun v omrežju.

5. ZAKLJUČEK

V prispevku je podana kratka predstavitev spletne aplikacije Supervizor+, način njenega delovanja ter uporabljenih tehnologij. Supervizor+ predstavlja praktičen primer uporabe javno

dostopnih podatkov na spletu ter demonstrira trenutne zmožnosti na področju avtomatske ekstrakcije podatkov.

6. VIRI IN LITERATURA

[1] Bing Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications), Springer, 2011

[2] Lovro Šubelj & Marko Bajec, Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction, Physical Review E, 83(3):036103, 2011

[3] Lovro Šubelj & Marko Bajec, Robust network community detection using balanced propagation, European Physical Journal B, 81(3):353-362, 2011

[4] http://jena.sourceforge.net/documentation.html

[5] http://www.w3.org/RDF/

[6] http://graphml.graphdrawing.org/