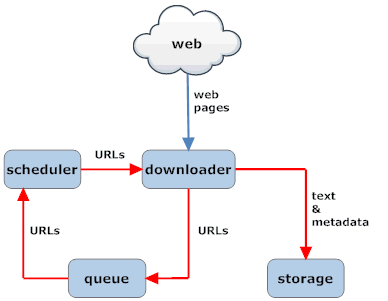
crawler

وجود حجم زیاد دیتا بدون ساختار در محیط وب و در دسترس نبودن آسان آنها باعث شده که پردازش آنها سخت شود. وب کرالر ها بعنوان یک تکنولوژی نقش مهمی در بدست آوردن اطلاعات از دیتاهای موجود در وب دارند. بعبارتی وجود وب کرالر ها باعث بهبود جست و جو در موتور جست و جویی مثل گوگل می شود.وجود وب کرالر ها باعث می شود که جست و جو را جهت دهد و در واقع یک موتور جست و جوی عمودی بوجود آورد. در واقع کرالر ها بعنوان هسته اصلی موتور های جست و جو هستند که هر چه دقیق تر و سریع تر باشند عملکرد آنها در استخراج اطلاعات بهتر می شود.

کرالر ها از 4 بخش اصلی تشکیل شده اند که در شکل زیر بخش های آنها مشخص شده است:

روند کار کرالر به این صورت است که schedulers آن URL که قرار است crawl شود را به downloader می دهد حال downloader از وب اطلاعات مورد نیاز مربوط به آن صفحه را پیدا می کند و به information extractors می دهد حال information extractors با توجه به دستورالعمل های استخراج اطلاعات ، اطلاعات مورد نیاز را استخراج می کند سپس url بعدی از صف url ها وارد schedular می شود.



معیارهای بررسی کیفیت کرالرها:

1. Distributed

قرار دادن کرالر ها در یک محیط توزیع شده

1. scalability

عکلکرد کرال بسیار کند است پس استفاده از ماشین های مختلف در این عملیات باعث بهبود عملکرد کرالینگ و افزایش پهنای باند می شود.

1. performance and efficiency
2. quality

کیفیت کرالر ها با کرال کردن درصد بیشتری از صفحات مورد نیاز کاربر تشخیص داده می شود.

1. freshness

با تغییر محتوای اینکه کرالر اطلاعات خود را بروز نگه دارد اهمیت دارد.

1. extensibility

Types of web crawlers:

1. generic web crawlers (traditional crawler)

کرالر های معمولی کل صفحات وب را بدون محدودیت خزش می کنند

1. focusing web crawlers,

کرالر هایی که همی صفحه وب را کرال نمی کنند بلکه فقط روی یک موضوع خاص تمرکز دارند. در این نوع کرالر دو نوع ماژول وجود دارد

1. web page judgment module

اولویت نمایش صفحات را بر اساس میزان ارتباط تاپیک وارد شده به محتوای سایت است.

1. URL link priority ranking module

اولویت دادن به url ها براساس میزان ارتباط به تاپیک مورد بررسی و نمایش مرتبط ترین آنها

1. incremental web crawlers,
2. در وب کرالر افزایشی وقتی یک صفحه آپدیت می شود نیازی نیست کل اطلاعات کرال شده بروز رسانی شود و فقط کافی است آنهایی که تغییر کرده اند بروز رسانی شوند. معمولا incremental web crawlers ها بسته به نوع تغییری که رخ می دهد ممکن است در موقعیت های زیر قرار گیرند:امدن یک صفحه جدید یا تغییر محتوا
3. distributed crawlers

Distributed web crawler بر روی مجموعه ای از کامپیوتر ها اجرا می شوند که هر کدام یک focusing web crawlers اجرا می کنند که بر اساس نحوه چیدمان این کامپیوترها سه نوع web crawler داریم:

* Master slave

یک کامپیوتر در head قرار دارد و سایر کامپیوترها با اون در ارتباطند ولی بین فرزندان ارتباطی وجود ندارد.

* autonomous mode

بین فرزندان می تواند ارتباط دایره ای یا شبکه بصورت پبچیده باشد.

* mixed mode

این مدل ترکیبی از دو مدل قبل است .یعنی هم ادمبن با فرزندان در ارتباط هم بین فرزندان ارتباط وجود دارد.

1. parallel crawlers

بدست آوردن همه ی اطلاعات از اینترنت با استفاده از یک کرالر سخت است.به همین علت فرایند های کرالینگ در وب باید بصورت موازی در زمان کوتاه انجام شود.

1. Internet of things web crawlers.

متد ذخیره سازی داده

بطور کلی گرالر ها دو متد ذخیره سازی دارند:

* ذخیره در فایل لوکال
* دخیره در دیتابیس

یکی از مهمترین دیتابیس های key-value برای کار با کرالر ها Redis است.

استراتژی های جست و جو در وب:

* اول عمق
* اول سطح
* best priority search strategy