

MAGİSTRATURA MƏRKƏZİ

Nemətli Nemət Vüqar oğlu

MAGİSTR DİSSERTASIYASI

f.-r.e.n.,dos. A.X.Abdullayev

f.-r.e.n.,dos. T.Ə.Əliyeva

Kafedra müdürü: akad. Ə.M.Abbasov

Bakı – 2018

M Ü N D Ə R İ C A T

GİRİŞ.....	3
Fəsil I. İnformasiya-axtarış sistemləri haqqında məlumat.....	7
1.1. İnformasiya-axtarış sistemlərinin tarixi.....	7
1.2. İnformasiya-axtarış sistemlərinin iş prinsipləri.....	14
1.3. İnformasiya-axtarış sistemlərinin tipləri.....	15
Fəsil II. İnformasiya-axtarış sistemlərinin arxitekturası.....	26
2.1. İnformasiya-axtarış sistemlərinin komponentləri.....	26
2.2. Veb axtarış robotları.....	28
2.3. İndeksləşdirmə modulu.....	44
2.4. Sorğu və sıralama modulu.....	60
2.5. İnformasiya-axtarış sisteminin verilənlər bazası.....	65
2.6. İstifadəçi interfeysi.....	68
Fəsil III. İnformasiya-axtarış sistemlərinin inkişaf istiqamətləri.....	71
3.1. Axtarış nəticələrinin yaxşılaşdırılması qaydaları.....	71
3.2. İnformasiya-axtarış sisteminin optimallaşdırılması.....	73
3.3. İnformasiya-axtarış sistemlərində istifadə olunan ən yeni alqoritmlər. Google və Yandex informasiya-axtarış sistemlərinin alqoritmləri.....	75
3.4. Müasir informasiya-axtarış sistemləri arasında fərqlər, axtarış nəticələrinin müqayisəsi.....	85
NƏTİCƏ VƏ TƏKLİFLƏR.....	97
ƏDƏBİYYAT SİYAHISI.....	101
PE3IOME.....	104
SUMMARY.....	105

GİRİŞ

Mövzunun aktuallığı. Hazırkı dövrdə yəni, “informasiya” əsrində çox az sayda insan lazım olan informasiyanı əldə etmək üçün həm yaxşı həm də ən asan vasitə olan informasiya-axtarış sistemləri olmadan həyatını təsəvvür edə bilər. İnsanların əksəriyyəti informasiya əldə etmək, araşdırmaq və məlumat mübadiləsi üçün etibarlı və birbaşa yollarla maraqlanır. Üstəlik, təqdim edilən informasiyalar hər zaman yenilənmiş və insanların tələblərinə uyğun olmalıdır. İnternet vasitəsilə bu gün böyük miqdarda məlumatları çox qısa bir müddət ərzində əldə etmək mümkündür. Lakin ilk vaxtlarda yəni, informasiyaların yaddaş qurğuları və ya serverlərdə yeni toplanmağa başladığı zamanlarda insanlar informasiyaları axtarmaq, tapmaq və sıralamaqda çətinliklə qarşılaşdılar. Bu da öz növbəsində informasiya-axtarış sistemlərinin yaranmasına təkan verdi.

Hər bir informasiya-axtarış sisteminin müvəffəqiyyəti istifadəçinin həmin informasiya-axtarış sistemində verdiyi axtarış sorğusuna müvafiq cavabın tapılıb-tapılmaması ilə müəyyən edilir. Buna görə də, hər bir informasiya-axtarış sisteminin ən vacib məqsədi axtarış nəticələrini daim yaxşılaşdırmaqdır. İstifadəçiləri müvafiq və düzgün axtarış nəticələri ilə təmin etmək üçün bir çox müxtəlif texnikalar, arxitekturalar, alqoritmlər və modellər hazırlanmış və tətbiq edilmişdir.

Hazırkı dövrdə internetdə daim artan informasiyaların miqdarı, fayl və saytların çoxalması nəticəsində istifadəçilərin özlərinə lazım olan informasiyaları əldə etməsi üçün informasiya-axtarış sistemlərinə çox böyük tələbat yaranmışdır. İnformasiya-axtarış sistemləri internetin sonsuz məlumat bazası içərisindən istifadəçiyə lazım olan informasiyanı ən sürətli və rahat formada istifadəçiyə təqdim edən kompleks sistemlərdir. Özünə göndərilən sorğulara cavab verə bilmək üçün hər bir informasiya-axtarış sistemi internetdəki məlumatları əvvəlcədən görmək məcburiyyətindədir. Başqa sözlə desək, istənilən informasiya-axtarış sistemi internetdə ancaq daha əvvəl gördüyü və xatırlaya bildiyi səhifələri istifadəçiyə təqdim edə bilər. Daha əvvəl görmədiyi və ya görsə də xatırlamadığı səhifələri istifadəçiyə təqdim edə bilməz.

Onlar eyniylə bir istifadəçinin brauzer vasitəsilə səhifələrə baxdığı kimi səhifələrə baxırlar və ordakı məlumatları anlamağa çalışırlar. Daha sonra isə məlumatları yaddaşlarında saxlayırlar və səhifədəki digər linkləri təqib edərək həmin səhifələrdə də eyni proseduranı tətbiq edirlər. Beləliklə, internetdə mümkün olduğu qədər çox sayda sayta baş çəkərək onları xatırlamağa çalışırlar. Səhifələri xatırlamaqda sözlərin böyük əhəmiyyəti vardır. Belə ki, informasiya-axtarış sistemləri sözlərin olduğu səhifələrin indekslərini yadda saxlayaraq hansı səhifədə hansı məlumatın olduğunu xatırlamağa çalışırlar. İnformasiya-axtarış sistemlərinin daha inkişaf etmiş formaları isə səhifədə olan sözlərin istifadə olunma sayını, sözlərin səhifədəki yerlərini, səhifə başlıqlarını və bu kimi məlumatlara diqqət yetirir və indekslərdə qeyd edir.

Səhifənin olduğu kimi saxlanması əvəzinə indekslənməsində məqsəq lazım olduqda səhifədəki məlumatlara əlçatanlığı asanlaşdırmaqdan, sürətli axtarışı təmin etməkdən və yaddaşa qənaət etməkdən ibarətdir. Məsələn, əlimizdə bir kitab var və bizə bu kitabda müəyyən bir mövzu lazımdır. Bu zaman biz necəki başlıqlara baxaraq özümüzə lazım olan səhifəni tapırıqsa informasiya-axtarış sistemlərində indeksləmə metodundan istifadə edərək axtarışı həyata keçirir. İstifadəçi tərəfindən informasiya-axtarış sisteminə sorğu daxil olduğu zaman axtarış sistemi indekslərinə baxır və sorğuda verilən sözlərin istifadə olunduğu səhifələri tapmağa çalışır. Daha sonra isə tapılan səhifələri müxtəlif kriteriyalara görə sıralayıb sonda istifadəçiyə təqdim edir.

Müasir cəmiyyətin informasiya tələbini ödəməkdə əsas rolu informasiya-axtarış sistemləri oynayır. Lakin çox az sayda insan gündəlik istifadə etdiyi bu sistemlərin arxitekturasından xəbərdardır. Bir çox informasiya-axtarış sistemlərinin arxitekturası haqqında informasiyalar cəmiyyətə təqdim edilmir. Buna səbəb həmin informasiyaların şirkət sirri olmasıdır. Başqa sözlə desək, informasiya-axtarış sistemlərinin arxitekturası aktual məsələlərdən biridir. Bu məqsədlə dissertasiya işi aktual bir məsələyə – informasiya-axtarış sistemlərinin arxitekturasının tədqiqinə həsr edilmişdir.

Tədqiqatın predmeti. Tədqiqatın predmetini müasir informasiya-axtarış sistemlərinin arxitekturasını təşkil edən əsas modul və alqoritmlərin araşdırılması,

onların üstün və mənfi cəhətlərinin öyrənilməsi, müxtəlif informasiya-axtarış sistemlərinin müqayisə edilməsi, informasiya-axtarış sistemlərinin optimallaşdırılması və axtarışın yaxşılaşdırılması haqda təklif və tövsiyələrin verilməsi təşkil edir.

Tədqiqatın obyektı. Tədqiqatın obyektini insanların gündəlik istifadə etdiyi müasir informasiya-axtarış sistemləri və onların müxtəlif növləri təşkil edir.

Tədqiqatın məqsədi. Tədqiqatın əsas məqsədi informasiya-axtarış sistemlərin tarixi inkişaf mərhələlərini araşdırmaq, müasir informasiya-axtarış sistemlərinin arxitekturasını təşkil edən modul və alqoritmləri tədqiq etmək, onların müsbət və mənfi xüsusiyyətlərini müəyyənləşdirməkdir.

Tədqiqatın informasiya bazası. Aparılan tədqiqatın informasiya bazasını Google, Bing, Yandex, Baidu, Yahoo! kimi ən çox istifadəçi sayına malik müasir informasiya-axtarış sistemləri haqqında statistik və tarixi məlumatlar, müxtəlif rus, ingilis və türk dilində olan elmi məqalə və ədəbiyyatlar, həmçinin müəllifin internetdə tədqiqat apararkən müxtəlif veb saytlardan əldə etdiyi məlumatlar təşkil edir.

Tədqiqatın elmi yeniliyi. Tədqiqatın elmi yeniliyi aşağıdakılardan ibarətdir:

- Ən çox istifadə olunan informasiya-axtarış sistemlərinin arxitekturası öyrənilmiş və əldə olunan məlumatlar ümumiləşdirilmişdir;
- İnformasiya-axtarış sistemlərinin keçmişdən müasir dövrə qədər olan tarixi araşdırılmışdır.
- Müxtəlif informasiya-axtarış sistemləri və onların mərkəzi komponentləri tədqiq edilmişdir.
- Həmçinin hər bir informasiya-axtarış sisteminin nüvəsini təşkil edən – axtarış robotları, indeksləşdirmə modulu, verilənlər bazası, istifadəçi interfeysi, sorğu və sıralama modulu kimi mühüm əhəmiyyətə malik komponentlərin hər biri ayrılıqda dərinlən araşdırılmış və əldə olunan biliklər ümumiləşdirilmişdir.
- Google və Yandex informasiya-axtarış sistemlərinin malik olduqları əsas alqoritmlər və onlarla birlikdə gələn yeniklər araşdırılmışdır.
- Google, Yandex, Yahoo!, Baidu və Bing kimi məşhur axtarış sistemləri arasında fərqlərin müəyyən edilməsi və onların qiymətləndirilməsi həyata keçirilmişdir.

- İnformasiya-axtarış sistemlərinin optimallaşdırılması məsələsinə diqqət yetirilmişdir.
- Axtarışın yaxşılaşdırılması üçün tövsiyyələr verilmişdir.

Tədqiqatın nəzəri və praktiki əhəmiyyəti. Magistr dissertasiya işi üçün aparılan tədqiqatın nəzəri və praktiki əhəmiyyəti müasir informasiya-axtarış sistemlərinin ayrı-ayrı növlərinin araşdırılması, mövcud sistemlər arasındakı fərqlərin təyin edilməsi, bu fərqlərə səbəb olan tərkib hissə və alqoritmlərin qiymətləndirilməsinə yönəlmişdir.

İşin strukturu və həcmi. Magistr dissertasiya işi giriş, üç fəsil, nəticə və təkliflər, ədəbiyyat siyahısından ibarət olmaqla kompüter yazısı ilə 105 səhifədir.

Giriş hissəsində mövzunun aktuallığı əsaslandırılmış, tədqiqatın predmeti, tədqiqatın obyektı və məqsədi, tədqiqatın informasiya bazası, onun elmi yeniliyi, nəzəri və praktiki əhəmiyyəti haqqında informasiya verilmişdir.

Magistr dissertasiya işinin “İnformasiya-axtarış sistemləri haqqında məlumat” adlanan I fəslində informasiya-axtarış sistemlərinin tarixindən, onların iş prinsipindən və müxtəlif tipli informasiya-axtarış sistemlərindən bəhs olunur.

Magistr dissertasiya işinin “İnformasiya-axtarış sistemlərinin arxitekturası” adlanan II fəsili məhz informasiya-axtarış sistemlərinin arxitekturasına həsr edilmişdir. Burada informasiya-axtarış sistemlərinin əsasını təşkil edən modullar haqqında məlumatlar, həmin modulların hansı işləri yerinə yetirməsi kimi məsələlər şərh edilmişdir.

Magistr dissertasiya işinin “İnformasiya-axtarış sistemlərinin inkişaf istiqamətləri” adlanan III fəslində axtarışın yaxşılaşdırılması qaydalarından bəhs edilmiş, informasiya-axtarış sisteminin optimallaşdırılması məsələsinə toxunulmuş, informasiya-axtarış sistemlərində istifadə olunan alqoritmlər haqqında məlumatlar verilmiş, müasir informasiya-axtarış sistemləri arasında fərqlərdən bəhs edilərək axtarış nəticələrinin müqayisəsi şərh edilmişdir.

Dissertasiya işinin sonunda nəticə və təkliflər və ədəbiyyat siyahısı verilmişdir.

Fəsil 1. İnformasiya-axtarış sistemləri haqqında məlumat

1.1. İnformasiya-axtarış sistemlərinin tarixi

Müasir informasiya-axtarış sistemlərində mövcud olan kompleks alqoritmlər nəticəsində onlar istifadəçinin sorğularını qəbul edir və adətən yüksək dəqiqliklə sorğuların nəticələrini istifadəçilərə təqdim edirlər. İnaxtarış sistemləri internetdəki geniş məlumat bazasında Data Mining edərək sorğulara cavab verirlər [11].

İnformasiya-axtarış sistemləri ilk prototiplərindən bu günə kimi uzun yol qət etmişdirlər. Müasir informasiya-axtarış sistemləri veb axtarış robotlarından və indeksləyicilərdən tutmuş, robot.txt kimi protokolların tanınmasına, səs və qrafiki formada axtarışların həyata keçirilməsinə, sırf müəyyən kriteriyara əsasən axtarış edilməsi kimi yeniliklərə sahibdirlər. Sürətlə inkişaf edən texnologiya əsrinin təsiri ilə müxtəlif funksiyalara malik ayrı-ayrı şirkətlər tərəfindən fərqli informasiya-axtarış sistemləri yaradıldı [26].

Alta Vista axtarış sistemi təbii dildə sorğularla işləyən ilk informasiya-axtarış sistemidir. Lycos əlaqə səviyyəsi siqnallarına görə təsnifat apararaq söz yaxınlığa sahib açar sözləri uyğunlaşdırmaqla lazımi informasiyaları tapırdı. Ask Jeeves faktiki istifadəçi axtarış sorğularında uyğun olaraq insan redaktorlardan istifadə etmişdir.

Veb axtarış sistemləri özləri 1990-cı ilin Dekabr ayında veb-in əsasını qoydular. “Who is” istifadəçi axtarışı 1982-ci ildə, “Knowbot İnformasiya Servisi” olan multi-şəbəkə istifadəçi axtarış sistemi 1982-ci ildə tətbiq olundular. Kontent fayllarını axtaran və ilk yaxşı strukturlaşdırılmış “Archie” adlı FTP faylları axtaran sistem 1990-cı ildə sentyabrın 10-unda işə salınmışdır.

1993-cü ilin Sentyabr ayının əvvəllərində yaradılmış “World Wide Web” yəni “www” tamamilə əl ilə indeksləşdirilmişdir. Orada Tim Berners-Lee tərəfindən redaktə olunmuş veb serverlərin siyahısı var idi və onlar CERN adlı veb serverdə yerləşdirilmişdilər. 1982-ci ilə aid veb serverlərin tarixi siyahısı mövcuddur lakin hər dəfə daha çox veb serverlər online olduqca mərkəzi siyahı bundan artıq veb serverlərin

adlarını özündə saxlaya bilmədi. NCSA saytında yeni serverlər “What’s New” başlığı adı ilə yerləşdirildilər.

İnternetdə ilk kontent axtarışı etmək üçün olan vasitə Archie adlanırdı. Onun adı arxiv mənaını verən “archive” sözündən götürülmüşdür, lakin burada “v” hərfi yox idi. Montrealda yerləşən McGill universitetinin kompüter elmləri fakültəsinin tələbələri Alan Emtage, Bill Heelan və J. Peter Deutsch tərəfindən yaradılmışdır. Proqram ictimai anonim FTP saytlarında yerləşən bütün faylların siyahısını yükləyib faylların adlarını axtara bilən verilənlər bazası yaratdı. Archie informasiya-axtarış sistemi saytların məzmununu indeksləşdirmirdi, çünki məlumatların miqdarı az idi və onları asanlıqla axtarmaq mümkün idi [26].

Gopher-in yaradılması (Mark McCahill tərəfindən Minnesota universitetində 1991-ci ildə yaradılmışdır) iki axtarış proqramı olan Veronica və Jughead-in yaradılmasına təkan verdi. Onlar Archie kimi Gopherin indeksləşirilmiş formada yaddaşda saxlanan faylların adlarını və başlıqlarını axtarırdılar. Veronica (Very Easy Rodent-Oriented Net-wide Index to Computerized Archives) Gopherin siyahısına daxil olaraq Gopherin menyu başlıqlarında açar sözlərlə axtarışı təmin edirdi. Jughead (Jonzy's Universal Gopher Hierarchy Excavation And Display) xüsusi və yaxud spesifik Gopher serverindən menyu informasiyasını əldə etmək üçün lazım olan vasitə idi.

1993-cü ilin yazında veb üçün heç bir informasiya-axtarış sistemi mövcud deyildi və çox sayda xüsusi kataloqlar vasitəsilə həmin axtarışlar təmin olunurdu. Geneva universitetindən olan Oscar Nierstrasz periodik olaraq səhifələrin sürətini çıxaran və onların standart formatda qeyd edən Perl dilində skriptlər seriyasını yazmışdır. “Wandex” adlanan veb axtarış robotunun əsas məqsədi “World Wide Web”-in ölçüsünü müəyyənləşdirmək idi ki, o bu işi 1995-ci ilin sonlarına qədər tamamladı. 1993-cü ilin Noyabrında ikinci axtarış sistemi olan “Aliveb” yaradıldı. Aliveb veb axtarış robotundan istifadə etmirdi, lakin o bunun yerinə veb administratorlar tərəfindən yaradılmış saytın müəyyən formatda olan indeks fayllarından xəbərdar idi.

NCSA şirkətinə məxsus Mosaic™ adlanan veb brauzeri ilk veb brauzer deyildi. Lakin o ilk böyük sıçrayış edən veb brauzer idi. 1993-cü ilin noyabr ayında Mosaic 1.0 versiyası yaradıldı. O özündə cəlbədicə interfeys, əlfəcinlər, şəkillər və.s kimi xüsusiyyətləri cəmləşdirməklə digər kiçik brauzerlərdən daha çox irəli getdi.

JumpStation adlı proqram 1993-cü ilin dekabrında Jonathon Fletcher tərəfindən yaradılmışdır. O veb səhifələri tapmaq və indekslərini yaratmaq üçün veb axtarış robotlarından istifadə edirdi və həmin veb formanı özünün sorğu proqramında interfeys olaraq istifadə edirdi. Beləliklə o “World Wide Web”-də resurs axtaran ilk vasitə idi ki, özündə axtarış, indeksləmə və araşdırmanı birləşdirirdi. Platformada mövcud olan məhdud resurslara görə onun indekslənməsi və dolayısı yolla axtarışları brauzerin səhifələrinin başlıqlarının məhdudlaşdırılması ilə yekunlaşırdı [26].

1994-cü ildə yaradılan “Veb Crawler”-in əsasını bütün mətnləri araşdırma təşkil edirdi. Digər axtarış sistemlərindən fərqli olaraq bu axtarış sistemi saytlarda hər hansı bir sözü tapmağa imkan yaradırdı və bu bütün informasiya-axtarış sistemləri üçün bir standartə çevrilmişdir. Bu da ictimaiyyət tərəfindən bilinən bir ilk olmuşdur. Həmçinin 1994-cü ildə “Lycos” (Carnegie Mellon Universitetində yaradılmış) böyük bir kommersiya çalışmasına çevrilmişdir.

Tezliklə bir çox informasiya-axtarış sistemləri meydana gəldi və populyarlıq qazandılar. Bunlara Magellan, Excite, Infoseek, İnttomi, Northern Light və AltaVista informasiya-axtarış sistemləri daxildir. Maraqlı veb səhifələri axtarıb tapmaq üçün ən populyar yollardan biri isə Yahoo idi. Lakin onun axtarış funksiyası veb səhifələrin tam mətn nüsxəsini deyil məhz veb kataloqunda axtarış etmək idi. İstifadəçilər açar sözlərdən istifadə etməklə axtarış etmək əvəzinə kataloqlara baxırdılar.

1996-cı ildə Netscape şirkəti Netscape veb axtarış brauzerinin köməkliyi ilə insanlara veb informasiya-axtarış sistemini eksklüziv formada təqdim etmək istədi. Netscape beş vacib informasiya-axtarış sistemləri ilə müqavilə bağladı. Onun bu beş informasiya-axtarış sistemi ilə müqavilə bağlamaqda əsas marağı ondan ibarət idi ki, hər il 5 milyon dollara qədər hər bir informasiya-axtarış sistemindən gəlir götürəcəkdə.

Həmin beş informasiya-axtarış sistemi Yahoo!, Magellan, Lycos, Infoseek və Excite idi.

Google şirkəti 1998-ci ildə goto.com adlı kiçik bir axtarış şirkətindən axtarış şərtlərini satın aldı. Və bunun nəticəsində internetdə ən sərfəli bizneslərdən biri olan informasiya-axtarış sistemləri biznesinin əhəmiyyətinə yüksək təsir göstərdi.

1990-cı illərdə internet sahəsinə olunan yatırımlardan ən karlısı məhz informasiya-axtarış sistemlərinə olunan yatırımlar idi. Bir çox şirkətlər bu sahədə çox böyük gəlirlər əldə etdilər. Bəziləri “Northern Light” kimi şirkətləri satın alırdılar.

2000-ci illərdə Google informasiya-axtarış sistemi önə çıxdı. Google şirkəti bir çox axtarışlar üçün PageRank adlı yeni bir innovasiyadan istifadə edirdi. Həmin bu yenilik Google şirkətinin yaradıcılarından olan Sergey Brin və Larry Page tərəfindən təklif olunmuşdur. Bu yeni PageRank adlanan alqoritm istənilən səhifəyə müəyyən dərəcə mənimsədirdi və daha çox baxış sayı olan səhifə və saytların informasiya-axtarış sisteminin nəticələr səhifəsində ilk sıralarda görünməsinə təmin edirdi. Google şirkəti özünün informasiya-axtarış sistemində çox sadə interfeys vermişdir. Google-un əksinə olaraq onun rəqibləri axtarış sistemlərinə veb portal yerləşdirdilər.

Yahoo! şirkəti 2000-ci ildə İntomi axtarış sistemində əsaslanan axtarış xidmətləri göstərmişdir. Yahoo! şirkəti 2002-ci ildə İntomi, 2003-cü ildə Overture şirkətlərini aldı. Yahoo! 2004-cü ilə qədər Google-un bazası əsasında işlədi, daha sonra satın aldığı şirkətlərin texnologiyalarına əsaslanan yeni texnologiyadan istifadə etdi.

Microsoft şirkəti 1998-ci ilin payızında İntomi şirkətinin axtarışlarının nəticələrindən istifadə edərək ilk dəfə olaraq “MSN Search” informasiya-axtarış sistemini yaratdı. 1999-cü ildən başlayaraq axtarış sistemi Looksmart, İntomi-nin nəticələrini qarışıq formada göstərməyə başladı, 1999-cü ildə qısa müddətliyə də olsa MSN informasiya-axtarış sistemi AltaVista sistemindən olan nəticələri istifadə etməyə başladı. 2004-cü ildən başlayaraq Microsoft özünün yeni texnologiyaları əsasında olan yeni axtarış sistemini istifadə etməyə başladı.

Microsoft-un yeni adlı axtarış sistemi olan – Bing informasiya-axtarış sistemi 1 iyun 2009-cu il tarixində işə salındı. 29 iyul 2009-cu ildə Yahoo! və Microsoft

şirkətləri yeni müqavilə bağladılar və müqavilə şərtlərinə əsasən Yahoo! şirkəti Microsoft Bing texnologiyası ilə təchiz olundu.

İnformasiya-axtarış sistemlərinin xronologiyasına aşağıdakı cədvəldən nəzər yetirmək olar:

Cədvəl 1. İnformasiya-axtarış sistemlərinin xronologiyası

Yaradıldığı il	Axtarış sistemi	Hazıki statusu
1993	W3Catalog	Aktiv deyil
	Aliweb	Aktiv deyil
	JumpStation	Aktiv deyil
	WWW Worm	Aktiv deyil
1994	WebCrawler	Aktivdir
	Go.com	Aktiv deyil
	Lycos	Aktivdir
	Infoseek	Aktiv deyil
1995	AltaVista	Aktiv deyil, Yahoo!-a yönləndirilib
	Daum	Aktivdir
	Magellan	Aktiv deyil
	Excite	Aktivdir
	SAPO	Aktivdir
	Yahoo!	Aktivdir
1996	Dogpile	Aktivdir
	İnktomi	Aktiv deyil, Yahoo! tərəfindən satın alındı
	HotBot	Aktivdir (lycos.com)
	Ask Jeeves	Aktivdir, yenidən adlandırıldı (ask.com)
1997	Northern Light	Aktiv deyil
	Yandex	Aktivdir

1998	Google	Aktivdir
	İxquick	Aktivdir
	MSN Search	Aktivdir, Bing-in tərkibində
	Empas	Aktiv deyil
1999	AlltheWeb	Aktiv deyil, Yahoo!-a yönləndirilib
	GenieKnows	Aktivdir, Yellowee.com olaraq yenidən adlandırıldı
	Naver	Aktivdir
	Teoma	Aktiv deyil, Ask.com-a yönləndirilib
	Vivisimo	Aktiv deyil
2000	Baidu	Aktivdir
	Exalead	Aktivdir
	Gigablast	Aktivdir
2001	Kartoo	Aktiv deyil
2003	Info.com	Aktivdir
	Scroogle	Aktiv deyil
2004	Yahoo! Search	Aktivdir, öz axtarış sistemini yaratdı
	A9.com	Aktiv deyil
	Sogou	Aktivdir
2005	AOL Search	Aktivdir
	GoodSearch	Aktivdir
	SearchMe	Aktiv deyil
2006	Soso	Aktiv deyil
	Quaero	Aktiv deyil
	Search.com	Aktivdir
	ChaCha	Aktiv deyil
	Ask.com	Aktivdir
	Live Search	Aktivdir Bing olaraq, MSN Search olaraq yenidən adlandırılmışdır

2007	Wikiseek	Aktiv deyil
	Sproose	Aktiv deyil
	Wikia Search	Aktiv deyil
	Blackle.com	Aktivdir, Google Search
2008	Powerset	Aktiv deyil, (Bing-ə yönləndirilmişdir)
	Picollator	Aktiv deyil
	Viewzi	Aktiv deyil
	Boogami	Aktiv deyil
	LeapFish	Aktiv deyil
	Forestle	Aktiv deyil, (Ecosia-a yönləndirilmişdir)
	DuckDuckGo	Aktivdir
2009	Bing	Aktivdir, Live Search olaraq yenidən adlandırılmışdır
	Yebol	Aktiv deyil
	Mugurdy	Aktiv deyil
	Scout (Goby)	Aktivdir
	NATE	Aktivdir
2010	Blekkko	Aktiv deyil, IBM-ə satıldı
	Cuil	Aktiv deyil
	Yandex (İngilis versiyası)	Aktivdir
2011	YaCy	Aktivdir, P2P veb axtarış sistemi
2012	Volunia	Aktiv deyil
2013	Qwant	Aktivdir
	Coc Coc	Aktivdir, Vyetnam dilində axtarış sistemi
	Egerin	Aktivdir, Kürd dilində axtarış sistemi
2015	Cliqz	Aktivdir, brauzerə integrasiya edilmiş axtarış sistemi

1.2. İnformasiya-axtarış sistemlərinin iş prinsipləri

İnformasiya-axtarış sistemləri istifadəçiləri onları maraqlandıran məlumatlarla bağlı və ya həmin məlumatlarla oxşarlıq təşkil edən bir qrup elementlər təqdim edən interfeyslə təmin edir. Kriteriyaya axtarış sorğusu deyilir. Mətn axtarış sistemlərinə nəzər saldıqda axtarış sorğuları adətən istənilən konsepsiyanı müəyyən edən bir sıra sözlər kimi ifadə edilir. Sıxılıqla fərqlilik göstərən bir neçə axtarış sorğusu söz düzülüş forması mövcuddur. İnformasiya-axtarış sistemlərindəki əvvəlki saytların adları da dəyişdirilə bilər. Bəzi mətn axtarış sistemləri istifadəçilərə ağ boşluqla ayrılmış iki və ya üç sözü daxil etməsini tələb edərkən, digər informasiya-axtarış sistemləri istifadəçilərə bütün sənədləri, səsləri və təbii dilin müxtəlif formalarını təyin etməyə imkan verə bilər. Bəzi informasiya-axtarış sistemləri sorğuların keyfiyyətini yaxşılaşdırmaq məqsədilə axtarış sorğularına təkmilləşdirmələr tətbiq edirlər. Sorgu anlayış metodları sorgu dillərini standartlaşdırmaq üçün istifadə edilə bilər [1].

Sorğunun müəyyən kriteriyalarına cavab verən nəticələr siyahısı adətən çeşidləndirilir və səviyyəyə görə sıralanır. Səviyyəyə görə sıralama ən yüksəkdən ən zəifə görə aparılır. Sıralama istifadəçiyə özünə lazım olan informasiyanı ən qısa müddətdə tapmağa imkan verir. Ehtimallı axtarış sistemləri oxşarlıq təşkil edən nəticələr sıralaması həyata keçirir. Boolean axtarış sistemləri əsasən boolean-stil söz düzümü edir yəni, AND, OR, NOT və XOR operatorlarından istifadə edir.

Bəzi kriteriyalara uyğun olaraq sıralanan uyğun nəticələrdən ibarət bir sıra təmin etmək üçün informasiya-axtarış sistemi adətən nəzərə alınan nəticələr qrupu haqqında meta-verilənləri toplayır. Hər bir veb saytın yerini informasiya-axtarış sisteminin yaddaşda saxlaması üçün kifayət qədər yaddaş tələb olunur, bu da sistemin iş sürətinə mənfi təsir edə bilər. Bu məqsədlə saytların ünvanları indekslənməmiş formada informasiya-axtarış sistemlərinin yaddaşında saxlanılır. Bu metod informasiya-axtarış sistemlərinin axtarış prosesindən sonra əldə etdikləri nəticələrin ünvanlarını istifadəçiyə ən sürətli formada təqdim etməyə imkan verir. Alternativ olaraq, informasiya-axtarış sistemi hər bir elementin sürətini keş yaddaşda saxlaya bilər, belə

ki indeksləşdirilmiş və ya arxiv məqsədləri üçün daha effektiv və tez işləməsini təmin edə bilər. Digər növ axtarış sistemləri indeks saxlamır [1].

Crawler və ya Bot tipli veb axtarış robotları axtarış sorğusunun elementlərini real vaxt ərzində toplaya və qiymətləndirə bilirlər. Meta-axtarış sistemləri nə indeks nə də ki keş məlumatları toplamırlar. Bunun əvəzinə onlar bir və ya daha çox informasiya-axtarış sistemlərinin nəticələrini bir yerə toplayıb emal edərək nəticə çıxarırlar və sonra nəticəni istifadəçiyə təqdim edirlər.

1.3. İnformasiya-axtarış sistemlərinin tipləri

İnformasiya-axtarış sistemlərinin tarixi xronologiyasına nəzər saldıqda məlum olur ki keçmişdən bu günə kimi müxtəlif tipli axtarış sistemləri yaradılmışdır. Həmin informasiya-axtarış sistemləri müxtəlif kriteriyalara görə bir-birlərindən fərqlənilirlər.

Mənşəyinə görə informasiya-axtarış sistemlərinin aşağıdakı tipləri mövcuddur [26]:

1. Fərdi kompüterlər üçün axtarış sistemləri
2. Birləşmiş axtarış sistemləri
3. İnsan köməyi ilə indekslənən axtarış sistemləri
4. Meta-axtarış sistemləri
5. İkili axtarış sistemləri
6. Veb axtarış sistemləri

Axtarılan məzmunun növünə görə informasiya-axtarış sistemlərinin tipləri aşağıdakılardır:

1. Tam mətn axtarış sistemləri
2. Şəkil axtarış sistemləri
3. Video axtarış sistemləri
4. Musiqi axtarış sistemləri

İnterfeysinə görə informasiya-axtarışının aşağıdakı tipləri mövcuddur:

1. Artan axtarış
2. Dərhal cavab
3. Semantik axtarış

4. Seçim əsaslı axtarış

Mövzuya görə axtarışın tiplərinə aşağıdakılar daxildir:

1. Biblioqrafik verilənlər bazasında axtarış
2. Müəssisə daxili axtarış
3. Tibbi ədəbiyyat axtarışı
4. Vertikal axtarış

➤ *Fərdi kompüterlər üçün axtarış sistemləri*

Fərdi kompüterlər üçün axtarış sistemləri, istifadəçinin internet mühitində axtarış etməyindən fərqli olaraq, onların kompüterdə axtarış etməsinə xidmət edirlər. Bu vasitələr veb-brauzer tarixçəsi, e-poçt arxivləri, mətn sənədləri, səs faylları, şəkillər və videolar da daxil olmaqla istifadəçinin fərdi kompüterində mövcud olan məlumatları əldə etmək üçün nəzərdə tutulmuşdur. Fərdi kompüterlər üçün axtarış sistemlərinin əsas üstünlüklərindən biri, axtarış nəticələrinin düzgün indekslərin istifadəsi nəticəsində qısa zaman müddətində istifadəçinin sorğusuna uyğun nəticənin nümayiş etdirilməsidir.

Müxtəlif fərdi kompüterlər üçün axtarış proqramları artıq mövcuddur. Fərdi kompüterlər üçün axtarış proqramlarının əksəriyyəti müstəqil tətbiqlərdir, bir çoxları da integrasiya olunmuş mühitdə axtarış imkanlarını təmin edirlər.

Fərdi kompüterlər üçün axtarış sistemləri iki əsas səbəbdən böyük firmalara narahatlıq yaratmışdır: səmərəsizlik və təhlükəsizlik. Bir tərəfdən, istifadəçilər müvafiq faylları tez-tez tapa bilməlidirlər, digər tərəfdən isə məhdud faylları əldə etməməlidirlər. Analitik firma Gartnerə görə, bəzi şirkətlərin məlumatlarının 80% -i strukturlaşdırılmış olmayan məlumatların içərisində kilidlənmişdir.

Tarixən fərdi kompüterlər üçün masaüstü axtarış sistemi ilk dəfə Apple şirkətinə məxsus fərdi kompüterlərdə 1990-cı illərin əvvəllərində tətbiq olunmuşdur. Apple kompüterlərində istifadə olunan ilk axtarış sistemi Sherlock adlanırdı və Mac əməliyyat sistemi üzərində işləyirdi.

Fərdi kompüterlər üçün axtarış sistemləri iri həcmli (əsasən GB-larla ölçülən) faylları axtararkən maksimum iş tezliyi əldə etmək məqsədilə indeksləmə prosesini həyata keçirir və indeksləri yaddaşda saxlayırlar. Əsasən indeksləşdirmə prosesi vaxt tələb edən prosesdir və fərdi kompüterin iş görmədiyi zamanlarda yerinə yetirilir.

Fərdi kompüterlər üçün axtarış sistemləri adətən fayllar haqqında üç növ məlumat toplayırlar:

- Fayl və qovluq adları
- Meta-verilənlər, MP3, PDF və JPEG kimi fayl tiplərində başlıqlar, müəlliflər, şərhlər
- Fayl məzmunu (yalnız dəstəklənən sənəd növləri üçün)

Fərdi kompüterlər üçün olan masaüstü axtarış sistemlərinə nəzər salaq:

- *Windows Əməliyyat Sisteminin axtarış sistemi*

Hazırda fərdi kompüterlər üçün olan masaüstü axtarış sistemlərindən ən geniş yayılmışı Microsoft Corporation tərəfindən yaradılan Windows Search-dır. Bu sistem fərdi kompüterdə yerləşən kataloq və faylların yerlərini indeksləşdirərək yaddaşda saxlayır. Bu sayədə istifadəçinin verdiyi axtarışlara tez və effektiv formada cavab verə bilər. İndeksləşdirmə metodu ilk dəfə 1996-cı ildə meydana gəldi və əsasən şəxsi və korporativ kompüter şəbəkələrində faylların axtarışını sürətləndirmək məqsədilə yaradılmışdır. İndeksləşdirmə xidməti, Microsoft veb serverlərindən istənilən sabit diskə faylları endirmək üçün istifadə olunurdu. İndeksləşdirmə fayl formatı ilə həyata keçirilmişdir. İstifadəçilərin təqdim etdiyi şərtləri istifadə edərək, fayl formatları içərisində olan məlumatlarda axtarışlar aparırdı. İndeksləşdirmə xidmətinin qarşılaşdığı ən böyük problem yeni fayl əlavə olunduqdan sonra indekslənməsinin vacib olmasıdır. İndeksləşdirmə prosesi RAM-da fayl və kataloqların indekslərini saxlayırdı və bu da öz növbəsində kompüterin məhsuldarlığını aşağı salırdı. Böyük həcimdə faylların indeksləşdirilməsi üçün güclü texnika və uzun müddət tələb olunurdu. Hal-hazırda Windows Search axtarış sistemində bu proseslər qısa bir zaman müddətində icra

olunur və axtarış verilən zaman ən qısa müddət ərzində axtarışın nəticəsi istifadəçiyə təqdim edilir. Buna səbəb sürətlə inkişaf edən texnologiya və müasir kompüterlərdə istifadə edilən RAM, prosessor kimi detalların keçmişdəkilə nəzrən daha güclü və məhsuldar olmasıdır. Windows Search axtarış sistemi çalışdığı zaman prosessoru gücünün böyük bir hissəsini istifadə edir.

Windows Vistanın yaradılmasından sonra Windows Search 3.1 meydana gəlmişdir. Sələfləri WDS və Windows Search 3.0-dan fərqli olaraq həm indekslənmiş həm də indekslənməmiş fayl və kataloqları problemsiz şəkildə axtara bilirdi. Əlavə olaraq Ram və prosessoru olan tələbatı böyük ölçüdə azaldıldığına görə indeksləşdirmə müddətində azaldı. Windows 7 və sonrakı versiyalarda Windows Search 4.0 və onun yeni nəsil modifikasiyaları istifadə olunur.

- *Mac Əməliyyat Sisteminin axtarış sistemi*

Mac Əməliyyat Sistemi Apple Search axtarış sistemi ilə fərdi kompüterlər üçün masaüstü axtarış sistemini tətbiq edən ilk əməliyyat sistemidir. Beləliklə bu sistem istifadəçilərə Macintosh kompüterlərdəki faylları format tipinə görə, bu fayllardakı meta verilənlər və fayllar içindəki məzmun da daxil olmaqla axtarış etməyə imkan verirdi. Apple Search axtarış sistemi müştəri-server proqramı idi və beləliklə əsas kompüterdən əlavə ayrı server tələb olunurdu. İlk yaradılan və Macintosh kompüterlərdə tətbiq olunan axtarış sistemi 5 MB RAM və MC68040 prosessoruna ehtiyac duyurdu.

1997-ci ildə Sherlock masaüstü axtarış sistemi Mac OS 8.5 ilə birlikdə buraxılmışdır. Sherlock (məşhur dedektiv Sherlock Holmesin adına) Mac Əməliyyat sisteminin brauzerinə də inteqrasiya edilmişdir. Beləliklə Sherlock axtarış sistemi, masaüstü axtarış funksiyasını internetə giriş əldə etməklə genişləndərək, istifadəçilərə həm kompüterin sərt diskində, həm də internetdə axtarış etməyə imkan verdi.

Spotlight axtarış sistemi Mac OS X 10.4 Tiger-in bir hissəsi olaraq 2005-ci ildə yaradıldı. Spotlight ilk vaxtlarda uzun başlama müddətinə sahibdi, lakin sərt disk indeksləşdirildikdən sonra bu müddət azaldı. Fayllar istifadəçi tərəfindən əlavə

edildikdə, indeksləmə minimal prosessor və RAM resurslarından istifadə edərək arxa planda hər zaman yenilənirdi.

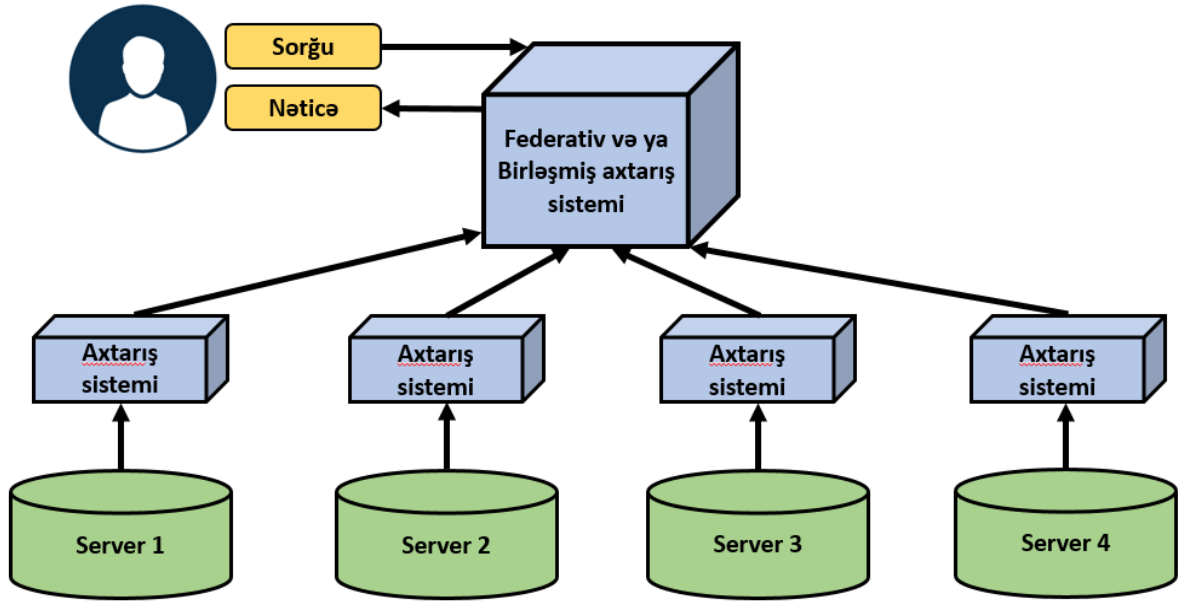
- *Linuks Əməliyyat Sisteminin axtarış sistemi*

Linuks istifadəçilərinin bacarıq səviyyələrindən və kompüterdə quraşdırılmış alətlərdən funksional əməllər ilə və ya brauzer əsaslı (adətən inkişaf etmiş skriptləmə seçimləri ilə) istifadəçi interfeysindən asılı olaraq kompüterdə geniş axtarış seçimləri təqdim edir. Əlavə olaraq bir çox istifadəçi verilmiş indeks bazasından özünün indeksləşdirmə paketlərini yarada bilər. Bir axtarış robotu paketdən çıxarılan PDF, DOC, DOCX, ODT fayllarının indeksləşdirilməsi işini reallaşdırarkən, digər axtarış robotu isə W/V, LDAP və digər kataloq verilənləri bazaları ilə işləyir.

- *Birləşmiş axtarış sistemləri*

Birləşmiş axtarış və ya federativ axtarış sistemi birdən çox axtarışda olan resursların eyni vaxtda axtarışına imkan verən informasiya-axtarış texnologiyasıdır. Bir istifadəçi birləşmədə və ya federasiyada iştirak edən axtarış sistemlərinə, məlumat bazalarına və ya digər sorğu sistemlərinə paylanan sorğu göndərir. Axtarışdan sonra axtarış sistemlərindən istifadəçiyə təqdimat üçün alınan nəticələr birləşdirilir. Bu tez-tez internetdə fərqli informasiya resurslarını birləşdirmək üçün istifadə olunan bir üsuldur. Bir çox məlumat mənbələrini böyük bir axtarış nəticəsi daxilində birləşdirmək üçün istifadə olunur. Bu axtarış sistemləri istifadəçiyə real vaxtda birdən çox verilənlər bazasında axtarış etməyə, müxtəlif verilənlər bazalarından axtarılan məzmunu uyğun nəticələr alan və nəticəni istifadəçiyə təqdim etməyə imkan verən bir sistemlərdir.

Beləliklə bir çox məlumat resurslarına vahid nöqtəyə çıxış təmin edilir və adətən məlumatları standart və ya qismən homogenləşdirilmiş formasında qaytarır. Müəssisə daxilində belə sistemlərin tətbiqi mümkündür. Məsələn, müəssisədə birdən artıq yerdə fayllar saxlanırsa bu zaman axtarış veriləndə axtarış sistemi hər iki yerə sorğu göndərir və onların nəticələrini birləşdirərək istifadəçiyə təqdim edir. Birləşmiş və ya federativ axtarış sisteminin strukturunu Şəkil 1-dəki kimi göstərmək olar:



Şəkil 1. Birləşmiş axtarış sisteminin strukturu

➤ *İnsan köməylə indekslənən informasiya-axtarış sistemləri*

İnsan köməylə indekslənən informasiya-axtarış sistemləri axtarış nəticələrini süzgəcdən keçirmək və istifadəçilərə onların axtarış tələbini aydınlaşdırmaqda yardım göstərmək üçün insan əməyindən istifadə edən axtarış sistemləridir. Məqsədi istifadəçilərə müvafiq nəticələr verə biləcək və ya olmaya biləcək çox sayda nəticə verən ənənəvi informasiya-axtarış sistemlərindən fərqli olaraq, məhdud sayda müvafiq nəticələr təqdim etməkdir. Bu informasiya-axtarış sistemləri veb səhifələrin siyahılarını əldə edir. Kataloqa saytın ünvanı, adı və qısa təsviri daxildir. Kataloqların üstünlüyü ondan ibarətdir ki, bütün resurslar əl ilə idarə olunur, buna görə də məzmunun keyfiyyəti digər növ sistemlər tərəfindən avtomatik olaraq alınan nəticələrdən daha yaxşı olacaqdır. Ancaq bir mənfi cəhəti vardır – kataloqda məlumatların yenilənməsi əl ilə aparılır və əhəmiyyətli dərəcədə vaxt tələb edə bilər. Səhifə sıralanması istənilən an dəyişdirilə bilməz.

İnsan köməylə indekslənən informasiya-axtarış sistemlərinin nümunələrinə ChaCha, Mahalo.Com və hazırda mövcud olmayan Nownow (Amazon.com-dan) və Sproose daxildir.

➤ *Meta-Axtarış sistemləri*

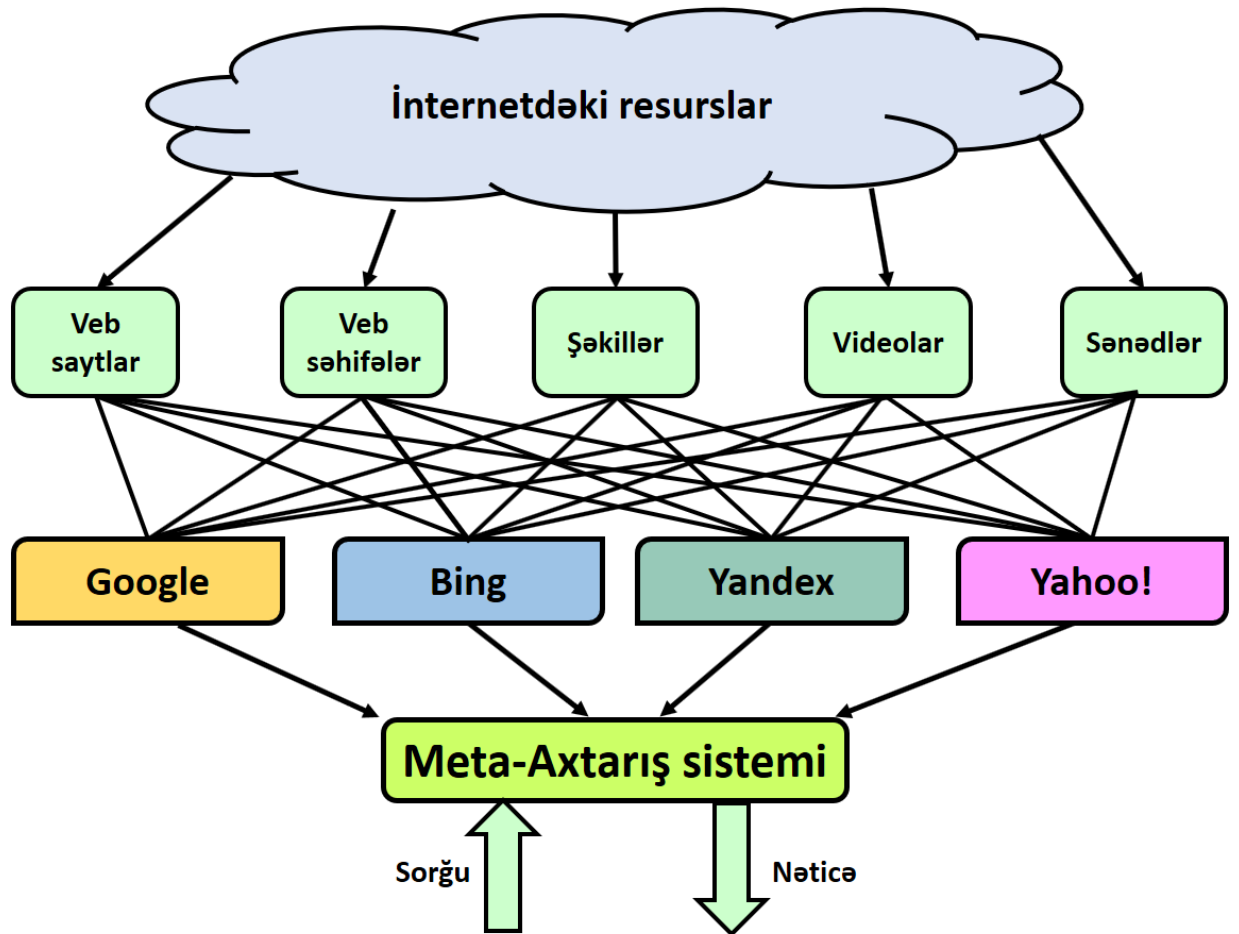
Meta-axtarış sistemləri internetdən öz nəticələrini toplamaq üçün başqa informasiya-axtarış sistemlərinin sorğularının nəticələrini istifadə edən axtarış sistemləridir. Meta-axtarış sistemləri istifadəçi tərəfindən sorğuları qəbul edərək üçüncü tərəf informasiya-axtarış sistemlərinə özü sorğular göndərir. Lazımı nəticələr toplandıqdan sonra hər bir informasiya-axtarış sistemi öz alqoritmlərinə görə nəticələri qiymətləndirərək onları sıralayır. Daha sonra onları meta-axtarış sisteminə göndərir. Meta-axtarış sistemi özü alınan nəticələri sıralayaraq istifadəçiyə təqdim edir.

Təbii ki, meta-axtarış sistemlərinin də müəyyən problemləri mövcuddur. İnformasiya-axtarış sistemlərində saxlanılan veb saytlar müxtəlif ola bilər. Bu da aidiyyatı olmayan nəticələrin meydana çıxmasına səbəb ola bilər. Digər problem isə spam-lardır. Spam kimi problemlər də axtarışın dəqiqliyini əhəmiyyətli dərəcədə azaldır.

İstifadəçilərin müəyyən bir sahəni əhatə edən xüsusi məlumatları əldə etməsinə imkan verən Meta-axtarış sistemlərinin bir çox növləri var. Bunlara Savvysearch, Search.com, Metaseek, Metacrawler, Dogpile, Startpage, Incrawler, Vroosh, Turboscout, Unabot, Drazz, Yippy, Allplus meta-axtarış sistemləri daxildir.

Meta-axtarış sistemlərinə sorğu göndərildiyində o digər informasiya-axtarış sistemlərinə həmin sorğunu yönəldir. Bunun ən böyük üstünlüyü ondan ibarətdir ki, istifadəçi minimal zaman müddəti ərzində sorğusuna uyğun ən yaxşı nəticəni əldə edir.

Bildiyimiz kimi hər bir informasiya-axtarış sistemi müxtəlif alqoritmlərdən istifadə edir. Bu da öz növbəsində sorğuların nəticəsinə təsir edir. Yəni informasiya-axtarış sistemləri fərqli olduğu üçün alınan nəticələrdə müxtəlif ola bilər. Meta-axtarış sistemi olmasaydı istifadəçi bütün axtarış sistemlərində tək-tək axtarış etməli olardı. Nəzərə almaq lazımdır ki bu uzun vaxt tələb edən prosesdir. Meta-axtarış sistemi istifadəçiləri bu kimi problemlərdən xilas edir, onlara sorğularına ən uyğun nəticələri göstərir. Meta-axtarış sisteminin strukturunu Şəkil 2-dəki kimi göstərmək olar:



Şəkil 2. Meta-axtarış sisteminin strukturu

Meta-axtarış sistemlərinin çatışmayan cəhətləridə mövcuddur. Meta-axtarış sistemləri sorğu formalarını şifrələməyə və sorğu sintaksisini tam tərcümə etməyə qadir deyil. Meta-axtarış sistemləri tərəfindən yaradılan linklərin sayı məhduddur. Buna görə də istifadəçiyə sorğunun tam nəticələrini verməyə bilər. Əlavə olaraq meta-axtarış sistemləri sponsorlu veb saytlara əsas üstünlük verir və onları ilk sıralarda göstərir.

Meta-axtarış sistemlərinin özlərinin verilənlər bazası mövcud deyil. Lakin sorğunun göndərdiyi digər informasiya-axtarış sistemləri özlərinə aid verilənlər bazalarına malikdirlər. Bu səbəblə birdən çox mənbədən məlumat inteqrasiyası üçün virtual verilənlər bazası yaradılır.

Hər bir informasiya-axtarış sistemi özlüyündə unikaldir və sıralanan məlumatları yaratmaq üçün fərqli alqoritmlərə malikdirlər, beləliklə eyni nəticə cütləri də yaranacaqdır. Eyni olanları aradan qaldırmaq üçün meta-axtarış sistemi bu nəticələrə

öz alqoritmini tətbiq edərək emal edir. Yenidən tərtib edilmiş siyahı istifadəçi üçün nümayişə hazırlanır. Meta-axtarış sistemi digər informasiya-axtarış sistemlərinə müraciət etdikdə, bu informasiya-axtarış sistemləri üç yolla cavab verəcəkdir:

1. Hər ikisi əməkdaşlıq edəcəklər, onlar meta-axtarış sisteminə öz indeks bazalarına giriş imkanı verəcək və meta-axtarış sistemi üçün tam interfeysə çıxışı təmin edəcəkdir. Onlar öz indeks bazalarındakı bütün dəyişikliklər barədə meta-axtarış sisteminə informasiya verəcəklər.
2. Digər informasiya-axtarış sistemləri meta-axtarış sistemi ilə əməkdaşlıq etməyəcəkdir, lakin meta-axtarış sisteminin onların interfeysinə daxil olmalarına mane olmayacaqlar.
3. İnformasiya-axtarış sistemi meta-axtarış sistemini tamamilə rədd edə bilər. Yəni, meta-axtarış sisteminin öz verilənlər bazasına müraciətini hüquqi müstəvidə qadağan edə bilər.

➤ *İkili axtarış sistemləri*

İkili axtarış sistemləri özündə həm digər informasiya-axtarış sistemlərinin həm də meta-axtarış sistemlərinin xarakteristikalarını saxlayan və əlavə olaraq istifadəçilər tərəfindən daha əvvəl təsnif edilmiş axtarış nəticələrinin toplusunu özündə birləşdirən sistemlərdir. Bu istifadəçiyə öz axtarış indeksindən və ya bir və ya birdən çox axtarış sistemlərindən, meta-axtarış sistemlərindən, verilənlər bazasından və ya hər hansı bir informasiya-axtarış sistemindən nəticə əldə etməyə imkan verir. İkili axtarış sistemləri, kompüter axtarış proqramlarının imkanlarını insan tərəfindən təsnif edilən nəticələrlə birləşdirən avtomatlaşdırılmış informasiya-axtarış sistemidir.

İkili axtarış sistemləri ənənəvi meta-axtarış sistemlərindən görünməyən rahatlıqla birdən çox axtarış botunun imkanlarını istifadə etmək üçün bir üsuldur.

İstifadəçiyə ikili axtarış sistemləri dəyişdirilə bilən axtarış robotu kimi görünə bilər, lakin, “səhnə arxası” texnologiyası onu axtarış prosesinə interfeys, həmçinin ikili axtarış sistemləri seansı ərzində insan tərəfindən təsnif edilmiş və avtomatik olaraq sənədlərin indeksinə daxil edilmiş nəticə toplusunu əldə etməyə imkan verir. Bir

çox informasiya-axtarış sistemlərində və meta-axtarış sistemlərində əlavə xüsusiyyətlər var, lakin fərdi axtarışların dəqiqliyini və uyğunluğunu artırmaqla və nəticələrini idarə etmək bacarığı ilə resursların axtarışı üçün lazım olan vaxtın azaldılması ideyası hər birində eynidir.

➤ *Veb Axtarış sistemləri*

İnformasiya-axtarış sistemi insanların internetdə və ya istifadəçinin kompüterində fərqli-fərqli məlumatları tapmasında ona kömək edən vasitədir. Bir çox istiqamətdə işləyən müxtəlif axtarış sistemləri vardır.

Veb axtarış sistemləri, istifadəçiyə internetdə olan məlumatları tapmağa kömək edir. Açar söz axtarışı ən az 2 istiqamətdən qiymətləndirilə bilər: Birincisi istifadəçi tərəfindən daxil edilən açar sözlə eyni olan sənədləri (və ya bir neçə açar sözdən ibarət olan bir cümlə) tapmaq və ya daxil edilən açar söz ilə əlaqədar sənədləri tapmaq. Son istifadəçi üçün informasiya-axtarış sistemi, axtarış açar sözlərini yazıb, nəticədə əlaqəli səhifələrə və ya sənədlərə əlaqələrin olduğu bir interfeys formasına sahib veb saytdır.

Nəticə səhifəsi ayrıca hər bir link üçün təsviri mətn məlumatı və adətən bir sıra aidiyyatı reklamlardan ibarət olur. Bu gün informasiya-axtarış sistemləri qapalı və kommersiya cəhətdən qorunan daxili quruluşa malik inkişaf etmiş proqram və avadanlıq sistemləridir. Qısacası, hər bir informasiya-axtarış sisteminə, hər bir linki izləyərək veb üzərindən səyahət edən və verilənlər bazalarına aid sənədlər haqqında məlumatları toplayan axtarış robotları (bot, crawler və ya spider) daxildir. Daha sonra məlumatlar sənədlərdən çıxarılır və xüsusi alqoritmlərlə indeksləşir. Hər bir sənəd və söz sıralanır və reytingə görə düzülür. Sonda istifadəçiyə, sorğusu ilə əlaqəli məlumatların yerləşdiyi saytlar xüsusi nizama görə sıralanmış veb səhifə siyahısı kimi ən yaxşı formada təqdim edilir.

Veb axtarış sistemləri içərisində ən geniş istifadə edilən informasiya-axtarış sistemi Google –dur. Ondan sonrakı yerləri isə ardıcılıqla, Bing, Yahoo!, Baidu, Yandex, Ask, AOL, Excite və digər veb informasiya-axtarış sistemləri tutur.

Google -un məşhurluğuna baxmayaraq, bəzi ölkələrdə yerli məşhur informasiya-axtarış sistemləri vardır. Misal üçün, Baidu Çində aparıcı informasiya-axtarış sistemidir. Yandex isə rusdilli ölkələrdə ən məşhur informasiya-axtarış sistemidir və bu ölkələrdə bazar payının 60%-dan çoxunu tutur.

Fərqli internet informasiya-axtarış sistemləri müəyyən bir açar söz üçün hansı veb səhifələrinin daha əlaqəli olduğunu və hansı veb səhifələrin axtarış sisteminin nəticə səhifəsinin ən üstündə görünməsi lazım olduğunu təyin etmək üçün fərqli alqoritmlər istifadə edirlər. İstifadəçilər təbii olaraq onlara ən yaxşı və ən əlaqəli nəticələri verəcək informasiya-axtarış sistemini seçirlər.

Unikal alqoritmləri ən əlaqəli nəticələri çıxarmağa çalışdığı üçün, informasiya-axtarış sistemləri çox vaxt axtarış alqoritmləri ilə müdafiə olunurlar. Məşhur informasiya-axtarış sistemləri sorğuya ən uyğun nəticələr təqdim edənlərdir.

Müasir informasiya-axtarış sistemləri istifadəçilər həm açar sözlərdən istifadə edərək həm də şəkil, video, səs vasitəsilə axtarış etməyə imkan verir. Bu cəhətdən informasiya-axarış sistemləri daima özlərini inkişaf etdirirlər.

Ümumiyyətlə axtarış sistemlərinin içərisində ən çox istifadə olunanı veb informasiya-axtarış sistemləridir. Gündəlik həyatda hər bir insan həmin sistemlərdən mütamadi olaraq istifadə etməkdədir. Veb informasiya-axtarış sistemləri mövcud informasiya cəmiyyətinin informasiya ehtiyacını ödəmək üçün yaradılmış vasitədir. Veb informasiya-axtarış sistemləri istifadəçiyə sadəcə sorğu xanasına istədikləri sözü daxil edərək saniyələr ərzində çoxlu nəticələr almağa xidmət edən sistemlərdir [4].

Fəsil 2. İnformasiya-axtarış sistemlərinin arxitekturası

2.1. İnformasiya-axtarış sistemlərinin komponentləri

Müasir veb informasiya-axtarış sistemlərinin əsas vəzifəsi istifadəçiyə lazım olan informasiyanı ona ən qısa müddət ərzində və dəqiqliklə çatdırmaqdan ibarətdir. Axtarış sistemlərinə olan əsas tələblər [1]:

1. Alınan nəticələrin dolğunluğu
2. Axtarış sisteminin dəqiqliyi
3. Alınan nəticələrin aktuallığı
4. Axtarış sürəti
5. Aydınlığı

Hazırda məşhur veb informasiya-axtarış sistemlərinin demək olarki hər biri indeksləşdirmə üsulunu dəstəkləyən prinsip üzrə işləyirlər. İndeksləşdirmə üsulunun tətbiqi nəticəsində informasiya-sistemlərinin axtarış sürətləri hiss ediləcək dərəcədə artdı. Belə ki, informasiya-axtarış sistemləri öz verilənlər bazalarında hər bir nəticəyə uyğun linkləri saxlamaq yerinə hər bir linki müvafiq alqoritmlərə əsasən indeksləşdirməklə yadda saxlayırlar. İnformasiya-axtarış sisteminə sorğu verildiyi zaman sorğu modulu həmin indeksləşdirmə moduluna müraciət edərək lazım olan linklərin bazadakı ünvanları götürür. Müvafiq olaraq reytingə yəni, baxış sayına uyğun olaraq sıralanaraq istifadəçi interfeysi vasitəsilə istifadəçiyə sıralanmış formada nəticələr təqdim edilir. Bütün müasir veb informasiya-axtarış sistemlərinin arxitekturası aşağıdakı komponentlərdən ibarət olurlar:

- Veb axtarış robotları - Crawler, Spider və ya Bot
- İndeksləşdirmə modulu
- Verilənlər bazası
- Sorğu və sıralama modulu
- İstifadəçi interfeysi

Veb axtarış robotlarının iş prinsipi internetdə mövcud olan bütün səhifə və veb saytları tək-tək gəzmək və orada olan məlumatları və linkləri yadda saxlayaraq indeksləşdirmə moduluna ötürməkdən ibarətdir. İnternetə bir dəqiqədə yüz minlərlə

məlumatlar daxil olunur. Bu səbəblə axtarış robotları dayanmadan bütün vebi gəzirlər və oradakı məlumatları yadda saxlayırlar. Təbii ki, axtarış robotları həmin prosesi yerinə yetirmək üçün müxtəlif kompleks alqoritmlərdən istifadə edirlər.

İndeksləşdirmə modulunun iş prinsipi axtarış robotları tərəfindən göndərilən məlumatları indeksləşdirməkdən ibarətdir. İndeksləşdirmə modulu xüsusi alqoritmlərlə indeksləmə prosesini həyata keçirir. Daha sonra indeksləşdirilmiş verilənlər informasiya-axtarış sisteminin verilənlər bazasına göndərilir. İstifadəçi tərəfindən sorğu verildiyi zaman isə verilənlər bazasından tələb olunan linkləri əldə etmək üçün əvvəlcədən bazaya yerləşdirilmiş indekslər bu modul vasitəsilə oxunur və lazım olan linklər toplanaraq sorğu və sıralama moduluna ötürülür.

Verilənlər bazasının iş prinsipi indeksləşdirilmiş verilənləri yadda saxlamaq və lazım olduqda indeksləşdirmə modulunu əvvəlcədən daxil edilmiş indekslərlə təmin etməkdən ibarətdir. Qeyd etmək lazımdır ki, verilənləri indeksləşdirilmiş formada yadda saxlamaqla həm verilənlər bazasında yer sərfiyyatını maksimum dərəcədə azaltmaq həm də bazaya müraciət zamanı indeksləşdirmə modulunun əvvəlcədən bazaya yerləşdirilmiş lazımı məlumatı tapmaq zamanını minimumlaşdırır.

Sorğu və sıralama modulunun iş prinsipi müvafiq olaraq aşağıdakı ardıcılıqdan ibarətdir:

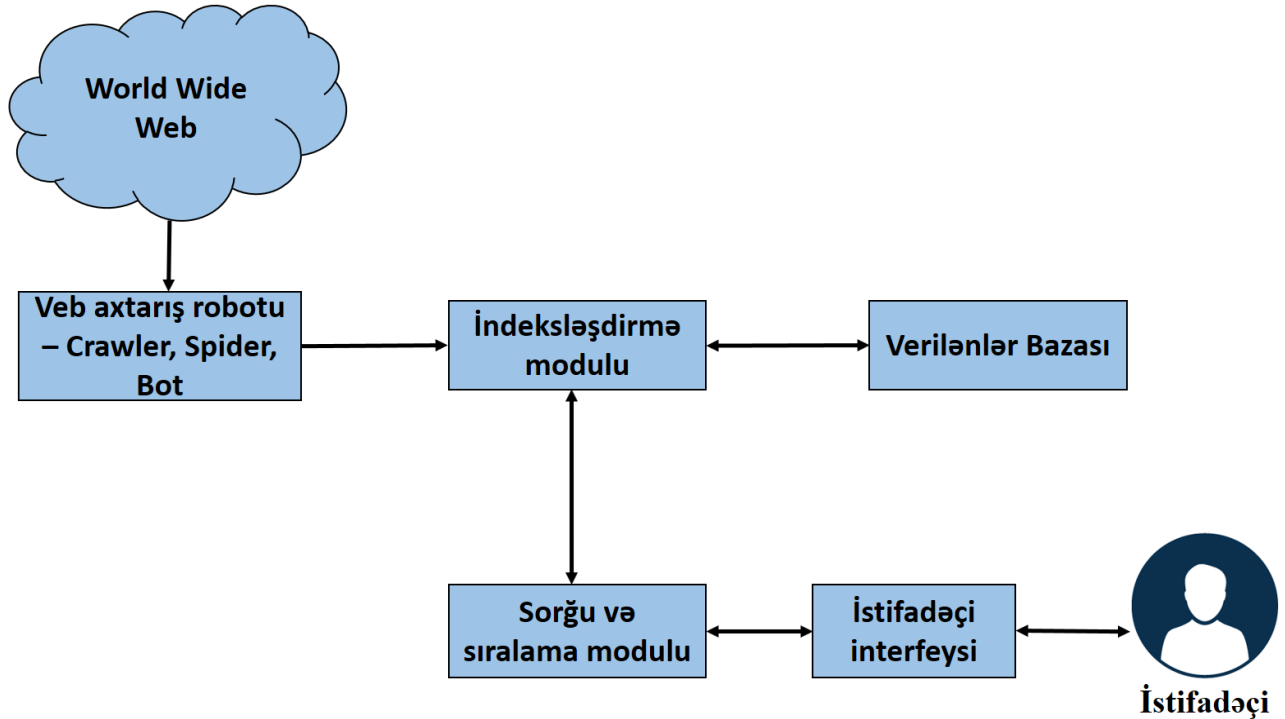
1. İstifadəçi tərəfindən verilən sorğuları qəbul etmək
2. Verilənlər bazasında axtarış etmək
3. Lazımi indeksləri tapmaq
4. Həmin indeksləri oxumaqla lazımi linkləri əldə etmək
5. Həmin linkləri müxtəlif alqoritmlərə uyğun olaraq sıralamaq
6. Nəticələri istifadəçi interfeysinə ötürmək

İstifadəçi interfeysinin məqsədi istifadəçi ilə informasiya-axtarış sistemi arasında əlaqə yaratmaqdan ibarətdir. Təbii ki, interfeys nə qədər rahat və sadə olarsa istifadəçilər tərəfindən informasiya-axtarış sisteminin istifadəsi bir o qədər rahat olar.

Hazırda istifadəçilər informasiya-axtarış sistemləri arasında seçim edərkən

əvvəlcə axtarışların nəticəsinə daha sonra isə axtarış sisteminin interfeysinə əsasən seçim edirlər.

İnformasiya-axtarış sistemlərinin arxitekturası Şəkil 3-də göstərilmişdir:



Şəkil 3. İnformasiya-axtarış sistemlərinin arxitekturası

2.2. Veb axtarış robotları

Veb axtarış robotları bəzən crawler bəzən də spider adlanır. Hansı ki, veb indeksləmə məqsədilə sistemli olaraq “World Wide Web”-i dayanmadan nəzərdən keçirir. Veb informasiya-axtarış sistemləri və bəzi digər saytlar veb məzmununu və ya başqalarının saytlarındakı veb məzmununun indekslərini yeniləmək üçün veb crawler və ya spider kimi süni intellektə əsaslanan proqramlardan istifadə edirlər. Veb axtarış robotları istifadəçilərin daha effektiv axtarışlar etməsi üçün yüklədikləri səhifələri skan edərək axtarış sisteminin indeksləşdirmə modulunda emal edirlər [38].

Axtarış robotları ziyarət edilən veb saytlarda və ya veb səhifələrdə yerləşən resursları istehlak edir və tez-tez icazə almadan saytlara daxil olurlar.

Bəzi mexanizmlər skan edilməsini istəməyən ictimai saytlar üçün axtarış robotlarına bildiriş göndərirlər. Məsələn, robots.txt faylı da daxil olmaqla bir saytın yalnız bəzi hissələrini indeksləşdirmək üçün botlara təklif göndərə bilər və ya botlardan heç bir şey istəməzlər.

İnternet səhifələrinin sayı olduqca çoxdur. Bu səbəblə hətta ən böyük crawlerlər belə əksiksiz bir indeks yaratmaqda zəif qalırlar. Bu səbəblə axtarış sistemləri 2000-ci ilə qədər yəni internetin ilk illərində uyğun axtarış nəticələri vermək üçün bir-birləri ilə rəqabətdə idilər. Hal-hazırda axtarılan mövzu ilə əlaqəli nəticələr demək olarki dərhal verilir. Crawlerlər hiperlinklər və HTML kodlarını dəstəkləyirlər. Onlar həmçinin veb-mining üçün də istifadə edirlər [38].

Web axtarış robotları baxış keçiriləcək linklərin bir siyahısı ilə işə başlayır. Veb axtarış robotu bu linklərə baxış keçirdikdən sonra həmin səhifələrdəki bütün linkləri skan etmə sərhəddi adı verilən yeni ziyarət ediləcək linklər siyahısına əlavə edir. Sonrakı linklərə bəzi siyasətlər qrupuna əsasən yenidən baxış keçirilir. Axtarış robotu saytların arxivləşdirilməsini həyata keçirirsə o bu zaman həmin məlumatları kopyalayır və olduğu kimi yadda saxlayır. Arxivlər ümumiyyətlə elə saxlanırlar ki, onlara veb-də olan kimi baxmaq, oxumaq və yerini dəyişmək lokal diskdə olduğu kimi asan olsun.

Arxiv baza kimi tanınır və veb səhifələrin toplanması və idarə olunması üçün nəzərdə tutulmuşdur. Baza yalnız HTML səhifələri özündə depolayır və bu səhifələr fərqli fayllar kimi saxlanılır. Baza müasir bir verilənlər bazası kimi verilənləri özündə saxlayan bir sistemə bənzəyir. Tək fərq ondadır ki, baza verilənlər bazası kimi hansısa verilənlər bazasının idarəetmə sisteminin köməyinə ehtiyac duymur. Baza axtarış robotu tərəfindən əldə edilən veb səhifələrin ən son versiyasını yaddaşda saxlayır.

Server proqram təminatı tərəfindən skan olunan mümkün URL ünvanların miqdarı eyni kontentin olmamasına çalışan axtarış robotunun işini çətinləşdirir.

HTTP GET (URL bazalı) parametrlərinin sonsuz birləşmələri mövcuddur ki, bunlardan yalnız kiçik bir hissəsi həqiqətən unikal məzmunu nəticə olaraq qaytaracaqdır. Məsələn, sadə bir onlayn foto qalereya URL ünvanında HTTP GET parametrləri ilə göstərildiyi kimi istifadəçilərə üç seçim təklif edə bilər. Şəkilləri

qruplaşdırmaq üçün dörd üsul varsa, iki fayl formatı bir istifadəçi tərəfindən təmin olunan kontent daha sonra isə eyni kontent 48 müxtəlif URL ünvanı tərəfindən həmin səhifəyə istinad verə bilər. Bu riyazi birləşmə axtarış robotu üçün problem yaradır, çünki lazım olan kontenti almaq üçün bu kontent nisbətən kiçik skriptləşdirilmiş sonsuz kombinasiyalarına görə qruplaşdırılmalıdır.

Axtarış robotları veb səhifələri skan edərkən müxtəlif siyasətlər əsasında baxış keçirirlər [38].

➤ *Veb Axtarış robotlarının axtarış siyasəti*

Axtarış robotunun işləmə prinsipi bəzi siyasətlər toplusunun nəticəsidir.

- Yüklənəcək səhifələri müəyyənləşdirən seçim siyasəti
- Səhifələrdəki dəyişikliklərin nə zaman yoxlanacağını müəyyənləşdirən yenidən ziyarət siyasəti
- Veb saytların həddindən artıq yüklənməsinin qarşısını almaq üçün olan nəzakət siyasəti
- Ayrı-ayrı veb axtarış robotlarının necə əlaqələndiriləcəklərini göstərən paralelləşdirmə siyasəti

➤ *Veb axtarış robotlarının seçim siyasəti*

Veb-in hal-hazırkı ölçüsü nəzərə alındıqda hətta ən böyük bazaya malik informasiya-axtarış sistemləri belə normal istifadəçilər üçün əlçatan olan hissənin yalnız müəyyən bir hissəsini təşkil edir. 2009-cu ildə aparılan araşdırmalar nəticəsində aydın olduki, böyük istifadəçi kütləsinə malik informasiya-axtarış sistemləri belə indeksləyə bilən veb-in sadəcə 40-70 faizini indeksləməyə müvəffəq olmuşdurlar.

Nəzərə almaq lazımdır ki, 1999-cu ildə Steve Lawrence və Lee Giles tərəfindən aparılan araşdırmalar nəticəsində təstiqləndi ki, həmin ildə heç bir informasiya-axtarış sistemi veb-in 16 faizdən çoxunu indeksləyə bilməmişdir. İstənilən axtarış robotu daima veb səhifələrin yalnız müəyyən bir bölümünü nəzərdən keçirdiyinə görə, həmin bu bölüm təsadüfə yox, ən uyğun səhifələrdən ibarət olması arzu olunandır. Bu

vəb səhifələr arasında öncəliyi təyin etmək üçün əhəmiyyətlik metrikasının mövcud olması məcburidir. Hər hansı bir səhifənin əhəmiyyəti onun daxili keyfiyyətindən, bağlantılarının və ya ziyarətlərin populyarlığından və hətta URL ünvanının funksiyalarından asılıdır. Mükəmməl seçim siyasəti yaratmaq əlavə çətinliklərə səbəb olur: Axtarış robotu veb üzərindəki informasiyaların müəyyən bir qismi ilə işləməlidir. Buna səbəb axtarış robotunun daima genişlənməkdə olan veb-dəki veb səhifələrin tamamını bilməməsidir.

Cho.J veb axtarış robotunun axtarış siyasəti üzərində ilk araşdırmalar aparan şəxs idi. Onların məlumat toplusu fərqli strategiyalarla skan edilmiş 180.000 səhifədən ibarət stanford.edu domeni idi. İstifadə edilmiş metrika ölçümləri “breadth-first” üsulu, backlink sayı və qismən PageRank hesablamaları idi. Nəticələrdən birində axtarış robotu yüksək Pagerank dəyərində sahib səhifələrini skan etmə prosesindən əvvəl endirmək istəyirsə onda PageRank strategiyası ən optimal variant hesab edilir, ardınca “breadth-first” üsulu və backlink sayı gəlir. Bununla belə bu nəticələr sadəcə bir ədəd domen üçündür.

Marc Najork and Janet Wiener ilk geniş miqyaslı axtarış prosesini həyata keçirərək 328 milyon səhifə üzərində “breadth-first” üsulunu tətbiq etmişdirlər. Skanetmənin başlanğıcında “breadth-first” üsulu yüksək Pagerank dəyərində malik olan səhifələri əhatə edən iri miqyaslı ilk axtarış prosesini həyata keçirdi. Bu nəticəyə əsasən müəlliflər tərəfindən təqdim edilən izah belə idi: “Bütün əsas səhifələrin müxtəlif istifadəçilər üçün çox sayda linkləri mövcuddur və onlar istifadəçinin axtarısa əsas səhifədən başlamasına baxmayaraq digərlərinə nisbətə tez tapıla bilərlər” [38].

Serge Abiteboul OPIC adlanan alqoritmə əsaslanan bir axtarış strategiyası hazırladı. OPIC (On-line Page Importance Computation) sözünün hərfi mənada tərcüməsi “Onlayn Səhifə Əhəmiyyətinin Hesablanması” deməkdir. OPIC alqoritmində hər bir səhifə ona verilən başlanğıc əl çatan səhifələr və ona yönələn bütün səhifələr arasında bərabər paylanır. O PageRank dəyəri ilə oxşardır lakin o daha sürətlidir və proses bir addımla yerinə yetilir. OPIC alqortimindən asılı axtarış robotu ilk növbədə yüksək əlçatanlıq dəyərində malik olan ilk səhifəni yükləyir. Təcrübələr

100.000 səhifəlik süni paylaşılmış linklərdən ibarət qraflar üzərində həyata keçirildi. Lakin bundan sonra başqa strategiyalarla müqayisə edilmədi və real veb-də təcrübələr həyata keçirilmədi.

Paolo Boldi 40 milyon səhifədən ibarət “.it” domenli vebin alt çoxluğundan və veb bazalı axtarış robotundan istifadə edərək 100 milyon səhifənin simulyasiyanı test etdi. Əvvəlcə “breadth-first” daha sora isə “depth-first”, “təsadüfi seçim” və “omniscient” strategiyalarını tətbiq etdi. Müqayisə səhifənin reytinginin nə qədər düzgün hesablanması, yəni axtarış robotu tərəfindən səhifənin təqribi hesablanmış reytinginin onun həqiqi reytinginə nə qədər yaxın olması prinsipi üzərində qurulmuşdur [23].

İzlənilən bağlantıları məhdudlaşdırılması

Axtarış robotu hər zaman HTML tipli səhifələrdə axtarış etmək istəyir və o internetdəki MİME tipli olan hər şeydən uzaq durmağa cəhd göstərir. MİME (Multipurpose İnternet Mail Extensions) tərcümədə “çoxməqsədli internet poçt əlavələri” sözlərinin qısaltmasıdır. MİME internet standartlarından biridir. E-poçt tətbiqləri ilə göndəriləcək poçtlara müxtəlif növ məzmunu əlavə etmək üçün istifadə olunan internet standartıdır. MİME sadə poçt mübadilə protokolunu yəni SMTP-ni (Simple Mail Transfer Protocol) həm mətn tipli həm də mətn tipli olmayan birdən çox məzmun əlavə edə biləcək şəkildə genişləndirir. Yəni MİME vasitəsi ilə e-poçtlara şəkil, səs və video tipli fayllar əlavə edilə bilirlər. E-poçt tətbiqlərinə əlavə olaraq axtarış robotları da müxtəlif MİME tiplərini dəstəkləyirlər. Bunun nəticəsində botlar səhifənin HTML tipli olub-olmamasını və ya səhifədə göstərə biləcəyi tipdə faylın olub-olmamasını anlayaraq nə etməli olduğuna qərar verə bilər. MİME standartı 1992-ci ildə İnternet Engineering Task Force (IETF) tərəfindən təyin edilmişdir və hazırda yeni versiyası olan S/MİME şifrələnmiş əlaqələri də dəstəkləməkdədir [38].

Axtarış robotu axtarış edən zaman sadəcə HTML resurslarını tələb edir, bütün veb resurslarını tələb etmədən öncə veb-dəki MİME tipli resursları ayırd etmək üçün HTTP HEAD tələbini irəli sürür. Çoxlu sayda HEAD tələbini önləmək üçün botlar

URL ünvanlarını yoxlayır və hər zaman sonu .html, .htm, .asp, .aspx, .php, .jsp, .jspx və ya bir ədəd “ / ” işarəsi ilə bitən URL ünvanlarını tələb edir. Bu strategiya bir çox HTML veb resurslarının istənmədən gözdə keçirilməsinə səbəb ola bilər. Bəzən URL ünvanlarının sonunda “?” işarəsi ola bilər. Bu zaman axtarış robotu həmin “?” işarəsini silməyə məcbur olur. Bu da öz növbəsində bəzi hallarda axtarış robotunun axtardığı ünvanı tapmasına mane ola bilər. Bu metod bəzi hallarda güvənilməzdir.

URL ünvanlarının normallaşdırılması

Ümumiyyətlə axtarış robotları bəzi veb səhifələrin bir dəfədən artıq skan edilməsinin qarşısını almaq məqsədi ilə URL ünvanlarının normallaşdırılması üsullarından geniş istifadə edirlər. URL ünvanlarının normallaşdırılması həmçinin URL kanonizasiyası da adlanır və o URL ünvanlarının əsaslı şəkildə yenidən dəyişdirilməsini və URL ünvanlarının ardıcıl olaraq standartlaşdırılması prosesini yerinə yetirir. URL kanonizasiyası informasiya-axtarış sistemləri tərəfindən eyni məzmunun iki fərqli link strukturu tərəfindən göstərilməməsi üçün veb axtarış robotunun hansı URL ünvanını istifadə edəcəyini müəyyənləşdirən üsuldur. Eyni məzmunun birdən çox URL ünvanında olması veb saytların ən arzu etmədikləri haldır. Bununla yanaşı, informasiya-axtarış sistemləri tərəfindən "dublikat" kimi qəbul edilir və kontentin sərbəst buraxılmasına səbəb olur. Təbii ki, eyni məzmunu malik iki ədəd URL ünvanı mövcuddur. Bu zaman həmin veb səhifələrin birinin sahibi öz URL ünvanını əsas kimi göstərə bilər. Bunun üçün o öz URL ünvanına kanonizasiya etiketləri əlavə etməlidir. Bundan sonra veb axtarış robotu həmin iki URL ünvanında olan məzmunları əldə etdiyi zaman hansında kanonizasiya etiketi mövcuddursa onu əsas URL ünvanı, digərini isə dublikat URL ünvanı kimi qəbul edəcəkdir. Sadəcə bir misal kimi kanonizasiya etiketi aşağıdakı formada verilə bilər:

```
<link rel="canonical" href="http://www.example.com"/>
```

URL ünvanlarının normallaşdırılması üsullarının aşağıdakı növləri vardır:

1. Hərfləri kiçiltmə. Məsələn:

HTTP://www.Example.com/ → http://www.example.com/

2. Sleş (/) işarəsi əlavə etmə. Məsələn:

`http://www.example.com → http://www.example.com/`

3. Əsas səhifə linkini yox etmə. Məsələn:

`http://www.example.com/default.asp → http://www.example.com/`

4. URL ünvanının bir hissəsini çıxarma. Məsələn:

`http://www.example.com/bar.html#pageheader → http://www.example.com/bar.htm`

5. Linkdəki nöqtələri aradan götürmə. Məsələn:

`http://www.example.com/./a/b/./c/./d.html → http://www.example.com/a/c/d.html`

6. “www” etiketini aradan götürmə. Məsələn:

`http://www.example.com/ → http://example.com/`

7. QueryString-ləri linkdən çıxarma. Məsələn:

`http://www.example.com/display?id=123&fakefoo=fakebar → http://www.example.com/display?id=123`

Fokuslanmış skanetmə

Bir veb səhifənin axtarış robotu üçün önəmi verilmiş sorğuya uyğun səhifə funksiyası kimi də istifadə oluna bilməsidir. Bir-birinə bənzəyən səhifələri yükləməyə çalışan botlara fokuslanmış axtarış robotları deyilir. Fokuslanmış skanetmə məfhumu Filippo Menczer və Soumen Chakrabarti tərəfindən irəli sürülmüşdür.

Fokuslanmış skanetmənin əsas problemi veb axtarış robotu mühitində səhifəni həqiqətən endirmədən əvvəl müəyyən bir səhifənin tərkibindəki mətnlərin bənzərliyini əvvəlcədən xəbər verməkdir. Mümkün öngörmə linklərin mətnləridir, bu internetin ilkin vaxtlarında yaradılan ilk veb boty olan Pinkerton-un yavaşması idi. Bu yavaşmada hələ ziyarət edilməmiş olan veb səhifələr arasında bənzərliyi təyin etmək üçün ziyarət edilmiş səhifələrin tam məzmununu istifadə etməyi təklif edirdilər. Fokuslanmış skanetmənin nailiyyəti əsasən əvvəlcədən axtarılan müəyyən xüsusi başlıq və mövzuların linklərinin zənginliyinə bağlıdır və fokuslanmış skanetmə ümumi olaraq başlanğıc nöqtələri təyin etmək üçün ümumi bir veb informasiya-axtarış sisteminə arxalanır [19].

Akademik fokuslanmış axtarış robotları

Fokuslanmış axtarış robotları nümunələrinə misal olaraq akademik axtarış robotlarını göstərmək olar. CiteSeerX informasiya-axtarış sisteminin botu olan citeSeerErxBot kimi pulsuz akademik sənədləri axtaran robotu qeyd etmək lazımdır.

Başqa akademik informasiya-axtarış sistemlərinə misal olaraq Google Scholar, Microsoft Academic Search və başqalarını göstərmək olar. Akademik sənədlər PDF fayl formatında saxlanıldıqları üçün akademik fokuslanmış axtarış robotu əsasən PDF skanlaması, PostScript faylları, Microsoft Word tipli faylların Zip formatda sıxışdırılmış formalarını skanlamaqla maraqlanırlar. Bu səbəblə də, Heritrix kimi ümumi açıq qaynaqlı axtarış robotları digər MIME növlərini süzgəcdən keçirmək üçün xüsusi hazırlanmalıdır və ya bu ara sənədləri çıxarmaq və onları hədəflənmiş skanetmə verilənlər bazası və məlumat bazasına idxal etmək üçün istifadə olunur. Bu sənədlərin akademik və ya akademik olmadığını müəyyənləşdirmək çətinidir və skanetmə prosesinə əhəmiyyətli bir yük əlavə edə bilər, beləliklə bu proses maşın təlimi və ya müntəzəm ifadə alqoritmlərini istifadə etməklə həyata keçirilir. Bu akademik sənədlər adətən fakültə və universitetlərin əsas səhifələrindən və ya tədqiqat institutlarının veb səhifələrindən əldə edilir. Akademik sənədlər bütün veb səhifələrindəki məzmunun kiçik bir hissəni tutduğundan yaxşı server seçimi bu veb axtarış robotlarının effektivliyini artırmada mühüm rol oynayır. Digər akademik axtarış robotları başlıqları, qəzet və abstraktlar kimi akademik sənədlərin meta-verilənlərini ehtiva edən düz mətn və HTML faylları yükləyə bilərlər. Bu sənədlərin ümumi sayını artırır, lakin əhəmiyyətli bir hissəni pulsuz PDF yükləmələri kimi təmin edilə bilmir.

➤ Veb axtarış robotlarının təkrar ziyarət siyasəti

Veb çox dinamik bir xarakterə malikdir və veb-in bir qismini skanetmək həftələr və ya aylarla çəkə bilər. Bir veb axtarış robotu skanetmə əməliyyatını tamamladığında, yeni faylların yaradılması, yenilənmələr və silinmələr də daxil olmaqla bir çox əməliyyat baş verə bilər.

Yenililik: Bu ikili ölçü yerli sürətin dəqiq olub-olmadığını göstərir. T zaman müddəti ərzində P səhifəsinin yeniliyi aşağıdakı kimi təyin edilir:

$$F_p(t) = \begin{cases} 1 & \text{əgər } p \text{ yerli nüsxəyə } t \text{ zaman müddətində bərabədirsə} \\ 0 & \text{əks halda} \end{cases}$$

Yaş: Bu ölçü yerli sürətin nə dərəcədə köhnə olduğunu göstərir. T zaman müddəti ərzində P səhifəsinin yaşı aşağıdakı kimi müəyyənləşdirilir:

$$A_p(t) = \begin{cases} 0 & \text{əgər } p \text{ } t \text{ zaman müddətində modifikasiya edilməyibsə} \\ t & \text{əks halda} \end{cases}$$

Burada t , p səhifəsinin modifikasiya zamanını ifadə edir.

Coffman yeniliyə bərabər bir veb axtarış robotunun hədəfini izah etməyə çalışdı, ancaq fərqli bir ifadə istifadə etmişdir: o, crawlerlərin səhifələrin köhnəlmə müddətlərini minimuma endirmələrini təklif etmişdir. Həmçinin veb skanetmə problemində veb brauzerin server olduğu və veb saytlar tərəfindən sifariş edilmiş bir sıra növbələrin olduğu bir server-sorgu sistemi kimi modelləşdirildiyini bildirir. Səhifə modifikasiyaları yəni, səhifə dəyişiklikləri istifadəçilərin gəlişidir və gediş müddətləri səhifəyə girişlə təkcə bir veb sayta baş çəkmə arasındakı müddətdir. Bu modeldə müştərilərin reyting sistemində ortalama gözləmə müddəti veb axtarış robot üçün orta yaşa bərabərdir.

Axtarış robotunun məqsədi bazasındakı səhifələrin ortalama yeniyini mümkün qədər yüksək saxlamaq və ya səhifələrin orta yaşını mümkün qədər aşağı saxlamaqdır. Bu məqsədlər ekvivalent deyildir: ilk halda, axtarış robotu neçə səhifənin köhnəlmiş səhifələrə mənsub olmağını müəyyən etməlidir, ikinci halda isə axtarış robotu səhifələrin yerli nüsxələrinin nə qədər köhnə olduğunu müəyyən etməlidir.

Cho və Garcia-Molina tərəfindən iki sadə təkrar ziyarət siyasəti öyrənilir:

- Eynicinsli siyasət: Bu dəyişiklik dərəcələrindən asılı olmayaraq toplanmaqda olan bütün səhifələri eyni tezliklə yenidən ziyarət etməyi nəzərdə tutur.

- Proporsional siyasət: Daha tez-tez dəyişən səhifələri yenidən ziyarət etməyi nəzərdə tutur. Ziyarət tezliyi (təxmini) dəyişiklik tezliyi ilə mütənasibdir.

Hər iki halda da səhifələrin təkrarlanan skanetmə qaydaları ya təsadüfi ya da sabit bir qaydada edilə bilər.

Cho və Garcia-Molina simulyasiya edilən veb-də və real veb-də skanlanmanın hər ikisində ortalama yenilik baxımından eynicinsli siyasətin proporsional siyasətdən daha üstün olduğunu təəccüblü şəkildə sübut etdilər. İntuitiv olaraq veb axtarış robotları müəyyən bir müddət ərzində neçə səhifəni skan edə biləcəkləri haqqında məhdudiyyətə malik olduqlarından daha az yenilənən səhifələrin hesabına sürətlə dəyişən səhifələr üçün çoxlu yeni skanlama aparacaq. Sürətlə dəyişən səhifələrin təzəliyi daha az dəyişən səhifələrdən daha qısa müddətə davam edir. Başqa sözlə, proporsional siyasət, tez-tez yenilənən səhifələrin skan edilməsi üçün daha çox resurs ayırır, lakin onlardan daha az yenilik müddətinə sahib olur.

Yeniliyi artırmaq üçün axtarış robotu çox tez-tez dəyişən elementləri məhdudlaşdırmalıdır. Optimal təkrar ziyarət siyasəti nə vahid siyasət nə də proporsional siyasətdir. Orta yeniliyi saxlamaq üçün ən uyğun metod çox tez-tez dəyişən səhifələri məhdudlaşdırmaq və orta yaşın aşağı səviyyədə saxlanması üçün ən uyğun olan hər bir səhifənin dəyişmə dərəcəsi ilə monoton olaraq artım göstərən giriş tezliklərini istifadə etməkdir. Hər iki halda optimal proporsional siyasətə nisbətən eynicinsli siyasət daha yaxındır. Coffman qeyd edir ki, gözlənilən köhnəlmə müddətini minimuma endirmək üçün hər hansı bir səhifəyə giriş mümkün qədər aralıq şəkildə saxlanmalıdır. Yenidən ziyarət siyasətinin açıq formullarına ümumiyyətlə çatmaq mümkün deyil, lakin onlar səhifə dəyişikliklərinin paylanmasıdan asılı olaraq ədədi olaraq əldə etmək mümkündür. Cho və Garcia-Molina eksponensial bölünmənin səhifə dəyişikliklərini təsvir etmək üçün uyğun olduğunu göstərmişlər, İpeirotis paylanmaya təsir göstərən parametrlərin aşkarlanması üçün statik alətləri qeyd edirdi. Burada ziyarətlər kefiyyətinə görə eyni cinsli kimi qəbul olunur. Amma bu reallıqdan uzaq olan bir ssenaridir və buna görə də veb səhifənin kefiyyəti haqqında əlavə məlumatın əldə olunması üçün daha mükəmməl skanlama siyasətinə ehtiyac vardır.

➤ *Veb axtarış robotlarının nəzakət siyasəti*

Axtarış robotları informasiya-axtarış sistemlərinə nəzərən məlumatları daha tez və dərin axtara bilər və bu da saytın işləməsinə mənfi təsir edir. Əgər bir axtarış robotu bir saniyədə bir neçə sorğu həyata keçirirsə və ya böyük həcmli fayllar yükləyirsə serverin işi çətinləşir.

Kosterin də qeyd etdiyi kimi, veb axtarış robotlarının müəyyən tapşırıqlar üçün istifadəsi faydalıdır lakin, ümumi birlik üçün qiyməti vardır. Veb axtarış robotlarının istifadəsinin xərclərinə daxildir:

- Veb axtarış robotları iş zamanı müəyyən buraxış imkanı və uzun müddət ərzində yüksək səviyyəli paralelləşdirmə ilə işlədikləri üçün onlar şəbəkə resurslarına tələbkardırlar;
- Əgər serverə müraciətlərin sayı çoxdursa server həddən artıq yüklənəcəkdir;
- Düzgün yazılmamış axtarış robotları serverləri və ya marşrutlaşdırıcıları sıradan çaxara bilərlər və ya öhdəsindən gələ bilməyəcəkləri səhifələri yükləyə bilərlər.
- Eyni zamanda fərdi axtarış robotları bir neçə istifadəçi tərəfindən istifadə olunarsa onlar şəbəkələrin və veb serverlərin işini poza bilər.

Bu problemlərin qismən həlli axtarış robotlarını istisna edən robots.txt kimi də tanınan protokoldur. Bu protokolda administratorlar veb serverlərin hansı hissələrinin axtarış robotları tərəfindən əlçatmaz olacağını qeyd edirlər. Bu standart özündə bir şəxs tərəfindən serverə eyni anda müraciətlərin miqdarını göstərən funksiyanı əhatə etmir.

Lakin bu funksiya serverin yüklənməsinin qarşısını ala biləcək ən effektiv üsuldur. Bu yaxınlarda Google, ASK, Jeeve, MSN və Yahoo informasiya-axtarış sistemləri robots.txt faylında əlavə bir parametr olan və istifadəçilərin sorğuları arasındakı vaxt fasiləsini təyin etməyə imkan verən “crawl-delay” komponentini əlavə ediblər.

İlk təklif olunan fasilə intervalı 60 saniyə olub. Əgər səhifələr bu interval ilə yüklənərlərsə və əgər səhifələrin sayı 100.000 olarsa onda bu tipli səhifələri yükləməyə sonsuz buraxış imkanı və iki ay vaxt lazım gələrdi.

Cho fasilə intervalı olaraq 10 saniyə istifadə edir və WIRE axtarış robotu isə susmaya görə 15 saniyə istifadə edir. MercatorWeb axtarış robotu adaptiv nəzakət siyasətindən istifadə edir və əgər verilən serverdən məlumatları oxumağa bir neçə saniyə tələb olunursa onda növbəti səhifənin yüklənməsinə tələb olunan fasilə intervalı olaraq o on saniyə seçir.

Veb axtarış robotlarını araşdırma məqsədi ilə istifadəsi üçün daha ətraflı analiz və resurs sərfi tələb olunur.

➤ *Veb axtarış robotlarının paralelləşdirmə siyasəti*

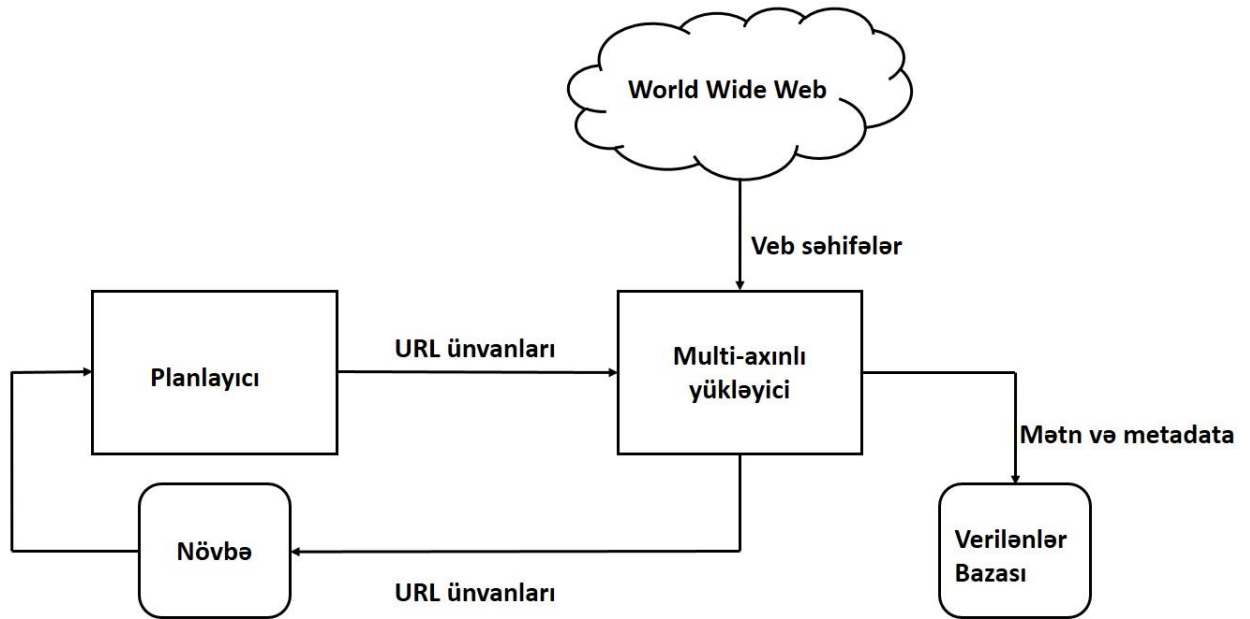
Paralel axtarış robotu paralel olaraq bir çox prosesləri idarə edən botlardır. Məqsəd paralelləşdirmədən yükü minimuma endirərkən və eyni səhifənin təkrar yüklənməsinin qarşısını almaq üçün yüklənmə dərəcəsini artırmaqdır. Eyni veb səhifəni birdən çox yüklənməsinin qarşısını almaq üçün, skanetmə sistemi eyni URL ünvanını iki müxtəlif skanlama proseduru ilə tapmaq mümkün olduğu üçün skanlama prosesi zamanı aşkar edilmiş yeni URL ünvanları təyin etməklə paralelləşdirmə siyasəti tələb edir.

Veb axtarış robotunun arxitekturası

Bir veb axtarış botu və ya robotu əvvəlki bölmələrdə də qeyd edildiyi kimi yalnız yaxşı skanetmə strategiyasına malik olmamalı, eyni zamanda yüksək optimallaşdırılmış arxitektura da sahib olmalıdır.

Shkapenyuk və Suel qeyd edir ki: Qısa bir müddət ərzində saniyədə bir neçə səhifəni yükləyə bilən axtarış botu yaratmaq olduqca asandır, lakin bir neçə həftə ərzində yüz milyonlarla səhifəni yükləyə biləcək yüksək göstəriciyə malik sistem yaratmaq sistem dizaynında giriş-çıxış şəbəkəsinin effektivliyi, etibarlılıq və idarə olunma baxımından bir sıra problemlər yaradır.

Veb axtarış robotlarının arxitekturası aşağıdakı Şəkil 4-də aydın formada verilmişdir:



Şəkil 4. Standart veb axtarış robotunun yüksək səviyyəli arxitekturası

Veb axtarış botları informasiya-axtarış sistemlərinin mərkəzi hissələrindən biridir və onların alqoritmləri və arxitekturası haqqında məlumatlar kommersiya sirləri kimi saxlanılır. Axtarış robotları dizaynlar edildikdə və hazırlandıqda, başqalarının onların oxşarını yaratmasının qarşısını almaq üçün çox vacib bir detal gizlədilir və ya haqqında səhf məlumat verilir. Böyük informasiya-axtarış sistemlərinin sıralama alqoritmlərinin işinə maneə törədən "axtarış sistemi spamları" ilə əlaqədar də narahatlıqlar mövcuddur.

Dərin Veb-də skanetmə

Veb saytların çox hissəsi dərin və ya görünməz şəbəkədə yerləşir. Belə saytlarda adətən yalnız verilənlər bazasına məlumat göndərmək olur və buna görə sadə istifadəçilər həmin saytları onlara link olmadan tapa bilməzlər. Google Sitemaps və “mod_oai” protokolları bu veb resursları aşkar olunması üçün nəzərdə tutulublar.

Veb saytların dərin skan olunması baxılmış veb linklərin sayını artırır. Bəzi botlar yalnız “< a href=“URL”>” formasında olan URL ünvanları qəbul edirlər.

Müəyyən hallarda Googlebot kimi botlar bütövlükdə veb səhifəni və ya hipermətnə olan mətni qəbul etməyə imkan verir.

Strateji yanaşma bütöv və dərin veb kontentə tətbiq oluna bilər. “Screen scraping” yəni? ekranın skripinqi adlanan texnologiya ilə proqram təminatı avtomatik və ya bir neçə dəfəlik veb formaya sorğu göndərməyə sazlan bilər. Bu tip proqram təminatı bir neçə veb səhifədə olan müvafiq miqdarda veb formanı əhatə etməyə imkan verir. Bir veb formadan alınan məlumatlar digər veb formaya tətbiq oluna bilər.

AJAX-da yaradılmış veb səhifələr veb axtarış botları üçün problem yaradan səhifələrdir. Google onların botunun tanıyıb indeksləyə biləcəyi AJAX çağırışları formatı təklif etmişdir.

Robots.txt fayllarının geniş miqyaslı təhlilinə əsaslanan yeni bir araşdırma müəyyən veb axtarış robotunun başqaları üzərində üstünlük təşkil etdiyini, Googlebotun ən çox seçilmiş veb axtarış botu olduğunun göstəricisidir.

Vizual və proqramlı axtarış botları

İnternetdə müəyyən vizual veb botları vardır ki, onlar səhifənin strukturundan yayınaraq istifadəçiyə lazım olan məlumatları sütun və sətir şəklində formalaşdırırlar.

Vizual və klassik botların əsas fərlərindən biri onların axtarış robotlarının sazlanması üçün lazım olan hissələrinin proqramlaşdırıla bilməsidir. Son nəsil axtarış robotlarından olan Diffbot, outwithub və İmport.io veb məlumatların skan olunması və tənzimənməsi üçün lazım olan funksiyaları silir.

Vizual yayınma metodu istifadəçi tərəfindən öyrədilmə və sonra yarım-strukturlaşdırılmış məlumatlar şablonuna əməl edərək məqsədə çatmaya əsaslanır.

Baxmayaraq ki, bu texnologiya yeni deyil və o Google tərəfindən alınmış Needlebase-in əsasını təşkil edir və hələ də investorlar tərəfindən bu sahədə davamlı böyümə və investisiyaların yatırımı davam edir.

Aşağıdakı Cədvəl 2 ümumi məqsədli axtarış robotları arxitekturalarının siyahısı (fokuslanmış veb axtarış robotları istisna olmaqla) və müxtəlif komponentlərə və qeyri-adi xüsusiyyətlərə verilən adların qısa təsvirini ehtiva edir:

Cədvəl 2. Ümumi məqsədli veb axtarış robotları

Bingbot	Microsoft-un Bing veb botunun adını daşıyır. O Msnbotu əvəz etdi.
FAST Crawler	FAST Crawler paylanmış axtarış botudur.
Googlebot	Google tərəfindən istifadə edilən, axtarış sistemləri üçün axtarışı mümkün edən indekslər yaratmağa xidmət edən, vebdən sənəd toplayan axtarış botudur
GM Crawl	SaaS rejimində istifadə edilə bilən yüksək yüksək səviyyəli axtarış robotudur.
PolyBot	"Skanlama meneceri", bir və ya daha çox "yükləyicilər"-ə sahib C ++ və Python-da yazılmış paylanmış botdur.
RBSE	İk yaradılan olunan veb axtarış robotu idi. O iki proqrama əsaslanırdı: Spider və Mite.
Swiftbot	Swifttype-in veb botudur və xüsusi bir axtarış sistemini yaratmaq üçün hazırlanmışdır, müəyyən veb saytların qrupunu indeksləşdirmək üçün nəzərdə tutulmuşdur.
WebCrawler	WebCrawler ilk ümumi açıq mənt indeksini yaratmaq üçün istifadə edildi. O "WWW"-ə əsaslanırdı. WebCrawler Google və Yahoo!-nın ən yaxşı axtarış nəticələrini qarışdıran bir axtarış robotu idi. WebCrawler istifadəçilərə şəkillər, audio, video, xəbərlər, sarı səhifələr və ağ səhifələrə baxmaq imkanı verir.
WebFountain	World Wide Web-də strukturlaşdırılmamış məlumatların öyrənilməsi üçün IBM tərəfindən yaradılmış internet analitik axtarış sistemidir.

WebRACE	Java-da tətbiq olunan və eRACE adlı daha ümumi sistemin bir hissəsi kimi istifadə edilən skanlama və keş moduludur
World Wide Web Worm	World Wide Web Worm sənəd başlığı və URL-lərinin sadə kataloq yaratmaq üçün istifadə olunan botdur.
Xenon	Fırıldaqcılığı aşkar etmək üçün dövlət vergi orqanları tərəfindən istifadə edilən bir veb axtarış robotudur.
Yahoo! Slurp	Yahoo!-nun Bingbotdan istifadə etmək üçün Microsoft ilə müqavilə imzalamasına qədərki dövrdə Yahoo -nun axtarış botunun adı idi.

Açıq mənbəli axtarış robotlarının siyahısı:

- Octopars
- Frontera
- Seeks
- Open Search Server
- Apache Nutch
- Sphinx (search engine)
- Xapian
- Scrapy
- StormCrawler
- GRUB
- ht://Dig
- news-please
- tkWWW Robot
- GNU Wget
- PHP-Crawler
- HTTrack
- Heritrix
- mnoGoSearch

2.3. İndeksləşdirmə modulu

Axtarış indeksi, sənədlər haqqında məlumatları ehtiva edən və informasiya-axtarış sistemlərində istifadə edilən məlumat strukturudur. İnformasiya-axtarış sistemi tərəfindən həyata keçirilən indeksləşdirmə, sürətli və dəqiq informasiya əldə etmək üçün məlumatların toplanması, təsnif edilməsi və saxlanması prosesidir. İndeksin formalaşması dilçilik, koqnitiv psixologiya, riyaziyyat, informatika və fizika konsepsiyalarını əhatə edir. Veb bazalı indeksləşdirmə, internetdə veb səhifələrin axtarışı üçün hazırlanmış informasiya-axtarış sistemləri kontekstində indeksləşdirmə prosesidir.

Populyar informasiya-axtarış sistemləri təbii dillərdə yazılmış sənədlərin tam mətn indekslənməsi üzərində işləyir, video, audio və qrafik kimi multimedia faylları da axtarışda iştirak edə bilirlər.

Meta-axtarış sistemləri digər axtarış xidmətlərinin indekslərindən istifadə edir və özündə indeksləri saxlamır. Bunun əksinə olaraq keş-bazalı axtarış sistemləri indeksləri korpusla birlikdə qalıcı olaraq yaddaşda saxlayırlar. Tam mətn indekslərindən fərqli olaraq, qismən mətn xidmətləri indeksin ölçüsünü azaltmaq üçün indeksin dərinliyini məhdudlaşdırır. İnkişaf etmiş informasiya-axtarış sistemləri tələb olunan vaxt və əməliyyat xərcləri səbəbindən əvvəlcədən müəyyən edilmiş bir vaxt aralığında indeksləşdirmə əməliyyatını həyata keçirirlər. Veb axtarış robotları və ya botlara əsaslanan informasiya-axtarış sistemləri isə indeksləri real vaxtda yaradırlar [39].

İndeksin istifadəsinin məqsədi axtarış sorğusu üçün müvafiq sənədlərin axtarılması sürətini artırmaqdır. İndeks olmadan informasiya-axtarış sistemi hər sənədi çox vaxt və çox emal gücü tələb edəcək şəkildə bazada axtarmaq məcburiyyətində qalacaq. Məsələn, 10,000 sənədin indeksləri milli-saniyələr içində sorğulana bilərkən, hər sözün 10,000 böyük sənəddə ardıcıl olaraq axtarılması saatlarla vaxt tələb edə bilər. İndeksin saxlanılmasına ayrılan əlavə yaddaş və indeks yeniləmək üçün lazım olan müddətin artırılması informasiya tapmaq üçün lazım olan vaxtın azaldılması ilə balanslaşdırılır [39].

İnformasiya-axtarış sistemini təkmilləşdirərkən aşağıdakı amillər nəzərə alınmalıdır:

1. Birləşmə faktorları

Məlumatlar mətn-korpus keçidində mətnin necə indeksə daxil edildiyini, sözlərin və ya mövzu xüsusiyyətlərinin necə indeksləşdirildiyini və birdən çox indekslərin avtosinxronizasiyasız fəaliyyət göstərib-göstərməməsini müəyyənləşdirməlidir. İndeksləşdirmə modulu əvvəlcə köhnə məzmunu yeniləməli olduğunumu yoxsa yeni məzmunun əlavə edilib-edilmədiyini yoxlamalıdır. Keçid adətən məlumatların toplanması siyasəti ilə bağlıdır. İnformasiya-axtarış sisteminin indekslərinin birləşdirilməsi SQL Merge əmrlərinə və digər birləşdirmə alqoritmlərinə bənzərdir.

2. Saxlama metodları

İndeksləşdirilmiş məlumatları necə yaddaşda saxlamaq lazımdır? Yəni saxlanılan məlumatın növünü sıxışdırılmış yoxsa filtirlənmiş olmasını müəyyənləşdirmək lazımdır.

3. İndeksin ölçüsü

İndeksi saxlamaq üçün nə qədər kompüter yaddaşına ehtiyac olduğunu müəyyənləşdirmək lazımdır.

4. Axtarış sürəti

Tərs çevrilmiş indeksdə sözü nə qədər tez tapa biləcəyinizlə xarakterizə olunur. Verilənlər sturukturunda girişi tapma sürəti onun yenilənmə və ya silinmə sürəti ilə müqayisə olunması kompüter elminin əsasını təşkil edir.

5. Saxlama

İndeksin uzun müddət necə saxlanılmalı olduğu və harada saxlanılmalı olduğu müəyyənləşdirilməlidir.

6. Xətaya davamlılıq

İnformasiya-axtarış sisteminin etibarlı olması vacib məsələlərdən biridir. Problemlər özündə zədələnmiş indekslərin emalını, korlanmış məlumatların izolyasiyalı emal olunmasının mümkünlüyünü, keş və birləşmiş bölünmə kimi

korlanmış aparat vasitələrinin bölmələrə və sxemlərə ayrılmasını saxlayır. Çoxaltma bir obyektin birdən çox nüsxəsinin məzmununu (məsələn, verilənlər bazasının məzmunu) sinxronizasiya etmək üçün istifadə olunan mexanizmdir. Çoxaltma, bir mənbədən digərinə kopyalama anlamına gələn bir prosesdir.

Verilənlərin indeks strukturları

İnformasiya-axtarış sisteminin arxitekturası indeksləşdirmə metodları və indekslərin saxlanması üsulları ilə fərqlənir. İndeksləşdirmə metodları aşağıdakı növlərdən ibarətdir:

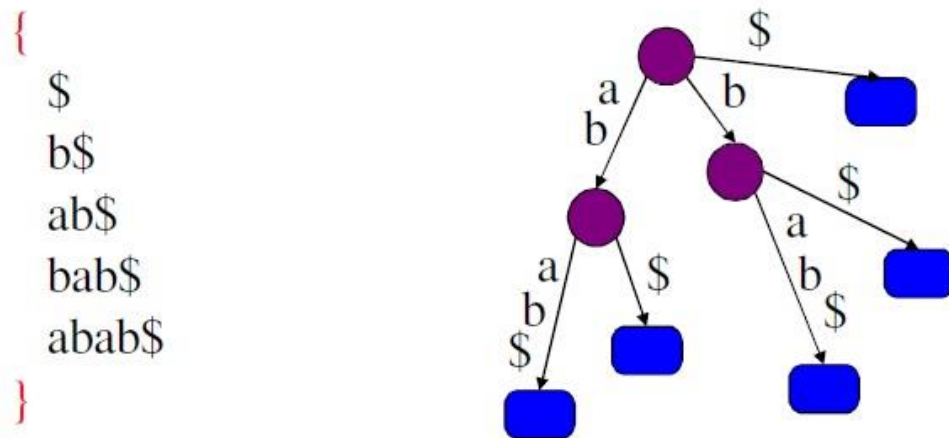
➤ Suffiks ağacı

Suffiks ağacı struktur olaraq ağac formasında qurulmuşdur və rəqəmsal axtarış müddətini dəstəkləyir. Sözüün şəkilçilərinin saxlanmasına əsaslanır. Ağaclar, informasiya-axtarış sisteminin indeksləşdirilməsi üçün vacib olan genişləndirilə bilən xəşləməni dəstəkləyir. DNT ardıcılığı və klasterləşdirilməsində nümunələri axtarmaq üçün istifadə olunur. Əsas mənfi cəhətdir odur ki, sözü ağacın içində saxlamaq, sözü saxlamaq üçün lazım olandan əlavə bir yer tələb edir. Daha az virtual yaddaş tələb olunur və məlumatlar blok-sıralama sıxışdırılmasını dəstəkləyir. Suffiks ağacı söz emal etmə alqoritmlərindən biridir. DNT fayllarının həcmi qıqabaytlarla ölçüldüyünə görə, əl ilə DNT analizini etmək mümkün deyil. Əslində DNT fayllarının kompüterin vasitəsilə emalı çox vaxt tələb edir. Suffiks ağacı bioloji məlumatlar, axtarış sistemləri, kompilyatorlar, əməliyyat sistemləri, verilənlər bazası və s. üçün istifadə olunur [18].

Suffiks ağacının mahiyyəti:

- Verilən mətn m uzunluğunda olan string (S)
- S üçün istifadə olunan zaman $O(m)$
- Tapılmalı olan string Q , onun uzunluğu n olsun
- Q -nün S -in içərisində axtarılması üçün lazım olan zaman $O(n)$

Misal üçün, Şəkil 5 –də $S=abab$ stringinə baxaq. S stringinin suffiks ağacı S -in bütün suffikslərini bir ağacda ehtiva etməsidir. \$ simvolu baxılan suffiksin sonunu işarə edir.



Şəkil 5. Suffiks ağacının strukturu

➤ *Tərs çevrilmiş indeks metodu*

Tərs çevrilmiş indeks hər bir axtarış meyarının adətən bir xəş cədvəl və ya ikili ağac şəklində olan siyahısını özündə saxlayır. Kompüter elmində tərs çevrilmiş indeks verilənlər bazasında və ya sənəddə söz və ya nömrələr kimi məzmunundan eyniləşməni özündə saxlayan kataloq məlumat strukturudur. Tərs çevrilmiş indeksin məqsədi, verilənlər bazasına yeni bir sənəd əlavə edildikdə, artan əməliyyat xərcləri ilə sürətli tam mətn axtarışlarını təmin etməkdir. Tərs çevrilmiş fayl indeksdən daha çox verilənlər bazası faylı ola bilər. İnformasiya-axtarış sistemlərində istifadə edilən ən məşhur məlumat strukturudur və axtarış sistemlərində geniş miqyasda istifadə edilir.

Bundan əlavə olaraq, bəzi ümumi məqsədli meynfreym bazasına əsaslanan verilənlər bazası idarəetmə sistemləri (VBİS) - ADABAS, DATACOM / DB və Model 204 daxil olmaqla tərs çevrilmiş siyahı arxitekturalarından istifadə etmişdir.

Tərs çevrilmiş indeksin iki əsas növü var:

- Qeydiyyat səviyyəli tərs çevrilmiş indeks - hər bir söz üçün sadəcə bir istinadlar siyahısını ehtiva edən kataloq.
- Söz səviyyəli tərs çevrilmiş indeks - əlavə olaraq hər bir sözün sənəd içərisində mövqeyini ehtiva edən kataloq.

Sonuncu forma daha çox funksionallıq təklif edir (söz axtarışları kimi), lakin daha çox emal gücünə və yaradılacaq sahəyə ehtiyac duyur.

Tərs çevrilmiş indeksə nümunə olaraq aşağıdakı misala baxaq:

Fərz edək ki, üç ədəd sənəd vardır və bu sənədlərin hər biri aşağıdakı cümlələrdən təşkil edilmişdir.

- ❖ Sənəd1: İnformasiya sistemlərinin tarixi.
- ❖ Sənəd2: İnformasiya-axtarış sistemlərinin arxitekturası.
- ❖ Sənəd3: İnformasiya kommunikasiya texnologiyaları.

Tərs çevrilmiş indeksləmənin tətbiqindən sonra:

informasiya: (1,0) , (2,0) , (3,0)

sistemlərinin: (1,12) , (2,20)

tarixi: (1,26)

axtarış: (2,12)

arxitekturası: (2,34)

kommunikasiya: (3,12)

texnologiyaları: (3,26)

Qeyd etmək lazımdır ki, mötərizə içərisində verilmiş ilk rəqəm axtarılan sözün hansı sənəddə yerləşdiyini, ikinci rəqəm isə axtarılan sözün həmin sənəddəki cümlənin harasında işləndiyini müəyyənləşdirir. Əgər informasiya-axtarış sisteminə “tarixi” sözünün sorğusu verilərsə, onda (2,26) indeksinə əsasən həmin sözün 2-ci sənəddə 26-cı yerdə olduğunu müəyyənləşdirərək axtarışın nəticəsini təqdim edəcək. Bu isə informasiya-axtarış sistemini sənədləri tək-tək skan etmək və buna uyğun olaraq əlavə vaxt israfına yol verməkdən xilas edir.

➤ *Elmi məqalələrə istinad indeksi*

Elmi məqalələrə istinad indeksi bibliografik göstəricinin bir növüdür. Bu nəşrlər arasındakı istinad indeksidir və istifadəçilərə hansı sənədlərin son və əvvəlinci olduğunu müəyyənləşdirməyə kömək edir.

1960-cı ildə Yucin Qarfild ilk akademik jurnal “Science Citation” jurnalı üçün olan istinad indeksini yaratdı. Daha sonra Social Sciences Citation Index və sonra Arts and Humanities Citation Index yaradıldı. İlk avtomatik istinad indeksləşdirilməsi 1997-ci ildə CiteSeer tərəfindən aparılmışdır. Bu tip məlumatların digər mənbələri kimi Google Scholar və Elseviers Scopus – u götürmək olar [39].

Ümumi qəbul olunmuş akademik istinadların indeksləri özündə aşağıdakıları saxlayır:

- “Clarivate Analytics” tərəfindən təklif olunan “Web of Science”.
- “Elsevier” tərəfindən təklif olunan “Scopus” yalnız internet üzərindən əlçatandır. Burada da analogi olaraq elmlərdə və sosial elmlərdə izləmə üsulu ilə fənləri birləşdirir.

Onlardan hər biri hansı sənədin hansı sənədə yönəldiyini təyin edən mexanizm və nəşrlər arasında əlaqəni təmin edən indeksləri təklif edirlər. Onlar publik sayılımlar və bir birindən fərqlənirlər. Web of Science və Scopus-a abunə olmaqla onlardan istifadə etmək olar.

Əlavə olaraq CiteSeer və Google Scholar internet üzərindən açıq istifadədədir. Eyni zamanda bir neçə digər açıq istifadədə olan indeksləmə xidmətləri vardır: INSPIRE-HEP yüksək enerji fizikasını əhatə edir, PubMed bütün təbiət və bio-tibbi mövzuları və fizikanı əhatə edir.

➤ *N-qram*

N-qram müəyyən bir ardıcılıqla təkrarlanma dərəcəsini müəyyənləşdirmək üçün istifadə edilən üsullardan biridir. Adı “*n*” və “*qram*” sözlərinin birləşməsindən ibarətdir. Burada *n*, təkrarın yoxlanıldığı dəyərdir, *qram* isə bu təkrarın ardıcılıq içərisindəki çəkisini ifadə etmək üçün istifadə edilmişdir. Yəni, baxılan sözlərin neçə-

neçə bölünəcəyini n təyin edir, q ram isə bir sözdən nə qədər olduğunu müəyyənləşdirməyə yarayır.

Əgər $n=1$ olarsa, onda 1-qram və ya uniqram adlanır,

Əgər $n=2$ olarsa, onda 2-qram və ya biqram adlanır,

Əgər $n=3$ olarsa, onda 3-qram və ya triqram adlanır.

Üçdən sonrakı bütün bölünmələrdə n -qram ifadəsi istifadə edilir. N-qram riyaziyyat, biologiya, kartoqrafiya, informasiya-axtarış sistemlərinin indeksləşdirilməsində, DNT hesablamalarında və genetik ardıcılığın tapılması sahələrində istifadə edilir.

N-qram hərflərə və ya sözlərə tətbiq edilə bilər. Hərflərə tətbiq olunan n-qrama misal olaraq “kommunikasiya” sözünü nəzərdən keçirək:

$n=1$, uniqram:

<i>k</i>	2	<i>k</i>	2
<i>o</i>	1	<i>a</i>	2
<i>m</i>	2	<i>s</i>	1
<i>m</i>	2	<i>i</i>	2
<i>u</i>	1	<i>y</i>	1
<i>n</i>	1	<i>a</i>	2
<i>i</i>	2		

Cədvəl 3

$n=2$, biqram:

<i>ko</i>	1	<i>ka</i>	1
<i>om</i>	1	<i>as</i>	1
<i>mm</i>	1	<i>si</i>	1
<i>mu</i>	1	<i>iy</i>	1
<i>un</i>	1	<i>ya</i>	1
<i>ni</i>	1		
<i>ik</i>	1		

Cədvəl 4

$n=3$, triqram:

<i>kom</i>	1	<i>kas</i>	1
<i>omm</i>	1	<i>asi</i>	1
<i>mmu</i>	1	<i>siy</i>	1
<i>mun</i>	1	<i>iya</i>	1
<i>uni</i>	1		
<i>nik</i>	1		
<i>ika</i>	1		

Cədvəl 5

Sözlərə tətbiq olunan n-qrama misal olaraq “informasiya-axtarış sistemlərinin əsas komponentləri” cümləsini nəzərdən keçirək:

n=1, unigram:

informasiya	1
axtarış	1
sistemlərinin	1
əsas	1
komponentləri	1

Cədvəl 6

n=2, bigram:

informasiya-axtarış	1
axtarış	1
sistemlərinin	1
sistemlərinin əsas	1
əsas komponentləri	1

Cədvəl 7

n=3, triqram:

informasiya-axtarış	1
sistemlərinin	1
axtarış	1
sistemlərinin əsas	1
sistemlərinin əsas komponentləri	1

Cədvəl 8

N-qram əsasən axtarış alqoritmlərində istifadə olunur. Axtarılan sözlə mövcud məlumatların müqayisəsi zamanı dəqiq nəticə əldə etmək üçün istifadə edilir. Çünki n-qram bizə təkrarlanan sözləri göstərir. Bunun hesabına təkrarlanma faizi hesablanı bilər. Nəticədə bu sayı uyğun olaraq axtarılan sözlə mövcud məlumatlar arasında oxşarlıq sayı hesablanı bilər və bu oxşarlıq sayına görə istifadəçiyə məlumatlar göstərilə bilər.

➤ Sənəd-termin matrisi

Sənəd-termin matrisi və ya termin-sənəd matrisi sənədlər toplusunda istifadə olunan sözlərlərin tezliyini müəyyən edən riyazi matrisdir. Sənəd-termin matrisindəki sətirlər bazada olan sənədləri, sütunlar isə terminləri ifadə edir.

Sənəd-termin matrisinə misal olaraq aşağıdakı nümunəyə baxaq:

İki ədəd qısa mətnli sənəd olduğu fərz edək:

- ❖ Sənəd1: Google ən çox istifadə edilən axtarış sistemidir.
- ❖ Sənəd2: Axtarış sistemləri içərisində ən məşhuru google axtarış sistemidir.

Yuxarıdakı mətnlər üçün sənəd-termin matrisi aşağıdakı kimi olacaq:

Termin	Sənəd 1	Sənəd 2
google	1	1
ən	1	1
çox	1	0
istifadə	1	0
edilən	1	0
axtarış	1	1
sistemidir	1	1
sistemləri	0	1
içərisində	0	1
məşhuru	0	1

Cədvəl 9

Paralel indeksləmə problemi

İnformasiya-axtarış sistemlərinin indeksləşdirmə modulunun əsas vəzifələrindən biri, ardıcıl hesablama proseslərinin idarə olunmasıdır. Burada bir çox əlaqəli uğursuzluqlar yarada biləcək vəziyyətlər mövcud ola bilər. Məsələn, bazaya yeni bir sənəd əlavə edilir və indeks yenilənməlidir, eyni zamanda indeks axtarış sorgularına cavab verməlidir. Bu, iki rəqib vəzifə arasında toqquşmaya səbəb olur.

Müəlliflərin informasiya istehsalçıları olduqlarını və veb axtarış robotlarının isə bu informasiyanın istehlakçı olduğunu, informasiyaları əldə edərək onu keş yaddaş (və ya korpus) halında saxlayacağını düşünün. Birbaşa indeks korpus tərəfindən yaradılan informasiyaların istehlakçısıdır və tərs çevrilmiş indeks birbaşa indekslə hazırlanmış informasiyaların istehlakçısıdır. Bu adətən istehsalçı-istehlakçı modeli kimi adlandırılır. İndeksləşdirmə modulu axtarıla bilən informasiyaların istehsalçısıdır və istifadəçilər axtarışa ehtiyacı olan istehlakçılardır. Nizamsız yaddaşda saxlama və paylanmış emal zamanı çətinlik dərəcəsi artır. Daha böyük həcmli indekslənmiş məlumatlarla işləmək üçün, informasiya-axtarış sistemi paylanmış hesablama

arxitekturasına ehtiyac duya bilər və bu da öz növbəsində informasiya-axtarış sistemi ilə birlikdə işləyən bir neçə qurğudan ibarət olmasına səbəb olacaqdır. Bu, məntiqsizlik ehtimalını artırır və tam sinxronlaşdırılmış, paylanmış, paralel bir arxitekturanı dəstəkləməyi çətinləşdirir.

➤ *Birbaşa indeks*

Birbaşa indeks hər sənəd üçün sözlərin siyahısını özündə saxlayır. Birbaşa indeksin sadələşdirilmiş formasına aşağıdakı kimi nümunə göstərmək olar:

Sənəd	Söz
Sənəd 1	Google, böyük, indeks, bazasına, malik axtarış sistemidir
Sənəd 2	Yandex, ən, çox, rusiyada, istifadə, olunur
Sənəd 3	Alqoritm, fərqi, axtarış, nəticələrinə, təsir, edir

Cədvəl 10

Birbaşa indeksin yaradılması zərurəti sözlərin gələcəkdə analiz olunaraq axtarış indeksinin yaradılması və bir dəfəlik sənəddə yadda saxlanması ilə izah olunur. Birbaşa indeksin formalaşdırılması özündə qismən dar tərs çevrilmiş indeksinin yenilənməsindən yayınan asinxron sistemli emal prosesini saxlayır. Birbaşa indeks tərs çevrilmiş indeksə çevirmək üçün sıralanır. Birbaşa indeks əsasən sənədlər əsasında sıralanmış sözlərdən və sənədlərdən ibarət cütlərinin siyahısıdır. Birbaşa indeksin tərs çevrilmiş indeksə çevrilməsi sadəcə olaraq sənədləri sözlərlə ayırmaqdan ibarətdir. Bu baxımdan tərs çevrilmiş indeks birbaşa indeksin sözləri ilə sıralanır.

➤ *Tərs çevrilmiş indeks*

Bir çox informasiya-axtarış sistemləri sorğudan sözlərdən ibarət olan sənədləri tez bir şəkildə tapmaq üçün axtarış sorğusunu qiymətləndirərkən tərs çevrilmiş indekstdən istifadə edir və sonra bu sənədləri aidiyyətinə uyğun olaraq sıralayır. Tərs

çevirilmiş indeks hər bir sözün olduğu sənədlərin siyahısını saxladığı üçün infomasiya-axtarış sistemi sorğudakı hər sözlə əlaqəli sənədləri tapmaq üçün istifadə edə və onları tez bir zamanda tapa bilər. Aşağıda tərs çevrilmiş indeksin sadələşdirilmiş nümunəsini göstərək:

Söz	Sənəd
google	Sənəd 1
indeks	Sənəd 1
axtarış	Sənəd 1, Sənəd 3
rusiyada	Sənəd 2
təsir	Sənəd 3
yandex	Sənəd 2

Cədvəl 11

Bu indeks yalnız bir sözün müəyyən bir sənəddə olub olmadığını müəyyən edə bilər, çünki sözün tezliyi və mövqeyi ilə bağlı heç bir məlumat saxlamır və bu səbəblə boolean indeksi hesab olunur. Tərs çevrilmiş indeks hansı sənədlərin axtarış tələbinə uyğun gəldiyini müəyyən edə bilər, lakin müvafiq sənədləri emal edə bilməz. Bəzi hallarda, tərs çevrilmiş indeks hər bir sənəddə hər bir sözün istifadə tezliyi və ya sənəddəki sözün mövqeyi kimi əlavə məlumatları ehtiva edə bilər. Mövqe məlumatları, axtarış alqoritminin ifadələr üçün axtarışın dəstəklənməsi üçün söz yaxınlığını müəyyən etməyə imkan verir. İstifadə tezliyi sorğuya kömək etmək üçün sənədləri sıralamaqda istifadə edilir.

İndekslərin birləşdirilməsi

Tərs çevrilmiş indeks birləşmə və ya yenidən qurma yolu ilə doldurulur. Yenidən qurma birləşməyə bənzəyir, ancaq əvvəlcə tərs çevrilən indeksin məzmununu silir. Bu arxitektura, əlavə olunacaq və ya yenilənəcək sənəd və ya sənədləri müəyyənləşdirir və sonra hər bir sənədi sözlərə ayıraraq artan indeksləməni

dəstəkləmək üçün nəzərdə tutula bilər. Texniki dəqiqlik üçün, birləşdirmə, bir və ya daha çox kompüterin sərt diskində yerləşən indeks keşi ilə adətən virtual yaddaşda yerləşən yeni indeksləşmiş sənədləri birləşdirir.

Təhlil əməliyyatı bitdikdən sonra indeksləşdirmə modulu, müəyyənləşdirilən sənədi uyğun olan sözlər üçün sənəd siyahısına əlavə edir. Daha böyük bir informasiya-axtarış sistemində, tərs çevrilmiş indeks üçün hər sözü tapmaq əməliyyatı çox vaxt tələb edə bilər. Bu səbəblə həmin proses adətən iki hissəyə ayrılır :

- Birbaşa indeksin inkişafı prosesi
- Birbaşa indeksin məzmunun tərs çevrilmiş indeksə çevirmə prosesi

Sənədlərin sintaksis analizi

Sənədlərin sintaksis analizi, sənədin hissələrini (sözləri) və ya digər daşıyıcı formanı birbaşa və tərs çevrilmiş indeksləmə üçün ayırır. Tapılmış sözlər token adlanır və informasiya-axtarış sisteminin indeksləşdirilməsi kontekstində və təbii dilin sintaksis analizində tokenləşdirmə olaraq adlandırılır. Sintaksis analizi bəzən qismən formatlaşdırma, morfoloji analiz, mətnin seqmentləşdirilməsi, kontentin analizi, mətnlərin analizi, mətnlərin intellektual analizi və ya nitqin seqmentləşdirilməsi kimi adlandırılır. "İndekslemə", "ayırma" və "tokentəşdirmə" terminləri korporativ dildə bir-birlərini əvəz edirlər.

Təbii dilin emalı daim araşdırılır və təkmilləşdirilir. Tokenləşdirmə, indeksləşdirmə problemlərində sənədlərdən lazımi məlumatları yüksək keyfiyyətli axtarışa dəstək olmaq üçün şərait yaradır. İndeksləşdirmə üçün tokenləşdirmə çoxsaylı texnologiyaların tətbiqi ilə məşğul olur və bunların tətbiqi çox vaxt korporativ sirr kimi qorunur.

Təbii dilin emalı prosesindəki problemlər:

1. Söz sərhədlərinin qeyri-müəyyənliyi

İlk baxışda tokenləşdirmə sadə bir məsələ kimi görünə bilər, ancaq xüsusilə çox dilli indeksləşdirmə modulu inkişaf etdirilərkən müəyyən problemlər baş verə bilər.

Rəqəmsal formada, Çin, Yapon və ya Ərəb dili kimi bəzi dillərdə yazılmış mətnlər qarışıq formaya malikdirlər, çünki sözlər bir boşluqla aydın şəkildə ayrılır.

Tokenləşdirmənin məqsədi istifadəçilər üçün axtarış edəcəkləri sözləri tanımaqdır. Hər bir dəstəklənən dilə müəyyənləşdirici hazırlamaq üçün söz sərhədlərini düzgün formada təyin etməyə kömək edən xüsusi-dil məntiqindən istifadə olunur.

2. Dilin qeyri-müəyyənliyi

Sənədlərin daha dəqiq formada sıralanması üçün bir çox informasiya-axtarış sistemləri, dil və ya istinad etdiyi nitqin bir hissəsi barədə əlavə məlumatı nəzərə ala bilər. Bu üsullar dildən asılıdır, çünki dillər arasında sintaksis fərqlidir. Tokenləşdirmə zamanı bəzi informasiya-axtarış sistemləri sənədin dilini avtomatik olaraq müəyyən etməyə çalışırlar

3. Müxtəlif fayl formatları

Hansı baytların sənədin simvollarını təmsil etdiyini düzgün şəkildə müəyyən etmək üçün fayl formatı düzgün təyin edilməlidir. Fərqli fayl formatlarını dəstəkləyən informasiya-axtarış sistemləri sənədləri düzgün formada açmalı, sənədə daxil olmalı və simvolları müəyyənləşdirməlidir.

4. Yaddaş səhvləri

Təbii dildəki məlumatlarının keyfiyyəti hər zaman mükəmməl ola bilməz. Bu zəiflik müvafiq fayl protokoluna əməl etməyən, xüsusilə internetdə bilinməyən bir sıra sənədlərdən qaynaqlanır. İkili simvollar səhvən sənədin müxtəlif hissələrində xətalı formada kodlana bilər. Bu simvolların tanınmaması və müvafiq əməliyyatların icra edilə bilməməsi, indeksin keyfiyyətini və ya indeksləşdirmə modulunun performansını aşağı salır.

Tokenləşdirmə

İnsanlardan fərqli olaraq, kompüterlər təbii dildə olan sənədlərin tərkibini anlamırlar, sözləri və cümlələri avtomatik olaraq tanımırlar. Bir kompüter üçün sənədin tərkibi yalnız baytlardan ibarət ardıcılıqdır. Kompüterlər boşluq simvolunun sənəddə sözləri ayırdığını bilmirlər. Bunun əvəzinə insanlar kompüter token adlandırılan və söz və ya sözləri nəyin təyin etdiyini müəyyən edəcək formada proqramlaşdırmalıdırlar. Belə proqramlar adətən “tokenizer” (tokenləşdirici) və ya “parser” (təyinedici) adlanır. Bir çox informasiya-axtarış sistemləri və digər təbii dil emalı proqramları, Lex və YACC kimi təyinetmə üçün əlverişli xüsusi proqramları dəstəkləyirlər.

Tokenləşdirmə zamanı təyinedici, sözləri və nöqtə kimi digər elementləri əks etdirən simvol sıralarını, bəziləri isə görünməyən simvolları müəyyən edir. Təyinedici, e-poçt ünvanları, telefon nömrələri və URL ünvanları kimi müəyyən obyektləri tanıya bilir. Hər bir token identifikasiya edildiyi zaman, tokenin vəziyyəti (yuxarı, aşağı, qarışıq, düzgün), dil və ya kodlaşdırma, leksik kateqoriya ('isim' və ya 'fel' kimi), cümlənin nömrəsi, cümlənin içindəki yeri, uzunluq və sətir nömrəsi kimi bəzi xüsusiyyətlər yaddaşda saxlanıla bilər.

Dilin tanınması

Informasiya-axtarış sistemi birdən çox dili dəstəkləyirsə, tokenləşdirmə zamanı ilk addım hər sənədin dilini müəyyən etməkdir. Sonrakı addımların bir çoxu dildən asılıdır. Dilin tanınması prosesi, bir proqramın sənədin dilini avtomatik olaraq təyin etmək və ya təsnif etmək üçün cəhd göstərdiyi prosesdir. Avtomatlaşdırılmış dilin tanınması təbii dilin emalında davam edən araşdırma mövzudur. Sözlərin hansı dilə məxsus olduğunu müəyyənləşdirmək hansısa dil tanıma cədvəlinin istifadəsini tələb edə bilər.

Sənədlərin format analizi

İnformasiya-axtarış sistemi birdən çox sənəd formatlarını dəstəkləyirsə, sənədlər tokenləşdirmə üçün hazırlanmalıdır. Problem ondan ibarətdir ki, bəzi sənəd formatları mətnin məzmununa əlavə olaraq formatlaşdırma haqqında məlumatları da özündə saxlayır. Məsələn, HTML sənədləri cümlə başlanğıcı, qalın yazılar, şifrlərin ölçüsü və stili kimi informasiyaları özündə saxlayan HTML etiketlərindən ibarətdir.

İnformasiya-axtarış sistemi məzmunla “işarələmə” arasındakı fərqi nəzərə almazsa, indeksdə lazımsız məlumatlar mövcud olacaq və bu xəta düzgün olmayan axtarış nəticələrinə səbəb olacaq. Format analizi verilmiş faylın daxilində olan sənədin formatının emal edilib, nümayiş olunması prosesidir. Format analizi həmçinin struktur analizi, teqlərin deskriptasiyası, mətnin formatı, mətnin normallaşdırılması kimi də adlandırılır. Sənədlərin format analizinin işi fərqli fayl formatlarının mövcudluğu ilə dahada çətinləşir. Bəzi fayl formatları əqli mülkiyyət hüquqları ilə qorunur, onlar haqqında az məlumatlar göstərilir, bəziləri isə əksinə yaxşı sənədləşdirilmişdir. Bir çox informasiya-axtarış sisteminin dəstəklədiyi ümumi, yaxşı sənədləşdirilmiş fayl formatları: HTML, ASCII, Microsoft Word, Microsoft Excel, Microsoft PowerPoint, LaTeX, Multimedia meta data format, UseNet, PostScript, PDF, IBM Lotus Notes, SGML.

Bəzi informasiya-axtarış sistemləri sıxılmış və ya şifrəli formatda saxlanılan faylları dəstəkləyir. Sıxılmış formatla işləyərkən indeksləşdirmə modulu əvvəlcə sənədi çıxarır. Bu addım bir və ya daha çox faylın alınmasına gətirib çıxara bilər, bu zaman hər biri indeksləşdirilməlidir. Aşağıdakı dəstəklənən sıxılmış fayl formatları mövcuddur: ZIP, RAR, CAB, Gzip, BZIP, Tape ARchive (TAR), TAR.Z, TAR.GZ və ya TAR.BZ2 [39].

Bölmənin tanınması

Bəzi informasiya-axtarış sistemləri tokenləşdirmədən əvvəl sənədin əsas hissələrini müəyyənləşdirən bölmənin tanınmasını həyata keçirirlər. Heç də bütün sənədlər düzgün yazılmış kitab kimi fəsillər və səhifələr kimi oxunmurlar. İnternetdəki

xəbər paylaşımları və korporativ hesabat kimi bəzi sənədlər əsas məlumat olmayan və yalnız məzmun və yan bloklardan təşkil olunurlar. Baxmayaraq ki, sənədin daxilində olan məlumatlar ekranda müxtəlif hissələrdə göstərilir amma mənbə kodu bu məlumatı ardıcıl şəkildə özündə saxlayır. Mənbə kodunda olan sözlər ardıcıl olaraq indeksləşdirilirlər baxmayaraq ki, onlar ekranın müxtəlif yerlərində əks etdirilirlər.

Informasiya-axtarış sistemləri sənədin bütün məzmununu indeksləşdirdiklərinə görə indeksləmə və axtarış keyfiyyəti aşağı düşə bilər. Aşağıdakı iki əsas problemi qeyd etmək lazımdır:

- Müxtəlif bölmələrdə yerləşməyə birləşmiş (əlaqəli) indeks kimi baxılır baxmayaraq ki, bu belə deyil.
- Yan panelin daxilində olan əlavələr indeksə daxilirlər. Lakin, sənədin həqiqi əhəmiyyətinə əks etdirmir. Buna görə də indeks sənəd haqqında faydasız informasiya ilə dolu olur.

Bölmənin analizi üçün informasiya-axtarış sisteminin vizuallaşdırma məntiqini reallaşdırması tələb olunur. Yəni hər sənəd üçün onun abstrakt təsəvvür olmalıdır və daha sonra sənəd əvəzinə onun haqqında olan təsəvvür indeksləşdirilməlidir. Məsələn: bəzən məzmunun internetdə nümayiş olunması üçün JavaScript – dən istifadə olunur.

Əgər sistem JavaScript – i “görmürsə” o zaman səhifənin indeksləşdirilməsi düzgün həyata keçirilmir. Nəzərə alsaq ki, bəzi informasiya-axtarış sistemləri vizuallaşdırma problemləri barədə düşünmürlər, veb proqramçılar məzmunun düzgün indeksləşdirilməsinə əmin olmaq üçün məzmunu JavaScript üzərindən göstərməməyə çalışırlar və ya NoScript-dən istifadə edirlər. Eyni zamanda bu faktordan sistemi bəzi gizli məlumatları “görməyə” məcbur etmək üçün də istifadə edirlər.

2.4. Sorğu və sıralama modulu

Axtarış sorğusu – istifadəçini maraqlandıran məlumatları özündə ehtiva edən saytların siyahısını əldə etmək üçün axtarış sahəsində daxil edilən söz, söz birləşməsi və ya cümlədir. Axtarış sorğusunun formatı həm informasiya-axtarış sistemindən, həm də axtarış edilən informasiyanın növündən asılıdır. Əsasən, axtarış sorğusu, söz axtarışı və ya informasiya-axtarış sisteminin sorgulama dilinin inkişaf etmiş xüsusiyyətlərini istifadə edərək, bəzəndə söz ya da ifadələr formasında verilir [17].

Axtarış sorğusunun növləri:

- İnformasiya məqsədli – hər hansı bir informasiyanı tapmaq üçün istifadə edilir;
- Tranzaksiyalı – istifadəçi müxtəlif əməliyyatları yerinə yetirmək üçün istifadə edir;
- Naviqasiyalı – hər hansısa bir yeri axtarmaq və ya sayt axtarmaq üçün istifadə olunur;
- Multimediya – şəkil, video və musiqi axtarmaq üçün istifadə olunur;
- Ümumi məqsədli – bu tip sorğularda istifadəçinin nə istədiyi bəlli olmur.

Sorğuların genişləndirilməsi – xüsusilə sorğunu anlama baxımından informasiya-axtarışı əməliyyatlarında axtarış performansını artırmaq üçün axtarış sorğusunun yenidən qurulması prosesidir. İnformasiya-axtarış sistemləri kontekstində sorğu genişləndirilməsi, istifadəçinin daxil etdiyi giriş məlumatlarının qiymətləndirməkdən ibarətdir. Sorğuların genişləndirilməsi aşağıdakı üsulları əhatə edir:

- Sözlərin sinonimlərini tapmaq və sinonimləri də axtarmaq;
- Axtarış sorğusundakı hər sözün müxtəlif morfoloji formalarını yaratmaq və axtarış etmək;
- Qrammatik səhvlərini düzəldərək, yeni formanı avtomatik olaraq axtarmaq və nəticələri istifadəçiyə təklif etmək;
- Orijinal sorğudakı terminləri yenidən dəyərləndirmək;

Sorğu genişləndirilməsi xüsusilə təbii dil emalı və informasiya-axtarış sistemləri sahələrində tədqiq olunan metodologiyadır.

Sorğu modulu son istifadəçi ilə qarşılıqlı əlaqəyə məsuliyyət daşıyır. O istifadəçidən axtarış sorğusunu qəbul edir, onu emal edir, indeksləşdirmə moduluna yönəldir və oradan alınan nəticələri istifadəçiyə təqdim edir. Müasir informasiya-axtarış sistemləri “avtomatik tamamlama” olaraq adlandırılan köməkçi funksiyaya malikdirlər. İstifadəçilər informasiya-axtarış sisteminə sorğuları daxil edərkən, sorğu modulu istifadəçiyə hal-hazırda daxil etdiyi sözləri, söz birləşmələrini, sözlərin müxtəlif formalarını, sorğu ilə əlaqədar və tez-tez verilən sorğuları təklif edir. Bu funksiya nəticəsində informasiya-axtarış sistemləri ümumi istifadəyə yararlı vəziyyətə gətirildi, insan-kompüter qarşılıqlı əlaqəsinin sürəti artırıldı. Bundan əlavə olaraq istifadəyə yararlı funksiyalara axtarış sorğusunu dəqiqləşdirməyə xidmət edən köməkçi xüsusi operatorları aid etmək olar. Misal üçün Google axtarış sistemində bu operatorlar:

- “-söz” – “unec” kimi yazsaq unec sözünün istifadə olunduğu bütün sənədləri axtarışın xaricində saxlayır;
- İkili operatorlar “or”, “and”;
- “sayt:url” – Müəyyən bir URL ünvanında və ya müəyyən bir domendə axtarışı həyata keçirir;
- “link:url” – Axtarılan URL ünvanına bağlı olan bütün veb səhifələrdə axtarış etmək;
- “filetype: query” – Sorğu tipindəki bütün sənədləri axtaracaq.

Axtarış nəticələrini sorğuya uyğun olaraq sıralamaq hər bir informasiya-axtarış sistemi üçün çox vacib bir vəzifədir. Bu məqsədlə müxtəlif faktorların çoxsaylı variantları nəzərə alınır. Bütün bunlar axtarış sisteminin indeksləşdirmə modulunda saxlanılır, bunlardan bəziləri:

- Sənəddəki hər bir sözün istifadə sayı (sözün istifadə sıxlığı olaraq adlandırılır)
- Sənəddəki müəyyən bir sözün yeri, yazı tipi və s.
- Digər səhifələrlə əlaqələri və bu əlaqələr haqqında mühüm məlumatlar (linklər haqqında mətn). Bu internetin modelinin yaradılması və hansı səhifənin haraya aid olduğunu tam olaraq bilmək üçün vacibdir.

Bu məlumatlara əsasən PageRank hesablanır. PageRank informasiya-axtarış sistemlərinin axtarış nəticələrini əlaqəlilik prinsipi əsasında sıralamaq üçün istifadə etdiyi texnologiyalardan ən əsasıdır.

Sıralama modulu, nəticələri ən uyğun nəticə və ya linklərin, informasiya-axtarış sisteminin nəticələr səhifəsində ilk sıralarda göstəriləcəyi formada sıralamanı həyata keçirir. Yəni, sıralama modulunun köməyi ilə informasiya-axtarış sistemi, istifadəçinin sorğusuna ən uyğun olan nəticələri istifadəçiyə sıralayaraq təqdim edir. Bunun əsas üstünlüyü ondadır ki, hər bir sorğuya uyğun bir çox əlaqəli nəticə və ya veb səhifə tapıla bilər və istifadəçi özünə əsas lazım olanı tapmaq üçün onların hər birinə tək-tək baxmalıdır. Bu da öz növbəsində müəyyən qədər vaxt itkisinə səbəb olacaq. Bunun əvəzinə istifadəçilər axtardıqları informasiyanı ən tez müddətdə əldə etmək istəyirlər. Məhz sıralama modulunun məqsədi, bütün bu yorucu və vaxt aparan əməliyyatlardan istifadəçiləri xilas etməkdir [24].

Müasir informasiya-axtarış sistemləri sıralamaq üçün açar sözləri tapmaq və hesablamaqdan əlavə olaraq daha mürəkkəb sıralama alqoritmlərdən istifadə edir. Bu alqoritmlər informasiya-axtarış sistemlərinin insan kimi düzgün və dəqiq nəticələr əldə etməsinə imkan verir.

İnformasiya-axtarış sistemi baxımından, iki məqam veb səhifələrinin sıralanması kimi qiymətləndirilə bilər:

- Veb səhifənin məzmunu termin sıxlığı baxımından nəzərdən keçirilə bilər,
- Link təhlili, səhifənin internetdə nə qədər vacib olduğunu hesablamaq üçün istifadə edilə bilər.

Termin sıxlığı – bir sənəddəki hər hansı bir sözün istifadə miqdarını nəzərə almaqla həmin sənədin önəmini təxmin etmək üçün istifadə edilən statik ölçüdür. Termin sıxlığının tətbiqi əlavə olaraq müəyyən bir sənədin mövzu və ya hətta yalnız bir paraqrafın müəyyən edilməsi üçün faydalıdır. Başqa sözlə desək, sənədi təsvir edən ən yaxşı açar sözləri tapmağa kömək edir.

Link təhlili optimal link parametrlərini üst səviyyəli rəqiblərin link strukturunu öyrənərək müəyyənləşdirir. Bənzər mövzulara sahib digər veb saytlardan yüksək

səviyyəli əlaqələri olan saytlar daha yüksək reytingə sahib olurlar. Optimal link strukturu axtarış reytinglərini əsaslı şəkildə artırır [10].

Link təhlilində istifadə olunan ən əsas üsullardan biri PageRank alqoritmidir. PageRank bir səhifənin "əhəmiyyətini" ölçmək üçün Google-un tətbiq etdiyi sıralama alqoritmidir. PageRank, Google –un qurucuları Larry Page və Sergey Brin tərəfindən yaradılmışdır. PageRank internetdə veb səhifənin əhəmiyyətini 0 –dan 10 –a qədər olan şkalada təyin etmək üçün Google tərəfindən istifadə edilən texnologiyadır [40].

Başlıq etiketi və açar sözlər kimi bütün digər amillər nəzərə alındıqda, Google daha "mühüm" saytları müvafiq olaraq istifadəçilərin axtarış nəticələrinə görə artan sıra ilə nəticələri istifadəçiyə təqdim etmək üçün PageRank alqoritmindən istifadə edir. Yəni, Google-da sıralama modulu aşağıdakı kimi işləyir:

- Açar sözlər ilə uyğun olan bütün səhifələri tapır.
- Açar sözləri "səhifə faktorları" kimi uyğun olaraq təşkil edir.
- Səhifələrdəki linklərin mətnini nəzərə alır.
- PageRank verilənləri ilə düzgün nəticələr əldə edir.

PageRank –ın təyin edilməsi: Əgər A səhifəsi B səhifəsinə istinad edirsə, onda A səhifəsi B səhifəsinin “əhəmiyyətli” səhifə olduğunu hesab edir. Əgər C səhifəsi B səhifəsinə istinad edirsə bu zaman A səhifəsi daha əhəmiyyətli hala gəlir. A səhifəsinə istinad edən səhifə sayı nə qədər çox olarsa A səhifəsinin reytingi bir o qədər yüksək olacaq. Təbii ki, burada əsas faktorlardan biri də A səhifəsinə baxış sayıdır. Həm səhifəyə baxış sayı həm də səhifəyə edilən istinadlar PageRank – ın təyin edilməsində əsas rol oynayır.

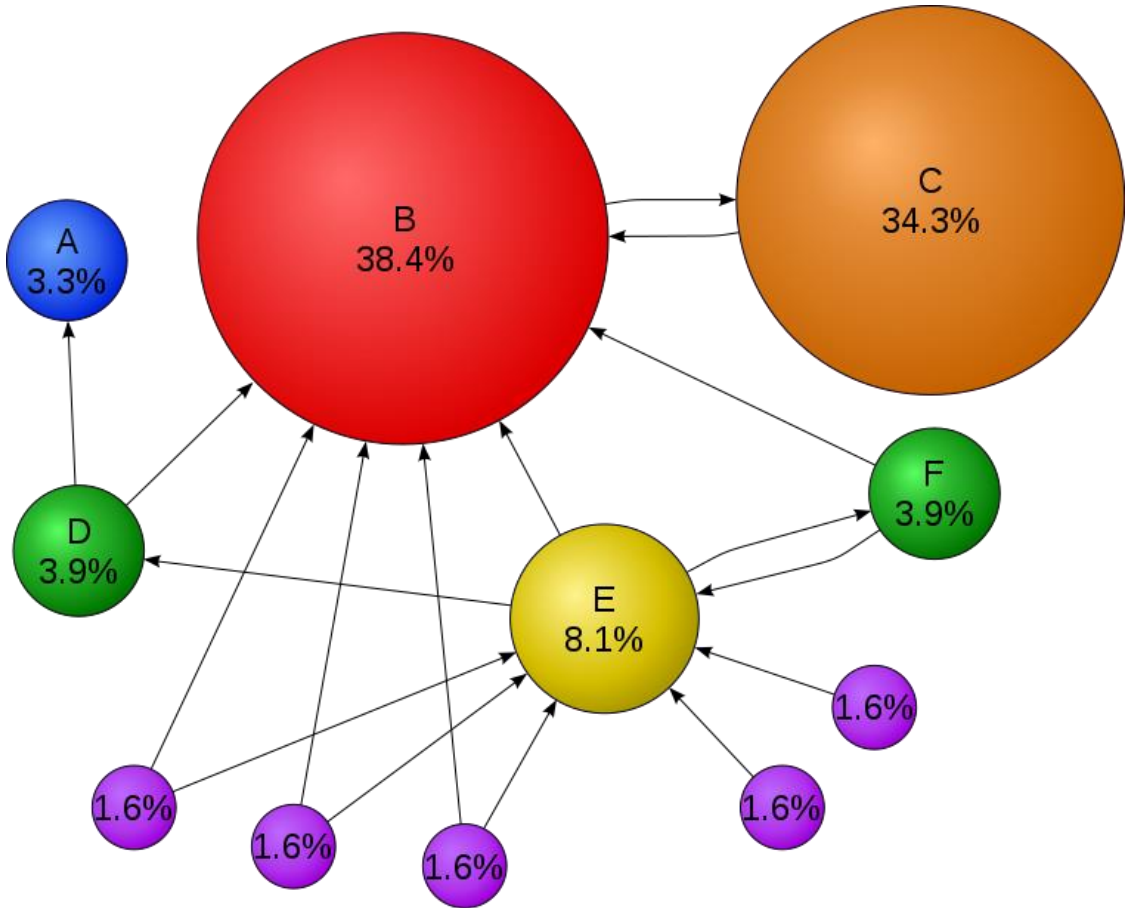
PageRank –a təsir edən amillər:

- Səhifəyə daxil olan istifadəçi sayı
- Səhifəyə edilən istinadlar
- İstinad edən səhifənin öz PageRank dəyəri
- Səhifəyə link verən saytların PagerRank dəyərləri
- Səhifədə istifadə olunan başlıqlar və açar sözlər.

PageRank aşağıdakı düstur vasitəsi ilə hesablanır:

$$PR(A) = (1 - d) + d \left(\frac{PR(T_1)}{C(T_1)} + \dots + \frac{PR(T_n)}{C(T_n)} \right)$$

Burada $PR(A)$ baxılan veb səhifənin PageRank dəyəri, T_i i -ci istinad edən veb səhifə, n veb səhifəyə istinad edən başqa veb səhifələrin sayı, C donor səhifədə olan kənar istinadların sayıdır, d donor səhifənin akseptor-səhifəyə ötürdüyü çəki hissəsini əks etdirən əmsal (adətən onun qiymətini 0,85 götürürlər, yəni donor səhifə akseptor səhifələrə öz çəkisinin 85%-ə qədərini verə bilər) və d 0 ilə 1 arasında qiymət ala bilər. PageRank – a sadə bir nümunə kimi aşağıdakı şəkil 6-nı göstərmək olar:



Şəkil 6. Riyazi olaraq PageRank – a nümunə

E səhifəsinin bəzi səhifələrə nisbətən daha yüksək PageRank dəyəri vardır, çünki onu bir çox səhifə istinad etmişdir. Baxmayaraq ki, C səhifəsi E səhifəsinə nisbətən daha az istinada malikdir, lakin C səhifəsi E səhifəsindən daha yüksək PageRank –a

sahibdir. Bunun səbəbi yüksək PageRank – a sahib B səhifəsinin C səhifəsinə bağlı olmasıdır.

Veb səhifə sahibləri üçün PageRank mühüm önəm daşıyır. Əgər səhifənin PageRank dəyəri aşağıdırsa axtarış olunduğu zaman həmin səhifə istifadəçiyə göstərilməyə bilər. Bu səbəblə veb səhifə sahibləri daima öz səhifələrinin PageRank dəyərini yüksək saxlamağa məcburdurlar.

Sorğu və sıralama modulu əldə olunan nəticələri sıralayarkən yüksək PageRank dəyərindən azalana doğru sıralayır. Yəni, daha yüksək PageRank – a sahib olan səhifələr daha yuxarıda özünə yer tapır.

2.5. İnformasiya-axtarış sisteminin verilənlər bazası

Əvvəlki mövzulardan aydın olduğu kimi axtarış robotu veb-i skan edərkən oradan topladığı informasiyaları indeksləşdirmə moduluna ötürür. Bu informasiyalar linklərdən, səhifə başlıqlarından, açar sözlərdən, səhifənin hansı serverdə yerləşdiyindən, məqalədə istifadə olunan əsas terminlərdən, faylların yerləşdiyi serverlərin ünvanlarından, faylların ölçülərindən, mətnə istifadə olunan söz sayından, HTML kodlarından və.s ibarət ola bilər. Bütün bu informasiyaları birbaşa informasiya-axtarış sisteminin verilənlər bazasında saxlamaq mümkün deyil. Çünki, bütün bu informasiyaların verilənlər bazasında saxlanması aşağıdakı problemləri yarada bilər [41]:

- Axtarışın sürətini fərq ediləcək qədər aşağıya endirəcəkdir;
- Axtarış nəticələrində anlaşılmazlıqlar yaranacaqdır;
- Axtarış sistemi bütün bu verilənləri emal etmək üçün yüksək iş tezliyi tələb edəcəkdir;
- Bəzi nəticələr istifadəçiyə təqdim edilməyə bilər;
- Axtarış müddəti həddən artıq uzanacaqdır;
- Verilənlər bazasında verilənlər arasında əlaqəsizlik problemi yarana bilər;
- Verilənlərin yenilənməsi zamanı müxtəlif problemlər baş verə bilər;

- Bütün informasiyaları verilənlər bazasında saxlamaq üçün həddən artıq yer tələb olunacaqdır. Nəzərə almaq lazımdır ki, internetin həcmi hər saniyə dayanmadan artır və bu qədər informasiyanı hansısa bir bazada toplamaq mümkünsüzdür.

Yuxarıda sadalanan problemlərlə qarşılaşmamaq üçün informasiya-axtarış sistemləri indeksləşdirmə modulundan istifadə edirlər. Yəni axtarış robotu tərəfindən əldə edilən informasiyalar indeksləşdirmə modulu vasitəsilə indekslənmiş hala salınır.

Əldə edilmiş indeksli verilənlər bundan sonra informasiya-axtarış sisteminin verilənlər bazasında saxlanılır. İndeksələnmiş verilənlərin indekslənməmiş verilənlərdən üstünlükləri aşağıdakılardır:

- Axtarış sorğusunun müddəti minimuma enəcəkdir. İstifadəçi öz sorğusunun nəticəsini ən qısa müddətdə əldə edəcəkdir;
- İnformasiya-axtarış sistemi verilənlər bazasında axtarış etmək üçün minimum iş tezliyinə ehtiyac duyacaqdır. Yəni, axtardığı indeks rahatlıqla tapacaqdır;
- Axtarışın sürətində yüksəlmə olacaqdır;
- Bazadakı indekslərin yenilənməsi asanlaşacaqdır;
- Bazadakı verilənlər indeksləndi üçün yaddaşda minimum yer tələb edəcək və bu da öz növbəsində verilənlər bazasının yaddaşından qənaət etməsinə səbəb olacaqdır;
- İndeksələnmiş halda olan verilənlər indekslənməmiş verilənlərə nisbətdə veb-in böyük bir hissəsinin təhlil edilməyə imkan verəcəkdir;
- İstifadəçilərə təqdim olunan nəticələr dolğun və dəqiq olacaqdır.

Bəzən informasiya-axtarış sistemlərinin verilənlər bazasını indekslər bazası olaraq adlandırırlar. Buna səbəb indeksləşdirmə modulunun indekslədiyi verilənləri həmin bazada toplamasıdır. Yəni informasiya-axtarış sisteminin verilənlər bazasının əsasını indekslər təşkil edir. Bu indekslər informasiya-axtarış sisteminə istifadəçi tərəfindən sorğu verildiyi zaman indeksləşdirmə modulunun verilənlər bazasından özünə lazım olanı əldə etməsi prosesini açıq şəkildə sürətləndirir.

Lakin qeyd etmək lazımdır ki, informasiya-axtarış sisteminin verilənlər bazası sadəcə indeksləri özündə saxlamır. Eyni zamanda istifadəçilər tərəfindən edilən axtarış

sorğularını da yadda saxlayır. O bununla istifadəçilərin daha çox nəyi axtardığını, onların daha çox nə ilə maraqlandıklarını müəyyən edir. Bundan əlavə olaraq həmin bazada istifadəçilərə təqdim olunan nəticələr haqqında keş fayllarda saxlanılır. Bu isə indeksləşdirmə modulunun üzərinə düşən yükü azaldır.

İnformasiya-axtarış sistemləri istifadəçi məlumatlarını bazada saxlamaq baxımdan iki növə ayrılır:

- İstifadəçi məlumatlarını özündə saxlayan – Google, Yandex, Bing, Ask.com, AOL, Lycos.
- İstifadəçi məlumatlarını özündə saxlamayan – DuckDuckGo, İxquick və.s

İnformasiya-axtarış sistemlərinin verilənləri böyük yaddaş tutumuna, yüksək iş tezliyinə, sürətli yeni nəsil prosessorlara sahib serverlərdə saxlanır. Bunlara misal olaraq Google şirkətinə məxsus çox saylı və regional olaraq paylanmış Data mərkəzlərini göstərmək olar. Qeyd etmək lazımdır ki, həmin Data mərkəzlərində çoxsaylı serverlər quraşdırılmışdır. Bu Data mərkəzlər ABŞ-da 8 ədəd, Şimali Amerikada 1 ədəd, Avropada 4 ədəd, Asiyada 2 ədəd olmaqla qurulmuşdur. Bütün bunlar bulud server arxitekturasını dəstəkləyir.

Bəzən istifadəçilər internetdə olan media faylları, saytlar, və digər informasiyaların həmin serverlərdə saxlanıldığını düşünürlər. Lakin bunun əksinə onlar müxtəlif fərqli serverlərdə saxlanılırlar. İnformasiya-axtarış sisteminin serverləri isə yalnız onların indekslərini, sorğu nəticələrini və bəzi istifadəçi məlumatlarını özündə saxlayır.

2.6. İstifadəçi interfeysi

İstifadəçi interfeysi informasiya-axtarış sistemi ilə istifadəçi arasında qarşılıqlı əlaqəni təmin edir. İstifadəçi interfeysi HTML kodlarını istifadəçilərin anlama biləcəkləri formaya salır. Yəni bir veb səhifənin HTML kodunda o səhifədəki başlıqların rəngi kod olaraq ağ mavi verilərsə, informasiya-axtarış sistemi o səhifənin başlıqlarını vizual olaraq istifadəçiyə mavi rəngdə təqdim edəcəkdir [17].

İstifadəçilər informasiya-axtarış sisteminin yalnız istifadəçi interfeysi ilə qarşılıqlı əlaqədə olurlar. İstifadəçilər açar sözləri istifadəçi interfeysi vasitəsilə informasiya-axtarış sistemində daxil edir və sorğunun nəticəsini istifadəçi interfeysindən əldə edir. Bu səbəblə istifadəçi interfeysi nə qədər sadə və rahat olarsa, istifadəçilər tərəfindən həmin informasiya-axtarış sistemində tələbat çox olar.

Hər bir informasiya-axtarış sistemi müxəlif istifadəçi interfeysinə malikdir. İstifadəçilər informasiya-axtarış sistemi seçərkən, əsasən istifadəçi interfeysi və düzgün nəticələr təqdim edən informasiya-axtarış sistemlərinə üstünlük verirlər. İstifadəçi interfeysi aşağıdakı xüsusiyyətlərə malik olmalıdır:

- İstifadəçi tərəfindən rahat anlaşılan olmalıdır;
- Naviqasiya paneli istifadəçiyə rahat olmalıdır;
- Arxa fon seçimi düzgün olmalıdır;
- Əlavə funksiya seçimlərinə malik olmalıdır;
- İnformasiya-axtarış sisteminin loqosu münasib yerdə olmalıdır;
- Nəticələr səhifələri arasında keçid rahat və sürətli olmalıdır;
- Nəticələr səhifəsində təqdim olunan şəkillərin keyfiyyəti optimallaşdırılmış olmalıdır ;
- Müxtəlif dil seçimlərini dəstəkləməlidir;
- Axtarış zamanı fərqli fayl tiplərinin seçimini dəstəkləməlidir;
- Digər informasiya-axtarış sistemlərinə nəzərən yeniliklər olmalıdır.

Hər bir informasiya-axtarış sisteminin interfeysi iki hissədən ibarətdir:

- Sorğu səhifəsi
- Nəticələr səhifəsi

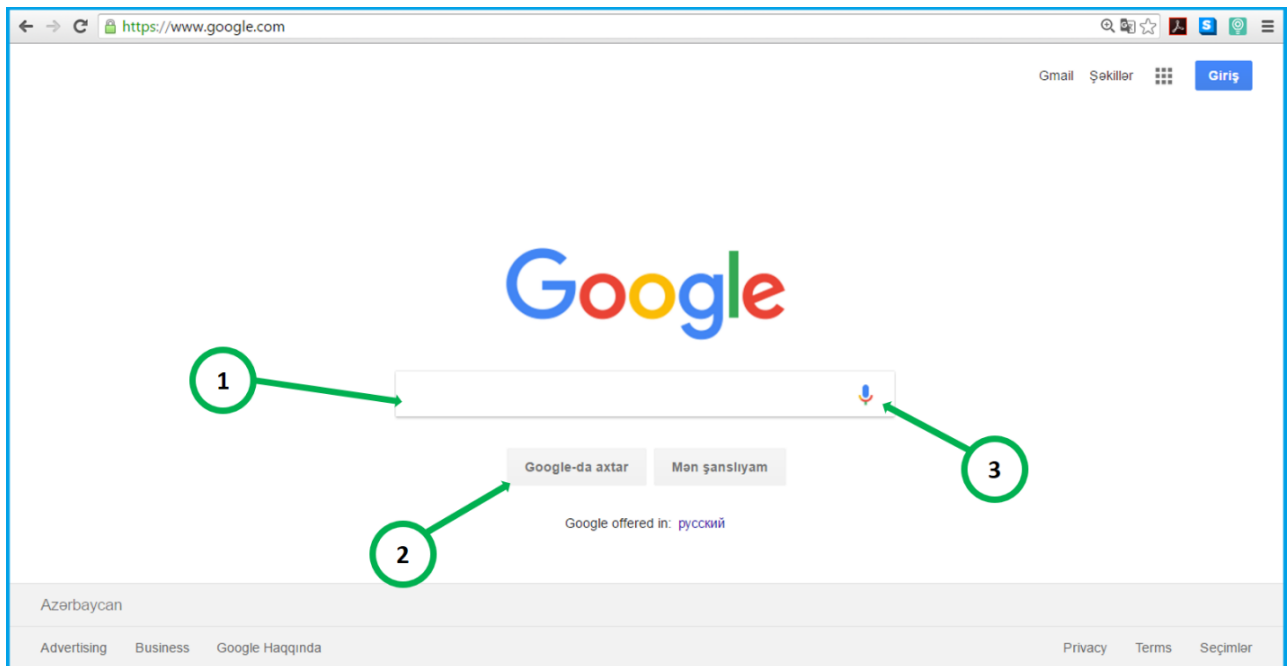
Sorğu səhifəsi istifadəçilərin öz sorğularını daxil etdikləri sahədir. Burada axtarış qutusu, “axtar” düyməsi, informasiya-axtarış sisteminin loqosu və əlavə seçimlər yerləşir.

Nəticələr səhifəsi isə informasiya-axtarış sisteminin sorğunu emal etdikdən sonra istifadəçiyə təqdim etdiyi nəticələrin göstərildiyi sahədir. Burada səhifələr arasında keçid imkanı, axtarış sisteminin loqosu, müxtəlif fayl tipləri üzrə ayırma (şəkillər, videolar, kitablar və.s) və əlavə seçimlər yerləşir.

İnformasiya-axtarış sistemləri aşağıdakı istifadəçi interfeysi vasitələrinə sahib ola bilərlər:

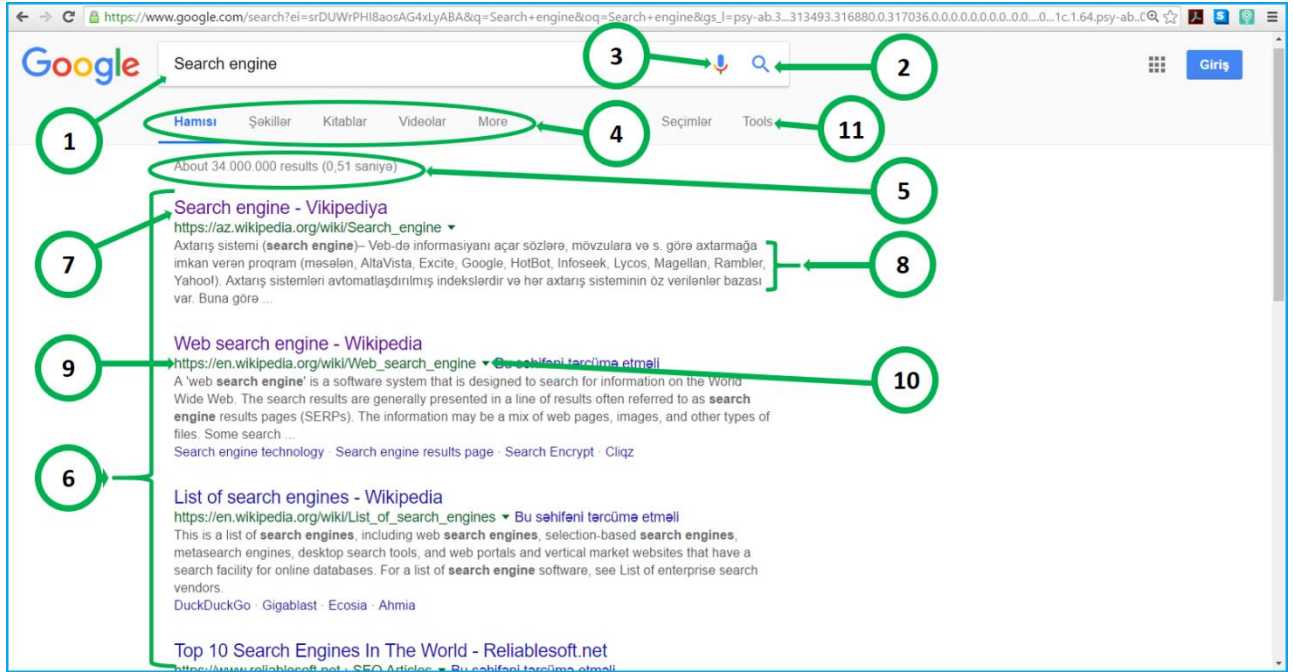
- ✓ İnformasiya-axtarış sistemi ilə istifadəçi arasında interfeys;
- ✓ Sistem resursları ilə informasiya-axtarış sistemi arasında interfeys;
- ✓ Şəbəkə resurslarına sahib istifadəçi interfeysi

Yuxarıda sadalananları aydın formada anlamaq üçün Google informasiya-axtarış sisteminin istifadəçi interfeysini nəzərdən keçirək:



Şəkil 7. Google informasiya-axtarış sisteminin sorğu səhifəsi

- (1) – Axtarış sorğusu sahəsi;
- (2) – Google axtarış sisteminin axtarış düyməsi;
- (3) – Sorğuların səsə daxil edilməsi düyməsi;



Şəkil 8. Google informasiya-axtarış sisteminin nəticələr sahifəsi

- (4) – Google axtarış sisteminin müxtəlif axtarış xidmətlərinə istinadları;
- (5) – Axtarış nəticəsində tapılmış sənədlərin sayı və axtarışa sərf edilmiş müddət;
- (6) – Axtarış nəticəsində Google tərəfindən tapılmış sənədlərin siyahısı, onlara istinad və onlar haqqında qısa məlumatlar;
- (7) – Tapılmış sənədin və ya veb saytın adı;
- (8) – Tapılmış sənədin məzmunundakı sorğu şərtləri, tərkibində rast gəlinən fraqmentlər;
- (9) – Axtarış nəticəsində tapılmış sənədin ünvanı;
- (10) – Bənzər sənədlərin axtarışı;
- (11) – Axtarışın sorğusunun sazlanması düyməsi.

Fəsil 3. İnformasiya-axtarış sistemlərinin inkişaf itiqamətləri

3.1. Axtarış nəticələrinin yaxşılaşdırılması qaydaları

İnternetdə hər gün çox sayda veb səhifələr yaradılır. Bu qədər veb səhifənin yaradılması internetdə informasiya kirliliyinə səbəb olur. Bu qarışıq informasiya toplusunun içərisindən lazım olan informasiyanı əldə etmək üçün istifadəçilərin yeganə vasitəsi informasiya-axtarış sistemləridir.

İnformasiya-axtarış nəticələrinin yaxşılaşdırılmasında iki əsas faktor rol oynayır:

1. Düzgün informasiya-axtarış sistemindən istifadə etmək;
2. Düzgün axtarış üsullarından faydalanmaq.

Yuxarıda sadalananların hər ikisinə də riayət edildikdə, istifadəçi həm zamandan qənaət edir həm də lazım olan ən yaxşı informasiyanı qısa yoldan əldə edir.

Axtarışa başlamazdan əvvəl istifadəçi axtaracağı informasiyaya uyğun informasiya-axtarış sistemini seçməlidir. Hazırkı dövrdə Google ən geniş istifadə olunan informasiya-axtarış sistemidir. Google sisteminin daxil olduğu ümumi məqsədli informasiya-axtarış sistemləri mətnlərin axtarılmasında ideal sistemlərdir. Yahoo! isə kətiqorik informasiya-axtarışında inkişaf etmişdir. Bunlardan əlavə olaraq bir neçə informasiya-axtarış sistemlərini eyni anda istifadə edən meta-axtarış sistemlərində istifadə etmək mümkündür. Misal üçün, Cophernic – dən istifadə edərək istifadəçi ən məşhur informasiya-axtarış sistemlərinin nəticələrindən eyni anda yararlanı bilər.

Düzgün axtarış üsullarına aşağıdakı metodları aid etmək olar:

- Axtarılacaq mövzu ilə əlaqəli açar sözlərin seçilməsi. Ən dəqiq nəticəni almaq üçün bir neçə açar sözdən istifadə etmək məqsədəuyğundur.
- İki və daha artıq açar sözü birləşdirərkən onların arasına “and”, “&” və ya “+” simvollarından birini yerləşdirmək.
- Sözləri hissə-hissə axtarmaq üçün “?” və “*” simvollarından istifadə etmək. Əgər sözün əvvəlinin tamamlamaq lazımdırsa sözün əvvəlinə “?” (?formasiya),

əksinə sözün sonunu tamamlamaq lazımdırsa “*” (info*) simvolundan istifadə etmək olar.

- Açar sözləri tək-tək axtarmaq.
- Sadəcə daxil edilən açar sözlərdən ibarət nəticələr əldə etmək üçün açar sözləri dırnaq işarəsi içərisində yazmaq lazımdır (“açar söz”).
- Açar sözlərdən göstərilməsini istəmədiyiniz sözlərin əvvəlinə “-” simvolu qoymaq.
- Nəticələrin daxilində mütləq olmasını istədiyinin açar sözlərin əvvəlinə “+” simvolunu yerləşdirmək.
- Açar sözləri daxil edərkən böyük və ya kiçik hərflərlə yazmağın mənası yoxdur.
- Sorğuları sadə şəkildə vermək.
- Əgər sosial şəbəkə istifadəçiləri axtarılacaqsa, onda adın qarışısına “+” və ya “@” simvolu əlavə etmək.
- Sorğuda istifadə edilən sözlərin sinonimlərindən istifadə etmək.
- Dəqiq fayl tipində olan fayl və sənədlərin tapılması üçün (filetype:pdf) operatorunu açar sözlərin əvvəlinə yazmaq.
- Əgər müəyyən tarixlərdə əlavə edilmiş informasiyalar axtarılırsa həmin tarixlərin arasına “..” simvolları əlavə edilir (2010..2018).
- Google informasiya-axtəriş sistemini hesablayıcı(kalkulyator) kimi istifadə etmək üçün sorğu sahəsinə hesablanacaq rəqəmləri daxil etmək kifayətdir (100*6).
- “Allintext: ” operatoru vasitəsilə axtarılan sözlərin hamısının mətnin daxilində olduğu səhifələr tapılır.
- “Cache: ” operatoru vasitəsilə Google – da əvvəlcədən edilmiş sorğular içərisində axtəriş etmək.
- “Site: ” operatorundan istifadə etməklə axtəriş yalnız həmin sayt daxilində aparılacaqdır.

3.2. İnformasiya-axtarış sisteminin optimallaşdırılması

İnformasiya-axtarış sistemin optimallaşdırılması yəni SEO (Search Engine Optimization) – veb səhifələrin informasiya-axtarış sistemlərinin kriteriyalarına uyğun hala gətirilərək axtarış nəticələrində ilk sıralarda təqdim edilməsinə xidmət edir. Müasir informasiya-axtarış sistemləri nəticələri sıraladığı zaman müxtəlif kriteriyalara görə tapılan nəticələr arasında seçim edirlər. Bu səbəblə veb proqramçılar yaratdıqları səhifərdə bu kriteriyalara diqqət edirlər. Adətən axtarış nəticələrində ən üst sıralarda olan veb səhifələr yüksək ziyarətçi sayına sahib olurlar. Bu da öz növbəsində həmin səhifənin reytingini yüksəldir [31].

İnformasiya-axtarış sistemlərinin optimallaşdırılması aşağıdakı məqsədlər üçün edilir:

- Açar sözlərin axtarışı zamanı səhifənin üst sıralarda yerləşməsi;
- Veb səhifənin reyting sıralamasında yüksəlməsi;
- Veb səhifəyə ziyarətçi sayının artması;
- Reklam performansını artırmaq;
- Sayta baxış sayının artırılması;
- Veb saytın maddi və mənəvi dəyərini yüksəltmək.

İnformasiya-axtarış sistemlərinin optimallaşdırılması iki istiqamətdə aparılır: daxili optimallaşdırma və xarici optimallaşdırma [33].

Daxili optimallaşdırma – informasiya-axtarış sistemlərinin veb səhifələri daha rahat skan etməsi və anlaması üçün səhifə daxilində edilən bəzi texniki dəyişikliklərdir. Daxili optimallaşdırma tədbirlərinə aşağıdakılar daxildir:

- Keyfiyyətli və orijinal məzmun hazırlamaq;
- Veb saytın sürətinin yüksəldilməsi;
- Mobil cihazlar üçün veb səhifənin uyğunlaşdırılması;
- HTML, CSS istifadəsi;
- Mövzu ilə əlaqəli açar sözlərin araşdırılması və düzgün seçilməsi;
- Rəqib veb səhifələrin müəyyən edilməsi;

- Səhifə başlıqlarının Meta etiketlərə etikətlənməsi;
- W3C standartlarının tətbiqi;
- Robot.txt faylından istifadə edilməsi;
- Uyğun səhifə başlıqlarının seçilməsi;
- Qrafik və şəkillərdə keyfiyyətin optimallaşdırılması;
- Domen – in düzgün seçilməsi;
- Lazımsız JavaScript – lərin veb saytdan yığılması.

Xarici optimallaşdırma üçün bir çox fərqli üsullar mövcuddur. Ən məşhur olanı isə veb saytın keyfiyyətli və yüksək backlink sayına sahib olmasıdır. Backlink – bir veb saytın digər veb saytlara verdiyi xarici linklərdir. Axtarış robotları backlink-ləri referans olaraq qəbul edir. Bundan əlavə aşağıdakı üsullar da xarici optimallaşdırmada istifadə edilir:

- Sosial şəbəkələr vasitəsilə sosial əlaqələrin artırılması;
- Başqa veb saytlardan trafikə təmin edilməsi;
- Müxtəlif reklamlar vasitəsilə veb sayta girişlərin artırılması;
- Keyfiyyətli və orijinal backlinklərin toplanılması üçün veb saytın hər zaman güncəllənməsi;
- Veb sayta birbaşa girişlərin artırılması.

Başqa sözlə desək informasiya-axtarış sistemlərinin optimallaşdırılması tədbirləri birbaşa sıralama moduluna təsir etmək üçün nəzərdə tutulmuş tədbirlər kompleksidir. İnformasiya-axtarış sistemlərinin optimallaşdırılması ən çox internet marketing sahəsində və veb sayt sahibləri tərəfindən geniş istifadə olunur. Buna səbəb onların öz məhsullarını (və ya saytlarını) axtarış zamanı ən üst sıralarda təqdim etməkdir.

3.3. İnformasiya-axtarış sistemlərində istifadə olunan ən yeni alqoritmlər.

Google və Yandex informasiya-axtarış sistemlərinin alqoritmləri.

Müxtəlif informasiya-axtarış sistemlərində eyni sorğuları verərkən alınan nəticələrdəki fərqliliklər kifayət qədər olur. Halbuki bütün informasiya-axtarış sistemləri eyni bazadan yəni internetdən yararlanırlar. Alınan fərqli nəticələrin səbəbi məhz həmin informasiya-axtarış sistemlərinin alqoritmlərindən qaynaqlanır. Hər bir informasiya-axtarış sistemi indeksləşdirmə, sorğu və sıralama modullarında fərqli texnologiyalar və alqoritmlərdən istifadə edir. Rusdilli ölkələr Yandex sistemindən geniş istifadə edirlər. Digər ölkələr üçün isə Google ən unikal informasiya-axtarış sistemi hesab olunur. Lakin istifadəçi sayını nəzərə aldıqda Google Yandex-i qabaqlayır. Bu Yandex –in Google –dan zəif olması demək deyil, sadəcə olaraq regional istifadə məsələsidir. Hər ikisi kifayət qədər istifadəçi sayına sahibdirlər.

Google link analizində PageRank alqoritmindən istifadə edir. Yandex isə bunun əvəzinə maşın təliminə əsaslanan MatrixNet texnologiyasında faydalanır.

➤ *Google informasiya-axtarış sisteminin alqoritmləri*



Şəkil 9. Google informasiya-axtarış sisteminin loqosu

Google informasiya-axtarış sistemləri bazarında əsas yeri tutan veb axtarış sistemidir. Bu danılmaz faktdır ki, o həm rusdilli həm də digər ölkələrdə geniş istifadəçi kütləsinə malikdir. Məhz bu qədər istifadəçinin Google-u seçməsi təsadüfi ola bilməzdi. Təbii ki, bunun əsas səbəblərindən biri də Google –un istifadəçiləri digər informasiya-axtarış sistemlərinə nəzərən daha dolğun və dəqiq nəticələrlə təmin etməsidir. Google –da dayanmadan öz informasiya-axtarış sistemində yeniliklər etmək

məqsədlə yeni alqoritmləri tətbiq edir. Google –un alqoritmlərinə və onların tətbiqi ilə gələn yeniliklərə nəzər salaq [36]:

1. “Caffeine” alqoritmi (10 avqust 2009)

Google Caffeine alqoritmi ilə öz informasiya-axtarış sistemi tarixindəki ən əhəmiyyətli yenilənməni təqdim etdi. Google şirkəti uzun müddət həmin alqoritm üzərində çalışmışdır və sonunda onu mükəmməl hala gətirmişdir.

Caffeine alqoritmi yeni indeksləmə alqoritmi idi. Bu yeni sistem Google –un informasiyaları daha məhsuldar formada təhlil etməsinə və verilənlər bazasında saxlamasına imkan yaradı. Google bu alqoritm vasitəsilə bazasındakı indeks miqdarının 50% -ə qədər artırdı. Həmin alqoritm fərqli axtarış robotlarını – crawlerləri göndərərək əldə etdiyi nəticələri yenidən təhlil edir, sonra təkrar indekslənməni həyata keçirirdi.

Caffeine alqoritmi ilə Google veb səhifələri təhlil etmə, verilənlərin toplanması və indekslənməsini saniyələr ərzində reallaşdırırdı.

2. “Panda” alqoritmi (24 fevral 2011)

Panda alqoritmi Google informasiya axtarış sistemi üçün yaradılmış sıralama alqoritmi idi. Bu alqoritm veb səhifələrə keyfiyyət əmsalı təyin edirdi. Daha sonra bu əmsalı uyğun olaraq sıralama faktoru kimi istifadə edirdi.

Panda alqoritmi əsasən plagiat məzmunların, istifadəçilər tərəfindən yaradılan spamların və açar sözlərin ifrad dərəcədə istifadəsini təyin etməyə xidmət edirdi.

Həmçinin Google –un Panda alqoritmının əsas məqsədi orijinal və keyfiyyətli məzmunu sahib olan veb səhifələri sıralama zamanı reyting cədvəlində yuxarıya çəkmək, keyfiyyətsiz veb səhifələri isə ən aşağıya endirmək idi. Bundan başqa Panda alqoritmının yeniliklərindən biri də istifadəçilərin davranışlarına nəzarət etməsidir. Yəni istifadəçinin veb saytda nə qədər qalması, internetdə nələrə maraqlanması, istifadə etdiyi linklər kimi əlavə parametrlərə diqqət yetirirdi.

3. “Penguin” alqoritmi (24 aprel 2012)

Penguin alqoritmi iki əsas yeniliyin tətbiqi ilə nəzərə çarpır:

1. Spam və ya lazımsız linklərin təyin edilməsi.
2. Həddindən artıq optimallaşdırılmış zəif və keyfiyyətsiz məzmunu sahib veb səhifələrin təyin edilib müəyyən tədbirlərin görülməsi.

Bu alqoritm yeniləməsi, sosial şəbəkələrdən istifadə edərək zəif məzmunlu, aşağı keyfiyyətli linklər yaradaraq səhifə reytinglərini artıran saytlara qarşı bir növ filtr vəzifəsini daşıyırdı. Penguin alqoritminin əsas məqsədi axtarış göstəricilərini manipulyasiya edən saytlar təyin edib tədbir görmək idi.

Google –un spam məzmununa qarşı çıxdığı Panda alqoritminə əlavə olaraq, Penguin alqoritmi veb saytlara daxil olan linklərin keyfiyyət qaydalarına uyğun olub-olmadığını müəyyən etməyə xidmət edir.

Google Penguin alqoritminin 4 versiyası mövcuddur. Ən son versiya olan Penguin 4.0 versiyası 2016-cı ildə tətbiq edilmişdir.

4. *“Pirate” alqoritmi* (10 avqust 2012)

Bu alqoritm müəllif hüquqlarını qorumaq məqsədi ilə Google tərəfindən yaradılan alqoritmdir. Film saytları əsas təsir obyektidir. Əgər hər hansısa bir veb sayta müəllif hüquqlarının pozulduğunu təyin edərsə dərhal həmin veb saytı silir. Bu alqoritm sadəcə avtomatik işləmir, eyni zamanda müəlliflərin şikayətləri əsasında təyin edilmiş veb saytları da silir.

5. *“Hummingbird” alqoritmi* (22 avqust 2013)

Hummingbird Google sisteminin əsas axtarış alqoritm bir hissəsidir. O biliklər qrafından istifadə edərək təbii dildə olan sorğulara cavab verməni sürətləndirirdi. Bu alqoritm tətbiqi ilə Google –da böyük yeniliklər olmuşdur. Hummingbird özündən əvvəlki alqoritm kimi spamlarla və ya keyfiyyətsiz linklərlə mübarizə etmək üçün nəzərdə tutulmamışdı. Onun məqsədi istifadəçilər tərəfindən verilən sorğuların daha yaxşı anlaşılması, onlara ən yaxşı nəticələrin təqdim edilməsi idi. Hummingbird –in

tətbiqi ilə artıq Google istifadəçilərin sorğularını daha aydın başa düşürdü. Bu da öz növbəsində istifadəçiyə lazım olan konkret nəticənin təqdim edilməsini reallaşdırırdı.

Hummingbird Google informasiya-axtarış sisteminə səsli axtarışların yerinə yetirilməsi üçün şərait yaradırdı. O istifadəçilərin sorğularını anlaya bilir, cümlələrin mənalarını dərk edirdi. Hummingbird əsasən istifadəçiyə sorğuları ilə daha əlaqəli nəticələr vermək üçün nəzərdə tutulmuşdur.

6. *“Pigeon” alqoritmi* (24 iyul 2014)

Google –un Pigeon alqoritmi ilə gələn əsas yenilik artıq sorğuların nəticələri sıralanarkən istifadəçinin yerləşdiyi yer nəzərə alınır. Pigeon Google Map (xəritə) proqramı ilə birgə işləyirdi. İstifadəçilər sorğuları daxil etdiyi zaman xəritədə onların yerləşdiyi koordinatlar nəzərə alınır, tapılan nəticələr lokal olaraq sıralanırdı. Bu alqoritm yerli bizneslə məşğul olanlara müsbət təsir etdi. Belə ki, istifadəçi fərqli ölkələrdə filialları olan şirkətin adını axtarısa verdikdə Google istifadəçinin yerini təyin edərək həmin koordinata yaxın olan şirkətin filialının nəticəsini Google Map vasitəsilə istifadəçiyə ekranın sağ küncündə təqdim edir. İstifadəçi bu səhifəyə daxil olarsa cihazdakı Google Map tətbiqi ilə həmin şirkətə asanlıqla gedə bilər.

7. *“Mobilegeddon” alqoritmi* (21 aprel 2015)

Google bu yeni alqoritm ilə artıq nəticələrin sıralanmasını mobil cihazlar üçün fərqli formada edirdi. Mobilegeddon alqoritminin tətbiqindən sonra Google mobil cihazlar ilə fərdi kompüterlərdən verilən sorğular arasında fərq qoymağa başladı. Əgər istifadəçi mobil cihazdan istifadə edərək Google –da axtarış edərsə bu zaman sıralama alqoritmi mobil cihazlar üçün optimallaşdırılmış veb saytlara üstünlük verirdi. Bu o deməkdir ki, əgər veb səhifə mobil qurğular üçün optimallaşdırılmayıbsa mobil cihazla axtarış edildiyi zaman həmin veb səhifə təqdim edilməyəcək. Google Mobilegeddon alqoritmi ilə veb səhifə sahiblərinin öz səhifələrini fərdi kompüter versiyalarından əlavə olaraq mobil cihazlar üçün uyğun hala gətirməyə məcbur etdi.

8. *“RankBrain” alqoritmi* (26 oktyabr 2015)

RankBrain alqoritmi Google Hummingbird alqoritminin bir hissəsidir. Bu alqoritm Google –un sorğuların mənanalarını anlamasına kömək edən maşın təlimi sistemidir və bu onun sorğulara cavab olaraq ən uyğun axtarış nəticələrinin tapılmasına xidmət edir. Google RankBrain alqoritmini üçüncü ən əhəmiyyətli sıralama faktoru kimi təqdim etmişdir. RankBrain həm sorğu həm də sıralama funksiyalarını reallaşdırırdı. Həmçinin bu alqoritm Google –un ilk maşın təliminə əsaslanan alqoritmi idi. Onun köməyi ilə artıq Google insan kimi düşünə bilir, istifadəçiləri daha rahat anlayır, onların nə ilə maraqlandıklarını müəyyən edə bilirdi. Yəni əgər istifadəçi sorğunu səhf daxil etsə belə, bu alqoritmin köməyi ilə Google istifadəçinin nə istədiyi müəyyənləşdirir və düzgün nəticəni təqdim edir.

9. *“Possum” alqoritmi* (1 sentyabr 2016)

Possum alqoritmi Google –un lokal reyting filtridir. Bu alqoritmin tətbiqindən sonra Google –da axtarış edən istifadəçinin fiziki koordinatlarına və sorğuda istifadə edilən ifadələrə bağlı olaraq daha geniş çeşiddə nəticələr sıralanırdı. Possum bundan əlavə olaraq həm də informasiya-axtarış sisteminin optimallaşdırılmasından istifadə edən veb saytlara təsir edirdi. Fərz edək ki, hər hansısa bir veb sayt informasiya-axtarış sistemini optimallaşdıraraq öz saytını nəticələrdə ən üstə yerləşdirsə belə bu fərqli şəhərlərdə axtarış edildiyi zaman işə yaramır. Çünki Possum alqoritmi istifadəçinin koordinatlarına uyğun olaraq ona münasib nəticələri təqdim edir.

10. *“Fred” alqoritmi* (8 mart 2017)

Google –un Fred alqoritminin vəzifəsi veb saytlardan süni şəkildə gəlir əldə etmək istəyənlərə mənfi təsir göstərməkdir. Google –un Fred yeniləməsindən təsirlənən bir çox zəif məzmunlu, keyfiyyətsiz və həddən artıq reklamlardan istifadə edən saytlar oldu. Qısacası, AdSense və ya digər reklam gəlirləri əldə etmək üçün yaradılan, istifadəçilərin maraqlarına uyğun olmayan veb saytlar hədəf kimi qəbul edilmişdir. Fred alqoritminin tətbiqindən sonra istifadəçilərin yararına olmayan veb

saytlar sıralamadan çıxarılaraq bundan sonra təqdim edilməməyə başladı. Bu alqoritmin əsas diqqət yetirdiyi kriteriyalar aşağıdakılardır:

- İstifadəçilərə daha keyfiyyətli xidmət təklif edən saytları müəyyənləşdirmək və sıralamada onları yuxarıya qaldırmaq.
- Kifayət qədər yüksək keyfiyyətli məzmunu sahib veb saytları təyin etmək və onları önə çəkmək.
- Lazımsız reklamlardan istifadə edən saytları müəyyən edib onlara tədbir görmək.

11. “Maccabees” alqoritmi (12 dekabr 2017)

Bu alqoritm bir çox veb saytın trafik həcminə ciddi təsir etmişdir. Yenilənmənin açar sözləri çox olan, lazımsız açılış səhifələrinə sahib olan saytlara təsir etdiyi müşahidə edilmişdir. Google illərdir öz alqoritmində davamlı olaraq dəyişikliklər etməkdədir, ancaq Maccabees alqoritmi kiçik dəyişikliklər çərçivəsində görülə biləcək bir yeniləmə deyil. Çünki bu yeniləmə mərkəzi alqoritmə əsaslı dəyişikliklər etdi. Əsas alqoritm əsasən aşağıdakı kimi işləyir:

1. Verilmiş sorğu ilə səhifənin əlaqəlilik dərəcəsini müəyyən edir.
2. İnformasiya-axtarış sisteminin nəticələr səhifəsindəki sıralamanın linklərdən necə təsir gördüyünü müəyyənləşdirir.
3. Hər hansı bir veb səhifədə olan məzmunun dəyərini müəyyən etdir.

Maccabees alqoritmi, açar sözlərin müxtəlif variantlarının həddən artıq istifadə olunduğu veb saytlara təsir etdi. Bu cür saytlarda bənzər məzmun üçün bir neçə açılış səhifəsi olur, lakin məzmunun başlığı və açar sözlər bir-birindən fərqlidirlər. Zəif məzmunlu, çox sayda reklamlara sahib bu veb saytlar Maccabees yeniləməsindən zərər çəkmişdirlər. Çünki Maccabees alqoritmi ilə Google, daha yüksək keyfiyyətə malik veb səhifələri irəliyə çəkmək məqsədi daşıyır.



Şəkil 10. Yandex informasiya-axtarış sisteminin loqosu

Yandex Rusiya tərəfindən yaradılmış informasiya-axtarış sistemidir. Rusdilli ölkələrdə informasiya-axtarış sistemləri üzrə bazar payının 60% - dən çoxu Yandex –in üzərinə düşür. Yalnız Rusiya üçün deyil, həmçinin Ukrayna, Qazaxıstan və Türkiyə üçün də nəzərdə tutulmuşdur. Yandex sistemindəki nəticələrin sıralama alqoritmı daim yenilənir. Onlara yeni funksiyalar əlavə edilir, mövcut xətlər aradan qaldırılır, istifadəçilərin tələblərinə uyğun hala salınır. Yandex şirkəti hazırladıqları alqoritmlərə şəhər adları verməklə məşhurdur. Yandex –in alqoritmləri və onlarla birlikdə gələn yeniliklərə nəzər salaq [35]:

1. *“Magadan” alqoritmi* (16 may 2008)

Magadan axtarış alqoritmı Yandex üçün inqilab idi. Bu alqoritmə sayt daxili unikallığın vacibliyi nəzərə alınmışdı. Sıralama zamanı reytingin hesablanmasında iki qat artım var idi. Latin əlifbası ilə daxil edilmiş rus sözlərini anlamaq üçün əlavələr edilmişdi. Digər maraqlı xüsusiyyətlərindən biri də artıq Yandex mətnləri daha yaxşı anlamağa çalışması və açar sözlərin kombinasiyalarından ibarət olmasalarda sıralamanı yetirə bilməsi idi. Magadan həmçinin ingilis dilində olan veb saytları indeksləməyə və sıralamağa çalışan alqoritm idi.

2. *“Naxodka” alqoritmi* (11 sentyabr 2008)

Magadan alqoritmindən 6 ay sonra Naxodka alqoritmı yaradıldı. Bu alqoritm köməkçi nitq hissələrini (ədat, bağlayıcı və.s) anlamaq və nəticələri sıralamaqda istifadə üçün nəzərdə tutulmuşdu. İnformasiya-axtarış sisteminin tezarus lüğəti

genişləndirilmişdi, bu da sözlərin daha aydın başa düşülməsinə və sıralanmasına şərait yaradırdı.

3. “Arzamas” alqoritmi (10 aprel 2009)

Arzamas, Yandex-ə yerli axtarışları gətirən ilk alqoritm idi. Bu alqoritmə bütün axtarış sorğularının iki növə təsnif edilməsi əsas götürülmüşdür: coğrafi asılı sorğular və coğrafi müstəqil sorğular. Əlavə olaraq, Arzamas pop-up və pop-under reklamı olan səhifələrdə çox yavaşlayan ilk sıralama alqoritmi idi. Bu alqoritm vasitəsilə Yandex rus dilini daha yaxşı anlamağı öyrənmişdir. Bu alqoritmə 19 əsas regiona görə axtarış nəticələri ayrılırdı. Yəni, eyni axtarış sorğusu verildiyi zaman 19 regionun hər birində təqdim olunan nəticələr fərqlənirdi. Arzamas nəticələri regionlara görə sırlayaraq təqdim edirdi.

4. “Snezhinsk” alqoritmi (17 noyabr 2009)

Bu alqoritmə çoxsaylı sıralama faktorları tətbiq edildi. Bu alqoritmə gələn əsas dəyişikliklər bunlardır:

1. Əlavə regional faktorlar əlavə edildi
2. MatrixNet texnologiyası (maşın təlimi) tətbiq edildi
3. Orijinal məzmunu sahib olan saytlar axtarış nəticələrində daha yuxarıda yer tutmağa başladı
4. AGS filtrləri əlavə edildi. Bu filtrlərin tətbiqi nəticəsində aşağı keyfiyyətli məzmunu sahib veb saytlar nəticələrin sıralanmasında aşağıda yer tutmağa başladı.

5. “Konakavo” alqoritmi (22 dekabr 2009)

Yandex –in bu yeni alqoritmi Snezhinsk –in yenilənmiş versiyası idi. Onun tətbiqi ilə axtarış nəticələri sadəcə 19 bölgə üçün yox, ölkədaxili 1250 şəhər üçün fərqli olaraq təqdim edilirdi.

6. *“Obninsk” alqoritmi* (13 sentyabr 2010)

Obninsk–in Yandex–ə gətirdiyi əsas fərq axtarış sistemlərinin optimallaşdırılması sahəsində oldu. Rusiyadakı istifadəçilər tərəfindən verilən sorğularda nəticələrin sıralanmasında yaxşılaşdırmalar, veb proqramçılar tərəfindən süni olaraq artırılan səhifə reytinginə limit qoyulması və digər yaxşılaşdırmalarda bu alqoritmın tətbiqi ilə reallaşdırıldı.

7. *“Krasnodar” alqoritmi* (15 dekabr 2010)

Krasnodar alqoritmi fərqli intellektual texnologiya olan “Spectrum” texnologiyasından istifadə edirdi. Spectrum istifadəçinin axtarış niyyətini anlamağa, çoxmənalı sözlər üçün daha düzgün nəticələri əldə etməyə imkan verirdi. Yandex Krasnodar alqoritmi ilə sorğuları müxtəlif xüsusiyyətərə görə təsnifləşdirir, onları obyektlərdən ayırır və istifadəçinin istək kateqoriyasını təyin edirdi. Əlavə sorğu sözləri Spectrum –a səhifələrin reytinglənməsində kömək edirdi. Təqdim edilən nəticələrin keyfiyyəti əvvəlkilərə nisbətə yüksəldilmişdi.

8. *“Reykjavik” alqoritmi* (17 avqust 2011)

Reykjavik istifadəçilərin axtarış niyyətləri anlamaqda növbəti addım idi. Bu alqoritmın köməyi ilə Yandex istifadəçinin dil seçimini öyrənməyi bacarır və bununla əlaqəli olaraq fərqli dillərdə olan sorğulara düzgün nəticələr təqdim edirdi. Bu alqoritm əsasən xarici dillər üçün axtarışlarla bağlı idi.

9. *“Kaliningrad” alqoritmi* (12 dekabr 2012)

Kaliningrad Yandex istifadəçilərinə fərdiləşdirilmiş axtarışların tətbiqi yeniliyini gətirdi. Bütün axtarış nəticələrinin 75-80% -i hər istifadəçi üçün Google –dan fərqli olaraq fərdi olmuşdur. Yəni sorğunun nəticəsi istifadəçilərin maraq dairəsi əsasında sıralanırdı və hər istifadəçinin sorğusuna müxtəlif nəticələr təqdim edilirdi. Yandex bunun üçün istifadəçinin əvvəlki axtarış sorğularını nəzərdən keçirir, ona uyğun olaraq nəticələri sıralayırdı.

10. “*Minusinsk*” *algoritmi* (15 aprel 2015)

Yandex Minusinsk algoritmi ilə informasiya-axtarış sistemlərinin optimallaşdırılması sahəsində müəyyən tədbirlər həyata keçirmişdir. Alqoritmın məqsədi süni reyting artımını təyin etmək və ona qarşı tədbir görməkdən ibarət idi.

11. “*Vladivostok*” *algoritmi* (2 fevral 2016)

Vladivostok alqoritmının tətbiqi ilə mobil axtarış nəticələrində reyting sıralamasında dəyişikliklər oldu. Yandex bu alqoritmə mobil qurğular üçün öz informasiya-axtarış sistemini optimallaşdırdı. Bundan sonra mobil cihazlar və kompüterlə verilən sorğuların nəticələri arasında fərqliliklər yarandı. Vladivostok –un hazırlanmasında əsas məqsəd axtarış nəticələrinin mobil cihazlar üçün adaptasiya edilməsi idi.

12. “*Palekh*” *algoritmi* (2 noyabr 2016)

Yandex –in Palekh algoritmi neyron şəbəkəsinə əsaslanırdı. O 1500 faktora əsasən səhifə reytinginin hesablanmasını təmin edirdi. Bu alqoritm istifadəçi sorğularını, səhifə başlıqlarını emal edərək özü öyrənməyə çalışırdı. Neyron şəbəkəsinin tətbiqi Yandex –in internetdəki məzmunları daha rahat anlamasına, istifadəçilərə konkret nələrin lazım olduğunu müəyyənləşdirməyə kömək edirdi.

13. “*Korolyov*” *algoritmi* (22 avqust 2017)

Korolyov yenilənmiş dərin neyron şəbəkə əsasında hazırlanmış alqoritmdir. O sələfi Palekh –in yeni versiyası idi. Korolyov alqoritmının sələfindən əsas üstünlüyü onun internetdəki bütün mətnlərin daxilindəki yazıları anlama bilməsi idi. Əvvəlki versiya olan Palekh algoritmi isə sadəcə başlıqları anlama bilirdi. Bunun müqabilində Korolyov alqoritmı sadəcə açar sözlər ilə yox həmçinin sözlərin mənaları ilə də axtarış edə bilir. Korolyov alqoritmı Yandex –in sahib olduğu ən müasir və effektiv alqoritmdir. Bu alqoritm istifadəçi sorğularını və nəticələri informasiya-axtarış

sisteminin verilənlər bazasında toplayaraq emal edir, buradan yeni nəticələr çıxarır və bununlada daima özünü inkişaf etdirir. Məhz elə bu səbəblə də Yandex informasiya-axtarış sisteminin sorğu nəticələri hər keçən gün dahada yaxşılaşır. Buna istifadəçilərdən toplanan informasiyalarında böyük təsiri vardır. Çünki bu alqoritm insanlarla əlaqədə olaraq özünü inkişaf etdirir [27].

Korolyov alqoritmi istifadəçilərin niyyətini daha yaxşı anlamaq üçün milyardlarla axtarış sorğusunu və yüklənən veb səhifələrini təhlil edir. Daha sonra semantik şəbəkə yaradaraq bu verilənləri dəyərlənirir. Bu dəyəərə müəyyən rəqəmlər mənimsədir. Bu alqoritm Yandex –in MatrixNet texnologiyasından yararlanaraq bir sıra amilləri nəzərdən keçirdikdən sonra istifadəçiyə sorğuları təqdim edir. MatrixNet –in daxilində çalışan və səhifə reytinginin hesablanmasında əsas rol oynayan isə TİC (thematic index citation) –dir. TİC veb səhifələrə 0-10.000 intervalında olan ədədlərdən birini mənimsədir.

3.4. Müasir informasiya-axtarış sistemləri arasında fərqlər, axtarış nəticələrinin müqayisəsi

Müasir informasiya-axtarış sistemləri bazar payına görə ölkələr üzrə fərqli nəticələr göstərir. Yandex Rusiya bazarında lider mövqeni tutarkən, dünya miqyasında 5-ci sıranı tutur. Baidu Çinin 1 nömrəli informasiya-axtarış sistemi olsa da, o da dünya miqyasında 4-cü sıranı tutur. Təbii ki, bazar payına regionların və istifadəçi sayının böyük təsiri var. Bu səbəblə Google Rusiya və Çin kimi dövlətlərdə 1-ci sıranı tutmasa da dünya miqyasında o lider mövqeyini qoruyur [29].

İnformasiya-axtarış sistemləri üzrə statistik məlumatları “gs.statcounter.com” saytından əldə etmək mümkündür. Bu saytdakı statistik məlumatlar aşağıdakı kimidir:



Şəkil 11. 2017-2018 illəri ərzində dünya miqyasında statistika

Bundan əlavə olaraq qeyd etmək lazımdır ki, Rusiyada ən çox istifadə olunan informasiya-axtarış sistemləri arasında sıralama belədir: Yandex 53.67% , Google 42.96%, Mail.ru 2.25%, Bing 0.45%, Yahoo! 0.35% təşkil edir.

Çində bu göstəricilər müvafiq olaraq: Baidu 73.02%, Shenma 10.9%, Haosou 7.9%, Sogou 4.46%, Google 1.89%, Bing 1.37% təşkil edir.

Ölkəmizdə isə bu göstəricilər müfaviq olaraq : Google 94.13%, Yandex 2.36%, Yahoo! 1.69%, Mail.ru 0.66%, Bing 0.56% təşkil edir.

Yuxarıdakı statistik məlumatlardan da göründüyü kimi informasiya-axtarış sistemlərinin istifadə həcmi regionlar üzrə fərqlilik təşkil edirlər. Bu səbəblə dünya miqyaslı ən geniş istifadə olunan informasiya-axtarış sistemlərini nəzərdən keçirmək, onların müsbət və mənfi cəhətlərini müəyyən etmək vacibdir.

Google informasiya-axtarış sistemi xüsusilə istifadə etməkdə olduğu Google+ sosial integrasiyasıdan əlavə olaraq etibarlı axtarış nəticələri ilə tanınır. Google informasiya-axtarış sistemi online reklam texnologiyaları, cloud computing və Gmail, ofis paketləri, anlıq mesajlaşma və digər proqram xidmətləri təklif edən Google Corporation –a aiddir. Şirkət, 1998-ci ildə Sergey Brin və Larry Page tərəfindən informasiyaların nizamlanmasını təşkil etmək və həm əlçatan həm də faydalı hala

gətirmək məqsədi ilə yaradılmışdır. Google şirkəti öz informasiya-axtarış sistemini illər ərzində onu daha çox sosial yönümlü axtarış sisteminə çevirən yeni əlavələrlə inkişaf etdirmişdir. Google vasitəsi ilə informasiya axtarmaq, informasiyaları başqaları ilə bölüşmək və digər istifadəçilərlə ünsiyyət qurmaq mümkündür. Mükəmməl xüsusiyyətlərə sahib olmasına baxmayaraq, Google mükəmməl deyil və qismən onunla rəqabət etməyə çalışan bir çox informasiya-axtarış sistemləri mövcuddur.

Üstün cəhətləri:

- ✓ Google, rahat istifadə edilə bilən, lazım olan bütün informasiyaları tapıb oxumaq üçün sadə və modern dizayna malikdir.
- ✓ İnkişaf etmiş tərcümə sistemi olan “Google Translate”-ə sahibdir.
- ✓ Google, yeni məzmunun müəyyənləşdirilməsi və indeksləşdirilməsini dayanmadan davam etdirdiyinə görə, daha böyük indeks bazasına malikdir.
- ✓ Digər informasiya-axtarış sistemləri ilə müqayisədə yüksək keyfiyyətli axtarış nəticələrinə malikdir. Bunun səbəbi, 1 gün əvvəl nəşr olunan yeni bir səhifəni tapa bilməsi üçün qabaqcıl indeksləşdirmə alqoritmlərindən istifadə etməsidir.
- ✓ İstifadəçiyə axtarış zamanı sözlərin müxtəlif formalarını və məşhur axtarışlar nəticələrini təklif etməklə kömək edir.
- ✓ Sosial integrasiya xüsusiyyətləri, alış-veriş etmək, qiymətlərin müqayisə edilməsi və digər istifadəçilərin şərhlərini oxumaq kimi ictimai fəaliyyətlər üçün mükəmməl informasiya-axtarış sistemidir.
- ✓ Ətraflı axtarış sorgularının və ani önizləmə vasitələrinin istifadəsi, aşağı keyfiyyətli veb-saytlardan və ya spamlardan qaçınmaqla keyfiyyətli nəticələr əldə etməyə imkan verir.
- ✓ Smartfonlar üçün tətbiqi mövcuddur.
- ✓ Google Scholar, Google Docs, Google Earth kimi əlavə tətbiqləri mövcuddur.
- ✓ Hazırda 123 fərqli dili dəstəkləyir.

Mənfi cəhətləri:

- Sosial integrasiya yalnız Google sosial platformaları üçündür və bu səbəblə daha az faydalı ola bilər.
- İstifadəçilər haqqında çoxsaylı informasiyaların toplanması.
- Bir çox program təminatı google hesabına giriş tələb edə bilər ki, bu da istifadəçilərdə narazılıq yarada bilər.
- Google -un şəkil axtarışı digər axtarışları kimi həssas deyildir.

Bing informasiya-axtarış sistemi Microsoft Corporation –a məxsusdur və həmin şirkət tərəfindən idarə olunur. Daha yeni olduğu üçün Google -dan daha kiçikdir. Bing 2009-cu ildə yaradılmışdır. Bing -in indeksləşdirmə alqoritmi Google-dan çox fərqlidir və daha yaxşı video, şəkil və xəbərlər axtarmaq xüsusiyyətlərinə sahibdir.

Onun idman, maliyyə, tərcümə, uçuşun izlənməsi, pul vahidlərin çevrilməsi, qabaqcıl hesablama və kalkulyatorlar və bu kimi inkişaf etmiş xüsusiyyətləri mövcuddur. Bing əlavə olaraq lokal məlumatlar verən bir xüsusiyyətdə malikdir. Bu xüsusiyyət trafik yenilikləri, bizneslə məşğul olan iş yerlərinin siyahıları, yerli restoran siyahıları, şəhərdəki otel siyahıları kimi məlumatları əhatə edir [43].

Bing-in ən böyük üstünlüyü öz sosial şəbəkələrindən əlavə başqa sosial media platformaları ilə də integrasiya olunmasıdır. Bu integrasiyalara Hotmail, Facebook, Apple və Windows integrasiyaları daxildir. Böyük xüsusiyyətlərinə baxmayaraq, Bing hələ də yeni informasiya-axtarış sistemidir və inkişaf etməsi üçün bir az vaxt lazımdır.

Üstün cəhətləri:

- ✓ Cəlbedici interfeysə malikdir.
- ✓ Daha keyfiyyətli nəticələr əldə etmək üçün regiona, veb səhifəyə və ya açar sözə görə nəticələri filtrləyəbilməyiniz üçün inkişaf etmiş axtarış xüsusiyyətinə sahibdir. Daha dəqiq nəticələr əldə etmək üçün cari yerinizdən də istifadə edə bilərsiniz.

- ✓ Arxa fon şəklinin gündəlik dəyişməsi.
- ✓ İnkişaf etmiş filtrlər və ya inkişaf etmiş sorğu operatorları – istifadəçilərə, şəkillərin ölçüsü, rəngli və ya ağ-qara olması kimi xüsusiyyətlərə əsaslanaraq axtarış nəticələrini dəqiqləşdirməyə imkan verir.
- ✓ Videoların axtarışında videonun uzunluğu, ekran ölçüsü, piksel və mənbənin təyini üçün tənzimləmələrə malikdir.
- ✓ Paketlərin izləməsi xidməti – istifadəçi daşınma şirkətinin adını və izləmə nömrəsi daxil etdikdə, Bing birbaşa izləmə məlumatlarını verəcəkdir.
- ✓ Windows 8, Windows 10 və Microsoft Edge -ə inteqrasiyası.
- ✓ Şəkillərin axtarılmasında müxtəlif parametrlərə görə xüsusi axtarışlara imkan verən əlavələr, daha mükəmməl şəkil axtarışının mövcud olması.

Mənfi cəhətləri:

- Daha yeni olduğu üçün Google kimi geniş indeks bazasına malik deyil.
- Google kimi rahat istifadə imkanı yoxdur.
- İstifadəçilərin məlumatlarının bazada saxlanması.
- Hələlik sadəcə 40 fərqli dili dəstəkləyir.
- Xəbər axtarışlarını ölkələrə görə filtləmək asan deyil.
- Şəkillə axtarış etmək Google və Yandex –dən fərqli olaraq yalnız URL ünvanları vasitəsilə mümkündür.
- Axtarış nəticələrinin digər informasiya-axtarış sistemlərinə nəzərən zəif olduğu məlumdur.

Baidu Çində ən geniş istifadə edilən informasiya-axtarış sistemidir. Bu informasiya-axtarış sistemi 2000-ci ildə yaradılmışdır. Çin dilində Google məhdud axtarış nəticələri verə bilər, çünki Google sırf çin dilində fəaliyyət göstərmək üçün nəzərdə tutulmayıb. Bu səbəblə Baidu sistemi yaradıldı. Kifayət qədər qısa bir müddətdə ölkədəki lider informasiya-axtarış sisteminə çevrildi. Buna səbəb onun sırf çinli istifadəçilər üçün dizayn edilməsi oldu.

Baidu həmçinin, Google-un artan biznes və şəxsi istifadə proqramlarına oxşar əlavə xidmətlər təqdim edir. Bunlara Baidu Baike, onlayn ensiklopediya, Baidu Musiqi, onlayn musiqi proqramı və Baidu Map xəritə xidmətlərini misal olaraq göstərmək olar. Dünyanın 6-cı ən böyük internet şirkəti və 660 milyondan çox aktiv istifadəçiyə sahib olan Çinin ən böyük 3-cü şirkətidir.

Üstün cəhətləri:

- ✓ Çinli istifadəçilər üçün daha geniş axtarış nəticələri təqdim edir.
- ✓ Şəkillə axtarış xidməti kifayət qədər yaxşıdır.
- ✓ Çində Google –u əvəz edə biləcək potensialı mövcuddur.
- ✓ Müxtəlif əlavə tətbiq və integrasiyalara sahibdir.

Mənfi cəhətləri:

- Yalnız çinli istifadəçi kütləsinə xidmətmət məqsədilə yaradıldığına görə dünyada az tanınır və başqa ölkələrdə istifadə edilmir.

Yahoo ən qədim informasiya-axtarış sistemlərindən biridir. Yahoo informasiya-axtarış sistemi Google kimi olmasa da Google ilə rəqabət aparmasına kömək edən bir çox funksiya sahibdir. Yahoo informasiya-axtarış sistemi David Filo və Jerry Yang tərəfindən 1995-ci ildə qurulan Yahoo Inc şirkətinə məxsusdur. İnformasiya-axtarış sistemindən başqa, Yahoo şirkəti veb portal, Yahoo news, Yahoo mail, Yahoo finance və Yahoo video paylaşma kimi xidmətlər təklif edir [42].

Üstün cəhətləri:

- ✓ Nəticələr Bing informasiya-axtarış sistemində olduğu qədər genişdir.
- ✓ Alış-veriş axtarışlarında digər informasiya-axtarış sistemlərindən daha çox variant təklif edir.
- ✓ Yahoo –nun mobil telefonunuza yükləyə biləcəyiniz mobil tətbiqi mövcuddur.
- ✓ Şəkil axtarışı zamanı müxtəlif kriteriyalara görə axtarışın təşkili.

- ✓ Ən son xəbərlərin birbaşa əsas səhifədə təqdim edilməsi.

Mənfi cəhətləri:

- Alış-veriş xüsusiyyətləri təcrid olunmuş haldadır və istifadəçini əsas axtarış sistemindən uzaqlaşdırır.
- Google və Bing kimi şəkil ilə axtarışın mümkünsüzlüyü.
- Həddindən artıq məlumatın əsas ekranda olması istifadəçi üçün çaşqınlıq yarada bilər. Bu baxımdan istifadəçilər əsasən sadə və rahat interfeysə malik sistemlərə maraq göstərirlər.

Yandex Rusiyada 20 il əvvəl İlya Segalovich və Arkady Volozh adlı iki rus proqramçı tərəfindən yaradılmışdır. Rusiyadan əlavə olaraq Türkiyə, Ukrayna, Qazaxıstan və Belarusiyada fəaliyyət göstərən şirkətdir [44].

İstifadəçilərini internetdə informasiya axtarmaqla təmin edən informasiya-axtarış sistemi və informasiya portalıdır. Yandex bunlardan əlavə xəbərlər, e-poçt, xəritə, oyun kimi bir çox müxtəlif onlayn xidmətlər də təqdim edir. Yandex bazar payına görə Rusiyada informasiya-axtarış sistemləri bazarında lider mövqeyini saxlayır. Öz antivirus və anti-spam texnologiyaları ilə istifadəçilərini zərərli proqramların olduğu saytlar haqqında xəbərdar edir.

Üstün cəhətləri:

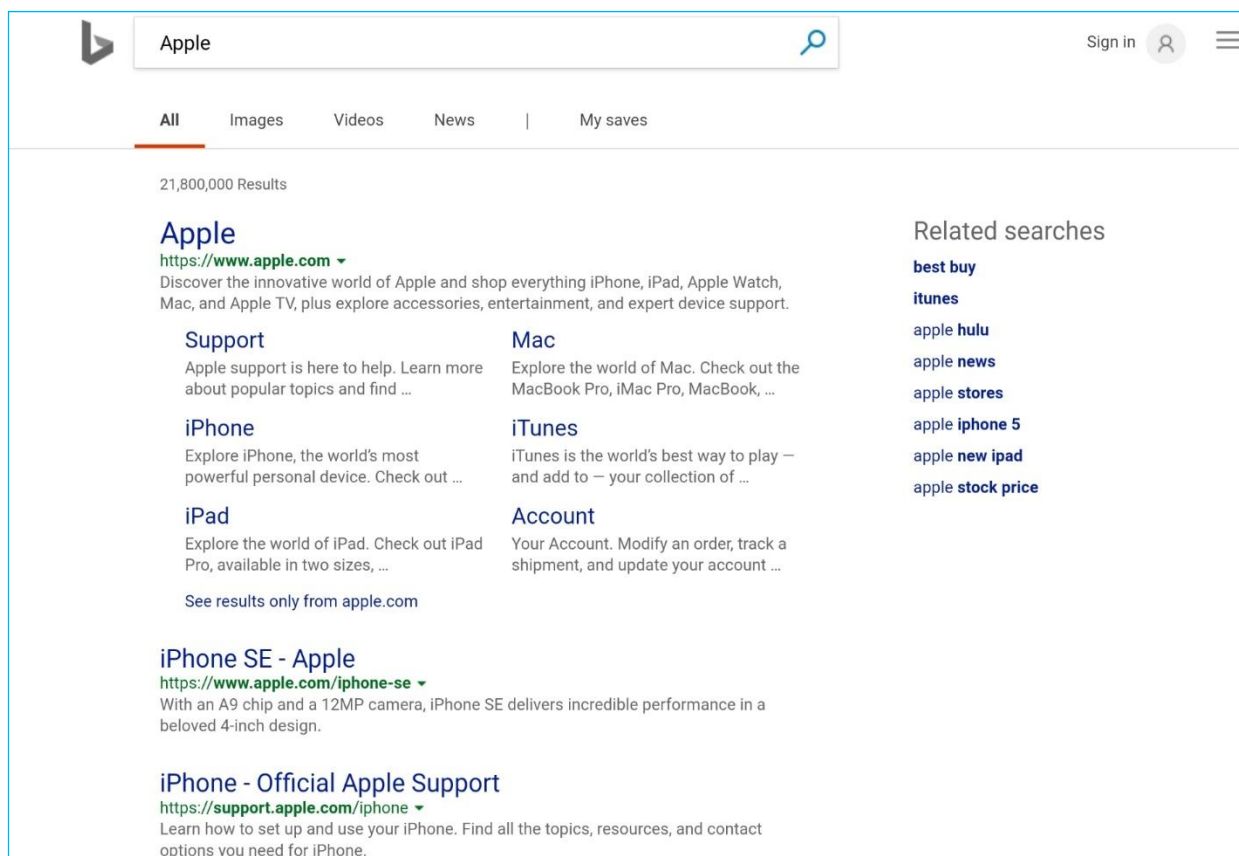
- ✓ Rusdilli ölkələr üçün daha yaxşı axtarış nəticələri verməsi.
- ✓ Nəticələri regionlara görə fərqli sıralaması
- ✓ İstifadəçi məlumatları ilə daima özünü inkişaf etdirməsi
- ✓ Dil problemlərinə Yandex tərəfindən daima yaxşılaşdırmalar davam etdirilir. Məsələn, istifadəçi kiril hərfləri ilə ingilis dilində olan sözləri yazı bilər və ya tərsinə rus dilində olan sözləri latın hərfləri ilə yazı bilər. Yandex kompleks axtarışları daha yaxşı başa düşür və buna görə daha çox nəticə verir.

- ✓ Regional axtarış nəticələri daha münasibdir. Google –dan fərqli olaraq, Yandex –də regional xəbərlər haqqında dəqiq məlumatı tapmaq ehtimalı daha yüksəkdir.
- ✓ Poçt, nəqliyyat, hava proqnozu, bulud hesabı üçün Yandex Disc, xəritələr, xəbərlər kimi Yandex xidmətlərinin integrasiyası. Bütün bunlar əsas səhifədə yerləşir və istifadəçi bir kliklə onlardan yararlanı bilər.
- ✓ Şəkillə axtarış etmək yararlı və daha əlaqəlidir.
- ✓ Fərqli ölkələr üçün xüsusi olaraq hazırlanmışdır.

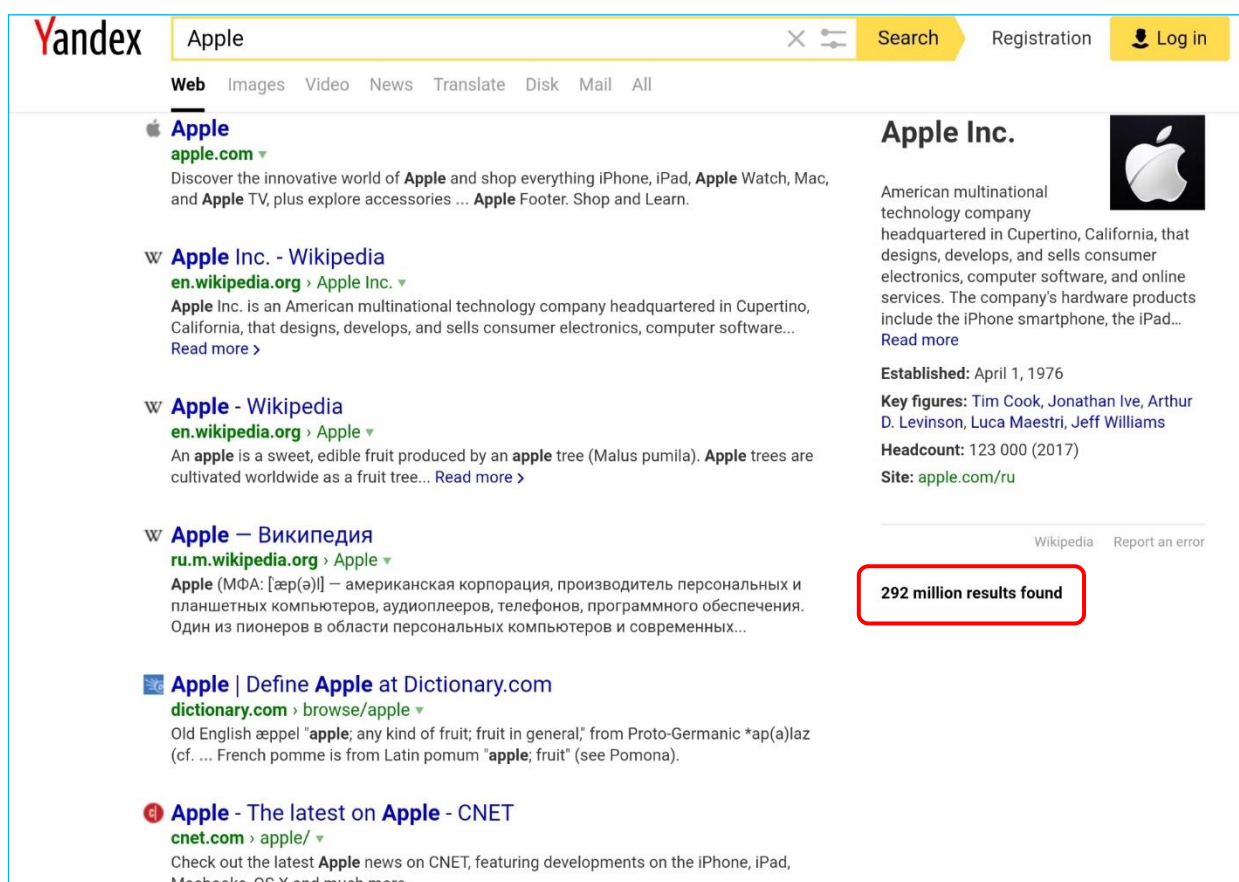
Mənfi cəhətləri:

- Yandex –də qeydiyyatdan keçən istifadəçilərin məlumatlarının toplanması.
- Qlobal olaraq tanınmamasına görə digər informasiya-axtarış sistemlərinə nəzərən az istifadəçi sayı.
- İstifadəçilər üçün maraqlı olmayan interfeysə sahib olması.
- Google –a nisbətə axtarış robotları və indeksləşdirmə modulunun sürətinin yavaş olması.

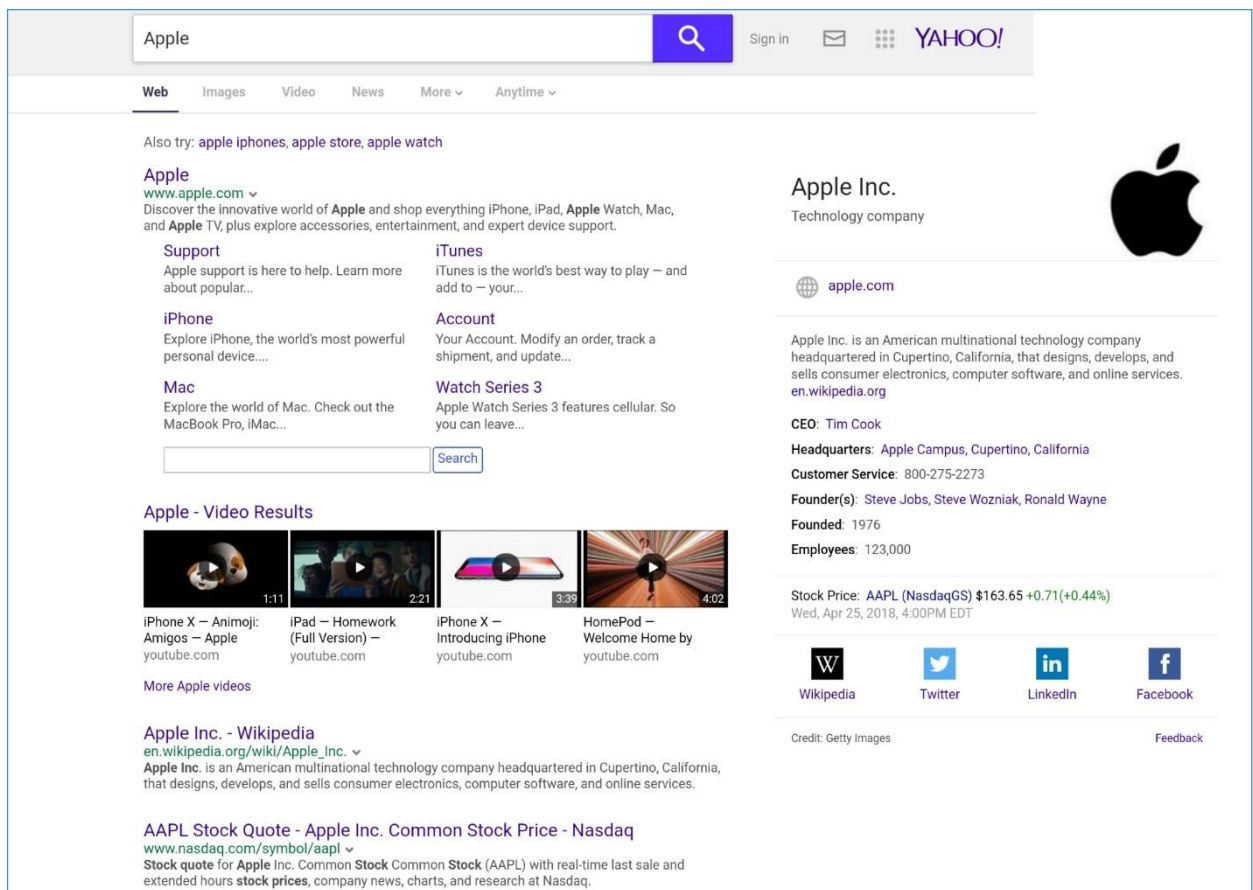
Dünya miqyaslı ən çox istifadəçi sayına malik olan informasiya-axtarış sistemlərinin yuxarıda sadalanan müsbət və mənfi tərəflərini görmək üçün onların hər birinə eyni açar sözü sorğu olaraq vermək lazımdır. Daha sonra isə nəticələrə əsasən onlardan hansıların daha münasib olduğuna qərar vermək olar. Bu məqsədlə Bing, Yandex, Yahoo! və Google informasiya-axtarış sistemlərinin axtarış sahəsinə “Apple” sözünü daxil edərək alınan nəticələrə nəzər salaq:



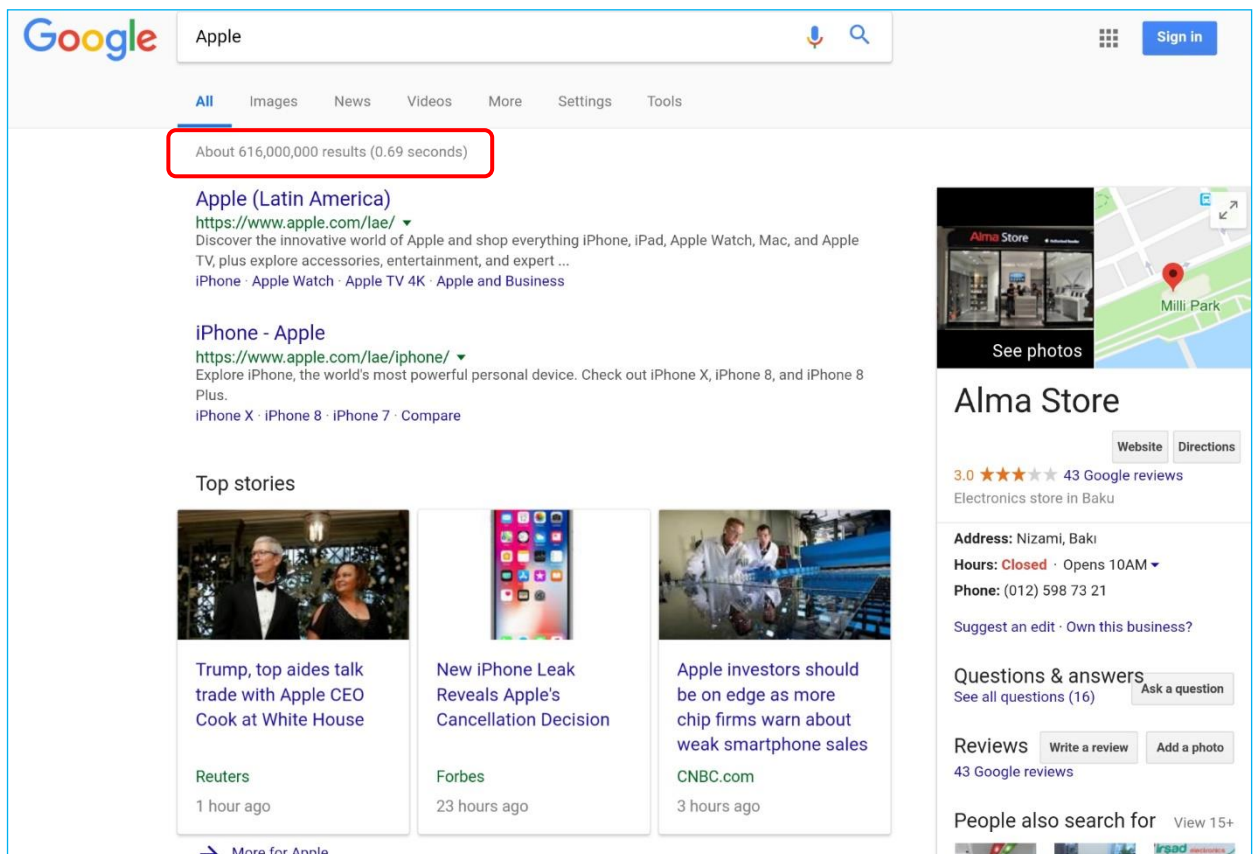
Şəkil 12. Bing –in nəticələr səhifəsi



Şəkil 13. Yandex –in nəticələr səhifəsi



Şəkil 14. Yahoo! –nün nəticələr səhifəsi



Şəkil 15. Google–un nəticələr səhifəsi

Yuxarıda ən çox istifadə edilən informasiya-axtarış sisteminin nəticələri təqdim olunmuşdur. Şəkillərdən də aydın olduğu kimi nəticələr və onların nizamlanması müxtəliflik təşkil edir. Bing informasiya-axtarış sistemi nəticələri təqdim edərkən ekranın sağ tərəfində istifadəçiyə sorğusunu dəqiqləşdirmək üçün müxtəlif variantlar təqdim edir. Bunlara əlaqəli axtarış nəticələri də deyilir. Bing digər informasiya-axtarış sistemlərinə nəzərən çox sadə axtarış nəticələri təqdim etmişdir. O sadəcə Apple şirkəti və onun məhsullarına fokuslanmış digər ehtimalları nəzərə almamışdır.

Yandex isə Bing -ə nəzərən dolğun nəticələr təqdim etmişdir. O həm Apple şirkəti həm də həmin söz ilə əlaqəli nəticələrə fokuslanmışdır. Çünki Yandex nəzərə alır ki, istifadəçi sadəcə Apple şirkətini nəzərdə tutmaya bilər. Bundan əlavə Yandex, Bing –dən fərqli olaraq “Apple” sorğusuna aid neçə nəticənin tapıldığını istifadəçiyə təqdim edir. Bu funksiya hələlik Google və Yandex –də mövcuddur. Həmçinin Yandex ekranın sağ tərəfində Apple şirkəti haqqında informasiyalar, yaradılma tarixi, işçi sayı kimi müxtəlif məlumatları da istifadəçiyə təqdim etmişdir. Lakin nə Yandex nə də Bing axtarış nəticələrində sorğu ilə bağlı şəkilləri təqdim etməmişdir. İstifadəçi bunun üçün şəkillər bölməsinə keçid etməlidir.

Yahoo! həm Bing həm də Yandex –in nəticələrini birlikdə təqdim etmişdir. O da Yandex kimi ekranın sağ tərəfində Apple şirkəti haqqında müxtəlif məlumatları əks etdirmiş, lakin bu məlumatlar əvvəlkinə nisbətən biraz dolğundur. Yahoo! bundan əlavə, axtarış sözü ilə əlaqəli olan bəzi videoları və müəyyən statistikaları, müxtəlif sosial şəbəkələrə keçidləri də axtarış nəticələrində əks etdirmişdir. Yahoo –da Bing kimi sadəcə Apple şirkətinə fokuslanmış, digər ehtimalları düşünməmişdir. Həmçinin Yandex kimi müvafiq axtarış sözünə uyğun neçə nəticənin tapıldığını və Google kimi axtarışa nə qədər vaxt sərf edildiyini göstərməmişdir.

Google digər informasiya-axtarış sistemlərinə nəzərən olduqca dolğun və dəqiq nəticələr təqdim edərək özünün dünyanın ən yaxşı informasiya-axtarış sistemi olduğunu sübut etmişdir. Google istifadəçinin Apple sözünü axtararkən sözün müxtəlif mənalarda ola biləcəyi ehtimalını nəzərə almışdır. Onun ən böyük üstünlükləri axtarışa sərf olunan vaxt və nə qədər nəticənin tapılması haqda məlumat və istifadəçinin

koordinatlarına nəzərən nəticələrin sıralanması oldu. O nəzərə alıb ki, əgər istifadəçi şirkət və ya mağaza ilə maraqlanırsa bu zaman ona ən yaxın Apple mağazasını tapsın. Google sadəcə mağazanı yox, həm də onun haqqında müxtəlif informasiyaları da təqdim etmişdir. Bütün bu informasiyaları ümumi nəticələrdən ayırmaq üçün onları ekranın sağ tərəfində əks etdirmişdir.

Baxılan dörd informasiya-axtarış sistemi arasında sadəcə Google və Yandex sorğuya uyğun nə qədər nəticə olduğunu göstərmişdir. Google 616 milyon, Yandex isə 292 milyon nəticə arasında sıralama etmişdir. Google 2 dəfədən artıq əlavə nəticələr tapmışdır. Google –un bu qədər çox nəticəyə sahib olmasının səbəbi onun informasiya-axtarış sistemləri arasında ən böyük indeks bazasına sahib olmasıdır. Digər səbəb isə Google –un sorğu və sıralama üçün istifadə etdiyi alqoritmlərin mükəmməl olmasıdır. Onun sorğu və sıralama alqoritmlərinin nəticəsində axtarış sorğusuna sadəcə 0,69 saniyə vaxt sərf olunmuşdur. Yəni Google 616 milyon nəticə arasında seçim edərkən çox qısa bir müddətdə ən əlaqəli nəticələri təqdim etmişdir. Və bununla Google, niyə informasiya-axtarış sistemləri bazarında lider mövqeyində olduğunu bir daha sübut edir.

NƏTİCƏ VƏ TƏKLİFLƏR

İnformasiya-axtarış sistemləri hazırkı informasiya cəmiyyəti üçün mühüm əhəmiyyət kəsb edirlər. İnsanlar informasiya mübadiləsi üçün əvvəla informasiyanın toplanması prosesini həyata keçirirlər. Keçmişdə informasiya əldə etmək indiki qədər asan deyildi. Lakin internetin bu qədər geniş yayılması ilə artıq, informasiya əldə etmə olduqca asanlaşmışdır. İstənilən istifadəçi internetdən istifadə etməklə özünə lazım olan məlumatı informasiya-axtarış sistemləri vasitəsilə saniyələr ərzində əldə etmək imkanına malikdir.

İnformasiya-axtarış sistemlərinin müxtəlif tipləri mövcuddur. Bu tiplərin bəziləri lokal axtarışları həyata keçirərkən, bəzi tipləri də internet vasitəsilə qlobal axtarışları reallaşdırır. İnformasiya-axtarış sistemlərinin ən geniş yayılmış növü veb informasiya-axtarış sistemləridir. Bu sistemlər digər axtarış sistemlərinə nəzərən çoxsaylı istifadəçi sayına və böyük inkişaf potensialına malikdir. Həmçinin, veb informasiya-axtarış sistemləri daima inkişaf edən sahələrdən biridir. Məhz bu səbəbdən də dissertasiya işində əsas yer veb informasiya-axtarış sistemlərinin araşdırılmasına verilmişdir.

Veb informasiya-axtarış sistemlərinin bir çox növləri mövcuddur. Bunların ən populyar nümayəndələrinə Google, Yandex, Bing, Yahoo, Baidu kimi axtarış sistemlərini misal göstərmək olar.

Tədqiqatın nəticəsində əldə olunan biliklərə əsaslanaraq demək olar ki, bütün informasiya-axtarış sistemlərinin mərkəzi komponentləri oxşardırlar. Lakin sual meydana çıxa bilər ki, bəs niyə eyni axtarış sorğusuna hər bir informasiya-axtarış sistemi fərqli nəticələr təqdim edir? Çünki burada əsas rolu informasiya-axtarış sistemlərinin tərkibi hissələrində istifadə olunan alqoritmlər oynayır.

Hər bir informasiya-axtarış sistemi veb axtarış robotlarından, indeksləşdirmə modulundan, sorğu və sıralama modulundan, verilənlər bazasından və istifadəçi interfeysindən ibarətdir. Bu modullar hər biri informasiya-axtarış sistemi üçün mühüm əhəmiyyət kəsb edir.

İnternetin tərkibinin daima genişləndiyini nəzərə alsaq, veb axtarış robotlarının əhəmiyyətini xüsusilə qeyd etmək yerinə düşərdi. Çünki, hər saniyədə internetə

minlərlə fayl və informasiyalar yüklənir. İnformasiya-axtarış sistemi həmin fayl və informasiyaları istifadəçiyə təqdim etmək üçün onları əvvəlcədən görməlidir. Bu işi veb axtarış robotları yerinə yetirirlər. Onlar daima dayanmadan bütün interneti gəzərək veb sayt, formalar, veb səhifələr, sosial platformalar kimi geniş və daima yenilənən informasiya bazalarını müşayət edirlər. Axtarış robotları internetdə olan bu informasiya bazalarına daxil olur, ordaki məlumatları anlamağa çalışır, məlumatların yeni olub-olmamasını aydınlaşdırır, əgər yeni informasiya əldə edərsə bu informasiyaları indeksləşdirmə moduluna göndərir. Axtarış robotları bu əməliyyatları dayanmadan təkrar edirlər. Bu məqsədlə onlar fərqli alqoritmlərdən istifadə edirlər, bəziləri isə süni intellektə əsaslanan müxtəlif texnologiyalardan yararlanırlar. Təbii ki, bu kimi informasiyalar şirkət sirri olduğu üçün cəmiyyətə təqdim olunmur. Bu səbəblə dissertasiya işi məhz informasiya-axtarış sistemlərinin arxitekturasının tədqiqinə həsr olunmuşdur və olduqca aktual məsələlərdən biridir.

Axtarış robotları tərəfindən göndərilən məlumatlar indeksləşdirmə modulu tərəfindən təhlil edilir və daha sonra indeksləşdirilir. İndeksləşdirmənin informasiya-axtarış sistemləri üçün mühüm əhəmiyyəti vardır. Belə ki, indeksləşdirilmiş məlumatlar həm bazada daha az yer tutur, həm də informasiya axtarış sürətini qat-qat yüksəldir. Hər bir informasiya-axtarış sistemi indeksləşdirmə modulunda müxtəlif texnologiya və üsullardan istifadə edir. Daha sonra indeksləşdirilmiş məlumatlar informasiya-axtarış sisteminin verilənlər bazasında saxlanılır.

İstifadəçi tərəfindən sorğu verildiyi zaman sorğu və sıralama modulu indeksləşdirmə moduluna müraciət edərək daxilində açar sözlərin istifadə olunduğu saytların indekslərini oradan götürür. Bu indekslər oxunaraq həmin saytların URL adresləri əldə olunur. Bundan sonra isə əldə olunan nəticələr sıralanaraq istifadəçiyə təqdim edilir. Sıralama informasiya-axtarış sistemi üçün vacib məsələlərdən ən əsasıdır. Çünki, informasiya-axtarış sisteminin müvəffəqiyyəti onun təqdim etdiyi nəticələrdən asılıdır. Bu səbəblə müxtəlif informasiya-axtarış sistemləri sıralama üçün fərqli alqoritmlərdən istifadə edirlər. Məsələn, Google PageRank texnologiyasından istifadə edir, Yandex isə maşın təliminə əsaslanan MatrixNet texnologiyasından

yararlanır. Təbii ki, bu müxtəlifliyin axtarış nəticələrinə böyük təsiri vardır. Google – un PageRank texnologiyası veb saytlara 0 ilə 10 aralığında ədəd mənimsədir. Bu ədədi təyin etmək üçün o səhifəyə baxış sayı, səhifədəki məlumatların yenilənməsi, səhifənin tərkibi, səhifəyə istinad edən veb saytların sayı kimi müxtəlif kriteriyalara əsaslanır. MatrixNet isə daha çox istifadəçinin internetdə nə ilə maraqlandığına fikir verir, əvvəllər axtardığı informasiyalara əsaslanaraq hər istifadəçi üçün sıralamanı fərqli formada edir. PageRank –ın əksinə olaraq MatrixNet daima özünü inkişaf etdirir.

Sorğu və sıralama modulu tərəfindən əldə olunan nəticələr istifadəçi interfeysi vasitəsilə təqdim edilir. İnformasiya-axtarış sisteminin istifadəçi interfeysi də vacibdir. Çünki, istifadəçilər özlərinə rahat olan sadə axtarış sistemlərinə üstünlük verirlər.

Dissertasiya işində informasiya-axtarış sistemlərinin optimallaşdırılması üçün görülən daxili və xarici tədbirlərdən bəhs edilmişdir. Ümumiyyətlə informasiya-axtarış sistemlərinin optimallaşdırılması birbaşa sıralama prosesinə təsir edən tədbirlər kompleksidir. İstənilən veb sayt sahibi bundan istifadə edərək öz veb saytını axtarış nəticələri səhifəsində ilk sıralara süni şəkildə çəkə bilər. Marketing sahəsində bu üsul geniş tətbiq edilmə potensialına malikdir.

Dissertasiya işinin III fəslində informasiya-axtarış sistemlərinin ən məşhur nümayəndələri araşdırılmış, müqayisə edilmiş, axtarış nəticələrində aralarındakı fərqlər müəyyən edilmiş, bu fərqlərin səbəbləri aydınlaşdırılmağa çalışılmış, Google və Yandex kimi iki fərqli sistemin alqoritmləri şərh edilmişdir. Hər bir informasiya-axtarış sistemi əvvəldə də qeyd olunduğu kimi müxtəlif alqoritm və texnologiyalara əsaslanır.

Aparılmış tədqiqatlar nəticəsində aşağıdakı təkliflərin irəli sürülməsini məqsədəuyğun hesab edirik:

- Əvvəla axtarılacaq informasiya və ya faylın növünə uyğun olaraq informasiya-axtarış sistemi düzgün seçilməlidir.
- Açar sözlər axtarılacaq informasiya ilə əlaqəli olmalıdır. Qeyd etmək lazımdır ki, ən yaxşı nəticəni əldə etmək üçün ən azı iki açar sözdən istifadə etmək

məqsədəuyğundur. Açar sözlərin sayının həddən çox olması axtarış nəticələrinə mənfi təsir edə bilər.

- Axtarışın yaxşılaşdırılması üçün müəyyən operatorlardan istifadə etməklə daha əlaqəli nəticələr əldə etmək mümkündür. Bu operatorlar sorğu moduluna təsir göstərir.
- Meta-axtarış sistemləri ilə bir neçə informasiya-axtarış sisteminin nəticələrindən faydalanmaq olar.
- İnformasiya-axtarış sistemlərinin optimallaşdırılması mühüm əhəmiyyətə malikdir. Belə ki, onun marketinqə tətbiqində geniş istifadə potensialı mövcuddur.
- Ən müasir informasiya-axtarış sistemlərinin qarşılaşdırılmasında məlum oldu ki, onların heç biri tamamilə mükəmməl deyildir, lakin nəticələrdə Google rəqiblərinə nisbətə daha yaxşı olduğunu sübut edir. Bu səbəblə informasiya axtarışında Google –un geniş istifadə olunmasını təklif etmək yerinə düşərdi.
- Veb saytların müəyyən robot.txt fayllarından istifadə edərək axtarış robotlarını yönləndirməsi mümkündür.
- İnformasiya-axtarış sistemlərinin optimallaşdırılmasından istifadə etməklə PageRank –a təsir etmək və axtarış nəticələrini süni şəkildə dəyişdirmək olar.

ƏDƏBİYYAT SİYAHISI

- 1) Qasimov V.Ə. İnformasiya axtarışı üsulları və sistemləri. Dərslik. Bakı: MTN-in Maddi-texniki Təminat Baş İdarəsinin Nəşriyyat-Poliqrafiya Mərkəzi. 2015, 288 s.
- 2) Мидоу Ч., Анализ информационно-поисковых систем, М., Мир, 1970
- 3) Маннинг, Кристофер Д. Введение в информационный поиск. // М.: Вильямс, 2011.
- 4) Храмов П. Информационно-поисковые системы Internet // Открытые системы. 1996. № 3.
- 5) Сычев А.В. Информационно-поисковые системы, 2006 г.
- 6) Астрахов, А.В. Технология информационного поиска. 2011.
- 7) Castillo, Carlos (2004). Effective Web Crawling (Ph.D. thesis). University of Chile. Retrieved January 26, 2014
- 8) Gianna M. Del Corso; Antonio Gullí; Francesco Romani (2005). "Fast PageRank Computation via a Sparse Linear System". Internet Mathematics. Lecture Notes in Computer Science. 2 (3): 118–130. doi:10.1007/978-3-540-30216-2_10. ISBN 978-3-540-23427-2. Archived from the original on 2014-02-09.
- 9) Segev, El (2010). Google and the Digital Divide: The Biases of Online Knowledge, Oxford: Chandos Publishing.
- 10) Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry (August 18, 1997). The PageRank Citation Ranking: Bringing Order to the Web. Retrieved April 5, 2014.
- 11) Christopher D. Manning; Raghavan, Prabhakar; Schütze, Hinrich (2008). Introduction to information retrieval. Cambridge, UK: Cambridge University Press. p. 108,109,111. ISBN 0-521-86571-9.
- 12) Shkapenyuk, V. and Suel, T. (2002). Design and implementation of a high performance distributed web crawler. In Proceedings of the 18th International Conference on Data Engineering (ICDE), pages 357-368, San Jose, California. IEEE CS Press.

- 13) C. C. Foster, Information retrieval: information storage and retrieval using AVL trees, Proceedings of the 1965 20th national conference, p.192-205, August 24–26, 1965, Cleveland, Ohio, United States
- 14) Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, Sriram Raghavan (2000). Effective Web Crawling. Stanford University. Retrieved January 26, 2014
- 15) René Pickhardt, 2013. Web Science MOOC – Search Engine Economy
- 16) Spetka, Scott. "The TkWWW Robot: Beyond Browsing". NCSA. Archived from the original on 3 September 2004. Retrieved 21 November 2010.
- 17) Shestakov, Denis (2008). Search Interfaces on the Web: Querying and Characterizing. TUCS Doctoral Dissertations 104, University of Turku
- 18) C. C. Foster, Information retrieval: information storage and retrieval using AVL trees, Proceedings of the 1965 20th national conference, p.192-205, August 24–26, 1965, Cleveland, Ohio, United States
- 19) Chakrabarti, S., van den Berg, M., and Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11–16):1623–1640.
- 20) Tomasic, A., et al.: Incremental Updates of Inverted Lists for Text Document Retrieval. Short Version of Stanford University Computer Science Technical Note STAN-CS-TN-93-1, December, 1993
- 21) C.J. van Rijsbergen, Information Retrieval, London, Butterworths, 1979
- 22) Clarke, C., Cormack, G.: Dynamic Inverted Indexes for a Distributed Full-Text Retrieval System. TechRep MT-95-01, University of Waterloo, February 1995.
- 23) Pant, Gautam; Srinivasan, Padmini; Menczer, Filippo (2004). "Crawling the Web" (PDF). In Levene, Mark; Poullovassilis, Alexandra. *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*. Springer. pp. 153–178. ISBN 978-3-540-40676-1.
- 24) Gleich, David F. (January 2015). "PageRank Beyond the Web". *SIAM Review*. 57 (3): 321–363. arXiv:1407.5107. doi:10.1137/140976649.

- 25) elibrary.bsu.az/books_400/N_406.pdf
- 26) https://en.wikiversity.org/wiki/User:Oleamm/Search_engines_history_and_architecture#Web_Search_Engine
- 27) <https://yandex.com/company/blog/new-intelligent-search-algorithm-korolyov/>
- 28) <http://novikov.com.ua/service/node/6575>
- 29) <https://www.searchenginejournal.com/9-biggest-differences-yandex-vs-google-seo/168628/>
- 30) <https://www.techhelp.ca/blog/google-bing-yahoo-compare/>
- 31) <http://www.celilcan.net/seo-nedir/>
- 32) <https://www.seoamo.net/single-post/2016/06/09/Site-D%C4%B1%C5%9F%C4%B1-SEO-Nedir>
- 33) https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%B8%D1%81%D0%BA%D0%BE%D0%B2%D0%B0%D1%8F_%D0%BE%D0%BF%D1%82%D0%B8%D0%BC%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F
- 34) <https://www.how2shout.com/tools/top-best-web-internet-search-engines-googles-alternative.html>
- 35) <https://www.russiansearchtips.com/2013/03/the-evolution-of-yandex-search-algorithm/>
- 36) <https://www.link-assistant.com/news/google-algorithm-updates.html>
- 37) [https://en.wikipedia.org/wiki/Search_engine_\(computing\)](https://en.wikipedia.org/wiki/Search_engine_(computing))
- 38) https://en.wikipedia.org/wiki/Web_crawler
- 39) https://en.wikipedia.org/wiki/Search_engine_indexing
- 40) <https://en.wikipedia.org/wiki/PageRank>
- 41) https://en.wikipedia.org/wiki/Database_index
- 42) https://en.wikipedia.org/wiki/Yahoo!_Search
- 43) [https://en.wikipedia.org/wiki/Bing_\(search_engine\)](https://en.wikipedia.org/wiki/Bing_(search_engine))
- 44) <https://en.wikipedia.org/wiki/Yandex>
- 45) https://en.wikipedia.org/wiki/Google_Search

Р Е З Ю М Е

Информационно-поисковых систем нынешнее представляют важное значение для информационного общества. Люди осуществляют процессом сбора информации прежде всего, для обмена информацией. В прошлом получении информации столько было нелегко как сейчас. Однако при широком использовании Интернета получение информации стало намного проще. Любой пользователь интернета с использованием себе информационно-поисковых систем имеет возможность получить необходимую информацию посредством за считанные секунды.

Веб информационно-поисковых систем существует множество видов. Из самых популярных представителей как поисковые системы можно привести в пример Google, Yandex, Bing, Yahoo, Baidu. Все информационно-поисковые системы аналогичны центральным компонентам, но каждая информационно-поисковая система предлагает разные результаты для одного и того же запроса. Потому что главную роль здесь играет в состав информационно-поисковых систем, используемые алгоритмы. Такая информация не предоставляется общественности, так как это секрет компании. По этой причине крайне тема диссертации является одним из актуальных вопросов.

Каждая информационно-поисковая система состоит из веб-поисковых роботов, модулей индексирования, модулей запросов и сортировки, базы данных и пользовательского интерфейса. Эти модули представляют важное значение для каждого из информационно-поисковая система.

S U M M A R Y

Information retrieval systems and search engines are important for the current information society. People firstly carry out the process of gathering information for the information exchange. In the past, it was not easy to get the information. However, with the wide use of the Internet, getting the information has become much easier now. Any user can obtain any information through web search engines within seconds.

There are many types of Web search engines: The most popular of these are Google, Yandex, Bing, Yahoo, Baidu.

The central component of all information retrieval systems are similar. But each retrieval system's offer to the same query are different. Main role here plays the part of the algorithms used in a information retrieval system. Such information is confidential. So they are not provided to the public. That's why the subject of the dissertation is actual.

Each information retrieval systems consist of web search engines, indexing module, query engine and ranking module, database and user interface. Each of these modules are important for retrieval system.