
بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

پروژه‌ی کارشناسی

مهندسی کامپیوتر ـ نرم‌افزار

عنوان:

پیاده‌سازی خزنده‌ی موتور جستجوی هوشمند کسب و کار فارسی

نگارش:

بهنام حاتمی ورزنه

استاد راهنما:

دکتر حمید بیگی

شهریورماه ۱۳۹۲

دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

پروژه‌ی کارشناسی

مهندسی کامپیوتر ـ نرم‌افزار

عنوان:

پیاده‌سازی خزنده‌ی موتور جستجوی هوشمند کسب و کار فارسی

نگارش:

بهنام حاتمی ورزنه

استاد راهنما:

دکتر حمید بیگی

امضای استاد ممتحن: امضای استاد راهنما: نمره:

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل ۱

پیش گفتار

با افزایش روز افزون حجم دانش ذخیره شده به صورت دیجیتالی و در قالب‌های مختلف نظیر اخبار، صفحات وب، صفحات شخصی، مقالات علمی، کتاب‌ها، تصاویر، فایل‌های صوتی و تصویری و شبکه‌های اجتماعی، فرآیند جستجو به دنبال بخش خاصی از مطالب که مدنظر است و یافتن آن، تبدیل به کاری دشوار شده است. بنابراین نیاز به داشتن ابزار محاسباتی جدید که امکان سازماندهی، جستجو و فهم این حجم انبوه از اطلاعات را بدهد، بیش از پیش حس می‌شود.

در حال حاضر، در مورد داده‌هایی که به صورت برخط ذخیره شده‌اند می‌توان از دو طریق جستجو و پیوند[۱] صفحه، به مطلب مورد نظر دست یافت. به این صورت که می‌توان مطلب مورد نظر را در قالب واژگان کلیدی در یک موتور جستجو وارد کرد و در پاسخ به آن، مجموعه‌ای از اسناد مرتبط با عبارت جستجو را دریافت کرد. اما بعضاً ممکن است شخص جستجو کننده به جای جستجو به دنبال یک سند خاص، به دنبال مطالب در یک زمینه‌ی موضوعی خاص و ارتباط آن‌ها با یکدیگر باشد. در این صورت لازم است تا شخص جستجو کننده قبل از جستجو با استفاده از واژگان کلیدی، ابتدا زمینه را پیدا کرده و مطالب مرتبط با آن را مطالعه کند. این زمینه و اسناد مرتبط با آن ممکن است در گذر زمان نیز تغییر کنند. بنابراین استفاده از ساختار معنایی اسناد و طبقه‌بندی آن‌ها با استفاده از این ساختار، روشی دیگر برای کاوش و استفاده از اسناد است.

در بسیاری از مجموعه‌های اسناد، به دلیل حجم بالای مطالب، نمی‌توان به طور کامل از قوای انسانی

---

[۱] link

برای خواندن همه‌ی اسناد و پیدا کردن ساختار معنایی آن‌ها و جستجو به دنبال سایر اسناد مرتبط با استفاده از واژگان کلیدی استفاده کرد. به همین منظور روش مدل‌سازی موضوع[۲] که مبتنی بر پردازش زبان طبیعی با استفاده از یادگیری ماشین است، به همراه جمع آوری و استخراج خودکار اطلاعات معرفی شده است.

روش مدل‌سازی موضوع، یک مدل آماری برای یافتن عناوین استفاده شده در یک مجموعه با حجم بالا از اسناد، با استفاده از اطلاعات معنایی و ساختار معنایی نهان اسناد است. فرض اصلی روش‌های مدل‌سازی عناوین، تشکیل شدن هر سند از تعداد اندکی از عناوین است که در آن هر عنوان، دارای توزیع مشخص و مرتبط با موضوع از کلمات است. بنابراین کلماتی که در توزیع احتمال مربوط به هر عنوان به کار رفته در سند، با احتمال بالایی حضور داشته باشند، با احتمال بالایی جزء کلمات تشکیل‌دهنده‌ی سند نیز می‌باشند. بنابراین با استفاده از روش‌های آماری و به صورت مشابه، این الگوریتم‌ها، کلمات استفاده شده در متن را به منظور یافتن زمینه‌های معنایی اصلی به کار رفته در متن و همچنین یافتن ارتباط این زمینه‌ها و تغییرات آن در گذر زمان، بررسی می‌کنند.

[۲] topic modeling

فصل ۲

معرفی مسئله

۲-۱ تعریف دقیق مسأله

هدف از انجام این پژوهش، پیاده سازی موتورهای جستجوی هوشمند کسب و کار فارسی است.

روزانه حجم بالایی از آگهی‌های استخدام، در فضای برخط و در قالب صفحات وب و یا صفحات شخصی افراد، منتشر می‌شوند. از طرفی تعداد این صفحات بسیار زیاد است و به روز رسانی صفحات معمولاً از سرعت بالایی (تقریباً هر روز) برخوردار است. از طرفی دیگر، اغلب این صفحات، آگهی‌هایی در همه‌ی زمینه‌های موضوعی و شغلی و همچنین شرایط مکانی نظیر شهر و استان محل کار را پوشش می‌دهند.

در حال حاضر، در چند مورد از سایت‌های فارسی که در زمینه‌ی استخدام فعالیت می‌کنند، امکان دسته بندی مطالب بر حسب نوع آگهی وجود دارد، اما این دسته‌بندی توسط انسان و بدون استفاده از روش‌های یادگیری انجام می‌شود و در بسیاری از موارد متأسفانه دسته‌بندی موجود، چندان کامل نیست. همچنین امکان جستجو اغلب به صورت جستجوی متنی در این سایت‌ها وجود دارد و امکان جستجو با توجه به مواردی همچون جنسیت فرد، نوع شغل و موقعیت مکانی آن وجود ندارد. همچنین تعامل آن‌ها با افراد با استفاده از روش‌هایی مانند عضویت و یا ارسال نظر و در مواردی اندک، ارسال رزومه است. اما در روش‌های تعاملی و گزینش خبرهای مرتبط با افراد نیز متأسفانه از

روش‌های هوشمند استفاده نمی‌شود و این کار با استفاده از نیروی انسانی صورت می‌گیرد.
با توجه به ویژگی‌های مطرح شده برای این صفحات وب، مشاهده و جستجوی روزانه در میان حجم انبوه اخبار و آگهی‌ها، بدون استفاده از روش‌های هوش مصنوعی و تنها با استفاده از نیروی انسانی هم برای یافتن افراد متناسب با شغل و گزینش با توجه به توانایی آن‌ها و هم برای فرد متقاضی، کاری بسیار دشوار است. بنابراین می‌توان از الگوریتم‌های یادگیری در قسمت دریافت اخبار و پیمایش صفحات وب و همچنین تعامل با متقاضی و همچنین دسته‌بندی آگهی‌ها و اخبار استفاده کرد.

در این پژوهش، با استفاده از تکنولوژی‌های موجود برای بازیابی و اختصاصی سازی آن‌ها، از سایت‌هایی که در زمینه‌ی استخدام فعالیت دارند، آگهی‌های استخدام استخراج می‌شود و با پردازش هرکدام، اطلاعات تخصصی مورد نیاز جداسازی شده و سپس شاخص بندی و برای اجرای انواع پرسمان‌ها توسط تکنولوژی‌های موجود آماده می‌گردد.

در این پژوهش، برای بازیابی اطلاعات از نرم افزار متن باز[۱] Nutch[۲]، استفاده می‌شود. این نرم افزار، دارای امکانات و ویژگی‌های خاص خود می‌باشد و امکان پیکربندی بالایی دارد. سپس اطلاعات بازیابی شده را به وسیله‌ی Lucene[۳]، شاخص بندی می‌کنیم. این نرم افزار یکی از بهترین نرم افزارهای متن باز در این زمینه است. سپس با استفاده از Solr[۴]، امکان پرسمان بر روی اطلاعات استخراج شده فراهم خواهیم کرد. سپس اطلاعات به دست آمده به پژوهش مکمل برای پردازش‌های بعدی داده می‌شود.

در پژوهش مکمل، از الگوریتم‌های یادگیری برای هوشمند کردن دسته‌بندی آگهی‌ها و اخبار استفاده می‌شود. این سامانه‌ی هوشمند، از اطلاعات پیمایش شده‌ی صفحات وب استفاده می‌کند، بنابراین ورودی مسئله تعدادی از آگهی‌های فارسی است. هدف دسته‌بندی آگهی‌ها بر اساس موضوع آن‌هاست، به گونه‌ای که هر آگهی بتواند در یک یا چند دسته با موضوع مرتبط با خود قرار بگیرد. این مسئله همانند مسئله‌ی مدل‌سازی عناوین است. به این صورت که تعدادی سند (در قالب آگهی) در اختیار داشته و هدف نهایی قرار دادن این اسناد در یک یا چند دسته و بدست آوردن این دسته‌هاست. بنابراین از دو الگوریتم LDA و PLSA که در ادامه شرح داده خواهد شد، برای

[۱] Open source
[۲] nutch.apache.org
[۳] lucene.apache.org
[۴] lucene.apache.org/solr

حل این مسئله استفاده می‌شود. البته باید توجه کرد که در مدل‌سازی عناوین، تاپیک‌ها به صورت هوشمند نام‌گذاری نمی‌شوند.
بنابراین نام‌گذاری مناسب دسته‌ها جزئی از راه حل مسئله محسوب می‌شود. در نهایت خروجی این مسئله، تعدادی موضوع با عناوینی همچون «استخدام بانک‌ها»، «استخدام نیروی انتظامی» و یا به تفکیک مکانی مانند «استخدام استان تهران» و «استخدام استان اصفهان» و همچنین اسناد مرتبط با هر یک از موضوعات می‌باشد.

۲-۲ کارهای مشابه

در زمینه‌ی کسب و کار هوشمند آنلاین، در زبان‌های دیگر کارهای مشابهی انجام شده است که از جمله آن‌ها می‌توان به صفحه‌ی وب Texkernel[۵] اشاره کرد. این سایت از ۶ قسمت اصلی تشکیل شده است که به صورت مجتمع در کنار یکدیگر قرار گرفته‌اند و از هر یک از این سرویس‌ها می‌توان به صورت جداگانه استفاده کرد. در زیر به اختصار به هر یک از این سرویس‌ها و ویژگی‌های آن‌ها اشاره می‌کنیم:

- قسمت استخراج که قسمت‌های مختلف رزومه را به صورت خودکار از روی کارنامک[۶] و یا صفحه‌ی کاربر در رسانه‌های اجتماعی و تکمیل پروفایل کاربر به صورت اتوماتیک استخراج می‌کند.
- قسمت منابع که به صورت اتوماتیک کارنامک و اطلاعات فرد در شبکه‌های اجتماعی را جدا کرده و به صورت گرافیکی در کنار رزومه‌ی اصلی فرد قرار می‌دهد و به کاربر امکان ویرایش و اضافه یا حذف اطلاعات از کارنامک خود در پایگاه داده‌ی سایت را می‌دهد. پس از این مرحله اطلاعات فرد در پایگاه داده‌ی صفحه ذخیره می‌شود تا در مراحل بعدی مورد استفاده قرار گیرد.
- قسمت جستجو امکان جستجو در میان رزومه‌های موجود در پایگاه داده برای یافتن افراد مرتبط با هر شغل و رتبه‌بندی آن‌ها را می‌دهد.

[۵] www.texkernel.com
[۶] Curriculum vitae (CV)

- قسمت خوراک شغل[۷] که به صورت خودکار به صورت روزانه در سایت‌های کسب و کار جستجو می‌کند و آگهی‌های جدید را پردازش کرده و قسمت‌های مورد نیاز را از آن استخراج می‌کند.

- قسمت وصل کردن که متن آگهی کار را دریافت کرده و به صورت خودکار، افراد متناسب با آن شغل بر روی پایگاه داده‌ها جستجو و به صورت فهرست بدست می‌آیند.

- قسمت برداشت که به صورت خودکار، شغل‌های متناسب با توانایی و شرایط کاربر که بر روی خوراک شغل قرار دارد را به او نشان می‌دهد.

هر یک از این بخش‌ها به صورت جداگانه قابل دسترسی و استفاده در صفحه مورد نظر هستند. اما متأسفانه هیچ یک از این بخش‌ها از زبان فارسی پشتیبانی نمی‌کند.
کار انجام شده در این پژوهش مشابه بخش خوراک شغل است و اطلاعات مورد نیاز را از آگهی‌های فارسی استخراج می‌کند.
از ویژگی‌های اصلی قسمت خوراک شغل سایت textkernel می‌توان به موارد زیر اشاره کرد:

- مقایسه هر آگهی با آگهی‌های دریافت شده در ۶ ماه اخیر و تشخیص شغل‌های یکتا و رفتار کارفرماها

- به روز رسانی و بررسی وضعیت شغل‌ها از نظر باز یا بسته بودن و همچنین ظرفیت باقیمانده از شغل به صورت روزانه

- داشتن پیوند به صفحه‌ی فرد در شبکه‌ی اجتماعی Linkedin[۸]

قابل ذکر است، که نوع کارنامک فرد، باید ساختاری مشابه ساختار Linkedin باشد، به همین علت، بررسی دقیق Linkedin برای این پروژه نیاز است.

---

[۷] JobFeed

[۸] www.linkedin.com

فصل ۳

موتورهای جستجو

با توجه به آمار جهانی اینترنت، در تاریخ ۳۱ام مارچ ۲۰۰۸، ۱/۴۰۷ میلیارد انسان، از اینترنت استفاده می‌نمایند. میزان نفوذ اینترنت به طور روز افزون در حال افزایش است. شبکه جهانی گسترده وب[۱] (که معمولاً به اختصار وب نامیده می‌شود)، یک سیستم از اسناد ابرمتن[۲] به هم متصل است که به وسیله‌ی اینترنت قابل دسترسی هستند. با استفاده از یک مرورگر، کاربر امکان مشاهده‌ی صفحات وب که دارای محتوای داده‌ای، عکس، فیلم و سایر امکانات چند رسانه‌ای است را دارد و می‌تواند توسط لینک‌ها، بین آن‌ها جابه‌جا گردد.

همان گونه که تعداد صفحات وب، به طور روزافزون در حال افزایش است، نیاز به موتور جستجو بیشتر احساس می‌گردد. در این فصل، ما توضیح مختصری در مورد المان‌های پایه‌ی هر سیستم جستجویی به همراه نحوه‌ی عملکرد آن المان را مورد بررسی قرار می‌دهیم. سپس، نقش خزنده‌های وب[۳]، که یکی از اصلی‌ترین بخش‌های اصلی هر سیستم جستجوی اینترنتی می‌باشد را مورد بررسی قرار خواهیم داد.

---

[۱] World Wide Web

[۲] Hyper text documents

[۳] Web crawlers

۳-۱ موتور جستجوی وب

محتوای بسیاری از شبکه جهانی گسترده‌ی وب، قابل استفاده برای میلیون‌ها نفر است. بسیاری از افراد، دسترسی به صفحات وب را از نقاط آغازی مانند، Yahoo[۴] و MSN[۵] و... آغاز می‌نمایند. اما بسیار از افراد نیازمند اطلاعات، برای شروع فعالیت اینترنتی خود از موتورهای جستجو آغاز می‌نمایند. در این حالت، کاربر یک پرسمان[۶] ارسال می‌نماید، که معمولاً به صورت لیستی از کلیدواژه‌ها[۷] است و در پاسخ، لیستی از صفحات وب که احتمالاً مرتبط با درخواست کاربر بوده (معمولاً صفحاتی که دارای آن کلیدواژه‌ها بوده است) را دریافت می‌کند. در زمینه‌ی وب، موتورهای جستجو، در واقع به جستجوگرهایی گفته می‌شود، که در یک پایگاه داده‌ای[۸] از فایل‌های وب، جستجوی خود را انجام می‌دهد.

۳-۲ انواع موتورهای جستجو

به طور کلی، سه نوع موتور جستجو وجود دارد:

- موتورهای جستجویی که به وسیله‌ی ربات‌ها اجرا می‌شوند (معمولاً به خزنده‌ها، مورچه‌ها[۹] یا عنکبوت‌ها[۱۰] معروفند).
- موتورهای جستجویی که بر اساس ارسال‌های کاربران اجرا می‌شوند.
- موتورهای جستجویی که بر اساس تلفیق دو نوع بالا به دست می‌آید.

دو نوع اصلی موتورهای جستجو در زیر به اختصار توضیح داده شده است:

---

[۴] www.yahoo.com
[۵] www.msn.com
[۶] Query
[۷] Keywords
[۸] Database
[۹] Ants
[۱۰] Spiders

۳-۲-۱ موتورهای جستجو مبتنی بر خزنده‌ها

چنین موتورهای جستجویی، از تعدادی عامل‌های[۱۱] نرم افزاری خودکار (که خزنده نامیده می‌شود) تشکیل شده است. این خزنده‌ها، صفحات وب را دریافت، اطلاعات و ابر تگ‌های[۱۲] آن را استخراج می‌کنند. همچنین برای دسترسی به تمام صفحات یک وب سایت و شاخص بندی[۱۳] آن‌ها، لینک‌های داخل صفحات را دنبال می‌کند. خزنده، تمام اطلاعات استخراج شده را، در یک مخزن مرکزی[۱۴] ذخیره می‌نماید. سپس داده‌ها در مخزن شاخص بندی می‌گردد. خزنده همچنین به طور متناوب به صفحات بازیابی شده مراجعه می‌نماید و در صورت تغییر اطلاعات خود را به روز رسانی می‌نماید. تناوب چنین کاری توسط مدیر سیستم، تنظیم می‌گردد.

۳-۲-۲ موتورهای جستجو مبتنی بر انسان

چنین موتورهای جستجویی، مبتنی است بر داده‌هایی که به مرور زمان به وسیله‌ی انسان، به سیستم ارسال می‌شود، شاخص بندی می‌گردد و دسته بندی[۱۵] می‌گردد. در این نوع موتور جستجوها، تنها داده‌هایی که ارسال شده است، در شاخص‌ها ذخیره می‌شود. چنین موتورهای جستجویی به ندرت در مقیاس بزرگ مورد استفاده می‌گردد، اما در سازمان‌هایی که با داده‌های با مقیاس کوچک روبرو هستند، بسیار پراستفاده است.

۳-۳ ساختار و نحوه‌ی کار موتورهای جستجو

ساختار پایه‌ی هر موتور جستجویی مبتنی بر خزنده، در شکل ۳-۱ نشان داده شده است. از این رو، فازهای اصلی هر موتور جستجویی عبارتند از:

[۱۱] Agents
[۱۲] Metatags
[۱۳] Indexing
[۱۴] Central Repository
[۱۵] Classified

۳-۳.۱ جمع آوری اطلاعات یا خزش

هر موتور جستجویی که بر پایه‌ی یک خزنده کار می‌کند، منابع اطلاعاتی خود را برای ارائه‌ی خدمات تأمین می‌کند. خزنده‌ها، نرم افزارهای کوچکی هستند که از طریق موتورهای جستجو به سایت‌ها سر می‌زنند، دقیقاً به همان روشی که انسان‌ها لینک‌های بین صفحات را دنبال می‌کنند. معمولاً در ابتدا، یک لیست ابتدایی از آدرس وب سایت‌ها به هر خزنده داده می‌شود. خزنده باید صفحه‌ی مربوط به هرکدام را دریافت نماید. پس از آن، لینک‌های داخل این صفحات بازیابی شده را استخراج نماید و اطلاعات استخراج شده را به واحد کنترل خزنده تحویل دهد. این واحد تصمیم می‌گیرد که چه لینک‌هایی در ادامه بازیابی گردد و لیست آن‌ها را برای خزنده ارسال می‌نماید. مراحل بیان شده را می‌توانید در شکل ۳-۲ ببینید.

۳-۳.۲ نگه داری پایگاه داده یا مخزن

همان طور که در شکل ۳-۱ می‌بینید، تمام داده‌های یک موتور جستجو، در یک پایگاه داده ذخیره می‌شود و تمام جستجوها و عملیات داده‌ای، به کمک این پایگاه داده انجام می‌پذیرد. این پایگاه داده نیاز دارد در طول زمان با توجه به تغییرهای بیرونی بروز رسانی گردد. در مرحله‌ی بازیابی و پس از اتمام مرحله دریافت اطلاعات به وسیله‌ی خزنده، موتور جستجو باید تمام اطلاعات جدید و مفید صفحات بازیابی شده را استخراج و در پایگاه داده ذخیره نماید. در بعضی از موتورهای جستجو، یک مخزن از صفحات ذخیره شده به صورت موقت بین این دو مرحله قرار می‌گیرد. حتی بعضی مواقع، موتورهای جستجو، یک حافظه‌ی سریع نهان[۱۶] از صفحاتی که بازیابی شده‌اند، نگه می‌دارد تا بتواند مرحله‌ی شاخص بندی را سریع‌تر انجام دهد و همچنین امکان جستجوی ابتدایی بر روی داده‌های دریافت شده را فراهم آورد.

۳-۳.۳ شاخص بندی

زمانی که صفحه‌ی بازیابی شده، در مخزن ذخیره می‌شود، کار بعدی موتور جستجو، ایجاد یک شاخص برای داده‌های ذخیره شده می‌باشد. واحد شاخص بندی، تمام کلمات را از هر صفحه

[۱۶] Cache

استخراج می‌نماید و آدرس صفحه‌ی مدنظر را به ازای هر کلمه‌ی استخراج شده، ذخیره می‌نماید. نتیجه کار، معمولاً یک لغت نامه‌ی بزرگ می‌باشد که می‌تواند آدرس تمام صفحه‌هایی را که در آن‌ها کلمه‌ی خاصی آمده‌اند را به ما بدهد. به وضوح صفحات به صفحاتی محدود می‌شود که در فاز قبلی بازیابی شده‌اند. همان طور که قبلاً ذکر شده بود، شاخص بندی متن، مشکلات و چالش‌های خاص خودش را دارد. از جمله‌ی آن می‌توان به سایز بزرگ آن و سرعت زیاد تغییرات در آن اشاره نمود. همچنین علاوه بر چالش‌های فوق‌الذکر، جستجو برای شاخص‌های نادر و کمتر رایج نیز خود چالش زا است. به طور مثال، واحد شاخص بندی، می‌تواند یک شاخص ساختاری از اتصالات بین صفحات تولید نماید.

۳-۴.۳ پرسمان

این واحد با پرسمان‌های کاربر سروکار دارد. واحد پرسمان، مسئول دریافت و پاسخ گویی به درخواست‌های جستجو از طرف کاربران می‌باشد. این واحد به صورت اساسی وابسته به شاخص‌های موجود و بعضی مواقع به مخزن صفحات ذخیره می‌باشد. به علت حجم زیاد وب، و وارد شدن عبارات جستجوی کوتاه به وسیله کاربران در حد یک یا دو کلیدواژه، مجموعه جواب موجود، بسیار زیاد می‌باشد.

۳-۵.۳ رتبه دهی

به علت اینکه مجموعه سندهای مرتبط با پرسمان وارد شده‌ی کاربر، بسیار زیاد است، یکی از مهم‌ترین وظایف موتورهای جستجو نمایش مرتبط‌ترین نتایج به کاربر است. برای اجرای کارآمد چنین امری، نتایج رتبه دهی می‌گردند. واحد رتبه دهی، به همین منظور وظیفه‌ی مرتب کردن نتایج را به گونه ای دارد که نتایج بالاتر احتمال بیشتری داشته باشند که همان اسنادی که کاربر به دنبال آن است باشند.

پس از پیدا کردن نتایج، به وسیله‌ی واحد رتبه دهی به هر یک از نتایج رتبه اختصاص داده شد، نتایج نهایی جستجو به کاربر نشان داده می‌شود. این روشی است که تقریباً تمام موتورهای جستجو مطابق آن کار می‌کنند.

۳-۴ نمونه‌ی موتورهای جستجو

تعدادی موتور جستجو در حال حاضر قابل استفاده است. در زیر لیستی از مهم‌ترین و مشهورترین موتورهای جستجو آورده شده است:

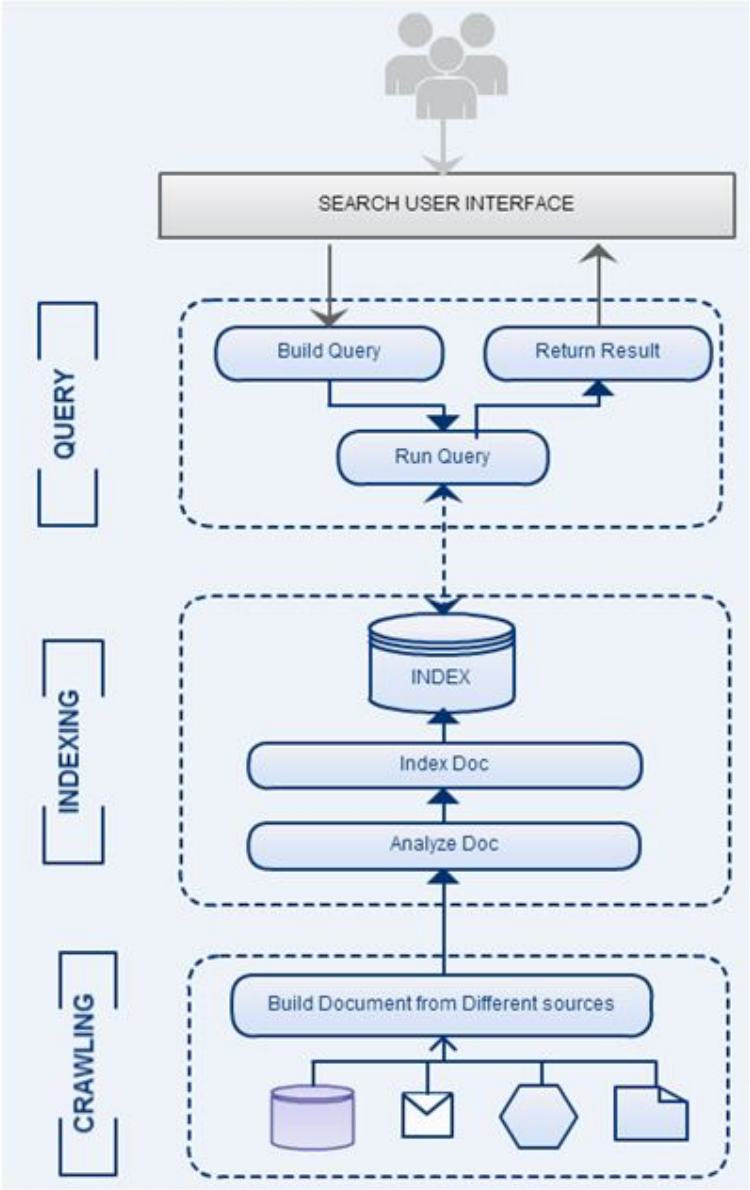
- Google[۱۷]
- Yahoo
- MSN
- E-Bay[۱۸]
- AOl[۱۹]

و تعداد بسیار زیادی موتور جستجوی دیگر در دسترس هست که کاربران را برای رسیدن به اطلاعات مدنظر یاری می‌نماید.

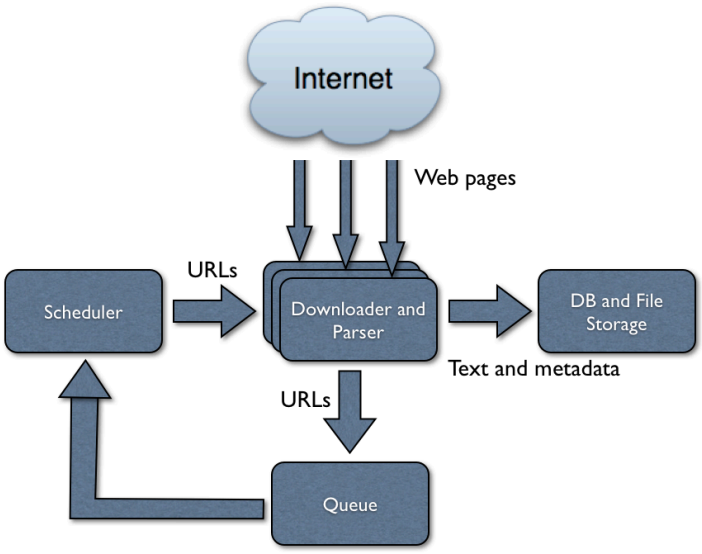
۳-۵ خلاصه‌ی فصل

موتورهای جستجو، به عنوان کلید اصلی ورود به جهان گسترده وب است. تکامل و اجزای موتورهای جستجو قسمتی مهمی از مطالعه‌ی جهان گسترده‌ی وب هستند. قسمت‌های ضروری موتور جستجو، عبارتند از خزنده، استخراج کننده، برنامه ریز و پایگاه داده. بعضی از مهم‌ترین موتورهای جستجوی پرکاربرد عبارتند از Google و MSN و ... .

[۱۷] www.google.com
[۱۸] www.ebay.com
[۱۹] www.aol.com



شکل ۳-۱: ساختار و نحوه‌ی کار یک موتور جستجو.



شکل ۳-۲: نحوه‌ی کار خزنده.

فصل ۴

موتور جستجوی Nutch

۴-۱ مقدمه

پروژه‌ی Nutch یک نرم افزار متن باز بر اساس زبان جاوا است، که امکان خزش و جمع آوری طیف مختلفی از داده‌ها را، از یک شبکه داخلی، بخشی از اینترنت یا کل جهان گسترده وب را دارد. به طور کلی قبل از پیاده سازی Nutch امکان تحلیل نتایج موتورهای جستجوی معروف، در برابر پرسمانی دلخواه، وجود نداشت و نتیجه‌ی جستجوی آن‌ها و نحوه‌ی رتبه بندی آن‌ها، با معیاری مناسب و منصفانه قابل مقایسه نبود. یکی از دلایل این امر، وجود الگوریتم جستجوی اختصاصی و منابع بسته[۱] در این شرکت‌ها می‌بودند. البته دلیل چنین کاری، علاوه بر تمایل انحصار طلبی و رقابت، جلوگیری از سوء استفاده‌ی منتشر کنندگان هرزنامه‌ها، در بالا بردن رتبه‌ی یک دامنه خاص بود. پروژه‌ی Nutch با متن باز بودن خود، سعی در برطرف کردن این معضل نمود. یکی از اهداف این موتور جستجو، ایجاد شفافیت و افزودن جزئیات به نحوه‌ی رتبه بندی صفحات وب بود. علاوه بر آن، ارائه‌ی یک موتور جستجوی جایگزین، برای افرادی که به وسیله‌ی محدود موتورهای جستجوی تجاری موجود امروزی، راضی نشده‌اند. همچنین ربات نظاره گر Nutch، این امکان را به مدیران سایت‌ها می‌دهد، که قسمت‌هایی از سایتشان که به وسیله‌ی این روش جمع آوری می‌گردد را تحت

[۱] Closed Source

مدیریت و کنترل خود داشته باشند.
ساختار پروژه‌ی Nutch به گونه‌ای طراحی شده است که، هم از لحاظ حجم جستجو و هم از لحاظ سرعت، قابل گسترش و بهبود مي باشد. به همین منظور، استفاده از روش‌های توازي در بازیابی اطلاعات، در پیاده سازی آن لحاظ شده است. بخش‌های اصلي Nutch شامل سه قسمت اصلی واحد خزنده، واحد شاخص بندی و واسط جستجو بر روی داده‌های است. خزنده‌ی Nutch به گونه‌ای طراحی شده که بر روی هر شبکه‌ی داخلی یا خارجی کار می‌کند. اطلاعات بازیابی شده توسط این قسمت، در یک پایگاه داده به نام WebDb برای استفاده‌های آتی ذخیره می‌شوند. خزنده، علاوه بر بازیابي و ذخیره سازی، با استفاده از نرم افزاری به نام Lucene برای شاخص بندی اطلاعات بازیابی شده، استفاده می‌کند. از شاخص‌های به دست آمده، برای بازیابی اطلاعات به وسیله‌ی واسط جستجو استفاده خواهد شد.
ویژگی بارز Nutch در برابر موتورهای جستجوی موجود، ساختار قابل گسترش آن می‌باشد. به طور مثال، Nutch برای زمانی که نیاز به بازیابی یک یا چند دامنه‌ی خاص قابل استفاده است، تا زمانی که بعضی از اطلاعات یک دامنه را می‌خواهیم از صافی عبور دهیم، قابل استفاده است. Nutch با استفاده از ساختار افزونه‌ای که توسط زبان نشانه گذاری[۲] به سیستم شناخته می‌شود امکان، چنین کاری را فراهم می‌آورد. چنین ساختاری که همانند ساختار Eclipse می‌باشد، این امکان را ایجاد می‌کند تا بتوان بدون تغییری بنیادی در کد، به اصلاح رفتار برنامه اقدام نمود.

۴-۲ نحوه‌ی عملکرد

۴-۲-۱ عملکرد کلی

طبق شکل ۴-۱، اجزای خزنده، شامل WebDb لیست واکشی، بازیاب‌ها و به‌روزرسان ها است. WebDb یک پایگاه داده‌ی اختصاصی شده است، که صفحات بازیابی شده به همراه لینک‌های ورودي و خروجي آن را در خود ذخیره می‌کند. همچنین در مورد هر صفحه، مجموعه‌ی کوچکی از اطلاعات، مانند آخرین زمان بازیابی را ذخیره می‌کند. لیست واکشی، با استفاده از اطلاعات WebDb تهیه می‌شود. این لیست، شامل لینک صفحاتی است که باید در این مرحله بازیابی

[۲] Markup language (XML)

گردد. بازیاب‌ها، با استفاده از این لیست، صفحات را بازیابی نموده و WebDb را متناسب با آن به‌روزرسانی می‌نمایند. در این مرحله، تغییر صفحه نسبت به بازیابی قبل، نیز برای کنترل تناوب بازیابی ذخیره می‌گردد. محتوای به دست آمده نیز برای جستجو استفاده می‌شود. این چرخه، به گونه ای طراحی شده است تا بتواند تا ابد اجرا گردد و همواره تصویری به روز از صفحات وب را ارائه دهد.

زمانی که صفحات وب، بازیابی گردید، Nutch امکان جستجو را با استفاده از Searcher خود فراهم می‌کند. در واقع پس از این مرحله، ابتدا واحد شاخص بندی، محتوای استخراج شده را در لیست‌های شاخص وارون[۳] ذخیره می‌نماید. هر سند، به تعدادی ناحیه‌ی شاخص بندی تقسیم می‌گردد و هر کدام در یک رویه‌ی جداگانه شاخص بندی می‌شود. در نهایت، استخری[۴] از کارگزارها[۵]، ارتباط بین کاربر و واحد جستجو را فراهم می‌کند. شکل کلی این رویه در شکل ۴-۱ آورده شده است.

۴-۲.۲ ساختار افزونه

ساختار افزونه پذیری ،Nutch کامل شبیه ساختار Eclipse می‌باشد. در واقع، Nutch یک سیستم مرکزی برای کنترل یک مجموعه از ابزارها که با یکدیگر کار می‌کنند می‌باشد، تا بتواند یک قابلیت را به آن اضافه نماید. بعد از مطالعه‌ی ساختار Eclipse و اعمال آن به ساختار افزونه‌ی Nutch به این نتیجه رسیدیم که مهم‌ترین المان‌های افزونه پذیری ،Nutch عبارتند از افزونه‌ها، نقاط گسترش پذیر و سیستم کنترل افزونه‌ها است. سیستم کنترل این امکان را به Nutch می‌دهد که بدان، کارایی اضافه شود. این کارایی به وسیله‌ی یک افزونه، به سیستم اضافه می‌گردد. هر افزونه، در واقع یک المان گذاردنی[۶] است که تعدادی نقاط گسترش را پیاده سازی می‌کند و این قابلیت‌ها، به وسیله‌ی سیستم کنترل مرکزی اجرا می‌شوند.

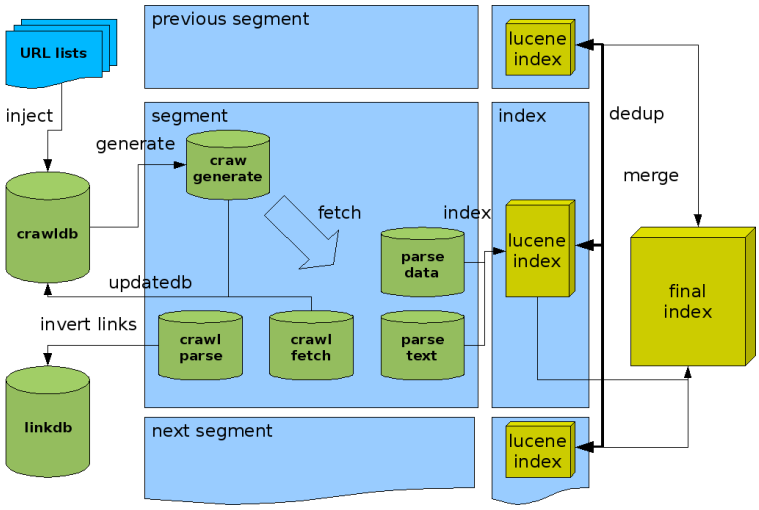
---

[۳] Inverted Index

[۴] pool

[۵] Web Servers

[۶] Pluggable



شکل ۴-۱: ساختار خزنده‌ی Nutch.

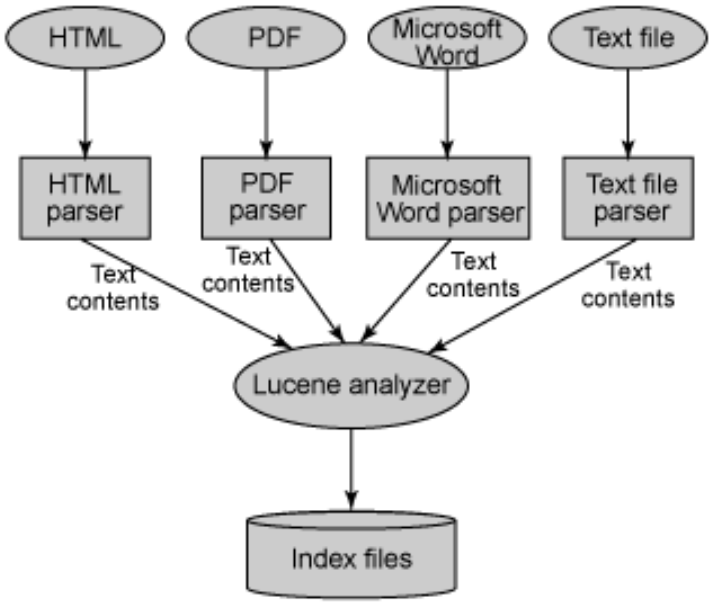
۴-۳ ساختار Lucene

۴-۳-۱ عملکرد کلی

کتاب خانه‌ی Lucene به افراد امکان افزودن امکان شاخص بندی و جستجو در نرم افزارهای خود را می‌دهد. این کتاب خانه امکان شاخص بندی و جستجو در هر نوع داده‌ای را، تا زمانی که قابلیت تبدیل شدن به متن را داشته باشد را، دارد. این بدان معنی است که این کتاب خانه امکان جستجو در صفحات وب، فایل‌های PDF و Words را دارد. به این خاطر، Lucene بهترین کتاب خانه برای نوشتن موتور جستجو است.

۴-۳-۲ شاخص بندی

شاخص بندی، در واقع مرحله‌ی تبدیل متن به شاخص می‌باشد. شاخص‌ها در واقع خود، داده ساختاری‌اند که سرعت عملیات بازیابی اطلاعات را، بهبود می‌بخشند. شاخص‌ها، جزء المان‌های اصلی Lucene می‌باشند.



شکل ۴-۲: ساختار Lucene.

برای شاخص بندی داده در این سیستم، داده باید به جویباری[۷] از تکه‌های متنی تبدیل گردند. بعد از آن، Lucene، داده‌ها را با تکه تکه کردن جویبار داده و اجرای عملیاتی بر روی آن، آن را برای شاخص بندی آماده می‌کند. به طور مثال، یکی از این عملیات می‌تواند، کوچک سازی حروف، برای از بین بردن حساسیت جستجو به بزرگی و کوچکی باشد. این مرحله به مرحله‌ی تحلیل معروف است. بعد از اینکه جویبارها، تحلیل گردیدند، داده‌ها آماده برای اضافه شدن به شاخص‌ها می‌باشند. مرحله‌ی شاخص بندی در شکل ۴-۲ متصور شده است.

این کتاب خانه، روشی نوآورانه برای نگهداری شاخص‌ها به کار می‌برد. این کتاب خانه مستقل از تعداد شاخص‌ها، برای هر شاخص از چند قطعه استفاده می‌کند. استفاده کردن از چند قطعه، اضافه کردن یک سند تازه شاخص بندی شده را به وسیله‌ی اضافه کردن آن به کوچک‌ترین قطعه که تازه ساخته شده است و ترکیت آن قطعه با بقیه‌ی قطعات به صورت متناوب، تسریع می‌بخشد. این رویه، اضافه کردن سند را بسیار کارا می‌کند، زیر شاخص‌هایی که در هر لحظه تغییر می‌کند را به شدت کاهش می‌دهد.

[۷] Stream

بعضی از سیستم‌های بازیابی، برای اضافه شدن یک سند، بیاز به بروز رسانی تمام شاخص دارند، اما Lucene به علت پشتیبانی از شاخص بندی افزایشی، چنین مشکلی را ندارد. چنین امری بدان معناست که Lucene امکان جستجو بر روی سند را بلافاصله بعد از شاخص بندی شدن آن بدون نیاز به بروز رسانی تمام شاخص، فراهم می‌سازد.

۴-۳.۳ تحلیل گر

همان طور که در قسمت قبل، بحث شد، تحلیل یکی از مهم‌ترین مراحل، شاخص بندی است. این مرحله، داده‌های متنی را به یکی از اساسی‌ترین نمایش شاخص یعنی واژه‌های بنیادی[۸]، تبدیل می‌نماید. این واژه‌های بنیادی برای تطابق پرسمان‌ها در مرحله‌ی جستجو با اسناد استفاده می‌شود.

۴-۴ خلاصه

در این فصل، ساختار Nutch به همراه مکانیزم افزونه پذیری مورد مطالعه قرار گرفت. همچنین ساختار شاخص بندی Lucene و تحلیل گر آن نیز به اختصار توضیح داده شد. در فصل آتی، پیاده سازی افزونه‌ها و تغییرات اعمال شده توضیح داده خواهد شد.

[۸] Terms

فصل ۵

پیاده سازی

۵-۱ مقدمه

در این فصل روند طي شدن پروژه توضیح داده مي شود.
در این پروژه، در ابتدا به بررسی‌های جامع در مورد وضع سایت‌های فعال کسب و کار در ایران، پرداختیم. البته بررسی‌ها به نمونه‌های داخلی محدود نگردید و بررسی‌هایی در مورد نمونه‌های خارجی موفق نیز انجام شد. با بررسی دقیق‌تر، سایت‌های هدف برای استخراج اطلاعات انتخاب شد. همچنین برای امکان ایجاد مشابه فارسی Linkedin و امکان استخراج اطلاعات طبقه بندی شده از Linkedin به بررسی شمای پایگاهی این سایت، اقدام نمودیم و تا حد خوبی، ساختار داده‌ای این سایت پرطرفدار خارجی را به دست آوردیم. سپس به مرحله‌ی پیاده سازی خزنده پرداختیم. در این مرحله، برای اجرای کار، نیاز به نوشته شدن، استخراج گر اختصاصی برای هر دامنه نیاز داشتیم. این استخراج گر، از هر صفحه‌ی آگهی داده‌های مفید آن را استخراج می‌نماید. این امر، باعث بهبود دقت جستجو در مراحل بعدی می‌گردد، زیرا داده‌های غیر مرتبط به آن آگهی را عملاً حذف می‌نماید. سپس با نوشتن افزونه برای تبدیل این داده‌های استخراج شده، به ساختاری سازگار با Lucene، امکان شاخص بندی داده‌ها را فراهم کردیم. سپس به بررسی میزان تأثیر گذاری این کار بر روی دقت نتایج جستجو پرداختیم.

۵-۲ پیش نیازها

برای سازگار سازی Nutch با سایت‌های فعال در زمینه‌ی استخدام، نیاز به پیاده سازی افزونه‌ی خاص هر سایت است. علاوه بر این، نیاز است مکانیزمی تعبیه گردد که با توجه به سایت، افزونه‌ی خاص خود را پیدا کند و آن را اجرا نماید. برای چنین کاری، از الگوی کارخانه[۱] استفاده گردید. ساختار این الگو در زیر آمده است. این الگو با توجه به نوع سایت، که از روی آدرس سایت مشخص می‌شود، در بین افزونه‌هایی که خود را در سیستم ثبت نام کرده‌اند، جستجو می‌کند و در صورت پیدا کردن مورد مناسب آن را اجرا می‌کند. چنین ساختاری، اضافه کردن یک افزونه‌ی جدید برای یک سایت جدید را بسیار راحت می‌کند و از اضافه شدن تغییرات اجتناب می‌نماید. پس از آن باید داده‌های استخراج شده را برای شاخص بندی به Lucene اضافه نماییم، که توسط افزونه شاخص بندی انجام می‌پذیرد.

۵-۳ مراحل پیاده سازی

۵-۳.۱ بررسی و انتخاب سایت‌های هدف

در این مرحله، با بررسی و جستجو در بین سایت‌های فارسی زبان که در زمینه‌ی استخدام و آگهی‌های مربوط به آن، فعالیت داشتند به دست آمد. به طور خاص، در مورد هر کدام، ویژگی‌های بارز استخراج شد. همچنین در این مرحله، به بررسی چند سایت خارجی که پیشروی در این زمینه، پرداخته شد. لیست سایت‌های بررسی شده در زیر آمده است:

۵-۳.۲ بررسی سایت LinkedIn

پس از بررسی سایت‌های فارسی، نوبت به بررسی سایت‌های موفق در زمینه‌ی کاریابی در زبان انگلیسی رسیدیم. یکی از بزرگ‌ترین شبکه های کاریابی، شبکه‌ی اجتماعی Linkedin می‌باشد. این شبکه در حال حاضر دارای ۲۰۰ میلیون کاربر فعال دارد. راه اندازی سایتی مشابه Linkedin

[۱] Factory method

جدول ۵-۱: لیست سایت‌های فعال کاریابی به زبان فارسی

آدرس سایت	نوع تعامل با کاربر	به‌روزرسانی	نوع جستجو
estekhtam.com	ارسال توضیح[۲]	هر روز	متن ساده
karyab.net	عضویت و ثبت کارنامک	هر روز	کارنامک، آگهی، نظرات
banki.ir	عضویت و ثبت رزومه	هر ۲ روز	کارجویان، شغل، تخصص و مکان
e-estekhdam.com	عضویت و ثبت رزومه	هر روز	مکان
bazarekar.ir	عضویت و ارسال کارنامک	ـ	مقطع تحصیلی، رشته، جنسیت، استان و شغل
estekhdamnews.com	عضویت و ارسال رزومه	هر روز	متن ساده
unp.ir/jobs.php	خبرنامه با ایمیل	هر روز	متن ساده
estekhdamcenter.ir	عضویت کارجویان و کارفرمایان، ثبت کارنامک و آگهی شغل	هر روز	متن ساده
estekhdami.org	دریافت آگهی و ارسال توضیح	هر روز	متن ساده

به زبان فارسی و حتی استخراج کارنامک افراد از این سایت یکی از اهداف پروژه برای پیشنهاد کار می‌باشد. به همین منظور استخراج مدل داده ای این سایت به عنوان اولین قدم در زمینه‌ی به دست آوردن اطلاعات طبقه بندی شده از این سایت می‌باشد. در شکل ۵-۱ شمای پایگاهی استخراج شده از این دامنه نشان شده است.

۵-۳.۳ آماده سازی Nutch

در این مرحله، باید ابتدا سایت Nutch را برای اجرا آماده ساری نمود. برای کار ما از نسخه‌ی ۲. ۲ این نرم افزار استفاده نمودیم. این نسخه نسبت به نسخه‌های گذشته تغییرات بنیادی داشته است. به طور مثال می‌توان به ساختار ارتباط با پایگاه آن اشاره نمود. از این نسخه به بعد، با استفاده از تکنولوژی gora این امکان را ایجاد کرده است که Nutch مستقل از نوع پایگاه داده کار کند. به همین منظور در این پروژه از پایگاه داده‌ی mysql[۳] استفاده گردید. برای راه اندازی اولین کارهایی که انجام شد تنظیم نوع پایگاه داده‌ی آن و ستون‌های جدول پایگاهی مربوطه بود.
پس از این مرحله، باید خود نرم افزار را تنظیم می‌کردیم. در زیر تنظیمات کلی نرم افزار آمده است. در اینجا ما اسم عامل خزنده و لیست مجاز آن‌ها را مشخص کرده‌ایم. همچنین استفاده از sql نیز در اینجا مشخص شده است. تنظیمات آخر نیز تنظیمات رمزنگاری ارتباط شبکه‌ای را مشخص می‌کند.

[۳] www.mysql.com

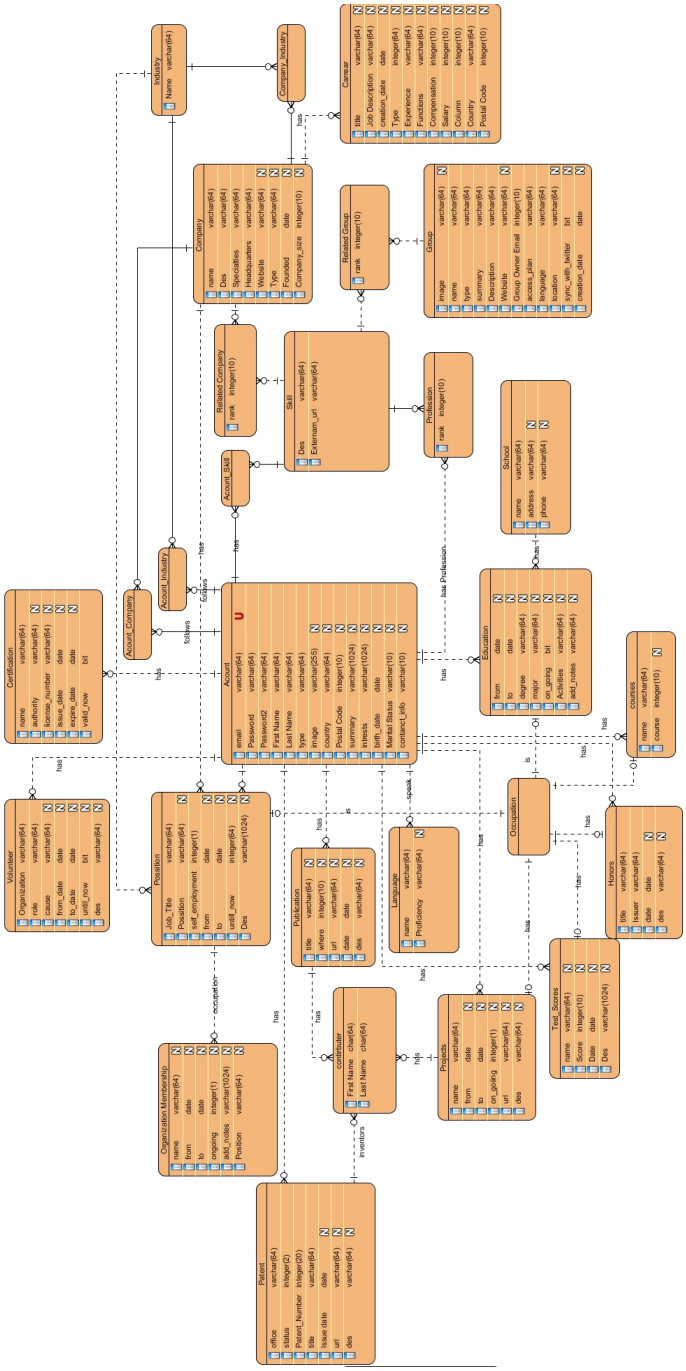
```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
	<property>
		<name>http.agent.name</name>
		<value>EmploySpider</value>
	</property>
	<property>
		<name>http.robots.agents</name>
		<value>EmploySpider,*</value>
	</property>
	<property>
		<name>storage.data.store.class</name>
		<value>org.apache.gora.sql.store.SqlStore</value>
	</property>
	<property>
		<name>parser.character.encoding.default</name>
		<value>utf-8</value>
	</property>
</configuration>
```

۴.۳-۵ پیاده سازی افزونه‌های تجزیه کننده

۵.۳-۵ پیاده سازی افزونه‌ی شاخص بندی

۴-۵ خلاصه



شکل ۵-۱: شماي داده اي سايت .Linkedin

فصل ۶

نتایج

۶-۱ مقدمه

۶-۲ محیط اجرای برنامه

۶-۳ روش به کار رفته

۶-۴ نتایج و بحث

۶-۴-۱ تعداد واژه‌ها

۶-۴-۲ زمان بازیابی اطلاعات

۶-۴-۳ نتایج جستجو

۶-۵ خلاصه

فصل ۷

نتيجه گيري

۷-۱ خلاصه

۷-۲ كارهاي آينده

سپاس

از استاد بزرگوارمان، دکتر حمید بیگی که با کمک‌ها و راهنمایی‌های بی‌دریغشان، ما را در انجام این پروژه یاری داده‌اند، تشکر و قدردانی می‌کنیم. هم‌چنین از آقای محمود نشاطی، به عنوان سرپرست پروژه و یاری دهنده در این مسیر صمیمانه سپاس‌گزاریم.

Bibliography

abstract

ToDO

Sharif University of Technology

Computer Engineering Department

B.Sc. Thesis

Computer Engineering - Software

Title:

Implementing Crawler Of Persian Business Search Engine

By:

Behnam Hatami Varaneh

Supervisor:

Dr. Hamid Beigi

Aguest 2013