Universitatea din Craiova

Facultatea de Automatică, Calculatoare şi Electronică

Proiect S.C.D.
Trasaturile, configuratia si modul de funtionare pentru Google

COORDONATOR:

PROF.DR.ING. Matei Vînătoru

STUDENTĂ:

Stănică Roxana
Master - A.S.C

1. Google - Introducere....3
Căutarea în WEB....3
Câteva sfaturi pentru o căutare eficientă:....6
Utilizatori....7
Administraţie....8
Funcţionarea Google – pe scurt....8
Hardware....8
Cum manevrează Google cererile de căutare?....10
Sistemul PageRank....10
Googlebot-ul....11
Sistemul de fişiere Google....11
Scalabilitate:....12
1.1 Motoare de căutare – Evoluţie: 1994-2000....13
1.2 Google: Evoluţie concomitentă cu Web-ul....13
1.3 Caracteristici ale configuraţiei....14
1.3.1 Capacitate îmbunătăţită de căutare....14
1.3.2. O cercetare academică a motorului de căutare....15
2. Trăsăturile sistemului....15
2.1 PageRank: Ordonarea Web-ului....16
2.1.1 Descrierea unui calcul PageRank....16
2.1.2 Explicarea intuitivă....17
2.2 Textul link-ului....17
2.3 Alte trăsături....18
3. Activităţi similare....18
3.1 Extragerea de informaţii....19
3.2 Diferenţe între Web şi colecţiile bine controlate....19
4. Configuraţia sistemului....20
4.1 Scurtă descriere a arhitecturii Google....21
4.2 Structuri majore de date....21
4.2.1 BigFiles....22
4.2.2 Biblioteca (Repository)....22
4.2.3 Indexul documentelor....22
4.2.4 Lexiconul....23
4.2.5 Listele de hit-uri (Hit Lists)....23
4.2.6 Indexul primar (Forward Index)....24
4.2.7 Indexul complementar....25
4.4 Indexarea Web-ului....25
4.5 Căutarea....27
4.5.1 Sistemul de clasificare....27
4.5.2 Scurta recapitulare....28
5. Rezultate şi mod de funcţionare....30
5.1 Parametri pentru stocare....30
5.2 Modul de funcţionare a sistemului....31
5.3 Procesul de căutare....31
6. Concluzii....32
6.1 Activităţi viitoare....32
6.2 Căutare la standarde înalte....32
6.3 Arhitectura scalabilă....33
Referinţe....34

1. Google - Introducere

Oricât de imparţiali încearcă jurnaliştii să fie, există o tendinţă clară în presă de a critica tot sau aproape tot ce face Microsoft şi de a reda într-o lumină placută şi optimistă realizările Google.

Este greu de spus când a început această modă şi care au fost motivele din spatele ei. Un ochi atent ar observa că până la urmă este vorba de doi giganţi din lumea IT care nu sunt atât de diferiţi, factorul primar care face diferenţa între cei doi fiind vârsta.

Iniţial, Google a reprezentat un proiect de cercetare conceput în 1996 de Larry Page şi Sergey Brin, doi doctoranzi de la Stanford. Cei doi au plecat de la ipoteza că un *motor de căutare* care analizează relaţiile dintre website-uri ar duce la obţinerea de rezultate mai bune decât cele furnizate de tehnicile folosite atunci, şi anume de câte ori se repetă cuvântul cheie pe pagină.

În martie 2001, Eric Schmidt a devenit membru al consiliului de directori, iar în august 2001, a fost numit CEO-ul companiei. Popularitatea noului motor de căutare s-a datorat design-ului aerisit, simplităţii şi acurateţii rezultatelor întoarse, iar numărul de utilizatori a crescut exponenţial. În timp, după succesul motorului de căutare clasic, Google şi-a extins braţele virtuale şi peste imagini, video, ştiri, grupuri, mail, cumpărături, hărţi si multe altele. În plus, există zvonuri că Google ar plănui o reţea Internet proprie în Statele Unite, formată cu ajutorul kilometrilor întregi de fibră optică rămaşi după boom-ul de la sfârşitul anilor '90. Această fibră optică, care străbate suprafeţe întinse ale Statelor Unite poartă numele de "fibra întunecată" (dark fiber).

Căutarea în WEB

Internetul poate fi considerat ca fiind cea mai mare bibliotecă disponibilă şi actualizată permanent. De aceea uneori găsirea informaţiei care ne este necesară poate fi dificilă.

World Wide Web (WWW sau Web) reflectă chiar prin numele sau trăsăturile sale esenţiale :

- este destinat căutării informaţiei în întregul Internet (world wide = în lumea întreagă)
- foloseşte hipertextul pentru organizarea informaţiei, ceea ce face ca aceasta să apară ca o pânză de păianjen (web) permiţând navigarea cu uşurinţă de la o pagină la alta.

Pentru regăsirea informaţiei în Web se pot folosi următoarele instrumente:

- **serviciul de navigare** – browserul, care permite accesarea informaţiilor prin introducerea adresei de web a paginii (numită URL - Uniform Resource Locator – identificator standard al locului în care se găseşte

resursa), sau prin urmărirea unei hiperlegături (*en*. hyperlink) dintr-un alt document (care conţine URL-ul paginii respective)

- **serviciul de căutare automată** – prin:
    - instrumente de căutare (*search tool*)
    - instrumente de indexare (*indexing tool*)
    - motor de căutare (*search engine*)

Browserul permite, în general, efectuarea următoarelor operaţii:

- navigarea printre paginile web şi vizualizarea lor.
- urmărirea legăturilor dintre documente care conţin hiperlegături.
- copierea informaţiilor din Internet pe propriul calculator.
- căutarea informaţiei în Internet.
- regăsirea rapidă a informaţiilor prin folosirea „semnelor de carte” („pagina favorită”) şi a istoricului.
- accesarea altor servicii Internet: poşta electronică, mesagerie instantanee etc.

Aşadar, browserul integrează accesul la mai multe servicii din Internet printr-o interfaţă accesibilă şi uşor de utilizat.

O alternativă mai rapidă pentru „răsfoirea” paginilor web în căutarea informaţiei dorite este apelarea la un ***serviciu de căutare***, adică un site web care conţine în general următoarele categorii de informaţii:

- instrucţiuni care arată modul în care trebuie folosit serviciul
- metoda prin care utilizatorul poate să propună un subiect pentru căutare

Portalurile sunt site-uri specializate care îndeplinesc funcţia cataloagelor dintr-o bibliotecă. Aceste site-uri aranjează pe categorii sau domenii diverse site-uri existente în Internet şi le ordonează în funcţie de anumite criterii în cadrul categoriilor stabilite (pe subiecte, după popularitate, etc.). De multe ori, portalurile oferă şi alte servicii pe lângă cel de căutare (poştă electronică, ştiri etc.).

Motoarele de căutare sunt site-uri care au rolul de a ajuta utilizatorul să găsească mai uşor şi mai direct informaţia în Internet. De cele mai multe ori, în cuprinsul unui site cu rol de căutare, se regăsesc ambele funcţionalităţi – indexare în catalog şi meniu de căutare. Motorul de căutare este în fapt o aplicaţie care „răsfoieşte" paginile web din Internet în căutarea cuvintelor sau frazelor cerute de utilizator. Pentru aceasta sunt folosite nişte programe automate care alcătuiesc liste de cuvinte din interiorul site-urilor. Rezultatele căutării sunt afişate în funcţie de relevanţa stabilită de motorul de căutare, utilizând indexarea termenilor din aceste liste.

De aceea, termenii căutaţi trebuie să fie cât mai definitorii pentru subiectul în cauză (*keywords* – „cuvinte-cheie").

Pentru a limita aria de căutare, se recomandă utilizarea modului de căutare avansată („*advanced search*"), care permite găsirea mai rapidă a informaţiei datorită criteriilor multiple de căutare.

De exemplu, ***Google*** afişează doar paginile care includ *toţi* termenii căutării. Nu este necesar să includem "and" între termeni. Pentru a rafina căutarea, se adaugă doar alţi termeni, iar rezultatele vor conţine un subset specific al paginilor întoarse de către cererea originală.

Putetem exclude un cuvânt din căutare prin scrierea semnului minus ("-") imediat înaintea termenului pe care vrem să îl evităm. De asemenea, se pot căuta fraze prin includerea acesteia între ghilimele.

Câteva sfaturi pentru o căutare eficientă:

1. Fiţi cât mai exact. Printr-o interogare precisă, se obţin mai puţine rezultate şi conţinutul relevant este mai uşor de găsit. De exemplu, dacă am căutat cuvântul *lege*, am obţinut aproximativ **3.580.000** de rezultate. Pentru căutarea *lege drept autor* am obţinut **12.000** de rezultate, în timp ce pentru fraza exactă **"lege privind dreptul de autor"** am obţinut **30** de rezultate.
2. Nu folosiţi cuvinte uzuale. Utilizaţi cuvinte cât mai adecvate subiectului căutat, altfel utilitarul de căutare va returna zeci de pagini web cu informaţii irelevante pentru dumneavoastră. Vezi diferenţa între **5.570.000** rezultate pentru **masă** şi **4.960** pentru **mijloace comunicare masă.**
3. Învăţaţi să adaptaţi interogarea. Dacă interogarea dumneavoastră returnează prea multe rezultate, restrângeţi aria de căutare. Dacă rezultatele returnate nu sunt suficiente, reformulaţi-o într-un mod mai general. Nu întotdeauna primele cuvinte -cheie sunt şi cele mai bune.
4. Folosiţi diferite forme ale cuvintelor. Puteţi utiliza diferite cuvinte care se referă la subiectul căutat pentru a obţine cât mai multe informaţii relevante pentru dumneavoastră.
5. Folosiţi sinonimele. De exemplu, scrieţi şi „alergare” şi ”jogging”. Dacă folosiţi un utilitar de căutare care acceptă combinaţii de cuvinte - cheie, separaţi sinonimele prin cuvântul cheie OR
6. Folosiţi citate între ghilimele. În cazul când căutaţi o anumită frază sau un titlu, plasaţile între ghilimele (de ex. „Internet pentru începători”) în formularul utilitarului de căutare.
7. Folosiţi majuscule atunci când este necesar. Majoritatea directoarelor şi indexurilor de căutare fac diferenţa între literele mici şi literele mari din

şirul de caractere căutat. Dacă textul introdus conţine numai litere mici, utilitarul de căutare va identifica numai textul scris fie cu litere mari, fie cu litere mici. În cazul în care scrieţi şi o litera mare, utilitarul de căutare presupune ca aceasta are o semnificaţie specială şi va afişa numai rezultatele care corespund exact şirului respectiv.

8. Aflaţi secretele utilitarului de căutare folosit. Unele utilitare de căutare oferă facilităţi speciale prin care conţinutul relevant poate fi găsit mai uşor.

Google este cel mai mare motor de căutare şi datorită parteneriatelor sale cu Yahoo!, Netscape şi altele este capabil să răspundă la mult mai multe interogări decât orice alt serviciu online asemănător.

- Interogări soluţionate în fiecare zi: peste 150 de milioane
- Pagini web parcurse: peste 2,4 miliarde.
- Tipuri de fişiere căutate
- HyperText Markup Language (html)
- Adobe Portable Document Format (pdf)
- Adobe PostScript (ps)
- Lotus 1-2-3 (wk1, wk2, wk3, wk4, wk5, wki, wks,θ wku)
- Lotus WordPro (lwp)θ
- MacWrite (mw)θ
- Microsoft Excel (xls)θ
- Microsoft PowerPoint (ppt)
- Microsoft Word (doc)
- Microsoft Works (wks,θ wps, wdb)
- Microsoft Write (wri)
- Rich Text Format (rtf)
- Text (ans,θ txt)
- Imagini: peste 330 milioane
- Mesaje Usenet: peste 700 de milioane

Utilizatori

Google.com este unul din primele zece dintre cele mai populare site-uri de pe Internet şi este folosit de oameni din lumea întreagă

- Media lunară a numărului de utilizatori: 55.6 milioane (Nilesen/NetRatings 5/02)
- Gradul de utilizare atât acasă, cât şi la birou: #4 (Nilesen/NetRatings 5/02)
- Limbile pentru care Google asigură o interfaţă: 81

- Limbile în care Google oferă rezultate: 35
- Utilizatori: mai mult de jumătate din traficul Google.com provine din afara USA.

Administraţie

Personalul Google include profesionişti cu experienţă în domeniul tehnologiei, iar compania este susţinută de fonduri provenite de la două firme ce se ocupă cu investiţiile de risc.

- Număr aproximativ de angajaţi: 400
- Deţinători ai titlului de doctor: peste 50

Proiectarea unui motor de cătare este o provocare. Motoarele de căutare indexează zeci sau chiar sute de milioane de pagini web, implicând un număr echivalent de termeni distincţi. Acestea raspund la zeci de milioane de întrebări în fiecare zi. Deşi importanţa acestor motoare de căutare pe web este mare, totuşi ele nu au constituit subiectul unei cercetări academice amănunţite. În plus, datorită gradului rapid de avansare a tehnologiei şi dezvoltării continue a web-ului, metoda de creare a unui motor de căutare este foarte diferită de cea folosită acum trei ani. Această lucrare oferă o descriere amanunţită a motorului de cautare.

În afara problemelor de măsurare a capacităţii motoarelor tradiţionale de căutare până la a putea suporta o cantitate importantă de date, există noi provocări tehnice corelate cu utilizarea informaţiilor adiţionale prezente în hipertext, cu scopul producerii de rezultate mai bune. Lucrarea de faţa pune aceasta întrebare, cum să construieşti un sistem practic de măsurare care poate folosi informaţia adiţională din hipertext. De asemenea, luăm în considerare modul de tratare efectiv al colecţiilor de hipertext, care nu pot fi în întregime controlate şi unde fiecare este liber să publice ceea ce doreşte.

Funcţionarea Google – pe scurt

Hardware

Pentru a putea oferi capacitate suficientă de service, structura fizică a Google-lui este alcătuită din *clustere* şi *computere* situate peste tot în lume, cunoscute sub numele de *ferme de servere*. Aceste ferme de servere sunt alcătuite dintr-un număr mare de computere de nivel jos pe care rulează sisteme bazate pe *Linux* care operează cu GFS (Sistemul de fişiere Google), iar cele mai mari dintre aceste ferme au peste 1000 de *noduri* de stocare şi peste 300 de *TB* de spaţiu de stocare pe disk.

S-a speculat că Google are cel mai mare computer din lume. S-a estimat că Google are în jur de:

- 899 de *rack-uri*
- 79, 112 maşini
- 158,244 Unităţi centrale de prelucrare (UCP)
- 316,448 GHz putere de procesare
- 158,224 Gb de RAM
- 6,180 Tb de spaţiu pe Hard

Clustere-le sunt folosite pentru a face ca 2 sau mai multe calculatoare să apară ca unul singur pentru o comunitate de utilizatori. Custerele sunt folosite pentru a oferi fiabilitate crescută, şi/sau pentru a creşte performanţa de pe un singur calculator.

Un cluster este un grup de calculatoare legate între ele care lucrează împreună ca un computer paralel. Una dintre cele mai populare implementări o reprezintă cluster-ul cu noduri pe care rulează Linux-ul ca şi sistem de operare şi software-ul Beowulf pentru a implementa paralelismul.

Microsistemele Sun şi-au lansat de asemenea un produs gen cluster numit Grid engine.

De asemenea denumită şi cluster de servere, fermă de computere sau ranch, o fermă de servere este un grup de servere legate prin reţea care sunt găzduite într-o singură locaţie. O fermă de servere modernizează procesele interne prin distribuirea lucrului între componentele individuale ale „ fermei” şi expediază procesele computaţionale prin folosirea puterii mai multor servere. „Ferma” se bazează pe software-ul „ load-balancing” care realizează sarcini cum ar fi urmărirea cererii pentru procesarea puterii de la diferite maşini, stabilind prioritatea temelor şi programându-le şi reprogramând-le în funcţie de prioritatea şi cererea pe care utilizatorii o pun pe reţea. Când un server din „ fermă” eşuează, un altul poate interveni ca şi back-up.

Un nod este un dispozitiv care este conectat ca şi parte a unei reţele de calculatoare. Nodurile pot fi computere, personal digital assistants (PDAuri), telefoane mobile sau variate dispozitive ale reţelei. Pe o reţea IP, un nod este orice dispozitiv cu o adresă IP. Nodurile sunt deseori conectate prin centrale, routere sau printr-un switch de reţea.

Linux-ul este un sistem de operare gratuit de tip Unix, creat cu originalitate de către Linus Torvalds cu ajutorul unor programatori din jurul lumii. Creat sub licenţa GNU General Public, codul sursă pentru Linux este disponibil gratuit pentru toată lumea.

Un rack de 19 inch este un sistem standard pentru încărcarea unor module electronice variate într-o stivă sau un „ grilaj” , cu o lărgime de 19 inch.

Cum manevrează Google cererile de căutare?

Când un user introduce o întrebare în căsuţa de căutare la Google.com, aceasta este transmisă în mod aleator la unul dintre clustere-le lui Google. Întrebarea va fi apoi manevrată în mod exclusiv de către cluster. Un *echilibrator de încărcare* monitorizează cluster-ul, apoi împrăştie cererea către serverele din cluster ca să asigure că încărcarea de pe hardware este egală.

Echilibratorul de încărcare are nevoie de informaţii de profil pentru indica cum să re-mapeze procesele şi canalele de comunicare pentru a optimiza performaţa. Acesta va fi un ehilibrator-optimizator static care va lua informaţia din uneltele de profil şi va produce echilibrare de încărcare pentru viitoarele rulări.

Căutarea propriu-zisă are loc în 2 faze:

- În prima fază, cuvintele din cerere sunt verificate cu o listă din index-urile serverelor care conţin detaliile de potrivire ale documentelor. Sistemul PageRank este angajat într-un set relevant de documente, ce sunt identificate de trimiteri la alte referinţe ale cuvintelor şi care vor determina scorul pentru fiecare document, dar la întoarcere afectează poziţia documentului pe pagina rezutatelor.
- Rezultatele de la serverele indexate sunt apoi trimise la servere de documente în formularul de identificatori de documente. Rezidentă în serverele document este o copie a lui *World Wide Web*, de unde se ia rezumatul conţinutului paginii web apoi le reîntoarce către server. Rezultatul este apoi procesat şi este redat înapoi ca un document *HTML* care poate fi afişat pe *browser*-ul user-ului ca *pagină web*.

Sistemul PageRank

PageRank, denumit astfel după Larry Page, care l-a inventat, este unul dintre modurile prin care Google determină importanţa paginii, care, la reîntoarcere, va decide unde anume pe lista de rezulate va apărea rezultatul.

Algoritmul exact a lui PageRank, aşa cum este extras din “The Anatomy of a Large-Scale Hypertextual Web Search Engine” este în felul următor:

Presupunem că pagina *A* are paginile *T1...Tn* care pointează spre ea. Parametrul *d* este un factor, care poate fi setat între 0 şi 1. De obicei îl punem pe *d* pe valoarea 0.85. Există mai multe detalii despre *d* în secţiunea următoare. De asemenea, *C(A)* este definit ca fiind numărul de link-uri care „ies” din pagina *A*. RankPage-ului paginii *A* este redat în felul următor:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

A se nota că PageRank-urile dintr-o probabilitate a distribuţiei peste paginile web este făcută astfel încât suma PageRank-urilor a tuturor paginilor web va fi una.

Googlebot-ul

Serverele Google sunt populate de către *web crawler*-urile Google-ului, Googlebot, care se mişcă de pe site pe site pe internet, făcând download a unor copii a paginilor web şi salvându-le în index-ul Google-ului (cunoscut sub numele de cache) pentru viitoarele refetinţe. Googlebot-ul însăţi este rezident pe multe calculatoare care accesează simultan pe sute de pagini web. Acesta emulează un browser, aşa că, mulţi dintre webmasteri îşi vor da seama că atunci când vizitează, acesta lasă un „marcaj" în secţiunea de browser a protocolului site-ului web mai degrabă decât în secţiunea de „păianjen" decât cele mai multe web crawler-uri care se înregistrază.

Un web crawler (cunoscut şi sub numele de web spider) este un program care deschide World Wide Web într-o manieră metodică, automată. Web crawlerii sunt în pricipal folosiţi pentru a crea o copie a tuturor paginilor web vizitate pentru procesări ulterioare de către un motor de căutare, care va indexa paginile download-ate pentru a oferi căutari mai rapide.

Există două metode pe care Googlebot le foloseţte pentru a găsi o pagină web: ori prin ajungerea la pagina web după ce „s-a târât" prin link-urile acesteia, ori ia pagina după ce a fost înscrisă de către webmaster. Prin înscrierea link-ului principal Googlebot va trece prin toate link-urile care se găsesc în indexul paginii şi în fiecare pagină următoare, până când întregul site a fost indexat.

Sistemul de fişiere Google

Sistemul de fişiere Google este un sistem propriu de administrare a fişierelor dezvoltat de către Sanjay Ghemawat, Shun-Tak Leung şi Urs Holyle pentru Google, ca un mijloc de a manevra un număr masiv de cereri peste un număr mare de clustere ale serverelor.

Sistemul a fost creat la fel ca majoritatea celorlalte *sisteme distribuite* de fişiere, pentru performanţă maximă, pentru a face faţă unui număr mare de utilizatori, *scalabilitate*, pentru a fi capabil să manevreze expansiunile inevitabile, siguranţă, pentru a asigura maximul de *uptime* şi *disponibilitate*, pentru a asigura faptul că, toate calculatoarele sunt disponibile pentru a trata întrebările.

Scalabilitate:

1. Cuvânt des folosit care se referă la felul în care un sistem hardware sau software se poate adapta pentru a creşte cererile.

2. Se referă la orice la care i se poate schimba mărimea

3. Când este folosit pentru a descrie un sistem al computerului, proprietatea de a rula mai mult de un procesor.

Uptime este o măsură a timpului de când un computer de sistem a fost pornit. A fost creat pentru a descrie antonimul cuvântului downtime, adică timpul cât un sistem a fost non-operaţional.

Disponibilitatea este o măsură a cât de mult timp o reţea sau o conexiune a rulat. În general, formula este: Timpul rulării/Timpul măsurat (timpul de rulare împărţit la timpul măsurat). Deşi, dacă se măsoară ceva pentru 20 de minute şi acesta a fost disponibil doar 19 din acestea, disponibilitatea va fi de 95%.

Din cauza deciziei lui Google de a folosi un număr mare de computere de nivel jos în loc să folosească un număr mai mic de sisteme de tip server, sistemul de fişiere Google a trebuit să fie creat astfel încât să poată trata eşecurile sistemului, să efectueze monitorizarea constantă a sistemelor, detectarea de erori, *toleranţa erorilor* şi recuperarea automată.

Toleranţa erorilor este disponibilitatea unui sistem de a răspunde unui eşec neaşteptat a dispozitivelor hardware sau software. Există multe nivele de tolerare a erorilor, cel mai jos fiind amabilitatea de a continua operaţile în cazul unei căderi de tensiune. Multe dintre sistemele de computere care tolerează erorile monitorizează toate operaţiile, aceasta este în cazul în care, fiecare operaţie este efectuată pe două sau mai multe sisteme duplicate, aşa că în cazul în care unul dintre acestea eşuează celălalt poate prelua controlul.

Aceasta însemna că, toate clustere-le ar trebui să reţină replici multiple a informaţiei create de web crawler-ii Google-lului. Aceasta este foarte relevant pentru Gmail care a fost recent implementat pe Google, unde email-ul personal al utilizatorilor trebuie să aibă backup pentru a preveni pierderea informaţiilor.

Din cauza mărimii *bazei de date* a Google-ului, sistemul a trebuit creat ca să poată trata fişiere enorme de mai mulţi giga-bytes totalizând mulţi terabytes ca şi mărime. A fost creat astfel datorită faptului că stocarea fişierelor în kilobytes ar însemna să existe miliarde de fişiere, care s-ar putea dovedi imposibil de manevrat. O altă metodă angajată să trateze fişiere de mărime enormă este că schimbările (cunoscute şi sub numele de mutaţii) sunt mai degrabă adăugate decât să aibă fişiere peste care s-a rescris, astfel se minimizează timpul de acces la fişiere.

GFS este un sistem propriu aşa că aplicaţiile software pot fi construite de obicei în jurul lui, asigurând faptul că Google are control maxim asupra sistemului, în acelaşi timp permiţând sistemului să rămână flexibil.

Web-ul creează noi provocări pentru obţinerea de informaţii. Cantitatea de informaţii de pe web creşte într-un ritm alert, pe măsura numărului de noi

utilizatori lipsiţi de experienţă în arta căutarii pe web. De obicei, oamenii navighează pe web folosind graficul acestuia de link-uri, adeseori începând cu indici superiori calitativ, mentinuţi de intervenţia umană, cum ar fi Yahoo sau cu motoare de căutare. Listele unde intervine mintea umană acoperă subiecte diverse şi populare, dar sunt subiective, costisitoare de intreţinut şi menţinut, greu de imbunătăţit şi nu pot acoperi toate subiectele ce ţin de domenii specializate. Motoarele de căutare automate, care se bazează pe potrivirea de cuvinte-cheie, oferă, în mod obişnuit, prea multe rezultate neconcludente. Pentru a înrăutăţi situaţia, unii agenţi publicitari încearcă să câstige atenţia prin diferite metode destinate să inducă în eroare motoarele automate de căutare. Ei au construit un motor de căutare pe scară largă care are răspuns pentru multe din problemele sistemului existent. Acesta utilizează, în special, structura adiţională din hipertext pentru a oferi rezultate concludente. Au numit acest sistem Google deoarece este un cuvânt asemănător cu googol sau 10100 şi corespunde cel mai bine intenţiei lor de a construi motoare de căutare pe scară largă.

1.1 Motoare de căutare – Evoluţie: 1994-2000

Tehnologia motoarelor de căutare a fost obligată să se reinventeze în mod constant pentru a putea ţine pasul cu expansiunea web-ului. În 1994, unul dintre primele motoare de căutare, World Wide Web Worm (WWWW), avea un index de 110,000 pagini web şi documente accesibile pe web. În noiembrie 1997, motoarele de căutare cele mai performante realizau indexarea de la 2 milioane (WebCrawler) până la 100 de milioane de documente web (din Search Engine Watch). Este previzibil că, până în anul 2000, un index complet al Web-ului să conţină peste un bilion de documente. În acelaşi timp, numărul interogărilor la care pot face faţă motoarele de căutare a crescut într-un ritm incredibil de rapid. În martie şi aprilie 1994, World Wide Web Worm primea în jur de 1500 de întrebări pe zi. În noiembrie 1997, Altavista pretindea că rezolva aproximativ 20 de milioane de interogări pe zi. Odată cu creşterea numărului de utilizatori ai web-ului şi sistemelor automate care puneau întrebări motoarelor de căutare, este posibil ca motoarele performante de căutare să poată raspunde la sute de milioane de întrebări pe zi până în anul 2000. Scopul sistemului este să rezolve multe dintre aceste probleme, atât din punct de vedere al calităţii, cât şi al scalabilităţii, coordonate introduse de tehnologia motoarelor de căutare care au atins cifre extraordinare.

1.2 Google: Evoluţie concomitentă cu Web-ul

Crearea unui motor de căutare care să ţină pasul cu web-ul zilelor noastre este subordonată multor provocări. Tehnologia rapidă de parcurgere este necesară pentru a ţine o evidenţă a documentelor web şi pentru a le actualiza în

permanenţă. Spaţiul de stocare trebuie folosit în mod eficient pentru organizarea indicilor şi, opţional, chiar a documentelor. Sistemul de indexare trebuie să proceseze cât mai bine sute de gigabiţi de date. Întrebările trebuie să aibă un răspuns într-un interval cât mai scurt, la o rată de sute până la mii pe secundă.

Aceste sarcini devin din ce în ce mai dificile pe măsură ce Web-ul se dezvoltă. Totuşi, performanţele hardware şi costurile au crescut şi ele în mod constant pentru a rezolva parţial această problemă. Există, bineînţeles, câteva excepţii importante de la acest progres cum ar fi timpul de căutare pe disc şi performanţele sistemului de operare. În proiectarea sistemului Google, s-au luat în considerare atât rata de evoluţie a Web-ului, cât şi schimbările tehnologice. Google este construit în aşa fel încât să parcurgă eficient mari cantităţi de date. În acelaşi timp, Google utilizează eficient spaţiul de stocare pentru organizarea indexului. Structurile sale informaţionale sunt optimizate pentru accesul rapid şi concludent (vezi secţiunea 4.2). Se speră că preţul indexării şi stocării de text sau HTML să devină în cele din urmă mai mic în raport cu cantitatea de date ce va fi disponibilă. Acest lucru va genera caracteristici mai performante de organizare pentru sistemele centralizate ca Google.

1.3 Caracteristici ale configuraţiei

1.3.1 Capacitate îmbunătăţită de căutare

Scopul principal a fost acela de a îmbunătăţi calitatea motoarelor de căutare. În 1994, existau păreri că un index complet de căutare va face posibilă găsirea rapidă a oricărui răspuns. În opinia Best of the Web 1994 – Navigators, “Cel mai bun sistem de navigare ar trebui să faciliteze găsirea unui răspuns la aproape orice întrebare pe Web (din momentul introducerii datelor)”. Totuşi, Web-ul anului 1997 este diferit. Oricine a utilizat recent un motor de căutare, poate oricând să spună că nu complexitatea indexului este singurul factor care asigură calitatea rezultatelor căutării. Rezultatele neconcludente (junk results) estompează de obicei relevanţa rezultatelor de care un utilizator este într-adevar interesat. De fapt, din noiembrie 1997, numai unul dintre cele 4 motoare de căutare comerciale îşi menţionează numele (returnează pagina proprie de căutare ca răspuns la numele sau în topul primelor 10 rezultate). Una dintre principalele cauze ale acestei probleme este aceea că numărul documentelor din indici a crescut progresiv acoperind diferitele grade de importanţă, în timp ce modul de percepere al utilizatorului asupra acestor documente nu s-a schimbat. Oamenii manifestă în continuare interes doar pentru primele 10 rezultate. Din acest motiv, pe măsura ce colecţia de date se măreşte, este nevoie de instrumente cu un grad ridicat de precizie (adică număr de documente relevante returnate, de exemplu în topul primelor 10 rezultate). Într-adevar, se vrea ca noţiunea de “relevant” să includă caracteristica de cel mai important ignorând chiar căutarea

în sine (numărul total de documente relevante pe care sistemul este capabil să le ofere. Există chiar un sentiment de optimism care se referă la utilizarea mai multor informaţii hipertextuale ce pot contribui la îmbunătăţirea căutării şi la alte aplicaţii. În particular, structura şi textul link-ului oferă multe informaţii pentru efectuarea de analize relevante şi filtrări de calitate. Google foloseşte atât structura link-ului cât şi textul acestuia (vezi sectiunile 2.1 si 2.2).

1.3.2. O cercetare academică a motorului de căutare

În afară de o creştere considerabilă, Web-ul a adoptat şi o caracteristică comercială de-a lungul timpului. În 1993, 1,5% din serverele web erau pe domenii .com. Numărul acestora a atins 60 de procente în 1997. În acelaşi timp, motoarele de căutare au părăsit academicul în favoarea comercialului. Până în prezent dezvoltarea majorităţii motoarelor de căutare a fost efectuată de companii care aveau prea puţin de-a face cu detaliile tehnice. Acest fenomen a determinat rămânerea în umbră a tehnologiei motoarelor de căutare şi orientarea acestora către comercial. Cu Google, scopul este acela de a orienta dezvoltarea şi perceperea acestora către sfera academicului.

O altă trăsătură importantă a acestei configuraţii era construirea de sisteme pe care mulţi oameni să le poată utiliza. Modul de utilizare era important pentru deoarece unele dintre cele mai interesante cercetări implică manipularea unei vaste cantităţi de informaţii care sunt disponibile pe sistemele web moderne. De exemplu, zilnic sunt iniţiate zeci de milioane de căutari. Totuşi, toate aceste date sunt foarte dificil de obţinut, în special pentru că se consideră că au valoare comercială.

Scopul final al acestei configuraţii era realizarea unei arhitecturi care să poată susţine activităţi inedite de căutare pe scară largă a informaţiei pe web. Pentru a sprijini aceste activităţi de căutare, Google stochează toate documentele în format comprimat pe care le parcurge. Unul dintre scopurile principale ale configuraţiei Google a fost acela de a stabili un mediu unde ceilalţi căutători puteau intra repede, de a procesa secţiuni considerabile ale web-ului şi oferi rezultate interesante care ar fi fost greu de oferit prin adoptarea unei alte metode. În intervalul scurt de timp în care sistemul a fost construit, existau deja câteva lucrări care foloseau bazele de date oferite de Google şi multe altele sunt realizate acum. O altă este aceea de a constitui un mediu asemănător cu un laborator spaţial unde cercetătorii sau chiar studenţii să poată propune şi chiar executa experimente interesante folosind informaţiile complexe ale web-ului.

2. Trăsăturile sistemului

Motorul de căutare Google este caracterizat de două trăsături importante care ajută la producerea de rezultate cu un grad ridicat de precizie. În primul

rând, Google se foloseşte de structura de link-uri a Web-ului pentru a calcula un indice calitativ al fiecarei pagini web. Această estimare a nivelului calitativ se numeşte *PageRank* În al doilea rând, Google utilizează link-urile pentru a îmbunătăţi rezultatele căutării.

2.1 PageRank: Ordonarea Web-ului

Graficul de link-uri al web-ului este o resursă importantă care a rămas în mare parte neutilizată de motoarele de căutare. S-au realizat hărţi conţinând nu mai puţin de 518 milioane din aceste hyperlink-uri, o mostră semnificativă a totalului. Aceste hărţi permit calcularea rapidă a PageRank-ului unei pagini web, o măsura obiectivă a importanţei link-urilor care corespunde cu ideea subiectivă de importanţă a oamenilor. Datorită acestei corespondenţe, PageRank reprezintă o metodă excelentă de stabilire a gradului de importanţă a rezultatelor căutărilor bazate pe cuvinte cheie. Pentru cele mai populare subiecte, un text simplu care se potriveşte cu căutarea şi care este limitat la titluri ale paginii web este foarte bine reprezentat atunci când PageRank stabileşte importanţa rezultatelor (demonstraţie disponibilă la google.stanford.edu). Pentru căutările ce au la bază un text integral în sistemul principal Google, PageRank este, de asemenea, de mare ajutor.

2.1.1 Descrierea unui calcul PageRank

Literatura de specialitate referitoare la link-uri a fost raportată la web, în general prin numerotarea link-urilor sau backlink-urilor unei pagini date. Acest lucru stabileşte cu aproximaţie importantă sau calitatea unei pagini. PageRank extinde această idee nu prin efectuarea unei numerotări a link-urilor din toate paginile, ci prin stabilirea numărului de link-uri dintr-o pagină. PageRank este definit după cum urmează:

Presupunem că pagina A este formată din paginile $T_1 \ldots T_n$ care se referă la aceasta (adică sunt link-uri). Parametrul d este un factor de nivelare care se află între 0 şi 1. De obicei, stabilim valoarea 0.85 pentru acest factor. Mai multe detalii despre d sunt oferite în secţiunea următoare. De asemenea, C(A) este definit ca un număr de link-uri care nu fac parte din pagina A. PageRank-ul paginii A este după cum urmează:

$$PR(A) = (1-d) + d(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

Trebuie reţinut că PageRank formează o distribuţie a probabilităţii paginilor web, astfel că suma tuturor paginilor web ce ţin de PageRank este 1. PageRank sau PR(A) poate fi calculat utilizând un simplu algoritm repetabil şi care corespunde principalului vector propriu al matricii link-ului normalizat al web-ului. De asemenea, un PageRank pentru 26 milioane de pagini web poate fi

calculat în câteva ore într-un punct de lucru de mărime medie. Există multe alte detalii care depăşesc intenţia acestei lucrări.

2.1.2 Explicarea intuitivă

PageRank poate fi considerat un model al comportamentului utilizatorului. Să presupunem că există un navigator oarecare care vizitează o pagină web aleasă la întamplare şi care accesează link-urile, fără a reveni la pagina iniţială: în cele din urmă se va plictisi şi se va orienta spre altă pagină web aleasă la întâmplare. Probabilitatea că acest navigator să viziteze o pagină este reprezentată de PageRank. Iar d, factorul de nivelare, reprezintă probabilitatea ca navigatorul să se plictisească la fiecare pagină accesată şi să continue căutarea paginilor la întamplare. O variaţie importantă este aceea de a adaugă doar factorul de nivelare d unei singure pagini sau unui grup de pagini. Acest lucru permite personalizarea şi poate face aproape imposibilă inducerea deliberată în eroare a sistemului pentru obţinerea unui calificativ superior. Pentru mai multe referiri la PageRank.

O altă explicaţie intuitivă este că o pagină poate avea un PageRank ridicat dacă există mai multe pagini care fac referire la aceasta sau dacă există câteva pagini care au un PageRank ridicat şi care o recomandă. În mod intuitiv, paginile la care se face referire din multe colţuri ale web-ului sunt considerate importante. De asemenea, paginile care probabil au o singură referire de la gazda Yahoo-ului sunt considerate importante. Dacă o pagina nu are un nivel calitativ ridicat sau are un link insuficient, este mai mult decât probabil ca pagina gazdă a Yahoo-ului nu va avea nici un link pentru aceasta. PageRank face faţă ambelor situaţii şi chiar mai mult de atât prin propagarea recursivă a gradului de importanţă în întreaga structură de link-uri a web-ului.

2.2 Textul link-ului

Textul link-ului este tratat într-un mod cu totul special de motorul de căutare. Majoritatea motoarelor de căutare asociază textul link-ului cu pagina de care se leagă. În plus, se asociază cu pagina pe care link-ul respectiv o indică. Acest sistem prezintă mai multe avantaje. În primul rand, link-urile oferă deseori descrieri mai precise ale paginilor web decât o fac paginile respective. În al doilea rând, link-urile pot face referire la documente care nu pot fi indexate de un motor de căutare bazat pe text, cum ar fi: imagini, programe sau baze de date. Acest lucru face posibilă returnarea paginilor web care nici macăr nu au fost parcurse. Trebuie reţinut că paginile care nu au fost parcurse pot cauza probleme din moment ce nu le-a fost niciodată verificată validitatea înainte de a fi oferite utilizatorului. În acest caz, motorul de căutare poate oferi o pagină care nu a

existat niciodată cu adevarat, dar care are hyperlink-uri care fac referire la ea. Totuşi, este posibil ca rezultatele să fie sortate, astfel ca această problemă apare rareori.

Ideea corelării textului link-ului cu pagina web la care se referă a fost implementată în World Wide Web Worm, în special pentru că ajută la căutarea informaţiei de tip non-text şi mareşte aria de acoperire a căutarii prin numărul mai redus de documente descărcate. Se foloseşte propagarea de link-uri deoarece textul link-ului poate contribui la oferirea de rezultate mai bune. Utilizarea eficientă a text-ului link-ului este dificilă din punct de vedere tehnic din cauza cantităţilor mari de date care trebuie procesate. În procesul de parcurgere a 24 milioane de pagini, s-au indexat peste 259 de milioane de link-uri.

2.3 Alte trăsături

În afară de PageRank şi de utilizarea textului link-ului, Google are şi alte trăsături. Prima este aceea că are informaţii de bază pentru toate căutarile şi astfel utilizează, în mod frecvent, proximitatea în procesul de căutare. A doua se referă la faptul că Google are în vedere detaliile vizuale ale prezentării cum ar fi mărimea fonturilor. Cuvintele scrise cu un font mai mare sau cu caractere îngroşate sunt percepute altfel decât celelalte cuvinte. A treia trăsătură este aceea că se ţine o evidenţă a întregului şir al paginilor HTML.

3. Activităţi similare

Cercetarea navigarii pe web are o istorie scurtă şi concisă. World Wide Web Worm (WWWW) a fost unul dintre primele motoare de căutare. Acesta a

fost ulterior urmat de alte câteva motoare de căutare academice, dintre care multe sunt acum companii publice. Comparativ cu dezvoltarea Web-ului şi importanţa motoarelor de căutare, există puţine documente valoroase refe-ritoare la motoarele de căutare recente. Conform lui Michael Mauldin (cercetător la Lycos Inc) "serviciile variate (inclusiv Lycos) urmăresc îndeaproape detaliile acestor baze de date". Totuşi, există studii despre trăsăturile specifice ale motoarelor de căutare. Un studiu bine realizat este acela care duce la obţinerea de rezultate prin procesarea ulterioară a rezultatelor motoarelor de căutare comerciale sau la producerea, la scară redusă, de motoare de căutare individualizate. Cercetări amanunţite au fost făcute pe sistemele care se ocupă cu extragerea de informaţii, în special pe colecţiile bine controlate. În următoarele două secţiuni, se va discuta despre zone unde această cercetare trebuie extinsă pentru a se putea lucra mult mai bine pe web.

3.1 Extragerea de informaţii

Cercetarea sistemelor care se ocupă cu extragerea de informaţii a fost iniţiată acum mulţi ani şi este acum bine dezvoltată. Totuşi, majoritatea cercetărilor asupra acestor sisteme se concentrează pe colecţiile omogene şi bine întretinute cum ar fi: colecţiile de lucrări stiintifice sau articole despre un subiect apropiat. Într-adevăr, prototipul extragerii de informaţii, The Text Retrie-val Conference foloseşte pentru teste colecţii mici dar bine controlate. Testul "Very Large Corpus" are doar 20 GB în comparaţie cu 147 GB ai celor 24 de milioane de pagini web. Tot ce funcţionează bine pe TREC nu oferă de obicei rezultate satisfăcătoare pe web. De exemplu, modelul de vector spaţial standard încearcă sa returneze documentele care se apropie cel mai mult de interogare, ţinând seama de faptul că atât interogarea cât şi documentul sunt vectori definiţi de ocurenţa cuvintelor lor. Pe web, această strategie oferă de obicei documente scurte care conţin interogarea şi câteva cuvinte în plus. De exemplu, s-a observat că un motor important de căutare returnează o pagina conţinând doar "Bill Clinton Sucks" şi o fotografie dintr-o interogare referitoare la Bill Clinton. Unii sunt de părere că, pe web, utilizatorii ar trebui să specifice cu mai multă acurateţe ceea ce doresc să afle şi să adauge mai multe cuvinte întrebării respective. Dacă un utilizator formulează o interogare de tipul "Bill Clinton" ar trebui să primească rezultate concludente din moment ce există o cantitate considerabilă de informaţii valoroase disponibile pentru acest subiect. Prin aceste exemple procesul standard de extragere de informaţii trebuie extins pentru a putea utiliza web-ul în mod eficient.

3.2 Diferenţe între Web şi colecţiile bine controlate

Web-ul reprezintă o colecţie vastă de documente eterogene. Documentele de pe web sunt caracterizate printr-o variaţie internă constantă a documentelor şi de asemenea prin meta informaţie externă care poate să fie disponibilă. De exemplu, documentele diferă din punct de vedere intern prin limbaj (atât uman, cât şi de programare), vocabular (adrese de e-mail, link-uri, coduri zip, numere de telefon, numere de producţie), tip de format (text, HTML, PDF, imagini, sunete) şi pot fi generate chiar de sistem (fişiere jurnal sau extrase din baza de date). Pe de altă parte, definim meta informaţia externă ca informaţia care poate fi dedusă despre un document, dar care nu este conţinută de acesta. Exemple de meta informaţie externă includ aspecte precum reputaţia sursei, frecvenţa actualizării, calitatea, popularitatea sau uzajul si link-urile. Nu numai că sursele posibile ale meta informaţiei externe sunt variate, dar lucrurile care sunt măsurate variază de asemenea din punct de vedere al gradelor de importanţa. De exemplu, dacă se compară informaţia uzuală de pe o pagină gazdă importantă ca Yahoo, care primeşte zilnic milioane de accesări de pagini, cu un articol istoric obscur care poate fi accesat o dată la zece ani. În mod evident, aceste două aspecte trebuie tratate în mod diferit de un motor de căutare.

O altă diferenţă majoră între web şi colecţiile tradiţionale este aceea că, practic, nu există nici un control asupra a ceea ce pot publica oamenii pe web. Corelând această flexibilitate de a putea publica orice cu influenţa considerabilă a motoarelor de căutare care pot afecta traficul (şi există diverse companii care manipulează deliberat motoarele de căutare pentru a obţine profit), ajungem la o problemă serioasă. Această problemă nu a fost pusă în cazul sistemelor tradiţionale de extragere de informaţii. De asemenea, este interesant de observat că eforturile depuse în cazul metadatelor au eşuat în cazul motoarelor de căutare, deoarece orice text de pe pagină care nu este direct oferit utilizatorului este supus manipulării efectuate de motoarele de căutare. Există chiar numeroase companii care s-au specializat în manipularea motoarelor de căutare în vederea obţinerii de profit.

4. Configuraţia sistemului

4.1 Scurtă descriere a arhitecturii Google

În această sectiune va fi o descriere despre modul de funcţionare al sistemului. În secţiunile următoare se va discuta despre aplicaţiile şi structurile de date care nu vor fi menţionate în această secţiune. Cea mai mare parte din Google este realizată în C sau C++ pentru eficienţă şi poate rula atât în Solaris, cât şi în Linux.

În Google parcurgerea web-ului (descărcarea de pagini) este făcută de mai multe crawlere diferite. Există un server URL care trimite listele de URL-uri ce trebuie găsite de crawlere. Paginile web care sunt găsite sunt apoi returnate serverului de stocare, care le memorează. Acesta comprimă paginile şi le depune într-o bibliotecă. Orice pagină web are un număr de identificare numit docID, care este oferit ori de câte ori un nou URL este analizat şi extras dintr-o pagină web. Funcţia de indexare este realizată de indexer şi de sorter. Indexer-ul indeplineşte o serie de funcţii. Citeşte documentele din bibliotecă, decomprimă documentele şi le analizează. Fiecare document este convertit într-o serie de asocieri de cuvinte numite hit-uri. Acestea înregistrează cuvantul şi poziţia sa în document, aproximează dimensiunea fontului şi tipurile de litere folosite. Indexer-ul distribuie aceste hit-uri într-o serie de categorii, creând un index parţial dezvoltat de sortare. Indexer-ul mai îndeplineşte şi o altă funcţie importantă. Analizează toate link-urile din fiecare pagina web şi stochează informaţii importante despre acestea într-un fişier de link-uri. Acest fişier conţine informaţii suficiente pentru a stabili unde ne îndreaptă link-ul respectiv, precum şi textul link-ului.

Sistemul de analizare a URL-urilor citeşte fişierul de link-uri şi converteşte URL-urile relative în URL-uri absolute şi, respectiv, în docID-uri. Plasează textul link-ului în indexul iniţial care este asociat cu docID-ul la care se referă link-ul. Acesta generează de asemenea o bază de date de link-uri care nu sunt altceva decât corespondentele docID-urilor. Această bază de link-uri este folosită pentru calcularea PageRank-urilor pentru toate documentele.

Sorter-ul preia categoriile care sunt sortate de docID şi le clasifică după wordID pentru a forma un index complementar (inverted index). Un program numit DumpLexicon preia această listă împreună cu lexiconul produs de indexer şi formează un lexicon nou care poate fi utilizat de searcher. Searcher-ul este rulat de un server şi foloseşte lexiconul construit de DumpLexicon împreună cu indexul complementar şi PageRank pentru a răspunde întrebărilor.

4.2 Structuri majore de date

Structurile de date ale Google sunt optimizate astfel încât o colecţie amplă de documente poate fi parcursă şi indexată cu puţin efort. Deşi CPU-urile şi majoritatea ratelor de input output s-au îmbunătăţit de-a lungul anilor o simplă căutare pe disc tot necesită 10 ms pentru a fi realizată. Google este proiectat să evite acest gen de căutări de câte ori este posibil, iar acest lucru a avut o influenţă considerabilă asupra formatului structurilor de date.

4.2.1 BigFiles

BigFiles reprezintă fişiere virtuale care parcurg multiple fişiere de sistem şi care sunt operaţionale cu 64 de biţi integrali. Împărţirea între multiplele fişiere de sistem este realizată în mod automat. Aceste fişiere suportă de asemenea alocarea şi dealocarea descriptorilor de fişiere, din moment ce sistemele de operare nu oferă destul pentru ceea ce avem noi nevoie. BigFiles suportă şi opţiuni simple de comprimare.

4.2.2 Biblioteca (Repository)

Biblioteca conţine HTML-ul integral al fiecarei pagini web. Fiecare pagină este comprimată prin folosirea zlib. Optarea pentru o tehnică de compresie reprezintă echilibrul între viteză şi proporţia comprimării. S-a ales viteza zlib dintr-o serie de îmbunătăţiri semnificative aduse comprimării de bzip. Rata compresiei bzip era de aproximativ 4 la 1 în bibliotecă, în comparaţie cu zlib care oferea o rată de 3 la 1. În bibliotecă, documentele sunt stocate unul după altul şi sunt prefixate de docID, precizându-li-se lungimea şi URL-ul. Biblioteca nu solicită alte structuri de date care să fie folosite pentru ca aceasta să fie accesată. Acest lucru contribuie la consistenţa informaţiei uşurând dezvoltarea; se poate reconstrui toate celelalte structuri de date doar din bibliotecă şi dintr-un fişier care listează erorile crawler-ului.

4.2.3 Indexul documentelor

Indexul documentelor păstrează informaţii despre fiecare document. Acesta este un index ISAM (Index sequential access mode) cu o lăţime fixă, ordonat de un docID. Informaţia conţinută de fiecare scurtă introducere include statutul curent al documentului, un indicator către bibliotecă, o evidenţă a documentului şi statistici variate. Dacă documentul a fost parcurs atunci conţine un indicator către un fiţier cu multe variabile numit docinfo şi care cuprinde URL-ul şi titlul documentului. În caz contrar, indicatorul se îndreaptă către lista URL-urilor care cuprinde numai URL-uri. Această hotărare de design a fost

luată în conformitate cu dorinţa de a avea o structură compactă de date, precum şi cu abilitatea de stabili un record de accesare unică a discului în timpul unei căutări.

În plus, există un fişier care este folosit în convertirea URL-urilor în docID-uri. Acesta conţine o lista cu URL-uri împreuna cu docID-ul corespunzător şi este sortat de sumă de control. Pentru a găsi docID-ul unui anume URL, suma de control a URL-ului este calculată şi o căutare binară este realizată pe fişierul de sume de control pentru identificarea docID-ului. URL-urile pot fi convertite în docID-uri luând mai multe simultan prin alipirea la acest fişier. Aceasta este tehnica pe care cel ce soluţionează URL-uri o foloseşte pentru a transforma URL-urile în docID-uri. Această metodă de abordare este importantă pentru că altfel trebuie să efectuam o căutare pentru fiecare link care, ţinând cont de disc, ar dura mai mult de o lună pentru baza de 322 milioane de link-uri.

4.2.4 Lexiconul

Lexiconul are mai multe forme diferite. O modificare esenţiala adusă, spre deosebire de sistemele anterioare, este ca lexiconul să poate ajusta memoria la un preţ rezonabil. În implementarea actuală putem păstra lexiconul în memorie într-un sistem cu 256 MB de memorie. Lexiconul actual conţine 14 milioane de cuvinte (deşi câteva cuvinte rare nu au fost adăugate). Lexiconul are două parti: o listă de cuvinte (înşiruite împreună, dar separate de zerouri) şi o serie de indicatori. Pentru a îndeplini diverse funcţii, lista de cuvinte are informaţii auxiliare care se aflş însă dincolo de scopul acestei lucrări.

4.2.5 Listele de hit-uri (Hit Lists)

O listă de hit-uri corespunde unei liste de apariţii ale unui anumit cuvânt într-un document, incluzând informaţii despre poziţia, fontul şi tipul de literă folosit. Listele de hit-uri explică cea mai mare parte a spaţiului utilizat atât în indicele primar (forward index), cât şi în indicele complementar (inverted index).

Din această cauză, este important să fie reprezente cât mai eficient posibil. Am luat în calcul mai multe alternative pentru poziţia de codificare, font şi tipul de literă – codificarea simplă (un grup de trei numere inetgrale), codificarea compactă (o serie de biţi optimizaţi manual) şi codificarea Huffmann. În final, am ales codificarea compactă optimizată manual deoarece necesită de departe mai putin spaţiu decat codificarea simplă şi mult mai puţină manipulare a biţilor decât codificarea Huffmann. Detalii despre aceste hit-uri sunt arătate.

Codificarea compactă foloseşte doi biţi pentru fiecare hit. Există două tipuri de hit-uri: hit-uri complexe (fancy hits) şi hit-uri simple (plain hits). Hit-urile complexe includ apariţia hit-urilor într-un URL, titlu, textul link-ului sau meta tag. Hit-urile simple includ restul. Un hit simplu constă dintr-un bit referitor la tipul de literă, marimea fontului şi 12 biţi de poziţii ale cuvântului într-un document (toate poziţiile ce depăşesc 4095 sunt catalogate 4096). Mărimea fontului este reprezentată relativ faţă de restul documentului utilizând 3 biţi (doar 7 valori sunt de fapt folosite deoarece 111 este simbolul care semnalează apariţia unui hit complex). Un hit complex constă într-un bit referitor la tipul de literă, mărimea fontului este setată la 7 pentru a indica că este vorba de un hit complex, 4 biţi pentru codificarea tipului de hit complex şi 8 biţi de poziţie. Pentru hit-urile de tip anchor, cei 8 biţi ai poziţiei sunt împărţiţi în 4 biţi pentru poziţie în link şi 4 biţi pentru conţinutul docID-ului în care link-ul apare. Aceasta ne oferă o sintagma redusă de căutare din moment ce nu există multe link-uri pentru un anumit cuvânt. Trebuie să se actualizeze metoda de stocare a hit-urilor anchor pentru permiterea unei rezoluţii mai mari în cadrul poziţiei şi câmpurilor de docID-uri. Se va folosi mărimea fontului în legătură cu restul documentului deoarece, atunci când căutam, nu dorim listarea diferită a unor documente identice doar pentru că unul din documente este scris cu un font mai mare.

Lungimea unei liste de hit-uri este stocată înainte chiar de hit-urile în sine. Pentru a economisi spaţiu, lungimea listei de hit-uri este combinată cu wordID-ul în indexul primar şi cu docID-ul în indexul complementar. Acest lucru o limitează la 8 şi respectiv 5 biţi (există o serie de trucuri care permit ca 8 biţi să fie imprumutaţi din wordID). Dacă lungimea este mai mare şi nu se poate încadra în respectivii biţi, atunci un cod de rezervă este folosit în aceşti biţi, iar următorii 2 biţi vor conţine lungimea actuală.

4.2.6 Indexul primar (Forward Index)

Indexul primar este deja parţial sortat şi este stocat într-o serie de categorii (s-au folosit 64). Fiecare categorie conţine o serie de wordID-uri. Dacă un document conţine cuvinte care ţin de un anumit barrel, docID-ul este intregistrat în categorie urmat de o listă de wordID-uri cu liste de hit-uri care corespund cuvintelor respective. Această schemă necesită mai mult spaţiu de stocare din cauza docID-urilor duplicate, dar diferenţa este foarte mică pentru un număr considerabil de categorii şi economiseşte timp şi complexitate de codificare în faza finală de indexare făcută de sorter. Mergând mai departe, în loc de a stoca wordID-urile actuale, se stochează fiecare wordID ca o diferenţă relativă de la wordID-ul minim care se găseşte în categoria în care se află şi wordID-ul. Astfel, se pot folosi 24 biţi pentru wordID-uri în categorii nesortate, lasând 8 biţi pentru lungimea listelor de hit-uri.

4.2.7 Indexul complementar

Indexul complementar constă din aceleaşi categorii ca şi indexul primar, cu diferenţa că acestea au fost procesate de sorter. Pentru fiecare wordID valid, lexiconul conţine un indicator către categoria în care wordID-ul este inclus. Acest indicator se referă la o listă de docID-uri luate împreună cu listele de hit-uri corespunzatoare. Această listă reprezintă toate apariţiile acelui cuvânt în toate documentele.

4.4 Indexarea Web-ului

— *Analiza*: orice sistem de analizare care este destinat să ruleze pe întreg Web-ul trebuie să facă faţă unei mari cantităţi de erori posibile. Acestea se pot întinde de la mici greşeli în sintagmele HTML până la kilobiţi de zerouri în mijlocul unei sintagme, caractere non-ASCII, sintagme HTML depozitate în interior şi o mare varietate de alte greşeli care pot provoca imaginaţia oricui pentru a-l determina să apară cu altele noi. Pentru viteza maximă, în loc să se folosească YACC pentru a genera un sistem CFG de analiză, se foloseşte un cablu electric pentru iniţierea unui analizator lexical *which* care se dotează cu propria informaţie. Dezvoltarea acestui sistem de analiză care rulează cu o viteză rezonabilă şi care este foarte solid a însemnat un timp îndelungat de muncă continuă.

— *Indexarea documentelor în categorii*: după ce fiecare document este analizat, este codificat într-o serie de categorii. Fiecare cuvânt este convertit într-un wordID utilizând un conţinut de tip in-memory (lexiconul). Adăugările care se fac conţinutului lexiconului sunt consemnate într-un fişier. Din momentul în care cuvintele sunt convertite în wordID-uri, apariţiile lor în documentul curent sunt traduse în liste de hit-uri şi sunt scrise în categoriile primare. Principala dificultate cu simultaneitatea fazei de indexare este aceea că lexiconul trebuie împărţit. În loc de împărţirea lexiconului, se scrie un logaritm pentru toate cuvintele auxiliare care nu se aflau în lexiconul de bază, care s-a fixat la 14 milioane de cuvinte. În acest fel, indexer-ele multiple pot rula în paralel şi apoi fişierul jurnal cu logaritmii cuvintelor auxiliare poate fi procesat de un singur indexer final.

— *Sortarea*: pentru a putea genera indexul complementar, sorter-ul preia fiecare categorie primară şi o sortează după wordID pentru a produce o categorie complementară pentru titlu şi hit-urile de tip anchor şi o categorie complementară de text integral. Acest proces se întâmplă cu fiecare categorie în parte, necesitând astfel un spaţiu mic de stocare temporară. De asemenea, se compară faza de sortare pentru a utiliza cât mai multe sisteme posibile prin

simpla rulare de sorter-e multiple, care pot procesa categorii diferite în acelaşi timp. Din moment ce categoriile nu sunt compatibile cu memoria principală, sorter-ul le subdivide mai departe în subcategorii care se potrivesc cu memoria bazată pe word ID şi docID. Apoi, sorter-ul încarcă fiecare subcategorie în memorie, o sortează şi îi transcrie conţinutul în categorii complementare mici şi în categorii complementare detaliate.O altă caracteristică importantă o reprezintă ordinea în care docID-urile trebuie să apară în listă. O soluţie simplă este să fie stocate în funcţie de docID. Acest lucru permite fuzionarea rapidă a diferitelor liste pentru întrebările cu multe cuvinte. O altă posibilitate este să fie stocate în funcţie de apariţia cuvântului în fiecare document. Acest lucru implică lipsa de valoare a răspunsurilor la întrebările alcătuite dintr-un singur cuvânt şi face posibil faptul că răspunsurile la întrebările cu multe cuvinte să se afle chiar la început. Totuşi, fuzionarea este cu mult mai dificilă. De asemenea, acest lucru face ca dezvoltarea să devină mult mai dificilă, în sensul că o schimbare în funcţia de ordonare necesită o reconstruire a indexului. S-a ales un compromis între aceste opţiuni, prin păstrarea a două serii de categorii complementare - o serie pentru listele de hit-uri care include titlul sau link-ul şi o altă serie pentru toate listele de hit-uri. În acest fel, se verifică primul set de categorii şi dacă nu se găsesc destule potriviri în cadrul acestora se verifică şi cele mai mari.

Rularea unui crawler web reprezintă o provocare. Există acţiuni riscante şi subiecte sigure şi, mult mai important, există subiecte sociale. Parcurgerea (crawling) reprezintă cea mai fragilă aplicaţie din moment ce implică interacţiunea cu sute de mii de servere web şi nume de servere diverse care se află dincolo de posibilitatea de control a sistemului.

Pentru a pacurge sute de milioane de pagini web, ***Google are un sistem rapid (fast distributed crawling)***. Un singur server URL oferă liste de URL-uri unui număr de crawlers (în general se folosesc în jur de 3). Atât server-ul URL, cât şi crawler-ele sunt realizate în Python. Fiecare crawler ţine în jur de 300 de conexiuni (connections) deschise simultan. Acest lucru este necesar pentru regăsirea paginilor web la o viteza suficient de rapidă. La viteze mari sistemul poate să parcurgă peste 100 de pagini pe secundă utilizând 4 crawlere. Acesta se ridică la aproximativ 600K de date pe secundă. O acţiune importantă este reprezentată de verificarea DNS. Fiecare crawler menţine un cache DNS propriu, astfel că nu este nevoie să se facă un control DNS înainte de parcurgerea fiecărui document. Fiecare dintre sutele de conexiuni se poate afla în stadii diverse: verificarea DNS, conectarea la gazdă, transmiterea solicitărilor şi primirea răspunsurilor. Aceşti factori fac din crawler o componentă complexă a sistemului. Acesta foloseşte IO asincron pentru a face faţă solicitărilor şi un număr de secvenţe pentru mutarea preluărilor de pagini din secţiune în secţiune. Se adevereşte astfel că rularea unui crawler care se conectează la mai mult de jumătate de milion de servere şi care generează zeci de milioane de fişiere jurnal implică o cantitate considerabilă de e-mailuri şi apeluri telefonice. Datorită numărului mare de persoane care sunt online, există întotdeauna aceia care nu

ştiu ce este un crawler deoarece acesta este primul pe care îl văd. Aproape în fiecare zi primim un mesaj de genul ‘V-aţi uitat la foarte multe pagini din site-ul meu. Ce parere aveţi despre el?'. Există de asemenea unele persoane care nu au aflat de existenţa protocolului de excluziune la roboţi şi se consideră că pagina lor ar trebui protejată împotriva unei indexari de genul ‘Această pagina este protejată şi nu va fi indexată‘, care este dificil de înteles de crawler-ele web. De asemenea, din cauza cantităţii uriaşe de date implicate, lucruri neaşteptate se pot întampla.

De exemplu, sistemul a încercat să parcurgă un joc online.

Rezultatul a constat într-o mulţime de mesaje iritante chiar în mijlocul jocului. Se pare că aceasta a fost o problemă uşor de rezolvat. Dar această problemş nu aparuse până în momentul în care s-au descărcat zeci de milioane de pagini. Datorită variaţiei ridicate în paginile web şi în servere, este practic imposibil să testezi un crawler fără să-l rulezi pe o parte considerabilă a Internetului. Invariabil, apar sute de probleme obscure care se pot ivi pe o singură pagină din tot web-ul şi pot cauza distrugerea crawler-ului sau mai rău, poate cauza o reacţie imprevizibilă sau incorectă. Sistemele care accesează parţi mari din Internet trebuie să fie foarte solide şi testate cu multă atenţie. Din moment ce sistemele complexe cum sunt crawler-ele vor duce în mod invariabil la apariţia problemelor, trebuie să existe resurse semnificative dedicate citirii de e-mail-uri şi rezolvării problemelor din momentul în care acestea apar.

4.5 Căutarea

Scopul căutarii este acela de a oferi rezultate concludente în timp util. Multe dintre motoarele de căutare comerciale par să fi facut progrese considerabile din punct de vedere al eficienţei. De aceea, se pune accentul mai mult pe calitate.

Pentru marcarea unei limite a timpului de răspuns, odată ce un anumit număr de documente care se potrivesc cu interogarea (40.000 de obicei) este găsit, cel care a iniţiat căutarea poate accesa paginile.

Aceasta înseamna că este posibil ca rezultate neconcludente să fie oferite în schimb. În prezent, se investighează alte metode pentru rezolvarea acestei probleme. În trecut, s-au sortat hit-urile în concordanţă cu PageRank, lucru care pare să fi imbunătăţit situaţia.

4.5.1 Sistemul de clasificare

Google păstreaza mult mai multe informaţii despre documentele web decât motoarele tipice de căutare. Fiecare listă de hit-uri include poziţia, fontul şi informaţii despre tipul de literă folosit. În plus, se iau în calcul hit-urile după textul link-ului şi PageRank-ul documentului. Combinarea tuturor acestor

informaţii într-un singur rezultat este dificilă. S-a conceput funcţia de ordonare astfel încât nici un factor particular să nu aibă o influenţă prea mare. Se iamai întâi cazul cel mai simplu - o interogare cu un singur cuvânt.

Pentru afişarea unui document folosind o interogare cu un singur cuvânt, Google parcurge toate listele de hit-uri ale documentului pentru cuvântul respectiv. Google consideră fiecare hit ca aparţinând unuia dintre diversele tipuri (titlu, link, URL, fonturi mari şi fonturi mici de text simplu etc.), fiecare dintre acestea având grade diferite de importanţă în funcţie de tipul din care face parte. Aceste grade de importanţă formează un vector indexat în funcţie de tip. Google numară hit-urile fiecărui tip din lista de hit-uri. Apoi fiecare poziţie este reorganizată într-un clasament în funcţie de importanţă. Gradele de importanţă cresc liniar în funcţie de primele poziţii, dar se reduc repede astfel încât este relevant numai un anume număr de apariţii. Se preia produsul scalar al vectorului de ponderi de apariţii împreună cu vectorul de ponderi de tipuri pentru a calcula un scor IR al documentului. În final, scorul IR este combinat cu PageRank pentru a oferi un rezultat final al documentului.

Pentru o interogare alcatuită din mai multe cuvinte, situaţia este şi mai complicată. În acest caz, listele multiple de hit-uri trebuie parcurse simultan astfel încât hit-urile care sunt apropiate într-un document sunt plasate pe poziţii superioare faţă de cele care sunt departate unele de altele. Hit-urile din listele multiple sunt potrivite astfel incat hit-urile apropiate sunt puse impreuna. Pentru fiecare set de potriviri de hit-uri, se calculează o apropiere. Această apropiere se bazează pe cât de departate sunt hit-urile în cadrul documentului (sau link-ului), dar este clasificată în 10 clase cu valori diferite, mergând de la o sintagmă apropiata până la ‘nu foarte aproape'. Se fac contorizări nu numai pentru fiecare tip de hit, dar şi pentru fiecare tip şi apropiere. Fiecare pereche de tip şi apropiere are o pondere tip-apropiere. Contorizările sunt clasificate în funcţie de ponderile de apariţii şi se preia produsul scalar pentru ponderile de apariţii şi ponderile de tip-apropiere pentru realizarea unui scor IR. Toate aceste numere şi matrice pot fi afişate odată cu rezultatele căutării folosind o metodă specială de corectare. Toate aceste afişări sunt de foarte mare ajutor în dezvoltarea sistemului de ordonare.

4.5.2 Scurta recapitulare

Funcţia de ordonare are mulţi parametri precum ponderile de tip şi ponderile tip-apropiere. Stabilirea valorilor exacte pentru aceşti parametri ţine de domeniul magiei. Pentru a face acest lucru, există un mecanism de feedback al utilizatorului în cadrul motorului de căutare. Un utilizator încrezător poate evalua opţional toate rezultatele care îi sunt returnate. Acest feedback este salvat. Apoi, când se modifică funcţia de ordonare, se observă impactul acestei schimbări asupra căutarilor anterioare care au fost ordonate. Deţi departe de

perfecţiune, acest lucru ne face sa observăm cum o schimbare în funcţia de ordonare poate afecta rezultatele căutării.

5. Rezultate şi mod de funcţionare

Cea mai importantă caracteristică a unui motor de căutare este calitatea rezultatelor căutării. În timp ce o evaluare completă a utilizatorului se află dincolo de scopul acestei lucrări, experienţa cu Google a dovedit că este posibil sa se obţina rezultate mult mai bune pentru majoritatea căutărilor decât ale celor mai importante motoare comerciale de căutare. Dacă se caută ‘bill clinton' se obţin rezultate ce evidenţiază câteva dintre trăsăturile Google. Rezultatele sunt grupate de server. Acest lucru este de mare ajutor atunci când se examinează seturile de rezultate. O serie de rezultate sunt din domeniul whitehouse.gov, care reprezintă rezultatul aşteptat pentru o căutare de acest gen. În mod obişnuit, majoritatea motoarelor de căutare comerciale nu oferă nici un rezultat din whitehouse.gov, cu atât mai puţin pe cele corecte. Trebuie observat că nici un titlu nu este afişat pentru primul rezultat, din cauză că nu a fost parcurs. În schimb, Google s-a bazat pe textul link-ului pentru a hotărî dacă acesta era un răspuns potrivit pentru întrebare. În mod asemănător, al cincilea rezultat este o adresă de e-mail, care, bineînţeles, nu poate fi parcursă (crawlable) şi care este, de asemenea, un rezultat al textului link-ului.

Toate rezultatele sunt pagini cu un înalt nivel calitativ şi, la o ultimă verificare, nici una dintre ele nu este un link irelevant (broken). Acest lucru se întâmplă din cauză că toate au un PageRank ridicat. PageRank-urile sunt marcate prin procentajele de culoare roşie. În concluzie, nu sunt rezultate despre un alt Bill în afara de Clinton sau despre un alt Clinton în afară de Bill şi asta pentru că se acordă o importanţă majoră aproximării sintagmelor în care apare un cuvânt. Bineînţeles că un adevărat test al calităţii pentru un motor de căutare ar trebui să implice şi un studiu complet despre utilizator sau o analiză a rezultatelor.

5.1 Parametri pentru stocare

În afară de calitatea căutării, Google este proiectat să măsoare capacitatea Web-ului pe măsuă ce acesta se dezvoltă. Un aspect al acestui fenomen îl reprezintă modul eficient de stocare. Datorită comprimării, mărimea totală a bibliotecii este de 53 GB, puţin peste o treime din totalul de date pe care le conţine. La preţurile actuale ale discurilor biblioteca poate fi considerată o sursă ieftină de date. Ceea ce este mai important - totalul datelor folosite de motorul de căutare necesitaă un spaţiu comparabil de stocare, adică în jur de 55 GB. Mai departe, un răspuns poate fi găsit pentru majoritatea interogarilor prin folosirea indexului complementar.

5.2 Modul de funcţionare a sistemului

Este important pentru un motor de căutare să parcurgaă şi să indexeze eficient. Astfel, informaţia poate fi permanent actualizată şi modificările majore aduse sistemului pot fi testate relativ repede. Pentru Google, operaţiunile importante sunt ***Crawling (parcurgerea), Indexing (indexarea) si Sorting (sortarea)***. Este dificil de măsurat cât a durat crawling-ul în total din cauză că discurile au fost în întregime completate, numele serverelor nu mai sunt funcţionale sau din cauza oricărei probleme care putea determina oprirea sistemului. În total a durat cam 9 zile să se descarce cele 26 milioane de pagini (inclusiv erorile). Totuşi, din moment ce sistemul rula fluent şi mai repede, s-au descărcat ultimele 11 milioane de pagini în doar 63 de ore, în medie 4 milioane de pagini pe zi sau 48,5 pagini pe secundă. S-au rulat indexer şi crawler simultan. Indexer-ul a rulat mai repede decât crawler-ele. Aceasta s-a întâmplat din cauză că s-a petrecut destul timp pentru optimizarea indexer-ului astfel încât să nu determine întârzierea dezvoltării sistemului. Aceste optimizări au inclus o serie de updatări ale indexului documentului şi plasarea structurilor precare de date pe un disc local. ***Indexer-ul rulează aproximativ 54 de pagini pe secundă. Sorter-ele pot fi rulate complet în paralel, utilizând 4 sisteme, întreg procesul de sortare durând în jur de 24 de ore.***

5.3 Procesul de căutare

Îmbunătăţirea procesului de căutare nu a reprezentat punctul principal de interes al cercetării al realizatorilor Google-ului până în acest moment. Versiunea curentă Google răspunde la cele mai multe din interogări între 1 şi 10 secunde. Acest interval este dominat în mare parte de IO-ul discului asupra NFS (din moment ce discurile sunt împărţite pe sisteme diferite). Mai departe, Google nu are nici un fel de optimizare cum ar fi cache de căutări, subindici pe termeni comuni şi alte optimizări obişnuite. *Intenţia lor este de a mări considerabil viteza Google prin distribuţie şi hardware, software şi îmbunătăţiri algoritmice.* Scopul este acela de a putea răspunde la mai multe sute de interogări pe secundă.

6. Concluzii

Google este proiectat să fie un motor de căutare scalabil. Scopul principal este acela de a oferi rezultate de calitate pe fondul dezvoltării rapide a World Wide Web. Google foloseşte o serie de tehnici pentru ameliorarea calităţii căutării incluzând page rank, textul link-ului şi alte informaţii apropiate. Mai departe, Google reprezintă o arhitectură completă pentru adunarea paginilor web, indexarea lor şi efectuarea de interogări asupra lor.

6.1 Activităţi viitoare

Un motor amplu de căutare este un sistem complex care rămâne deschis completărilor. Scopurile noastre imediate sunt acelea de a imbunătăţi căutarea şi de a afişa aproximativ 100 de milioane de pagini web. Câteva îmbunătăţiri simple aduse eficienţei includ cache de interogări, alocare optimizată a discului şi subindici. O altă secţiune care necesita o cercetare amanunţită este reprezentată de updatări. Trebuie să avem algoritmi potriviţi pentru a decide ce pagini mai vechi ar trebui reparcurse şi ce pagini noi ar trebui parcurse. O zonă promiţătoare de cercetare este aceea de a utiliza un *cache proxy* pentru construirea de baze de date cu scopul căutării din moment ce acestea sunt influenţate de cerere. Se plănuieşte adăugara de trăsături simple ce sunt suportate de motoarele comerciale de căutare cum ar fi operatorii de tip boolean, negaţia şi implicaţia. Totuşi, alte trăsături cum ar fi feedback-ul relevanţei şi clustere (Google suportă în mod obişnuit clusterele bazate pe numele de domeniu) de-abia acum încep să fie explorate. Se intenţionează de asemenea susţinerea contextul de utilizator (cum ar fi locaţia utilizatorului) şi rezumarea rezultatelor. Apare şi preocuparea extinderii utilizării structurii de link-uri şi a textului link-ului. Experimentele simple indică faptul că PageRank poate fi personalizat prin amplificarea ponderii homepage-ului unui utilizator sau a bookmark-urilor acestuia. În ceea ce priveşte textul link-ului, se experimentează utilizarea textului link-urilor învecinate alături de textul link-ului în sine. Un motor de căutare pe Web repezintă un mediu bogat pentru iniţiativele de cercetare.

6.2 Căutare la standarde înalte

Cea mai mare problemă cu care se confruntă astăzi utilizatorii de motoare de căutare o reprezintă calitatea rezultatelor pe care le primesc. Pe când rezultatele sunt deseori amuzante şi lărgesc orizontul utilizatorului, ele pot deveni şi frustrante şi pot consuma timp pretios. De exemplu, primul rezultat pentru o căutare ‘Bill Clinton' pe unul dintre cele mai populare motoare de

căutare comerciale a fost Bill Clinton Joke of the Day: April 14, 1997. Google este destinat să ofere rezultate de o calitate superioară astfel încat Web-ul să continue să se dezvolte rapid, iar informaţia să poata fi gasită uşor. Pentru a putea realiza acest lucru, Google utilizează frecvent informaţia hipertextuală ce constă din structura de link-uri şi din textul link-urilor. Google foloseşte de asemenea aproximarea şi informaţia despre fonturi. În timp ce evaluarea unui motor de căutare este dificilă, s-a aflat în mod intuitiv că Google returnează rezultate la un înalt nivel calitativ spre deosebire de motoarele comerciale de căutare. Analizarea structurii de link-uri prin PageRank permite Google să evalueze calitatea paginilor web. Utilizarea textului link-ului ca o descriere a ceea ce indică link-ul contribuie la relevanţă şi, într-o anumită măsură, la înaltul standard calitativ al rezultatelor. În cele din urmă, utilizarea unor informaţii asemănătoare ajută la mărirea gradului de relevanţă al multor interogări.

6.3 Arhitectura scalabilă

În afară de calitatea căutării, Google este destinat să măsoare. Trebuie să dea dovadă de eficienţă atât în spaţiu, cât şi în timp, iar factorii constanţi sunt foarte importanţi atunci când vorbim de Web. Prin implementarea Google, s-au observat întârzieri în CPU, accesarea memoriei, capacitatea memoriei, căutările pe disc, conţinutul discului, capacitatea discului şi IO de reţea. Google a evoluat pentru a depăşi o serie din aceste probleme prin intermediul unor operaţiuni variate. Structurile majore de date ale Google utilizează în mod eficient spaţiul de stocare. Mai departe, operaţiunile de parcurgere, indexare şi sortare sunt suficient de eficiente pentru a construi un index al unei părţi substanţiale a Web-ului – 24 de milioane de pagini în mai puţin de o săptămână.

Referinţe

Google Guide (2003) “How Google Works” *Google Guide* Retrieved 8 Sept, 2004 from: googleguide.com

Sullivan, R., (2004) “Google and Googlebot Information” *searchengineposition.com* Retrieved 8 Sept, 2004 from: searchengineposition.com

Brin, S., Page, L. (2000) “The Anatomy of a Large-Scale Hypertextual Web Search Engine” Retrieved 8 Sept, 2004 from: stanford.edu

Barroso, L. A., Dean, J., Hölzle, U. (2003) “Web search for a planet: The Google cluster architecture” *Micro, IEEE*, Volume: 23 , Issue: 2, pp.22-28 Ghemawat, S., Gobioff, H., and Leung, S. T., (2003) “The Google File System” *SOSP’03*, October 19–22

Traducere de pe pagina: http://wiki.media-culture.org.au/index.php/Google-How_It_Works.