6 darbas. Informacijos paieška WWW

Darbo tikslas – įsisavinti informacijos paieškos pasauliniame WWW tinkle principus .

1. Teorinė dalis

Pasaulinis WWW tinklas

WWW apjungia visas Interneto paslaugas į vieningą sistemą, ir informacijos mainai tarp Interneto vartotojų labai supaprastėja. WWW sistemos šerdis yra hiperteksto dokumentai(WWW dokumentai), kurie yra saugomi Interneto tinklo kompiuteriuose ir yra prieinami kiekvienam WWW sistemos vartotojui. WWW dokumente galima talpinti ne vien tik tekstinę, bet ir grafinę, video ir audio informaciją.

WWW dokumentas

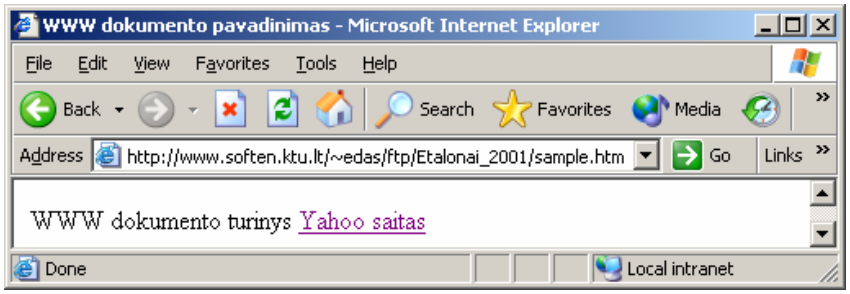
WWW dokumente (1 pav.) yra matomi(tekstas, paveikslai, rėmeliai, fonas…) ir nematomi(HTML operatoriai) objektai. “<>” skliaustuose yra HTML operatoriai.

```
<HTML>
<HEAD>
<TITLE>WWW dokumento pavadinimas</TITLE>
</HEAD>
<BODY>

WWW dokumento turinys

<A HREF="http://www.yahoo.com/"  > Yahoo saitas
 </BODY>

</HTML>
```

1 pav. WWW dokumentas HTML kalba

WWW naršyklės

Darbui su WWW yra naudojamos WWW naršyklės. Jos yra 2 tipų – grafinės ir tekstinės. Grafinės naršyklės (Netscape Navigator, Internet Explorer) aplinkoje galima atidaryti WWW dokumentus su tekstine, grafine, audio ir video informacija.
Tekstinės naršyklės (Lynx) dirba tiktai tekstiniame režime. Ši naršyklė naudojama IBM, VAX ir UNIX šeimų kompiuteriuose. Šiame darbe bus panaudotos grafinės naršyklės. 2 pav. matome kaip atrodo WWW dokumentas Internet Explorer naršyklės lange.



2 pav. Internet Explorer aplinkoje atidarytas WWW dokumentas .

Paieškos programos (Search engines)

Pažodinis vertimas – paieškos variklis. Paieškos programos uždavinys yra naršyti po Internetą ir ieškoti vartotojo nurodytos informacijos. Paieškos programų objektai yra WWW dokumentai ir failai, į kuriuos yra saitai(link) minėtuose dokumentuose. Paieškos programas galima suskirstyti į 2 pagrindines grupes - **Search engines** ir **Directories.**

Search engines **.** Šio tipo programos dar vadinamos **"**vorais**" (spiders)**. Toks pavadinimas panaudotas todėl, kad Search engines pastoviai lanko visas WWW vietas(sites) ir kuria WWW dokumentų katalogus(**index**). Tokiu būdu yra sukaupiamas milžiniškas dokumentų katalogas. Pagrindinis trūkumas yra tas, kad paieškos rezultate gauname daug nenaudingos informacijos -"šiukšlių".

Directories**.** Šio tipo programose WWW katalogas yra kuriamas ne automatiškai, o interaktyviai. Žmogaus dalyvavimas informacijos klasifikavimo procese sąlygoja tikslesnį informacijos paskirstymą. Paieškos rezultatai paprastai būna geresni negu spiders tipo programose. Pagrindinis trūkumas yra tas, kad paieška vykdoma tiktai tarp užregistruotų WWW dokumentų.

Mišrios paieškos programos. Šio tipo programose yra realizuota mišri paieška. Dalis WWW puslapių yra interaktyviai klasifikuota. 1 lentelėje pateiktos 10 populiariausių paieškos programų. Kiekviena paieškos programa turi savo URL adresą. Paieška vykdoma panaudojant WWW naršyklę. Nurodžius URL adresą, naršyklė atidaro WWW dokumentą, kuriame įvedame duomenis ir stebime paieškos rezultatus.

1 lentelė

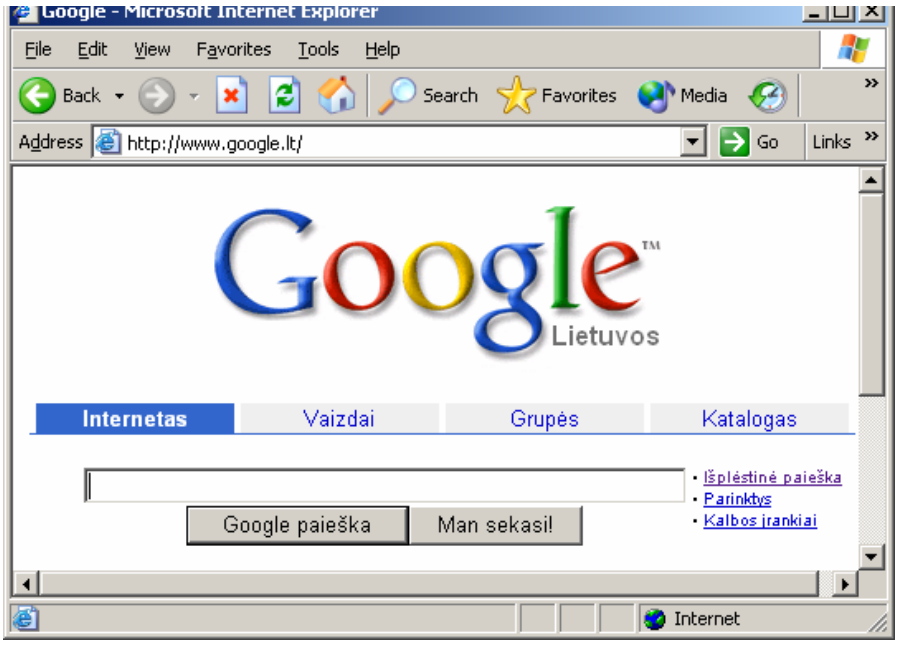
Eil. Nr.	Paieškos programa	URL adresas	Pastaba
1	Google	http://www.google.com	Search engines paieška
2	Yahoo	http://www.yahoo.com	Directories tipo paieška
3	Ask Jeeves	http://www.askjeeves.com	Mišri paieška
4	MSN Search	http://search.msn.com	Naudoja kelias programas
5	AOL	http://search.aol.com	Mišri paieška
6	Alta Vista	http://altavista.digital.com	Search engines paieška
7	Look Smart	http://www.looksmart.com	Directories tipo paieška
8	Netscape Search	http://search.netscape.com	Naudoja kelias programas
9	Info Space	http://www.infospace.com	Naudoja kelias programas
10	Overture (GoTo)	http://www.overture.com	Search engines paieška

Deleted: /www.

Paieškos rakto konstravimas**.** Paieškos raktu vadinsime žodžių seką, kuri bus panaudota WWW dokumento paieškai. 3 pav. yra Google paieškos programos WWW dokumentas, atidarytas Internet Explorer naršyklės lange. Paieškos raktas yra įvedamas į lauką virš „Google paieška" klavišo mažosiomis raidėmis.

Paprasta paieška. Priklausomai nuo to, kokie panaudoti skyrikliai, galimi tokie paprastos paieškos variantai:

Žodžio(žodžių) paieška. Jei paieškos raktas vienas žodis, bus ieškoma tų WWW dokumentų, kurių tekste yra bent vienas paieškos rakte nurodytas žodis arba tas žodis turi tokią pat pradžią. Tokius žodžius toliau vadinsime panašiais žodžiais, pvz., žodžiai computer ir computers. 2 lentelėje pateikti minėtų žodžių paieškos rezultatai.



3 pav. Google paieškos programos WWW dokumentas atidarytas Internet Explorer naršyklės lange

2 lentelė. Žodžio paieška

Paieškos programa	Paieškos raktas	Rasta dokumentų/data
Google	computer	58800000/02.06.12
	computers	21300000/02.06.12
Yahoo	computer	21206/02.06.12
	computers	20622/02.06.12

Kaip matome, žodžio computers paieškos rezultatas yra mažesnė dokumentų aibė. Reikia vengti bendrinių žodžių naudojimo. Net ir tuo atveju kai tikrai žinome, kad tame dokumente yra bendrinis žodis, tai duoda maža naudos, nes dokumentų aibė būna labai didelė ir praktiškai neaprėpiama. Labiau tinka tikriniai žodžiai arba sutrumpinimai.

Kai paieškos rakte panaudoti keli žodžiai, jų skyriklis yra tarpas. Paieškos rezultatas yra WWW dokumentų, kuriuose yra visi rakte nurodyti žodžiai arba panašūs žodžiai, aibė.

3 lentelė. Žodžių paieška

Paieškos programa	Paieškos raktas	Rasta dokumentų/data
Google	Alta	4760000/02.06.12
	Vista	6390000/02.06.12
	Alta Vista	1400000/02.06.12
	Search engine Alta Vista	238000/02.06.12
Yahoo	Alta	162/02.06.12
	Vista	636/02.06.12
	Alta Vista	30/02.06.12
	Search engine Alta Vista	1/02.06.12

Deleted: Google
Deleted: 129996
Deleted: 183426
Deleted: Search engine
Deleted: 4918309
Deleted: a
Deleted: v
Deleted: 5216061
Deleted: 556/
Deleted: 1069/
Deleted: Search engine
Deleted: 1792/2936
Deleted: 80783/
Deleted: Lycos,Northern Light

Plus (+) paieška. "+" ženklas analogiškas tarpo ženklui, padėtam prieš žodį paieškos rakte. Paieškos rezultatas yra WWW dokumentai, kuriuose yra visi žodžiai, paieškos rakte turintys "+" ženklą . 4 lentelėje pateikti keli plus paieškos rezultatai.

4 lentelė. Plus (+) paieška

Paieškos programa	Paieškos raktas	Rasta dokumentų/data
Google	Valdymo +centras	3870/02.06.13
	valdymo +centras +Litnet	192/02.06.13
Yahoo	Valdymo +centras	3260/02.06.13
	Litnet +valdymo +centras	184/02.06.13

Visi 192 (Google) WWW dokumentai patenka į paieškos pagal pirmą raktą skaičių (3870).

Minus (-) paieška. "-" ženklas, padėtas prieš žodį paieškos rakte, reiškia, kad ieškomi WWW dokumentai, kuriuose nėra žodžių, turinčių paieškos rakte "-" ženklą. 5 lentelėje pateikti keli tokios paieškos rezultatai.

5 lentelė. Minus (-) paieška

Paieškos programa	Paieškos raktas	Rasta dokumentų/data
Google	Lithuania	3080000/02.06.13
	KTU	153000/02.06.13
	Lithuania -KTU	492000/02.06.13
Yahoo	Lithuania	201/02.06.13
	KTU	17/02.06.13
	Lithuania -KTU	190/02.06.13

Visi 190 (Yahoo) WWW dokumentų turi bent vieną žodį „Lithuania" ir neturi nei vieno žodžio „KTU". Papildomas žodžio su "-" ženklu panaudojimas mažina WWW dokumentų aibę.

Frazės (" ") paieška. Frazė yra žodžių seka tarp kabučių(""). Jei paieškos rakte nurodyta frazė, randami WWW dokumentai, kuriuose būtinai yra bent viena tokia frazė. 6 lentelėje pateikti keli frazės paieškos rezultatai.

6 lentelė. Frazės paieška

Paieškos programa	Paieškos raktas	Rasta dokumentų/data
Google	"valdymo centras"	3260/02.06.13
	"litnet valdymo centras"	184/02.06.13
Netscape Search	"valdymo centras"	318/02.06.17
	"litnet valdymo centras"	191/02.06.17

Kuo ilgesnė frazė, tuo paprastai mažesnė surastų WWW dokumentų aibė.

Kombinuota paieška. Šiuo atveju paieškos rakte naudojami žodžiai, frazės bei "+" ir "-" ženklai. 7 lentelėje pateikti kombinuotos paieškos rezultatai

7 lentelė. Kombinuota paieška

Paieškos programa	Paieškos raktas	Rasta dokumentų/data
Netscape Search	"valdymo centras" +Kaunas	892/02.06.18
	"valdymo centras" +Kaunas -LITNET	793/02.06.18
	"valdymo centras" +Kaunas +LITNET	99/02.06.18
Yahoo	"valdymo centras" +Kaunas	10/02.06.18
	"valdymo centras" +Kaunas -LITNET	9/02.06.18
	"valdymo centras" +Kaunas +LITNET	1/02.06.18

Naudojant kombinuotą paiešką reikia pradžioje gauti kuo didesnę WWW dokumentų aibę, o po to stengtis ją mažinti. Tai ypač aktualu tuo atveju, kai mes tiksliai nežinome, kokie žodžiai yra panaudoti dokumento pavadinime ir pačiame dokumente. Labai gerai, kai žinome, kad dokumente yra tikrai panaudotas vienas ar kitas žodis arba frazė. Jeigu tiksliai nežinome, koks žodis, tada reikia naudoti kelis jo sinonimus. Mūsų tikslas - gauti realiai aprėpiamą WWW dokumentų aibę ir neprarasti informacijos. Praradimas reiškia, kad paieškos rakte buvo panaudota bloga frazė arba nereikalingas žodis su (-) ženklu. Tada atitinkami WWW dokumentai nepatenka į paieškos rezultatų aibę.

Pagerinta paieška. Naudojant paprastą paiešką ne visada pavyksta surasti reikiamą WWW dokumentą. Dabar aptarsime pagerintą paiešką, kai naudojami loginiai operatoriai ir raktiniai žodžiai.

Loginė paieška. Šiuo atveju paieškos rakte naudojami loginiai operatoriai AND, OR, NOT ir skliaustai "(", ")". Loginiai operatoriai rašomi didžiosiomis raidėmis.

AND operatorius. AND operatorius, jungiantis žodžius ar frazes, reiškia, kad rastuose WWW dokumentuose visi tie žodžiai arba frazės tikrai pasitaikys.

OR operatorius. OR operatorius**,** jungiantis žodžius ar frazes, reiškia, kad rastuose WWW dokumentuose būtinai pasitaikys bent vienas iš tų žodžių arba frazių.

NOT operatorius. NOT operatorius, panaudotas prieš žodį arba frazę, reiškia, kad rastuose WWW dokumentuose tikrai nebus žodžių ar frazių, prieš kuriuos buvo panaudotas NOT operatorius.

Operatoriniai skliaustai. Operatoriniai skliaustai panaudojami sudėtingų loginių išraiškų konstravimui ir nurodo loginių operatorių panaudojimo eilės tvarką. 8 lentelėje pateikti loginių operatorių panaudojimo pavyzdžiai.

Deleted: litnet
Deleted: 2948
Deleted: litnet
Deleted: 1206
Deleted: lvc
Deleted: 1198
Deleted: "litnet valdymo centras" +lvc –noc
... [1]
Deleted: "litnet valdymo centras" +Kaunas
Deleted: 4
Deleted: 4
Deleted: "litnet valdymo centras" +lvc
Deleted: "litnet valdymo centras" +lvc –noc

8 lentelė. Loginė paieška

Paieškos programa	Paieškos raktas	Rasta dokumentų/data
AOL	Universitetas	4062/02.06.18
	Kaunas	37586/02.06.18
	Universitetas OR Kaunas	1244/02.06.18
	Universitetas AND Kaunas	1219/02.06.18
	Universitetas NOT Kaunas	50/02.06.18
	Universitetas AND (Vilnius OR Kaunas)	296/02.06.18
	Universitetas NOT (Vilnius OR Kaunas)	44/02.06.18

Deleted: 183
Deleted: 2945
Deleted: 3088
Deleted: 40
Deleted: Vilnius
Deleted: 137
Deleted: 76

Paieška panaudojant raktinius žodžius. Raktinis žodis nurodo, kad yra įjungtas specialus paieškos režimas:

Raktinis žodis **title** paieškos rakte prieš žodį arba frazę reiškia, kad paieška vykdoma tiktai dokumentų pavadinimuose. HTML kalboje dokumentų pavadinimai (1 pav.) įvedami panaudojant TITLE operatorių:

<TITLE>WWW dokumento pavadinimas</TITLE>

Atidarytame dokumente jie yra nematomi. (2 pav.). Tokio pavadinimo WWW dokumento paieškos raktas užrašomas taip:

title:"WWW dokumento pavadinimas"

Raktinis žodis **host** paieškos rakte prieš žodį reiškia, kad paieška bus vykdoma tuo žodžiu pavadintame kompiuteryje. Pavyzdžiui, nurodymas host: soften.ktu.lt reiškia, kad ieškoma kompiuteryje soften.ktu.lt.

Raktinis žodis **link** paieškos rakte prieš žodį reiškia, kad paieškos rezultatas yra WWW dokumentai, kuriuose yra bent vienas saitas panašus į duotą žodį. Pavyzdžiui, link:yahoo reiškia ieškoti WWW dokumentų, kuriuose yra saitas į Yahoo paieškos programą.

2. Tipinė užduotis

1. Paprasta paieška. Duotos naršyklės aplinkoje, panaudodami nurodytas paieškos programas atlikite iš anksto žinomo WWW dokumento paiešką. Savo paieškos rezultatus įrašykite į 9 lentelę greta jau esančių skaičių, kurie gauti 2002.06.20. Pradiniai duomenys:
   WWW naršyklė - Netscape;
   pirma paieškos programa - Netscape Search (tipas - Search engines, žr. 1 lentelę);
   antra paieškos programa - Yahoo (tipas - Directories, žr. 1 lentelę);
   WWW dokumentas - WWW tinkle patalpintas dokumentas adresu http://www.litnet.lt/index.html.
   a) Pradžioje atidarykite nurodytą dokumentą ir susipažinkite su jo turiniu. Rekomenduojame atidaryti 2 naršyklės langus. Viename lange atidarykite WWW dokumentą adresu http://www.litnet.lt/index.html, kitame pirmos paieškos programos WWW dokumentą adresu http://search.netscape.com.
   b) Lentelėje nurodyta eilės tvarka konstruokite paieškos raktus ir rezultatus įrašykite į lentelę. Paieškos raktas sukonstruotas sėkmingai, jeigu rezultatas neviršyja pastabose nurodyto skaičiaus.
   c) Pasitikrinkite, ar tarp paskutinio bandymo rezultatų yra užduotyje nurodytas WWW dokumentas.

Deleted: /www.

d) Naujame naršyklės lange atidarykite antros paieškos programos WWW dokumentą adresu http://www.yahoo.com.
e) Panaudojant sukonstruotus raktus pakartokite paieškas su antra programa. Rezultatus įrašykite į lentelę. Paieškos raktas būtinai turi likti tas pats.
f) Pasitikrinkite, ar tarp paskutinio bandymo rezultatų yra užduotyje nurodytas WWW dokumentas.

9 lentelė. Paprastos paieškos rezultatai

Eil. Nr.	Paieškos rakto variantas	Sukonstruotas paieškos raktas	Rasta dokumentų/data		Pastaba
			Pirma programa	Antra programa	
1	Vienas žodis	LITNET	7460/ 02.06.20	1/ 02.06.20	<1000
2	Plus(+)	LITNET +valdymo +centras	192/ 02.06.20	70/ 02.06.20	<100
3	Minus(-)	LITNET –Vilnius –Klaipėda	38/ 02.06.20	11/ 02.06.20	<500
4	Frazė (“”)	“LITNET valdymo centras”	192/ 02.06.20	184/ 02.06.20	<10
5	Mišrus	“LITNET valdymo centras” - servisas –mokslo	83/ 02.06.20	54/ 02.06.20	<10
Rastų WWW dokumentų URL adresai		http://www.litnet.lt/litnet/apielitnet.html			

Pastaba. Lentelėje nurodytos pastabos galioja tiktai pirmai paieškos programai.

2. Pagerinta paieška. Duotos naršyklės aplinkoje, panaudodami nurodytą paieškos programą atlikite duotą bendrinį žodį turinčių WWW dokumentų paiešką. Panaudokite loginius operatorius sumažinti WWW dokumentų aibei iki nurodyto skaičiaus. Paieškos rezultatus įrašykite į 10 lentelę greta jau esančių skaičių, kurie gauti 2002.06.20. Pradiniai duomenys:
   WWW naršyklė - Netscape;
   paieškos programa - AOL (1 lentelė);
   bendrinis žodis - dažnai naudojamas žodis computer.
   a) Atidarykite paieškos programos WWW dokumentą (http://search.aol.com).
   b) Atlikite paiešką pagal duotą bendrinį žodį (computer) ir rezultatą įrašykite į lentelę.
   c) Panaudodami loginį (AND arba OR) operatorių patikslinkite bendrinį žodį tokiu būdu, kad žymiai (bent 10 kartų) sumažėtų WWW dokumentų aibė. Lentelėje panaudotas žodis – notebook. Paieškos rezultatą įrašykite į 10 lentelę.
   d) Kartokite paiešką patikslindami paieškos raktą ne daugiau 10 kartų. 10 lentelėje yra 6 bandymų rezultatai.
   e) Atidarykite bet kurį paskutinio bandymo dokumentą ir patikrinkite, ar tikrai įvykdytos paieškos rakte nurodytos loginės sąlygos..

10 lentelė. Pagerintos paieškos rezultatai

Eil. Nr.	Paieškos rakto variantas	Sukonstruotas paieškos raktas	Rasta dokumentų/data
1	Bendrinis žodis	computer	23769984 /02.06.20
2	AND arba OR	computer AND notebook	574596 / 02.06.20
3	Skliaustai ()	computer AND notebook AND (IBM OR ThinkPad)	17687/02.06.20
4	Bet kuris prieš tai panaudotas	computer AND notebook AND (IBM OR ThinkPad) AND portfolio	1140 / 02.06.20
5	Frazė	computer AND notebook AND (IBM OR ThinkPad) AND portfolio AND “Protection System”	14/ 02.06.20
6	NOT	computer AND notebook AND (IBM OR ThinkPad) AND portfolio AND “Protection System” AND NOT black	3/ 02.06.20
Rastų WWW dokumentų URL adresai		http://www.thinkpaddepot.com/	

Pastaba. Bendras bandymų skaičius ne daugiau 10. Paieškos rezultatas ne daugiau 10 WWW dokumentų, kuriuose yra patikslintas bendrinis žodis. Nepamirškite įrašyti į 10 lentelę bent vieno rasto WWW dokumento URL adresą.

3. Kontrolinė užduotis

Kiekvieno užduoties punkto rezultatus pademonstruokite dėstytojui!

1. Paprasta paieška. Duotos naršyklės aplinkoje, panaudodami nurodytas paieškos programas atlikite iš anksto žinomo WWW dokumento paiešką. Paieškos rezultatus įrašykite į 11 lentelę. WWW naršyklė - Internet Explorer; pirma paieškos programa – Google; antra paieškos programa - Look Smart; WWW dokumentas - http://www.lrytas.lt/index.htm.

11 lentelė

Eil. Nr.	Paieškos rakto variantas	Sukonstruotas paieškos raktas	Rasta dokumentų/data		Pastaba
			Pirma programa	Antra programa	
1	Vienas žodis				<1000
2	Plus(+)				<100
3	Minus(-)				<500

4	Frazė (“”)				<10
5	Mišrus				<10
Rastų WWW dokumentų URL adresai					

Pastaba. Lentelėje nurodytos pastabos galioja tiktai pirmai paieškos programai. Nepamirškite įrašyti į 11 lentelę bent vieno rasto WWW dokumento URL adresą.

2. **Pagerinta paieška.** Duotos naršyklės aplinkoje, panaudodami nurodytą paieškos programą atlikite duotą bendrinį žodį turinčių WWW dokumentų paiešką. Panaudodami loginius operatorius sumažinkite WWW dokumentų aibę iki nurodyto skaičiaus. Paieškos rezultatus įrašykite į 12 lentelę. WWW naršyklė – Netscape; paieškos programa – Infoseek; bendrinis žodis – Window.

12 lentelė

Eil. Nr.	Paieškos rakto variantas	Sukonstruotas paieškos raktas	Rasta dokumentų/data
1	Bendrinis žodis	Window	
2	AND arba OR		
3	Skliaustai ()		
4	Bet kuris prieš tai panaudotas		
5	Frazė		
6	NOT		
Rastų WWW dokumentų URL adresai			

Pastaba. Bendras bandymų skaičius ne daugiau 10. Paieškos rezultatas ne daugiau 10 WWW dokumentų, kuriuose yra patikslintas bendrinis žodis. Nepamirškite įrašyti į 12 lentelę bent vieno rasto WWW dokumento URL adresą.

Rekomenduojama literatūra

1. J. Adomavičius ir kt. Informatika (I dalis). Kaunas, Technologija, 2001
2. A Webmaster's Guide To SearchEngines.
http://www.searchenginewatch.com/webmasters/index.html

	“litnet valdymo centras” –noc	1