به نام خدا

پژوهشکده ارتباطات و فناوری اطلاعات

مطالعه و بررسی جویشگرهای متنی

پروژه: مدیریت طرح

کد پروژه: ۹۳۳۲۰۱۲

مجری:	علیرضا یاری
تهیه‌کننده:	تیم مدیریت طرح
کد گزارش:	P-PD-VAS-SBM-S-013-1.02
تاریخ ارائه:	۹۴/۰۴/۲۰
نسخه/وضعیت:	نهایی

شناسنامه گزارش

عنوان: : بررسی جویشگرهای متنی		**شماره نسخه:** ۱,۰۲
کد:P-PD-VAS-SBM-S-013-1.02	**نوع گزارش:** فنی	**تاریخ ارائه گزارش:** ۹۴/۰۴/۲۰
نام پروژه: مدیریت طرح		**نوع پروژه:** پژوهشی - کاربردی
تاریخ شروع: ۹۳/۱۱/۲۰		**تاریخ پایان:** ۹۶/۱۱/۲۰ (۳۶ ماه)
نام گروه: طرح جویشگر		
کد پروژه: ۹۳۳۲۰۱۲		**شماره و تاریخ قرارداد:** ۹۳/۱۱/۲۰
مجری: علیرضا یاری		**ناظر / ناظرین:** کامبیز بدیع، رامین شکری‌پور و روح‌اله رحمانی
تهیه کننده/ تهیه کنندگان: تیم مدیریت طرح		
نام و نشانی مجری: تهران، انتهای خیابان کارگر شمالی، مرکز تحقیقات مخابرات ایران، کد پستی ۱۴۳۹۹۵۵۴۷۱، تلفن: ۸۴۹۷۴۲۴		
نام و نشانی حمایت کننده: تهران، خیابان شریعتی، وزارت ارتباطات و فناوری اطلاعات		
ملاحظات:		

چکیده:

تحقیقات نشان می‌دهد بیش از ۸۰٪ کاربران اینترنتی وب سایت‌های جدید را از طریق موتور جستجو پیدا می‌کنند. جویشگرها سرویس جستجو را بر روی اقلام مختلف موجود در وب در قالب خدمات مجزائی نظیر جویشگر محتوای متنی، جویشگر تصاویر، جویشگر اصوات، جویشگر ویدئوها و غیره، ارائه می‌دهند. در سرویس جستجوی متنی وب که در آن عمدتا جستجو بر روی محتوای متون موجود در در صفحات وب انجام می‌شود، معمولا انتظار بیشتری از جویشگر وجود دارد و آن اینست که با شناسایی هدف کاربر در صورت نیاز از سایر سرویسهای موجود خود نیز برای پاسخگویی به نیاز کاربر استفاده کند.

هدف از این پژوهش بررسی جویشگرهای متنی موجود و تجارب موفق سایر کشورها در زمینه بومی‌سازی آن می‌باشد. تا در نهایت به یک جمع‌بندی از عوامل موفقیت این جویشگرها برسیم.

کلمات کلیدی: جویشگر متنی، خزش، نمایه‌سازی، رتبه‌بندی

وضعیت گزارش: نهایی	**زبان گزارش: فارسی**
وضعیت دسترسی: عادی	**تعداد صفحات:** ۳۳

چکیده

تحقیقات نشان می‌دهد بیش از ۸۰٪ کاربران اینترنتی وب سایت‌های جدید را از طریق موتور جستجو پیدا می‌کنند. جویشگرها سرویس جستجو را بر روی اقلام مختلف موجود در وب در قالب خدمات مجزائی نظیر جویشگر محتوای متنی، جویشگر تصاویر، جویشگر اصوات، جویشگر ویدئوها و غیره، ارائه می‌دهند. در سرویس جستجوی متنی وب که در آن عمدتا جستجو بر روی محتوای متون موجود در در صفحات وب انجام می‌شود، معمولا انتظار بیشتری از جویشگر وجود دارد معمولا انتظار بیشتری از جویشگر وجود دارد و آن پردازش هوشمند پرس و جو و نیاز اطلاعاتی کاربر و استفاده از بهترین منابع موجود برای پاسخگویی به نیاز او می‌باشد.

هدف از این پژوهش بررسی جویشگرهای متنی موجود و تجارب موفق سایر کشورها در زمینه بومی‌سازی آن می‌باشد. تا در نهایت به یک جمع‌بندی از عوامل موفقیت این جویشگرها برسیم.

اطلاعات مرتبط

مستندات مرتبط

شماره مستند	نوع مستند	نام مستند

تغییرات اعمال‌شده در نسخه‌های پیشین

شماره نسخه	تاریخ	تغییرات اعمال شده
۱,۰	۹۴/۰۴/۱۴	آماده‌سازی گزارش
۱,۰۲	۹۴/۰۶/۲۲	چکیده و نتیجه گیری باتوجه به فیدبک ناظرین اصلاح شد.

تأییدکنندگان

	نام و نام خانوادگی	تاریخ	امضاء	ملاحظات
مجری پروژه	علیرضا یاری			
تهیه کننده/ تهیه کنندگان	تیم مدیریت طرح			
ناظر پروژه	کامبیز بدیع، رامین شکری‌پور و روح‌اله رحمانی			
مدیر گروه				
مسئول مستندات پژوهشکده	مانا روزی طلب			
رئیس پژوهشکده/ معاون پژوهشی	علیرضا یاری			

اسامی اعضای تیم مدیریت طرح بر اساس حروف الفبا	
۱. محمد آزادنیا	۱۱. مهدی عمادی
۲. محمد مهدی اثنی اشری	۱۲. مژگان فرهودی
۳. شهره جهانبخش	۱۳. محمد مهدی کیخا
۴. مونا داوودی	۱۴. مریم محمودی
۵. غزاله رحمانی فرزین	۱۵. پویان مسعودی فر
۶. فرزانه رحمانی	۱۶. اکبر مقدر
۷. فرزاد زرگری	۱۷. امین میرزائی
۸. محمد صادق زاهدی	۱۸. محمدرضا میرصراف
۹. علی شریفی	۱۹. حمیدرضا نصیری آسایش
۱۰. معصومه عظیم زاده	۲۰. علیرضا یاری
۱۱. طاهره علوی زرگر	

سرفصل مطالب

موتورهای جستجوی متنی وب ۹
۱,۱ موتور جستجوهای متنی فارسی ۱۱
۱,۱,۱ موتور جستجوی متنی پارسیجو ۱۲
۱,۱,۲ موتور جستجوی متنی یوز ۱۴
۱,۱,۳ موتور جستجوی متنی ریسمون ۱۶
۱,۱,۴ موتور جستجوی متنی زال ۱۸
۲,۱ موتور جستجوی متنی بومی سایر کشورها ۱۹
۱,۱,۵ موتور جستجوی بومی روسیه ۱۹
۱,۱,۶ موتور جستجوی بومی کره جنوبی ۱۹
۱,۱,۷ موتور جستجوی بومی چین ۲۰
۱,۱,۸ موتور جستجوی بومی ویتنام ۲۰
۱,۱,۹ موتور جستجوی بومی جمهوری چک ۲۱
۳,۱ موتور جستجوهای متنی جهانی ۲۱
۱,۱,۱۰ موتور جستجوی گوگل ۲۱
۱,۱,۱۱ موتور جستجوی یاهو ۲۲
۱,۱,۱۲ موتور جستجوی بینگ ۲۲
۴,۱ شاخصهای ارزیابی جویشگرهای متنی ۲۳
۵,۱ جمعبندی ۳۱

فهرست جداول

جدلول ۱: نتایج متناظر سه پرس وجو..... ۲۷
جدلول ۲: خلاصه ارزیابی و معیارهای سنجش آنها ۳۰

فهرست اشکال

شکل ۱:نمای کلی سیستم جستجوی اطلات وب ۱۰

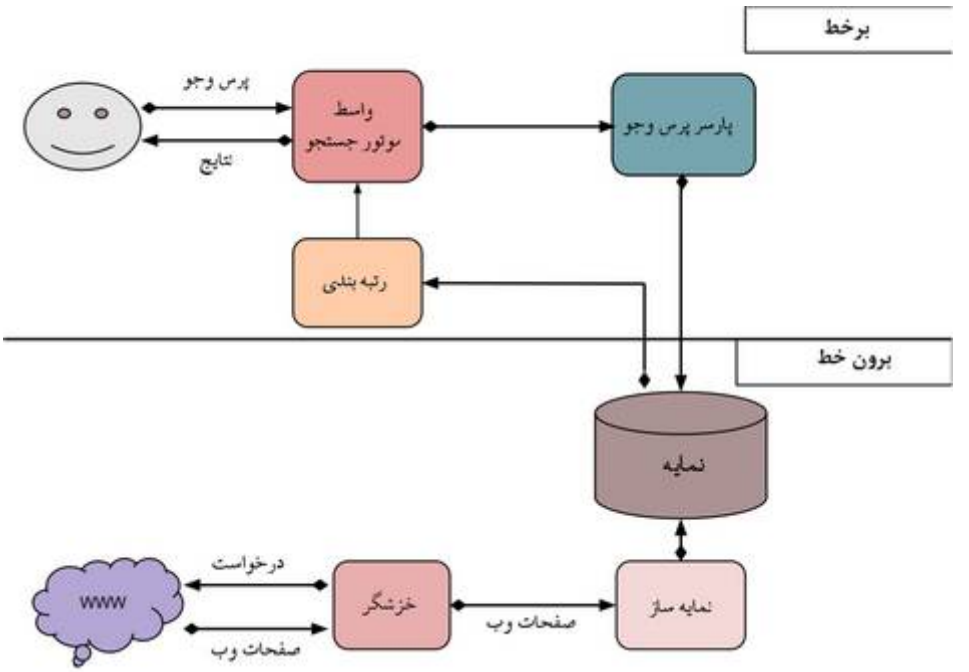
موتورهای جستجوی متنی وب

وب عمده‌ترین محیط خدماتی اینترنت است که امکانات چندرسانه‌ای برای دستیابی به داده‌ها، اطلاعات، فایل‌ها و دانش را در اختیار کاربران قرار می‌دهد و در دنیای امروزی تبادل اطلاعات بین افراد در سراسر جهان را بسیار ساده و آسان نموده است. استفاده از وب در سالیان اخیر، رشد زیادی داشته است. یکی از دلایل اصلی این مهم، رفع نیازهای اطلاعاتی کاربران در حداقل زمان ممکن است. موتورهای جستجو ازجمله ابزارهای بسیار کارآمد در این زمینه می‌باشند؛ چنانکه موتورهای جستجوی مشهوری مانند گوگل و یاهو نقطه ورود بسیاری از کاربران برای جستجو در وب هستند.

شِمای یک سیستم جستجوی اطلاعات وب را نشان می‌دهد. ابتدا صفحات وب توسط خزشگر[۱] جمع‌آوری شده و در مخزن اسناد قرار می‌گیرند. بعد از عمل خزش ساختار وب شامل پیوندها (گراف) و ابرپیوندها[۲] ایجاد خواهد شد. به دنبال خزش واحد نمایه‌ساز، اسناد جمع آوری شده را شاخص‌گذاری می‌کند. سیستم در پاسخ به پرسش کاربر اسناد مرتبط را از نمایه سیستم استخراج کرده و طبق معیارِ رتبه‌بندی، مرتب نموده و به کاربر نشان می‌دهد.

---

[۱] Crawler
[۲] Hyper Link



شکل ۱:نمای کلی سیستم جستجوی اطلات وب

تحقیقات نشان می‌دهد که بیش از ۳,۷ میلیارد از کاربران در هر ماه به موتورهای جستجو مراجعه می‌کنند. درحقیقت بیش از ۸۰٪ کاربران اینترنتی وب سایت‌های جدید را از طریق موتور جستجو پیدا می‌کنند. بر اساس اطلاعات منتشره در وب سایت معتبر الکسا، موتورهای جستجوی گوگل، یاهو و بایدو به ترتیب رتبه‌های اول، چهارم و پنجم ترافیک جهانی وب را به خود اختصاص داده‌اند.

این جویشگرها دسترسی به محتوای موجود و رفع نیاز اطلاعاتی کاربر در وب را در کمترین زمان ممکن از طریق تنوعی از سرویس‌ها تسهیل می‌کنند. در واقع جستجو بر روی اقلام مختلف موجود در وب در قالب خدمات مجزائی نظیر جویشگر محتوای متنی، جویشگر تصاویر، جویشگر اصوات، جویشگر ویدئوها و غیره، ارائه می‌شود.

در سرویس جستجوی متنی وب که در آن عمدتا جستجو بر روی محتوای متون موجود در در صفحات وب انجام می‌شود، معمولا انتظار بیشتری از جویشگر وجود دارد و آن اینست که در صورتی که کاربر در جستجوی اقلامی دیگری نیز باشد جویشگر بتواند با تشخیص هوشمند پرس و جو سرویس مناسب را شناسایی نموده و در بین نتایج بازگشتی حداقل یک پاسخ مناسب از سرویس شناسایی شده را نیز ارائه دهد.

از منظر گستره پوشش کاربران موتورهای جستجوی موجود را می‌توان در دو گروه جهانی و بومی دسته‌بندی نمود. موتورهای جستجوی گوگل، یاهو و بینگ از نوع جهانی می‌باشند و با هدف خدمات‌دهی به کلیه کاربران وب جهانی توسعه یافتند. موتورهای جستجوی بومی توسط برخی از کشورهای جهان برای مقابله با نفوذهای فرهنگی و امنیتی موتورهای جستجو و نیز بهره‌مندی اقتصادی توسعه یافتند که به عنوان نمونه می‌توان به یاندکس روسیه،

ناور کره و بایدو چین اشاره نمود. در ادامه ابتدا به بررسی وضعیت موتورهای جستجوی متنی فارسی پرداخته و سپس چند نمونه موتور جستجوی بومی موفق که در سایر کشورها توسعه داده شده و طیف وسیعی از کاربران را به خود جذب نموده تشریح می‌شود. و در نهایت چند موتور جستجو که در سطح بین‌المللی خدمات ارائه می معرفی خواهند شد.

مطالعات و بررسی‌های به عمل آمده در خصوص شاخصهای عمده موفقیت موتورهای جستجوی بومی نشان می‌دهد که علل عمده موفقیت این موتورهای جستجو علاوه بر فراهم‌ساختن پوشش مناسب و کیفیت مطلوب در ارائه نتایج، توجه به فرهنگ، خط، زبان و محتوای بومی از یک سو و ارائه خدمات متنوع و جذاب از سوی دیگر بوده است.

با توجه به اینکه تمرکز این گزارش موتورهای جستجوی متنی می‌باشند در ادامه به بررسی موتورهای جستجوی متنی مطرح خواهیم پرداخت. علاوه بر اینکه وضعیت این موتورهای جستجو را در کشور را تشریح خواهیم نمود به بررسی موتورهای جستجوی متنی بومی سایر کشورها و همچنین موتورهای جستجوی جهانی خواهیم پرداخت.

۱,۱ موتور جستجوهای متنی فارسی

موتورهای جستجوی متنی فارسی موجود به دو دسته جویشگرها و فراجویشگرها قابل تقسیم‌بندی هستند. جویشگرها موتورهای جستجویی هستند که از ابتدا در کشور توسعه داده شده و خود به خزش وب و نمایه‌سازی و رتبه‌بندی صفحات گردآوری شده می‌پردازند. اما در فراجویشگرها، موتورجستجو خزشی انجام نداده و ترکیبی از بازرتبه‌بندی نتایج منتخب سایر موتورهای جستجوی موجود را به عنوان نتایج رتبه‌بندی خود به کاربر ارائه می‌دهد.

مهمترین موتورهای جستجوی متنی توسعه داده شده در داخل کشور موتور جستجوی پارسی‌جو و یوز می‌باشند که در حال حاضر در قالب شرکت‌های دانش‌بنیان به فعالیت خود ادامه می‌دهند. این دو موتور جستجو با پوشش بیش از پانصد میلیون صفحه وب فارسی خدمات جستجوی متنی را به کاربران فارسی‌زبان ارائه می‌نمایند.

پارسی‌جو و یوز علاوه بر سرویس جویشگر متنی دارای تنوعی از خدمات ارزش افزوده و سایر خدمات جستجو هستند. به عنوان نمونه پارسی‌جو شامل خدماتی نظیر جستجوی تصویری، ویدئو، نقشه، علمی، آوا و دانلود و غیره و یوز نیز علاوه بر جستجوی متنی شامل جستجوی پدیده‌های وبلاگی و خبری، تازه‌های تصویری، اخبار مهم و ورزشی می‌باشد. هر دو جویشگر متنی با پردازش هوشمند پرس و جو و شناسایی نیاز کاربر در صورت نیاز از سایر خدمات نیز نتیجه‌های مرتبط را به کاربر ارائه می‌دهند.

علاوه بر پارسی‌جو و یوز دو موتور جستجوی دیگر که نمایه‌های مستقلی دارند عبارتند از زال و ریسمون که در بخشهای مربوطه به تشریح مشخصات دقیق‌تر آنها خواهیم پرداخت.

در زمینه فراجویشگرها نیز موتورهای جستجویی نظیر سلام، پارسیک، جاماسپ و بیاب مطرح هستند. این موتورهای جستجو نتایج موتورهای جستجوی موجود نظیر گوگل و بینگ را ترکیب نموده و در اختیار کاربران قرار می‌دهند. با توجه به اینکه در طرح جویشگر ایجاد موتورجستجوی مستقل از نمونه‌های خارجی مد نظر بوده است شامل این گونه فراجویشگرها نمی‌شود و در این گزارش به آنها نخواهیم پرداخت.

۱،۱،۱ موتور جستجوی متنی پارسی‌جو

شرح:

موتور جستجوی پارسی‌جو" با هدف پوشش ۵۰۰ میلیون صفحه فارسی به همراه سرویس‌های ارزش افزوده با قابلیت جذب و پاسخگویی به دومیلیون پرس‌وجو در روز از سال ۸۹ توسعه داده شده است. در واقع هدف پارسی جو تنها ارائه جویشگر متنی نیست و برای رسیدن به تعداد کاربر قابل قبول باید سایر خدمات جذاب را نیز برای کاربر فراهم نماید. اما با توجه به تمرکز این بخش روی جویشگر متنی به ارائه مشخصه‌های این بخش اکتفا می‌کنیم.

- شمای کلی پارسی جو:

پارسی جو مانند هر جویشگر متنی دیگر از سه مولفه اصلی خزشگر، نمایه‌ساز و جویشگر یا بازیابی اطلاعات تشکیل شده است. مشخصه‌های مرتبط برای پیاده سازی یک موتور جستجوی با پوشش بیش از پانصد میلیون صفحه و پاسخگویی به میلیون‌ها پرس‌وجو در روز در ادامه ارائه می‌گردد.

خزشگر ویژگیهایی مانند خزش بیش از پانصد میلیون صفحه، نگهداری بیش از یک میلیارد آدرس صفحه را به همراه تازگی هوشمند دارد. بخش پارسر و نمایه سازی برای مدیریت و پردازش در مقیاس بالا در محیط کاملا توزیع شده انجام می‌شود. بخش جستجوی برخط مسئول پاسخگویی به میلیون‌ها پرس‌وجو در روز می‌باشد. پارسی‌جو برای پاسخ دهی بیشتر چند کپی از نمایه ایجاد نموده است. با توزیع سازی نمایه‌ها، بهینه سازی و گذاشتن نمایه داخل حافظه تلاش نموده به سرعت قابل قبولی در پاسخگویی به نیاز کاربران دست یابد. علاوه بر موارد ذکر شده زیرسیستمهای زیر نیز برای پاسخگویی بهتر به نیاز کاربران در نظر گرفته شده است.

- زیرسیستم پیشنهاد دهنده پرس و جو:

بخش پیشنهاد دهنده پرس‌وجو یکی از بخش‌هایی است که استفاده کاربران را از سیستم راحت میکند. در این بخش تمام پرس‌وجوهای کاربران بررسی شده و با توجه به زمان و فرکانس آنها به کاربر ارائه می‌شود.

- زیرسیستم خطایابی فارسی:

این ماژول بر پایه ی داده های وب پرس و جوی کاربر را اصلاح نموده و او را به مقصود نهایی جستجویش نزدیک می‌کند.

- زیرسیستم کشینگ:

هدف کشینگ نگهداری تمام صفحات داخل موتور جستجوی میباشد. به عبارت دیگر لازم است پانصد میلیون سند را در خود ذخیره سازی نماید.

مشخصه‌های فنی مورد انتظار :

خلاصه مشخصه‌های فنی مورد انتظار این محصول که مشخصه‌های اعلامی مجری پروژه می باشد عبارتند از:

- فرمت اسناد مورد پشتیبانی : متنی شامل HTML, RTF, TEXT و XHTML
- زبان مورد پشتیبانی: انگلیسی و فارسی
- تعداد کاربران همزمان (به ازای پرسشهای سه کلمه‌ای): ۱۰۰ کاربر
- نوع جستجو: جستجوی منطقی (Boolean Search)، جستجوی عبارت (Phrase Search) و نیز جستجو روی دامنه مشخص اسناد
- زمان پاسخ: کمتر از یک ثانیه
- میزان دقت: ۸۵٪ دقت Google
- ارایه سرویس خطایاب فارسی
- نوع نمایه‌سازی: Inverted Index
- نوع رتبه‌بندی: استفاده از ترکیب محتوای صفحات و گراف وب بصورت نهایی
- نرخ بروز رسانی اسناد جمع‌آوری شده: تطبیقی بر حسب نوع محتوا و نوع سایت
- بهره‌مندی از مکانیسم Caching
- امکان ارایه خدمات جستجو بصورت وب سرویس

وضع موجود:

ویژگی‌ها و خدمات موتور جستجوی پارسی‌جو به شرح زیر است:

۱-سرویس جستجوی متنی: موتور جستجوی متنی تا کنون ۵۰۰ میلیون صفحه فارسی را پوشش داده است.

۲-سایر خدمات:

- سرویس جستجوی تصویری
- سرویس جستجوی آوا
- سرویس خبری
- سرویس جستجوی علمی
- سرویس جستجوی نقشه

نحوه دسترسی:

این سامانه در حال حاضر از طریق آدرس زیر در دسترس می‌باشد:

www.parsijoo.ir

متولی:

توسعه دهنده این سامانه شرکت وب‌پردازان پارسی‌جو می‌باشد. این پروژه با حمایت پژوهشگاه ارتباطات و فناوری اطلاعات (مرکز تحقیقات مخابرات ایران) اجرا شده‌است.

۱،۱،۲ موتور جستجوی متنی یوز

شرح:

یوز یک موتور جستجوی ایرانی است که بر منابع فارسی موجود در فضای وب تمرکز دارد. طراحی و تولید موتور جستجوی یوز از اواخر سال ۱۳۸۸ با تلاش نیروهای متخصص داخلی آغاز شده است. یوز بنابر ادعای مجری پروژه تاکنون توانسته است بیش از یک میلیارد صفحه را پوشش دهد. یوز همچنین دارای خدمات جستجوی خبر، وبلاگ و عکس می‌باشد.

ویژگی‌های موتور جستجوی یوز:

- تمرکز بر زبان فارسی
- سرعت بالا با هدف گذاری پاسخ‌دهی سریع به کاربران با میانگین تأخیر کمتر از ۱ ثانیه

- تحلیل نیازهای متداول کاربران و پاسخ‌دهی مستقیم به چندین نوع از جستجوهای کاربران
- نمایه‌سازی بی‌درنگ: قابل جستجو نمودن صفحات جدید چند دقیقه پس از خزش
- معماری مقیاس‌پذیر به نحوی که برای افزایش پوشش صفحات وب، فقط کافیست ماشین‌های جدید به خوشه‌ها اضافه شود

مشخصه‌های فنی مورد انتظار :

مشخصه‌های فنی مورد انتظار این محصول در واقع مشخصاتی است که برای نسخه اولیه این محصول دیده شده است و لزوما نشان‌دهنده وضعیت موجود آن نمی‌باشد. خلاصه‌ای از این مشخصات در ادامه ارائه شده است:

- فرمت اسناد مورد پشتیبانی : متنی شامل HTML, RTF, TEXT و XHTML
- زبان مورد پشتیبانی: انگلیسی و فارسی
- تعداد کاربران همزمان (به ازای پرسشهای سه کلمه‌ای): ۵۰۰ کاربر
- نوع جستجو: جستجوی منطقی (Boolean Search)، جستجوی عبارت (Phrase Search) و نیز جستجو روی دامنه مشخص اسناد
- زمان پاسخ: کمتر از یک ثانیه
- میزان دقت: ۷۰٪ دقت Google
- ارایه سرویس خطایاب فارسی
- نوع نمایه‌سازی: Inverted Index پس از حذف Stop-Words، شناسایی صفحات Spam و نیز Stemming
- نوع رتبه‌بندی: استفاده از ترکیب محتوای صفحات و گراف وب بصورت نهایی (مثلاً روش HITS)
- تنوع کدینگهای اسناد و پرس‌وجوها: Unicode، UTF-۸، Windows-۱۲۵۲ و Windows-۱۲۵۶
- نرخ بروز رسانی اسناد جمع‌آوری شده: تطبیقی بر حسب نوع محتوا و نوع سایت
- امکان فیلتر گذاری بر حسب فهرست سایتها یا موضوعات مورد نظر
- بهره‌مندی از مکانیسم Caching
- دسته‌بندی وب سایت‌هایی که نمایه‌سازی می‌شوند.
- امکان ارایه خدمات جستجو بصورت وب سرویس

وضع موجود:

- خدمات: موتور جستجوی متنی با پوشش بیش از یک میلیارد صفحه فارسی

- سایر خدمات: جستجوی خبر، وبلاگ و عکس

نحوه دسترسی:

این سامانه در حال حاضر از طریق آدرس زیر در دسترس می‌باشد:

www.yooz.ir

متولی:

توسعه دهنده این سامانه آزمایشگاه پردازش زبان طبیعی دانشگاه تهران می‌باشد. این پروژه با حمایت پژوهشگاه ارتباطات و فناوری اطلاعات (مرکز تحقیقات مخابرات ایران) اجرا شده‌است.

۱،۱،۳ موتور جستجوی متنی ریسمون

شرح:

ریسمون یکی از قدیمی‌ترین جویشگرهای وب فارسی است. این جستجوگر همه وب سایتهای فهرست Link.ir را در دوره های زمانی یک ماهه می پیماید و محتویات و مطالب آنها را نمایه سازی می کند و در بانک اطلاعاتی خود جهت ارائه خدمات جستجو به مراجعه کنندگان نگهداری می نماید. پروژه جستجوگر ریسمون از مهرماه ۱۳۸۳ در شرکت رادکام آغاز گردیده است.

- شمای کلی ریسمون:
- مؤلفه پیمایش و نمایه سازی :

این مؤلفه سامانه ای است که با خزیدن در وب سایت ها، محتویات آنها را برای فرایند نمایه سازی و قراردادن در بانک اطلاعاتی، پالایش می کند. این سامانه معمولا برای یافتن وب سایت ها، از فهرست مرجع خود استفاده می کند. موتورهای جستجوگر غول پیکری مانند Google،دیگر تنها به فهرست مرجع خود متکی نیستند(البته در ابتدا متکی بود) بلکه همه پیوندهای موجود در هر وب سایتی را به صورت زنجیره ای دنبال می کنند. اما ریسمون فهرست مرجعی دارد که سامانه پیمایش و نمایه سازی تنها به نمایه سازی وب سایتهای آن اقدام می نماید. این فهرست مرجع، وب سایت Link.ir است.

- مؤلفه پویش و جستجو:

این مؤلفه، بانک اطلاعاتی را که بوسیله مؤلفه پیمایش و نمایه سازی پر شده است را برای عبارات مورد نظر، جستجو می‌کند. در واقع هسته اصلی جستجوگر در این این مؤلفه نهفته است. در مورد این مؤلفه آنچه از همه مهمتر است زمان پاسخ آن است و باید الگوریتم های مناسب و نیز زیرساختهای سخت افزاری کارا، برای به حداقل رساندن زمان پاسخ این مؤلفه به کار گرفته شود.

- مؤلفه پایش و نظارت:

این مؤلفه نقش نظارتی و تحلیلی بر خدمات موتور جستجوگر دارد. اطلاعات آماری که از این مؤلفه استخراج می گردد بسیار ارزشمند است و به عنوان مثال مشخص می کند که مردم بیشتر به چه چیزی علاقه دارند و به دنبال چه می گردند. علاوه بر این امکان ارائه اطلاعات آماری جغرافیایی نیز وجود دارد. تحلیل این اطلاعات آماری از دیدگاه های مختلف تجاری، فرهنگی و اجتماعی و حتی امنیت ملی بسیار مهم و قابل توجه است.

قابلیتهای عمومی مورد انتظار ریسمون عبارتند از:

- نمایه سازی اسناد و پرونده هایی با قالبهای Open Office, Text RTF, Microsoft Office, PDF, HTML با پشتیبانی کامل از زبان فارسی.
- نمایه سازی محتویات فارسی بصورت جامع ، بطوریکه صفحاتی که با کاف و یای عربی تولید شده اند نیز به صورت فارسی نمایه سازی می شوند و در خروجی جستجو ظاهرمی گردند.
- نمایه سازی وب سایتهایی که با پروتکل SSL امن شده اند.
- نمایه سازی اسناد و پرونده هایی که در FTP سایت قراردارند.
- بهره مندی از Caching برای نگهداری اسناد و صفحات نمایه سازی شده.
- زمانبندی نمایه سازی مجدد.
- دسته بندی وب سایتهایی که نمایه سازی می شوند و امکان ارائه خدمات جستجو بصورت یک وب سرویس به وب سایتهای تابعه پورتالها دارد.
- قابلیت اعتبارسنجی برای ورود به بخشهای اینترانتی که برای دسترسی به آنها به گذرواژه نیاز است.
- پشتیبانی از stop words برای مشخص کردن کلمات و یا عباراتی که لازم نیست در نتیجه جستجو ظاهر شوند، مانند حروف اضافه "و"، "از"، "به" ، ...
- جستجوی منطقی بصورت ترکیب عطفی، فصلی و یا نفی از منطق بولی.
- بهره مندی از جستجوی پیشرفته با قابلیتهایی نظیر محدود کردن جستجو به موضوع مورد نظر.
- اجرای چند نسخه از برنامه نمایه سازی و عملکرد همزمان آنها برای تسریع در امر نمایه سازی.
- بهره مندی از روال رتبه بندی نتایج جستجو به طوریکه صفحات و اسناد با ارتباط بیشتر، رتبه بالاتری در خروجی جستجو دارند.
- نمایش تعداد کل نتایج یافت شده.
- محاسبه و نمایش زمانی که صرف جستجو شده است.

وضع موجود:

- خدمات: موتور جستجوی متنی با پوشش تخمینی ۲ میلیون صفحه وب فارسی
- سایر خدمات: اخبار(نتایج بروز نیستند)، اطلاعات ۱۱۸ (قطع است)، شعر و ادب

نحوه دسترسی:

این سامانه در حال حاضر از طریق آدرس زیر در دسترس می‌باشد:

http://www.rismoon.com

متولی:

شرکت رادکام.

۱,۱,۴ موتور جستجوی متنی زال

شرح:

زال یک موتور جستجوی فارسی است که توسط شرکت بیان با هدف ارتقا کیفیت خدمات جستجو برای کاربران فارسی زبان در حال توسعه است. در حال حاضر استفاده از زال تنها از طریق فراجستجوگر سلام امکان پذیر است.

ویژگی‌ها و قابلیت های مورد انتظار زال عبارتند از:

- تشخیص هرز نوشته ها و تله های تبلیغاتی
- تشخیص محتوای غیر اخلاقی
- تکمیل خودکار عبارات
- ریشه یابی لغات
- پیشنهادات مشابه
- ابهام زدایی
- دسته بندی مفهومی نتایج
- تاریخچه مصور صفحات

وضع موجود و نحوه دسترسی:

جویشگر زال در فراجویشگر سلام به عنوان یکی از موتورهای جستجویی است که از ترکیب نتایج آنها برای پاسخگویی به نیاز کاربر استفاده می‌شود و تا کنون خروجی تحت وب مستقلی از آن ارائه نگردیده است. همچنین از طریق آدرس www.zal.ir فقط معرفی اجمالی از آن موتور جستجو ارائه شده است و از طریق این پیوند خدمت جستجو ارائه نمی‌کند:

http://www.zal.ir

http://www.salam.ir

متولی:

توسعه دهنده این سامانه شرکت بیان می‌باشد.

۲,۱موتور جستجوی متنی بومی سایر کشورها

۱,۱,۵ موتور جستجوی بومی روسیه

یاندکس که در سال ۱۹۹۷ در روسیه پایه‌گذاری شد، موتور جستجوی بومی روسیه و اکراین است. این موتور جستجو ۵۶ میلیون کاربر دارد که اکثرا از کشورهای روسیه، اکراین، قزاقستان و بلاروس می‌باشند. یاندکس در حال حاضر بیش از ۲ بیلیون صفحه را نمایه‌سازی کرده است که رقم چشمگیری محسوب می‌شود. ۶۰ درصد کاربران روسی از یاندکس و ۳۲ درصد از گوگل استفاده می‌کنند. دلیل اصلی موفقیت این موتور را می‌توان در موارد زیر خلاصه کرد:

- پردازش کامل زبان روسی
- ارائه سرویس‌های ویژه بومی به روسیه (تلویزیون، موسیقی)
- پوشش قابل توجه وب

Mail.ru نیز اولین انتخاب روس‌ها در زمینه پست‌الکترونیک است. این سایت نیز طرفداران فراوانی در کشورهای دیگر دارد.

۱,۱,۶ موتور جستجوی بومی کره جنوبی

ناور که موتور جستجوی بومی کره‌جنوبی می‌باشد، در سال ۱۹۹۹ توسط شرکت سامسونگ در کره جنوبی پایه گذاری شده است. ۷۳ درصد کاربران اینترنت کره‌ای از ناور و فقط دو درصد از گوگل استفاده می‌کنند! و توانسته گوگل را با شکست مواجه کند. این درگاه در ازبکستان، بلاروس، قزاقستان و اکراین نیز طرفدار دارد. پرتال ناور در کره‌جنوبی مرجع بسیاری از کاربران برای رفع نیازمندی‌هایشان است. خدماتی نظیر موتور جستجو، وبلاگ، پست‌الکترونیک و بازی‌های آنلاین، ناور را به پنجمین وب‌سایت پربازدید در کره‌جنوبی تبدیل کرده است. دلیل اصلی موفقیت این موتور را می‌توان در موارد زیر خلاصه کرد:

- عِرق ملی کره‌ای‌ها در استفاده از محصولات بومی

- ارائه خدمات ویژه بومی مطابق با فرهنگ کشور کره (به ویژه صفحه اصلی ناور با توجه به فرهنگ کره‌ای طراحی شده است)

وب‌سایت‌های دیگری نظیر DAUM نیز با دارابودن امکاناتی نظیر پست‌الکترونیک، جستجو و بخش‌های خبری نقش مهمی در برآورده‌سازی نیازهای اینترنتی کاربران کره‌ای دارد.

۱،۱،۷ موتور جستجوی بومی چین

موتور جستجوی بایدو در کشور چین در سال ۲۰۰۰ پیاده‌سازی شده است. این موتور در چین رتبه یک، در کشور هنگ‌کنگ و کره‌جنوبی رتبه شش، در کشور تایوان رتبه هشت و در کشور ژاپن رتبه ۱۴ برترین وب‌سایت‌های اینترنتی را به دست آورده است. نکته قابل توجه آن است که این موتور جستجو توانسته در دنیا رتبه ۶ را به خود اختصاص دهد. در چین موتور جستجوی بایدو ۷۰ درصد سهم بازار و گوگل ۲۰ درصد سهم بازار را در دست دارد.

موتور بایدو دارای ۵۷ سرویس متنوع می‌باشد که باعث شده است ۵۸ درصد ترافیک جستجوی چین را به سمت خود جلب کند. بایدو درحال حاضر ۷۴۰ میلیون صفحه وب، ۸۰ میلیون تصویر و ده میلیون فایل مالتی‌مدیا را نمایه‌سازی کرده است. بایدو امروزه توانسته در عرصه جستجوی اینترنتی بازار پر رونق چین را به خود اختصاص دهد، به طوری که توانسته درصد بالایی از مشتریان و تبلیغ‌دهندگان را که پیش از این از کاربران گوگل بودند، اکنون به سوی خود جذب کند. به همین خاطر است که اخیراً گوگل برای حضور در بازار چین حاضر شده تمامی شرط و شروط دولت چین را در زمینه محتوای ارائه شده به کاربران بپذیرد.

جستجو تنها یکی از خدمات بایدو می‌باشد و این شرکت علاوه بر آن تنوعی از خدمات را به کاربرانش ارائه می‌دهد. به نحوی که توانسته به راحتی نیاز ۴۹۰ میلیون کاربر چینی را برآورده می‌سازد. دلیل اصلی موفقیت این موتور را می‌توان در موارد زیر خلاصه کرد:

- سرویس‌های متنوع بومی
- پشتیبانی کامل زبان چینی

۱،۱،۸ موتور جستجوی بومی ویتنام

ویتنام یکی از بازارهایی است که در آن گوگل با یک رقیب قدرتمند در زمینه جستجوی وب مواجه است. کمپانی ویتنامی Coc Coc در ژانویه ۲۰۱۳ موتور جستجوی رایگان خود را با انگیزه تسخیر بازار ویتنام روانه بازار کرد و علاوه بر این موتور جستجو، یک مرورگر، یک اپلیکیشن موبایل مبتنی بر مکان و یک سیستم تبلیغاتی مناقصه‌ای نیز ارائه نمود. این موتور جستجو در مدت ۲ سال از گوگل پیشی گرفته و رتبه اول جستجو در

ویتنام گردیده است. موتور جستجوی Coc Coc که به عنوان جدی‌ترین رقیب گوگل در ویتنام فعالیت می‌کند، توسط شرکتهایی چون Yandex موتور جستجوی بومی روسیه، Mail.ru ارائه دهنده خدمات ایمیل و اینترنت بومی روسیه و شرکت سرمایه‌گذاری بین‌المللیDigital Sky Technologies یکی از سرمایه‌گذاران فیسبوک حمایت می‌شود.

این موتور جستجو، بزرگترین پایگاه داده ویتنام را با بیش از ۲,۱ میلیارد صفحه اینترنتی، در اختیار دارد که در آن میزان داده با نام دامنه اختصاصی ویتنام شامل .com.vnو.vn دو برابر بیشتر از گوگل است. ادعای Coc Coc این است که در طراحی موتور جستجویش از الگوریتم‌هایی استفاده کرده که امکان جستجو به زبان ویتنامی را با کیفیت و سرعت بیشتری به کاربر میدهد و البته در کنار این دقت و سرعت، دسترسی به منابع و اطلاعات محلی نیز، برای آن یک مزیت رقابتی با ارزش محسوب می‌شود.

۱,۱,۹ موتور جستجوی بومی جمهوری چک

سزنم در در سال ۱۹۹۶ توسط ایوو در پراگ به‌عنوان اولین درگاه وب در جمهوری چک تاسیس شده است. در مقایسه با بقیه اتحادیه اروپا که در آن گوگل حاکم است در جمهوری چک گوگل رتبه دوم را داراست.

سزنم در ابتدا با ذخیره ۵۰۰۰۰ شرکت تجاری مانند یک کتاب زرد شروع بکار نمود. سپس ایمیل را به سبد محصول خود افزود. تا سال ۲۰۰۱ سزنم شامل اخبار سیاسی ، اجتماعی و اقتصادی شد. در ماه مه ۲۰۰۸، سزنم جویشگر اینترنتی غالب در جمهوری چک با آمار (٪۵۹,۸۹) در مقایسه با جویشگر گوگل (٪۳۱,۶۱) بوده است. همچنین در سال ۲۰۱۲ سزنم ٪۴۲,۸۴ از جستجوی کاربران کشور چک را به خود اختصاص داد. سزنم مانند پنج شرکت دیگر در جهان، هنوز هم جویشگر شماره یک در منطقه خود است.

سزنم: ۷۰ درصد سهم بازار چک

۳,۱ موتور جستجوهای متنی جهانی

۱,۱,۱۰ موتور جستجوی گوگل

موتور جستجوی گوگل در سال ۱۹۹۶ در دانشگاه استنفورد پیاده‌سازی و در سال ۱۹۹۸ کار خود را به عنوان یک شرکت مستقل آغاز نمود. موتور گوگل در اکثر کشورها بیشترین استفاده کننده را به خود اختصاص

داده است. مطابق گزارش ComScore درسال ۲۰۱۳، ۶۵ درصد جستجوها در گوگل انجام شده است. دلیل اصلی موفقیت این موتور را می‌توان در موارد زیر خلاصه کرد:

- رتبه‌بندی مناسب: رتبه‌بندی به صورت عادلانه و با توجه به شعار "شیطان نباش". به عبارت دیگر معیارهای خارجی مانند پرداخت پول، در بالابردن رتبه وب‌سایت‌ها تأثیر نخواهد داشت.
- واسط کاربر ساده
- پردازشگر پرس‌وجوی چند زبانه

در آمد اصلی موتور گوگل (۹۹٪) حاصل از تبلیغات می‌باشد که با استفاده از مدل تجاری PPC و تبلیغات به صورت Adwords انجام می‌شود. در این مدل، تبلیغات با استفاده از کلمات کلیدی موجود در پرس‌وجوی کاربر نمایش داده می‌شود.

۱،۱،۱۱ موتور جستجوی یاهو

موتور جستجوی یاهو نیز ابتدا در دانشگاه استنفورد پیاده‌سازی شد و سپس در سال ۱۹۹۵ در قالب یک شرکت، کار خود را بصورت رسمی در آمریکا شروع کرد. بنا بر آمار کلی، موتور یاهو از نظر تعداد کاربران بعد از گوگل قرار دارد. لازم به ذکر است که یاهو قبل از به وجود آمدن گوگل بیشترین ترافیک را به خود اختصاص داده بود. از سال ۲۰۰۰ الی ۲۰۰۴ سرویس جستجوی یاهو توسط گوگل انجام می‌گرفت. از جمله دلایل اصلی موفقیت این موتور در بعضی کشورها، می‌توان به واسط بومی متناسب با کشور مربوطه اشاره نمود. یاهو برای کشورهای مختلف واسط بومی و متفاوت ارائه می‌کند که باعث جذب هر چه بیشتر کاربران آن ناحیه می‌گردد (برای مثال در ژاپن، تایوان و هنگ‌کنگ، یاهو رتبه یک را دارا می‌باشد).

دلیل اصلی عقب افتادگی یاهو از گوگل، را می‌توان بوجود آمدن رقیبی قدرتمند همچون گوگل و همچنین استفاده یاهو از گوگل برای سرویس جستجو به مدت چهار سال دانست.

۱،۱،۱۲ موتور جستجوی بینگ

موتور جستجوی بینگ توسط شرکت مایکروسافت ایجاد و در سال ۲۰۰۹ بر روی وب قرار گرفت. در حقیقت بینگ ویرایش جدید موتور جستجوی قبلی مایکروسافت یعنی ام.اس.ان و لایو محسوب می‌شود. گرچه در ابتدا بینگ نتوانست درصد قابل توجهی از کاربران را به خود اختصاص دهد ولی در سال ۲۰۱۱ پیشرفت چمشگیری از خود نشان داده است. برای مثال در مارس ۲۰۱۱ در آمریکا، گوگل ٪۶۴ و بینگ ٪۳۰ ترافیک را به خود اختصاص داده است. با این روند احتمال می‌رود که در آینده نزدیک درصد قابل توجهی از کاربران را به خود جذب کند. دلایل اصلی موفقیت این موتور را می‌توان در موارد زیر خلاصه کرد:

- واسط کاربری پویا و جالب که مرتباً نیز در حال تغییر است.
- پیشنهاد دهنده پرس‌وجوی قوی
- شرکت مایکروسافت
- ویژگی‌های جدید محلی مانند جستجوی رستوران‌ها، هتل‌ها و غیره
- رتبه‌بندی خوب

۴,۱شاخص‌های ارزیابی جویشگرهای متنی

به منظور ارائه سرویس جستجویی با کیفیت مطلوب به کاربران نهایی آن، فرآیند آزمون از اهمیت زیادی برخوردار است. دو دسته معیار برای ارزیابی هر سیستم نرم‌افزاری تحت عنوان نیازهای کارکردی و غیرکارکردی مطرح هستند. با توجه به اینکه موتور جستجو یک سامانه نرم‌افزاری بزرگ با تعداد زیاد کاربر است بنابراین علاوه بر نیازهای کارکردی، ملاحظات غیرکارکردی نیز از اهمیت بالایی برخوردار است. همچنین ارزیابی رفتار کاربر و آمار بازدید از موتور جستجو معیار مهمی در سنجش کسب موفقیت سیستمهای با تعداد کاربر زیاد محسوب می‌گردد. بنابراین تحلیل‌ها و اطلاعات بدست آمده در این بخش کمک بزرگی به سنجش میزان موفقیت موتور جستجو و تاثیرگذاری نیازهای کارکردی و غیرکارکردی خواهد نمود. در ادامه چارچوب ارزیابی موتور جستجو مبتنی بر دو منظر نیازهای کارکردی و غیرکارکردی مطرح شده است.

۱- معیارهای ارزیابی نیازهای کارکردی

- دقت:

وظیفه‌ی موتور جستجو پاسخ‌گویی به پرس‌وجوهایی می‌باشد که کاربران به آن ارسال می‌کنند. در پاسخ به هر پرس‌وجو، موتور جستجو لیستی از اسناد را به صورت رتبه‌بندی شده، در اختیار کاربر قرار می‌دهد. هر چقدر که اسناد موجود در لیست بازگردانده شده دارای کیفیت مناسب‌تری باشند گفته می شود که موتور جستجو از دقت بالاتری برخوردار می‌باشد. منظور از کیفیت مناسب برای یک سند این است که سند مذکور دارای محتوای مرتبط[۳] با پرس‌وجوی مطرح شده باشد. هر چقدر که به‌صورت متوسط، میزان ارتباط اسناد موجود در لیست نتایج بازگردانده شده توسط یک موتور جستجو، بیشتر از

[۳] Relevant

این متوسط ارتباط برای اسناد بازگشتی توسط یک موتور جستجوی دیگر باشد، آن‌گاه گفته می‌شود که موتور جستجوی اول دارای دقت بالاتری می‌باشد.

برای ارزیابی میزان مرتبط بودن نتایج یا دقت موتور جستجوی متنی معیارهای متنوعی وجود دارد که به نوع پرس و جوی ورودی به موتور جستجو وابسته است. پرس و جوهای اطلاعاتی و پیمایشی دو دسته بزرگ و عمده از پرس و جوهای کاربران را تشکیل میدهند بنابراین معیارهای هریک در ادامه به تفکیک ارائه خواهد شد:

- ارزیابی دقت موتور جستجو برای پرس و جوهای اطلاعاتی:

- هدف از این معیار سنجش میزان مرتبط بودن نتایج موتورهای جستجو در پاسخ به پرس و جوهای اطلاعاتی می‌باشد. معیارهای معمول برای کیفیت‌سنجی نظیر NDCG [۴]، MAP [۵] و غیره برای این معیار قابل محاسبه می‌باشند. در ادامه شرح مختصری بر هر یک از این معیارها ارائه شده است.

- معیار P@N

با توجه به اینکه کاربران تنها به بررسی nسند بالایی بسنده می‌کنند، در روش p@n نسبت تعداد اسناد مرتبط در nنتیجه اول ارائه شده نسبت به n محاسبه می‌شود. در این روش دقت روی n پاسخ نخست تعیین و تحلیل می‌شود.

$$P@n = \frac{\#\ ofrelevantdocumentsintopnresults}{n}$$

- معیار MAP

- استانداردترین معیار که در نشست TREC[۶] برای ارزیابی موتورهای جستجو از آن استفاده می‌شود معیار میانگین‌گیری از متوسط دقت می‌باشد که یک معیار تک مقداره در میان سطوح مختلفی از فراخوانی است. برای هر پرس‌وجو متوسط دقت عبارت است از میانگین مقادیر دقت در k سند بالایی هنگامی که اسناد مرتبط بازیابی می‌شوند. سپس برای محاسبه MAP از میانگین دقت‌های پرس‌وجوهای مختلف میانگین‌گیری می‌شود. اگر اسناد مرتبط برای مجموعه پرس‌وجوهای $q_i \epsilon Q$، $(d_1, \ldots dm_j)$باشد و R_k یک مجموعه اسناد مرتبط از بالاترین اسناد بازیابی شده توسط موتور جستجو باشد وقتی که مجموعه اسناد d_k باشد داریم:

[۴] normalDiscounted Cumulative Gain
[۵] Mean Average Precision
[۶] Text Retrieval Conference

$$MAP(Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{m}\sum_{j=1}^{m} Precision(R_{ij})$$

لازم به ذکر است که اگر هیچ سند مرتبطی بازیابی نشود میزان دقت در معادله بالا صفر می‌شود (Christopher D. Manning ۲۰۰۸).

- معیار NDCG

معیاری برای ارزیابی میزان کارایی الگوریتم‌های موتورهای جستجو یا الگوریتم‌های رتبه‌بندی است که اغلب در سیستم‌های بازیابی اطلاعات مورد استفاده قرار می‌گیرد. از آنجایی که کاربران علاقمند به رؤیت صفحات مرتبط در بالای لیست هستند، در این روش نتایجی که ارتباط بیشتری با پرس‌وجوی کاربر دارند در صورتی که رتبه بالاتری داشته باشند از اهمیت بیشتری در محاسبه برخوردارند. در محاسبه NDCG ابتدا باید مقدار CGرا محاسبه کنیم برای این منظور مجموع میزان ارتباط هر نتیجه ارائه شده در لیست نتایج با پرس‌وجوی کاربر را به دست می‌آوریم.

$$CG_p = \sum_{i=1}^{p} rel_i$$

که در اینجا rel_i میزان ارتباط نتیجه iام با پرس‌وجو می‌باشد.

عددی که به کمک این تابع به دست می‌آید رتبه‌بندی سیستم را مد نظر قرار نمی‌دهد و تنها معیاری برای مرتبط بودن هر سند است، بنابراین اگر نتیجه‌ای که ارتباط کمتری با پرس‌وجو دارد در رتبه بالاتری نسبت به نتیجه مرتبط‌ترقرار بگیرد در مقدارCG تفاوتی ایجاد نمی‌شود. با توجه به این مشکل از DCG برای لحاظ نمودن رتبه‌بندی استفاده می‌کنیم.

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

به این ترتیب اگر نتایجی که ارتباط بیشتری دارند در رتبه پایین‌تر قرار بگیرند سودمندی و کارایی سیستم کاهش می‌یابد. با توجه به اینکه اندازه نتایج جستجو برای پرس‌وجوهای مختلف متفاوت است بنابراین برای مقایسه کارایی موتورهایی جستجو باید DCG را نرمال کنیم.

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

در اینجا IDCGp همان DCG ایده‌آل در نقطه p است.

مزیت NDCG را می‌توان در موارد زیر خلاصه کرد:

- این معیار میزان ارتباط اسناد با رتبه نهایی آنها را ترکیب می‌کند.

- تفسیر آن ساده است.
- این معیار خیلی به نتایج پرت وابسته نیست. چون که مقادیر CG از ابتدای نتایج تا یک نقطه دلخواه محاسبه می‌شود.

برای روشن تر شدن مثال زیر را ارائه می‌کنیم:

فرض کنید رتبه بندی یک موتور جستجو به شکل مقابل است:

$D_۱,D_۲,D_۳,D_۴,D_۵,D_۶$

کاربر میزان ارتباط نتایج با پرس‌وجوی خود را با اعداد بین ۰ تا ۳ امتیاز دهی می‌کند:

۳،۲،۳،۰،۱،۲

به این ترتیب CG به صورت مقابل اندازه‌گیری می‌شود:

$$CGp = \sum_{i=1}^{\sum} reli = ۳ + ۲ + ۳ + ۰ + ۱ + ۲ = ۱۱$$

با تغییر جایگاه هر نتیجه تغییری در CGp بوجود نمی‌آید. برای حل این مشکل DCG_p را محاسبه می‌کنیم:

I	rel_i	$\log_۲ i$	$\frac{rel_i}{\log_۲ i}$
۱	۳	۰	
۲	۲	۱	۲
۳	۳	۱,۵۹	۱,۸۸۷
۴	۰	۲	۰
۵	۱	۲,۳۲	۰,۴۳۱
۶	۲	۲,۵۹	۰,۷۷۲

$$DCG_۶ = rel_۱ + \sum_{i=۲}^{\sum} \frac{rel_i}{\log_۲ i} = ۳ + (۲ + ۱.۸۸۷ + ۰ + ۰.۴۳۱ + ۰.۷۷۲) = ۸.۰۹$$

به این دلیل که اندازه نتایج برای هر پرس‌وجو متفاوت است DCG را نرمال می‌کنیم. برای نرمال سازی DCGباید نتایج را بصورت ایده‌آل رتبه‌بندی کنیم؛ یعنی نتایج را به نحوی مرتب کنیم که نتایج مرتبط‌تر در رتبه‌های بالاتر قرار گیرند. در این مثال رتبه بندی ایده‌آل بصورت زیر است:

$D_۱,D_۳,D_۲,D_۶,D_۵,D_۴$

IDCGایده‌آل برابر است با:

$$IDCG_۶ = ۸.۶۹۳$$

و نهایتاً nDCGبرای این پرس‌وجو:

$$nDCG_۶ = \frac{DCG_۶}{IDCG_۶} = \frac{۸.۰۹}{۸.۶۹۳} = ۰.۹۳۰۶$$

در نهایت با یک میانگین‌گیری از nDCG تعدادی پرس‌وجو می‌توان کارایی و صحت نتایج موتورهای جستجوی مختلف را به دست آورد.

- ارزیابی دقت موتور جستجو برای پرس و جوهای پیمایشی:

پرس و جوهای پیمایشی یکی از انواع رایج پرس‌وجوهای کاربران موتور جستجو، هستند که از نظر کاربر تنها یک پاسخ مرتبط دارند و کاربر نیز از پاسخ مورد انتظار آگاهی دارد. معیار مورد استفاده برای سنجش عموما MRR [۷] می‌باشد.

- معیار MRR

رتبه دوجانبه برای نتایج پرس‌وجو، معکوس موقعیت اولین سند مرتبط می‌باشد. با میانگین‌گیری رتبه متقابل برای چند پرس‌وجو MRR به دست می‌آید.

$$MRR = \frac{۱}{|Q|}\sum_{\sum}^{|\sum|}\frac{۱}{rank\sum}$$

برای مثال فرض کنید ۳ پرس‌وجو با نتایج متناظر با آنها را داریم:

جدول ۱: نتایج متناظر سه پرس وجو

پرس‌وجو	لیست نتایج	رتبه‌ای اولین سند مرتبط	رتبه متقابل

[۷] Mean Reciprocal Rank

الف	۳,۲,۱	۳	۱/۳
ب	۳,۲,۱	۲	۱/۲
ج	۳,۲,۱	۱	۱

با استفاده از این سه نمونه ما می‌توانیم مقدار MRR را به صورت زیر به دست بی‌آوریم:

MRR = (۱+۱/۲+۱/۳)/۳ = ۱۱/۱۸

اگر غیر از اولین رتبه مرتبط سایر نتایج هم حائز اهمیت باشند از روش MRR استفاده نمی‌شود.

- پوشش: پوشش به معنای نسبت تعداد صفحات نمایه‌سازی شده به تعداد کل صفحات موجود در وب می‌باشد. با توجه به اینکه حجم و کیفیت صفحات نمایه‌شده تاثیر زیادی بر بهبود عملکرد موتور جستجو و کیفیت نتایج بازگشتی دارد این معیار به عنوان یکی از شاخص‌های اصلی برای ارزیابی موتورهای جستجو مطرح است؛ بنابراین لازم است به‌صورت دقیق و قابل اطمینانی اندازه‌گیری شود.
- تازگی: یکی از عواملی که تاثیر زیادی بر رضایت کاربر دارد، میزان بروز بودن نتایج ارائه شده از سوی موتور جستجو است. بنابراین سنجش این معیار از اهمیت زیادی برخوردار است. علاوه بر اینکه میزان پوشش صفحات موجود در وب تاثیر زیادی بر بروز بودن نمایه دارد، فرکانس بروزرسانی آنها نیز از اهمیت بالایی برخوردار است. به عنوان نمونه برخی سایتها مانند سایتهای خبری با فرکانس بالاتری بروز می‌شوند، بنابراین خزشگر وب باید سیاست مناسبی برای شناسایی و جمع‌آوری آنها داشته باشد. این مسئله کمک می‌کند تا تازگی[۸] اسناد حفظ شود. همچنین امکان بروزرسانی نمایه در کمترین زمان ممکن بر اساس صفحات جدید و بروز وجود داشته باشد.

۲- معیارهای ارزیابی نیازهای غیرکارکردی

- دسترس‌پذیری[۹]
  - میزان پاسخگویی سیستم در قبال پرس‌وجوهای ارسال‌شده
  - نداشتن شکست در سیستم

[۸] Freshness
[۹] Availability

- کارایی[۱۰]
  - زمان پاسخگویی به پرس‌وجوها
  - توانایی در پاسخگویی به کاربر همزمان
- آزمون گرافیک واسط کاربری[۱۱]
- قابلیت همکاری[۱۲]
  - داشتن وب‌سرویس مناسب
  - داشتن افزونه برای مرورگر IE
  - داشتن افزونه برای مرورگر firefox
- سازگاری[۱۳]
  - سازگاری با مرورگر IE
  - سازگاری با مرورگر firefox

۳-تحلیل رفتار کاربران و آمار بازدید:

با توجه به اینکه نتایج کارایی سیستم در بخشهای کارکردی و غیرکارکردی در میزان اقبال کاربران و رفتار آنها با موتور جستجو قابل ردیابی و پیگیری است. بنابراین تحلیل‌های ارائه شده در این بخش از اهمیت بالایی برخوردار است. به عبارتی بررسی وضعیت، میزان پیشرفت و مقبولیت یک جستجوگر نیازمند بررسی و تحلیل آمار بازدیدکنندگان و تعداد جستجوهای صورت گرفته توسط کاربران در جستجوگر است. بنابراین با پردازش و تحلیل داده‌هایی ذخیره شده در لاگ موتور جستجو که بیانگر نحوه تعامل کاربران با موتور جستجو است، می‌توان وضعیت موتور جستجو را از جهات مختلف مانند: تحلیل پرس‌وجو، تحلیل نشست، تحلیل کلیک کاربران، تحلیل آی پی، تحلیل کاربران ماندگار تحلیل و بررسی نمود. اطلاعاتی که می‌توان کسب نمود شامل تعداد کاربران روزانه، تعداد کاربران ماندگار، تعداد نشست‌ها و میانگین زمان نشست‌ها، لیست وب سایت‌های پر بازدید، لیست پرس‌وجوهای پرکاربرد، میانگین کلیک کاربران برای هر جستجو و نشست می‌باشد.

خلاصه شاخصهای ارزیابی و معیارهای سنجش آنها در جدول زیر ارائه گردیده است:

---

[۱۰] Performance

[۱۱] Gui Testing

[۱۲] Interoperability

[۱۳] Compatibility

جدلول ۲: خلاصه شاخصهای ارزیابی و معیارهای سنجش آنها

سامانه		**جستجوی متنی**
دقت/کیفیت		nDCG، MAP، MRR، P@N و Precision
پوشش		تعداد اسناد نمایه شده
تازگی		عددی مابین ۰ و ۱ که مقدار بیشتر نشان دهنده تازگی بیشتر است.
شاخص‌های مرتبط با آمار بازدید	**تحلیل در سطح پرس و جو**	تحلیل طول پرس و جوها متوسط تعداد پرس وجوها در بازه های زمانی مختلف (روزانه، ماهانه و غیره) تعداد پرس وجوها بر حسب نوع سرویس جستجو تعداد پرس وجوهای بدون پاسخ
	تحلیل آمار بازدید	تعداد بازدیدکنندگان به تفکیک سرویسهای مختلف
	تحلیل در سطح نشست	متوسط تعداد پرس وجوها در یک نشست متوسط مدت زمان یک نشست
	تحلیل در سطح کاربران	میانگین کلیک به ازای هر پرس و جو تعداد کاربران کل/ ماندگار/ جدید
	تحلیل در سطح IP	تعداد IP های یکتا
نیازمندیهای غیرکارکردی		سرعت پاسخ گویی دسترس پذیری زمان پاسخگویی به کاربران همزمان امنیت کاربر پسندی سازگاری با مرورگرهای مختلف امکانات راهنمایی و یا گزارش مشکلات یا پیشنهادات

۵,۱جمع‌بندی

مهمترین بخش موتور جستجو سرویس جستجوی متنی می‌باشد. در سرویس جستجوی متنی وب که در آن عمدتا جستجو بر روی محتوای متون موجود در صفحات وب انجام می‌شود، معمولا انتظار بیشتری از جویشگر وجود دارد و آن پردازش هوشمند پرس و جو و نیاز اطلاعاتی کاربر و استفاده از بهترین منابع موجود برای پاسخگویی به نیاز او می‌باشد. بنابریان از دو منظر کلی باید موتورهای جستجو تقویت شوند از یک منظر اتکا بر منابع و پردازشهای هوشمند برای درک پرس و جو با توجه به نوع و نیاز اطلاعاتی و از منظر دیگر بازیابی اطلاعات با استفاده از الگوریتمهای مناسب و رتبه‌بندی کارا که در آن هم شخصی‌سازی تا حد امکان در نظر گرفته شده باشد و هم تنوع و تعدد خدمات مورد استفاده پس‌زمینه لحاظ شده باشد.

به علاوه موتورهای جستجوی بزرگ در کنار پردازش قوی و هوشمند اطلاعات باید از جنبه طراحی و معماری نیز توانایی پاسخگویی به تعداد قابل توجهی از کاربران را داشته باشند. به عبارتی ویژگیهای توزیع-شدگی و مقیاس‌پذیری نیز باید در موتور جستجو در نظر گرفته شود.

موتورهای جستجوی موفق بومی و جهانی عمدتا دارای ویژگیهای زیر بوده‌اند که می‌توان در توسعه موتور جستجوی متنی مورد توجه قرار داد:

- پردازش قوی زبان
- پوشش مناسب و با کیفیت صفحات
- رتبه‌بندی مناسب
- پیشنهاد دهنده پرس‌وجوی قوی
- واسط کاربر ساده
- پردازشگر پرس‌وجوی چند زبانه
- بازیابی اطلاعات مکان محور
- توجه به ویژگی‌های جدید محلی مانند جستجوی رستوران‌ها، هتل‌ها و غیره
- زیرساختهای مطمئن
- معماری مقیاس‌پذیر

Information Technology Institute

and Virtual Environments Group IT Platforms

Technical Report

Project Name:

Project Director	
Author(s)	
Document Code	
Preparing Date	۹۴,۰۴,۱۴
Status/Version	Final/۱,۰