

رسالة محمد

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 پایان نامه با حمایت های 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

معلمانم 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 پایان نامه از راهنمایی‌ها و مساعدت‌های اساتید و دوستان عزیز بهر بهره‌برده‌ام 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 پایان نامه 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 معلمانم و دبیرانم که هر یک 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 در نظر گرفتن ویژگی‌های زبان فارسی و مسائل مرتبط با آن، به بررسی گسترش پرس‌وجو در موتور جستجوی فارسی می‌پردازیم. برای این کار از هستان‌شناسی عمومی فارسی و شبکه مفهوم استفاده می‌کنیم که در برگزیده 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

صفحه	عنوان
۱-----	۱ مقدمه
۱۱-----	۲ مروری بر پژوهش‌های مرتبط انجام شده
۱۱-----	۱-۲ گسترش پرس‌وجو
۱۴-----	۲-۲ گسترش پرس‌وجوی خودکار
۱۴-----	۱-۲-۲ گسترش پرس‌وجوی خودکار: براساس نتایج جستجو
۱۹-----	۲-۲-۲ گسترش پرس‌وجوی خودکار: براساس ساختارهای دانش وابسته به مجموعه
۲۴-----	۳-۲ گسترش پرس‌وجوی تعاملی
۲۶-----	۱-۳-۲ گسترش پرس‌وجوی تعاملی: براساس نتایج جستجو
۲۷-----	۲-۳-۲ گسترش پرس‌وجوی تعاملی: براساس ساختارهای دانش وابسته به مجموعه
۲۹-----	۳-۳-۲ گسترش پرس‌وجوی تعاملی: براساس ساختارهای دانش مستقل از مجموعه
۳۰-----	۴-۲ هستان‌شناسی
۳۴-----	۱-۴-۲ گسترش پرس‌وجوی مبتنی بر احتمال
۳۸-----	۲-۴-۲ گسترش پرس‌وجوی مبتنی بر هستان‌شناسی
۴۱-----	۵-۲ شبکه مفهوم
۴۴-----	۱-۵-۲ گسترش پرس‌وجو مفهومی با استفاده از شبکه مفهوم
۴۵-----	۲-۵-۲ کارهای انجام شده در زمینه گسترش پرس‌وجو با استفاده از شبکه مفهوم
۵۳-----	۶-۲ برخی ملاحظات عمومی در گسترش پرس‌وجو
۵۳-----	۱-۶-۲ بازخورد مرتبط
۵۵-----	۲-۶-۲ واژگان پرس‌وجو
۵۶-----	۳-۶-۲ الگوریتم‌های رتبه‌بندی
۶۰-----	۴-۶-۲ قضاوت‌های برخط در مورد مرتبط بودن
۶۳-----	۳ روش پیشنهادی: گسترش پرس‌وجو در موتور جستجوی فارسی

۱-۳	ساخت شبکه مفهوم براساس هستان شناسی فارس نت	۶۴
۱-۱-۳	ریشه یابی..	۶۸
۲-۳	دریافت پرس وجوی کاربر، تجزیه و ریشه یابی	۶۹
۳-۳	استخراج زیرگراف	۷۱
۴-۳	وزن دهی به زیر گراف و رتبه بندی مفهوم های زیرگراف	۷۴
۵-۳	انتخاب k واژه مناسب و افزودن به پرس وجوی اولیه کاربر	۷۸
۶-۳	نتایج ارزیابی روش پیشنهادی	۷۹
۷-۳	مقایسه روش پیشنهادی با روش های قبلی	۸۱
۴	نتیجه گیری و پیشنهادات	۸۵
	واژه نامه فارسی به انگلیسی	۸۹
	واژه نامه انگلیسی به فارسی	۹۴
	منابع....	۹۹

فهرست شکل‌ها

عنوان شکل	صفحه
شکل ۱-۱: معماری یک موتور جستجوی ساده	۱
شکل ۲-۱: گسترش پرس‌وجو (روش‌ها و منابع)	۷
شکل ۱-۲: یک نمونه از شبکه مفهوم	۴۱
شکل ۲-۲: روابط مفهوم‌ها در شبکه مفهوم	۴۲
شکل ۳-۲: رابطه مفهوم و کلمات در شبکه مفهوم	۴۲
شکل ۴-۲: شبکه مفهوم	۴۷
شکل ۵-۲: شبکه معنایی برای معنی کلمه bus	۵۰
شکل ۶-۲: اشتراک بین شبکه معنایی برای معنی کلمات mountain و top	۵۰
شکل ۷-۲: مثالی از الگوریتم رتبه‌بندی PageRank	۶۰
شکل ۱-۳: شبکه مفهوم استخراج شده از هستان‌شناسی	۶۷
شکل ۲-۳: شبکه مفهوم استخراج شده از هستان‌شناسی فارسی	۷۰
شکل ۳-۳: زیرگراف به ازای واژه پرس‌وجوی «کلمه ۳» و رابطه Hypernym (ریشه‌ها)	۷۱
شکل ۴-۳: زیرگراف به ازای واژه پرس‌وجوی «کلمه ۳» و رابطه Hypernym (تا سطح ۲)	۷۲
شکل ۵-۳: زیرگراف به ازای واژه پرس‌وجوی «کلمه ۳» و رابطه Hypernym (کامل)	۷۳
شکل ۶-۳: زیرگراف به ازای واژه پرس‌وجوی «کلمه ۳» و رابطه Synonym (کامل)	۷۳
شکل ۷-۳: نتایج بدست آمده برای روش پیشنهادی	۸۰
شکل ۸-۳: نتایج بدست آمده برای گسترش به ازای هر واژه و گسترش به ازای کل پرس‌وجو	۸۱
شکل ۹-۳: مقایسه نتایج بدست آمده برای روش پیشنهادی و روش نوینگیلی	۸۲

فهرست جدول‌ها

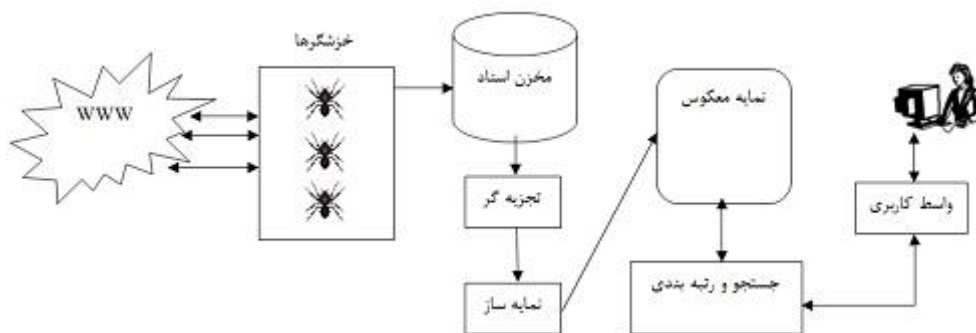
عنوان جدول	صفحه
جدول ۱-۲: مقایسه نتایج روش خانم فرهودی.....	۴۱
جدول ۲-۲: نتایج بدست آمده به ازای کلماتی که فقط دارای یک معنی می‌باشند.....	۵۲
جدول ۳-۲: نتایج بدست آمده به ازای کلمات رفع ابهام شده.....	۵۲
جدول ۱-۳: درجه اهمیت برای هر کدام از روابط هستان‌شناسی.....	۷۶
جدول ۲-۳: نتایج بدست آمده برای روش پیشنهادی.....	۸۰
جدول ۳-۳: نتایج بدست آمده برای گسترش به ازای هر واژه و گسترش به ازای کل پرس‌وجو.....	۸۰
جدول ۴-۳: مقایسه نتایج بدست آمده برای روش پیشنهادی و روش نویگی.....	۸۲
جدول ۵-۳: نتیجه گسترش چند پرس‌وجو با استفاده از روش پیشنهادی و روش نویگی.....	۸۳

فصل اول

مقدمه

۱ مقدمه

در اوایل پیدایش وب به دلیل کم بودن اسناد، تعیین محل دسترسی آنها بدون نیاز به موتور جستجو^۱ امکان پذیر بود به عنوان مثال: ساختار فعلی موتور جستجوی Yahoo که از روش دایرکتوری استفاده می کند. اما با توجه به رشد روزافزون وب، امروزه این کار امکان پذیر نیست. لذا موتورهای جستجو تبدیل به ابزارهای مهمی جهت بازیابی اطلاعات از این محیط پهناور شده اند به طوری که فقط روزانه ۹/۶ میلیارد جستجو توسط موتورهای جستجوی اصلی در ایالت متحده امریکا انجام می شود. هدف اصلی موتور جستجو ارائه نتایج مرتبط و با کیفیت، با توجه به پرس و جوی ارسالی توسط کاربر می باشد.



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

شکل ۱-۱ ساختار کلی یک موتور جستجو را نشان می دهد [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 می دهد. تجزیه گر^۴ پس از استخراج متن صفحات و انجام پیش پردازش های لازم، آنها را به نمایه ساز^۵ تحویل می دهد. نمایه ساز پس از استخراج کلمات متن، آنها را در

¹ Search engine

² Crawler

³ Download

⁴ Parser

⁵ Indexer

ساختمان داده‌ای به نام نمایه^۱ قرار می‌دهد. در اکثر موتورهای جستجو از ساختار نمایه معکوس استفاده می‌شود. نمایه معکوس^۲ 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^۳

¹ Index

² Inverted index

³ Retrieval engine