موتور جستجوگر چگونه کار می کند؟

وقتي جستجويي در يک موتور جستجوگر انجام و نتايج جستجو ارايه مي شود، کاربران در واقع نتيجه کار بخش هاي متفاوت موتور جستجوگر را مي بينند. موتور جستجوگر قبلا "پايگاه داده اش را آماده کرده است و اين گونه نيست که درست در همان لحظه جستجو، تمام وب را بگردد. بسياري از خود مي پرسند که چگونه امکان دارد گوگل در کمتر از يك ثانيه تمام سايت های وب را بگردد و ميليون ها صفحه را در نتايج جستجوی خود ارايه کند؟
نه گوگل و نه هيچ موتور جستجوگر ديگري توانايي انجام اين کار را ندارند .همه آنها در زمان پاسخ گويي به کاربران، تنها در پايگاه داده اي که در اختيار دارند به جستجو مي پردازند و نه در وب !
موتور جستجوگر به کمك بخش های متفاوت خود، اطلاعات مورد نياز را قبلا" جمع آوري، تجزيه و تحليل مي کند و آنرا در پايگاه داده اش ذخيره مي نمايد و به هنگام جستجوی کاربر تنها در همين پايگاه داده می گردد .

بخش هاي مجزاي يك موتور جستجوگر عبارتند از:

- Spider •يا عنکبوت
- Crawler •يا خزنده
- Indexer •يا بايگانی کننده
- Database •يا پايگاه داده
- Ranker •يا سيستم رتبه بندی

الف) -Spider عنکبوت

اسپايدر يا روبوت(Robot) ، نرم افزاري است که کار جمع آوري اطلاعات مورد نياز يك موتور جستجوگر را بر عهده دارد. اسپايدر به صفحات مختلف سر مي زند، محتواي آنها را مي خواند، اطلاعات مورد نياز را جمع آوري مي کند و آنرا در اختيار ساير بخش هاي موتور جستجوگر قرار مي دهد .

کار يك اسپايدر، بسيار شبيه کار کاربران وب است. همانطور که کاربران، صفحات مختلف را بازديد مي کنند، اسپايدر هم درست اين کار را انجام مي دهد با اين تفاوت که اسپايدر کدهاي HTML صفحات را مي بيند اما کاربران نتيجه حاصل از کنار هم قرار گرفتن اين کدها را.

اما يک اسپايدر آنرا چگونه می بيند؟
براي اين که شما هم بتوانيد دنياي وب را از ديدگاه يك اسپايدر ببينيد، کافي است که کدهاي HTML صفحات را مشاهده کنيد.

آيا اين دنياي متني برای شما جذاب است؟

اسپايدر، به هنگام مشاهده صفحات، از خود بر روي سرورها رد پا برجای مي گذارد. شما اگر اجازه دسترسي به آمار ديد و بازديدهاي صورت گرفته از يک سايت و اتفاقات انجام شده در آنرا داشته باشيد، مي توانيد مشخص کنيد که اسپايدر کدام يک از موتورهاي جستجوگر صفحات سايت را مورد بازديد قرار داده اند.
يکی از فعاليتهاي اصلی که در SEM انجام می شود تحليل آمار همين ديد و بازديدها می باشد.

اسپايدرها کاربردهاي ديگري نيز دارند، به عنوان مثال عده اي از آنها به سايت هاي مختلف مراجعه مي کنند و فقط به بررسي فعال بودن لينك هاي آنها مي پردازند و يا به دنبال آدرس پست الکترونيکي (Email) مي گردند.

ب) Crawler -خزنده

کراولر، نرم افزاري است که به عنوان يك فرمانده براي اسپايدر عمل مي کند. آن مشخص مي کند که اسپايدر کدام صفحات را مورد بازديد قرار دهد. در واقع کراولر تصميم مي گيرد که کدام يك از لينك های صفحه ای که اسپايدر در حال حاضر در آن قرار دارد، دنبال شود. ممکن است همه آنها را دنبال کند، بعضي ها را دنبال کند و يا هيچ کدام را دنبال نکند.

کراولر، ممکن است قبلا "برنامه ريزي شده باشد که آدرس های خاصی را طبق برنامه، در اختيار اسپايدر قرار دهد تا از آنها ديدن کند. دنبال کردن لينك های يک صفحه به اين بستگي دارد که موتور جستجوگر چه حجمي از اطلاعات يک سايت را مي تواند در پايگاه داده اش ذخيره کند و همچنين ممکن است اجازه دسترسي به بعضي از صفحات به موتورهاي جستجوگر داده نشده باشد .

شما به عنوان دارنده سايت، همان طور که دوست داريد موتورهاي جستجوگر اطلاعات سايت شما را با خود ببرند، مي توانيد آنها را از بعضي از صفحات سايت تان دور کنيد و اجازه دسترسي به محتواي آن صفحات را به آنها ندهيد. تنظيم ميزان دسترسي موتورهاي جستجوگر به محتوای يک سايت توسط پروتکل Robots انجام مي شود که در مقالات ديگر سايت به آن پرداخته شده است. به عمل کراولر ، خزش (Crawling) مي گويند.

ج) Indexer -بايگانی کننده

تمام اطلاعات جمع آورش شده توسط اسپايدر در اختيار ايندکسر قرار مي گيرد. در اين بخش اطلاعات ارسالي مورد تجزيه و تحليل قرار مي گيرند و به بخش هاي متفاوتي تقسيم مي شوند. تجزيه و تحليل بدين معني است که مشخص می شود اطلاعات از کدام صفحه ارسال شده است، چه حجمي دارد، کلمات موجود در آن کدام است، کلمات چندبار تکرار شده است، کلمات در کجاي صفحه قرار دارند و . ...

در حقيقت ايندکسر، صفحه را به پارامترهای آن خرد می کند و تمام اين پارامترها را به يک مقياس عددی تبديل می کند تا سيستم رتبه بندي بتواند پارامترهای صفحات مختلف را با هم مقايسه کند.

در زمان تجزيه و تحليل اطلاعات، ايندکسر براي کاهش حجم داده ها از بعضي کلمات که بسيار رايج هستند صرفنظر مي کند. کلماتي نظير a ، an، the، www، isو ... . از اين گونه کلمات هستند.

د) DataBase - پايگاه داده

تمام داده هاي تجزيه و تحليل شده در ايندکسر، به پايگاه داده ارسال مي گردد. در اين بخش داده ها گروه بندي، کدگذاري و ذخيره مي شود. همچنين داده ها قبل از آنکه ذخيره شوند، طبق تکنيکهای خاصی فشرده مي شوند تا حجم کمي از پايگاه داده را اشغال کنند .

يك موتور جستجوگر بايد پايگاده داده عظيمي داشته باشد و به طور مداوم حجم محتوای آنرا گسترش دهد و البته اطلاعات قديمي را هم به روز رسانی نمايد .بزرگي و به روز بودن پايگاه داده يك موتور جستجوگر براي آن امتياز محسوب مي گردد . يکي از تفاوتهاي اصلي موتورهاي جستجوگر در حجم پايگاه داده آنها و همچنين روش ذخيره سازي داده ها در پايگاه داده است .

ه) Ranker -سيستم رتبه بندی

بعد از آنکه تمام مراحل قبل انجام شد، موتور جستجوگر آماده پاسخ گويي به سوالات کاربران است. کاربران چند کلمه را در جعبه جستجوي (Search Box) آن وارد مي کنند و سپس با فشردن Enter منتظر پـاسخ مي مانند .

برای پاسخگويي به درخواست کاربر، ابتدا تمام صفحات موجود در پايگاه داده که به موضوع جستجو شده، مرتبط هستند، مشخص مي شوند. پس از آن سيستم رتبه بندي وارد عمل شده، آنها را از بيشترين ارتباط تا کمترين ارتباط مرتب مي کند و به عنوان نتايج جستجو به کاربر نمايش می دهد.

حتي اگر موتور جستجوگر بهترين و کامل ترين پايگاه داده را داشته باشد اما نتواند پاسخ هاي مرتبطي را ارايه کند، يك موتور جستجوگر ضعيف خواهد بود. در حقيقت سيستم رتبه بندي قلب تپنده يك موتور جستجوگر است و تفاوت اصلي موتورهاي جستجوگر در اين بخش قرار دارد.

سيستم رتبه بندي براي پاسخ گويي به سوالات كاربران، پارامترهاي بسياري را در نظر مي گيرد تا بتواند بهترين پاسخ ها را در اختيار آنها قرار دارد. حرفه اي هاي دنياي SEM به طور خلاصه از آن به) Algo الگوريتم) ياد مي كنند . الگوريتم، مجموعه اي از دستورالعمل ها است كه موتور جستجوگر با اعمال آنها بر پارامترهاي صفحات موجود در پايگاه داده اش، تصميم مي گيرد كه صفحات مرتبط را چگونه در نتايج جستجو مرتب كند. در حال حاضر قدرتمندترين سيستم رتبه بندي را گوگل در اختيار دارد.

مُي توان با ادغام كردن اسپايدر با كراولر و همچنين ايندكسر با پايگاه داده، موتور جستجوگر را شامل سه بخش زير دانست كه اين گونه تقسيم بندی هم درست می باشد:

- كراولر
- بايگانی
- سيستم رتبه بندی

تذكر- برای سهولت در بيان مطالب بعدی هر گاه صحبت از بايگانی كردن (شدن) به ميان می آيد، مقصود اين است كه صفحه تجزيه و تحليل شده و به پايگاه داده موتور جستجوگر وارد مي شود.

براي آنكه تصور درستي از نحوه كار يك موتور جستجوگر داشته باشيد داستان نامتعارف زير را با هم بررسي مي كنيم.

داستان ما يک شكارچی دارد. او تصميم به شكار می گيرد:

-كار كراولر :
او قصد دارد براي شكار به منطقه حفاظت شده ابيورد، واقع در شهرستان درگز (شمالی ترين شهر خراسان بزرگ) برود .

-پروتكل: Robots
ابتدا تمام محدوديت هاي موجود براي شكار در اين منطقه را بررسي مي كند:

- آيا در اين منطقه می توان به شكار پرداخت؟
- كدام حيوانات را می توان شكار كرد؟
- حداكثر تعداد شكار چه ميزانی است؟
- و . ...

فرض می كنيم او مجوز شكار يک اوريال (نوعي آهو) را از شكارباني منطقه دريافت مي كند .

-كار اسپايدر
او اوريالي رعنا را شكار مي كند و سپس آنرا با خود به منزل می برد.

-كار ايندكسر
شكار را تكه تكه كرده، گوشت، استخوان، دل و قلوه، كله پاچه و ... آنرا بسته بندي مي كند و بخش های زايد شكار را دور می ريزد .

-كار پايگاه داده
بسته هاي حاصل را درون فريزر قرار داده، ذخيره می كند .

-کار سیستم رتبه بندی

مهمانان سراغ او مي آيند و همسر او بسته به ذائقه مهمانان براي آنها غذا طبخ مي كند. ممكن است عده اي كله پاچه، عده اي آبگوشت، عده اي جگر و ... دوست داشته باشند. پخت غذا طبق سليقه مهمانان كار سختی است. ممكن است همه آنها آبگوشت بخواهند اما آنها مسلما" بامزه ترين آبگوشت را مي خواهند!

نکته ها:

•شکارچی می توانست برای شکار کبک یا اوریال و یا هر دو به آن منطقه برود همانطور که موتور جستجوگر می تواند از سرور سایت شما انواع فایل (عكس، فايل متنی، فايل اجرايي و ...) درخواست کند.

•غذای خوشمزه را می توانید با نتایج جستجوی دقیق و مرتبط مقایسه کنید. اگر شکارچی بهترین شکار را با خود به منزل ببرد اما غذايي خوشمزه و مطابق سليقه مهمانان طبخ نگردد، تمام زحمات هدر رفته است.

•به عنوان آخرین نکته این مقاله یاد آوری می کنم که به شکار اوریالی رعنا آن هم در منطقه حفاظت شده ابیورد، اصلا فکر نکنید. اما توصیه می شود که حتما از طبیعت بکر آن دیدن فرمایید (بدون اسلحه.(!