عنوان زیرپروژه:

بهینه‌سازی استفاده از موتورهای جستجو در پیکره‌های متنی زبان فارسی

فهرست مطالب

عنوان **شماره صفحه**

عنوان **شماره صفحه**

عنوان **شماره صفحه**

1. مقدمه

گسترش اسناد الکترونیکی در سطح سازمان‌ها، مراودات روزمره و مهمتر از همه در سطح وب، با وجود تمام مزایا و جنبه‌های کارآمد آن، باعث بروز مشکلاتی از جمله یافتن اطلاعات مورد نیاز در میان انبوه اطلاعات است. وقتی حجم وتنوع اطلاعات عرضه شده بالا باشد، افراد برای پیدا کردن اطلاعات مورد نیاز خود – حتی اگر این اطلاعات مهم و برجسته هم باشند – باید اطلاعات عرضه شده را با روش‌های خاصی جستجو و پالایش کند. به همین منظور موتورهای جستجوی پیشرفته عرضه شدند که قادر هستند تا اطلاعات مورد نیاز کاربر را از میان ملیون‌ها سند یافته و نتایج را به ترتیب اولویت به کاربر ارایه دهند. وقتی صحبت از سند (document) می‌شود، منظور ما، اسناد غیرساخت‌یافته هستند که شامل انبوهی از کلمات و جملات در مورد موضوعی خاص هستند و هیچ طبقه‌بندی موضوعی برای اسناد انجام نشده است.

کاربر برای پیدا کردن سند یا مطلب مورد نظر خود در موتور جستجو، از «کلمات کلیدی» یا «کلیدواژه» استفاده می‌کند. موتور جستجو نیز برای یافتن اسناد مرتبط از کلیدواژه‌ها کمک می‌گیرد. اما اینکه چقدر الگوریتم رتبه‌بندی موتور جستجو به الگوریتم ذهنی کاربر، در بیان کلمات کلیدی نزدیک باشد، خود مقوله مفصلی است که در این تحقیق به آن خواهیم پرداخت.

در این میان زبان فارسی ویژگیهای خاصی دارد که بازیابی اطلاعات برای آن را مشکل ساخته است [1]. به نظر نویسنده، کابران فارسی‌زبان برای یافتن مستندات و مطالب مورد نظر خود در صفحه اول موتورهای جستجو، اغلب با شکست مواجه می‌شوند و معمولا زمان بیشتری را – نسبت به زبان‌های دیگر – برای جستجو در وب اختصاص می‌دهند. البته بخشی از آن مربوط می‌شود به کمبود منابع فارسی در وب اما بخش عمده آن در اثر عدم وجود قانون همه‌گیر برای دستور خط فارسی است؛ این مشکل باعث شده تا کلمات کلیدی انتخاب شده توسط کاربر و کلمات موجود در مستندات سازگار نبوده و از طرفی الگوریتم‌های رتبه‌بندی، درک درستی از مرز کلمات و عبارات و نیز رفتار کاربر در انتخاب کلمه کلیدی نداشته باشند.

برای کاستن از مشکلات عدم تفاهم بین کاربر و موتورهای جستجو در زبان فارسی، برخی از روش‌ها پیشنهاد و انجام شده‌اند. در یک نمونه بارز در سایت *پارسیک* [2]، با ایجاد یک واسط بین پرسش کاربران و موتورهای جستجو و تغییر و تبدیل کلمات فرستاده شده، با استفاده از سرویس وب موتورهای جستجو، نتایج متفاوتی به کاربر داده می‌شود.

وقتی اصلاح «بهینه‌سازی» مطرح می‌شود، چندین مفهوم از آن برداشت می‌شود که این تفاوت در برداشت، از تفاوت بین مفهوم «بازیابی اطلاعات» و «بازیابی داده» نشات می‌گیرد. داده‌ها ابهام ندارند اما اطلاعات نیاز به تفسیر دارد و در نتیجه مبهم می‌شوند. سیستم بازیابی داده نیاز به رفع این ابهام‌ها را ندارد اما در سیستم بازیابی اطلاعات باید هر چه بهتر اطلاعات را مدل کنیم تا ابهام ها در درک اطلاعات توسط سیستم کمتر شوند. در بازیابی داده به روش‌های اندیس‌گذاری و انباره‌های ذخیره‌سازی پرداخته می‌شود، اما در بازیابی اطلاعات به تفسیر داده‌ها و ارتباط آن با نیاز کاربر پرداخته می‌شود. در سیستم های بازیابی داده، کارایی سیستم از نظر سرعت و فضا به عنوان معیار ارزیابی در نظر گرفته می‌شود، و در سیستم های بازیابی اطلاعات، معیار دقت (precision) و بازخوانی (recall) و شبیه به آن، به عنوان معیار ارزیابی سیستم به کار می‌روند.

1-1 حوزه مساله

در این گزارش ما به مسایل مربوط با بازیابی داده نمی‌پردازیم. همه مسایل مربوط به آن در [3] به طور کامل حل شده است؛ آنچه که بدان می‌پردازیم بهینه‌سازی در بازیابی اطلاعات است که منظور از آن بهبود پارامترهای ارزیابی مرتبط با سیستم‌های بازیابی اطلاعات است. بطور خلاصه منظور ما از «بهینه‌سازی استفاده از موتور جستجو» این است که ترتیب نتایج بازگردانده شده‌ی آن، یا صفحه اول نتایج، شامل درصد قابل قبولی از مستندات مرتبط با نیاز کاربر باشند.

البته اصلاح «بهینه‌سازی موتور جستجو» یا «Search Engine Optimization»، (یا بطور اختصاری SEO)، در ادبیات عمومی کامپیوتر، به تکنیک‌هایی گفته می‌شود که طراح سایت آنها را بکار می‌برد تا رتبه سایت مورد نظر در نتایج موتورهای جستجو به رتبه‌های بالاتر، یا صفحه اول نتایج، آورده شود. این تکنیک‌ها از الگوریتم‌های موتورهای جستجو به نفع خود استفاده می‌کنند و سایت خود را طوری طراحی می‌کنند که در اثر ورود کلیدواژه‌های خاصی بیشترین امتیاز را بگیرد و در صفحه اول ظاهر شود. اگرچه الگوریتم‌های رتبه‌بندی موتورهای جستجوی پر طرفدار، از دسترس عموم مخفی است، اما همیشه راه‌هایی برای حدس زدن این الگوریتم‌ها وجود دارد. این تکنیک‌ها موتورهای جستجو را گمراه می‌کنند و این باعث می‌شود تا سایت‌های بی‌اهمیت هم‌سطح سایت‌های مهم نمایش داده شوند، اما با این وجود، موتورهای جستجو علاقه‌ای به مقابله با این مساله را ندارند.

در این گزارش در کنار مسایل مربوط به بازیابی اطلاعات، بطور خلاصه به مبحث SEO خواهیم پرداخت؛ چرا که مسایل مطرح شده در آن بی‌ارتباط با الگوریتم‌های رتبه‌بندی نیستند و می‌توانند برای طراحان موتور جستجو نیز مفید باشند.

در ادامه این گزارش در بخش 2 ما به بیان پیش‌نیازهای لازم از جمله تشریح ساختار موتور جستجو خواهیم پرداخت. در این بخش پارامترهای بهینه‌سازی معرفی شده و در مورد آنها بحث می‌شود. در بخش 3 در مورد الگوریتم‌ها و پارامترهای رتبه‌بندی صحبت می‌کنیم. در بخش‌های بعدی به مسایلی که به نحوی با بهینه‌سازی فارسی سروکار دارند می‌پردازیم.

در بخش 4 به معرفی کلمات عمومی فارسی می‌پردازیم. بخش 5 اختصاص دارد به ریشه‌یابی در فارسی. در بخش 6 بازیابی تحمل‌پذیر را بررسی نموده و در بخش 7 رفتار کاربر را مورد توجه قرار می‌دهیم. بخش 8 به متاجستجوگر تخصیص یافته و در نهایت بخش 9 به بررسی چگونگی بهینه‌سازی برای موتور جستجو می‌پردازد.

2. پیش‌نیازها

در بازیابی اطلاعات (Information Retrieval) هدف این است تا اسناد یا رکوردهای مرتبط با نیاز کاربر از میان انبوه رکوردهای اطلاعاتی استخراج شوند. کاربر نیاز خود را توسط تعدادی از کلمات کلیدی به موتور جستجو می‌دهد و موتور جستجو با گرفتن پرسش (query) از کاربر، مستنداتی را که می‌تواند پاسخی به نیاز اطلاعاتی وی باشند، به کاربر نمایش می‌دهد. جمع‌آوری و نگهداری مستندات قابل دستیابی در یک ساختار بهینه، تشخیص مشابه‌ترین سند به پرسش کاربر و نمایش مستندات به ترتیب میزان مرتبط بودن، از نکات اصلی در معماری یک موتور جستجو است.

موتور جستجو در اساس از سه بخش تشکیل شده است.

- بخش اول «خزنده وب» است که وظیفه‌ی گردآوری مجموعه اسناد موجود را به عهده دارد.
- بخش دوم «نمایه‌ساز» (Indexer) موتور جستجو است که مجموعه اسناد کاوش شده توسط کاوشگر را به نمایه‌های (Index) قابل استفاده تبدیل می‌نماید.
- بخش سوم، «مدل‌های بازیابی اطلاعات و الگوریتم رتبه‌بندی» است که هسته اصلی موتور جستجو را تشکیل می‌دهد. منظور از رتبه‌بندی، اولویت در نمایش مستندات و صفحات بازیابی شده است.

در ادامه، هر کدام از بخش‌ها را بررسی می‌کنیم.

2-1 خزنده وب

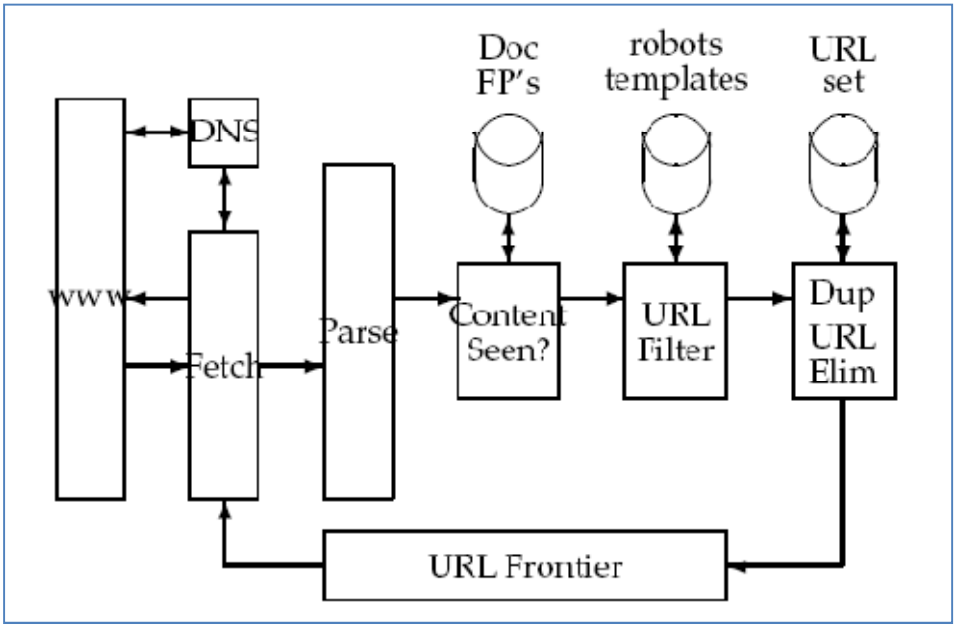
کاوشگر وب وظیفه‌ی انتقال صفحات از وب به موتور جستجو را به عهده دارد. برای این منظور از نرم‌افزاری موسوم به خزنده (crawler) استفاده می‌نماید. این قسمت، نرم‌افزاری است که به صفحات مختلف وب سر می‌زند و اطلاعات مورد نیاز موتور جستجوگر را جمع‌آوری می‌کند و آنرا در اختیار سایر بخش‌ها قرار می‌دهد.

در وقاع خزنده وب در درون یک گراف بزرگ از اتصالات بین صفحات پیش می‌رود و صفحات جدید را دریافت کرده و محتوان آنرا استخراج کرده و به سایر بخش‌ها می‌دهد. در این پیمایش خزنده نباید

URLهای تکراری را دریافت کند لذا باید لیست صفحات مرور شده در جایی نگهداری شوند. در واقع خزنده از یک لیست ابتدایی از URLها شروع کرده و به ترتیب پیش می‌رود. هر URLی که دریافت شد حذف می‌شود و URLهای موجود در آن صفحه به لیست اضافه می‌شوند. به این لیست « URL Frontier» می‌گویند.

از نظر تئوری، کاوشگر وب می‌تواند از یک صفحه در وب شروع کند، تمامی پیوندهای آن صفحه را بگیرد و آنها را به نوبت کاوش نماید و این کار را آنقدر ادامه دهد تا تمامی صفحات اینترنت کاوش شوند. اما مشکل این ایده در عدم دستیابی به تمام صفحات وب از یک نقطه‌ی شروع است، زیرا بسیاری از صفحات، به صفحات دیگر پیوندی ندارند، بنابراین کاوشگر قبلا توسط موتورهای جستجو برای آدرس‌های خاصی برنامه‌ریزی می‌شود.

خزنده وب معمولا از چندین ریسمان (thread) تشکیل شده است و این ریسمان‌ها به صورت همروند کار می‌کنند (در معماری‌های توزیع شده به صورت توزیع‌شده عمل می‌کنند). در شکل 1 معماری یک خزنده ساده نشان داده شده است [4] .



شکل 1 - معماری یک خزنده ساده [4]

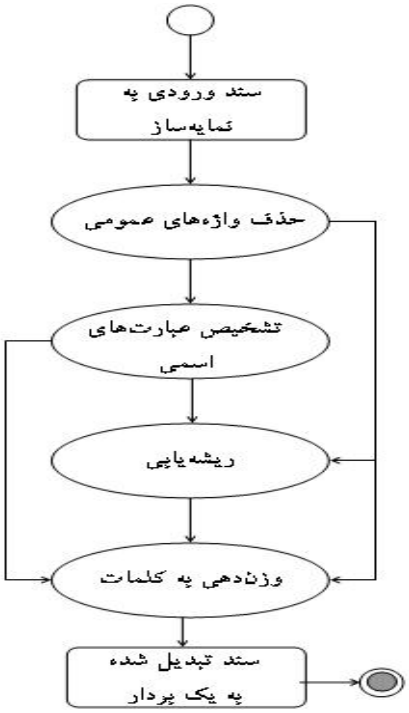
2-2 نمایه‌ساز

تمام اطلاعات جمع‌آوری شده توسط کاوشگر وب بعد از طی مراحل ذخیره‌سازی در مخزن، در اختیار نمایه‌ساز قرار می‌گیرد. در این بخش، اطلاعات ارسالی مورد تجزیه و تحلیل قرار می‌گیرد. تحلیل صفحات به این معنی است که اطلاعات از کدام صفحه هستند، کلمات کلیدی صفحه کدامند، وزن هر یک چقدر است، واژه‌ها در کدام قسمت صفحه به کار رفته‌اند و ... . در حقیقت نمایه‌ساز، صفحات را به پارامترهای آن تجزیه می‌کند و تمام این پارامترها را به یک مقیاس عددی تبدیل می‌کند تا سیستم رتبه‌بندی بتواند پارامترهای مختلف صفحات را با هم مقایسه کند.

نمایه‌سازی، فرایند تحلیل محتوای اطلاعاتی سند به منظور استخراج کلید واژه‌ها به همراه ارزش آن با زبان ویژه نظام نمایه‌سازی است. از آنجا که حضور همه‌ی کلمات متن در نمایه‌سازی سربار زیادی برای سیستم دارد، به نظر می رسد یکی از مشکل‌ترین فعالیت ها در روند نمایه سازی انتخاب کلید واژه‌هایی است که نشان‌دهنده‌ی محتویات سند باشد. هنگام ذخیره اطلاعات به صورت الکترونیکی باید براساس ویژگی های هر زبان، از قواعد و دستور زبان خاصی پیروی کرد تا تشخیص محتوا به گونه‌ای صحیح انجام گیرد، لذا ضرورت تحقیق در مورد نمایه‌سازی خودکار متون فارسی به تفاوت نحوه‌ی ساخت و نیز بکارگیری واژگان در متون این زبان در مقایسه با سایر زبانهای طبیعی برمی‌گردد.

کلمات کلیدی، عناصر بسیار مهمی در جست و جو و دسترسی به اطلاعات هستند. آن‌ها می‌توانند به عنوان مجموعه‌ی کلمات (یک کلمه یا مجموعه‌ای از کلمات) تشریح‌کننده‌ی سند در طی عملیات جست و جو مد نظر قرار گیرند. به عبارت دیگر، هر عبارت مهمی که محتویات داخل سند را تشریح کند، کلمه کلیدی گفته می‌شود.

برای استخراج کلمات کلیدی یک سری پیش‌پردازش‌هایی باید روی متن باید انجام بگیرد. یکی از این پیش‌پردازش‌ها، تعیین کلمات است. معمولا برای تعیین‌کردن کلمات از فضای خالی، علامات آخر جمله استفاده می‌کنند. در زبان فارسی استفاده از فضای خالی می‌تواند مشکل‌ساز شود، چون بعضی از کلمات فارسی چندبخشی هستند و ممکن است با این مکانیزم یک‌کلمه، چندین کلمه متمایز تشخیص داده شود.



شکل 2 - نمودار فعالیت برای نمایه‌سازی متن

برای ایجاد نمایه مناسب از متن معمولا مجموعه فعالیت‌ها مطرح شده در نمودار فعالیت شکل 2 انجام می‌گیرد [5]. البته لازم به ذکر است که برخی از فعالیت‌های مطرح شده در این شکل در برخی نمایه‌سازی‌ها انجام نمی‌شود. عمده فعالیت‌هایی که در نمایه‌سازی انجام می‌شود عبارتند از:

- حذف واژه‌های عمومی (stopword)
- استخراج عبارت‌های اسمی (noun phrase)
- ریشه‌یابی (stemming)
- وزن‌دهی به واژه‌ها و عبارت‌ها
- استخراج کلمات

در زیر هر کدام از موارد فوق توضیح داده شده‌اند.

2-2-1 حذف واژه‌های عمومی

واژه‌های عمومی، کلماتی هستند که با تکرار بالایی در متون وجود دارند و در ارزیابی متن تاثیر مثبت ندارند یا به اصطلاح در تفکیک (discrimination) نقش ندارند. به منظور کاهش حجم پردازش در اولین فعالیت، این واژه‌ها را از مجموعه کلمات سند، حذف می‌کنیم.

2-2-2 استخراج عبارت‌های اسمی

عبارت‌های اسمی، واژه‌های ترکیبی از دو یا چند اسم هستند که در کنار هم به کار می‌روند و به صورت یک عبارت معرفی می‌شوند. استخراج عبارت‌های اسمی باعث بالا بردن دقت بازیابی (recall) می‌شود.

2-2-3 ریشه‌یابی

بسیاری از کلمات به کار رفته در متن، حالت‌های دستوری متفاوتی از یک ریشه هستند. به منظور کاهش حجم نمایه و بالا بردن معیار بازیابی، معمولا از ریشه کلمات به جای حالت‌های دستوری متفاوت آنها استفاده می‌شود.

2-2-4 وزن‌دهی به واژه‌ها و عبارت‌ها

یکی از موارد مهم در نمایه‌سازی که نقش کلمات را از نظر میزان تاثیر آنها به عنوان کلمات کلیدی متن مشخص می‌کند، وزن کلمه است. در این مرحله با استفاده از الگوهای مختلف وزن‌دهی، به هر کلمه یا عبارت استخراج شده وزنی نسبت داده می‌شود. این وزن بیانگر میزان تاثیر کلمه در موضوع اصلی متن در مقایسه با سایر کلمات به کار رفته در متن است.

2-2-5 استخراج کلمات

در نهایت کلمات و عبارت‌های استخراج شده به همراه وزن آنها به صورت نمایه، معرفی می‌شود.

پس از تعیین کلمات، کلمات عمومی را حذف کرده و بقیه متن را ریشه‌یابی می‌کنیم و سپس کلمات را وزن‌دهی کرده و تبدیل به بردار می‌کنیم و با اعمال آستانه، لیست کلمات کلیدی استخراج می‌شود. در ادامه، مراحل استخراج کلمات کلیدی را تشریح می‌کنیم.

نمایه‌ساز یکی از قسمت‌های اساسی و مهم در موتورهای جستجو است بطوریکه نمایه‌سازی مناسب می‌توانند تاثیر قابل توجهی در بالا بردن کارایی موتور جستجو داشته باشد. متاسفانه پیشرفت واحدهای تحقیقاتی در نمایه‌سازی خودکار متون فارسی کند بوده است. به این معنی که برخلاف متون غیرفارسی، نمایه‌ساز خودکاری که با هزینه‌ی مناسب قابل دستیابی باشد و بتواند یک متن فارسی را به نمایه‌های آن تجزیه کند در دسترس نیست. در [6] یک نمایه‌ساز خودکار فارسی با قابلیت ریشه‌یابی کلمات، طراحی و پیاده‌سازی شده است. در [7] نیز یک پیاده‌سازی برای نمایه‌ساز متون فارسی انجام شده است.

2-3 مدل‌های بازیابی

اولین گام جهت طراحی سیستم بازیابی اطلاعات این است که مدلی برای توصیف و تعیین مشابهت‌های موجود میان اطلاعاتی که در اختیار دارد با نیازهای اطلاعاتی کاربر تعریف کند. در این بخش مدل یا مدل‌های مورد استفاده‌ی موتور جستجوگر، برای بازیابی اطلاعات و رتبه‌بندی آنها بیان می‌شود.

یکی از نکات اصلی که برای کاربر اهمیت زیادی دارد نحوه‌ی رتبه‌بندی نتایج بدست آمده توسط موتور جستجوگر است. تفاوت در کارایی موتورهای جستجو ناشی از الگوریتم‌ها و مدل‌های مختلفی است که در این قسمت از موتور جستجو پیاده‌سازی شده‌اند. یکی دیگر از نکات این مدل‌ها رفتار متفاوت آنها در زبان‌های مختلف و مجموعه اسناد مختلف است. به این معنی که مدل‌های بازیابی اطلاعات که در موتورهای جستجو به منظور یافتن مشابه‌ترین سند به پرسش کاربر از میان اسناد موجود استفاده می‌شود، باید برای زبان‌های متفاوت (انگلیسی، فارسی و ...) پیاده‌سازی وارزیابی شوند تا بتوان برای زبان مقصد بهترین مدل را انتخاب و استفاده کرد.

حاصل تحقیقات گسترده در بازیابی اطلاعات، طراحی و معرفی مدل‌های مختلفی برای سیستم‌های بازیابی اطلاعات است. برخی از مهم‌ترین آنها، مدل فضای برداری (Vector-Space)، دودویی (Binary)، احتمالی-آماری، شبکه عصبی، فازی، N-gram و شبکه‌های استنتاجی هستند. این مدل‌ها با توجه به مجموعه داده‌های مورد استفاده و زبان مقصد کارایی متفاوتی دارند. مدل‌های فوق را می‌توان در سه کلاس زیر طبقه‌بندی کرد:

- مدل‌های جبری: مانند مدل دودویی (Boolean)،
- مدل‌های تئوری مجموعه‌ای: مانند مدل فضای برداری (Vector Space)،
- مدل‌های احتمالی-آماری (Probabilistic Models)

این مدل‌ها با توجه به مجموعه داده‌های مورد استفاده و زبان مقصد کارایی متفاوتی دارند.

2-3-1 مدل دودویی

در مدل دودویی، نیاز اطلاعاتی کاربر به صورت عبارتی منطقی با عملگر های AND، OR و NOT بیان می‌شود و هر سندی که این عبارت در مورد آن صحیح باشد بازیابی می‌شود. مثلاً اگر نیاز اطلاعاتی به صورت Iran AND Oil بیان شود، تمامی اسنادی که کلمه‌ی Iran و Oil را با هم دارند به کاربر نمایش داده می‌شوند. متاسفانه در مدل دودویی سند یا باربط است یا نیست، و هیچ معیاری برای سنجش میزان ربط وجود ندارد. مثلا دو سندی که یکی تماما در باره ایران و نفت بحث می‌کند، و دیگری در مورد اقتصاد جهانی صحبت می‌کند و فقط از نام ایران و نفت به عنوان مثالی در یک جمله استفاده کرده است، از نظر سیستم تفاوتی نیست. در صورتیکه در واقع سند اول بیشتر به نیاز کاربر مربوط است.

استراتژی جست و جوی دوارزشی، اسنادی را بازیابی می‌کند که برای پرس وجو مقدار True را داشته باشند. این فرموله سازی زمانی قابل توجیه است که پرس وجو به صورت کلمات شاخص (کلمات کلیدی) و ترکیب این کلمات با استفاده از عملگرهای منطقی معمول مثل AND, OR, NOT نمایش داده شود.

برای مثال اگر پرس وجو $Q = (K_1 ANd\ k_2) OR (K_3 ANd\ (Not\ K_4))$ باشد، جست و جوی دوارزشی تمام اسنادی را بازیابی خواهد کرد که با استفاده از K_1 و K_2 شاخص شده باشند و هم‌چنین اسنادی که با استفاده از K_3 شاخص شده و با K_4 شاخص نشده باشند را نیز بازیابی خواهد کرد.

2-3-2 مدل برداری

در مدل برداری، هر مستند را به صورت برداری از کلمات در نظر می گیریم و فضایی چند بعدی که ابعاد آنرا کلمات تشکیل می‌دهند ایجاد می کنیم. سپس هر سند در این فضا به صورت یک بردار نمایش داده می‌شود. مولفه‌های این بردار سند، در واقع وزن هایی هستند که نشان می‌دهند هر یک از کلمات چقدر در متمایز کردن آن سند دخیل هستند. در مدل احتمالاتی ، به هر سند احتمالی اختصاص داده می‌شود که مربوط بودن آن مستند را به نیاز کاربر به صورت احتمال بین صفر و یک بیان می‌کند.

در مدل برداری، برای سنجش میزان ربط اسناد و نیاز اطلاعاتی کاربر، سیستم دقیقا به مانند قبل نیاز اطلاعاتی کاربر را هم به فضای چندبعدی از کلمات می‌برد و در نتیجه برای سنجش میزان شباهت میان این دو بردار می‌توان از زاویه‌ای که این دو بردار با هم می سازند استفاده کرد. اسنادی که با نیاز اطلاعاتی کاربر دقیقا هم جهت هستند مسلما نسبت کلماتشان به همان نسبت کلمات نیاز اطلاعاتی است و در نتیجه مرتبط‌تر خواهند بود. برتری این مدل این است که به ما درجه‌ای از ربط را می‌دهد.

مدل فضای برداری پایه‌ای‌ترین مدل در سیستم‌های بازیابی اطلاعات است که در [8] معرفی شد. در این مدل ابتدا سند به برداری تبدیل می‌شود که حاوی کلمات مهم متن سند، به همراه وزن هر کلمه بر اساس میزان تاثیرگذاری کلمه بر محتوی متن در مقایسه با سایر کلمات است. تهیه بردار برای هر سند بر اساس تکنیکی به نام نمایه‌سازی صورت می‌گیرد. در نمایه‌سازی ابتدا کلمات عمومی از متن حذف می‌گردند و کلمات باقی مانده ریشه‌یابی می‌شوند. سپس بر اساس پارامترهای مختلفی مانند تعداد تکرار کلمه در متن، تعداد تکرار کلمه در متن، تعداد تکرار کلمه در اسناد مجموعه و مولفه‌های نرمال‌سازی وزنی به هر کلمه نسبت داده می‌شود. همین فعالیت‌ها برای پرسش کاربر نیز تکرار می‌شود. به این ترتیب هر سند از مجموعه‌ای از کلمات به برداری تبدیل می‌شود که در فضای جدیدی به نام فضای برداری قرار دارد. در این فضا که بسته به تعداد کلمات مجموعه یک فضای n بعدی است، بردار هر سند ترسیم می‌شود. پرسش کاربر نیز بعد از اعمال فعالیت‌های نمایه‌سازی به برداری تبدیل می‌شود که در فضای جدید ترسیم می‌گردد. در این فضا هر سندی که به پرسش کاربر نزدیک‌تر باشد سند مرتبط شناخته می‌شود و بازیابی می‌گردد. معیار نزدیکی در این فضا زاویه‌ای است که بردار پرسش با هر یک از بردارهای سند می‌سازد. این میزان نزدیکی، معمولا با رابطه زیر که به نام مشابهت کسینوسی شناخته می‌شود، محاسبه می‌گردد:

$$sim(q_i, d_j) = \frac{\vec{q}_i . \vec{d}_j}{|\vec{q}_i| \times |\vec{d}_j|} = \frac{\sum_{k=1}^{t} w_{ki} \times w_{kj}}{\sqrt{\sum_{k=1}^{t} w_{ki}^2} \cdot \sqrt{\sum_{k=1}^{t} w_{kj}^2}}$$

در این رابطه q_i بردار پرسش کاربر، d_j بردار سند jام، w_{ki} وزن کلمه k ام در پرسش کاربر و w_{kj} وزن کلمه kام در سند d_j است.

2-3-3 مدل احتمالاتی

در مدل احتمالاتی هم به ازای هر نیاز اطلاعاتی، تمامی اسناد بر اساس احتمال این که این سند با نیاز اطلاعاتی مرتبط باشد، مرتب می‌شوند و لیست اسناد در نهایت به صورت درجه بندی شده (مانند مدل برداری) به کاربر نمایش داده می‌شود به نحوی که اولین سندی که کاربر می بیند از همه بیشتر احتمال دارد که به نیاز او ربط داشته باشد.

بعد از تعریف این مدل، سیستم اکنون آماده است که نیاز اطلاعاتی کاربر را دریافت کند. معمولا کاربران نیاز اطلاعاتی خود را در قالب چندین کلمه یا عبارات معمولی به سیستم بیان می‌کنند. سیستم سپس بر اساس مدلی که اطلاعات را در آن مدل کرده است، میزان ربط هر سند را با نیاز اطلاعاتی کاربر محاسبه می‌کند و آن سندهایی را که از همه باربط تر تشخیص داده شده اند به عنوان خروجی باز می گرداند.

2-4 معیارهای ارزیابی

برای ارزیابی سیستم‌های بازیابی اطلاعات و یا مقایسه دو سیستم بازیابی اطلاعات، معیارهایی وجود دارند. این معیارها را در این بخش معرفی می‌کنیم. در نهایت هدف طراح سیستم بهبود این پارامترها است. در واقع در این گزارش، منظور از بهینه‌سازی، بهبود این پارامترها می‌باشد.

2-4-1 دقت

بیانگر قابلیت سیستم برای ارائه فقط موارد مربوط به درخواست کاربر است. این مقدار نسبت مستقیمی با میزان تعیین‌کنندگی کلمات در متن دارد. برای محاسبه پارامتر دقت بر اساس رابطه، نسبت تعداد اسناد مرتبط بازیابی شده بر کل اسناد بازیابی شده برای پرس و جو، محاسبه می‌شود.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

در برخی ترجمه‌ها [9] از این معیار به عنوان «مانعیت» نام برده شده است.

2-4-2 بازخوانی

بیانگر قابلیت سیستم برای ارائه موارد مربوط به درخواست کاربر است. این مقدار نسبت مستقیمی با میزان دربرگیری کلمات در متن دارد. برای محاسبه بازخوانی نسبت تعداد اسناد مرتبط بازیابی شده بر کل اسناد مرتبط با پرس و جو محاسبه می‌شود. رابطه‌ی زیر برای محاسبه پارامتر بازخوانی به کار می‌رود.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

در برخی ترجمه‌ها [9] از این معیار به عنوان «جامعیت» نام برده شده است.

2-4-3 پارامتر Fall-out

این پارامتر بیانگر نسبت میزان خطا می‌باشد. و با محاسبه نسبت تعداد اسناد نامرتبط بازیابی شده بر کل اسناد نامرتبط با پرس و جو محاسبه می‌شود. رابطه زیر محاسبه پارامتر Fall-out را نشان می‌دهد.

$$\text{fall-out} = \frac{|\{\text{non-relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{non-relevant documents}\}|}$$

4-4-2 پارامتر $F_{measure}$

در حقیقت این پارامتر میانگین هارمونیک پارامترهای بازخوانی و دقت می‌باشد. هدف در سیستم‌های بازیابی اطلاعات بیشینه کردن این معیار می‌باشد. $F_{measure}$ بر اساس رابطه زیر محاسبه می‌شود.

$$F = \frac{2 \cdot \mathrm{precision} \cdot \mathrm{recall}}{(\mathrm{precision} + \mathrm{recall})}.$$

مقدار دو پارامتر «دقت» و «بازخوانی» غالبا نسبت معکوس با هم دارند و بهبود یکی باعث افت دیگری می‌شود. با توجه به این که عملیات ارزیابی با استفاده از مجموعه‌ای از پرس و جوها انجام می‌شود، روشی به عنوان روش برتر در نظر گرفته می‌شود که برای مجموعه پرس و جوها میانگین بهتری داشته باشد. به عنوان مثال، اگر مقدار هر دو پارامتر در پرس و جوی A بیش‌تر از پرس و جوی B باشد، نتایج پرس و جوی A بهتر خواهد بود.

برای سنجش کارآمدی بازیابی اطلاعات، میزان جامعیت، مانعیت و ریزش (Fallout) معیارهای عملکرد کارآمدی نظام‌های بازیابی اطلاعات به شمار می‌روند.

3. الگوریتم‌های رتبه‌بندی

امروزه به قدری الگوریتم‌های رتبه‌بندی پیچیده شده‌اند که تعداد پارامترهای رتبه‌بندی در یک جستجو به هزاران مورد می‌رسد. لذا تحلیل و بررسی این الگوریتم‌ها در این تحقیق ممکن نیست. در اینجا به بیان برخی از مشهورترین پارامترهای آنها می‌پردازیم.

3-1 پارامترهای رتبه‌بندی

در رتبه‌بندی پارامترهای زیادی دخیل هستند. این پارامترها را می‌توان در 3 دسته طبقه‌بندی کرد:

1. کلمات (تعداد و موقعیت)
2. لینک‌ها (تعداد ارجاعات)
3. آمار کاربران (کلیک یا رای کاربر)

بیشترین پارامترهای رتبه‌بندی مربوط می‌شوند به کلمات. در بخش بعد روش‌های وزن‌دهی به کلمات و عبارات را بررسی می‌کنیم. تعداد و تنوع این پارامترها بسیار زیاد است در این تحقیق فقط به بیان کلی برخی از پارامترهای مهم بسنده می‌کنیم.

هرچه تعداد لینک‌های داده شده به یک سایت در اینترنت بیشتر باشد، نشان دهنده اهمیت آن سایت می‌باشند. پارامترهای مربوط به لینک‌های اینترنت تاثیر به سزایی در رتبه یک سایت دارند. اولین بار در گوگل این پارامتر استفاده شد [10]. در این مورد در بخش 9 صحبت می‌کنیم.

اخیرا تکنیک‌های پیشرفته‌ای برای رتبه‌بندی ابداع شده‌اند که از رفتار کاربر به عنوان یک پارامتر استفاده می‌کنند. تعداد کلیک‌هایی که بر روی یک لینک در نتایج جستجو می‌شود می‌تواند نشان دهنده ارتباط بیشتر آن با کلمه جستجو باشد. همچنین اخیرا در موتور جستجوی گوگل امکان نظردهی بر نتایج توسط کاربر وجود دارد. در این مورد در بخش 7 صحبت می‌کنیم.

در ادامه در مورد پارامترهای رتبه‌بندی مربوط به کلمات صحبت می‌کنیم.

3-2 روش‌های وزن‌دهی به کلمات و عبارات

در این مرحله با استفاده از الگوهای مختلف، به هر یک از واژه‌های استخراج‌شده وزنی نسبت داده می‌شود. این وزن، بیانگر میزان تاثیر کلمه به موضوع متن در مقایسه با سایر کلمات به کار رفته است. وزن‌دهی به کلمات بر اساس اهمیت آن‌ها در متن انجام می‌گیرد. اهمیت کلمات را می‌توان بر پایه شرایط زیر مشخص کرد [4, 5]:

- وزن آماری کلمه: بر پایه‌ی تکرار کلمات در متن، بر پایه‌ی توزیع کلمات در متن
- مکان قرارگیری کلمه در متن: اهمیت کلماتی که در عنوان متن، زیر عنوان، بدنه متن و یا چکیده متن باشد متفاوت است. می‌توان از موقعیت کلمه برای ارزش دهی به کلمه استفاده کرد.
- مفهوم هر کلمه، که بیانگر ارتباط کلمه با کلمه‌های دیگر است (کلمات مترادف و متضاد).
- کاربرد خاص کلمه: مثلاً، اسامی در سیستمی که به دنبال اسامی خاص می‌گردد دارای اهمیت بیش‌تر است.

در بسیاری از روش‌های معمول، برای استخراج کلمات کلیدی از وزن‌دهی به کلمات بر اساس معیار فراوانی کلمات در متن استفاده می‌شود. فراوانی کلمات نیز به دو صورت زیر در اسناد بررسی می‌شود:

- فراوانی مطلق (Absolute Frequency)
- فراوانی نسبی (Relative Frtequency)

در فراوانی مطلق، فقط تعداد تکرار کلمه در یک سند سنجیده می‌شود ولی در فراوانی نسبی، تعداد تکرار کلمه در یک سند به همراه تکرار سایر کلمات در آن سند و تعداد تکرار کلمه در سایر اسناد مورد ارزیابی قرار می‌گیرد.

3-2-1 ارزیابی کلمات کلیدی

برای ارزیابی کلمات کلیدی استخراج شده از متن که از آستانه تعیین شده برای وزندهی عبور می‌کنند، باید معیارهای زیر را در نظر داشت:

3-2-1-1 جامعیت (Exhaustivity)

بیانگر میزانی است که همه کلمات متن در استخراج کلمات کلیدی ظاهر شده‌اند. در واقع هر چه کلمات بیش‌تری از متن در استخراج کلمات کلیدی به کار روند، میزان جامعیت کلمات کلیدی و نیز نسبت آیتم‌هایی که با آن می‌توانند بازیابی شوند زیاد خواهد بود.

3-2-1-2 تعیین کنندگی (Specifity)

یعنی هر کلمه‌ی کلیدی تا چه حد دقیق، متن‌های مربوط را مشخص می‌کند. کلمه کلیدی که دارای سطح بالایی از تعیین‌کنندگی است، موارد نامربوط را به کلمات به کار رفته در آن نگاشت نمی‌کند.

3-2-2 پارامترهای وزن‌دهی

پارامترهای وزن‌دهی به کلمات زیاد می‌باشند که ما در زیر برخی از آن‌ها را برمی‌شماریم.

3-2-2-1 پارامتر *tf.idf*

یکی از پرکاربردترین روابط در حوزه بازیابی اطلاعات، پارامتر *tf.idf* می‌باشد [5]، که از حاصل‌ضرب فراوانی کلمه در فراوانی معکوس سند به دست می‌آید. این روش یک روش مبتنی بر چند سند می‌باشد، که در آن منظور از فراوانی کلمه، فقط تعداد تکرار کلمه در یک سند خاص است. هم‌چنین منظور از فراوانی معکوس سند، تعداد اسنادی است که این کلمه خاص در آن اسناد ظاهر شده است. دلیل مقبولیت این روش نسبت به سایر روش‌ها را می‌توان با توجه به سهولت در استفاده از این روش، محاسبات کم و نتایج قابل قبول دانست.

3-2-2-2 پارامتر سیگنال و نویز

در این روش از تئوری اطلاعات استفاده شده است [5]. در این تئوری، هر چه احتمال رخداد کلمه بیش‌تر باشد، بار اطلاعاتی کم‌تری برای آن در نظر می‌گیرند. کلمات با اهمیت که دارای توزیع متمرکز هستند، یعنی تنها در بعضی از اسناد متنی ظاهر شده اند میزان نویز کمی دارند.

3-2-2-3 پارامتر مقدار تمایز

در این روش، برای وزن‌دهی کلمات از قدرت تمییزدهندگی کلمات بین اسناد مختلف استفاده می‌شود [4, 5]. مقدار تمایز (Discrimination Value) را با استفاده از معیارهای مشابهت محاسبه می‌کنند. استفاده از کلمه‌ای از سند به عنوان کلمه‌ی کلیدی که باعث کاهش مشابهت این سند با سایر اسناد می‌شود. هر چه مقدار تمایز بیش‌تر باشد، بیانگر تخصصی‌تر بودن این کلمه و اهمیت بیش‌تر آن در متمایز کردن سندی که در آن ظاهر شده، از سایر اسناد است. در واقع انتخاب کلمه‌ای از یک سند با مقدار تمایز زیاد به عنوان کلمه کلیدی، باعث کاهش شباهت این سند با سایر اسناد می‌شود. برای تعریف شباهت بین دو سند متنی از معیارهای مشابهت استفاده می‌شود.

3-2-3 وزن‌دهی در یک نمایه‌ساز فارسی

در [7] یک نمایه‌سازی متون فارسی معرفی شده است. در اولین گام کلمات عمومی بر اساس یک لیست از قبل آماده شده حذف می‌شوند. برای کلمات عمومی یک لیست 180 کلمه‌ای بر اساس تعداد تکرار کلمه در سند ایجاد شده است. برای ریشه‌یابی کلمات از یک روش مبتنی بر حذف پس‌وند و پیش‌وند استفاده کرده اند. نمایه‌ساز سینا از چهار روش وزن‌دهی tfidf , Lnu , ltn, ntc استفاده کرده است. برای پیکره از 450 متن شامل چکیده مقالات مرتبط با کامپیوتر استفاده شده است. الگوهایی که برای وزن‌دهی به واژگان در نمایه‌ساز *سینا* [7] پیاده‌سازی شده‌اند در جدول زیر آمده‌اند:

نام روش	الگوی وزن‌دهی
tf.idf	
Lnu	
ltn	
ntc	

پارامترهای به کار رفته در جدول به شرح زیر می‌باشند:

N: تعداد کل سندها می‌باشد.

idf: معکوس تعداد اسنادی است که کلمه در آنها به کار رفته است.

NUT: تعداد واژه‌های واحد در سند می‌باشد.

Slope: شیب منحنی در نرمالسازی اسناد مجموعه است.

Pivot: میانگین تعداد واژه‌های واحد در مجموعه مستندات می‌باشد.

با مقایسه‌ای که در استفاده از روشهای مختلف وزن‌دهی و نتایج از حاصل از آن در [7] انجام شده است، مشخص گردید که دو روش *Lnu* و *tf.idf* سایر الگوها مناسب تر عمل می‌کند. دلیل بهینه‌تر بودن این دو روش تاثیر پارامترهای سایر واژه‌های سند در وزن‌دهی به یک واژه است. به عبارت دیگر وزن هر واژه علاوه بر پارامترهای مربوط به آن نسبت به سایر واژه های موجود در سند وزن‌دهی می‌شود. برای ارزیابی از معیارهای بازخوانی و دقت استفاده کرده‌اند. میانگین پارامتر دقت با ریشه‌یابی 66% و بدون ریشه‌یابی 54% بوده است.

4. کلمات عمومی فارسی

به عنوان اولین فعالیت در روند نمایه‌سازی، واژه‌های عمومی در متن ورودی حذف می‌گردند. با توجه به شیوه استفاده از واژگان زبان، بعضی واژه‌ها در همه متون با تکرار زیاد وجود دارند. این گونه واژه‌ها، واژه‌های عمومی زبان نامیده می‌شوند. در واقع این واژه‌ها، واژه‌هایی مثل ضمایر، قیود، حروف اضافه و ربط هستند که در بازیابی، تاثیری بر ارزش محتوایی سند ندارد. واژه‌های عمومی زبان یا توسط زبان شناسان معرفی می‌شود و یا بر اساس نرخ تکرار در هر سند بدست می‌آیند.

بعضی از کلمات در همه‌ی متون با فراوانی زیاد وجود دارند که ارزش محتوایی ندارند، مثل ضمایر، قیود، حروف اضافه و ربط و بعضی از افعال پرتکرار. به این کلمات، کلمات عمومی گفته می‌شود. با حذف کلمات عمومی در متن کاوی آماری میزان محاسبات کم شده و کارآیی روش‌ها نیز بیش‌تر می‌شود. در جدول زیر برخی کلمات عمومی فارسی آمده‌اند.

برخی از کلمات عمومی فارسی
امروز، گفتم، اکنون، خواهند، آر، آقا، آقای، آقایان، آمد، آمده، آن، آنان، آن‌جا، آنچه، آنکه، آن‌ها، آیا، اخیر، از، است، اسلامی، اش، افزود، اگر، اگرچه، الا، البته، الی، ام، اما، امروزه، اند، اندی، او، اولین، ای، ایران، ایشان، ایم، این، این‌جا، اینکه، این گونه، با، بااین، بااینکه، بار، باز، باشد، باشید، باشیم، بالاخره، باید، بجز، بدهید، بدون، بر، برای، براین، برخی، برلزوم، بسیار، بسیاری، به طور، بعد، بکنید، بگذاریم، بگوییم، بلکه، بماند، به، بود، بودند، بوده، بی، بیش، بین، پس، پی، پیش، تا، تر، تری، تمامی، تو، توسط، توی، جا، جز، ، چرا، چنان، چند، چنین، چه، چو، چون، چونکه، حال، حالی، حالی که، حتی، حدود، حقیقتا، خانم، خانمها، خواهد، خود، خودم، خودمان، خویش، داخل، داد، دادم، دادند، داده، دار، دارای، دارد، دارند، داریم، داشت، داشته، داشته، داند، دانند، در، درآن، دراین، درباره، دربر، دربعد، دربین، درپی، درجای، درحال، درحالی، درحالی که، دردو، درکل، درین، دور، دیگر، را، رسیده، رو، روی، زدند، ساعت، سر، سعی، سو، سوی، شامل، شد، شدن، شدند، شده، شما، شود، طی، علیرغم، علیه، غیر، فقط، کرد، کردم، کردن، کردند، کرده، کنار، کند، کنم، کنند، کنید، کنیم، که، گذاری، گرچه، گردند، گرفت، گرفته، گفت، گفتند، گفته، لزوم، ما، مانند، متوالی، مثلا، من، می، می‌شود، میان، میتواند، میخواهیم، میدانند، میرسد، میشود، میکنم، میکنند، ندارد، ندارم، ندارند، نداشته، نشدند، نظر، نماید، نموده، نمی، نمیکند، نیز، نیست، نیستند، ها، های، هایی، هر، هریک، هست، هستم، هستند، هستید، هستیم، هم، همان، همه، همین، هنوز، هیچ، و، وجود، ولی، وی، یا، یافت، یعنی، یکدیگر، یکم، یکی

در [7]، برای واژه‌های عمومی زبان فارسی یک فهرست 180 واژه‌ای بر اساس تعداد تکرار واژه در سند و بالاتر بودن نرخ تکرار از آستانه تعریف شده، تهیه گردیده است. در اولین گام کلمات متن ورودی بعـد از مقایسه با این واژه‌ها در صورت برابری حذف می‌شوند.

5. ریشه‌یابی در فارسی

در این بخش به ریشه‌یابی یا «ریخت‌شناسی» کلمات در زبان فارسی می‌پردازیم. ریخت‌شناسی بخشی از علم پردازش زبان طبیعی است که به ساختارهای کلمات و ریشه‌یابی واژگان می‌پردازد. به عمل بیرون‌آوردن ریشه اصلی یک واژه؛ ریشه‌یابی (stemming) گویند. در واقع ریخت‌شناسی به علم شناختن اجزای معنی‌دار از یک واژه گویند که آن واژه را می‌سازد؛ به این اجزای معنی‌دار تکواژ (morpheme) گویند.

در ریخت‌شناسی، واژه‌ها به دو طریق بسط می‌یابند: تصریف (inflection) و اشتقاق (derivation). در تصریف، از ترکیب یک واژه با اجزای دستوری دیگر، واژه‌ای جدید در همان نوع و ردهٔ واژهٔ قبلی ایجاد می‌گردد. به عنوان مثال علامت جمع «ها» در فارسی که با اضافه‌کردنش به هر اسمی یک اسم جدید به وجود می‌آید؛ مثلاً واژهٔ «کتاب» با اضافه‌شدن «ها» به «کتاب‌ها» تبدیل می‌شود که در این صورت، هم کتاب از نوع دستوری اسم است و هم کتاب‌ها. روش دوم، روش اشتقاق است. در اشتقاق با افزودن یک جز دستوری به یک واژه، یک واژه در رده جدیدی به وجود می‌آید. به عنوان مثال اگر تکواژِ «-ِش» را به واژهٔ مصدری «کن» اضافه کنیم، واژهٔ کنش به وجود می‌آید که واژه جدید دیگر از نوع مصدر نیست و یک اسم است. در ریشه‌یابی موتورهای جستجو ما فقط به ریشه‌یابی تصریفی می‌پردازیم.

5-1 طبقه‌بندی روش‌های ریشه‌یابی

الگویتم‌های گوناگون زیادی برای ریشه‌یابی در زبان‌های مختلف از جمله زبان فارسی برای ریشه‌یابی داده شده است. این الگوریتم‌ها را می‌توانیم با توجه به نحوه عملکرد و میزان دقت آنها در دسته‌های جداگانه طبقه‌بندی کنیم. این دسته‌ها را در ادامه بیان می‌کنیم.

5-1-1 ریشه‌یاب جدولی

ساده‌ترین روشی که برای ریشه‌یابی به نظر می‌رسد، نگهداری ریشه هر واژه در یک جدول است. در این روش با جستجوی واژه در این جدول، ریشه واژه مشخص می گردد. هر چند از این روش می توان نتایج خوبی گرفت، اما نگهداری این جدول سربار زیادی برای سیستم خواهد داشت و تنها محدود به کلمات از پیش تعیین شده هستیم.

ساده‌ترین روش از جنبه پیاده‌سازی در بین ریشه یا بن‌ها است. در این روش ریشه کلمات در یک جدول نگهداری می‌شود. و برای یافتن ریشه، کلمه مورد نظر را از جدول جست و جو کرده و ریشه متناظر را مشخص می‌کند. در واقع این روش بیش‌تر شبیه یک عمل جست و جو است تا ریشه‌یابی. ریشه‌یاب جدولی بهترین نتایج را در بین ریشه‌یاب‌ها دارند. ولی از معایب آن سربار زیاد برای نگهداری جدول و هم‌چنین در دسترس نبودن این جدول برای واژگان فارسی می‌باشد.

5-1-2 ریشه‌یابی بر اساس الگوریتم پورتر

روش پورتر (Porter) یک روش توانمند و در عین حال یکی از قدیمی‌ترین روش‌های ریشه‌یابی در زبان انگلیسی است. این روش بر پایه زبان‌شناسی و دسته‌بندی کلمه‌ها به کمک واج‌ها و هجاها بنا نهاده شده است. پس از آن وندهای کلمات درون گردایه به طور خودکار برداشته می‌شوند [11, 12].

دسته‌ای دیگر از کارها، شامل الگوریتم‌های ریشه‌یابی هستند که بر اساس قوانین ریخت‌شناسی زبان مربوطه کار می‌کنند [13-15]. در این الگوریتم‌ها، برنامه از درون ساختاری تصمیم‌گیرنده مانند یک فلوچارت عبور کرده و با افزودن و کاستن وندها با رعایت قواعد املایی و دستوری، سعی در یافتن ریشه کلمات یا بطور خاص افعال دارد. این کارها عمدتا مشابه الگوریتم پورتر هستند که برای زبان انگلیسی طراحی شده است. مشکل این دسته از الگوریتم‌ها برای کلمات جدا از هم است. با توجه به اینکه در فارسی مرز دقیق کلمات مشخص نیست، برای کلمات چندپاره این روش‌ها خوب عمل نمی‌کنند. در [16] یک سیستم ساده مبتنی بر قانون برای افعال فارسی ارایه شده است.

این نوع ریشه‌یاب‌ها با حذف پیشوند و یا پسوند به ریشه واژه می‌رسند. الگوریتم این ریشه‌یاب‌ها از تعدادی قوانین تشکیل می‌شود که با یافتن اولین قانون مناسب و ممکن برای حذف پیشوند و یا پسوند، آن قانون مورد استفاده قرار می‌گیرد. اکثر ریشه‌یاب‌های موجود از این نوع، اقدام به زدودن بزرگترین

دنباله ممکن از حروف واژه بر طبق قوانین می‌نماید. این فرایند آنقدر ادامه می یابد تا هیچ حرف دیگری نتواند زدوده شود.

زبان فارسی برای اشتقاق و ساخت کلمات از الحاق پسوندها و پیشوندها استفاده می‌کند. بنابراین ریشه‌یابی در زبان فارسی فرایند حذف این الحاقات است. از طرفی متأسفانه قانون مدون و کلی برای ساخت واژگان اشتقاقی زبان فارسی وجود ندارد و در مورد هر پسوند و پیشوند استثناهای زیادی یافت می‌شود که کار ریشه‌یابی را بس مشکل می‌کند. بنابراین در مورد هر قانون باید استثنائات آن را شناسایی و نگهداری کرد، تا دقت ریشه یاب خودکار بهبود یابد.

5-1-3 ریشه‌یابی بر اساس مدل حالت متناهی

ریخت‌شناسی بر اساس مدل حالت-متناهی، روش متداولی است که در [17] و [18] نمونه‌ای از آن را می‌توان دید. اساس کار آنها بر یک مدل زبان جهانی است که در [19] ارایه شده است. تعریف الگو بر اساس عبارات منظم انجام می‌شود و پیاده سازی آن بر اساس مدل ماشین حالت-متناهی است. طراحی پردازشگر ریخت‌شناسی حالت-متناهی را می‌توان به دو بخش مجزا تقسیم کرد: بخش مربوط به «طرح زبانی» و بخش مربوط به «طرح رایانه‌ای» [17]. منظور از طرح زبانی، ارایه توصیف نظری جامع و کامل و مانع از ریخت‌شناسی افعال در زبان فارسی است. ارایه توصیف جامع از ریخت‌شناسی در زبان فارسی به گونه‌ای که قابل کاربرد در برنامه‌های رایانه‌ای باشد، نخستین گام جهت طراحی برنامه‌های کاربردی است. این توصیف می‌بایست تمام صورت‌های تصریفی فعل در زبان فارسی را ارایه دهد. بخش دوم، طرح رایانه‌ای است. در این بخش نیازهای سخت‌افزاری و نرم‌افزاری پردازشگر تعریف می‌شود، طرح زبانی پیاده‌سازی شده و ویژگی‌ها و ساختار داخلی فایلهای برنامه تشریح می‌گردد.

5-1-4 ریشه‌یابی به کمک روش‌های آماری

در این دسته از روش‌ها یک گردایه‌ی بزرگ از کلمه‌ها با ساخت‌های گوناگون گردآوری می‌شود. هرچه این گردایه بزرگ‌تر و کامل‌تر باشد این ریشه‌یاب‌ها بهتر کار می‌کنند. در این روش تحلیل آماری به کار گرفته می‌شود. با روش آماری وندهایی که در کلمه‌ها تکرار شده اند، شناسایی می‌گردند. این روش به زبان بستگی ندارد و این بزرگ‌ترین برتری این روش می‌باشد. در بیش‌تر زبان‌های هند و اروپایی، اغلب

بر پایه‌ی وند اشتقاق انجام می‌شود. اگر این روش بتواند برای زبان انگلیسی پاسخ شایسته‌ای بدهد؛ گسترش آن به دیگران زبان‌های دسته‌ی هند و اروپایی ساده خواهد بود. این روش با سه مشکل بزرگ روبروست:

- در این روش به یک گردایه‌ی بزرگ از کلمه‌ها نیاز است. این گردایه باید کامل باشد و کلمات درون آن نیز درست باشند. وجود کلمات نادرست در گردایه بر کارایی این ریشه‌یاب اثر بسیار بد می‌گذارد و آن را گمراه می‌کند. گردآوری گردایه‌ی بزرگی از کلمات صد در صد درست فارسی نیز، ناممکن می‌نماید.
- هنوز این روش‌ها در حال آزمایش هستند و کارایی آن‌ها چشم‌گیر نیست.
- این روش‌ها نیاز به رایانه‌های با سرعت زیاد و حافظه بزرگ دارند و اجرای برنامه‌های نوشته شده بر پایه‌ی این روش‌ها بسیار زمانبر است. برای اجرای این روش‌ها با رایانه‌های در دسترس باید تعدادی از آن‌ها با هم موازی شوند و شاید برای یک بار اجرا، چند روز زمان گرفته شود. گرچه در پیاده‌سازی این روش‌ها بهتر می‌توان به نیازهای آن‌ها پی برد.

2-5 کارهای انجام‌شده در ریشه‌یابی فارسی

از جمله کارهای انجام شده در زمینه ریشه‌یابی کلمات فارسی می‌توان به پروژه بن [20]، ریشه‌یاب آماری [21] و [22] اشاره نمود.

در [20] یک ریشه‌یاب خاص زبان فارسی طراحی گشته است که به عنوان جزئی از یک موتور بازیابی مورد استفاده قرار می‌گیرد. الگوریتم این ریشه‌یاب شبیه ریشه‌یاب Porter است. اولین قدم الگوریتم پیدا کردن زیر رشته‌ای از لغت ورودی است که در لیست پس‌وندهای فارسی (که از روی گرامر فارسی تهیه شده است) وجود داشته باشد. اگر بیش‌تر از یک پس‌وند برای لغت پیدا شد، الگوریتم طولانی‌ترین پس‌وندی را انتخاب می‌کند که تعداد حروف ریشه (بخش اصلی لغت) را کم‌تر از حد مجاز نکند. (مثلاً در اینجا کم‌ترین تعداد حروف برای ریشه 3 کاراکتر است) مثلاً برای لغت «دستشان» می‌توان دو پس‌وند «ان» و «شان» را دید که «شان» طولانی‌تر است و چون حروف باقی مانده «دست»، 3 حرف یا بیش‌تر هستند،مشکلی برای انتخابش وجود ندارد. در این کار برای تعیین پس‌وند آخر لغت از یک DFA استفاده

شده است که ورودی آن وارون شده‌ی رشته‌ی (کلمه‌ی) ورودی است و همه‌ی حالت‌ها در آن حالت نهایی‌اند.

بن [20]، یک ریشه‌یاب «حذف وند» است. یعنی در هر قدم پس‌وندها یا پیش‌وندهایی را برمی دارد تا به لغت اصلی برسد. دیکشنری بن شامل مصدر و بن مضارع فعل‌هاست. الگوریتم بن به این صورت است که بیش‌ترین کاراکترهای ممکن را از لغت برمی دارد (برمبنای قواعدی) و این کار را آنقدر تکرار می‌کند تا دیگر امکان‌پذیر نباشد. ولی با این روش ریشه‌ی به دست آمده ممکن است صحیح نباشد. مثلاً با برداشتن پس‌وند «ی» از لغت «خانگی»، ریشه‌ی «خانگ» به دست می‌آید. برای حل این مشکل، بن از روش Recoding استفاده می‌کند که تبدیلی به شکل «$AXC \to AYC$» است و در آن A و C زمینه تبدیل را مشخص می‌کنند و X رشته‌ی ورودی و Y رشته تغییر یافته است.

ریشه‌یاب طراحی شده در نمایه‌ساز سینا مشابه ریشه‌یاب Porter برای زبان انگلیسی است [23]. هر دو ریشه یاب کلمه را با یک سری پیشوندها و پسوندها در چند مرحله تطابق می‌دهند تا پسوندها و پیشوندها حذف شوند و ریشه کلمه به دست آید. تفاوت این ریشه‌یابها به تفاوت زبان آنها برمی‌گردد. الگوریتم Porter الگوهایی از حروف صدادار و بی‌صدا برای تخمین محتوای اطلاعات مشخص می‌کند. در فارسی بسیاری از حروف صدادار نوشته نمی شوند. لذا ریشه یاب نمی تواند از آنها استفاده کند. در این ریشه‌یاب برای رفع این مشکل از روش تعریف حداقل طول ریشه استفاده کرده‌ایم. تفاوت دیگر این ریشه یاب با ریشه یاب Porter در تشخیص پیشوند است، ریشه‌یاب می‌تواند پیشوندها را مشخص کند در حالیکه ریشه‌یاب Porter الگوریتمی برای تشخیص پیشوند ارائه نداده است.

6. بازیابی تحمل‌پذیر

منظور از بازیابی تحمل‌پذیر (tolerant retrieval) اینست که موتور جستجو بتواند اشتباهات کاربر در ورود کلیدواژه یا عبارات را جبران نموده یا پیشنهاد اصلاح آنرا به کاربر ارایه نماید. از این روش می‌توان غلط‌های املایی کاربر را اصلاح نمود یا عبارات مشابه متداول را به کاربر ارایه کرد. چندین روش در بازیابی تحمل‌پذیر وجود دارد که در این بخش آنها را معرفی می‌کنیم.

6-1 غلطیابی املایی

دو روش عمده برای غلط‌یابی املایی استفاده می‌شود که در اینجا به معرفی آنها خواهیم پرداخت. این دو روش عبارتند از:

- فاصله ویرایشی (edit distance)
- همپوشانی k-gram (k-gram overlap)

البته الگوریتم‌های دیگری مانند الگوریتم‌های آوایی «phnetic» وجود دارند که به دلیل مشکل زبان فارسی در اعراب گذاری زیاد کارایی ندارند و به آنها نخواهیم پرداخت.

قبل از اینکه به این روش‌ها بپردازیم نگاهی می‌اندازیم به عملکرد موتور جستجو در قبال غلط‌های املایی از نگاه کاربر.

ما بر روی دو شیوه خاص غلط‌یابی متمرکز می‌شویم. این دو عبارتند از:

- کلمه مجزا (isolated word)
- حساس به متن (context-sensitive)

در غلط‌یابی «کلمه مجزا» ما هر بار بر روی یک کلمه متمرکز می‌شویم. یعنی اگر پرسش شامل چند کلمه باشد، عمل غلط‌یابی را هر بار بر روی کلمات آن به طور جدا گانه انجام می‌دهیم. در غلط‌یابی «حساس به متن»، مجاورت کلمات از نظر تشکیل عبارات متداوط بررسی می‌شود. به عنوان مثال اگر کاربر پرسش «فروشگاه مهرآباد تهران» را وارد کند، هر سه کلمه‌ی تشکیل دهنده آن، درست است؛ اما

احتمالا منظور کاربر «فرودگاه مهرآباد تهران» بوده است. این کار توسط الگوریتم‌های حساس به متن انجام می‌شود.

در ادامه ابتدا به الگوریتم‌های کلمه مجزا یعنی «فاصله ویرایشی» و «همپوشانی k-gram» خواهیم پرداخت و سپس الگوریتم‌های «حساس به متن» را توضیح می‌دهیم.

6-2 بکار بردن غلط‌یاب در موتور جستجو

موتور جستجو امکان غلط‌یابی املایی را در چند طریق مختلف بکار می‌برد:

1. زمانی که کلمه غلط در پرسش کاربر وارد شد، کلمات صحیح متناظر آنرا پیدا کن و به همراه کلمه غلط به مرحله‌ی بعدی بفرست. مثلا اگر کلمه «ارتبات» وارد شد، مستنداتی را بازیابی کن که در آنها کلمه «ارتبات» یا اصلاح‌شده آن مانند «ارتباط» و «ارتداد» موجود باشند.
2. مانند حالت اول عمل کن فقط به شرطی که کلمه وارد شده در لغت‌نامه نباشد. یعنی اگر کلمه «ارتبات» در واژه‌نامه نبود به مانند حالت 1 عمل کن.
3. مانند حالت اول عمل کن فقط به شرطی که تعداد مستندات یافته‌شده در اثر پرسش واردشده – مثلا «ارتبات» - کمتر از مقدار از پیش تعیین شده‌ای باشد.
4. وقتی که پرسش وارد شده تعداد مستنداتی کمتر از مقدار از پیش تعیین شده‌ای را بازگرداند، در اینصورت موتور جستجو پیشنهادی برای اصلاح کلمه واردشده به کاربر می‌دهد.

6-3 الگوریتم فاصله ویرایشی

فاصله ویرایشی بین دو رشته کاراکتر عبارت است از تعداد اعمالی که لازم است تا یکی را به دیگری تبدیل کند. این اعمال می‌توانند شامل حذف و درج و جابجایی باشند (بسته به الگوریتم این اعمال فرق می‌کنند). تعدادی الگوریتم برای تعریف یا محاسبه فاصله ویرایشی وجود دارند که به صورت زیر هستند:

- Hamming distance
- Levenshtein distance
- Damerau-Levenshtein distance
- Jaro-Winkler distance
- Wagner-Fischer edit distance
- Ukkonen
- Hirshberg

یکی از الگوریتم‌های مهم، الگوریتم Levenshtein است. که از روش برنامه سازی پویا برای محاسبه فاصله بین دو رشته استفاده می‌کند.

```
m[i,j] = d(s1[1..i], s2[1..j])

m[0,0] = 0
m[i,0] = i, i=1..|s1|
m[0,j] = j, j=1..|s2|

m[i,j] = min(m[i-1,j-1]
             + if s1[i]=s2[j] then 0 else 1 fi,
             m[i-1, j] + 1,
             m[i, j-1] + 1 ), i=1..|s1|, j=1..|s2|
```

مثال: فاصله بین دو کلمه «kitten» و «sitting» در الگوریتم Levenshtein برابر 3 است. مراحل تبدیل در زیر دیده می‌شود:

1. kitten → sitten (substitution of 's' for 'k')
2. sitten → sittin (substitution of 'i' for 'e')
3. sittin → sitting (insert 'g' at the end)

4-6 الگوریتم مجاورت کا-گرم

مدل N-gram علاوه بر کاربردهای دیگر، برای بررسی مجاورت دو رشته استفاده می‌شود. مجموعه N-gram شامل دنباله‌های N تایی یک رشته است.

مثال: رشته information را در نظر بگیرید. 4-gram های آن بصورت زیرند:

```
Info, nfor, form, orma, rmat, mati, atio, tion
```

روش کلی به این صورت است که ابتدا تمام N-gram های کلمات دیکشنری را تولید می‌کنیم و آنها را اندیس گذاری می‌کنیم. وقتی یک کلمه اشتباه را می‌خواهیم اصلاح کنیم همین کار را با آن کلمه می‌کنیم.

دو روش وجود دارد:

- ابتدا N-gram های کلمه پیدا می‌کنیم و آنها را با N-gram های دیکشنری مقایسه می‌کنیم. فرض بر اینست که کلمه اشتباه فقط دو یا سه کاراکتر اشتباه یا گم شده یا تغییر یافته دارد. با مقایسه N-gram ها می‌توان نزدیکترین کلمه درست را پیدا کرد.
- ابتدا کلمات مشابه کلمه‌ی اشتباه را با استفاده از الگوریتم Levenshtein برای یک فاصله ویرایشی معین، پیدا می‌کنیم سپس برای هر کدام از آنها، N-gram ها را تولید میکنیم. هر کدام از کلمات که تعداد بیشتری N-gram مشابه با کلمه غلط داشت را به عنوان پیشنهاد ارایه می‌کنیم.

مساله‌ای که در پیاده‌سازی وجود دارد اندیس گذاری N-gram ها است که نیازمند یک اندیس گذاری متن کامل است.

الگوریتم N-gram برای کشف غلط های ناشی از جای خالی (space) نیز کار می‌کند. برای اینکار می‌توانیم در تولید مشابه های نزدیک کلمه ، جای خالی را بین حروف قرار دهیم (علاوه بر افزودن و کاستن و جابجایی).

6-5 غلط‌یابی حساس به متن

در مواردی که عبارت وارد شده توسط کاربر شامل کلمات صحیح از نظر املا باشد، نیز ممکن است اشتباهی از سوی کاربر در وارد کردن عبارت صورت گرفته باشد. اشتباهی که منجر شود یک کلمه به کلمه درست دیگری تبدیل شود. مثلا فرض کنید کاربر عبارت «فروشگاه مهرآباد تهران» را وارد کرده

باشد؛ اگرچه این عبارت درست است اما ممکن است منظور کاربر «فرودگاه مهرآباد تهران» بوده باشد و ممکن هم هست که اینطور نباشد. برای ارایه چنین اصلاحات یا پیشنهاداتی نمی‌توان از الگوریتم‌های «کلمه مجزا» استفاده کرد. برای این منظور از الگوریتم‌های حساس به متن کمک می‌گیریم. در اینجا تکنیک‌های حساس به متن برای غلط‌یابی عبارت‌ها معرفی می‌شوند.

6-5-1 روش اول

ساده‌ترین روش این است که برای هرکدام از کلمات عبارت وارد شده را به طور جداگانه، کلمات مشابه را با استفاده از یکی از روش‌های «کلمه مجزا» مانند «فاصله ویرایشی» و «کا-گرم» پیدا کنیم و ترکیبات مختلف آنها را تشکیل دهیم. سپس هر کدام از عبارات تشکیل شده را بازیابی کرده و هرکدام که تعداد نتایج بیشتری را باز گرداند به عنوان پیشنهاد به کاربر ارایه دهیم. برای مثال در عبارت «فروشگاه مهرآباد تهران» می‌توانیم عبارت معادل هر کدام از کلمات مانند «فرودگاه» و «نهرآباد» و «مهران» را با ترکیبات مختلف بازیابی کنیم.

این روش می‌تواند سربار زیادی ایجاد کند مخصوصا وقتی تعداد کلمات مشابه برای هر کدام از کلمات زیاد باشد.

6-5-2 روش دوم

می‌توان از برخی از روش‌های تشخیصی برای بهبود نتایج جستجو استفاده کرد. به جای اینکه تمام ترکیبات ممکن با کلمات مشابه را تولید کنیم، متداول‌ترین ترکیبات را تولید می‌کنیم. ترکیبات متداول را از روی آمار همنشینی دوکلمه‌ای (biwords) بدست می‌آوریم و آنرا برای سه کلمه گسترش می‌دهیم. مثلا عبارت «فرودگاه مهرآباد» بسیاد متداول‌تر از «فروشگاه مهرآباد» است. همچنین عبارت «مهرآباد تهران» متداول‌تر از «مهرآباد مهران» است. لذا ترکیب «فرودگاه مهرآباد تهران» محتمل‌تر است.

دو منبع متداول برای بدست آوردن آمار همنشینی‌های دو کلمه‌ای وجود دارد:

- همنشینی کلمات در اسناد نمایه‌گذاری شده
- همنشینی کلمات در پرسش‌های وارد شده توسط کاربران

7. بررسی رفتار کاربر

وقتی کاربر به دنبال موضوعی خاص می‌گردد، در نهایت منظور خود را در قالب واژه‌ها یا عبارات به موتور جستجو وارد می‌کند و انتظار دارد تا مطالب مورد نظر خود را بیابد. طبیعی است که همه کاربران مثل هم فکر نمی‌کنند و در نتیجه وقتی دو کاربر مختلف به دنبال موضوعی یکسان می‌گردند ممکن است از کلمات کلیدی متفاوتی استفاده کنند. میزان موفقیت کاربر از نظر سرعت و دقت، بستگی به هوش و طرز فکر و دریافت ذهنی وی از عملکرد موتور جستجو دارد.

تجربه نشان داده است که کاربر پس از مدتی به رفتار موتور جستجو آشنا می‌شود و کلماتی را که انتخاب می‌کند، با مشاهده نتایج جستجو تغییر می‌کند. عملکرد موتور جستجو در رفتار کاربر در انتخاب کلمات کلیدی تاثیر می‌گذارد.

تحلیل رفتار کاربر می‌تواند الهام‌بخش طراحی الگوریتم‌های موفق برای موتورهای جستجو باشد، لذا در این بخش به این موضوع خواهیم پرداخت. در این بخش ابتدا مفهوم ربط از دیدگاه کاربر و سیستم‌های بازیابی را بررسی می‌کنیم. سپس روش نظرخواهی از کاربر را برای رتبه‌بندی معرفی می‌کنیم و سپس چالشی تحقیقی برای کاربران فارسی زبان مطرح می‌کنیم.

7-1 مفهوم ربط

بیان درخواست دقیق بر پارامترهای جستجوی جامعیت و مانعیت تأثیر می‌گذارد. کلیدواژه‌ها را بایستی با شکل صحیح و در قالبی مناسب وارد کرد و در انتظار پاسخ از سوی موتور جستجو بود. امّا آیا همیشه کاربر می‌تواند آنچه را در تفکر خود دارد در قالب کلیدواژه‌های مناسب به نظام عرضه کند؟ آنچه مسلم است این است که کاربران تجارب، دانش، و مهارت‌های متفاوتی با یکدیگر دارند. یک نظام بازیابی آرمانی باید قادر باشد کمال مطلوب کاربرانی با شرایط مختلف را مهیا کند.

برای اینکه کاربر بتواند نیاز خود را با زبانی قابل فهم برای نظام تبیین کند باید مهارت‌ها و دانش خاصی را نیز به کار بگیرد. سه دانش ذهنی و فنی و معنایی را برای رسیدن به مقصود برای کاربر ضروری به نظر می‌رسد [9]:

- دانش ذهنی: دانش مورد نیاز برای تبدیل یک نیاز اطلاعاتی به یک درخواست قابل جستجو است.
- دانش معنایی: چگونه و کی قابلیت‌های موجود در نظام را باید بکار برد؟
- دانش فنی: مهارتهای اساسی بکارگیری رایانه و ترکیب درخواست‌های وارد شده به‌عنوان عبارت-های جستجوی خاص.

هر یک از سه دانش فوق تأثیر شایانی بر میزان جامعیت نتایج بازیابی شده می‌گذارد چرا که بکارگیری این سه نوع دانش، افزایش میزان اسناد بازیابی شده را سبب می‌شود.

نکته قابل توجه اینکه نیاز کاربر همیشه همان چیزی نیست که در قالب سؤال آن را مطرح می‌کند. همه کاربران قادر نیستند تا فضاهای خالی ذهن خود را از یک مسأله به خوبی توصیف کنند. جهل کاربر نسبت به یک مسأله عمدتاً مرزی مشخص ندارد و به همین دلیل است که رفتار کاربران در حین جستجو تا حدی غیر قابل پیش‌بینی می‌شود و ما از برخی از ابزارها برای مطالعه رفتار آنها استفاده می‌کنیم.

ربط عامل حاکم بر تأثیر هر فرآیند ارتباطی است. از آنجا که هدف بازیابی اطلاعاتی برقراری ارتباط است، از این رو ربط هم کلید جدایی ناپذیر بازیابی مؤثر است. ربط را می‌توان ملاک توفیق بازیابی دانست. ربط مقیاس مؤثر بودن میان منبع اطلاعات و دریافت کننده است. ربط کیفیتی انتزاعی است، کیفیتی یگانه میان فرد و مدرکی معین که پشتیبان این پذیره است که آن را تنها کاربر اطلاعات می‌تواند داوری کند. ربط کیفیتی فردی دارد که به وضعیت شناختی کاربر، مشکلی باید گشوده شود، دانش قبلی از همان موضوع، فوریت کاربرد دانش جستجو شده و ارزشی که به اطلاعات نهاده میشود بستگی دارد [9].

در [9]، ملاک های ربط از نظر کاربر و نظام‌های بازیابی اطلاعات وب مقایسه شده است

ربط از نظر کاربر

- وضعیت شناختی کاربر
- ارزشی که به اطلاعات نهاده می‌شود
- فوریت کاربرد دانش جستجو شده
- دانش قبلی از همان موضوع
- مشکلی که باید گشوده شود

ربط از نظر سیستم‌های بازیابی

- محل کلیدواژه
- بسامد نسبی
- وجود کلیدواژه در متاتگ
- محبوبیت وب سایت

در این دو مدل دیده می‌شود که آنچه ملاک ربط برای کاربر است با آنچه ملاک ربط از نظر نظام است به طور آشکار متفاوت است.

بنابراین مشکل اصلی نظام‌های بازیابی اطلاعات، سنجش میزان ربط اطلاعات ذخیره شده یا ارتباط بین اطلاعات درخواست شده و اطلاعات بازیابی شده است. به عبارتی دیگر، با ارائه یک سؤال به نظام، نظام بازیابی باید بررسی کند که آیا اطلاعات ذخیره شده مربوط به پرسش است یا نه. اما ایهام و استعارات پشت واژگان و نقص بیان مفاهیم با برخی واژگان، این ارتباط را مشکل می‌سازد.

7-2 نظرخواهی از کاربر در رتبه بندی

برای برطرف کردن مشکل سوء تفاهم در مفهوم ربط بین ذهن کاربر و الگوریتم‌های موتور جستجو، اخیرا موتورهای جستجو از الگوریتم‌های پیشرفته‌ای برای رتبه‌بندی استفاده می‌کنند که در آنها نظر کاربر به عنوان یک پارامتر لحاظ می‌شود. پیشتاز این روش گوگل است که برای اولیل بار امکان نظرخواهی از کاربر را در نتایج جستجو برای کاربران فراهم کرد.

استفاده از نظر کاربران برای رتبه‌بندی نتایج جستجو موضوع تحقیقات گسترده‌ای است که مدتی است رواج پیدا کرده است. یکی از جدیدترین موارد که در آن مجموعه روش‌ها مورد مطالعه قرار گرفته‌اند در [24] آمده است.

3-7 کاربران فارسی زبان

تجربه نشان داده است که ضعف عملکرد موتورهای جستجو در زبان فارسی، کاربران را واداشته تا در انتخاب کلمات کلیدی از الگوهای خاصی استفاده کنند. این الگوها باید مورد مطالعه قرار گیرد. در این تحقیق مجال پرداختن به این موضوع به طور مستقل وجود ندارد و صرفا به عنوان یک چالش معرفی می‌شود.

در مورد رفتار کاربران فارسی زبان هیچ کاری تا کنون گزارش نشده است و به نظر می‌رسد این موضوع به عنوان یکی از زمینه‌های مطالعاتی جای زیادی برای کار دارد.

8. متا جستجوگر

یک متاجستجوگر یا «metasearch engine» سایتی است که بطور واسطه بین کاربر و موتورهای جستجو قرار می‌گیرد؛ پرسش کاربر را دریافت می‌کند و آنرا پالایش کرده و با استفاده از سرویس وب موتورهای جستجو، نتایج را از چندین موتور جستجو دریافت، و حاصل را ترکیب کرده و به کاربر ارایه می‌دهد. استفاده از این روش باعث می‌شود دامنه جستجو وسیع شود و نتایج بهتری حاصل شود. مجموعه نکات لازم برای ساخت و طراحی کارامد متاجستجوگر را در [25] می‌توان یافت.

استفاده از متا جستجوگرها در زبان فارسی در برخی کارها گزارش شده است مشهورترین آن سایت پارسیک [2] است. این امکان وجود دارد که با استفاده از متاجستجوگرها برخی از نقص‌های موتورهای جستجوگر در مورد زبان فارسی را پوشش داد و نتایج بهتری به دست آورد.

با آمدن گوگل فارسی [26] بسیاری از مشکلات مربوط به رسم‌الخط و فاصله‌گذاری در فارسی حل شد.

مشکلات مربوط به زبان فارسی در بازیابی اطلاعات در [1] برشمرده شده‌اند.

ایده استفاده از سایت واسطه برای بهبود نتایج جستجوی زبان فارسی، هنوز آنطور که باید، جا نیافتاده و بسیاری از مسایل مربوط به آن هنوز بررسی نشده‌اند. ما این ایده را راهگشای برخی از مسایل موجود در زبان فارسی می‌دانیم. اگرچه این ایده ابتدا در پارسیک [2] پیاده شد، اما این پیاده‌سازی بسیار محدود است و فقط شامل اصلاح کدینگ‌های کاراکترها می‌شود. بسیاری از کارهای دیگر را می‌توان در یک متاجستجوگر فارسی انجام داد.

مجموعه کارهایی که می‌توان در یک ابرجستجوگر فارسی انجام داد به صورت زیر هستند:

- اصلاح کدگذاری کاراکترها
- ایجاد ترکیبات منطقی با واژه‌های هم معنی
- ریشه‌یابی و فرستادن حالت‌های تصریفی مختلف یک کلمه
- اصلاح فاصله‌گذاری

در ادامه به این موارد خواهیم پرداخت.

8-1 اصلاح کدگذاری

قبل از آمدن گوگل فارسی، کاربران فارسی زبان مشکلات عمده‌ای در یافتن مستندات فارسی در اینترنت داشتند. مستندات فارسی اینترنت با کدینگ‌های مختلفی نوشته شده بودند که نه کاربران و نه موتورهای جستجو بر روی آن توافق داشتند. به عنوان مثال تفاوت در کد کاراکتر «ک» و «ی» که باعث می‌شد یک کلمه، به کلمه‌ای دیگر تبدیل شود. برای این منظور برخی از سایت‌ها مانند پارسیک [2] برای رفع برخی از مشکلات ابداع شدند. در واقع در سایت پارسیک برای رفع مشکلات کدینگ فارسی، کلمات وارد شده توسط کاربر با چندین کدگذاری به چندین موتور جستجو ارسال می‌شدند و نتایج با هم ترکیب می‌شدند. در واقع این سایت به صورت یک سایت «metasearch engine» عمل می‌کند.

8-2 ترکیب با واژه‌های هم‌معنی

در صورتی که موتور جستجو در الگوریتم رتبه‌بندی خود از کلمات هم‌معنی زبان آگاه نباشد، می‌توان با استفاده از این روش یعنی ترکیب با واژه‌های هم‌معنی، نتایج بهتری را توسط متاجستجوگر فراهم کرد.

در موتور جستجو امکان تولید عبارت‌های منطقی با AND و OR فراهم است. می‌توان برای کلماتی که توسط کاربر وارد می‌شود، کلمات هم‌معنی آنرا یافت و با ترکیبات منطقی مقتضی برای موتور جستجو فرستاد. به عنوان مثال وقتی کاربر به دنبال کلمه «سپاهان» می‌گردد، معادل آن، «اصفهان» را با آن ترکیب منطقی کرد.

در توسعه‌ی این روش می‌توان از جایگزینی عبارات معادل نیز استفاده نمود؛ مثلا وقتی کاربر عبارت «رسول اکرم» را وارد می‌کند می‌توان عبارت معادل آنرا، مثلا «حضرت محمد»، تولید کرد و برای موتور جستجو فرستاد.

البته در روش‌های فوق می‌توان یا از ترکیب منطقی استفاده کرد یا اینکه پرسش را در دو وهله مختلف از موتور جستجو انجام داد و نتایج را در متاجستجوگر ترکیب کرد و برای کاربر فرستاد.

هنوز تحقیق مستقلی در مورد کارایی این روش در زبان فارسی انجام نشده است و در این گزارش صرفا به بیان این ایده پرداخته شده است. البته طبیعی است که بهبودی در نتایج حاصل خواهد شد اما میزان آن باید اندازه‌گیری شود.

3-8 ریشه‌یابی و تصریف

زمانی که یک موتور جستجوگر از نحو و نگارش و دستور خط زبان فارسی آگاه نیست، وقتی یک سایت فارسی را نمایه‌گذاری می‌کند، به کلمات موجود در داخل آن به صورت توده‌ای از رشته‌ها نگاه می‌کند. و این باعث می‌شود که به پارامترهای رتبه‌بندی خدشه وارد شود. مثلا پارامتر تعداد تکرار کلمات به درستی محاسبه نمی‌شود. اگر نمایه‌گذار نتواند تشخیص دهد که کلمه «لباس» با «لباسها» یکی است، تعداد تکرار کلمه به درستی محاسبه نمی‌شود.

حال فرض کنید ما در شرایطی هستیم که در طراحی و ساخت یک موتور جستجو هیچ دخل و تصرفی نمی‌توانیم داشته باشیم. در این حالت تنها را چاره استفاده از یک واسطه برای بهبود نتایج فوق است. در مورد مثال فوق فرض کنید یک سایت واسط پرسش کاربر را «لباس» دریافت کند، و آنرا به صورت «لباس OR لباس‌ها»، یا با ترکیبات مختلف منطقی به موتور جستجو بفرستد، می‌تواند نتایج بهتری دریافت کند.

البته گوگل فارسی [26] اغلب مشکلات مربوط به ریشه‌یابی کلمات را مرتفع کرده است و به نظر می‌رسد که دیگر ریشه‌یابی و تصریف در واسط کمکی به بهبود نتایج نکند. اغلب مشکلاتی که اکنون باقی مانده‌اند مربوط می‌شود به نحوه فاصله یا نیم‌فاصله گذاری در بین بخش‌های یک کلمه.

4-8 اصلاح فاصله‌گذاری

این مشکل تشخیص مرز کلمه را برای نمایه‌گذار دشوار می‌کند. در این گونه موارد می‌توان شکلهای مختلف اتصال بخش‌های کلمه را تولید کرد و به موتور جستجو فرستاد.

یکی از راه حل‌های مشکل فاصله گذاری در فارسی، استفاده از روش‌های انـدیس‌گـذاری n-gram اسـت. این روش که در زبان انگلیسی معمولا برای غلبه بر مشکل حذف «کلمات عمومی» و همچنین تشخیص «عبارات متداول» استفاده می‌شود، می‌تواند در زبان فارسی برای غلبه بر مشکل فاصله گـذاری اسـتفاده شود. یک مثال کلاسیک در مورد کلمات عمومی در زبان انگلیسی وجود دارد که وقتی کـاربر بـه دنبـال عبارت «to be or not to be» می‌گردد چون تمام آنها کلمات عمومی هستند باید حذف شـوند؛ امـا ایـن اتفاق نمی‌افتد و با استفاده از روش‌های اندیس‌گذاری n-gram می‌توان اسناد مرتبط را یافت. کاربرد دیگر در مورد عبارات متداول و اصلاح املایی است. در فارسی نیز می‌تـوان بـه کـارایی ایـن روش بـرای حـل مشکل فاصله‌گذاری اتکا کرد. با این وجود کاربرد این روش برای این منظور باید بومی‌سازی شود.

با توجه به مشکلات ساختاری در رسم‌الخط زبان فارسی و نبود یـک قـانون فراگیـر بـرای آن، کلمـات و عبارات به صورت‌های گوناگونی نوشته می‌شوند و تعیین مرز کلمات کار دشواری است. در نتیجـه وقتـی کاربر در موتور جستجو به دنبال موضوعی خاص می‌گردد، نمی‌تواند به درستی کلمه کلیدی را انتخاب و وارد کند. به عنوان مثال کاربر به دنبال کلمه «آبسردکن» می‌گردد و اگر در بسیاری از سایت‌ها این کلمه به صورت‌های فاصله‌دار مثلا «آب‌سردکن» نوشته شده باشد، کاربر نمی‌تواند به نتیجه دلخواه خود دست یابد. همچنین نحوه اتصال پیشوندها و پسوندها در زبان فارسی چندان قانونمند نیست و این مساله باعث می‌شود که دقت موتورهای جستجو در ریشه‌یابی و اندیس‌گذاری کاهش یابد.

در استفاده از روش n-gram برای زبان فارسی باید حالتهای مختلف تصـریف فارسـی را در نظـر گرفـت. اگرچه گوگل فارسی اخیرا نتایج قابل قبولی برای کلمات فاصله‌دار فارسی ماننـد «آب سـرد کـن» ارایـه می‌دهد، اما هنوز جای بهبود برای آن وجود دارد.

5-8 جمع‌بندی

در جمع‌بندی باید گفت که اگرچه استفاده از سایت‌های واسط یا ابرجستجوگرها می‌توانند به بهبود نتایج جستجوی فارسی کمک کنند اما بهبود الگوریتم‌های نمایه‌گذاری در داخل موتور جستجو بسیار کارامدتر از آن است.

9. بهینه‌سازی برای موتور جستجو

منظور از «بهینه‌سازی برای موتور جستجو» یا «Search Engine Optimization» که به اختصار «SEO» نامیده می‌شود، این است که سایت خود را طوری طراحی کنیم که رتبه آن در موتورهای جستجو افزایش یافته و در صفحات اول نمایش داده شود و از این طریق ترافیک سایت یا تعداد بازدید کنندگان آنرا افزایش یابد.

به عبارت دیگر بهینه سازی سایت وب این است که در نتایج یک موتور جستجوی مشهور، بیشترین امتیاز را داشته باشد. اهمیت این موضوع از آنجا ناشی می شود که اکثر مردم از موتورهای جستجو برای رسیدن به مطلب یا محصول مورد نظر خود استفاده میکنند و اکثر مردم فقط به صفحه ی اول نتایج جستجو نگاه میکنند. بنابراین برای داشتن ترافیک بالا از طرف موتورهای جستجو، این مسئله الزامی است که سایت مزبور در صفحه‌ی اول نتایج جستجو قرار گیرد.

علم بهینه‌سازی برای موتور جستجو، شاخه‌ای است از بازاریابی بر پایه موتور جستجو برای بالا بردن تعداد و کیفیت بازدیدکنندگان یک سایت به کمک موتور جستجو. در بهینه‌سازی موتور جستجو از تکنیک‌هایی استفاده می‌شود که الگوریتم‌های موتور جستجو را اصطلاحا فریب داده و باعث می‌شود که سایت مورد نظر در نتایج در صفحه اول قرار گیرد. علم بهینه‌سازی موتور جستجو، در مورد روشهای فنی مانند عنوان صفحه‌ی مناسب، تگ‌ها و متاتگ‌ها، کلیدواژه‌ها و عبارات کلیدی و توضیحات مناسب سایت و کلا محتوایی که موتورهای جستجو آنها را دوست دارند، مطالعه می‌کند.

امروزه مبحث بهینه‌سازی برای موتور جستجو چنان گسترده شده که تحقیقات مبسوطی بر روی آن صورت می‌گیرد و کتاب‌ها و مقالات زیادی در این زمینه منتشر شده‌اند. در [27] می‌توانیم نمونه‌ی جامعی از اصطلاحات و روش‌ها و تکنیک‌های بهنه‌سازی را بیابیم.

موتورهای جستجو صفحات وب را به وسیله نرم‌افزار خزنده پیدا و فهرست‌بندی و توسط نمایه‌گذار آنرا امتیازدهی می‌کنند. طبیعی است که تمام موتور جستجو از یک الگوریتم استفاده نمی‌کنند. برای مثال اگر صفحه وبی در یکی از موتورهای جستجو امتیاز بالایی داشته باشد، ممکن است در دیگر موتورها این چنین نباشد. الگوریتم‌های رتبه‌بندی از اسناد محرمانه موتورهای جستجو به حساب می‌آیند اما یکی از کارهایی که متخصصان بیهنه‌سازی انجام میدهند، پیگیری و حدس تمام تغییرات عملکرد داخل

موتورهای جستجو است تا بتوانند صفحات وب را بر طبق این تغییرات بهینه‌سازی کنند. بـه عـلاوه آنهـا همراه با تغییرات موتورهای جستجوی مختلف خود را تابع این موتورها قرار می‌دهند.

تکنیک‌هایی که در این حوزه مورد استفاده قرار می‌گیرند بی‌ارتباط با الگوریتم‌های رتبه‌بندی موتورهـای جستجو و همچنین رفتار کاربران در استفاده از کلمات کلیدی نیستند. همچنین طراحان موتور جسـتجو برای کیفیت بخشیدن به نتایج جستجو باید با این تکنیک‌ها مقابله کنند تا سایت‌های بی‌اهمیت خـود را در صدر نتایج جا نزنند. لذا در اینجا پرداختن به مبحث «بهینه‌سازی برای موتور جسـتجو» را بـی‌فایـده نمی‌دانیم.

در این بخش تکنیک‌های مورد استفاده در حوزه SEO را معرفی می‌کنیم. بیشتر منابع (کتابها و مقالات) موجود در این زمینه بیشتر برای زبان انگلیسی هستند و به طور خاص برای زبان فارسـی منبعـی یافـت نشد. البته بیشتر روش‌های بکار رفته برای تمام زبان‌ها مشترک هستند اما برخی از تکنیک‌ها که مـرتبط با ساختار زبان هستند - و عمدتا مربوط می‌شوند به انتخاب کلمات کلیدی - خاص زبان هستند. در این بخش سعی می‌شود تکنیک‌های مزبور با گرایش زبان فارسی توضیح داده شوند.

در ادامه این بخـش، ابتـدا بـه اهمیـت تبلیغـات و فعالیـت در اینترنـت مـی‌پـردازیم و دلیـل محبوبیـت بهینه‌سازی را بیان می‌کنیم. در بخش‌های بعدی به بیان تکنیک‌های آن می‌پردازیم.

9-1 اهمیت بهینه‌سازی سایت

روزانه ملیونها کاربر اینترنتی با میلیون‌ها صفحات اینترنتی با حجم انبوهی از اطلاعات مواجه می‌شوند که دستیابی به این اطلاعات بدون موتورهای جستجو تقریبا امکان ناپذیر است، به طبع گردانندگان سایت با دانستن الگوریتم‌های رتبه‌بندی موتورهای جستجو می‌توانیم به هدف خود که افزایش رتبه سـایت‌هـا در موتورهای جستجو و در نتیجه آن افزایش بازدید کنندگان است، دسـت یابنـد. از ایـن رو تـلاش تمـامی وبسایت نویسان، قرار گرفتن در بالاترین رتبه هنگام جستجو می‌باشد. در نهایت یک سایت بهینـه شـده، برابر با بینندگان و بازدید کنندگان بیشتر و در نهایت برابر با هزاران دلار تبلیغات مجانی اسـت. از سـوی دیگر موتورهای جستجو با ارائه الگوریتم‌های متفاوت، در صدد نمایش نتـایج مطلـوب و مـرتبط‌تـری بـه کاربران هستند و این امر باعث رضایت کاربران از موتورهای جستجو خواهد گردید.

سایتهایی که در موتورهای جستجو حضور ندارند یا به نوعی حضورشان ضعیف می باشد، اشباع نشده نامیده می شوند. استفاده از لغت اشباع به این معنا که نتوانسته اند به گونه ای در اینترنت اشباع و حل شده و در جاهای مختلف حضور داشته باشند که در صفحات ابتدائی آمده یا در معرض دید عموم قرارگیرند؛ در حقیقت به علت بی توجهی به فعالیتی که کاربران اینترنتی که بر روی اینترنت انجام می‌دهند، این سایتها در حال از دست دادن فعالیت تجاری خود می باشند. امروزه موتورهای جستجو جز اولین و قویترین ابزار اطلاع رسانی در اینترنت مطرح هستند.

اساس کار به این ترتیب است که اکثر بازاریابهایی که در زمینه اینترنتی فعالیت می کنند البته بازاریابهایی که به صورت بین المللی کار می کنند نه صرفا ایران، چون در ایران این گونه مسائل به صورت سنتی انجام می شود. متد مورد استفاده آنها، تحقیقات بروی کلمات کلیدی است که می تواند آن تجارت را تکان دهد. ممکن است پیدا کردن کلمات کلیدی خیلی سخت باشد و در عین حال کار طاقت فرسایی می باشد و نیاز به دانش خاص خود دارد. کسی که بتواند در انتخاب کلمات کلیدی دقت به خرج دهد و توانایی پیدا کردن قسمتهای نهفته و کلیدواژه‌های موفقی که به آنها خیلی توجه نمی شود را داشته باشد، موفق‌تر است.

2-9 تکنیک‌های بهینه‌سازی برای موتور جستجو

ما تکنیک‌های بهینه‌سازی را در دو دسته کلی طبقه‌بندی می‌کنیم:

- کلمات و جایگاه آن
- لینک‌های بین صفحات

البته برخی تکنیک‌ها در هیچ کدام از دسته‌های فوق قرار نمی‌گیرند؛ مثلا:

- نحوه طراحی و چینش سایت
- فرمت فایل‌ها
- استفاده از تکنیک‌های پیشرفته طراحی وب
- تعداد صفحات سایت

- امکان RSS در سایت
- ...

موارد فوق ممکن است به بهینه‌سازی سایت کمک کند؛ اما این‌ها حدس‌هایی است که در کتاب‌های بهینه‌سازی زده شده‌اند [27, 28] و تا حدی خارج از موضوع این تحقیق هستند و ما به آنها نمی‌پردازیم.

در ادامه این بخش با تکنیک‌های بهینه‌سازی برای موتور جستجو می‌پردازیم [27, 28]. این تکنیک‌ها را در دو بخش «کلمات و جایگاه آن» و «لینک‌های بین صفحات» بررسی می‌کنیم. سپس به معرفی ابزارهای بهینه‌سازی می‌پردازیم.

3-9 کلمات و جایگاه کلمات

مهمترین کار در بهینه‌سازی سایت، انتخاب کلمات کلیدی مناسب و مربوط است. انتخاب کلمات کلیدی به قدری اهمیت دارد که اشتباه کردن در آن باعث از دست دادن برتری در رتبه‌بندی می‌شود. طبیعی است که کلمه کلیدی باید با توجه به موضوع و فعالیت سایت انتخاب شود.

9-3-1 انتخاب کلمه

در انتخاب کلمه کلیدی باید توجه داشت که کاربر وقتی به دنبال موضوع یا محصولی خاص می‌گردد، از کدام کلمات کلیدی استفاده می‌کند. باید کلماتی را انتخاب کرد که به طرز فکر کاربران نزدیکتر باشد. ممکن است کاربران از میان قشر عمومی اجتماعی باشند، در اینصورت به کاربردن کلمات رسمی و علمی خوب جواب نمی‌دهد. برعکس ممکن است مخاطبان از میان طبقه تحصیل‌کرده باشند که در این صورت باید از اصطلاحات فنی استفاده کرد. برای انتخاب کلمات درست تعدادی از روش‌های تشخیصی در [27] معرفی شده‌اند. به طور کلی انتخاب کلمات کلیدی قانون خاصی ندارد و نیاز به زیرکی و مهارت دارنده سایت است.

9-3-2 چگالی کلمات کلیدی

نکته‌ای که قابل بحث است، اینست که چگالی کلمات کلیدی چقدر باید باشد. اگر صفحه‌ای 1000 کلمه داشته باشد و 10 بار کلمه‌ی کلیدی به کار رفته باشد، می‌گوییم که چگالی کلمه 1 درصد است. اینکه چگالی کلمه چقدر باشد تا سایت بهینه شود، دقیقا معلوم نیست. برخی رقم 5% و 7% را گفته‌اند [27].

چیزی که باید از آن اجتناب کرد، تکرار بیش از حد کلمه است. اگر کلمه‌ی کلیدی بیش از حد مشخصی تکرار شود، از طرف موتور جستجو به عنوان سایت فریبکار شناسایی شده و به لیست سیاه می‌رود. در این صورت مدت‌ها طول می‌کشد تا بتوان آنرا از لیست سیاه بیرون آورد.

9-3-3 جایگاه کلمات کلیدی در سند

محل قرار گرفتن کلمه در سند مهم است. مثلا اگر کلمه کلیدی اگر در بین تگ‌هایی مانند <title> و <h1> باشد ارزش‌گذاری بیشتری می‌شود. در مورد ارزش و اهمیت تگ‌ها گمانه‌زنی‌های زیادی می‌شود. نمونه‌ای از ارزش و اهمیت تگ‌ها را می‌توان در [29] یافت. نکته‌ی قابل توجهی در مورد تگ‌های متا وجود دارد. در ابتدا موتورهای جستجو اهمیت زیادی به این تگ‌ها میدادند و باعث شد تا بسیاری از تکنیک‌های بهینه‌سازی برای سوء استفاده از این روش ایجاد شود. الگوریتم‌های امروزی توجه بسیار کمی به تگ‌های متا دارند و امتیاز بسیار کمی به این کلمات می‌دهند.

امروزه الگوریتم‌های رتبه‌بندی بسیار پیشرفته شده‌اند بطوریکه طراح سایت نباید در نمایش کلمات کلیدی اغراق کند و باید سایت طوری طراحی شود که همه‌چیز طبیعی به نظر برسد.

9-3-3-1 اندازه صفحه

اندازه صفحه ممکن است در امتیازهی موثر باشد. صفحاتی که طول کمتر از 100 داشته باشند کم ارزش تلقی می‌شوند [28] و صفحاتی که بیش از 250 داشته باشند اگر به چند صفحه تقسیم شوند می‌توانند امتیاز بیشتری برای سایت به همراه داشته باشند [28].

9-4 لینک‌های بین صفحات

لینک‌ها تاثیر به‌سزایی در تعیین رتبه یک سایت دارند. هرچه تعداد لینک‌ها به یک سایت در اینترنت بیشتر باشد، رتبه‌ی آن بالاتر خواهد بود.

9-4-1 لینک به سایت

به عنوان یک اصل، هرچه تعداد لینک‌های وارد شونده به یک سایت هرچه بیشتر باشند، اهمیت سایت بیشتر می‌شود. اما شرط آن این است که لینک‌ها از سایت‌های درست و معتبر وارد شوند. در اینترنت سایت‌هایی وجود دارند که از این امر سوء استفاده کرده و به صورت مدور به هم لینک می‌دهند تا امتیاز همدیگر را زیاد کنند. همچنین سایت‌های بدرفتاری وجود دارند که سرویس لینک را ارایه می‌دهند. امروزه این سایت‌ها توسط الگوریتم‌های پیشرفته موتور جستجو، به عنوان سایت‌های فریبکار شناسایی می‌شوند و در لیست سیاه قرار می‌گیرند. لذا استفاده از این روش لینک دهی درست نیست و لینک‌ها باید از سایت‌های معتبر باشند. هرچه رتبه‌ی سایت لینک‌دهنده بیشتر باشد، امتیاز سایت لینک گرفته بیشتر است. البته در الگوریتم‌های پیشرفته موضوع سایت‌ها نیز اهمیت دارد. اگر کلمات کلیدی سایت‌های لینک‌دهنده بیشتر باشد، امتیاز داده شده به آن بیشتر است.

این مسایل باعث شده که الگوریتم‌های امتیازدهی بر اساس لینک‌ها بسیار پیچیده شوند. در واقع تحلیل پیچیده‌ای بر روی گراف اینترنت انجام شود.

9-4-2 لینک به بیرون

لینک‌های بیرون رونده از سایت نیز در امتیازدهی به سایت تاثیر دارند. وجود تعدادی لینک بیرون‌رونده می‌تواند مثبت تلقی شود. البته اگر لینک‌ها به سایت‌های مرتبط (یا مشابه) داده شوند نشان‌دهنده حساب شده بودن لینک‌ها است. در مورد تعداد لینک‌های بیرون‌رونده، نظرات مختلفی وجود دارد. اگر تعداد لینک‌ها از تعداد معینی بیشتر باشد. اثر منفی روی امتیاز صفحه دارد. احتمال اینکه کاربرانی که صفحه مزبور را تماشا می‌کنند از صفحه خارج شوند زیاد است. در واقع صفحاتی که فقط شامل لینک

هستند هیچ امتیازی نمی‌گیرند. در [28] تعداد حداقل 2 و حداکثر 15 لینک برای هر صفحه پیشنهاد شده است.

9-4-3 لینک‌های داخلی

لینک‌های درون صفحات یک سایت به همدیگد، در امتیازدهی نقش دارند. هرچه تعداد این لینک‌ها بیشتر باشد، احتمال اینکه بازدیدکننده سایت، درون سایت باقی بماند بیشتر است. همچنین وجود لینک‌های داخلی نشان‌دهنده گستردگی اطلاعات ارایه شده است. در مورد امتیاز منفی لینک‌ها داخلی گمانه‌زنی خاصی وجود ندارد اما بهتر است تعداد این لینک‌ها در حد معقول باشد.

9-4-4 لینک‌های نامعتبر

لینک‌های نامعتبر لینک‌هایی هستند که کار نمی‌کنند؛ یعنی یا اشتباه تایپ شده‌اند یا اینکه سایت مقصد منقضی شده است.

9-5 ابزارهای بهینه‌سازی

ابزارهای بسیار برای کمک به امر بهنه‌سازی وجود دارند که در این بخش برخی از نرم‌افزارهای منبع‌باز را معرفی می‌کنیم.

9-5-1 نرم‌افزارهای منبع‌باز

نرم‌افزار NickeBot [30]: ابزار یافتن کلمات کلیدی مناسب

نرم‌افزار SERPS [31]: سایتی است که به یافتن رتبه در گوگل و سایر موتورهای جستجوی مهم کمک می‌کند.

نرم‌افزار Meta Tag Analyzer [32]: اطلاعات متا را از نظر میزان ارتباط به متن سایت ارزیابی می‌کند.

نرم‌افزارهای تجاری بسیار زیادی نیز وجود دارند که به آنها نمی‌پردازیم.

10. خلاصه

در این مقاله به مساله بهینه‌سازی رتبه‌بندی در موتور جستجو پرداختیم و از جنبه‌های مختلفی آنرا بررسی کردیم. در مورد جنبه‌های مربوط به زبان، «ریشه‌یابی» و «کلمات عمومی» را توضیح دادیم. برای ایجاد ارتباط بهتر بین مفهوم ذهنی کاربر و موتور جستجو، عملکرد کاربر را بررسی کردیم. در واقع هدف بهینه‌سازی ایجاد یک ارتباط بهتر بین کاربر و موتور جستجو است.

به مساله بازیابی تحمل‌پذیر پرداختیم و مکانیزم اصلاح اشتباهات کاربر و ارایه اصلاحات در موتور جستجو را توضیح دادیم.

متاجستجوگر روش دیگری برای برقرای ارتباط بهتر بین کاربر و موتور جستجو می‌باشند که در مورد جنبه‌های قابل انجام برای زبان فارسی راه‌هایی پیشنهاد کردیم.

مساله بهینه‌سازی برای موتور جستجو، باعث شده تا موتورهای جستجو الگوریتم‌های خود را بر اساس آن تعدیل کنند لذا پرداختن به تکنیک‌های آن باعث درک بهتر از کار موتور جستجو می‌گردد که در بخش 9 به آن پرداختیم.

مراجع

[1] ل. مرتضایی, "مسایل خط و زبان فارسی در ذخیره‌سازی و بازیابی اطلاعات," فصلنامه اطلاع رسانی, دوره 17, 1380.

[2] پارسیک. "جستجوگر وب و اخبار فارسی پارسیک," http://www.parseek.com.

[3] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition*: Morgan Kaufmann, 1999.

[4] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*: Cambridge University Press, 2008.

[5] R. Baeza-Yates, and B. Ribeiro-Neto, *Modern Information Retrieval*: Addison Wesley, 1999.

[6] م. تشکری, "ساخت یک نمایه‌ساز خودکار برای متون فارسی," دانشکده مهندسی کامپیوتر, دانشگاه امیرکبیر, تهران, 1380.

[7] ح. بشیری, ف. کربلایی, و ش. موسوی, "طراحی و ارزیابی نمایه‌ساز خودکار متون فارسی," in یازدهمین کنفرانس بین‌المللی کامپیوتر انجمن کامپیوتر ایران, تهران, 1384.

[8] G. Salton, and M. J. McGill, *Introduction to Modern Information Retrieval*, New York: Mc Graw Hill, 1983.

[9] م. ش. اژه‌ای, و س. امیدی‌فر, "بررسی مولفه های موثر بر میزان بازیابی اطلاعات و دقت بازیابی اطلاعات در نظام های بازیابی اطلاعات وب مدار," در همایش چالش‌های علم اطلاعات, اصفهان, 1386.

[10] S. Brin, and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, no. 1-7, pp. 107-117, 1998.

[11] J. E. Hopcroft, R. Motwani, and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, 2nd ed.: Addison Wesley, 2000.

[12] T. Winograd, *Language As a Cognitive Process: Syntax*: Addison-Wesley, 1982.

[13] R. Hessami-Fard, and G. Ghasem-Sani, "Stemmer Algorithm Design for Persian Language," in 11th International CSI Computer Conference (CSICC'2006), Tehran, Iran, 2006.

[14] M. I. Mobarakeh, and B. Minaei-Bidgoli, "Verb Detection in Persian Corpus," *International Journal of Digital Content Technology and its Applications* vol. 3, pp. 58-65, 2009.

[15] A. Mokhtaripour, and S. Jahanpour, "Introduction to a new Farsi stemmer," in Proceedings of the 15th ACM international conference on Information and knowledge management, Arlington, Virginia, USA, 2006.

[16] Megerdoomian, and Karine, "Developing a Persian Part-of-Speech Tagger," in First Workshop on Persian Language and Computers, Tehran University, Iran, 2004.

[17] ا. دفتری‌نژاد, "ساختواژه حالت-متناهی: روشی مناسب برای طراحی پردازشگر ساختواژی," در هفتمین همایش زبانشناسی ایران, 1386.

[18] K. Megerdoomian, "Finite-State Morphological Analysis of Persian," in Workshop on Computational Approach to Arabic Script-Based Languages, 2004.

[19] K. Beesley, and L. Karttunen, *Finite State Morphology*: Stanford, CSLI Publications, 2003.

[20] M. Tashakori, M. Meybodi, and F. Oroumchian, "Bon: The Persian stemmer," *Lecture Notes in Computer Science (LNCS), Springer Verlag*, vol. 2510, pp. 487-494, 2002.

[21] م. نصیری, م. ش. اسماعیلی, و ک. ابولحسنی, "یک ریشه‌یاب آماری برای زبان فارسی," در مجموعه مقالات یازدهمین کنفرانس بین‌المللی کامپیوتر, 1384.

[22] K. Taghva, R. Beckley, and M. Sadeh, "A Stemming Algorithm for the Farsi Language," in Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume I - Volume 01, 2005.

[23] M. F. Porter, "An algorithm for suffix stripping," *Program 14(3)*, pp. 130-167, 1980.

[24] W. Ng, L. Deng, and D. L. Lee, "Mining User preference using Spy voting for search engine personalization," *ACM Trans. Internet Technol.*, vol. 7, no. 4, pp. 19, 2007.

[25] W. Meng, C. Yu, and K.-L. Liu, "Building efficient and effective metasearch engines," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 48-89, 2002.

[26] گوگل. "موتور جستجوی فارسی گوگل," 2009; http://www.google.com/intl/fa.

[27] J. Ledford, *SEO: Search Engine Optimization Bible*: Wiley 2007.

[28] H. Davis, *Search Engine Optimization*: O'Reilly, 2006.

[29] B. S. Konia, *Search Engine Optimization with WebPosition Gold 2*, p.^pp. 338 : Wordware Publishing, 2002.

[30] "NickeBot," http://www.nichebot.com.

[31] "SERPS," http://www.seo-guy.com/seo-tools/se-pos.php.

[32] "Meta Tag Analyzer," http://www.seo-guy.com/seo-tools/se-pos.php.