
بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

پروژه‌ی کارشناسی

مهندسی کامپیوتر ـ نرم‌افزار

عنوان:

پیاده‌سازی 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 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

فصل ۱

پیش گفتار

فصل ۲

معرفی مسئله

۲-۱ مقدمه

۲-۲ تعریف دقیق مسأله

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، از سایت‌های آگهی استخدام موجود، اطلاعات استخراج شد و سپس شاخص بندی و آماده برای اجرای انواع پرسمان‌ها در اطلاعات استخراج شده می‌باشد. سپس اطلاعات به دست آمده به پژوهش مکمل برای استفاده 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 می‌شود، 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 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.

۲-۳ 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 :http //www.textkernel.com/ 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 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. قسمت خوراک شغل 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

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 نحوه ی 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 بندی

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:

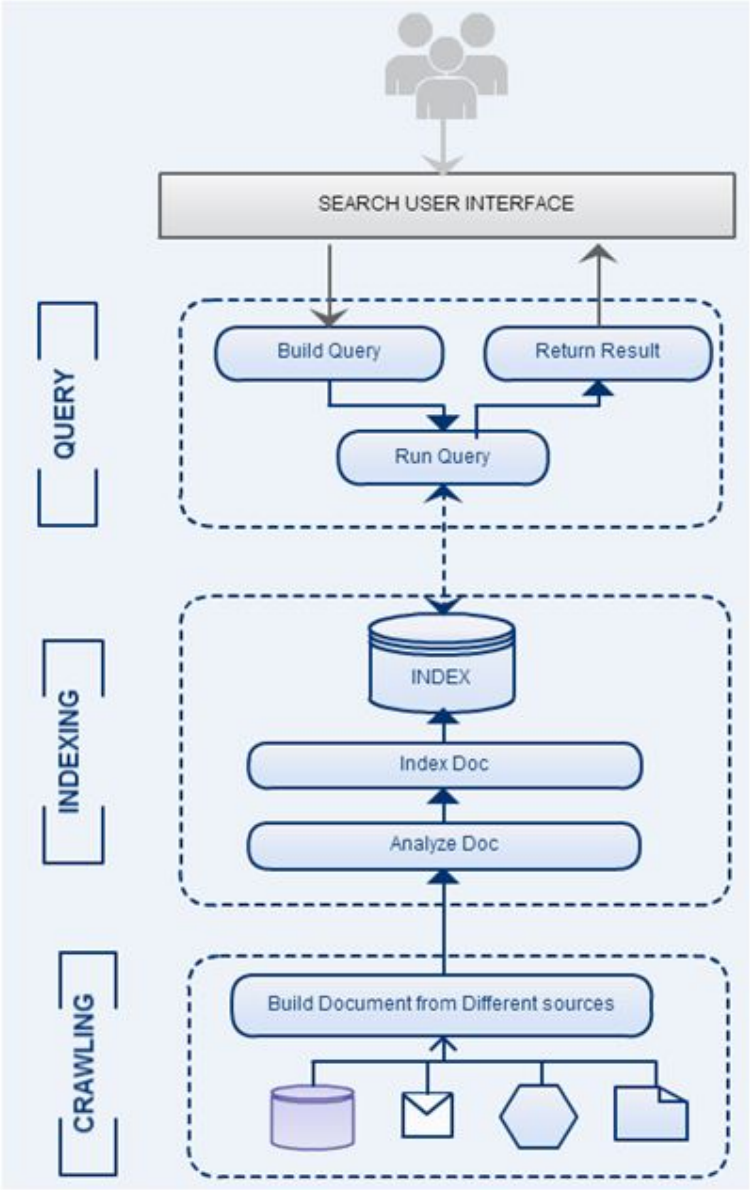
- Google[۱۷]
- Yahoo
- MSN
- E-Bay[۱۸]
- AOl[۱۹]

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

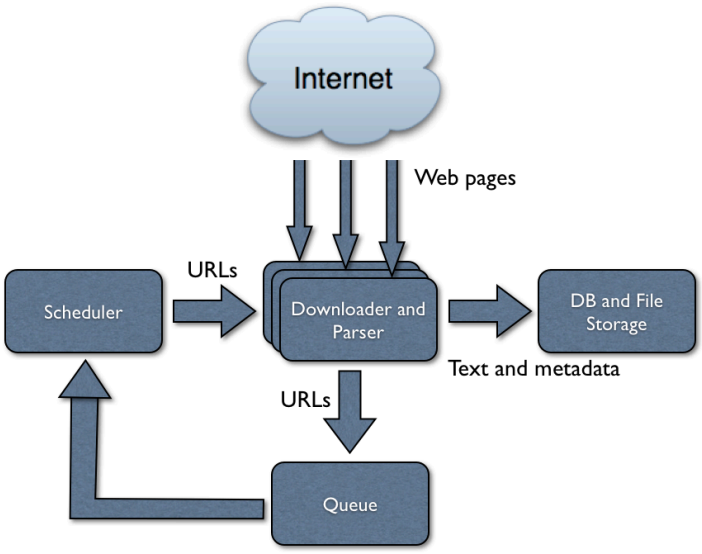
۳-۵ خلاصه‌ی فصل

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 نحوه ی کار یک موتور جستجو.



شکل ۳-۲: نحوه ی کار خزنده.

فصل ۴

موتور جستجوی Nutch

فصل ۵

پیاده سازی

فصل ۶

نتایج و نتیجه گیری

سلام[۱]

[۱] Decision Problem

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 Persian Business Search Engine

By:

Behnam Hatami Varaneh

Supervisor:

Dr. Hamid Beigi

Aguest 2013