Romana Grilj
Katarija 6a, 1251 Moravče, Slovenija
Študijski program: Računalništvo in informatika, MAG
Vpisna številka: 63080523

Komisija za študijske zadeve
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Večna pot 113, 1000 Ljubljana

Vloga za prijavo teme magistrskega dela
Kandidat: Romana Grilj

Romana Grilj, študent/-ka magistrskega programa na Fakulteti za računalništvo in informatiko, zaprošam Komisijo za študijske zadeve, da odobri predloženo temo magistrskega dela z naslovom:

Slovenski: **Primerjava uspešnosti odprtokodnih in komercialnih orodij za luščenje podatkov**
Angleški: **Performance comparison of open source and commercial information extraction tools**

Tema je bila že potrjena lani in je ponovno vložena: ***DA***

Izjavljam, da so spodaj navedeni mentorji predlog teme pregledali in odobrili ter da se z oddajo predloga strinjajo.

Magistrsko delo nameravam pisati v slovenščini.

Za mentorja/mentorico predlagam:

Ime in priimek, naziv: doc. dr. Slavko Žitnik
Ustanova: Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Elektronski naslov: slavko.zitnik@fri.uni-lj.si

V Ljubljani, 7. januar 2023.

PREDLOG TEME MAGISTRSKEGA DELA

1 Področje magistrskega dela

slovensko: analiza podatkov, spletno rudarjenje
angleško: data analysis, Web Mining

2 Ključne besede

slovensko: Analiza podatkov, ekstrakcija podatkov, strukturni podatki, spletno rudarjenje
angleško: Data analysis, Information Retrieval, structural data, Web Mining

3 Opis teme magistrskega dela

Pretekle potrditve predložene teme:
Predložena tema je bila oddana in potrjena v preteklih letih.

3.1 Uvod in opis problema

Cilj avtomatske ekstrakcije podatkov s spleta (ang. web content mining) je odkriti koristne informacije in znanje iz spletnih povezav, vsebine strani in uporabnih podatkov. Ekstrakcija podatkov je pomembna, da pridemo do željenih podatkov. Dober primer je spletna stran ceneje.si, ki iz različnih spletnih strani pridobiva podatke o cenah izdelkov. V magistrski nalogi bom preverila kakšni so načini pridobivanja podatkov iz spleta. Pregledala bom aplikacije in metode, ki jih nudijo podjetja na tem področju. Posebej me bodo zanimala podjetja, ki so se razvila iz akademskih sfer, pri katerih bom lahko našla pristope, ki jih uporabljajo za pridobivanje podatkov z spleta. Naredila bom analizo funkcionalnosti, ki jih je mogoče izvajati (npr. ekstrakcija imenskih entitet, sentimenta, storitve, ki jih podjetja nudijo). Izbrala bom tri različne pristope, katere bom implementirala s pomočjo javno dostopnih orodij in primerjala njihovo uspešnost na realnih podatkih iz spleta s podatki, ki jih obljubljajo podjetja oz. so objavljeni v znanstvenih člankih na označenih korpusih. Pridobljeni podatki so lahko v strukturirani ali nestrukturirani obliki. Primer nestrukturiranih podatkov so zapisi v text-u, brez definirane strukture. Strukturirani podatki pa imajo skladno v naprej definirano obliko, kot na primer Wikipedija ali javni registri.

3.2 Pregled sorodnih del

Pregled sorodnih del sem razdelila na dva dela, kjer prvi del vsebuje pregledovanje spleta in drugi del predstavlja metode ekstrakcije.

Liu [1] je napisal obsežno besedilo o spletnem rudarjenju, ki je sestavljen iz dveh delov. Prvi del opisuje temelje podatkovnega rudarjenja in strojnega učenja, kjer so predstavljeni vsi bistveni koncepti, algoritmi podatkovnega rudarjenja in strojnega učenja. Drugi del pa zajema ključne teme spletnega rudarjenja, kjer uporabljamo metode prečesavanja spleta, iskanje, analiza družabnih omrežij, pridobivanje strukturiranih podatkov, integracija informacij, analiziranje mnenj in analiza razpoloženja, rudarjenje po spletnih straneh, rudarjenje poizvedb, računsko oglaševanje in priporočevalni sistemi obdelane tako v širino kot v globino. Spletni pajek (angl. Web crawler) [2], je interni agent ki sistematično brska po svetovnem spletu, običajno za namene spletnega indeksiranja.

Raziskovalci [3] so pregledali strukturirano pridobivanje podatkov s spletnih strani. Pridobljeni strukturni podatki so lahko kasneje integrirani in ponovno uporabljeni v širokem spektru aplikacij, kot so portali za primerjavo cen, orodja za poslovno inteligenco in druge. Spodbujajo industrijo in akademike da iščejo samodejne rešitve za reševanje problemov. Predstavili so nov pristop strukturiranega pridobivanje podatkov, ki temelji na združevanju vizualno podobnega spleta elementov strani. Njihova metoda ClustVX združuje vizualne in čiste HTML značilnosti spletnih strani, da združi vizualno podobne strani ter izvleče strukturirane spletne podatke. Pridobi lahko strukturirane podatke, na katerih je prisoten več kot en podatkovni zapis. Z obsežno eksperimentalno oceno treh referenčnih naborov ClustVX dosega boljše rezultate kot druge najsodobnejše metode.

Izvleček spletnih podatkov, aplikacij in tehnik opisuje pet različnih tehnik [4], ki se uporabljajo za reševanje problema pridobivanja spletnih podatkov. Uporabniki po spletnih straneh lahko brskajo na vizualni ali besedilni način ali pa ga preprosto vnesejo v sistem s spletnim naslovom dokumentov, ki vsebujejo informacije. Generacijo ovoja zaenkrat definirajo kot postopek za pridobivanje nestrukturiranih informacij iz vira podatkov in jih preoblikujejo v strukturirane podatke. Avtomatizacija in razporejanje je ena izmed najpomembnejših lastnosti ekstrakcije sistemov spletnih podatkov. Sposobnost ustvarjanja makrov za izvedbo več primerkov iste naloge, vključno z možnostjo simulacije klika uporabnika, izpolnjevanje obrazcev in izbiranje menijev ter gumbov ter podpora AJAX sistema za asinhrono posodabljanje strani, to so le nekatere najpomembnejše lastnosti avtomatizacije. Preoblikovanje podatkov je pomembno področje, kajti podatke lahko pridobimo iz več različnih virov. Pomembno je tudi čiščenje podatkov, da uporabniki dobijo homogene informacije v edinstveni strukturi.

Članek o raziskavi nenadzorovanih tehnik za pridobivanje spletnih podatkov [5] predstavlja sistem za pridobivanje spletnih informacij, ki so razviti in razvrščeni v ročne, nadzoro-

vane, polnadzorovane in nenadzorovane tehnike. Podrobneje so opisali delovanje različnih algoritmov za raziskovanje. Roadrunner alogaritem deluje na zbirki spletnih dokumentov in vhodnih strani kjer primerja strani z vhodnimi podatki eno z drugo. ExALG poizkuša iskati žetone, kateri so v vhodnem dokumentu in kateri so lahko dodani v predloge, tako da lahko generirajo ekstrakcije s pomočjo meril za razlikovanje in gnezdenje. FivaTech uporablja algoritem ujemanje po drevesu (A tree-matching algorithm) [6] za generiranje predlogov. Metoda Trinity uporablja trinarno drevo, katero je uporabljeno za generiranje regularnih izrazov in potrebuje veliko manj časa za ekstrahiranje v primerjavi z ostalimi tehnikami, zaradi česar je bolj učinkovita.

Spletni pajek [7] je računalniški program, ki brska po svetovnem spletu na metodičen, avtomatiziran ali urejen način. Spletni pajki so iskalniki s polnim besedilom, ki upravljalniku pomagajo pri krmiljenju po spletu. Članek vsebuje pregled različnih tehnik spletnih pajkov. Predstavili so štiri pomembna pravila prečesavanje spleta:
- Politika, ki navaja strani za prenos.
- Pravilnik o ponovnem obisku spletne strani, ki določa kdaj je potrebno preveriti spremembe na straneh.
- Pravilnik vljudnosti, ki določa, kako se izogniti preobremenjenosti spletnih mest.
- Politika paralelizacije, ki določa kako uskladiti porazdeljene pajke po spletu.

Članek opisuje RCrawler [8], ki je dodatni paket statističnega programskega jezika R za prečesavanje spleta in luščenje podatkov. Program razčleni in shrani strani, kjer izvleče vsebino ter ustvari podatke, ki jih aplikacije lahko uporabijo za pridobivanje spletnih vsebin. Sam program je zelo prilagodljiv in ga je mogoče prilagoditi drugim aplikacijam. Glavne značilnosti programa so prečesavanje spleta v več nivojih, ekstrakcija vsebine in zaznavanje podvojenih vsebin. RCrawler ima močno optimiziran sistem in lahko prenese veliko število strani na sekundo, hkrati pa je odporen na določene tipe programskih zrušitev in pasti.

Supervizor [9] je spletna aplikacija, ki omogoča iskanje povezav med poslovnimi subjekti, ter osebami v Sloveniji. Povezave so vizualizirane s pomočjo grafa, kjer vozlišča predstavljajo pravne ali fizične osebe.

Parlameter [10] je aplikacija, ki vsebuje velik nabor metapodatkov o govorcih (spol, sta rost, izobrazba, strankarska pripadnost) in je jezikovno označen (lematizacija, tegiranje), kar omogoča številne raziskave s področja digitalne humanistike in družboslovja. V prispevku pokažejo koruptično analitične tehnike za raziskovanje političnih razprav

3.3 Predvideni prispevki magistrske naloge

1. V okviru magistrske naloge bom podrobneje raziskala področje obstoječih aplikacij in metod, vezanih na pregledovanje spleta in ekstrakcijo podatkov, ki jih nudijo podjetja. Usmerila se bom v odprtokodne programske sisteme, katere je mogoče nadgraditi in uporabiti v namene spletne ekstrakcije podatkov, kot tudi za namene ekstrakcije slik, čiščenja podatkov, nestrukturirane ekstrakcije podatkov, ločeno zbiranje podatkov, itd. Na osnovi pregleda bom poskušala ugotoviti kakšno tehnologijo uporabljajo podatja.

3. S pomočjo predlaganih korpusov in ročnim/pol-avtomatskim pregledom uspešnosti storitev bom izbrala tri ustrezne projekte, s katerimi bom implementirala ključne metode ekstrakcije podatkov iz realnih spletnih strani, ter primerjala evalvacije z obljubami podjetij in rezultati v znanstvenih člankih. Analizirala bom uporabnost med plačljivimi ter odprtokodnimi tešitvami. S primerjanjem bom analizirala uspešnost implementiranih metod s storitvami podjetij, ki se ukvarjajo z nudenjem teh storitev (Google, Azure, AWS, spaCy).

3.4 Metodologija

1. Pripravila bom analizo že obstoječih storitev, ki jih nudijo ponudniki povezani na pregledovanje spleta in ekstrakcijo podatkov s spleta in s tem pridobila podatke katere aplikacije so bolj primerne za kateri jezik, ter katero aplikacijo je najboljše nadgraditi. Usmerila se bom v odprtokodne programske sisteme [11], ki jih je mogoče nadgraditi, dober primer so Textractor [12], Scrapy [13], Tabula [14].

2. Med izbranimi ponudniki bom preverila, kateri omogočajo največ različnih storitev in izbrala tri, kateri bodo imeli največ funkcionalnosti. Po trenutno opravljeni analizi so najaktualnejše storitve:
- izvleček ključnih besed [15] (Samodejna identifikacija izrazov, ki najbolje opisujejo temo dokumenta.)
- prepoznavanje imenskih entitet [16] (Proces ekstrakcije informacij katere cilj je nestrukturiran tekst vnaprej definirane kategorije kot na primer imena oseb, imena organizacij, lokacij ter druge.)
- analiza sentimenta [17] (Tehnika obdelave naravnega jezika, ki določa ali je del vsebine pozitiven, negativen ali nevtralen.)
- povezovanje entitet [18] (Identificira in razjasni identiteto entitet najdenih v besedilu. Na primer v stavku »Prejšnji teden smo šli v Seattle.« bi bila identificirana beseda »Seattle«. Seattle je beseda, vezana na specifičen primerek v obstoječi bazi znanja, npr. WikiData.)
- ekstrakcija razmerja [19] (Indeficira smiselne povezave med pojmi, ki so omenjeni v be-

sedilu. Na primer, relacija »čas stanja« se najde tako, da se ime stanja poveže s časom ali med okrajšavo in/ali celotnim opisom.)

3. Po opravljeni analizi bom izbrala ustrezne odprtokodne projekte, s katerimi bom implementirala ključne metode ekstrakcije podatkov s spleta, ki jih sicer nudijo podjetja, identificirana v analizo.

4. Za analizo bodo uporabljeni korpusi pridobljeni s spletne strani:
- izvleček ključnih besed: »Keyword Extraction on SemEval 2010 Task 8« [20]
- prepoznavanje imenskih entitet: »Named Entity Recognition on WNUT 2017« [21]
- sentimenta razmerja: »Sentiment Analysis on Amazon Review Full« [22]
- povezovanje entitet: »Entity Linking on OKE-2015« [23]
- ekstrakcija razmerja: »Relation Extraction on Adverse Drug Events (ADE) Corpus« [24]

5. Naredila bom analizo uspešnosti že obstoječih implementiranih metod s pomočjo korpusov ali pridobitev rezultatov ter ročen/pol-avtomatski pregled uspešnosti, pri čemer bom primerjala evalvacije z obljubami podjetij in rezultati v znanstvenih člankih. Naredila bom analizo uspešnosti s storitvami podjetij, kot so na primer Google, Azure in AWS, ki se specifično ukvarjajo z nudenjem teh plačljivih storitev.

3.5 Literatura in viri

[1] L. Bing, Web data mining, IEEE Trans. Pattern Anal. Mach. Intell. (2011).

[2] Web crawler, `https://en.wikipedia.org/wiki/Web_crawler` (Nov. 2022).

[3] G. Tomas, C. Antanas, Unsupervised structured data extraction from template-generated web pages (2014).

[4] F. Emilio, F. Giacomo, B. Robert, Web data extraction, applications and techniques: A survey (2010).

[5] P. Disha, T. Dr. Ankit, A survey of unsupervised techniques for web data extraction (2015).

[6] A. M. Pinzón, M. H. Hoyos, J.-C. Richard, L. Flórez-Valencia, M. Orkiszb, A tree matching algorithm: Application to airways in ct images of subjects with the acute respiratory distress syndrome (2017).

[7] S. Amaduha, M. Phill, Web crawler for mining web data (2017).

[8] K. Salim, F. Mohamed, Rcrawler: An r package for parallel web crawling and scraping (2017).

[9] M. Bajec, Štefan Furlan, A. Kumer, L. Šubelj, S. Žitnik, D. Lavbič, Supervizor (2011).

[10] D. Fišer, N. Ljubešić, T. Erjavec, Parlameter a corpus of contemporary slovene parliamentary proceedings (2018).

[11] Odprtokodni programski sistemi, https://www.goodfirms.co/blog/ (Nov. 2022).

[12] Textractor, https://opensource.com/article/18/7/textricator (Nov. 2022).

[13] Scrapy, https://scrapy.org/ (Nov. 2022).

[14] Tabula, http://webdata-scraping.com/extracting-data-pdfs-using-tabula/ (Nov. 2022).

[15] Kljucne besede, https://en.wikipedia.org/wiki/Keyword_extraction (Jan. 2023).

[16] Entitete, https://www.geeksforgeeks.org/named-entity-recognition/ (Nov. 2022).

[17] Sentimentalna analiza, https://dynamics.microsoft.com/en-us/ai/customer-insights/what-is-sentiment-analysis/ (Jan. 2023).

[18] Povezovanje entitet, https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/entity-linking/overview (Jan. 2023).

[19] Ekstrakcija razmerja, https://learn.microsoft.com/en-us/azure/cognitive-services/text-analytics-for-health/concepts/relation-extraction (Jan. 2023).

[20] Keyword extraction, https://paperswithcode.com/sota/keyword-extraction-on-semeval-2010-task-8 (Jan. 2023).

[21] Named entity recognition, https://paperswithcode.com/sota/named-entity-recognition-on-wnut-2017 (Jan. 2023).

[22] Sentiment analysis, https://paperswithcode.com/sota/sentiment-analysis-on-amazon-review-full (Jan. 2023).

[23] Entity linking, https://paperswithcode.com/sota/sentiment-analysis-on-amazon-review-full (Jan. 2023).

[24] Relation extraction, https://paperswithcode.com/sota/relation-extraction-on-ade-corpus (Jan. 2023).