Računalno otkrivanje mreže povezanih pojmova

Computerized Discovery of Network of Interrelated Terms

Mentor: Prof. dr. sc. Branko Jeren

Zadatak

U sklopu diplomskog zadatka potrebno je istražiti stanje znanosti i tehnike u području automatiziranog prikazivanja znanja pomoću generirane mreže povezanih pojmova važnih za neko područje. Naglasak staviti na jednostavnost upotrebe za korisnike s minimalnim informatičkim vještinama te na dobavljivost i troškove korištenja alata. Potrebno je prilagoditi jedan takav alat i istražiti njegovu upotrebljivost za učenje i istraživanje u nekom tehničkom području.

U vezi dodatnih informacija obratiti se predmetnom nastavniku.

Kronologija

1. tjedan (5.-11.3.2012.)

Izvještaj sa prezentacije Maltego alata

- U alatu Maltego moguće je stvarati entitete i pisati transformacije; entiteti se sastoje od imena i svojstava

- Moguće je grafički prikazati veze između entiteta

- Transformacije su programi pisani u proizvoljnom programskom jeziku (najčešće Perlu ili Pythonu) koji za ulaz/izlaz imaju određen oblik xml datoteke; transformacije primaju entitet, a vraćaju entitete ili mrežu entiteta

- Svaku transformaciju korisnik mora sam pokrenuti

Ideja: Maltego bi se mogao iskoristiti kao alat za učenje; unošenjem jednog pojma moglo bi biti moguće pronaći udruge, knjige, časopise, predmete, sveučilišta, autore koji se bave tim pojmom ili ga sadrže te ih međusobno povezati. U tu svrhu trebalo bi napisati transformacije kao što su npr. transformacija koja pretražuje akademsku bazu podataka i baze knjiga (npr. Amazon).

- Spajanje na baze podataka je jednostavno isprogramirati u PHP-u, još jednostavnije u Pythonu

- Maltego već ima dostupne pluginove za Facebook, Twitter i LinkedIn te transformacije kao što je npr. traženje povezanih pojmova (To related phrases)

Plan:

- Instalirati Maltego Community Edition, proučiti dokumentaciju

- Potražiti postojeće transformacije na zadanu temu na Webu...

- Pretražiti postojeće alate/projekte slične Maltego-u na Webu (npr. “clustering search engines”, “bussiness intelligence”, RapidMiner, RapidNet, RapidAnalytics...)

Izvještaj

Maltego

http://www.paterva.com/web5/

Maltego je aplikacija otvorenog koda za forenzičke radnje koja pruža sučelje za prikupljanje informacija za prikaz u jednostavno razumljivom obliku. Zahvaljujući grafičkim bibliotekama, aplikacija omogućuje identificiranje ključnih veza između informacija i identificiranje nepoznatih veza među njima. Aplikaciju je jednostavno instalirati, a dostupna je za operacijske sustave Microsoft Windows, Mac i Linux.

Alat Maltego koristi klijent/poslužitelj arhitekturu u svrhu prikupljanja podataka za određivanje odnosa u stvarnom svijetu i veze među podacima kao što su:

- Osobe
- Grupe ljudi (socijalne mreže)
- Tvrtke
- Organizacije
- Web stranice
- Internet infrastruktura (poput domena, DNS imena, IP adresa i sl)
- Izraze
- Članstva
- Dokumenti i spisi

Što je TDS? Paterva TDS (Transform Distribution Server) je community-based poslužitelj čija svrha je dijeljenje, upravljanje i implementacija lokalnih i prilagođenih transformacija diljem Interneta.

Video koji objašnjava TDS:
http://www.youtube.com/watch?v=dl5-53AgCJ4&feature=player_embedded.

TDS link: https://cetas.paterva.com/TDS/

Pisanje transformacija: http://www.youtube.com/watch?v=XR6Sxe3wlDE&feature=player_embedded

Altervative alatu Maltego:

• **NetGlub**

http://www.netglub.org/ Još je u beta verziji, iznimno sličan Maltego-u samo što nema sve njegove mogućnosti.

• **Sentinel Visualizer**

http://www.fmsasg.com/Products/SentinelVisualizer/

http://www.fmsasg.com/News/PressRoom/Visualizer%20Overview.pdf Nije freeware, potrebno ga je naručiti.

• **RapidNet**

- Program otvorenog koda koji omogućava vizualizaciju relacija i struktura podataka vezanih uz određenu tvrtku

- Program je interaktivno rješenje za istraživanje strukture unutar entiteta, daje uvid u kompleksnu strukturu tvrtke i relacija između:

- osoba (korisnika)
- isporuka i destinacija
- komponenata poslovnih procesa

- Programski alat koji omogućava:

- vizualizaciju proizvoljnih KPI-eva na čvorovima
- prikaz geografskih informacija i regija definiranih korisnicima na kartama

Ostali alati:

Cluuz search

http://www.cluuz.com

Cluuz pretraga, primjer: “firewall”

- Web tražilica koja omogućava pretraživanje pojma, ispis top rezultata, prikaz mreža pojmova unutar istraživanja, ispis najvažnijih Web stranica vezanih uz pojam te entiteta vezanih uz svaku od Web stranica

- Pretraživanje je moguće prilagoditi tj. odrediti koji se od sljedećih pojmova trebaju pretražiti:

- osobe
- tvrtke/organizacije
- telefonski brojevi
- e-mail adrese
- adrese
- domene
- datumi
- geografija
- ostalo

“Cluuz radi neki posao za vas. Standardna tražilica kao rezultat prikazuje popis linkova. Cluuz pretražuje Web stranice, izdvaja bitne pojmove i slike, oblikuje ih u skupine i prikazuje u obliku grafa povezanih pojmova (semantičkog grafa) gdje možete kliknuti na bilo koji entitet te dodatno usredotočiti pretraživanje.”

- Nisu pronađeni dostupni API-ji za Cluuz search

Yippy search

http://twotrees.search.yippy.com/

- Rezultate pronađenih Web stranica slaže u mape po temama

Collarity search

http://www.collarity.com/about-us.html

- Prikazuje sljedeće skupine rezultata: Web, Video, Twitter, Images, News, Related searches, Blog

iSeek

http://www.iseek.com/iseek/home.page

- Tražilica nove generacije - postoje dva područja pretraživanja, Web i Education

- Moguće je postavljati upite na engleskom jeziku

- Ne stvara grafove, ali pronalazi povezane riječi, Web stranice, logo-e poznatih udruga i sveučilišta koja se bave traženom temom te prikazuje sljedeće skupine rezultata: Source, Topics, Subject, Resource Type, Grade Level, Places, People, Organizations, Standards by State

Microsoft Academic Search

http://academic.research.microsoft.com/

- Osim pretraživanja i ispisa rezultata po temama nudi sljedeće mogućnosti vizualizacije:

- Academic Map – prikazuje geografski organizacije i autore vezane uz određenu domenu
- CFP Calendar – pretražuje konferencije po domenu, vremenu i geografskoj lokaciji
- Domain Trend – vizualizacija istraživačkih trendova u području računalne znanosti
- Organization Comparison – uspoređivanje dvaju organizacija
- Co-author Graph – prikaz relacija između traženog autora i povezanih istraživača

- Co-author Path – prikaz relacija između dva istraživača preko zajedničkog koautora
- Genealogy Graph - ne radi
- Paper Citation Graph - nije jasno kako koristiti opciju ili ne radi

- Ima detaljno opisane API-je

TouchGraph

http://www.touchgraph.com/navigator

- Primjer rezultata u TouchGraph-u:

“Touchgraph je alat za vizualizaciju (pisan u Javi) za istraživanje veza između povezanih Web stranica. Unesite pojam ili naziv Web stranice da biste vidjeli mapu, a možda i nove Web stranice koje bi vas mogle zanimati. Interaktivna Facebook verzija preglednika pokazuje veze između Facebook prijatelja, fotografija i lokacija.”

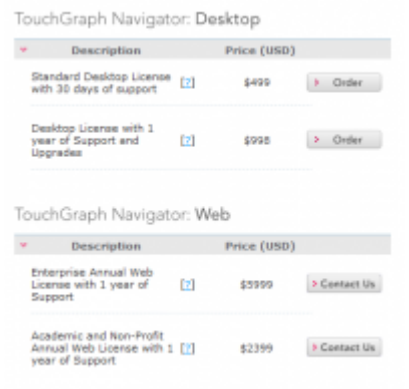
- Demo je dostupan na: http://www.touchgraph.com/seo

- Program je za komercijalnu upotrebu, moguće je preuzeti trial verziju koja ističe za 30 dana

- Touch Graph Navigator 2: http://www.touchgraph.com/assets/navigator/help2/application.html

- Potencijalno zanimljivo, ovo je potrebno još istražiti

- Cijene TouchGrapf Navigator alata za desktop i web platforme:



AskKen search

http://askken.heroku.com/

- Visual Knowledge Browser – Omogućava pretraživanje pojmova i vizualizaciju vezanih širih pojmova

- Upisivanjem pojma nudi se odabir različitih značenja istog pojma (npr. Integral – matematički pojam, film, europska agencija, glazbeni sastav itd.)

Zaključak:

- Izrađena je mapa područja

- Istraženi su postojeći slični alati na Webu

- Instalirani su alati Maltego i RapidNet

- Nisu pronađene korisne transformacije vezane uz alat Maltego na Webu

Plan za daljnji rad:

- Potrebno je istražiti još alata te fokusirati pretragu

- Pokušati pronaći korisne transformacije za Maltego ako postoje i ako su dostupne

- Istražiti na koji način funkcionira RapidNet

2. tjedan (12.-18.3.2012.)

Izvještaj sa sastanka, ponedjeljak, 12.3.2012., 10:00

Prezentirani su alati pronađeni u prethodnom tjednu. Plan za idući tjedan je:

- Nastaviti istraživanje i pronaći još sličnih alata
- Klasificirati i sistematizirati pronađene alate i navesti njihova svojstva
- Postupno dodavati pojmove, alate, organizacije, sveučilišta itd. u mapu područja

Izvještaj

Instalirani su alati RapidNet i RapidMiner:

RapidMiner

http://rapid-i.com/content/view/181/190/

- Open source aplikacija za rudarenje podataka i poslovnu inteligenciju

- Pisan je u Javi te podržava Windows, Linux i Mac OS X operativne sustave

- Programski alat je sposoban za OLAP obradu podataka iz opsežnih skladišta podataka

RapidMiner sadrži preko pet stotina modula za transformiranje, ekstrahiranje i analizu podataka; zatim statističko prognoziranje, vizualizaciju, OLAP obradu, module vezane uz poslovnu inteligenciju i brojne druge. Također podržava skripte i plug-inove te razmjenu obrađenih podataka putem XML standarda. Postoje i komercijalne inačice RapidMinera s još više značajki, dok je open source inačica došla do svoje pete iteracije. Sučelje RapidMinera veoma je pregledno i funkcionalno.

- Alat je kompliciran i iako ima pregledno vizualno sučelje, pogodan je za data-mining te nešto kompliciraniju analizu podataka

- Pretraženi su plug-inovi i nije pronađen nijedan koristan plug-in

RapidNet - pogodan prvenstveno za simulaciju mrežnih protokola, najvjerojatnije ne može biti iskorišten u svrhu diplomskog rada

Zatim, pronađene zanimljive Web tražilice:

Mnemomap

http://www.mnemo.org/

- Web tražilica koja je još u početnom stadiju razvoja - Njezina funkcionalnost se dijeli na 3 dijela:

- Atomic-Tree
- Query List
- Tabs

Atomic tree:

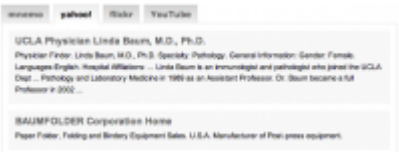
Ispisuje sinonime, tagove i prijevode. Svaku od pronađenih riječi moguće je postaviti kao centar atomskog stabla ili dodati kao ključnu riječ u pretraživanje.

Query list:

Popis upita, moguće je postaviti status svake riječi na “aktivno” ili “pasivno”, gdje “aktivno” označava da se uključuje u pretraživanje, a “pasivno” da ne pripada.

Tabs:

Svaki tab predstavlja svoj izvor rezultata pretraživanja.



Spezify

http://spezify.com/

- Web tražilica koja prikazuje web stranice kao slike

TouchGraph SEO Browser

http://www.touchgraph.com/seo

- Web tražilica napisana u Javi

- Uključuje pretragu te ispis povezanih pojmova, web stranica i top domena

- Ispisuje top domene u obliku interaktivnog grafa

- Pojedinu domenu je moguće postaviti kao korijen grafa ili sakriti

WikiMindMap

http://www.wikimindmap.org/

- WikiMindMap je alat za vizualno pretraživanje sadržaja Wikipedije - Ispisuje rezultate u obliku kognitivne mape - Mapu je moguće preuzeti u FreeMind formatu

VisionLink Search

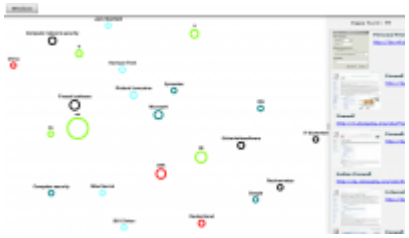
http://www.pikko-software.com/information-visualization-tool.html

- Web tražilica te alat za analizu i mapiranje informacija

- Rezultat pretrage je vizualni i intuitivni prikaz rezultata pretrage ili text-mininga prikazan pomoću interaktivnih relacijskih grafova

- Sve karte koje generira VisionLink temelje se na Flash tehnologiji te se mogu prikazati interaktivno unutar web preglednika

- Primjer pretrage za “firewall”:



Rezultat pretrage su kružići i na desnoj strani Web stranice prikazane u obliku slika. Prelaskom miša preko kružića iscrtavaju se veze sa ostalim kružićima. Različite vrste rezultata su određene bojom, a veličina označava “broj” pronađenih rezultata. Npr. domene su označene svijetlozelenom bojom, zemlje, crvenom, osobe svijetloplavom, organizacije/udruge/tvrtke tamnoplavom bojom, a pojmovi crnom bojom.

Demo: http://demo.pikko-software.com/exalead/

Thinkpedia

http://thinkpedia.cs.auckland.ac.nz/

- Tražilica za vizualno pretraživanje Wikipedije - Rezultat je graf povezanih pojmova, grane sa pojmovima su grupirane po zemljama, tvrtkama, društvenim oznakama, organizacijama, pozicijama, osobama itd.

Thinkbase

http://thinkbase.cs.auckland.ac.nz/

- Tražilica slična Thinkpediji, još je u razvoju, omogućava pretragu po temama (na engleskom jeziku)

Silobreaker

http://www.silobreaker.com/

- Web tražilica koja sadrži opciju Network za ispis rezultata u obliku mreže povezanih pojmova, obuhvaća tvrtke, organizacije, ljude, gradove, ključne riječi i proizvode - Primjer pretraživanja osobe “Stieg Larsson” je na slici:

Nadalje, pronađeni programski paketi/alati:

Bibex - Bibliographic Exploration Tool

http://www.dama.upc.edu/technology-transfer/bibex

http://www.youtube.com/watch?v=OKLN8_NA3y8&feature=player_embedded

- BIBEX je inovativni program za istraživanje bibliografskih repozitorija i dokumenata7

- Omogućava brze odgovore na sofisticirane upite nad velikom količinom podataka

- Proizvod pet godina istraživanja i razvoja na UPC-u (Politehničkom Sveučilištu u Kataloniji)

- Temelji se na upitima postavljenim nad DEX bazama podataka

- DEX je sustav baza podataka temeljen na grafovima napisanim u Javi i C++ - http://sparsity-technologies.com/dex - Osnovne karakteristike DEX sustava su učinkovito pretraživanje

i analiza velikog broja podataka uz male zahtjeve nad skladišnim prostorom za spremanje podataka. Kompatibilan je sa Windows, MAC i Linux operativnim sustavima.

- BIBEX omogućuje izvođenje složenih bibliografskih upita i prikaz rezultata svojih upita kao kombinaciju grafova i tekstova

- Alat omogućava istraživanje odnosa između autora, ključnih riječi i radova na takav način da korisnik može obavljati složene bibliografske pretrage te tim načinom istražiti tko je tko u pojedinim područjima istraživanja

- Demo (Social network search) dostupan je u obliku za pretraživanja autora i osoba vezanih uz traženog autora:

HistCite

http://thomsonreuters.com/products_services/science/science_products/a-z/histcite/

http://interest.science.thomsonreuters.com/forms/HistCite

http://en.wikipedia.org/wiki/Histcite

- HistCite je programski paket za provođenje bibliometrijskih analiza i organizaciju rezultata pretraživanja baze Web of Science

- Razvio ga je Eugene Garfield, osnivač Instituta za znanstvene informacije i izumitelj važnih alata za pronalaženje informacija, kao što su Current Contents i Science Citation Index

- Uz pomoć programa moguće je otkriti predmetnu strukturu, povijest i međusobne veze objavljene literature

- Besplatna probna verzija programskog paketa dostupna je na Web stranicama

- HistCite se pokreće u Internet Explorer pregledniku i zahtijeva korisnički računa na Web of Knowledge stranici

- Isproban je na primjerku HistCite bibliografskog zapisa

VOSviewer

http://www.vosviewer.com/

- VOSviewer je besplatan alat prvenstveno namijenjen za analizu bibliometrijskih mreža

- Program se može primjerice koristiti za izradu mape publikacija, autora ili časopise na temelju mreže citiranja ili stvoriti karte ključnih riječi na temelju mreže pojava u pojedinim bibliometrijskim zapisima

- Ulazni podaci moraju biti u .txt formatu i podržava samo zapise na engleskom jeziku

- Primjer mape:

Novi pojam na koji sam naišla je Open-source intelligence (OSINT).

http://en.wikipedia.org/wiki/Open_source_intelligence

Open-source intelligence (OSINT) je oblik prikupljanja i upravljanja podacima koji uključuje pronalazak, odabir i pribavljanje informacija sa javno dostupnih izvora te analizu podataka u svrhu oblikovanja novih saznanja. Definicija „Open source intelligence“ se mijenjala tijekom vremena, što se tiče organizacija i država. Najjednostavnije rečeno ona je „neklasificirana informacija“. Još jedna upotrebljiva definicija je „informacija s potencijalom mogućeg značaja za obavještajne službe koja je dostupna javnosti“.

Alat koji je vezan uz taj pojam:

Cogito Intelligence Platform

http://www.osint.it/english/cogito-intelligence-platform-osint.asp

http://www.osint.it/pdf/Europe_Cogito_Intelligence_Platform.pdf

- Cogito Intelligence Platform je rješenje za Open-source Intelligence (OSINT)

- Programski paket koji primjenjuje semantičku analizu podataka za podršku u radu analitičara i osoba koje se bave područjem upravljanja znanjem u svim fazama ciklusa prikupljanja podataka te omogućuje otkrivanje uzoraka u informacijama i međusobnih veza s mogućnošću vizualizacije

- Razvila ga je udruga Expert System koja se bavi proizvodnjom programskih alata vezanih uz semantiku

- Moguće je zatražiti demo verziju alata: http://www.expertsystem.net/demo_prodotti.asp

Također, pri pretraživanju je pronađeno i nekoliko open-source alata za vizualizaciju podataka koji će možda biti korisni u daljnjem radu, kao što su: Axiis, Flare, Tulip, Impure, Prefuse...

Otkriveni su i Python moduli korisni za Maltego transformacije i web-mining:

- Pymaltego - http://code.google.com/p/pymaltego/ - Python programski okvir za Maltego transformacije
- Pattern - http://www.clips.ua.ac.be/pages/pattern - Modul za programski jezik Python koji se može koristiti u svrhu web-mininga. Predstavlja nakupinu alata za prikupljanje podataka sa stranica kao što su: Google, Twitter, Wikipedia putem njihovih API-ja, spominje se i HTML DOM parser. Sadrži dijelove za analizu teksta, clustering i klasifikaciju i vizualizaciju podataka (mreže grafova). Sadrži više od 30 primjera skripti i 350 testova.

Zaključak:

- Otkrivene su nove domene pretraživanja

- RapidMiner i RapidNet su previše složeni alati za otkrivanje pojmova između entiteta, barem u odnosu na Maltego, treba tražiti dalje

- Pronađeno je još alata i tražilica:

- Tražilice: Cluuz, Yippy, Collarity, iSeek, Microsoft Academic Search, AskKen, Mnemomap, Spezify, TouchGraph SEO Browser, WikiMindMap, VisionLink Search, Thinkpedia, Thinkbase, Silobreaker...
- Programski alati: Maltego, RapidMiner, RapidNet, NetGlub, Sentinel Visualizer, TouchGraph, BIBEX, HistCite, VOSviewer, Cogito Intelligence Platform...

- Tražilice je moguće podijeliti u one sa vizualnim rezultatima i ostale te klasificirati po svojstvima kao što su: izgled ispisa rezultata, domena pretraživanja, jezik pretraživanja

- Programske alate je moguće klasificirati po svojstvima kao što su: dostupnost, namjena, mogućnosti te podržane platforme

Plan za daljnji rad:

- Tražiti još mogućih alata...

- Detaljnije klasificirati pronađene alate

3. tjedan (19.-25.3.2012.)

Izvještaj sa sastanka, ponedjeljak, 19.3.2012., 12:00

Prezentiran je rezultat rada prošlog tjedna; komentirani su pronađeni alati.

Dogovoreni zadaci za idući tjedan:

- Istražiti postoji li još takvih alata, posebno potražiti alate slične Maltegu
- Sistematizirati do sada pronađene alate, svrstati ih i opisati njihova obilježja (npr. u tablici)

Izvještaj

Pronađeni alati:

Hakia

http://googlesystem.blogspot.com/2007/01/hakia-knowledge-search-engine.html

http://hakia.com/

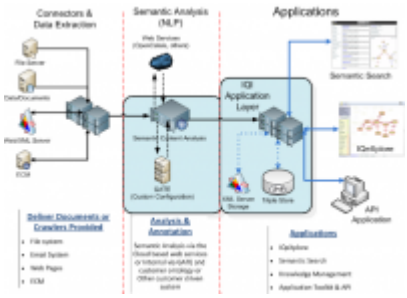
Web semantička tražilica koja ne ispisuje rezultate temeljeno samo na podudaranju traženih riječi na Web stranici već koristi mehanizme kojima pokušava oponašati ljudske kognitivne sposobnosti prepoznavanja onog što je bitno u Web stranicama. Također svrstava rezultate u grupe kao što su: službene stranice, biografije, slike, vijesti i intervjui, fan stranice, nagrade, govori, mitovi, polemike, resursi, inovacije, statistike, literatura i filmografija. Upiti se pišu na engleskom jeziku.

IQExplore

http://innovativequery.com/products-and-services/professional-services.html

Profesionalni programski paket. Nema mogućnost preuzimanja verzije za procjenu, potrebno ga je naručiti.

“Što se vanjskih izvora podataka tiče, možemo dodati mogućnost preuzimanja podataka s bilo kojeg tekstualnog izvora - Web stranica, rezultata pretraživanja, RSS feed-ova, blogova, PDF ili Word datoteka, sustava za upravljanje dokumentima, portala i plaćenih informacijskim uslugama.”



Primjer analize raznih izvora podataka; modeliranje veza između entiteta; opisan je jedan use-case scenarij korištenja IqeXplore-a: http://innovativequery.com/blog/4-augmented-governamce.html

Centrifuge Visual Analytics Network

http://www.centrifugesystems.com/index.php

Zatražena je verzija za procjenu alata te članci vezani uz projekt. (Evaluation Period: 3/22/2012 - 4/5/2012)

Visual Analytics Network se pokreće pokretanjem Web preglednika, povezivanjem s bilo kojim brojem izvora podataka i interaktivnom vizualizacijom rezultata u obliku grafova, tablica i geoprostornih pogleda. Izvori ulaznih podataka s kojima je moguće povezati alat su: Microsoft Excel, Microsoft Access, Microsoft SQL Server, Oracle Thin, PostgreSQL, Text, XML.

Dostupnost: po narudžbi. Verzija za procjenu može se zatražiti besplatno i traje 14 dana, proizvod je pod licencom i ova verzija se može koristiti samo za evaluaciju i demonstraciju proizvoda. Potrebno je imati Adobe Flash Player 10, podržani preglednici su Internet Explorer 7, Chrome 5, Mozilla Firefox 3.

MarketVisual

http://www.marketvisual.com/

Tražilica, pretraga obuhvaća profesionalne životopise, uključujući detaljne položaje, datume i prateću interaktivnu vizualizaciju mreže pojmova. Neograničen pristup svim mogućnostima ovog Web alata moguće je dobiti za $19,99 mjesečno.

XANALYS Link Explorer

http://www.xanalys.com/solutions/linkexplorer.html

http://www.xanalys.com/documents/XANALYSLinkExplorerWhitePaper.pdf

Programski paket je potrebno naručiti. Xanalys Link Explorer je dostupan za 30 dana procjene. Kao ulaz prima uređene podatke (npr. database dokumente, Microsoft Excel datoteke) te na osnovu tih podataka gradi grafikone i mreže povezanih pojmova.

Link XANALYS Explorer obrađuju podatke s raznih izvora; moguće je preuzeti podatke iz baza podataka kao što su ODBC baze podataka ili uvesti vlasite podatke u XANALYS Link Explorer internu bazu podataka. U programskom alatu svi se podaci interpretiraju kao skup objekata i poveznica. Korisnici mogu definirati način na koji se podaci prevode u objekte i poveznice.

Također, pronađen je i paket libraryja koji služi kao pomoć pri izgradnji mreže podataka na Webu, moglo bi biti korisno u daljnjem radu:

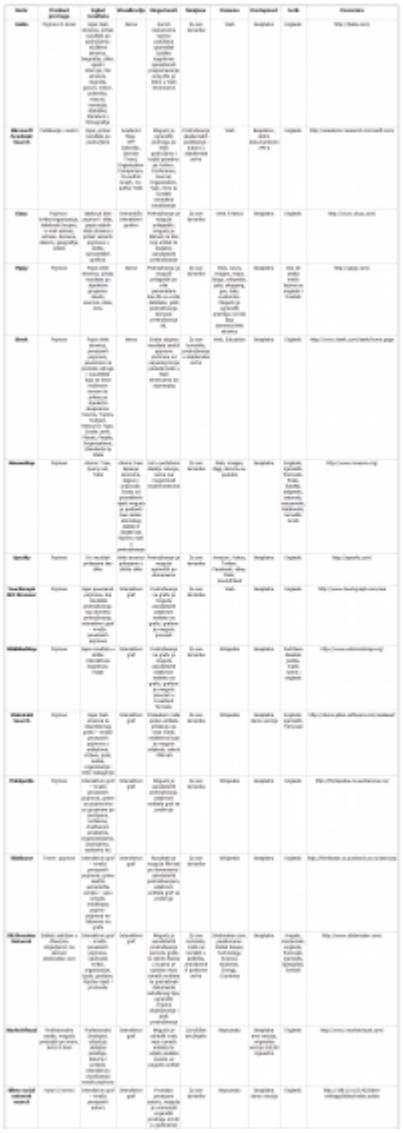
Cytoscape Web

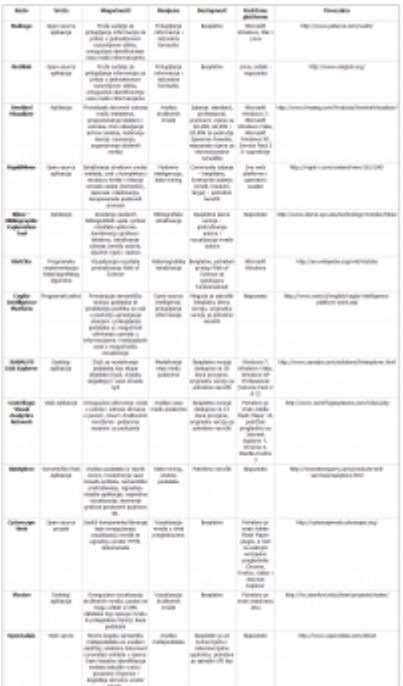
http://cytoscapeweb.cytoscape.org/

Cytoscape Web je skup knjižnica za vizualizaciju mreža koje je moguće ugraditi u Web stranicu. Cytoscape nije samostalan program već besplatan alat koji služi programerima pri prikazu mreža na Webu. Svojstva projekta:

- Predstavlja višekratne komponente koja omogućavaju ugradnju vizualizacije mreža unutar HTML dokumenata
- Omogućava jednostavno integriranje u HTML pomoću svojih Javascript API-ja
- Moguće je prilagoditi učitavanja i prikaz podataka

Tablice svih tražilica i programskih alata:





Napravljene u: HTML Table Factory - http://www.bdvnotepad.com/tablefactory_en.htm

Plan za daljnji rad:

- Konzultirati se s mentorom oko odabira alata/platformi za podrobnije istraživanje/analizu
- Poboljšati/proširiti tablicu pronađenih alata

4. tjedan (26.3.-1.4.2012.)

Izvještaj sa sastanka, ponedjeljak, 26.3.2012., 12:00

Zaključci:

- Alate je potrebno klasificirati
- Izlistati ključne riječi po kojima su se odvijale pretrage
- Upoznati se detaljnjije s alatom Maltego i pisanjem transformacija

Izvještaj

Još programskih alata...:

Vizster

http://hci.stanford.edu/jheer/projects/vizster/

Programski alat napisan u Javi za vizualizaciju društvenih mreža. Vizster ne sadrži podatke sa društvenih mreža; podaci se mogu učitati iz XML datoteke koji opisuje mrežu ili prilagođene MySQL baze podataka.

Alati korisni pri analizi metapodataka:

OpenCalais

http://www.opencalais.com/about

Web servis OpenCalais automatski stvara bogatu semantiku metapodataka za unešeni sadržaj. Korištenjem obrade prirodnog jezika (NLP), strojnog učenja i drugih metoda, analizira dokument i pronalazi entitete u njemu. Osim klasične identifikacije entiteta, Calais također vraća i povezane činjenice i događaje skrivene unutar teksta.

Besplatan je za komercijalnu i nekomercijalnu upotrebu; potrebno je zatražiti API key.

Document Viewer → http://viewer.opencalais.com/

Sličan alat: **Foca**

http://www.informatica64.com/foca.aspx

Program za analizu metapodataka u tekstu.

Online verzija alata → http://www.informatica64.com/foca/

Domene:

VisualComplexity

http://www.visualcomplexity.com/vc/ - Domena posvećena vizualizaciji složenih mreža

SEO Tools - Search optimization tools

http://tools.seobook.com/ - Domena posvećena programskim alatima za optimizaciju online pretraživanja

Također sam naletjela na sljedeći članak:

Social Networking Special Ops: Extending data visualization tools for faster Pwnage

http://www.defcon.org/images/defcon-18/dc-18-presentations/Suggmeister/DEFCON-18-Suggmeister-Social-Net-Special-Ops-WP.pdf

Ovaj članak opisuje kako alati za vizualizaciju podataka mogu biti prilagođeni analizi društvenih mreža. Mogućnosti takvih alata uglavnom obuhvaćaju data-mining, prepoznavanje entiteta i vizualizacije. Dvije studije slučaja opisuju te tehnike u kontekstu društvenog umrežavanja.

*Tablica iz članka koja opisuje neke pronađene alate:

	Visualization	Interactive Visualization	NER
Processing	Y	Y	N
Graphviz	Y	Y	N
OpenCalais	N	N	Y
Maltego	Y	Y	Y
touchgraph	Y	Y	N
mindraider	Y	Y	N
Vizster	Y	N	N

NER = Name Entity Recognition

Popis važnijih ključnih riječi korištenih u pretraživanju:

Top queries:

Data visualization, Exploratory analysis and visualization of network data, Social network analysis, Link analysis and visualization, Intelligence and forensics tool, Multigraphing, Clustering web search results, Visual search engines, Visual browsing, Entity-relation modeling, Semantic search engines, Information visualization, Information/intelligence gathering, Business intelligence tools, Web crawling, Innovative search engines, Knowledge discovery, Discovering and clustering related terms, Academic search engines, Bibliographic/bibliometric exploration, Interactive visualization, Data

mining, Contextual network graphing, Semantic association identification, Linked/Related/Interrelated data, Information mapping, Data relationships, Open-source intelligence, Network graphing, Graph visualization, Knowledge management, Exploration, Browsing, Search optimization

Various combinations with:

Tool, Software, Application, Engine, Framework

Also:

Alternatives, Similar tools

Grafički prikaz alata:

(Pri kraju dokumenta)

Napravljeno u: Gliffy - http://www.gliffy.com/

Zaključak:

- Pronađeno je još alata od kojih su neki vezani uz analizu metapodataka, smatram da nisu potencijalno korisni za ovaj diplomski rad, ali sam ipak opisala jedan primjer
- Pronađeni alati su osvježeni u tablicama
- Svi alati, tražilice, skupovi knjižnica itd. su prikazani u obliku grafa
- Izlistan je popis ključnih riječi koje su najviše korištene pri pretraživanju
- Nove javno dostupne transformacije su dodane u Maltego sa TDS-a
- Pronađeni su neki primjeri open-source transformacije na Webu
- Počela sam s istraživanjem transformacija i njihovih specifikacija opisanih na http://www.paterva.com/web5/

Plan za daljnji rad:

- Dodati pronađene članke u mapu područja
- Nastaviti rad na Maltego transformacijama

Mapa područja

Napravljena u: FreeMind - http://freemind.sourceforge.net/wiki/index.php/Main_Page

5. tjedan (2.-8.4.2012.)

Izvještaj

Pisanje transformacija u Maltegu

Transformacije je moguće pisati lokalno (skripta se može nalaziti bilo gdje na računalu) ili ih dodati u vlastiti TransformSeed na Paterva TDS serveru (https://cetas.paterva.com/TDS/). Da bi to bilo moguće, transformacija se mora nalaziti na Web serveru u *cgi-bin* folderu. Dodavanje transformacije u vlastiti seed omogućava lakše otkrivanje transformacije iz Maltego GUI-a bez prethodne konfiguracije alata, međutim, nije nužno za testiranje same transformacije.

Pripisanju lokalne transformacije korišteni su sljedeći linkovi:

- Specifikacije lokalnih transformacija: http://www.paterva.com/web5/documentation/localTransforms-SpecIII.pdf
- Integracija s Maltegom: http://www.paterva.com/web5/server/Integration%20options%20with%20Maltego%20v3.pdf
- Paterva forum na kojem je moguće pronaći Maltego libraryje za jezike Ruby, Python i PHP: http://www.paterva.com/forum/

Instalirano je izdanje Pythona je 2.7.2 (http://www.python.org/getit/), razvojno okruženje Eclipse (http://www.eclipse.org/) sa pluginom za Python PyDev (http://pydev.org/) te pomoćni libraryji.

Napisana je jednostavna lokalna transformacija *ToGoogleURLs.py* koja kao ulazni parametar prima *Phrase* entitet, a vraća listu *URL* entiteta. Komunikacija se odvija tako da Maltego ispisuje entitete na standardni izlaz, a prima entitete putem standardnog ulaza u XML formatu. Transformacija u primjeru obavlja pretragu putem Google-a i vraća 10 prvih rezultata. Broj rezultata je moguće promijeniti u kodu (op. community edition alata Maltego dopušta ispis najviše 12 entiteta). Prikaz izvođenja:

Pri pisanju transformacije korišteni su:

- Python Maltego library (*MaltegoTransform*): http://www.paterva.com/forum//index.php/topic,94.0.html
- Google search library za Python (*google*): http://breakingcode.wordpress.com/2010/06/29/google-search-python/

Zaključci:

- Dio pisanja transformacije vezan uz komunikaciju s Maltegom je uspješno obavljen
- Python prog. jezik je odabran zbog dobro napisanih modula za komunikaciju s Maltegom
- Također, najpogodniji je za komunikaciju s raznim tražilicama koje će biti potrebno koristiti u daljnjem radu

Plan za daljnji rad:

- Započeti rad na implementaciji nešto složenijih pretraga
- Pronaći API-je za neke druge tražilice osim Google-a, npr. Amazon
- Istražiti na koji način je moguće postaviti skriptu na Web tako da može biti dodana u vlastiti TransformSeed

6. tjedan (9.-15.4.2012.)

Izvještaj

Počela sam sa radom na transformaciji koja pristupa Amazonu. Amazon ima novi API od studenog 2011. godine te stoga stariji Python moduli ne funkcioniraju: http://pypi.python.org/pypi/python-amazon-product-api/0.2.5, http://aws.amazon.com/code/Python/134 itd.

Jedini Python modul koji radi sa novim API-jem je *Python Simple Product Amazon API* - http://pypi.python.org/pypi/python-amazon-simple-product-api/1.0.0 i može se preuzeti na sljedećoj stranici: https://github.com/yoavaviram/python-amazon-simple-product-api

Koristeći taj modul napisala sam transformaciju *ToAmazonBooks.py* koja pretražuje Amazon i vraća naslove knjiga. Transformacija prima entitet *Phrase* kao ulaz, a vraća listu entiteta *BookTitle* (koji ne postoji u Maltegu već sam ga sama dodala). Za pristup Amazonu bilo je potrebno registrirati se na stranici http://aws.amazon.com/s3/ i kao Amazon Associate te pribaviti sljedeća 3 koda za pristup:

- Access Key ID: AKIAJNVI5O2JEA4PPY3A
- Secret Access Key: 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
- Unique Associates ID: evey0c-20

Ti kodovi se nalaze u *config.py* modulu i koriste se u transformaciji. Primjer izvođenja transformacije vidi se na slici:

Zaključci:

- Osim samog naslova knjige, pretraga vraća i sljedeća svojstva knjige kao što su: cijenu i valutu, EAN - The European Article Number, URL slike, izdavača i dimenzije knjige
- Python Simple Product Amazon API može obaviti pretragu samo po nazivu i ASIN-u knjige. Nigdje se ne spominje autor kao svojstvo knjige zbog čega bi trebalo istražiti slične module napisane u drugim programskim jezicima (npr. Perlu ili PHP-u)

Plan za daljnji rad:

- Omogućiti preuzimanje dodatnih svojstava knjige: entitet *BookTitle* bi mogao sadržavati dodatna polja kao što su *EAN* i *Publisher*
- Istražiti na koji način je moguće preuzeti i autora knjige kao rezultat pretrage

7. tjedan (16.-22.4.2012.)

Izvještaj sa sastanka, ponedjeljak, 16.4.2012., 09:30

Novi planovi...

- Staviti opise i objašnjenja pojmova naznačenih u grafu pronađenih alata
- Pronaći neke Web alate za pretragu akademskih pojmova kao što su: članci, konferencije, sveučilišta
- Istražiti njihovu dostupnost; na koje načine im se može pristupiti, putem kojih API-ja ili protokola
- Nastaviti s pisanjem transformacija koje prikupljaju podatke s Amazona

Izvještaj

Grafički prikaz alata:

Objašnjenja nekih pojmova u grafu...

- Tools – Programski alati
- Libraries – Programske biblioteke
- Applications – Primijenjena programska potpora / programi
- Search Engines – Web tražilice

Business Intelligence

Poslovna inteligencija (eng. Business Intelligence, BI) je skup metodologija i koncepata za prikupljanje, analizu i distribuciju informacija uz pomoć različitih programskih alata. Poslovna inteligencija je jedna od tehnika poslovnog izvještavanja, koja omogućuje pronalaženje informacija potrebnih za lakše i točnije donošenje poslovnih odluka. Neke od metoda poslovne inteligencije uključuju rudarenje podataka (Data Mining), skladištenje podataka (Data Warehousing) i OLAP mrežnu analitičku obradu podataka. OLAP (Online Analytical processing) je vrsta obrade podataka koja daje brzi odgovor na višedimenzijske upite. Rudarenje podataka (eng. Data mining) je sortiranje, organiziranje ili grupiranje velikog broja podataka i izvlačenje relevantnih informacija. Skladištenje podataka (eng. Data warehousing) je metoda kojom se analizira i obrađuje velika količina podataka za potporu odlučivanju i upravljanju u poduzeću.

OSINT

„Open-source intelligence“ označava prikupljanje „neklasificiranih informacija“, odnosno, informacija s potencijalom mogućeg značaja za obavještajne službe koja je dostupna javnosti.

Bibliographic Exploration

Bibliografsko istraživanje: pronalaženje i pretraživanje autora, publikacija, knjiga u bibliografskim izvorima u kojima se mogu naći informacije o pojedinim zapisima.

Metadata Analysis

Analiza metapodataka, označava i prikupljanje metapodataka iz digitalnih datoteka (npr. tekstualnih). Metapodatci su podatci koji opisuju karakteristike nekog izvora u digitalnom obliku. Korisni su kod pregledavanja, prijenosa i dokumentiranja informacijskog sadržaja. U digitalnom smislu to su strukturirani podatci koji opisuju, objašnjavaju, lociraju ili na neki drugi način omogućavaju lakše upravljanje resursima.

Simple Data Visualization

Grafički prikaz strukturiranih skupova podataka i veza među podacima pohranjenim u određenom datotečnom obliku (bazi podataka, XML zapisu, Excell datoteci). Ne omogućuje naknadno otkrivanje veza među podacima.

Semantic Technology

Označava uporabu semantičkih tehnologija pri izvođenju programa ili pretraživanju. Semantika se kao grana lingvistike bavi proučavanjem značenja jezičnih znakova, dok s druge strane informacijske znanosti nisu usredotočene na značenje informacija. Međutim područja umjetne inteligencije (engl. artificial intelligence) kao što su obradba prirodnog jezika (engl. natural language processing), strojno učenje (engl. machine learning) i strojno prevođenje (engl. machine translation) zanima obradba prirodnog jezika na svim jezičnim razinama, pa i na semantičkoj. Semantičke tehnologije ne zanima struktura poveznica među različitim mrežnim sjedištima, odnosno dokumentima nego odnosi među elementima i njihovim svojstvima. Utvrđivanjem odnosa među elementima i njihovim svojstvima s pomoću metapodataka omogućava se strukturiranje nestrukturiranih ili polustrukturiranih podataka na mreži. Semantičke tehnologije uključuju alate za automatsko prepoznavanje tema i koncepata, ekstrakciju značenja i informacija te kategorizaciju podataka. Informacijske znanosti promatraju informacije po strukturalističkome principu. Izvor:
http://www.mensa.hr/glavna/misli-21-stoljeca/652-semanticki-web

Pisanje transformacija [pretraživanje Amazona]

Promijenila sam modul koji koristi Amazon Product API tako da prikuplja i imena autora.

Dokumentacija API-ja nalazi se na stranici:
http://webservices.amazon.com/AWSECommerceService/AWSECommerceService.wsdl

U Maltegu sam stvorila entitete *Book* i *Author* i dodala ih u vrstu entiteta *Academic*.

Book Properties

- Value: Vrijednost entiteta, naziv knjige, ispisuje se u grafu u Maltegu
- Full Title: Pun naziv knjige
- Authors: Popis autora
- EAN: Internacionalni broj članka
- Price and Currency: Cijena i valuta

Author Properties

- Value: Puno ime

- First Name: Ime autora, često uz inicijal srednjeg imena
- Last Name: Prezime

Napisala sam 3 transformacije:

1. **To Amazon Books** - input: *Phrase*, output: *Books*
2. **To Amazon Authors** - input: *Phrase*, output: *Authors*
3. **To Amazon Books By Authors** - input: *Author*, output: *Books*
4. **To Amazon Authors By Books** - input: *Book*, output: *Authors*

Neki primjeri izvođenja u Maltegu...

Transformacije koje vraćaju *Books* kao osnovnu vrijednost entiteta vraćaju 60 znakova naslova knjige zbog ljepšeg prikaza na grafu. Pun naslov knjige se nalazi u svojstvu entiteta *Full Title*.

Problemi pri ispisu podataka u Maltegu su uglavnom nastajali zbog toga što Amazon podržava i ne-ASCII znakove dok ih Maltego ne podržava. To je za sada riješeno ignoriranjem ne-ASCII znakova u podacima. Također, iako su podaci većine knjiga ispravno upisane u Amazon, postoje neke vrijednosti koje su krivo upisane, npr. imena autora skupa s naslovom knjige u polju gdje bi se trebao nalaziti samo naslov, dva autora upisana u polje za jednog autora, prvo upisano prezime a zatim ime itd.

Sljedeći zadatak: probati poboljšati parser za imena autora koji za sada radi za formate u kojima je upisana većina autora: Parser imena za sada pretvara sva imena u *TitleCase*, pronalazi ime (sastoji se od prvog imena i inicijala) i prezime za formate imena u kojima je prezime na zadnjem mjestu (većina autora na Amazonu).

Pretraživanje akademskih pojmova

Academic databases and search engines:
http://en.wikipedia.org/wiki/Academic_databases_and_search_engines

Scirus

http://www.scirus.com/

Pretraživanje članaka

Google Scholar

http://scholar.google.hr/

Pretraživanje akademskih članaka. Od ožujka 2012. nije dostupan putem Google AJAX API-ja.

Programska biblioteka za PHP:
http://code.google.com/p/bioguid/source/browse/trunk/www/scholar_ris.php?spec=svn112&r=112

Microsoft Academic Search

http://academic.research.microsoft.com/ - Pretraživanje publikacija i autora

Library of Congress

http://www.loc.gov/index.html

Može se pristupiti putem Z39.50 protokola: http://www.loc.gov/z3950/gateway.html#about

Z39.50

- http://old.cni.org/pub/NISO/docs/Z39.50-brochure/50.brochure.part01.html
- http://www.oclcpica.org/content/1400/pdf/z3950_handbook_paper.pdf
- http://code.google.com/p/yaz4python/
- http://pypi.python.org/pypi/PyZ3950/
- http://perl.z3950.org/docs/visit.html
- http://zoom.z3950.org/api/
- https://github.com/asl2/PyZ3950
- http://www.panix.com/~asl2/software/PyZ3950/zoom.html

Plan za daljnji rad:

- Vidjeti može li se napisati još koja korisna transformacija za Amazon
- Implementirati Python parser za ljudska imena: http://code.google.com/p/python-nameparser/
- Pronaći još Web alata za pretragu akademskih pojmova
- Istražiti upotrebu Z39.50 protokola te pronaći korisne programske biblioteke koje implementiraju korištenje Z39.50

8. tjedan (23.-29.4.2012.)

Izvještaj sa sastanka, ponedjeljak, 23.4.2012., 09:00

Plan za dalje...

- Što označava EAN na Amazonu - ISBN i koja verzija
- Dodati svojstva *Publisher* i *Year/Date* u entitet *Book* (godina izdanja je naznačena u svojstvu *Publisher*)
- Istražiti izvode li se transformacije na TDS-u ili na lokalnom serveru jednom kada su dodane u vlastiti Transform Seed

- Otkriti načine plaćanja komercijalne verzije Maltega
- Obaviti izvoz konfiguracije Maltega i testirati pokretanje lokalnih transformacija na udaljenom računalu (koje module je potrebno instalirati - Python 2.7, lxml, bottlenose, BeautifulSoup...)
- Početi na pisanju transformacije za Library of Congress (pomoću Z39.50 protokola) te nakon toga otkriti najbolje načine za pretraživanje akademskih članaka i ustanova

Izvještaj

Dodana su svojstva *Publisher* (izdavač) i *Release/Date* (datum izdanja) u entitet *Book*.

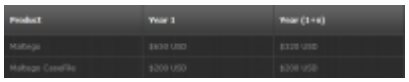
EAN za knjige na Amazonu označava ISBN-13 te sam stoga preimenovala svojstvo *EAN* u *ISBN-13* - http://www.amazon.com/gp/seller/asin-upc-isbn-info.html

Buying Maltego / Maltego CaseFile

What do I get?

- A license key that is valid for one year on a single host gives immediate access to run transforms.
- Unlimited use of transforms on commercial server, shared only by other commercial users.
- Email-based support and integrated documentation.
- Free to use with any external transforms and/or transform application servers.

Plaćanjem se dobije licencni ključ koji vrijedi jednu godinu na jednom serveru, neograničen pristup svim transformacijama na komercijalnom serveru, e-mail podršku i integriranu dokumentaciju i slobodno korištenje svih vanjskih transformacija i njihovih servera.

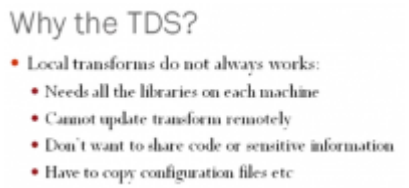


Maltego radi na Windowsima, Linux-based operativnim sustavima i Apple operativnim sustavima. Zahtijeva Java Runtime Environment 1.6 od Oracle-a (prijašnjeg Sun Microsystemsa).

Transform Distribution Server

Uvod u Transform Distribution Server:

https://www.youtube.com/watch?v=dl5-53AgCJ4



TDS je Web aplikacija putem koje korisnici omogućuju dijeljenje vlastitih lokalnih transformacija sa ostalim korisnicima putem Transform Seeda. Međutim, lokalne transformacije se postavljaju i pokreću na lokalnom serveru i nije ih moguće postaviti na TDS.

9. tjedan (30.4.-6.5.2012.)

Izvještaj

Library of Congress

Instalirala sam pomoćne Python biblioteke za Zoom koje podržavaju Z39.50 protokol:

- http://www.panix.com/~asl2/software/PyZ3950/
- http://code.activestate.com/pypm/pyz3950/

te PLY (Python Lex-Yacc): http://www.dabeaz.com/ply/

Zoom dokumentacija: http://zoom.z3950.org/api/zoom-current.html

Napisala sam skriptu koja preuzima rezultate pretrage Library of Congress spajanjem putem Z39.50 protokola. Pronašla sam parser za MARC podatke i pomoću njega ostvarila čitanje podataka o imenu autora knjige. Napisala sam transformaciju *To LoC Authors* koja kao ulaz prima entitet *Phrase*, a kao izlaz vraća *Author* enitete.

Za sada sam otkrila kako obaviti pretraživanje autora i titlova knjiga te kako prikupiti podatke o autorima.

Plan za daljnji rad:

- Otkriti kako pretražiti knjige te kako prikupiti podatke o pojedinoj knjizi iz MARC zapisa (prilagoditi parser)
- Otkriti na koji se način Z39.50 protokolom prikupljaju MARC zapisi (by relevance, by date, by title, random...) te zapisi od njih su najvažniji - idealno bi bilo da su već poredani po važnosti
- Napisati transformacije za pretraživanje knjiga po ključnoj riječi te za pretraživanje knjiga po autoru

10. tjedan (7.-13.5.2012.)

Izvještaj sa sastanka, ponedjeljak, 7.5.2012., 13:00

Plan za daljnji rad:

- Istražiti sustave koji podržavaju Z39.50 protokol (Koha, NSK baza podataka itd.)
- Nakon toga: istražiti pretraživanje članaka preko nekih od baza na: http://www.online-baze.hr/, projekte i predmete na sveučilištima, videozapise (YouTube)

Izvještaj

Library of Congress

Napisala sam još 3 transformacije za Loc. Sve transformacije:

1. **To LoC Authors** - input: *Phrase*, output: *Authors*
2. **To LoC Books** - input: *Phrase*, output: *Books*
3. **To LoC Books By Authors** - input: *Author*, output: *Books*
4. **To LoC Authors By Books** - input: *Book*, output: *Authors*

Također, dodala sam i transformaciju **To Amazon Authors By Books**.

MARC zapisi koju se prikupljaju putem Z39.50 iz Library of Congress nisu sortirani i ponavljaju se. Ispis prikladnih rezultata je za sad riješen u **To LoC Authors** gdje sam implementirala *Counter* koji grupira rezultate po autoru i sortira ih prema broju ponavljanja u rezultatima. Izvođenje transformacije je zbog toga nešto sporije, ali se u Maltegu ispisuju relevantni autori. Stoga bi se nešto slično trebalo implementirati i u ostalim transformacijama.

Nadalje, transformacije koje vraćaju Book entitete za sad vraćaju samo naziv knjige, puni naziv i autore. Potrebno je dodati parsiranje ostalih svojstava entiteta uz pomoć: http://www.loc.gov/marc/bibliographic/ecbdhome.html

Problem koji se javljao pri parsiranju svih mogućih znakova koji nisu ASCII niti utf-8 je riješen dekodiranjem te ponovnim kodiranjem svakog zapisa.

Konačno, mislim da bi bilo dobro preimenovati sve transformacije u oblik *To Entity [UsingSearch Engine]*, npr. *To Books [Using Amazon]* jer su uglavnom tako imenovane transformacije, iako nije pravilo.

Online baze podataka

http://www.online-baze.hr/

NSK

- http://www.nsk.hr/
- http://katalog.nsk.hr/F?RN=865770126

Integrirani knjižnični sustav NSK:

- http://iks.nsk.hr/

Z39.50 serveri

- http://staff.library.mun.ca/staff/toolbox/z3950hosts.htm
- Z-BRARY: http://www.z-brary.com/n.html

Podaci za pristup NSK katalogu kao Z39.50 serveru postoje na Z-BRARY, ali mi nije uspjelo spajanje na server → poslati e-mail:

Hostname: katalog.nsk.hr

Port: 7090

Database: voyager

Syntax: Opac

Saznala sam da se Koha može postaviti kao Z39.50 server.

Podaci za pristup Z39.50 serveru knjižnice Filozofskog fakulteta su javno dostupni na službenoj stranici knjižnice. Uspjela sam se spojiti na njihov server i prikupiti zapise:

Što se tiče pristupa knjižničnom sustavu FER-a nisam uspjela pronaći podatke o spajanju na Koha-u → potrebno je poslati e-mail CIP-u.

11. tjedan (14.-20.5.2012.)

Izvještaj sa sastanka, četvrtak, 17.5.2012., 9:30

Poslati upite za:

- FER-ov Koha sustav (marijana.glavica@ffzg.hr, dobrica.pavlinusic@ffzg.hr)
- Online baze podataka (danijel.namjesnik@irb.hr, jadranka.stojanovski@irb.hr)
- Pristup integriranom knjižničnom sustavu http://iks.nsk.hr/ (Aleph_projektni_tim@nsk.hr)

Izvještaj

Upiti su poslani.

EBSCOHost http://support.epnet.com/knowledge_base/detail.php?id=2591

Z39.50 Access: Users can integrate EBSCO databases into a library ILS using the Z39.50 protocol. EBSCO provides a full range of Z39.50 compliant functionality, including:

Federated Search: *Serial Solutions (360 Search)*, *MetaLib (Ex Libris)*

Bibliographic Management software: *EndNote (Thomson Reuters)*

Library Catalog Portals: *ENCompass (Endeavor)*, *Horizon Information Portal (SirsiDynix)*, *One Search (Follett)*

OVID
http://ovidsupport.custhelp.com/app/answers/detail/a_id/1502/kw/z3950/session/L3RpbWUvMTMzNzI5

NTQ0NS9zaWQvdHVvbTNuWWs%3D

Ista greška se javlja pri spajanju na OVID i EBSCO. OVIS i EBSCO su navedeni u popisu baza podataka kojima je moguće pristupiti putem klijentskog programa EndNote, a podaci o spajanju putem Z39.50 se također odnose na postavke za EndNote program.

http://www.lib.vt.edu/endnote/z3950.html

The databases below are from commercial vendors who allow direct connection to some or all of their databases via EndNote connection files or for which there is an available import filter or where there is direct export functionality.

Moguća rješenja → isprobati još mogućih kombinacija koristeći Python Zoom, pokrenuti Z39.50 klijenta napisanog u nekom drugom programskom jeziku kao što je Java (međutim, čak i ako povezivanje proradi to povlači sa sobom jako puno posla vezanog uz parsiranje i spajanje na Maltego), pronaći bazu članaka i publikacija koja je Z39.50 compliant ili koja ima dostupan API.

Servis za prikupljanje Z39.50 zapisa: https://github.com/dpavlin/Biblio-Z3950

Google Scholar

Google Scholar nema API, ali sam našla modul kojim se Google Scholar može pretraživati po naslovu i autorima:

- http://www.icir.org/christian/scholar.html
- http://www.icir.org/christian/downloads/scholar.py

Can extract publication title, main online URL, number of citations, number of online versions, link to Google Scholar's main cluster for the work, and Google Scholar's cluster of all works referencing the publication.

Dodala sam novi entitet *Article* i napisala 2 transformacije:

1. **To Articles [Using Google Scholar]** - input: *Phrase*, output: *Articles*
2. **To Articles By Author [Using Google Scholar]** - input: *Author*, output: *Articles*

FER Koha

Ostvarila sam pristup, moguće je pretraživati po naslovu i autoru.

Hostname: lib.fer.hr

Port: 9999

Database: biblios

Encoding: UTF-8

Syntax: MARC21/USMARC

Transformacije:

1. **To Books [Using FER Koha]** - input: *Phrase*, output: *Books*
2. **To Authors [Using FER Koha]** - input: *Phrase*, output: *Authors*
3. **To Books By Author [Using FER Koha]** - input: *Author*, output: *Books*

FFZG Koha

Ostvarila sam pristup, moguće je pretraživati po naslovu i autoru.

Hostname: koha.ffzg.hr

Port: 9999

Database: biblios

Syntax: MARC21

Encoding: utf-8

Transformacije:

1. **To Books [Using FFZG Koha]**
2. **To Authors [Using FFZG Koha]**
3. **To Books By Author [Using FFZG Koha]**

Plan za dalje: dodati vraćanje svojstava u transformacije za LoC, dovršiti transformacije za FER Kohu, saznati je li potrebno pisati transformacije za FFZG, promijeniti imena svih transformacija u Maltegu, saznati Z39.50 za NSK ili Integrirani knjižnični sustav NSK (ako postoje).

12. tjedan (21.-27.5.2012.)

Izvještaj

Dovršene su transformacije za **Amazon**, **Library of Congress**, **FER Kohu**, **FFZG Kohu Google Scholar**, a stare transformacije su preimenovane. Dodana je transformacija **To Authors By Book** koja vraća autore knjige. Transformacija je bitna zbog načina na koji Maltego iscrtava veze među entitetima.

- **To Books [Using Amazon]** - input: *Phrase*, output: *Books*
- **To Authors [Using Amazon]** - input: *Phrase*, output: *Authors*
- **To Books By Author [Using Amazon]** - input: *Author*, output: *Books*

- **To Books [Using LoC]** - input: *Phrase*, output: *Books*
- **To Authors [Using LoC]** - input: *Phrase*, output: *Authors*
- **To Books By Author [Using LoC]** - input: *Author*, output: *Books*

- **To Books [Using FER Koha]** - input: *Phrase*, output: *Books*
- **To Authors [Using FER Koha]** - input: *Phrase*, output: *Authors*
- **To Books By Author [Using FER Koha]** - input: *Author*, output: *Books*

- **To Books [Using FFZG Koha]** - input: *Phrase*, output: *Books*
- **To Authors [Using FFZG Koha]** - input: *Phrase*, output: *Authors*
- **To Books By Author [Using FFZG Koha]** - input: *Author*, output: *Books*

- **To Articles [Using Google Scholar]** - input: *Phrase*, output: *Articles*
- **To Articles By Author [Using Google Scholar]** - input: *Author*, output: *Articles*

- **To Authors By Book** - input: *Book*, output: *Authors*

Maltego ne može pročitati ne-ASCII znakove pa se trenutno dijakritički znakovi prevode u upitnike. Saznati je li moguće prevesti te znakove.

Izvještaj sa sastanka, petak, 25.5., 12:30

Zaključci:

- Pokušati pronaći podatke za Z39.50 za sljedeće baze podataka: Current Contents, Web of Science, Springer Link, Scopus, Inspec
- Normalizirati dijakritičke znakove i ostale ako je moguće
- Urediti kod, komentirati bitnije dijelove
- Dogovoren je okvirni izgled samog diplomskog rada. Rad bi se trebao sastojati od sljedećih dijelova:

1. Uvod u problematiku područja
2. Istraživanje, opisi alata, usporedbe, tablice
3. Motivacija za korištenje Maltega
4. Opis rada na transformacijama
5. Općenite upute za pisanje transformacija
6. Zaključak

- Početi pisati diplomski rad

13. tjedan (28.5.-3.6.2012.)

Izvještaj

Što se baza podataka sa Centra za online baze podataka tiče, niti jednoj nije moguće pristupiti pomoću improviziranog Z39.50 klijenta već samo pomoću programa namijenjenih pristupu kao što je EndNote.

Current Contents, *Web of Science*, *SpringerLink* – pomoću EndNote-a, *Scopus* – nije moguće pristupiti putem Z39.50, *Inspec* – moguće putem podataka za Ovid također pomoću EndNote-a

Riješen je problem normalizacije za dijakritičke znakove, a kako se čini i za većinu ostalih znakova. I dalje se još znaju javljati greške pri pretraživanjima, doduše rijetko. Potrebno je još malo testirati. Kod koji radi najbolje do sada:

```
def downcode(string):

string = string.decode('utf-8', 'ignore')

string = unicode(string)
```

```
string = unicodedata.normalize('NFKD', string).encode('ascii', 'ignore')

return string
```

Update: Čini mi se da sam ispravila sve greške vezane uz parsiranje znakova (i zapisa općenito). Neke transformacije se ponekad ne izvrše, ali samo ako ih je puno pokrenuto odjednom, dok se inače normalno izvode.

Moduli su posloženi po direktorijima, a kod je uređen i komentiran. Ostavila sam neki višak u komentarima koji bi mi još mogao poslužiti pri otkrivanju potencijalnih grešaka te zakomentirane “testne” dijelove koda. Njih ću maknuti prije predaje.

Jedan primjer izvođenja transformacija na izrazu “matematika 3” sam spremila u *matematika 3.mtgx* datoteku.

Preuzela sam predložak i upute za pisanje diplomskog rada.

Izvještaj sa sastanka, petak, 1.6., 12:30

14. tjedan (4.-10.6.2012.)

Izvještaj sa sastanka,

Izvještaj

15. tjedan (11.-17.6.2012.)

Izvještaj sa sastanka,

Izvještaj