دانـــشگاه یـــــزد
دانشکده مهندسی برق و کامپیوتر
گروه مهندسی کامپیوتر

پایان نامه
برای دریافت درجه کارشناسی ارشد
مهندسی فناوری اطلاعات- شبکه‌های کامپیوتری

خزش و رتبه‌بندی کارا مبتنی بر ویژگی‌های گراف وب

استاد راهنما:
دکتر علی‌محمد زارع‌بیدکی

استاد مشاور:
دکتر ولی درهمی

پژوهش و نگارش:
محمدامین گلشنی

اسفندماه ۱۳۹۰

این پایان‌نامه با حمایت‌های مالی
مرکز تحقیقات مخابرات ایران
به انجام رسیده است.

تقدیم به:

گوهران پاکی که روشنایی بخش دیدگانم در تمام زندگی‌ام بودند

پدرم و مادرم

آنها که گرمای وجودشان پشتیبان من

در سردترین روزگاران بودند

برادران فداکارم

سپاسگزاری

این ناچیز مجموعه ای است که با الطاف و عنایات حضرت احدیت به سرانجام رسیده که لازم میدانم دست شکر به آستانش گشوده، رحمت بی کرانش را بستایم تا مزید ابر رحمانی او را به نظاره بنشینم و بر سبیل ((من لم یشکر المخلوق لم یشکر الخالق)) از زحمات بی دریغ استاد راهنمایم دکتر علی محمد زارع بیدکی که در طی این مدت با راهنمایی های ارزنده و صبر و شکیبایی خود، مرا به مسیر پرافتخار علم و پژوهش هدایت نموده اند کمال تشکر را داشته باشم و نیز از اساتید بزرگوار جناب دکتر ولی درهمی (استاد مشاور)، جناب دکتر محمد قاسم زاده (داور داخلی) و جناب دکتر سیدابوالفضل شاهزاده فاضلی (داور خارجی) که در تنیل بدین سو یاریم نمودند تشکر و قدردانی نمایم.

در پایان از کلیه عزیزانی که مرا در انجام این پایان نامه یاری نمودند صمیمانه تشکر می نمایم.

امید است در سایه توجهات ساحت اقدس الهی سهای دلمان نوری از منبع لایزال الهی گیرد.

چکیده

امروزه وب جهان گستر به عنوان بهترین محیط برای تولید اطلاعات، انتشار و دسترسی به دانش مورد نیاز کابران تبدیل شده است. نیاز به اشتراک گذاری داده‌ها و هزینه پایین تولید در دسترس قرار دادن صفحات وب، منجر به پویایی وب شده به طوری که رشد بی‌رویه ساختار و پوشش متنوع موضوعات خود را نمایان می‌سازد. نکته حائز اهمیت در این محیط روش یافتن اطلاعات مورد نیاز کاربر می‌باشد. آمارهای منتشر شده حاکی از آن است که اکثر کاربران اینترنتی، از موتورهای جستجو به عنوان درگاه دسترسی و جستجو در این حجم عظیم از اطلاعات استفاده می‌کنند. با توجه به وابستگی بالای کاربران به موتور جستجو، این ابزار به جزئی لاینفک از وب تبدیل شده و در حال حاضر بر اساس فناوری‌های موجود می‌توان ادعا کرد جایگزین مناسب‌تری جهت استخراج اطلاعات از این محیط وسیع و غنی وجود ندارد.

در این پایان نامه جهت پوشش مناسب صفحات وب (پیدا کردن سریع صفحه‌ها مهم)، الگوریتم خزشی به نام IECA[1] مبتنی بر چندین ویژگی نظیر درجات خروجی، درجات ورودی ، فاصله لگاریتمی و خاصیت ساختاری گراف وب ارائه گردید. جهت ارزیابی دقت الگوریتم IECA از چهار گراف مختلف وب ایران، ایتالیا، انگلستان و دانشگاه برکلی استفاده می‌گردد. نتایج آزمایشات حاکی از کاراتر بودن الگوریتم IECA نسبت به سایر الگوریتم‌های بررسی شده می‌باشد. همچنین یک روش رتبه‌بندی مبتنی بر انتشار[2] امتیاز محبوبیت[3] صفحات وب و اطلاعات موجود در آدرس صفحات وب (تعداد اسلش‌های، "/"، موجود در آدرس صفحه) ارائه و جهت بررسی کارایی الگوریتم ارائه شده مجموعه داده LETOR 3 مورد استفاده قرار گرفت. بر اساس نتایج آزمایشات استفاده از آدرس صفحات در عمل انتشار اهمیت صفحات در افزایش دقت فرآیند رتبه‌بندی موثر می‌باشد.

کلمات کلیدی: موتور جستجو، خزشگر، گراف وب ، رتبه‌بندی، انتشار وابستگی، تعداد اسلش‌های موجود در آدرس

[1] Intelligent Effective Crawling Algorithm
[2] Propagation
[3] Popularity

فهرست مطالب

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فهرست جداول

عنوان **صفحه**

فهرست علائم اختصاری

مخفف عبارت	عبارت انگلیسی
AP	Average Precision
HITS	Hypertext-Induced Topic Search
HS	Hyperlink-based Score Propagation
HSS	Hyperlink-based Slash-Score Propagation
HST	Hyperlink-based Slash-Term Propagation
HT	Hyperlink-based Term Propagation
IDF	Inverse Document Frequency
TF	Term Frequency
TSPR	Topic-Sensitive PageRank
UO	Uniform Out-link
WI	Weighted In-link
WO	Weighted Out-link

فصل اول

مقدمه

۱-۱ شرح پایان‌نامه

موضوع پایان‌نامه به خزش و رتبه‌بندی صفحات وب اختصاص داده شده است. خزش و رتبه‌بندی از بخش‌های مهم در موتورهای جستجو می‌باشند که به ترتیب وظیفه‌ی بارگذاری صفحات با کیفیت و رتبه‌بندی نتایج بازگشتی به کاربر را بر عهده دارند. تا کنون الگوریتم‌های خزش متنوعی پیشنهاد شده‌اند، اما یا بازدهی مناسبی نداشته‌اند یا از پیچیدگی بالایی برخوردار هستند. لذا الگوریتم خزش کارایی به نام(IECA(Intelligent Effective Crawling Algorithm پیشنهاد می‌گردد که نسبت به الگوریتم‌های فعلی از کارایی بالاتر و پیچیدگی پایین‌تری برخوردار می‌باشد. در روش پیشنهادی اهمیت صفحات بر اساس چندین ویژگی نظیر درجات ورودی ، فاصله لگاریتمی و خاصیت ساختاری گراف وب (بالابودن درجه ورودی در صفحات با کیفیت) تعیین می‌گردد. نتایج آزمایشات بر روی چهار گراف مختلف وب (ایران، انگلستان، ایتالیا و گراف وب دانشگاه برکلی) حاکی از کاراتر بودن الگوریتم پیشنهادی نسبت به سایر الگوریتم‌های خزش می‌باشد. همچنین بخشی از پایان‌نامه به مبحث رتبه‌بندی اختصاص گرفته و مدلی مبتنی بر انتشارشامل دو الگوریتم به نام‌های Hyperlink-based Slash-Score propagation(HSS) و Hyperlink-based Slash-Term propagation(HST) جهت تعیین رتبه‌ی صفحات پیشنهاد گردید. جهت ارزیابی مدل رتبه‌بندی پیشنهادی LETOR 3 مورد استفاده قرار گرفت. بر اساس آزمایشات دو الگوریتم پیشنهادی نسبت به سایر الگوریتم‌های رتبه‌بندی انتشاری از دقت بهتری برخوردار هستند.

۱-۲ مقدمه

امروزه با توجه به رشد روز افزون اطلاعات و محتوای موجود در وب و همچنین تغییرات زیاد در اطلاعات موجود، موتورهای جستجو نقش مهمی در بازیابی اطلاعات از اینترنت ایفاء می نمایند. قابل ذکر است که حدود ۸۰٪ از افراد از طریق موتورهای جستجو به سایت‌ها و اطلاعات مورد نظرشان دسترسی پیدا می‌کنند [۱].

در روزهای اولیه‌ی پیدایش وب به دلیل کم بودن اسناد، تعیین محل دسترسی آنها بدون نیاز به موتور جستجو (به روش دایرکتوری مثل ساختار فعلی‌موتور جستجویYahoo) امکان پذیر بود اما امروزه با توجه به رشد چشمگیر وب، این کار امکان پذیر نیست. لذا موتورهای جستجو تبدیل به ابزاری مهم جهت بازیابی اطلاعات از این محیط پهناور شده‌اند به طوری که ماهانه حدود ۹/۶ میلیارد جستجو توسط موتورهای جستجوی اصلی در ایالت متحده امریکا انجام می‌شود [۲]. جدول۱-۱ سهم هر موتور جستجو در انجام پرس‌وجوهای کاربران و تعداد پرس‌وجوهای دریافتی از کاربران را نشان می‌دهد.

جدول ۱-۱:سهم موتور جستجوهای اصلی در پاسخ به پرس‌جوهای کاربران در امریکا [۲].

موتور جستجوهای اصلی	سهمی که هر موتور جستجو در پرس‌وجوهای

	دریافتی داشته(درصد)		
	نوامبر ۲۰۰۷	دسامبر ۲۰۰۷	مقایسه بین دو ماه
کل پرس‌وجوهای انجام شده	٪۱۰۰	٪۱۰۰	٪۰.۰
موتور جستجوی Google	٪۵۸/۶	٪۵۴/۴	-٪۰/۲
موتور جستجوی Yahoo	٪۲۲/۴	٪۲۲/۹	٪۰/۵
موتور جستجوی Microsoft sites	٪۹/۸	٪۹/۸	٪۰/۰
موتور جستجوی Time Warner Network	٪۴/۵	٪۴/۶	٪۰/۱
موتور جستجوی ASK	٪۴/۶	٪۴/۳	-٪۰.۳

هدف اصلی موتور جستجو ارائه‌ی نتایج مرتبط و با کیفیت با توجه به پرس‌وجوی ارسالی توسط کاربر است. برای یک موتور جستجو مشکلاتی وجود دارد که باید از عهده آنها برآید، در ادامه به این مشکلات اشاره شده است[۳].

۱) اولین مشکل مشخص کردن صفت "مرتبط" می‌باشد که باعث شده موتورهای جستجو اهمیت زیادی به الگوریتم‌های رتبه‌بندی بدهند. بازیابی صفحه‌ی مرتبط با پرس‌وجوی کاربر در نتایج نخست و بخصوص در صفحه‌ی اول بسیار پر اهمیت است چرا که معمولا کاربران کمی عجول بوده و فقط ۱۰ نتیجه‌ی اول را بررسی می‌کنند.

۲) مسئله‌ی دیگری که مطرح می‌شود مسئله‌ی پهنش رتبه[1] می‌باشد. با توجه به راه‌اندازی تجارت الکترونیک در محیط وب برای شرکت‌های تجاری مهم است که در نتایجی که موتور جستجو به کاربر نشان می‌دهد جزء نتایج اولیه باشند. لذا برای رسیدن به این هدف به پهنش رتبه رو می‌آورند، پهنش رتبه به دو صورت تغییر محتوا و یا افزایش تعداد پیوندها انجام می‌گیرد. در روش تغییر محتوا دسته‌ای ازکلمات خاص مدام در صفحه تکرار می‌شوند که ممکن است رنگ فونت اختصاص داده شده به این دسته از کلمات همرنگ با رنگ پس زمینه باشد یا این دسته از کلمات دارای فونت کوچکی باشند که کاربر متوجه آنها نشود.

۳) مسئله‌ی دیگر که در رابطه با الگوریتم‌های مبتنی بر پیوند مطرح می‌شود، مسله‌ی غنی‌تر شدن اغنیا[2] می‌باشد. یعنی اینکه صفحات محبوب در صدر نتایج قرار می‌گیرند و صفحات محبوب تازه متولد شده که فعلا کسی آنها را ندیده و اشاره‌ای به آنها نمی‌شود در دید کاربران قرار نمی‌گیرند، در نتیجه پیوند به صفحات محبوب بیشتر شده و محبوب‌تر می‌شوند.

[1] Spam
[2] Rich get richer

۴) مسئله دیگری که در رابطه با موتور جستجو مطرح می‌شود، مسئله‌ی پوشش[۱] و تازگی[۲] است. این دو معیار در جهت مخالف یکدیگر عمل می‌کنند چرا که اگر خزشگر هر چه صفحات بیشتری بارگذاری کند کمتر می‌تواند به تازگی صفحات بارگذاری شده بپردازد. بزرگترین موتور جستجو در سال ۱۹۹۹یعنی Northen Light در حدود ۱۶٪ از وب را تحت پوشش قرار می‌داد و در سال ۲۰۱۰ موتور جستجوی گوگل در حدود ۷۰٪ از وب را تحت پوشش قرار داده است.

جهت حل مسئله‌ی پوشش می‌توان از فراجویشگرها[۳] استفاده کرد به این صورت که پرس‌وجوی کاربر را به چندین موتور جستجو ارسال می‌کنند و نتایج بازگشتی را با یکدیگر ادغام می‌کنند اما چون روش‌های رتبه‌بندی موتور جستجوهای مختلف با یکدیگر متفاوت است ادغام نتایج بازگشتی از موتورهای جستجو کار راحتی نبوده و از آنجا که موتورهای جستجو دارای سرعت متفاوتی هستند، ممکن است عمل جستجو کند گردد.

۵) مشکل کلیدی دیگری که می‌توان به آن اشاره کرد انتظارات افراد مختلف از یک پرس‌وجو است مثلا هدف یک محقق از پرس‌وجوی *موتور جستجو* می‌تواند مقالات علمی باشد در حالی که هدف یک دانش‌آموز از یک پرس‌وجو می‌تواند یافتن اطلاعاتی ساده در مورد موتور جستجو باشد.

۶) مسئله‌ی دیگر مبهم بودن یکسری کلمات استفاده شده در پرس‌وجو می‌باشد، در مواقعی که کاربر نتواند درخواست خود را با چندین عبارت بیان کند. به عنوان مثال کلمه Python را در نظر بگیرید که دارای سه معنی مختلف، غیبگو، افعی و زبان برنامه‌نویسی است.

مطابق با مباحث ذکر شده، یکی از بخش‌های مهم موتور جستجو، بخش رتبه‌بندی می‌باشد. رتبه‌بندی فرآیندی است که کیفیت یک صفحه از لحاظ ارتباط با پرس‌وجوی کاربر توسط موتور جستجو تخمین زده می‌شود. با توجه به اینکه به ازای هر پرس‌وجوی کاربر معمولا هزاران صفحه‌ی مرتبط وجود دارد، لازم است آنها را اولویت بندی کرده و بهترین‌ها به کاربر نشان داده شود. مسائلی مانند حوصله کم کاربران (مشاهده‌ی ۱۰ تا ۲۰ نتیجه‌ی اول)، کم بودن کلمات پرس‌وجو، حجم زیاد اطلاعات، فرآیند رتبه‌بندی را با مشکلاتی جدی روبه‌رو می‌کند [۳]. در حال حاضر سه روش مهم رتبه‌بندی مبتنی بر محتوا (استفاده شده در روش‌های سنتی بازیابی اطلاعات)، مبتنی بر ساختار (پیوندهای بین صفحات وب)و ترکیبی از محتوا و ساختار وجود دارد. در روش‌های استفاده شده در بخش سنتی بازیابی اطلاعات مدل‌هایی مانند بولی[۴]، احتمالی و فضای برداری[۵] جهت رتبه‌بندی اسناد بر مبنای محتوای آنها ارائه شده است. در فصل دوم مروری بر الگوریتم‌های رتبه‌بندی انجام خواهد گرفت.

---

[1] Coverage
[2] Freshness
[3] Meta-Search engines
[4] Boolean
[5] Polysemy

۳-۱ ساختار وب

می‌توان وب را به صورت یک گراف در نظر گرفت که در آن صفحات به منزله‌ی گره‌های گراف و پیوند بین صفحات به منزله‌ی یال‌های موجود در گراف می‌باشند. بر اساس تحقیقات انجام شده بیشتر از ۹۰٪ وب دارای ساختاری مانند شکل ۱-۱ است که از چهار قسمت اصلی تشکیل شده است [۵-۳].

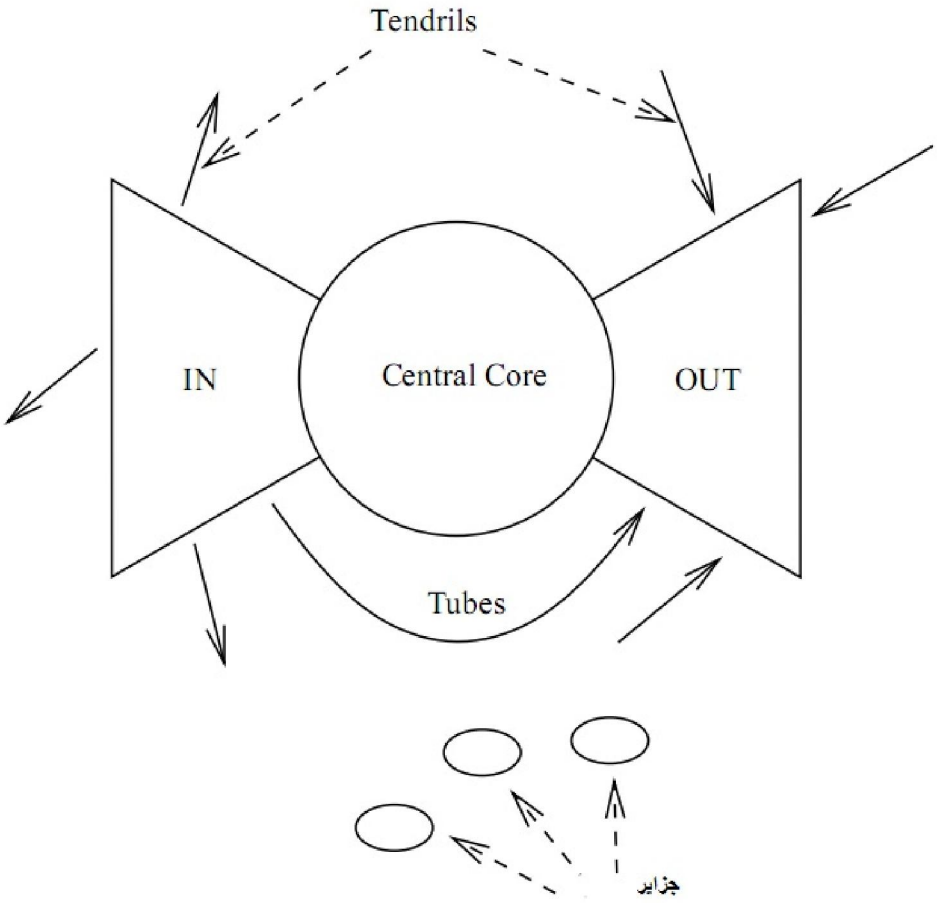
۱) هسته‌ی مرکزی[1]. شامل اسنادی است که به هم متصل‌اند و تمام اسناد موجود در این بخش قابل دسترسی‌اند.

۲) بخش IN. این بخش شامل اسنادی است که از بخش هسته‌ی مرکزی به آنها پیوندی وجود ندارد اما از این بخش به هسته‌ی مرکزی پیوند وجود دارد. می‌توان گفت که در این بخش اسنادی قرار می‌گیرند که یا تازه متولد شده‌اند، یا ناشناخته‌اند یا اسنادی‌اند که دارای اهمیت زیادی نمی‌باشند. اما این نکته نباید از دید پنهان باشد که این اسناد، **نقاط خوبی برای شروع عمل خزش**[2] می‌باشند.

۳) بخش OUT. این بخش متشکل از اسنادی است که از هسته‌ی مرکزی به آنها پیوند وجود دارد اما از اسناد موجود در این بخش هیچ پیوندی به اسناد موجود در هسته‌ی مرکزی وجود ندارد. به عبارتی می‌توان گفت اسناد قرار گرفته در این بخش، اسنادی‌اند که شامل پیوند به یکدیگر می‌باشند. مثلا یک وب سایت که از صفحات دیگر به آنها اشاره شده است اما آنها به صفحات دیگری اشاره نکرده‌اند و فقط به یکدیگر پیوند دارند.

۴) بخش دیگر پیچک‌ها[3] می‌باشند که جزء بخش IN یا بخش OUT می‌باشند اما به هسته‌ی مرکزی پیوندی ندارند. قسمت دیگر لوله یا تونل[4] می‌باشد که به اسنادی گفته می‌شود که در بخش IN بوده و به بخش OUT متصل شده‌اند و پیوندی به بخش هسته‌ی مرکزی ندارند و در نهایت آخرین قسمت جزایرند[5] که توسط هیچ یک از اسنادی که در این ساختار پاپیونی شکل وجود دارد، پیوندی به آنها وجود ندارد.

[1] Central Core
[2] Seed
[3] Tendrils
[4] Tube
[5] Island



شکل ۱-۱: ساختار پایونی وب[۵-۳]

۱-۴ موتور جستجو در یک نگاه

موتور جستجو از چهار قسمت اصلی تشکیل شده است (که ممکن است این بخش‌ها با یکدیگر ادغام شوند یا به بخش‌های بیشتری تقسیم شوند)[۳].

- **خزشگر[۱]- ربات[۲]، عنکبوت[۳]،راه‌رونده[۴] یا سرگردان[۵]**

برنامه‌ای است که وظیفه‌ی بازیابی صفحات از وب و ذخیره کردن آنها را در مخزن بر عهده دارد[۷, ۶, ۴, ۳]. خزشگر دو هدف تازه‌سازی و پوشش بالارا به دنبال دارد که با

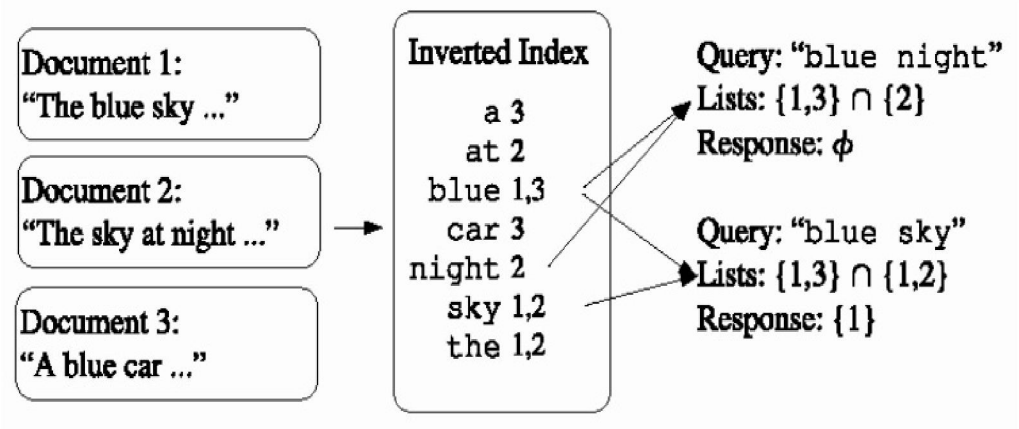
[1] Crawler
[2] Robots
[3] Spider
[4] Walker
[5] wanderer

یکدیگر رابطه عکس دارند، بعد از اتمام عمل پیمایش علاوه بر صفحات، گراف وب نیز قابل استخراج می‌باشد.

- **نمایه‌ساز**

این واحد اسناد ذخیره شده در مخزن‌ها را پردازش کرده و نمایه سازی[1] می‌کند. جهت بالا بردن سرعت دسترسی از نمایه‌سازی معکوس استفاده می‌شود. در نمایه سازی معکوس به جای اینکه مشخص شود که یک سند شامل چه کلماتی می‌باشد، مشخص می‌گردد که یک کلمه در چه سندهایی ظاهر شده است. نمایه‌ساز معکوس از دو بخش واژگان، که شامل تمام کلمات موجود در همه‌ی اسناد و لیست مکانی آنها شامل تمام اسنادی که آن کلمه در آنها ظاهر شده است می‌باشد.

شکل زیر نمایه‌ساز معکوس را نشان می‌دهد که شامل حروف اضافه نیز می‌باشد.



شکل ۱-۲: مثالی از نمایه سازی معکوس[۳].

- **موتوربازیابی**[2]

موتور بازیابی رابطه‌ی مستقیم با دو بخش نمایه‌سازی و واسط کاربر دارد. وظیفه‌ی اصلی این واحد رتبه‌بندی جواب‌ها با استفاده از نتایج تهیه شده توسط واحد نمایه‌سازی و گراف تهیه شده توسط خزشگر می‌باشد. با توجه به اینکه کاربر پرس‌وجوی را به زبان طبیعی بیان می‌کند از کارهای مهم واحد پردازش پرس‌وجو، عملیات پردازش زبانی پرس‌وجو مانند نرمال سازی[3] و بسط پرس‌وجو می‌باشد. از وظایف دیگر این واحد، انجام عملیاتی مانند مدل کردن کاربر، و نیز اِعمال بازخورد[4] خواهد بود. بعد از پردازش پرس‌وجو و بدست

[1] index
[2] Retrieval Engine
[3] Normalization
[4] Relevance Feedback