دانشگاه یزد

دانشکده مهندسی برق و کامپیوتر

گروه مهندسی کامپیوتر

پایان نامه

برای دریافت درجه کارشناسی ارشد

مهندسی فناوری اطلاعات - شبکه‌های کامپیوتری

گسترش پرس‌وجو در موتور جستجوی فارسی

استاد راهنما:

دکتر علی محمد زارع بیدکی

استاد مشاور:

دکتر کیارش میزانیان

پژوهش و نگارش:

سیامک ساعدی

شهریور ۱۳۹۰

این پایان‌نامه با حمایت‌های مالی

مرکز تحقیقات مخابرات ایران

به انجام رسیده است.

تقدیم به سه وجود مقدس:

آنان که ناتوان شدند تا ما به توانایی برسیم. . .

موهایشان سپید شد تا ما روسفید شویم. . .

و عاشقانه سوختند تا گرمابخش وجود ما و روشنگر راهمان باشند. . .

پدرم

مادرم

معلّمانم و استادانم

تقدیر و تشکر

حمد و سپاس خداوند رحمان و رحیم را که صفاتش را نهایتی نیست، خداوندی که توفیق تلاش در عرصه علم آموزی و معرفت اندوزی را به ما عنایت فرمود و مرا در تمام مراحل زندگی یای و همراهی فرمود.

در طول دوران تحصیل و تهیه این پایان‌نامه از راهنمایی‌ها و مساعدت‌های اساتید و دوستان عزیزی بهره بُرده‌ام که در اینجا لازم است از همه ایشان مراتب سپاس قلبی و تشکر خالصانه خود را داشته باشم.

از استاد فرهیخته جناب آقای دکتر علی محمد زارع بیدکی بسیار سپاسگزارم، چرا که بدون راهنمایی‌های ایشان این پایان‌نامه بسیار مشکل می‌نمود. از استاد فرهیخته جناب آقای دکتر کیارش میزانیان به دلیل مشاوره‌های بی دریغ‌شان، که بسیاری از سختی‌ها را برایم آسان‌تر نمودند سپاسگزارم. از همه معلّمانم و دبیرانم که هریک با تمام وجود خویش بر آموخته‌هایم افزودند، قدردانی می‌نمایم.

چکیده :

پرس‌وجو برای بیان نیازهای کاربران به اطلاعات موجود در صفحات وب و سایر منابع، مطرح می‌شود. پرس‌وجوهای کاربران معمولاً «بسیار کوتاه» و شامل دو یا سه کلمه می‌باشند. قابلیت زبان‌های مختلف در بیان یک مفهوم با چندین کلمه مترادف و وجود یک کلمه با بیش از یک معنی، منجر به مشکل «عدم تطابق کلمه» و «مبهم بودن پرس‌وجو» می‌شوند. در نتیجه تعداد زیادی اسناد غیرمرتبط بازیابی شده و دقت کاهش می‌یابد. برای رفع این مشکلات از تکنیک گسترش پرس‌وجو استفاده می‌شود که از طریق پیشنهاد و افزودن واژگان مناسب به پرس‌وجوی کاربر، موجب بهبود دقت بازیابی خواهد شد. مهمترین مسئله، تشخیص و انتخاب واژه خوب برای افزودن به پرس‌وجوی کاربر می‌باشد. در مباحث گسترش پرس‌وجو دو عامل مهم «روش انتخاب واژگان گسترش» و «منابع واژگان گسترش» وجود دارد.

در این پایان‌نامه با درنظر گرفتن ویژگی‌های زبان فارسی و مسائل مرتبط با آن، به بررسی گسترش پرس‌وجو در موتور جستجوی فارسی می‌پردازیم. برای این کار از هستان‌شناسی عمومی فارس‌نت و شبکه مفهوم استفاده می‌کنیم که در برگیرنده روابط میان مفاهیم و کلمات می‌باشد. پس از دریافت پرس‌وجوی کاربر از طریق مطابقت آن با شبکه مفهوم، گروهی از کلمات مرتبط با پرس‌وجوی کاربر استخراج می‌شوند و در مرحله بعد مطابق الگوریتم ارائه شده، رتبه‌بندی می‌شوند. در پایان تعدادی از کلمات بسیار مرتبط با پرس‌وجوی کاربر به صورت خودکار به پرس‌وجوی اولیه کاربر افزوده می‌شوند و پرس‌وجوی گسترش یافته مورد جستجو قرار می‌گیرد. نتایج آزمایش‌های انجام شده، نشان دهنده بهبود عملکرد و افزایش دقت بازیابی می‌باشد.

کلمات کلیدی: موتور جستجو، پرس‌وجو، گسترش پرس‌وجو، هستان‌شناسی، شبکه مفهوم

فهرست مطالب

عنوان **صفحه**

فهرست شکل‌ها

عنوان شکل **صفحه**

فهرست جدول‌ها

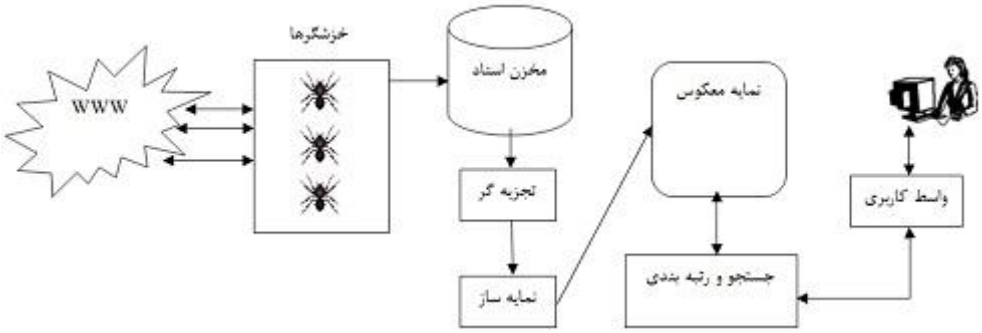
عنوان جدول **صفحه**

فصل اول

مقدمه

۱ مقدمه

در اوایل پیدایش وب به دلیل کم بودن اسناد، تعیین محل دسترسی آنها بدون نیاز به موتور جستجو[۱] امکان‌پذیر بود به عنوان مثال: ساختار فعلی موتورجستجوی Yahoo که از روش دایرکتوری استفاده می‌کند. اما با توجه به رشد روزافزون وب، امروزه این کار امکان‌پذیر نیست. لذا موتورهای جستجو تبدیل به ابزارهای مهمی جهت بازیابی اطلاعات از این محیط پهناور شده‌اند به طوری که فقط روزانه ۹/۶ میلیارد جستجو توسط موتورهای جستجوی اصلی در ایالت متحده امریکا انجام می‌شود. هدف اصلی موتور جستجو ارائه نتایج مرتبط و با کیفیت، با توجه به پرس‌وجوی ارسالی توسط کاربر می‌باشد.



شکل ۱-۱: معماری یک موتور جستجوی ساده [۶]

شکل ۱-۱ ساختار کلی یک موتور جستجو را نشان می‌دهد[۶]. هر موتور جستجو دارای یک یا چند خزشگر[۲] است که صفحات وب را پیمایش نموده و پس از بارگذاری[۳] صفحات، آنها را در مخزن اسناد قرار می‌دهد. تجزیه‌گر[۴] پس از استخراج متن صفحات و انجام پیش پردازش‌های لازم، آنها را به نمایه‌ساز[۵] تحویل می‌دهد. نمایه‌ساز پس از استخراج کلمات متن، آنها را در

[1] Search engine
[2] Crawler
[3] Download
[4] Parser
[5] Indexer

ساختمان داده‌ای به نام نمایه[1] قرار می‌دهد. در اکثر موتورهای جستجو از ساختار نمایه معکوس استفاده می‌شود. نمایه معکوس[2] از دو بخش تشکیل شده است : واژگان و لیست محل رخداد واژگان. واژگان شامل کلمات موجود در همه اسناد می‌باشد و لیست محل رخداد واژگان، شامل شناسه تمامی اسناد حاوی آن واژه می‌باشد. ماژول جستجو و رتبه بندی پس از جستجوی پرس‌وجو بر روی نمایه معکوس و بازیابی اسناد مرتبط، آنها را با استفاده از یک الگوریتم رتبه‌بندی، مرتب نموده و نتایج را در قالب یک واسط کاربری به کاربر نمایش می‌دهد.

به طور کلی موتور جستجو از چهار قسمت اصلی تشکیل شده است و ممکن است این بخش‌ها با یکدیگر ادغام شوند یا به بخش‌های بیشتری تقسیم شوند [۶] :

۱. **خزشگر**

برنامه‌ای است که وظیفه‌ی بازیابی صفحات از وب و ذخیره کردن آنها را در مخزن بر عهده دارد. همچنین بعد از اتمام عمل پیمایش علاوه بر صفحات، گراف وب را نیز قابل استخراج است.

۲. **نمایه‌ساز**

این واحد اسناد ذخیره شده در مخزن‌ها را پردازش کرده و نمایه سازی می‌کند. جهت بالا بردن سرعت دسترسی، از نمایه معکوس استفاده می‌شود. در نمایه معکوس به جای اینکه مشخص کنیم که یک سند شامل چه کلماتی می‌باشد، مشخص می‌کنیم که یک کلمه در چه سندهایی ظاهر شده است. نمایه معکوس شامل دو بخش است:

- بخش واژگان : شامل تمام کلمات موجود در همه اسناد
- لیست مکانی واژگان : شامل لیست اسناد حاوی یک کلمه

۳. **موتور بازیابی**[3]

[1] Index
[2] Inverted index
[3] Retrieval engine