
بسم الله الرحمن الرحيم

دانشکده مهندسی برق و رباتیک

رشته مهندسی رباتیک

پایان نامه کارشناسی ارشد

شناسایی، طبقه بندی و یادگیری اشکال و تصاویر در بستر وب

توسط یک موتورجستجو

نگارنده : مهرداد صالحی

استاد راهنما

دکتر علیرضا احمدی فرد

شهریور ۱۳۹۶

دانشگاه صنعتی شاهرود

دانشکده مهندسی برق و رباتیک

گروه مهندسی رباتیک

پایان نامه کارشناسی ارشد آقای مهرداد صالحی

تحت عنوان: شناسایی، طبقه بندی و یادگیری اشکال و تصاویر در بستر وب توسط یک موتورجستجو

در تاریخ ۱۳۹۶/۶/۱۴ توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی ارشد مورد ارزیابی و با درجه ..... مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
نام و نام خانوادگی : علیرضا احمدی فرد		نام و نام خانوادگی :	
نام و نام خانوادگی :		نام و نام خانوادگی :	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
نام و نام خانوادگی : سید علی سلیمانی ایوری		نام و نام خانوادگی : مسعود مهدی زاده رخی	
نام و نام خانوادگی : امیر حسین نایبی آستانه			
نام و نام خانوادگی :			
نام و نام خانوادگی :			

تقدیم به :

پدر و مادر عزیز و مهربانم

که در سختی‌ها و دشواری‌های زندگی همواره یاوری دلسوز و فداکار و پشتیبانی محکم و مطمئن برایم بوده‌اند.

تشکر و قدردانی :

جناب آقای دکتر علیرضا احمدی فرد که در تمام مراحل این **پایان نامه** بنده را مورد راهنمایی قراردادند.

تعهد نامه

اینجانب **مهرداد صالحی** دانشجوی دوره کارشناسی ارشد رشته **مهندسی رباتیک** دانشکده **مهندسی برق و رباتیک** دانشگاه صنعتی شاهرود نویسنده پایان نامه **طبقه بندی و یادگیری اشکال و تصاویر در بستر وب توسط یک موتورجستجو** تحت راهنمائی **دکتر علیرضا احمدی‌فرد** متعهد می شوم .

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

چکیده :

با افزایش روزافزون تولید داده دربستر وب نیاز به یافتن اطلاعات مورد نیاز با صرف کمترین زمان و هزینه، اهمیت موتورهای جستجو را روز به روز پررنگتر می‌کند. البته یافتن لیستی بلند بالا از تصاویر و نتایج نامرتبط با موضوع مورد جستجو همواره باعث ناراحتی کاربران شده و باید زمان و وقت بسیاری را صرف کنند تا بتوانند موضوع مورد نیاز خود را بیابند. یکی از مهمترین کاربردهای موتورجستجو بررسی صفحات وب می‌باشد، برنامهی هوشمند که بعد از جستجوی و استخراج اطلاعات موجود دربستر وب به پردازش و گروه بندی اطلاعات می‌پردازد و نتایج رابه ترتیب شباهت واهمیت با موضوع مورد جستجو با یک اولویت خاص به کاربرنمایش می‌دهد.

راه حل پیشنهادی این پروژه شامل استخراج متن (متون نزدیک به تصویر) و تصویر از صفحات وب با استفاده از یک موتور جستجوی ساخته شده محلی و یادگیری تصاویر با یک شبکه عصبی عمیق می‌باشد در این جا چالش های مانند ایجاد یک پایگاه داده اولیه از بستر وب و یادگیری شبکه بدون ناظر و گروه بندی متون و تصاویر وجود دارد.دراین پروژه به معرفی چالش‌های یک موتور جستجوی وب با محوریت تصویر و روش هایی برای حل این مشکلات به همراه نوآوری‌های انجام گرفته در آن می‌پردازیم و در انتها با استفاده از یادگیری عمیق به استخراج ویژگی از تصاویر و متون داخل صفحات وب پرداخته و به طبقه بندی این نتایج جهت دسترسی سریعتر به اطلاعات و گروه بندی اشیا می‌پردازیم.

کلمات کلیدی :

موتور جستجو ، یادگیری عمیق، طبقه بندی تصاویر، خزنده وب، پردازش موازی، موتور جستجوی تصویر.

فهرست مطالب

عنوان صفحه

فهرست جدول ها

عنوان صفحه

فهرست اشکال

عنوان صفحه

فهرست فرمول‌ها

عنوان صفحه

فصل اول:

پیش گفتار

۱-۱مقدمه

۱-۱-۱ پیش گفتار

با افزایش روز افزون تولید اطلاعات اینترنت[۱] طبقه بندی و دستیابی سریع به داده ها اهمیت بسیاری پیدا کرده است. در این حجم داده جدا سازی اطلاعات مفید از این حجم داده بسیار سخت می‌باشد. هر روزه در دنیا میلیون ها سایت[۲] با موضوعات و اطلاعات مختلف ایجاد می‌گردد که ما از آنها هیچ اطلاعاتی نداریم. پس چطور در این فضا اطلاعاتی را که مورد نظر ما است را کسب کنیم. برای این راهکار موتورهای جستجو[۳] متولد شدند. وظیفه اصلی یک موتور جستجو جمع آوری اطلاعات از بستر اینترنت، طبقه بندی داده های دریافتی در یکسری گروه مشخص و نمایش لیست نتایج مرتبط با موضوع مورد جستجوی به کاربر درخواست کننده می‌باشد. در حقیقت وظیفه اصلی یک موتور جستجو جمع آوری سیل عظیم این داده ها در غالب مشخص و سهولت دستیابی اطلاعات بستر وب می‌باشد.

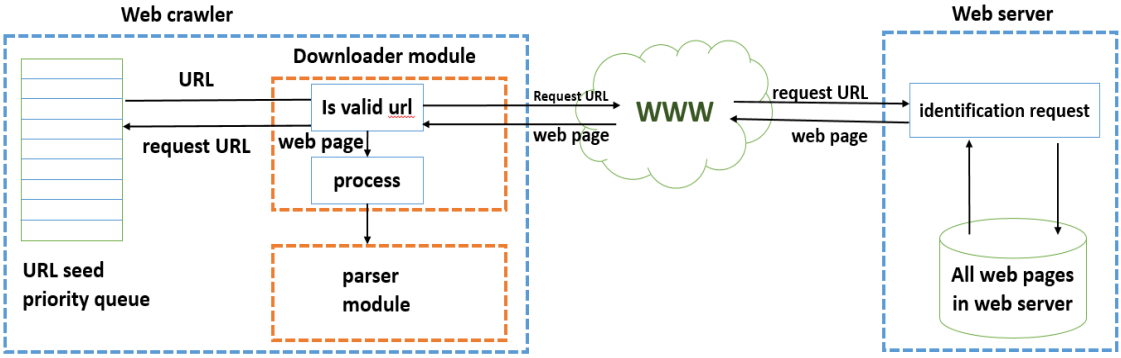
۱-۱-۲ صفحات وب[۴]

منظور از صفحات وب همان مستندات و فایل‌های است که در بستر وب و توسط مرورگرهای وب[۵] قابل نمایش هستند. کنسرسیوم جهانی وب[۶] قالب استاندارد HTML را برای نمایش این مستندات در مرورگرهای وب معین کرده است. صفحات وب حاوی اطلاعات متنوعی میباشند از متن و تصویر گرفته تا فایل‌های تصویری. صفحات وب از نظر محتواهم به چندین دسته تقسیم می‌شوند، مثلا بعضی صفحات وب مطالب کلی در باره هر موضوعی دارند و یا خاص منظوره اند به عنوان مثال صفحه ای درباره آخرین اخبار و اطلاعات ورزشی حاوی اطلاعات است و صفحه دیگری شامل جدیدترین اطلاعات سخت افزار کامپیوتر.

---

۱ internet
۲ site
۳ web crawler
۴ web page
۵ web browser
۶ world wide web (w3c)

یکی از وظایف موتورهای جستجو شناسایی موضوعیت کلی صفحه وب می‌باشد تا بهتر بتواند صفحه وب را مورد بررسی قرار دهد. به عنوان مثال اگر یک صفحه وب جدید ترین اطلاعات را درباره علوم ستاره شناسی نمایش می‌دهد پس شانس یافتن اطلاعات درباره جدیدترین روش هایی درمان سرطان در این صفحه وب کم و یافتن اطلاعات درباره یافتن جدیدترین کهکشان های کشف شده زیاد می‌باشد. در شکل ۱-۱ می‌توانید یک ساختار اولیه و کلی از موتور جستجو را ببینید.



شکل ۱-۱ نمای کلی یک موتور جستجوی وب

HTML[۱] ۱-۱-۳

زبان نشانه گذاری ابر متن یا به اختصار **اچ‌دی‌اِم‌اِل** در کنار سی‌اِس‌اِس[۲] هسته فناوری ساخت صفحه‌های وب هستند. اچ‌تی‌ام‌ال زبان توصیف ساختار صفحه‌های وب است.زبانی‌است برای نشانه‌گذاری ابرمتن که برای تدوین قالب و طراحی صفحه‌های وب به کار می‌برند. دستورالعمل‌های این زبان، برچسب[۳] نام دارند که محتوای یک صفحه وب، با آن‌ها، نشانه‌گذاری شده و بدین‌ترتیب، نحوه نمایش آن صفحه برای مرورگرهای وب، توصیف می‌شود. آخرین نسخه این زبان در حال حاظر نسخه ۵ می‌باشد. در واقع زبانی‌که

---

hypertext markup language[۱] cascading style sheets[۲] tag[۳]

موتورهای جستجو آنرا می‌فهمند و از طریق آن به جمعآوری اطلاعات می‌پردازند همین زبان HTML می‌باشد. هر یک از برچسب‌هایHTML، معنا و مفهوم و کاربرد خاصی دارند و تأثیر مشخصی بر محتوای صفحه می‌گذارند. به عنوان مثال برچسب‌های برای تغییرظاهری متن، نظیر درشت و ضخیم کردن یک کلمه یا برقراری پیوند به صفحات دیگر درHTML تعریف شده‌اند . ویا برچسب های برای نمایش فایل های صوتی تصویری مانند کلیپ های ویدیوی و حتی بعضی برچسب ها جهت تعریف متن و توضیحات درHTML تعریف شده‌اند.

یک سندHTML، یک پرونده مبتنی بر متن[۱] است که معمولاً با پسوند htm. یا html. نام‌گذاری شده و محتویات آن از برچسب‌های اچ‌تی‌ام‌ال تشکیل می‌شود. مرورگرهای وب، که قادر به درک و تفسیر برچسب‌های HTML هستند، تک‌تک آن‌ها را از داخل سند HTML خوانده و سپس محتوای آن صفحه را نمایان‌سازی[۲]می‌کنند.

HTML زبان برنامه‌نویسی نیست، بلکه زبانی برای نشانه‌گذاری ابرمتن است و اساساً برای ساختارمند کردن اطلاعات و جدایش اجزای منطقی یک نوشتار نظیر عناوین، تصاویر، فهرست‌ها، بندها و جدول‌ها به کار می‌رود. از سوی دیگر، HTML را نباید به عنوان زبانی برای صفحه‌آرایی یا نقاشی صفحات وب به کار برد. این وظیفه اکنون بر دوش فناوری‌های دیگری همچون CSS است. در موتورهای جستجو صفحات HTML توسط یک بخش مشخص به نام تجزیه کننده وب[۳] مورد بررسی قرار گرفته و اطلاعات و مطالب مهم از این صفحات استخراج می‌گردد. در فصل های بعد به اختصار درباره هر کدام از این بخش ها صحبت خواهد شد و همینطور به معرفی تگ‌ها و معرفی آنها می‌پردازیم.

---

text–based[۱] render[۲] webparser[۳]

۲-۱ اهداف پروژه

در این قسمت اهداف کلی این پروژه به طور مختصر عنوان می‌شود.

طراحی یک موتور جستجوی تحت وب :

۱- جستجو در بستر وب وجمع آوری اطلاعات از روی متن وتصاویر مرتبط با پرس‌وجو :

در این قسمت هدف دریافت صفحات وب و استخراج اطلاعات از داخل آن می‌باشد. و سپس طبقه بندی اطلاعات و سرانجام نمایش نتایج.

۲- دادن امتیاز به صفحات بررسی شده جهت جستجوهای بعدی :

در اینجا به صفحات بازدید شده بر حسب ارزش و اطلاعاتی که به سوال مورد جستجوی ما داده اند امتیازی داده می‌شود تا در جستجو های بعدی درباره همین سوال یا سوالات و پرس‌وجوهای مشابه (سوال های مشابه در یک طبقه) با سرعت و اطمینان بیشتری جستجو انجام شود.

۳- لیست اولویت :

این لیست حاوی پیوندهای[۲] مورد جستجو می‌باشد که بر حسب اولویت به موتور جستجو جهت بررسی و دانلود داده می‌شوند.

۴- مبدل مساله[۳] :

استفاده از یک مبدل مساله جهت تولید سوالات مشابه با سوال پرسیده شده از موتور جستجو و شاخص گذاری همه این سوالات در یک پایگاه داده جهت ساده سازی پرس‌وجوی بعدی.

---

webParser[۱] link[۲] problem generator[۳]

۵- مقایسه نتایج جستجو :

مقایسه نتایج جستجو با نتایج و ویژگی های استخراج شده موجود در پایگاه داده و به روز رسانی ویژگی های مربوط به سوال پرسیده شده از موتور جستجو. بررسی پیوند های جدید بدست آمده از جستجو در صفحه وب جاری که بعضی از این پیوند ها مرتبط با موضوع مورد جستجوی ما هستند

و بعضیدیگر ارجاع به صفحات تبلیغات و فاقد محتوا هستند. تشخیص اهمیت و دادن اولویت به این صفحات از وظایف موتور جستجو است.

۶- استخراج ویژگی :

کسب اطلاعات از داخل تصاویر و متون موجود در صفحه مورد بررسی (بعد از دانلود صفحه وب از لیست صفحات مورد بررسی، باید ویژگی های مورد نیازرا از داخل این صفحه استخراج کرده و همینطورپیوند های متصل به صفحات دیگر موجود درصفحه جاری را با توجه به اولویت به صف اولویت موتور جستجو اضافه کنیم.)

- بررسی چالش ها، بررسی راهکارهای ارایه شده و معرفی نوآوری های پروژه :

در فصل سوم این پایان نامه به معرفی لیست چالش های یک موتور جستجوی تحت وب پرداخته سپس راهکار های ارایه شده تا کنون را مورد بررسی قرار می‌دهیم و در انتها به معرفی نوآوری های انجام گرفته در پروژه می‌پردازیم. در فصل چهارم هم به بررسی اجمالی از مراحل مختلف یک موتور جستجو می‌پردازیم.

فصل دوم:

معرفی موتورجستجوگر وب

۲-۱ معرفی

یک موتور جستجو در حقیقت نرم افزاری جهت بررسی و کسب اطلاعات درباره یک موضوع یا پرس‌و جوی خاص در صفحات وب، جمع آوری اطلاعات و نمایش نتایج می‌باشد. و از ابزار و قسمتهای زیر تشکیل شده است.

- صف اولویت[۱]
- خزنده وب[۲]
- پایگاه داده : جهت ذخیره اطلاعات از قبیل صفحات وب و شاخص گذاری پرسش های انجام شده از موتور جستجو.

به طور خلاصه یک موتور جستجو از طریق یک نرم افزار به نام خزنده وب و با استفاده از یک صف شامل لیست آدرس صفحات وب به جستجو در صفحات وب پرداخته و نتایج بدست آمده را در یک پایگاه داده ذخیره می‌کند تا در جستجو های بعدی مطالب بدست آمده را در اختیار جستجو کننده قرار دهد.

۲-۲ خزنده وب چیست؟

خزنده وب به برنامه کامپیوتری گفته می‌شود که وب جهان گستر را به صورت مرتب و سلسله مراتبی بر حسب یکسری اولویت خاص بازدید کرده و اطلاعات آن را مورد پردازش قرار می‌دهد. از خزنده وب با عناوین دیگری مانند spiders ant و web robots نیز یاد می‌شود. البته نباید نامی مانند خزنده وب باعث نگرانی از این موضوع شود که با یک ویروس کامپیوتری طرف هستیم.

---

[۱] priority queue [۲] web crawler

۲-۳ کاربردهای خزنده وب

۲-۳-۱ موتورهای جستجو

در موتورهای جستجو، خزنده وب نقش بررسی کننده و کسب کننده اطلاعات را بازی می‌کند. خزنده وب سایت ها را به فرم دوره‌ای و بر حسب یک لیست اولویت بازدید کرده و نتایج جستجو را به موتور جستجو می‌فرستد. که بعد از پردازش و کسب اطلاعات و طبقه بندی داده‌ها و همینطور شاخص گذاری، نتایج را به کاربران نمایش می‌دهند.

۲-۳-۲ مدیریت فنی وب سایت

در یک وب سایت با هزاران صفحه و مطلب، و صدها کاربر که هر روز میلیون ها اطلاعات جدید را در سایت درج می‌کنند . مدیریت داده ها توسط نیروی انسانی کاری بسیار مشکل و حتی غیر ممکن به نظر می‌رسد. در اینجا هم وظیفه خزنده وب استخراج ویژگی ها از صفحات جهت سهولت مدیریت سایت می‌باشد. از طرف دیگر مدیریت فنی وب سایت شامل یافتن لینک‌های شکسته[۱]، اعتبار سنجی[۲] کدهای HTML ، فایل‌های css و ... می‌باشد. در اینجا منظور از لینک شکسته در صفحه وب اشاره به لینک‌های است که در واقع بدون مقصد هستند یعنی وقتی که شما بر روی یک عبارت لینک شده کلیک می‌کنید، صفحه‌ی مربوط به آنها بارگذاری نمی‌شود و امکان دارد با خطای ۴۰۴ یا Not Found مواجه شوید. به این دلیل که یا صفحه به طور کامل از سایت حذف شده یا مشکلی برای صفحه رخ داده است و یا بحث مربوط به امنیت صفحه که مثلا توسط یک کاربر خاص قابل دیدن نیست و برای مشاهده اطلاعات داخلی این صفحه وب باید به اهراز هویت پرداخته شود.

---

broken links[۱] validation[۲]

در حقیقت یکی از وظایف مهم خزنده وب تشخیص این لینک های شکسته است تا در بازدید های بعدی (به فرم دوره ای موتور جستجو هر چند وقت یکبار صفحات بررسی شده را دوباره دنبال می‌کند) در لیست لینک های بررسی متوجه این لینک های معیوب شده و زمان و وقت خود را برای بررسی این لینک ها به هدر ندهد.

۲-۳-۳ جمع آوری اطلاعات خاص

کاربرد دیگر خزنده‌های وب جمع‌آوری اطلاعات خاصی مانند آدرس‌های ایمیل است. معمولا هدف از اینکار جلوگیری از ارسال هرزنامه می‌باشد. و یا حتی جمع‌آوری تصاویر از صفحات وب که موضوع این پروژه هم دراین مورد می‌باشد.

۲-۴ نحوه کارکرد یک خزنده وب چیست؟

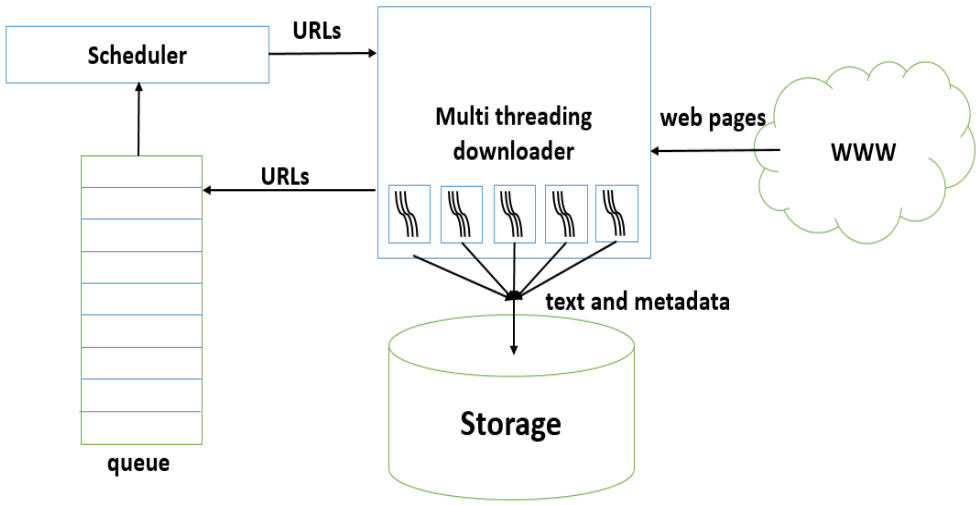
به صورت عمومی نحوه کارخزنده وب به این صورت است که ابتدا لیستی از URL[۱] ها (آدرس‌های وب) که به عنوان seed شناخته می‌شوند را برای بازدید پردازش می‌کند. البته این آدرس ها بر حسب یکسری امتیازها اولویت بندی می‌شوند. در مرحله بعدی صفحه مورد نظر دانلود شده و در مکانی مشخص ذخیره می‌شود. سپس باید اطلاعات داخل صفحه پردازش شوند هنگام پردازش آدرس‌ها، لیست لینک ها و آدرس های موجود در صفحات مورد جستجو را گردآوری کرده و به لیست ابتدایی اضافه می‌کنند. که اینجا هم صفحات اضافه شده به لیست باید برحسب اولویت در صف قرار بگیرند بقیه اطلاعات را نیز با توجه به نیاز و هدف خود ذخیره و پردازش می‌کند.

---

[۱] uniform resource locator

۲-۴-۱ معماری

خزنده وب بخش اصلی و مرکزی هر موتور جستجو را تشکیل می‌دهد. به همین جهت الگوریتم و معماری آن ها به شدت مخفی نگه داشته می‌شود. با این وجود معماری سطح بالای[۱]آن در شکل ۲-۱ مشخص می‌شود.



شکل ۲-۱ معماری سطح بالای خزنده وب [1]

۲-۵ فرایندهای پیش پردازش یک خزنده وب

در زیر لیستی از فرایندها و پیش پردازش‌های یک خزنده وب را مشاهده می‌کنید.

۲-۵-۱ عادی کردن آدرس[۲]

منظور از عادی کردن آدرس، یکی کردن آدرس‌هایی می‌باشد که دارای خروجی یکسانی هستند. هدف از این کار جلوگیری ازجمع آوری اطلاعات یکسان از چندین URL است. همانطور که قبلا هم گفته شده ما

---

۱ high level architecture

۲ url normalize

لیست آدرس های صفحات وب را در یک صف اولویت قرار دادیم و اگر مدام آدرس های با مقادیر متفاوت ولی صفحه‌وب یکسان را در صف اضافه کنیم چیزی جز افزایش بارکاری برای موتور جستجو ندارد. بهتر است این آدرس‌ها یکی شوند.آدرس در وب برای دستیابی راحت‌تر کاربران به یک سایت یا محتوای خاص می‌باشد. هر صفحه وب دارای یک آدرس منحصر بفرد می‌باشد به عنوان مثال برای استفاده از سایت ویکی‌پدیا ما از آدرس منحصر بفرد https://wikipedia.org استفاده می‌کنیم. چند مرحله از عادی سازی یک آدرس در زیر نمایش داده شده است.

- تبدیل آدرس به حروف کوچک : هدف تبدیل تمام کاراکترهای بزرگ به کاراکترهای کوچک می‌باشد. به دلیل جلوگیری از تکرار آدرس به عنوان مثال تمام آدرس‌های زیر به یک صفحه وب اشاره می‌کنند

  - http://www.gamefa.com
  - HTTP://WWW.GAMEFA.COM
  - http://www.GAMEFA.com

در زیر یک مثال از این فرایند را می‌بینید.

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

- افزودن / به انتهای آدرس در صورت نیاز : در زیر یک مثال از این فرایند را می‌بینید.

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

- حذف آدرس ایندکس دایرکتوری : آدرس ایندکس فقط برای مدیریت صفحه وب می‌باشد و نتیجه آدرس بدون ایندکس با ایندکس یکی می‌باشد. پس بهتر است برای ایجاد آدرس‌های تکراری بالا ایندکس ها حذف گردد.

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

- بزرگ کردن حروف encode شده : بهتر است حروفِ انکُد شده یا همان حروف بعد از علامت ٪ حذف شوند.

http://www. example.com/a%c2%b1b → http://www.example.com/a%C2%B1b

حذف بخش زاید : بخش زاید بخشی است که معمولا شامل بخش های فیلتری و محدود کننده می‌باشد. مثلاً در یک صفحه شامل ۱۰۰ تصویر اگر در انتهای آدرس عبارت type=book ? باشد شاید منظور این است که فقط تمام تصاویر با نوع کتاب را نمایش بدهد که با حذف این مورد ما به جای لیست فیلتر شده لیست کامل را ببینیم. و یا بخش‌های مانند اندیس که قسمتی از صفحه وب است که با علامت # شروع می‌شود و معمولا کار خاصی انجام نمی‌دهد به عنوان مثال برای رفتن به بالای صفحه وب.

http://www.example.com/bar.html#top → http://www.example.com/bar.html

- حذف و تبدیل آی پی[۱] به دامنه : هر آدرس در بستر وب به یک آدرس آی‌پی (IP) تبدیل می‌شود این آدرس از یکسری عدد تشکیل شده است و کاملاً منحصر بفرد است. در حقیقت آدرس URL و آی پی هر دو یک آدرس را نمایش می‌دهند ولی آدرس URL عمومیت راحتتر و ساده‌تری دارد. پس بهتر است آدرس URL را بدست بیاوریم.

http://208.77.188.166/ → http://www.example.com/

- اعمال محدودیت بر روی پروتکل‌ها مانند تبدیل https به http : دلیل این کار این است که معمولا سایت‌ها در بستر اینترنت جهت جلوگیری از خرابکاری و هک[۲] از پروتکل‌های امنیتی مختلفی در

---

[۱] internet protocol (IP)

[۲] hack

سایت استفاده می‌کنند. مثلا آدرس سایتی که با HTTPS شروع می‌شود غالبا از نظر امنیتی بسیار سخت گیر بوده و کار خزنده وب را برای جستجو در سایت دشوار می‌کند ولی آدرسی که با HTTP شروع می‌شود غالبا محدودیت های کمتری به خزنده وب اعمال می‌کند و مجوز جستجوی بیشتری را به خزنده وب می‌دهند.

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

- حذف پورت پیش فرض(پورت ۸۰ به صورت پیش فرض برای http می‌باشد) : پورت در آدرس وب نمایشدهنده یک آدرس منحصر بفرد در سروری می‌باشد که سایت در آن قرار دارد (ممکن است خدمات دهنده تحت وب[۱] چند سایت را در خود داشته باشد که هر کدام پورت مخصوص به خود را دارند) ولی نیازی به وارد کردن آن در هنگام ورود به صفحه وب نیست پس جهت خلاصه سازی آدرس بهتر است مقدار پورت آدرس حذف گردد. از طرف دیگر ممکن است خدمات دهنده آدرس پورت را بنا به دلایلی عوض کند مثلا از پورت ۱۲۳ به ۲۴۵ تغییر دهد پس خزنده وب دچار گمراهی می‌شود.

http://www.example.com:80/bar.html → http://www.example.com/bar.html

http://www.example.com:80/bar.html → http://www.example.com/bar.html

- حذف www از اول دامنه : این کلمه مشخص کننده دارا بودن لایسنس وب گستر جهانی سایت می‌باشد و بهتر است از آدرس حذف گردد.

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

---

web server[۱]

- حذف متغیرهای اختیاری از رشته پرس‌وجو[1]: در آدرس منظور از رشته پرس‌وجو یک پرس‌و جوی مشخص است که کار خاصی را روی صفحه وب انجام داده و محتوای سایت را تغییر می‌دهد بعضی از عناصر این پرس‌وجو تغییر خاصی در نتایج انجام نمی‌دهند مثلا وجود پارامتر&sort=ascending یعنی نتایج را به ترتیب صعودی مرتب کنید که تاثیری در نتایج نمی‌دهد پس بهتر است از آدرس حذف شود. و یا مثلا در صورتی که هیچ عبارت پرس‌وجوی در آدرس نیست بهتر است علامت ? از آدرس حذف گردد.

http://www.example.com/display?id=&sort=ascending → http://www.example.com/display

۲-۵-۲ بررسی آدرس صفحات وب

- معرفی غالب دامنه[2] و صفحات وب : دامنه یک آدرس منحصر بفرد می‌باشد که نشان دهنده یک سایت خاص است و هر وب سایت دارای یک دامنه می‌باشد. این دامنه در سرویس خدمات دهنده قرار دارد غالب یک دامنه را در زیر می‌بینید. در این جا قسمت‌های که داخل کروشه ([]) قرار دارند می‌توانند در یک دامنه باشند و یا وجود نداشته باشد.

protocol://[subDomains.]domainName.netType[:portNumber]

حال به معرفی قسمت‌های مختلف یک دامنه می‌پردازیم.

- پروتکل اینترنت[3](protocol): پروتکل اینترنتی دربردارنده یک قرار دادی است جهت مبادله اطلاعات در شبکه‌های اینترنتی و هر پروتکل مشخص کننده نوع اهمیت صفحه وب می‌باشد به عنوان مثال پروتکل http لایه امنیتی پایین‌تری نسبت به پروتکل https دارد.

---

query string [1] domain[2] internet protocol[2]

- نام دامنه (domainName) : نام دامنه در حقیقت اسم وب سایت می‌باشد و به طور کامل منحصر بفرد است.
- زیر دامنه (subDomain) : ممکن است یک دامنه شامل چندین زیر دامنه هم باشد که در حقیقت وظیفه قسمت کردن و گروه‌بندی اطلاعات وب‌سایت را بر عهده دارند. به عنوان مثال اگر یک دامنه با نام http://www.gamefa.com داشته باشیم یکی از زیر دامنه‌های آن مثلا http://www.cinema.gamefa.com می‌باشد.
- نوع نِت (netType) : این قسمت مشخص کننده استاندارد نت می‌باشد مثلا نوع نت com. نشان دهنده یک سایت عمومی و یا ir. نشان‌دهنده وجود دامنه در محدوده ایران می‌باشد.
- شماره پورت (portNumber) : هر خدمات دهنده وب شامل یک یا چند دامنه است که هر کدام را با یک شماره پورت منحصر بفرد مشخص می‌کند.

- حال به معرفی ساختار یک آدرس وب می‌پردازیم. در این جا قسمت‌های که داخل کروشه ([]) قرار دارند می‌توانند در یک صفحه وب باشند و یا نه.

domain/path/[?query][#hash]

- دامنه (domain) : این قسمت همان دامنه است که در قسمت قبل به طور کامل معرفی گردید.
- مسیر (path) : این قسمت مشخص کننده مکان مشخصی در دامنه است و شامل یک مسیر مشخص در دامنه است به عنوان مثال https://a-z-animals.com/animals/dolphin/ مسیر مشخصی را در دامنه معرفی می‌کند که ما را به بخش حیوانات (/animal) و در نهایت صفحه وب معرفی دُلفین (/dolphin) هدایت می‌کند.
- پرس‌وجو (query) : این قسمت جهت فیلتر کردن یک صفحه خاص مورد استفاده قرار می‌گیرد به عنوان مثال در یک صفحه شامل ۱۰۰ تصویر درباره کتاب‌های مختلف که با استفاده از پرس‌وجوی

مثل type=computer? فقط کتاب‌های با عنوان کامپیوتر را مشاهده می‌کنیم. بهتر است جهت جلوگیری از فیلتر و حذف قسمت‌های از صفحه وب این قسمت به طور کامل از آدرس حذف گردد.

- اندیس (hash) : این قسمت از بخش‌های اختیاری یک آدرس وب می‌باشد. این بخش فقط جهت ارجاع به یک قسمت مشخص از صفحه وب می‌باشد و کاربرد خاص دیگری ندارد به عنوان مثال وجود عبارت film# ما را به بخش معرفی فیلم در صفحه وب می‌برد. بهتر است این قسمت هم از صفحات وب حذف شود.

۲-۶ قسمت های مهم یک خزنده وب :

۲-۶-۱ شاخص گذاری[۱]

در این قسمت هدف نگاشت نتایج به دست آمده از صفحات وب به پرس‌وجوهای پرسیده شده از موتورجسجو می‌باشد. این کار به منظور سهولت در دستیابی به اطلاعات بدست آمده از صفحات وب می‌باشد. به عنوان مثال موتور جستجوی گوگل اطلاعات بیش از هزاران میلیارد سایت را در یکصد میلیون گیگابایت شاخص گذاری کرده است. البته بعضی از پرس‌وجوها دارای شباهت های از نظر معنای هستند که در یک گروه اندیس می‌شوند مانند دو پرس‌وجوی **"اطلاعاتی درباره انواع میزها"** و **"معرفی میزهای مختلف"** که در حقیقت به یک موضوع مشابه اشاره دارند.

---

indexing[۱]

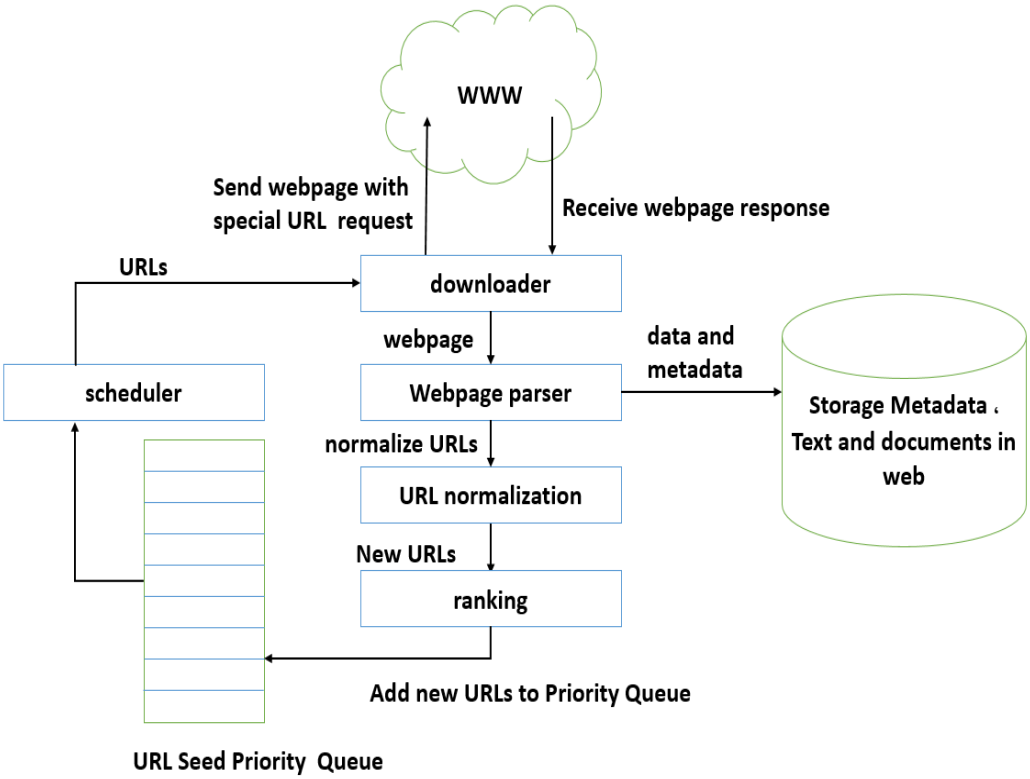
۲-۶-۲ پرس‌وجو[۱]

مفهوم پرس‌وجو در خزنده وب و بررسی موضوعات مشابه با شاخص ها (جهت سرعت در بررسی نتایج) است. در حقیقت پرس‌وجو یک نوع طبقه بندی نتایج صفحات وب است. زمانی که کاربر سوالی از موتور جستجو می‌پرسد. این جستجو در نهایت به یکی از طبقه بندهای موتور جستجو (یا چند تا با درجه اولویت مختلف) وصل شده و موتور جستجو بر حسب اولویت اطلاعات طبقه مورد بحث نتایج را نمایش می‌دهد. از طرف دیگر یکی از کاربردهای پرس‌وجو ایجاد یکسری سوال و درخواست توسط بخشی در موتور جستجو به نام مبدل مساله[۲] است. به این صورت که وقتی کاربر سوالی از موتور جستجو می‌پرسد مبدل مساله درخواست های مشابه با آن ایجاد می‌کند. این کار باعث گسترش و پیشرفت بهتر موتور جستجو می‌شود. در این پروژه همانطور که در فصل‌های جلوتر خواهیم دید ما از ۸۰۰ پرس‌وجوی مختلف برای تشخیص اشیاء استفاده می‌کنیم.

2-6-3 رتبه بند[۳]

هدف از این قسمت دادن امتیاز به صفحات بررسی شده توسط موتور جستجو می‌باشد تا در بررسی‌های بعدی جهت به روز رسانی اطلاعات درباره یک موضوع یا موضوع‌های مشابه با آن برحسب این امتیاز صفحات اولویت‌بندی شوند. آردس صفحات وب بسیار زیاد می‌باشند و در بعضی مواقع هیچگونه شباهتی با موضوع مورد جستجو ندارند پس بهتر است به هر آدرس یک امتیاز خاص درباره یک موضوع خاص داده شود. در فصل های بعدی به معرفی راهکارهای برای دادن امتیاز به صفحات وب بر حسب یک روش پیشنهادی می‌پردازیم. در شکل ۲-۲ تصویری از معماری یک خزنده وب مشاهده می‌شود.

Query[۱] problem generator[۲] ranking[۳]



شکل ۲-۲ معماری کامل یک خزنده وب [1]

۲-۶-۴ Frontier Crawler

این بخش شامل لیستی از آدرس صفحات وب معتبر[۱] می‌باشد. این لیست شامل مجموعه ای از آدرس صفحات وب با نامه دانه[۲] می‌باشد. که یا توسط کاربران موتور جستجو معرفی شده اند و یا توسط خزنده وب در صفحات وب یافت شده‌اند. البته استخراج آدرس بر حسب اولویت می‌باشد. در پروژه ما ابتدا به معرفی یک لیستی از آدرس دامنه های مشخص پرداخته و خود موتور جستجو با استخراج آدرس های جدید از صفحات بررسی شده می‌پردازد.

valid URL[1] seed[2]

۲-۶-۵ دریافت کننده[۱]

مهمترین کار این بخش همانطور که از اسمش مشخص است دریافت صفحه وب مشخصی از لیست Frontier می‌باشد (بر حسب اولویت). بنابراین این قسمت از خزنده یک درخواست HTTP برای صفحه مورد نظر به سروری که شامل این صفحه وب است فرستاده و نتیجه ارسال شده از سرور را که همان صفحه وب است را دریافت می‌کند اطلاعاتی که این بخش دریافت می‌کند همان صفحه وب با استاندارد HTML می‌باشد. و سپس به استخراج اطلاعاتی مانند متن، تصویر و آدرس صفحات دیگر داخل صفحه دریافتی می‌پردازیم.

۲-۶-۶ مخزن[۲]

این قسمت جهت ذخیره و مدیریت لیست بزرگی از داده ها استفاده می‌شود. البته مخزن تنها فقط صفحات وب استاندارد (html) را شامل می‌شوند به همراه تاریخچه‌ای از آنها. در اینجا منظور از تاریخچه اشاره به دوره‌های مختلفی است که موتور جستجو یک صفحه وب را دریافت می‌کند زیرا اطلاعات و ساختار صفحات وب دچار تغییرات دوره ای می‌شود که بسته به نوع صفحه وب کاملاً متغیر می‌باشد به عنوان مثال صفحه اصلی یک سایت

خبری تغریبا هر ۱۰ دقیقه یکبار تغییر می‌کند ولی یک صفحه آموزشی ممکن است هفته‌ها دچار تغییر نشود پس باید در دوره های مختلف به دریافت این صفحات وب پرداخت.

۲-۷ انواع موتورهای جستجوی وب :

در اینجا به معرفی انواع موتورهای جستجو می‌پردازیم. شکل۲-۳ الگوریتم کامل یک خزنده وب را نمایش می‌دهد.

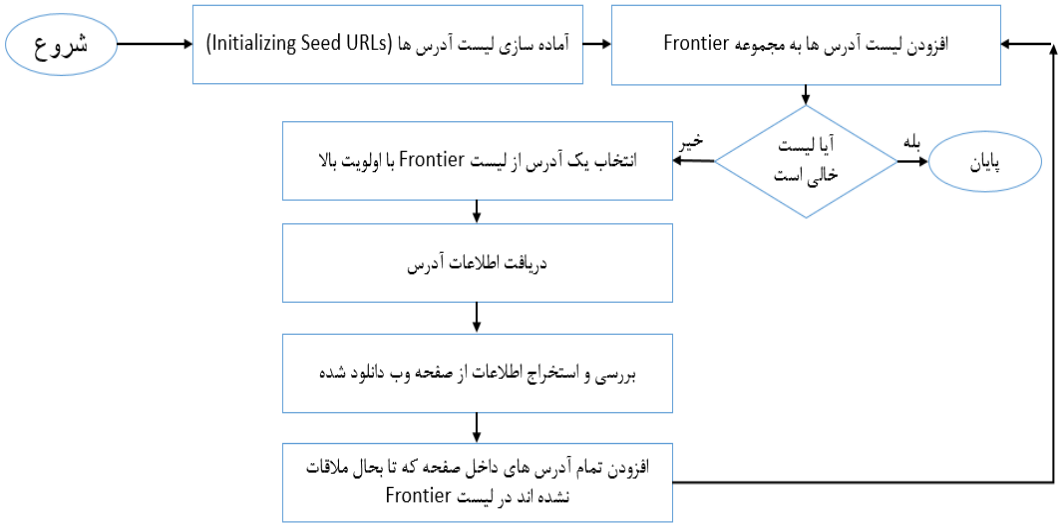
---

۱ Downloader ۲ repository

۱-۷-۲ خزنده وب موازی[1]:

این نوع از خزنده وب شامل چندین پردازنده جهت بررسی صفحات وب به فرم موازی است (دریافت صفحات وب و ذخیره آنها به فرم محلی و استخراج لینک و اطلاعات از صفحات وب به فرم موازی توسط چند خزنده وب انجام

می‌گیرد). در این نوع موتورهای جستجو باید هریک از فرایندها با بقیه ارتباط داشته باشند جهت به روز رسانی موتور جستجو و یکی شدن اطلاعات هر کدام از فرایندها.



شکل ۲-۳ الگوریتم کامل مراحل یک خزنده وب [1]

- مزایا

- پردازش موازی و سریع تر داده ها

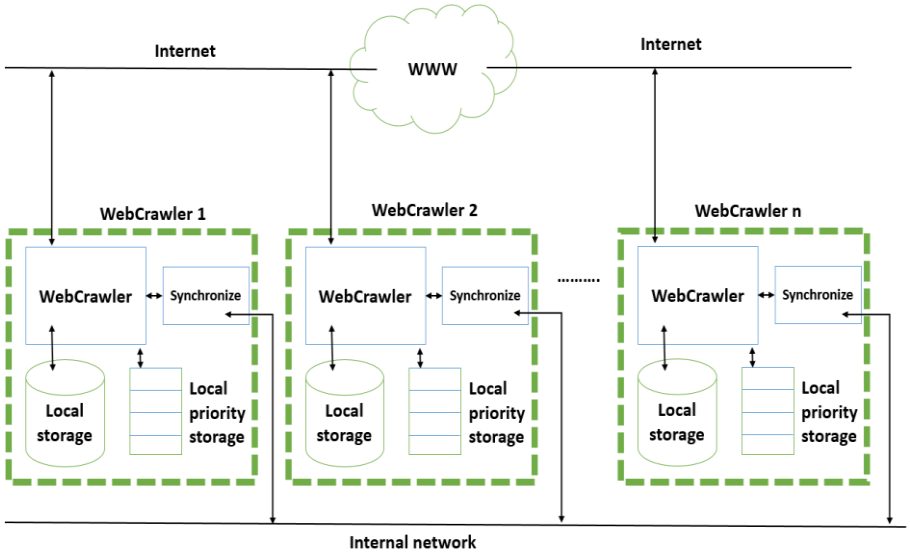
Parallel[1]

- مقیاس پذیری[1] : تعداد صفحات وب به فرم نمایی در حال رشد است پس برای یک خزنده وب تنها غیر ممکن است بدست آوردن و بررسی تمام صفحات وب در یک زمان معقول پس به فرم همزمان از چند موتور جستجو استفاده می‌شود.
- قابلیت مقایسه نتایج جستجو و اشتراک گزاری

- معایب

- داده های تکراری : به علت عملکرد موازی ممکن است چند خزنده وب یک صفحه را به فرم جدا بررسی کنند و نتایج مشابه زیادی در موتور جستجو و مخزن آن ذخیره شود.

شکل ۲-۴ یک نمایش بلوک دیاگرامی از موتور جستجوی موازی را نمایش می‌دهد.



شکل ۲-۴ خزنده وب موازی [1]

۲-۷-۲ خزنده وب توزیع شده[2]

Scalability[1] distributed[2]

این خزنده وب صفحات وبی را که با هم رابط دارند دریافت کرده. این صفحات با موضوع مورد جستجوی خاص موتور جستجو هم ارتباط دارند. به فرم تخصصی این نوع خزنده از نمونه‌ها و کاوش در وب از طریق ارتباطات هدایت شده یاد می‌گیرد و شامل ۳ مولفه زیر است.

۱- کلاسه بند[۱] : این قسمت جهت طبقه‌بندی مستندات وب به چند گروه مختلف استفاده می‌شود و اطلاعات به دست آمده از صفحه‌وب را در گروه‌های مختلف ادغام می‌کند.

۲- تقطیرکننده[۲]: به فرم خلاصه وظیفه این قسمت بررسی ارزش صفحه با موضوع مورد بررسی می‌باشد. البته این ارزش روی صفحات و لینک‌های مرتبط با صفحه انتخابی هم تاثیر دارد، و ممکن است ارزش این صفحات را تغییر دهد.

۳- خزنده : جهت بررسی صفحه از این قسمت استفاده می‌شود. و وظیفه آن دریافت صفحه وب است.

- مزایا

- یادگیری هوشمند و گسترده تر داده ها : به خاطر طبقه بندی اطلاعات.
- سهولت دادن امتیاز به صفحات وب به خاطر استفاده از سیستم تقطیر کننده

- معایب

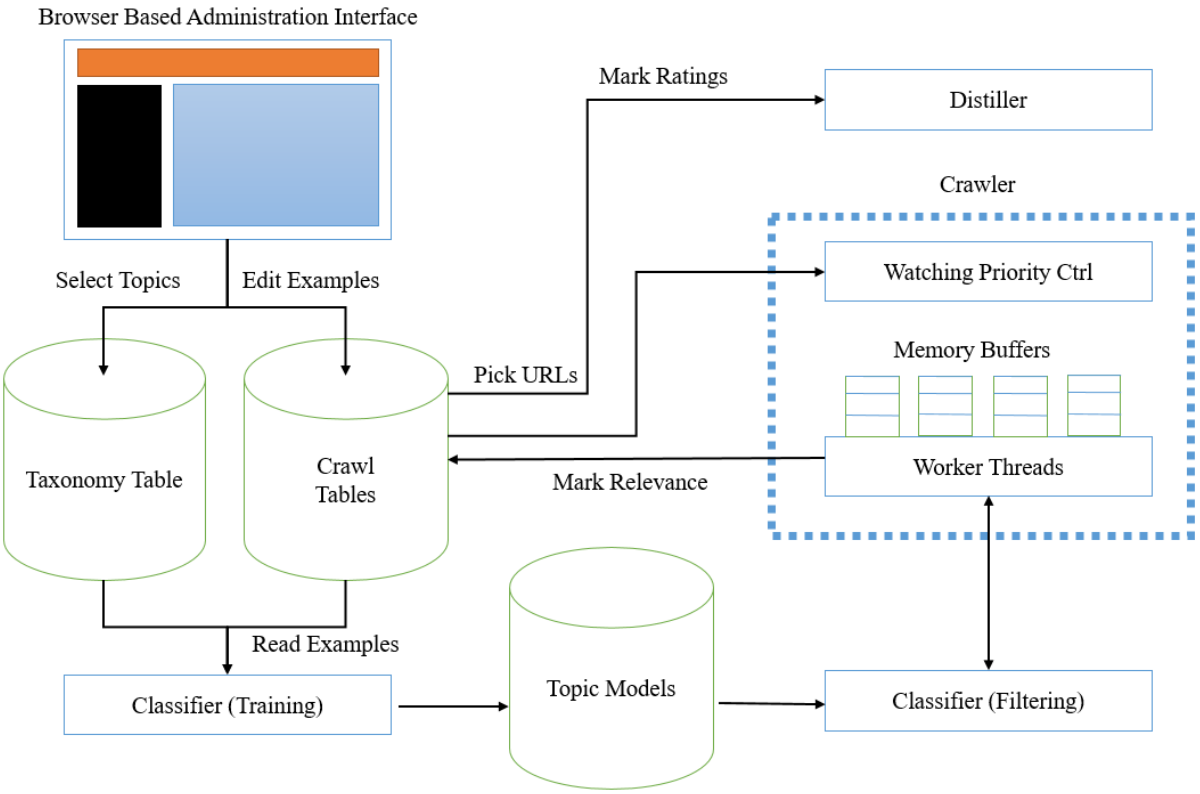
- عدم توانایی بررسی صفحات مخفی : منظور از صفحات مخفی صفحاتی است که جهت ورود به آنها نیازمند تشخیص هویت می‌باشد.

شکل ۲-۵ یک نمایش بلوک دیاگرامی از موتور جستجوی توزیع شده را نمایش می‌دهد.

---

۱ Classifier

۲ distille



شکل ۲-۵ خزنده وب توزیع شده [1]

۳-۷-۲ خزنده وب افزایشی[1]

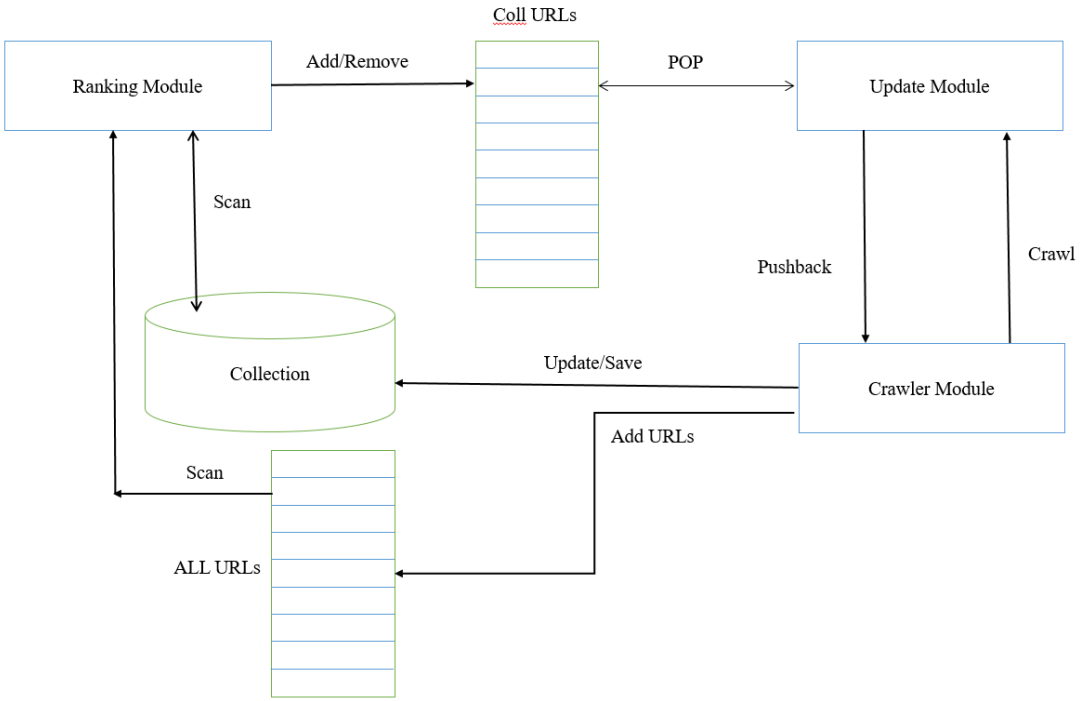
در این نوع از خزنده وب صفحات وب با یک دوره زمانی مشخص بررسی می‌شوند. در این روش حتی امکان پاکسازی بعضی صفحات، جهت آزاد سازی فضا برای دریافت صفحات جدید است. در این نوع از خزنده وب اهداف زیر تحقق می‌یابد[1].

- تازه نگهداشتن اطلاعات محلی : با بررسی دوره‌ای صفحات امکان تازه نگهداشتن اطلاعات و کسب اطلاعات بهتر بیشتر می‌شود. به دلیل تغییر مداوم اطلاعات در بستر وب بسیار مهم است که موتور جستجو به سرعت از تغییرات صفحات وب آگاه شود.

Incremental[1]

- بهبود کیفیت مجموعه محلی : با بررسی دوره‌ای صفحات وب با صفحاتی که دارای اطلاعات با ارزش‌تری هستند آشنا می‌شویم. مثلا اگر صفحه ای مدام ارزش اتصالش در بررسی ها پایین برود به دلایل مختلف مثل کاهش بازدید و یا شباهت کمتر با موضوعات پرس‌وجوی ما (در این پروژه ۸۰۰ گروه مختلف داریم) پس در دوره های بعدی ارزش کمتری در لیست اولویت می‌گیرد.
- معایب
  - نتایج ممکن است خیلی با موضوع بررسی خزنده وب هم‌خوانی نداشته باشند
  - بار شبکه : به علت بررسی مداوم و تکراری صفحات یک بار اضافه به شبکه تحمیل می‌شود.
  - عدم توانایی بررسی صفحات مخفی

شکل ۲-۶ یک نمایش بلوک دیاگرامی از موتور جستجوی افزایشی را نمایش می‌دهد.



شکل ۲-۶ خزنده وب افزایشی [1]

2-7-4 خزنده وب مخفی

به فرم معمول موتور جستجو قادر به بررسی صفحاتی که نیازمند تشخیص هویت‌اند نیست از این جهت از خزنده های وب مخفی استفاده می‌شود [3] که با استفاده از یک جدول به نام انتساب دهنده ارزش[۱] که به اختصار LVS گفته می‌شود و ۴ مولفه زیر اقدام به تشخیص هویت و ورود به صفحات وب مخفی می‌کند.

- تجزیه و تحلیل فرم[۲] : بررسی اطلاعات داخل فرم ثبت نام درون وب سایت. مثلا فرم حاوی یک المنت ورودی ایمیل است و یک فیلد تعیین جنسیت و ...
- انتساب دهنده ارزش[۳] : در این مرحله ما از جدول LVS که نوعی نگاشت دهنده از جدول به المنت های داخل فرم است استفاده کرده و فیلدهای داخل فرم را پر می‌کنیم.هر المنت داخل جدول حاوی لیستی از مقادیر متفاوتی است به عنوان مثال فیلد ایمیل شامل لیستی از ایمیل های معتبر می‌باشد که انتساب دهنده ارزش یکی از آنها را در فرم قرار می‌دهد.
- ارسال کننده فرم[۴] : این قسمت مسئول ارسال فرم پر شده به سرور وب سایت مشخص شده می‌باشد
- تجزیه و تحلیل پاسخ[۵] : این قسمت مسئول بررسی پاسخ دریافتی از سرور می‌باشد و رفع مشکلات احتمالی (مثلا اگر ایمیل قبلا در سرور ثبت شده دوباره از جدول LVS یک ایمیل جدید در فیلد مربوطه قرار دهد و دوباره فرم را به سرور ارسال کند.)

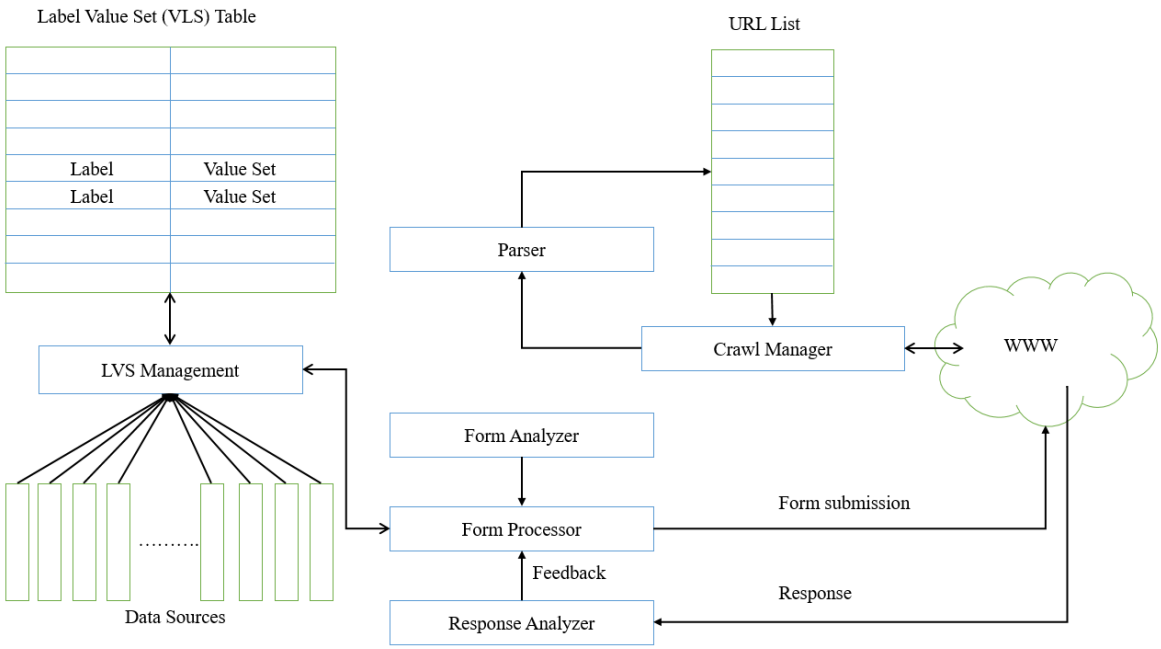
---

label value set [۱] form analysis[۲] value assignment[۳]

form submission[۴] response analysis[۵]

- معایب
  - بار شبکه : به علت ارسال و بررسی مرتب فرم با جدول LVS نوعی بار اضافه در شبکه ایجاد می‌شود.

شکل ۲-۷ یک نمایش بلوک دیاگرامی از موتور جستجوی مخفی را نمایش می‌دهد.



شکل ۲-۷ خزنده وب مخفی [1]

۲-۸ معرفی چند موتور جستجوی مهم

بعضی موتورهای جستجو کاربردخاص منظوره دارند به عنوان مثال فقط در زمینه تشخیص هرزنامه‌ها و یا تشخیص تصاویر در صفحات وب استفاده می‌شوند ولی بعضی از موتورهای جستجو کاربرد های عمومی‌تر دارند و در همه موارد استفاده می‌شوند. جدول ۲-۱ لیست از ۵ موتور جستجوی محبوب سال ۲۰۱۶ را نمایش می‌دهد البته بعضی موتور های جستجو حالت کاربردی در علوم مختلف را هم ایفا می‌کنند به عنوان

مثال یافتن اطلاعات درباره یک موضوع خاص یا حتی جمع آوری داده برای یک موضوع مشخص مانند دریافت لیست تصاویر درباره صورت انسان جهت پردازش تصویر و یا ایجاد یک پایگاه داده اولیه برای یک فرایند یادگیری. در این پروژه هم با استفاده از استخراج متن و تصاویر به گروه بندی و تشخیص اشیاء می‌پردازیم.

جدول ۲-۱، ۵ موتور جستجوی محبوب وب

ردیف	نام	لوگوی موتور جستجو
۱	Google	
۲	Yahoo	
۳	Mahalo	
۴	webopedia	
۵	Bing	

۲-۹ معرفی فایل robots.txt

این فایل برای دادن اطلاعات اولیه در زمینه وب سایت مورد پردازش به موتور جستجو استفاده می‌گردد. به عنوان مثال با این فایل خدمات‌دهنده صفحه وب می‌تواند میزان دسترسی موتور جستجو به بعضی زیر شاخه ها را مشخص کند. به عنوان مثال دستورات زیر که داخل فایل robots.txt قرار دارد از دسترسی موتور جستجو

به مسیر[1] /tmp/ جلوگیری می‌کند:

```
User-agent: *
Disallow: /tmp/
```

فایل robots.txt یک استاندارد می‌باشد. به همین جهت موتور جستجو (معمولا موتورهای جستجو مخرب) می‌تواند آن را نادیده بگیرد. ولی این فایل بیشتر جهت راهنمایی یک موتورجستجو برای دسترسی سریع‌تر و با اطمینان بالاتری به اطلاعات کلی صفحات وب یک دامنه خاص است.

---

[1] Path

فصل سوم:

معرفی چالش ها،

راه حل ها و نوآوری های

موتورجسجو

۳-۱ مقدمه

دراین فصل ابتدا به معرفی ساختار داخلی صفحات وب پرداخته و در ادامه به بررسی چند مشکل و راه‌کارهای برای آنها در موتور جستجو می‌پردازیم.

۳-۲ دریافت و بررسی صفحات وب

۳-۲-۱ مقدمه

همانطور که گفته شد صفحات وب در کنسرسیوم جهانی وب قالب استاندارد HTML می‌باشد. اما غالب یک HTML به چه صورت می‌باشد. در فصل اول و به طور مختصر با مفهوم زبان HTML آشنا شدیم در اینجا بهتر است قبل از جمع آوری اطلاعات از یک صفحه وب با غالب صفحه وب و کلیاتی از آن آشنا شویم.

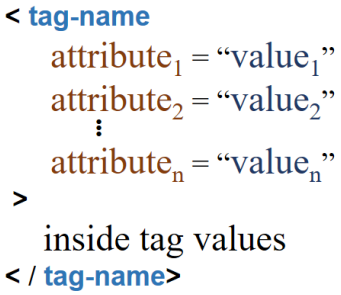
۳-۲-۲ معرفی غالب یک صفحه وب

یک صفحه ابر متن از یکسری المنت با نام تگ[۱] تشکیل می‌شود غالب کلی یک تگ را می‌توانید در شکل ۳-۱ ببینید. همانطور که در تصویر مشخص است هر تگ دارای یک نام منحصر به فرد است و یکسری ویژگی که هر کدام یک نام منحصر به فرد و مقدار دارند. هر تگ مخصوص به کار خاصی است مثل یکسری تگ برای درج متن، بعضی برای نمایش فایل‌های صوتی و تصویری و غیره استفاده می‌شوند. بعضی ویژگی‌ها نیز در تگ‌ها برای ظاهر و غالب تگ تعریف می‌شوند و بعضی ها مقادیر مشخصی به تگ می‌دهند به عنوان مثال ویژگی کلاس[۲] در هر تگ برای تعریف یک غالب و ظاهرو یا ویژگی value به مقدار تگ اشاره می‌کند. در جدول ۳-۱ لیستی از چند تگ مهم در صفحات وب معرفی شده است. شناسایی این تگ ها در صفحات

---

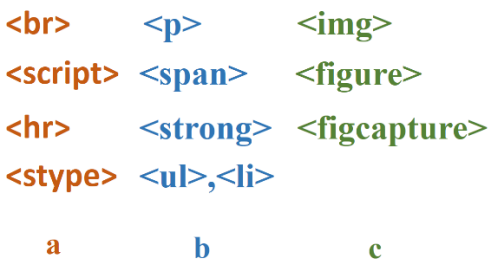
[۱] Tag [۲] class

وب از اهمیت بالای برخوردار است به این دلیل که با شناسایی و کاربرد هر تگ در صفحات وب می‌شود از ماهیت و کاربردهای یک صفحه وب خاص



شکل ۳-۱ غالب کلی یک تگ HTML

مطلع گردید و اهمیت صفحه درباره یک موضوع خاص مورد پرس‌وجو پررنگتر شود. مثلا اگر در یک صفحه خاص تگ شکل زیادی موجود باشد می‌شود در بررسی تصاویر بیشتر به این صفحه مراجعه کرد و یا اگر صفحه ای دارای لینک پیوندی زیاد است به احتمال زیاد این صفحه یک صفحه خبری و یا حتی تبلیغاتی می‌باشد. البته باید به کمک تکنیک‌های هوشمند و روش‌هایی امتیاز دهی دستی (رای کاربران در بازدید سایت بعد از درخواست یک سوال و انتخاب لینک نتیجه) در گرفتن امتیاز صفحه در یک مورد خاص اهمیت داد. در شکل ۳-۲ لیستی از چند تگ پرکاربرد صفحات وب را می‌توان دید همانطور که در این شکل مشخص است تگ‌های قسمت (a) حاوی اطلاعات مفیدی نیستند و بیشتر به منظور تغییر ظاهر صفحه وب استفاده می‌شوند. از طرفی شانس مشاهده متن در تگ‌های قسمت (b) در شکل بالا است و همچنین احتمال مشاهده یک تصویر در تگ‌های قسمت (c) بالا است.



شکل ۳-۲ چند تگ مهم و پرکاربرد

جدول ۳-۱ لیست تگ‌های پر کاربرد صفحات وب

ردیف	نام تگ	توضیحات	نام چند ویژگی مهم		
۱	<img>	شامل یک تصویر در صفحه وب و فاقد فرزندی می‌باشد	src width height alt	= آدرس تصویر = عرض تصویر = ارتفاع تصویر = توضیحات تصویر	
۲	< button>	تعریف یک دکمه قابل کلیک در صفحه وب	name	= نام کلیک	
۳	<figure>	شبیه تگ img فقط دارای یک بخش توضیح هم است.	figcaption src	= توضیحات تصویر = آدرس تصویر	
۴		نمایش توضیحات و متن در صفحات وب	value	= مقدار	
۵	<menu>	نمایش یک لیست انتخاب شوند در وب (غالبا در صفحات وب برای نمایش منو ها و اتصال به صفحات دیگر از این تگ استفاده می‌شود)	menuitem value	= مقداریک عنصر = مقدار	
۶	<p>	نمایش یک پاراگراف در صفحات وب (این تگ بیشتر برای نمایش توضیحات و پاراگراف ها در وب استفاده می‌شود)			
۷	<meta>	این تگ جهت توضیحات صفحه وب بکار می‌رود و به عنوان مثال ویژگی keywords شامل لیستی از کلمات کلیدی صفحه وب فوق می‌باشد که برای راهنمایی موتور جستجو و یا فریب آن استفاده می‌شود.			

۸	<a>	این تگ جهت ارتباط با دیگر صفحات وب استفاده می‌شود و بیشتر در صفحاتی با غالب خبری و تبلیغاتی استفاده می‌شود.	href = آدرس صفحه وب دیگر

۳-۲-۳ دریافت کننده[۱] صفحه وب : در شکل ۳-۳ دیاگرامی از وظایف این بخش نمایش داده شده است. حالا به معرفی قسمت‌های مختلف این قسمت از موتور جستجو می‌پردازیم.

۱- درخواست آدرس صفحه وب با بالاترین اولویت از صف seed و بررسی آن : منظور از بررسی آدرس اطمینان از عدم مشکل یا هرزنامه[۲] بودن آدرس می‌باشد. این کار به صورت تجربی و بررسی ارزش صفحه وب بدست می‌آید.

۲- ارسال درخواست صفحه وب انتخابی به خدمات دهنده[۳] آن و انتظار دریافت نتیجه از خدمات دهنده

۳- بررسی دریافت موفق صفحه وب : آیا صفحه وب با موفقیت از خدمات دهنده آن دریافت گردیده است (در صورت دریافت صحیح قسمت دریافت کننده پیغام ۲۰۰ را که نماینده جواب صحیح است دریافت می‌کند). در جدول ۳-۲ لیست نتایج برگردانده شده از خدمات دهنده دارای آدرس صفحه وب را نمایش می‌دهد. این نتایج در تعیین امتیاز یک صفحه وب کمک زیادی می‌کنند به عنوان مثال با ایجاد یک خطا در دریافت صفحه وب می‌شود امتیاز آن را کاهش داد.

---

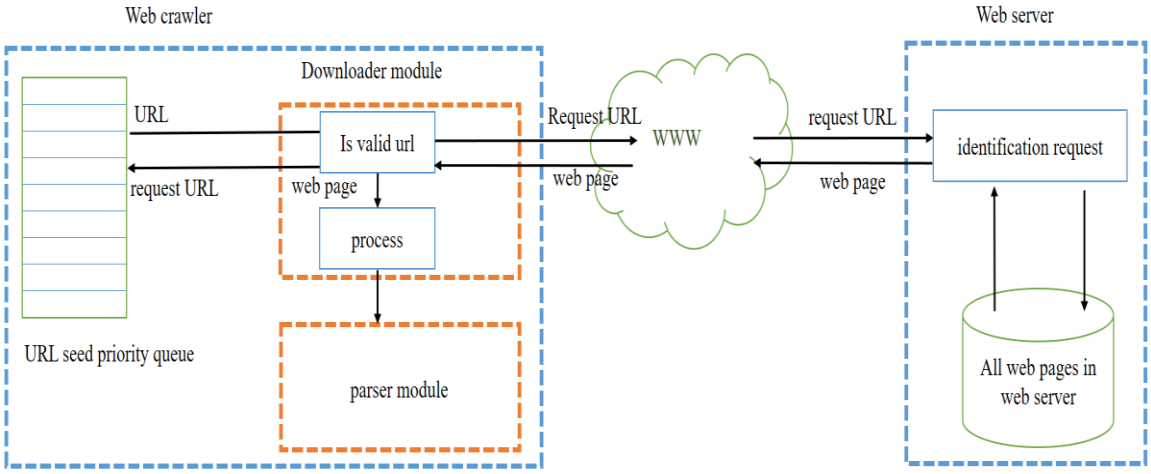
web page downloader[۱] spam[۲] web server[۳] redirection[4]

جدول ۳-۲ لیست نتایج درخواست یک صفحه وب

ردیف	نام تگ	توضیحات
۱	۲۰۰	صفحه وب با موفقیت از خدمات دهنده دریافت شد.
۲	۴۰۴	صفحه وب مورد نظر یافت نشد. (به دلایل مختلفی مانند حذف موقت صفحه وب به منظور به روز رسانی اطلاعات آن یا حذف دائمی آن و یا حتی تغییر آدرس صفحه‌وب[۴]
۳	۵۰۱	جهت مشاهده اطلاعات صفحه وب نیاز به اعمال اهراز هویت می‌باشد
۴	۴۰۶	خدمات دهنده به فرم موقت قادر به ارسال پاسخ (صفحه وب) نیست به دلایلی مانند حجم بالای درخواست های ارسالی به خدمات دهنده

پیش‌پردازش روی صفحه وب دریافتی : قبل از ارسال صفحه وب دریافتی به قسمت تجزیه کننده باید یکسری پیش‌پردازش روی صفحه وب انجام بگیرد :

- بررسی عدم خرابی تگ ها : در html باید تمام تگ ها بسته شوند در غیر این صورت در هنگام تجزیه در ماژول تجزیه کننده ممکن است مشکلی ایجاد گردد. منظور از بسته بودن تگ داشتن انتها برای آن است (</tagname>) به دلیل غالب درختی یک صفحه وب.
- بررسی صحیح غالب صفحه وب : صفحه وب دریافتی از نظر غالب و سلسله مراحل تگ ها درست است. به عنوان مثال تگ <body> باید بعد از تگ <head> در صفحه وب قرار داشته باشد زیرا تگ head یکسری اطلاعات مهم را قبل از بارگذاری تگ body در صفحه وب دریافت می‌کند.



شکل ۳-۳ بلوک دیاگرام ماژول دریافت کننده

۳-۲-۴ تجزیه کننده[۱] صفحه وب

وظیفه مهم یک تجزیه کننده بعد از دریافت یک صفحه وب تجزیه کردن و استخراج اطلاعات مهم از داخل صفحه وب می‌باشد. حالا به معرفی چند مشکل در این ماژول و راهکارهای برای آنها می‌پردازیم.

۳-۲-۴-۱ شناسایی و اولویت بندی و حذف تگهای بی استفاده

یکسری از تگ ها در صفحات وب به هیچ عنوان جنبه اطلاعات نداشته و فقط برای کارهای مانند زیباسازی صفحات وب، کارهای پیش زمینه[۱] (مثلا وقتی روی یک دکمه کلیک می‌کنید یک کار خاص انجام بگیرد) و ... استفاده می‌شوند. یکی از وظایف تجزیه کننده شناسایی این تگ ها در صفحه وب و حذف آنها است مطابق با شکل ۳-۲ قسمت (a). و یا حتی باید به تگ ها اولویت داده شود مثلا در یک صفحه خاص بعضی تگ ها دارای اهمیت و اولویت بالای هستند به عنوان یک مثال فرض کنید وارد یک صفحه وب شدید که کاملا در زمینه معرفی گُل‌ها است پس محتوای تگ های مانند <img> و <figure> در این صفحه باید

Parser[۱]

اولولیت بالای را بگیرند (البته اهمیت درباره پرس‌وجوهای با موضوعیت گل هم دارای اهمیت می‌باشد). چه راهکاری برای این امر باید انجام شود و تجزیه کننده چگونه باید تگ های بدون استفاده را از صفحه جدا کرده و برای هر تگ یک اولویت را در نظر بگیرد.

راه حل : بعضی از تگ ها فاقد محتوای مفید هستند مانند تگ style همانطور که درشکل ۳-۴ نمایش داده شده این تگ فقط جهت بهبود قالب صفحات وب می‌باشد و هیچگونه اطلاعات مفیدی در باره صفحه وب به ما نمی‌دهد اما تگ span نسبتا مفید است وحاوی اطلاعاتی درباره صفحه وب.

```
<style>
    .view-input{
        border-left:none;
        border-right:none;
        border-top:none;
        border-bottom:1px solid blue;
    }
</style>
<span
  type="text"
  style="color:#34ED29"
  class="view-input"
>
    محتوای این سایت درباره هوش مصنوعی است
</span>
```

شکل ۳-۴ مثالی از HTML

نوآوری

راهکار دیگر که در این پروژه استفاده شده است حذف تگ‌های می‌باشد که به مرور زمان ارزش و اهمیت خود رادرباره یک پرس‌وجوی خاص از دست می‌دهند. به عنوان مثال اگر موضوع مورد بررسی موتور جستجو درباره اخبار باشد طبق بررسی های انجام گرفته معمولا تگهای های مانند <p> و <h1> اولویت و داده های بیشتری نسبت به تگ دارند. به عنوان مثال دیگر در صفحات وبی مانند pinterest.com اهمیت تگ <img> که برای نمایش تصاویر می‌باشد بسیار بالا است و این صفحه غالبا از تصاویر تشکیل گردیده است.

ولی در صفحه وبی مانند www.cnn.com ما می‌بینیم که تگ <a> که یک تگ اتصالی بین صفحات وب

---

[۱] Background

است بسیار زیاد می‌باشد که غالباً هم همینطور است به دلیل ماهیت خبری این صفحه باید لینک اتصال به صفحات خبری دیگر هم زیاد باشد البته راهکار فقط روی زیاد بودن یک تگ در صفحه متمرکز نیست بلکه با استفاده از تکنیک یادگیری عمیق[۱] و شبکه های عصبی[۲] به مرور زمان متوجه می‌گردد که در صفحه وب مشخص مثلا تگ <h1> چه اهمیتی دارد. با بالا رفتن اولویت یک تگ در یک دامنه ارزش اطلاعاتی آن تگ در آن دامنه در یک پرس‌وجوی خاص هم بالا می‌رود.

پس تجزیه کننده برای هر صفحه یک جدول از لیست تگ های مهم با اولویت را مشخص می‌کند. و این لیست مدام در حال به روز رسانی می‌باشد. بعد از بررسی صفحات و استخراج ویژگی ها و با استفاده از پاداشی که از کاربران بازدید کننده می‌گیرد و آموزش های که می‌بیند این اطلاعات به روز می‌شود. پس این تگ در این صفحه به مرور زمان اهمیت بالای می‌گیرد.

در حقیقت ما در این پروژه و در این قسمت از یک تکنیک یادگیری عمیق برای امتیاز دهی هر صفحه استفاده می‌کنیم. البته بعضی از تگ ها و ویژگی های آنها هم در حذف و یا کاهش اولویت آنها اهمیت دارد که یا به فرم تجربی قابل استفاده است و یا با روش‌های یادگیری ماشین و شبکه عصبی. به عنوان مثال فرض کنید یک تگ شکل دارای ویژگی طول ۲ پیکسل می‌باشد بدیهی است این اندازه کوچک کمکی به موتور جستجو خصوصا در زمینه پردازش تصویر نمی‌کند و یا مثلا یک تگ پاراگراف که

فاقد محتوا است هم کمکی به پردازش نمی‌کند. و یا در زمینه یادگیری فرض کنید ۱۰۰ بار موتور جستجو صفحه مورد نظر را بررسی کرده است و به مرور متوجه شده ارزش تگی مانند <li> (این تگ برای لیست

بندی صفحات بسیار مهم است) در صفحه بسیار مورد توجه کاربران قرار گرفته و یا تگی به نام <meta>

---

[۱] deep learning

[۲] neural network

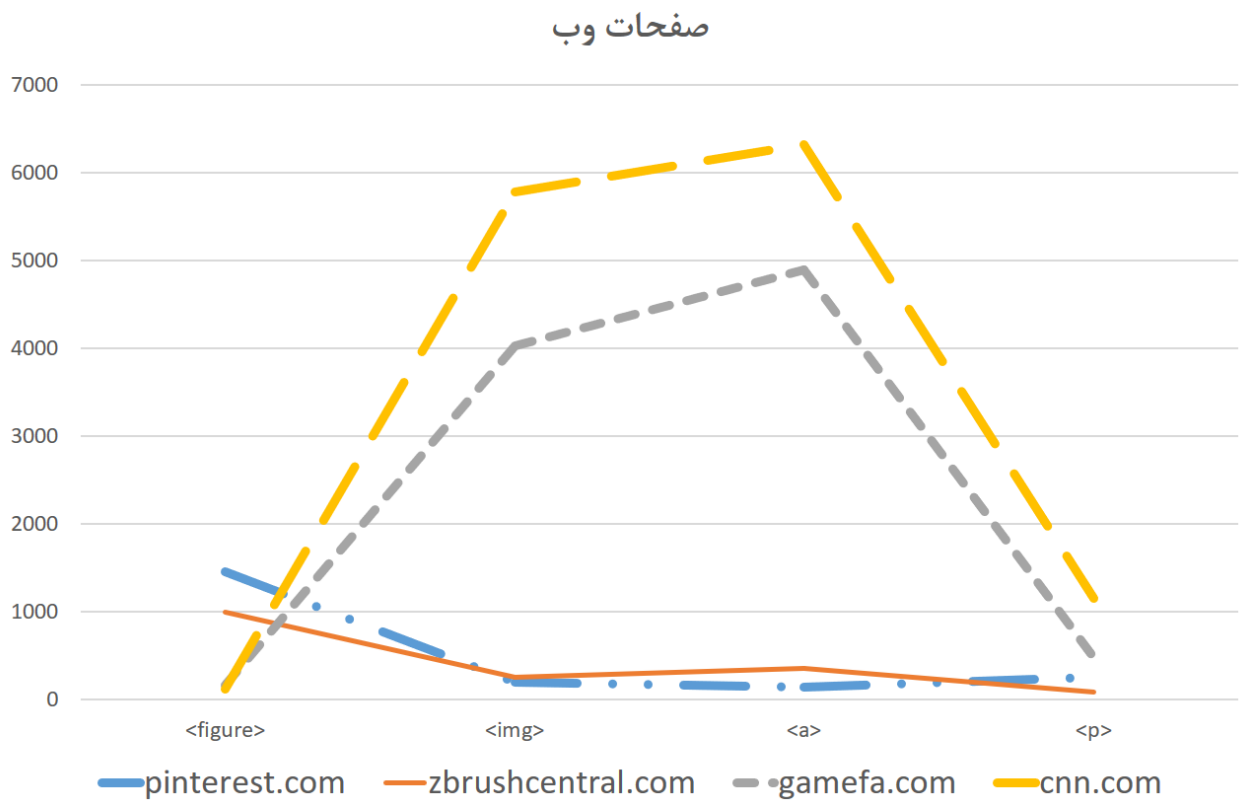
با ویژگی به نام keyword که در صفحه وب مشخص کننده کلمات کلیدی سایت است در صفحه این کلمات با موضوع صفحه ارتباطی ندارند پس به مرور زمان ارزش این تگ در این صفحه کمرنگ تر می‌گردد.

پس موتور جستجو به طور مداوم ارزش هر تگ را به فرم دوره ای در هر صفحه وب تغییر می‌دهد.

در این پروژه ما بعد از استخراج متن از نزدیک تصاویر صفحات وب به ایجاد ۸۰۰ گروه مختلف (پرس‌وجو) با استفاده از یادگیری عمیق و متن می‌پردازیم. و شناسایی و گروه بندی اشیاء در صفحه وب با کمک این ۸۰۰ گروه مشخص می‌شود.

در شکل ۳-۵ به دلخواه ۴ دامنه انتخاب شده‌اند ۲ دامنه با محوریت اخبار (gamefa.com، cnn.com) و ۲ دامنه دیگر با محوریت تصویر (pinterest.com، zbrushcentral.com).

همچنین ۴ تگ خاص از نظر تعداد تکرار و شباهت با پرس‌وجوهای مشخص شده در پروژه در این دامنه‌ها بررسی شده‌اند. (تگ تصویری<img> و <figure>، تگ ارتباطی <a>، تگ پاراگراف <p>) البته این ۴ دامنه شامل بیش از هزاران صفحه وب داخلی‌اند و اطلاعات مفید در زمینه های مختلف را در این ۴ دامنه بررسی شده است، همانطور که می‌بینید اهمیت تگ های تصویری و اتصالی در صفحات مربوط به گالری تصاویر بسیار بالا بوده ولی در صفحات اخباری تگ های پاراگراف و اتصال از تکرار و اهمیت موضوعی بالای برخوردارند. در اینجا ما ۸۰۰ گروه مشخص شده را با تمام متن‌ها و تصاویر نزدیک آنها در هر صفحه وب دامنه ها مقایسه کرده و درجه شباهت آنها را بدست می‌آوریم.



شکل ۳-۵ بررسی تگ ها در صفحات وب

اهمیت مکان تگ ها

آیا قرار گرفتن تگ در مکان خاص بر اهمیت آن می‌افزاید و یا تاثیر در آن ندارد. به عنوان مثال اگر یک تگ در ابتدای صفحه بیاید از نظر ارزش و اهمیت با تگی که در انتهای تصویر می‌آید تفاوتی با هم دارند.

البته بیشتر اهمیت یک تگ روی محتوای داده آن است ولی مکان قرار داشتن تگ هم تاثیری هر چند ناچیز در اهمیت تگ ایفا می‌کند. به عنوان مثال تگ <h1> که معمولاً در ابتدای صفحات وب می‌آید و بیشتر برای نمایش سرتیتر اطلاعات است از اهمیت بسیار بیشتری به نسبت تگ مشابهی که در انتهای صفحه می‌آید دارد. زیرا در ابتدای صفحه جنبه یک عنوان و در انتهای صفحه بیشتر جنبه تبلیغ و بی ارزشتری را دارد. در اینجا منظور از مکان تگ فاصله تگ تا ابتدای صفحه وب می‌باشد. در شکل ۳-۶ مشاهده می‌کنید که به عنوان مثال تگ <p>

در فاصله ۵ و تگ در فاصله ۳ تا ریشه صفحه وب (تگ <body>) قرار گرفته اند.

شکل ۳-۶ مکان تگ در درخت html

نوآوری : در این پروژه برای هر تگ در هر صفحه یک مکان قرار داده می‌شود (منظور افزودن یک ویژگی است که مکان این تگ را در صفحه وب معین می‌کند) و بعد از بررسی تگ ها و داده ها به فرم یادگیری عمیق و به

مرور زمان متوجه اهمیت مکان تگ در یک صفحه خاص می‌شویم. یعنی بعد از استخراج متن و تصویر از تگ ها و بررسی درصد شباهت آنها با پرس‌وجوهای داخل موتور جستجو به اهمیت و شباهت آنها با یک پرس‌وجوی خاص می‌رسیم.

۳-۲-۴-۲ استخراج ویژگی از تگ

بعد از تجزیه صفحه وب و حذف تگ های غیرضروری (پیش پردازش) و دادن امتیاز به هر تگ باید تگ ها مورد پردازش قرار بگیرند و بردار ویژگی آنها برای مراحل بعدی استخراج گردد. دراینجا هدف یافتن تصاویر و متون مفید و مرتبط با پرس‌وجوهای مشخص شده می‌باشد. البته همانطور که گفته شد هر تگ حاوی یکسری اطلاعات و ویژگی می‌باشد که در صورت نیاز این اطلاعات از تگ مربوطه استخراج می‌شود.

نوآوری :

در این پروژه ما یک ماشین یادگیری بدون ناظر داریم (تصاویر فاقد برچسب و گروه مشخص هستند و فقط از صفحات وب گرفته شده اند. همینطور ۸۰۰ پرس‌وجوی ایجاد شده نیز از متون نزدیک صفحات وب استخراج شده‌اند) پس باید ماشین به فرم تجربی بیاموزد و قابلیت شناسایی تصاویر مفید را داشته باشد. در اینجا به جز گرفتن تصاویر و استخراج بردار ویژگی برای آنها ما از تکنیک استخراج ویژگی از متون نزدیک تصویر هم استفاده کرده‌ایم. متون نزدیک به تصویر را با استفاده از یک تکنیک شبکه عصبی چند لایه و پردازش متن مورد پردازش قرار داده و بعد از یادگیری می‌شود به هر متن نزدیک تصویر یک درجه اهمیت داد بدین صورت که چقدر این تصویر بر اساس این متن با یک پرس‌وجوی داخل موتور جستجو شباهت دارد. البته این قسمت فقط درصدی از اهمیت تصویر را معین می‌کند. خود تصویر و استخراج ویژگی های آن و درصد شباهت آن با دیگر تصاویر یگ پرس‌وجوی خاص در این فرایند تاثیر دارد.

به عنوان مثال در سایت pinterest.com بعد از بررسی مشخص شد ویژگی alt که در تگ <img> قرار دارد و خلاصه ای از موضوع تصویر مورد جستجو را به ما می‌دهد کمک شایانی در تشخیص گروه تصویر می‌کند ولی در سایت دیگری مانند http://www.flowerpicturegallery.com که نمایش دهند تصاویری در زمینه گُل می‌باشد. تگ پدر تصویر از نوع دارای توضیحات خوبی درباره نوع و ارزش تصویر می‌باشد و تگ alt خود تصویر مطلب خوبی درباره تصویر نمی‌دهد (بیشتر به معرفی خود دامنه پرداخته است). پس باید برای هر صفحه وب بر حسب اهمیت هر تگ به محتوای متنی آن اهمیت و ارزشی بدهیم. دقت شود که برای هر تصویر ما چند لایه بالاتر آن را هم مورد بررسی قرار می‌دهیم و متون داخل آنها را با ارزش و اهمیت ها ی مختلف جمع آوری می‌کنیم.

بعد از استخراج متون از این تگ ها به استخراج ویژگی از خود تصویر می‌پردازیم.

۳-۴-۲-۳ استخراج ویژگی از تصویر

قبل از استخراج ویژگی‌های یک تصویر لازم است درباره ماهیت تصویر اطلاعاتی کسب کنیم.

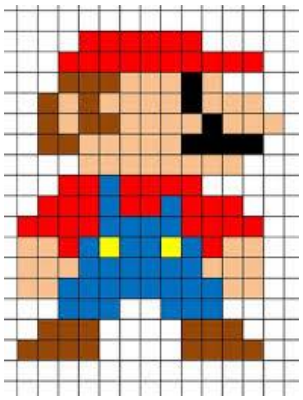
۱-۳-۴-۲-۳ معرفی تصویر

تصویر دیجیتال، از تعداد بسیار زیادی از مربع‌های کوچک به نام پیکسل[۱] تشکیل شده‌است. هر پیکسل دارای یک شماره رقمی[۲] است که بیان‌گر مقدار روشنایی آن پیکسل است. مقدار عددی پیکسل هم از صفر (سیاه) تا ۲۵۵ (سفید) متغیر است. به این نوع از تصاویر، تصاویر رستری[۳] هم می‌گویند. هر تصویر رستری از تعدادی سطر و ستون تشکیل شده است . در زیر چند نکته در مورد تصویر به صورت مختصر آورده شده است[17] .

معرفی پیکسل : برای مشخص کردن رنگ یک پیکسل، روش‌های مختلفی استفاده می‌شود. آنچه که متداول‌تر است RGB[۴] (آبی قرمز و سبز ۳ رنگ اصلی) است، که ۳ کانال مختلف برای ۳ رنگ قرمز، سبز و آبی در نظر می‌گیرند. اما در پردازش تصویر از فضاهای رنگی دیگر استفاده بیشتری می‌شود. برای مثال فضای رنگ HSV. هر کانال دارای ۲۵۶ حالت خواهد بود در نتیجه هر پیکسل می‌تواند ۱۶۷۷۷۲۱۶ رنگ مختلف را نشان دهد.

تفکیک‌پذیری تصویر : تفکیک‌پذیری تصویر به تعداد پیکسل‌ها در طول و عرض تصویر بستگی دارد. مانند شکل 3-7 که همانطور که می‌بینید تعداد پیکسل ها نمایش داده شده است[17] .

Pixel[۱] digital number[۲] raster[۳] red-green-blue[۴]



شکل ۳-۷ تصویر به تفکیک پیکسل [17]

مشکل در پردازش تصویر : در موضوعاتی مانند پردازش تصویر تشخیص چهره یا هر موضوع خاص دیگری استخراج ویژگی توسط یک الگوریتم مشخص انجام می‌گیرد مانند SIFT[۱] یا PCA[۲] و ... ولی در اینجا ما با تصاویر با فرمت ها و اندازه هایی مختلف مواجه هستیم و از طرفی خروجی ما نیز درصد شباهت تصویر با لیست پرس‌وجوهای مختلف می‌باشد که از متون نزدیک تصاویر در صفحات وب بدست می‌آید. برای استخراج ویژگی و پردازش تصویر و متون [10] در این پروژه ما از شبکه‌های عصبی کانولوشن[۳] و یادگیری عمیق[۴] استفاده کرده‌ایم.

۲-۳-۴-۲-۳ شبکه های عصبی کانولوشن

شبکه های عصبی کانولوشن شبیه شبکه های عصبی مصنوعی معمولی می‌باشند. این نوع شبکه ها متشکل از نورونهایی[۵] با وزنها و بایاسهای[۶] قابل یادگیری هستند. [20]هر نورون تعدادی ورودی دریافت کرده و

Scale-invariant feature transform[۱] Principal component analysis [۲] convolution neural network (CNN)[۳]
deep learning[۴] neurons[۵] bios[۶]

سپس حاصل ضرب وزنها در ورودی ها را محاسبه کرده و در انتها با استفاده از یک تابع تبدیل (فعال سازی[1]) غیرخطی درصد شباهت ورودی با بردار خروجی (لیست پرس‌وجو ها) را تعیین می‌کند. همینطورشبکه شامل یک تابع امتیاز مشتق پذیر [2]را شامل می‌شود. که در یک طرف آن پیکسل‌های خام تصویر ورودی و یا متن نزدیک تصویر و در طرف دیگر آن امتیازات مربوط به هر پرس‌وجو قرار دارد. این نوع شبکه ها هنوز یک تابع هزینه مثل SVM یا Softmax (درلایه آخر) تمام مرتبط[3] دارند و تمامی نکات مطرح در مورد شبکه های عصبی معمولی در اینجا هم صادق است. پس تفاوت این دو شبکه در چیست؟ معماری های شبکه های عصبی کانولوشن بصورت صریح فرض می‌کنند که ورودی های آنها تصاویر هستند , با این فرض ما می‌توانیم ویژگی های مشخصی را درون معماری تعبیه کنیم.با این عمل تابع پیشرو را میتوان بصورت بهینه تر پیاده سازی کرد و همینطور با این کار میزان پارامترهای شبکه نیز بشدت کاهش پیدا میکند . در شکل ۳-۸ یک شبکه کانولوشن تصویرنمایش داده شده است.

شکل ۳-۸ شبکه عصبی CNN تصویر [20]

activator function[1] differentiable[2] fully connected[3]

خلاصه معماری شبکه عصبی کانولوشن :

شبکه های عصبی یک ورودی دریافت می‌کنند (در قالب یک بردار) و سپس آنرا از تعدادی لایه مخفی[1] عبور می‌دهند. و در نهایت یک خروجی که نتیجه پردازش لایه های مخفی است در لایه خروجی نمایش داده می‌شود که نمایش دهنده درصد شباهت ورودی با یکی از پرس‌وجوهای خروجی می‌باشد. هر لایه مخفی از تعدادی نورون تشکیل شده که این نورون ها به تمام نورون‌های لایه قبل از خود متصل می‌شوند . نورونهای هر لایه بصورت مستقل عمل کرده و هیچ ارتباطی با یکدیگر ندارند [20]. آخرین لایه تماماً متصل به لایه خروجی معروف است و معمولا نقش نمایش دهنده امتیاز هر طبقه را ایفا می‌کند.

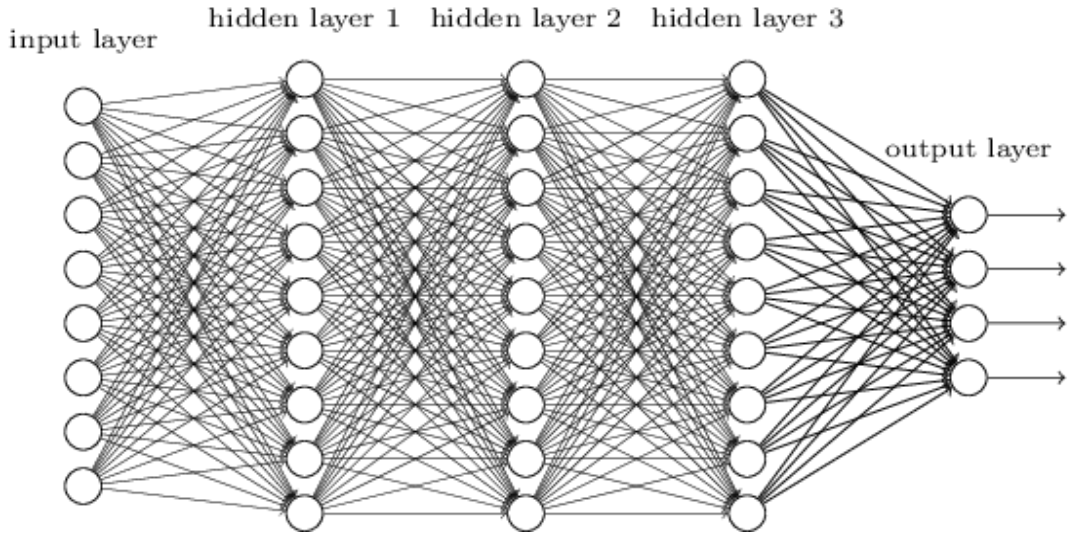
شبکه های عصبی معمولی برای تصاویر بخوبی مقیاس پذیر نیستند. بعنوان مثال فرض کنیم دارای مجموعه ای از تصاویر با اندازه ای 32*32* 3 هستیم (۳۲ پیکسل عرض, ۳۲ پیکسل ارتفاع و ۳ کانال رنگ). بنابراین یک نورون با اتصال کامل[1] در لایه مخفی اول یک شبکه عصبی معمولی به ۳۰۷۲ وزن نیاز خواهد داشت. این مقدار شاید در نظر اول مقدار قابل توجهی بنظر نیاید اما بطور واضح مشخص است که این معماری تماماً متصل قابل استفاده برای تصاویر بزرگتر نخواهد بود مخصوصا در صفحات وب که هر تصویر اندازه متفاوت دارد. برای مثال یک تصویر با اندازه متعارف تر مثل ۲۰۰*۲۰۰*۳ باعث میشود که یک نورون 200*200* ۳= 120,000وزن داشته باشد. علاوه بر این ما قطعا خواهان تعداد بیشتری از این نورون ها خواهیم بود. مشخص است این تعداد اتصال باعث افزایش نمایی پردازش شده و با افزایش تعداد پارامترها باعث ایجاد overfitting خواهد شد [20].

---

[1] hidden layer

توده های سه بعدی از نورونها[1]

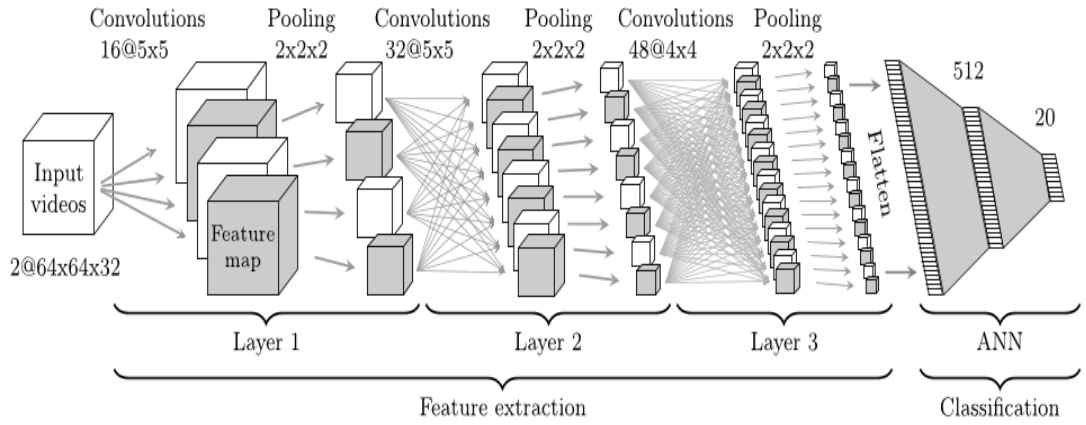
شبکه های عصبی کانولوشن از این واقعیت که ورودی شامل تصاویر است استفاده کرده و معماری شبکه را به روش معقولی محدود کردند. برخلاف یک شبکه عصبی, لایه های یک شبکه عصبی کانولوشن شامل نورونهایی است که در سه بعد عرض, ارتفاع و عمق قرار گرفته اند. کلمه عُمق در اینجا اشاره به بُعد سوم یک توده فعال سازی[2]دارد و به معنای عمق یک شبکه عصبی کامل که به معنای تعداد لایه‌های موجود در آن است نمی‌باشد. هر نورون در هر لایه بجای اتصال با تمام نورون ها در لایه قبل تنها به ناحیه کوچکی از لایه قبل از خود متصل است [20]. با رسیدن به انتهای معماری شبکه ما اندازه تصویر را کاهش می‌دهیم بگونه ای که در انتها تصویر کامل ورودی ما به یک بردار حاوی امتیاز طبقه‌ها کاهش پیدا میکند (۸۰۰ پرس‌وجوی انتخابی) و ما با یک بردار که حاوی امتیاز هر طبقه است مواجه خواهیم بود. شکل ۳-۹ یک شبکه چند لایه و در شکل ۳-۱۰ یک شبکه کانولوشن ۳بعدی را نمایش می‌دهند.



شکل ۳-۹ شبکه عصبی چند لایه [20]

---

[1] 3d volumes of neurons [2] activation volume



شکل ۳-۱۰ شبکه کانولوشن ۳ بعدی [20]

حالا به معرفی لایه‌های شبکه کانولوشن می‌پردازیم.

- **لایه کانولوشن :** در این لایه ما بردار ورودی را در بردار ویژگی ها[۱] کانوال می‌کنیم. در حقیقت در این لایه هر قسمت از لایه قبلی با یک پنجره مشخص کانولوشن شده و نتیجه به لایه بعدی داده می‌شود [20].
- **لایه RELU یا لایه نرمال کننده**[1] : ممکن است در لایه کانولوشن یکسری از داده ها مقادیر نامتعارفی (مثلا منفی) بگیرند. پس بهتر است بردار نرمالایز شود. در این جا برای نرمال از RELU استفاده شده است. مطابق با فرمول ۳-۱ که جهت جلوگیری از کسب مقادیر منفی توسط لایه‌های مختلف [20].

$$X_i = \text{Maximum}(X_i , 0) \qquad \text{(فرمول ۳-۱)}$$

- **لایه Pool[۲] :** در این مرحله ما به کاهش اندازه بردار ورودی می‌پردازیم (مثلا میانگین گیری ۹ پیکسل مجاور هم). این کار باعث دستیابی به ویژگی‌های‌سطح بالا[۳] می‌شود. (در پردازش تصویر منظور از ویژگی های سطح پایین همان شناسایی لبه ها[۴] و شناسایی منحنی ها[۵] و ... است و ویژگی های سطح

---

rectified linear unit layer [۱] pooling layer [۲] high level features[۳] edgs[۴] curve[۵]

بالا مثلا شناسایی دست، بینی و همینطور در متن منظور گروه‌ها و کلمات کلیدی و مهم هستند). یکی دیگر از اهداف این لایه نگاشت بردار ورودی به بردار خروجی و نتیجه می‌باشد (در این پروژه همان ۸۰۰ پرس‌وجوی خروجی).

- **لایه اتصال کامل :** این لایه همان امتیاز دهنده بردار ورودی به بردار خروجی و داخل موتور جستجو می‌باشد و تصویر بر حسب درصد شباهت با پرس‌وجوها در یکی یا چند تا از این پرس‌وجوها قرار می‌گیرد.

۴-۴-۲-۳ استخراج ویژگی از متن

در این پروژه علاوه بر خود تصاویر ما از متون صفحات وب هم برای کمک به طبقه‌بندی و شناسایی اشیاء استفاده می‌کنیم. بخاطر ماهیت بدون ناظر بستر وب. متون نزدیک به تصویر در صفحات وب می‌توانند کمک خوبی برای شناسایی پرس‌وجوها باشند. حال به معرفی چالش‌های استخراج یک لیست پرس‌وجو از داخل متن می‌پردازیم

چالش‌ها

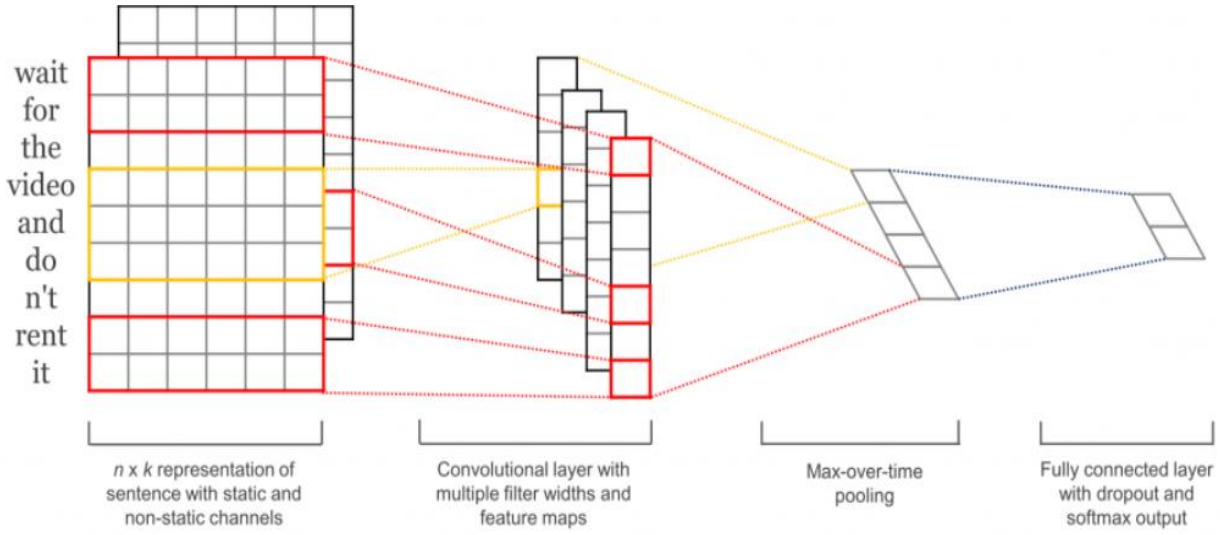
- چگونه تشخیص دهیم متن درون صفحه وب با تصویر داخل صفحه وب مرتبط است.

- چگونگی تشخیص اهمیت و اولویت متن با تصویر : در نزدیک تصویر (در سطوح مختلف درخت html) متونی یافت می‌شود به چه نحوی باید به هر کدام از این متون اولویتی داده شود.
- چگونگی ایجاد بردار ورودی از متن استخراج شده از صفحات وب.

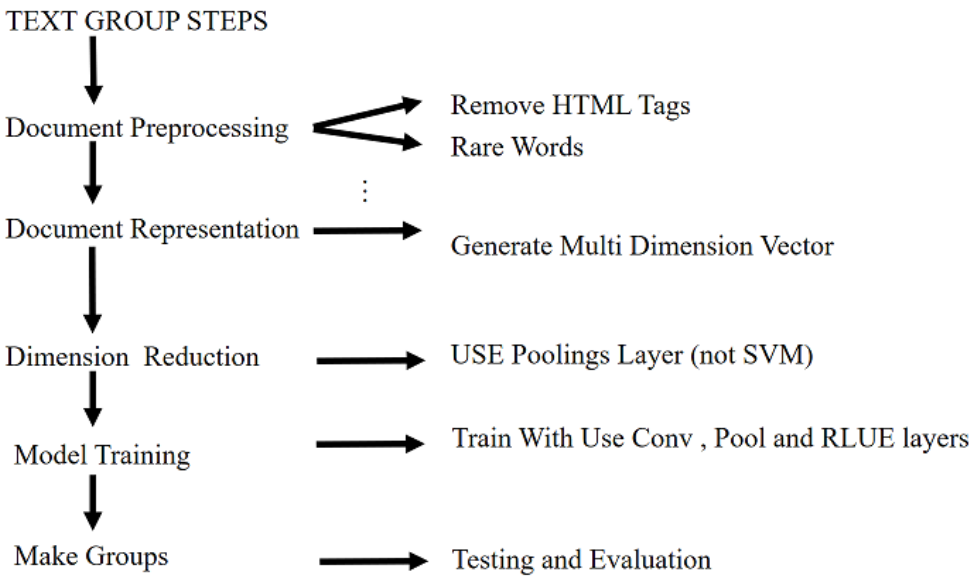
نوآوری

در اینجا ما به استخراج متون نزدیک تصاویر پرداخته و سپس این متون را در یک پاراگراف جمع کرده و پیش پردازش‌های را روی این متن انجام می‌دهیم مانند حذف کلمات تکراری و حذف کاراکترهای ویرایش مانند خط جدید (\n) و... در مرحله بعدی به ایجاد بردار ویژگی از متن پرداخته و این بردارها را به یک شبکه عصبی کانولوشن داده و نتیجه آن یک بردار حاوی ۸۰۰ طبقه مختلف از اشیاء می‌باشد. (در این پروژه ما فقط از ۸۰۰ گروه مختلف استفاده کرده‌ایم) [21]. این فاز جهت ایجاد طبقات اصلی پرس‌وجو در موتور جستجو اجرا می‌شود. در این فاز از ۱۰ دامنه و بیش از ۱۰۰۰۰ صفحه وب استفاده شده است. خروجی این مرحله لیست گروه‌های اولیه مورد جستجو می‌باشد.

سپس با استفاده از شبکه عصبی کانولوشن و یادگیری عمیق ما میزان شباهت هر متن به دست آمده نزدیک تصویر را با تمام پرس‌وجو های داخل موتور جستجو بررسی می‌کنیم. از طرفی به مرور زمان و گرفتن امتیاز هر متن میزان اهمیت مکان متن در یک صفحه وب خاص و درمورد یک پرس‌وجوی خاص مشخص می‌شود. مطابق شکل ۳-۱۱ متن ورودی به یک شبکه عصبی چند لایه کانولوشن به صورت یک بردار داده می‌شود و خروجی آن میزان شباهت متن با هر پرس‌وجو می‌باشد. برای بردار ویژگی ورودی هم بجای تصویر ما از ماتریس متنی استفاده می‌کنیم هر سطر این ماتریس از حروف یک کلمه تشکلیل شده است. بردار خروجی این شبکه نیز به تعداد پرس‌وجوهای پایگاه داده و مقدار شباهت متن ورودی با آنها می‌باشد. در شکل ۳-۱۲ نیز مراحل استخراج پرس‌وجوهای اصلی موتور جستجو نمایش داده شده است.



شکل ۳-۱۱ بردار ورودی شبکه کانولوشن متن [21]

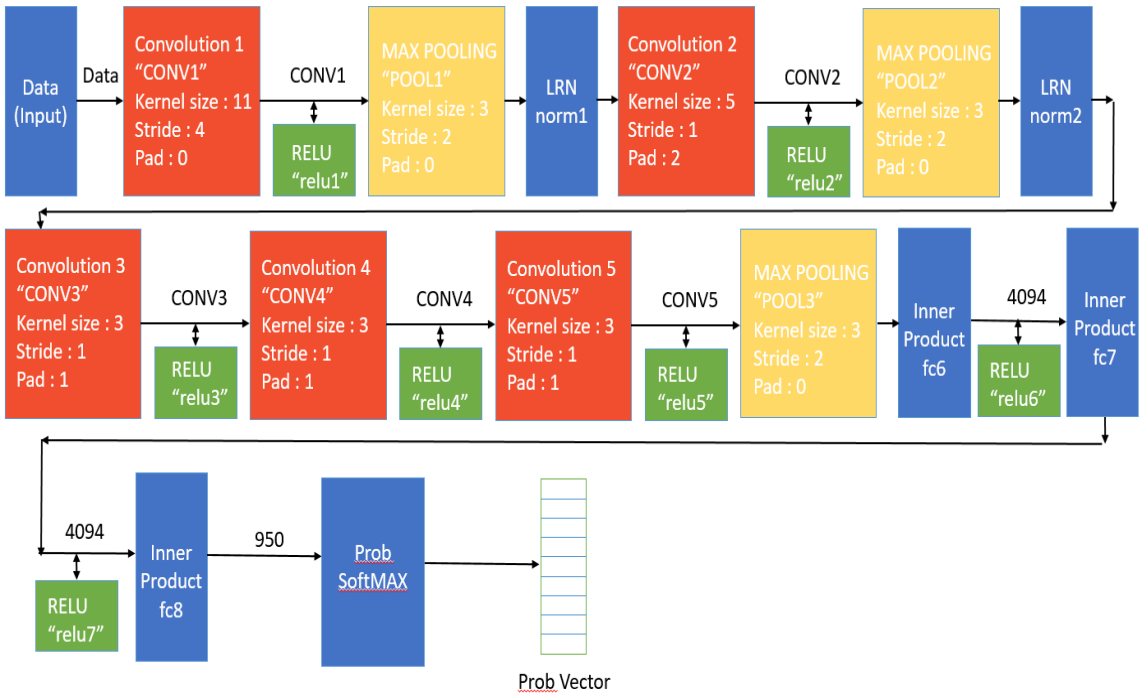


شکل ۳-۱۲ بردار ورودی شبکه کانولوشن متن [21]

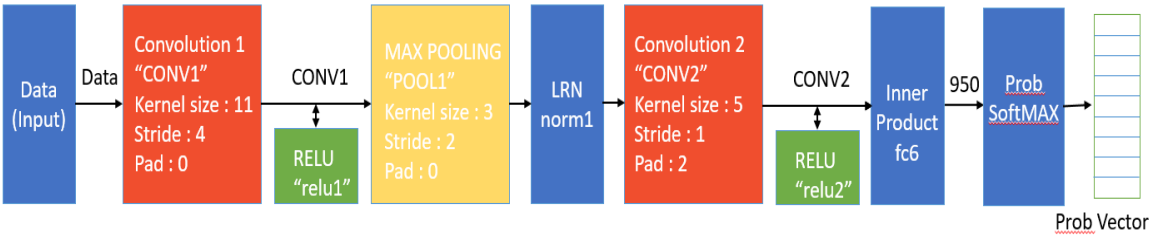
معرفی فریم ورک Caffe

در این قسمت به معرفی یکی از برنامه‌های بسیار عالی‌جهت یادگیری عمیق با نام Caffe (کافی) می‌پردازیم. این برنامه برای سیستم عامل‌های (ویندوز، لینوکس و یونیکس) و زبان‌های برنامه نویسی مختلف (c++، python و ...) ارایه گردیده است. این برنامه توسط آقای Yangqing Jia توسعه داده شد و در حال حاضر بصورت

کاملا رایگان توسط دانشگاه برکلی توسعه داده می‌شود . در این پروژه برای یادگیری تصاویر و متون صفحات وب از دو شبکه عصبی عمیق استفاده کرده‌ایم. برای شبکه عصبی تصویر مطابق با شکل ۳-۱۳ از یک شبکه عصبی بزرگ استفاده شده است که حاوی چندین لایه کانولوشن، Pooling و RELU به همراه روش های مانند softmax می‌باشد. همچنین برای انتخاب و یادگیری متن (ایجاد ۸۰۰ گروه پرس‌وجو) هم از یک شبکه عصبی کوچکتر مطابق با شکل های ۳-۱۴ استفاده گردیده است. برای این برنامه بعد از مشخص کردن مراحل و لایه‌های شبکه عصبی آنها را در فایل های با پسوند prototxt ذخیره می‌کنیم. در این فایل به معرفی لایه های مختلف تعریف شده در شبکه عصبی کانولوشن پرداخته و پارامترهای یادگیری را مشخص می‌کنیم. در شکل ۳-۱۵ یک نمونه از این تعریف را می‌بینید.



شکل ۳-۱۳ بلوک دیاگرام یادگیری تصویر در کافی



شکل ۳-۱۴ بلوک دیاگرام یادگیری متن در کافی

```
name: "CaffeNet"
layer {
  name: "data"
  type: "Input"
  top: "data"
  input_param { shape: {
     dim: 10 dim: 1 dim: 255 dim: 255
  } }
}
layer {
  name: "conv1"
  type: "Convolution"
  bottom: "data"
  top: "conv1"
  convolution_param {
    num_output: 96
    kernel_size: 11
    stride: 4
  }
}
```

شکل ۳-۱۵ نمونه‌ای از فایل prototxt در کافی

همانطور که در تصویر ۳-۱۵ مشخص است. هر لایه در شبکه کانولوشن با لایه های بالای و پایینی خود ارتباط دارد به عنوان مثال در شکل لایه کانولوشن اول ما با نام conv1 از طرفی به لایه ورودی متصل است.

در مرحله بعدی و قبل از شروع یادگیری ما باید مجموعه داده آموزش و آزمون خود را بسازیم که برای این منظور ابتدا تصاویر را از پایگاه داده استخراج کرده (تصاویر دریافت شده در صفحات وب) و برای سادگی یک پیش پردازش مطابق زیر انجام می‌دهیم

- **یکی کردن اندازه تصاویر** :

به دلیل عدم یکی بودن اندازه تصاویر و نوع آنها (ممکن است بعضی تصاویر ۳ کاناله و بعضی تک کاناله

و خاکستری باشند) در این پیش پردازش و به منظور ایجاد ثابت کردن اندازه پنجره ورودی تصاویر را به اندازه ثابت ۲۵۵ *۲۵۵ به منظور ثابت کردن پنجره ورودی تبدیل می‌کنیم. البته می‌توان تصویر را به قطعه های مختلف با اندازه ثابت تبدیل کرد و آنرا به عنوان نتیجه به شبکه یادگیری فرستاد که به علت محدودیت سخت افزاری این کار انجام نگردید.

- **خاکستری کردن تصاویر:**

به منظور ساده سازی پردازش و کاهش اندازه ورودی بجای استفاده از یک تصویر با ۳ کانال آنرا به یک کانال خاکستری تبدیل کردیم (البته ممکن است تصاویر ۳کاناله و خاکستری در صفحات وب وجود داشته باشند که به منظور یکی کردن غالب تصاویر همه را به فرمت خاکستری تبدیل کردیم).

- **تغییر همه تصاویر به یک فرمت :**

در بستر وب تصاویر با فرمت های مختلف وجود دارند (فرمت های مانند PNG، JPG، WEBM و...) به این منظور تصاویر را به یک فرمت مشخص که در اینجا فرمت انتخابی ما JPEG است به دلیل فشرده سازی خوب و بالای تصویر و عدم حذف زیاد جزییات از داخل آن انتخاب کرده‌ایم.

- **حذف تصاویر تکراری :**

در این بخش تصاویر تکراری حذف می‌گردند. جهت حذف به راحتی می‌توان تصاویری را که بیش از ۱۰ بار در پایگاه داده تکرار شده اند از داخل آن حذف کرد. دلیل این امر این است که به احتمال بالا این تصاویر یکی از موارد زیر می‌باشند.

- یک تصویر عمومی در صفحه وب می‌باشد مانند لوگوی یک دامنه خاص که با نرخ تکرار بالای در صفحات وب قرارمی‌گردد.
- تصویر تبلیغاتی می‌باشد که در صفحات زیادی و به تکرار دیده می‌شود.

بعد از این پیش پردازش در گام بعدی ما متن ها و تصاویر به همراه فایل آموزش به برنامه کافی جهت آموزش می‌دهیم. در این پروژه تعداد ۱۰۰۰۰ تکرار را برای این شبکه عصبی به همراه ۸۰۰ گروه مشخص تعیین کرده ایم. و بعد از هر ۲۰۰ تکرار یک مرحله تست را هم داریم. مشخصات کامل آموزش در جدول ۳-۳ قابل مشاهده است. در انتها و بعد از فاز آموزش برنامه کافی یک فایل حاوی تمام وزنها (یادگیری شده) و لایه های مختلف شبکه عصبی کانولوشن را با پسوند caffeprototype به عنوان خروجی می‌دهد. که برای تست تصاویر ورودی و متن ها در مراحل بعدی استفاده می‌شود. به این صورت که از این مرحله به بعد تصاویر و متون بدست آمده از صفحات وب به این شبکه داده شده و خروجی میزان شباهت متن و یا تصویر ورودی به پرس‌وجوهای صفحه وب را نمایش می‌دهد (بردار خروجی).

جدول ۳-۳ توضیحات فاز آموزشی

ردیف	نوع	توضیحات
۱	مشخصات سخت افزار آموزش	• 2 عدد کارت گرافیکی مدل Geforce 840M با حافظه ۸ گیگا بایت (تکنیگ SLI) و ۸۰۰ واحد کُوداَ • یک عدد پردازنده core i5 4600U با فرکانس کاری GHZ۱۸۰۰ و ۴ هسته • ۱۲ گیگا بایت حافظه مدل DDR3
۲	تعداد تکرار فاز آموزشی	۱۰۰۰۰ بار تکرار
۳	تعداد گروه مشخص شده اولیه	۸۰۰ گروه مشخص گردیدکه به عنوان پرس‌وجوهای موتور جستجو مشخص شده‌اند
۴	تعداد تصاویر	۵۰۰۰۰ تصویر از بیش از ۱۹۲۰ دامنه و ۲۳۳۵۰ صفحه وب متفاوت
۵	فاز تست بعد از چند آموزش	۲۰۰
۶	زمان آموزش	زمان تکرار ۱۰۰۰۰ تکرار در این پروژه حدوداً چهل روز به طول انجامید

۵-۲-۳ صف اولویت آدرس‌ها

همانطور که قبلا هم گفته شد موتور جستجو جهت دریافت هر صفحه وب باید درخواستی به خدمات دهنده آن صفحه بفرستد متن درخواست هم باید آدرس صفحه مورد بررسی باشد. در زیر لیست مشکلات و راه حل‌های این قسمت را بررسی می‌کنیم.

- جمع آوری آدرس صفحات وب : به چه صورت آدرس یک صفحه وب را بدست آورد و متوجه شد آیا این آدرس معتبر است. (حاوی اطلاعات است یا ارجاع به یک صفحه تبلیغاتی یا حتی ویروسی باشد)

راه حل : برای این منظور ابتدا لیستی از دامنه‌های مشخص را به موتور جستجو نمایش می‌دهیم (در حافظه آن ذخیره می‌کنیم). موتور جستجو بعد از استخراج آدرس از درون این صفحات و بررسی اطلاعات آنها توسط قسمت تجزیه کننده خودش به استخراج آدرس‌های جدید می‌پردازد. برای تشخیص عدم مشکل و تبلیغاتی بودن یک صفحه وب هم می‌شود به بررسی سابقه دامنه‌ای که آدرس از آن آمده است پرداخت به این صورت که اگر صفحات داخل این دامنه فاقد محتوای مفید درباره پرس‌وجوهای مورد بررسی موتور جستجو باشند این صفحات از لیست صفحات وب حذف شوند و به همین صورت ارزش دامنه این صفحات و صفحاتی از دامنه‌های دیگر که به این دامنه ارجاع داده‌اند (به این دامنه و محتوای داخلش لینک پیوندی داده‌اند) هم کاهش می‌یابد. در زیر چند مزیت کسب آدرس‌های معتبر را بررسی می‌کنیم [5].

- جلوگیری از صفحات تبلیغاتی : بعضی صفحات وب فقط جنبه تبلیغاتی دارند و هیچگونه اطلاعات مفیدی برای موتور جستجو ندارند و از طرفی به علت ورود اطلاعات نا صحیح به زیان موتور جستجو هستند به دلیل افزایش پردازش داده‌های که فاقد محتوای مفید هستند.
- کاهش فضای جستجو : با حذف صفحات وب تکراری بار پردازشی موتور جستجو کاهش می‌یابد.

- چگونگی تشخیص اولویت یک آدرس : به دلیل رشد نمایی آدرس صفحات وب، چه راهکاری برای تشخیص اهمیت آدرس صفحه وب در مورد یک پرس‌وجوی خاص در موتور جستجو است.

راه حل : سیستم رتبه بندی مهمترین قسمت یک موتور جستجو است و تفاوت اصلی در موتورهای جستجو در همین بخش نهفته است. سیستم رتبه بندی برای پاسخگویی به کاربران پارامترهای زیادی را در نظر می‌گیرد تا بهترین پاسخ را ارائه کند. روش هایی رتبه بندی صفحات وب به دو نوع زیر تقسیم می‌شوند [1].

- مبتنی بر ساختار[۱] : مانند الگوریتم هایpage ranking algorithm، الگوریتم کلونی
- مبتنی بر محتوا[۲].

در روش های مبتنی بر ساختار به ساختار و مقادیر تگ های یک صفحه وب امتیاز می‌دهیم. به عنوان مثال اگر یک صفحه وب دارای تگ‌های تصویر زیادی است می‌شود این احتمال است که این صفحه وب دارای اطلاعات مفیدی درباره تصویر باشد (یا حتی شامل اطلاعات غیر مفید و تبلیغاتی می‌باشد). در روش‌هایی مبتنی بر محتوا بیشتر تمرکز روی محتوا و صحت داده ها می‌باشد. در این پروژه ما از ترکیب این دو روش استفاده کرده‌ایم که رتبه صفحه وب ترکیبی از غالب صفحه وب و متن و تصاویر مشابه با پرس‌وجو های داخل موتور جستجو بدست می‌آید در بخش‌های آیند به طور کامل در این مورد بحث می‌کنیم.

۱-۵-۲-۳ روش آقای ساجین گوپتا[۳]

---

structure-based [۱] content-based [۲] sachin gupta[۳]

آقای گوپتا با استفاده از فرمول ۳-۲ به امتیاز دهی به آدرس‌های وب توسط موتور جستجو پرداختند[22] .

$$PR(u) = (1 - d) + d * \sum_{v \in B(u)} (PR(v) * W^{in}(v, u) * W^{out}(v, u)) \quad \text{(فرمول ۳- ۲)}$$

حال به معرفی پارامترهای این فرمول می‌پردازیم.

- W^{in} : لینک های ورودی به صفحه‌وب u (تعداد تمام صفحاتی که به صفحه u ارجاع دارند و از داخل آن صفحات می‌شود وارد این صفحه شد) و تمام لینک‌های ورودی ارجاع داده شده از لینک خروجی v (تمام لینک‌های که صفحه v به صفحه u یا پدران و فرزندان صفحه u متصل شده‌اند).
- W^{out} : لینک های خروجی از صفحه u و تمام لینک‌های خروجی از همه صفحات ارجاع داده شده از لینک ورودی v.
- d یا فاکتور میرایی[۱] : که یک ثابت در این فرمول می‌باشد و آقای گوپتا پیشنهاد داده‌اند از عدد ۰,۸۵

در آن استفاده شود. البته در مقاله اشاره به آموزش و تغییر این فاکتور به مرور زمان شده است.

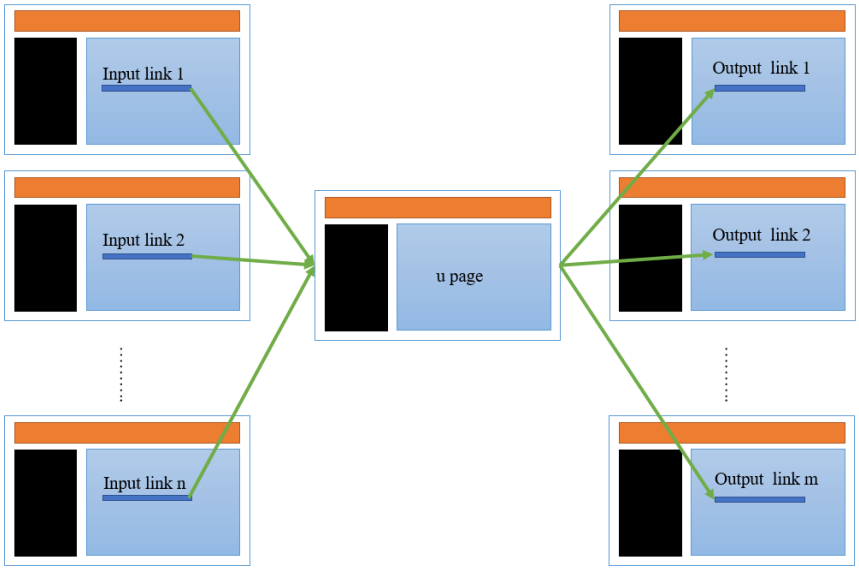
- PR[۲] : تابع محاسبه امتیاز جهت محاسبه امتیاز لینک u استفاده می‌شود.
- صفحات وب : در اینجا u همان لینک صفحه وب جاری است و v مجموع لینک‌های هستند که به نحوی به صفحه u متصل هستند مثلا یا از صفحه u به آنها ارجاع داریم که به آن لینک ورودی گویند یا صفحه v به صفحه u یک ارجاع داده است که به آن لینک خروجی گویند.
- تابع B(u) : این تابع از مجموعه لینک‌های موجود لینک‌های را انتخاب می‌کند که یا به صفحه u ارجاع داده‌اند یا صفحه u به آنها ارجاع داده است. در حقیقت ما باید لیست لینک‌های ورودی و خروجی هر صفحه وب را بیابیم. جهت انجام این کار ما باید تمام صفحات وب و لینک های متصل

---

۱ damping factor ۲ process ranking

به آنها را بیابیم و سپس به محاسبه مقادیر این اتصالات بپردازیم. به این منظور از یک جدولی به نام جدول رابطه[1] در پایگاه داده (در فصل چهارم به طور کامل درباره پایگاه داده[2] استفاده شده در موتور جستجو صحبت می‌کنیم) استفاده کرده‌ایم که وظیفه آن ذخیره لیست ارجاعات به صفحات وب می‌باشد.

در شکل ۳-۱۶ به مفاهیم ارجاع ورودی و خروجی صفحات وب اشاره می‌کنیم.



شکل ۳-۱۶ لینک‌های ورودی و خروجی صفحه وب[1]

۳-۲-۵-۲ روش موتور جستجوی گوگل

البته این موتور جستجوی گوگل به مرور زمان امتیاز دهی خود را به صفحات وب را تغییر می‌دهد. ولی یکی از روش‌هایی امتیاز دهی به صفحات وب این موتور جستجو مطابق فرمول ۳-۳ می‌باشد [1].

$$PR(u) = (1-d) + d\sum_{i=1}^{n} \frac{PR(T^i)}{C(T^i)}$$ (فرمول ۳-۳)

---

[1] relation table

[2] data base

حال به معرفی پارامترهای این فرمول می‌پردازیم.

- d یا فاکتور میرایی: که یک ثابت در این فرمول می‌باشد و از عدد ۰,۸۵ در آن استفاده شده است.
- T^i : مجموعه صفحاتی که به صفحه u ارجاع داده اند.
- $C(T^i)$: تعداد لینک های که از صفحه T^i خارج شده اند (منظور تعداد صفحات وبی است که صفحه T^i به آنها ارجاع داده است).
- PR : جهت محاسبه امتیاز لینک از این تابع استفاده می‌شود.
- تابع B(u) : این تابع از مجموعه لینک های موجود لینک های را انتخاب می‌کند که یا به صفحه u لینک ورودی دارند یا لینک خروجی.

یکی از مزایای روش موتور جستجوی گوگل تغییر اولویت یک صفحه وب با افزایش اهمیت صفحات متصل به آن می‌باشد. و از معایب این روش هم می‌شود به این مورد اشاره کرد که بین صفحاتی که واقعاً اطلاعات خوب و مهمی دارند و صفحاتی که فاقد اطلاعات با ارزشند و فقط به منظور افزایش بازید از یکسری کلمه کلیدی استفاده کرده اند تفاوتی نمی‌گذارد.

۳-۵-۲-۳ روش page ranking developed algorithm

این روش گونه توسعه یافته روش PageRank است و نحوه امتیاز دهی به صفحات وب در این روش به فرم زیر است [5].

- اهمیت دادن به لینک های ورودی و خروجی صفحه وب جاری
- امتیار رتبه بندی مبتنی بر محبوبیت صفحات توزیع شده. (این محبوبیت یا از طریق بازدید کاربران و کلیک روی لینک آن در نتایج صفحاتوب می‌باشد یا به فرم افزایش امتیاز صفحه در موتور جستجو

درباره یک پرس‌وجوی خاص) روشهای مختلفی برای پیاده سازی این الگوریتم وجود دارد مثل شبکه عصبی و مارکوف.

نوآوری : در این پروژه برای هر دامنه یک امتیاز در نظر گرفته می‌شود. در ابتدا که صفحات وب فاقد امتیاز هستند طبق الگوریتم page ranking می‌باشد ولی به مرور زمان و بازدید از صفحات امتیاز صفحات با توجه به نوع پرس‌وجوی پرسیده شده از آن تغییر می‌کند. هر صفحه امتیازی از یک پرس‌وجوی خاص می‌گیرد (همان پرس‌وجوهای داخل موتور جستجو). پس برای هر آدرس یک جدول از امتیازات پرس‌وجو را ذخیره می‌کنیم. در مورد این روش به فرم کامل در فصل چهرم صحبت می‌کنیم.

۳-۳ مدیریت پایگاه داده

منظور از این بخش اطلاعاتیست که بعد از دریافت توسط موتور جستجو جهت استفاده‌های بعدی در موتور جستجو ذخیره می‌شوند. حذف اطلاعات و ویژگی های تکراری یا بی ارزش در پایگاه داده جهت افزودن اطلاعات جدید. به علت محدود بودن پایگاه داده قسمت مدیریت پایگاه داده باید اطلاعات تکراری یا فاقد ارزش را حذف کند. در موتور جستجو باید لیستی از سوالات و پرس‌وجو های پرسیده شده از موتور جستجو ذخیره شوند.

چالش : چه روشی برای اندیس گذاری نتایج باید استفاده شود. در حالت خلاصه منظور از اندیس کردن نتایج نوعی نگاشت نتایج به پرس‌وجوی موجود در موتور جستجو می‌باشد. و این اندیس ها چگونه در پایگاه داده ذخیره شوند.

نوآوری :

باید نتایج بدست آمده در صفحه وب به نحوی اندیس گذاری شوند که با صرف کمترین هزینه و وقت

نتایج آماده نمایش شوند. از طرفی یک نتیجه ممکن است با چندین موضوع مرتبط باشد به عنوان مثال فرض کنید یک تصویر با موضوعیت برج در پایگاه داده داریم و دو پرس‌وجوی "نمایه زیبا از **برج** میلاد" و "**برج** آزادی" از موتور جستجو پرسیده شود. پس در هر دو این پرس‌وجو ها باید در لیست نتایج نمایش داده شود. این کار بعد از مراحل یادگیری موتور جستجو و یا انتخاب کاربران بازدید کننده از سایت تغییر می‌کند. در این جا با عبور پرس‌وجوها از شبکه عصبی موتور جستجو درصد شباهت پرس‌وجوی جدید با لیست پرس‌وجوهای داخل موتور جستجو مشخص می‌شود. باید برای هر پرس‌وجو یک اندیس داشته باشیم. و از

طرفی بعد از مشخص شدن اهمیت پرس‌وجو در یک صفحه وب به دادن امتیاز به تصاویر و متون نزدیک تصاویر در صفحه وب می‌پردازیم.

۴-۳ مدیریت پرس‌وجو

ایجاد و طبقه بندی پرس‌وجو ها در یک موتور جستجو از اهمیت بالایی برخوردار می‌باشد. ممکن است کاربر سوالی را از موتور جستجو بپرسد که در لیست طبقات پرس‌وجوی موتور جستجو وجود نداشته باشد. ولی متن پرس‌وجوی جدید با یک یا چندتا از پرس‌وجوهای ذخیره شده در موتور جستجو شباهت داشته باشد. در این مرحله ما ابتدا پرس‌وجوی جدید را بعد از مرحله پیش پردازش و حذف کلمات تکراری و علامت‌های ویرایشی به یک بردار تبدیل کرده و به شبکه عصبی موتور جستجو می‌دهیم و خروجی این قسمت میزان شباهت پرس‌وجوی جدید با پرس‌وجوی جدید با پرس‌وجوهای داخل موتور جستجو است.

کاربردها

- بررسی مشکلات نحوی پرس‌وجو:

ممکن است شما در پرس‌وجوی خود کلمه ای را به اشتباه وارد کرده باشید مثلا در جمله show best dug کلمه dog را به اشتباه dug وارد کرده‌اید. که وظیفه موتور جستجو تشخیص این اشکالات نحوی، لغوی و ارسال مجدد صحیح پرس‌وجو می‌باشد. که به راحتی در شبکه عصبی میزان شباهت dug با dog مشخص می‌گردد و می‌شود نتیجه را به عنوان کلمه صحیح به کاربر نمایش داد.

- نمایش لیست پرس‌وجو های مشابه و پر بازدید :

هدف از این قسمت نمایش لیست چند پرس‌وجوی مشابه با پرس‌وجوی خود که نتایج خوبی داشتند است.

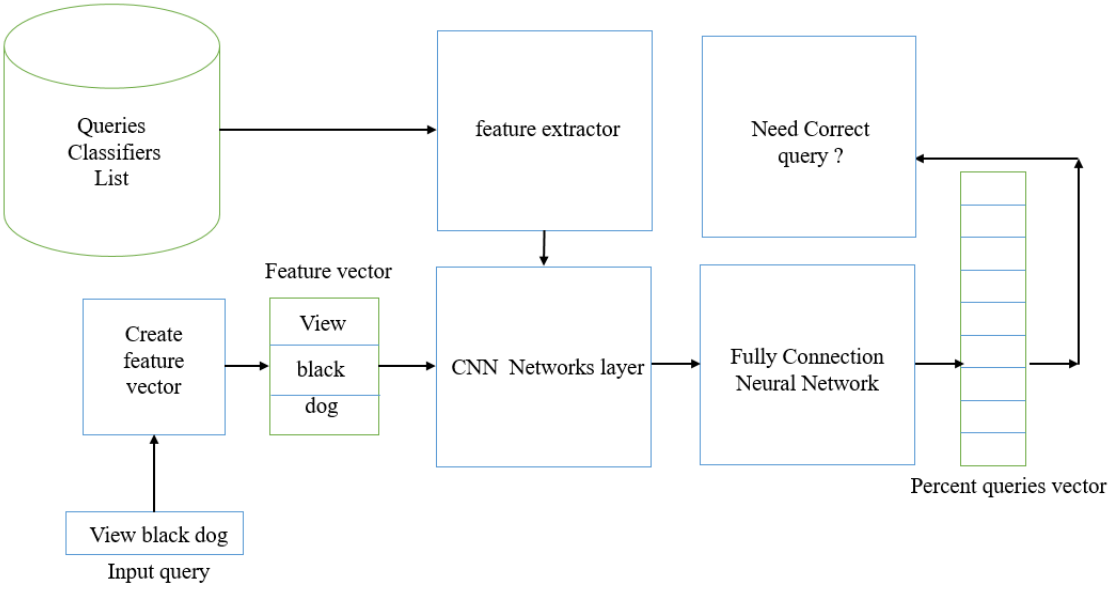
- چگونگی بررسی شباهت سوال پرسیده شده از موتور جستجو با پرس‌وجوهای موجود در هر طبقه و دادن یک

امتیاز به عنوان امتیاز شباهت به آن و یا حتی ایجاد یک طبقه بند جدید.

نوآوری :

در این روش و با استفاده از تکنیک یادگیری عمیق به بررسی شباهت متن پرس‌وجوی داده شده به موتور جستجو با پرس‌وجوهای ذخیره شده در موتور جستجو پرداخته و در نهایت خروجی این شبکه عصبی لیست

احتمال شباهت متن پرس‌وجوی ورودی با طبقات مختلف پرس‌وجو می‌باشد که در نهایت پرس‌وجو یا به یک یا چند طبقه خاص بر حسب احتمال شباهت متصل می‌گردد و یا یک طبقه جدید ایجاد می‌گیرد. همانطور که در شکل ۳-۱۷ مشاهده می‌شود.



شکل ۳-۱۷ شبکه CNN برای محاسبه طبقه بندی پرس‌وجوی جدید [21].

همانطور که در شکل ۳-۱۷ مشخص شده. ابتدا ما بردار ویژگی پرس‌وجوی ورودی را به دست آورده و آنرا به شبکه عصبی چند لایه کانولوشن می‌دهیم و سپس نتیجه بدست آمده به شبکه عصبی جهت طبقه بندی پرس‌وجو

داده می‌شود که در نهایت یک بردار حاوی احتمال شباهت پرس‌وجوی ما از سایر پرس‌وجو ها بدست می‌آید. اندازه این بردار به اندازه طبقات پرس‌وجوی موجود در پایگاه داده می‌باشد. در صورت شباهت پرس‌وجوی جدید با یکی یا چند نتیجه در بردار ویژگی این پرس‌وجو به آنها اضافه می‌شود و اگر پرس‌وجوی امتیاز خوبی برای شباهت با این پرس‌وجو را نداشته باشد یک طبقه جدید ایجاد می‌گردد و سر انجام بردارهای ویژگی هم برای مراحل بعدی به روز رسانی می‌گردند. در مرحله آخر هم می‌شود با نتایج بدست آمده (بردار احتمال خروجی شباهت متن با پرس‌وجو های داخل موتور جستجو) و درصد شباهت پرس‌وجو با لیست نتایج به شباهت پرس‌وجوی ورودی با پرس‌وجوهای موجود رسید و همینطور به بررسی مشکلات نحوی و لغوی پرس‌وجو پرداخت البته می‌شود قبل از ایجاد بردار ویژگی به بررسی مشکلات لغوی هم پرداخت (انجام یک پیش پردازش).

فصل چهارم:

معرفی الگوریتم موتور جستجوی

وب

۴-۱ مقدمه

در فصل قبل ما درباره چالش ها و راه حل های پیش روی یک موتور جستجو بحث کردیم حالا و در ادامه به بررسی مراحل پیاده سازی موتور جستجوی با محوریت تصویر به صورت مرحله به مرحله پرداخته و هر قسمت را معرفی می‌کنیم.

۴-۲ دریافت لیست صفحات وب

در این مرحله مدیریت صف آدرس درخواستی به پایگاه داده فرستاده و لیست تمام آدرس های وب را از آن می‌گیرد. برای قرار دادن آدرس ها در لیست فوق باید ابتدا اولویت هر آدرس مشخص گردد. بعضی از این آدرس ها جدید بوده و تا بحال پردازش نشده اند و فاقد اولویت هستند پس چگونه اولویت این صفحات را مشخص کنیم. در صورتی که صفحه وب از قبل در پایگاه داده مقدار اولویت نداشته باشد (به تازگی این آدرس بدست آمده است) باید برای بدست آوردن اولویت آن مراحل زیر را انجام داد. فرض کنید صفحه جاری که فاقد امتیاز است با نام unPriorityP شناخته می‌شود.

- اگر آدرس unPriorityP جز یک دامنه خاص باشد که در پایگاه داده موجود است می‌توان لیست کلیه صفحاتی که با این صفحه ارتباط دارند را یافت و میانگین این امتیازات را به عنوان امتیاز صفحه unPriorityP قرارداد [1].

در صورتی که آدرس unPriorityP جز یک دامنه خاص در پایگاه داده نیست ابتدا کلیه صفحاتی که دارای اولویت هستند را گروهبندی می‌کنیم. گروهبندی به این صورت می‌باشد که کلیه صفحاتی که متعلق به یک دامنه‌اند را در یک گروه قرار می‌دهیم به عنوان مثال دو صفحه www.hardwarenews.com/cpu ،www.hardwarenews.com/ram هر دو به یک دامنه اصلی به نام www.hardwarenews.com اشاره دارند پس هر دو در یک گروه قرار می‌گیرند مانند شکل ۴-۱.

- حال در هر گروه به دنبال صفحاتی می‌گردیم که با این صفحه ارتباط دارند یعنی به این صفحه ارجاع داده‌اند. سپس برای هر صفحه در گروه به محاسبه امتیازش می‌پردازیم که به فرم جمع کردن تمام اتصالات به همراه وزنشان می‌باشد.

در مرحله آخر با میانگین گیری امتیاز هر صفحه در یک گروه خاص امتیاز آن گروه خاص به صفحه وب مشخص می‌گردد. مطابق فرمول ۴-۱.

$$PR(group_i) = (\sum_{P \in group_i} ScoreLink(P) * P(priority)) / count(group_i) \quad \text{(فرمول ۴-۱)}$$

شکل ۴-۱ گروه بندی دامنه ها

حال به معرفی پارامترهای فرمول می‌پردازیم :

- PR : تابع محاسبه امتیاز یک گروه خاص ($group_i$) به صفحه بدون امتیاز می‌باشد.
- P : صفحه که مطعلق به یک گروه خاص می‌باشند. (دامنه یکسانی با سر گروه دارد)
- ScoreLink : وطیفه این تابع محاسبه امتیاز ارتباطی بین صفحه P با صفحه unPriorityP می‌باشد. یعنی محاسبه تعداد لینک ارتباطی صفحه P و صفحه unPriorityP . در صورت عدم ارتباط این دو صفحه نتیجه این تابع صفر می‌باشد.
- priority : همان مقدار ارزش صفحه P است.
- count : این تابع وظیفه محاسبه تعداد صفحاتی است که در یک دامنه‌اند و با صفحه unPriorityP ارتباط دارند.

- حال که امتیاز هر گروه به صفحه unPriorityP مشخص گردید. باید با یک روش به محاسبه امتیاز اولیه صفحه وب unPriorityP بپردازیم.
  - روش میانگین گیری : در این روش اهمیت و اولویت صفحه با میانگین گیری از صفحات مرتبط به این صفحه بدست می‌آید (البته وزن اتصال هم در این مورد اهمیت دارد) [1].
  - روش تصادفی[1] تکی : در این روش یکی از دامنه ها را به تصادف انتخاب کرده و ارزش آنرا به این صفحه می‌دهیم. البته می‌شود از روش هایی مانند چرخ رولت[2] هم استفاده کرد (اهمیت هر دامنه بر حسب امتیاز ارتباطش با صفحه unPriorityP می‌باشد).

روش تصادفی چند گانه : این روش تغریبا شبیه روش تصادفی تکی است با این تفاوت که ابتدا ما n گروه از دامنه ها را به فرم تصادفی انتخاب می‌کنیم و با میانگین‌گیری از آنها نتیجه را به‌عنوان امتیاز به صفحه می‌دهیم. در این پروژه از روش تصادفی چندگانه استفاده می‌کنیم.

---

[1] stochastic method

[2] roulette wheel

حالا تمام صفحات امتیاز دارند و به ترتیب اولویت به صف آدرس اضافه می‌گردند.

حال که ارزش دامنه همه صفحات مشخص شد بهتر است به روش پیشنهادی که برای دادن امتیاز به یک دامنه خاص در این پروژه استفاده شده است نیز اشاره داشته باشیم.

۴-۲-۱ محاسبه امتیاز دامنه

حالا به محاسبه امتیاز یک دامنه خاص و درباره یک موضوع پرس‌وجوی خاص می‌پردازیم. در این قسمت با استفاده از یادگیری ماشین و رگرسیون به آموزش و محاسبه ارزش یک دامنه می‌پردازیم.

در این روش ابتدا به معرفی چند پارامتر ورودی می‌پردازیم.

- domain[i,j] : منظور محاسبه امتیاز دامنه i ام و پرس‌وجوی j ام است.
- imagesquery[i,j] : این پارامتر تعداد تصاویر دیده شده در دامنه i ام که با پرس‌وجوی j ام شباهت دارد را نمایش می‌دهد.
- seeimages[i] : این پارامتر تعداد تصاویر دیده شده در دامنه i ام را نمایش می‌دهد.
- textquery[i,j] : این پارامتر تعداد متون دیده شده در دامنه i ام که با پرس‌وجوی j ام شباهت دارد را نمایش می‌دهد.
- seetexts[i] : این پارامتر تعداد متون دیده شده در دامنه i ام و نزدیک تصاویر را نمایش می‌دهد.
- successpages[i] : این پارامتر تعداد صفحات وبی را در دامنه i ام که به طور صحیح از خدمات دهنده آن دریافت شده است را نمایش می‌دهد.
- allpages[i] : این پارامتر تعداد صفحات وبی را در دامنه i ام قرار دارد را نمایش می‌دهد.

حال با استفاده از این پارامترها به ایجاد ۳ پارامتر ورودی مطابق پارامترهای فرمول ۴-۲ می‌پردازیم.

$$X_0 = 1 \quad X_1 = \frac{\text{imagesquery[i , j]}}{\text{seeimages[i]}} \quad X_2 = \frac{\text{textquery[i , j]}}{\text{seetexts[i]}} \quad X_3 = \frac{\text{successpages[i , j]}}{\text{allpages[i]}}$$ (فرمول ۴-۲)

در مرحله بعدی و با استفاده از وزن‌های ذخیره شده در پایگاه داده درباره دامنه iام و پرس‌وجوی jام به محاسبه امتیاز دامنه می‌پردازیم. مطابق با فرمول ۴-۳.

$$\text{domain[i,j]} = X_0 * W_0 + X_1 * W_1 + X_2 * W_2 + X_3 * W_3$$ (فرمول ۴-۳)

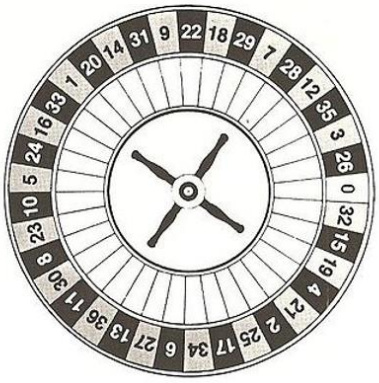
حال به محاسبه احتمال لیست دامنه ها می‌پردازیم در اینجا با استفاده از روش softmax به محاسبه این احتمال مطابق فرمول ۴-۴ می‌پردازیم. در اینجا ما فرض کرده‌ایم تعداد دامنه ها D تا است.

$$\text{domainrank[i,j]} = \frac{e^{domain[i.j]}}{\sum_{d=1}^{D} e^{domain[d.j]}}$$ (فرمول ۴-۴)

در فرمول ۴-۴ منظور از domainrank[i,j] محاسبه احتمال انتخاب دامنه iام درباره پرس‌وجوی jام می‌باشد. حالا با استفاده از چرخ رولِت به انتخاب یکی از این دامنه ها می‌پردازیم مطابق با شکل ۴-۲ [4].

$$\text{Probs} = [\ \text{domainrank}[1, j]\ ,\ \text{domainrank}[2\ , j]\ ,\ \dots\ ,\ \text{domainrank}[D\ , j]\ ]$$

roulette wheel



شکل ۴-۲ انتخاب دامنه درمورد یک پرسوجوی خاص[4]

بعد از انتخاب دامنه و صفحه وب به دریافت اطلاعات می‌پردازیم. و در نهایت بعد از دریافت اطلاعات به محاسبه نتیجه صفحه وب و به روز رسانی وزن ها و مقدار امتیاز دامنه در یک پرس‌وجوی خاص می‌پردازیم. در این پروژه وزنها و مقدار امتیاز قبلی در پایگاه داده ذخیره می‌گردد.

۳-۴ ماژول دریافت کننده صفحه وب

بعد از پر شدن صف آدرس و مقدار دهی اولویت به کلیه صفحات باید ماژول دریافت کننده را آگاه کرد تا اولین آدرس را از صف اولویت دریافت کرده و با ارسال درخواستی به خدمات دهنده آدرس به دریافت صفحه وب بپردازد. بعد از دریافت صفحه توسط ماژول دریافت کننده و قبل از ارسال به مرحله بعد یکسری پیش پردازش مطابق مراحل زیر صورت می‌گیرد.

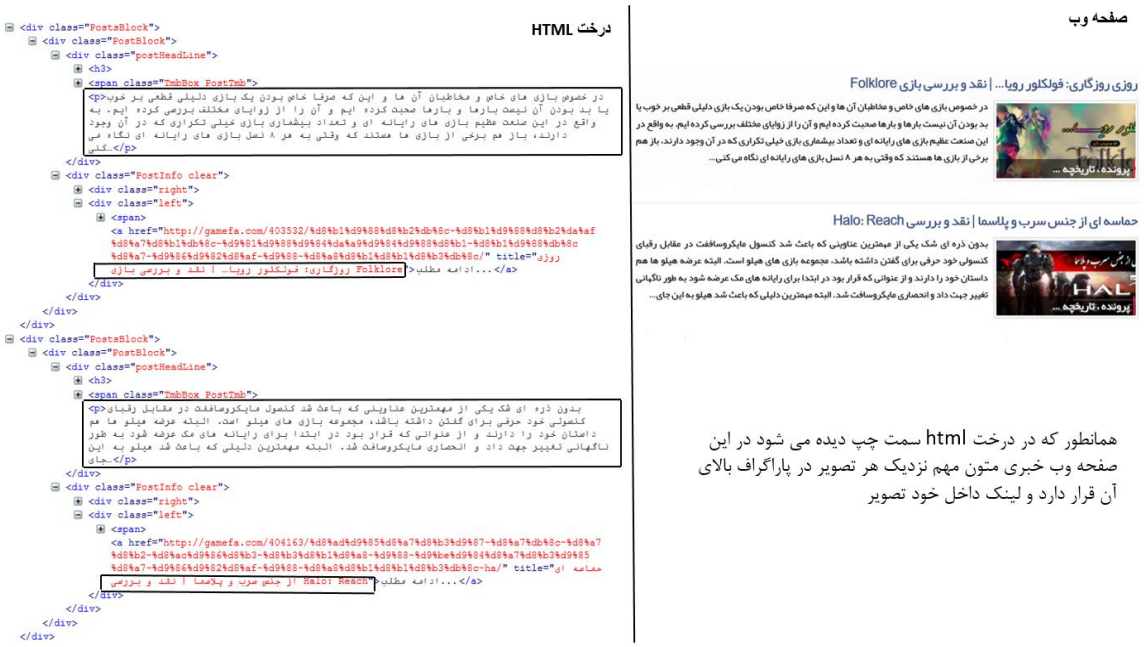
- **بررسی صحت و درستی صفحه وب** : همانطور که در فصل قبل توضیح داده شد آیا صفحه وب دریافتی صحیح است و به طور کامل از طرف خدمات دهنده[1] به موتور جستجو ارسال گردیده است. در صورت بروز خطا یکسری کد مخصوص ایجاد می‌شود که موتور جستجو متوجه صحت و درستی نتایج یا خرابی آنها می‌گردد و این نتایج در امتیاز صفحه وب و دامنه آن تاثیر گذار می‌باشد.
- **تغییر امتیاز صفحه وب** : در صورت اشکال در صفحه وب می‌شود امتیاز صفحه وب (و حتی دامنه آن) را کاهش داد. (دلیل عدم حذف آنی صفحه بجای کاهش امتیاز این است که ممکن است صفحه وب موقتا غیر فعال شده باشد و در آینده دوباره روی خدمات دهنده قرار بگیرد). البته در اینجا کلیه آدرسها در پایگاه داده هم که با این صفحه ارتباط لینکی دارند با کاهش امتیاز مواجه می‌شوند.
- **بررسی قالب صحیح صفحه وب** : آیا قالب صفحه با استاندارد صفحات وب HTML سازگار است یا مشکلی برای آن به وجود آمده است. در اینجا ماژول دریافت کننده قالب صفحه وب را با استاندارد صفحه وب مقایسه

می‌کند و در صورت عدم مشکل این محتوا را برای قسمت تجزیه کننده می‌فرستد و در غیر این صورت دوباره به کاهش امتیاز صفحه وب و همینطور صفحات وب مرتبط با آن می‌پردازد (صفحاتی که به صفحه وب دریافتی ارجاع داده‌اند).

۴-۴ قسمت تجزیه کننده

در مرحله قبل ما صفحه وب را دریافت کرده و صحت آنرا مورد بررسی قرار دادیم. وظیفه اصلی این قسمت تجزیه کردن صفحه وب و استخراج داده های مورد نیاز برای پردازش اطلاعات می‌باشد و همینطور بررسی آنها از قبیل متون، تصاویر و لینک های ارتباطی.

۴-۴-۱ استخراج تصاویر و متن از صفحه وب : در این مرحله ما به دریافت تصاویر در صفحه وب و لیست متون نزدیک به تصویر می‌پردازیم. منظور از متون نزدیک به تصویر لیست متونی است که در یکی از سطوح



شکل ۴-۳ متون نزدیک تصویر

صفحه‌وب (درخت HTML) و نزدیک به تصویر قرار دارند. این متون بر حسب یک روش خاص از صفحه‌وب استخراج شده‌اند (در مراحل قبلی پردازش صفحه وب با روش هایی یادگیری‌عمیق صفحه وب متوجه اولویت متون نزدیک به صفحه شده است به عنوان مثال در یک صفحه وب ممکن است متوجه شده باشیم که متن موجود در داخل سر صفحه وب[1] و یا خود تصویر صفحه از اهمیت بالای برخوردارند ولی در یک صفحه وب دیگر ممکن است متن با اهمیت نزدیک صفحه درپاراگراف بالای آن باشد و ...) یا به فرم تصادفی (اگر اولین بار است یکسری متون را از داخل آن استخراج می‌کنیم) شاید راهکار منطقی تر استفاده از الگوی صفحه ای است که بیشترین شباهت را با صفحه فوق دارد (معمولا صفحات دیگری از دامنه که با صفحه فوق ارتباط دارند). از طرفی هر صفحه یک الگوی استخراج متون نزدیک به تصویر مخصوص به خود دارد که با استفاده از یادگیری عمیق به مرور زمان یاد می‌گیرد. در شکل ۴-۳ می‌توان تصویر و متون نزدیک به آن را مشاهده کرد. بعد از دریافت تصویر و متون داخل صفحه وب در این مرحله ما به ایجاد بردارتصویر و بردارمتن پرداخته و هر کدام را به شبکه عصبی کانولوشن خودشان برای پردازش می‌دهیم.

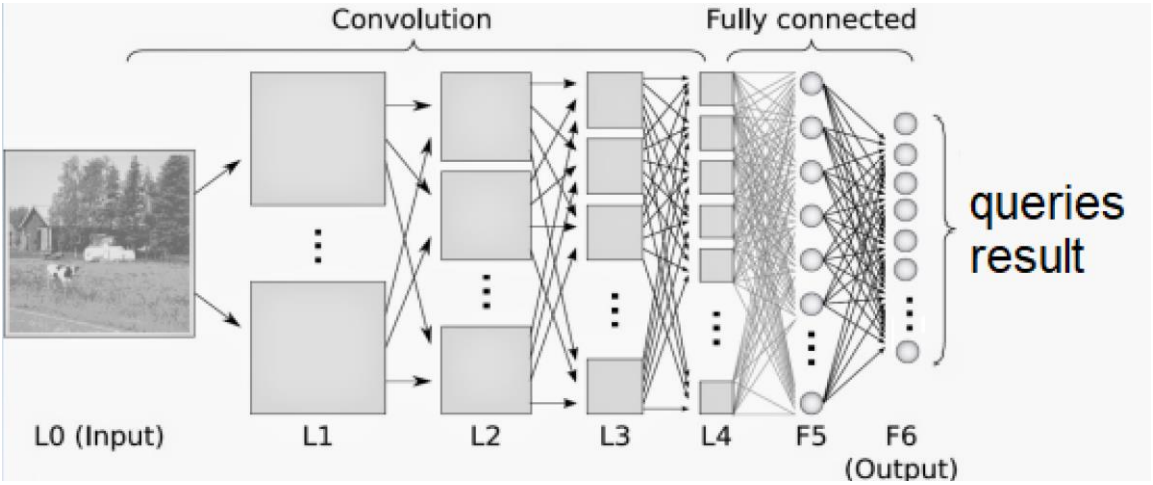
۴-۴-۲ بررسی تصویر و متن داخل صفحه وب

در گام بعدی ما به ایجاد بردار ویژگی تصویر و متون نزدیک تصویر پرداخته و آنها را به شبکه‌های کانولوشن خودشان مطابق شکل های ۴-۴ و ۴-۵ می‌دهیم. خروجی برای هر کدام از این شبکه ها یک بردار احتمال[2] به طول کلیه پرس‌وجوهای موتور جستجو می‌باشد. و احتمال شباهت هر پرس‌وجو با تصویر و متن مشخص می‌گردد. در این مرحله با توجه به متن استخراج شده و بردار احتمال به روز رسانی مکان تصاویر مهم در صفحه وب جاری می‌پردازیم. (به عنوان مثال ممکن است به مرور زمان مکان پر اهمیت متن نزدیک یک تصویر از پاراگراف نزدیک آن به برچسب فرزند آن منتقل گردد).

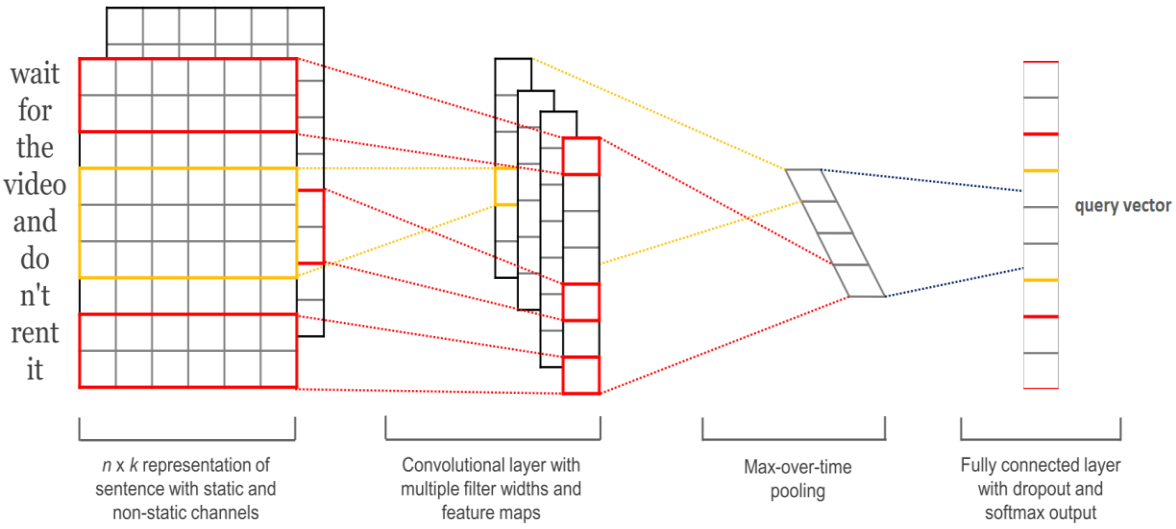
---

[1] header

[2] stochastic vector



شکل ۴-۴ شبکه کانولوشن تصویر [20]



شکل ۴-۵ شبکه کانولوشن متن [21]

حال بر حسب بردار ویژگی بدست آمده از روی تصویر می‌شود به اهمیت گروه تصویر پی برد یک راه حل برای این منظور اتصال تصویر به گروه ها با توجه به وزن بدست آمده است هر چقدر وزن یک پرس‌وجو بالاتر باشد احتمال شباهت تصویر با آن گروه هم به همان نسبت افزایش می‌یابد. حال با بروز رسانی بردارهای ویژگی پرس‌وجو ی خاص با استفاده از تصویر و وزن ارتباطی بین آنها می‌پردازیم.

در گام بعدی با استفاده از بردار ویژگی بدست آمده از متن می‌شود به بررسی پرس‌وجو ها پرداخت و برای ایجاد یک پرس‌وجوی جدید یا الحاق پرس‌وجو ها اقدام کرد. پرس‌وجو ها هم همانطور که قبلا گفته شد با هم اتصال دارند با یکسری وزن خاص حالا که بردار متن بدست آمده می‌شود وزن پرس‌وجوها را به روز رسانی کرد و ممکن است پرس‌وجوی جدید ایجاد شده که یا به یک پرس‌وجوی سطح بالاتر متصل می شود یا خود تشکیل یک طبقه پرس‌وجوی جدید را می‌دهد.

حال به به روز رسانی امتیاز و اولویت صفحه وب جاری و صفحات مرتبط با آن می‌پردازیم (تمام صفحات مرتبط با صفحه جاری چه از لحاظ ارتباط لینکی و هم از لحاظ یکسان بودن دامنه).

۳-۴-۴ استخراج لینک های اتصالی

دراین مرحله ما به استخراج کلیه لینک‌های اتصالی صفحه‌وب جاری به صفحات وب دیگر می‌پردازیم. در حقیقت هدف از این کار محاسبه امتیاز صفحه وب جاری (یا به روز رسانی) و صفحات مرتبط با آن می‌باشد. بعد از بدست آوردن لیست لینک ها در چند مرحله زیر به اقداماتی می‌پردازیم.

- **نرمالایز کردن :**

همانطور که در فصل دوم اشاره شد جهت کاهش بار پردازشی موتور جستجو با استفاده از فرایند نرمال سازی باعث کاهش و مرتب سازی آدرس‌های وب می‌شویم. و این کار باعث کاهش ترافیک شبکه نیز خواهد گردید.

- **به روز رسانی اولویت**

در این مرحله هر کدام از این ارجاعات استخراج شده را با صفحات موجود در پایگاه داده مقایسه می‌کنیم. و اگر در پایگاه داده این آدرس وجود داشته باشد بر حسب امتیازات بردار ویژگی صفحه جاری و صفحه بدست آمده اولویت صفحه تغییر می‌کند. دقت شود فرض کنید در لیست آدرس‌ها، آدرسی به

نام newURL داریم بدیهی است تمام صفحاتی در پایگاه داده که به نحوی با newURL ارتباط دارند (چه از لحاظ داشتن دامنه یکسان، دارا بودن رجاع به صفحات داخل پایگاه داده و یا خود آدرس و همینطورشباهت بالا در پرس‌وجوهای موتور جستجو) دچار تغییر امتیاز خواهند شد.

۴-۵ قسمت کنترل پایگاه داده [۱]

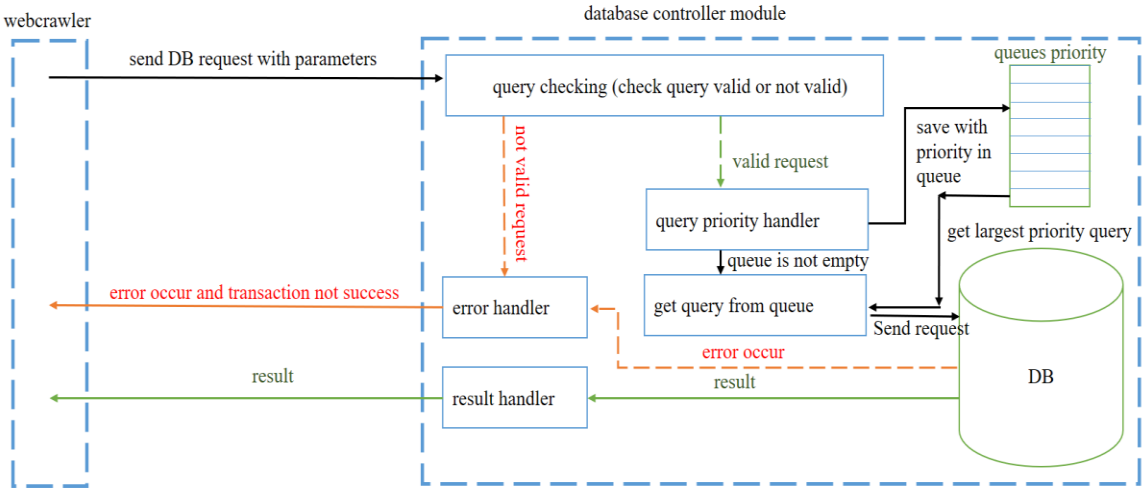
وظیفه اصلی این قسمت ذخیره و بازیابی اطلاعات در پایگاه داده می‌باشد. از طرفی به دلیل درخواست‌های زیاد و مکرر موتور جستجو از یک طرف و تغییر و به روز رسانی های بسیار زیاد موتور جستجو روی پایگاه داده از طرف دیگر باید این قسمت از لحاظ تراکنشی[۲] بسیار قوی و با اطمینان عمل کند. و کوچکترین اشکالی دراین امر منجر به گمراه شدن موتور جستجو می‌گردد به عنوان مثال در هنگام به روز رسانی بردار پرس‌وجوی یک آدرس وب اگر یک فرایند دیگر اقدام به به روز رسانی همان آدرس بکند اطلاعات با ارزش از بین خواهد رفت. در شکل ۴-۶ می‌توانید بلوک دیاگرام قسمت کنترل پایگاه داده را ببینید. درادامه به بررسی قسمت‌های مختلف این کنترلر می‌پردازیم.

- **بررسی درخواست[۳]**

این قسمت بررسی می‌کند آیا درخواست پرسیده شده از پایگاه داده معتبر است یا نه. دلیل این امرهم جلوگیری از تزریق سوالات تزریقی[۴] و درخواست‌های که فاقد مجوز لازم هستند (به عنوان مثال هر ماژول در موتور جستجو مجوز خاص خود را دارد مثلا ماژول دریافت کننده صفحات وب مجوز تغییر بردار اولویت هیچ صفحه وبی را ندارد)

---

۱ database controller module ۲ transaction ۳ query checking ۴ sql injection



شکل ۴-۶ بلوک دیاگرام ماژول کنترل پایگاه داده

- **بررسی اولویت**[۱]

این قسمت بعد از تایید درخواست توسط قسمت بررسی قرار دارد. و به بررسی اولویت درخواست می‌پردازد و بر حسب اولویت در صف درخواست ها قرار می‌دهد. این قسمت از اهمیت بسیاری در این ماژول قرار دارد و باید با یک سیستم هوشمند به تشخیص اهمیت درخواست بپردازد. به عنوان مثال اگر دو درخواست در صف داریم که یکی دریافت لیست اولویت کلیه صفحات وب و دیگری تغییر در مقدار اولویت یک صفحه‌وب باشد بدیهی است که ابتدا بهتر است اولویت آدرس تغییر کرده و سپس کل اولویت آدرس‌ها را به موتور جستجو بفرستیم.

- **اجرای درخواست**[۲]

این قسمت بالاترین درخواست را از صف اولویت استخراج کرده و دستورات را جهت اجرا به پایگاه داده می‌فرستد. و پاسخ را دریافت کرده و نتیجه را به موتور جستجو می‌فرستد البته اگر تراکنش نا موفق بود خطا را برای مدیریت به قسمت کنترل خطا می‌فرستد.

---

priority handler[۱] run handler[۲]

- **کنترل خطا**[1]

وظیفه این قسمت ارسال خطا به موتور جستجو را بر عهده دارد چنانچه خطای در نتیجه پایگاه داده رخ بدهد و یا درخواست موتور جستجو معتبر نباشد این قسمت وظیفه کنترل این خطا را بر عهده دارد.

۴-۶ پایگاه داده موتور جستجو :

در موتور جستجو برای ذخیره داده‌ها از یک پایگاه داده رابطه‌ای استفاده کرده‌ایم. دلایل استفاده از پایگاه داده رابطه‌ای را می‌توانید درلیست زیر ببینید.

- به علت ارتباط بالای جداول پایگاه داده با یکدیگر(به عنوان مثال جدول تصاویر ارتباط نسبتاً بالای با جدول صفحات وب دارد و هر تصویر به یک صفحه وب متصل می‌باشد) یک پایگاه داده رابطه‌ای بسیار بهتر و سریعتر می‌تواند جداول را با یکدیگر مرتبط کند.
- مدل رابطه‌ای پایگاه داده به طور پیش فرض بسیار سریعتر از مدل‌های دیگر پایگاه داده مانند مدل ساختاری می‌باشد. و به دلیل حجم بالای داده در یک موتور جستجو سرعت پاسخگویی پایگاه داده و ارتباط آن با موتور جستجو از اهمیت بالای برخوردار است.

در این پروژه از یک پایگاه داده رابطه‌ای به نام webCrawlerDB استفاده می‌کنیم. در این پایگاه داده از چندین جدول که با یکدیگر رابطه دارند استفاده گردیده است. حال به معرفی جداول پایگاه داده می‌پردازیم.

۱- **جدول دامنه**[2] : همانطور که از نام این جدول مشخص می‌باشد وظیفه این جدول ذخیره دامنه‌هایست که موتور جستجو پیدا می‌کند. در جدول ۴-۱ می‌توانید لیست ستون‌های این جدول را ببینید. و در شکل ۴-۷ دیاگرام این جدول را مشاهده می‌کنید.

---

[1] error handler [2] domain

domain		
id	char(36)	PRIMARY KEY UNIQUE
domain	varchar(1000)	
vote	float	

شکل ۴-۷ جدول دامنه

جدول ۴-۱ مشخصات جدول دامنه

ردیف	نام ستون	نوع	توضیحات
۱	id	char(36)	کلید اصلی جدول و مقداری منحصربه‌فرد برای هر سطر جدول .
۲	domain	varchar(1000)	ذخیره دامنه جدول
۳	vote	float	وظیفه این ستون جدول ذخیره امتیاز دامنه می‌باشد. برای فرایند انتخاب دامنه در موتور جستجو از این بخش استفاده می‌شود.

۲- جدول امتیاز دامنه[۱] :

این جدول وظیفه نگهداری وزن‌های هر دامنه جهت به روز رسانی‌امتیاز آنها را دارد. برای به روز رسانی امتیاز هر دامنه با استفاده از رگرسیون خطی برای هر دامنه از این جدول استفاده می‌گردد که وظیفه ذخیره پارامترهای و به روز رسانی دامنه را به عهده دارد. در جدول ۴-۲ می‌توانید لیست ستون‌های این جدول را ببینید. و در شکل ۴-۸ دیاگرام این جدول را مشاهده می‌کنید.

[۱] domain rank

DOMAINRANK		
id	char(36)	PRIMARY KEY UNIQUE
domainID	char(36)	FOREIGN KEY DOMAIN(id)
preHypothesis	float	
newHypothesis	float	
W0	float	
W1	float	
W2	float	
W3	float	
learningRatio	float	

شکل ۴-۸ جدول امتیازدامنه

جدول ۴-۲ مشخصات جدول امتیازدامنه

ردیف	نام ستون	نوع	توضیحات
۱	id	char(36)	کلید اصلی‌جدول و مقداری منحصربه‌فرد برای هر سطر جدول
۲	domainID	char(36)	کلید اصلی جدول‌دامنه (جهت ارتباط این دو جدول با یکدیگر)
۳	preHypothesis	float	ذخیره امتیاز قبلی دامنه در این سلول قرار دارد.
۴	newHypothesis	float	ذخیره امتیاز کنونی دامنه در این ستون قرار دارد
۵	W۰	float	ذخیره وزن پارامتر اول در دامنه
۶	W۱	float	ذخیره وزن پارامتر دوم در دامنه
۷	W۲	float	ذخیره وزن پارامتر سوم در دامنه
۸	W۳	float	ذخیره وزن پارامتر چهارم در دامنه
۹	learningRatio	float	نرخ یادگیری برای هر دامنه در روش رگرسیون

۳- جدول صفحات وب[۱]:

وظیفه این جدول ذخیره آدرس تمام صفحات وب داخل یک دامنه خاص می‌باشد. هر دامنه شامل چندین صفحه با آدرس مختلف می‌باشد که همه آنها در این جدول قرار دارند. در جدول ۴-۳ می‌توانید لیست

[۱] Page table

ستون‌های این جدول را ببینید. و در شکل ۴-۹ دیاگرام این جدول را مشاهده می‌کنید.

PAGE		
id	**char(36)**	**PRIMARY KEY UNIQUE**
domainID	**char(36)**	**FOREIGN KEY DOMAIN(id)**
uri	**varchar(2000)**	
vote	**float**	

شکل ۴-۹ جدول صفحه وب

جدول ۴-۳ مشخصات جدول دامنه

ردیف	نام ستون	نوع	توضیحات
۱	id	char(36)	کلید اصلی‌جدول و مقداری منحصربه‌فرد برای هر سطر جدول
۲	domainID	char(36)	کلید اصلی جدول دامنه (جهت ارتباط این دو جدول با یکدیگر)
۳	uri	Varchar(2000)	وظیفه این ستون جدول ذخیره آدرس منحصر بفرد و نرمالایز شده هر صفحه وب می‌باشد.
۴	vote	float	وظیفه این ستون جدول ذخیره امتیاز صفحه وب می‌باشد. برای فرایند انتخاب صفحه وب در موتور جستجو از این بخش استفاده می‌شود.

۴- جدول تاریخچه[۱] :

این جدول وظیفه نگهداری تاریخچه دریافت یک صفحه وب را دارد. بعد از دریافت هر صفحه وب محتوای صفحه وب در این جدول به همراه تاریخ دریافت ذخیره می‌شود. در جدول ۴-۴ می‌توانید لیست ستون‌های این جدول را ببینید. و در شکل ۴-۱۰ دیاگرام این جدول را مشاهده می‌کنید.

---

history table[۱]

HISTORY		
id	char(36)	PRIMARY KEY UNIQUE
pageID	char(36)	FOREIGN KEY PAGE(id)
response	int	
content	text	
seeImages	smallint	
getImages	smallint	
seeLinks	smallint	
normalLinks	smallint	
date	timestamp	

شکل ۴-۱۰ جدول تاریخچه

۵- جدول رابطه :

این جدول وظیفه ذخیره رابطه بین صفحات وب را بر عهده دارد. در حقیقت در این جدول لیست از ارتباطات صفحات مختلف وب را داریم. به عنوان مثال اگر صفحه وب X با صفحه وب Y ارتباط داشته باشد در این جدول یک سطر حاوی این ارتباط قرار دارد. در جدول ۴-۵ می‌توانید لیست ستون‌های این جدول را ببینید و در شکل ۴-۱۱ دیاگرام این جدول را مشاهده می‌کنید.

RELATION		
id	char(36)	PRIMARY KEY UNIQUE
inputlink	char(36)	FOREIGN KEY PAGE(id)
outputlink	char(36)	FOREIGN KEY PAGE(id)

شکل ۴-۱۱ جدول رابطه

جدول ۴-۴ مشخصات جدول تاریخچه

ردیف	نام ستون	نوع	توضیحات
۱	id	char(36)	کلید اصلی این جدول و مقداری منحصربه‌فرد برای هر سطر جدول است
۲	pageID	char(36)	کلید اصلی جدول صفحه. (جهت ارتباط این دو جدول با یکدیگر) هر صفحه وب شامل چندین تاریخچه از خود می‌باشد.
۳	response	int	امتیاز نتیجه دریافت هر صفحه وب در یک زمان مشخص می‌باشد. همینطور که قبلا گفته شد به عنوان مثال اگر صفحه وب با موفقیت دریافت شود این امتیاز ۲۰۰ می‌باشد.
۴	content	text	وظیفه این ستون جدول ذخیره خود صفحه وب می‌باشد که در غالب HTML می‌باشد.
۵	seeImages	smallInt	نمایش تعداد تصاویر دیده شده در یک صفحه وب خاص در یک زمان مشخص می‌باشد.
۶	getImages	smallInt	نمایش تعداد تصاویری که در یک صفحه وب خاص و در یک زمان مشخص با موفقیت دریافت شده‌اند.
۷	seeLinks	smallInt	تعداد آدرس صفحات وب دیده شده در یک صفحه وب مشخص و در یک زمان مشخص.
۸	normalLink	smallInt	تعداد آدرس صفحات وب نرمال شده در یک صفحه وب مشخص و در یک زمان مشخص.
۹	date	date	تاریخ دریافت صفحه وب

جدول ۴-۵ مشخصات جدول رابطه

ردیف	نام ستون	نوع	توضیحات
۱	id	char(36)	کلید اصلی این جدول و مقداری منحصربه‌فرد برای هر سطر جدول است
۲	inputLink	char(36)	کلید اصلی جدول صفحه (جهت ارتباط این دو جدول با یکدیگر)
۳	outputLink	char(36)	کلید اصلی جدول صفحه (جهت ارتباط این دو جدول با یکدیگر)

۶- جدول تصویر :

این جدول وظیفه ذخیره تمامی تصاویر یک صفحه وب مشخص را دارد. در جدول ۴-۶ می‌توانید لیست ستون‌های این جدول را ببینید. و در شکل ۴-۱۲ دیاگرام این جدول را مشاهده می‌کنید.

جدول ۴-۶ مشخصات جدول رابطه

ردیف	نام ستون	نوع	توضیحات
۱	id	char(36)	کلید اصلی این جدول و مقداری منحصربه‌فرد برای هر سطر جدول است
2	pageID	char(36)	کلید اصلی جدول صفحه وب (جهت ارتباط تصویر با صفحه وب)
۳	blob	text	این ستون جهت ذخیره محتوای تصویر دریافتی توسط موتور جستجو استفاده می‌شود.
۴	w	int	این ستون در بردارنده تعداد پیکسل سطرهای یک تصویر مشخص است
۵	h	int	این ستون در بردارنده تعداد پیکسل ستون‌های یک تصویر مشخص است
۶	src	text	این ستون دربرگیرنده آدرس تصویر دریافتی می‌باشد.

IMG		
id	char(36)	PRIMARY KEY UNIQUE
pageID	char(36)	FOREIGN KEY PAGE(id)
blob	text	
w	int	
h	int	
src	text	

شکل ۴-۱۲ جدول تصویر

۷-جدول متن[۱]:

این جدول وظیفه ذخیره تمامی متن‌های نزدیک یک تصویر مشخص به همراه امتیاز آن متن می‌باشد. در جدول ۷-۴ می‌توانید لیست ستون‌های این جدول را ببینید. و در شکل ۴-۱۳ دیاگرام این جدول را مشاهده می‌کنید.

TEXT		
id	char(36)	PRIMARY KEY UNIQUE
imgID	char(36)	FOREIGN KEY IMG(id)
text	text	
distance	Varchar(100)	
tag	text	

شکل ۴-۱۳ جدول متن

در انتها و در شکل ۴-۱۴ می‌توانید لیست کامل جدول‌های پایگاه داده به همراه ارتباط آنها با یکدیگر را ببینید.

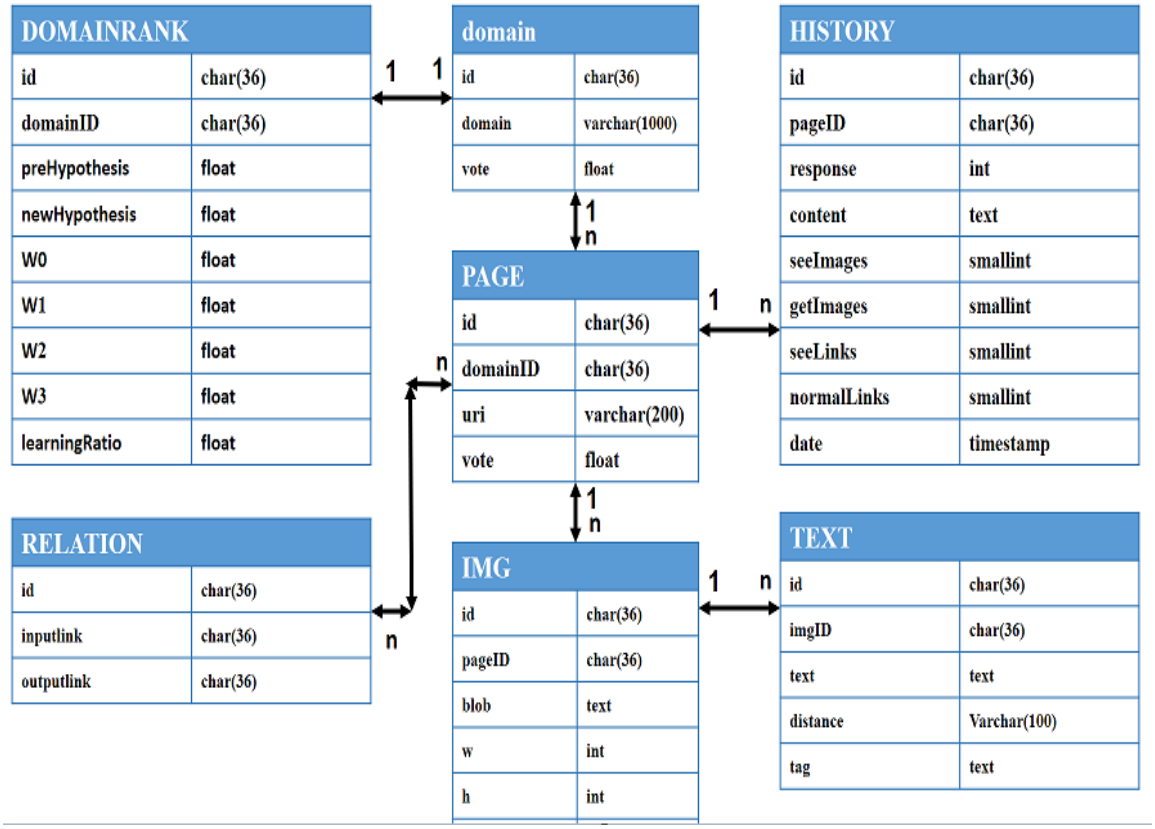
در ارتباطات عبارات ۱وn نمایشگر تعداد رابطه بین جداول است به عنوان مثال و مطابق شکل هر سطر از جدول page با تعداد زیادی از جدول history ارتباط دارد.

---

text table[۱]

جدول ۴-۷ مشخصات جدول متن

ردیف	نام ستون	نوع	توضیحات
۱	id	char(36)	کلید اصلی این جدول و مقداری منحصربه‌فرد برای هر سطر جدول است
۲	imgID	char(36)	کلید اصلی جدول تصویر (جهت ارتباط تصویر با صفحه وب)
۳	text	text	این ستون جهت ذخیره یکی از متون نزدیک یک تصویر مشخص می‌باشد.
۴	distance	Varchar(100)	این ستون در بردارنده فاصله متن از تصویر در درخت HTML می‌باشد
۵	tag	text	این ستون در بردارنده نام تگی است که این متن را دربر گرفته است.



شکل ۴-۱۴ لیست روابط در پایگاه داده

فصل پنجم:

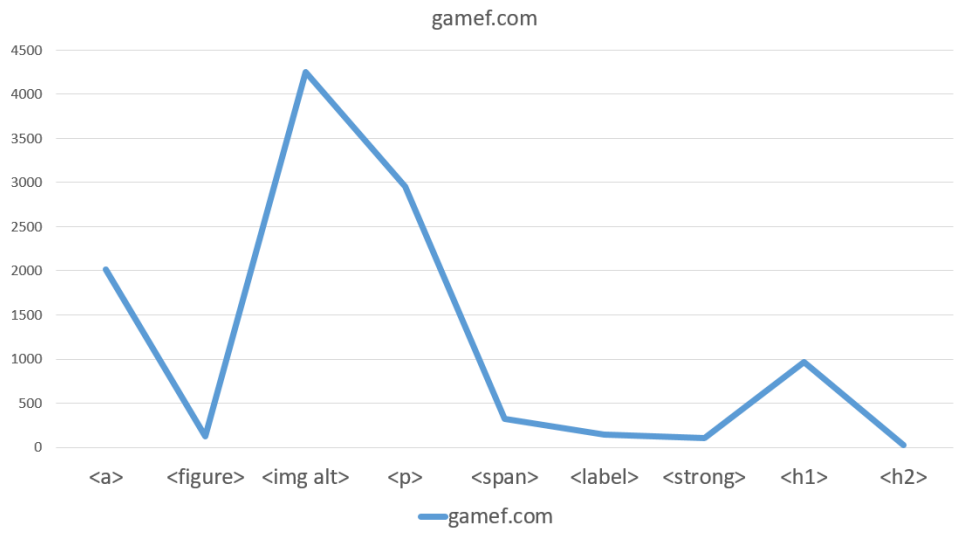
نتیجه گیری

۵-۱ مقدمه

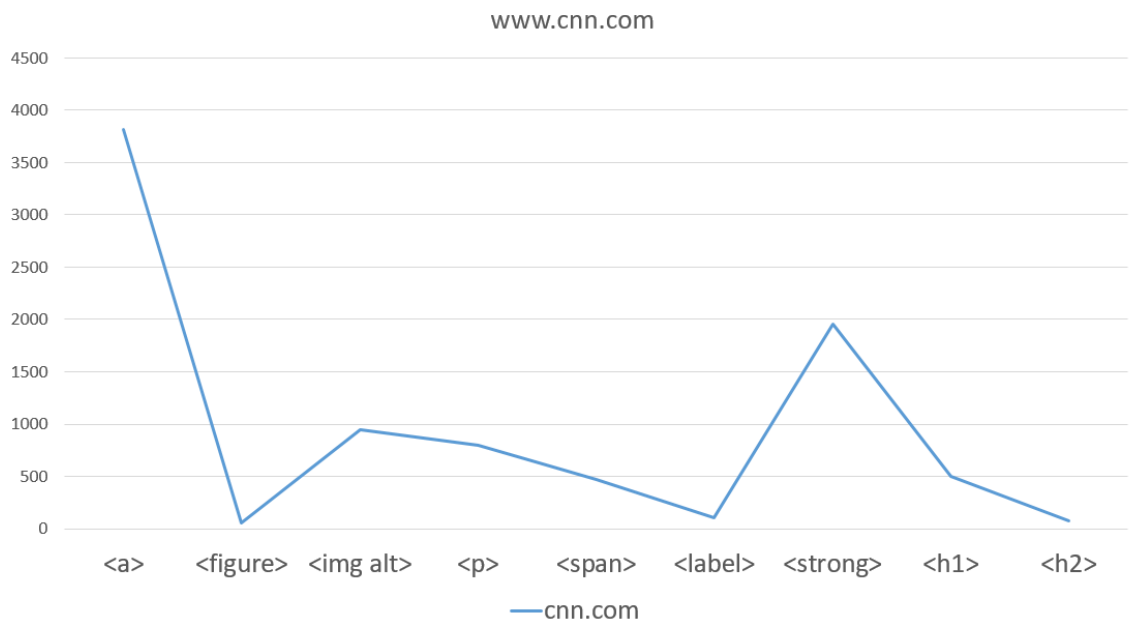
در فصول قبل ابتدا به بررسی صفحات وب و انواع آنها پرداختیم. سپس در فصل سوم به بررسی چالش ها و راهکار های برای آنها پرداختیم. و در فصل چهارم هم به بررسی اجمالی از روند کار موتور جستجو پرداختیم در این فصل که فصل نتیجه گیری می‌باشد حالا به بررسی چند خروجی و مقایسه روش‌هایی مختلف در پروژه می‌پردازیم.

۵-۲ اهمیت مکانی متون در صفحه وب

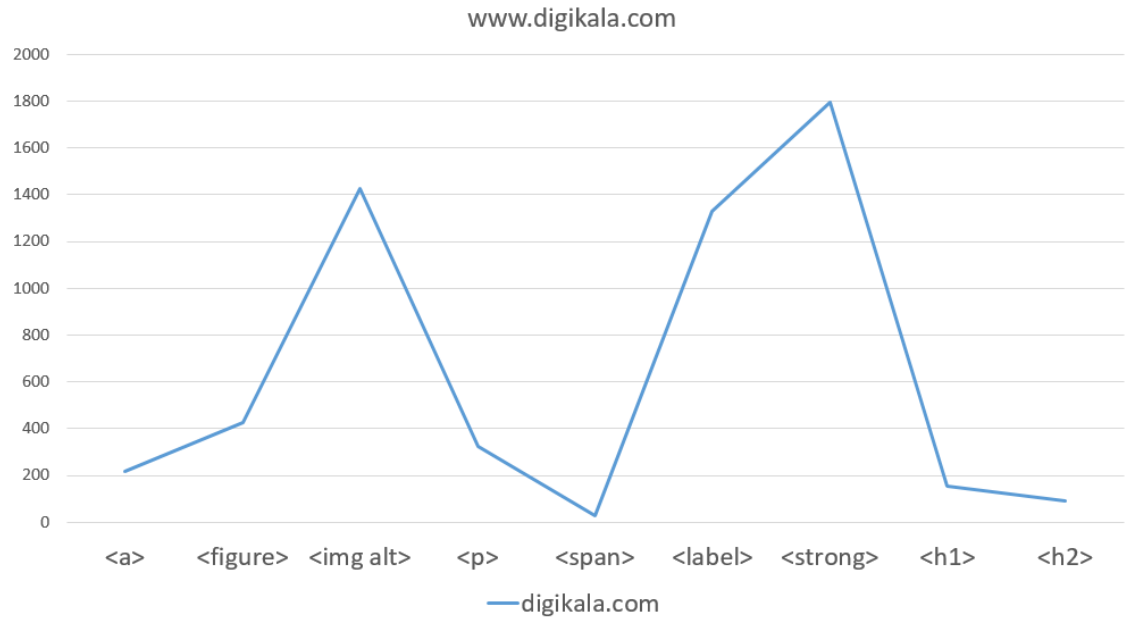
خوشبختانه صفحات وب (کلیه صفحات یک وب سایت در غالب یک دامنه) از لحاظ غالب و تصویر حالت یکسانی دارند. به عنوان مثال در شکل‌های ۵-۱ الی ۵-۶ به بررسی ۶ دامنه مختلف و مکان قرار گرفتن بهترین متون نزدیک تصاویر را در آنها ببینید (در اینجا منظور از تگ های مفید تگ‌های است که اطلاعات داخلی آنها بیشترین شباهت را با لیست پرس‌وجوهای داخل موتور جستجو دارند چه از نظر متنی چه از نظر تصویری). همانطور که در تصاویر مشخص است هر سایت تقریبا یک غالب مشخص دارد.



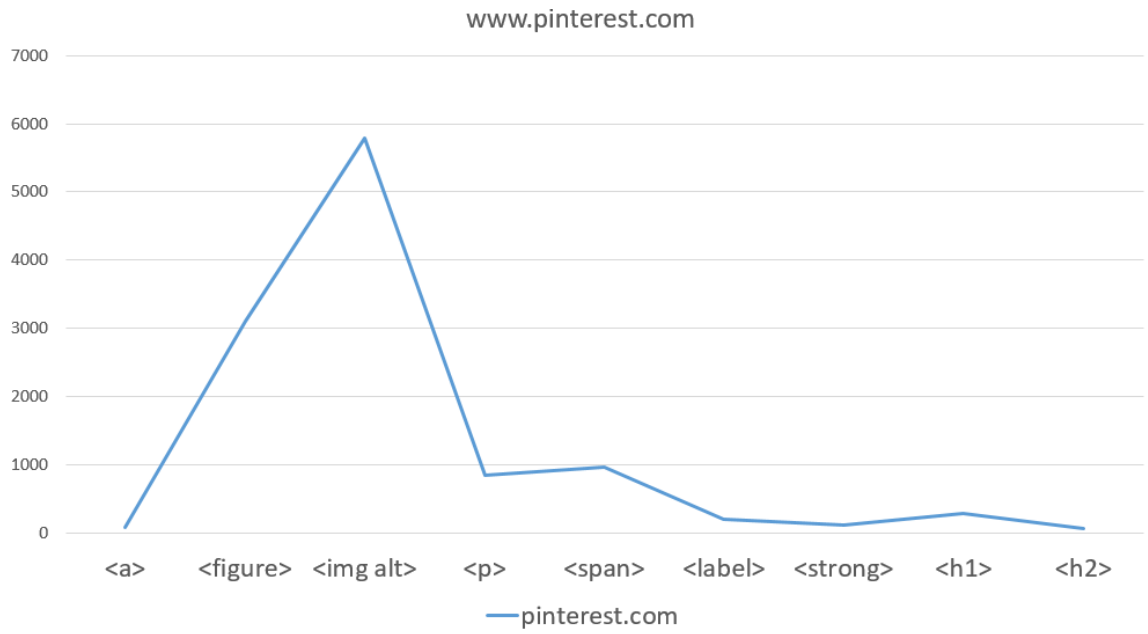
شکل ۵-۱ مقایسه مکان تصاویر دردامنه -۱



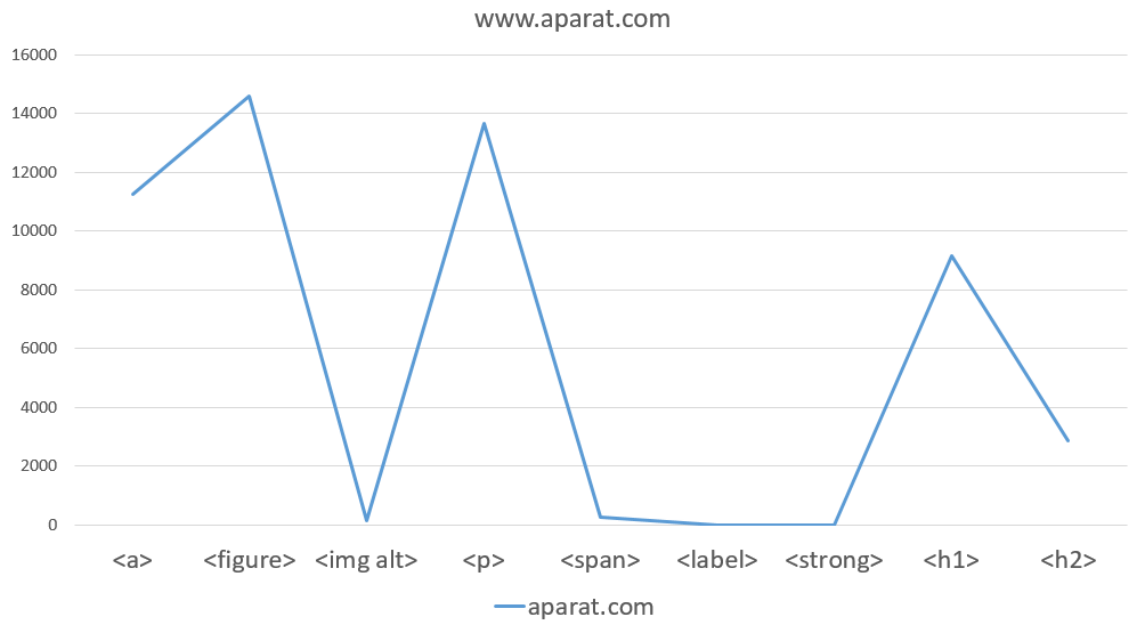
شکل ۵-۲ مقایسه مکان تصاویر دردامنه -۲



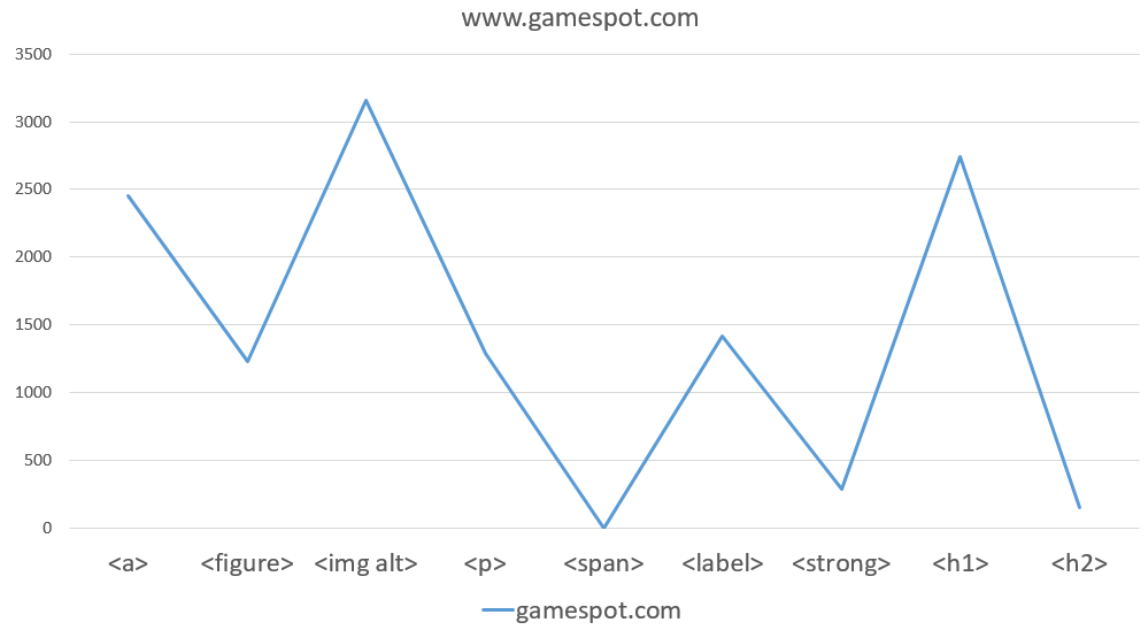
شکل ۵-۳ مقایسه مکان تصاویر دردامنه -۳



شکل ۵-۴ مقایسه مکان تصاویر دردامنه -۴



شکل ۵-۵ مقایسه مکان تصاویر دردامنه -۵



شکل ۵-۶ مقایسه مکان تصاویر دردامنه -۶

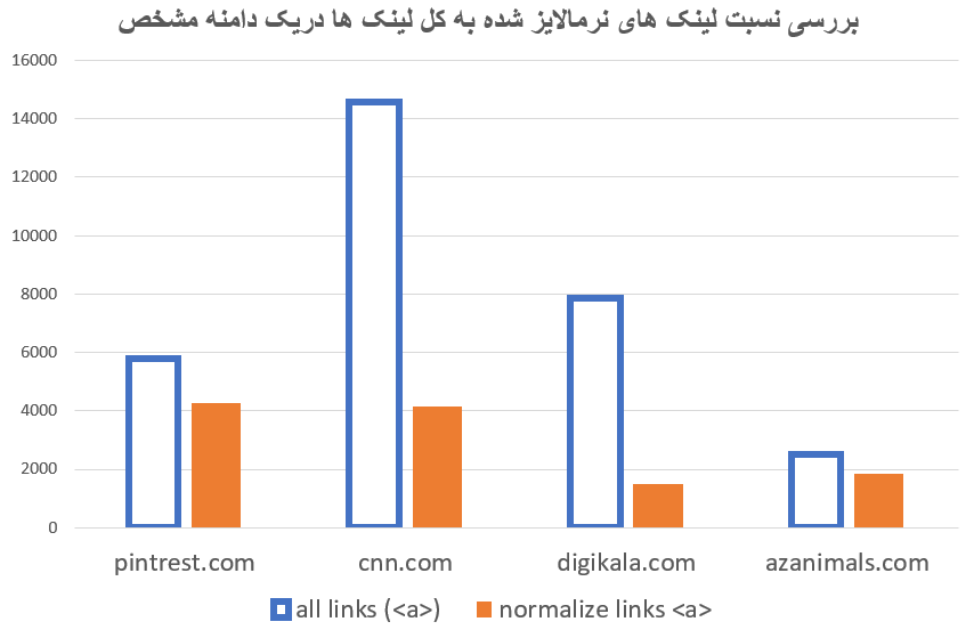
در این جا منظور بررسی متون نزدیک تصویر می‌باشد که اطلاعات خوبی درباره تصویر را در خود دارد و همینطور خود تصاویر. به عنوان مثال در سایت gamefa.com بیشترین متون مفید نزدیک تصویر (مشابه پرس‌وجوهای موتور جستجو) در غالب پاراگراف بالای آن و متن توضیحی داخل خود تصویر و توضیحات داخل اتصالات (تگ <a>) می‌باشد. پس می‌شود نتیجه گرفت که با بررسی در هر دامنه خاص به اهمیت مکان متون و تصاویر با ارزش درباره یک پرس‌وجوی خاص مشخص می‌شود. البته بهتر است این مقادیر با دیدن صفحات جدید به روز شود. ولی در کل معمولا غالب صفحات وب در یک دامنه یکسان است.

به عنوان مثالی دیگر در دامنه digikala.com همانطور که مشخص است تگ‌های و <label> بیشترین شباهت را با پرس‌وجوهای داخل موتور جستجو دارند.

۵-۳ تاثیر نرمال سازی در کاهش بار پردازشی صفحات وب

در اینجا به بررسی تاثیر نرمال سازی آدرس صفحات در یک دامنه خاص می‌پردازیم. برای این بررسی ما بیش از ۵۰۰ صفحه وب از ۳۰ دامنه متفاوت را مورد بررسی قرار داده و از همه این ۵۰۰ صفحه تمام

لینک‌های داخلی آنرا بدست آوردیم، سپس تمام این آدرس ها را نرمالایز کردیم نتیجه بدست آمده حیرت آور است. بعد از ارسال موتور جستجو به این صفحات وب و استخراج تمام لینک ها بیش از ۷۸۳۹ لینک تولیدگردید است که بعد از نرمالایزه کردن و حذف دامنه های تکراری، تعداد این لینک ها به ۲۰۹۳عدد $(\frac{7839}{2093} = 3.74)$ کاهش می‌یابد. که بسیار عالی می‌باشد. پس می‌توان نتیجه گرفت که برای کاهش ترافیک شبکه و همینطور کاهش بار پردازش موتور جستجو بهتر است قبل از هر فرایندی آدرس را نرمالایز کنیم. در شکل ۵-۷ به بررسی ۴ دامنه پرداختیم که برای هر دامنه ۱۰۰۰ صفحه وب را بررسی کرده و تعداد لینک بدست آمده را با تعداد لینک بعد از نرمالایز کردن مقایسه نموده‌ایم.



شکل ۵-۷ مقایسه نتیجه نرمال سازی صفحات وب

۵-۴ تایین امتیاز هر صفحه وب.

تغییر امتیاز هر صفحه وب همانطور که قبلا هم گفته شد در چند مرحله اتفاق می‌افتد. ممکن است امتیاز صفحه افزایش یا کاهش بیابد و این تغییر امتیاز صفحه روی صفحاتی که با این صفحه ارتباط دارند (به

صفحه فوق لینک ورودی داده اند و از صفحه خود به این صفحه یک اتصال ایجاد کرده اند) هم تاثیر دارد و باید امتیاز آنها هم دچار تغییر گردد.

۱-۴-۵ دلایل اصلی کاهش امتیاز یک صفحه وب

۱- عدم دسترسی به صفحه وب :

دلیل این اتفاق یکی از عوامل زیر می‌باشد

- **عدم داشتن مجوز لازم جهت ورود به صفحه وب** : ممکن است صفحه وب قرار داشته روی یک خدمت‌گذار مجوز عمومی برای دسترسی به یک صفحه وب را ندهد و برای ورود به این صفحه وب باید اقداماتی مانند اهراز هویت صورت بگیرد. شاید یک راه حل برای رفع این مشکل استفاده از پروتکل http بجای https باشد و یا حتی استفاده از موتورهای جستجوی مخفی.
- **حذف کامل صفحه وب از روی خدمت گذار** : ممکن است صفحه وب انتخابی بنا به دلایلی (مانند انقراض و قدیمی شدن اطلاعات صفحه وب) از خدمت‌گذار خود حذف شده باشد.
- **حذف موقت صفحه وب از خدمت گذار** : ممکن است صفحه وب انتخابی بنا به دلایلی (تغییر محتوای صفحه وب مانند غالب یا اطلاعات) به طور موقت از روی خدمت‌گذار خود حذف شده باشد. و بعد از مدتی دوباره بر روی خدمت‌گذار خود قرار بگیرد.

راه حل : برای موارد بالا بهتر است بعد از بروز مشکل مدتی صفحه وب از لیست صفحات وب ما خارج گردد

و بعد از یک مدت دوباره در لیست بررسی قرار بگیرد. همانطور که در جدول ۵-۱ می‌بینید ما یک سایت خاص (gamefa.com) را مورد بررسی قرار دادیم و همانطور که می‌بینید به عنوان مثال در صورت عدم دسترسی به یک صفحه وب معمولا غالب سایت در حال به روز رسانی می‌باشد و بعد از مدت تغریبا ۳۵ ساعت دوباره در دسترس قرار می‌گیرد. و تغریبا در صورت حذف یک مطلب بعد از ۴ بار بررسی می‌شود نتیجه گرفت مطلب به کلی از سایت و دامنه فوق حذف گردیده است. پس اگر موتور جستجو این مطلب را متوجه شود بعد از یک دوره زمانی ۳۵ ساعته دوباره آدرس فوق را در لیست بررسی خود قرار می‌دهد و در بررسی مشاهده گردید که اغلب سایت ها در این مورد یک استاندارد نسبتا ثابتی را رعایت می‌کنند. در جدول ۵-۱ ما به بررسی چند زیر دامنه از دامنه gamefa.com پرداخته و همانطور که می‌بینید زمان غیر قابل از دست دادن صفحات در آنها مشخص شده است.

جدول ۵-۱ بررسی ۸ زیر دامنه در یک دامنه خاص متوسط در دسترس بودن بعد از ۴ بار بررسی

ردیف	نام آدرس	زمان دردسترس بودن
۱	http://cinema.gamefa.com/	۳۱ ساعت و ۲۴ دقیقه
۲	http://gamefa.com/category/platform/playstation/۳	۳۶ ساعت و ۲۹ دقیقه
۳	http://gamefa.com/category/tech/	32ساعت و ۱۸ دقیقه
۴	http://gamefa.com/category/tech/hardware/	۴۱ ساعت و ۱۲ دقیقه
۵	http://gamefa.com/category/featured/news-game/	۲۹ ساعت و ۱۴ دقیقه
۶	http://cinema.gamefa.com/category/news/	۱۲ ساعت و ۳۵ دقیقه
۷	http://cinema.gamefa.com/category/tv-series-news/	۱۹ ساعت و ۴۹ دقیقه
۸	http://cinema.gamefa.com/writers/	۳۹ ساعت و ۲۳ دقیقه

مهمترین کاربرد این بخش در کاهش بار ترافیکی و پردازشی موتور جستجو خود را نشان می‌دهد. به عنوان مثال اگر ما بعد از مشاهده یک خطا در دریافت یک صفحه وب دوباره اقدام به دریافت این صفحه کنیم با

یک ترافیک بالا و بی‌فایده طرف خواهیم بود که مدام در حال دریافت صفحات وبی می‌باشد که به طور موقت غیر فعالند و هیچ داده‌ای نیز جمع آوری نمی‌شود.

نوآوری :

در این جا بعد از بروز مشکل عدم دسترسی علاوه بر کاهش امتیاز صفحه وب و دامنه آن یک دوره استراحت برای آن در نظر گرفته می‌شود (البته برای هر دامنه یک مقدار) و بعد از اتمام زمان بروز مشکل اگر کماکان صفحه وب در دسترس نبود علاوه بر کاهش امتیاز صفحه وب و دامنه آن، زمان استراحت دامنه را هم افزایش می‌دهیم با استفاده از فرمول ۱-۵ .

$$RestTime_{domain} = RestTime_{Domain} + \alpha * PageRank_{selected\ web\ page} \qquad \text{(فرمول ۵-۱)}$$

حالا به معرفی پارامترهای این فرمول می‌پردازیم.

- $RestTime_{domain}$: منظور زمان استراحت کلی یک دامنه است بر حسب میلی ثانیه بعد از وقوع یک مشکل.
- $PageRank_{selected\ web\ page}$: امتیاز صفحه وب جاری
- α : ضریب میرایی که ثابتی است که به فرم تجربی بدست می‌آید و برای هر دامنه مقدار مختلفی دارد.

در اینجا به مرور زمان، زمان استراحت صفحه وب افزایش می‌یابد پس برای جلوگیری از این مشکل بعد از اتمام زمان استراحت و در صورت در دسترس بودن صفحه وب (بعد از زمان استراحت و اقدام به دریافت دوباره صفحه وب با موفقیت دریافت می‌شود) انتخابی زمان استراحت را با همان فرمول بالا کاهش می‌دهیم فقط ثابت α ضریب منفی می‌گیرد. از طرفی بعد از چند دفعه غیر فعال شدن صفحه وب می‌شود نتیجه گرفت صفحه به طور کامل

از سایت حذف گردیده است و می‌شود با یک اطمینان بالای آنرا از لیست آدرسها حذف کرد. مورد آخر این که در بعضی از دامنه ها کل دامنه یک جا و به دلایل مختلفی از لیست خدمات‌دهنده برای یک مدت مشخص خارج می‌شود. یعنی با ملاحظه خرابی حتی یک صفحه از دامنه می‌شود به این نتیجه رسید که دامنه موقتا از کار افتاده است (معمولا جهت به روز رسانی دامنه).

۲- کاهش امتیاز به دلیل لینک دادن به صفحه وبی که امتیازش کاهش یافته است :

فرض کنید صفحه‌وبی به نام X از یک دامنه خبری به نام Y به یک صفحه‌وب به نام Z از یک سایت دیگر به نام W لینک اتصالی داده است (به عنوان مثال لینک خبر را از سایت فوق گرفته است) و حالا صفحه وب Z بنا به دلایلی از دامنه W حذف گردیده است. پس به نحوی موتور جستجو باید امتیاز صفحه X را هم کاهش بدهد. دلیل این امر مشکل در لینک اتصالی است که فقط یک بار پردازشی اضافه روی موتور جستجو می‌گذارد. البته ممکن است صفحه وب موقتاً از خدمت‌گذار حذف گردیده باشد پس بهتر است بعد از برگشت به صفحه وب امتیاز صفحاتی که به آن لینک داده بودند نیز افزایش یابد.

۵-۵ بررسی اهمیت متون نزدیک تصاویر در صفحه وب

شاید مهمترین نوآوری این پروژه را بتوان ادغام پردازش تصویر و متن‌های داخل صفحه وب با پرس‌وجوهای مورد جستجو نامید. در این مورد جدول زیر مشخص کننده صحت این مطلب می‌باشد. همانطور که می‌بینید در این جدول ۱۰ دامنه را انتخاب کرده و در هر کدام لیست تصاویری که در کنارشان حداقل یک متن مطابق با تصویر داشته باشد در مقابل تصاویری که حتی یک متن مشابه موضوع تصویر در آنها باشد را مشاهده می‌کنیم. به عنوان مثال دامنه blaxup.com تقریبا نزدیک هر تصویر ۲ متن مشابه با آن وجود دارد (مشابه پرس‌وجوهای داخل موتور جستجو). این مطلب را می‌توان در شکل ۵-۸ هم مشاهده کرد. همانطور که می‌بینید در کنار تصویر ۲ مطلب کاملا مرتبط باتصویر وجود دارد. همانطور که قبلا هم بررسی کردیم

غالب دامنه های وب که شامل صد ها یا حتی هزاران صفحه وب هستند معمولا یکی است. پس می‌شود نتیجه گرفت که به عنوان مثال در دامنه blaxup.com و نزدیک هر تصویر حداقل دو متن مرتبط با آن وجود دارد. این روش کمک شایانی در افزایش سرعت پردازش داده دارد به عنوان مثال وقتی موتور جستجو یک صفحه وب جدید را از دامنه blaxup.com دریافت می‌کند با یک درجه بالای میتواند به متون نزدیک تصویر و محتوای آنها اعتماد بکند یا نه. خوشبختانه در غالب سایت می‌شود به این مطلب رسید و دید که این نرخ تقریبا در دامنه سایت یکسان می‌ماند. به عنوان مثال در جدول ۲-۵ لیست ۵ صفحه وب جدید اضافه شده به دامنه pinterest.com را می‌بینید که نرخ متن مرتبط با تصویر آنها بدون محاسبه نزدیک نرخ دامنه سایت می‌باشد ($\frac{7480}{8412}$ = ۰,۸۸). پس می‌شود حتی بدون محاسبه و با دیدن لینک صفحه وب جدید متوجه تعداد متون نزدیک یک تصویر گردید ولی برای بروز رسانی اطلاعات محاسبه اطلاعات هر صفحه و اعمال روی دامنه کلی اجباری است.

شکل ۵-۸ متون نزدیک و مرتبط به تصویر در یک دامنه خاص

جدول ۵-۲ بررسی نسبت متون مرتبط با تصاویر در ۱۰ دامنه مختلف

ردیف	دامنه	متن مرتبط با تصویر / تصویر
۱	http://edition.cnn.com/	$۱,۶۸ = \frac{3409}{2019}$
۲	http://p۳۰download.com/	$۱,۶۵ = \frac{19127}{11489}$
۳	http://gamefa.com/	$۲,۴۲ = \frac{8652}{3568}$
۴	http://www.blaxup.com/	$۲,۰۱ = \frac{2981}{1478}$
۵	http://www.aparat.com/	$۱,۱۷ = \frac{21774}{18592}$
۶	www.intelcapital.com	$۰,۸۷ = \frac{4198}{4785}$
۷	https://twitter.com	$۱,۳ = \frac{14796}{11325}$
۸	https://www.forbes.com	$۱,۴۴ = \frac{1420}{984}$
۹	pinterest.com	$۰,۸۸ = \frac{7480}{8412}$
۱۰	www.vajehyab.com	$۰,۵۸ = \frac{1028}{1756}$

۵-۶ بررسی نتایج تست موتور جستجو

در اینجا بعد از فاز آموزشی توسط برنامه کآفی خروجی ما یک شبکه وزنی بزرگ می‌باشد که جهت تست تصاویر ورودی از آنها استفاده می‌گردد. در تصاویر ۵-۹ الی ۵-۱۹ نمایش از تست تعداد ۹ پرس‌وجوی بدست آمده از شبکه عصبی خروجی(به فرم تصادفی) با ۱۰ صفحه وب که درباره این موضوع مطالبی دارند نمایش داده شده است. در تصاویر به عنوان مثال اگر در مجموع این صفحات حاوی ۴۰ تصویر مشابه با گروه انتخابی باشند. اگر

شبکه ما تعداد ۳۰ تا از این تصاویر را درست و مطابق این گروه انتخاب کرده باشد دقت این شبکه عصبی در این گروه ۷۵٪ می‌باشد.

P1 : https://en.wiktionary.org/wiki/magpie

P2 : http://www.abc.net.au/news/2017-08-31/make-friends-with-magpies-to-avoid-swooping-expert-says/8856438

P3 : https://www.rspb.org.uk/birds-and-wildlife/bird-and-wildlife-guides/bird-a-z/m/magpie/

P4 : https://www.collinsdictionary.com/dictionary/english/magpie

P5 : http://www.arkive.org/magpie/pica-pica/

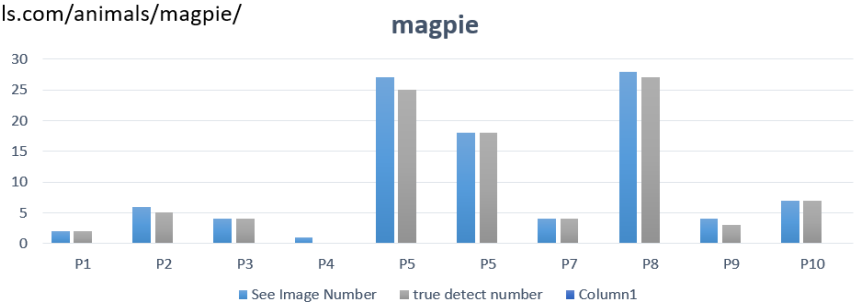
P6 : http://what-when-how.com/birds/black-billed-magpie-birds/

P7 : https://www.britannica.com/animal/magpie

P8 : http://nzbirdsonline.org.nz/species/australian-magpie

P9 : http://www.birdsinbackyards.net/species/Cracticus-tibicen

P10: https://a-z-animals.com/animals/magpie/



شکل ۵-۹ بررسی ۱۰ صفحه درباره کلاغ زاغی

P1 : http://en.wikipedia.org/wiki/Chlamydosaurus/

P2 : http://www.nationalgeographic.com/animals/reptiles/f/frilled-lizard/

P3 : https://www.britannica.com/animal/frilled-lizard/

P4 : https://a-z-animals.com/animals/frilled-lizard/

P5 : http://easyscienceforkids.com/all-about-frilled-lizards/

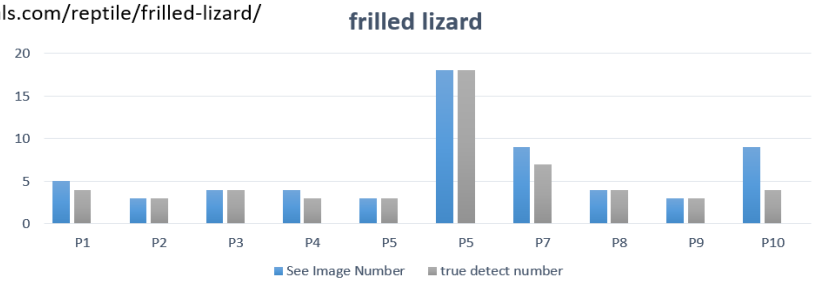
P6 : http://www.arkive.org/frilled-lizard/chlamydosaurus-kingii/

P7 : http://www.outback-australia-travel-secrets.com/frilled_lizard.html/

P8 : http://reptilepark.com.au/animals/reptiles/dragons/frilled-neck-lizard/

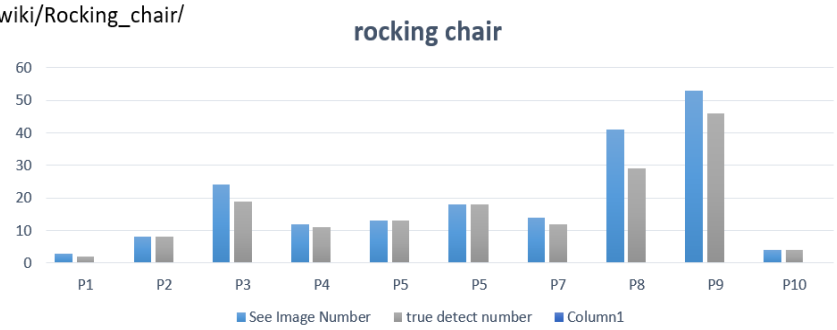
P9 : http://www.reptilesmagazine.com/Care-Sheets/Frilled-Lizard-Care-Sheet/

P10: https://www.aboutanimals.com/reptile/frilled-lizard/



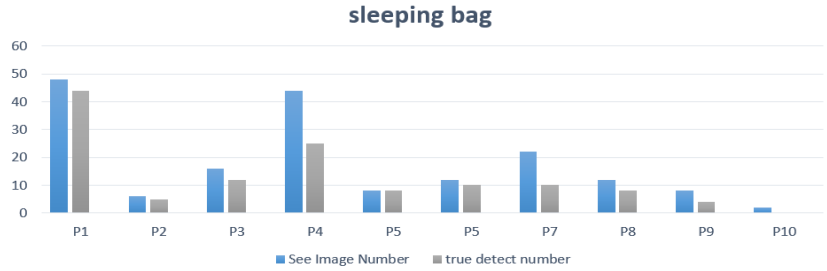
شکل ۵-۱۰ بررسی ۱۰ صفحه درباره مارمولک فرشته

P1: https://www.lowes.com/pd/Garden-Treasures-Black-Patio-Rocking-Chair/3881543/
P2 : https://www.wayfair.com/furniture/sb0/rocking-chairs-c413906.html
P3 : https://www.houzz.com/photos/rocking-chairs/
P4 : http://www.ikea.com/us/en/catalog/products/S89861010/
P5 : http://www.garyweeks.com/rocking_chairs.htm/
P6 : https://www.polywoodoutdoor.com/styles/rocking-chairs.html/
P7 : https://www.amazon.com/Chairs-Rocking-Living-Room-Furniture/s?ie=UTF8&page=1&rh=n%3A3733491%2Cp_n_feature_three_browse-bin%3A3486058011/
P8 : http://shop.crackerbarrel.com/home-furniture/indoor-furniture/rocking-chairs/
P9 : https://www.martinandmacarthur.com/koa-furniture/living-room/rocking-chairs/
P10 : http://en.wikipedia.org/wiki/Rocking_chair/



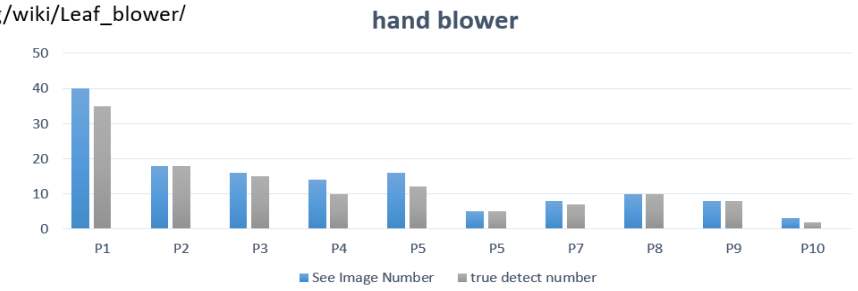
جدول ۵-۱۱ بررسی ۱۰ صفحه درباره صندلی راحتی

P1 : http://www.amazon.com/camping-sleeping-bags/b?ie=UTF8&node=2204498011/
P2 : https://en.wikipedia.org/wiki/Sleeping_bag/
P3 : https://www.thenorthface.com/shop/equipment-sleeping-bags/
P4 : https://www.rei.com/c/sleeping-bags-and-accessories?r=c&origin=web&ir=category%3Asleeping-bags-and-accessories&page=1/
P5 : http://www.sportsdirect.com/camping/sleeping-bags/
P6 : http://www.gooutdoors.co.uk/camping/sleeping-bags/
P7 : http://www.huskyeu.eu/sleeping-bags/
P8 : http://featheredfriends.com/down-sleeping-bags/three-season.html/
P9 : http://sleepingbags-cumulus.eu/uk/categories/sleeping-bags/down-sleeping-bags/
P10: http://haeld.com/



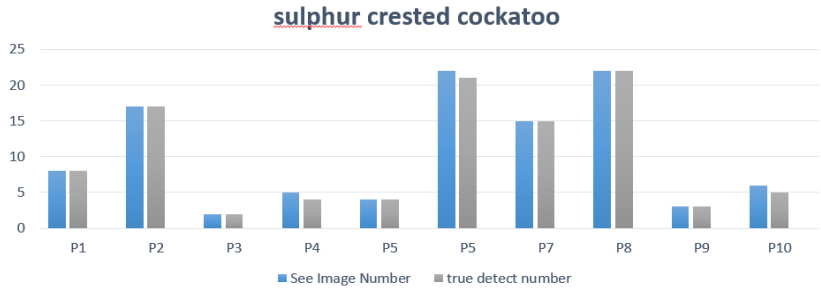
شکل ۵-۱۲ بررسی ۱۰ صفحه درباره کیسه خواب

P1: https://www.homedepot.ca/en/home/categories/outdoors/outdoor-power-equipment/leaf-blowers.html/
P2 : http://dir.indiamart.com/impcat/hand-blower.html/
P3 : http://www.popularmechanics.com/home/tools/reviews/g168/which-leaf-blower-is-the-best-we-test-8-to-find-out/
P4 : http://www.popularmechanics.com/home/tools/g2195/best-battery-powered-leaf-blowers/
P5 : http://www.husqvarna.com/us/products/leaf-blowers/
P6 : https://www.stihlusa.com/products/blowers-and-shredder-vacs/
P7 : https://www.stihlusa.com/products/blowers-and-shredder-vacs/homeowner-blowers/
P8 : https://www.amazon.in/ELECTRIC-HAND-BLOWER-EBC-ELECTREX/dp/B00RXQGGJI/259-5963314-6545724?_encoding=UTF8&%2AVersion%2A=1&%2Aentries%2A=0&portal-device-attributes=desktop/
P9 : https://www.consumerreports.org/cro/leaf-blowers/buying-guide/
P10 : http://en.wikipedia.org/wiki/Leaf_blower/



شکل ۵-۱۳ بررسی ۱۰ صفحه درباره دمنده دستی

P1 : http://www.rollandbirdsfamily.com/available-parrots.html?gclid=Cj0KCQjwiqTNBRDVARIsAGsd9MqruUUUAn2PqYNGx22Z4IKNFYj4n0VHvUY_XBVBT5MbAXgvZPA-5jsaAjSOEALw_wcB/
P2 : http://www.thespruce.com/sulphur-crested-cockatoos-390714/
P3 : http://www.birdsinbackyards.net/species/Cacatua-galerita/
P4 : http://animaldiversity.org/accounts/Cacatua_galerita/
P5 : http://cockatoo-info.com/species/sulphur-crested-cockatoo/
P6 : http://www.bizrate.com/cockatoo-crested-sulphur/
P7 : http://www.oiseaux-birds.com/card-sulphur-crested-cockatoo.html/
P8 : http://nzbirdsonline.org.nz/species/sulphur-crested-cockatoo/
P9 : http://www.billabongsanctuary.com.au/sulphu/
P10: http://en.wikipedia.org/wiki/Sulphur-crested_cockatoo/



شکل ۵-۱۴ بررسی ۱۰ صفحه درباره نوعی طوطی سفید

P1 : https://en.wikipedia.org/wiki/Dumbbell/

P2 : http://www.fitness-superstore.co.uk/gym-equipment/dumbbells-kettlebells/dumbbells/

P3 : https://www.amazon.com/dumbbells-dumbbell-set/b?ie=UTF8&node=3408401/

P4 : https://dumbbell-exercises.com/

P5 : http://anilhasehalteresmusculacao.com.br/dumbbells/

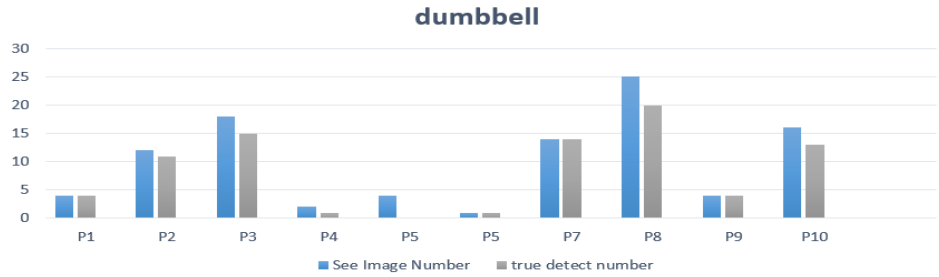
P6 : https://en.wiktionary.org/wiki/dumbbell/

P7 : https://www.roguefitness.com/conditioning/strength-equipment/dumbbells/

P8 : http://www.sportsdirect.com/fitness-and-exercise-equipment/gym-equipment/weights-and-dumbbells/

P9 : https://www.amazon.in/Dumbbells/b?ie=UTF8&node=3404717031/

P10: https://www.flipkart.com/sports-fitness/exercise-fitness/dumbbells/pr?sid=dep,xnh,f92/



شکل ۵-۱۵ بررسی ۱۰ صفحه درباره دَمبِل

P1 : http://www.mingnuoclean.com/en/

P2 : http://en.wikipedia.org/wiki/Garbage_truck/

P3 : http://www.mingnuoclean.com/en/

P4 : http://en.wiktionary.org/wiki/garbage_truck/

P5 : http://www.hyundai.com/worldwide/en/special-cv/garbage-truck/roll-pack/

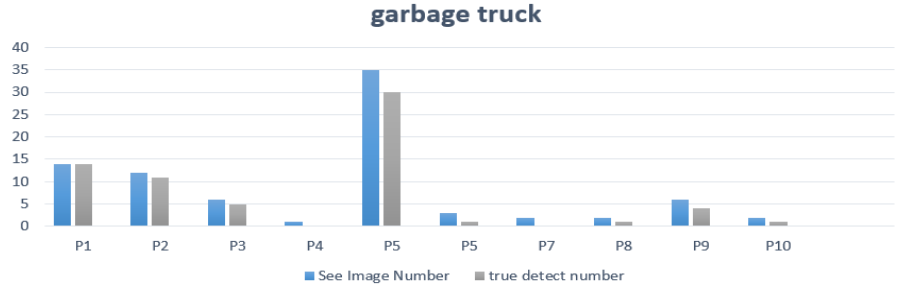
P6 : http://www.trackyourtruck.com/fleet-tracking-systems/industry/garbage-trucks/

P7 : http://www.presstelegram.com/general-news/20170830/firefighters-put-out-garbage-truck-blaze-at-long-beach-va/

P8 : http://shop.lego.com/en-LU/Garbage-Truck-10519/

P9 : http://www.spoonflower.com/tags/garbage%20truck/

P10: http://www.pepperl-fuchs.us/usa/en/12393.htm/



شکل ۵-۱۶ بررسی ۱۰ صفحه درباره ماشین حمل زباله

P1 : https://www.apple.com/lae/ipod-touch/

P2 : https://www.apple.com/shop/buy-ipod/ipod-touch/

P3 : https://en.wikipedia.org/wiki/IPod_Touch/

P4 : https://www.theguardian.com/technology/askjack/2017/aug/10/apple-discontinuing-ipods-alternative-mp3-player-podcasts-audible-audio-books/

P5 : http://www.argos.co.uk/static/Browse/ID72/33012867/c_1/1%7Ccategory_root%7CTechnology%7C33006169/c_2/2%7C33006169%7CiPod%2C+MP3+and+headphones%7C33008472/c_3/3%7Ccat_33008472%7CiPod%7C33012867.htm/

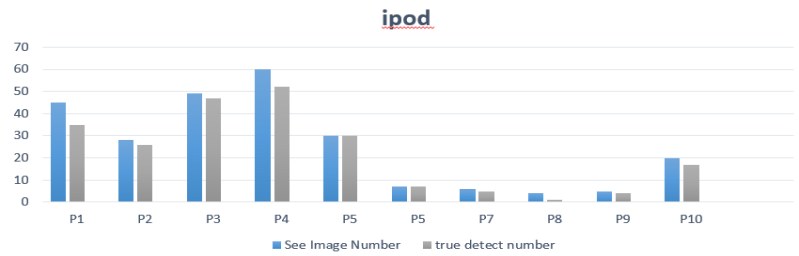
P6 : https://www.walmart.com/ip/Apple-iPod-touch-16GB/42608121/

P7 : http://www.ifixit.com/Device/iPod/

P8 : http://www.cnet.com/products/apple-ipod-classic-7th-generation/review/

P9 : http://www.businessinsider.com/apple-kills-off-ipod-nano-ipod-shuffle-2017-7/

P10: https://www.tesco.com/direct/technology-gaming/ipod/cat3376298.cat?source=others/



شکل ۵-۱۷ بررسی ۱۰ صفحه درباره ipod

فهرست مراجع لاتین

[1] B. Pinkerton - “WebCrawler: finding what people want”, University of Washington “ - 2000.

[2] A.K.Sharma, J. P. Guspta , D. P. Agarwal- “Augment Hypertext Documents suitable for parallel crawlers” - Feb2000.

[4] Gyanendra Kumar, Neelam Duahn - “ Page Ranking Based on Number of Visits of Web Pages” - 2011.

[5] Ali Mohammad Zareh Bidoki, Nasser Yazdani - “DistanceRank: An intelligent ranking algorithm for web pages” - June 2007

[6] L. Wood – “Programming the Web: the W3C DOM specification” - Jan/Feb 1999

[7] W3C. Retrieved - “HTML & CSS” - 10 Feb 2012

[8] Dave Raggett, Arnaud Le Hors, Ian Jacobs - “HTML 4.01 Specification - Abstract”. W3C Recommendation. W3C” - 10 Feb 2000.

[9] Ravindra D.Kale – “Study of web crawler and its different types” - 2014

[10] Liwei Wang , Yin Li – “Learning Deep Structure Preserving Image text Embeddings” - 2016

[11] S. Hameed – “An Optimal Algotithm for HTML page Building Process” - 2004

[12] Luis Gravano – “Query vs. Crawling based lassification of Searchable web DataBase” - 2013

[13] Yuki Endo , Satoshi lizuka – “Deep Prop : Extracting Deep Features from a Single Image For Edit Propagation” –Dec 2011

[14] Dilip Kummar Sharma – “A Comparative Analysis of Web Page Ranking Algorithms” - 2001

[15] Josip krapac – “improving web image search results using query relative classifiers” - 2015

[16] Anthony Rios, Ramakanth Kavuluru – “Convlution Neural Network for Biomedical Text Classification : Application in indexing Biomedical Articles” - 2015

[17] Springer : 978-1-4471-2502-0 – “Image Acquisition” - 2001

[18] Larry Brown – “Deep Learning For Image Classification” - 2015

[19] Aliaksei Serveryn , Alessandro Moschitti - ”Learning to Rank Short Text Pairs with Convolution Deep Neural Networks” - 2013

[20] Isha Mahajan , Sachin Gupta - “Extended Weighted Page Rank Based on VOL by Finding User Activities Time and Page Reading Time, Storing them Directly on Search Engine DataBase Server” - 2017

Abstract :

With growing information on the Web, importance of a search engine which enables to provide desired information at low cost and minimum time is vital. An internet user at least once every day uses a search engine to find the required information. Of course, finding a huge-profile list of irrelevant images and results during a search is always a source of inconvenience.

One of the most important use of web search engine is extracting existing information on the web, processes them and categorizes information and eventually displays relevant results to the search subject. The result must be reported in order of similarity and relevance to the subject of the search with a particular priority to the user.

In this thesis, first different type of search engines are reviewed. Then, we present the challenges for a web search engine. We propose a method using deep learning to extract features from images and texts from web pages to categorize the results for quicker access to desired information.

Keywords :

search engine, deep learning, image categorization, web crawler, parallel processing, image search engine,innovations

Shahrood University of Technology

Faculty of Electrical Engineering and Robotic

M.Sc. Thesis in Robotic Engineering

Identification, Classification and Learning Shapes and Images on the web Platform by a Search Engine

By: Mehrdad Salehi

Supervisor:

Dr. Alireza AhmadyFard

September 2017