Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Matic Jazbec

Analiza dimenzij kakovosti informacij spletnih strani slovenskih podjetij

diplomsko delo

univerzitetni študijski program
prve stopnje
računalništvo in informatika

Mentor: doc. dr. Dejan Lavbič

Ljubljana, 2016

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Podjetja na svetovnem spletu na različne načine uveljavljajo svojo prednost in razlike pred konkurenco. Zagotovo so kvalitetno predstavljene informacije, ki jih stranka o podjetju najde na svetovnem spletu, zelo pomembne in prispevajo pri odločanju med množico konkurenčnih možnosti. V okviru diplomske naloge se osredotočite na slovenska podjetja in analizirajte njihove lastne spletne strani, prisotnost na socialnih omrežjih ter ostale izbrane atribute kakovosti. Na podlagi množice merljivih podatkov določite dimenzije kakovosti informacij spletnih strani slovenskih podjetij in prisotnost na socialnih omrežjih ter poskušajte identificirati statistične razlike glede na velikost podjetja in dejavnost, ki jo opravljajo. Še posebej bodite pozorni na obstoj določenih trendov. V okviru diplomske naloge razvijte prototip orodja za pridobivanje podatkov iz javno dostopnih virov in izvedite analizo pridobljenih podatkov ter rezultate kritično ovrednotite.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
AJPES	The Agency of the Republic of Slovenia for Public Legal Records and Related Services	Agencija Republike Slovenije za javnopravne evidence in storitve
EFA	exploratory factor analysis	eksploratorna faktorska analiza
HTML	hypertext markup language	označevalni jezik za oblikovanje večpredstavnostnih dokumentov
HTTP	hypertext transfer protocol	protokol za prenos hiperteksta
IQ	information quality	kakovost informacij
LINQ	language integrated query	jezikovno vgrajeno povpraševanje
MSA	measure of sampling adequacy	mera ustreznosti vzorca
PCA	principal component analysis	analiza glavnih komponent
PRS	Slovenian Business Register	Poslovni register Slovenije
SKD	standard classification of activities	standardna klasifikacija dejavnosti
SQL	structured query language	strukturiran povpraševalni jezik
URL	uniform resource locator	enotni naslov vira
XML	extensible markup language	razširljivi označevalni jezik
XPATH	XML path annotation	jezik za naslavljanje delov dokumentov XML

Povzetek

Naslov: Analiza dimenzij kakovosti informacij spletnih strani slovenskih podjetij

Avtor: Matic Jazbec

Svetovni splet je v zadnjih letih postal eden najbolj priljubljenih medijev, preko katerega se podjetja predstavljajo in oglašujejo svoje izdelke ali storitve. Spletna predstavitev lahko prinese odločilno konkurenčno prednost, zato je pomembno, da je le-ta karseda kakovostna. Načinov za predstavitev je na spletu mnogo, od posameznih podjetij pa je odvisno, katere bodo uporabili in kako. Nekatera podjetja se odločijo za predstavitev na enem izmed spletnih portalov, specializiranih za povezavo med kupci in ponudniki storitev ali izdelkov, nekatera držijo stik s strankami preko socialnih omrežij, spet tretja postavijo svoje spletne portale in porabnikom ponudijo edinstveno izkušnjo ob stiku s podjetjem. Namen diplomske naloge je razvoj spletnega pajka, ki pridobi informacije s spletnih strani slovenskih podjetij in njihovih profilov na socialnih omrežjih, ter na omenjenih podatkih ovrednoti različne dimenzije kakovosti informacij. Rešitev je izdelana kot konzolna .NET aplikacija, ki uporablja knjižnice za pridobivanje spletnih vsebin z zahtevki HTTP in njihovo razčlenjevanje s tehnologijami, kot sta xPath in LINQ. Za shranjevanje podatkov je uporabljen strežnik Microsoft SQL Server. Na podlagi izmerjenih karakteristik spletnih strani je bila opravljena statistična analiza, kjer se izkaže, da se vrednosti atributov dimenzij kakovosti med slovenskimi podjetji statistično značilno razlikujejo glede na velikost podjetja in dejavnost, ki jo opravljajo.

Ključne besede: efektivnost spletnih strani, kakovost informacij, spletni pajki, luščenje podatkov s spleta, razčlenjevanje.

Abstract

Title: Information quality dimensions analysis of Slovenian companies' websites

Author: Matic Jazbec

Over the last decade the World Wide Web became the most popular media through which the companies introduce themselves and advertise their products or services. Web presentation can bring about the crucial competitive advantage, therefore, it is imperative that it is of the highest quality possible. There are various ways of presenting oneself on the internet and it depends on the individual companies which approach they will take. Some companies decide on the website portals' advertising specialized in the connection between the customer and provider of the service or product, others keep in contact with their clients via social media, and some create their own website portals and offer their users a unique contact experience. The purpose of this thesis is to develop a web crawler that would collect information from the websites of Slovenian companies and their profiles on social networks and apply a model for evaluating the quality of information. The solution is implemented as a .NET console application that uses HTTP client to obtain web content and query languages such as xPath and LINQ, to extract the relevant data. Extracted data is stored in Microsoft SQL Server. Statistical analysis is then conducted to identify the relations between measured features and join them into common information quality dimensions. It turns out that we can find and explain the variance in information quality among the websites of Slovenian companies based on their size and business type.

Keywords: web site effectiveness, information quality, web crawlers, web scraping, parsing.

Poglavje 1

Uvod

Svetovni splet je kot eden najpomembnejših delov interneta nepogrešljiv tudi v svetu trženja, saj s svojo dostopnostjo in številčno bazo uporabnikov ponuja podjetjem nove priložnosti za zaslužek. Kolikšno konkurenčno prednost si ustvarijo, pa je odvisno od posameznih podjetij in njihovih investicij v spletne predstavitve. Če je v začetku devetdesetih let veljalo, da že sama prisotnost na svetovnem spletu pomeni dodano vrednost za podjetje, je ta v prihodnjih letih z eksplozijo komercialnih spletnih strani popolnoma izginila [30]. Dodano vrednost za podjetje zato prinaša dostopnost spletne strani, kakovost informacij na njej in uporabniška izkušnja. Če je del uporabniške izkušnje odvisen tudi od percepcije uporabnika, so na drugi strani tudi elementi, ki odločno vplivajo na efektivnost spletne predstavitve in jih je mogoče pridobiti in ovrednotiti programsko.

V diplomskem delu bomo tako na podlagi prosto dostopnih podatkov merili kakovost informacij na spletnih straneh podjetij. Na kakovost informacij in podatkov pogosto gledamo kot na večdimenzijski koncept z mnogimi karakteristikami, ki pa so močno odvisne od konteksta, v katerem so te informacije objavljene [21]. V kontekstu komercialnih spletnih strani se pri izdelavi instrumenta zato osredotočamo na informacije, ki so ključne tako za uporabnika, ki išče določen izdelek ali storitev, kot za podjetje, ki ta izdelek ali storitev ponuja. Da lahko te informacije služijo kot mere za posamezno

dimenzijo kakovosti, jih je potrebno ustrezno pretvoriti v številske atribute - kvantificirati.

Podatke, ki bodo služili kot posamezni atributi dimenzij, bomo pridobili s pomočjo spletnega pajka, ki bo razčlenil podane spletne strani. Poleg razčlenitve lastnih strani podjetij, bo pajek zbral tudi informacije s Facebook profilov podjetij ter nekaterih spletnih portalov, ki se že ukvarjajo z oceno spletišč, in tako pridobil karseda celovito oceno predstavitve podjetja na spletu. Za izhodišče in primerjavo bodo uporabljeni podatki iz poslovnega registra Slovenije Agencije za javnopravne evidence in storitve, ki hrani podatke o obliki, dejavnosti in velikosti posameznih podjetij ter njihove kontaktne podatke, v kolikor se zastopnik podjetja strinja z njihovo objavo. Za primere, ko podjetje nima vpisanega spletnega naslova, bomo razvili algoritem za iskanje le-tega s spletnim brskalnikom. Vsi algoritmi za identifikacijo in luščenje podatkov s spletnih strani bodo implementirani v jeziku C# ter z uporabo poizvedovalnih jezikov xPath in LINQ, navigacija po spletu pa bo potekala preko HTTP odjemalca. Pridobljene podatke bomo nato obdelali in analizirali v okolju R.

Cilj diplomske naloge je z uporabo razvite programske opreme iz spleta izluščiti lastnosti spletnih strani, izdelati okvir za ocenjevanje njihove kakovosti in na vzorcu slovenskih podjetij preveriti njegovo napovedno moč. Pridobljene podatke bomo statistično analizirali in najprej opisali posamezne dimenzije kakovosti. Na koncu bomo na podlagi identificiranih dimenzij preverili, kako se kakovost spletnih strani razlikuje med dejavnostmi in velikostjo podjetij ter opisali statistično značilne razlike med skupinami. Razvito rešitev bo moč uporabiti tudi kot orodje za oceno lastne spletne strani in na podlagi rezultatov posameznih dimenzij identificirati možne izboljšave spletne predstavitve.

Poglavje 2

Raziskovalna področja

2.1 Kakovost informacij

Z eksplozijo svetovnega spleta je nastalo mnogo spletnih vsebin z velikim številom informacij, katerih kakovosti se medsebojno močno razlikujejo. Zato se je pojavila potreba po kriterijih za vrednotenje le-teh. Kakovost lahko vrednotimo z različnih vidikov, eni izmed njih so dostopnost, enostavnost uporabe in točnost informacij [17]. Pomembnosti te problematike so se kmalu začela zavedati tudi podjetja, ki so bodisi že investirala bodisi načrtovala investicijo v svojo spletno predstavitev in elektronsko poslovanje. Spletna stran, ki se izkaže kot zahtevna za uporabo in je ciljna publika ne sprejme, negativno predstavlja organizacijo in s tem slabša njen položaj na trgu [17].

V zadnjih letih je bilo veliko študij posvečenih prav oblikovanju spletnih strani za namene elektronskega poslovanja [40]. Največ pozornosti je bilo namenjene komercialnim oziroma poslovnim spletnim stranem, izobraževalnim stranem, spletnim stranem bank ter vladnih organizacij. Sprva so bile raziskave ozko usmerjene v iskanje ključnih prednosti spletnih strani za posamezno dejavnost podjetja, ali pa so upoštevale le del faktorjev, ki vplivajo na kakovost predstavitve. Kasneje so se na podlagi predhodnih raziskav [22, 29], začeli pojavljati tudi instrumenti za bolj celovito merjenje kakovosti ali teoretična ogrodja za izdelavo le-teh. Za merjenje kakovosti se uporabljajo trije

pristopi [16]:

- računalniško merjenje,
- mnenja ekspertov,
- mnenja uporabnikov.

Računalniško merjenje kakovosti uporablja programsko opremo za avtomatizirano zajemanje ključnih karakteristik spletnih strani. Običajno temelji na spletnem pajku, ki kakovost oceni tako na podlagi vsebine strani, kot na podlagi tehničnih lastnosti, ki jih razbere iz izvorne kode. Tovrsten pristop je uporaben pri analizi velikega števila strani, a po drugi strani iz svoje ocene popolnoma izpušča uporabnikovo dojemanje spletne strani [11].

Analize ekspertov se običajno začnejo z identifikacijo ključnih karakteristik strani določene domene. Nato pregledajo manjšo množico strani iz pripadajoče domene in jih ocenijo na podlagi prej razvitega okvirja za ocenjevanje. Rezultati tovrstnega dela so pogosto konceptualni okvirji za ocenjevanje kakovosti, ki so lahko uporabljeni na drugih primerih ali pa služijo kot podlaga za nadaljnje raziskovanje.

Zadnja možnost je, da za mnenje oziroma oceno vprašamo uporabnika spletne strani. Ta pristop, za razliko od zgornjih dveh, gleda na spletno stran z vidika uporabnika, ki igra ključno vlogo pri njenem uspehu. Za ta namen je bil razvit instrument WebQual [25], ki skozi vprašalnik od uporabnika pridobi oceno za 12 dimenzij kakovosti spletne strani:

- primernost informacij,
- komunikacija po meri,
- zaupanje,
- odzivnost,
- razumljivost,
- intuitivnost operacij,

- izgled,
- inovativnost,
- čustvena nota,
- konsistentna podoba,
- spletna celovitost,
- dodana vrednost k interakciji.

Podobno kot WebQual, tudi ostale instrumente sestavljajo dimenzije merjenja kakovosti. Tabela 2.1 predstavlja najbolj pogosto uporabljene dimenzije in njihove opise po Wang-u in Strong [38], razvrščene po pogostosti pojavitev.

V sodobnejši literaturi se namesto raziskav taksonomij kakovosti informacij, pogosto pojavlja diskusija o objektivnem in subjektivnem pogledu na kakovost informacij. S stališča objektivnega pogleda je kakovost neodvisna od opazovalca in situacije [18] ter definirana kot vrednost informacij, na voljo za izbran namen ali kot stopnja, do katere so informacije uporabne za določeno uporabnikovo aktivnost [12]. Običajno gre tu za specifikacije, namenjene optimalnemu shranjevanju podatkov na način, da se izognemo pomanjkljivostim pri predstavitvi informacij iz resničnega sveta v informacijskem sistemu [37]. Po drugi strani je subjektiven pogled na informacije ravno tako pomemben. Subjektiven vidik informacij je opredeljen kot odvisen od uporabnika in situacije in prikazan s podatki, ki z vidika nekoga ali nečesa ustvarijo razliko [18]. Iz te perspektive kakovost informacij predstavlja posameznikovo presojo o njihovi uporabnosti. Oba pogleda sta relevantna in upoštevana v literaturi o večkriterijskem odločanju, a avtorji s težavo potegnejo ločnico med njima [31].

Tabela 2.1 predstavlja zgolj konceptualne razlage dimenzij, od raziskovalca pa je odvisno, katere bo uporabil. Te morajo biti izbrane glede na kontekst generiranja in uporabe informacij, katerih kakovost ocenjujemo, saj

Dimenzija	Opis
Natančnost	V kolikšni meri so podatki pravilni, zanesljivi in zagotovo brez napak.
Konsistentnost	V kolikšni meri so informacije predstavljene v enaki obliki in v skladu s prejšnjimi podatki.
Varnost	V kolikšni meri je dostop do podatkov omejen na način, da zagotavlja varnost.
Pravočasnost	V kolikšni meri so informacije dovolj ažurne za dano nalogo.
Celovitost	V kolikšni meri so informacije prisotne ter podane dovolj široko in globoko za dano nalogo.
Jedrnatost	V kolikšni meri so informacije predstavljene kratko in jedrnato.
Zanesljivost	V kolikšni meri so informacije pravilne in zanesljive.
Dostopnost	V kolikšni meri so informacije lahko in hitro pridobljive.
Razpoložljivost	V kolikšni meri so informacije fizično dostopne.
Objektivnost	V kolikšni meri so informacije nepristranske.
Relevantnost	V kolikšni meri so informacije uporabne za dano nalogo.
Uporabnost	V kolikšni meri so informacije jasne in enostavne za uporabo.
Razumljivost	V kolikšni meri so podatki jasni, nedvoumni in lahko razumljivi.
Količina informacij	V kolikšni meri je količina podatkov, ki je na voljo, primerna.
Kredibilnost	V kolikšni meri veljajo podatki za resnične.
Navigacija	V kolikšni meri je podatke lahko najti in povezati.
Ugled	Kolikšen je ugled vira ali vsebine.
Učinkovitost	V kolikšni meri so podatki hitro zadostni za dano nalogo.
Dodana vrednost	V kolikšni meri podatki koristni in prinašajo prednosti.

Tabela 2.1: Pogosto uporabljene dimenzije kakovosti.

se s kontekstom spreminjajo tudi atributi kakovosti informacij [35]. Veliko opravljenih študij [26, 32, 36] obravnava ocenjevanje kakovosti informacij na spletu, zlasti na Wikipediji, saj je ta odprta in dostopna za vse. Članki so bili ocenjeni z dveh perspektiv - notranje, kjer so kot ocenjevalci nastopali uredniki člankov in zunanje, kjer so kakovost ocenjevali bralci. Uredniki člankov so skušali identificirati mere, ki članke ločijo med t.i. „izbranimi članki“ (članki, ki jih uredniki Wikipedije ocenijo kot najboljše) in tistimi, ki to niso. V sklopu raziskave [36] so bili identificirani novi, za Wikipedijo specifični atributi, ki vplivajo na oceno kakovosti, kot so število urejanj članka, število urednikov, zunanje povezave in dolžina članka. Sorodne raziskave so obogatile taksonomijo kakovosti informacij in prispevale k globljem razumevanju ocenjevanja kakovosti informacij [39]. Pri javnem pregledu (iz zunanje perspektive) so kot ocenjevalci pogosto nastopili študenti. V raziskavi, ki sta jo izvedla Arazy in Kopak [9], so obravnavali dimenzije kakovosti na treh naključno izbranih člankih. Rezultati študije so razkrili, da se bralci niso strinjali o kakovosti člankov, zaradi česar dimenzij kakovosti niso mogli izmeriti. Do tovrstnih ugotovitev so prišli tudi avtorji [39], saj so v raziskavi naleteli na dve različni interpretaciji števila urejanj posameznega članka. Nekateri bralci so to prepoznali kot pozitivno lastnost, saj naj bi urejanja bogatila vsebino članka, spet drugi so bili mnenja, da veliko število urejanj negativno vpliva na kakovost. Merjenje kakovosti na svetovnem spletu se torej izkaže za precejšen izziv, saj kot prvo ne obstaja standard za preverjanje kakovosti objavljene vsebine, kot drugo pa morajo uporabniki, ki iščejo in uporabljajo informacije sami presoditi, kakšna je njihova kakovost [34].

Ko raziskovalec identificira končnega uporabnika spletne aplikacije, mora glede na izbran pristop in domeno identificirati karakteristike spletne strani, ki jih bo ocenil in z njimi opisal posamezno dimenzijo. Dimenzijam je nato potrebno določiti prioriteto glede na njihovo pomembnost, nujnost in ceno. Zadnji korak pri izdelavi instrumenta je razvoj metrik za oceno izbranih dimenzij [22]. Kljub zadostni količini literature o kakovosti informacij, je večina raziskav teoretičnih in jim manjkajo metode ali vsaj predlogi za kvan-

tifikacijo predlaganih konceptov. Izziv, ki se ga lotevamo v tem diplomskem delu je torej poleg identifikacije ključnih lastnosti strani za oceno dimenzij kakovosti, tudi transformacija le-teh v tako obliko, da bodo primerne za algoritme strojnega učenja. Zhu in Gauch [41] sta za izboljšavo spletnega pajka, uporabljenega za iskanje informacij po spletu, uporabila mere kakovosti informacij na obravnavanih spletnih straneh. Za kvantificiranje karakteristik spletnih strani sta ubrala enostaven pristop. Ažurnost informacij sta npr. izmerila kot časovni žig zadnje spremembe dokumenta, popularnost kot število povezav, ki kažejo na obravnavano stran, razpoložljivost pa kot razmerje med nedelujočimi in vsemi povezavami na strani. Podoben problem sta obravnavala tudi Eppler in Muenzenmayer [13], ki sta objavila seznam dimenzij (glej tabelo 2.2) kakovosti informacij ter za vsakega podala način meritve in predlagano orodje.

Iz tabele 2.2 je razvidno, da nekaterih dimenzij vseeno ni mogoče dobro oceniti z uporabo programske opreme in je potrebno sodelovanje uporabnika. Pri pristopu, ki smo ga izbrali v diplomskem delu se torej pojavijo težave z obravnavo dimenzij, vezanih na uporabnikovo dojemanje strani, saj zaradi obsega raziskave anketiranje ne bi bilo učinkovito. Načela, da moramo pri oceni upoštevati tudi uporabnikov pogled, se bomo vseeno držali in obravnavali tudi profile podjetij na socialnih omrežjih, kjer je razvidno agregirano mnenje uporabnikov. Opazimo tudi, da je poleg same vsebine spletnih strani pomembna tudi infrastruktura, uporabljene tehnologije in implementacija spletne aplikacije, saj lahko le-ta močno vpliva na uporabniško izkušnjo z vidika hitrosti, dostopnosti, priročnosti idr. Te dimenzije pogosto uvrščamo pod kakovost medija [13] ali kakovost storitve [19].

2.2 Luščenje podatkov s spleta

2.2.1 Ekstrakcija podatkov

Potrebne podatke za oceno kakovosti spletne strani bomo pridobili z ekstrakcijo podatkov s spleta. Ekstrakcija podatkov je proces pridobivanja rele-

Dimenzija	Spletni indikator	Orodje za merjenje
Dostopnost	število nedelujočih povezav	analizator strani
Konsistentnost	število strani s stilskim odstopanjem	analizator strani
Pravočasnost	število (pre)velikih strani ali datotek, z dolgim časom nalaganja	analizator strani
Jedrnatost	število strani globoko v hierarhiji	analizator strani
Vzdrževanost	število strani z manjkajočimi meta-podatki	analizator strani
Aktualnost	čas od zadnje spremembe vsebine strani	analizator strani
Uporabnost	število neobiskanih ali nepovezanih strani, ocena uporabnika	analizator strani in prometa, raziskava med uporabniki
Priročnost	zahtevnost navigacije - število izgubljenih sledi pri navigaciji	analizator prometa, orodja za spletno rudarjenje
Hitrost	strežniški in omrežni odzivni čas	spremljanje strežnikov in omrežja, analizator strani
Celovitost	ocena uporabnika	raziskava med uporabniki
Jasnost	ocena uporabnika	raziskava med uporabniki
Natančnost	ocena uporabnika	raziskava med uporabniki
Sledljivost	število strani brez navedenega avtorja ali vira	analizator strani
Varnost	število šibkih prijav, privzetih poverilnic	analizator strani, pregledovalnik vrat (ang. port scanner)
Pravilnost	ocena uporabnika	raziskava med uporabniki
Interaktivnost	število obrazcev in personaliziranih strani	analizator strani

Tabela 2.2: Kvantitativne mere posameznih dimenzij in predlagana orodja za merjenje.

vantnih podatkov iz običajno nestrukturiranih ali slabo strukturiranih virov (med katere spada tudi svetovni splet) za potrebe nadaljnje obdelave ali analize. Če se je zgodovinsko s pojmom ekstrakcija podatkov ciljalo predvsem na spremembe formatov podatkov zaradi strojne opreme, se danes ta pojem uporablja predvsem za ekstrakcijo iz nestrukturiranih virov in drugačnih formatov v programski opremi. Proces ekstrakcije podatkov s svetovnega pogosto najdemo tudi pod izrazom luščenje podatkov (ang. Web Scraping).

Pojem luščenje podatkov s spleta je tesno povezan s spletnim indeksiranjem, ki uporablja spletnega pajka za pregledovanje spletnih strani. Spletno indeksiranje v veliki večini uporabljajo spletni brskalniki, medtem ko je luščenje podatkov precej bolj splošno uporabno. Luščenje podatkov povezujemo tudi z avtomatizacijo spleta (ang. Web Automation) [8], kjer programska oprema simulira človeško brskanje po spletu za namen testiranja ali pridobivanja informacij. Druge uporabe luščenja podatkov s spleta med drugim vključujejo zaznavo sprememb na spletni strani, primerjavo cen, iskanje podatkov o kontaktih in integracijo spletnih podatkov.

Čeprav je danes večina informacij na spletu v obliki delno strukturiranih HTML dokumentov, njihova ekstrakcija ni trivialna. HTML dokumenti so namreč včasih napisani ročno, včasih generirani z namensko programsko opremo, v obeh primerih pa pogosto pride do napak nepravilno oblikovanih dokumentov. Kljub predpisani strukturi tudi pravilno oblikovani HTML dokumenti niso idealni, saj so primarno namenjeni opisu vizualne predstavitve podatkov na strani in ne programski ekstrakciji [28]. Pri luščenju naletimo tudi na prepreke v obliki težje dostopnih podatkov, objavljenih na primer v obliki slike ali Flash aplikacije, ali v obliki sistemov za preprečevanje avtomatiziranega brskanja, kot je CAPTCHA. Vse od naštetih se seveda da zaobiti s primernimi algoritmi, vendar je pri tem potrebno upoštevati pravne vidike. Legalnost luščenja podatkov se po svetu namreč precej razlikuje, v splošnem pa je lahko v nasprotju s pogoji uporabe posameznih spletnih strani, a je njihova sodna izvršljivost nejasna.

2.2.2 Tehnike luščenja

Glede na strukturo vira in razpoložljivo opremo, lahko izbiramo med več tehnikami luščenja podatkov. Trenutne rešitve [1, 4] rangirajo od ad-hoc rešitev, ki zahtevajo sodelovanje uporabnika, do popolnoma avtomatiziranih komercialnih sistemov za pridobivanje podatkov.

Luščenje z uporabnikom

V nekaterih primerih programska oprema ne more nadomestiti uporabnika bodisi zaradi sistemskih omejitev (blokiran dostop) bodisi zaradi potrebe po pregledu in preverbi podatkov. V tem primeru je edina rešitev, da podatke izlušči kar uporabnik in jih ročno pretvori v zahtevano strukturo. Tovrsten pristop je smiseln tudi v primeru, da je čas, ki ga uporabnik skupno porabi za luščenje bistveno krajši od časa, potrebnega za implementacijo namenskega algoritma.

Regularni izrazi

Med bolj enostavne pristope spada luščenje z uporabo regularnih izrazov. Regularni izrazi (ang. regular expression, regex) so posebne kombinacije znakov, ki se uporabljajo za napredno iskanje ujemanj v besedilih. Prednost tega pristopa je, da uporabnik ne potrebuje programerskega ozadja, saj je podpora za regularne izraze vgrajena že v nekaterih operacijskih sistemih in urejevalnikih besedil. Podprti so tudi v večjem delu programskih jezikov in jih je zato moč uporabiti tudi v kombinaciji z naprednejšimi tehnikami.

HTTP zahtevki

Izvorno kodo spletnih strani programsko običajno pridobimo s pošiljanjem HTTP zahtevkov. Ti so pogosto enostavni GET zahtevki, ki v odgovoru vrnejo HTML kodo zahtevane strani. V primerih, da spletna stran s katere želimo izluščiti podatke, zahteva avtentikacijo uporabnika, ali pa uporablja mehanizme za validacijo zahtevkov, pa dostop do njih ni tako trivialen. V tem

primeru mora implementator preučiti postopek avtentikacije in mehanizmov v ozadju in s svojimi zahtevki posnemati delovanje spletnega brskalnika. Z opazovanjem HTTP zahtevkov in odgovorov pri nalaganju podatkov na spletni strani se pogosto odkrijejo tudi enostavnejši načini pridobitve le-teh od razčlenjevanja HTML-ja. Vsebina se namreč iz podatkovnega vira na stran pošlje preko ločenih zahtevkov, ki podatke vračajo bolj strukturirano, kot so kasneje prikazani - najpogosteje v formatu JSON ali XML.

Razčlenjevanje HTML

Najbolj pogost način pridobivanja informacij s spleta je razčlenjevanje HTML-ja strani. Za ta namen obstajajo mnoge knjižnice, ki nad HTML dokumenti implementirajo različne poizvedovalne jezike, kot sta xPath in xQuery. Največji problem kontinuiranega razčlenjevanja spletnih strani je dinamičnost spleta, saj se struktura HTML dokumenta pogosto nenapovedano spremeni in podatek, ki ga iščemo, ni več na voljo z isto poizvedbo. Zato moramo paziti, da lokacijo podatka v HTML dokumentu definiramo čim manj eksplicitno in čim bolj robustno, sicer vsaka sprememba na spletni strani terja tudi popravke v programski kodi.

Razčlenjevanje DOM-a

V primeru, da se vsebina strani generira dinamično na strani odjemalca, leta ne bo vidna v HTML kodi. V tem primeru nam pomagajo knjižnice za vključitev spletnega brskalnika kot sta Mozilla Firefox ali Internet Explorer v aplikacijo. Za, v vključenem brskalniku, prikazano stran se namreč ustvari objektni model dokumenta (ang. Document Object Model, DOM), ki elemente HTML dokumenta predstavi kot drevesno strukturo. V drevesu je vsako vozlišče predstavljeno kot objekt in kot tako dostopno tudi programsko.

Namenska programska oprema

Na voljo je veliko orodij, namenjenih luščenju podatkov, kot so Mozend a [4], Webhose.io [7], Visual Scraper [6] idr. Delujejo tako, da samodejno prepoznajo strukturo spletne strani ali pa uporabniku ponudijo vmesnik, preko katerega označi, kateri podatki na spletni strani ga zanimajo in na podlagi tega generirajo skripto.

Prepoznava semantičnih oznak

Luščenje nam lahko olajšajo metapodatki ali semantične oznake na spletni strani. Če so semantične oznake vdelane v strani, lahko na to tehniko gledamo kot poseben primer razčlenjevanja objektnega modela dokumenta. V drugih primerih so semantične oznake lahko organizirane v semantični sloj in shranjene ločeno od dejanskih spletnih strani, da programi za luščenje podatkov pridobijo podatkovno shemo in navodila s tega sloja pred dejansko ekstrakcijo s strani [27].

Računalniški vid

Raziskave se ukvarjajo tudi z razvojem algoritmov, ki bi z uporabo računalniškega vida in strojnega učenja prepoznale in izluščile informacije iz spletnih strani z vizualno interpretacijo le-teh - enako, kot to počne človek.

V diplomskem delu bomo od zgoraj naštetih tehnik uporabljali pošiljanje in manipulacijo HTTP zahtevkov za dostop do vsebine ter razčlenjevanje HTML-ja strani tako z uporabo poizvedovalnih jezikov, kot z uporabo objektnega modela dokumenta. V nekaterih funkcijah si bomo pomagali tudi z regularnimi izrazi.

2.3 Metode analize podatkov

Kot rezultat avtomatiziranega zbiranja podatkov z uporabo spletnega pajka bomo podatke predstavili v obliki matrike, kjer vrstice predstavljajo posame-

zna podjetja, stolpci pa izmerjen atribut kakovosti. Za analizo pomembnosti izmerjenih atributov in korelacij med njimi, bomo izvedli faktorsko analizo. Ko ugotovimo, katere kombinacije atributov kakovosti najbolje opišejo naš nabor podatkov, pa bomo poskusili poiskati še statistično značilne razlike med skupinami podjetij različnih velikosti in dejavnosti.

2.3.1 Faktorska analiza

Faktorska analiza je metoda za redukcijo podatkov. Gre za niz matematično-statističnih postopkov, katerih cilj je zmanjšati število medsebojno povezanih spremenljivk. Pridobljen manjši nabor spremenljivk, imenovanih tudi faktorji, pojasnjuje medsebojno povezanost spremenljivk, ki ga tvorijo [15]. Ker dimenzij kakovosti informacij ne moremo neposredno izmeriti, bomo vsako dimenzijo opisali z nekaj izmerljivimi atributi, nato pa s faktorsko analizo poizkusili odkriti povezave med izbranimi atributi in jih združiti v smiselne dimenzije. Skupne, nemerljive spremenljivke imenujemo tudi latentne spremenljivke. Faktorska analiza je povezana z analizo glavnih komponent, katere cilj je določiti čim manjše število linearnih kombinacij spremenljivk tako, da z njimi opišemo čim večji del celotne variance. Kljub povezavi pa metodi nista identični [10]. V preteklosti je bilo že precej razprav na temo razlik med omenjenima tehnikama, jasno pa je, da analiza glavnih komponent predstavlja manj kompleksno različico eksploratorne faktorske analize. Postopek faktorske analize se običajno začne z raziskavo medsebojnih povezanosti med opazovanimi spremenljivkami. V ta namen uporabimo koeficient korelacije kot mero povezanosti in pripravimo tabelo korelacij med spremenljivkami. Iz matrike lahko opazimo razmerja med množicami obravnavanih spremenljivk in ocenimo, ali je podlaga primerna za nadaljevanje raziskave. K analizi lahko pristopimo na dva načina. Prvi je že omenjena eksploratorna analiza (EFA), katero izberemo, če nimamo predstave o tem, koliko faktorjev tvorijo obravnavani atributi. Kot rezultat skušamo iz celotnega nabora spremenljivk na podlagi kovariance ugotoviti manjši nabor hipotetičnih (latentnih) faktorjev. Drugi, potrditveni pristop raziskovalec uporabi, če razpolaga z izdelano

teorijo o raziskovanem problemu in želi preveriti pravilnost svojih hipotez. Raziskovalec lahko npr. predvideva, da nabor opazovanih spremenljivk v bistvu temelji na dveh dimenzijah ter, da nekatere spremenljivke pripadajo eni, nekatere pa drugi dimenziji. Če je faktorska analiza namenjena testiranju hipotez in ne raziskovanju skupnih dimenzij, se imenuje potrditvena faktorska analiza [14]. Poleg različnih tipov analize, obstaja tudi več načinov ekstrakcije faktorjev. V eksploratorni faktorski analizi se običajno uporablja že omenjena analiza glavnih komponent (PCA), kjer se uteži faktorjev izračunajo na način, da pojasnijo največjo možno varianco. Za določitev najbolj primernega števila faktorjev obstaja mnogo metod, njihovi rezultati pa se običajno med seboj razlikujejo. Med enostavnejšimi je Kaiserjev kriterij [20], ki določa, da obdržimo le faktorje z lastno vrednostjo večjo od 1, saj je pri manjših lastnih vrednostih delež variabilnosti premajhen. Uporaba tega kriterija kot edinega za določitev števila faktorjev je sicer odsvetovana, saj pogosto izbere preveč faktorjev [23]. Drug enostaven kriterij je izbiranje faktorjev glede na odstotek pojasnjene variance. Raziskovalci se običajno zadovoljijo s številom faktorjev, ki skupno pojasni 80% - 90% variance. Število faktorjev pa lahko določimo tudi vizualno, s testom drobirja (ang. scree test). Na x os grafa umestimo faktorje, na y pa pripadajoče lastne vrednosti, urejeno padajoče po lastnih vrednostih. Kriterij pravi, da opustimo vse faktorje, ki se na grafu pojavijo desno od „kolena“, kjer začne krivulja padati bolj položno. Faktorski model je potrebno z analizo še rotirati. V nerotirani matriki je namreč najbolj uteženih prvih nekaj faktorjev, običajno pa posamezen element utežuje več kot en faktor, kar otežuje interpretacijo. Da bi bili rezultati bolj razumljivi torej izvedemo rotacijo matrike, s čimer skušamo priti do vzorca uteži, kjer posamezni elementi močno utežujejo le po en faktor. Največkrat se v EFA uporabi VARIMAX rotacija, ki maksimizira varianco kvadratov uteži v vsakem faktorju. Na ta način dobimo faktorje, ki jih posamezna spremenljivka bodisi močno bodisi zelo šibko utežuje, kar olajša povezavo spremenljivke z enim samim faktorjem [33]. Če želimo začetni nabor spremenljivk izraziti z novimi faktorji, moramo izračunati fak-

torske vrednosti. Če matrika X predstavlja standardizirane originalne vrednosti spremenljivk in je B matrika faktorskih uteži, je $C = X \times B$ matrika standardiziranih faktorskih vrednosti. Z drugimi besedami, standardizirane vrednosti spremenljivk pomnožimo s faktorskimi utežmi in jih seštejemo po posameznih faktorjih.

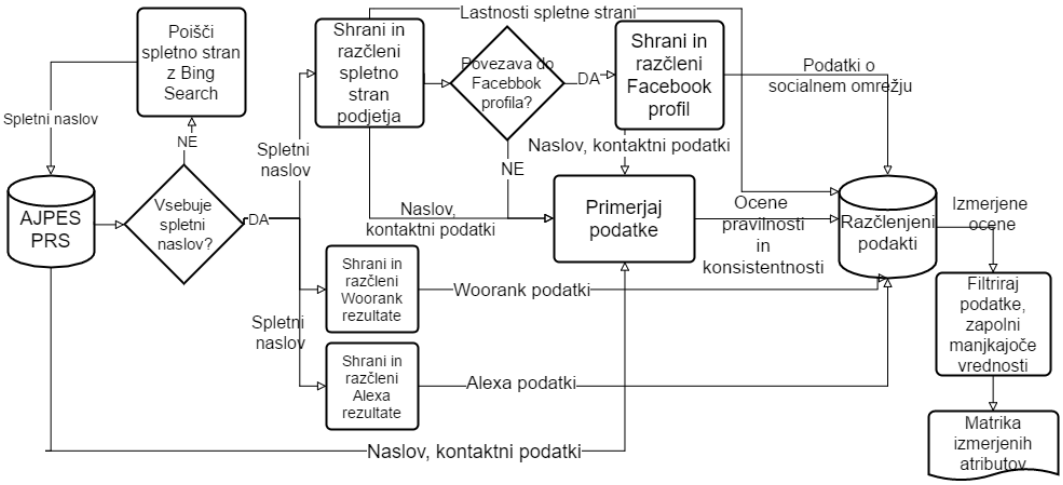
Poglavje 3

Orodje za merjenje kakovosti informacij

3.1 Ideja

V zadnjih 20 letih se je svetovni splet (ang. World Wide Web, WWW) uveljavil kot ključno okolje za upravljanje, izmenjavo in objavo informacij. Poslovno priložnost so začutile tudi organizacije, ki so preko tega medija želele čim bolje predstaviti tako sebe kot svoje izdelke in storitve. Z eksplozijo spleta je močno naraslo tudi število komercialnih spletnih strani, katerih kakovost pa je zaradi pomanjkanja izvršljivih standardov pogosto vprašljiva. Ideja je izdelati aplikacijo, ki bo znala oceniti kakovost spletnih strani in informacij na le-teh, za podan nabor podjetij. Aplikacija bo iz baze podatkov prebrala ime podjetja in spletni naslov ter na podlagi tega skušala preveriti kakovost spletne predstavitve podjetja na spletu. Za primere, ko v izvornih podatkih ne bo podanega naslova spletne strani podjetja, bomo implementirali tudi algoritem za iskanje le-tega z uporabo spletnega brskalnika. V primeru, da program na lastni spletni strani podjetja najde tudi naslov Facebook profila, bo razčlenil tudi tega. Spletni naslov podjetja bo kot parameter podan še na nekatere spletne portale, ki se ukvarjajo z ocenjevanjem spletnih strani. Tako bomo pridobili še nekatere informacije o spletnih straneh,

ki so sicer težje izmerljive. Prebrani podatki bodo brez obdelave shranjeni v podatkovno bazo. Z uporabo enostavne aplikacije bomo nato pridobljene podatke prebrali iz podatkovne baze, jih združili v smiselne mere kakovosti, kvantificirali in shranili v format, primeren za analizo - datoteko csv. Podatke v tej datoteki bomo podrobno analizirali in v kombinaciji s splošnimi podatki o podjetjih skušali najti zakonitosti in vzorce. Opisan podatkovni tok je grafično prikazan na sliki 3.1, posamezni podatkovni viri in implementacija razčlenjevalnikov pa so podrobneje razloženi v poglavjih 3.2 in 3.4.



Slika 3.1: Podatkovni tok in vključeni procesi.

3.2 Viri podatkov

Za sestavo karseda celovite ocene predstavitve podjetja na spletu, bomo podatke pridobili z več virov. Struktura in vsebina posameznih virov je opisana v tem poglavju.

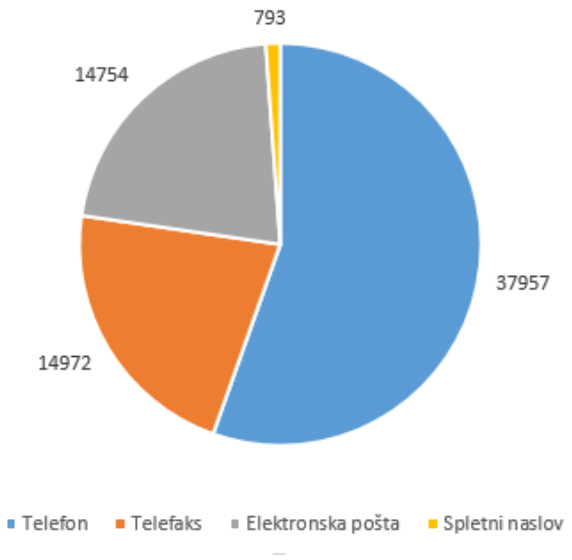
3.2.1 Poslovni Register Slovenije, AJPES

Za izhodišče bomo uporabili Poslovni Register Slovenije, ki ga je v obliki Access datoteke posredovala Agencija za javnopravne evidence in storitve Republike Slovenije - AJPES. Poslovni register Slovenije (PRS) je osrednja baza podatkov o vseh poslovnih subjektih s sedežem na območju RS, ki opravljajo pridobitno ali nepridobitno dejavnost, o njihovih delih in o podružnicah tujih podjetij. V Poslovnem registru Slovenije je bilo z avgustom 2016 vpisanih približno 207.000 poslovnih subjektov, ki opravljajo registrirane dejavnosti ali s predpisom oziroma aktom o ustanovitvi določene dejavnosti [5]. V tej raziskavi smo se omejili zgolj na podjetja, ki po pravnoorganizacijski obliki spadajo med delniške družbe (d.d.) ali družbe z omejeno odgovornostjo (d.o.o.), saj bi sicer bila ta preobsežna. Prav tako smo se omejili na podjetja, za katera ima AJPES evidentiran vsaj en kontaktni podatek (telefon, naslov elektronske pošte ali spletna stran) in ga sme objaviti. Razlog za to omejitev je, da brez kontaktnih podatkov ne moremo preveriti pravilnosti nobenega kontaktnega podatka, ki ga je podjetje objavilo na spletni strani ali socialnem omrežju in zato ocena takega spletišča ni smiselna. Celoten nabor smo tako omejili na 34.456 podjetij, delež kontaktnih podatkov, s katerimi razpolaga AJPES, pa je prikazan na sliki 3.2. Delež podjetij, ki so vpisala oziroma dovolila objaviti svoj spletni naslov je zelo majhen, približno 1%, kar pomeni, da se bomo pri veliki večini morali zanesti na rezultate spletnega iskalnika. Poleg kontaktnih podatkov nas bodo zanimali še naslednji podatki o posamezni enoti:

- matična številka,
- popolno ime,
- kratko ime,
- oblika,
- naslov,

- pošta in poštna številka,
- velikost po EU klasifikaciji,
- dejavnost po SKD klasifikaciji.

Velikost in dejavnost sta vključena zgolj za namene analize rezultatov in ne bosta vključena v algoritmih za ocenjevanje kakovosti informacij. Za vse podatke, ki jih je posredoval AJPES bomo privzeli, da so ažurni in pravilni in jih bomo kot take uporabili kot podlago za preverjanje pravilnosti podatkov, objavljenih na drugih straneh.



Slika 3.2: Struktura objavljenih kontaktnih podatkov.

3.2.2 Lastna spletna stran podjetja in Facebook profil

Kot osrednji predmet analize bo obravnavana lastna spletna stran vsakega podjetja. Do strani bomo dostopali preko povezave v PRS ali povezave, ki jo bo kot primerno vrnil algoritem za iskanje z uporabo spletnega iskalnika (delovanje algoritma je podrobneje predstavljeno v poglavju 3.4.2). Na spletni

strani bo program poiskal in primerjal kontaktne podatke, naslov ter povezave do profilov na socialnih omrežjih. Preveril bo tudi prisotnost obrazca za pošiljanje pošte in zemljevida, zabeležil nivo na katerem so objavljene ključne informacije ter preveril stilske lastnosti strani. Med najdenimi povezavami do profilov na socialnih omrežjih, bo pajek obiskal Facebook profil podjetja, saj so ostali (Google Plus, Twitter, LinkedIn) premalo razširjeni. Za oceno aktivnosti podjetja na socialnem omrežju, nas bo zanimal čas zadnje objave, za oceno priljubljenosti med uporabniki pa ocena profila in število prejetih ocen. Iz profila bomo izluščili tudi kontaktne podatke in naslov podjetja in jih primerjali s tistimi iz AJPES-a.

3.2.3 Alexa.com

Alexa je spletni portal z orodji za digitalni marketing, s katerimi lahko uporabniki analizirajo prednosti in slabosti svojih spletnih strani. Ponuja analizo optimizacije za spletne brskalnike ter pogled v demografijo obiskovalcev spletne strani in prometa. Za dostop do vseh funkcionalnosti portal sicer ponuja plačljiv dostop, a se bomo za potrebe diplomskega dela zadovoljili z omejeno, a brezplačno različico orodja Site Overview. Gre za enostavno orodje, ki od uporabnika zahteva le vnos povezave do spletne strani, katero želi analizirati in prikaže stran z rezultati. Iz portala Alexa bomo poleg nekaj tehničnih podatkov izluščili podatke o obnašanju uporabnika na strani:

- stopnja odboja (ang. Bounce Rate) - število uporabnikov, ki spletišče po ogledu naslovne strani zapustijo,
- število ogledov na obiskovalca,
- čas na strani,
- čas nalaganja strani,
- velikost strani,
- odstotek obiskovalcev, ki pridejo na stran preko spletnega iskalnika,

- država registracije domene.

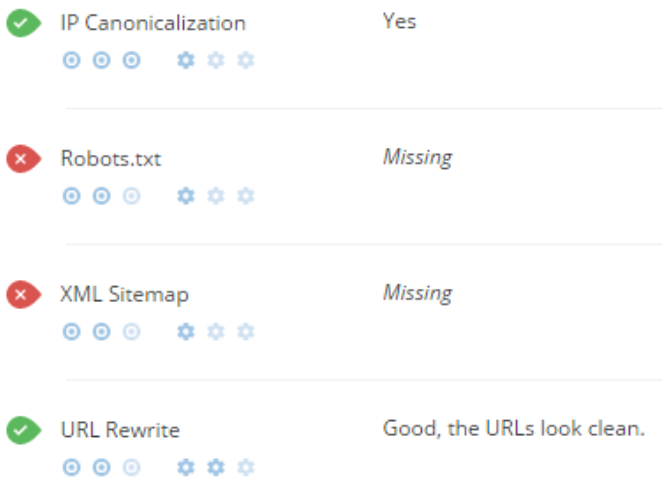
3.2.4 Woorank.com

Woorank je po obliki in namembnosti zelo podoben Alexi, le, da ta ponuja manj demografskih podatkov in bistveno več tehničnih informacij o spletišču. Za vsak podatek je podana tudi informacija o stopnji vpliva na kakovost strani ter o zahtevnosti reševanja problema, v primeru, da ta obstaja. Razdeljena je na tri večje sklope, optimizacija, promocija in meritve, vsak od teh pa je razdeljen na več manjših. Optimizacija tako ponuja podatke o optimizaciji strani za spletne iskalnike, mobilne naprave in uporabljene spletne tehnologije, promocijski sklop pa na primer ponuja podatke o povratnih povezavah (straneh, ki kažejo na obravnavano spletno stran) ter socialnih omrežjih. Portal Woorank objavlja tudi podatke storitve Web of trust[1] in podajajo oceno o tem, kako zaupanja vredna je spletna stran. Ker za veliko obravnavanih atributov stran ponuja le tekstovni opis (glej sliko 3.3), iz katerega ne moremo sestaviti številske ocene, bomo za te shranili zgolj vrednost 1, če je atribut ocenjen pozitivno in 0, če zahteva popravke. Vrednosti bomo nato po sklopih sešteli in dobili smiselne ocene za vsakega od njih. Celovito oceno spletišča na podlagi števila uspešnih testov, števila atributov, ki zahtevajo popravke in števila napak, sestavi tudi Woorank.

3.3 Arhitektura informacijske rešitve

Aplikacija za zajem in ekstrakcijo podatkov je implementirana v okolju .NET. Osrednji del je namizna konzolna C# aplikacija, ki z uporabo knjižnjic za podporo protokola HTTP in razčlenjevanje HTML dokumentov skrbi za shranjevanje spletnih strani in njihovo razčlenjevanje. Sestavljena je iz jedra, ki skrbi za komunikacijo s podatkovno bazo in datotečnim sistemom, ter razčlenjevalnikov posameznih strani (glej sliko 3.4). V jedru aplikacije se

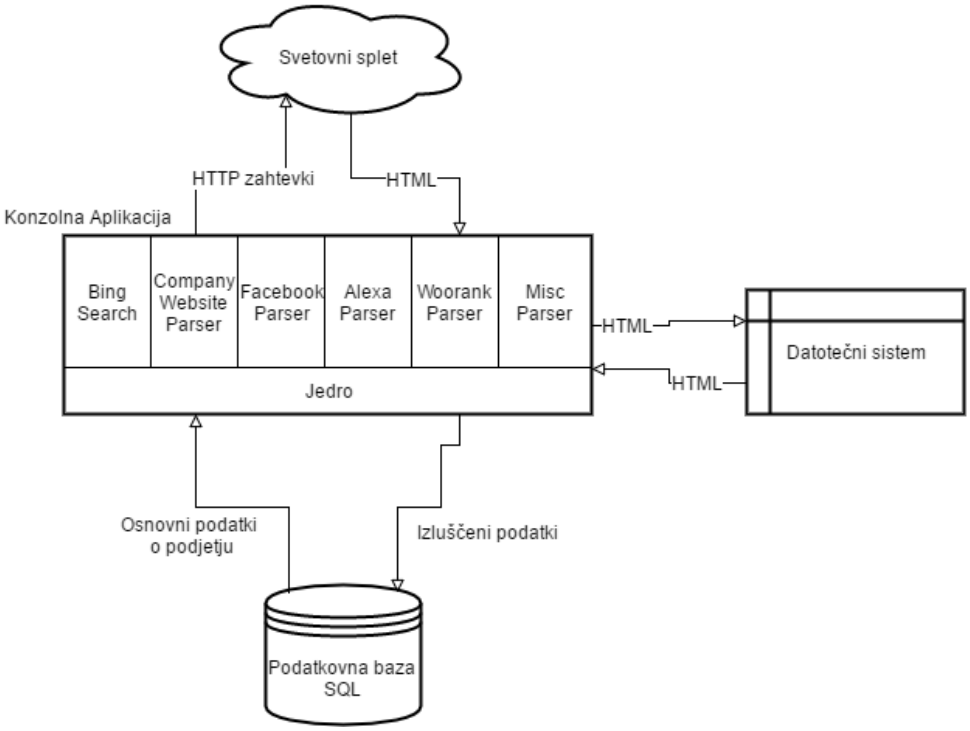
[1] https://www.mywot.com/



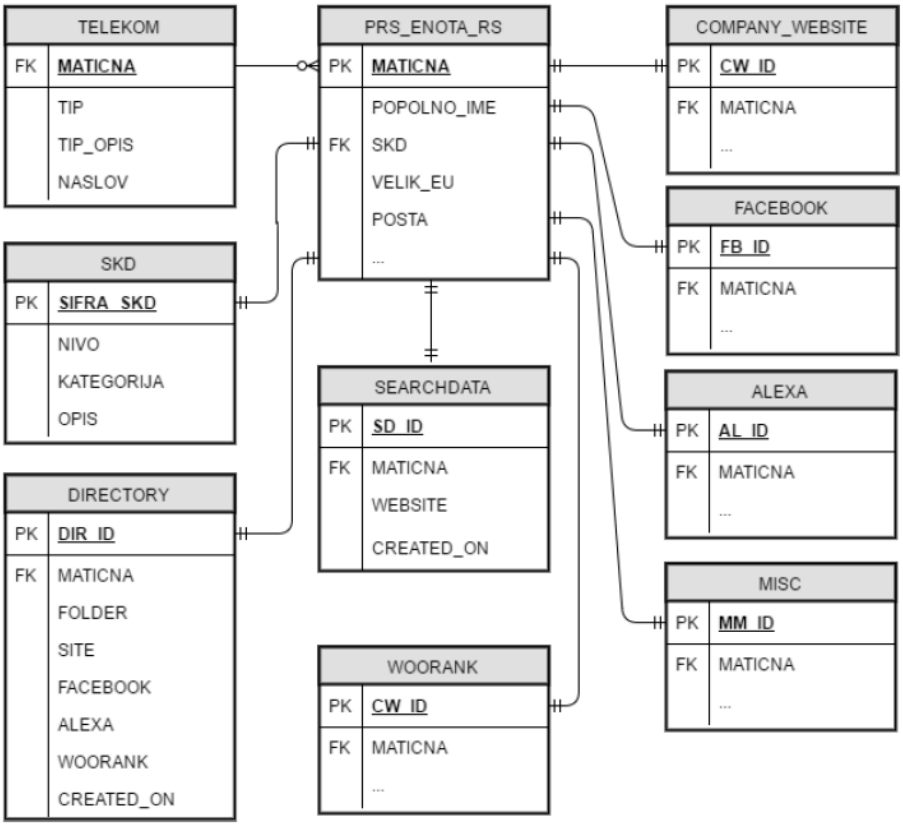
Slika 3.3: Izpis informacij o atributih spletne strani na portalu Woorank.

torej za posamezno podjetje iz podatkovne baze prebere informacija o njegovi spletni strani ter direktoriju, kjer so, ali bodo v prihodnosti, shranjene datoteke z izvorno kodo strani. Jedro poskrbi za klice razčlenjevalnikov (ang. parsers) posameznih virov, ki s prej opisanimi tehnikami pridobijo izvorno kodo strani s spleta (v kolikor te še ni na disku) in iz nje izluščijo relevantne informacije. Pod razčlenjevalnike štejemo tudi razred, ki skrbi za iskanje spletnih naslovov podjetij, ki niso objavljeni v PRS in jih posledično nimamo v podatkovni bazi. Čeprav ta ne vrača podatkov, ki bodo neposredno vključeni v oceno spletišča, je po strukturi in uporabi enak ostalim razčlenjevalnikom - ob klicu iz jedra aplikacije pošlje zahtevek na spletni iskalnik, ovrednoti rezultate in v primeru uspeha vrne rezultat, ki se zapiše v podatkovno bazo. Za podatkovno bazo smo uporabili Microsoft SQL Server, kjer podatkovni model sestavlja deset tabel, predstavljenih na sliki 3.5. Tabeli PRS_ENOTA_RS in TELEKOM sta del izpisa širšega nabora podatkov iz Poslovnega registra Slovenije, ki jih je dne 4.7.2016 posredoval AJPES. Tabela PRS_ENOTA_RS vsebuje osnovne podatke o poslovnih subjektih, TELEKOM pa kontaktne podatke. Prav tako je AJPES vir šifranta standardne klasifikacije dejavnosti (SKD), ki vsebuje opise in šifre dejavnosti podjetij. Omenjene tabele

predstavljajo izhodiščne podatke za analizo, medtem ko so ostale namenjene hranjenju rezultatov in poti do shranjenih datotek. V tabelo DIRECTORY se ob prenosu na disk zapišejo poti do shranjenih datotek za posamezno podjetje, razčlenjevalniki pa na podlagi teh zapisov datoteke odprejo in izluščijo podatke. SEARCHDATA je namenjena hranjenju spletnih naslovov podjetij, ki jih pajek pridobi s pomočjo spletnega iskalnika, v ostalih tabelah pa so podatki, izluščeni iz virov. Med njimi je tudi tabela MISC, ki hrani ocene, ki niso direktno povezane s samo enim virom, temveč so nastale kot rezultat izračunov na podlagi več virov. Tak podatek je na primer stopnja pravilnosti naslova, ki upošteva tako podatke, izluščene iz spletne strani podjetja, kot tiste iz Facebook profila.



Slika 3.4: Arhitektura aplikacije



Slika 3.5: Model podatkovne baze

3.4 Tehnične podrobnosti implementacije

V tem poglavju se bomo posvetili implementaciji pajka in tehničnim podrobnostim ekstrakcije podatkov iz posameznih virov.

3.4.1 Jedro aplikacije

V jedro aplikacije spadajo razredi s skupnimi metodami za delo s podatkovno bazo ter datotečnim sistemom, ki se uporabljajo ob vsakem zajemu podatkov. Za programsko interakcijo s podatkovno bazo SQL je uporabljena tehnologija LINQ to SQL. Gre za komponento ogrodja .NET, ki ponuja vmesno ogrodje za preslikavo objektov relacijske podatkovne baze v objekte. Tako lahko

razvijalec upravlja s podatki iz podatkovne baze v izbranem podprtem programskem jeziku. Ob klicu se koda preslika v SQL ukaze, ko podatkovna baza vrne rezultate, pa so ti znova preslikani v objekte. Na drugi strani za interakcijo s svetovnim spletom preko protokola HTTP skrbi razred HttpClient, ki podpira asinhrono pošiljanje in prejemanje poljubnih HTTP zahtevkov, za razčlenjevanje HTML dokumentov pa bomo uporabili knjižnico HtmlAgilityPack [3]. Jedro programa je tako precej enostavno - iz podatkovne baze se z uporabo LINQ-a preberejo matične številke in imena podjetij za obravnavo, opcijsko tudi spletna stran. V veliki večini primerov, ko podatka o spletni strani nimamo, se najprej izvede iskanje le-te z uporabo spletnega iskalnika Bing Search. Implementacija tega bo podrobneje opisana v naslednjem poglavju. Če iskalnik spletno stran najde, se izvajanje nadaljuje s pridobivanjem podatkov iz posameznih virov, sicer tega podjetja ne moremo obravnavati in bo izločeno iz raziskave.

3.4.2 Bing Search

Zaradi pomanjkanja informacij o spletnih naslovih v Poslovnem registru Slovenije, se je pred začetkom ekstrakcije podatkov o spletnih straneh pojavila potreba po implementaciji razčlenjevalnika in algoritma za iskanje le-teh. Ker iskalnik Google zazna, da zahtevke pošilja programska oprema in ne človeški uporabnik, to ustavi z mehanizmom Captcha. Mehanizem zahteva, da uporabnik prepozna in prepiše znake s slike in tako potrdi, da gre za človeškega uporabnika, ne robota. Tudi tovrstne prepreke so sicer premostljive, a so lahko legalno sporne, poleg tega pa zahtevajo veliko truda. Zaradi široke izbire iskalnikov sem se enostavno odločil za izbiro drugega - Bing Search. Izkazalo se je, da ta na noben način ne blokira spletnih pajkov, poleg tega pa rezultate izpiše v enostavnejši strukturi od Googla. Stran z rezultati iskanja pridobimo z enostavnim GET zahtevkom, ki mu kot parameter podamo ime podjetja. Izmed vseh vrnjenih rezultatov mora program zaznati kateri, če sploh, je dejanski spletni naslov podjetja. Za to skrbi algoritem za prepoznavanje, ki v prvi fazi iz kratkega imena podjetja sestavi seznam ključnih besed.

Ime podjetja se razdeli po posameznih besedah in nato ustrezno prefiltrira - iz seznama se odstranijo vezniki ter besede in kratice, ki se pogosto pojavljajo v imenih podjetij (npr. „podjetje“, „d.o.o“). Algoritem nato preveri, če se katera od preostalih ključnih besed iz seznama nahaja v imenu gostitelja in v primeru, da se, vrne spletni naslov kot pravi naslov podjetja. V nekaterih primerih je seveda nemogoče ugotoviti, ali je algoritem našel pravi naslov, saj imena podjetij še zdaleč niso unikatna. Zato bomo po luščenju podatkov iz ostalih virov opravili še filtriranje spletnih naslovov na podlagi lokacije, ki jo zazna portal Alexa.

3.4.3 Alexa

Orodje Site Overview v okviru spletnega mesta Alexa kot vhodni parameter prejme le URL naslov strani, ki jo želimo analizirati. Rezultati so prikazani na strani s spletnim naslovom oblike „http://www.alexa.com/siteinfo/*imeGostitelja*“, v katerem lahko ime gostitelja prilagajamo posameznimi spletni strani podjetja in tako do rezultatov dostopamo neposredno. Portal Alexa podatke, ki jih potrebujemo za raziskavo, večinoma objavlja v enako strukturirani obliki. Številski podatki so objavljeni kot tekst v svoji vrstici, pred dejansko vrednostjo pa je naveden še naslov. Izjemi sta zgolj podatka o času nalaganja ter številu vhodnih povezav, ki imata okoli številskega podatka še besedilo, posledično je drugačna tudi HTML struktura. Algoritem se sprehodi po vseh vozliščih z naslovi, nato pa iz sledečega vozlišča prebere še številsko vrednost podatka. Postopek združuje uporabo objektnega modela HTML dokumenta ter poizvedovalnega jezika xPath in je delno prikazan v izpisu 3.1.

Alexa vsako spletno stran uvrsti na lestvico priljubljenosti v svetovnem merilu ter v državi registracije strani. Ker slovenska podjetja na svetovni lestvici ne dosegajo vidnejših rezultatov, te lestvice za nas niso uporabne. Je pa uporabno ime države, v kateri Alexa rangira stran, saj lahko na podlagi tega podatka potrdimo, da smo s spletnih iskalnikom dejansko našli pravo stran podjetja.

```
foreach (var node in
    doc.DocumentNode.SelectNodes("//h4[@class=\"metrics-title\"]"))
{
    string title = node.ChildNodes[0].InnerHtml.Trim();
    HtmlNode datanode =
        node.SelectSingleNode("following-sibling::div/strong");

    string valueString =
        (from child in datanode.ChildNodes
        where child.NodeType == HtmlNodeType.Text
        && child.InnerText.Trim() != ""
        select child.InnerText.Trim()).FirstOrDefault();

    //Pisanje vrednosti v pripadajoce stolpce v bazi

}
```

Izpis 3.1: Razčlenjevanje podatkov s portala Alexa.

3.4.4 Woorank

Woorank uporablja skoraj identičen pristop za dostop do strani z rezultati, kot Alexa. Osnovnemu URL naslovu https://www.woorank.com/en/ se tudi tu pripne URL strani, ki jo želimo oceniti. Rezultati, ki jih želimo pridobiti, so podani v različnih strukturah. Najbolj pogosto bomo za posamezen atribut strani pogledali le zastavico, ki označuje, ali je le-ta prestal Woorankov test primernosti. Atributi, kot so varnost, zaupanje, hitrost izvajanja na mobilni napravi, so grafično prikazani na lestvici (z diskretnimi ali zveznimi vrednostmi), nekateri pa imajo definiran specifičen prikaz in zato zahtevajo razčlenjevanje ločeno od ostalih podatkov na strani. Vrednosti dvajsetih atributov lahko preberemo iz zastavic, kot so prikazane na sliki 3.3. Vsak atribut je v HTML dokumentu opremljen z identifikatorjem „criterium-*ime_atributa*“, zato nabor vozlišč hitro pridobimo z xPath izrazom, izpisanem v izpisu 3.2.

Če gre za vozlišče z zastavico, pogledamo le še CSS razred ikone in na

```
//div[@class=\"module-content\"]/div[starts-with(@id,\"criterium-\")]
```

Izpis 3.2: xPath izraz, ki vrne vozlišča z rezultati testov.

podlagi tega nastavimo vrednost atributa na 1 ali 0. Vrednosti, ki so predstavljene na lestvici, lahko zavzamejo številske (od 0 do 100) ali opisne vrednosti (zelo slabo, slabo, dobro, zelo dobro). Za potrebe grafičnega prikaza lestvice pa imajo vse vrednosti kot atribut podano tudi absolutno vrednost, ki je za številske vrednosti enaka, pri opisnih pa opis preslika na interval (0,100), kar nam prihrani nekaj dela pri kvantifikaciji vrednosti.

3.4.5 Facebook profil

S Facebook profila bomo pridobili informacije o aktivnosti podjetja na socialnem omrežju in o zadovoljstvu uporabnikov. Preverili bomo tudi pravilnost objavljenih kontaktnih podatkov. Ob raziskavi izvorne kode opazimo, da Facebook podatke o naslovu ter oceni hrani tudi v formatu JSON-LD. JSON-LD je metoda za kodiranje povezanih podatkov z uporabo JSON-a in omogočanje izvajanja semantičnih poizvedb. Zasnovan je okoli konteksta informacij in ponuja dodatne preslikave iz formata JSON v model RDF (Resource Description Framework, družina W3C specifikacij) [2]. Objava v tem formatu nam, v skladu z idejo semantičnega spleta, omogoči enostavno deserializacijo podatkov, zato se z ekstrakcijo niti ni potrebno ukvarjati. Primer izpisa podatkov za podjetje Abanka d.d. je v izpisu 3.3. Kontaktni podatki so na spletni strani zapisani v tabeli v zavihku „About“ in jih izluščimo iz tabele. Čas zadnje objave pa preberemo iz časovnice podjetja in sicer z xPath izrazom poiščemo prvi element z atributom „data-utime“. Ta atribut predstavlja časovno značko objave v Unix času - številu sekund, ki so pretekle od polnoči 1.1.1970. Za razliko od ostalih atributov, časa zadnje objave ni smiselno luščiti iz datoteke, shranjene na disku, saj se ta spreminja. Ob vsaki

```
{
    "@context":"http://schema.org",
    "@type":"Organization",
    "name":"Abanka",
    "address": {
        "@type":"PostalAddress",
        "streetAddress":"Slovenska cesta, 58",
        "addressLocality":"Ljubljana, Slovenia",
        "addressRegion":"Central Slovenia Statistical Region",
        "postalCode":"1000 Ljubljana"
        },
    "aggregateRating": {
        "@type":"AggregateRating",
        "ratingValue":3.9,
        "ratingCount":"56"
        }
}
```

Izpis 3.3: Format JSON-LD

ponovitvi se zato časovnica naloži in razčleni neposredno iz spletne strani. Ker sam časovni žig težko uporabimo kot metriko kakovosti, bomo raje merili število preteklih dni od zadnje objave. Izračunali jih bomo kot razliko med časovnim žigom razčlenjevanja strani in časovnim žigom zadnje objave.

3.4.6 Lastna stran podjetja

V primerjavi z ostalimi viri, se ekstrakcija podatkov iz lastnih strani podjetij izkaže za svojevrsten izziv. Če smo pri prejšnjih virih imeli v veliki meri opravka le z opisovanjem pozicije podatka na spletni strani, si tu s tem pristopom ne moremo pomagati. Struktura HTML dokumenta se namreč razlikuje od strani do strani, zato moramo razviti univerzalne tehnike za prepoznavo podatkov, ki jih želimo izluščiti. Namesto po drevesni strukturi HTML dokumenta, bomo tu podatke iskali po besedilu spletne strani. Zanimalo nas bo, ali ima podjetje na svoji spletni stani objavljen naslov, telefonsko številko, na-

slov elektronske pošte ter povezave do profilov na socialnih omrežjih. Povrh tega bomo preverili tudi, ali ima podjetje svojo lokacijo prikazano na zemljevidu, ter ali je možno elektronsko pošto poslati neposredno iz spletne strani z uporabo „mailto“ povezave. Za vse naštete podatke bomo preverili tudi na katerem hierarhičnem nivoju so objavljeni. Najdene kontaktne podatke bomo ocenili tako z vidika pravilnosti (primerjava z AJPES-om in Facebookom), kot z vidika konsistentnosti njihovega zapisa (npr. število različnih oblik zapisa telefonske številke). Ker so nekatera spletišča zelo obsežna, smo se v raziskavi omejili na prenos in razčlenjevanje prvih dveh nivojev strani.

Prepoznava naslova

Prepoznave naslova se lotimo na podlagi pravega naslova podjetja, ki je vpisan v PRS. V besedilu najprej iščemo kombinacije poštne številke in poštnega kraja, nato še kombinacije ulice in hišne številke. Z besedo kombinacije mislimo na različne oblike zapisa, ki se pogosto pojavljajo (npr. *pošta, kraj*, *kraj, pošta*, *hišna številka, ulica*). Ker se zapisi istega naslova lahko med seboj razlikujejo tudi po vsebini (dvojezična imena krajev, spuščena ali okrajšana beseda ulica, cesta), jih ne smemo primerjati neposredno. V ta namen uporabimo Levenshteinovo razdaljo, ki predstavlja minimalno število operacij potrebnih za preoblikovanje enega niza v drugega [24]. Naslov torej prepoznamo tako, da iz imen ulic najprej odstranimo prej omenjene pogoste besede, nato pa izračunamo Levenstheinovo razdaljo med nizom na spletni strani in pravim naslovom. Razdaljo, pri kateri niza še priznamo za „enaka“, določimo relativno na dolžino niza. Pravilnost naslova na strani sem številsko ovrednotil glede na najdene komponente naslova, uteženo po lestvici v tabeli 3.1. Komponente naslova so ovrednotene glede na to, kako natančno opišejo lokacijo. V primeru, da na strani identificiramo le kraj in poštno številko, je pravilnost 40 odstotna, medtem ko je ob identificiranem kraju, poštni številki in ulici 80 odstotna. Iskanje in vrednotenje zgolj števil (poštne in hišne številke) ne bi bilo smiselno, zato jih iščemo le v kombinaciji z ulico oziroma pošto.

Komponenta	Vrednost
Kraj	1
Kraj in poštna številka	2
Ulica	2
Ulica in hišna številka	3

Tabela 3.1: Številsko vrednotenje najdenih delov naslova, glede na pomembnost.

Prepoznava telefonske številke

Poleg naslova podjetja, moramo v tekstu prepoznati tudi telefonsko številko. Prepoznavanje se izkaže za zelo zahtevno, saj je možnih oblik zapisa mnogo. Številka je lahko mobilna ali stacionarna, vsebuje predpono omrežne skupine ali klicno kodo države, zapisana je lahko strnjeno ali pa z različnimi znaki ločena v skupine števil. Uporaba regularnih izrazov je sicer možna, a se ne izkaže za najbolj učinkovito tehniko. Iz vsakega tekstovnega elementa na spletni strani sem zato najprej odstranil vse znake, ki niso alfanumerični, z izjemo pik. Odstranil sem tudi nize, ki se pogosto pojavljajo neposredno ob telefonskih številkah (npr. „Telefon“, „GSM“, „Tel.“ ...). Iz pridobljenega nabora nizov ohranimo le tiste, ki jih sestavljajo številke in pike. Pike v številskih nizih nam pomagajo pri ugotavljanju, ali gre morda za datum ali decimalno število. Če ugotovimo, da gre za celo število, preverimo še njegovo dolžino in v kolikor obsega med 7 in 12 znaki, niz obravnavamo kot telefonsko številko. Algoritem seveda ni popolnoma zanesljiv, saj se na strani lahko pojavljajo zapisi, za katere iz strukture ne moramo sklepati, ali gre za telefonsko številko. Se pa je algoritem skozi iteracije izkazal za dovolj zanesljivega, da ga lahko uporabimo v raziskavi. Med prepoznanimi telefonskimi številkami merimo pravilnost in konsistentnost njihovega zapisa. Konsistentnost izmerimo kot delež različnih formatov zapisa med vsemi najdenimi številkami. Formate telefonskih številk med seboj primerjamo glede na obliko zapisa klicne kode države ter glede na ločila med posameznimi sklopi številk. Pravilnost

ocenimo kot ujemanje zadnjih 6 števk s tistimi, v AJPES-u. Celotne številke ne primerjamo zaradi prej opisanih razlik v zapisu.

3.4.7 Ostale lastnosti

Zaradi bolj standardne oblike, lahko za iskanje povezav do profilov na socialnih omrežjih in elektronske pošte uporabimo regularne izraze. Pri socialnih omrežjih, zemljevidih in „mailto“ povezavah iščemo povezave (href) specifične oblike, elektronsko pošto pa iščemo z regularnim izrazom, ki išče nize oblike *nekaj@foo.bar* in je prikazan v izpisu 3.4.

```
\b[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}\b
```

Izpis 3.4: Regex za ujemanje naslovov elektronske pošte.

Za vsak najden podatek sproti zabeležimo še, na katerem nivoju je bil najden in na podlagi teh podatkov sestavimo skupno oceno najdljivosti podatkov.

Poglavje 4

Analiza rezultatov

V tem poglavju se bomo posvetili analizi podatkov, pridobljenih z implementiranim spletnim pajkom. Pajek je za izhodiščne podatke uporabil podatke 34.456 podjetij iz poslovnega registra Slovenije, ki po pravnoorganizacijski obliki spadajo med delniške družbe (d.d.) ali družbe z omejeno odgovornostjo (d.o.o.). Dodaten pogoj, da podjetje vključimo v raziskavo je bil, da je za posamezno podjetje v registru vpisan vsaj en kontaktni podatek, katerega pravilnost preverjamo. Končni nabor podatkov za analizo z odstranjenimi osamelci in podjetji, katerih spletnega naslova nismo našli, obsega 19.055 podjetij, opisanih s 25 spremenljivkami v tabeli 4.1. Spremenljivkam smo zaradi lažjega prikaza poleg imen dodali še oznake x1-x25. Atribute bomo najprej smiselno združili v dimenzije kakovosti z uporabo faktorske analize. V osrednjem delu analize bomo skušali poiskati korelacije med kakovostjo informacij in velikostjo podjetja ter ugotoviti, kako je kakovost odvisna od dejavnosti podjetij. Atributi, ki so bili v fazi analize odstranjeni iz nabora, so označeni z „zvezdico“ (*).

4.1 Identifikacija dimenzij kakovosti

Pred ugotavljanjem napovedne moči izmerjene kakovosti informacij, je bilo potrebno pridobljene mere smiselno razvrstiti v dimenzije kakovosti, ki smo

Oznaka	Ime atributa	Razlaga
x1	BounceRate	odstotek obiskovalcev, ki so stran zapustili brez obiska podstrani
x2	TimeOnSite	dolžina obiska strani, v milisekundah
x3	ViewsPerVisitor	število obiskov na obiskovalca
x4	*SearchVisits	število obiskov iz spletnega iskalnika
x5	AddressCorrect	stopnja pravilnosti naslova
x6	ConsistentNumbers	stopnja konsistentnega zapisa telefonskih številk
x7	CorrectNumber	stopnja pravilnosti telefonskih številk
x8	CorrectEmail	stopnja pravilnost naslova elektronske pošte
x9	ParagraphStyles	stopnja odstavkov s stilskim odstopanjem
x10	Wot_ChildSafety	odstotek primernosti vsebine za otroke
x11	Wot_Trust	odstotek vrednosti zaupanja
x12	*Wot_Vendor	odstotek primernosti ponudnika
x13	*Wot_Privacy	odstotek zasebnosti
x14	PageSize	velikost spletne strani
x15	LoadTime	čas nalaganja spletne strani, v milisekundah
x16	Facebook.rating	ocena profila na Facebooku
x17	*Facebook.rates.count	število ocen profila na Facebooku
x18	DaysSinceLastFbPost	število dni od zadnje objave na Facebooku
x19	ContactChannels	število objavljenih kontaktnih kanalov
x20	HasMailToAndEmail	celovitost podajanja naslova elektronske pošte
x21	HasAddressAndMap	celovitost podajanja lokacije
x22	PresentationLevel	razmerje v količini informacij, podanih na prvem in drugem nivoju
x23	SEO	stopnja Optimizacije vsebine za spletne iskalnike
x24	Mobile	odstotek primernosti prikaza na mobilnem aparatu
x25	Technologies	stopnja primernosti uporabljenih tehnologij

Tabela 4.1: Končni nabor pridobljenih atributov.

jih opisali v poglavju 2. Tega smo se lotili s faktorsko analizo, katere cilj je najti novo, manjšo množico spremenljivk, ki predstavljajo to, kar je skupnega spremenljivkam iz začetnega nabora.

Ker uporabljeni viri ne vsebujejo popolnih informacij za vsako od obravnavanih strani, je bilo v začetnem naboru veliko manjkajočih podatkov. Število izpisanih karakteristik na portalu Alexa in Woorank je namreč odvisno od obiskanosti obravnavane strani, saj za bolj priljubljene strani lahko pridobimo več podatkov. Prav tako niso imela vsa podjetja objavljene povezave na svoj Facebook profil (ali pa ga sploh nimajo), zato so se tudi pri atributih povezanih s Facebookom pojavile manjkajoče vrednosti. Kjer je bilo smiselno, smo manjkajoči podatek nadomestili z najnižjo oceno tistega atributa. Če podjetje na primer nima Facebook profila, smo ga ocenili (Facebook.Rating) z 0. Pri atributih, kjer zapolnjevanje manjkajočih vrednosti na opisan način ni bilo smiselno (npr. LoadingTime, PageSize), smo namesto manjkajočih vrednosti uporabili povprečno vrednost atributa.

4.1.1 Preučevanje spremenljivk

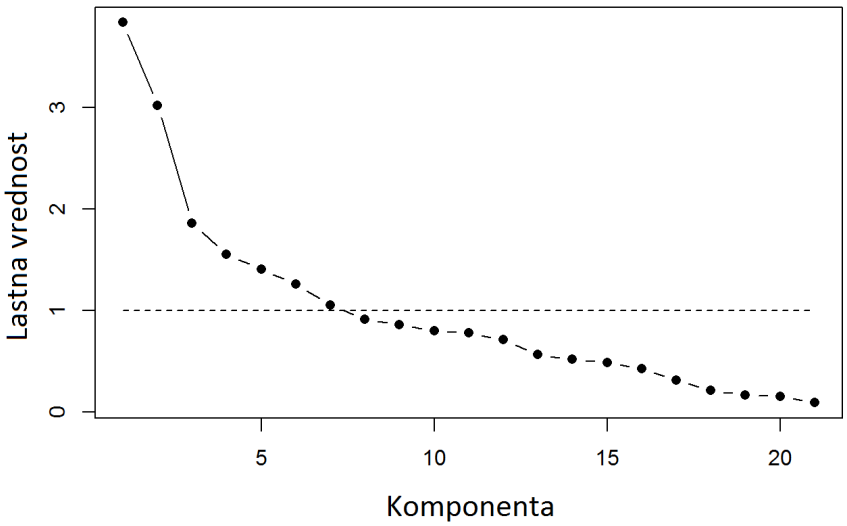
Pregled spremenljivk začnemo z vizualizacijo in izračunom matrike korelacij (glej tabelo 4.2. Iz te opazimo, da med spremenljivkami obstajajo številne statistično značilne odvisnosti. Pregled korelacij med obravnavanimi spremenljivkami pokaže, da trije od štirih atributov storitve Web of Trust (Vendor, Child Safety, Privacy) med seboj popolnoma korelirajo, zato smo dva od njih odstranili iz nadaljnje obravnave. Ker je za uspešno faktorsko analizo potrebno upoštevati predpostavko, da mora med spremenljivkami obstajati nekaj korelacije (vendar ta ne sme biti prevelika), iz nabora izločimo tudi elemente, katerih korelacija ni statistično značilna. To preverimo z MSA testom in izločimo spremenljivki x2 in Facebook.ratings.count, katerih faktor pade pod priporočeno mejo 0.5. Skupen MSA je nato enak 0.721, kar je sprejemljivo. Za oceno skupne statistične značilnosti matrike uporabimo tudi Bartlettov test, ki v našem primeru znaša $1.608 \times 10^5 (p = 0)$.

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20	x21	x22	x23	x24	x25
x1	1.00	0.21	-0.34	-0.35	-0.02	-0.01	-0.02	-0.01	0.00	-0.02	-0.02	-0.02	-0.02	0.01	0.01	-0.03	0.01	0.02	0.01	-0.01	-0.03	0.00	0.01	-0.03	-0.02
x2		1.00	-0.21	-0.12	0.02	0.00	0.01	0.00	-0.01	-0.13	-0.13	-0.13	-0.13	-0.04	0.03	-0.01	-0.01	0.03	-0.05	0.00	0.02	-0.02	-0.02	-0.11	-0.05
x3			1.00	0.57	0.01	0.01	0.01	0.01	0.01	0.08	0.07	0.07	0.07	0.03	0.00	-0.01	0.00	0.00	0.02	0.00	0.01	0.02	0.01	0.07	0.03
x4				1.00	0.00	0.02	0.01	0.00	0.02	0.08	0.08	0.08	0.08	0.02	0.00	0.02	0.00	-0.02	0.03	0.00	0.01	0.00	0.02	0.07	0.04
x5					1.00	0.15	0.43	0.36	0.11	-0.13	-0.14	-0.14	-0.14	-0.05	0.01	0.01	-0.01	0.04	0.07	0.26	0.83	0.38	0.05	-0.13	0.04
x6						1.00	0.27	0.06	0.21	0.03	0.04	0.04	0.04	0.06	0.03	0.08	0.00	-0.10	0.57	0.39	0.20	0.49	0.12	0.05	0.11
x7							1.00	0.25	0.07	-0.07	-0.08	-0.08	-0.08	-0.02	0.01	0.02	0.00	0.02	0.13	0.15	0.38	0.27	0.03	-0.07	0.02
x8								1.00	0.06	-0.06	-0.07	-0.07	-0.07	-0.03	0.01	0.01	0.00	0.03	0.06	0.18	0.31	0.20	0.03	-0.06	0.03
x9									1.00	0.03	0.03	0.03	0.03	-0.01	0.00	0.03	0.01	-0.05	0.24	0.31	0.13	0.24	0.10	0.02	0.11
x10										1.00	0.90	0.90	0.90	0.19	-0.05	0.06	0.04	-0.17	0.24	0.01	-0.11	0.06	0.18	0.56	0.22
x11											1.00	1.00	1.00	0.18	-0.01	0.07	0.04	-0.17	0.25	0.02	-0.11	0.07	0.20	0.63	0.24
x12												1.00	1.00	0.18	-0.01	0.07	0.04	-0.17	0.25	0.02	-0.11	0.07	0.20	0.63	0.24
x13													1.00	0.18	-0.01	0.07	0.04	-0.17	0.25	0.02	-0.11	0.07	0.20	0.63	0.24
x14														1.00	0.06	0.04	0.01	-0.08	0.17	0.04	-0.01	0.11	0.09	0.17	0.08
x15															1.00	0.03	-0.01	-0.02	0.05	0.03	0.03	0.06	0.03	0.03	0.02
x16																1.00	0.08	-0.43	0.26	0.06	0.03	0.14	0.10	0.08	0.11
x17																	1.00	-0.05	0.03	0.01	-0.01	0.01	0.01	0.03	0.02
x18																		1.00	-0.34	-0.08	0.01	-0.19	-0.11	-0.15	-0.12
x19																			1.00	0.57	0.15	0.74	0.28	0.23	0.29
x20																				1.00	0.28	0.68	0.14	0.02	0.15
x21																					1.00	0.45	0.08	-0.09	0.05
x22																						1.00	0.18	0.09	0.16
x23																							1.00	0.25	0.79
x24																								1.00	0.25
x25																									1.00

Tabela 4.2: Matrika korelacij.

4.1.2 Izbira faktorjev

Ko obdržimo le spremenljivke s pomenljivimi medsebojnimi povezavami, moramo ugotoviti, koliko komponent bomo obdržali za nadaljnjo analizo, glede na odstotek variance, ki jo pojasni. Po kriteriju testa drobirja (glej sliko 4.1), bi bilo primerno obdržati zgolj 3 komponente, saj postane s četrto lastno vrednostjo krivulja bolj položna. A ker je odstotek pojasnjene variance s tremi komponentami le 41,52% (glej tabelo 4.3), poleg tega pa bi izvirne spremenljivke težko smiselno razvrstili v tako majhno število dimenzij, raje uporabimo Kaiserjev kriterij. Obdržimo torej 7 spremenljivk z lastnimi vrednostmi večjimi od 1.



Slika 4.1: Scree plot.

Komponenta	Lastna vrednost	% variance	% variance skupno
1	3.84	18.27	18.27
2	3.02	14.39	32.66
3	1.86	8.86	41.52
4	1.55	7.40	48.92
5	1.41	6.70	55.62
6	1.26	6.00	61.62
7	1.05	5.01	66.64
8	0.91	4.35	70.99
9	0.86	4.10	75.09
10	0.80	3.80	78.89
11	0.78	3.70	82.60
12	0.71	3.39	85.99
13	0.56	2.68	88.67
14	0.52	2.47	91.14
15	0.49	2.33	93.47
16	0.43	2.03	95.50
17	0.32	1.51	97.01
18	0.21	1.01	98.02
19	0.17	0.80	98.82
20	0.16	0.74	99.56
21	0.09	0.44	100.00

Tabela 4.3: Prispevki posameznih komponent k varianci.

Za izračun faktorskih uteži uporabimo metodo glavnih komponent. Metoda vrne faktorje razvrščene po pomembnosti glede na količino variance. Ključno vlogo pri interpretaciji faktorjev igrajo uteži, ki nam povedo, kako močno ter v kateri smeri posamezne spremenljivke vplivajo na faktor. Da za vsako spremenljivko maksimiziramo utež le na enem faktorju in s tem olajšamo interpretacijo, izvedemo na matriki še VARIMAX rotacijo. Kot rezultat dobimo spremenljivke, ki močno utežujejo en faktor in zelo šibko ostale. Za bolj pregleden prikaz faktorskih uteži smo v tabeli 4.4 prikazali le statistično značilne uteži, ki so večje od 0.4. Komunaliteta v zadnjem stolpcu nam pove, kako dobro je spremenljivka pojasnjena z izbranimi faktorji. Tabela 4.5 povzema lastne vrednosti in odstotek pojasnjene variance skozi posamezne faktorje.

Spremenljivka	F1	F2	F3	F4	F5	F6	F7	Komunaliteta
ContactChannels	0.81							0.83
PresentationLevel	0.79							0.79
HasMailToAndEmail	0.79							0.66
ConsistentNumbers	0.72							0.54
ParagraphStyles	0.50							0.33
Wot_Trust		0.94						0.90
Wot_ChildSafety		0.93						0.87
Mobile		0.75						0.62
AddressCorrect			0.89					0.82
HasAddressAndMap			0.86					0.78
CorrectNumber			0.62					0.41
CorrectEmail			0.58					0.34
ViewsPerVisitor				0.83				0.70
TimeOnSite				0.83				0.69
BounceRate				-0.68				0.47
SEO					0.92			0.89
Technologies					0.92			0.89
Facebook.Rating						0.85		0.72
DaySinceLastFbPost						-0.81		0.70
LoadTime							0.80	0.66
PageSize							0.58	0.40

Tabela 4.4: Faktorske uteži po VARIMAX rotaciji.

	F1	F2	F3	F4	F5	F6	F7	Komunaliteta
Vsota kvadratov uteži (lastne vrednosti)	2.80	2.51	2.44	1.86	1.78	1.51	1.10	13.99
Odstotek lastnih vrednosti	13.34	11.97	11.61	8.84	8.48	7.18	5.23	66.64

Tabela 4.5: Pregled faktorjev.

4.1.3 Interpretacija faktorjev

Z identificiranimi sedmimi faktorji torej pojasnimo 66.635 odstotkov variance med podatki. Pri identifikaciji dimenzij smo se uprli na pretekle raziskave,

opisane v poglavju 2 in glede na spremenljivke, ki utežujejo faktorje iz njih ustvarili naslednje dimenzije:

1. Faktor 1 (F1): **Celovitost**: ContactChannels, PresentationLevel, HasMailToAndEmail, ConsistentNumbers, ParagraphStyles,

2. Faktor 2 (F2): **Varnost in dostopnost**: Trust, ChildSafety, Mobile,

3. Faktor 3 (F3): **Natančnost**: AddressCorrect, HasAddressAndMap, CorrectNumber,CorrectEmail,

4. Faktor 4 (F4): **Uporabniška izkušnja**: TimeOnSite, ViewsPerVisitor, BounceRate,

5. Faktor 5 (F5): **Implementacija**: SEO, Technologies,

6. Faktor 6 (F6): **Socialno omrežje**: FacebookRating, DaysSinceLastFbPost,

7. Faktor 7 (F7): **Pravočasnost**: LoadTime, PageSize.

Opazimo, da ima spremenljivka DaysSinceLastFbPost (število dni od zadnje objave na Facebooku) negativen predznak, kar pomeni da na kakovost udejstvovanja podjetja na socialnem omrežju vpliva v obratni smeri. V praksi to pomeni, da manj, kot je podjetje aktivno na socialnem omrežju, slabša je njegova ocena v tej dimenziji. Podobno ima negativen predznak BounceRate, saj visoka vrednost pomeni, da veliko uporabnikov hitro zapusti spletno stran. S previdnostjo moramo obravnavati tudi Pravočasnost - LoadTime (čas nalaganja) in PageSize (velikost strani) jasno korelirata, a v smislu pravočasnosti pomeni višja vrednost slabšo kakovost.

V nadaljnji analizi opustimo začetne spremenljivke in uporabljamo identificirane dimenzije kakovosti, izračunane kot linearna kombinacija uteži.

4.2 Primerjava faktorjev z lastnostmi podjetij

Da bi ugotovili, kako so dimenzije kakovosti informacij povezane s splošnimi lastnostmi podjetij, smo jih ročno razvrstili v gruče glede na njihovo dejavnost in velikost ter izračunali povprečje faktorjev v vsaki gruči. Poleg tega je vsaki povprečni vrednosti v grafu dodan 95% interval zaupanja, ki je bil pridobljen s pomočjo bootstrap postopka.

Analiza po dejavnostih podjetij

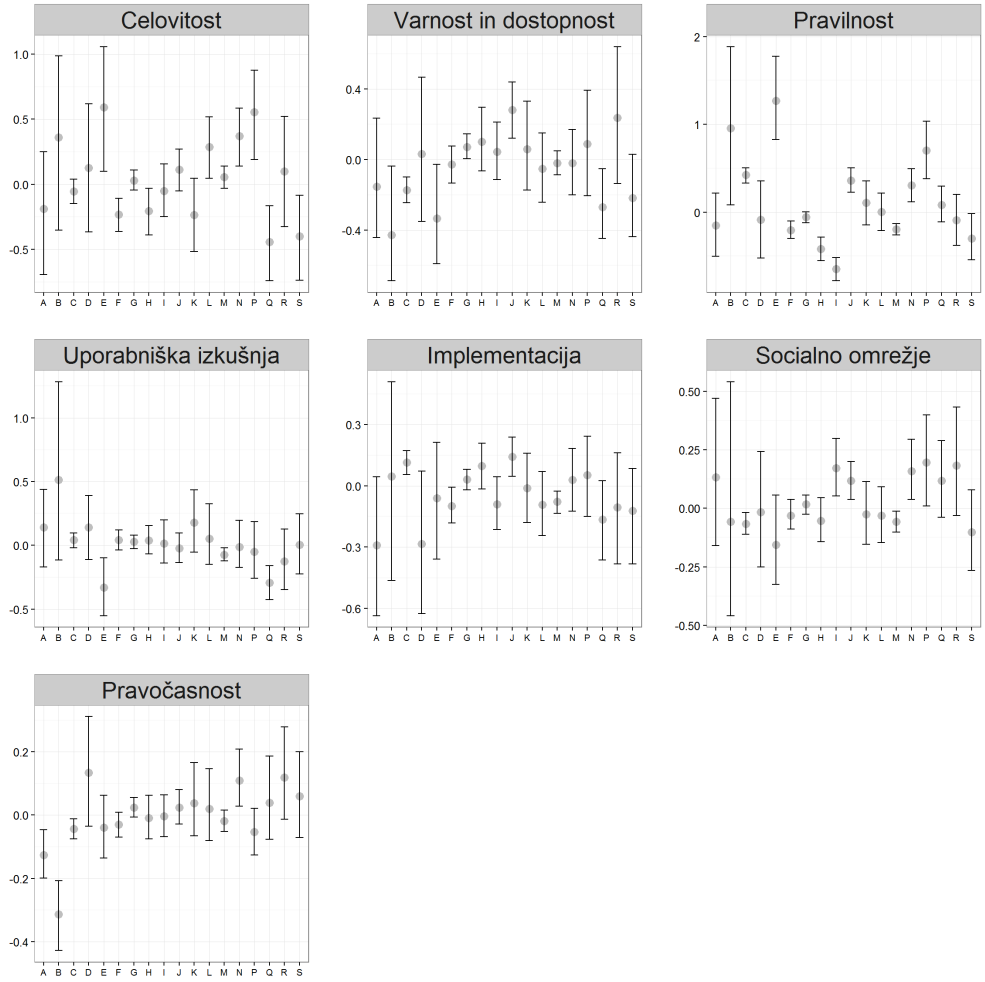
Na sliki 4.2 so prikazane povprečne vrednosti posameznih faktorjev z dodanimi intervali zaupanja, po standardni klasifikaciji dejavnosti. Klasifikacija je predstavljena v tabeli 4.6, vodoravne črte pa dejavnosti ločujejo v štiri gospodarske sektorje. Ker vzorec za dejavnost kategorije O (dejavnost javne uprave in obrambe, dejavnost obvezne socialne varnosti) sestavljajo le 4 podjetja, smo to skupino odstranili. Zaradi atributov, ki negativno korelirajo in standardiziranih vrednosti, je lahko skupna vrednost tudi manjša od 0. Opazimo, da med dejavnostmi podjetij najbolj varirajo kakovosti dimenzij celovitosti, varnosti in dostopnosti ter pravilnosti podatkov. Uporabniška izkušnja, kakovost implementacije in socialnega omrežja ter pravočasnost, med posameznimi dejavnosti večinoma ne kažejo statistično značilnih razlik.

V smislu celovitosti informacij je največje odstopanje opazno med podjetji ki se ukvarjajo z izobraževalno dejavnostjo in podjetji, ki se ukvarjajo z zdravstvom. Da sta skupini statistično različni, s stopnjo značilnosti 0.05 potrdi tudi Mann-Whitney-Wilcoxov test ($W = 60994$, $p = 1.282 \times 10^{-5}$).

Varnost in dostopnost spletišča je pri podjetjih kategorije J (informacijske in komunikacijske dejavnosti) boljša od polovice ostalih kategorij in ima najvišjo povprečno oceno. Povprečje je pri kategoriji J najvišje tudi v smislu kakovosti implementacije spletnega portala, s statistično značilnostjo pa lahko potrdimo tudi, da je kakovost boljša kot pri petih ostalih kategorijah. Glede na to, da gre za podjetja, ki se ukvarjajo z informacijsko dejavnostjo,

so ta odstopanja pričakovana, saj podjetja z izdelavo lastne spletne strani demonstrirajo kakovost svojih storitev.

V kakovosti profila na socialnem omrežju so gostinska podjetja (kategorija I) boljša od tistih, ki se ukvarjajo s predelovalno dejavnostjo (kategorija C), gradbeništvom (kategorija F), prometom in skladiščenjem (kategorija H), pa tudi od tistih, ki se ukvarjajo z znanstveno dejavnostjo (kategorija M). Gostinstvo namreč zahteva bistveno bolj pogosto interakcijo s strankami, veliko prednost pa podjetju prinesejo tudi uporabniki, ki na socialnem omrežju delijo pozitivne izkušnje in s tem skrbijo za promocijo.



Slika 4.2: Povprečne vrednosti dimenzij po SKD.

oznaka	dejavnost
A	kmetijstvo in lov, gozdarstvo, ribištvo
B	rudarstvo
C	predelovalne dejavnosti
D	oskrba z električno energijo, plinom in paro
E	oskrba z vodo, ravnanje z odplakami in odpadki, saniranje okolja
F	gradbeništvo
G	trgovina, vzdrževanje in popravila motornih vozil
H	promet in skladiščenje
I	gostinstvo
J	informacijske in komunikacijske dejavnosti
K	finančne in zavarovalniške dejavnosti
L	poslovanje z nepremičninami
M	strokovne, znanstvene in tehnične dejavnosti
N	druge raznovrstne poslovne dejavnosti
O	dejavnost javne uprave in obrambe, dejavnost obvezne socialne varnosti
P	izobraževanje
Q	zdravstvo in socialno varstvo
R	kulturne, razvedrilne in rekreacijske dejavnosti
S	druge dejavnosti
T	dejavnost gospodinjstev z zaposlenim hišnim osebjem, proizvodnja za lastno rabo
U	dejavnost eksteritorialnih organizacij in teles

Tabela 4.6: Standardna klasifikacija dejavnosti.

Statistično značilne razlike v uporabniški izkušnji lahko opazimo med finančno in zavarovalniško dejavnostjo (kategorija K) ter na primer okoljskimi dejavnostmi (kategorija E) in zdravstvom (kategorija Q). Pri slednjih upo-

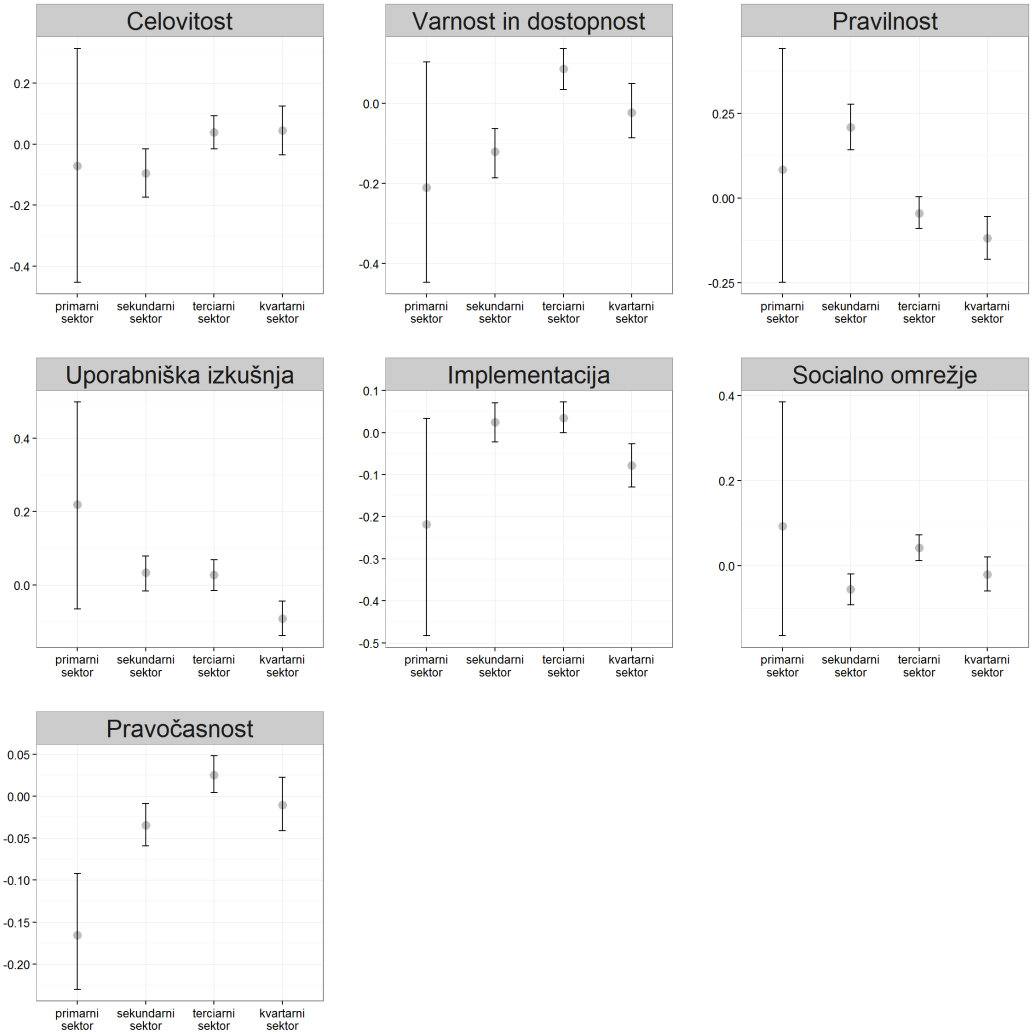
rabniki stran obiščejo manjkrat in jo tudi prej zapustijo. Razlog lahko iščemo v tem, da finančna in zavarovalniška podjetja nudijo na spletu širok nabor informacij o svojih storitvah, mnogo pa jih nudijo tudi prek spleta, zaradi česar obiski strani trajajo dlje. Na drugi strani pri podjetjih v kategorijah E in Q uporabniki pogosto poiščejo le ključno informacijo, ki jo potrebujejo in nato stran zapustijo. V zdravstvu je to zelo pogosto delovni čas in telefonska številka, saj večina komunikacije poteka preko telefona.

Da bi ugotovili ali obstajajo statistično značilne razlike tudi med gospodarskimi sektorji, združimo kategorije dejavnosti v naslednje skupine:

- primarni sektor (A, B),
- sekundarni sektor (C, D, E, F),
- terciarni sektor (G, H, I, J, K, L, M, N),
- kvartarni sektor (O, P, Q, R).

Povprečne vrednosti so skupaj s 95% intervali zaupanja prikazane na sliki 4.3. Opazimo, da so spletne strani podjetij storitvenega (terciarnega) sektorja z vidika celovitosti, varnosti in dostopnosti ter socialnih omrežij bolj kakovostne od tistih iz sekundarnega sektorja. Obratno pa lahko trdimo pri dimenziji pravilnosti podatkov, kjer dosegajo podjetja iz sekundarnega sektorja višjo oceno od tistih z terciarnega in kvartarnega. Sklepamo lahko, da želijo podjetja storitvenega sektorja narediti svoje storitve kar se da dostopne in varne. Pri podjetjih, ki svojih spletnih strani ne uporabljajo zgolj za predstavitev, ampak preko njih poslovanje tudi izvajajo (spletno bančništvo, spletne trgovine), je faktor varnosti še posebej pomemben. Prav tako je v zadnjem času opaziti vedno večji poudarek na dostopnosti storitev, kar se kaže s podporo izvajanja le-teh na mobilnih napravah. Pri dostopnosti v smislu podpore strankam pa ima močno vlogo tudi socialno omrežje. Vse to so lahko razlogi, da dajejo podjetja storitvenega sektorja večji poudarek na celovitost, varnost, dostopnosti ter socialna omrežja, medtem, ko se podjetja iz sekundarnega sektorja, ki svojih storitev ne morejo nuditi preko spleta,

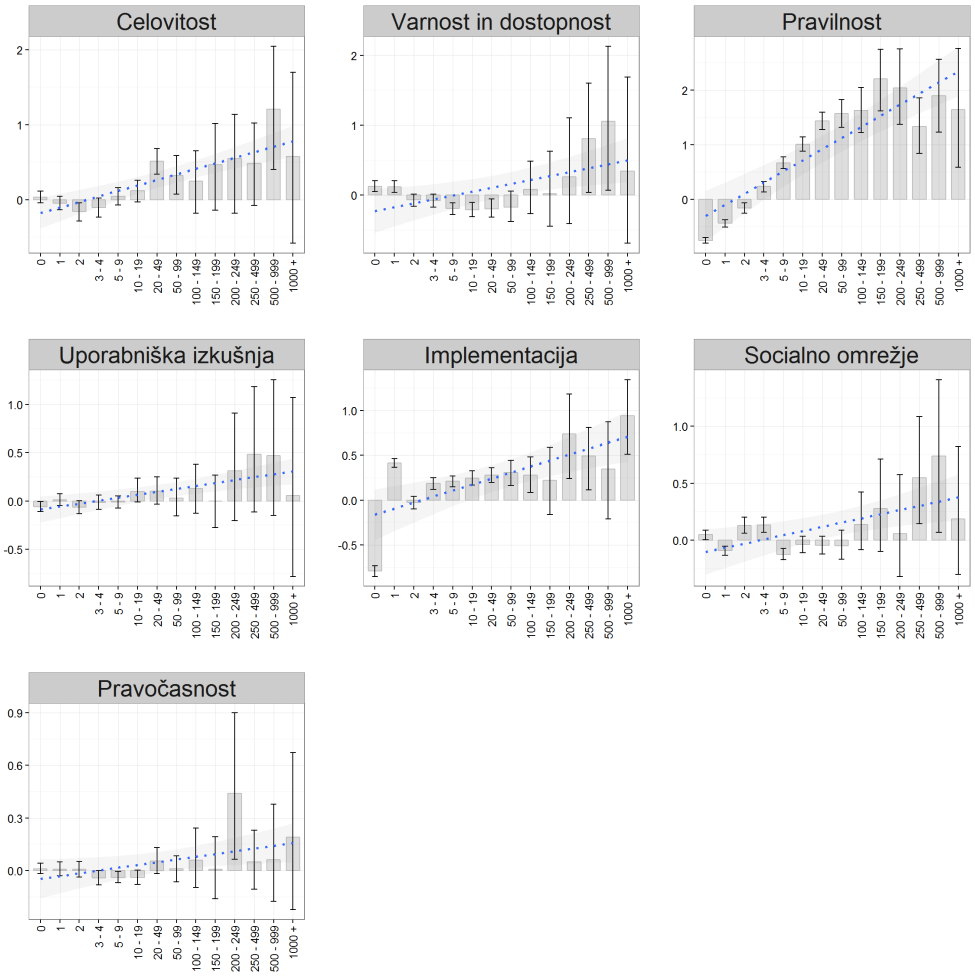
koncentrirajo predvsem na jasen in pravilen prikaz svojih kontaktnih podatkov.



Slika 4.3: Povprečne vrednosti dimenzij po gospodarskih sektorjih.

Analiza po velikosti podjetja

Povprečja dimenzij kakovosti po velikosti so prikazana na sliki 4.4. V vseh dimenzijah kakovosti je opaziti vsaj majhno rast z večanjem velikosti podjetja, kar lahko razložimo s sorazmerno večjim številom uporabnikov in posledično višjimi investicijami v spletne predstavitve. Trend je najlepše viden pri pravilnosti informacij, kjer je rast na intervalu od 0 do 20 zaposlenih praktično linearna, z neprekrivanjem intervalov zaupanja pa lahko potrdimo, da so razlike med skupinami na tem intervalu statistično značilne.



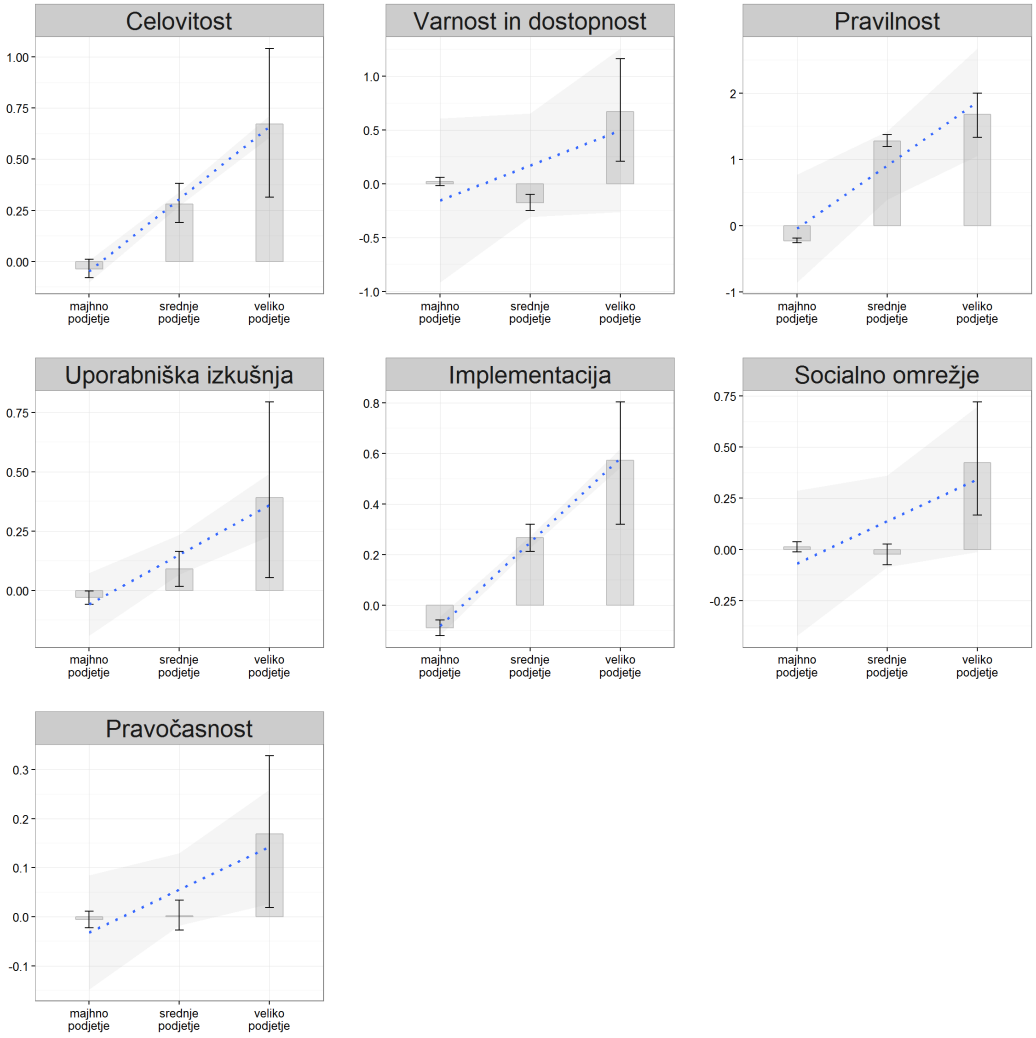
Slika 4.4: Povprečne vrednosti dimenzij po številu zaposlenih.

V dimenziji varnosti in dostopnosti opazimo upad kakovosti pri srednje velikih podjetjih, medtem ko je za velika podjetja stopnja varnosti zelo visoka. To razložimo z večjim številom obiskovalcev, na podlagi katerih je spletno stran spoznana za zaupanja vredno. Močno odstopanje velikih podjetij je opazno tudi pri primerjavi profilov na socialnih omrežjih, kar pomeni, da se večje korporacije bolj zavedajo pomena oglaševanja in komunikacije s strankami na socialnem omrežju.

Ker je število velikih korporacij v Sloveniji majhno, so sorazmerno majhni tudi vzorci podjetij večjih velikostnih razredov. Skupine zato združimo v tri velikostne razrede:

- majhno podjetje (do 9 zaposlenih),
- srednje veliko podjetje (od 10 do 199 zaposlenih),
- veliko podjetje (od 200 zaposlenih).

Na sliki 4.5 so prikazane povprečne ocene dimenzij kakovosti za zgoraj opisane skupine. Z združitvijo velikostnih razredov se znebimo nekaterih statističnih odstopanj, ki so bila prej prisotna pri podjetjih v največjih štirih velikostnih razredih (200 in več zaposlenih). Z združitvijo skupin postane trend rasti kakovosti z velikostjo podjetja še bolj jasen, znova z izjemo dimenzije varnosti in dostopnosti pri srednje velikih podjetjih. Označen je z modro premico, ki predstavlja zglajeno krivuljo povprečnih ocen posamezne dimenzije po velikostnih skupinah. Povprečja so zglajena z uporabo funkcije posplošenega linearnega modela (ang. generalized linear model, glm), siv pas okoli premice pa predstavlja 90% interval zaupanja.



Slika 4.5: Povprečne vrednosti dimenzij po združenih velikostnih razredih podjetij.

Poglavje 5

Sklepne ugotovitve

V diplomski nalogi smo najprej identificirali vire potencialno uporabnih prosto dostopnih podatkov za oceno kakovosti informacij na spletnih straneh. Vzorec so sestavljale spletne strani podjetij, ki po pravnoorganizacijski obliki spadajo med delniške družbe ali družbe z omejeno odgovornostjo in imajo v poslovnem registru Slovenije zabeležen vsaj en kontaktni podatek. V okolju .NET smo izdelali spletnega pajka, ki je iz spleta izluščil informacije za omenjen nabor podjetij in jih pretvoril v primerno obliko za analizo. Nabor izluščenih podatkov vključuje informacije o pravilnosti podatkov, zapisanih na lastnih spletnih staneh, celovitosti predstavitev informacij, tehnične podrobnosti o spletnih straneh ter podatke o obiskovalcih. V sklopu analize podatkov smo ta nabor zožili na 7 faktorjev, ki jih sestavljajo med seboj povezane spremenljivke iz začetnega nabora. V teh faktorjih smo identificirali uveljavljene dimenzije kakovosti iz znanstvene literature in na koncu poskusili razložiti razlike med kakovostjo informacij pri podjetjih različnih velikosti in panog.

Ugotovili smo torej, kako lastnosti spletnih strani izmeriti (številsko oceniti) ter kako te enostavne lastnosti povezati s koncepti kakovosti informacij. Izdelana rešitev poleg analize trenutnega stanja kakovosti informacij na straneh slovenskih podjetij, ponuja tudi orodje za celovito in avtomatizirano oceno lastne spletne strani. Izdelana programska oprema potrebuje za iz-

delavo ocene spletišča le ime podjetja in pravilne podatke o podjetju, na podlagi katerih lahko oceni pravilnost. Priporočena je tudi vključitev spletnega naslova, saj ne moremo z gotovostjo trditi, da bo algoritem s spletnim iskalnikom našel pravo spletno stran. Rezultati analize lahko lastniku spletne strani pomagajo pri odločanju o nadaljnjih investicijah v spletno predstavitev podjetja, saj izpostavijo prednosti in slabosti spletišča.

Pokazali smo, da obstajajo statistično značilne razlike v kakovosti informacij tako med dejavnostmi podjetij, kot med podjetji različnih velikosti. Ugotovili smo, da kakovost spletnih predstavitev z velikostjo podjetja raste skoraj linearno. Z vidika dejavnosti podjetij pa smo uspeli pokazati, da podjetja, ki se ukvarjajo z informacijsko in komunikacijsko dejavnostjo, dejansko izdelujejo kakovostne rešitve tudi za lastno predstavitev. Pokazali smo tudi statistično značilne razlike med podjetji sekundarnega in terciarnega sektorja, ki se zaradi narave svojih dejavnosti kažejo predvsem v dimenzijah, ki so povezane z elektronskim poslovanjem. Na splošno je med slovenskimi podjetji še veliko prostora za izboljšave pri uporabi socialnih omrežij za namen trženja ter pri investicijah v izdelavo spletnih predstavitev z bogatejšo uporabniško izkušnjo.

Delo vseeno pušča še precej prostora za izboljšave in razširitve, tako v smislu izdelane programske opreme, kot pri analizi. Za oceno kakovosti informacij je bilo v delu ocenjenih le nekaj atributov spletnih strani. Če izvzamemo zaznavanje naslova in številk, je bila vsebinska analiza besedila na spletni strani izvzeta iz raziskave. To dopušča možnost dodajanja novih atributov kakovosti, ki na primer ocenjujejo zahtevnost besedila na strani, jedrnatost vsebine ter ostale metrike, ki jih lahko izmerimo z naprednejšimi metodami tekstovne analize. Poleg naprednejše analize teksta, bi lahko v raziskavo vključili tudi analizo grafične podobe strani, ki zagotovo močno vpliva na uporabnikovo percepcijo. Z vidika analize bi lahko, ob prisotnosti ustreznih finančnih podatkov, skušali poiskati korelacije med kakovostjo informacij ter finančnimi izidi podjetij in tako obravnavati hipotezo, da kakovost spletnih strani vpliva na uspešnost podjetja.

Literatura

[1] Diffbot Is Using Computer Vision to Reinvent the Semantic Web. http://www.xconomy.com/san-francisco/2012/07/25/diffbot-is-using-computer-vision-to-reinvent-the-semantic-web/#. Dostopano: 4. 9. 2016.

[2] JSON-LD, A JSON-based Serialization for Linked Data. https://www.w3.org/TR/json-ld/. Dostopano: 4. 9. 2016.

[3] Knjižnjica HtmlAgilityPack. https://htmlagilitypack.codeplex.com/. Dostopano: 4. 9. 2016.

[4] Mozenda.com. http://www.mozenda.com/. Dostopano: 4. 9. 2016.

[5] Spletna stran AJPES. http://www.ajpes.si/Registri. Dostopano: 16. 8. 2016.

[6] Visual Scraper - Free web scraping tool. http://www.visualscraper.com/. Dostopano: 4. 9. 2016.

[7] Webhose.io, Crawled web data for your business. https://webhose.io/. Dostopano: 4. 9. 2016.

[8] Wikipedia: Web Scraping. https://en.wikipedia.org/wiki/Web_scraping. Dostopano: 4. 9. 2016.

[9] Ofer Arazy and Rick Kopak. On the measurability of information quality. *Journal of the American Society for Information Science and Technology*, 62(1):89–99, 2011.

[10] David J Bartholomew, Fiona Steele, Jane Galbraith, and Irini Moustaki. *Analysis of multivariate social science data*. CRC press, 2008.

[11] Christian Bauer and Arno Scharl. Quantitive evaluation of web site content and structure. *Internet Research*, 10(1):31–44, 2000.

[12] Martin J Eppler. *Managing information quality: increasing the value of information in knowledge-intensive products and processes*. Springer Science & Business Media, 2006.

[13] Martin J Eppler and Peter Muenzenmayer. Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology. In *IQ*, pages 187–196. Citeseer, 2002.

[14] Benjamin Fruchter. *Introduction to factor analysis.* Van Nostrand, 1954.

[15] A. Fulgosi. *Faktorska analiza*. Udžbenici Sveučilišta u Zagrebu. Školska knjiga, 1988.

[16] Mouzhi Ge and Markus Helfert. A review of information quality research—develop a research agenda. In *Paper presented at the International Conference on Information Quality 2007*. Citeseer, 2007.

[17] Layla Hasan and Emad Abuelrub. Assessing the quality of web sites. *Applied Computing and Informatics*, 9(1):11 – 29, 2011.

[18] Birger Hjørland. Information: objective or subjective/situational? *Journal of the American Society for Information Science and Technology*, 58(10):1448–1456, 2007.

[19] Beverly K Kahn, Diane M Strong, and Richard Y Wang. Information quality benchmarks: product and service performance. *Communications of the ACM*, 45(4):184–192, 2002.

[20] Henry F Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

[21] Barbara D. Klein. User perceptions of data quality: Internet and traditional text sources. *Journal of Computer Information Systems*, 41(4):9–15, 2001.

[22] Shirlee-ann Knight and Janice M Burn. Developing a framework for assessing information quality on the world wide web. *Informing Science: International Journal of an Emerging Transdiscipline*, 8(5):159–172, 2005.

[23] C.E. Lance and R.J. Vandenberg. *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences*. Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences. Routledge, 2009.

[24] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, February 1966.

[25] Eleanor T Loiacono, Richard T Watson, and Dale L Goodhue. Webqual: A measure of website quality. *Marketing theory and applications*, 13(3):432–438, 2002.

[26] Brendan Luyt, Tay Chee Hsien Aaron, Lim Hai Thian, and Cheng Kian Hong. Improving wikipedia's accuracy: Is edit age a solution? *Journal of the American Society for Information Science and Technology*, 59(2):318–330, 2008.

[27] S. K. Malik and S. Rizvi. Information extraction using web usage mining, web scrapping and semantic annotation. In *Computational Intelligence and Communication Networks (CICN), 2011 International Conference on*, pages 465–469, Oct 2011.

[28] Jussi Myllymaki. Effective web data extraction with standard {XML} technologies. *Computer Networks*, 39(5):635 – 644, 2002.

[29] Felix Naumann and Claudia Rolker. Assessment methods for information quality criteria. Number 138 in Informatik-Berichte. Institut für Informatik, 2000.

[30] Anne Ellerup Nielsen. *Rhetorical features of the company website.* centre for Internet Research, 2002.

[31] Richard J Ormerod. Critical rationalism in practice: Strategies to manage subjectivity in or investigations. *European Journal of Operational Research*, 235(3):784–797, 2014.

[32] Peter Pirolli, Evelin Wollny, and Bongwon Suh. So you know you're getting the best possible information: a tool that increases wikipedia credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1505–1508. ACM, 2009.

[33] Denise F Polit and Cheryl Tatano Beck. *Resource manual for nursing research.* Wolters Kluwer Health/lippincott Williams & Wilkins,, 2012.

[34] Soo Young Rieh. Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, 53(2):145–161, 2002.

[35] G Shankar and Stephanie Watts. A relevant, believable approach for data quality assessment. In *IQ*, pages 178–189, 2003.

[36] Klaus Stein and Claudia Hess. Does it matter who contributes: a study on featured articles in the german wikipedia. In *Proceedings of the eighteenth conference on Hypertext and hypermedia*, pages 171–174. ACM, 2007.

[37] Yair Wand and Richard Y Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95, 1996.

[38] Richard Y Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

[39] Eti Yaari, Shifra Baruchson-Arbib, and Judit Bar-Ilan. Information quality assessment of community generated content: A user study of wikipedia. *Journal of Information Science*, 37(5):487–498, 2011.

[40] Ping Zhang and G. von Dran. Expectations and rankings of web site quality features: results of two studies on user perceptions. In *System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference*, 2001.

[41] Xiaolan Zhu and Susan Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 288–295. ACM, 2000.