Žad Deljkić

Programski alati za automatsko mapiranje tematskih područja

Software tools for automatic mapping of subject areas

Zadatak

U svrhu brzog obuhvaćanja temeljnih informacija za neko tematsko područje, potrebno je predložiti programske alate koji bi automatski prikupljati standardne podatke o tematskom području te grafički prikazivali međuzavisnosti na način prilagođen korisniku za daljnju analizu. Naglasak staviti na programsku dogradnju i korištenje programskih alata u širokoj uporabi, poput internetskih preglednika, i rješenja zasnovana na otvorenom kodu.

Dnevnik rada	GitHub repozitorij

Plan rada

1. Napisati upute za uspješno prevođenje i pokretanje NetGlub master-a, slave-a i klijenta (GUI)
2. Napraviti i dokumentirati što veći broj različitih transformacija
3. Napisati detaljne upute za izradu transformacija - od izrade Python skripte do integriranja u NetGlub sistem
    1. puno o izradi transformacija za netglub se može naučiti iz resursa za Maltego (primjer) - jer na kraju krajeva Netglub je baš kopija Maltega
4. ~~Uspješno prevesti i pokrenuti klijent-a na Windows OS-u, napisati upute za to~~
5. ~~Isprobati cijeli sistem preko mreže - master i slave(ovi) na Ubuntu OS-u, klijenti na Ubuntu i Windows OS-u~~
6. ~~Napisati popis budućih projekata u kontekstu ovog rada~~
7. Ubaciti dobivanje ISBN-a u tranformacije s knjigama
    1. trebao bi biti posebni parametar da/ne koji nas pita jel želimo ISBN - zato što dobivanje ISBN-a može biti sporo (otvaranje nove stranice za svaki rezultat)
8. Slično za fraze, treba napraviti mehanizam (Python modul) koji se bavi validacijom fraza - on će nam reći jel fraza koju smo generirali ima smisla (primjerice da proba naći tu frazu u naslovu/sažetku knjiga i radova)
    1. Onda kada imamo taj mehanizam, treba postojati parametar koji nas pita je li želimo velidaciju fraza prilikom neke od transformacija koja generira fraze - jer je opet validacija relativno spora (otvaranje nove stranice)
9. Ostalo:
    1. ~~Provjeri je li NetGlub smatra da su dva entiteta sa istim primarnim (npr. naslov knjige), ali različitim sekundarnim atributima (npr. autori, izdavač) jednaki (je li ih “merge-a”)~~ saznato - zapisano u poglavlje “Entiteti” niže
    2. PhraseToSimilarPhrasesW - makni rezultate sa “sidebar” i “nav” u nazivu
    3. Transformacija ili mehanizam koji provjerava je li fraza “legitimna” - pokuša naći knjige, članke, i sl. i ako ništa ne nađe onda nije?

Upute za NetGlub

- Općenito o NetGlub-u, osnovno korištenje NetGlub-a
- Prevođenje i pokretanje NetGlub master-a, slave-a i klijenta na Ubuntu 14.04.
- Prevođenje i pokretanje NetGlub klijenta na Windows 7
- Izrada NetGlub transformacije i integriranje u cijeli sistem
- Pokretanje NetGlub master-a i slave-a za mrežni rad

Entiteti

- Phrase - default entitet, najčešće ulazni
- Book
- Article
- Conference
- Author

Zapisivanje imena autora:

U entitetima Book i Article bi u polju “authors” imena autora trebala biti ovako zapisana:

<prezime1>, <ime1>; <prezime2>, <ime2>; ...

Prezime bi trebali biti puno prezime, dok ime može biti puno ime, inicijali i slično - ovisi što nam je dostupno kod baze koju pretražujemo.

U različitim stranicama/bazma je ime autora različito zapisano. U nekim slučajevima u istoj bazi ime autora može biti zapisano u različitim formatima. Zadatak transformacije koja daje knjigu ili članak kao izlaz da pokuša pročitati i zapisati imena autora u ispravnom formatu.

Sam entitet autora ima jedno obavezno polje - prezime. Tako da i ako je na nekim stranicama zapisano puno ime, na nekima samo inicijali, a na nekima primjerice puno prvo ime i inicijali srednjeg imena, autor će se i dalje prepoznati kao isti. Nedostatak toga je rubni slučaj kada postoji više autora sa istim prezimenom u istom području, ali to je prihvatljiva posljedica za dobivenu fleksibilnost.

Isti (spojeni entiteti):

Netglub smatra da su dva entiteta jednaka ako su im svi obavezni atributi (optional=false) jednaki. Službeno ne postoji nikakav “primarni” atribut - no konvencija među default entitetima je da svaki entitet ima atribut nazvan “value” koji je obavezan (uz još ostale obvezne/neobvezne atribute).

Tu nastaje dilema - ako je za knjige i članke jedini obvezni atribut naslov, netglub neće moći razlikovati dve knjige s istim naslovom. Zato je potrebno uvesti još neki atribut koji će ih razlikovati.

Ja sam stavio da to bude polje autora. U tom slučaju se može dogoditi da netglub smatra da su dvije zapravo iste knjige različite jer imaju različiti tekst u polju autora (npr. “Doe John” i “Doe J”), no u praksi je to manji problem od prvog problema kojeg ovo rješava (različite knjige - isti naslov).

Izvor: datoteka qng/src/graph.cpp, funkcija Node::isSimilitarTo

Transformacije

U sve transformacije bi trebalo implementirati mehanizam da ne “zaspammaju” stranice koje pretražuju - primjerice ako tražimo članke s nekom temom iz samo 2015 godine, mogli bi pretražiti 1000 stranica Google Scholara i naći samo par rezultata. Ukoliko je korisnik zadao da npr. pronađemo 20 takvih rezultata, transformacija ih ne smije beskonačno tražiti već pretražiti razumnu/“pristojnu” količinu.

Gotove transformacije:

- PhraseToSimilarPhrasesW - uzima frazu i nalazi slične fraze pomoću Wikipedije
- PhraseToBookLOC - uzima frazu i nalazi relevantne knjige pomoću Library of Congress
- PhraseToArticlesGS - uzima frazu i nalazi relevantne članke pomoću Google Scholar-a
- PhraseToConferenceCA - uzima frazu i nalazi relevantne konferencije pomoću www.conferencealerts.com
- PhraseToBookArticleKoha - uzima frazu i stranicu koja koristi Koha sistem kao parametar, vraća relevantne članke i knjige. Primjeri stranica sa Koha sistemom:
  - FER Središnja Knjižnica
  - Knjižnica FFZG
- Book/ArticleToAuthor - transformacija koja će izvući autore iz knjiga/članaka, bila bi korisna kako bi se mogli grafički viditi koko kojih autora se “grupiraju” knjige i članci (tj. tko je potencijalno “bitniji” unutar nekog područja)
- PhraseToArticleIEEE - uzima frazu, vraća članak iz IEEE Xplore baze

Buduće transformacije:

- PhraseToBook/Article - različite transformacije koje uzimaju frazu i vraćaju relevantne knjige/članke koristeći sljedeće baze:
  - Amazon
  - NSK
  - Hrčak
  - Crosbi
  - Web of Science
  - Online baze (koristeći proxy?)
  - ACM Digital Library
- PhraseToConference - različite transformacije koje uzimaju frazu i vraćaju relevantne konferencije koristeći sljedeće baze:
  - IEEE Conference and Events
- Transformacije u suprotnom smjeru (autor → knjige/članci → fraze)
- Transformacija koja povezuje osobu s FER-a i zavod koristeći https://www.fer.unizg.hr/imenik
- PhraseToSimilarPhrases - transformacija za nalaženje sličnih fraza samo mijenjanjem individualnih riječi u frazi njenim sinonimima + možda mehanizam koji za svaki dobiveni rezultat provjerava jel dobar/zadovoljavajući

Lokalno testiranje transformacija:

Primjer pozivanja transformacije iz ljuske:

./transform phrase value test “” NbResult 5

Gornji primjer poziva transformaciju s ulaznim entitetom tipa “phrase”, čiji je atribut “value” jednak

“test” i s parametrom “NbResult” jednakim “5”

Bugovi/greške kod Netglub-a

Netglub za komunikaciju koristi XML-RPC - specifikacija (link) može pomoći kod debuggiranja.

Općenito

- Master nezna reagirati na timeout od mysql baze
  - ako se nista ne desava neko vrijeme (8 sati po defaultu kod mysql-a) master ce izgubit konekciju i kod sljedeceg pokusaja konekcije od klijenta ce se pojaviti serial not valid
  - rjesenje za sad je povecati wait_timeout na maksimalnu vrijednost od 1 godine kod mysql-a i restartat master bar 1 godisnje
  - bolje rješenje - promjena source koda, dvijee opcije:
  - a) main.cpp: setconnectoptions(mysql_opt_reconnect) ne cini se da radi
  - b) ntgsessionmanager.cpp: db.isopen() i isactive() navodno ne funkcionira https://bugreports.qt.io/browse/qtbug-223
  - implementirati provjeru jel db ziv na neki drugi nacin?
- Kod netglub klijenta, ako se unese URL i konkretno port koji drop-a konekcije (obično firewall), klijent će se “zamrznuti” i vječno čekati odgovor
  - Što je još gore, ako je bio odabran “Remember settings”, svaki put kod paljenja će se isto dogoditi i jedini način za spriječini to je ručno naći gdje su pohranjeni settingsi
  - Na windowsima su settingsi pohranjeni u registry-u pod HKEY_CURRENT_USER\Software\Diateam\qng (URL i serial)
- Na nekim računalima kod netglub klijenta, klijent se nemože spojiti i baca grešku -32300 HTTP request failed
  - Gledajući kroz source, čini se da klasa QHttp (relevantni source, funckija QHttpPrivate::_q_slotError) interno koristi QTcpSocket i taj socket se ne uspjeva spojiti i baca neku grešku (popis mogućih greška)
  - Ne znam još koju točno grešku baca, znam samo po source-u koje ne baca (0, 1, 2 sa popisa)
- Transformacija PersonToEmailSE ima par greška:
  - vraća dobre rezultate za “john smith” ali ne i za “John Smith” → ubaciti pretvaranje u lowercase prije
    - values = [v.lower() for v in values] nakon values
  - regex za email u dijeli koji matcha domenu bez top dijela treba biti greedy (pretvorit *? u samo *)
    - inače će e-mail primjer@pod.domena.com prepoznati kao primjer@pod.domena
  - kad se dogodi exception ne napravi ništa (pass) i nastavi dalje, expection mogu biti razne greške, u najmanju ruku treba ispisati poruku - ali ne sa write_error jer takvo ispisivanje implicira prekid transformacije

Sigurnost

- Master na istom interfaceu/ip adresi očekuje i slaveove i klijente - to bi trebalo biti moguće odvojiti

    - npr. slave-ovi samo na 127.0.0.1, a klijenti na 0.0.0.0
- Master nikako ne autorizira slaveove - postoji “kostur” te funkcionalnosti ali nije implementirana
- Datoteke u /etc/netglub/ obično svi korisnici mogu čitati - a između ostalog sadrže username i password za bazu podataka, privatne ključeve certifikata...

Budući projekti

Općenito

- Alat koji prebacuje datoteku u NetGlub (.ntg) formatu u mind mapu u Freemind formatu (.mm)
    - .ntg je zippani XML, .mm je čisti XML - jednostavna python skripta bi bila dovoljna
- Alat koji potpuno zaobilazi NetGlub i “ručno” korisničko prtraživanje direktno koristi transformacije kako bi od ulazne fraze automatski generirao mapirano područje (možda u obliku mind mape)
    - Alat bi prvo mogao naći slične fraze pomoću PhraseToSimilarPhrases i filtrirati dobre pomoću nekog mehanizma
    - Zatim bi našao knjige, članke i slično pomoću već postojećih transformacija
    - Na kraju bi rezultate nekako lijepo prikazao - možda u obliku Freemind mind mape (.mm)
- Izrada alternativnog NetGlub klijenta koji radi unutar internet preglednika
- Pokrenuti javno dostupan server koji pokreće Netglub infrastrukturu na koju se ljudi mogu spojiti samo sa klijentom, bez potrebe da sami doma imaju pokrenut master i slave
    - Možda je za tu svrhu moguće dobiti VPS od srca?

Nadogradnja Netglub-a

- Izrada dodatnih transformacija (vidi “Buduće transformacije” gore)
- Izrada funkcionalnosti lokalnih transformacija u klijentu (već je započeta, postoji kostur)