«آناتومی یک موتور جستجو وب فوق متنی در مقیاس وسیع»

خلاصه:

در این بخش، به گوگل خواهم پرداخت، یک نمونهٔ اصلی از یک موتور جستجوی در مقیاس وسیع که استفاده وسیعی از ساختار اراده شده در فوق متنی می کند. گوگل برای جستجو و یافتن (Crawl) و شاخص بندی وب به طور مؤثر و تولید نتایج هرچه رضایت بخش تر نسبت به سیستم های موجود طراحی شده است. این نمونه اصلی با پایگاه داده ای متشکل متن و فوق پیوند کامل ۲۴ میلیون صفحه در http://google.standard.edi/ موجود می باشد. مهندسی یک موتور جستجو یک وظیفهٔ چالش آور است. موتورهای جستجو دهها تا صدها میلیون صفحه وب متشکل از تعداد قابل ملاحظه ای موضوعهای متفاوت را شاخص بندی می کنند و پاسخ گوی دهها میلیون پرس و جو به صورت روزانه هستند. بر خلاف اهمیت بالای موتورهای جستجوی برروی وب تحقیقات آکادمیک بسیار اندکی برروی آنها صورت گرفته است (در کشور عزیز ما دقیقاً هیچ مطالعه و تحقیقی صورت نگرفته است). علاوه بر این به دلیل سرعت پیشرفت تکنولوژی وب، امروزه ساخت یک موتور جستجو مسبت به سه سال پیش بسیار متفاوت است. این بخش به بررسی و توصیف عمقی این موتور جستجوی وب در مقیاس وسیع می پردازد. جدای از مشکلات تغییر مقیاس تکنیکهای جستجوی قدیمی داده با این وسعت، چالشهای تکنیکی جدیدی در زمینه استفاده از اطلاعات اضافی ارائه شده در فوق متن برای تولید نتایج جستجوی بوجود آمده است. این بخش به این که چگونه می توان یک سیستم در مقیاس وسیع عملی که بتواند اطلاعات اضافی ارائه شده در فقو متن را استخراج کند را تولید کرد، پاسخ

خواهد گفت. همچنین ما به این مشکل که چگونه می توان با مجموعه های فوق متن کنترل نشده (هر کسی می تواند هر چه خواست بنیسد) کنار آمد، نیز دقت خواهیم کرد.

1. معرفی

وب چالشهای جدیدی برای بازیابی اطلاعات ایجاد می کند. حجم اطلاعات موجود برروی وب به سرعت در حال افزایش است و به همان نسبت تعداد کاربران جدید که در جستجوی وب بی تجربه هستند افزایش می یابد. مردمی که احتمالاً وب را از طریق گراف پیوند آن مرور می کنند، اغلب کار خود را با شاخصهای ذخیره شده با کیفیت بالای انسانی مانند یاهو! یا موتورهای جستجو شروع می کنند. لیتهاس ذخیره و نگهداری شده توسط انسانی موضوعهای معروف را به طور موثری پوشش می دهند اما شخصی بودن، گران و پرهزینه بودن برای ساخت و نگهداری، کندی در پیشرفت و ناتوانی در پوشش موضوعهای مبهم و پیچیده از عیبتهای عمده آنها محسوب می شود. موتورهای جستجو بر پایهٔ هم خوانی کلمات کلیدی معمولاً نتیج را با کیفیت بسیار پایین برمی گرداند. برای بهتر شدن شرایط، بعضی شرکتهای تبلیغاتای تلاش وسیعی برای بدست آوردن نظر مردم از طریق گمراه کردن موتورهای جستجوی اتوماتیک می کنند. اقایان سرگی برین و لاورنس پیج موتور جستجوی در مقیاس وسیعی ساخته اند که به تعداد زیادی از مشکلات سیستم های موجود پرداخته است. و آن استفاده وسیعی از این ساختمام ارائه شده در فوق متن می کند به منظور فراهم کردن نتایج جستجوی با کیفیت بالاتر، اسیم این سیستم، گوگل، انتخاب شده است. زیرا گوگل تلفظ معمول googol یا 10^{100} است و بسیار مناسب هدف ما برای ساختن یک موتور جستجوی بسیار در مقیاس وسیع است.

۱.۱ موتورهای جستجوی وب – گسترش یافتن: 1994-2001

تکنولوژی موتورهای جستجو باید به میزان زیادی تغییر پیدا می کرد تا بتواند هماهنگی خود را با گسترش وب حفظ کند. در 1994، یکی از اولین موتورهای جستجوی وب یعمی کرم وب گسترهٔ جهانی (WWWW) شاخصی از ۱۱۰/۰۰۰ صفحه وب و اسناد در دسترس وب داشت. از نوامبر 1998 موتورهای جستجوی برتر ادعای شاخص بندی از ۲ میلیون (WebCrawler) تا ۱۰۰ میلیون (از Search Engine Watch) صفحه وب و سند را داشتند. قابل پیش بینی است که تا سال 2001 یک شاخص جامع از وب شامل بیش از دو میلیارد سند باشد. در همان زمان تعداد پرس و جوهایی که موتورهای جستجو اداره می کنند به طور شگفت آوری افزایش می یابد. در ماه مارس و آوریل 1994، کرم وب گستره جهانی (wwww) به طور روزانه حدوداً ۱۵۰۰ پرس و جو را دریافت می کرد. در ماه نوامبر 1998، آلتاویستا (Altavista) اظهار داشت که روزانه حدود ۲۰ میلیون پرس و جو را اداره می کند. با افزایش تعداد کاربران وب و سیستمهای اتوماتیک که از موتورهای جستجو پرس و جو می کنند به نظر می رسد که تا سال 2001 موتورهای جستجو صدها میلیون پرس و جو را اداره خواهند کرد. هدف سیستم گوگل توجه به بسیاری از مشکلات کیفیتی و مقیاس پذیری است که با عرضه تکنولوژی موتورهای جستجوی اینترنتی به میزان زیادی گسترش یافته اند.

1.2.1 گوگل: تغییر دادن وب

این موتور جستجوایی که در سطح وب امروز باشد چالشهای بسیاری را پدید می آورد. تکنولوژی جستجو و یافتن سریع برای جمع آوری و به روز رسانی سندهای وب لازمی می باشد. فضای ذخیره سازی بهید به طور کارآمدی برای ذخیره شاخصها و به طور اختیاری خود سندها بکار

گرفته شود. سیستم شاخص بندی باید صدها گیگا بایت داده را به طور کارآمد پردازش کند. پرس و جحوها باید به سرعت اداره شوند (با نرخ صدها تا هزاران پرس و جو در ثانیه).

همان گونه که وب گسترش می یابد این وظایف نیز به طور صعودی مشکل می شوند. اگرچه عملکرد سخت افزار و هزینه ها به طور چشمگیری بهبود یافته اند و تا حدی از این سختی را تعدیل کرده اند. با این وجود تعدادی استثنای قابل اشاره نیز مانند زمان استوانه یابی دیسک و قابلیت ادامه کار در شرایط غیرمنتظره سیستم عامل وجود دارند. در طراحی گوگل هر دو مسئلهع گسترش وب و تغییرات تکنولوژیک در نظر گرفته شده اند. گ.گل برای تغییر مقیاس دادن مجموعه داده ها به خوبی طراحی شده است و از فضای ذخیره سازی به طور مؤثری استفاده می کند. ساختمان داده های آن برای دسترسی سریع بهینه سازی شده اند (به بخش 4.2 نگاه کنید). علاوه بر این، هزینه شاخص بندی و ذخیره متن یا HTML نهایتاً بستگی نمسبی به میزان در دسترسی آنها دارد و این بر تغییر مقیاس منتاسب برای سیستم های متمرکز شده مانند گوگل تاثیرگذار است.

3.1. اهداف طراحی

1.3.1. کیفیت جستجوی بهینه شده

هدف اصلی در طراحی گوگل بهینه کردنم موتورهای جستجوی وب است. در سال 1994، بعضی از مردم تصور می کردند یک شاخص جستجوی کامل امکان یافتن هر چیزی را میسر می سازد. بر طبق مقالهٔ بهترینهای وب 1994 – پیمایشگرها و «بهترین سرویس پیمایشی باید امکان یافتن تقریباً هر چیزی را به آسانی فراهم کند (هنگامی که تمام داده ها وارد شدند)». اگرچه وب 1999 کاملاً متفاوت است. هر کسی که اخیراً از یک موتور جستجو استفاده کرده باشد به

سادگی در می یابد که کامل بودن شاخص تنها عامل مؤثر بر کیفیت نتایج جستجو نمی باشد. «نتایج آشغال» اغلب تمام نتایج مورد علاقه کاربر را خراب می کنند. در حقیقت در نوامبر ۱۹۹۹، تنها یکی از چهار مکوتور تجاری برتر نتایج را خودش می یابد (در پاسخ در ده نتیجه برتر، صفحه جستجو شده خودش را برمی رگداند). یکی از دلایل اصلی این مشکل این است که تعداد سندهای موجود در شاخصها به دلایل روشنی افزایش پیدا کرده اند اما توانایی کاربر بریا یافتن و نگاه کردن اسناد پیشرفت نکرده است. مردم هنوز خواستار نتیجه اول جستجو هستند. به همین دلیل، همان طور کهئ اندازهٔ مجموعه گسترش می یابد، به ابزارهایی که دقت بسیار بالایی دارند نیاز بیشتری پیدا می شود (تعداد اسناد مربوط و مناسب برگردانده شده، در بین ده نتیجه برتر می آید). در واقع، گوگل می خواهد مفهوم «مناسب» فقط شامل بهترین اسناد باشد درحالیکه ممکن است، ده ها هزار سند تقیرباً وجود داشته باشد. خوش بینی های جدیدی در زمینه بهبود عملکرد موتورهای جستجو و سایر برنامه های اجرایی با استفاده بیشتر از اطلاعات فوق متنی بوجود آمده است [Kleinberg 98]. علی الخصوص، ساختمان پیوندها [Page 98] و نوشته پیوندها اطلاعات زیادی برای قضاوت مناسب و فیلترینگ کیفیت فراهم می کند. گوگل از هر دوی ساختمان پیوند و متن انکر استفاده می کند.

2.3.1. تحقیقات موتور جستجوی آکادمیک

جدای از گسترش بسیار زیاد، وب به طور افزایشی در طول زمان حالت تجاری به خود گرفته است، در سال 1993، ۱/۵٪ از سرویس دهندگان وب بر دامنه com. قرار داشتند. این مقدار در سال 1998 به ۶۰٪ رسید. در همان زمان، موتورهای جستجو از حوزهٔ آکادمیک به تجاری کوچ

کردند. تا امروز اغلب پیشرفتهای موتورهای جستجو در شرکتهایی صورت می گیرد که حداقل میزان انتشار جزئیات را دارند. این باعث می شود تکنولوژی موتور جستجو تا حد زیادی مثل جادوی سیاه مخفی باقی بماند و گرایش تبلیغاتی پیدا کند. با گکوگل، سعی شده است تا پیشرفت و فهم بیشتری در قلمرو آکادمیک صورت گیرد.

یکی دیگر از اهداف طراحی ساخت سیستمهایی بود که تعداد قابل قبولی از مردم می توانند استفاده کنند. قابلیت کاربری در طراحی بسیار مهم بوده است زیرا بنظر می آید که اغلب تحقیقات جالب شامل تأثیر استفاده گسترده از سیستمهای مدرن وب در دسترس هستند می باشد. برای مثال، هر روز دهها میلیون جستجو اجرا می شوند. اگرچه، بدست آوردن این داده ها مشکل است، بیشتر به این دلیل که با توجه به جوانب اقتصادی این داده ها ارزشمند هستند.

هدف نهایی طراحی گوگل ساخت یک معماری که قابلیت پشتیبانی از فعالیتهای تحقیق نوظهور بررسی داده های در مقیاس وسیع وب را داشته بوده است. برای پشتیبانی از استانداردهای تحقیقاتی نوول، گ.گل تمام اسناد فعلی را که جستجو می کند و می یابد به صورن فشرده ذخیره می کند. یکی از اهداف اصلی طراحی گوگل بوجود آوردن محیطی بود تا سایر محققات بتوانند به سرعت وارد شده، قسمت بزرگی از وب را پردازش کرئه و نتایج جالب توجهی را تولید کنند که در غیر این صورت تولدی آنها غیر ممکن باشد. در مدت زمان کوتاهی سیستم به جایی رسید که تعداد زیادی مقاله و تحقیق با استفاده از پایگاه داده گ.گل ایجاد شده بودند و بسیاری دیگر، در دست اقدام هستند. هدف دیگر بوجود آوردن یک محیط لابراتوار مانند بود که محققان و حتی دانشجویان بتوانند تجربیات جالب و پیشنهادات مفیدی برروی داده های وب در مقیاس وسیع گوگل داشته باشند.

2. ویژگیهای سیستم

موتور جستجوی گوگل دو ویژگی مهم دارد که به تولید نتایج با وضوح و دقت بالا کمک می کند. اول، گوگل از ساختار پیوند وب برای محاسبه رتبه بندی کیفیت برای هر صفحه وب استفاده می کند. این رتبه بندی، رتبه صفحه نامیده می شود. دوم، گوگل از پیوند برای بهبود نتایج جستجو بهره می گیرد.

1.2- رتبه صفحه: نظم بخشیدن به وب

گراف فراخوانی (پیوند) وب یک منبع بسیار مهم است که توسط موتورهای جستجوی وب کنونی بی استفاده مانده است. گوگل نقشه هایی شامل بیش از یک میلیارد از این فقو پیوندها که نمونه ای چشمگیر از کل هسته را بوجود آورده است. این نقشه ها اجازه محاسبه سریع «رتبه صفحه» یک صفحه وب را می دهند، یک معیار عینی که اهمیت اشاره به آن برابر با تصویر ذهنی مردم از اهمیت است. بخاطر این تطابق، رتبه یک صفحه راه عالی برای اولویت دادن به نتایج جستجوهای کلمه کلیدی در وب. برای اغلب موضوعهای معروف یک نوشته ساده متناظر با جستجو است به این معنی که محدود به تیترهای صفحات باشد یعنی زمانی که نتایج جتوسط رتبه بندی صفحه اولویت بندی می شوند به طور قابل تحسینی اجرا می شوند. برای جستجوهای کاملاً متنی نیز در سیستم اصلی گوگل رتبه بندی صفحه کمک قابل ملاحظه ای می کند.

1.2.2. توصیف محاسبه رتبه صفحه

منابع نوشته آکادمیک در وب عمدتاً از طریق شمارش نوشته ها یا پیوندهای بازگشتی به یک صفحه خاص به کار گرفته شده اند. این کار تقریبی از اهمیت یا کیفیت صفحه به دست می دهد. رتبه بندی صفحه این مفهوم را از طریق نرمال سازی بوسیله تعداد پیوندها در یک صفحه و

نه شمارش پیوندها به طور مساوی در تمام صفحات، گسترش می دهد، رتبه بندی صفحه به صورت زیر تعریف می شود:

در نظر بگیرید که صفحات $T_1 \ldots T_N$ به صفحه a اشاره می کند (یعنی منبع هستند). پارامتر d یک گامل محدود ساز است که می تواند بین 0 تا 1 تنظیم شود و اغلب d با مقدار 0.85 تنظیم می شود. توضیحات بیشتر در مورد d در بخش بعید ارائه می شود. بنابراین C(A) به عنوان تعداد صفحاتی که از صفحه A خارج می شوند، تعریف می شود. رتبه صفحه A به صورت زیر داده می شود.

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

توجه کنید که رتبه های صفحه یک توضیح احتمالی برروی صفحات می دهد، بنابراین مجموع رتبه های تمام صفحات وب یک (1) خواهد بود.

رتبه صفحه یا PR(a) می تواند بوسیلهٔ یک الگوریتم تکرار ساده محاسبه شود و با بردار خاص اصلی از ماتریس پیوند نرمال شده از وب تطابق داده شود. بنابراین، رتبه بندی صفحه 26 میلیون صفحه وب می تواند در کمتر از چند ساعت برروی یک ایستگاه کاری متوسط محاسبه شود. بسیاری جزئیات دیگری هستند که از محدوده این مقاله خارج است.

2.1.2. توجیه شهودی

رتبه صفحه می تواند به عنوان یک مدل از رفتار عملکرد کاربر فرض شود. فرض می کنیم که یه «مرورگر تصادفی» وجود دارد چکه یک صفحه به طور تصادفی به او داده می شود و او برروی پیوندها کلیک می کند و هیچگاه دکمه (BACK) را نمی زند اما سرانجام خسته می شود و از یک صفحه تصادفی دیگر کار خود را ادامه می دهد. احتمال اینکه این مرورگر تصادفی یک صفحه را ملاقات کند رتبه آن صفحه می باشد و d یعنی عامل محدودساز احتمال این است که

آن «مرورگر تصادفی» از هر نسخهع خسته شود و تقاضای یک صفحه تصادفی دیگر بکند. تفاوت مهم این است که عامل محدودساز d را تنها یک صفحه، یا گروهی از صفحات اضافه کنیم. این کار امکان شخصی سازی را ایجاد می کند و تقریباً گمراه کردن عمدی سیستم به منظور بدست آوردن یک رتبه بالاتر را غیرممکن می سازد. گوگل انشعابات متعدد دیگری برای رتبه بندی صفحه دارد که از محدوده این نوشته خارج است.

توجیه شهودی دیگر این است که یک صفحه می توان یک رتبه صفحه بالا داشته باشد اگر صفحات زیادی به آن اشاره کنند یا صفحاتی وجود دارند که به آن اشاره می کنند و خود رتبه صفحه بالایی دارند. به ضوح، صفحاتی که به خوبی از جاهای محتلفی از وب تکرار می شوند ارزش نگاه کردن دارند. همچنین، صفحاتی که ممکن است یک احضار از طرف جایی مانند صفحه خانگی یاهوا داشته باشند عموماً ارزش نگاه کردن دارند. اگر یک صفحه کیفیت بالایی نداشته باشد یا یک پیوند شکسته شده باشد به احتمال زیاد صفحه خانگی یاهوا! به آن پیوند نمی شود. ضمناً رتبه بندی صفحه هر دوی این حالات و حالات دیگر را با وزن دهی تبلیغی به طور بازگشتی از طریق ساختار پیوند وب انجام می دهد.

2.2. متن انکر (Anchor)

در موتور جستجوی گوگل با نوشتهٔ پوندها به شیوه های خاصی برخورد می شود. اغلب موتورهای جستجو نوشته یک پویند را به صفحه ای که پیوند در آن است مربوط می سازند. گ.گل علاوه بر این نوشته پیوند را به صفحه ای که به آن اشاره می کند نمیز مربوط می سازد. این کار منافع زیادی دارد. اول، انکرها اغلب توصیف دقیق تری از صفحات وب نسبت به خود صفحات ارائه می دهند. دوم، انکرها ممکن است برای سندهایی که نمی توانند توسط موتورهای جستجوی بر پایه

متن شاخص بندی شوند وجود داشته باشندذ. مانند عکسها، برنامه ها، و پایگاه ها داده. این کار در حقیقت امکان بازگرداندن صفحاتی را که عمل جستجو و دانلود (Crawl) برروی آنها صورت نگرفته است را می دهد. توجه کنید که صفحاتی که عمل جستجو و دانلود برروی آنها صورت نگرفته است می توانند ایجاد مشکل کنند از آنجا که آنها هیچ گاه برای صحت و اعتبار منطقی قبل از برگردانده شدن به کاربر چک نمی شود. در این حالت موتور جستجو حتی می تواند صفحه ای را که اصلاً وجود ندارد اما فوق پیوندها به آن اشاره می کنند بازگرداند. اگرچه امکان دسته بندی نتایج وجوود دارد درنتیجه این مشکل خاص به ندرت اتفاق می افند.

ایده متن انکر تبلیغاتی به صفحه ای که به آن باز می گرئ توسط کرم وب گسترده جهانی (WWWW) تحقق پیدا کرد. زیرا این متن به جستجوی اطلاعات غیرمتنی و گسترش دامنه جستجو با سندهای دانلودی کمتر کمک می کند. گوگل به این دلیل از انکر تبلیغاتی استفاده می کند که متن انکر می تواند در فراهم کردن کیفیت بهتر نتایج کمک کند. استفاده مفید از متن انکه به دلیل حجم بالای که باید پردازش شود از نظر تکنیکی مشکل است. در مجموعه جستوجو و یافته شده حال حاضر گوگل که شامل 240 میلیون صفحه است بیش از دو و نیم میلیارد انکر شاخص بندی شده وجود دارد.

3.2. ویژگیهای دیگر

جدار از رتبه صفحه (PageRank) و استفاده از متن انکر، گکوگل ویژگیهای متعدد دیگری دارد. اول، اطلاعات مکانی تمام بهترینها (Hits) را دارد و بنابراین استفاده وسیعی از اطلاعات مجاورتی در جستجو می کند. دوم، گوگل جزئیات بعضی بخشهای دیداری مانند اندازه فونتهای

کلمات را نگهداری می کند. به کلماتی که بزرگتر نوشته شده اند یا پررنگتر هستند وزن بالاتری داده می شود. سوم، HTML کل و خام هر صفحه در انباره موجود می باشد.

3. کارهای مربوطه

تحقیقات جستجو برروی وب تاریخچه کوتاه و موجزی دارد. کرم وب جهانی (wwww) یکی از اولین موتورهای جستجو وب بوده است. این حرکت متعاقباً توسط موتورهای جستجوی آکادمیک متعددی دنبال شد که بسیاری از آنها هم اکنون تبدیل به شرکتهای تجاری شده اند. در مقایسه با گسترش وب و اهمیت موتورهای جستجو سمدهای اندکی در مورد موتورهای جستجو اخیر وجود دارد. به عقیده مایکل ماولدین (سرمحقق شرکت Lycos)، سرویسهای مختلف (شامل Lycos) به سختی از جزئیات پایگاه داده هایشان محافظ می کنند. اگرچه کار قابل توجهی برروی ویژگیهای خاصی از موتورهای جستجو صورت گرفته است. به خصوص کار و تحقیقی که بیشتر نمودار است و بارز است کاری است که برروی عملیات بعد از پردازش برای بدست آوردن نتایج در موتورهای جستجوی تجاری فعلی صورت گرفته است و در ایجاد موتورهای جستجوی در مقیاس کوچک «شخص شده» کاربرد دار. در نهایت تحقیقات زیاد چبرروی سیستمهای بازیافت اطلاعات صورت گرفته است به خصوص بر مجموعه هایی که نظارت درستی بر آنها اعمال می شود.

1.3. بازیافت اطلاعات

کار برروی سیستم های بازیافت اطلاعات به سالها قبل باز می گردد و پیشرفت قابل توجهی کردهخ است. اگرچه، اغلب تحقیقات برروی سیستم های بازیافت بروی مجموعه های کوچک و همگن به خوبی کنترل شده صورت گرفته است مانند مجموعه های مقالات علمی یا داستانهای

اخباری برروی موضوع قابت و به همین صورت. آزمایش کارایی (benchmark) اولیه بازیافت اطلاعات، کنفرانس بازیافت متن، از یک مجموعه واقعاً کوچک و کاملاً کنترل شده برای سنجش مارائیهایش استفاده می کرده است. مثیزان کارایی کوپوس بسیار بزرگ {TREC96] تنها ۲۰ گیگابایت است کیسه با ۱۴۷ گیگابایت جستجو و یافته شده از 240 میلیون صفحه وب گوگل بسیار محدود است. مواردی که برروی TREC به خوبی کار می کنند اغلب برروی وب نتایج مناسبی ایجاد نمی کنند. برای مثال بردار استاندارد مدلب فضا سعی در بازگرداندن مشابه ترین سندها به پرس و جو را دارد* با در نظر گرفتن اینکه هر دوی پرس و جو و سند بردارهایی تعریف شده بر اساس کاربرد کلمه هستند. اما این استراتژی برروی اغلب سندها ببسیار کوتاه را بر می گرداند که در خقیقتا خودپرس و جو به اضافه چند کلمه محدذود هستند. فی المثل ما شاهد بودیم که یک موتور جستجوی مهم صفحه ای را شامل چند جملهٔ «جورج بوش کندزد» و تصویری از پرس و جوی «جورج بوش» «جرج بوش» برگردانده است.بعضی ها استدلال می کنند که کاربران برروی وب باید چیزی را که می خواهند دقیق تر مشخص کنند و در حقیقت کلمات بیشتری به پرس و جوهایی که ایجاد می کنند ، اضافه کنند. گوگلی به شدت به شدت با این نظر مخالف است. اگر کاربری پرس و جویی مانند «جورجو بوش» را صادر کند ، آنها باید تا زمانی که حجم بالایی از طلاعات در دسترس با کیفیت بالا برروی این موضوع وجود دارد ، نتایج معقولی برگردانند. با توجه به مثالهایی اینچنین، ما باور داریم که استاندارد بازیافت اطلاعات برای تقابل بهتر با وب نیاز به گسترش فراوانی دارد.

2.3.2.3. تفاوتهای وب با مجموعه های کنترل شده

وب مجموعه ای از سندهای کاملاً نامتجانس و کنترل نشده است. اسناد موجود برروی وب از نظر شکل داخلی و همچنین فرااطلاعات خارجی موجود تفاوتهای فراوانی دارند. برای مثال، استاندارد از نظر داخلی تفاوتهایی مانند زبان ایجاد (هر دو حالت انسانی و برنامه نویسی)، اصطلاحات واژگان (آدرسهای ایمکیل، پیوندها، کدهای آدرس، شماره های تلفن، شماره های تولیدات)، نوع یا فرمت (متن، Html، PDF، تصویر، صدا) دارند و حتی ممکن است تولیدات ماشینی باشند (فایلهای گزارشی یا خروجی یک پایگاه داده). از طرف دیگر، فرا اطلاعات خارجی به عنوان اطلاعات نتیجه گرفته شده از یک سند تعریف می شوند، اما شامل اطلاعات درونی آن نیستند. مثالهای فرا اطلاعات خارجی شامل مواتردی مانند اعتبار و شهرت منبع، تناوب به روز رسانی، کیفیت، تعداد دفعات اجرا و منابع استناد است. نه تنها منابع ممکن فرا اطلاعات خارجی تنفاوت دارند بلکه مواردی که شامل تفاوت می شوند بسیار گوناگون هستند. برای مثال، اطلاعات استعمال یک صفحه خانگی مهم مانند صفحه خانگی یاهو را که میلیونها بازدید را در حال حاضر دریافت می کند با اطلاعات استعمال یک مقاله گمنام تاریخی که ممکن است هر ده سال یکبار بازدید شود مقایسه کنید. مسلماً این دو مورد باید به نوع متفاوتی در موتورهاید جستجو برخورد شوند.

تفاوت بزرگ دیگر بین وب و مجموعه های به خبی کنترل شده قدیمی این است که به طور منطقی کنترلی بر این که مردن چه چیزی برروی وب قرار می هند وجود ندارد. انعطاف پذیری تولید تمام متنهای دلخواه را با نفوذ شدید موتورهای جستجو ترکیب کنید تا قدرت هدایت ترافیک به مسیری خاص توسط شرکتهایی که برای سود بیشتر نتایج موتورهیا جستجو را دستکاری می کنند، بدست آید. که تبدیل به مشکل بزرگی شده است. این مشکل در سیستمهای بازیافت اطلاعات قدیمی مورد توجه قرار نگرفته بود. همچنین جالب است اشاره شود

که حاصل کار فرا داده برای موتورهای جستجو به طور عمده غیر قابل استفاده و شکست خورده است. دلیل این امر سوء استفاده از هر نوع متن در صفحات وب است که به طور غیر مستقیم به کاربر ارائه شده باشد به منظور دستکاری در موتورهای جستجو حتی شرکتهای متعددی وجود دارمد که در زمینه دستکاری در موتورهای جستجو برای سود بیشتر تخصص دارند.

4. آناتومی سیستم

در ابتدا یک مباحثه سطح بالا از معماری سیتم ارائه می شود. سپس توصیفی عمقی از ساختمان داده هاتی مهم سیستم خواهیم داشت. در نهایت، بخشهای کاربلردی مهم مانند: جستجو دانلود (Vrawling) ، شاخص بندی به طور عمقی توضیح داده می شوند.

1.4. نگاهی کلی به معماری گوگل

در این بخش، یک نکاه اجمالی سطح بالا به عملکرد سیستم همان طور که در شکل ۱ نشان داده شده است خواهیم داشت. بخشهای بعدی برنامه های کاربردی و ساختمان داده هیا اشاره نشده در این بخش را توصیف می کنند. اغلب تستهای گوگل با C و C++ برنامه ریزی شده است به دلیل بازدهی بهتر و امکان اجرا برروی هر دو سیستم لینوکس و سولاریس.

در گوگل، عمل Crawling (دانلود کردن صفحات وب) وب توسط برنامه های جستجو کننده و یابنده (Crawler) متعدد دستبندی شده صورت می گیرد. یک سرویس دهنده URL (URL server) وجود دارد که لیستهای URL ها را جهت واکنشی به Crawler می فرستد. صفحه های وب که واکنشی شدند به سرویس دهنده انباره فرستاده می شوند. سپس صفحه های وب توسط سرویس دهنده انباره فرشده می شود و درون مخزن قرار می گیرند. هر صفحه وب یک شماره شناسه مربوطه دارد که docID نامیده یم شود و زمانی به آن اختصاص داده می شود که

URL جدید از یک صفحه وب تجزیه و استخراج می شود. تابع شاخص بند و دسته بند اجرا می شود. شاخص بند تعدادی تابع را اجرا می کند، مخزن را می خواند، اسناد را از حالت فشرده خارج، و تجزیه می کند. هر سند به مجموعه ای از رویدارهای کلمه تبدیل می شود که هر کدام از آنها «بهترینها» نام دارد. بهترینها خود کلمه، مکان در سند، تقریبی از اندازه فونت و حالت بزرگ نویسی را ذخیره می کند. شاخص بند تمام بهترینها را درون مجموعه ای از «مخزنه» توزیع می کند و یک شاخص پیشرفته و مرتب شده را ایجاد می کند. شاخص بند یک کار مهم دیگر را نیز انجام می دهد، تمام پیوندهای موجود در هر صفحه وب را تجزیه و استخراج می کند و اطلاعات مهم مربوط به آنها را درون یک فایل انکر ذخیره می سازد. این فایل حاوی اطلاعات کافی برای تشخیص مکانی که هر پیوند به آن اشاره می کند و یا از آن اشاره می شود، و همچنین نوشته پویند می باشد.

تجزیه گر URL فایل انرک را می خواند و URL های مربوط را به URL های قطعی و کامل تبدیل می کند و در نهایت docID ها را می سازد. نوشته انکر را درون شاخص پیشرو قرار می دهد که وابسته docID ای است که انکر به ان اشاره می کند. همچنین پایگاه داده ای از پیوندها که در حقیقت جفتهایی از docID هستند را تولید می کند. پایگاه داده پیوندها برای محاسبه رتبه صفحه تمام اسناد بکار می رود.

دسته بندی مخازم را که بر اساس docID مرتب شده اند می گیرد (این یک کثال ساده است، به بخش 5.4..2 مراجعه کنید) و آنها را بر اساس کلمه (wordID) دوباره مرتب سازی می کند و با این کار شاخص معکوس را تولید می کند. این کار به صورت درجا صورت می گیرد درنتیجه به فضای موقت اندکی برای انجام این عملیات نیاز داریم. دسته بند همچنین یک لیست از شناسه های کلمه و آفستها ایجاد می کند و از آنها برای تولید شاخص معکوس کمک می گیرد.

یک برنامه به نام «روبرداشت واژگان» این لیست را با واژه نامه تولید شده توسط شاخص بند با هم می گیرد تا یک واژه نامه جدید که توسط جستجوگر مورد استفاده قرار می گیرد را تولید کند. جستجوگر توسط یک سرویس دهنده وب اجرا می شود و از واژه نامه تولید شده توسط روبرداشت واژگان و از شاخص معکوس و رتبه صفحه با هم برای پلسخگویی به پرس و جو ها استفاده می کند.

2.4. ساختمان داده های مهم

ساختمان داده های گوگل بهینه شده هستند بنابراین یک مجموعه سند بزرگ می تواند با هزینه ای کم دنبال گشته و دانلود شود، شاخص بندی شود و در نهایت مورد جستجو قرار گیرد. اگرچه، cpu ها و میزان سرعت ورودی و خروجی انبوه به طور چشمگیری در سالهای اخیر بهبود یافته اند، زمان استوانه جویی در دیسک هنوز به حدود ۱۰MS زمان برای کامل شدن احتیاج دارد. گویی به گونه ذی طراحی شده است که تا جای ممکن از استوانه جویی در دیسک اجتناب کند و این کار تاثیر قابل ملاحظه ای برروی طراحی ساختمانهای داده داشته است.

1.2.4. فایلهای بزرگ

فایلهای بزرگ (Big Files) فایلهای مجازی هستند که در طول سیستمهای فایل چند گانه گسترش داده شده اند و قابل آدرس دهی به صورت ۶۴ بیتی هستند. تخصیص حافظه بین سیستمهای فایل چندگانه به صورت اتوماتیک اداره می شود. بسته فایلهای بزرگ همچنینی تخصیص و بازپس گیری حافظه از توصیفگر فایل را بر عهده دارد و این کار از آنجا صورت می گیرد که سیتمهای عامل نیازهای سیستم گوگل را برطرف نمی کنند. فایلهای بزرگ همچنین گزینه های مقدماتی فشرده سازی را پشتیبانی می کنند.

2.2.4. مخزن

مخزن، HTML کامل هر صفحه وب را شامل می شود. هر صفحه با استفاده از (RFC 1950) zlib فشرده می شود . انتخاب تکنیک مورد استفاده گوگل در فشرده سازی توازنی است بین سرعت و درجه فشرده سازی. گوگل سرعت zlib را به همراه بهبود چشمگیر در فشرده سازی که توسط bzip ارائه می شود، انتخاب کرده است. درجه فشرده سازی bzip را به همراه بهبود چشمگیر در فشرده سازی که توسط bzip ارائه می وشد، انتخاب کرده است. درجه فشرده سازی bzip به وطر تقریبی ۴ به ۱ می باشد. که در مقایسه با فشرده سازی ۳ به ۱ zlib برروی مخزن بهینه می باشد. در مخزن سندها به صورت پی در پی ذخیره می شوند و بر اساس docID، طول و URL عنوان بندی می شوند (شکل ۲). مخزن به هیچ نوع ساختملن دادهٔ دیگری که به منظور دستیابی به آن مورد استفاده قرار گیرد، نیاز ندارد. این حالت به سازگاری این ساختمان داده کمک می کند و گسترش آن را نیز ساده می سازد. بنابراین گوگل می تواند تمامی ساختمان داده های دیگر را تنها از مخزن و یک فایل که شامل خطاهای Crawler است بازسازی کند.

3.2.4. شاخص سند

شاخص سند اطلاعات مربوط به هر سند را نگهداری می کند. این شاخص ISAM است که گستردگی اصلاح شده دارد و بر اساس docID مرتب شده است. اطلاعات ذخیره شده در هر مدخل شامل وضعیت شند، یک اشاره گر به مخزن، یکگ جمع مقابله ای از سند و آمارهای مختلف است. سند جستجو دانلود شده شامل یک اشلره گر به یک فایل گسترده متغیر که docinfo (اطلاعات سند) نامیده می شود و خود URL آن سند و تیتر آن را در بر دارد می

باشد. در غیر این صورت آن اشاره گر به یک لیست URL که تنها شامل URL مورد نظر می باشد اشاره می کند. این نوع طراحی به جهت فراهم آوردن یک ساختمان داده فشرده معقول و همچنین ایجاد قابلیت واکنشی یک رکورد در یک استوانه جویی دیسک برای هر جستجو اتخاذ شده است.

به علاوه فایلی وجود دارد که برای تبدیل URL ها به docID مورد استفاده قرار می گیرد. و شامل لیستی از جمعهای مقابله ای URL می باشد به همراه docID های معادل آنها و بر اساس جمع مقابله ای مرتب شده است. به منظور یافتن docID یک URL خاص، جمع مقابله ای آن URL محاسبه می شود و یک جستجوی دودویی برروی فایل جمعهای مقابله ای صورت می گیرد تا docID آن پیدا می شود.

URLها ممکن است با انجام یک الگوریتم ادغام با فایل جچمع مقابله به صورت گروهی به docID ها تبدیل شوند. این تکنیکی است که تجزیه گر URL برای تبدیل URL ها به docID ها مورد استفاده قرار می دهد. این حالت به روزرسانی گروهی بسیار مهم است زیرا در غیر این صورت باید برای هر پیوند یک استوانه جویی انجام شود که در این صورت چمع آوری یک مجموعه داده ۳۰۰ میلیونی برروی یک دیسک بیشتر از یک ماه طول خواهد کشید.

4.2.4 واژه نامه

واژه نامه اشکال گوناگونی دارد. مهم ترین تغییر نسبت به سیستم های اولیه این است که با صرف هزینه ای معقول می توان واژه نامه را در حافظه جا دادا. در شیوه اجرایی جاری می توان واژه نامه را برروی حافظه اصلی ۲۵۶ مگابایتی یک سیتم نگهداری کرد. .واژه نامه فعلی شامل ۱۴ میلیون کلمه می باشد (البته بعضی کلمات نادر به واژه نامه اضافه نشده اند). واژه نامه در دو

بخش عملی می شود – لیستی از کلمات (ظاهراً بهم پیوسته اند اما بوسیله کاراکترهای null از هم جدا شده اند) و یک جدول هش از اشاره گرها، برای کارهای مختلف، لیست کلمات اطلاعات کمکی دیگری نیز دارد که توضیح آنها خارج از محدوده این مقاله است.

5.2.4. لیستهای بهترینها

یک لیست بهترینها معادل لیستی است از رویدادهای یک کلمهٔ خاص در یک سند خاص به همراه اطلاعات موقعیت، فونت و بزرگ نویسی. اکثر فضای اشغال شده توسط هر دوی شاخصهای پیشرو و معکوس مربوط به لیستهای بهترینها می باشد. به همین دلیل لازم است آنها را تا حد ممکن مؤثر و کار را ارائه کنیم. گوگل انتخابها و جانشینهای متعددی برای کد کردن موقعیت، فونت و بزرگ نویسی در نظر گرفته است - - کد بندی ساده (اعداد سه تایی)، کدبندی فشرده (تخصیص بهینه شده بیتها به صورت دستی) و کدبندی هاف من. در نهایت گوگل از یک نوه کدبندی فشرده بهینه دستی استفاده می کند به این دلیل که به فضای کمتری نسبت به کدبندی ساده و دساکاری بیتی بسیار کمتری نسبت به کدبندی هافمن احتیاج دارد. جزئیات بهترینها در شکل ۳ نشان داده شده است.

کدبندی فشرده گوگل از دو بایت برای هر یک از بهترینها استفاده می کند. دو نوع بهترین وجود دارد: بهترینهای شگفت و بهترینهای آشکار. بهترینهای شگفت شامل بهترینهایی است که در یک URL، تیتر، نوشته انکر یا فوق تک ظاهر می شوند. بهترینهای آشکار شامل بقیه موارد می شود. یک بهترین آشکار شامل یک بیت بزرگ نویسی، اندازه فونت و ۱۲ بیت برای موقعیت کلمه

در یک سند (تمام موقعیتهای بالای ۴۰۹۶ با ۴۰۹۶ نشان داده می شوند) می شود. اندازه فونت که در مقایسه با بقیه سند مربوطه نشان داده می شود با استفاده از ۳ بیت صورت می گیرد (تنها تا عدد ۷ را می توان برای اندازه فونت استافده کرد زیرا 111 نشانه نمایی است کهع وجود یک بهترین شگفت را نشان می دهد). یک بهترین شگفت شامل بیت بزرگ نویسی، اندازه فونت تنظیم شده در عدد ۷ که نشان دهنده یک بهترین شگفت است، ۴ بیت برای کدبندی نوع بهترین شگفت و ۸ بیت برای موقعیت. برای بهترینهای انکر، ۸ بیت مخصوص موقعیت به ۴ بیت برای موقعیت در انکر و ۴ بیت برای یک هش از docID مه انکر در آن واقع شده است، تقسیم می شود. این کار قابلیت جستجوی محدود برروی عبارات را تا زمانی که تعداد بسیار زیاید انکر برای یک کلمهٔ خاص وجود نداشته باشد به ما می دهد. گوگل همواره شیوه ذخیره سازی بهترینهای انکر را به منظور دستیابی به وضح و دقت بیشتر در زمینه های موقعیت و هش docID به روز رسانی می کند. گوگل اندازه فونت را در مقایسه با بقیه سند مورد استفاده قرار می دهد زیرا در هنگام جستجو کسی انتظار ندارد که سندهای همانند و یکسان به صورت متفاوت رتبه بندی شوند تنها به این دلیل که اندازه فوت یکی از آنها بزرگتر است.

طول یک لیست بهترینها قبل از خود بهترینها ذخیره می شود. برای صرفه جویی در فضا، طول لیست بهترینها را به ۸ و ۵ بیت محدود می سازد (حقه هایی وجود دارد که اجازه می دهد ۸ بیت از شناسه کلمه قرض گرفته شود). اگر طول لیست بیشتر از آن چیزی باشد که بتواند در این تعداد بیت جای گیرد از یک کد فرار در آن بیتها استفاده میشود و دو بایت بعدی شامل طول واقعی لیست خواهند بود.

6.2.4. شاخصهای پیشرو

شاخص پیشرو در واقع از قبل ساخته شده است. این شاخص در تعدادی مخزن ذخیره شده است (گوگل از 64 مخزن استفاده می کند). هر مخزن یک رنج از شناسه هیا کلمات را نگهداری می کند اگر یک سند شامل کلماتی باشد که در مخزن خاصی ریخته شده باشند، شناسه آن سند در آن مخزن ذخیره می شود. این طرح و شما به اندکی فضای ذخیره سازی بیشتر احتیاج دارد که به دلیل شناسه های سند تکراری رخ می دهد. اما این تفاوت برای تعداد معقولی از مخازن بسیار اندک است و باعث صرفه جویی در زمان قابل ملاحظه ای می شود و از پیچیدگی کد بندی در فاز آخر شاخص بندی که توسط برنامه مرتب ساز صورت می گیرد می کاهد. علاوه بر این، به جای ذخیره محض شناسه های کلنات شند، گوگل هر شناسه که به عنوان یک تفاوت مناسب از کمترین شناسه که در آن مخزن شناسه کلمه مورد نظر در آن قرار دارد، ذخیرهع می کند. با این روش، سیستم گوگل تنها از ۲۴ بایت برای شناسه های کلمات در مخازن مرتب نشده استفاده می کند و ۸ بیت دیگر را برای ذخیره طول لیست بهترینها باقی می گذارد.

7.2.4. شاخص معکوس

شاخص معکوس همان مخازن شاخص پیشرو است، تفاوت آنها در این است که مخازن توسط ترتیب بند مرتب شده اند. برای هر شناسه که معتبر، واژه نامه اشاره گری دارد که به مخزنی که شناسه کلمه در آن قرار دارد اشاره می کند. و درون مخزن به یک لیست سند که متشکل شده اند از docID (شناسه سند) و لیست بهترینهای معادل آن است، اشاره می کند. این لیست سند تمام موارد موجود بودن و در حقیقت رویدادهای کلمه را در تمام سندها ارائه می دهد.

مسئله مهم چگونگی قرارگیری و ظاهر شدن docID ها در لیست سند است. یک راه حل ساده ذخیره کردن سند به صورت مرتب شده بر اساس docID است. این کار اجازه ادغام سریع

لیستهای مختلف سند را با هم برای پرس و جوهای چند کلمه ای می دهد. حالت دیگر ذخیره کردن لیست سندها به صورت مکرتب شده بر اساس رتبه ای از رویدادهای هر سند است.

این کار پرس و جوهای تک کلمه ای را بی مایه و پیش پا افتاده می کند و پاسخ گویی به پرس و جوهای چند کلمه ای را تا حد کمال بالا می برد. اگرچه ادغام کردن در این حالت بسیار مشکل تر است. بنابراین، این کار گسترش و پیشرفت مجموعه را بسیار مشکل می سازد، زیرا هر تغییر در عملیات رتبه بندی احتیاج به ساخت دوباره شاخص دارد. گوگل حد وسطی از این دو حالت را انتخاب کرده است به این صورت که دو مجموعه از مخازن معکوس را نگهداری می کند - - یک مجموعه برای لیستهای بهترینها که شامل تیتر یا بهترینهای انکر است و مجموعه دیگر برای تمام لیستهای بهترینها. به این صورت، گوگل ابتدا اولین مجموعه مخازن را چک می کند و اگر جفتها معادلهایی کافی در آن وجود نداشت آنگاه به سراغ مجموعهٔ بزرگتر می رود.

3.4. جستجو و دانلود کردن وب

راه اندازی یک Crawler وب وظیفه ای چالش آور است. در این زمینه مسائل قابلیت اعتماد و اعتبار و عملکرد گول زنک سایتها و صفحات وب و حتی مهم تر از آن مسائل قانوین و اجتماعی مؤثر هستند عمل Crowling ظریف ترین کار سیستم است از آنجائیکه باید صدها هزار سویس دهنده وب و سرویس دنده های نام مختلف تقابل داشته باشد که تمام آنها خارج از کنترل سیستم هستند.

به منظور مقیاس بندی صدها میلیون صفحه وب، گوگل سیستم Crawling سریع و توزیع شده ای دارد. یک سروی دهنده URL لیستی شمال URLها را در اختیار تعدادی از Crawler ها قرار می دهد (گوگل معمولاً حدود ۳ تا راه اندازی می کند). هر دوی سرویس

دهنده URL و Crawler توسط زبان Python عملی می شوندو هر Crawler به تخمین حدود ۳۰۰ ارتباط باز را یک جا نگهداری می کند. بازیابی صفحات وب لازم است که با سرعت بالایی صورت گیرد. در اوج سرعت، سیستم گوگل می تواند بیش از ۱۰۰ صفحه وب را در ثانیه با استفاده از چهار Craqler، جستجو و دانلود کند. این حجم به طور تقریبی ۶۰۰ کیلوبایت داده در ثانیه می باشد. یکی از تأکیدهای عملیاتی مهم یافتن DNS است. هر Crawler حافظه نهان DNS خودش را نگهداری می کند. بنابراین قبل از جستجو و دانلود هر سند احتیاجی به یافتن DNS ندارد. هر کدام از صدها ارتباط یافته شده می توانند در وضعیتهای مختلف باشد. مانند DNS، ارتباط به میزبان، فرستادن درخواست و دریافت پاسخ. این عوامل Crawler را تبدیل به یک جزء پیچیدهد از سیستم یم سازد. Cvrawler از IO ناهمگام برای کنترل رویدادها و از تعدادی صف برای جابجایی واکنشهای صفحات در حالتی به حالت دیگر استفاده می کند.

به نظر می رسد راه اندازی یک Crawler که به بیش از نیم میلیون سرویس دهنده متصل است و دهها میلیون مدخل گزارشی تو.لید می کند خود باعث تولید تعداد قابل ملاحظه ای نامه الکترونیکی و تماس تلفنی می شود. به خاطر خیل عظیم مردمی که همه روزه بر خط می شوند، همیشه آنهایی هستند که نمی دانند Crawler چیست زیرا برای اولین بار است که آن را می بینند. تقرذیباً همه روزه نامه الکترونیکی دریافت می کنیم که شامل جملاتی مانند «بسیار عالی، شما صفحات زیادی از سایت مرا نگاه کردید. به نظرتان چطور بود؟» می باشد. همچنین مردمی هستند که چیزی در مورد پروتکل ربات مانع نمی دانند و فکر می کنند صفحه آنها باید با جمله ای مانند، «این صفحه کپی رایت شده است و نباید شاخص بندی شود.» در مقابل شاخص بندی شدن محافظت شودو که البته لازم به گفتن نیست که درک آن برای Crawler مشکل است.

همچنین به خاطر حجم بالای اطلاعات درگیر در این کار حوادث غیرمنتظره ای رخ می دهند. برای مثال، سیستم گوگل با یک بار تلاش می کند تا یک بازی بر خط را جستجو دانلود کند. این کار پیغامی آشغال زیادی را در حین بازی ایجاد کرد! به نظر می رسد که این مشکل به سادگی حل شود. اما این مشکل تا زمانی که دهها میلیون صفحه دانلود نشده بود خود را نشان نداد. بهع خاطر نوسانات و تغیریات وسیع در صفحات وب و سرویس دهنده ها، تست کردن یک Crawler بدون راه اندازی آن در قسمت بزرگی از اینترنت، به صورت مجازی غیرممکن می باشد. همواره صدها مشکل پیچیده و گنگ وجود دارد که ممکن است تنها برروی یک صفحه از تمام وب رخ دهند و موجب از کار افتادن Crawler یا بدتر از آن موجب وقوع رفتار و عکس العمل غیرمنتظره با غلط شوند. سیستمهایی که به قسمتهیا بزرگی از اینترنت دسترسی دارند لازم است به گونه ای طراحی شوند که قابلیت ادامه کار در شرایط غیرمنتظره را داشته باشند و همچنین به خوبی تست شوند. تا زمانی که سیستمهای پیچیدع بزرگ مانند Crawler پویشته مشکل ایجاد می کنند، لازم است منابع قابل ملاحظه ای به بررسی نامه های الکترونیکی و حل این گونه مشکلها اختصاص داده شود.

4.4. شاخص بندی وب

تجزیه کردن - - تمام تجزیه گرهایی که برای اجرا برروی تمام وب طراحی شده اند باید یک آرایه عظیم از خطاهای ممکن را اداره کنند. رنج این خطاها از خطای تایپی در تکهای HTML تا مجموعه کیلوبایتهای صفر در وسط یک تک تا کاراکترهای غیر اسکی تا تگهای HTML که صدها بار تودرتو هستند و مجموعه عظیمی از خطاهای دیگر که ذهن هر کسی را به چالش می کشد و همسان با آن خطاهای نادر را که در موارد خاص ایجاد می شوند را شامل می شود. برای

بدست آوردن بیشترین سرعت گوگل از یک تحلیل گر واژگانی که با پشته خودش مجهز شده است استفاده می کند. گسترش و پیشرفت این تحلیل گر که با سرعتی محقول کار می کند و قابلیت ادامه کار در شرایط غیر منتظره را دارد کار بسیار زیاد برده است.

شاخص زنی سندها به مخازن – بعد از تجزیه هر سند، آن سند به صورت تعدادی مخزن کدبندی می شود. هر کلمه به یک شناسه کلمه با استفاده از یک جدول هش درون حافظه یا – واژه نامه—تبدیل می شود. کلمات جدید اضافخ شده به جدول هش واژه نامه درون یک فایل گزارش داده می شوند. هنگامی که کلمات به شناسه تبدیل شدند، رویدادهای آنها در سند حاضر به صورت لیستهای بهترینها ترجمه می شود و درون مخازن تبدیل پیشرو نوشته می شوند. مشکل اصلی در موازی سازی فاز شاخص بندی این است که واژه باید به صورت اشتراکی درآید. به جای اشتراکی کردن واژه نامه، گوگل به روشی دست یافته است که در آن گزارش از تمام کلمات اضافی که در واژه نامه پایه نیستند را به دست می دهد که در گوگل تعداد آنها در ۱۴ میلیون ثابت شده است. به این صورت چندین شاخص بند می تواند به صورت موازی احرا شوند و در نهایت آن فایل گزارشی از کلمات اضافی می توانند توسط آخرین بند پردازش شود.

* مرتب سازی – به منظور تولید شاخص معکوس، برنامه مرتب ساز هر کدام از مخازن پیشرو را می گیرد و آن را بر اساس کلمه مرتب می کند. تا یک مخزن معکوسی برای تیتر و بهترینهای انکر و یک مخزن معکوسی برای متن بدست آید. این پردازه به صورت تنها یک مخزن در آن واحد صورت می گیرد در نتیجه به فضای موقتی ذخیرع سازی کمی احتیاج دارد. همچنین گوگل فاز مرتب سازی را به صورت موازی در می آورد تا از تمام ماشینهایی که دارد تنها با راه اندازی چندین مرتب ساز که می توانند در آن واحد مخزنهای متفاوتی را پردازش کنند استفاده کند. هنگامی که مخازن در حافظه اصلی جا نمی گیرد، برنامه مرتب ساز آنها را هرچه بیشتر به

سیدهایی که حتماً در حافظهع جا بگیرد تقسیم می کند که این سبدها بر پایه شناسه کلمه و docID (شناسه سند) هستند. سپس مرتب ساز هر کدام از این سبدها را درون حافظه بارگذاری می کند، آن را مرتب می کند و محتوایش را درون مخزن معکوس خلاصه و مخزن معکوس کامل می نویسد.

5.4. جستجو کردن

هدف از جستجو فراهم آوردن نتایج جستجوی با کیفیت به طور مؤثر است. بسیاری از موتورهای جستجوی تجایر بزرگ در زمینه بازرهی و راندمان توسعه و پیشرفت زیادی داشته اند. بنابراین در حقیقت گوگل تمرکز بیشتر برروی کیفیت جستجو بوده است. اگرچه طراحان گوگل باور دانرد که راه حلهای آنها می توانتند با اندکی تلاش بیشتر در مقیاس تجاری قابل انطباق باشند. ارزیابی پروسه پرس و جوی گوگل در شکل ۴ نشان داده شده است.

۱. پرس و جو را تجزیه کن
۲. کلمات را به شناسه های کلمه تبدیل کن
۳. از اول لیست سند در مخزن کوتاه جستجو را برای هر کلمه شرع کن.
۴. عمل مرور را در طول لیستهای سنمد تا زمانی که سندی پیدا شود که با تمام شرایط جستوجو همانگ باشد ادامه بده
۵. برای پرس و جو رتبه سند مورد نظر را محاسبه کن
۶. اگر در مخزن گوناه هستیم و در انتهای لیست سند هستیم ابتدای لیست سند را در مخزن کامل جستجو کن

۷. اگر در انتهای هر لیست سند نیستیم به مرحله ۴ برو.

۸. سندهایی را که توسط سیستم رتبه بندی پیدا شده اند k تای اولشان را برگردان

شکل ۴ – ارزیابی سیستم پرس و جوی گوگل

برای قرار دادن یک محدودیت بر زمان پاسخگویی، هنگامی که تعداد معینی (در حال حاضر ۴۰۰۰) از سندهای هماهنگ یافت شدند، جستجوگر به صورت اتوماتیک به مرحله ۸ شمل ۴ می رود. این کار به این معنی است که احتمالاً زیر مجموعه ای از مناسب ترین و بهترین نتایج برگردانده می شوند. گوگل در حال مطالعه راه حلهایی برای حل این مشکل است. در گذشته گوگل نتایج را بر اساس رتبه صفحه مرتب می کرد که به نظر می رسید که نتایج را بهبود می دهد.

1.5.4. سیستم رتبه بندی

گوگل نسبت به سایر موتورهای جستجوی معمول اطلاعات بیشتری در مورد سندهای وب نگهداری می کند. هر لیست از بهترینها شامل اطلاعات موقعیت، فونت و بزرگ نویسی می شود. علاوه بر این، گوگل بهترینها را بر اساس نوشته انکر و رتبه صفحه درجه بندی می کند. ترکیب تمام این اطلاعات برای بدست آوردن رتبه مار سختی است.

سیستم رتبه بندی گوگل طوری طراحی شده است تا هیچ عامل خاصی تأثیر فوق العاده نداشته باشد. اول، ساده ترین حالت را در نظر بگیرید – یک پرس و جوی یک کلمه ای، به منظور رتبه بندی یک سنمد برای یک پرس و جوی یک کلمه ای، گوگل لیست بهترینهای آن سند را برای آن کلمه نگاه می کند. گ.گل هر کدام از بهترینها را از انواع مختلفی در نظر می گیرد (تیتر، انکر، URL، متن ساده با فونت بزرگ، متن ساده با فونت کوچک و ...) که

هر کدام از این وزن نوع (type-woight) مخصوص خود را دارد. وزن های انواع یک بردار می سازند که بر اساس نوع شاخص بندی شده است. گوگل تعداد بهترینهای هر نوع را در لیست بهترینها محاسبه می کند. سپس هر کدام از این اعداد به یک تعداد وزن (Count-Weight) تبدیل می شود. تعداد اوزان به طور طولی با بالا رفتن تعداد نوع بالا می روند ولی سریعاً افت می کنند بنابراین بیشتر از یک تعداد خاص تعداد خاص کارساز نخواهد بود. گوگل نتیجه روشن بردار تعداد اوزالن را با انواع اوزان برای محاسبه یک نموره IR برای هعر سند درنظر می گیرد. در نهایت نمره IR با رتبه صفحه ترکیب می شود تا رتبه نهایی سند بدست آید.

برای یک جستجوی چند کلمه ای، شرایط پیچیده تر است. در این حالت چندین لیست بهترینها باید به صورت یکجا مرور شوند بنابراین بهترینهایی که نزدیک به یکدیگر یافت می شوند بالاتر از بهترینهایی که از هم دور هستند در یک سند وزن داده یم شوند. بهترینهایی که از لیستهای بهترینهای چندگانه بدست می آیند، با هم هماهنگ هستند. یک تقریب برای هر کدام از مجموعه های بهترینهای هماهنگک محاسبه می شود. این تقریب بر پایهٔ میزان فاصله بهترینها در سند (یا انکر) است اما به دو روش متفاوت «بین» کلاس بندی می شوند که محدوده آنها از یک سند کاملاً هماهنگ تا «حتی مشابه نیست» رنج بندی می ودو تعداد موجود نه تنها برای هر نوع از بهترینها محاسبه می شود بلکه برای هر نوع و تقریب محاسبه یم شود. هر جفت نوع و تقریب یک وزن نوع – تقریب دارد. تعداد موجود به وزن تعداد تبدیل می شود و گوگل نتیجه واضح تعداد اوزان و نوع – تقریب اوزان را برای محاسبه یک نمره ID درنظر می گیرد. تمام این اعداد و ماتریسها می توانند با نتایج جستجو توسط یک

حالت اشکالزایی خاص نمای شداده شوند. این نمایش برای بهبود کارکرد سیستم رتبه بندی بسیار مؤثر بوده اند.

2.5.4. بازخور

عملیات رتبه بندی عوامل زیادی مانند وزن نوع و وزن نوع – تقریب دارد. پیدا کردن مقادیر درست برای آنها چیزی مانند جادوی سیاه است. برای انجام این کار، گوگل یک مکانیزم بازخور کاربر در موتئر جستجو دارد. یک کاربر قابل اعتماد ممکن است به طور انتخابی تمام نتایج برگردانده شده را ارزیابی کنتد. این بازخوذ ذخیره می شود. سپس با استفاده از این بازخورها تابع رتبه بندی اصلاح و تغیری داده می شود. این کار به ما اجازه یم دهد از اینکه هر تغییر چه تأثیری برروی نتایج جستجو می گذارد می دهد که البته از کمال بسیار دور است.

5 عملکرد و نتایج

مهم ترین معیار و محک یک موتور جستجو کیفیت نتایج آن است. با وجود اینکه ارزیابی یک کاربر فراتر از محدوده این مقاله است، تجربه هود گوگل نشان داده است که گوگل نتایج بهتری نسبت به موتورهای جستجوی تجاری برای اغلب جستجوها بدست می دهد. به عنوان مثالی که استفاده رتبهع صفحه و نوشته انکر و تقریب را نشان دهد در شکل ۴ می توانید نتایج جستجو برروی «جرج بوش» را ببینید. این نتایج بعضی از ویژگیهای گوگل را نشان می دهد. نتایج بر اساس سرویس دهنده دسته بندی شده اند. این کار در هنگام بررسی کردن مجموعه نتایج کمک بسیار بزرگی است. تعدادی از نتایج از دامنه white house.gov هستند که به طور محتمل از چنین جستجویی انتظار می رود. در حال حاضر، اغلب

موتورهـای جسـتجوی تجـاری هـیچ کـدام از نتـایج خـود را در چنـین جسـتجویی از whitehouse.gov برنمی گردانند و بسیار کمتر نتایج درست را برمی گردانن. ایـن حالـت به این خاطر است که این نتیجه جستجو و دانلود نشده بود. در عوض گوگل با تکیه بر نوشته انکر متوجه شده است که می تواند جواب خوبی بای این پرس و جو باشـد. همچنـین مشـابه آن نتیجه پنجم یک نامه الکترونیکی است کـه مسـلماً قابـل دانلـود کـردن نیسـت و نتیجـه استفاده از نوشته انکر است.

تمام نتایج در حد معقول صفحات با کیفیت بالا هستندذ و در بررسی نهایی هیچ کدام از آنهـا پیوند شکسته نبوده اند که به این دلیل است که تمام آنها رتبـه صـفحه بـالای یدارنـد. رتبـه صفحات درصد قرمز بودن نوار گراف است. در پایان، هیچ نتیجه ای در مورد یک جرجـی کـه بوش نباشد یا در مورد یک بوش که جرج نباشد وجود ندارد که به این دلیل است که گوگـل اهمیتد بالایی برروی تقریب موراد موجود بودن کلمه قرار داده است. مسلماً یک تست واقعـی از کیفیت یک موتور جستجو شامل مطالعه وسیع در مورد کاربر خواهد بود و یا تجزیه نتـایج که در این مقاله جایی برای آنها نیست.

1.5. احتیاجات منبع ذخیره سازی

جدای از کیفیت جستجو، گوگل به گونه ای طراحی شده است تا هزینه ها با اندازه وب همان طور که گسترش می یابد مقیاس پذیر باشند. یـک جنبـه ایـن کـار اسـتفاده مـوثر از فضـای ذخیره سازی است. جدول ۱ ریز ارقام شاخص آماری و بعضی احتیاجات فضای ذخیره سـازی گوگل را نشان یم دهد. به دلیل فشرده سازی، اندازه کلی انباره حدود ۵۳ گیگابایت است کـه یک سوم کل داده های ذخیره شده است. با هزینه های کنونی دیسک، انباره یک منبـع ارزان

از داده های مفید است. مهم تر از آن، کلیه داده های مورد استافده موتور جستجو به فضای ذخیره سازی در حدود ۵۵ گیگابایت احتیاج دارد. گذشته از این، اغلب پرس و جوها می توانند با استفاده از شاخص معکوس کوتاه پاشخ گفته شدند. با کدبندی و فشرده سازی بهتر شاخص سند، یک موتور جستجوی با کیفیت بالا می تواند در یک درایو ۷ گیگاباتی یک کامپیوتر شخصی جای گیرد.

2.5. عملکرد سیستم

برای هر موتور جستجو Crawl کردن و شاخص بندی به طور موثر بسیار مهم است. به این صورت اطلاعات به روز نگه داشته می شوند و تغییرات اساسی که برروی سیستم اعمال می شوند، می توانند با سرعت مناسب تست شوند.

برای گوگل عملیات مهم عبارتند از Croqling، شاخص بندی و مرتب سازی، بسیار سخت است که زمان گلی عمل Crowling را بتوان بدست آورد زیرا احتمال سرریز دیسک، از کار افتادن سرویس دهندهٔ نام یا تعدادی دیگر مشل وجود دارد که می توانند کار سیستم را متوقف کنند. در حالت کلی تقریباً ۹ روز برای دانلود کردن ۲۶ میلیون صفحه (با در نظر خطاها) طول می کشد. اگر چه یکبار که سیستم به نرمی کار می کرد و با سرعت بیشتری کار می کرد دانلود ۱۱ میلیون صفحه در تنها ۶۳ ساعت صورت گرفت که بوطر متوسط ۴ میلیون صفحه در روز یا ۴۸/۵ صفحه در ثانیه می شود. گوگل عملکرد روی شاخص بند و Crowler را آزمایشس کرده است. شاخص بند کاملاً سریع تر Crawler اجرا می شود. این حالت بیشتر به این خاطر است که گوگل زمان زیادی برای بهینه کردن شاخص بند صرف کرده است بنابراین هیچ محدودیتی در کارش ندارد. این بهینه سازیها شامل به روز رسانیهای

کلی در شاخص سند و مکان قرارگیری ساختمان داده هیا اساسی برروی دیسک محلی است. این شاخص بند به طور تقریبی با سرعت ۵۴ صفحه در ثانیه کار می کند. ترتیب بندها می توانند به صورت کاملاً موازی اجرا شو.ند با استفادهد از چهار دستگاه که پروسه مرتب سازی با این چهار دستگاه حدوداً ۲۴ ساعت طول می کشد.

3.5. عملکرد جستجو

بهبود عملکرد جستجو تا به این نقطه هدف اصلی تحقیق آقایان سرگی برین و لاورنس پیج نبوده است. در حال حاضر گوگل اغلب پرس و جوها را بین ۱ تا ۱۰ ثانیه پاسخ می دهد. این زمان بیشر صرف عمل ۳/۵ دیسک