UNIVERZA V LJUBLJANI
FAKULTETA ZA DRUŽBENE VEDE

David Primc

Metode rangiranja spletnih strani

Diplomsko delo

Ljubljana, 2015

UNIVERZA V LJUBLJANI
FAKULTETA ZA DRUŽBENE VEDE

David Primc
Mentor: doc. dr. Damjan Škulj

Metode rangiranja spletnih strani

Diplomsko delo

Ljubljana, 2015

Metode rangiranja spletnih strani

Ljudje za iskanje informacij po spletu uporabljajo spletne iskalnike, katerih namen je lažje in hitrejše pridobivanje podatkov. Iskalniki omogočajo prikaz kakovostnih spletnih strani, ki jih razvrščajo glede na uporabnikove poizvedbe. Da pa lahko iskalniki prikazujejo spletne strani glede na uporabniško poizvedbo, potrebujejo dober in učinkovit algoritem rangiranja. Zaradi tega se v diplomski nalogi posebej posvetimo delovanju treh najbolj znanih metod rangiranja spletnih strani, PageRank, DistanceRank in VisualRank. Med temi je najbolj poznana prva, saj jo za rangiranje spletnih strani uporablja tudi Google, ki je trenutno najbolj uporabljen spletni iskalnik. Osredotočimo pa se tudi na delovanje spletnih iskalnikov in njihovo zgodovino. Za ilustracijo delovanja metod rangiranja smo razvili tudi lastni preprost spletni iskalnik, ki uporablja algortitem za rangiranje spletnih strani. Ta temelji na algoritmu PageRank z nekaterimi dodatki. Za pravilno delovanje potrebuje algoritem dostop do lastne podatkovne baze MySQL, ki se nahaja na lokalnem strežniku. Grafična podoba spletnega iskalnika je preprosta, saj vsebuje le vnosno polje in gumb. Tako iskalnik kot algoritem sta napisana v programskem jeziku PHP.

Ključne besede: spletni iskalnik, metode rangiranja spletnih strani, PageRank.

Website ranking methods

For online information retrieval people use search engines, whose purpose is easier and faster data retrieval. Search engines can display high-quality websites that are ranked according to the user's query. In order to display websites based on user query, they need a good and efficient ranking algorithm. This is the reason why, in this thesis we focus on the functioning of the three most popular methods of websites ranking, PageRank, DistanceRank and VisualRank. The first one being the best known among them, because it has been used by Google, which is one of the most popular Internet search engines. We also focus on the functioning of web search engines and their history. For the illustration of the website ranking methods, we developed our own simple search engine that implements a website ranking algorithm. It is based on PageRank algorithm with certain modifications. For proper functioning, the algorithm requires access to its own MySQL database, which is located on local server. The graphic design of the search engine is simple, as it only contains an input field and a button. The search engine as well as the algorithm are written in PHP.

Keywords: search engine, website ranking methods, PageRank.

KAZALO

1 UVOD 9

2 SPLETNI ISKALNIKI 11

2.1 KAKO DELUJEJO SPLETNI ISKALNIKI 11

2.2 POMEMBNOST POVRATNIH POVEZAV V SPLETNIH ISKALNIKIH 14

2.3 ČASOVNI RAZVOJ SPLETNIH ISKALNIKOV 16

3 METODE RANGIRANJA 19

3.1 PAGERANK 19

3.1.1 Definicija PageRank-a 19

3.1.2 Poenostavljen algoritem 20

3.1.3 Faktor dušenja 21

3.2 DISTANCERANK 22

3.3 VISUALRANK 24

4 LASTNI SPLETNI ISKALNIK 25

4.1 IZDELAVA 25

4.2 DELOVANJE LASTNEGA SPLETNEGA ISKALNIKA 27

4.2.1 Podatkovna baza 27

4.2.2 Algoritem za rangiranje spletnih strani 30

4.3 TESTIRANJE LASTNEGA SPLETNEGA ISKALNIKA 33

5 SKLEP 36

6 LITERATURA 38

KAZALO SLIK

Slika 3.1: Izsek iz mreže spletnih strani 23
Slika 4.1: Grafični prikaz lastnega spletnega iskalnika 26
Slika 4.2: Rezultati spletnega iskalnika na iskani niz 27
Slika 4.3: Zgradba tabele spletne_strani 28
Slika 4.4: Prikaz vhodnih povezav spletnih strani 29
Slika 4.5: Zgradba tabele vrednosti 30
Slika 4.6: Preverjanje ujemanja iskanega niza z URL-jem 31
Slika 4.7: Preverjanje ujemanja iskanega niza z naslovom strani in s ključnimi besedami 31
Slika 4.8: SQL stavek za seštevek vseh vrednosti v stolpcu ocena 32
Slika 4.9: SQL stavek za izpis vseh strani glede na iskani niz 32
Slika 4.10: Prikaz rezultatov iskanja v lastnem spletnem iskalniku in iskalniku Google. 34
Slika 4.11: Prikaz rezultatov na mobilnem telefonu 35

SEZNAM KRATIC

HTML – označevalni jezik za izdelavo spletnih strani (ang. Hyper Text Markup Language)

CSS – stilska podloga, ki skrbi za grafično obliko spletnih strani (ang. Cascading Style Sheets)

MySQL – sistem za upravljanje s podatkovnimi bazami (ang. My Structured Query Language)

PHP – odprtokodni programski jezik (ang. PHP Hypertext Preprocessor)

FTP – standardni omrežni protokol, ki se uporablja za prenos datotek med računalniki (ang. File Transfer Protocol)

TCP/IP – je množica protokolov, ki izvaja protokolski sklad prek katerega teče internet (ang. Transmission Control Protocol/Internet Protocol)

SEO – optimizacija spletnega iskalnika (ang. Search Engine Optimization)

URL – naslov spletnih strani v svetovnem spletu (ang. Uniform Resource Locator)

Spletni pajek – del iskalnega algoritma, ki obiskuje spletne strani in jih indeksira (ang. Crawler)

Protokol Gopher - je TCP/IP aplikacijski nivo protokola, namenjen za distribucijo, iskanje in pridobivanje dokumentov preko interneta. Protokol Gopher je bil predstavljen kot alternativa svetovnemu spletu v začetni fazi, vendar je na koncu HTTP postal prevladujoč protokol

Poizvedba – beseda ali besedna zveza, ki jo uporabnik išče na spletu (ang. Query)

Inktomi - je ameriško podjetje s sedežem v Kaliforniji, ki je omogočilo programsko opremo ponudnikom internetnih storitev

Looksmart - je ameriško spletno oglaševalno podjetje, ustanovljeno leta 1995

Picsearch - je švedsko podjetje, ki razvija in zagotavlja slikovne iskalne storitve za večja spletna mesta

Direct Hit – podjetje s spletnim iskalnikom, ki je omogočalo storitve spletnega iskanja večjim spletnim mestom in upravljalo javni iskalnik na spletnem mestu DirectHit.com

Teoma – spletni iskalnik, ustanovljen leta 2000

Računalništvo v oblaku – je slog računalništva, pri katerem so dinamično razširljiva in pogosto virtualizirana računalniška sredstva na voljo kot storitev preko interneta (ang. Cloud computing)

Uporabniški agent – programska oprema, ki je neposredno vpletena v pomoč uporabnikom na spletu (ang. User agent)

Hiperpovezava – referenca do dokumenta, ki ji lahko uporabnik spleta direktno sledi preko grafike ali niza besed (ang. Hyperlink)

Spam – nezaželjena pošta je pošiljanje enakih ali podobnih sporočil na veliko število naslovov

Sidro besedila – ponavadi daje uporabniku ustrezne opisne ali kontekstualne informacije o vsebini povezave (ang. Anchor text)

Farme povezav – gre za spletno stran ali skupino strani, ki umetno ustvarja zunanje povezave, koristne pri optimizaciji strani (ang. Link farms)

1 UVOD

Spletni iskalniki obstajajo že vse od leta 1990 in od takrat naprej se je njihovo število le zviševalo, njihova kakovost pa se je stalno izboljševala. Vsi poznamo najbolj znan spletni iskalnik Google, kateri si lasti več kot 70% iskalnega trga. Prav Google je za večino ljudi glavni vir informacij, ki jih dobijo preko spleta.

Glavni cilj spletnih iskalnikov je poiskati in organizirati podatke, ki so na voljo na spletu. Njihov namen je, da bi ljudje lažje brskali po spletu. Tako je eden od najpomembnejših izzivov v kateremkoli spletnem iskalniku najti kakovostne spletne strani in jih razvrstiti od najbolj pomembnih do manj pomembnih glede na uporabnikovo poizvedbo. Za iskanje spletnih strani skrbijo roboti, ki jih imenujemo pajki. Te se plazijo od strani do strani in jih indeksirajo. Tako pridemo do problema rangiranja oziroma razvrščanja spletnih strani, ki temeljijo na zahtevah ali preferencah uporabnikov. Da bi naredili splet bolj zanimiv in produktiven, potrebujemo dober in učinkovit algoritem rangiranja. To bo omogočilo spletnim iskalnikom prikazati najboljše strani za uporabnika, glede na njihove poizvedbe.

Seveda pa se algoritmi za rangiranje razlikujejo glede na iskalnik in glavni razlog za to je, da so vsi algoritmi med najbolj varovanimi skrivnostmi podjetij, ki posedujejo določen iskalnik. To je torej razlog, da prihaja do razlik v rangiranju strani med spletnimi iskalniki. A treba je vedeti, da vsi algoritmi temeljijo na analiziranju HTML kode spletne strani. Ko govorimo o analizi HTML kode mislimo na, kje na strani se pojavijo določene ključne besede, katere besede so napisane z uporabo krepkega besedila in v kakšni velikosti, koliko povezav se pojavi na strani, itd. Lahko bi našteli še kar nekaj primerov, a kot lahko vidimo iz zgornjih primerov, algoritmi upoštevajo ogromno elementov pri analizi spletne strani in s tem posledično tudi pri rangiranju spletnih strani.

Namen moje diplomske naloge je predstaviti delovanje spletnih iskalnikov, algoritme rangiranja, ki potekajo v ozadju iskalnikov, izdelati lastni spletni iskalnik in obenem implementirati vanj lastni algoritem za rangiranje spletnih strani.
Nalogo sem razdelil na tri dele. V prvem delu sem se posvetil spletnim iskalnikom. Predstavil sem njihovo delovanje, vse od plazenja pajkov, indeksiranja spletnih strani, ki jih odkrijejo, do iskanja teh strani v spletnih iskalnikih. Nato sem se osredotočil na povezave, ki jih odkrijejo prej omenjeni pajki, jih opisal in razložil zakaj so te povezave pomembne. Za konec prvega dela sem pregledal še časovni razvoj spletnih iskalnikov, in sicer vse od leta 1990, torej od prvega iskalnika, naprej.

V drugem delu sem se posvetil trem algoritmom rangiranja spletnih strani, in sicer algoritmom PageRank, DistanceRank in VisualRank. Najbolj sem se osredotočil na algoritem PageRank, saj je znano, da ga uporablja tudi najbolj znan iskalnik Google in zaradi tega, ker sem ga uporabil tudi v zadnjem delu diplomske naloge.

Tretji in hkrati zadnji del diplomske naloge sem namenil izdelavi lastnega spletnega iskalnika in poleg tega implementiral še lastni algoritem za rangiranje spletnih strani. Algoritem kot tudi iskalnik sem izdelal v programskem jeziku PHP in ga na koncu tudi testiral.

2 SPLETNI ISKALNIKI

Spletni iskalnik je namenjen iskanju informacij na spletu, katerih iskalni izidi so običajno prikazani v obliki seznama. Količina vsebine in posodabljanje podatkov se opira zgolj na to, kako pogosto se posodablja baza podatkov. Poizvedovalni algoritmi (računalniško programirane metode, ki razvrščajo rezultate iskanja), katerega vsak izmed večjih spletnih storitev uporablja za razvrščevalne namene, so tudi precej edinstveni za vsako posamezno storitev.

Splet raste eksponentno. Raziskave, ki so bile opravljene v letu 2013, so odkrile, da je bilo to leto dodanih približno 103 milijone spletnih strani.Zato ni mogoče, da bi kadarkoli kakšen iskalnik imel celoten splet na svojem trdem disku in ga posodabljal vsak dan. To je velik problem za iskalnike, saj preprosto ne morejo dohajati rasti spleta in nenehnih sprememb, ki so v teku na spletnih straneh. (Facts Hunt 2014)

2.1 KAKO DELUJEJO SPLETNI ISKALNIKI

Naslednje je povzeto po Make use of: How do search engines work (Bruce 2013). Delovanje spletnega iskalnika lahko razdelimo na tri osnovne stopnje. Kljub temu, da razčlenimo anatomijo iskalnika na tri različne komponente, zaradi lažjega razumevanja procesa, moramo poudariti, da to ni linearen proces. Veliko stvari se dogaja ob istem času in nekatere komponente so bolj tesno povezane kot druge.

1. plazenje (ang. crawling) – odkrivanje vsebine v svetovnem spletu,
2. indeksiranje (ang. indexing) – analiza in hranjenje v ogromnih podatkovnih bazah,
3. iskanje (ang. retrieval) – uporabniška poizvedba pridobi seznam ustreznih spletnih strani.

Prvi korak, plazenje robotov oziroma pajkov, je torej začetek procesa. Je pridobivanje podatkov o spletni strani. Iskalniki ohranjajo svoje metode za plazenje in razvrstitev spletnih strani kot poslovne skrivnosti. Vsak iskalnik ima svoj edinstven sistem. Čeprav se algoritmi, ki jih iskalniki uporabljajo, lahko razlikujejo od enega iskalnika do drugega, obstaja veliko podobnosti v načinu kako gradijo svoje indekse. Obstajajo številni zapleti, s katerimi se srečujejo pajki iskalnikov, kot so velikost spleta, njegova nenehna rast in spreminjajoče se okolje. Informacije na spletu, so podane preko milijonov različnih spletnih strežnikov. To pomeni, da je treba informacije najprej zbrati in nato sistematično postaviti v velike zbirke, preden se prenesejo naprej za indeksiranje.

Pajki so avtomatska programska oprema, ki so upravljani najpogosteje s strani iskalnikov. Te pregledujejo splet s sledenjem hiperpovezav na spletnih straneh in zbiranjem sprva tekstovnih podatkov in nato še ostalih podatkov za ustvarjanje indeksov. Čeprav je plazenje dejansko zelo hiter proces, konceptualno, pajek počne isto stvar kot obiskovalec spleta. Pajek najprej pridobi URL in se nato poveže z oddaljenim strežnikom, kjer stran gostuje. Zatem izda zahtevek, da naloži stran ter njeno tekstovno vsebino, nato pregleda povezave, ki jih stran vsebuje in se postavi v čakalno vrsto za nadaljnje plazenje. Ker pajek deluje samodejno in prenaša tekstovne podatke, kot so besedila, metapodatke, komponente HTML, alternativne oblike datotek in URL-je za analizo in nadaljnje plazenje, je sposoben skočiti iz ene strani na drugo preko povezav, ki jih je pregledal ob zelo visokih hitrostih. Ko pride do strani brez povezav, katerim bi lahko sledil, skoči na povezave, ki jih je morda pred tem zamudil ali pa na povezave, ki čakajo v čakalni vrsti za prihodnje plazenje. Postopek se ponavlja iz strežnika na strežnik, dokler ni več strani, ki bi jih lahko prenesel. Pajki uporabljajo tradicionalne grafe za pregledovanje spleta. Graf je sestavljen iz tako imenovanih vozlišč in robov. Vozlišča so URL-ji, robovi pa so povezave, vgrajene na straneh. Izhodne povezave (ang. outbound links) so povezave z vaše spletne strani, ki kažejo na druge strani. Vhodne povezave (ang. inbound links) pa so tiste, ki kažejo nazaj na vašo stran od nekje drugje.

Iskalnik Google, ki ima največji tržni delež med spletnimi iskalniki, v letu 2014 indeksiral več kot 30 miljard spletnih strani. (Statistic Brain 2015)

Pri drugem koraku je potrebno vse te strani razvrstiti glede na vsebino ter druge faktorje in tako ustvariti indeks vseh spletnih strani. Informacije, ki jih pajek najde na vsaki strani, se analizirajo, razčlenijo, ustvari se seznam besed in besednih zvez v dokumentu, pri čemer se upošteva:

- kolikokrat je beseda ali besedna zveza uporabljena na spletni strani (iskalniki zabeležijo spletno stran, na kateri je beseda ali besedna zveza, ki se prevečkrat uporablja, za nezaželeno (ang. spam))
- teža besede ali besedne zveze (teža besede ali besedne zveze poveča vrednost glede na to, kje se nahaja: v vrhu dokumenta, naslovu, podnaslovu, besedilnih povezavah, meta oznakah, ipd.)

Indeksirane informacije so shranjene v bazi podatkov in čakajo na nekoga, da prične z iskanjem. Ko prične z iskanjem v iskalniku, se besede, ki so jih vpisali v iskalno polje primerjajo z indeksiranimi podatki v podatkovni bazi iskalnika. Tako se ustvari seznam spletnih strani, ki je najprimernejši glede na iskano poizvedbo. Namen shranjevanja indeksa je optimizirati hitrost in učinkovitost pri iskanju ustreznih dokumentov za iskalno poizvedbo. Brez indeksa, bi iskalnik pregledal vsak dokument, kar bi zahtevalo veliko časa in računalniške zmogljivosti. Medtem, ko lahko indeks 10.000 dokumentov poizveduje v nekaj milisekundah, lahko zaporedno pregledovanje vsake besede v 10.000 velikih dokumentih traja ure.

Zadnji korak pa lahko vidimo tudi sami, in sicer je to prikaz rezultatov strani v spletnem iskalniku. Vanj vnesemo poizvedbo in iskalnik poskuša prikazati najbolj ustrezne dokumente, ki jih najde v celotnem indeksu in kateri se ujemajo z našo poizvedbo. To je najbolj zapleten korak, ampak tudi najpomembnejši. To je tudi področje, na katerem se iskalniki najbolj razlikujejo med seboj. Nekateri delajo s ključnimi besedami, drugi vam omogočajo, da zastavite vprašanje ali vključujejo napredne funkcije, kot so ključne besede glede na bližino ali filtriranje po starosti vsebine. Algoritem rangiranja preveri

vašo poizvedbo med milijardami strani in poizkuša ugotoviti kako pomembna je katera stran glede na vašo poizvedbo. Ta operacija je tako kompleksna, da podjetja skrbno varujejo svoje algoritme rangiranja, in sicer kot patentirane industrijske skrivnosti.

2.2 POMEMBNOST POVRATNIH POVEZAV V SPLETNIH ISKALNIKIH

Naslednje je povzeto po The Importance of Backlinks (WebConfs 2013). Povratne povezave so povezave, ki so usmerjene proti spletni strani. Poznane so tudi kot vhodne povezave. Število povratnih povezav je pokazatelj priljubljenosti ali pomembnosti te spletne strani. Povratne povezave so pomembne za optimizacijo spletnega iskalnika, saj nekateri iskalniki, namenjajo večjo vrednost spletnim stranem, ki imajo precejšnje število kakovostnih povratnih povezav in bodo obravnavali te spletne strani kot bolj pomembne od drugih v svojih straneh z rezultati glede na iskalno poizvedbo.

Ko iskalniki izračunajo ustreznost strani glede na ključno besedo, upoštevajo število kakovostnih vhodnih povezav na to stran. Zaradi tega avtorji spletnih strani ne smejo biti zadovoljni zgolj s pridobivanjem čim večjega števila vhodnih povezav, ampak morajo gledati predvsem na kakovost vhodnih povezav.

Iskalnik obravnava vsebino spletnih strani za določitev kakovosti povezav. S prihodom vhodnih povezav na določeno spletno stran iz drugih strani, katerih vsebina se navezuje na prvo stran, se te vhodne povezave obravnavajo kot bolj pomembne za prvo stran.

V primeru, da se na spletni strani najdejo vhodne povezave, katerih vsebina ni povezana s to stranjo, se štejejo za manj pomembne. Višja je pomembnost vhodnih povezav, večja je njihova kakovost. Na primer, če ima nekdo spletno stran o košarki in prejme povratno povezavo iz druge spletne strani o košarki, potem bi bila ta povezava bolj pomembna pri oceni iskalnika, kot pa povezava, ki bi prišla s spletne strani o hrani. Bolj kot je pomembna stran, ki ima povezavo nazaj na vašo spletno stran, boljša je kakovost povratne povezave.

Spletni iskalniki želijo, da imajo vse spletne strani enake pogoje pri ustvarjanju povratnih povezav in da bi se le-te ustvarjale počasi skozi določeno časovno obdobje. Čeprav je dokaj enostavno manipulirati povezave na spletno stran, za doseg čim višje razvrstitve, je veliko težje vplivati na iskalnik z zunanjimi povratnimi povezavami iz drugih spletnih strani. To je tudi razlog, da je faktor povratnih povezav v algoritmih iskalnikov tako visoko. Kriterij iskalnika glede kakovosti vhodnih povezav se je še zaostril, zaradi brezvestnih avtorjev spletnih strani, ki se trudijo za dosego teh vhodnih povezav z varljivo ali zahrbtno tehniko, kot so skrite povezave ali samodejno ustvarjene strani, katerih edini namen je, da zagotovijo vhodne povezave na spletne strani. Te strani se imenujejo »farme povezav« (ang. link farms) in ne le, da jih iskalniki zanemarijo, ampak lahko te povezave tudi blokirajo oziroma prepovedo spletno stran v celoti. Še eden izmed razlogov, da se dosežejo kakovostne povratne povezave je, da bi avtorji spletnih strani pritegnili obiskovalce, kateri bi obiskali njihovo stran. Težko je pričakovati, da bodo ljudje sami našli spletno stran, brez da bi jim avtor strani pokazal kako priti do nje. Eden od načinov, kako priti do strani je namreč preko medsebojnih povezav.

V številnih posodobitvah iskalnika Google so medsebojne povezave eden od ciljev najnovejšega filtra iskalnika. Veliko avtorjev spletnih strani se je dogovorilo o izmenjavi medsebojnih povezav, da bi tako zvišali uvrstitev svoji spletni strani s samim številom vhodnih povezav. V izmenjavi povezav avtor spletne strani postavi povezavo na svojo spletno stran, ki kaže na drugo stran in obratno. Mnoge od teh povezav preprosto niso bile pomembne in zaradi tega niso bile upoštevane. V izdelavi je tudi patent iskalnika Google, ki se bo ukvarjal ne le z priljubljenostjo spletnih strani na katere kažejo povezave, ampak tudi, kako zaupanja vredne so te strani. Ker iskalnik Google skrbno skriva podatke o vseh stvareh, ki se nanašajo na iskalnik in algoritme, še ni jasno ali je patent iskalnika Google, ki se bo ukvarjal s tem kako zaupanja vredne so strani, že v uporabi.

Avtorji spletnih strani pa morajo biti previdni tudi glede povezovanja spletnih strani z istim IP naslovom. Avtorjem večih sorodnih spletnih strani lahko povezava do vsake od teh spletnih strani škodi, saj bo v tem primeru iskalnik mislil, da poskušajo narediti nekaj

sumljivega. Veliko avtorjev spletnih strani poskuša manipulirati s povratnimi povezavami prav na ta način. Preveč povezav do spletnih strani z istega IP naslova, imenujemo bombardiranje povratnih povezav.

Obstaja še en način, s katerim lahko avtorji spletnih strani pridobijo kakovostne povratne povezave na njihovo spletno stran, in to imenujemo »sidro besedila« (ang. anchor text). Ko povezava vključuje ključne besede v besedilu hiperpovezave, pravimo temu kakovostno sidro besedila. Namesto, da se uporabijo besede, kot so »kliknite tukaj«, se raje uporabijo besedne zveze, kot so »Rezultati tekem«. To je veliko boljši način, da se izkoristi hiperpovezava.

2.3 ČASOVNI RAZVOJ SPLETNIH ISKALNIKOV

Razvoj spletnih iskalnikov se je začel v začetku leta 1990. Ker se je število spletnih strani na spletu povečevalo iz dneva v dan, so za namen lažjega iskanja podatkov, datotek in drugih stvari, na spletu razvili iskalnike.

1990 – Archie -> Prvi spletni iskalnik se je razvil kot šolski projekt, katerega je ustvaril Alan Emtage, študent na Univerzi McGill v Montrealu. Prvotni namen imena je bil »archive«, vendar pa je bil nato skrajšan na Archie. Spletni iskalnik Archie je indeks računalniških datotek, shranjenih na anonimnih FTP spletnih straneh v določenem omrežju računalnikov. Zaradi omejenega prostora, so bili na voljo le seznami in ne vsebina za vsako spletno stran. Iskanje je potekalo po podatkovni bazi na podlagi imena datoteke, spletni iskalnik Archie pa vam je povedal, v katerem omrežju se nahaja imenik, ki vsebuje kopijo datoteke, ki jo želite. (The history of SEO 2011)

1991 – Veronica -> Išče imena datotek in naslove, s pomočjo protokola Gopher.
Na svojem vrhuncu uporabe, je spletni iskalnik Veronica iskal preko baze podatkov z več kot 5.500 Gopher strežniki in več kot 10 milijonov Gopher dokumenti, katerih naslovi so vsebovali vašo ključno besedo. (The history of SEO 2011)

1992 – Vlib -> Ustvarjen s strani Tim Berners-Lee-ja kot virtualna knjižnica. Ljudje so lahko dostopali do velikega števila dokumentov iz več gostujočih strežnikov. Iskanje je bilo zelo osnovno, a je bilo vsebine, ki se je nahajala na strežnikih, ogromno v primerjavi z drugimi iskalniki, ki so se pojavili pred spletnim iskalnikom VLib. Te strežniki so sčasoma postali dostopni za ljudi po vsem svetu in zaradi tega je spletni iskalnik VLib postal znan kot zelo zgodnji temelj svetovnega spleta. (The history of SEO 2011)

1993 (Februar) – Excite -> Ustvarjen s strani šestih dodiplomskih študentov z univerze Stanford. Te so imeli idejo o uporabi statistične analize besednih odnosov, da bi naredili iskanje bolj učinkovito. (The history of SEO 2011)

1994 (Januar) – InfoSeek -> Ni prinesel veliko inovacij v svet spletnih iskalnikov, vendar so kljub vsemu ponudili nekaj dodatkov. Ena izmed teh je bila, da so omogočili avtorju strani, da objavi spletno stran v indeks iskanja v realnem času. (WordStream 2015)

1994 (April) – Yahoo! Directory -> Ustvarila sta ga David Filo in Jerry Yang. Prvotno je bil to zelo cenjen imenik spletnih strani, katere so bile katalogizirane s strani urednikov. Zaradi vedno večjega števila povezav so ga morali preoblikovati in narediti iskalni imenik. (WordStream 2015)

1994 (April) – WebCrawler -> Prvi pajek, ki je indeksiral celotne strani. Kmalu je postal tako priljubljen, da ga podnevi ni bilo mogoče uporabljati. (WordStream 2015)

1994 (Julij) – Lycos -> Lycos je bil prvotno razvit za izračun velikosti spleta z uporabo pajka. Ta se je plazil po internetu od strani do strani preko spletnih povezav. (Wall 2015)

1995 (December) – AltaVista -> Prinesel veliko pomembnih funkcij na spletno prizorišče. Prvi iskalnik, ki je omogočal poizvedbe v svojem (naravnem) jeziku, napredne tehnike iskanja in omogočanje uporabnikom, da so dodali ali odstranili svoj spletni

naslov v roku 24 ur. Zagotovili so tudi številne nasvete za iskanje in napredne iskalne funkcije. (WordStream 2015)

1996 (Maj) – HotBot -> To je bil zelo priljubljen iskalnik v 90-ih letih, bil je živih barv in vračal je odlične rezultate. (Wall 2015)

1997 (April) – Ask Jeeves/Ask.com -> Zagnan kot spletni iskalnik, ki je omogočal poizvedbe v naravnem jeziku. Imel je več kot 100 urednikov, ki so spremljali kaj so ljudje iskali, nato pa ročno izbirali strani, ki so se jim zdele najprimernejše glede na poizvedbo. (WordStream 2015)

1998 (September) – Google -> Google-ova sposobnost analiziranja povezav s spleta je pomagala proizvesti novo generacijo zelo pomembnih rezultatov, ki temeljo na plazenju robotov. Danes najbolj priljubljen iskalnik v uporabi. (The history of SEO 2011)

1998 – MSN -> Spletni iskalnik MSN Search je bila storitev ponujena kot del omrežja Microsoftovih spletnih storitev. V letu 2007 so spletni iskalnik preimenovali v Live Search. Kasneje, leta 2009, so še drugič spremenili ime spletnega iskalnika, in sicer v Bing. (The history of SEO 2011)

1998 – Overture -> Prej poznan kot GoTo.com. V letu 2001 so spremenili ime v Overture. Prvi, ki so začeli izvajati storitev imenovano »razvrstitev glede na plačilo« (ang. paid placement), kjer so bile strani razvrščene glede na to, koliko so bila podjetja pripravljeni plačati za svojo spletno stran. (WordStream 2015)

3 METODE RANGIRANJA

Metode rangiranja so računalniški procesi in formule, ki sprejmejo vaša vprašanja in jih spremenijo v odgovore. So skrbno varovane skrivnosti, in sicer iz dveh razlogov: podjetja spletnih iskalnikov želijo zaščititi svoje metode rangiranja pred svojimi konkurenti in želijo otežiti delo avtorjem spletnih strani, da le-te ne morejo manipulirati z razvrstitvijo svojih strani v iskalnikih. Spodaj sem opisal tri metode rangiranja, ki so med bolj znanimi.

3.1 PAGERANK

PageRank je dobil ime po Larry Page-u, enemu od ustanoviteljev iskalnika Google. Deluje tako, da izračuna število in kakovost vhodnih povezav na spletno stran in tako določi grobo oceno o tem, kako pomembna je spletna stran. Osnovna predpostavka je, da bolj pomembne so spletne strani, več vhodnih povezav sprejmejo od drugih spletnih strani. To ni edini algoritem, ki ga uporablja Google za razvrščanje rezultatov iskalnika, vendar pa je prvi algoritem, ki ga je podjetje uporabljalo, in zato tudi najbolj znan. Velja za eno izmed prvih tehnik analize, ki temeljijo na osnovi povezave, ki bi povečala učinkovitost pridobivanja informacij na spletu.

3.1.1 Definicija PageRank-a

Torej bi lahko postavili začetno in intuitivno definicijo PageRank-a, ki pravi, da je PageRank strani *k*, nenegativno realno število, ki ga določi sistem enačb:

$$P_r(k) = \sum_{h \to k} \frac{P_r(h)}{o(h)}, \qquad k = 1,2,\dots,n$$

kjer je $P_r(h)$ PageRank strani *h*, $o(h)$ je število izhodnih povezav strani *h* in vsota je razširjena na vse spletne strani *h*, ki kažejo na stran *k*; *n* je število strani na spletu. Če ima stran *h* veliko izhodnih povezav na isto stran *k*, vse te izhodne povezave štejejo kot ena. Po tej definiciji rang strani ni odvisen le od števila strani, ki kažejo na to stran, ampak prav tako od njihove pomembnosti. (Preto 2002)

Algoritem PageRank izhaja iz verjetnostne porazdelitve, ki se uporablja za prikazovanje verjetnosti, da bo oseba, ki naključno klika na povezave prispela do katere koli strani. PageRank je mogoče izračunati za zbirko dokumentov vseh velikosti. Verjetnost je izražena kot številčna vrednost med 0 in 1. Verjetnost 0,5 je običajno izražena kot »50% možnost«, da se nekaj zgodi. Torej, PageRank 0,5 pomeni, da obstaja 50% verjetnost, da bo oseba, s klikom na naključno povezavo usmerjena k dokumentu z PageRank vrednostjo 0,5. (Preto 2002)

3.1.2 Poenostavljen algoritem

Naslednje je povzeto po PageRank (Wikipedia 2015). Da bi lažje razumeli algoritem PageRank, spodaj podajam poenostavljem algoritem oziroma poenostavljen izračun vrednosti algoritma PageRank. Predpostavimo majhen nabor štirih spletnih strani: A, B, C in D. PageRank je nastavljen na enako vrednost za vse strani. PageRank predpostavlja verjetnostno porazdelitev med 0 in 1, kar pomeni, da je začetna vrednost za vsako stran 0,25.

Vrednost PageRank-a se prenese iz določene strani do ciljev njenih izhodnih povezav, ob naslednji ponovitvi pa se enakomerno porazdeli med vse izhodne povezave. Če bi bile edine povezave v sistemu s strani, B, C in D, do strani A, bi vsaka povezava prenesla 0,25 vrednosti PageRank-a na naslednjo ponovitev, za skupno vrednost 0,75.

$$PR(A) = PR(B) + PR(C) + PR(D)$$

Denimo, namesto da bi stran B imela povezavo do strani C in A, bi imela stran C povezavo do strani A in stran D bi imela povezavo do vseh treh strani. Torej, po prvi ponovitvi, bi stran B prenesla polovico obstoječih vrednosti, ali vrednost 0,125 na stran A in drugo polovico, ali vrednost 0,125 na stran C. Stran C bi prenesla vse svoje obstoječe vrednosti, 0,25, na edino stran s katero je povezana, torej na stran A. Ker ima stran D tri izhodne povezave, bi prenesla eno tretjino svoje obstoječe vrednosti ali približno vrednost 0,083, na A. Ob zaključku te ponovitve, bo stran A imela vrednost PageRank-a 0,458.

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}$$

V splošnem primeru je lahko vrednost PageRank-a za vsako stran *u* izražena kot:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)},$$

tj. vrednost PageRank-a za stran *u* je odvisna od vrednosti PageRank-a za posamezno stran *v* vsebovano v naboru B_u (nabor vsebuje vse strani, ki so povezane na stran *u*), deljeno s številom povezav iz strani *v* (L(v)).

3.1.3 Faktor dušenja

Teorija PageRank meni, da bo namišljeni uporabnik spleta, ki naključno klika na povezave, sčasoma končal s klikanjem. Verjetnost, na kateremkoli koraku, da bo oseba nadaljevala, je faktor dušenja *d*. Različne študije so preizkusile različne faktorje dušenja, vendar se na splošno domneva, da se bo faktor dušenja ustavil oziroma nastavil pri okoli 0,85. Faktor dušenja se odšteje od 1 (in v nekaterih različicah algoritma, se rezultat deli s

številom dokumentov (*N*) v zbirki) in ta izraz nato dodamo k produktu faktorja dušenja in vsoti dohodnega PageRank rezultata. (Wikipedia 2015)

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

$p_1, p_2, \ldots, p_n$ – strani, katerih pomembnost nas zanima

M(p_i) – množica strani, katerih izhodne povezave kažejo na stran p_i

L(p_j) – število odhodnih povezav na stran p_j

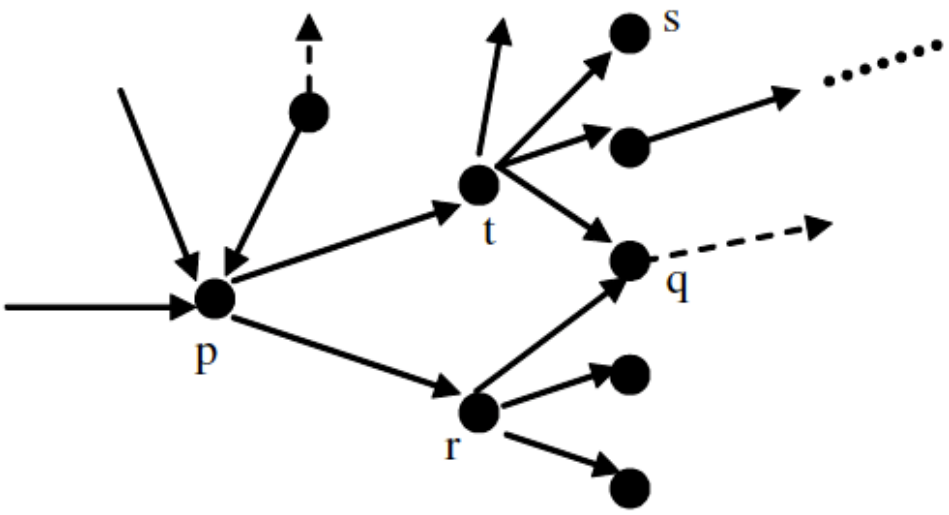
N – število vseh strani

3.2 DISTANCERANK

Razdalja je definirana kot število »povprečnih klikov« med dvema stranema. Cilj je zmanjšati razdaljo, tako da ima stran z manjšo razdaljo višjo uvrstitev. Eksperimentalni rezultati kažejo, da DistanceRank prekaša druge algoritme v razvrščanju strani in razporejanju plazenja. Poleg tega je računska kompleksnost algoritma DistanceRank nizka. (Zareh Bidoki in Yazdani 2008)

Razdalja med stranmi velja za kazen. Pri tej metodi je glavni cilj zmanjšati vsoto prejetih kazni (razdaljo) s strani uporabniškega agenta, tako da bo stran z nizko razdaljo bila višje rangirana. (Zareh Bidoki in Yazdani 2008)

Slika 3.1: Izsek iz mreže spletnih strani



Vir: Zareh Bidoki in Yazdani (2008)

Definicija 1: Če stran *i* kaže na stran *j*, je teža povezave med *i* in *j* enaka $\log_{10} O(i)$, kjer O(*i*) prikazuje izhodno stopnjo povezav strani *i*. (Zajko 2012)

Definicija 2: Razdalja med dvema stranema *i* in *j* je teža najkrajše poti (pot z najmanjšo vrednostjo) od *i* do *j*. Temu pravimo logaritemska razdalja in jo označujemo z d_{ij}.
Primer (slika 3.1): teža izhodnih povezav na straneh *p*, *r* in *t* je enaka $\log(2)$, $\log(3)$ in $\log(4)$ v tem zaporedju in razdalja med *p* in *q* je enaka $\log(2) + \log(3)$ če je pot *p*-*r*-*q*, najkrajša pot med *p* in *q*. Kot kaže slika 1, je razdalja med *p* in *s* $\log(2) + \log(4)$. Čeprav sta obe *s* in *q* na isti ravni povezave iz *p* (dva klika), je *q* bližje *p*. (Zareh Bidoki in Yazdani 2008)

Definicija 3: Če d_{ij} prikazuje razdaljo med dvema stranema *i* in *j* kot Definicija 2, nato d_i označuje povprečno razdaljo strani *j* in je definirana kot:

$$d_i = \frac{\sum_{j=1}^{V} d_{ij}}{V}$$

kjer *V* prikazuje število spletnih strani. (Zajko 2012)

3.3 VISUALRANK

VisualRank je sistem za iskanje in razvrstitev slik z analiziranjem in primerjanjem njihove vsebine, namesto da iščejo imena slik, spletne povezave ali drugo besedilo.

Čeprav je iskanje slik postala priljubljena funkcija na številnih spletnih iskalnikih, vključno z Yahoo!, MSN, Google, itd., večina slikovnega iskanja uporablja zelo malo, če sploh kaj, informacij glede slik. Glede na uspeh, ki temelji na besedilnem iskanju spletnih strani, večina iskalnikov vrne slike, ki temeljijo izključno na besedilu strani, iz katerih so povezane slike. (Ying in Baluja 2008)

VisualRank je sistem, za izboljšanje Googlovih rezultatov iskanja slik s poudarkom na zanesljivejšem in učinkovitejšem izračunu slikovnih podobnosti, ki velja za veliko število poizvedb in slik. (Ying in Baluja 2008)

VisualRank uporablja intuicijo slučajnega sprehoda za razvrstitev slik, ki temeljijo na vizualnih hiperpovezavah med slikami. Intuicija uporabe teh vizualnih hiperpovezav je, da če si uporabnik ogleduje sliko, so lahko druge sorodne (podobne) slike tudi v interesu. Še posebej, če ima slika *u* vizualno povezavo do slike *v*, potem obstaja verjetnost, da bo uporabnik prišel od slike *u* do slike *v*. Slike, ki so obiskane bolj pogosto, veljajo za pomembnejše. (Ying in Baluja 2008)

4 LASTNI SPLETNI ISKALNIK

4.1 IZDELAVA

Cilj diplomske naloge je bil ustvariti lastni spletni iskalnik in vanj implementirati algoritem za rangiranje spletnih strani. Iskalnik je narejen povsem minimalistično, saj prvi del služi le za vnos iskanih besed in drugi del za prikaz spletnih strani. Narejen je v programskem jeziku PHP, označevalnem jeziku za izdelavo spletnih strani, imenovanem HTML in slogovno predlogo CSS, katera omogoča, da lahko vsakemu elementu v datoteki HTML določimo slog oziroma stil.

Algoritem za rangiranje spletnih strani je napisan v programskem jeziku PHP. Ta se zažene, ko uporabnik vnese iskani niz v vnosno polje spletnega iskalnika in pritisne na gumb Iskanje. Nato se v drugem delu spletnega iskalnika spletne strani razvrstijo na podlagi ocene, ki jo pridobijo s strani algoritma za rangiranje.

Za pravilen izračun ocene mora algoritem imeti dostop do podatkov, kateri se nahajajo v sistemu za upravljanje s podatkovnimi bazami MySQL. Do le teh pa lahko dostopa preko lokalnega strežnika, katerega sem vzpostavil s pomočjo spletnega strežnika Apache.

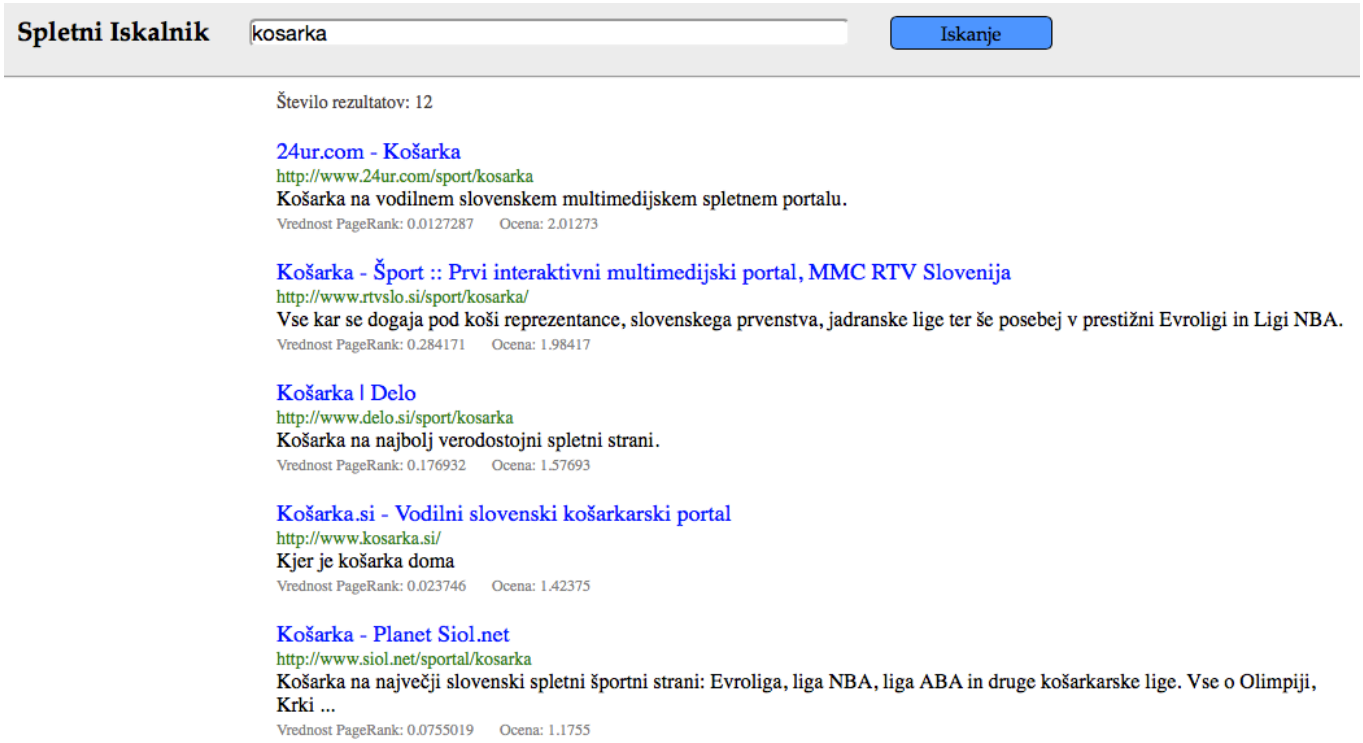
Na sliki 4.1 lahko vidimo izgled prvega dela spletnega iskalnika. Po izgledu vidimo, da je iskalnik povsem preprost, ima samo eno vnosno polje za vnos iskanih besed in pod njim gumb za začetek iskanja.

Slika 4.1: Grafični prikaz lastnega spletnega iskalnika

Po vnosu iskane besede v vnosno polje, se izvede algoritem, ki razvrsti spletne strani na podlagi ocene. Ta je odvisna od števila ključnih besed, spletnega naslova, imena strani, števila izhodnih povezav in vrednosti PageRank-a. Vsi našteti podatki podajo končno oceno.

V zgornjem delu slike 4.2 lahko vidimo iskano besedo, ki smo jo vnesli v vnosno polje. Pod vnosnim poljem je prikazano število vseh rezultatov, ki smo jih dobili za vneseni niz in pod njim imamo razvrščene vse rezultate. Rezultati so sestavljeni iz spletnega naslova, spletne povezave, opisa spletne strani, vrednosti PageRank-a in skupne ocene. Na podlagi skupne ocene so spletne strani tudi razvrščene.

Slika 4.2: Rezultati spletnega iskalnika na iskani niz



4.2 DELOVANJE LASTNEGA SPLETNEGA ISKALNIKA

Lastni spletni iskalnik ima torej implementiran algoritmem za rangiranje spletnih strani, kateri temelji na algoritmu PageRank z določenimi dodatki. V nadaljevanju bom opisal kako točno deluje algoritem za rangiranje spletnih strani. (PHP/ir 2009)

4.2.1 Podatkovna baza

Kot že omenjeno v točki 4.1. se ob pritisku na gumb Iskanje v spletnem iskalniku zažene algoritem za rangiranje spletnih strani. Ta se sprva poveže s podatkovno bazo znotraj ketere imamo ustvarjene tri tabele, in sicer: *spletne_strani*, *povezave* in *vrednosti*.

Na sliki 4.3 lahko vidimo vnose iz tabele *spletne_strani*. Vnešene imamo podatke o vsaki posamezni spletni strani. Kot lahko vidimo so prav vse strani v slovenskem jeziku

oziroma so vse slovenske spletne strani. Vse podatke sem pridobil preko spletne aplikacije Keyword Density Analyzer. (SEO Book 2015)

Izjema je bil le podatek število izhodnih povezav, katerega sem pridobil preko spletne aplikacije Link Extractor. Prav vsi podatki so pomembni, saj sem jih moral primerjati z iskanim nizom. (Get Rank 2015)

Slika 4.3: Zgradba tabele spletne_strani

id	title	description	keywords	url	outgoing_links
ucture	Košarka.si - Vodilni slovenski košarkarski portal	Kjer je košarka doma		http://www.kosarka.si/	58
2	Slovenska košarka	Portal o slovenski košarki	košarka, slovenska, klubi, mladinci, kadeti, člani...	http://www.slovenska-kosarka.si/	0
3	Košarkarska zveza Slovenije: Naslovnica	Uradna stran Košarkarske zveze Slovenije, organiza...		http://www.kzs.si/	34
4	Košarka - Šport :: Prvi interaktivni multimedijski...	Vse kar se dogaja pod koši reprezentance, slovensk...	Košarka, Šport, EP v košarki 2005, SP v košarki 20...	http://www.rtvslo.si/sport/kosarka/	82
5	ABA League	ABA league, official web site of Adriatic Basketba...	aba liga, aba league, adriatic basketball associat...	http://www.abaliga.com/	19
6	Košarka - Planet Siol.net	Košarka na največji slovenski spletni športni stra...		http://www.siol.net/sportal/kosarka	23
7	24ur.com - Košarka	Košarka na vodilnem slovenskem multimedijskem sple...	Košarka	http://www.24ur.com/sport/kosarka	138
8	Ekipa24.si	Ekipa24.si.	Ekipa24, Novice, Šport, Slovenija, Svet, Nogomet, ...	http://www.ekipa24.si/rubrika/kosarka	0
9	Košarka	Dnevnik	Spletno mesto Dnevnik.si sodi med najbolj obiskane...	Košarka	http://www.dnevnik.si/sport/kosarka	0
10	Košarka	Delo	Košarka na najbolj verodostojni spletni strani.	košarka, košarkarske tekme, olimpija, krka, evroli...	http://www.delo.si/sport/kosarka	52
11	Košarka - zurnal24	Košarka na portalu Zurnal24.si. Najnovejše globaln...	košarka	http://www.zurnal24.si/kosarka	13
12	Košarka - Wikipedija, prosta enciklopedija			http://sl.wikipedia.org/wiki/Košarka	10

Tabela *povezave* je pomembna predvsem za izračun vrednosti PageRank-a, saj vsebuje podatke o vhodnih povezavah. Te podatke sem pridobil s pomočjo brezplačnega programa SEO SpyGlass.

Iz slike 4.4 je lepo razvidno, da do spletne strani z id številko 1 (http://www.kosarka.si) kažejo povezave iz spletnih strani z id številkami 1, 2, 3, 4 in 8.

Slika 4.4: Prikaz vhodnih povezav spletnih strani

id	from_page_id	url
1	1,2,3,4,8	http://www.kosarka.si/
2	2,3	http://www.slovenska-kosarka.si/
3	1,2,3,4,5,8,10,12	http://www.kzs.si/
4	3,9	http://www.rtvslo.si/sport/kosarka/
5	1,2,3,4,5,9,10	http://www.abaliga.com/
6	4,5	http://www.siol.net/sportal/kosarka
7	4	http://www.24ur.com/sport/kosarka
8	0	http://www.ekipa24.si/rubrika/kosarka
9	8,9	http://www.dnevnik.si/sport/kosarka
10	3,4	http://www.delo.si/sport/kosarka
11	3,4,10	http://www.zurnal24.si/kosarka
12	12	http://sl.wikipedia.org/wiki/Košarka

Še zadnja tabela v podatkovni bazi je tabela imenovana *vrednosti*, ki vsebuje tudi končno oceno spletnih strani, na podlagi katere so spletne strani razvrščene v iskalniku.

Slika 4.5: Zgradba tabele vrednosti

id	page_rank	odhodne_povezave	url_value	title_value	keywords_value	ocena
1	0.023746	0.4	0	0	0	0.423746
2	0.0469372	0	0	0	0	0.0469372
3	0.0672505	0.1	0	0	0	0.16725
4	0.284171	0.7	0	0	0	0.984171
5	0.184568	0.1	0	0	0	0.284568
6	0.0755019	0.1	0	0	0	0.175502
7	0.0127287	1	0	0	0	1.01273
8	0.0127287	0	0	0	0	0.0127287
9	0.0425147	0	0	0	0	0.0425147
10	0.176932	0.4	0	0	0	0.576932
11	0.0601933	0.1	0	0	0	0.160193
12	0.0127287	0.1	0	0	0	0.112729

Vsi podatki v tabeli, ki jih lahko vidimo na sliki 4.5, so pomembni za algoritem rangiranja, saj jih le ta potrebuje pri končnem razvrščanju spletnih strani.

4.2.2 Algoritem za rangiranje spletnih strani

Algoritem za rangiranje spletnih strani je glavni del spletnega iskalnika, saj glede na iskani niz uporabnika določi ujemanja s podatki, ki se nahajajo v tabeli *spletne_strani*.
Na sliki 4.6 lahko vidimo del kode, ki poišče ujemanje med iskanim nizom in URL-jem. Na sliki 4.7 pa vidimo del kode, ki preveri ujemanje med naslovom strani in ključnimi besedami spletne strani. V primeru, da ujemanje obstaja lahko vidimo, da se vrednosti v podatkovni bazi v tabeli *vrednosti* spremenijo.

Slika 4.6: Preverjanje ujemanja iskanega niza z URL-jem

```
$st_vrstic = 0;
foreach ($domena as $vrednost) {
    $st_vrstic++;
    $a = parse_url($vrednost);
    $b = $a['host'];
    $c = $a['path'];
    if ((stripos($b, $iskani_niz)) || (stripos($c, $iskani_niz)) !== false) {
        mysql_query("UPDATE vrednosti SET url_value = 1 WHERE id = $st_vrstic");
    }
    else {
        mysql_query("UPDATE vrednosti SET url_value = 0 WHERE id = $st_vrstic");
    }
}
```

Slika 4.7: Preverjanje ujemanja iskanega niza z naslovom strani in s ključnimi besedami

```
// ISKANI NIZ SE PRIMERJA Z NASLOVOM STRANI
$st_vrstic = 0;
foreach ($naslov as $vrednost) {
    $st_vrstic++;
    if (stripos($vrednost, $iskani_niz) !== false) {
        mysql_query("UPDATE vrednosti SET title_value = 0.7 WHERE id = $st_vrstic");
    }
    else {
        mysql_query("UPDATE vrednosti SET title_value = 0 WHERE id = $st_vrstic");
    }
}


// ISKANI NIZ SE PRIMERJA S KLJUČNIMI BESEDAMI
$st_vrstic = 0;
foreach ($kljucne_besede as $vrednost) {
    $st_vrstic++;
    if (stripos($vrednost, $iskani_niz) !== false) {
        mysql_query("UPDATE vrednosti SET keywords_value = 0.4 WHERE id = $st_vrstic");
    }
    else {
        mysql_query("UPDATE vrednosti SET keywords_value = 0 WHERE id = $st_vrstic");
    }
}
```

V zadnji stolpec imenovan *ocena* v tabeli *vrednosti* bo algoritem zapisal skupno vrednost vseh preostalih stolpcev v tej tabeli, in sicer tako da bo seštel vse vrednosti po vrsticah med seboj. To naredimo s pomočjo naslednjega stavka SQL:

Slika 4.8: SQL stavek za seštevek vseh vrednosti v stolpcu ocena

```
mysql_query("UPDATE vrednosti SET ocena = (page_rank + odhodne_povezave + url_value + title_value + keywords_value)");
```

V naslednjem in zadnjem koraku algoritem izpiše vse spletne strani, ki nas zanimajo glede na iskani niz in jih s pomočjo naslednjega stavka SQL razvrsti padajoče glede na končno oceno:

Slika 4.9: SQL stavek za izpis vseh strani glede na iskani niz

```
        $rezultat = "SELECT spletne_strani.id, spletne_strani.title, spletne_strani.url, spletne_strani.keywords,
spletne_strani.description, vrednosti.page_rank, vrednosti.ocena
            FROM spletne_strani
            INNER JOIN vrednosti ON (spletne_strani.id = vrednosti.id)
            HAVING keywords LIKE '%$iskani_niz%' OR url LIKE '%$iskani_niz%' OR title LIKE '%$iskani_niz%'
            ORDER BY vrednosti.ocena DESC";
```

4.3 TESTIRANJE LASTNEGA SPLETNEGA ISKALNIKA

Spletni iskalnik sem testiral v spletnem brskalniku Safari. Ker nam program MAMP omogoča, da lahko neposredno preko naslova IP testiramo spletni iskalnik tudi na drugih napravah, sem se odločil, da ga preizkusim še na mobilnem telefonu iPhone 6.

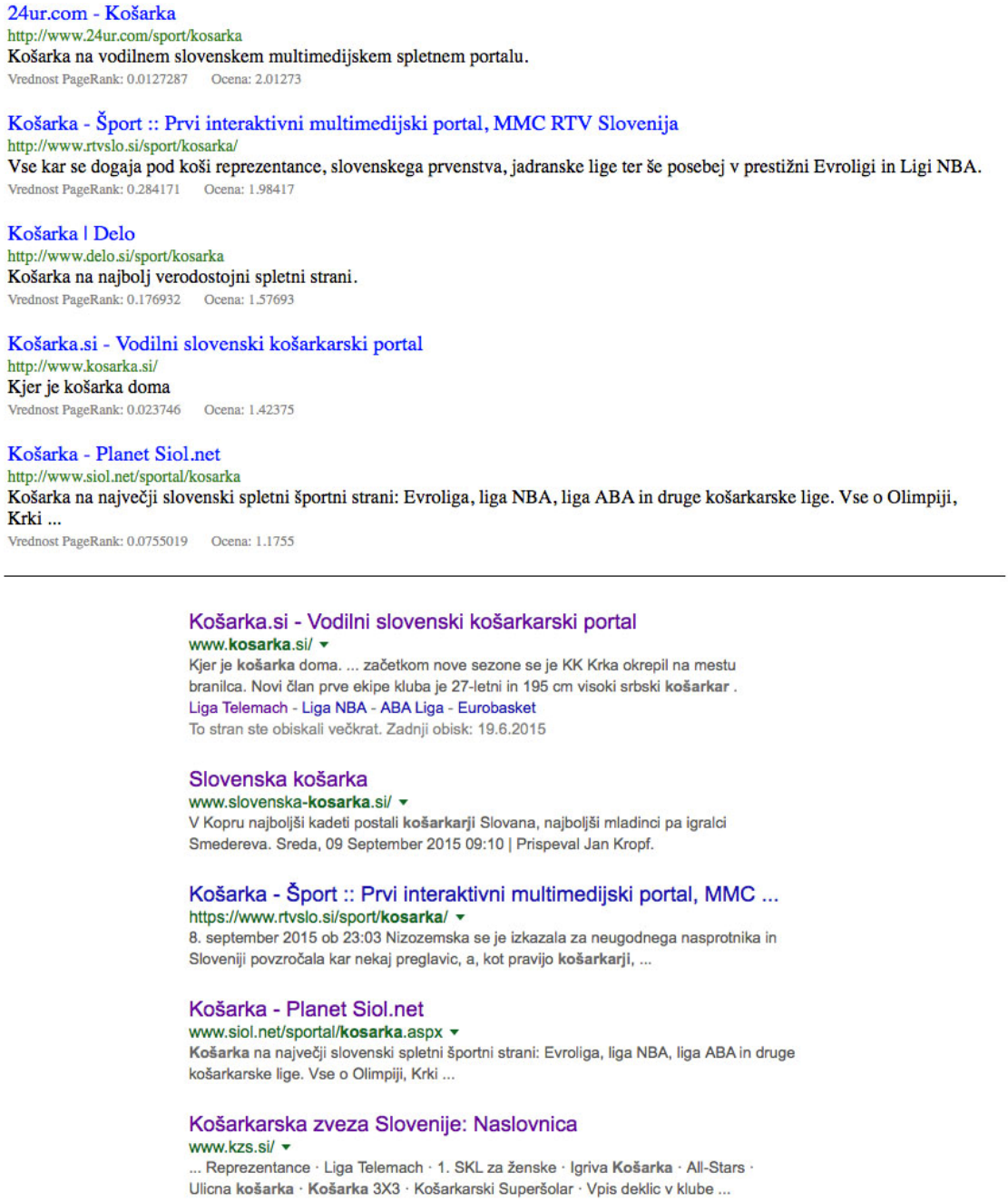
Spletni iskalnik je deloval brez težav in prav tako njegov grafični prikaz. Izakazal pa se je tudi na mobilnem telefonu, saj je bil tudi tam njegov grafični prikaz povsem enak tistemu na osebnem računalniku.

Pri testiranju smo se osredotočili predvsem na delovanje vnosnega polja in razvrstitev rezultatov. Pri vnosnem polju smo se osredotočili na vnos šumnikov, velikih in malih črk, pri razvrstitvi rezultatov pa smo poskušali doseči čim bolj podobno razvrstitev kot jo dosežemo v spletnem iskalniku Google. Razvrstitev rezultatov je temeljila, kot smo že povedali v točki 4.2., na algoritmu PageRank in nekaterih dodatkih.

Slika 4.10 prikazuje primerjavo rezultatov lastnega spletnega iskalnika in iskalnika Google na iskani niz *kosarka*. Potrebno je povedati še, da je bil iskalnik Google s pomočjo naprednega iskanja omejen le na spletne strani v slovenskem jeziku, saj sem tudi sam imel le strani v slovenščini.
Iz slike lahko vidimo, da so v zgornjem delu prikazani rezultati na iskani niz kosarka iz lastnega spletnega iskalnika, v spodnjem delu pa so prikazani rezultati iskanja s spletnega iskalnika Google.

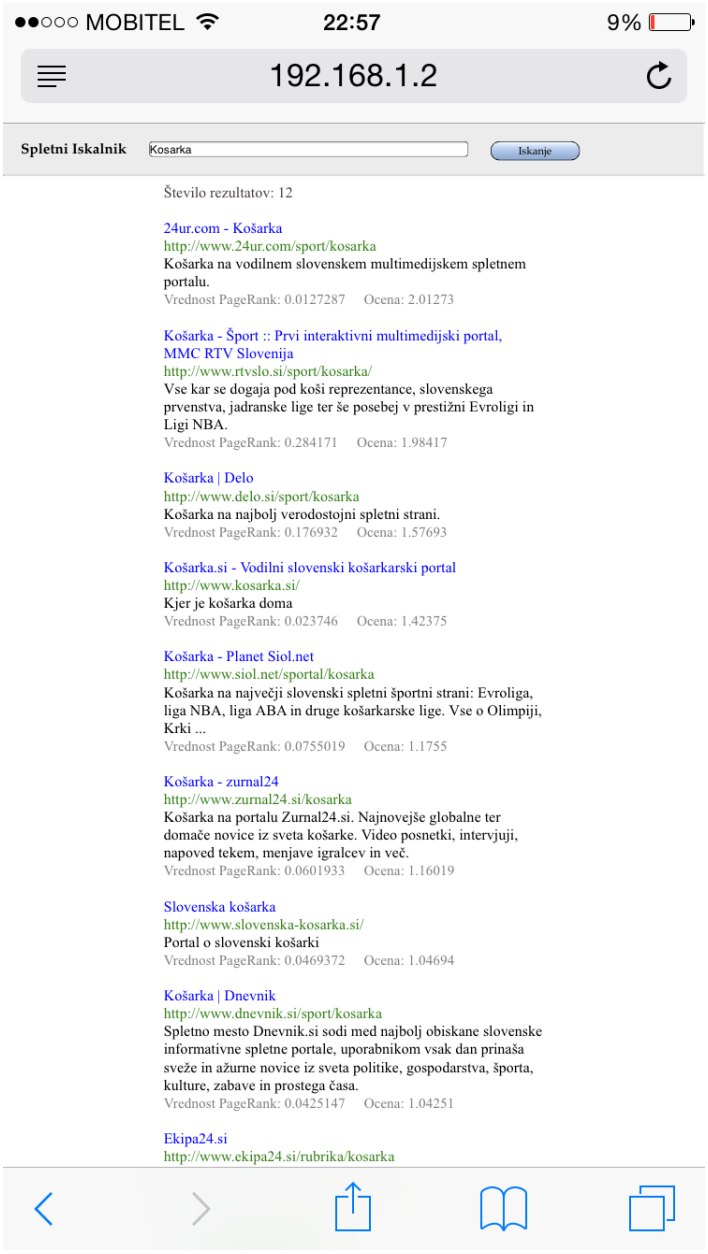
Vidimo, da se rezultati malce razlikujejo, a kljub vsemu so izmed petih rezultatov kar trije isti. Na prvem mestu v lastnem iskalniku je stran www.24ur.com/sport/kosarka, medtem ko je v iskalniku Google na prvem mestu stran www.kosarka.si. Tri strani, so v obeh iskalnikih, in sicer www.rtvslo.si/sport/kosarka, www.kosarka.si in www.siol.net/sportal/kosarka. Nobena od teh treh strani pa se ne nahaja na istem mestu v iskalnikih.

Slika 4.10: Prikaz rezultatov iskanja v lastnem spletnem iskalniku in iskalniku Google



Iz slike 4.11 pa lahko vidimo še prikaz rezultatov na mobilnem telefonu iPhone 6. Če pogledamo grafični prikaz iskalnika na mobilnem telefonu lahko opazimo, da se te popolnoma nič ne razlikuje od iskalnika na osebnem računalniku. Rezultati pa so na iskalni niz *kosarka* tudi razvrščeni popolnoma enako, saj ustvarjen algoritem deluje enako, ne glede na napravo na kateri se nahaja iskalnik.

Slika 4.11: Prikaz rezultatov na mobilnem telefonu



5 SKLEP

Kljub temu, da so algoritmi za rangiranje spletnih strani dobro varovane poslovne skrivnosti podjetji, ki so lastniki iskalnikov, morajo algoritme ves čas spreminjati in jih dopolnjevati oziroma izboljševati. Vsi podatki o algoritmih, ki jih najdemo na spletu, so stvar dolgotrajnega raziskovanja in objavljanja rezultatov teh raziskav na spletu. To je tudi razlog, da podjetja ves čas izboljšujejo svoje algoritme za rangiranje spletnih strani in s tem onemogočajo uporabnikom le teh, da bi na lahek način izboljšali razvrstitev svojih spletnih strani v iskalniku.

V diplomski nalogi sem se ukvarjal z razvojem lastnega spletnega iskalnika in implementacijo algoritma za rangiranje spletnih strani. Oba sem ustvaril v programskem jeziku PHP, poleg tega pa sem moral ustvariti tudi podatkovno bazo, tako da je lahko algoritem dostopal do podakov o spletnih straneh, na podlagi katerih je strani tudi razvrščal. Lastni algoritem temelji na algortimu PageRank, kateremu sem dodal nekaj dodatkov. Za izdelavo tako spletnega iskalnika kot tudi algoritma, sem moral prebrati kar nekaj literature in izboljšati svoje znanje o programskem jeziku PHP, a nedvomim v to, da mi bo to znanje prav prišlo tudi v prihodnje.

Ko je bil lastni iskalnik z implementiranim algoritmom končan, sem se odločil, da njegovo delovanje primerjam z najbolj znanim spletnim iskalnikom, ki je trenutno na trgu, iskalnikom Google. Primerjal sem prvih pet rezultatov obeh iskalnikov, a sem pri iskalniku Google s pomočjo naprednega iskanja omejil rezultate le na spletne strani v slovenskem. Razlog za to je bil, da sem tudi sam imel vse strani, katere sem vnesel v podatkovno bazo v slovenskem jeziku. Rezultati so bili podobni, a ne toliko kot sem morda pričakoval. Med prvimi petimi rezultati so se znašli trije isti, a v malo drugačnem vrstnem redu.

Brez dvoma se bodo algoritmi tudi v prihodnosti še naprej spreminjali in izboljševali, tako da bo uporabnikom vedno težje ugotoviti na kakšen način iskalnik rangira spletne

strani oziroma kateri elementi spletne strani so pomembnejši v primerjavi z ostalimi za boljšo razvrstitev strani v iskalniku.

6 LITERATURA

1. Bruce, James. 2013. *Make use of: How do search engines work*. Dostopno prek: http://www.makeuseof.com/tag/how-do-search-engines-work-makeuseof-explains/ (28. avgust 2015).

2. Facts Hunt. 2014. *Total number of Websites & Size of Internet as of 2013*. Dostopno prek: http://www.factshunt.com/2014/01/total-number-of-websites-size-of.html (8. september 2015).

3. Funding Universe. 2005. *Yahoo! Inc. History*. Dostopno prek: http://www.fundinguniverse.com/company-histories/yahoo-inc-history/ (29. avgust 2015).

4. Get Rank. 2015. *Link extractor*. Dostopno prek: http://www.getrank.org/tools/link-extractor/ (29. avgust 2015).

5. Grehan, Mike. 2002. *How search engine works.* New York: Incisive Media.

6. Jing, Yushi in Shumeet Baluja. 2008. VisualRank: Applying PageRank to Large-Scale Image Search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30 (11): 1877 – 1889.

7. PHP/ir. 2009. *PageRank in PHP*. Dostopno prek: http://phpir.com/pagerank-in-php/ (29. avgust 2015).

8. Pretto, Luca. 2002. A Theoretical Analysis of Google's PageRank. *Lecture Notes in Computer Science* (2476): 131 – 144.

9. SEO Book. 2015. *Keyword density analyzer*. Dostopno prek: http://tools.seobook.com/general/keyword-density/ (29. avgust 2015).

10. Statistic Brain. 2015. *Total Number of Pages Indexed by Google*. Dostopno prek: http://www.statisticbrain.com/total-number-of-pages-indexed-by-google/ (12. september 2015).

11. The History of SEO. 2011. *Short history of early search engines*. Dostopno prek: http://www.thehistoryofseo.com/The-Industry/Short_History_of_Early_Search_Engines.aspx (28. avgust 2015).

12. Wall, Aaron. 2015. *Search engine history: History of search engines*. Dostopno prek: http://www.searchenginehistory.com (28. avgust 2015).

13. WebConfs. 2013. *The Importance of Backlinks*. Dostopno prek: http://www.webconfs.com/importance-of-backlinks-article-5.php (7. september 2015).

14. Wikipedia. 2015. *PageRank*. Dostopno prek: https://en.wikipedia.org/wiki/PageRank (29. avgust 2015).

15. Word Stream. 2015. *The history of search engines – An infographic*. Dostopno prek: http://www.wordstream.com/articles/internet-search-engines-history (28. avgust 2015).

16. Zajko, Marko. 2012. *Algoritmi za inteligentno rangiranje spletnih strani*. Diplomsko delo. Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko.

17. Zareh Bidoki, A. Mohamed in Nasser Yazdani. 2008. DistanceRank: An intelligent ranking algorithm for web pages. *Information Processing and Management* 44: 877 – 892.