Љиљана Чабаркапа

ПРЕТРАЖИВАЧКИ СИСТЕМИ

Обим информација које се чувају путем Интернета је огроман и мери се десетинама терабајта. На серверима мреже чува се више од 2 милијарде Web-страница. Користећи Интернет корисник приступа не само текстуалним документима, већ и фото, аудио и видео-материјалима.

У овом мору информација поред актуелних наићи ћете на много „смећа“ – застарелих и безвредних информација и вулгарних реклама. Интернет је најдемократскији извор информација без централизоване контроле и скоро без цензуре. Свако може поставити свој сајт и на њему изнети своје мишљење. Нема чак ни ограничења којима би се избегло понављање информација. Зато, ради што лакшег проналажења информација Интернет поседује претраживачке системе и каталоге. Моћни петраживачки системи и каталози су сложени технички комплекси у које је интегрисано неколико десетина супербрзих рачунара, које опслужују стотине специјалиста. Претраживачки системи (енглески - Search Engines) у жаргону мреже се називају *претраживачи*.

Да би се коришћењем претраживача пронашао неки документ мора се формирати упит, на основу кога ће се изабрати документи у дистрибуираној бази података, која се чува на Интернету. Упит се формира помоћу кључних речи (једне или више њих). Резултат претраге се презентира у виду листе адреса (хипервеза) и кратке анотације (сижеа) у вези садржаја који се налази на наведеној адреси.

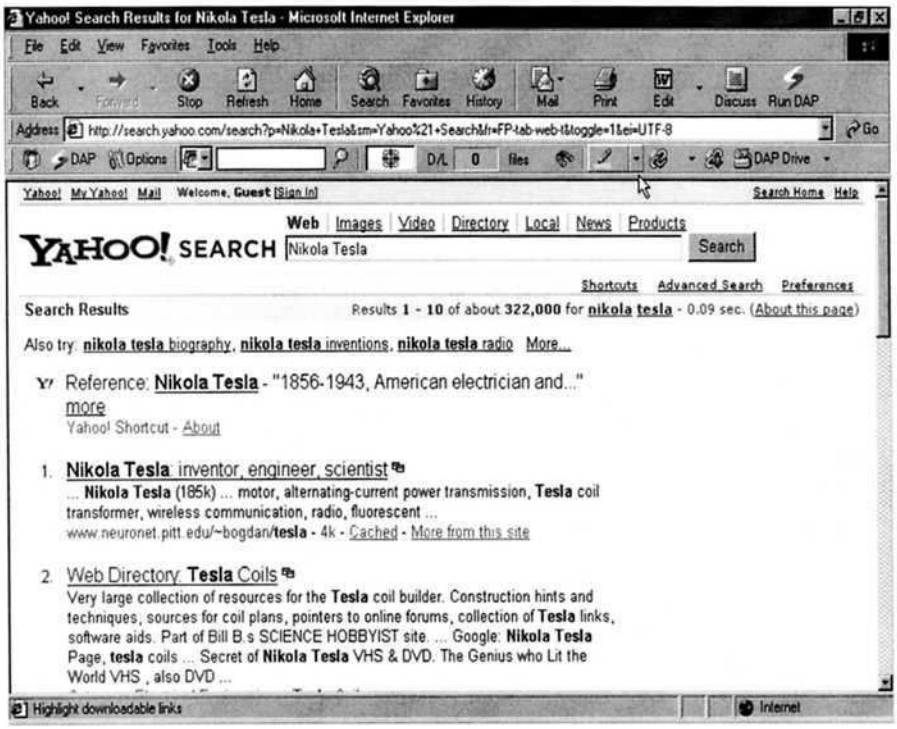
Кључна реч је појам који у што је могуће већој мери одражава садржај траженог документа. Разликујемо две врсте упита: просте и сложене. Упит састављен само од кључних речи и састава од кључних речи назива се прост. Прости упити повезани логичким и другим операторима називају се сложени упити.

Претрага помоћу кључне речи се реализује тако што се у пољу претраживача под именом `Search, Find, Go, Go Get it` (назив зависи од избора претраживача) унесе кључна реч, а затим активира претрага кликом на дугме `Search` (или `Submit`).

Како раде претраживачки системи

Претраживачки систем чине три основне компоненте:

- *паук* (робот, црв, агент) је програм, који аутоматски периодично „посећује“ сајтове, прикупљајући информације о садржају и адреси. Дакле, пауци повремено скенирају мрежу, евидентирају где се шта налази, да би кориснику,



Сл. 1

када је то потребно, могли указати тачно место где се тражени документ чува (тј. његову домену адресу);

- *база података*, која се формира на основу прикупљених информација и користи да би се одговорило на упит корисника;
- *интерфејс* претраживачког система са механизмом претраге по бази података.

Да би се процес тражења документа убрзао, претраживач активира процес *индексирања*. Овим процесом се сваком документу, који се налази на Интернету, кореспондира скуп кључних речи и региструје у бази података. Дакле, у бази података претраживачког система се не региструје цео документ, већ само његов део. Погрешно је мишљење неискусних корисника претраживача да се након постављања упита претражује по целој мрежи. Претраживање је само унутар базе претраживачког система.

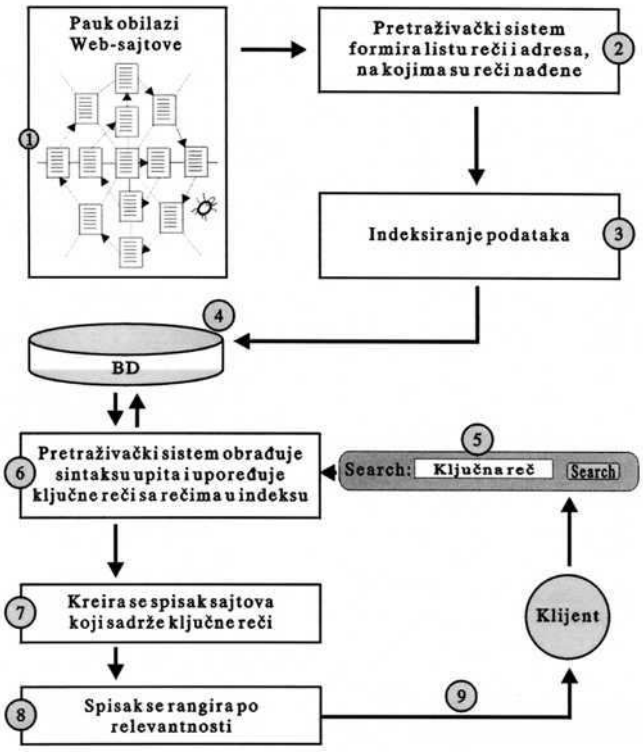
Када се опслужује конкретан упит којим се тражи нека информација, претраживач пореди кључне речи, које је унео корисник, са кључним речима који су у поступку индексирања сачуване у бази претраживача. У случају да се кључне речи поклапају, кориснику се издаје адреса датог документа, тј. указује на место где се чува на мрежи.

Овај процес претраге по кључним речима подсећа на поступак проналажења

коришћењем алфабетског индекса у књизи. Читалац помоћу алфабетског индекса одређује број странице на којој се налази тражена кључна реч. Аналогија резултату ове претраге на Интернету је адреса Web-странице.

Креирање индекса

Када паук нађе неки Web-сајт, фиксира речи, слике, хипервезе и друге елементе који се на њему налазе. За реч се евидентира где је смештена: у наслову (`title`), поднаслову (`subtitle`), метатагу (`meta tags`) или другим местима. Метатагови су ознаке које стављају власници страница да би истакли кључне речи и тематику по којима ће се обавити индексирање. Ово је посебно важно када кључне речи имају више значења. Тада метатагови усмеравају претраживачки систем да између неколико значења речи изберу правилно. Међутим, несавесни власници Web-сајтова метатаговима означавају најпопуларније речи на Интернету, које немају никакве везе са темом сајта да би к себи привукли што више посетилаца и подигли рејтинг посећености сајта. Препознавање и елиминисање таквих сајтова је још један задатак који треба да испуни добар претраживачки систем.



Сл. 2

Након што су прикупљене информације са Web-сајтова, обавља се индексирање добијених података. У том процесу се врши рангирање по важности речи нађених на страници. Пожељно је да у индекс уђу оне речи које су у тесној вези са темом сајта. Зато се речима додељују тежински коефицијенти у зависности од тога колико често и где се срећу (у заглављу странице, на почетку или крају странице, хипервези, метатагу итд). Сваки претраживачки систем има свој алгоритам додељивања тежинских коефицијената – због чега претраживачки системи за исту кључну реч дају различите спискове ресурса (докумената).

При креирању индекса води се рачуна и о елиминисању дупликата и „скоро дупликата“ – докумената који су врло слични, на пример, разликују им се само заглавља, а текст се дуплира. Таквих докумената је врло много, на пример, исти документ се може појавити на сајту неког симпозијума и сајту аутора текста.

Тенденција развоја петраживачких система је таква да најбржи петраживачки системи теже да произведу индексирање целог документа, а не само његовог назива и првих реченица. Најсавршенији роботи при индексирању скенирају не само прву страницу, већ по хипервезама залазе у дубину сајта.

Механизам претходног индексирања даје добре резултате у случају када је добро формиран упит (са добро погођеним кључним речима). Непромишљено постављање упита може изазвати велики број веза (линкова), тако да је у „информационој шуми“ тешко наћи тражени документ, или га уопште није могуће наћи.

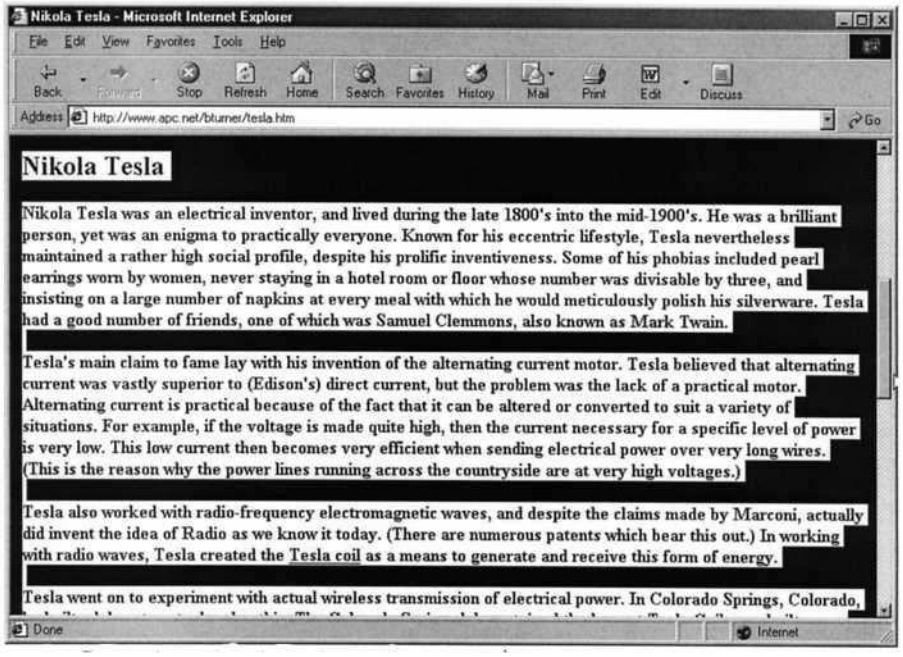
Ефикасност претраге се повећава коришћењем филтера. Они омогућавају:

- рестрикцију избора докумената коришћењем логичких оператора (реализује се сложена претрага);
- ограничавање простора претраге специфицирањем типа протокола, помоћу којих је креиран документ (претрага на Web-сајтовима или телеконференцијама);
- рестрикцију избора докумената према интервалу креирања документа (на пример, од 12. јула 2002. год. до 12. августа 2003. год);
- рестрикцију избора докумената према језику на ком су састављени (енглески, српски);
- рестрикцију избора докумената према територији размештаја сервера (на пример, само Европа);
- рестрикција дефинисањем одређеног дела документа (заглавље, адреса);
- рестрикција избором фразе са задатим поретком кључних речи.

Вежба 1. Коришћењем програма Internet Explorer активирати претраживач Google (`http://www.Google.com`) и у поље `Google Search` као кључне речи за претраживање унети `Nikola Tesla`. Из списка сајтова који се добијају као резултат претраге изабрати један.

Вежба 2. Изабрати део текста биографије (са сајта из вежбе 1) и копирати у неки фајл.

1. Селектовати текст за копирање.



Сл. 3

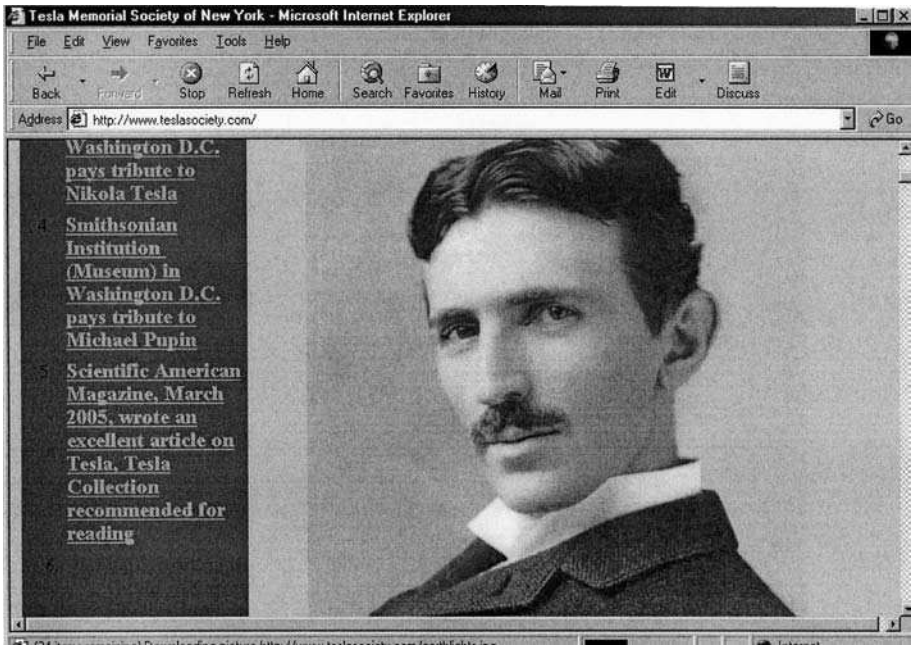
2. Кликнути на текст десним дугметом миша.
3. Из контекстног менија изабрати опцију `Copy`.
4. Покренути текстуални едитор.
5. У главном менију изабрати опцију `Edit/Paste`.
6. Након што је учитан текст изабрати `File/Save as ...`
7. У дијалогу `Save as ...` изабрати фолдер у коме треба сачувати фајл, а затим у поље `File Name` унети име фајла.

Вежба 3. Слику Николе Тесле са сајта копирати у фајл.

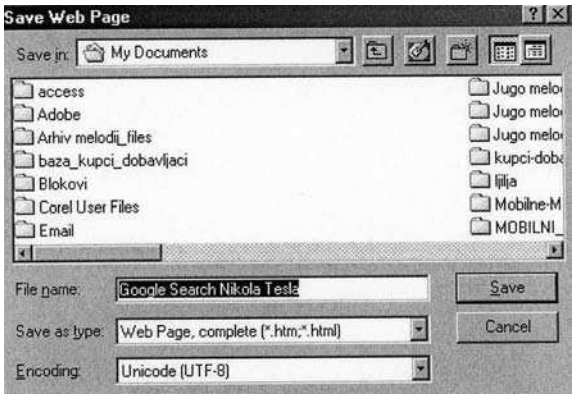
1. Кликнути на слику десним дугметом миша.
2. Из контексног менија изабрати опцију `Save Picture As ...`
3. У дијалогу за чување слике изабрати фолдер и име фајла под којим желите да сачувате слику.
4. Кликнути на дугме `Save`.

Вежба 4. Цео сајт копирати избором у менију браузера `File/Save as ...`

1. Избором `File/Save as ...` у браузеру активирајте дијалог `Save Web Page`.
2. У дијалогу изаберите фолдер, име фајла и тип чувања. Ако изаберете тип чувања `Web Page, complete`, биће сачувани текст, слике и све референце.


Tesla Memorial Society of New York - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss
Address http://www.teslasociety.com/
Go
Washington D.C. pays tribute to Nikola Tesla
Smithsonian Institution (Museum) in Washington D.C. pays tribute to Michael Pupin
Scientific American Magazine, March 2005, wrote an excellent article on Tesla, Tesla Collection recommended for reading

Сл. 4


Save Web Page
Save in: My Documents
access
Adobe
Arhiv melodii_files
baza_kupci_dobavljaci
Blokovi
Corel User Files
Email
Jugo melo
Jugo melo
Jugo melo
kupci-dob
lilja
Mobilne-M
MOBILNI_
File name: Google Search Nikola Tesla
Save as type: Web Page, complete (*.htm;*.html)
Encoding: Unicode (UTF-8)
Save
Cancel

Сл. 5