Дәріс 7.

Интернеттің іздеу технологиялары.

Қажетті ақпаратты іздеу принциптері. Іздеу машинасының жұмыс механизмі. Индексті құру.

Web-тен іздеу

Интернетте миллиондаған сайттар бар, соның ішінде өзекті ақпаратпен қоса көптеген ескі қорлар орналыстырылған. Интернет – белгілі бір басқарушысы жоқ демократиялық ақпарат көзі болып табылады. Кез келген адам желіге өзінің қорын орналастыра алады. Қорытындылап келгенде, интернетте ақпараттың қайталанбауына, оның стандартқа сай келуіне көп адамдар мән бере бермейді. Желіде барлығы бар екені белгілі, бірақ желіден қажетті ақпаратты алу қиын. Яғни, мәліметті табу үшін, оны жақсы іздей білу керек. Осы бөлімде интернет желісімен жұмыс істейтін іздеу аспаптары сипатталып, іздеу жүйесінің жұмыс механизмі түсіндірілген, іздеу оптимизациясына практикалық түсініктеме берілген.

Интернетте ақпаратты іздеуге арналған мынадай әртүрлі аспаптар бар: іздеу машиналары (поисковиктер), индекстелген каталогтер (рубрикаторлар), рейтингілер, метаіздеуіш жүйелер және тематикалық сілтемелердің тізімі, онлайн энциклопедиялары мен анықтамалар. Осы кезде әр түрлі үлгідегі ақпаратты табуда іздеу аспаптарының түрлі категорияларын қолдану тиімді болып келеді. Әр категорияны жеке қарастырайық.

Индекстелген каталогтер

Каталог дегеніміз тақырыптары бойынша топтастырылған иерархиялық құрылым түрінде берілетін мәліметтер. Иерархиялық құрылымның бірінші деңгейіндегі тематикалық бөлімі «спорт», «демалыс», «ғылым», «дүкендер» сияқты кең тараған тақырыптардан тұрады. Ал әр бөлімнің бөлімшелері болады. Осылайша, біртіндеп каталог бұтақтары арқылы саяхат жасап, іздеу облысын кішірейте отырып, сіз өзіңізге керекті облысты дәл анықтай аласыз. Мысалы оқу орындарын іздеу барысында мынадай тізбек пайда болуы мүмкін: *Білім-> Оқу орындары -> Жоғары оқу орындары ->Институттар*. Қажетті ішкі катологты тапқаннан кейін, одан сілтемелер жинағын аласыз. Катологтерді программалар емес, адамдар құрастырғандықтан, катологтегі барлық сілтемелер профильді болып табылады. Егер сіз ортақ тақырыпта жалпы ақпарат іздесеңіз, онда каталогке қатынаған дұрыс. Ал егер сізге нақты бір құжатты табу керек болса, онда каталог тиімсіз іздеу құралы болып табылады.

Желіде ортақ қолданылатын каталогтардан басқа, ерекшеленген каталогтар да бар. Егер де бір каталогта өте көп қор орналасса, онда оларды кең таралуына байланысты бірнеше бөліктерге бөлуге (ражнирование) болады. Мысалы, Яндекс каталогында бөліктеу басқа сайттардың біздің сайттағы сілтемелерінің индексімен жүргізіледі.

Желіде каталогтардан басқа рейтингтер де бар. Каталогтан рейтингтің айырмашылығы, мұнда қорларды тікелей оның иесі суреттесе, ал каталогта - авторы, демек оның редакторлары суреттейді.

Сілтемелердің тематикалық жинағы

Сілтемелердің тематикалық жинағы – бұл кәсіби топтармен немесе жеке жинақтаушылармен құрылған тізімдер. Шектелген кәсіби тақырыпты ірі каталогтің жұмыскерлер тобына қарағанда сол жұмыстың кәсіби маманы жақсы ашуы мүмкін.

Домендік атты теру

Каталог – бұл ыңғайлы іздеу жүйесі, бірақ егер сізге Intel немесе IBM компаниясының сервері керек болса, сіз каталогке қатынай алмайсыз. Сәйкес сайттардың атын табу қиын болмайды: www.intel.com, www.ibm.com.

Сол сияқты, сізге егер ауа райына арналған сайт қажет болса, оны www.weather.com серверінен іздеген дұрыс болады. Көп жағдайларда кілттік сөз арқылы сайтты табу мәтінде көп кездесетін сөзден тұратын құжатты тапқаннан ыңғайлы.

Іздеу кезінде танымалы емес компаниялардың адресінің атын интуитивті ойдан тергенде, бірде-бір іздеу жүйесінде тіркелмеген сервермен байланыс орнатуы мүмкін болғандықтан, ол іздеудің басқа түрлерімен табысты бәсекелесе алады. Осыған ұқсас іздеулер тиімсіз, сондықтан ізделінетін сайттың атын таба алмайтын болсаң, іздеу машинасын қолдану керек.

Іздеу машиналары

Сұранысқа жауап ретінде сіз әдетте құжаттардың ұзын тізімін аласыз, оның көбі сіздің сұрағыңызға жауап бермейді және сол тақырыпқа ешқандай қатысы болмайды. Сондай құжаттар релевантты емес (ағылшын сөзінен шыққан, *relevant*-лайықты, қатысты) деп аталады, ізденіс бойынша табылған құжаттар релевантты құжаттар деп аталады.

Табылған сілтемелердің тізіміндегі релевантты құжаттардың проценті сұраныстың дұрыс қойылуына байланысты болады.

Іздеу машинасы тапқан барлық құжаттардың ішіндегі релевант құжаттардың бөлігін іздеу дәлдігі деп атайды. Релевантты емес құжаттарды шуы бар құжаттар деп атайды. Егер табылған құжаттардың барлығы релевантты болып келсе (шуы жоқ құжаттар), іздеу дәлдігі 100% құрайды. Егер барлық релевантты құжаттар табылса, онда іздеу толымдығы 100% тең.

Сайып келгенде, іздеу сапасы екі өзара тәуелді параметрлермен анықталады: дәлдікпен және іздеу толықтығымен . Толықтықтың артуы дәлдікті төмендетеді және керісінше.

Іздеу машинасының жұмыс механизмі

Іздеу жүйелерін анықтама қызметімен салыстыруға болады, онда агенттер кәсіпорындарды аралап, мәліметтерді деректер базасына жинайды. Клиент анықтама қызметіне жолыққанда ақпарат сол деректер базасынан алынады. Мәліметтер базада ескіріп отырады, сондықтан агенттер оларды оқтын-оқтын жаңартады. Кейбір кәсіпорындар мәліметтерді өздері жібереді, сондықтан агенттердің оларға барудың қажеті болмайды. Басқаша айтқанда, анықтама қызметінің екі функциясы болады : жасау және деректер базасын тұрақты жаңарту және клиент сұранысы бойынша базадан хабар іздеу.

Сол сияқты, іздеу машинасы да екі бөлімнен тұрады: робот - ол берілген серверлерді аралап деректер базасын қалыптастырады, және іздеу механизмі. Робот терминінің көптеген синонимдері бар, роботтан басқа оны желілі агент немесе торапта жүргеніне байланысты құрт немесе өрмекші дейді.

Робот базасы негізі роботтың өзімен (робот өзі жаңа қорларға сілтемелер тауып алады) және аз дәрежеде өз сайттарын іздеу машиналарында тіркейтін қор иелеріменқалыптасады. Деректер базасын қалыптастыратын роботтан басқа табылған сілтемелердің рейтингісін анықтайтын программа бар.

Іздеу машинасының жұмыс принципі пайдаланушы көрсеткен кілттік сөздер арқылы ішкі каталогтан (деректер базасы) релеванттігі бойынша сұрыпталған сілтемелер тізімін беру арқылы іске асырылады.

Іздеу жүйесі тек ішкі каталогтармен операциялайтынын атап айтқан жөн. Іздеу машинасының мәліметтер базасы жүйедегі түйіндік адрестерді сұрау арқылы әрдайым жаңартылатынына қарамастан, іздеу машинасының ішкі қорларын және желі қорларын салыстыруға келмейді, сондықтан әрқашан машина ескірген адрес немесе қажетсіз ресурс табатыны өте ықтимал. Проблема тек қана ішкі қорлардың шектілігінде ғана емес, тағы роботтың жылдамдығының шектілігінде тұр. Іздеу машинасының ішкі қорларының көбеюі проблеманы шешпейді, себебі аралау жылдамдығы ақырлы.бірақ іздеу машинасының ішінде каталогтарға бөлінген Интернеттің кіріс қорларының белгілі бөлігінің көшірмесі болады деуге болмайды. Толық ақпарат (кіріс құжаттар) бәрі бірдей

сақталмайды, көбіне жиі тек оның бөлігі – индекстенген тізім немесе индекс деп аталатын, құжат жолынан шағын бөлігі сақталады.

Индекс құрау үшін кіріс мәліметтер қор көлемі минималды, ал іздеу тез әрі максималды пайдалы ақпарат беретіндей түрлендіріледі. Индекстелген тізімді түсіндіру үшін оның қағаз аналогы – конкорданс, яғни сөздікті келтіруге болады, онда белгілі жасушымен қолданылатын сөздер алфавиттік тәртіпте болады, және де жазушы шығармасында келтірілгеніне сілтеме болады.

Айтпаса да түсінікті, конкорданс (сөздік) шығарма тексінен шағын және одан керекті сөзді іздеген кітапты түгел парақтағаннан көп жеңіл.

Индексті құру

Желілік агенттер немесе робот-өрмекшілер Желі бойымен «өрмелейді», Web – беттерді талдайды және не әрі қай парақта табылғаны туралы ақпарат жинайды. Кезекті HTML-парақтарды табысымен көптеген іздеу машиналары (әр іздеу машиналарында әртүрлі) сөздерді, суреттерді, сілтемелерді және де басқа да элементтерді белгілейді. Сөздердің парақта барлығы ғана емес, әрі оның орналасуы, яғни бұл сөздің қайда орналасқаны: тақырыпта (title), тақырыпшаларда (subtitles), метатэгте (meta tags) немесе басқа орындарда. Әдетте негізгі сөз ескеріледі де, шылау мен одағайлар: «ал»,«бірақ» және «немесе» еленбейді. Метатегтер парақ иелерінің өзіне сол арқылы ізделінетін кілттік сөздер мен тақырыпты анықтауға мүмкіндік береді. Бұл әсіресе кілттік сөздің бірнеше мағынасы болғанда қажет. Метатегтер іздеу машинасын сөздердің бірнеше мағынасынан дұрысын таңдауға көмектеседі. Алайда метатегтер адал толтырылғанда ғана сенімді жұмыс істей алады.Web-парақтардың кейбір иелері өздерінің метатегтеріне Желіде көп аталатын өз сайт тақырыбына қатысы жоқ сөздермен толтырады, сол арқылы өзінің жаңа келушілерін тарту әрі қор қатысуы рейтингісін жоғарылату үшін жасайды. Іздеуден осы сияқты сайттарды шығару – жақсы іздеу жүйесінің тағы бір тапсырмасы. әрбір роботтың өз қараниетті жарнама үшін жазаланған қор тізімі бар.

Тапсырма берілген Web-парақтарда ақпарат жиналғаннан кейін алынған мәліметтерді индекстеу жүреді. Робот-өрмекшілер Web-парақтардың ақпараттарын қарастырып, кілттік сөздер арқылы индекстенген іздеу базасын құрады, содан кейін пайдаланушы сұранысы арқылы жүйе дұрыстығына (релевантты) қарай сайттар тізімін береді. Айқын, егер сіз сайтты «гүл» деген кілттік сөзбен іздесеңіз, онда іздеу машинасы сол сөз бар парақтарды тауып қана қоймай, бұл сөздің қай жерде сайт тақырыбына қатыстылығын анықтай алуы керек. Сөздің Web-парақтың профиліне қатыстығын анықтау үшін оның парақта қаншалықты жиі ұшырасатынын, берілген сөз туралы сілтемелердің бар-жоқтығын бағалау керек. Қысқаша айтқанда, парақта табылған сөздерді маңыздылық дәрежесіне қарай рангілеу керек.

Сөздерге салмақтылық коэфициенттері оның қанша және қайда кездесетініне қарай (парақ тақырыбында, беттің басы не аяғында, сілтемеде, метатегте және т.б) меншіктеледі. әрбір іздеу механизмі салмақ коэфициенттерін берудің өз алгоритмдері бар – бұл әртүрлі іздеу машиналарының бір кілттік сөз арқылы сұрауға әртүрлі қорлар тізімін берудің бір себебі. Парақтар әрдайым жаңартылып отыратындықтан, онда индекстеу үрдісі де жиі орындалып отырылуы керек. Робот-өрмекшілер сілтемелерді аралай жүріп, индекстен тұратын файлды құрады, ол үлкен болуы мүмкін. Оның көлемін азайту үшін ақпарат көлемін минимизациялау мен файлды сығуға жүгінеді. өңделгеннен кейін мәліметтер үнемі жанарып отыратын базада сақталады. Бірнеше роботтары бар іздеу машинасы секундына жүздеген парақтарды өңдей алады. Бүгінде мықты іздеу машиналары жүздеген миллион парақты сақтайды және күніне ондаған миллион сұранысты қабылдайды.

Индексті құруда көшірмелердің санын азайту тапсырмасы да шешіледі – қатесіз салыстыру үшін алдымен құжаттың кодировкасын анықтау қажеттігін ескерсек, тапсырма оңай емес. Бұдан да қиын тапсырмаға өте ұқсас құжаттарды айыру жатады (оларды «көшірме дерлік» деп атайды), мысалы оларға мазмұны бір ал тақырыбы әртүрлі болып келеді. Бұл сияқты құжаттар Желіде өте көп – мысалы біреу рефератты көшіріп алып өз

сайтында басқа атпен басып шығаруы мүмкін. Қазіргі заманғы іздеу машиналары барлық бұл проблемаларды шешуге мүмкіндік береді.

Индекс арқылы іздеу

Индекс арқылы іздеу мынадан құралады, яғни пайдаланушы сұраныс құрастырып оны іздеу машинасына береді. Бірнеше кілттік сөздерді қолдануда сұраныс тілін пайдаланған пайдалы, оның негізін буль операторлары құрайды.

Ең жиі қолданылатын буль операторлары:

- AND – бұл арқылы біріктірілген барлық терминдер ұсынылған құжатта қатысуы керек. Кейбір іздеу жүйелері «+» белгісін «AND» орнына қолданады;
- OR – кем дегенде бір кілттік сөз «OR »-мен қатысты, ізделінетін құжатта болуы керек;
- NOT- «NOT»-тан кейінгі кілттік сөз ізделінетін құжатта кездеспеуі керек. Кейбір іздеу жүйелері «-» белгісін «NOT» орнына пайдаланады;
- FOLLOWED BY – кілттік сөздер бірінен кейін бірі кезектесіп келуі керек;
- NEAR – сөздердің бірі екінші сөзден белгілі санды сөздерден кейін келуі керек;
- Тырнақшалар – тырнақша ішіндегі сөздер- бұл текст фрагменті құжат немесе файл ішінде кездесуі тиіс. Айта кетейік, сұраныс тілі семантикасы нақты бір іздеу машиналарында бір біріне ұқсамауы мүмкін, әдетте ол туралы іздеу машинасының нұсқауында мәлімет келтіріледі.

Шектерінде логикалық комбинация анықталатын мәтін *іздеу бірлігі* деп аталады. Бұл сөйлем, абзац не бүкіл құжат болуы мүмкін. Түрлі іздеу жүйелерінде әртүрлі іздеу бірліктері қолданылуы мүмкін. Сөйлем шегіндегі іздеу тек индексінде толық мекенжай (адрес) бар жүйелерде ғана мүмкін.

Пайдаланушы іздеу жүйесіне сұраныс жібергеннен кейін, ол сұраныс синтаксисін өңдейді, кілттік сөздерді индекстегі сөздермен салыстырады. Содан кейін сұранысқа жауап беретін сайттар тізімі релеванттылығына қарай рангіленіп, пайдаланушыға берілетіндей іздеу нәтижесі құрастырылады.

анықтамалар. Осы кезде әр түрлі үлгідегі ақпаратты табуда іздеу аспаптарының түрлі категорияларын қолдану тиімді болып келеді. Әр категорияны жеке қарастырайық.

Индекстелген каталогтер

Каталог дегеніміз тақырыптары бойынша топтастырылған иерархиялық құрылым түрінде берілетін мәліметтер. Иерархиялық құрылымның бірінші деңгейіндегі тематикалық бөлімі «спорт», «демалыс», «ғылым», «дүкендер» сияқты кең тараған тақырыптардан тұрады. Ал әр бөлімнің бөлімшелері болады. Осылайша, біртіндеп каталог бұтақтары арқылы саяхат жасап, іздеу облысын кішірейте отырып, сіз өзіңізге керекті облысты дәл анықтай аласыз. Мысалы оқу орындарын іздеу барысында мынадай тізбек пайда болуы мүмкін: *Білім-> Оқу орындары -> Жоғары оқу орындары ->Институттар*. Қажетті ішкі катологты тапқаннан кейін, одан сілтемелер жинағын аласыз. Катологтерді программалар емес, адамдар құрастырғандықтан, катологтегі барлық сілтемелер профильді болып табылады. Егер сіз ортақ тақырыпта жалпы ақпарат іздесеңіз, онда каталогке қатынаған дұрыс. Ал егер сізге нақты бір құжатты табу керек болса, онда каталог тиімсіз іздеу құралы болып табылады.

Желіде ортақ қолданылатын каталогтардан басқа, ерекшеленген каталогтар да бар. Егер де бір каталогта өте көп қор орналасса, онда оларды кең таралуына байланысты бірнеше бөліктерге бөлуге (ражнирование) болады. Мысалы, Яндекс каталогында бөліктеу басқа сайттардың біздің сайттағы сілтемелерінің индексімен жүргізіледі.

Желіде каталогтардан басқа рейтингтер де бар. Каталогтан рейтингтің айырмашылығы, мұнда қорларды тікелей оның иесі суреттесе, ал каталогта -

авторы, демек оның редакторлары суреттейді.

Сілтемелердің тематикалық жинағы

Сілтемелердің тематикалық жинағы – бұл кәсіби топтармен немесе жеке жинақтаушылармен құрылған тізімдер. Шектелген кәсіби тақырыпты ірі каталогтің жұмыскерлер тобына қарағанда сол жұмыстың кәсіби маманы жақсы ашуы мүмкін.

Домендік атты теру

Каталог – бұл ыңғайлы іздеу жүйесі, бірақ егер сізге Intel немесе IBM компаниясының сервері керек болса, сіз каталогке қатынай алмайсыз. Сәйкес сайттардың атын табу қиын болмайды: www.intel.com, www.ibm.com.

Сол сияқты, сізге егер ауа райына арналған сайт қажет болса, оны www.weather.com серверінен іздеген дұрыс болады. Көп жағдайларда кілттік сөз арқылы сайтты табу мәтінде көп кездесетін сөзден тұратын құжатты тапқаннан ыңғайлы.

Іздеу кезінде танымалы емес компаниялардың адресінің атын интуитивті ойдан тергенде, бірде-бір іздеу жүйесінде тіркелмеген сервермен байланыс орнатуы мүмкін болғандықтан, ол іздеудің басқа түрлерімен табысты бәсекелесе алады. Осыған ұқсас іздеулер тиімсіз, сондықтан ізделінетін сайттың атын таба алмайтын болсаң, іздеу машинасын қолдану керек.

Іздеу машиналары

Сұранысқа жауап ретінде сіз әдетте құжаттардың ұзын тізімін аласыз, оның көбі сіздің сұрағыңызға жауап бермейді және сол тақырыпқа ешқандай қатысы болмайды. Сондай құжаттар релевантты емес (ағылшын сөзінен шыққан, *relevant*- лайықты, қатысты) деп аталады, ізденіс бойынша табылған құжаттар релевантты құжаттар деп аталады.

Табылған сілтемелердің тізіміндегі релевантты құжаттардың проценті сұраныстың дұрыс қойылуына байланысты болады.

Іздеу машинасы тапқан барлық құжаттардың ішіндегі релевант құжаттардың бөлігін іздеу дәлдігі деп атайды. Релевантты емес құжаттарды шуы бар құжаттар деп атайды. Егер табылған құжаттардың барлығы релевантты болып келсе (шуы жоқ құжаттар), іздеу дәлдігі 100% құрайды. Егер барлық релевантты құжаттар табылса, онда іздеу толымдығы 100% тең.

Сайып келгенде, іздеу сапасы екі өзара тәуелді параметрлермен анықталады: дәлдікпен және іздеу толықтығымен . Толықтықтың артуы дәлдікті төмендетеді және керісінше.

Іздеу машинасының жұмыс механизмі

Іздеу жүйелерін анықтама қызметімен салыстыруға болады, онда агенттер кәсіпорындарды аралап, мәліметтерді деректер базасына жинайды. Клиент анықтама қызметіне жолыққанда ақпарат сол деректер базасынан алынады. Мәліметтер базада ескіріп отырады, сондықтан агенттер оларды оқ және деректер базасын тұрақты жаңарту және клиент сұранысы бойынша базадан хабар іздеу. Сол сияқты, іздеу машинасы да екі бөлімнен тұрады: робот - ол берілген серверлерді аралап деректер базасын қалыптастырады, және іздеу механизмі. Робот терминінің көптеген синонимдері бар, роботтан басқа оны желілі агент немесе торапта жүргеніне байланысты құрт немесе өрмекші дейді.

Робот базасы негізі роботтың өзімен (робот өзі жаңа қорларға сілтемелер тауып алады) және аз дәрежеде өз сайттарын іздеу машиналарында тіркейтін қор иелерімен қалыптасады. Деректер базасын қалыптастыратын роботтан басқа табылған сілтемелердің рейтингісін анықтайтын программа бар.

Іздеу машинасының жұмыс принципі пайдаланушы көрсеткен кілттік сөздер

арқылы ішкі каталогтан (деректер базасы) релеванттігі бойынша сұрыпталған сілтемелер тізімін беру арқылы іске асырылады.

Іздеу жүйесі тек ішкі каталогтармен операциялайтынын атап айтқан жөн. Іздеу машинасының мәліметтер базасы жүйедегі түйіндік адрестерді сұрау арқылы әрдайым жаңартылатынына қарамастан, іздеу машинасының ішкі қорларын және желі қорларын салыстыруға келмейді, сондықтан әрқашан машина ескірген адрес немесе қажетсіз ресурс табатыны өте ықтимал. Проблема тек қана ішкі қорлардың шектілігінде ғана емес, тағы роботтың жылдамдығының шектілігінде тұр. Іздеу машинасының ішкі қорларының көбеюі проблеманы шешпейді, себебі аралау жылдамдығы ақырлы.бірақ іздеу машинасының ішінде каталогтарға бөлінген Интернеттің кіріс қорларының белгілі бөлігінің көшірмесі болады деуге болмайды. Толық ақпарат (кіріс құжаттар) бәрі бірдей сақталмайды, көбіне жиі тек оның бөлігі – индекстенген тізім немесе индекс деп аталатын, құжат жолынан шағын бөлігі сақталады.

Индекс құрау үшін кіріс мәліметтер қор көлемі минималды, ал іздеу тез әрі максималды пайдалы ақпарат беретіндей түрлендіріледі. Индекстелген тізімді түсіндіру үшін оның қағаз аналогы – конкорданс, яғни сөздікті келтіруге болады, онда белгілі жасушымен қолданылатын сөздер алфавиттік тәртіпте болады, және де жазушы шығармасында келтірілгеніне сілтеме болады.

Айтпаса да түсінікті, конкорданс (сөздік) шығарма тексінен шағын және одан керекті сөзді іздеген кітапты түгел парақтағаннан көп жеңіл.

Индексті құру

Желілік агенттер немесе робот-өрмекшілер Желі бойымен «өрмелейді», Web – беттерді талдайды және не әрі қай парақта табылғаны туралы ақпарат жинайды. Кезекті HTML-парақтарды табысымен көптеген іздеу машиналары (әр іздеу машиналарында әртүрлі) сөздерді, суреттерді, сілтемелерді және де басқа да элементтерді белгілейді. Сөздердің парақта барлығы ғана емес, әрі оның орналасуы, яғни бұл сөздің қайда орналасқаны: тақырыпта (title), тақырыпшаларда (subtitles), метатэгте (meta tags) немесе басқа орындарда. Әдетте негізгі сөз ескеріледі де, шылау мен одағайлар: «ал»,«бірақ» және «немесе» еленбейді. Метатегтер парақ иелерінің өзіне сол арқылы ізделінетін кілттік сөздер мен тақырыпты анықтауға мүмкіндік береді. Бұл әсіресе кілттік сөздің бірнеше мағынасы болғанда қажет. Метатегтер іздеу машинасын сөдердің бірнеше мағынасынан дұрысын таңдауға көмектеседі. Алайда метатегтер адал толтырылғанда ғана сенімді жұмыс істей алады. Web-парақтардың кейбір иелері өздерінің метатегтеріне Желіде көп аталатын өз сайт тақырыбына қатысы жоқ сөздермен толтырады, сол арқылы өзінің жаңа келушілерін тарту әрі қор қатысуы рейтингісін жоғарылату үшін жасайды. Іздеуден осы
сияқты сайттарды шығару – жақсы іздеу жүйесінің тағы бір тапсырмасы. әрбір роботтың өз қараниетті жарнама үшін жазаланған қор тізімі бар.

Тапсырма берілген Web-парақтарда ақпарат жиналғаннан кейін алынған мәліметтерді индекстеу жүреді. Робот-өрмекшілер Web-парақтардың ақпараттарын қарастырып, кілттік сөздер арқылы индекстенген іздеу базасын құрады, содан кейін пайдаланушы сұранысы арқылы жүйе дұрыстығына (релевантты) қарай сайттар тізімін береді. Айқын, егер сіз сайтты «гүл» деген кілттік сөзбен іздесеңіз, онда іздеу машинасы сол сөз бар парақтарды тауып қана қоймай, бұл сөздің қай жерде сайт тақырыбына қатыстылығын анықтай алуы керек. Сөздің Web-парақтың профиліне қатыстығын анықтау үшін оның парақта қаншалықты жиі ұшырасатынын, берілген сөз туралы сілтемелердің бар-жоқтығын бағалау керек. Қысқаша айтқанда, парақта табылған сөздерді маңыздылық дәрежесіне қарай рангілеу керек.

Сөздерге салмақтылық коэфициенттері оның қанша және қайда

кездесетініне қарай (парақ тақырыбында, беттің басы не аяғында, сілтемеде, метатегте және т.б) меншіктеледі. әрбір іздеу механизмі салмақ коэфициенттерін берудің өз алгоритмдері бар – бұл әртүрлі іздеу машиналарының бір кілттік сөз арқылы сұрауға әртүрлі қорлар тізімін берудің бір себебі. Парақтар әрдайым жаңартылып отыратындықтан, онда индекстеу үрдісі де жиі орындалып отырылуы керек. Робот-өрмекшілер сілтемелерді аралай жүріп, индекстен тұратын файлды құрады, ол үлкен болуы мүмкін. Оның көлемін азайту үшін ақпарат көлемін минимизациялау мен файлды сығуға жүгінеді. өңделгеннен кейін мәліметтер үнемі жанарып отыратын базада сақталады. Бірнеше роботтары бар іздеу машинасы секундына жүздеген парақтарды өңдей алады. Бүгінде мықты іздеу машиналары жүздеген миллион парақты сақтайды және күніне ондаған миллион сұранысты қабылдайды.

Индексті құруда көшірмелердің санын азайту тапсырмасы да шешіледі – қатесіз салыстыру үшін алдымен құжаттың кодировкасын анықтау қажеттігін ескерсек, тапсырма оңай емес. Бұдан да қиын тапсырмаға өте ұқсас құжаттарды айыру жатады (оларды
«көшірме дерлік» деп атайды), мысалы оларға мазмұны бір ал тақырыбы әртүрлі болып келеді. Бұл сияқты құжаттар Желіде өте көп – мысалы біреу рефератты көшіріп алып өз сайтында басқа атпен басып шығаруы мүмкін. Қазіргі заманғы іздеу машиналары барлық бұл проблемаларды шешуге мүмкіндік береді.

Индекс арқылы іздеу

Индекс арқылы іздеу мынадан құралады, яғни пайдаланушы сұраныс құрастырып оны іздеу машинасына береді. Бірнеше кілттік сөздерді қолдануда сұраныс тілін пайдаланған пайдалы, оның негізін буль операторлары құрайды.
Ең жиі қолданылатын буль операторлары:

- AND – бұл арқылы біріктірілген барлық терминдер ұсынылған құжатта қатысуы керек. Кейбір іздеу жүйелері «+» белгісін «AND» орнына қолданады;
- OR – кем дегенде бір кілттік сөз «OR »-мен қатысты, ізделінетін құжатта болуы керек;
- NOT- «NOT»-тан кейінгі кілттік сөз ізделінетін құжатта кездеспеуі керек. Кейбір іздеу жүйелері «-» белгісін «NOT» орнына пайдаланады;
- FOLLOWED BY – кілттік сөздер бірінен кейін бірі кезектесіп келуі керек;
- NEAR – сөздердің бірі екінші сөзден белгілі санды сөздерден кейін келуі керек;
- Тырнақшалар – тырнақша ішіндегі сөздер- бұл текст фрагменті құжат немесе файл ішінде кездесуі тиіс. Айта кетейік, сұраныс тілі семантикасы нақты бір іздеу машиналарында бір біріне ұқсамауы мүмкін, әдетте ол туралы іздеу машинасының нұсқауында мәлімет келтіріледі.

Шектерінде логикалық комбинация анықталатын мәтін *іздеу бірлігі* деп аталады. Бұл сөйлем, абзац не бүкіл құжат болуы мүмкін. Түрлі іздеу жүйелерінде әртүрлі іздеу бірліктері қолданылуы мүмкін. Сөйлем шегіндегі іздеу тек индексінде толық мекенжай (адрес) бар жүйелерде ғана мүмкін.

Пайдаланушы іздеу жүйесіне сұраныс жібергеннен кейін, ол сұраныс синтаксисін өңдейді, кілттік сөздерді индекстегі сөздермен салыстырады. Содан кейін сұранысқа жауап беретін сайттар тізімі релеванттылығына қарай рангіленіп, пайдаланушыға берілетіндей іздеу нәтижесі құрастырылады.

Бақылау сұрақтары :

1. Интернетте қажетті ақпаратты іздеудің жалпы принциптері қандай?
2. Іздеу машинасының жұмыс механизмі қандай?
3. Интернетте ақпарат іздеу технологияларында индекс ұғымы нені білдіреді?
4. Индексті құру принциптері қандай?