
بسم الله الرحمن الرحيم

فصل اول

مقدمه

۱-۱-پیشینه

نخستین فردی که در سال ۱۹۹۵ این ایده را در اختیار داشت، دانیل داریلینگر از ایالت کلرادو بود .او فراجویشگر SearchSavvy معرفی کرد که به کاربران اجازه می‌داد تا ۲۰ جویشگر مختلف و دایرکتوری را بلافاصله جستجو کنند .اگرچه سریعا این فراجویشگر به جستجوهای ساده محدود شد و بنابراین بسیار قابل اعتماد نبود.(Meng et al., ۲۰۰۲)

دانشجوی دانشگاه واشنگتن اریک سلبرگ، یک نسخه به‌روزتر را که MetaCrawler نامیده می‌شود، منتشر کرد(Meng et al., ۲۰۰۲).

این فراجوشگر بر روی دقت SearchSavvy's با اضافه کردن نحو جستجوی خود در پشت صحنه ها و تطبیق دادن نحو با استفاده از موتور جستجو که در حال کاوش بود، بهبود یافت.(Meng et al., ۲۰۰۲)

Metacrawler تعداد جویشگرها را به ۶ تا کاهش داد، اما اگر چه نتایج دقیق‌تری تولید کردند، هنوز هم به اندازه جستجو در یک موتور منفرد دقیق در نظر گرفته نشده است.(Meng et al., ۲۰۰۲)

HotBot فراجویشگری دیگر بود که در ۲۰ مه ۱۹۹۶ ساخته شد .که در آن زمان متعلق به Wired بود، که از نتایج جستجو پایگاه‌داده‌های Inktomi و Direct Hit استفاده می‌کرد .پس از اینکه توسط Lycos در سال ۱۹۹۸ خریداری شد، توسعه برای موتور جستجو متوقف شد و سهم بازار آن به شدت کاهش یافت .پس از گذراندن چند تغییر، HotBot به یک رابط جستجوی ساده طراحی شده و ویژگی‌های آن در طراحی مجدد وب سایت Lycos گنجانده شده است(Meng et al., ۲۰۰۲) .

یک فراجویشگر متنی دیگر با نام Anvish توسط B. Shu و Subhash Kak در سال ۱۹۹۹ ساخته شدکه در آن نتایج جستجو با استفاده از شبکه‌های عصبی آموزش دیده طبقه‌بندی شدند؛ این کار بعدا در یک موتور فراجویشگر دیگر به نام Solosearch ترکیب شد(BoShu et al. , ۱۹۹۹).

Ixquick یک فراجویشگر است که اخیرا به خاطر سیاست حریم خصوصی‌اش شناخته شده است که این فراجویشگر در سال ۱۹۹۸ توسط David Bodnick توسعه داده شده است. (Hassanpour et al. , ۲۰۱۲)

در سال۲۰۰۳(۱۳۸۱) فراجویشگر فارسی پارسیک توسط آقای علیرضا شیرازی کار خود را آغاز کرد.که از جویشگرهایی همچون گوگل و یاهو برای جستجو و دریافت اطلاعات استفاده می‌کند. (Tazehkandi et al. , ۲۰۲۰)

در سال ۲۰۰۵ فراجویشگر Dogpile معرفی شد که متعلق به شرکت infospace بود و هدف آن این بود که هم‌پوشانی و رتبه‌بندی جویشگرهای وب را اندازه گیری کند. (Jansen et al., ۲۰۰۷)

در سال ۲۰۰۷ فراجویشگر جدیدی به نام WebFusion معرفی شده است ؛WebFusion مهارت جویشگرهای اصلی را در یک گروه خاص براساس ترجیحات کاربران یاد میگیرد. این فراجویشگر همچنین از click-through (مفهوم کلیک کاربران) استفاده می‌کند که مفهوم click-through یک بازخورد ضمنی از ترجیحات کاربران است که به عنوان یک سیگنال تقویت‌کننده در فرآیند یادگیری استفاده می‌شود و زمان جستجو را در لیست نتایج بازگشتی کاهش می‌دهد. (Dr keyhanipour et al., ۲۰۰۷)

در سال ۲۰۱۰ یک فراجویشگر شخصی چندعامله با استفاده از شبکه‌های مفهومی فازی خودکار معرفی شد که هدف آن استفاده از شبکه‌های مفهومی فازی اتوماتیک برای شخصی‌سازی نتایج یک فراجویشگر بود که با یک معماری چندعامله برای جستجو و بازیابی سریع تهیه شده است. (Batool Arzanian et al., ۲۰۱۰)

در سال ۲۰۱۲ فراجویشگر SEReleC برای پالایش و طبقه‌بندی نتایج جویشگرها فراهم می‌کند تا نتایج جستجو را بصورت پیوسته پیوند زده و منجر به کاهش چشمگیر تعداد صفحات شود (Raval et al., ۲۰۱۲)

در سال ۲۰۱۳ یک شیوه رتبه‌بندی تجمیعی برای فراجویشگرها معرفی شد که در این روش ویژگی‌های شایسته روش بوردا و مقیاس‌پذیری روش فوترول را باهم ترکیب می‌کند که آزمایشات نشان داده که این روش بهبود یافته توانسته جایگزین خوبی برای بسیاری از روش‌های قبلی باشد. (Junliang Feng et al., ۲۰۱۴)

درسال ۲۰۱۵ فراجویشگر MetaSurfer معرفی شد که براساس FAHP (فرآیند سلسله مراتبی تحلیلی فازی) و عملگر اصلاح شدهEOWA (میانگین وزنی سفارش یافته توسعه یافته) بود FAHP .برای مقایسه زوجی اسناد در جویشگرهای پایه استفاده می‌شود. بنابراین اپراتور اصلاح شده EOWA برای تجمیع نمرات اسناد بدست آمده از جویشگر به منظور بدست آوردن رتبه نهایی اسناد استفاده می‌شود.

در سال ۲۰۱۶ فراجویشگر MetaFusion معرفی شد که یک فراجویشگر کارآمد مبتنی بر الگوریتم ژنتیک بود این فراجویشگر با ترکیب AHP فازی و الگوریتم ژنتیک سعی در این داشت که نتایج جامع‌تر و بهینه‌تری کسب کند. دقت MetaFusion در مقایسه با Dogpile و Infospace بیشتر است. (Devendra K. Tayal et al., ۲۰۱۴)

در سال ۲۰۱۸ IM search معرفی شد که یک فراجویشگر شخصی مبتنی بر عامل بود . این موتور فراجویشگر براساس معماری چندعاملی برای بهبود سطح شخصی سازی پیاده سازی شد. (Meijia Wang et al., ۲۰۱۸)

در سال ۲۰۱۹ یک الگوریتم رتبه‌بندی چندشرطی براساس روش VIKOR برای فراجویشگرها معرفی شد. با استفاده از روش VIKOR و معیارهای استخراج شده ، صفحات وب رتبه بندی می شوند و ۱۰ نتیجه برتر برای کاربر ارائه می‌شود.(Mastoreh Haji۱ et al., ۲۰۱۹)

۱-۲-تعریف مساله

اینترنت منبع عظیمی از اطلاعات مختلف است و هر روز به حجم این اطلاعات در وب‌سایت‌های گوناگون اضافه می‌شود. مسلما کاربران برای یافتن اطلاعات موردنیاز خود در اینترنت به ابزار مناسب جستجو و یا همان جویشگرها نیاز دارند.

جویشگرهایی مانند گوگل این روزها ابزار جستجو را در زبان‌های مختلف از جمله فارسی در اختیار کاربران قرار می‌دهند اما در این میان جویشگرهای فارسی هم در اینترنت راه اندازی شده‌اند که با اینکه در ابتدای‌کار هستند و هنوز فاصله‌زیادی با رقبای قدرتمند خود دارند اما کم‌کم در میان کاربران فارسی زبان می‌توانند طرفدارانی داشته باشند.

جویشگرهای متنی فارسی‌زبان موجود به دو دسته جویشگرها و فراجویشگرها قابل تقسیم‌بندی هستند . (Meng et al., ۲۰۰۲)

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، جویشگر خزشی انجام نداده و ترکیبی از بازرتبه‌بندی نتایج منتخب سایر جویشگرهای موجود را به عنوان نتایج رتبه‌بندی خود به کاربر ارائه می‌دهد(Meng et al., ۲۰۰۲) .

مهمترین جویشگرهای متنی توسعه داده شده در داخل کشور جویشگرهای پارسی‌جو، یوز، زال ،ریسمون و جس‌جو می‌باشند که در حال حاضر در قالب شرکت‌های دانش‌بنیان به فعالیت خود ادامه می‌دهند.

در زمینه فراجویشگرها نیز موتورهای‌جستجویی نظیر سلام و پارسیک مطرح هستند؛ این فراجویشگرها نتایج جویشگرهای موجود نظیر گوگل و بینگ را ترکیب نموده و در اختیار کاربران قرار می‌دهند.

در ادامه تعدادی از معروف‌ترین جویشگرهای‌بومی را معرفی و آنها را طبق معیارهای ارزیابی سایت الکسا که در ادامه شرح می‌دهیم تجزیه‌وتحلیل می‌کنیم. این آمار مربوط به تاریخ ۱۳۹۸/۱۰/۱۲ است که از سایت الکسا بازیابی شده است.

حال میخواهیم طبق معیارهای ارزیابی سایت الکسا جویشگرهای‌فارسی را ارزیابی کنیم. سایت الکسا به طور مداوم انواع اطلاعات را از وب‌سایت‌ها جمع می‌کند و خدمات ارزیابی رایگان ارایه می‌دهد.

الکسا رتبه‌بندی ترافیک را با تجزیه‌وتحلیل میزان استفاده میلیون‌ها کاربر و داده‌های بدست‌آمده از سایر منابع داده‌های مختلف ترافیک را محاسبه می‌کند. رتبه ترافیک معیاری از محبوبیت وب سایت است. رتبه با استفاده از ترکیبی از میانگین بازدیدکنندگان روزانه و بازدید از صفحه در طی سه‌ماه گذشته محاسبه

می‌شود. سایتی که بیشترین میزان ترکیبی از بازدیدکنندگان و بازدید از صفحه را دارد ، در رده اول قرار دارد. (Erfanmanesh et al., ۲۰۱۰)

میانگین زمان ماندن در سایت عامل دیگری است که میزان توجه کاربران و میانگین دقیقه هایی را که کاربر در طول روز در سایت در طی سه ماه گذشته صرف کرده است را شامل می‌شود. (Erfanmanesh et al., ۲۰۱۰)

تعداد صفحات بازدید شده برای هر کاربر نیز توسط الکسا محاسبه می‌شود که کیفیت و تنوع اطلاعات در سایت را نشان می‌دهد. (Erfanmanesh et al., ۲۰۱۰)

معیار دیگر تعداد پیوندهایی است که یک وب سایت از وب سایت‌های دیگر دریافت می‌کند و محبوبیت آن را نشان می‌دهد. (Erfanmanesh et al., ۲۰۱۰)

علاوه بر این عوامل ، سایت الکسا در مورد درصد افرادی که از یک وب‌سایت بازدید می‌کنند (بازدیدکنندگان ملی و بین المللی) اطلاعاتی را ارائه می‌دهد. (Erfanmanesh et al., ۲۰۱۰)

از آنجا که از تعاریف شاخص‌های الکسا استنباط می‌شود ، این شاخص ها بیشتر مربوط به ترجیح کاربران است و این بدان معناست که اگر یک وب‌سایت بر اساس این فهرست‌ها عملکرد مناسبی داشته باشد ، محبوب‌تر و موردعلاقه کاربران وب است. (Erfanmanesh et al., ۲۰۱۰)

حال ما می‌خواهیم چندتا از معروفترین جویشگرها و فراجویشگرهای فارسی را طبق معیارهای الکسا بررسی و مقایسه کنیم که این معیارها رتبه ترافیک، میانگین تعداد صفحات بازدید شده توسط کاربر، زمانی که توسط هر کاربر در آن جویشگر گذرانده می‌شود، تعداد لینک‌هایی که از وب‌سایت‌های دیگر دریافت می‌کند و درصد بازدیدکنندگان ایرانی و بین المللی هستند.

ترافیک جستجو: درصد مراجعه به این جویشگر برای جستجوی اساسی(Erfanmanesh et al., ۲۰۱۰)

ترافیک جستجوی سایت رابطه مستقیمی با افزایش بازدید و جستجوی آن سایت در شبکه‌جهانی وب دارد که هرچه مراجعه کاربران دنیای وب به این سایت بیشتر باشد و کاربران بیشتری این سایت را مورد جستجو و بازدید قرار داده باشند درصد ترافیک جستجوی آن سایت بالاتر می‌رود. (Erfanmanesh et al., ۲۰۱۰)

تعداد پیوندهای سایت‌هایی که به این سایت رجوع کرده‌اند: سایت‌هایی که به این سایت مراجعه داشته‌اند که درصد آنها بصورت هفتگی در سایت الکسا محاسبه می‌شود و ما در این قسمت طبق معیارهای سایت الکسا درصد سایت‌هایی که به سایت مورد نظر مراجعه داشته‌اند را بررسی و بیان می‌کنیم. (Erfanmanesh et al., ۲۰۱۰)

رتبه الکسا سایت در ۹۰ روز اخیر: که این عدد در ترافیک‌جهانی و پیوندهای‌کلی اینترنت محاسبه می‌شود. (Erfanmanesh et al.,۲۰۱۰)

زمان سپری شده روزانه (دقیقه/ثانیه): متوسط زمان به واحد دقیقه و ثانیه که هر کاربر مراجعه‌کننده در طول روز در این سایت می‌گذاراند(Erfanmanesh et al., ۲۰۱۰)

کاربران ایرانی: درصد کاربران ایرانی مراجعه کننده به این سایت(Erfanmanesh et al., ۲۰۱۰)

کاربران غیر ایرانی: درصد کاربران غیرایرانی مراجعه کننده به این سایت (Erfanmanesh et al., ۲۰۱۰)

ردیف	موتورجستجو	آدرس
۱	یوز	www.yooz.ir
۳	زال	www.zal.ir
۴	ریسمون	www.rismoon.com
۵	پارسی‌جو	www.parsijoo.ir
۶	جس‌جو	www.jasjoo.com

جدول شماره۱: معرفی چند موتورجستجوی بومی

ردیف	موتور فراجویشگر فارسی	آدرس
۱	سلام	www.salam.ir
۲	پارسیک	www.parseek.com

جدول شماره۲: معرفی چند موتور فراجویشگر فارسی

موتورجستجو	ترافیک جستجو	بازدید	ارجاع‌هایی که به این سایت از سایت های دیگر انجام شده است	رتبه‌الکسا	زمان سپری شده روزانه (دقیقه/ثانیه)	کاربران ایرانی	کاربران غیر ایرانی
یوز	٪ ۲۰,۹	٪ ۴۲,۸	۱,۳۰۵	۴۸,۷۸۴	۱:۳۶	٪ ۸۵,۳	٪۹,۱(آآم ریکا)
زال	٪ ۵۰	٪۸۲,۴	۱۴,۶۲۲	۱,۳۵۸,۷۷۹	۰۰:۲۸	٪ ۱۰۰,۰	-
ریسمون	٪ ۲۰,۸	٪۲۹,۸	۴۵۸	۲۵۵,۷۴۵	۱:۳۳	٪ ۵۵	-
پارسی‌جو	٪ ۱۱۰,۸	٪ ۴۶,۹	۱,۳۰۱	۱۱,۲۹۷	۳:۵۶	٪ ۹۳,۳	٪ ۱,۳هند)
جس‌جو	٪ ۸۷,۶	٪ ۶۵,۷	۱,۱۵۳	۸۵,۹۴۲	۱:۴۰	٪ ۹۲,۰	-
گوگلر	٪ ۲۳,۱	٪ ۵۶,۳	۷۲	۹۸۶,۱۶۶	۰:۶۰	٪ ۱۰۰,۰	-

جدول شماره۳: مقایسه موتورهای جستجوگر فارسی مطابق با اطلاعات سایت الکسا (www.alexa.com)

موتور فراجویش گر	ترافیک جستجو	بازدید	سایت هایی که به این سایت رجوع کرده اند	رتبه الکسا	زمان سپری شده روزانه (دقیقه/ثانیه)	کاربران ایرانی	کاربران غیر ایرانی
سلام	٪ ۸,۱	٪ ۶۸,۱	۱۵,۱۴۷	۱۲۲,۴۱۱	۱:۱۱	٪ ۸۹,۲	-
پارسیک	٪ ۱۶,۹	٪ ۵۰,۸	۱,۸۴۶	۲۷,۹۳۷	۳:۱۰	٪ ۶۰,۴	٪ ۲۷,۸ (آمریکا)
							٪ ۳,۰ (ترکیه)

جدول شماره۴: مقایسه موتورهای فراجویشگر فارسی مطابق با اطلاعات سایت الکسا(www.alexa.com)

حال طبق معیارهای ارزیابی سایت الکسا جویشگرهای فارسی مورد بحث در تحقیق‌مان را در جدول شماره ۵ ترتیب‌دهی می‌کنیم:

ترافیک جستجو	بازدید	سایت هایی که به این سایت رجوع کرده اند	رتبه الکسا	زمان سپری شده روزانه (دقیقه/ثانیه)	کاربران ایرانی	کاربران غیر ایرانی
جس‌جو	زال	زال	زال	پارسی‌جو	زال	یوز
زال	جسجو	یوز	گوگلر	جس‌جو	گوگلر	پارسی‌جو
گوگلر	گوگلر	پارسی‌جو	ریسمون	یوز	پارسی‌جو	
یوز	پارسی‌جو	جس‌جو	جس‌جو	ریسمون	جس‌جو	
ریسمون	یوز	ریسمون	یوز	گوگلر	یوز	
پارسی‌جو	ریسمون	گوگلر	پارسی‌جو	زال	ریسمون	

جدول شماره ۵: ترتیب‌دهی موتورهای جستجوگرفارسی طبق اطلاعات سایت الکسا(www.alexa.com)

با توجه به ترتیب‌دهی متوجه می‌شویم که در اکثر موارد زال رتبه اول را دارد این درحالی است که موتورجستجوی زال بطور مستقل کاربردی ندارد بلکه به این دلیل رتبه اول را کسب کرده است که در فراجویشگر سلام که یکی از بهترین فراجویشگرهای زبان فارسی است استفاده می‌شود به همین دلیل است که سایت های زیادی به آن لینک پیوند می‌دهند و بازدید زیادتری نیز نسبت به دیگر جویشگرهای فارسی دارد.

حال طبق معیارهای ارزیابی سایت الکسا موتورهای فراجویشگر فارسی مورد بحث در تحقیق‌مان را در جدول شماره ۶ ترتیب‌دهی می‌کنیم:

ترافیک جستجو	بازدید	سایت هایی که به این سایت رجوع کرده اند	رتبه الکسا	زمان سپری شده روزانه (دقیقه/ثانیه)	کاربران ایرانی	کاربران غیر ایرانی
پارسیک	سلام	سلام	سلام	پارسیک	سلام	پارسیک
سلام	پارسیک	پارسیک	پارسیک	سلام	پارسیک	

جدول شماره۶: ترتیب‌دهی موتورهای فراجویشگر فارسی مطابق با اطلاعات سایت الکسا (www.alexa.com)

حال در این قسمت درباره هرکدام از جویشگرها و فراجویشگرهای فارسی معرفی شده در تحقیق توضیح بیشتری می‌دهیم:

جویشگر متنی پارسی‌جو

هدف موتورجستجوی پارسی‌جو پوشش ۵۰۰ میلیون صفحه فارسی و پاسخگویی به دو میلیون پرس‌وجو در روز است که از سرویس‌های ارزش‌افزوده برای جذب بیشتر کاربران استفاده می‌کند و از سال ۸۹ توسعه داده شده است. پارسی‌جو مانند هر جویشگر متنی دیگر از سه مولفه اصلی خزشگر، نمایه‌ساز و جویشگر یا بازیابی اطلاعات تشکیل شده است.(علیرضا یاری ،۱۳۹۴)

با استفاده از بخش خزشگر می‌تواند بیش از ۵۰۰ میلیون صفحه و بیش از یک‌میلیارد آدرس صفحه را با تازگی‌هوشمند نگه دارد. با استفاده از بخش پارسر و نمایه‌سازی می‌توانیم کارهای مدیریت و پردازش در مقیاس بالا را در محیطی کاملا توزیع‌شده انجام دهیم. (علیرضا یاری ،۱۳۹۴)

بخش جستجوی برخط مسئولیت پاسخگویی به میلیون‌ها پرس‌وجو در روز را دارد. پارسی‌جو برای پاسخ‌دهی بهتر چند کپی از نمایه ایجاد می‌کند و با توزیع‌سازی نمایه‌ها ، بهینه‌سازی و گذاشتن نمایه‌ها در حافظه تلاش می‌کند به سرعت خوبی برای پاسخ‌دهی به پرس‌وجوهای کاربران برسد. (علیرضا یاری ،۱۳۹۴)

علاوه بر موارد ذکر شده در پارسی‌جو یک‌سری زیرسیستم‌ها برای پاسخ‌دهی بهتر در نظر گرفته شده‌اند:

زیرسیستم پیشنهاددهنده پرس‌وجو با بررسی تمام پرس‌وجو ها با توجه به زمان و فرکانس آنها پرس‌وجو هایی را پیشنهاد می‌دهد. .(علیرضا یاری ،۱۳۹۴)

زیر سیستم خطایابی فارسی با توجه به داده‌های موجود در وب پرس‌وجوی کاربر را برای نیل به مقصود اصلاح می‌کند. (علیرضا یاری ،۱۳۹۴)

زیرسیستم کشینگ که هدف آن نگه‌داری تمام صفحات موتورجستجو می‌باشد. (علیرضا یاری ،۱۳۹۴)

پارسی‌جو علاوه بر سرویس جویشگر متنی دارای تنوعی از خدمات ارزش‌افزوده و سایر خدمات جستجو است .به عنوان نمونه پارسی‌جو شامل خدماتی نظیر جستجوی‌تصویری، ویدئو، نقشه، علمی می‌باشد .(علیرضا یاری ،۱۳۹۴)

جویشگر متنی پارسی‌جو با پردازش هوشمند پرس‌وجو و شناسایی نیاز کاربر در صورت نیاز از سایر خدمات نیز نتیجه‌های مرتبط را به کاربر ارائه می‌دهند .(علیرضا یاری ،۱۳۹۴)

از مشکلات موتورجستجوی پارسی‌جو می‌توان گفت که با دسترسی محدود به اطلاعات رو به رو است و نتایج آن در اکثر موارد ویکی‌پدیا فارسی را نشان می‌دهد همچنین حتی زمانی که کاربر نام یک لیست یا برند یک شرکت را جستجو کند نمی‌تواند مستقیما به سایت موردنظر رجوع کند.(علیرضا یاری،۱۳۹۴)

جویشگر متنی یوز

موتورجستجوی یوز از اواخر سال ۱۳۸۸ با تلاش نیروهای متخصص داخلی آغاز شده است. یوز ادعا می‌کند که تاکنون توانسته بیش از یک میلیارد صفحه را پوشش دهد. یوز همچنین دارای خدمات جستجوی خبر، وبلاگ و عکس می‌باشد.

موتورجستجوی یوز دارای ویژگی‌های زیر می‌باشد: (علیرضا یاری ،۱۳۹۴)

تمرکز بر زبان فارسی (علیرضا یاری ،۱۳۹۴)

سرعت بالا با هدف گذاری پاسخ‌دهی سریع به کاربران با میانگین تأخیر کمتر از ۱ ثانیه(علیرضا یاری ،۱۳۹۴)

تحلیل نیازهای متداول کاربران و پاسخ‌دهی مستقیم به چندین نوع از جستجوهای کاربران (علیرضا یاری ،۱۳۹۴)

نمایه‌سازی بی‌درنگ : قابل جستجو نمودن صفحات جدید چند دقیقه پس از خزش (علیرضا یاری ،۱۳۹۴)

معماری مقیاس‌پذیر به نحوی که برای افزایش پوشش صفحات وب، فقط کافیست ماشین‌های جدید به خوشه‌ها اضافه شود. (علیرضا یاری ،۱۳۹۴)

از مشکلات اساسی موتورجستجوی یوز می‌توان گفت که موتورجستجویی با اطلاعات کم است و توان نشان دادن اطلاعات جامع را ندارد و نتایجی که به کاربر نشان می‌دهد با پرس‌وجوی انجام شده تفاوت دارد و بعضا اطلاعات نامرتبط را در نتایج نمایش می‌دهد(علیرضا یاری ،۱۳۹۴)

جویشگر متنی ریسمون

ریسمون یکی از قدیمی‌ترین جویشگرهای وب فارسی است. این جویشگر همه وب‌سایتهای فهرست را در دوره‌های زمانی یک‌ماهه می‌پیماید و محتویات و مطالب آنها را نمایه‌سازی می‌کند و در Link.irبانک‌اطلاعاتی خود را جهت ارائه خدمات جستجو به مراجعه‌کنندگان نگهداری می‌نماید .پروژه جستجوگر ریسمون از مهرماه ۱۳۸۳ در شرکت رادکام آغاز گردیده است(علیرضا یاری ،۱۳۹۴)

جویشگر متنی زال

زال یک جویشگر فارسی است که توسط شرکت بیان با هدف ارتقا کیفیت خدمات‌جستجو برای کاربران فارسی زبان در حال توسعه است .در حال حاضر استفاده از زال تنها از طریق فراجویشگر سلام امکان پذیر است. (علیرضل یاری ،۱۳۹۴)

جویشگر زال دارای قابلیت‌های زیر می‌باشد: (علیرضا یاری ،۱۳۹۴)

تشخیص هرز نوشته‌ها و تله های (علیرضا یاری ،۱۳۹۴)

تشخیص محتوای غیراخلاقی (علیرضا یاری ،۱۳۹۴)

تکمیل خودکار عبارات (علیرضا یاری ،۱۳۹۴)

ریشه‌یابی لغات (تعلیرضا یاری ،۱۳۹۴)

پیشنهادات مشابه (علیرضا یاری ،۱۳۹۴)

ابهام زدایی (علیرضا یاری ،۱۳۹۴)

دسته‌بند مفهومی نتایج(علیرضا یاری ،۱۳۹۴)

تاریخچه مصور صفحات(علیرضا یاری ،۱۳۹۴)

جویشگر زال در فراجویشگر سلام به عنوان یکی از جویشگرهایی است که از ترکیب نتایج آنها برای پاسخگویی به نیاز کاربر استفاده می‌شود و تا کنون خروجی تحت‌وب مستقلی از آن ارائه نگردیده است. فقط معرفی اجمالی از آن جویشگر ارائه شده است و از طریق www.zal.ir همچنین از طریق آدرس این پیوند خدمات جستجو ارائه نمی‌کند. (علیرضا یاری ،۱۳۹۴)

جویشگر متنی جس‌جو

از اولین جویشگرهای فارسی است. در قسمت دیکشنری و مترجم شما می‌توانید از زبان های فارسی ، انگلیسی ، اسپانیایی ، آلمانی ، فرانسوی و ایتالیایی استفاده کنید که همه این زبان‌ها دو طرفه هست. همچنین می‌توانید همزمان در صفحات یکی از این زبان‌ها جستجو کنید مثلا با تایپ کلمه سلام و گذاشتن تیک بر روی زبان آلمانی کلمه شما در صفحات آلمانی جستجو می‌شود. (علیرضا یاری ،۱۳۹۴)

این فراجویشگر در اردیبهشت ۱۳۸۱ با هدف ایجاد خدمات جستجو برای فارسی‌زبانان تاسیس شده است و از بانک‌اطلاعات دیگر جستجوگرها مانند گوگل و یاهو استفاده می‌کند.این سایت توسط آقای علیرضا شیرازی که مدیر بلگفا هست مدیریت می‌شود و ظاهر سایت مانند بلاگفا ساده و پرسرعت هست که کاربرانی که از اینترنت کم سرعت استفاده می‌کنند به راحتی می‌تواند از آن استفاده کنند. (علیرضا یاری ،۱۳۹۴)

فراجویشگر سلام

سلام یکی از بهترین و قدرتمندترین جویشگرهای ایرانی است که با هوش‌مصنوعی کارآمد خود، عبارت مورد جست‌وجو را در بزرگ‌ترین جویشگرهای جهان جستجو می‌کند و در کمترین زمان، نتایج آن را ارائه می‌دهد. سلام عبارت جست‌وجو شده را از موتورهایی مانند گوگل ، بینگ ، اسک ، بلکو ، یاهو ، یاندکس و زال جستجو کرده و بهترین و مربوط‌ترین جواب‌ها را نشان می‌دهد.فراجویشگر سلام تنها

توانایی آن را دارد که اطلاعات را از خارج کشور جمع‌آوری کرده و به کاربر نشان می‌دهد و با قطع اینترنت‌ملی و در صورت بومی شدن اینترنت عملا فعالیت خاصی ندارد. (علیرضا یاری ،۱۳۹۴)

حال در جدول شماره ۷ ما یکسری دیگر از معیار های ارزیابی دو فراجویشگر سلام و پارسیک را نشان می‌دهیم که معایب و کمبودهای آنها می‌تواند کمک شایانی برای ایده های کارهای آینده باشد که با رفع آن معایب بتوانیم فراجویشگری کامل‌تر پیاده سازی کنیم. (علیرضا یاری ،۱۳۹۴)

معیار	شاخص مورد استفاده		پارسیک	سلام
پوشش			۲ موتور جستجو (گوگل و بینگ)	۴ موتور جستجو (گوگل و بینگ و یاهو و زال)
قابلیت جستجو	جستجوی پیشرفته		×	×
	تصحیح خطاهای نوشتاری		×	√
	پیشنهاد تکمیل پرس‌وجو		×	√
	تعیین عبارات در پرس‌وجو		×	×
	استفاده از عملگرهای منطقی در پرس‌وجو		√	√
	جستجو براساس	نوع‌سند	×	√
		تاریخ	×	√
		عنوان	×	×
		آدرس	×	×
		محدود به یک آدرس خطی	√	√
	جستجو کلمات‌عمومی		√	√
	گروه‌بندی نتایج		×	√
	پیشنهاد پرس‌وجوی مرتبط		×	√
	نمایش تعداد نتایج		√	×

نمایش نتایج	نمایش مدت زمان پاسخگویی		×	×
	مشخص کردن لغات پرس‌وجو در نتایج		√	×
	قابلیت جستجو انواع محتوا	متن	√	√
		تصویر	√	×
		ویدئو	×	×
		علمی	√	×
		نقشه	×	×
	قابلیت‌های پردازشی زبان‌فارسی		√	√

جدول شماره۷: معیارهای ارزیابی موتورهای فراجویشگر فارسی

بدلیل اینکه حجم داده‌هایی که نیازمند ایندکس شدن در پایگاه‌داده هستند، بسیار زیاد است، فرآیند ذخیره‌سازی و الگوریتم رتبه‌بندی از اهمیت بالایی برخوردارند و در بسیاری از پرس‌وجوهای کاربران استفاده از پردازش زبان‌طبیعی، یادگیری‌ماشین و روش های مختلف هوش‌مصنوعی از اهمیت بالایی برخوردار است. جویشگرهای بومی علاوه بر الزامات نرم‌افزاری و هوش‌مصنوعی نیازمند تجهیزات سخت‌افزاری پیشرفته نیز می‌باشد.

جویشگرها براساس کلمات‌کلیدی که کاربران جستجو می‌کنند توانایی بدست آوردن اطلاعات زیادی از رفتار مردم در جامعه را دارند و از دیدگاه‌ملی استفاده کاربران از جویشگرهای‌بومی باعث کاهش نگرانی استفاده سرویس‌های‌خارجی از اطلاعات تجمیعی کاربران می‌شود و از سوی دیگر نیز کاربران اکثرا برای حفظ حریم‌شخصی ترجیح می دهند از موتورهای‌جستجوی خارج از کشور استفاده کنند که بایدبتوانیم نگرانی کاربران در این مورد را رفع کنیم.

همانطور که در بحث قبلی گفتیم ، پیشرفت‌های زیادی در مورد راه‌حل‌های کارآمد و دقیق برای مشکل پردازش پرس‌وجوها در محیط فراجویشگرها حاصل شده است. با این حال ، به عنوان یک موضوع در حال ظهور ، بسیاری از مشکلات برجسته باقی مانده باید حل شود در این بخش چند چالش ارزشمند در این زمینه را ذکر می‌کنیم.

از معایب فراجویشگرها می‌توان به موارد زیر اشاره کرد: (Jaime Arguello ,۲۰۱۱)

اگر برای دقیق‌تر شدن جستجوی خود از یک سینتکس خاص استفاده کنیم ممکن است که آن فراجویشگر توانایی اینکه آن سینتکس را به سینتکس قابل فهم هریک از جویشگرها تبدیل کند نداشته باشد که این باعث می‌شود که دقت نتایجی که از فراجویشگر بازیابی می‌شود از نتایج حاصل از جستجوی پیشرفته در یک جویشگر کمتر باشد به این معنی که فراجویشگر از تمام ظرفیت همه جویشگرها استفاده نمی‌کند. (Jaime Arguello ,۲۰۱۱)

موضوع دیگر منابعی هستند که فراجویشگرها از آنها استفاده می‌کنند که اگر این منابع معیار میزان مرتبط بودن نتایج با پرسش کاربر را در نظر نگرفته باشند و فراجویشگر نیز ارزیابی و پردازشی روی آنها نداشته باشد کاربر با نتایجی روبه‌رو می‌شود که ارتباط چندانی با پرس‌وجوی مدنظر وی ندارند(Jaime Arguello ,۲۰۱۱

یکی دیگر از مشکلات و چالش‌ها این است که معمولا تعداد رکوردهایی که می‌توانیم از هر جویشگر بازیابی کنیم محدود هستند و این امر در کمیت و کیفیت فراجوشگر تاثیرگذار است و بنابراین ممکن است که نتایج کاملی را به کاربر نشان ندهند. (Jaime Arguello ,۲۰۱۱)

اکثر موتورهای فراجویشگر موجود روی تعداد کمی جویشگر عمل می‌کنند که این امر باعث کاهش کیفیت نتایج می‌شود. (Jaime Arguello ,۲۰۱۱)

تغییرات لحظه ای پارامترهای ارتباطی و فرمت نمایش نتایج جویشگرهای پایه نیز یکی دیگر از چالش های موجود می‌باشد. (Jaime Arguello ,۲۰۱۱)

مشکل انتخاب پایگاه داده : یعنی شناسایی جویشگرهایی که باید پرس‌وجوی کاربر را برای دریافت نتایج به آنها ارسال کنیم. (Jaime Arguello ,۲۰۱۱)

ادغام سیستم‌های محلی با استفاده از روش‌های مختلف نمایه‌سازی: استفاده از تکنیک های مختلف نمایه‌سازی در سیستم‌های محلی مختلف می‌تواند تأثیر جدی بر سازگاری شباهت‌های محلی داشته باشد. استفاده از تکنیک‌های مختلف نمایه‌سازی در واقع می‌تواند روی دقت تخمین در هر یک از سه مؤلفه نرم افزاری (یعنی انتخاب پایگاه‌داده، انتخاب اسناد و ادغام نتیجه) تأثیر بگذارد. (Jaime Arguello ,۲۰۱۱)

یکپارچه‌سازی سیستم های محلی که داده‌هایی که پشتیبانی می‌کنند متفاوت است(به عنوان مثال نمایش داده بولی در مقابل نمایش داده‌های فضای بردار) به احتمال زیاد هنگام تلفیق سیستم‌های محلی که هم از نمایش داده های بولی و هم از فضای برداری پشتیبانی می‌کنند ، با مشکلات جدی روبرو خواهیم شد بنابراین باید روشهای ادغام موثرتر را توسعه دهید. (Jaime Arguello ,۲۰۱۱)

روشهای جدید که ویژگی‌های خاص محیط موتور فراجویشگر را در نظر می‌گیرند ، باید طراحی و ارزیابی شوند. یکی از ویژگی‌های خاص این است که وقتی یک سندی توسط جویشگر بازیابی نمی‌شود ، ممکن است به این دلیل باشد که این سند توسط جویشگر ایندکس نمی‌شود. (Jaime Arguello ,۲۰۱۱)

حال ما در این تحقیق قصد داریم موتور فراجویشگری طراحی کنیم که پرسش کاربر را دریافت می‌کند و در مرحله اول از بین جویشگرهای فارسی مرتبط‌ترین ها را انتخاب می‌کند و پرسش کاربر را به آنها می‌فرستد و بعد نتایج دریافت شده را طبق الگوریتم‌هایی ترکیب می‌کند و مرتبط‌ترین‌هایشان را به کاربر نشان می‌دهد.

۱-۳-بازیابی اطلاعات عمودی

هدف بازیابی‌اطلاعات (IR) رتبه‌بندی موارد با توجه به نیازاطلاعات کاربر است که با استفاده از یک پرس وجو بیان شده است.

یک سرویس جستجوی عمودی برای برآوردن نوع خاصی از اطلاعات مورد نیاز (به عنوان مثال ، جستجوی تصویر ، جستجوی محصول ، جستجوی شغل) طراحی شده است. به همین دلیل ، عمودی‌های مختلف اغلب انواع مختلفی از نتایج (مثلاً تصاویر ، توضیحات محصول ، لیست شغل) را بازیابی می‌کنند و از الگوریتم های بازیابی متفاوت استفاده می‌کنند.تحقیقات جستجوی فدرال غالباً بین محیطی اشتراکی و غیراشتراکی فرق می‌گذارد. (Jaime Arguello ,۲۰۱۱)

بیشتر سیستم‌های جستجوی تجمیعی از معماری خط‌لوله با سه وظیفه متعاقب پیروی می‌کنند (Jaime Arguello ,۲۰۱۱)

اولین زیروظیفه (vertical selection) پیش‌بینی می‌کند که کدام عمودی (در صورت وجود) مربوط به پرس‌وجو است. می‌توان وظیفه انتخاب عمودی را به عنوان وظیفه تصمیم‌گیری در مورد اینکه کدام عمودی‌ها باید بدون توجه به موقعیتشان در SERP(search engine results pages) نمایش داده شوند ،

مشاهده کرد. به همین دلیل ، بیشتر رویکردها برای مرحله انتخاب‌عمودی پیش‌بینی های خود را با استفاده از شواهد قبل از بازیابی (به عنوان مثال ، جستجوی عبارت "خبر" را شامل می‌شود) را پایه گذاری می‌کند. زیروظیفه دوم (انتخاب نتایج‌عمودی) پیش‌بینی نتایج حاصل از یک عمودی خاص برای ارائه در SERP است. وظیفه انتخاب نتایج‌عمودی دارای یک هدف دوگانه است. هدف اصلی رضایت کاربر به طور مستقیم با نتایج‌عمودی است که در SERP جمع می‌شوند. هدف ثانویه دارای اهمیت بیشتری است .(Jaime Arguello ,۲۰۱۷

برخی از عمودها قابلیت جستجوی مستقیم دارند .اگر کاربر متوجه شود که ممکن است اطلاعات مربوط به عمودی وجود داشته باشد ، می‌تواند به سمت عمودی حرکت کند ، نتایج‌عمودی بیشتری را بررسی کند و حتی درخواستهای جدید را نیز به موتور جستجوی عمودی صادر کند. (Jaime Arguello ,۲۰۱۷)

در این راستا ، هدف ثانویه از انتخاب نتایج‌عمودی ، انتقال چگونگی محتوای عمودی زیرین است. بیشتر سیستم‌های جستجوی‌تجمیعی منتشر شده شرح داده شده است ، انتخاب نتایج‌عمودی را انجام نمی دهند و به سادگی چند نتیجه برتر را که در پاسخ به پرس‌وجو نمایش داده شده است ، نشان می‌دهند. وظیفه فرعی سوم و آخر (ارائه عمودی) تصمیم‌گیری در مورد ارائه کدام عمودی انتخابی است. عمودی‌های مختلف به طور معمول با بازنمایی های مختلف جانشین همراه است. به عنوان مثال ، نتایج تصویر با استفاده از ریزعکسها نمایش داده می‌شود ، در حالی که نتایج اخبار با استفاده از عنوان مقاله ، منبع ، تاریخ انتشار نمایش داده می‌شود و ممکن است شامل یک تصویر اختیاری از مقاله زیر باشد. به دلایل زیبایی‌شناختی و برای انتقال بهتر چگونگی داشتن محتوای‌مناسب برای کاربر فعلی ، نتایج‌عمودی معمولاً با هم جمع می‌شوند (یا به صورت افقی یا عمودی) روی SERP جمع شده. (Jaime Arguello ,۲۰۱۷

هدف از ارائه عمودی نمایش مناسب‌ترین عمودی‌ها با روشی برجسته‌تر است. یک رویکرد رایج این است که آنها را در SERP بالاتر نشان دهید (مثلاً بالاتر از اولین نتیجه وب). ارائه عمودی پس از صدور پرس‌وجو به عمودی اتفاق می‌افتد. بنابراین ، رویکردها برای نمایش عمودی می‌تواند پیش‌بینی‌های خود را با استفاده از شواهد قبل از بازیابی و همچنین شواهد پس از بازیابی (به عنوان مثال ، تعداد نتایج برگشت یافته توسط عمودی ، نمرات بالاترین بازیابی یا تعداد عبارات پرس‌وجو در نتایج برتر نشان دهد).

هدف از جستجوی فدرال(federated search) ارائه جستجوی یکپارچه در میان چندین مجموعه از اسناد متنی است که به آنها به عنوان منبع نیز گفته می‌شود. (Jaime Arguello ,۲۰۱۷)

هدف از بخش انتخاب عمودی (vertical selection) این است که تصمیم بگیرید کدام عمودها را در پاسخ به یک سؤال ارائه دهید. انتخاب عمودی در اصل یک وظیفه طبقه بندی چند کلاسه است. (Jaime Arguello,۲۰۱۷)

هدف از فن آوری جستجوی تجمیعی(aggregated search) ارائه جستجوی یکپارچه در طیف گسترده ای از سیستم‌های جستجوی بسیار تخصصی در یک رابط یکپارچه است - یک جعبه پرس‌وجو جستجو و یک نمایش مشترک از نتایج. (Jaime Arguello,۲۰۱۷)

تا به امروز ، بیشتر تحقیقات در جستجوی‌تجمیعی بر دامنه جستجوی‌وب متمرکز شده است؛ به همین دلیل ، بیشتر تحقیقات مورد بررسی در این مقاله نیز بر حوزه جستجوی‌وب متمرکز خواهد شد. پورتال‌های جستجوی‌وب تجاری مانند Google ، Bingو Yahoo! علاوه بر جستجوی‌وب ، به طیف گسترده‌ای از خدمات جستجوی‌تخصصی دسترسی پیدا کنید. این خدمات جستجوی تخصصی به خدمات جستجوی‌عمودی یا به صورت عمودی گفته می‌شود. (Jaime Arguello ,۲۰۱۷)

دو روش برای دستیابی کاربر به عمودی‌ها وجود دارد. اگر کاربر نتایج را از یک عمودی مشخص بخواهد و اگر عمودی قابلیت جستجوی مستقیم را دارد ، کاربر می‌تواند پرس‌وجو را مستقیماً به حالت عمودی صادر کند. با این حال ، در موارد دیگر ، کاربر ممکن است بداند که یک عمودی دارای محتوای مرتبط است ، یا ممکن است نتایج حاصل از چندین عمودی را به طور همزمان بخواهد .به همین دلیل ، یک کار مهم برای ارائه‌دهندگان جستجوی تجاری به پیش‌بینی و ادغام محتوای عمودی مربوطه در کنار نتایج جستجوی وب اصلی تبدیل شده است. (Jaime Arguello ,۲۰۱۷)

شکل شماره ۱، یک صفحه نتیجه جستجوی تجمیع(SERP) را در دامنه‌وب را نشان می‌دهد. در پاسخ به پرس‌وجوی "Saturn" ، یک سیستم جستجوی‌تجمعی تصمیم گرفت علاوه بر نتایج اصلی‌وب ، نتایج‌عمودی اخبار ، تصویر و ویدیو را به نمایش بگذارد. در این حالت ، سیستم پیش‌بینی می‌کند که بیشترین عمودی‌ها به ترتیب مربوط به اخبار ، تصاویر و عمودی‌های ویدیویی هستند. (Jaime Arguello ,۲۰۱۷)

شکل شماره۱: صفحه نتیجه یک جستجوی تجمیعی. (Jaime Arguello ,۲۰۱۷)

هدف از جستجوی‌فدرال ارائه جستجوی یکپارچه در میان چندین مجموعه از اسناد متنی است که به آنها به عنوان منبع نیز گفته می‌شود .مشابه جستجوی کل ، جستجوی فدراسیون معمولاً در سه وظیفه فرعی تجزیه می‌شود. (Jaime Arguello ,۲۰۱۷)

اولین وظیفه فرعی (نمایش منابع resource representation) ساخت توضیحی در مورد هر منبع توزیع شده است که می‌تواند برای پیش‌بینی کدام یک از آنها در پاسخ به یک سؤال جستجو کند. دومین وظیفه فرعی (انتخاب منبع resource selection) پیش‌بینی اینکه منابع در جستجوی پاسخ به یک پرس‌وجو جستجو می‌کنند. سومین وظیفه‌فرعی (ادغام نتایجresults merging) پیوند دادن نتایج حاصل از منابع مختلف انتخاب شده در یک رتبه بندی واحد است. (Jaime Arguello ,۲۰۱۷)

۱-۳-۱-مرور بر الگوریتم های جستجوی‌تجمیعی

موفق‌ترین روش‌ها برای انتخاب و ارائه عمودی از یادگیری‌ماشین استفاده می‌کند تا طیف گسترده‌ای از شواهد را به‌عنوان ویژگی‌های ورودی به مدل ترکیب کند. ویژگی‌ها را می‌توان از پرس‌وجو ، از حالت عمودی یا از جفت پرس‌وجو عمودی ایجاد کرد. به عنوان مثال ، یک نوع ویژگی پرس‌وجو ممکن است در

نظر بگیرد که آیا پرس‌وجو حاوی کلمه کلیدی "اخبار" است ، ممکن است یک نوع ویژگی عمودی تعداد کلیک‌های اخیر بر روی نتایج‌عمودی را در نظر بگیرد و ممکن است یک نوع پرس‌وجوعمودی تعداد اسناد مرتبط در مجموعه عمودی زیرین پرس‌وجو را تخمین بزند.مؤثرترین رویکردها برای انتخاب و ارائه‌عمودی ، استفاده خلاقانه از منابع مختلف شواهد موجود در سیستم ، از جمله داده‌های پرس‌وجو مربوط به مشخصات عمودی ، اسناد عمودی نمونه بردار و تعامل کاربر قبلی با محتوای عمودی را ایجاد می‌کند. (Jaime Arguello ,۲۰۱۷)

در حالی که ادغام شواهد یک مسئله کلیدی و مهم برای جستجوی تجمیعی است ، اما دو چالش اصلی را نیز مطرح می‌کند. (Jaime Arguello ,۲۰۱۷)

اولین چالش این است که همه ویژگی‌ها ممکن است برای همه عمودی‌ها در دسترس نباشد. به عنوان مثال ، برخی از عمودها را نمی‌توان مستقیماً توسط کاربران جستجو کرد. (Jaime Arguello ,۲۰۱۷)

چالش دوم این است که ، حتی اگر یک ویژگی برای همه عمودی‌ها در دسترس باشد ، ممکن است به طور یکسان در بین عمودی‌ها پیش‌بینی نشود. به عنوان مثال ، برخی از عمودی‌ها بیش از سایرین کلیک می‌شوند. به عنوان‌مثال ، یک عمودی خبر احتمالاً کلیک بیشتری نسبت به یک عمودی آب‌وهوا دارد ، که برای نمایش اطلاعات لازم به طور مستقیم در SERP طراحی شده است. (Jaime Arguello ,۲۰۱۷)

با توجه به دو چالش ذکر شده در بالا ، رویکردهای انتخاب عمودی به طور معمول یک مدل متفاوت را برای هر عمودی کاندید یاد می‌گیرد. به این ترتیب ، هر مدل می‌تواند نمایه‌ای از ویژگی‌های متفاوتی را اتخاذ کند و می‌تواند یک رابطه عمودی خاص بین مقادیر ویژگی و ارتباط عمودی خاص را بیاموزد. ارائه عمودی نیاز به حل‌وفصل اختلاف بین عمودی‌های مختلف برای نمایش در SERP دارد. به بیان دیگر ، ارائه عمودی نیاز به پیش‌بینی میزان ارتباط یک عمودی نسبت به نتایج وب و نسبت به سایر عمودی‌ها برای نمایش دارد. (Jaime Arguello ,۲۰۱۷)

رویکردهای ارائه عمودی را می‌توان به دو نوع طبقه‌بندی کرد: pointwise و pairwise interleaving. (Jaime Arguello ,۲۰۱۷)

روش های نقطه‌ای(pointwise) یاد می‌گیرند که میزان ارتباط هر بلوک یا ماژول عمودی را در پاسخ به یک پرس‌وجو پیش‌بینی کنند. بلوک های عمودی با توجه به اهمیت پیش‌بینی شده آنها با پرس‌وجو قرار می‌گیرند. (Jaime Arguello ,۲۰۱۷)

روش های Pairwise یاد می‌گیرند که ارتباط نسبی بین جفت بلوک‌ها یا ماژول‌های عمودی و / یا وب را پیش‌بینی کنند. بلوک‌های عمودی به گونه‌ای قرار گرفته‌اند که حداکثر با ترجیحات زوج پیش‌بینی شده توسط سیستم سازگار هستند. (Jaime Arguello ,۲۰۱۷)

جدیدترین روشها برای انتخاب و ارائه عمودی طیف گسترده ای از شواهد را برای پیش‌بینی ترکیب می‌کند. یک راه مناسب برای ترکیب شواهد ، آموزش یک مدل با استفاده از یادگیری ماشین است. الگوریتم‌های یادگیری‌ماشین یاد می‌گیرند که با استفاده از مجموعه مثالهای مثبت و منفی پیش‌بینی کنند. به عنوان مثال ما می‌توانیم یک مدل انتخاب عمودی برای عمودی خبر با استفاده از مجموعه‌ای از پرس‌وجوهای نمونه برای هر سیستم یاد بگیریم که کدام‌ها انتخاب شوند یا انتخاب نشوند. با این حال ، طراح سیستم وظیفه دارد با استفاده از مجموعه اقدامات یا ویژگی‌ها ، تصمیم بگیرد که چگونه جفت‌های پرس‌وجوهای عمودی را نشان دهد. ویژگی‌های خوب آنهایی است که بیشترین ارتباط عمودی را با یک پرس‌وجو دارند و ویژگی‌های بد آنهایی هستند که ارتباطی ندارند. بسیاری از خلاقیت‌ها به سمت طراحی ویژگی‌های مؤثر می‌روند. (Jaime Arguello ,۲۰۱۷)

روش‌های زیادی وجود دارد که با استفاده از آن ، سیستم جستجوی تجمعی می‌تواند پیش‌بینی کند که یک عمود خاص به یک پرس‌وجو مربوط می‌شود. وظیفه پیش‌بینی اینکه آیا یک عمودی خبر مرتبط است را در نظر بگیرید. اگر پرس‌وجو شامل اصطلاح خبر باشد ، تقریباً مطمئن است که مربوط به آن است. به طور مشابه ، یک سیستم ممکن است تعیین کند که عمودی خبر مربوط به پرس‌وجوی "انتخابات ریاست جمهوری" است زیرا بسیاری از اسناد موجود در این مجموعه شامل این عبارات پرس‌وجو هستند. سرانجام ، یک سیستم پیش‌بینی می‌کند که اگر عمودی خبر مربوط پرس‌وجو شبیه به پرس‌وجوهای اخیر پشت سر هم باشد که مستقیماً توسط کاربر اعلان می‌شود ، عمودی خبر مرتبط است. موفق ترین روش‌ها برای انتخاب و ارائه عمودی از یادگیری‌ماشین استفاده می‌کند تا شواهد گسترده‌ای را به‌عنوان ویژگی‌های ورودی به یک مدل ترکیب کند. در یادگیری انواع مختلفی از ویژگی‌ها ، کمک به آگاهی از شباهت‌ها و تفاوت‌های آن‌ها مفید است. به عنوان‌مثال ، درک منابع موردنیاز برای

تولید هر ویژگی مفید است و اینکه آیا تولید یک ویژگی نیاز به صدور پرس‌وجو کامل به یک عمودی خاص دارد. (Jaime Arguello ,۲۰۱۷)

۱-۴-پرس و جو

۱-۴-۱-ویژگی‌ها

ویژگی‌ها را می‌توان در دو بعد مشخص کرد. بعد اول مربوط به این است که آیا مقدار ویژگی به پرس‌وجو ورودی ، عمودی مورد نظر یا جفت پرس‌وجو و عمودی بستگی دارد. ویژگی‌های پرس‌وجو ویژگی‌های پرس‌وجو ورودی را توصیف می‌کنند و مقادیر آنها مستقل از عمودی در نظر گرفته شده است. ممکن است یک نوع ویژگی پرس‌وجو در نظر بگیرد که آیا پرس‌وجو شامل کلمه کلیدی "اخبار" است. ویژگی‌های عمودی ویژگی‌های یک عمودی خاص را توصیف می‌کنند و مقادیر آنها مستقل از پرس‌وجو است. (Jaime Arguello,۲۰۱۷)

ویژگی‌های پرس‌وجو توسط $\emptyset_q^*$ نشان داده شده است ، ویژگی‌های عمودی با $\emptyset_v^*$ نشان داده می شوند ، و ویژگی‌های پرس‌وجو عمودی توسط $\emptyset_{q,v}^*$ نشان داده می‌شوند. (Jaime Arguello ,۲۰۱۷)

با استفاده از این نماد ، می‌توانیم یک جفت پرس‌وجو-عمودی را به عنوان یک بردار از ویژگی‌ها در نظر بگیریم که ویژگی‌های مختلف از پرس‌وجو را شرح می‌دهد؛ زوج عمودی کاندید و پرس‌وجوی عمودی: ((Jaime Arguello ,۲۰۱۷

$$[\emptyset_q^* , \emptyset_v^* , \dots\dots , \emptyset_{q,v}^*]$$

ویژگی‌های پرس‌وجو

ویژگی‌های پرس‌وجو از پرس‌وجو تولید می‌شوند و مستقل از عمودی هستند که در نظر گرفته می‌شوند. ((Jaime Arguello ,۲۰۱۷

ویژگی‌های رشته‌ای پرس‌وجو

ویژگی‌های رشته پرس‌وجو ، وجود یا عدم وجود برخی از کلمات کلیدی موجود در پرس‌وجو را در نظر می‌گیرد. (Jaime Arguello ,۲۰۱۷)

به عنوان مثال ، عبارات پرس‌وجو مانند " image "، " picture" و"pics" نشان می‌دهد که عمودی image مرتبط است ، در حالی که عبارات پرس‌وجو مانند " buy "، "price" ,"shop" نشان می‌دهد که عمودی shopping مرتبط است. (Jaime Arguello ,۲۰۱۷)

ویژگی‌های پرس‌وجو خصوصیاتی از پرس‌وجو مانند طول پرس‌وجو ، ظرفیت یا حضور کاراکترهای خاص را توصیف می‌کنند. (Jaime Arguello ,۲۰۱۷)

ویژگی‌های کلاس بندی

می‌توان پرس‌وجوها را به کلاس‌های مختلفی مرتبط با قصد کاربر طبقه‌بندی کرد. (Jaime Arguello) ,۲۰۱۷

به‌عنوان مثال ، نمایش داده‌های وب مشخص شده به سه کلاس: navigation (هدف این است که به یک صفحه خاص برسید) ، informational (هدف این است که اطلاعات موجود در یک یا چند صفحه را پیدا کنید) و transaction (منظور از انجام یک تراکنش با واسطه است. برخی از کلاس‌های پرس‌وجو بیشتر از سایرین از نتایج‌عمودی بهره مند می‌شوند. به عنوان‌مثال ، پرس‌وجوهای informational بیشتر از پرس‌وجو های navigation از نتایج‌عمودی بهره‌مند می‌شوند. در زمینه جستجوی تجمیعی ، کار قبلی ویژگی‌های مرتبط با احتمال navigation بودن یک پرس‌وجو را در نظر گرفته است (احتمالاً سؤالی است که سیستم نباید نتایج‌عمودی را نشان دهد. (Jaime Arguello ,۲۰۱۷)

ویژگی‌های دسته بندی

یکی از مؤثرترین ویژگی‌های پرس‌وجو مورد استفاده در کارهای قبلی ، ویژگی‌های دسته بندی پرس‌وجو است که میزان تمایل پرس‌وجو به یک مجموعه از پیش تعریف شده از عناصر محلی را اندازه گیری می‌کند. ویژگی‌های دسته بندی پرس‌وجو به دو دلیل موفقیت آمیز بوده‌اند. در مرحله اول ، بسیاری از عمودی‌های مورد بررسی در انتخاب عمودی و ارایه عمودی قبلی مورد بررسی قرار گرفته‌اند (به عنوان مثال ، امور مالی ، بهداشت ، فیلم ، ورزش ، ورزش ، مسافرت). (Jaime Arguello ,۲۰۱۷)

دوم ، طبقه‌بندی پرس‌وجو به طور گسترده‌ای در زمینه سایر وظایف بازیابی‌اطلاعات مانند رتبه‌بندی اسناد مورد مطالعه قرار گرفته است

بنابراین ، سیستم‌های جستجوی‌تجمعی می‌توانند از رویکردهای طبقه‌بندی پرس‌وجو به خوبی آزمایش شده به منظور تولید ویژگی استفاده کنند. (Jaime Arguello ,۲۰۱۷)

بگذارید C مجموعه‌ای از عناوین گروه‌های موردنظر (مثلاً امور مالی ، بهداشت ، فیلم ، ورزش ، مسافرت ، و غیره) را مشخص کند و بگذارید R_q^n نتایج جستجوی بازگشت داده شده را که در پاسخ به سؤال q از جمع‌آوری اسناد از پیش طبقه‌بندی شده برگردانده شود ، نشان دهد. وابستگی query q به طبقه c €C می‌تواند به صورت زیر محاسبه شود: (Jaime Arguello ,۲۰۱۷)

$$P(c|q)=\frac{1}{Z}\sum_{d\in R_q^n} P(c|q) \times score(d,q),$$

جایی که P (c | d) مقدار اطمینان پیش‌بینی را نشان می‌دهد که سند d متعلق به طبقه c است ، نمره (d ، q) نمره بازیابی سند d را در پاسخ به query q و فاکتور نرمال سازی (q ، d) Rnq ۲d P = Z (Jaime Arguello ,۲۰۱۷)

در فرمول فوق ، P (c | q) با میانگین تمایل اسناد در Rnq به طبقه c متناسب است. با این حال ، میانگین یک میانگین وزنی است ، جایی که وزنها با نمره بازیابی نرمال هر سند برتر همراه است. این شبیه به چگونگی تخمین احتمال شرایط در یک مدل زبان ارتباط است (Jaime Arguello ,۲۰۱۷)

ویژگی‌ها

ویژگی‌های Corpus شواهدی را از اسناد عمودی نمونه برداری یا اسناد خارجی که بطور اکتشافی با هر عمودی در ارتباط هستند استخراج می‌کند. ایده این است که از ارتباط پیش‌بینی شده اسناد عمودی استفاده شود تا به پیش‌بینی ارتباط عمودی کمک کند.(Jaime Arguello ,۲۰۱۷)

ویژگی‌های Query-log : این ویژگی‌ها از شباهت بین پرس‌وجو و مواردی که مستقیماً توسط کاربران صادر می‌شود ، استفاده می‌کنند ، و انواع نیازهای اطلاعاتی را که توسط عمودی برآورده می‌شود ، توصیف می‌کنند. (Jaime Arguello ,۲۰۱۷)

ویژگی‌های پرس‌وجو شواهدی را از رشته پرس‌وجو به دست می آورند. برخلاف ویژگی‌های corpus و ویژگی‌هایquery-log مقدار ویژگی پرس‌وجو مستقل از عمودی است که مورد بررسی است. به

عنوان‌مثال ، این ویژگی‌ها از سوالی برای پرس‌وجو یا وجود یک نوع خاص از نام خاص استفاده می‌کنند . (Jaime Arguello ,۲۰۱۷)

ویژگی‌های corpus

ویژگی‌های Corpus به پیش‌بینی ارتباط عمودی بر اساس ارتباط (پیش‌بینی شده) اسناد مرتبط با عمودی کمک می‌کند. ما دو روش برای پیوند اسناد با عمودی را بررسی می‌کنیم. (Jaime Arguello) ,۲۰۱۷

یک گزینه این است که نتایج را مستقیماً از حالت عمودی نمونه‌برداری کنید. در جستجوی federated، نمونه‌برداری از منابع معمولاً برای نمایش منابع مورد استفاده قرار می‌گیرد که منبع با نمونه‌ای از اسناد در نظر گرفته می‌شود که معمولاً برای اسناد مشاهده نشده در منبع باشد. ایجاد ویژگی‌های corpus از اسناد نمونه بردار عمودی محدودیت بالقوه‌ای دارد. در هسته ، ویژگی‌های corpus شواهد مربوط به عمودی از ارتباط نمونه را به دست می‌آورد. با این حال ، بیشتر روش‌ها برای پیش‌بینی ارتباط نمونه ، مبتنی بر متن است (یعنی ، آنها بر شباهت متن-پرس‌وجو نمونه متمرکز هستند). این مورد برای عمودی‌هایی است که در بازیابی متن ضعیف هستند مشکل است (به عنوان مثال تصاویر ، فیلم ، نقشه). (Jaime Arguello ,۲۰۱۷)

به همین دلیل ، علاوه بر این که ویژگیهای corpus را از اسنادی که مستقیماً از عمودی نمونه برداری شده‌اند ، استخراج می‌کنیم ، از ویژگی‌های corpus نیز از اسناد غنی از متن که خارج از عمودی هستند با استفاده از heuristics دستی با عمودی ، استخراج می‌کنیم. (Jaime Arguello ,۲۰۱۷)

۱-۵-انتخاب عمودی(Vertical Selection)

هدف از انتخاب عمودی این است که تصمیم بگیرید کدام عمودها را در پاسخ به یک سؤال ارائه دهید. انتخاب عمودی در اصل یک کار طبقه بندی چندکلاسه است. با توجه به پرس‌وجو ، سیستم باید تصمیم باینری را برای هرعمودی کاندید شده بگیرد. انتخاب عمودی مانند بخش انتخاب منبع در federated search است.(Shokouhi et al. , ۲۰۱۱)

۱-۶-تعریف وظیفه رسمی

اگر V مجموعه ای از تمام عمودی‌های کاندید را مشخص کند و Q تمام مجموعه پرس‌وجوها را مشخص کند. فرض می‌کنیم که هر پرس‌وجو q€ Q با مجموعه ای از عمودی‌های Vq € V مرتبط است. در جستجوی تجمیعی ، امکان پذیر است که هیچ عمودی به پرس‌وجوی q مربوط نباشد. در این حالت ، انتخاب مطلوب برای انتخاب کننده عمودی این است که هیچ عمودی مربوط را پیش‌بینی نکنید و به سادگی نتایج وب را نشان دهید. ما عدم وجود هرگونه عمودی مربوطه را توسط Vq = Ø بیان می‌کنیم. به طور کلی ، ممکن است که پرس‌وجو q با چندین عمودی واقعی مرتبط باشد. (Jaime Arguello) ,۲۰۱۱

روش تک مدرکه (Single-evidence)

ساده‌ترین روش انتخاب‌عمودی از یک منبع واحد از شواهد برای پیش‌بینی اینکه کدام عمودها در پاسخ به یک پرس‌وجو انتخاب کنند استفاده می‌کند.رویکردهایsingle evidence تنها شامل دو مرحله است: (۱) محاسبه اندازه‌گیری برای هر عمودی کاندید و (۲) استفاده از یک یا چند آستانه برای پیش‌بینی اینکه کدام عمودی‌ها را انتخاب کنید و کدام یک را سرکوب کنید. (Jaime Arguello ,۲۰۱۷)

یک گزینه دیگر استفاده از آستانه یکسان برای همه عمودی کاندید است. این گزینه جایگزین منطقی خواهد بود اگر ما معتقدیم که اندازه‌گیری single evidence به طور مستقیم در تمام عمودی های کاندید قابل مقایسه است. (Jaime Arguello ,۲۰۱۷)

روش دیگر ، ما می‌توانیم از آستانه های مختلف برای عمودی‌های مختلف کاندید استفاده کنیم. (Jaime) Arguello ,۲۰۱۷

آرگلو و همکاران در سال ۲۰۰۹ یک پیش‌بینی کننده‌ی تک شواهد حاصل از داده‌های query-log عمودی را ارزیابی کرد (یعنی ، از پرس‌وجوها بطور مستقیم برای عمودی توسط کاربر استفاده می‌شود). این روش نمره احتمال پرس‌وجو داده شده توسط یک مدل زبان تولید شده از querylog عمودی را اندازه گیری کرد. یک استانه مشابه اعمال می‌شد تا پیش‌بینی دودویی برای هر عمودی کاندید تولید شود. آستانه با استفاده از داده‌های آموزش در قالب مجموعه‌ای ازپرس‌وجوهای داده شده با داوری‌های مربوط به هر عمودی کاندید تنظیم شد. (Jaime Arguello ,۲۰۱۷)

Duarte-Torres و همکاران یک روش دیگر برای رویکرد single-evidence معرفی کردند که شامل نمره ReDDE ، نمره وضوح و نمره احتمال پرس‌وجو بود. نمرات ReDDE و نمره احتمال پرس‌وجو مؤثرتر از نمره Clarity بود ، احتمالاً به این دلیل که نمرات وضوح به طور مستقیم قابل مقایسه با عمودی نیستند. (Jaime Arguello ,۲۰۱۷)

$\emptyset_{q.v}^{Redde} = \frac{\lvert v \rvert}{\lvert v_s \rvert} \sum_{d \in R_{q,csi}^{n}} I(d \in v_s)$	معادله ReDDE
$\emptyset_{q,v}^{clarity} = \sum_{w} P(w\lvert\theta_{q,v}) \log(\frac{P(w\lvert\theta_{q,v})}{P(w\lvert\theta_G)})$	معادله نمره وضوح
$\emptyset_{q,v}^{large-doc} = \frac{1}{z_q} \prod_{w \in q} P(w\lvert\theta_{v_s})$	معادله نمره احتمال پرس‌وجو

جدول شماره۸: معادلات روش تک مدرکه(Jaime Arguello ,۲۰۱۷)

دیاز در سال ۲۰۰۹ یک رویکرد single-evidence را ارائه داد که از داده های کلیک عمودی (کلیک های عمودی قبلی و جست و خیزها) برگرفته شده است. (Jaime Arguello ,۲۰۱۷)

باز هم ، اجازه دهید C_q^v تعداد دفعاتی باشد که سیستم عمودی v را برای پرس‌وجوی q انتخاب کرده است و کاربر روی آن کلیک کند ، و اجازه دهید S_q^v تعداد دفعاتی را که سیستم عمودی v را برای پرس‌وجوی q انتخاب کرده است و کاربر روی آن کلیک نکرده است.

همانطور که در معادله ۲,۶ توضیح داده شده است ، نرخ کلیک برای عمودی v و پرس‌وجو q ($\emptyset_{q,v}^{click}$) تعداد دفعاتی است که بر روی عمودی کلیک شده است (C_q^v) از تعداد دفعات نمایش داده شده در SERP ($C_q^v + S_q^v$). (Jaime Arguello ,۲۰۱۷)

سرانجام ، سیستم می‌تواند تصمیم بگیرد که اگرتاریخچه نرخ کلیک از یک آستانه فراتر رود ، می‌تواند عمودی v را در پاسخ به q انتخاب کند. (Jaime Arguello ,۲۰۱۷)

محدودیت اصلی این رویکرد این است که نیاز به یک تطابق دقیق بین پرس‌وجوها دارد. از نظر تئوری ،عمودی v باید برای سؤالات مشابه نرخ کلیک مشابه‌ای داشته باشد. به عنوان مثال ، عمودی اخبار باید نرخ کلیک مشابه را برای پرس‌وجوهای " انتخابات ریاست جمهوری آمریکا "و" انتخابات ریاست جمهوری ایالت‌متحده " داشته باشد. (Jaime Arguello ,۲۰۱۷)

دیاز در سال ۲۰۰۹ پیشنهاد کرد تا با استفاده از نرخ‌کلیک در جهت عمودی v و سؤالات مشابه q (نرم افزار ۲,۷) میزان کلیک را از طریق عمودی v و پرس‌وجوی q یکنواخت کنید. ایده اصلی این است که آمار نرخ کلیک را از طریق سؤالاتی که احتمالاً هدف مشابهی دارند ، به اشتراک بگذارید. (Jaime Arguello ,۲۰۱۷

همانطور که قبلاً نیز گفته شد ، شباهت بین دو پرس‌وجو به روشهای مختلفی قابل محاسبه است ، به عنوان مثال ، براساس همپوشانی بین اصطلاحات پرس‌وجو ، براساس همپوشانی بین نتایج برتر از یک مجموعه خارجی ، یا براساس شباهت بین مدل‌های مربوطه است.

دیاز همبستگی Bhattacharyya را بین مدل‌های ارتباط مرتبط با هر دو سؤال محاسبه کرد: (Jaime Arguello ,۲۰۱۷

$$Sim(q,q')=\sum_w \sqrt{P(w|\theta_q)P(w|\theta_{q'})}$$

پیش‌بینی کننده‌های single-evidence ساده و شهودی هستند. با این حال ، آنها دو نقص اصلی دارند. اول ، آنها نیاز دارند که منبع شواهد برای کلیه عمودی‌های کاندید در دسترس باشد. شواهد query-log عمودی برای عمودی‌هایی که قابلیت جستجوی مستقیم ندارند در دسترس نخواهد بود. به همین ترتیب ، داده های کلیک عمودی برای عمودی‌هایی که برای کلیک کردن طراحی نشده‌اند ، در دسترس نخواهند بود. (Jaime Arguello ,۲۰۱۷)

دوم ، پیش‌بینی کننده‌های تک‌شواهد برای تهیه پیش‌بینی ها به یک منبع واحد شواهد متکی هستند. تحقیقات قبلی نشان داده است که رویکردهایی که منابع مختلف شواهد را با هم ترکیب می‌کنند عملکرد بهتری دارند. (Jaime Arguello ,۲۰۱۷)

بعد ، ما رویکردهایی را مرور می‌کنیم که چندین منبع شواهد را برای پیش‌بینی اینکه عمودی مربوط به یک سؤال است ، ترکیب می‌کند(Jaime Arguello ,۲۰۱۷)

روش چند مدرکه (Multiple Evidence)

موفق‌ترین روش ها برای انتخاب‌عمودی با استفاده از یادگیری‌ماشین برای ترکیب چندین منبع شواهد به عنوان ویژگی‌های ورودی به یک مدل. (Jaime Arguello ,۲۰۱۷)

ترکیب شواهد برای انتخاب عمودی دو چالش‌اصلی را ایجاد می‌کند.اول ، ممکن است ویژگی‌های خاصی برای برخی از عمودی‌ها در دسترس نباشد.به عنوان مثال ، عمودی بدون قابلیت جستجوی مستقیم فاقد ویژگی‌های query-log عمودی است.دوم ، انتخاب عمودی مستلزم یادگیری یک رابطه عمودی خاص بین ویژگی‌های خاص و ارتباط یک عمودی خاص است. (Jaime Arguello ,۲۰۱۷)

به عنوان مثال ، ویژگی‌های دسته‌بندی پرس‌وجو ممکن است برای عمودی‌هایی که به صورت موضعی متمرکز شده‌اند مؤثرتر از عمودی باشد که طیف گسترده ای از موضوعات را پوشش می‌دهد ، مانند عمودی پرسش و پاسخ جامعه. (Jaime Arguello ,۲۰۱۷)

در کلیه کارهای قبلی تا به امروز ، هر دو این چالش‌ها با آموزش طبقه‌بندی‌های باینری مستقل (یک در هر عمودی) برطرف شده است؛ از این نظر ، هر طبقه‌بندی می‌تواند نمایه‌ای از ویژگی‌های متفاوتی را اتخاذ کند و بر ویژگی‌هایی تمرکز کند که برای عمودی مربوطه به طور منحصر به فرد پیش‌بینی می‌شوند. (Jaime Arguello ,۲۰۱۷)

طبقه‌بندی‌گرهای یادگیری‌ماشین برای یادگیری یک مدل پیش‌بینی از داده‌های‌آموزش استفاده می‌کنند. در زمینه انتخاب عمودی ، داده‌های آموزش در قالب مجموعه‌ای از نمایش داده شد Qv با داوری‌های مرتبط با توجه به عمودیv (Jaime Arguello ,۲۰۱۷)

الگوریتم یادگیری ماشین از داده‌های آموزش برای یادگیری رابطه پیش‌بینی‌کننده بین مجموعه ویژگی‌های ورودی و ارتباط عمودی استفاده می‌کند. ما می‌توانیم از یک مدل انتخاب عمودی به شرح زیر فکر کنیم: (Jaime Arguello ,۲۰۱۷)

$$f(q, v) = g(\emptyset_{(q,v)}, \theta_v),$$

جایی که f(q,v) مقدار اطمینان مدل را نشان می‌دهد که v عمودی مربوط به پرس‌وجوی q است ، $\emptyset_{(q,v)}$ یک بردار ۱ × m از m را نشان می‌دهد ، و v پارامترهای مدل را نشان می‌دهد. تابع g و تعریف دقیق θ_v به الگوریتم یادگیری مورد استفاده بستگی دارد. یک تصمیم مهم استفاده از طبقه بندی خطی یا غیرخطی است(Jaime Arguello ,۲۰۱۷.)

طبقه بندی کننده خطی

در یک طبقه بندی‌خطی ، هر ویژگی به پیش‌بینی نهایی مدل کمک می‌کند ، اما مدل از تعامل بین ویژگی‌ها استفاده نمی‌کند. (Jaime Arguello ,۲۰۱۷)

به عنوان مثال ، مدل نمی‌تواند بیاموزد که اگر مقدار ویژگی i به عمودی v مرتبط تر از مقدار ویژگی j (یا برعکس)i مرتبط‌تر است. به عنوان نمونه‌ای از یک طبقه‌بندی‌کننده خطی ساده ، یک طبقه‌بندی‌کننده پرسپترون با توجه به تابع زیر پیش‌بینی می‌کند که v عمودی مربوط به q است. (Jaime Arguello) ,۲۰۱۷

$$f(q,v)=\begin{cases}1 & if\ \emptyset_{(q,v)}\ .\theta_v > 0\\ 0 & otherwise\end{cases}$$

در این حالت ، θ_v به عنوان یک بردار ۱ × m از وزن ویژگی‌ها تعریف می‌شود(Jaime Arguello) . ,۲۰۱۷

این الگوریتم پارامترهای θ_v را می‌آموزد که باعث می‌شود دقت طبقه بندی در مجموعه آموزش QV به حداقل برسد. (Jaime Arguello ,۲۰۱۷)

طبقه‌بندی خطی محبوب که در کار انتخاب عمودی قبلی مورد استفاده قرار می‌گیرد ، رگرسیون لجستیک است. (Jaime Arguello ,۲۰۱۷)

در مورد رگرسیون لجستیک ، $\emptyset_{(q,v)}$همچنین به عنوان یک بردار ۱ × m از وزن ویژگی‌ها تعریف می‌شود ، و f (q,v) ، توسط. : (Jaime Arguello ,۲۰۱۷)

$$F(q,v)=\frac{\exp(\emptyset_{(q,v)}.\theta_v)}{1+\exp(\emptyset_{(q,v)}.\theta_v)}$$

طبقه‌بندی کننده‌های غیرخطی

روش‌های دیگر برای انتخاب عمودی از طبقه‌بندهای غیرخطی استفاده شده است که قادر به استفاده از تعامل ویژگی هستند. (Jaime Arguello ,۲۰۱۷)

از الگوریتم GBTD استفاده شده است. مؤلفه اصلی مدل GBDT یک درخت رگرسیون است. یک درخت رگرسیون یک درخت باینری ساده است. هر گره داخلی به یک ویژگی و یک وضعیت تقسیم که داده‌ها را تقسیم می‌کند ، مطابقت دارد. (Jaime Arguello ,۲۰۱۷)

هر گره پایانه با یک مقدار پاسخ ، مقدار خروجی پیش‌بینی شده مطابقت دارد. (Jaime Arguello) ,۲۰۱۷

GBDT درختان رگرسیون را در یک چارچوب تقویت‌کننده ترکیب می‌کند تا یک مدل پیچیده‌تر شکل بگیرد. در طول آموزش ، هر درخت رگرسیون اضافی روی بقایای پیش‌بینی فعلی آموزش داده می‌شود. طبقه بندی کننده‌های غیرخطی مانند GBDT دارای مزایا و معایبی هستند. مزیت اصلی این است که آنها می‌توانند از تعامل پیچیده بین ویژگی‌ها استفاده کنند. (Jaime Arguello ,۲۰۱۷)

با این وجود ، این الگوریتم می‌تواند روی داده‌های اموزشی هم تمرین کند و آزمایش شود. اگرچه ممکن است در ابتدا غیرشهودی به نظر برسد ، یک مدل انعطاف پذیرتر که قادر به طبقه‌بندی کامل داده‌های آموزش است ممکن است کمتر بتواند به خوبی به داده‌های جدید تعمیم داده شود. (Jaime Arguello) ,۲۰۱۷

انتخاب بین یک طبقه‌بندی کننده خطی یا غیرخطی ممکن است به عوامل مختلفی چون اندازه داده‌های آموزش بستگی داشته باشد (Jaime Arguello ,۲۰۱۷)

مدلهای تطبیقی

ارتباط عمودی‌ها با یک پرس‌وجو احتمالاً با گذشت زمان تغییر می‌کند. این خصوصاً برای موارد عمودی که روی رویدادهای اخیر مانند اخبار متمرکز شده‌اند ، اتفاق می‌افتد. (Jaime Arguello ,۲۰۱۷)

به‌عنوان‌مثال ، پرس‌وجوی "بوستون" ممکن است یک سؤال از اخبار باشد که در طی برخی مقاطع زمانی (هنگامی که یک رویداد مهم اتفاق می‌افتد) باشد ، اما در موارد دیگر نیست. در حالت ایده‌آل ، ما

یک سیستم انتخاب عمودی را می‌خواهیم که بتواند با تغییرات در خواست‌های کاربران سازگار باشد. ((Jaime Arguello ,۲۰۱۷

۱-۷-نمایش(ارایه) عمودی‌ها

هدف از ارائه عمودی این است که تصمیم بگیرید که مکان‌های عمودی انتخاب شده را نسبت به نتایج وب و یکدیگر ارائه دهید. به طور کلی ، ارائه عمودی به دلایل مختلف دشوارتر از انتخاب عمودی است. در مرحله اول ، اگر فرضیه درجه‌بندی را فرض کنیم ، هدف سیستم ارائه نتایج‌عمودی یا وب مرتبط‌تر به روشی برجسته‌تر است. در عمل ، این ترجمه می‌شود به ارائه نتایج مرتبط‌تر با بالاترین درجه در SERP. (Jaime Arguello ,۲۰۱۷)

بنابراین ، سیستم باید میزان ارتباط یک عمودی به یک پرس‌وجو را پیش‌بینی کند. دوم ، سیستم باید در تصمیم‌گیری در مورد ارائه کدام عمودی انتخابی ، عوامل مختلفی را در نظر بگیرد. حداقل ، سیستم باید هدف عمودی پرس‌وجو و همچنین کیفیت نتایج برگشت یافته توسط عمودی را در نظر بگیرد. به عنوان مثال ، در حالی که جستجوی "خرید iphone "به طور واضح قصد عمودی خرید را دارد ، یک سیستم ممکن است تصمیم بگیرد اگر نتایج ضعیف باشد نتایج خرید را کمتر در SERP ارائه دهد. در حقیقت ، در بعضی موارد ، یک سیستم حتی ممکن است این امکان را داشته باشد که با توجه به شواهد پس از بازیابی ، یک عمودی که قبلاً انتخاب شده است را نمایش ندهد. سرانجام ، سیستم های ارائه عمودی باید با این واقعیت مقابله کنند که عمودی‌های مختلف با سطوح مختلفی از بینایی(بصری بودن) مرتبط هستند. (Jaime Arguello ,۲۰۱۷)

بنابراین ، برای مثال ، نمایش تصاویر غیرمرتبط در وسط SERP ممکن است تأثیر منفی بیشتری نسبت به نمایش نتایج خبری غیرمرتبط در همان موقعیت داشته باشد. روش‌های پیشنهادی قبلی برای ارائه عمودی را می‌توان به دو نوع مختلف طبقه‌بندی کرد: رویکردهای نقطه(pointwise) و دو جهت (pairwise) (Jaime Arguello ,۲۰۱۷)

در هر دو حالت ، هر v عمودی انتخاب‌شده مربوط به یک بلوک است - دنباله ای از نتایج t_v که باید در SERP ارائه شود. در سیستم‌های فعلی ، برخی از عمودی‌ها (به عنوان مثال ، اخبار) نتایج را به صورت عمودی درون بلوک ترتیب می‌دهند ، در حالی که سایر عمودی‌ها (به عنوان مثال ، تصاویر) نتایج را به صورت افقی سازمان دهی می‌کنند. (Jaime Arguello ,۲۰۱۷)

۱-۷-۱-روش نقطه ای (pointwise)

رویکردهای نقطه‌ای به طور مستقیم میزان ارتباط هر بلوک عمودی به یک پرس‌وجو را پیش‌بینی می‌کنند. از این نظر ، رویکردهای نقطه جهت ارائه عمودی می‌تواند بسیار شبیه روش‌های انتخاب عمودی باشد. (Jaime Arguello ,۲۰۱۷)

ما می‌توانیم طبقه بندیگرهای مستقل و مستقیمی را آموزش دهیم تا میزان ارتباط یک عمودی به یک پرس‌وجو را پیش‌بینی کنیم و از مقادیر اطمینان پیش‌بینی از طبقه بندی کننده های مختلف استفاده کنیم تا تصمیم بگیریم که در آن هر یک از عمودی‌های انتخاب شده را ارائه دهیم. (Jaime Arguello) ,۲۰۱۷

با آموزش طبقه بندیگرهای مستقل ، می‌توانیم که هر طبقه بندی از نمایه‌ای از ویژگی‌های متفاوت استفاده کند و رابطه عمودی ویژه بین مقادیر ویژگی و ارتباط عمودی مربوطه را بیاموزد. چندین رویکرد نقطه‌ای که در کار قبلی مورد بررسی قرار گرفته‌اند فرض می‌کنند که بلوک‌های عمودی فقط می‌توانند در اسلات‌های خاص در نتایج وب ارائه شوند. (Jaime Arguello ,۲۰۱۷)

۱-۷-۲-روش زوجی (Pairwise)

رویکردهای Pairwise یاد می گیرند که نسبیت نسبی بین جفت بلوک کاندیداها را که در SERP نمایش داده می‌شود ، پیش‌بینی کنند. یک رویکرد زوجی پیشنهاد شده که به شرح زیر است. (Jaime) Arguello ,۲۰۱۷

بگذارید Bq مجموعه‌ای از بلوک‌های عمودی و بلوک‌های وب را نشان دهد که در پاسخ به سؤال q نمایش داده می‌شود. علاوه بر این ، فرض کنید بلوک‌های عمودی فقط در اسلات‌های خاص در SERP نمایش داده می‌شوند. (Jaime Arguello ,۲۰۱۷)

اگر چهار شکاف تصویر شده در شکل شماره ۲ را فرض کنیم ، Bq شامل یک بلوک برای هر عمودی انتخاب شده در پاسخ به پرس‌وجو q و سه بلوک وب است که همیشه نمایش داده می‌شوند (با عنوان w۱ ، w۲ و w۳ نشان داده می شوند). (Jaime Arguello ,۲۰۱۷)

شکل شماره ۲: روش زوجی(Jaime Arguello ,۲۰۱۷)

آموزش یک طبقه‌بندی باینری برای هر جفت نوع بلوک بود. در اینجا ، یک نوع بلوک به سیستم جستجوی تولیدکننده بلوک (به عنوان مثال ، عمودی خاص یا موتور جستجوی وب) اشاره دارد. (Jaime Arguello ,۲۰۱۷)

اگر ما N تا عمودی مختلف را فرض کنیم، سپس ما می‌توانیم $\binom{n}{۲}$+۲ طبقه بندی کننده مختلف را آموزش دهیم. (Jaime Arguello ,۲۰۱۷)

اصطلاح اول مربوط به طبقه‌بندی کننده‌های آموزش‌دیده برای پیش‌بینی ارتباط نسبی بین بلوک‌ها از دو عمودی متفاوت است ، در حالی که اصطلاح دوم مربوط به طبقه‌بندی کننده‌های آموزش دیده برای پیش‌بینی ارتباط نسبی بین بلوک‌ها از یک عمودی خاص و یک بلوک وب است. (Jaime Arguello ,۲۰۱۷

هر طبقه‌بندی می‌تواند از ویژگی‌های خاص خود استفاده کند ، که می‌توان آن را به عنوان اتصال بین دو بردار ویژگی تصور کرد: این ویژگی‌ها برای نوع بلوک اول پیش‌بینی می‌شوند و ویژگی‌هایی که برای نوع بلوک دوم پیش‌بینی می‌شوند. (Jaime Arguello ,۲۰۱۷)

۱-۸- روشهای یادگیری رتبه بندی

در یادگیری‌ماشین ، الگوریتم‌های یادگیری رتبه‌بندی (LTR) یاد می‌گیرند که موارد را به‌عنوان تابعی از مجموعه‌ای از ویژگی‌ها مرتب کند. در زمینه بازیابی اطلاعات ، از الگوریتم‌های LTR بیشتر برای رتبه‌بندی اسناد در پاسخ به یک سؤال استفاده شده است. در این حالت ویژگی‌های پیش‌بینی به طور معمول از جفت پرس‌وجو و سند (مستقل از پرس‌وجو) تولید می‌شوند. روشهای LTR موجود می‌توانند به

سه نوع طبقه‌بندی شوند. روش‌های نقطه‌ای(به عنوان مثال ، درختان تصمیم‌گیری توسعه یافته) یاد می گیرند که درجه ارتباط یک سند مستقل را از اسناد دیگرپیش‌بینی کنند. (Jaime Arguello ,۲۰۱۷)

روش های pair-wise (به عنوان مثال ، RankSVM) یاد می‌گیرند که آیا یک سند نسبت به سند دیگر مرتبط است یا خیر. (Jaime Arguello ,۲۰۱۷)

روش هایListwise (به عنوان مثال ، AdaRank) به طور مستقیم اندازه‌گیری ارزیابی IR مانند NDCG را بهینه می‌کند ، که کیفیت رتبه بندی را به عنوان یک کل در نظر می‌گیرد. روشهای LTR همچنین برای سایر کارهای IR مانند رتبه بندی پرس‌وجوهای پیشنهادی ، رتبه بندی پرس‌وجوهای کاندید ، و رتبه‌بندی مقالات مربوط به اخبار برای یک مقاله ورودی استفاده شده است. استفاده از LTR برای نمایش عمودی دو چالش اصلی را ایجاد می‌کند. (Jaime Arguello ,۲۰۱۷)

رویکردهای LTR نیاز به ویژگی مشترک بازنمایی دارند. در زمینه ارائه عمودی ، ممکن است بلوک‌های خاصی از ویژگی‌های خاصی در دسترس نباشند. به عنوان مثال ، برخی از عمودی‌ها نمرات بازیابی را ارائه نمی‌دهند. اگر می‌خواهیم از نمرات بازیابی عمودی برای تولید ویژگی‌های پس از بازیابی استفاده کنیم ، این ویژگی‌ها برای برخی از انواع بلوک در دسترس نیست. دوم ، برخی از رویکردهای LTR یک رابطه پیش‌بینی کننده ثابت بین ویژگی‌ها و ارتباط یک مورد فرض می‌کنند. به طور خاص ، این مورد در مورد مدل‌های خطی LTR صادق است. (Jaime Arguello ,۲۰۱۷)

۱-۹-ارزیابی

ارزیابی برای همه زیرمجموعه‌های بازیابی اطلاعات بسیار مهم است ، و همین مسئله در مورد جستجوی‌کل نیز صادق است. ارزیابی ، مقایسه هدف بین منابع مختلف شواهد را برای پیش‌بینی ارتباط عمودی ، الگوریتم‌های مختلف برای ترکیب منابع شواهد ، و تنظیمات پارامترهای مختلف برای یک سیستم خاص تسهیل می‌کند. (Jaime Arguello ,۲۰۱۷)

همانطور که قبلاً ذکر شد ، جستجوی تجمعی شامل دو کار زیر است: (Jaime Arguello ,۲۰۱۷)

۱- پیش‌بینی اینکه کدام عمودها در پاسخ به پرس‌وجو نمایش داده شود (انتخاب عمودی) (Jaime Arguello ,۲۰۱۷

۲- پیش‌بینی اینکه عمودی‌های انتخاب شده در کجای وب و چگونه نمایش داده شوند(ارایه عمودی) (Jaime Arguello ,۲۰۱۷)

انتخاب عمودی شامل پیش‌بینی اینکه کدام عمودی برای ارائه و کدام عمودی برای سرکوب است. (Jaime Arguello ,۲۰۱۷)

ارائه عمودی شامل حل مغایرت بین عمودی‌های مختلف انتخاب شده و ارائه مناسب‌ترین عمودی‌ها به روشی برجسته‌تر است. در عمل ، این به طور معمول به معنای نمایش مناسب‌ترین عمودی بالاتر در SERPجمع شده است. در این حالت ، ارزیابی بر توانایی سیستم برای پیش‌بینی اینکه کدام عمودها مربوط به یک پرس‌وجو هستند و کدام عمودی‌ها نیست ، متمرکز است. (Jaime Arguello ,۲۰۱۷)

۱-۹-۱-ارزیابی انتخاب عمودی

هدف از انتخاب عمودی این است که پیش‌بینی کند که کدام عمودها مربوط به یک پرس‌وجو هستند. با توجه به پرس‌وجو ، یک سیستم انتخاب عمودی پیش‌بینی باینری را برای هر عمودی کاندید ایجاد می‌کند: عمودی انتخاب شود یا خیر. (Jaime Arguello ,۲۰۱۷)

یک سیستم انتخاب عمودی خوب ، سیستم‌هایی است که تمام عمودی‌های مربوطه را به درستی انتخاب می‌کند و همه موارد غیر مرتبط را به درستی سرکوب می‌کند. (Jaime Arguello ,۲۰۱۷)

ما معیارهای ارزیابی انتخاب عمودی را با استفاده از نماد زیر مرور می‌کنیم. بگذارید Q مجموعه ای از پرس‌وجوهای ارزیابی را مشخص کند و V مجموعه ای از عمودی‌های کاندید را مشخص کند. همانطور که اغلب در IR اتفاق می افتد ، ما به طور معمول به عملکرد متوسط ، یا با میانگین سؤالات در Q یا عمودی در V اهمیت می دهیم. (Jaime Arguello ,۲۰۱۷)

برای تسهیل هر دو گزینه ، اجازه دهید Qv مجموعه ای از پرس‌وجوهای ارزیابی را نشان دهد که برای آنها vertical v مرتبط است و $Qv^{\sim}$ مجموعه ای از درخواستهای ارزیابی را نشان می‌دهد که سیستم برای آن پیش‌بینی کرده است که v مرتبط باشد. به همین ترتیب ، اجازه دهید Vq مجموعه ای از عمودی‌ها را که مربوط به query q و $V^{\sim}q$ است از مجموعه ای از عمودی که پیش‌بینی سیستم مربوط به q است نشان دهد. (Jaime Arguello ,۲۰۱۷)

اساساً ، انتخاب عمودی یک مسئله طبقه‌بندی چندکلاسه است. بنابراین ، تمام معیارهایی که مربوط به طبقه‌بندی چندکلاسی هستند نیز برای انتخاب عمودی کاربرد دارند. یک متریک ارزیابی که به طور گسترده مورد استفاده قرار می‌گیرد در طبقه بندی چندکلاس است. دو نوع پیش‌بینی درست وجود دارد که یک سیستم انتخاب عمودی می‌تواند در پاسخ به یک سؤال ایجاد کند. سیستم می‌تواند به درستی پیش‌بینی کند که یک عمودی خاص مرتبط است (یک پیش‌بینی مثبت صحیح) یا به درستی پیش‌بینی کند که یک عمودی خاص مرتبط نیست (یک پیش‌بینی منفی صحیح). (Jaime Arguello ,۲۰۱۷)

در زمینه انتخاب عمودی ، دقت درصد پیش‌بینی‌های صحیح مثبت و صحیح منفی صحیح را در تمام نمایش داده‌ها و عمودی‌ها اندازه گیری می‌کند: (Jaime Arguello ,۲۰۱۷)

$$\mathrm{A} = \frac{۱}{|Q| \times |V|} \sum_{q \in Q} \sum_{v \in V} I(v \in V_q \wedge v \in V'_q) \vee I(v \notin V_q \wedge v \notin V'_q)$$

مؤلفه اول یک پیش‌بینی مثبت صحیح را با توجه به پرس‌وجو q و عمودی v نشان می‌دهد ، و مؤلفه دوم پیش‌بینی منفی صحیح را با توجه به q و v بیان می‌کند. دقت دو اشکال اصلی دارد. (Jaime Arguello ,۲۰۱۷)

اشکال اول این است که دقت ، با طراحی ، انواع خطاهای ایجاد شده را مخفی می‌کند. در بعضی موارد، ممکن است بدانیم که آیا سیستم پیش‌بینی های مربوط به عمودی مثبت کاذب یا منفی کاذب را بیشتر می‌کند یا خیر. (Jaime Arguello ,۲۰۱۷)

اشکال دوم این است که تفسیر مقادیر دقت ممکن است دشوار باشد. برای مثال ، سیستمی که هر عمودی را برای هر پرس‌وجو انتخاب می‌کند (یا هر عمودی را برای هر پرس‌وجو سرکوب می‌کند) تقریباً مطمئناً یک مقدار دقت بیشتر از صفر خواهد داشت. در حقیقت ، با توجه به یک پرس‌وجو ، فقط چند عمودی (در صورت وجود) مرتبط هستند. بنابراین ، سیستمی که هر عمودی را برای هر پرس‌وجو سرکوب می‌کند ، می‌تواند به یک مقدار دقت بالایی دست یابد. (Jaime Arguello ,۲۰۱۷)

دقت(Precision) و صحت(recall) پرس‌وجوها در حالت عمودی بصورت زیر محاسبه می‌شوند: (Jaime Arguello ,۲۰۱۷)

$$P_Q = \frac{1}{|Q|}\sum_{q\in Q}\frac{|V_q \cap v'_q|}{|V'_q|}$$

$$R_Q = \frac{1}{|Q|}\sum_{q\in Q}\frac{|V_q \cap v'_q|}{|V'_q|}$$

در این حالت ، برای پرس‌وجو داده شده q ، دقت(precision) توانایی سیستم برای رد کردن عمودی‌های غیرمرتبط از مجموعه پیش‌بینی شده را اندازه گیری می‌کند ، در حالی که صحت(Recall) توانایی سیستم را برای درج عمودی‌های مربوطه در مجموعه پیش‌بینی شده اندازه گیری می‌کند. (Jaime Arguello ,۲۰۱۷

دقت(Precision) و صحت(recall) میانگین کلان در سطح عمودی بصورت زیر محاسبه می‌شوند: (Jaime Arguello ,۲۰۱۷)

$$\sum_{v\in V}\frac{|Q_q \cap Q'_q|}{|Q'_q|}$$

$$R_Q = \frac{1}{|V|}\sum_{v\in V}\frac{|Q_q \cap Q'_q|}{|Q'_q|}$$

در این حالت ، برای یک v عمودی مشخص ، دقت(precision) توانایی سیستم در سرکوب یک عمودی در هنگام عدم اندازه گیری ، اندازه گیری می‌شود ، در حالی که صحت(Recall) ، توانایی سیستم در انتخاب یک عمودی را در زمان مناسب اندازه گیری می‌کند.

در بعضی موارد ممکن است یک متریک واحد بخواهیم که تعادل بین دقت و فراخوان را اندازه گیری کند. در این حالت ، اندازه گیری f معادل میانگین هارمونیک دقت(pricision) و صحت(Recall) است: (Jaime Arguello ,۲۰۱۷)

$$F_* = \frac{2 \times P_* \times R_*}{P_* + R_*}$$

۱-۱۰-منحنی Precision-Recall

طبقه بندی‌گرهای یادگیری ماشین معمولاً علاوه بر تصمیم دودویی ، مقدار اطمینان پیش‌بینی را نیز به همراه می‌آورند. در چنین مواردی می‌توان یک پارامتر آستانه τ را معرفی کرد. ایده اصلی این است که سیستم فقط در صورتی عمودی v را به عنوان پاسخ پرس‌وجوی q برمی‌گرداند که مقدار پیش‌بینی اطمینان آن بیشتر از آستانه باشد. (Jaime Arguello ,۲۰۱۷)

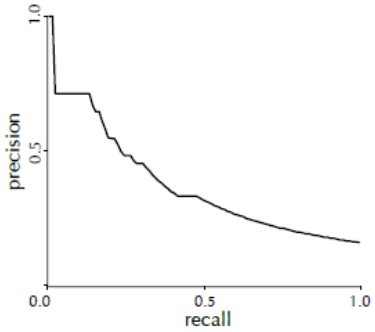
پارامتر τ را می‌توان تنظیم کرد تا دقت بیشتری در مورد recall(صحت) یا برعکس انجام شود. اگر فرض کنیم که مقادیر اطمینان طبقه‌بندی شده در دامنه [۰،۱] است ، با مقادیر بالاتر نشان دهنده اطمینان بالاتری است که v به q مربوط می‌شود ، پس می‌توانیم برای بالا بردن دقت نسبت به صحت(recall) (τ=۰,۹۰) تنظیم کنیم. ، یا می‌توانیم برای کم کردن صحت نسبت به دقت (τ=۰,۱۰) را تنظیم کنیم . (Jaime Arguello ,۲۰۱۷)

پارامتر τ از دو طریق می‌تواند در فرایند ارزیابی عمودی وارد شود. یکی از گزینه‌های دیگر تنظیم پارامتر τ با استفاده از یک مجموعه اعتبارسنجی است. این فرایند شامل سه مرحله است: (۱) ارزیابی مقادیر مختلف τ استفاده از یک مجموعه اعتبار سنجی ، (۲) انتخاب مقدار τ با بهترین عملکرد از نظر برخی از معیارهای انتخابی (۳) ارزیابی سیستم با بهترین مقدار پارامتر در مجموعه آزمون برگزار شده.

منحنی precision-recall(یا منحنی PR) گرافیکی است که دقت (در محور y) را به عنوان تابعی از صحت (در محور x) تجسم می‌کند. منحنی PR، یک تصویر کامل تر از مبادله سیستم بین دقت و صحت ارائه می‌دهد. (Jaime Arguello ,۲۰۱۷)

با توجه به دو سیستم رقیب ، مورد ایده آل این است که یک سیستم برای دستیابی به مقادیر بالاتر دقت نسبت به همه مقادیر صحت اقدام کند. (Jaime Arguello ,۲۰۱۷)

در این حالت ، غیرقابل تردید است که سیستم با مساحت بیشتر زیر منحنی PR بهتر باشد. از طرف دیگر ، می‌توانیم روی مقادیر دقیق مرتبط با سطح صحت(Recall) که فکر می‌کنیم برای کاربران مهمتر است تمرکز کنیم. اگر فکر می‌کنیم که کاربران معمولاً می خواهند هر عمودی را که مرتبط است ، ببینند ، می‌توانیم روی مقادیر دقیق مرتبط با سطح recall توجه کنیم. (Jaime Arguello ,۲۰۱۷)



شکل شماره۳: نمودار P-R(۲۰۱۷, Jaime Arguello)

۱-۱۱-ضرورت اجرا

از آنجا که جویشگرهای فارسی هنوز در ابتدای راه هستند و دارای ضعف‌هایی هستند و به قدرتمندی جویشگرهایی همچون گوگل نیستند؛ بنابراین اگر بتوانیم از مزایای همه آنها در کنار همدیگر استفاده کنیم می‌توانیم نتایج بهتر و سودمندتری بدست آوریم؛ پس ما قصد داریم که پرسش کاربر را به چندین جویشگر فارسی مانند پارسی‌جو و یوز و جس‌جو بدهیم و بعد از بین نتایج بازیابی شده بهترین نتایج‌ها و مرتبط‌ترین آنها را جدا کرده و باهم ترکیب کنیم.

از مهم‌ترین مزایای تولید یک فراجویشگر بومی که زیرساخت‌های آن نیز جویشگرهای بومی باشند می‌توان به موارد زیر اشاره کرد:

آسان‌سازی استفاده از اینترنت برای کاربران فارسی‌زبان

تسهیل در دسترسی به اطلاعات مناسب

توسعه کسب‌وکار در فضای اینترنت

مقابله با تهدیدهای ناشی از تحریم و بحران‌های سیاسی

جمع آوری و در دسترس قراردادن محتوای متناسب با فرهنگ اسلامی و زبان فارسی

محدود نمودن دسترسی به اطلاعاتی که در تقابل با فرهنگ ایرانی و اسلامی می‌باشد

کاهش هزینه‌های پهنای‌باند

کاهش درز اطلاعات کاربران و سازمان‌های ایرانی به خارج از مرزهای کشور

ایجاد اعتماد و فرهنگ مناسب درخصوص کاربردهای فنآوری اطلاعات

۱-۱۲-سؤالات تحقیق

۱-۱۲-۱-سؤالات اصلی

در حوزه زبان‌فارسی:

برای پرسش کاربر از بین جویشگرهای موجود کدام جویشگرها مناسب‌تر و مرتبط‌تر هستند ؟

از بین اسناد بازیابی شده از جویشگرها کدام اسناد مهم‌تر و مرتبط‌تر هستند؟

با چه الگوریتم‌ها و چه شیوه‌هایی می‌توان این نتایج بازیابی شده را ترکیب و به کاربر نشان داد؟

آیا می‌توان کارآیی این فراجویشگرها را به نسبت کارهای انجام شده پیشین بهبود داد؟

کدام الگوریتم و روش یادگیری ماشین میتواند دقت و کیفیت فراجویشگرها را بهبود بخشد؟

۱-۱۲-۲-سؤالات فرعی

برای طراحی این فراجویشگر بهتر است از جویشگرهای عمومی استفاده کنیم یا خصوصی یا ترکیبی از آنها؟

مقیاس‌پذیری فراجویشگر ما به چه صورت باشد؟

به چه طریق می‌توانیم ادغام بهینه‌ای داشته باشیم؟

۱-۱۳-فرضیات

جویشگرهای پایه که به عنوان پایگاه‌داده ها استفاده می‌شوند نتایج خوب و مفیدی را برای ما تولید می‌کنند.

جویشگرهای پایه دارای دقت کافی هستند.

جویشگرهای پایه دارای زبان پاسخگویی مناسب و سریعی هستند.

جویشگرهای پایه پوشش مناسب و کافی دارد که صفحات وب را پوشش می‌دهد.

۱-۱۴-جنبه‌های نوآورانه‌ی پژوهش

مدل‌سازی رابطه بین مراحل مختلف فراجویشگر برای ارزیابی جهت مهمی برای تحقیقات آینده است.

استفاده از جویشگرهای بومی برای ایجاد فراجویشگر مدنظر

استفاده از روش‌های یادگیری‌ماشین و ترکیب‌اطلاعات

استفاده از خوشه‌بندی برای سوالات کاربر و جواب‌های مرتبط

مروری بر آخرین الگوریتم‌ها و روش‌ها به منظور استفاده در حوزه فراجویشگرها

۱-۱۵-اهداف پژوهش

۱-۱۵-۱-اهداف علمی

ایجاد یک فراجویشگر بومی که هدف آن ارتقای سطح کیفیت و بهبود عملکرد جستجو با رویکرد یادگیری‌ماشین و بکارگیری روش‌های ترکیب اطلاعات، در جهت رفع و کاهش مشکلات فراجویشگرهای فعلی و افزایش کیفیت جستجو از جمله اهداف این تحقیق به شمار می‌روند.

۱-۱۵-۲-اهداف نظری

استخراج الگوریتم مفید و تعیین میزان اثربخشی رویکرد یادگیری‌ماشین و خوشه‌بندی در فراجویشگر بومی

۱-۱۵-۳-اهداف کاربردی

هدف کاربردی طراحی و پیاده سازی یک فراجویشگر بومی است که بتواند سوال کاربر را دریافت کرده و آنرا به جویشگرهای زیرین مرتبط با سوال کاربر بفرستد و از بین نتایج بازیابی شده، بهترین‌ها و مرتبط‌ترین آنها را انتخاب کرده و با روش‌های مناسبی ترکیب کند و به کاربر نشان دهد.

و دیگر اهداف کاربردی:

تقویت وارتقاء جستجوپذیری محتویات فارسی برروی شبکه

فراهم کردن اعتبار ملی، منطقه‌ای و بین‌المللی

توسعه‌ی کسب‌وکارهای مبتنی برفناوری‌اطلاعات

توسعه‌ی خدمات بومی و ملی مبتنی برفناوری‌اطلاعات

فرآهم شدن محتوی وکاربردهای بومی مبتنی برفناوری‌اطلاعات

استقلال ملی درفضای سایبری

توسعه‌ی دانش وفناوری بومی

بهره‌برداری‌های اطلاعاتی و امنیتی

کسب درآمد و رسیدن به سهم مناسبی از بازار بین‌المللی

رقابت با سرویس‌های خارجی

تأثیر در افزایش سطح رضایت‌مندی شهروندان از خدمات‌الکترونیک

محدود نمودن دسترسی به اطلاعاتی که در تقابل با فرهنگ ایرانی و اسلامی می‌باشد

تولید ایده‌های اقتصادی مناسب و بلندمدت

حجم، مقیاس و گستردگی جستجو

ارائه آمار و تحلیل‌های آماری

امکان جایگزینی در صورت تحریم بین‌المللی

مجهز به الگوریتم مناسب برای پوشش هرچه بیشتر سایت‌های فارسی

تطبیق مناسب با فرهنگ و خط و زبان فارسی

ابعاد اقتصادی:

اطلاع‌رسانی و تبلیغ مناسب در جهت توسعه نام تجاری شناخته شده

توسعه بسترهایی جهت تبلیغات مناسب برای درآمدزایی

توسعه بسترهای مالی مناسب جهت جذب سرمایه‌گذار

فصل دوم

اصول و نظریه بازیابی اطلاعات و ترکیب اطلاعات

۲-۱-اصول و مبانی نطریه بازیابی متن:

بازیابی متن (اطلاعات) با مشکل یافتن اسناد مرتبط (مفید) برای هر گونه پرس و جو از مجموعه اسناد متنی سروکار دارد. فناوری بازیابی متن تأثیر عمیق و مستقیمی بر موتورهای جستجوی وب دارد. در واقع ، موتورهای جستجوی نسل اول (حدود ۱۹۹۵-۱۹۹۷) تقریباً کاملاً بر اساس فناوری بازیابی متن سنتی ساخته شده بودند که در آن صفحات وب به عنوان اسناد متنی مورد بررسی قرار می گرفتند.

در این بخش ، ما یک مرور کلی از برخی مفاهیم اساسی در بازیابی متن کلاسیک ارائه می دهیم .نمای کلی در درجه اول بر اساس مدل فضای بردار است که در آن اسناد و پرسش های کاربر به عنوان بردار اصطلاحات با وزن نشان داده می شوند.

۲-۱-۱-معماری سیستم

معماری یک سیستم بازیابی متن اصلی در شکل زیر نشان داده شده است اسناد موجود در مجموعه اسناد یک سیستم بازیابی متن برای شناسایی اصطلاحات نماینده هر سند ، برای جمع آوری آمار خاصی از شرایط و سازماندهی اطلاعات در قالب خاصی که تسهیل کننده است ، جهت محاسبه سریع شباهت هر سند با توجه به هر پرس و جو پیش پردازش شده است.(Salton and McGill, ۱۹۸۳)

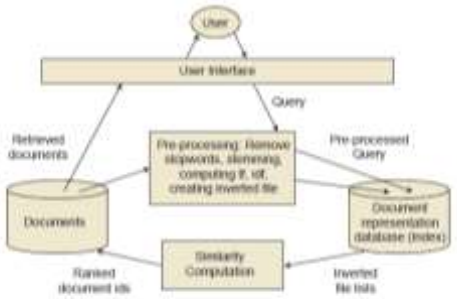


Figure 1.1: Architecture of a Basic Text Retrieval System.

هنگامی که درخواست کاربر دریافت می شود ، سیستم بازیابی متن پرس و جو را با شناسایی اصطلاحات نمایان کننده پرس و جو و سپس محاسبه وزن عبارات ، که اهمیت اصطلاحات را در نمایش محتوای پرس و جو نشان می دهد ، پردازش می کند.(Croft et al., ۲۰۰۹)

سپس سیستم با استفاده از شاخص از پیش ساخته شده ، شباهت اسناد را با پرس و جو محاسبه می کند و اسناد را به ترتیب نزولی از شباهت های آنها رتبه بندی می کند .جزئیات بیشتر در مورد این مفاهیم و عملکردها در بخشهای فرعی بعدی ارائه خواهد شد.(Croft et al., ۲۰۰۹)

۲-۱-۲-نمایش اسناد

محتویات یک سند ممکن است با کلمات موجود در آن نشان داده شود .برخی از کلمات مانند "a" ، "of" و "is" حاوی اطلاعات محتوای موضوعی نیستند .این کلمات را کلمات توقف می نامند و اغلب استفاده نمی شوند؛ تغییرات یک کلمه ممکن است در یک اصطلاح ترسیم شود .به عنوان مثال ، کلمات " compute" ، "computing " و "computation " را می توان با عبارت "comput " نشان داد .این را می توان با یک برنامه بنیادی ، که پسوندها را حذف می کند یا کاراکترهای دیگر را جایگزین آنها می کند ، به دست آورد. (Baeza-Yates and Ribeiro-Neto, ۱۹۹۹)

پس از حذف کلمات توقف و ریشه ، هر سند را می توان به صورت منطقی با بردار n اصطلاح نشان داد ، در صورتی که n تعداد کل اصطلاحات مجزا در مجموعه همه اسناد در مجموعه اسناد است، فرض کنید سند d را با بردار (d۱,d۲,…dn) نشان میدهیم که di نشاندهنده وزن اصطلاح iام در نمایش محتویات سند است.اگر یک اصطلاح در سند نباشد پس وزن آن در بردار صفر نمایش داده میشود؛ هنگامی که یک اصطلاح درسند d وجود دارد ، وزن آن معمولاً بر اساس دو عامل فرکانس اصطلاح (tf) و ضریب فرکانس سند.(df) نمایش داده میشود. (Baeza-Yates and Ribeiro-Neto, ۱۹۹۹)

tfیک عبارت در یک سند تعداد دفعاتی است که این عبارت در سند ظاهر می شود. از نظر بصری ، هرچه tf یک عبارت بیشتر باشد ، این واژه از اهمیت بیشتری برخوردار است. در نتیجه ، اصطلاح وزن فرکانس (tfw) یک عبارت در یک سند معمولاً یک تابع افزایش دهنده یکنواخت tf آن است df .یک عبارت عبارت است از تعداد اسناد و مدارک موجود در کل مجموعه اسناد که حاوی آن عبارت هستند. معمولاً هرچه df یک عبارت بیشتر باشد ، اهمیت واژه در تمایز اسناد مختلف کمتر است؛ بنابراین ، وزن

یک عبارت با توجه به df معمولاً یک تابع کاهش یکنواخت df آن است و وزن فرکانس سند معکوس (idfw)نامیده می شود. (Baeza-Yates and Ribeiro-Neto, ۱۹۹۹)

وزن یک عبارت در یک سند می تواند حاصل ضرب وزن فرکانس آن اصطلاح در وزن فرکانس معکوس آن باشد ، یعنیtfw * idfw ؛ وزن یک اصطلاح در یک سند ممکن است تحت تأثیر عوامل دیگری قرار گیرد مانند مواردی که در سند آمده است .به عنوان مثال ، اگر عبارت در عنوان سند ظاهر شود ، وزن ممکن است افزایش یابد؛ یک پرس و جو معمولی برای بازیابی متن نیز بصورت متن نوشته می شود. بنابراین ، می توان آن را مانند یک سند تلقی کرد و با استفاده از روشی که در بالا توضیح داده شد به یک بردار n بعدی تبدیل کرد. (Baeza-Yates and Ribeiro-Neto, ۱۹۹۹)

۲-۱-۳-تطبیق سند-پرس و جو

بعد از اینکه همه اسناد و یک پرس و جو به عنوان بردارهای یک بعد نشان داده شد ، می توان شباهت بین بردار پرس و جو و هر بردار سند را با استفاده از یک تابع شباهت محاسبه کرد. سپس اسنادی که بردارهای متناظر آنها شباهت زیادی با بردار پرس و جو دارند ، بازیابی می شوند؛ اجازه دهید q=(q۱,q۲,…,qn) بردار پرس و جو و d=(d۱,d۲,…,dn) بردار سند باشد؛ یک تابع شباهت ساده تابع ضرب نقطه ای(داخلی) است که به صورت زیر نمایش داده میشود. (Salton and McGill, ۱۹۸۳)

$$\text{dot (q,d)}=\sum_{i=۱}^{n} q_i * d_i$$

این تابع به ما اسنادی رابرمیگرداند که بیشترین شباهت را به اصطلاحات پرس و جوی کاربر دارند. یکی از مشکلات این تابع شباهت ساده این است که به نفع اسناد طولانی است ، زیرا به احتمال زیاد آنها احتمال اینکه اصطلاحات موجود در پرس و جوی کاربر را داشته باشند بیشتر است.یکی از راه های غلبه بر مشکل فوق استفاده از تابع کسینوسی به شکل زیر است: (Salton and McGill, ۱۹۸۳)

$$\text{Cos(q,d)}=\frac{\sum_{i=۱}^{n} q_i * d_i}{\sqrt{\sum_{i=۱}^{n} q_i^{۲}} * \sqrt{\sum_{i=۱}^{n} d_i^{۲}}}$$

تابع کسینوس دو بردار یک تفسیر هندسی دارد - این کسینوس زاویه بین دو بردار است .به عبارت دیگر ، تابع Cos فاصله زاویه ای بین بردار پرس و جو و بردار سند را اندازه گیری می کند. هنگامی که بردارها دارای وزن های منفی هستند ، تابع کسینوس همیشه مقداری را در [۱ ، ۰] برمی گرداند. مقدار تابع زمانی ۰ است که هیچگونه شباهتی بین بردار پرس و جو و بردار سند وجود نداشته باشد(یعنی زاویه ۹۰

درجه) و مقدار تابع زمانی ۱ است که بردارها یکسان باشند(یعنی دو بردار دارای زاویه ۰ باشند). (Robertson and Sparck Jones, ۱۹۷۶; Yu and Salton, ۱۹۷۶)

توابع شباهت دیگری نیز وجود دارد و برخی از آنها مجاورت اصطلاحات پرس و جو را در یک سند در نظر می گیرند. برای پشتیبانی از مطابقت مبتنی بر مجاورت ، برای هر سند و اصطلاح مشخص ، موقعیت های عبارت در سند باید جمع آوری شده و به عنوان بخشی از فهرست جستجو ذخیره شود. چندین مدل بازیابی متن دیگر نیز وجود دارد .در مدل بازیابی اولیه بولی ، اسناد بر اساس اینکه آیا عبارتهای پرس و جو را در خود دارند یا نه ، بازیابی می شوند و وزن اصطلاحات در نظر گرفته نمی شود. (Robertson and Sparck Jones, ۱۹۷۶; Yu and Salton, ۱۹۷۶)

یک پرس و جو بولی می تواند شامل یک یا چند عملگر بولی (AND، ORو NOT)باشد. در مدل احتمالی ، اسناد به ترتیب نزولی احتمال داده می شوند که یک سند مربوط به یک پرس و جو باشد؛ یکی از پرکاربردترین توابع شباهت بر اساس مدل احتمالی ، تابع Okapi است. در این رویکرد ، برای یک پرس و جو معین ، ما احتمال می دهیم که پرس و جو بر اساس هر سند ایجاد شود و سپس اسناد را به ترتیب نزولی احتمالات رتبه بندی کنیم. (Robertson and Sparck Jones, ۱۹۷۶; Yu and Salton, ۱۹۷۶)

۲-۱-۴-ارزیابی پرس و جو

محاسبه شباهت بین یک پرس و جو و هر سند به طور مستقیم ناکارآمد است زیرا اکثر اسناد هیچ اصطلاح مشترکی با یک پرس و جو ندارند و محاسبه شباهت های این اسناد اتلاف منابع است. برای بهبود کارایی ، یک فهرست پرونده معکوس از قبل ایجاد می شود. برای هر اصطلاح ti یک لیست وارونه در قالب زیر با یک هدر تولید و ذخیره میشوند.(Turtle and Flood, ۱۹۹۵)

i)… I(ti)=[(, $D_{i_1}, w_{i_1} i$).. (, $D_{i_k}, w_{i_k} i$)]

این در حالی است که D_{i_j}, نشاندهنده (شناسه) سند حاوی ti است و w_{i_j} وزن اصطلاح ti در سند D_{i_j} است که $1 \leq j \leq k$ است و k تعدا اسناد حاوی اصطلاح ti است. علاوه بر این ، یک جدول هش ، که یک ساختار داده شبیه به جدول است ، برای ترسیم هر عبارت پرس و جو در سربرگ لیست معکوس این عبارت استفاده می شود. فایل معکوس و جدول هش امکان محاسبه کارآمد شباهت های همه اسنادی را که شباهت های غیر صفر با هر پرس و جو دارند ، فراهم می کند.بطور خاص یک پرس و جو با m اصطلاح را درنظر بگیرید.(Turtle and Flood, ۱۹۹۵)

برای هر عبارت پرس و جو ، از جدول هش برای تعیین لیست معکوس این عبارت استفاده می شود. فهرستهای معکوس اساساً شامل تمام اطلاعات مورد نیاز برای محاسبه شباهتهای بین پرس و جو و همه اسناد حاوی حداقل یک عبارت پرس و جو است؛ یک استراتژی ارزیابی پرس و جو که به طور گسترده مورد استفاده قرار می گیرد ، استراتژی سند در زمان است که شباهت یک سند را در هر زمان محاسبه می کند و فقط اسنادی که حداقل یک عبارت پرس و جو را شامل می شود در نظر گرفته می شود. ایده اصلی این استراتژی به شرح زیر است.(Turtle and Flood, ۱۹۹۵)

در بسیاری از سیستم های بازیابی متن ، فایل معکوس بسیار بزرگ است و نمی تواند در حافظه اصلی نگهداری شود و بنابراین بر روی دیسک ذخیره می شود. اگر فایل معکوس روی دیسک باشد ، در ابتدای ارزیابی یک پرس و جو ، لیست های وارونه از همه عبارت های پرس و جو ابتدا به حافظه اصلی آورده می شوند. سپس شباهت اسناد ، که هر کدام حداقل دارای یک عبارت پرس و جو هستند ، محاسبه می شود (یک سند در یک زمان). (Turtle and Flood, ۱۹۹۵)

فرض کنید که یک پرس و جو شامل m اصطلاح است. هر اصطلاح مربوط به یک لیست معکوس است که در آن اسناد حاوی عبارت دارای شناسه های خود به ترتیب صعودی هستند؛ شباهت با انجام ادغام-m wayاز این لیست های معکوس محاسبه می شود که در مثال زیر زیر نشان داده شده است. با توجه به اسکن همزمان لیستهای معکوس ، یک اسکن از هر یک از لیستهای معکوس واژه های پرس و جو برای ارزیابی پرس و جو کافی است. .(Turtle and Flood, ۱۹۹۵)

مثال۱) شکل زیر ماتریس اصطلاح سند را برای یک سند نمونه از مجموعه ای از اسنادبرای سادگی ، وزنها فرکانسهای اصطلاح خام هستند و تابع محصول نقطه به عنوان تابع شباهت در اینجا استفاده می شود. با پنج سند و پنج عبارت مجزا نشان می دهد. (Turtle and Flood, ۱۹۹۵)

	t_1	t_2	t_3	t_4	t_5
D_1	2	1	1	0	0
D_2	0	2	1	1	0
D_3	1	0	1	1	0
D_4	2	1	2	2	0
D_5	0	2	0	1	2

از ماتریس در شکل فوق ما میتوانیم لیست معکوس زیر را به دست آوریم: .(Turtle and Flood, ۱۹۹۵)

I (t۱) = [(D۱, ۲), (D۳, ۱), (D۴, ۲)],

I (t۲) = [(D۱, ۱), (D۲, ۲), (D۴, ۱), (D۵, ۲)],

I (t۳) = [(D۱, ۱), (D۲, ۱), (D۳, ۱), (D۴, ۲)],

I (t۴) = [(D۲, ۱), (D۳, ۱), (D۴, ۲), (D۵, ۱)],

I (t۵) = [(D۵, ۲)].

که q یک پرس و جو با دو اصطلاح t۱ و t۳ است که وزن هر دو ۱ است یعنی هرکدام یکبار ظاهر شده اند.(Turtle and Flood, ۱۹۹۵)

ما در حال حاضر از استراتژی سند در زمان استفاده می کنیم تا شباهت اسناد را با توجه به q محاسبه کنیم. ابتدا لیست فایل های وارونه را برای دو عبارت پرس و جو در حافظه اصلی واکشی می کنیم.سپس عبارت زیر واکشی میشود: .(Turtle and Flood, ۱۹۹۵)

I (t۱) = [(D۱, ۲), (D۳, ۱), (D۴, ۲)] and I (t۳) = [(D۱, ۱), (D۲, ۱), (D۳, ۱), (D۴, ۲)]

اولین سند در هر دو لیست D۱ است(یعنی D۱ شامل هر دو عبارت t۱ و t۳ است و شباهت آن را با پرس و جو می توان از وزن ۲ برای t۱ و ۱ برای t۳ و پرس و جو با استفاده ضرب نقطه ای بصورت زیر محاسبه کرد: (Turtle and Flood, ۱۹۹۵)

dot(q,D۱) = ۱ ∗ ۲ + ۱ ∗ ۱ = ۳

موارد بعدی در دو لیست عبارتند (D۳,۱) ، (D۲و۱) ، هنگامیکه D۳> D۲ (از نظر شناسه سند)می توان تشخیص داد که D۲ حاوی t۱ نیست. بنابراین ، شباهت D۲ را می توان تنها بر اساس(D۲,۱) , dot(q, ۱ = ۱ ∗ ۱ = (D۲ محاسبه کرد؛ پس از پردازش(D۲, ۱) مورد بعدی در I (t۳) ، یعنی (D۳, ۱) در نظر گرفته می شود که منجر به محاسبه شباهت D۳ با استفاده از اطلاعات I (t۱) و I (t۳) بصورت زیر می شود.(Turtle and Flood, ۱۹۹۵)

dot(q,D۳) = ۱ ∗ ۱ + ۱ ∗ ۱ = ۲.

به طور مشابه ، شباهت D۴ را می توان به عنوان dot(q,D۴) = ۱ * ۲ + ۱ * ۲ = ۴ محاسبه کرد. از آنجا که(D۴, ۲) آخرین مورد در هر دو لیست است ، فرآیند محاسبه شباهت به پایان می رسد. (Turtle and Flood, ۱۹۹۵)

۲-۱-۵-اقدامات اثربخشی بازیابی

هدف از بازیابی متن ، یافتن اسناد مرتبط (مفید) برای شخصی است که پرس و جو را ارسال می کند و این اسناد را در لیست نتایج در بالا قرار می دهد. اثربخشی بازیابی یک سیستم بازیابی متن اغلب با یک جفت مقدار معروف به فراخوانی(recall) و دقت(precision) اندازه گیری می شود. فرض کنید ، برای یک پرسش کاربر مشخص ، مجموعه ای از اسناد مربوطه در مجموعه اسناد شناخته شده است. سپس فراخوان(recall) نسبت اسناد مربوطه بازیابی شده و دقت(precision) نسبت مرتبط بودن اسناد بازیابی شده مربوطه است. (Voorhees and Harman ,۲۰۰۵)

به عنوان مثال ، فرض کنید ۱۰ سند مرتبط به یک پرس و جو از بین ۲۰ سند بازیابی شده وجود دارد و ۶ مورد مربوط است .سپس برای این پرس و جو ، فراخوانی ۱۰/۶ = ۰,۶ و دقت ۲۰/۶ = ۰,۳ است؛ برای ارزیابی اثربخشی یک سیستم بازیابی متن ، اغلب از مجموعه پرسش های آزمایشی استفاده می شود. برای هر پرس و جو ، مجموعه ای از اسناد مربوطه از قبل مشخص شده است .برای هر پرس و جو آزمایشی ، یک مقدار دقیق برای هر مقدار فراخوانی مجزا بدست می آید. معمولاً فقط یازده مقدار فراخوانی ، ۰,۰ ,۰,۱ ,... ,۱,۰ در نظر گرفته می شود. (Voorhees and Harman ,۲۰۰۵)

هنگامی که مقادیر دقت در هر مقدار فراخوانی در تمام پرس و جوهای آزمایشی به طور متوسط باشد ، منحنی دقت فراخوانی متوسط بدست می آید؛ همچنین بسیاری از اقدامات موثر بازیابی دیگر برای سیستم های بازیابی متن وجود دارد. در زمینه موتورهای جستجو ، غالباً نمی توان مجموعه کاملی از اسناد مربوطه را برای پرسش های آزمایشی دانست. (Voorhees and Harman ,۲۰۰۵)

۲-۲-اصول و مبانی نظریه ترکیب اطلاعات:

ادغام داده ها در وب به ادغام لیست اسناد رتبه بندی شده در یک لیست واحد ، اشاره می کند که در پاسخ به درخواست کاربر توسط بیش از یک موتور جستجوی وب بازیابی می شوند. در این قسمت ، تکنیک های ادغام معرفی می شوند که نه تنها موقعیت های رتبه ، بلکه عنوان و خلاصه اسناد بازیابی شده را نیز در نظر می گیرد. آزمایشات ارزیابی ما نشان می دهد که تکنیک های ادغام فوق باعث بهبود

اثربخشی می شوند و اثربخشی آنها با رویکردی که لیست های رتبه بندی شده را با بارگیری و تجزیه و تحلیل اسناد وب ادغام می کند ، قابل مقایسه است. .(Lawrence, et al,۱۹۹۸)

ظهور شبکه جهانی وب با انفجار مقدار اطلاعاتی که به راحتی در دسترس هستند همراه شد. ابزارهای غالب جستجو در اینترنت استفاده از موتورهای جستجو است که سیستم های بازیابی اطلاعات مبتنی بر پرس و جو (IR) هستند که اسناد وب را فهرست بندی و بازیابی می کنند. این تکنیک های ادغام موقعیت اسناد بازیابی شده ، عنوان آنها و خلاصه مختصر همراه آنها را در نظر خواهد گرفت، اگرچه برخی از توابع ادغام از اسناد سند در محاسبات خود استفاده می کنند ، اما به دلایل زیر این نمرات با تکنیک های ادغام ما در نظر گرفته نمی شوند. .(Lawrence, et al,۱۹۹۸)

ابتدا ، نمرات سند توسط موتورهای جستجوی شرکت کننده رتبه بندی می شوند و آنها معمولاً با استفاده از مدل های مختلف IR محاسبه می شوند و بنابراین نمی توانند به طور مستقیم مقایسه شوند، ثانیاً ، حتی اگر دو موتور جستجو از یک مدل IR مشابه استفاده کنند ، نمرات سند همچنان قابل مقایسه نیست ، زیرا در محاسبه آنها ، آماری (tf & idf) وابسته به مجموعه اسناد وب نمایه شده توسط هر موتور جستجو وجود دارد. بخش زیر تکنیک های ادغام ما را معرفی می کند. .(Lawrence, et al,۱۹۹۸)

روش ۱ فقط رتبه اسناد را در نظر می گیرد ؛ روش ۲ عنوان و خلاصه آنها ؛ روش ۳ عنوان و خلاصه آنها و مدل سازی فرآیند تلفیق داده ها با استفاده از نظریه شواهد .Dempster-Shafer روش ۴ موقعیت های رتبه ، عنوان و خلاصه اسناد بازیابی شده را در نظر می گیرد و در نهایت روش ۵ با بارگیری و نمایه سازی اسناد وب لیست های رتبه بندی شده ادغام شده را تولید می کند .این روشهای ادغام در ادامه به تفصیل شرح داده شده است. .(Lawrence, et al,۱۹۹۸)

روش ۱: ادغام با استفاده از موقعیت های رتبه بندی

ساده ترین روشی که می توان در فرآیند ادغام لیست های جداگانه اسناد وب بازیابی شده به کار برد ، این روش است که فقط موقعیت اسناد را در نظر می گیرد. این روش به طور ضمنی اطلاعات مربوط به محتوای سند را در بر می گیرد ، زیرا موقعیت های رتبه توسط موتورهای جستجو که سند را بازیابی می کنند تعیین می شوند و ترتیب بر اساس شرایط نمایه سند تعیین می شود. در این روش ، اسناد تکراری به طور خلاصه رتبه بندی می شوند و بقیه اسناد به هم متصل می شوند .بنابراین ، روش ۱ اسناد بازیابی

شده توسط بیش از یک موتور جستجو را ترجیح می دهد .این روش ساده است ، زیرا بر حداقل اطلاعات ارائه شده متکی است و به عنوان پایه ای برای رویکرد تجربی ما عمل می کند. .(Lawrence, et al,۱۹۹۸)

روش ۲: ادغام با استفاده از عنوان و خلاصه اسناد بازیابی شده.

این روش عنوان و خلاصه اسناد بازیابی شده را در نظر می گیرد و سعی می کند بررسی کند که آیا می توان با نمایه سازی اسناد بازیابی شده با استفاده از آنها ، اثربخشی را بهبود بخشید. عنوان سند وب همانطور که در لیست اسناد بازیابی شده ارائه شده توسط هر یک از موتورهای جستجو شرکت کننده عنوان عنوان صفحه واقعی وب است که این سند با آن مطابقت دارد. خلاصه توسط موتور جستجویی که سند را بازیابی می کند ایجاد می شود و معمولاً شامل تکه هایی از سند است که شامل شرایط پرس و جو در پاسخ به این سند است. بنابراین ، این عناصر متنی اطلاعات مربوط به محتوای سند را در بر می گیرند و می توانند به عنوان نمایندگی آن مورد استفاده قرار گیرند ، زیرا متن کامل اسناد بازیابی شده مستقیماً توسط موتور فرا جویشگر نمایه نمی شود. .(Lawrence, et al,۱۹۹۸)

برای این روش ، یک مجموعه واحد از اسناد بازیابی تشکیل می شود که شامل تمام اسناد وب در لیست ارائه شده توسط موتورهای جستجوی شرکت کننده است. سپس می توان اسناد موجود در این مجموعه را با استفاده از اصطلاحات موجود در عنوانها و خلاصه فهرست بندی کرد و به عنوان بردارهای این اصطلاحات نمایشی ارائه کرد. وزنها به این اصطلاحات اختصاص داده می شود تا مشخص شود کدامیک از متمایز کننده های خوب محتوای این سند وب هستند. طرح وزن دهی متداول [I] tf x idf را نمی توان در اینجا به کار برد ، زیرا اصطلاحاتی که ما به عنوان تمایز دهنده خوب سند وب در نظر می گیریم ، عبارت های پرس و جو هستند و به احتمال زیاد در اکثر اسناد وجود دارند ، زیرا در پاسخ به این پرس و جو بازیابی شد. در نتیجه ، idf (t)، زمانی که t = عبارت پرس و جو ، به احتمال زیاد برابر با ۰ باشد. (Lawrence, et al,۱۹۹۸).

به جای آن از روش وزن دهی متفاوتی برای هر اصطلاح t در سند d استفاده می شود. اسناد به صورت d $\{w_{1,d} \ . \ . \ w_{k,d}\}$=نمایش داده می شوند ، جایی که wi وزن اصطلاح شاخصi ام در سند است که فقط با استفاده از فرکانس اصطلاح آن tf (t.d) محاسبه می شود. سپس اسناد با استفاده از شباهت نمایش اسناد و پرس و جو Q دوباره مرتب می شوند. و تابع شباهت مورد استفاده بصورت زیر است: (Lawrence, et al,۱۹۹۸).

$$\sum_{t \in Q} w(t,d)$$

اسناد تکراری دارای URL یکسانی هستند ، اما خلاصه های متفاوتی دارند زیرا این خلاصه ها توسط موتورهای جستجو که آنها را بازیابی کرده اند ، ایجاد می شود. برای هر سند ، خلاصه موارد تکراری آن ذخیره و به هم متصل می شوند. به این ترتیب ، اسناد تکراری با طولانی ترین خلاصه ها در بین اسناد بازیابی شده مرتبط هستند و مکانیزم تنظیم مجدد آنها را مطلوب می کند ، زیرا هیچ نرمالسازی سازی در عملکرد رتبه بندی اعمال نمی شود. این نشان می دهد که مدارک بازیابی شده توسط بیش از یک موتور جستجو باید رتبه بالاتری داشته باشند. (Lawrence, et al,۱۹۹۸)

روش ۳: ادغام با استفاده از نظریه شواهدDempster –Shafer

یک رویکرد رسمی را می توان با تعریف یک الگوریتم ادغام ، که عنوان و خلاصه اسناد بازیابی شده را در نظر می گیرد و بر اساس نظریه شواهد Dempster –Shafer است ، معرفی کرد. این نظریه امکان نمایش صریح عدم دقت ، جهل و ترکیبی از شواهد را فراهم می آورد و به عنوان ابزاری برای ادغام عدم قطعیت هنگام ادغام لیست های جداگانه اسناد بازیابی شده توسط موتورهای جستجوی شرکت کننده مورد استفاده قرار می گیرد.(Lawrence, et al,۱۹۹۸)

نظریه شواهد Dempster –Shafer ، بسط نظریه احتمالات است و امکان نمایش صریح عدم قطعیت و ترکیب شواهد را فراهم می آورد. ترکیبی از شواهد به عنوان یک ویژگی اساسی توسط قانون ترکیبی Dempster، که اجازه تجمیع را می دهد ، ثبت شده است. تجمیع مفهوم اساسی زیرین فرایند ادغام است و در نتیجه این نظریه امکان مدل سازی هر دو نمایندگی از اسناد بازیابی شده و خود استراتژی ادغام را فراهم می آورد. .(Lawrence, et al,۱۹۹۸)

با توجه به چارچوب Dempster-Shafer ، مجموعه ای که دامنه همه ارزشهای ممکن را که گزاره ها می توانند نشان دهند ، نشان می دهد ، قاب تشخیص نامیده می شود. سپس گزاره ها به عنوان زیر مجموعه های این مجموعه نمایش داده می شوند .(Lawrence, et al,۱۹۹۸)

باورها را می توان به گزاره هایی اختصاص داد تا عدم قطعیت خود را بیان کنند. باورها معمولاً بر اساس تابع چگالی m محاسبه می شوند: $m:\delta(U)\rightarrow[۰،۱]$ ،که به آن تخصیص احتمال ابتدایی (bpa) گفته می شود ، به گونه ای که: (Lawrence, et al,۱۹۹۸)

$M(\cdot)=\cdot , \sum_{A\subseteq U} m(A) = ۱$

m (A) دقیقاً به A متعهد است ، این شواهد معتقد است که مقدار u در A است .اگر شواهد مثبتی برای مقدار u در A وجود داشته باشد ، $m(A)> \cdot$ و A را کانونی می نامند .عناصر کانونی و bpas مربوط به آنها مجموعه ای از شواهد را مشخص می کند.(Lawrence, et al,۱۹۹۸)

این نظریه در واقع تعمیمي از نظریه احتمال در مدلسازي عدم قطعیت موجود در پدیدههاي فیزیکي میباشد .در نظریه احتمال کلاسیک، احتمال هر پیشامد به صورت حاصل جمع احتمالات جزئي تشکیلدهنده آن پیشامد در نظر گرفته میشود .در این نوع مدلسازي، نایقیني بصورت جهل نسبت به وقوع یک پیشامد ویا عدم وقوع یک پیشامد تعریف میشود که توصیف چندان دقیقي از آن ارائه نمیدهد .اما در نظریه دمپستر -شفر مدلسازي نایقیني بگونهاي است که جهل ما نسبت به پیشامدها نیز مورد انداز هگیري قرار میگیرد و کمیتی که معادل میزان احتمال وقوع، معرفی میشود، خود دارای عدم قطعیت و حد بالا و پائین است.(Lawrence, et al,۱۹۹۸)

$M:۲^{\theta}\rightarrow[۰،۱]$ $\quad \forall x \in ۲^{\theta}, \cdot \leq m(x) \leq ۱$

$M(\phi) = \cdot$ $\quad \sum_{x\in ۲^{\theta}} m(x) = ۱$

برای هر پیشامد x که زیر مجموعه از θ باشد،مقادیر اعتقاد و تعمق به عنوان حدود بالا و پایین برای میزان وقوع این پیشامد به صورت زیر تعریف میشوند که در آن $\bar{x}$ مکمل مجموعه x است: (Lawrence, et al,۱۹۹۸).

$Bel(x) = \sum_{y\subseteq x} m(y)$

$Pls(x) = ۱ - Bel(\bar{x})$

نهایتاً از قانون ترکیب دمپستر جهت ترکیب داده به اینصورت استفاده میشود که فرض کنید که دو تخصیص پایه $m_۱$ و $m_۲$ متناظر با دو منبع مستقل اطلاعاتي داریم که ترکیب این دو یک تخصیص پایه جدید بنام $m=m_۱\oplus m_۲$ را تشکیل میدهد که به اینصورت محاسبه میگردد: .(Lawrence, et al,۱۹۹۸)

$M(z)=k\times \sum_{x_1\cap x_2=z} m_1(x_1)m_2(x_2)$

$K=(1-\sum_{x_1\cap x_2=\emptyset} m_1(x_1)m_2(x_2))^{-1}$

ترکیب با معادلات بالا تنها در صورتی امکان پذیر است که چارچوب تمایز برای دو پایه m۱و m_2

یکسان باشند ،در غیر این صورت برای بدست آوردن یک چارچوب تمایز واحد، باید از نگاشتهای

مربوطه استفاده کرد که خارج از بحث حاضراست .در چنین مواردی پیچیدگی‌های محاسباتی قانون ترکیب دمپستر یک معضل اساسی در فرآیند ترکیب داده محسوب میشود. .(Lawrence, et al,۱۹۹۸)

روش ۴: ادغام با استفاده از موقعیت های رتبه ، عنوان و خلاصه اسناد بازیابی شده.

تا کنون ، روش ۱ از رتبه بندی اسناد استفاده می کند ، در حالی که روش ۲ و روش ۳ از اطلاعات موجود در عنوان و خلاصه اسناد بازیابی شده استفاده می کنند. اگر فهرستهای رتبه بندی ایجاد شده توسط این روشها به عنوان ورودی روش ۱ ارائه شوند ، می توانند بیشتر با هم ترکیب شوند. بنابراین ، با ادغام لیست های ایجاد شده توسط روش ۱ و روش ۲ ، عملیات همجوشی اطلاعات بیشتری را در نظر می گیرد. همین امر در مورد لیست روش ۱ و روش ۳ نیز صادق است. با معرفی این روش ، هدف ما بررسی این موضوع است که آیا ترکیب اطلاعات بیشتری که توسط موتورهای جستجو بازگردانده می شود (موقعیت های رتبه ، عناوین ، خلاصه ها) منجر به بهبود اثربخشی می شود یا خیر. (Lawrence, et al,۱۹۹۸).

روش ۵: ادغام با بارگیری اسناد وب.

تمام روشهای ادغام که در بخشهای قبلی مورد بحث قرار گرفت ، به اطلاعات ارائه شده توسط موتورهای جستجوی کمک کننده که اسناد وب را بازیابی می کنند ، بدون روشهای دسترسی به متن کامل این اسناد ، متکی است. با این حال ، اگر صفحات وب واقعی بارگیری و تجزیه و تحلیل شوند ، اعتقاد بر این است که رتبه بندی لیست نهایی ادغام شده می تواند بهبود یابد. این رویکرد بر اساس این واقعیت است که از آنجا که کل محتوای سند در دسترس است ، عملکرد ادغام می تواند این اطلاعات را در نظر گرفته و یک لیست رتبه بندی ادغام شده موثرتر ایجاد کند .علاوه بر این ، بارگیری صفحات منجر به شناسایی صفحاتی می شود که دیگر وجود ندارند. اشکال این روش این است که نیاز به بارگیری و تجزیه و تحلیل

اسناد وب در زمان واقعی دارد. در نتیجه به پهنای باند اضافی نیاز دارد ، تقاضای بیشتری برای عملکرد رایانه دارد و زمان پرس و جو را افزایش می دهد. با این حال ، موتور جستجوی Inquirus نشان داد که تجزیه و تحلیل زمان واقعی اسناد بازگشتی از موتورهای جستجوی وب امکان پذیر است و بنابراین می توان از این روش استفاده کرد.(Lawrence, et al,۱۹۹۸)

این الگوریتم ادغام مشابه روش ۲ است با این تفاوت که از متن کامل سند ، به جای خلاصه آن ، برای نمایش محتوای آن استفاده می شود. با این حال ، هنگامی که یک سند تکراری شناسایی شد ، تنها یک نسخه از آن ذخیره می شود. الگوریتم های رتبه بندی توصیف شده در روش ۲ ، بجای یک تابع رتبه بندی IR پیچیده تر استفاده می شوند ، به طوری که نتایج قابل مقایسه با روش های دیگر است. روش ۵ به عنوان پایه عمل می کند و مقایسه اثربخشی آن با سایر روشهای ادغام منجر به نتیجه گیری در مورد اینکه آیا در نظر گرفتن تنها اطلاعات ارائه شده توسط موتورهای جستجو کافی است ، بدون نیاز به دسترسی به خود اسناد وب ، می شود. .(Lawrence, et al,۱۹۹۸)

فصل سوم

ابرموتورهای جستجوی وب

۳-۱-مقدمه:

یک فراجویشگر یک سیستم جستجو است که راهی یکپارچه برای دسترسی به چندین موتور جستجوی موجود ارائه می دهد.فراجویشگرها با بازیابی اطلاعات توزیع شده ارتباط تنگاتنگی دارند. (Craswell, N., ۲۰۰۰)

مفهوم فراجویشگرها در وب از اوایل دهه ۱۹۹۰ مطرح شد .یکی از اولین ، اگر نه قدیمی ترین ، موتورهای فراجویشگر MetaCrawler(http://www.metacrawler.com/) است که برای اولین بار در سال ۱۹۹۴ توسعه یافت. از آن زمان ، تعداد زیادی از موتورهای فراجویشگر توسعه یافته و مورد استفاده قرار می گیرند.

۳-۲-موتورهای جستجوی وب:

قدیمی ترین موتورهای جستجوی وب اساساً سیستم های بازیابی متن برای صفحات وب بودند. با این حال ، محیط وب دارای برخی ویژگی های خاص است که ساخت موتورهای جستجوی مدرن را به طور قابل توجهی متفاوت از ساخت سیستم های بازیابی متن سنتی می کند.

۳-۲-۱-مشخصات ویژه وب

در زیر برخی از ویژگی های اصلی محیط وب آورده شده است که تأثیر بسزایی در توسعه موتور جستجو دارد(Craswell, N., ۲۰۰۰) .

۱-صفحات وب در تعداد زیادی از سرورهای وب مستقل ذخیره می شوند .برای یافتن آنها و واکشی آنها به روشی نیاز است تا بتوان آنها را برای جستجوی بعدی پردازش کرد(Craswell, N., ۲۰۰۰) .

۲-بیشتر صفحات وب در قالب HTML (HyperText Markup Language) هستند و برچسب های HTMLاغلب اطلاعات غنی را در مورد اصطلاحات موجود در این صفحات منتقل می کنند. به عنوان

مثال ، اصطلاحی که در عنوان یک صفحه ظاهر می شود یا اصطلاحی که با قلم خاص برجسته می شود ، می تواند این نکته را به شما نشان دهد که این اصطلاح در نمایش محتوای صفحه مهم است (Craswell, N., ۲۰۰۰) .

۳-صفحات وب به یکدیگر پیوند خورده اند. پیوندی از صفحه P۱ به صفحه P۲ به کاربر وب اجازه می دهد تا از صفحه P۱ به صفحه P۲ حرکت کند. چنین پیوندی همچنین شامل چندین قطعه اطلاعات است که برای بهبود اثربخشی بازیابی مفید است. اول ، این پیوند احتمال همبستگی مطالب دو صفحه را نشان می دهد. ثانیا ، نویسنده صفحه P۱ صفحه P۲ را با ارزش می داند .سوم ، متن قابل کلیک مرتبط با پیوند ، به نام anchor text of the link ، معمولاً توضیح کوتاهی از صفحه پیوند داده شده ارائه می دهد . (Craswell, N., ۲۰۰۰)

دو نوع موتور جستجو داریم:۱-موتور جستجو با هدف عمومی که قابلیت جستجو در تمام صفحات وب را فراهم میکنند. ۲- موتور جستجو با هدف اختصاصی که تمرکز بر جستجوی اسناد در یک سازمان یا موضوع خاصی را دارند.(Meng et al., ۲۰۰۲)

۳-۲-۲-خزنده وب:

خزنده وب یک برنامه برای واکشی صفحات وب از سرورهای وب از راه دور است که طور گسترده ای برای ایجاد مجموعه صفحات وب برای یک موتور جستجو استفاده می شوند. خزنده در وبسایت های موجود میگردد و محتوای آنها را در موتور جستجو ذخیره میکند و موقعی که در یک صفحه مشغول خزش ست لینک های موجود در آن صفحه را نیز وارد شده و جستجو میکند.سپس خزنده دو مرحله زیر را تکرار می کند تا زمانی که یا نشانی اینترنتی جدیدی پیدا نشود یا صفحات کافی واکشی نشده باشد: (Meng et al., ۲۰۰۲)

۱) نشانی اینترنتی بعدی را از لیست URL ها بردارید ، اتصال به سروری را که صفحه وب در آن قرار دارد ، برقرار کنید و با صدور درخواست HTTP (پروتکل انتقال ابرمتن) به سرور ، صفحه وب مربوطه را از سرور خود دریافت کنید(Meng et al., ۲۰۰۲) .

۲)URL های جدید را از هر صفحه وب واکشی شده استخراج کرده و اضافه کنید آنها را در لیست قرار دهید یک خزنده ممکن است صفحات وب را یا عرض اول یا عمق اول را واکشی کند. با عرض اولیه

خزیدن ، فهرست URL ها به صورت صف اجرا می شوند URL-های جدید همیشه در انتهای لیست اضافه می شوند .با خزیدن عمیق ، لیست URL به عنوان یک پشته پیاده سازی می شود URL-های جدید همیشه در ابتدای لیست اضافه می شوند(Meng et al., ۲۰۰۲) .

اغلب صفحات وب صفحات HTML هستند که شامل تگ های زیادی هستند. ما میتوانیم از فراوانی اصطلاح و فرکانس سند یک عبارت برای محاسبه وزن اصطلاح در یک سند استفاده می کند. ما همچنین می توانیم از اطلاعات برچسب برای تأثیر وزن یک اصطلاح استفاده کنیم. (Meng et al., ۲۰۰۲)

۳)سازماندهی نتایج:اکثر موتورهای جستجو نتایج بازیابی شده را به ترتیب نزولی مطلوبیت برآورد شده خود، نسبت به یک پرس و جو معین می کنند. مطلوبیت یک صفحه برای یک پرس و جو می تواند به روش های مختلف مانند شباهت صفحه با پرس و جو یا اندازه گیری ترکیبی از جمله شباهت و رتبه صفحه باشد.(Laender and Ribeiro-Neto, ۲۰۰۲; Chang et al., ۲۰۰۶) .

یک موضوع مربوط به نحوه نمایش رکورد نتایج جستجو (SRR) ، که مربوط به یک صفحه وب بازیابی شده است ، در یک صفحه نتیجه است.که نتایج بازیابی شده و نمایش داده شده در صفحه نتایج معمولا شامل سه بخش هستند..(Laender and Ribeiro-Neto, ۲۰۰۲; Chang et al., ۲۰۰٦)

عنوان صفحه وب ، آدرس صفحه وب ، و خلاصه ای از صفحه وب سایر اطلاعات مانند زمان انتشار و اندازه صفحه وب نیز توسط برخی از موتورهای جستجو در SRR ها گنجانده شده است. و یک متن که شامل حدودا ۲۰ کلمه است که توصیفی از صفحه وب را میدهد. (Laender and Ribeiro-Neto, ۲۰۰۲; Chang et al., ۲۰۰۶).

برخی از مسائلی که هنگام پیاده سازی الگوریتم خوشه بندی آنلاین نتایج باید مورد توجه قرار گیرد شامل موارد زیر است.(Laender and Ribeiro-Neto, ۲۰۰۲; Chang et al., ۲۰۰۶)

۱)از چه اطلاعاتی (عناوین ، نشانی های اینترنتی ، قطعات در مقابل اسناد کامل) برای انجام خوشه بندی باید استفاده کرد؟ در حالی که اطلاعات بیشتر ممکن است کیفیت خوشه ها را بهبود بخشد ، استفاده بیش از حد از اطلاعات ممکن است به دلیل محاسبه بالا و ارتباطات زیاد باعث تأخیر طولانی کاربران شود.(Laender and Ribeiro-Neto, ۲۰۰۲; Chang et al., ۲۰۰۶) .

۲)برای انجام خوشه بندی از چه معیارهایی باید استفاده کرد؟ این می تواند بر اساس شباهت بین SSR ها باشد ، یعنی نتایج بسیار مشابه باید با هم گروه بندی شوند .همچنین می تواند بر اساس تفسیر پرس و جو باشد ، یعنی نتایج مطابقت با تفسیر یکسان باید با هم گروه بندی شوند(Laender and Ribeiro-Neto, ۲۰۰۲; Chang et al., ۲۰۰۶).

(۳)چگونه می توان یک برچسب کوتاه و در عین حال معنی دار برای هر گروه تهیه کرد؟ (Laender and Ribeiro-Neto, ۲۰۰۲; Chang et al., ۲۰۰۶).

(۴)چگونه می توان گروه ها را سازماندهی کرد؟ آنها می توانند به صورت خطی یا به صورت سلسله مراتبی مرتب شوند .در مورد قبلی ، ترتیب خطی باید چگونه باشد؟ در حالت دوم ، چگونه می توان سلسله مراتب را ایجاد کرد؟ برخی از مسائل هنوز به طور فعال در حال تحقیق هستند (Laender and Ribeiro-Neto, ۲۰۰۲; Chang et al., ۲۰۰۶).

۳-۳- دلایل پیدایش فراجویشگرها:

در این بخش ، ما سعی می کنیم یک تجزیه و تحلیل جامع از مزایای بالقوه موتورهای فراجویشگر نسبت به موتورهای جستجو ارائه دهیم و بر مقایسه موتورهای فراجویشگر و موتورهای جستجو تمرکز می کنیم .(Meng et al., ۲۰۰۲)

-افزایش پوشش جستجو: با فراهم شدن دسترسی یکپارچه به موتورهای جستجوی زیرین یک فراجویشگر میتواند هر سندی را که حداقل توسط یک موتورجستجوی زیرین نمایه شده است را بازگردانی کند.از اینرو پوشش فراجویشگرها خیلی بهتر از موتورهای جستجوی جزیی است. این مزیت ، انگیزه اصلی پشت فراجویشگرهای اولیه است و این هنوز شناخته شده ترین مزیت آنها است.عبارت "افزایش پوشش جستجو" برای این دو رویکرد تا حدودی معانی متفاوتی دارد .برای اولی ، می توان از دو جنبه به آن نگاه کرد(Meng et al., ۲۰۰۲) .

اول ، این امر به طور گسترده ای پذیرفته شده است و با شواهدی قوی پشتیبانی می شود که موتورهای جستجوی اصلی مختلف مجموعه های مختلف صفحات وب را فهرست بندی می کنند ، هر چند که همه آنها سعی می کنند کل وب را فهرست بندی کنند. این بدان معناست که یک فراجویشگر با چندین موتور جستجوی اصلی به عنوان اجزاء پوشش وسیع تری نسبت به هر موتور جستجوی تک جزئی

دارد.ثانیاً ، موتورهای جستجوی مختلف اغلب از روش های مختلف ارائه اسناد و رتبه بندی نتایج استفاده می کنند ، و در نتیجه ، اغلب مجموعه های متفاوتی از نتایج برتر را برای یک پرس و جو کاربر مشابه باز می گردانند(Meng et al., ۲۰۰۲) .

-دسترسی به وب عمیق راحت تر است : وب شامل دو قسمت است؛ وب سطحی و وب عمیق ، و اندازه وب عمیق بسیار بزرگتر از وب سطحی است. موتورهای جستجو اصلی محتویات خود را با تکیه عمدتا بر خزنده های وب سنتی که صفحات وب را با پیوندهای URL دریافت می کنند ، به دست می آورند. این خزنده ها فقط می توانند به محتویات موجود در سطح وب دسترسی داشته باشند ، به این معنی که موتورهای جستجوی اصلی عمدتا وب سطحی را پوشش می دهند. در سال های اخیر ، خزنده های وب عمیق که بتوانند محتوای عمیق وب را بدست آورند در حال توسعه هستند و موفقیت هایی نیز کسب کرده اند. خزیدن در عمق وب اساساً با ارسال پرس و جو به موتورهای جستجوی عمیق وب و جمع آوری اطلاعات از نتایج برگشتی انجام می شود. محدودیت اصلی این تکنیک این است که به دست آوردن مطالب کامل از موتورهای جستجوی عمیق وب بسیار دشوار است زیرا تقریباً غیرممکن است که از تعداد معقولی از پرس و جوها برای بازیابی همه مطالب از یک موتور جستجوی وب عمیق استفاده شود. (Meng et al., ۲۰۰۲)

-کیفیت محتوای بهتر: کیفیت محتوای یک موتور جستجو را می توان با کیفیت اسناد نمایه شده توسط موتور جستجو اندازه گیری کرد کیفیت یک سند را می توان به روشهای مختلفی مانند غنای و قابلیت اطمینان محتوا اندازه گیری کرد. تجزیه و تحلیل ها بر اساس نحوه جمع آوری صفحات وب توسط موتورهای جستجوی اصلی و نحوه دسترسی موتورهای جستجو به محتوای موتورهای جستجو است . موتورهای جستجوی اصلی اسناد را از وب باز می کنند که شامل اسناد با کیفیت بالا (اسناد جدی با محتوای مفید) و اسناد بی کیفیت است زیرا همه اسناد میتوانند در وب منتشر شوند. با توجه به تعداد زیاد اسناد وب (حدود ۳۵ میلیارد برای گوگل) ، تضمین کیفیت اسناد خزیده شده برای این موتورهای جستجو بسیار دشوار است .در نتیجه ، ممکن است نتایج بد کیفیت توسط موتورهای جستجوی اصلی بازگردانده شوند. در مقابل ، موتورهای جستجوی تخصصی دارای اسناد با کیفیت بهتر هستند زیرا اغلب بر محتویات آنها کنترل بیشتری وجود داردبه عنوان مثال ، بسیاری از موتورهای جستجوی تخصصی فقط از اسنادی استفاده می کنند که از منابع معتبر برخوردار هستند ، مانند موتورهای جستجو که توسط

روزنامه ها و ناشران اداره می شود و محتواها معمولاً توسط نویسندگان حرفه ای یا نویسندگان قراردادی با کنترل ویراستاری نوشته می شوند. از آنجایی که فراجویشگرهای بزرگ اغلب فقط از موتورهای جستجوی تخصصی به عنوان موتورهای جستجوی اجزای خود استفاده می کنند ، محتویاتی که می توانند جستجو کنند نیز باید کیفیت بهتری داشته باشند (Meng et al., ۲۰۰۲) .

موتورهای جستجوی اصلی برای جمع آوری اسناد از سرورهای متعدد وب به خزنده های خود تکیه می کنند. با این حال ، این خزنده ها نمی توانند با توجه به تعداد زیاد صفحات وب و سرورهای وب درگیر ، و همچنین ماهیت متغیر در حال تغییر وب ، از محتوای سریع وب در حال تغییر باشند. به طور معمول چند روز تا چند هفته طول می کشد تا مطالب تازه به روز شده یا اضافه شده .در نتیجه ، محتویات فهرست بندی شده توسط موتورهای جستجوی اصلی معمولاً به طور متوسط چند روز قدیمی هستند.(Raghavan and Garcia-Molina, ۲۰۰۱; Madhavan et al., ۲۰۰۸)

در مقابل ، برای موتورهای جستجوی تخصصی حفظ تازگی مطالب بسیار آسان تر است زیرا از مجموعه اسناد بسیار کوچکتر استفاده می کنند و محتویات آنها اغلب در سرورهای محلی ذخیره می شود. بنابراین ، فراجویشگرها با استفاده از رویکرد موتور جستجوی با مقیاس بزرگ ، شانس بیشتری برای بازیابی اطلاعات به روزتر نسبت به موتورهای جستجوی اصلی و فراجویشگر که با موتورهای جستجوی اصلی ساخته شده اند ، دارند.(Raghavan and Garcia-Molina, ۲۰۰۱; Madhavan et al., ۲۰۰۸) .

-پتانسیل خوبی برای اثربخشی بازیابی بهتر : دو دلیل اصلی وجود دارد که یک موتور فراجویشگر با استفاده از رویکرد موتورهای جستجوی اصلی ساخته شده است تا بهتر از موتورهای جستجوی اصلی عمل می‌کند؛ به دلیل این واقعیت که موتورهای جستجوی اصلی مختلف دارای پوشش متفاوت و الگوریتم های مختلف رتبه بندی اسناد هستند ، به احتمال زیاد نتایج منحصر به فرد تری به دست می آید ، حتی در میان آنهایی که دارای رتبه بالایی هستند. (Raghavan and Garcia-Molina, ۲۰۰۱; Madhavan et al., ۲۰۰۸).

نتیجه ادغام کننده فراجویشگر میتواند با استفاده از این واقعیت که مجموعه اسناد موتورهای جستجوی اصلی دارای همپوشانی قابل توجهی هستند ، نتایج بهتری را ایجاد کند. این بدان معناست که بسیاری از اسناد به اشتراک گذاشته شده این شانس را دارند که برای موتورهای جستجوی مختلف در هر پرس و جو مشخص رتبه بندی شوند. اگر سند مشابهی توسط چندین موتور جستجوی بازیابی شود ، احتمال مرتبط

بودن سند با پرس و جو به میزان قابل توجهی افزایش می یابد زیرا شواهد بیشتری برای ارتباط آن وجود دارد. به طور کلی ، اگر یک سند توسط موتورهای جستجوی بیشتری بازیابی شود ، به احتمال زیاد سند بر اساس مشاهدات مهم زیر مرتبط خواهد بود. سیستم های مختلف جستجو تمایل دارند مجموعه ای مشابه از اسناد مربوطه اما مجموعه های مختلف اسناد بی ربط را بازیابی کنند. (Raghavan and Garcia-Molina, ۲۰۰۱; Madhavan et al., ۲۰۰۸).

اگرچه این مشاهدات بر اساس اعمال الگوریتم های رتبه بندی متفاوت برای مجموعه ای از اسناد یکسان انجام شده است ، اما وقتی مجموعه اسناد موتورهای جستجوی مختلف دارای همپوشانی بالایی هستند ، هنوز قابل استفاده است.(Meng et al., ۲۰۰۲)

-استفاده بهتر از منابع: موتورهای فراجویشگر از موتورهای جستجوی جزیی برای انجام جستجوی اولیه استفاده می کنند .این به آنها امکان می دهد از منابع ذخیره سازی و محاسبه این موتورهای جستجو استفاده کنند(Meng et al., ۲۰۰۲) .

در نتیجه ، فراجویشگرها از هزینه های زیر که برای راه اندازی موتور جستجو لازم است اجتناب می کنند:

۱-خزیدن و ذخیره مجموعه اسناد ، ۲- نمایه سازی اسناد جمع آوری شده ، و ۳- جستجوی پایگاه داده فهرست (Meng et al., ۲۰۰۲)

برای موتورهای جستجوی بزرگ ، فقط هزینه خرید رایانه های مورد نیاز ، نگهداری آنها و نگهداری عملیات آنها (شامل نگهداری نرم افزار/سخت افزار و مصرف برق) می تواند بسیار بالا باشد (Meng et al., ۲۰۰۲) .

اگرچه موتورهای فرا جویشگر نیز برای انجام کارکردهای خود مانند انتخاب موتور جستجو ، تولید نماینده و ادغام نتایج به زیرساخت های خاص خود نیاز دارند ، اما نیاز آنها به زیرساخت در مقایسه با موتورهای جستجو در همان مقیاس بسیار کمتر است؛ در حالی که موتورهای فرا جویشگر بزرگ دارای مزایای فوق نسبت به موتورهای جستجوی اصلی هستند ، اما دارای معایب ذاتی نیز هستند. (Meng et al., ۲۰۰۲)

اول ، طول می کشد تا یک موتور فراجویشگر نتایج را به موتورهای جستجوگر اصلی خود بازگرداند زیرا موتور فراجویشگر باید هر پرس و جو را به موتورهای جستجوی اجزای منتخب منتقل کند ، منتظر بماند تا پرس و جو توسط آنها پردازش شود و در نهایت منتظر بمانید تا نتایج از آنها بازگردانده شود. موتورهای جستجوی اصلی کمتر از یک ثانیه به ارزیابی یک پرس و جو نیاز دارند در حالی که موتورهای فرا جویشگر اغلب ۲-۵ ثانیه طول می کشد تا نتایج را بازگردانند. این تفاوت در زمان پاسخگویی بین موتورهای جستجو و موتورهای فرا جویشگر در آینده با افزایش سرعت اینترنت به احتمال زیاد کاهش می یابد. (Meng et al., ۲۰۰۲)

دوم ، موتورهای جستجوی اصلی کنترل کاملی بر الگوریتم های رتبه بندی اسناد خود دارند و شانس بیشتری برای استفاده از اطلاعات پیوند در صفحات وب دارند.(Meng et al., ۲۰۰۲)

زمان پاسخگویی ، کیفیت نتایج ، و کیفیت قطعات نتیجه این موتورهای جستجو می تواند بر عملکرد موتورهای فرا جویشگر تأثیر بسزایی بگذارد. (Meng et al., ۲۰۰۲)

۳-۴-معماری فراجویشگرها:

کاربر پرس و جوی موردنظر خود را به یک فراجویشگر میفرستد سپس آن فراجویشگر پرس و جو را به تعدادی موتور جستجو میفرستد و نتایج مناسب را از این موتورهای جستجو بازگردانی میکند که به این موتور های جستجوی زیرین موتور های جستجوی جزیی گفته میشود. موتورهای جستجوی مختلف فرمت های پرس و جوی مختلفی را دریافت میکنند که آن فراجویشگر باید فرمت پرس و جو را متناسب با آنها تغییر دهد و بفرستد.(W Meng, CT Yu , ۲۰۱۰)

سپس فراجویشگر باید نتایج بازیابی شده از موتورهای جستجوی جزیی را با هم دیگر به بهترین شکل ممکن ادغام کند و یک لیست رتبه بندی شده مناسب را به کاربر نشان دهد.این لیست میتواند لیستی از اسناد یا اینکه لیستی از ادرس های صفحات وب باشد؛در اینجا ما نیاز به یک تابع برای یافتن شباهت بین اسناد و پرس‌وجوی کاربر داریم.که شبیه ترین اسناد به پرس و جو را بازیابی و ادغام کند؛ که ممکن است موتورهای جستجوی مختلف تابع شباهت های متفاوتی داشته باشند ولی در نهایت فراجویشگر از یک تابع شباهت کلی استفاده میکند.(W Meng, CT Yu , ۲۰۱۰)

انتخاب کننده پایگاه داده:اگر تعداد موتورهای جستجوی جزیی یک فراجویشگر کم باشد منطقی است که پرس و جوی کاربر را به همه آنها ارسال کنیم اما اگر که این تعداد بسیار زیاد باشد آنگاه ارسال پرس و جوی کاربر به همه موتورهای جستجوی جزیی اصلا روش منطقی نیست بلکه باید از بین آنها مناسب ترین ها انتخاب شوند و پرس و جوی کارب به آنها ارسال گردد.(W Meng, CT Yu , ۲۰۱۰)

فرض کنید که یک فراجویشگر بیش از صد موتور جستجوی زیرین دارد و تنها نیاز به ۱۰ تا بهترین اسناد را دارد، ارسال پرس و جوی کاربر به موتورهای جستجوی جزیی نامناسب چندین مشکل اساسی دارد:

۱-درگیری منابع نامناسب برای پرس و جو ۲-ترجمه پرس و جو برایموتورهای جستجوی جزیی نامناسب که کاری زمانبر است.۳-ایجاد ترافیک زیادو نامناسب در شبکه(W Meng, CT Yu , ۲۰۱۰)

یک انتخاب کننده پایگاه مناسب باید از بین همه موتورهای جستجوی جزیی یک فراجویشگر مناسب ترین های آنها را نسبت به پرس و جوی کاربر انتخاب کند.(W Meng, CT Yu , ۲۰۱۰)

انتخاب کننده سند: هر موتور جستجوی جزیی که با انتخاب کننده پایگاه انتخاب شده متناسب با پرس و جوی کاربر تعدای سند برمیگرداند که انتخاب کننده سند با استفاده از الگوریتم های خاص خود تشخیص میدهد کدام اسناد مرتبط تر هستند و باید بازگردانی شوند.(W Meng, CT Yu , ۲۰۱۰)

ارسال کننده پرس و جو: ارسال کننده پرس و جو مسئول برقراری ارتباط با سرور هر موتور جستجوی انتخاب شده و ارسال پرس و جو به آن است .(HyperText Transfer Protocol) HTTP برای اتصال و انتقال داده (ارسال پرس و جو و دریافت نتایج) استفاده می شود؛ هر موتور جستجو الزامات خاص خود را در مورد روش درخواست) HTTP به عنوان مثال ، روش GET یا روش POST و قالب پرس و جو (به عنوان مثال ، نام جعبه پرس و جو خاص) دارد. فرستنده پرس و جو باید الزامات هر موتور جستجو را به درستی دنبال کند(W Meng, CT Yu , ۲۰۱۰) .

ادغام کننده پرس و جو: پس از اینکه هر موتور جستجو جزیی بهترین اسناد مرتبط با پرس و جوی کاربر را بازگردانی کرد حال در این مرحله موتور فراجویشگر باید نتایج بازیابی شده موتورهای جستجو جزیی را ادغام کرده و در یک لیست رتبه بندی شده به کاربر نشان دهد که اسنادی که در لیست ادغام شده در بالاتر قرار دارند اسناد مرتبط تر و مفیدتری هستند. یک ادغام کننده نتیجه خوب باید همه اسناد

برگشتی را به ترتیب نزولی از نظر شباهت جهانی با پرس و جو کاربر طبقه بندی کند (W Meng, CT Yu , ۲۰۱۰) .

درادامه سه بخش اصلی انتخاب کننده پایگاه داده و انتخاب کننده سند و ادغام کننده را بطور کامل شرح میدهیم.(W Meng, CT Yu , ۲۰۱۰)

۳-۴-۱-انتخاب کننده پایگاه داده:

هنگامی که یک فراجویشگر یک پرس و جو را از کاربر دریافت میکند باید از بین همه موتورهای جستجوی جزیی که دارد مرتبط ترین های انها با پس و جوی کاربر را با استفاده از انتخاب کننده پایگاه داده انتخاب کند. یک الگوریتم خوب انتخاب پایگاه داده باید پایگاه های داده بالقوه مفید را به طور دقیق شناسایی کند. روش های زیادی برای مقابله با مشکل انتخاب پایگاه داده پیشنهاد شده است. حال ما نیاز به رویکرد ها و الگوریتم هایی داریم که بتوانیم بفهمیم کدام موتورهای جستجوی جزیی با پرس و جوی ما ارتباط نزدیک تری دارند و مفیدتر هستند؛ این رویکردها در به سه دسته تقسیم شده اند. (W Meng, CT Yu , ۲۰۱۰)

Rough representative approaches: -سخت

در این رویکردها ، محتویات پایگاه داده محلی(موتورهای جستجوی جزیی) اغلب با چند کلمه یا پاراگراف کلیدی انتخاب شده نشان داده می شود؛ چنین نماینده ای فقط می تواند یک ایده کلی در مورد پایگاه داده ارائه دهد ، و در نتیجه روش های انتخاب پایگاه داده با استفاده از نمایندگان پایگاه داده خشن(سخت) در برآورد مفید بودن واقعی پایگاه های داده در رابطه با داده های دقیق چندان دقیق نیستند؛ این نمایندگان اغلب به صورت دستی ایجاد می شوند. هنگام تعیین میزان مناسب بودن مجموعه اسناد برای درخواست کاربر ، مجموعه ها بر اساس تطابق نمایندگان آنها با پرس و جو رتبه بندی می شوند(W Meng, CT Yu , ۲۰۱۰) .

Statistical representative approaches: -آماری

این رویکردها معمولاً محتویات یک پایگاه داده را با استفاده از اطلاعات آماری نسبتاً دقیق نشان می دهند. به طور معمول ، نماینده یک پایگاه داده شامل برخی اطلاعات آماری برای هر اصطلاح در پایگاه داده مانند فراوانی سند اصطلاح و میانگین وزن اصطلاح در بین تمام اسنادی است که این اصطلاح را

دارند. آمارهای تفصیلی امکان برآورد دقیق تر مفید بودن پایگاه داده را با توجه به هر گونه پرسش کاربر فراهم می کند. مقیاس پذیری چنین رویکردهایی به دلیل میزان اطلاعاتی که باید برای هر پایگاه داده ذخیره شود ، مسئله مهمی است؛ نماینده آماری یک پایگاه داده معمولاً هر اصطلاح در هر سند موجود در پایگاه داده را در نظر می گیرد و یک یا چند قطعه اطلاعات آماری را برای هر عبارت ذخیره می کند (W Meng, CT Yu , ۲۰۱۰) .

روشهای زیادی بر اساس نمایندگان آماری پیشنهاد شده است .در این بخش ، ما پنج رویکرد را شرح می دهیم.

روش D-WISE: D-WISE(فهرست وب و موتور جستجو) یک مرکز متمرکز است، D-WISEیک موتور جستجوی فراجویشگر با تعدادی موتورهای جستجوی زیربنایی توزیع شده است؛ در D-WISE ، نمایشگر یک موتور جستجوی جزیی شامل تعداد اسناد هر عبارت در پایگاه داده جزیی و همچنین تعداد اسناد موجود در پایگاه داده است. بنابراین ، نماینده پایگاه داده با n اصطلاح متمایز علاوه بر n عبارت ، شامل n +۱کمیت(فرکانس n سند و کاردینالیتی پایگاه داده)خواهد بود. اجازه دهید n_i تعداد اسناد موجود در پایگاه داده ith را نشان دهد و df_{ij} فرکانس سند عبارت t_j در پایگاه داده ith باشد (W Meng, CT Yu , ۲۰۱۰) .

فرض کنید q یک درخواست کاربر است .نمایندگان همه پایگاه های داده برای محاسبه نمره رتبه بندی هر موتور جستجوی جزء با توجه به q استفاده می شوند. نمرات مفید بودن نسبی همه پایگاه های داده را با توجه به q اندازه گیری می کنند؛ اگر امتیاز پایگاه داده A بالاتر از پایگاه داده B باشد ، پس پایگاه داده Aنسبت به پایگاه داده B به q مرتبط تر است. نمرات رتبه بندی به شرح زیر محاسبه می شود (W Meng, CT Yu , ۲۰۱۰) .

ابتدا ، اعتبار نشانه هر عبارت پرس و جو ، مثلاً عبارت t_j، برای پایگاه داده ith iام ، cv_{ij} ، با استفاده از فرمول زیر محاسبه می شود: (W Meng, CT Yu , ۲۰۱۰)

$$cv_{ij} = \frac{\frac{df_{ij}}{n_i}}{\frac{df_{ij}}{n_i} + \frac{\sum_{k \neq i}^{N} df_{kj}}{\sum_{k \neq i}^{N} nk}}$$

که در آن N تعداد کل پایگاه های داده های جزیی موجود در موتور فراجویشگر است. بصورت بصری ، cv_{ij} درصد اسناد موجود در پایگاه داده که حاوی عبارت t_j، است را نسبت به سایر پایگاه های داده اندازه گیری می کند؛ اگر پایگاه داده ith دارای درصد بالاتری از اسناد حاوی t_jدر مقایسه با پایگاه های داده دیگر باشد ، cv_{ij} آن پایگاه داده مقدار بیشتری دارد؛در مرحله بعد ، واریانس cv_{ij} از هر عبارت پرس و جو t_j برای همه پایگاه های داده مولفه ، cvv_j، به شرح زیر محاسبه می شود: (W Meng, CT Yu , ۲۰۱۰)

$$cv_{ij} = \frac{\sum_{i=1}^{N}(CV_{ij} - ACV_j)^2}{N}$$

جایی که ACV_j میانگین همه CV_{ij} برای همه پایگاه های داده مولفه است. (W Meng, CT Yu , ۲۰۱۰)

روشCORI Net:

در این روش نمایش هر پایگاه به ازای هر واژه با دو کمبت نشان داده میشود یکی فراوانی مستندات ودیگری تعداد پایگاه داده های شامل آن واژه است. در CORI Net ، برای یک پرس و جو q ، یک تکنیک رتبه بندی سند معروف به شبکه استنتاج که در سیستم بازیابی اسناد INQUERY استفاده می شود ، گسترش می یابد تا همه موتورهای جستجوی جزء با توجه به q رتبه بندی شوند. InCORINet، امتیاز رتبه بندی موتورهای جستجو در رابطه با پرس و جو q تخمینی است که موتور جستجو حاوی اسناد مفید است. این باور اساساً احتمال ترکیبی است که موتور جستجو برای هر عبارت جستجویی حاوی اسناد مفید است(W Meng, CT Yu , ۲۰۱۰) .

فرض کنید جستجوی کاربر شامل k تا اصطلاحt۱.....tk است و N تعدا موتورهای جستجوی جزیی است و df_{ij} فرکانس سند اصطلاح jام t_j در موتور جستجوی جزیی iام است S_i است و cf_j مجموعه فرکانس های t_j است .(W Meng, CT Yu , ۲۰۱۰)

$$p(tj\ |Si) = c۱ + (۱ - c۱) * Tij * Ij$$

$$Tij = c۲ + (۱ - c۲) * \frac{df_{ij}}{df_{ij} + k}$$

یک فرمول برای محاسبه اصطلاح وزن فرکانس t_j در سند فوق مربوط به Si است (W Meng, CT Yu , ۲۰۱۰).

$$Ij = \frac{\log(\frac{N+0.5}{cf_j})}{\log(N+0.1)}$$

روش GGLOSS: این روش یک مدل اولیه تحقیقاتی است و هر موتور جستجوی جزیی توسط یک جفت(df_i , W_i)نشان داده میشود که df_i فرکانس سند اصطلاح i-th و W_i مجموع وزن t_i در تمام اسناد موجود در موتور جستجوی جزء است؛ یک آستانه T با هر پرس و جو در gGlOSS همراه است تا نشان دهد که کاربر فقط به اسنادی علاقه دارد که شباهت آنها با پرس و جو بیشتر از T است؛ به طور خاص ، در gGlOSS ، مفید بودن موتور جستجوی S در رابطه با پرس و جو q و آستانه شباهت T به شرح زیر تعریف می شود: (W Meng, CT Yu , ۲۰۱۰)

$$usefulness(S, q, T) = \sum_{d \in S \cap sim(d,q) > T} sim(d, q)$$

جایی که sim(d, q) شباهت بین یک سند d و پرس و جوی q را نشان می دهد .از مفید بودن هر موتور جستجوی جزء به عنوان نمره رتبه بندی موتور جستجو استفاده می شود در حال حاضر ما باید مفید بودن هر موتور جستجوی جزیی مشخص شده توسط فرمول را تخمین بزنیم. (W Meng, CT Yu , ۲۰۱۰)

به نظر می رسد که برآورد مستقیم دشوار است .در gGlOSS ، دو روش برای برآورد سودمندی بر اساس دو فرض زیر ارائه شده است(W Meng, CT Yu , ۲۰۱۰) :

فرض همبستگی بالا :برای هر موتور جستجوی جزیی داده شده ، اگر عبارت پرس و جو وزن t_i حداقل در تعداد اسناد به عنوان عبارت جستار tj ظاهر می شود ، در این صورت هر سند حاوی عبارت tj نیز دارای عبارت ti است؛ فرض عدم پیوند: برای هر موتور جستجوی جزء معین ، برای هر دو عبارت جستجوی ti و tj ، مجموعه اسناد حاوی ti از مجموعه اسناد حاوی tj جدا نمی شود. (W Meng, CT Yu , ۲۰۱۰)

Learning-based approaches: یادگیری-

در این رویکردها ، دانش مربوط به پایگاه های داده ای که به احتمال زیاد اسناد مفید را به انواع پرس و جوها از تجربیات بازیابی گذشته آموخته است. چنین دانشی سپس برای تعیین مفید بودن پایگاه های داده برای پرس و جوهای آینده استفاده می شود. تجربیات بازیابی را می توان با استفاده از پرس و

جوهای آموزشی قبل از استفاده از الگوریتم انتخاب پایگاه داده و/یا پرس و جوهای کاربر واقعی در حالی که انتخاب پایگاه داده در حال استفاده فعال است ، بدست آورد. تجربیات به دست آمده در پایگاه داده ذخیره می شود. (W Meng, CT Yu , ۲۰۱۰)

۱-روش mrdd

این روش از یادگیری ایستا ومجموعه ای از جستجوهای آموزشی استفاده میکند و هر جستجوی آموزشی به هریک از موتورهای جستجوی جزیی ارسال میگردد. به ازای همه پاسخهای دریافتی از موتورهای جستجوی جزیی یک بردار به شکل $< r_1, r_2, \ldots r_n >$ به ما میدهد که در آن r_i یک عدد صحیح مثبت و بیانگر این است که برای دستیابی به i مستند جستجو باید r_i مستند برتر لیست رابازیابی کرد. (W Meng, CT Yu , ۲۰۱۰)

با در اختیار داشتن همه بردارهای ذکر شده به ازای همه جستجوهای آموزشی و همه پایگاه های داده هنگامی که یک پرس و جو به سیستم ارسال میگردد با همه جستجوهای آموزشی مقایسه شده و k عبارت جستجو که از همه به آن پرس و جو نزدیک تر هستند بازیابی میشوند؛ برای هر پایگاه داده D بردار توزیع میانگین روی k نزدیکترین عبارت جستجوی آموزشی و پاسگاه داده D حاصل میشود. (W Meng, CT Yu , ۲۰۱۰)

در این روش هر پایگاه داده بصورت مجموعه ای از بردارهای توزیع برای همه جستجوهای اموزشی نمایش داده میشود؛(W Meng, CT Yu , ۲۰۱۰)

ضعف اصلی این روش این است که به ازای هر جستجوی آموزشی، یادگیر بصورت دستی انجام میگیرد و همچنین تعیین جستجوهای اموزشی مناسب دشوار است و در صورت تغییر محتواهای پایگاه داده این اطلاعات دقیق نخواهند بود.(W Meng, CT Yu , ۲۰۱۰)

۲-روش SavvySearch

فراجویشگر savvysearch از یادگیری پویا استفاده میکند و رتبه هر موتور جستجوی جزیی نسبت به یک عبارت جستجو براساس تجارب بازیابی های قبلی حاصل از واژه های بکار گرفته شده در این پرس و جو محاسبه میگردد.(W Meng, CT Yu , ۲۰۱۰)

واحد انتخاب پایگاه داده به ازای موتورهای جستجوی زیرین یک بردار وزن بصورت($w_1, w_2 ... w_n$) نگهداری میکند که w_i منتناظر واژه i ام پایگاه داده است که در ابتدا تمام وزن ها با صفر مقداردهی میشوند و هنگامی که یک عبارت جستجو شامل واژه t_i به موتور جستجوی جزییD ارسال میگردد وزن w_i براساس نتایج مقداردهی میگردد و اگر آن موتور جستجوی جزیی هسچ نتسجه ای برنگرداند وزن متناظر آن به میزان $\mathbf{1}/_{\boldsymbol{k}}$ کاهش میابد که منظور از k تعداد واژه های بکار رفته در عبارت جستجوی ارسالی است و در مقابل اگر حداقل کاربر یکی از پاسخها را مشاهده کند وزن به میزان $\mathbf{1}/_{\boldsymbol{k}}$افزایش خواهد یافت. در نتیجه وزن مثبت زیاد w_i یعنی پایگاه داده D درگذشته پاسخهای مناسبی در مقابل واژه t_i داشته است.(W Meng, CT Yu , ۲۰۱۰)

همچنین در SavvySearch کارآیی فعلی هر کدام از موتور جستجوی جزئی براساس h (یعنی متوسط تعداد مستندات بازیابی شده برای پنج جستجوی آخر) و نیز (r متوسط زمان پاسخ این موتور جستجوی جزئی به پنج جستجوی آخر) در نظر گرفته میشود؛ اگر h پایین تر از آستانه Th (بصورت پیش فرض یک در نظر گرفته مبشود) باشد آنگاه جریمه ای بصورت زیر برای موتور جستجوی جزیی در نظر گرفته میشود.(W Meng, CT Yu , ۲۰۱۰)

$$p_h=\frac{(T_{h-h})^2}{T_h{}^2}$$

حال در مقابل اگر میانگین زمان پاسخ r بزرگتر از آستانه قدیمی Tr (پیش فرض ۱۵ ثانیه) باشد یک جریمه $p_r=\frac{(r-Tr)^2}{(r_0-Tr)^2}$ در نظر گرفته میشود که در آن $r_0 = 45\,\text{sec}$بیشترین زمان مجاز برای پاسخ یک موتور جستجوی جزیی میباشد؛ برای یک عبارت جستجوی جدید q متشکل از واژه های $t_1, t_2, \ldots t_n$، رتیه پایگاه داده D بصورت زیر محاسبه میشود:(W Meng, CT Yu , ۲۰۱۰)

$$r(q,D)=\frac{\sum w_{t_i}.\log(N/f_i)}{\sqrt{\sum_{i=1}^{k}|w_i|}}$$

که در فرمول فوق $\log(N/f_i)$ ؛ وزن عکس فراوانی پایگاه داده برای واژه t_i ، D عبارت است از تعداد پایگاه های داده ای که برای واژه t_i وزن مثبت دارند؛ هزینه‌های ذخیره به روز نگهداشتن نمایش هر موتور جستجوی جزئی درSavvySearch نسبتاً قابل قبول است .یکی از نقاط ضعفSavvySearch عملکرد ضعیف آن نسبت به جستجوهای جدید یا جستجوهای نادر است .از سوی دیگر، فرآیند دریافت بازخورد از

کاربر چندان دقیق نیست و ممکن است به راحتی منجر به تشخیص نادرست پایگاه‌های داده مفید گردد . اغلب کاربران، تنها نتایج با رتبه‌های بالا را صرفنظر از میزان مناسب بودن آنها بررسی میکنند (W Meng, CT Yu , ۲۰۱۰)

روش ProFusion

این فراجویشگر برای انتخاب پایگاه داده از روش ترکیبی استفاده میکند.دانش در این فراجویشگر به سیزده طبقه به شرح زیر تقسیم بندی میشود:(W Meng, CT Yu , ۲۰۱۰)

(۱) Science and Engineering, (۲) Computer Science, (۳) Travel, (۴) Medical and Biotechnology, (۵) Business and Finance, (۶) Art, (۷) Social and Religion, (۸) Society, Law and Government, (۹) Food, (۱۰) Animals and Environment, (۱۱) History, (۱۲) Music, (۱۳) Recreation and Entertainment.

موضوع هر طبقه با تعدادی واژه، مشخص میگردد .برای هر طبقه ازقبل، تعدادی جستجوی آموزشی، مشخص شده است .هدف از این کار، ایجاد یک یادگیری مقدماتی در خصوص این مطلب است که هر کدام از این موتورهای جستجوی جزئی چگونه به جستجوهای در حوزه های متفاوت، پاسخ میدهند .برای هر طبقه C و پایگاه داده D همه جستجوهای اموزشی به Dارسال میگردد .از بین ۱۰ پاسخ برتر ارسالی، مستندات مرتبط، شناسایی میشوند .سپس یک امتیاز که بیانگر میزان کارآییD در پاسخ به جستجو در طبقه Cمیباشد بر اساس فرمول زیر محاسبه میگردد(W Meng, CT Yu , ۲۰۱۰) :

$$C\times\frac{\sum_{i=1}^{10} N_i}{10}\times\frac{R}{10}$$

که در آن c یک ضریب ثابت است و N_iبرای مستند iام در صورتی که مرتبط باشد مقدار $1/i$ و در غیر اینصورت مقدار صفر را میگرد و R نیز عبارت است از تعداد مستندات مرتبط از بین ده مستند برتر به عنوان پاسخ.(W Meng, CT Yu , ۲۰۱۰)

۳-۴-۲-انتخاب کننده سند:

پس از اینکه انتخاب کننده پایگاه داده موتورهای جستجوی جزیی متناسب با پرس و پوی کاربر را انتخاب کرد و پرس و جو را به آنها فرستاد آنگاه نوبت این است که انتخاب کننده سند بهترین و مرتبط ترین اسناد را به کاربر نشان دهد؛ یک موتور جستجوی جزیی معمولاً اسناد را به ترتیب نزولی شباهت های محلی بازیابی می کند .در نتیجه ، مشکل انتخاب اسناد برای بازیابی از پایگاه داده های جزیی را می توان به یکی از دو مشکل زیر تبدیل کرد(W Meng, CT Yu , ۲۰۱۰) :

۱-تعیین تعداد اسناد برای بازیابی از پایگاه داده جزیی؛ اگر قرار است k سند از پایگاه داده جزیی بازیابی شود ، k سند با بیشترین شباهت های محلی بازیابی می شوند(W Meng, CT Yu , ۲۰۱۰) .

۲-تعیین یک آستانه محلی برای پایگاه داده جزیی به گونه ای که یک سند از پایگاه داده جزیی تنها در صورتی بازیابی شود که شباهت محلی آن با پرس و جو از آستانه بیشتر باشد (W Meng, CT Yu , ۲۰۱۰) .

برای هر یک از مشکلات ، هدف ما همیشه بازیابی همه یا حداکثر اسناد بالقوه مفید از هر پایگاه داده جزیی در حالی که بازیابی اسناد بی فایده را به حداقل رساند .ما رویکردهای پیشنهادی برای مشکل انتخاب سند را در چهار دسته زیر طبقه بندی می کنیم. (Meng et al., ۲۰۰۲)

۱-تعیین کاربر: موتور فراجویشگر به کاربر جهانی اجازه می دهد تا تعداد اسناد را از هر پایگاه داده اجزا بازیابی کند(Meng et al., ۲۰۰۲) .

۲-تخصیص وزن: تعداد اسناد برای بازیابی از پایگاه داده جزیی به نمره رتبه بندی (یا رتبه) پایگاه داده جزیی نسبت به نمرات رتبه بندی (یا رتبه) سایر پایگاه های داده بستگی دارد .در نتیجه ، به طور نسبی اسناد بیشتری از پایگاه های داده جزیی که رتبه بالاتری دارند بازیابی می شود(Meng et al., ۲۰۰۲) .

۳-رویکردهای مبتنی بر یادگیری: این رویکردها بر اساس تجربیات بازیابی گذشته پایگاه داده جزیی ، تعداد اسناد برای بازیابی از پایگاه داده جزیی را تعیین می کند.(Meng et al., ۲۰۰۲)

۴-بازیابی تضمینی: هدف این نوع رویکرد تضمین بازیابی همه اسناد مفید با توجه به هر گونه پرس و جوی کاربر است(Meng et al., ۲۰۰۲) .

۳-۴-۳-ادغام کننده:

هدف ادغام کننده نتایج این است که نتایج همه موتورهای جستجوی جزیی را باهم ادغام کند و یک لیست از بهترین نتایج را به ما نشان دهد. از دهه ۱۹۹۰ بسیاری از الگوریتم های ادغام نتایج پیشنهاد و مطالعه شده اند .این الگوریتم ها را می توان در چندین بعد طبقه بندی کرد(Meng et al., ۲۰۰۲) .

به عنوان مثال ، یک بعد میزان اطلاعات مربوط به هر نتیجه است که برای انجام ادغام استفاده می شود؛ این ممکن است از استفاده از رتبه های محلی هر نتیجه از موتورهای جستجو که آن را بازیابی کرده اند ، گرفته تا استفاده از رکوردهای نتیجه جستجو و استفاده از سند کامل نتیجه باشد. بعد دیگر میزان همپوشانی اسناد در موتورهای جستجو است که برای پاسخ به پرس و جو استفاده می شود. این رنج می تواند از عدم همپوشانی ، همپوشانی جزئی و احتمالاً مجموعه اسناد یکسان متغیر باشد (Meng et al., ۲۰۰۲) .

الگوریتم های ادغام نتایج اولیه فرض می کرد که همه موتورهای جستجوی جزیی برای هر نتیجه بازیابی شده با توجه به درخواست کاربر داده شده ، شباهت محلی را برمی گردانند. شباهت های محلی به دلایل مختلف ممکن است مستقیماً قابل مقایسه نباشند. به عنوان مثال ، یک دلیل این است که موتورهای جستجوی جزیی مختلف ممکن است شباهت های خود را در محدوده های مختلف نرمالسازی کنند ، به عنوان مثال ، یکی در [۱و۰] و دیگری در [۱۰۰۰و۰]. دلیل دیگر این است که وزن اسناد در موتورهای جستجوی مختلف بر اساس آمار مجموعه های مختلف (به عنوان مثال ، فرکانس اسناد) محاسبه می شود. (Meng et al., ۲۰۰۲)

برای مقایسه بیشتر شباهت های محلی ، الگوریتم ادغام شباهت های محلی را به یک محدوده مشترک نرمالسازی می کند ، به عنوان مثال [۱و۰]. (W Meng, CT Yu , ۲۰۱۰)

یک روش پیچیده ترکه SSL نامیده میشود و برای نرمالسازی شباهت های محلی از شباهت های جهانی با استفاده از یک تابع نگاشت یادگیری استفاده میشود. یک نمونه از مجموعه اسناد متمرکز CSDبا ترکیب اسناد نمونه برداری از هر موتور جستجو بدست می آید CSD .به عنوان نماینده مجموعه جهانی حاوی کلیه اسناد در تمام موتورهای جستجو در نظر گرفته می شود و شباهت های محاسبه شده

بر اساس CSD و عملکرد مشابه جهانی به عنوان شباهت های جهانی در نظر گرفته می شود. (W Meng, CT Yu , ۲۰۱۰)

هر پرس و جو به هر موتور جستجوی جزیی منتخب ارسال می شود تا برخی از اسناد و شباهت های محلی آنها را بازیابی کندشباهت جهانی هر سند بازیابی شده که در CSD ظاهر می شود نیز محاسبه می شود. سپس ، بر اساس چندین جفت شباهت جهانی و محلی برخی از اسناد از هر موتور جستجو ، یک تابع نقشه برداری می تواند بر اساس رگرسیون بدست آید. (W Meng, CT Yu , ۲۰۱۰)

هر موتور جستجو ، یک تابع نقشه برداری را می توان بر اساس رگرسیون بدست آورد. برخی از این الگوریتم های ادغام با در نظر گرفتن سودمندی یا کیفیت برآورد شده هر موتور جستجو ، شباهت های محلی نرمال شده را بیشتر تنظیم می کند تا به نتایج بازیابی شده از موتورهای جستجو مفیدتری برسند. نمره رتبه بندی موتور جستجو که در مرحله انتخاب موتور جستجو محاسبه می شود ، مفید بودن یک موتور جستجو را برای یک پرس و جو مشخص می کند. در نهایت ، نتایج بدست آمده از موتورهای جستجوی مختلف به ترتیب نزولی از شباهت های تعدیل شده توسط ادغام نتایج رتبه بندی می شوند. (W Meng, CT Yu , ۲۰۱۰)

در CORI Net ، تنظیم(adjustment) به شرح زیر عمل می کند. (W Meng, CT Yu , ۲۰۱۰)

اجازه دهید rs نمره رتبه بندی موتور جستجوی جزء S و av_rs میانگین نمرات رتبه بندی همه موتورهای جستجوی جزیی انتخابی باشد و Ls شباهت محلی نتیجه r از D باشد؛ پس شباهت تعدیل شده سند d با ls * (۱ + N * (rs - av_rs)/av_rs) محاسبه می شود ، جایی که N تعداد موتورهای جستجوی جزیی انتخاب شده برای پرس و جو داده شده است. بدیهی است که این تعدیل شباهت های محلی نتایج به دست آمده از موتورهای جستجو را افزایش می دهد که رتبه بندی آنها بالاتر از میانگین رتبه بندی است. در ProFusion ، شباهت محلی هر نتیجه به سادگی با ضرب آن در نمره رتبه بندی موتور جستجو که نتیجه را بازیابی کرده است ، تنظیم می شود. (W Meng, CT Yu , ۲۰۱۰)

اکثر موتورهای جستجوی امروزی شباهت هایی را برای نتایج بازیابی شده نشان نمی دهند. در نتیجه ، روشهای نرمال سازی و تعدیل شباهت محلی فوق برای موتورهای فراجویشگری که موتورهای جستجوی جزیی آنها مستقل و غیرهمکار هستند ، دیگر قابل استفاده نیست. (W Meng, CT Yu , ۲۰۱۰)

در ادامه ما بر الگوریتم های ادغام نتایج که از شباهت های محلی نتایج بازیابی شده استفاده نمی کنند ، تمرکز می کنیم؛ به طور کلی ، برای نتیجه ای که توسط موتور جستجوی جزء S بازیابی می شود ، اطلاعات زیر را می توان بدست آورد و برای ادغام نتایج استفاده کرد: (W Meng, CT Yu , ۲۰۱۰)

-سند کامل r :سند کامل را می توان با استفاده از URL صفحه وب ، که معمولاً در SRR نتیجه موجود است ، بارگیری کرد. (W Meng, CT Yu , ۲۰۱۰)

-رتبه محلی :r این رتبه موقعیت r در بین نتایج برگشت داده شده توسط S برای پرس و جو کاربر است.

-عنوان :r این عنوان صفحه وب برای r است که معمولاً در SRR نتیجه درج می شود. (W Meng, CT Yu , ۲۰۱۰) .

- URL r: این آدرس صفحه وب برای r است که معمولاً در SRR نتیجه درج می شود. توجه داشته باشید که ما نه تنها می توانیم صفحه وب را با استفاده از URL بارگیری کنیم ، بلکه اغلب می توانیم متوجه شویم که کدام سازمان/شخص صفحه وب را از آدرس اینترنتی منتشر کرده است. (W Meng, CT Yu , ۲۰۱۰)

-تکه ای از :r این یک متن کوتاه از صفحه وب برای r است که معمولاً در SRR نتیجه درج می شود. (W Meng, CT Yu , ۲۰۱۰)

-زمان انتشار :r این زمانی است که صفحه وب برای r منتشر شد.این اطلاعات اغلب در SRR نتیجه هنگامی که نتیجه به زمان حساس است ، گنجانده می شود. به عنوان مثال ، موتورهای جستجوی اخبار اغلب زمان انتشار مقالات خبری بازیابی شده در SRR ها را شامل می شوند .اگر این اطلاعات در SRR ارائه شده باشد ، می توان به جای آن از آخرین_زمان اصلاح شده صفحه وب استفاده کرد. (W Meng, CT Yu , ۲۰۱۰)

اندازه :r این تعداد بایت های صفحه وب برای r است که معمولاً در ART نتیجه گنجانده می شود. (W Meng, CT Yu , ۲۰۱۰)

-نمره رتبه بندی :S این نمره رتبه بندی S با توجه به درخواست کاربر است و نمره توسط انتخاب کننده موتور جستجو در مرحله انتخاب موتور جستجو محاسبه می شود. (W Meng, CT Yu , ۲۰۱۰)

همه اطلاعات فوق توسط الگوریتم های ادغام نتایج موجود استفاده نشده است .در حقیقت ، اکثر الگوریتم های ادغام نتایج فعلی از زیر مجموعه کوچکی از اطلاعات فوق استفاده می کنند. (W Meng, CT Yu , ۲۰۱۰)

در این بخش ، الگوریتم های ادغام نتیجه را بر اساس انواع اطلاعاتی که برای انجام ادغام استفاده می کنند ، طبقه بندی کرده و بر اساس این طبقه بندی ارائه می دهیم. (W Meng, CT Yu , ۲۰۱۰)

ادغام بر اساس محتوای اسناد کامل

بعد از اینکه یک موتور جستجوی جزیی ، درخواست کاربر ارسال شده از موتور فراجویشگر را پردازش می کند ، لیستی از SRR ها را به موتور فراجویشگر برمی گرداند. به منظور استفاده از محتوای کامل سند هر نتیجه برای انجام ادغام نتایج ، باید اسناد کامل همه این نتایج را از وب سایتهایی که آنها را با استفاده از URLهای این SRR ها میزبانی می کنند ، تهیه کنید. پس از وصول اسناد کامل ، ادغام نتیجه می تواند از هر تابع شباهت جهانی برای محاسبه شباهت های جهانی آنها با پرس و جو استفاده کند. (W Meng, CT Yu , ۲۰۱۰)

موردی را که در آن تابع شباهت جهانی تابع Cosine است در نظر بگیرید و فرکانس اسناد جهانی هر عبارت برای موتور فراجویشگر شناخته شده است (توجه داشته باشید که اگر موتورهای جستجوی جزیی انتخاب شده هیچ همپوشانی نداشته باشند ، فرکانس سند جهانی یک اصطلاح را می توان تقریباً به عنوان مجموع فرکانس های سند در تمام موتورهای جستجو انتخاب کرد. (W Meng, CT Yu , ۲۰۱۰)

پس از بارگیری سند ، می توان فرکانس هر عبارت را در سند بدست آورد. در نتیجه ، همه آمارها) یعنی فراوانی اصطلاح tf و فرکانس سند df هر عبارت (که برای محاسبه شباهت جهانی سند در دسترس است و شباهت جهانی را می توان محاسبه کرد. پس از محاسبه شباهت های جهانی همه اسناد بازیابی شده ، ادغام نتایج ، نتایج بدست آمده توسط موتورهای جستجوی جزیی مختلف را به ترتیب نزولی شباهت های جهانی آنها رتبه بندی می کند. (W Meng, CT Yu , ۲۰۱۰)

الگوریتم ادغام نتیجه در موتور فراجویشگر Inquirus شباهت جهانی هر سند بارگیری شده d را برای یک پرس و جو معین با استفاده از تابع شباهت زیر محاسبه می کند: (W Meng, CT Yu , ۲۰۱۰) .

$$sim(d, q) = c_1 N_p + (c_2 - \frac{\sum_{i=1}^{N_p-1} \sum_{j=i+1}^{N_p} \min(d(i,j), c_2)}{\sum_{k=1}^{N_p-1} (N_p - k)}) / (\frac{c_2}{c_1}) + \frac{N_t}{c_3}$$

N_p: تعداد واژه های متمایز پرس و جوی q که در سند d ظاهر شده اند. (W Meng, CT Yu , ۲۰۱۰)

N_t:تعداد کل اصطلاحات پرس و جو در سند d (W Meng, CT Yu , ۲۰۱۰) .

d(i, j): حداقل فاصله (بر حسب تعداد کاراکترها) i امین وjامین عبارتهای query در d است (W Meng, CT Yu , ۲۰۱۰) .

c_1: یک ثابت است که اندازه کلی sim(d, q) را کنترل می کند. (W Meng, CT Yu , ۲۰۱۰)

c_2 :یک ثابت است که حداکثر فاصله بین عبارتهای پرس و جو را مشخص می کند. (W Meng, CT Yu , ۲۰۱۰)

c_3 :ثابت است که اهمیت فرکانس اصطلاحات را مشخص می کند. (W Meng, CT Yu , ۲۰۱۰)

در Inquirus ، این تنظیمات به شرح زیر است: (W Meng, CT Yu , ۲۰۱۰) .

c۱ = ۱۰۰, c۲ = ۵۰۰۰, and c۳ = ۱۰∗c۱.

اگر q فقط یک عبارت داشته باشد ، Inquirusبه سادگی از فاصله ابتدای سند تا اولین بار استفاده از این عبارت به عنوان شاخص ارتباط استفاده می کند. تابع شباهت فوق نه تنها اصطلاحات مشترک بین پرس و جو و سند ، بلکه مجاورت اصطلاحات پرس و جو در سند را نیز ضبط می کند. (W Meng, CT Yu , ۲۰۱۰)

الگوریتم ادغام پیشنهاد شده توسط رسولوفو و همکاران در سال ۲۰۰۳ ابتدا اسناد کامل نتایج بازیابی شده از تمام موتورهای جستجوی انتخاب شده را بارگیری می کند تا مجموعه ای از اسناد را تشکیل دهد ، سپس هر سند را به عنوان بردار اصطلاحات با وزن نشان می دهد ، جایی که وزن ها بر اساس tf*idf محاسبه می شوند. در نهایت ، از یک تابع شباهت جهانی برای محاسبه شباهت جهانی هر سند با پرس و جو کاربر استفاده می شود ، گویی یک سیستم بازیابی متن برای مجموعه سند تشکیل شده وجود دارد. سپس نتایج به ترتیب نزولی از شباهت های جهانی رتبه بندی می شوند. (W Meng, CT Yu , ۲۰۱۰)

الگوریتم OptDocRetrv یک روش مبتنی بر سند کامل است که ترکیبی از انتخاب سند (به عنوان مثال ، تعیین تعداد نتایج بازیابی از هر موتور جستجوی انتخاب شده و ادغام نتایج است.فرض کنید که m بیشترین سند مشابه در کل موتورهای جستجوی جزیی با توجه به یک پرس و جوی معین برای عدد صحیح مثبت m مورد نیاز است. (W Meng, CT Yu , ۲۰۱۰)

اول ، برای برخی از عدد صحیح مثبت کوچک) به عنوان مثال ، kمی تواند از ۲ شروع شود (، هر یک از موتورهای جستجو با رتبه برتر جستجو می شود تا شباهت واقعی واقعی مشابه ترین سند خود را بدست آورد. (W Meng, CT Yu , ۲۰۱۰)

این ممکن است مستلزم بارگیری برخی اسناد از این موتورهای جستجو باشد .اجازه دهید min_sim حداقل این K شباهت ها باشد. در مرحله بعد ، از این k موتورهای جستجو ، همه اسنادی که شباهت های واقعی جهانی آنها بیشتر یا برابر با آستانه آزمایشی min_sim است ، بازیابی می شوند. (W Meng, CT Yu , ۲۰۱۰)

اگر m سند یا بیشتر بازیابی شده باشد ، این روند متوقف می شود. در غیر این صورت ، موتور جستجوی رتبه بندی شده بعدی در نظر گرفته می شود و مشابه ترین سند آن بازیابی می شود؛ سپس شباهت جهانی واقعی این سند با min_sim مقایسه می شود و حداقل این دو شباهت به عنوان یک آستانه جهانی جدید برای بازیابی همه اسناد از این موتورهای جستجو k+۱ که شباهت های واقعی جهانی آنها بیشتر یا مساوی این مورد است ، مورد استفاده قرار می گیرد. این فرایند تا زمان بازیابی m یا بیشتر سند تکرار می شود. سرانجام ، اسناد بازیابی شده به ترتیب نزولی از شباهت های واقعی جهانی خود رتبه بندی می شوند. برای کاهش احتمال فراخوانی چندین بار جستجو در چندین فرایند فوق ، هنگام فراخوانی اولیه موتور جستجو می توان تعداد بیشتری از نتایج را ذخیره کرد. (W Meng, CT Yu , ۲۰۱۰)

الگوریتم OptDocRetrv دارای ویژگی زیر است: اگر موتورهای جستجو به طور مطلوب رتبه بندی شوند و بیشترین شباهت اسناد را از موتورهای جستجو بدست آوریم ، این الگوریتم حداکثر l+۱ موتورهای جستجو را فرا می خواند تا بیشترین اسناد مشابه را به دست آورد. (W Meng, CT Yu , ۲۰۱۰)

بارگیری اسناد و تجزیه و تحلیل آنها می تواند یک کار گران قیمت باشد ، به ویژه هنگامی که تعداد اسناد قابل بارگیری زیاد است و اندازه اسناد نیز زیاد است.برااای یان مشکل تعدادی راه حل وجود دارد: (W Meng, CT Yu , ۲۰۱۰) .

اول: بارگیری از سیستم های مختلف محلی را می توان به صورت موازی انجام داد (W Meng, CT Yu , ۲۰۱۰) .

دوم: برخی از اسناد را می توان ابتدا تجزیه و تحلیل کرده و به کاربر نمایش داد تا در حالی که کاربر نتایج اولیه را می خواند ، تجزیه و تحلیل بیشتری انجام شود؛نتایج نمایش داده شده در ابتدا ممکن است به درستی رتبه بندی نشده باشند و هنگام تجزیه و تحلیل اسناد بیشتر ، رتبه کلی باید تنظیم شود. (W Meng, CT Yu , ۲۰۱۰)

سوم ، ما ممکن است فقط بخش ابتدایی هر سند (بزرگ) را برای تجزیه و تحلیل بارگیری کنیم. با افزایش پهنای باند اینترنت ، تاخیر ناشی از بارگیری اسناد در حال انجام باید کمتر و کمتر شود. (W Meng, CT Yu , ۲۰۱۰)

از سوی دیگر ، روشهای مبتنی بر بارگیری نیز دارای مزایای واضحی از جمله موارد زیر هستند. (W Meng, CT Yu , ۲۰۱۰)

اول ، هنگام تلاش برای بارگیری اسناد ، URLهای منسوخ را می توان شناسایی کرد. در نتیجه ، اسناد با URLهای منسوخ را می توان از لیست نتایج نهایی حذف کرد. (W Meng, CT Yu , ۲۰۱۰)

دوم ، با تجزیه و تحلیل اسناد بارگیری شده ، اسناد بر اساس محتوای فعلی آنها رتبه بندی می شوند. در مقابل ، شباهت های محلی ممکن است بر اساس نسخه های قدیمی این اسناد محاسبه شود. (W Meng, CT Yu , ۲۰۱۰)

سوم ، شرایط پرس و جو در اسناد بارگیری شده را می توان در صورت نمایش بدون تأخیر بیشتر به کاربر نمایش داد زیرا این شرایط قبلاً هنگام پردازش این اسناد برای محاسبه شباهت های جهانی آنها مشخص شده است. (W Meng, CT Yu , ۲۰۱۰)

ادغام بر اساس سوابق نتایج جستجو

سوابق نتایج جستجو (SRR) که توسط اکثر موتورهای جستجوی امروزی بازگردانده می شوند حاوی اطلاعات غنی در مورد نتایج بازیابی شده هستند. به ویژه ، عناوین SRRحاوی اطلاعات محتوای با کیفیت بالا هستند که ارتباط اسناد مربوطه را با توجه به موضوع پرس و جو منعکس می کند. (Rasolofo et al, ۲۰۰۳).

اولا ، بر هیچ کس پوشیده نیست که موتورهای جستجوی امروزی در عنوان یک صفحه نسبت به اصطلاحات صفحه وزن بیشتری دارند. ثانیاً ، قطعه قطعات معمولاً به طور خاص برای درخواست کاربر ارسال شده ایجاد می شوند و اغلب قطعه (های) متنی در اسناد هستند که با پرس و جو مطابقت بیشتری دارند. در نتیجه ، عنوان و قطعه ای از یک نتیجه می تواند سرنخ های خوبی در مورد اینکه آیا سند مربوطه به پرس و جو مربوط است ، ارائه دهد. .(Rasolofo et al, ۲۰۰۳)

چندین الگوریتم ادغام نتیجه برای انجام ادغام بر اساس اطلاعات موجود در SRR های بازیابی شده ، به ویژه عناوین و قطعاتی در SRR ها پیشنهاد شده است .در زیر برخی از این الگوریتم ها را معرفی می کنیم. .(Rasolofo et al, ۲۰۰۳)

۱-روش TSR: TSRعنوان و قطعه ای از هر نتیجه بازیابی شده را در یک سند نماینده واحد ترکیب می کند. اگر نتیجه ای توسط چندین موتور جستجو بازگردانده شود ، نماینده عنوان و تمام قطعات نتیجه را در بر می گیرد. شباهت بین پرس و جو و نماینده ، مجموع وزن عبارات پرس و جو در نماینده است .در لیست ادغام شده ، نتایج به ترتیب نزولی از شباهت نمایندگان آنها رتبه بندی می شود. (Rasolofo et al, ۲۰۰۳).

۲-روش **TSRDS:** این روش یک مدل TSR است به این دلیل که فرایند ادغام نتیجه را با استفاده از نظریه شواهد Dempster-Shafer مدل می کند. وجود یک عبارت پرس و جو در نماینده به عنوان شواهدی در مورد ارتباط نتیجه با پرس و جو تلقی می شود و نتایج به ترتیب نزولی شواهد ترکیبی آنها رتبه بندی می شود. بر اساس فرکانس سند اصطلاح بین نمایندگان برای نتایج موتور جستجو ، یک وزن به شواهد هر عبارت برای هر موتور جستجو اختصاص داده می شود. .(Rasolofo et al, ۲۰۰۳)

برای یک قطعه از متن T (به عنوان مثال ، عنوان یا قطعه ای) از یک SRR و یک پرس و جو q ، شباهت بین T و q ، که به صورت sim(T , q) مشخص می شود ، به شرح زیر تعریف می شود: (Rasolofo et al, ۲۰۰۳).

$$Sim(T, q) = \begin{cases} ۱۰۰۰۰۰ * |T \cap q| / \sqrt{|T|^۲ + |q|^۲} & if\ T \cap q \neq \emptyset \\ ۱۰۰۰ - Rank & if \end{cases}$$

$|X|$: طول x در تعداد اصطلاحات است. .(Rasolofo et al, ۲۰۰۳)

Rank: رتبه محلی SRR است. .(Rasolofo et al, ۲۰۰۳)

فقط ۱۰ نتیجه برتر از هر موتور جستجوی جزء برای شرکت در ادغام استفاده می شود. (Rasolofo et al, ۲۰۰۳).

$۱۰۰۰ - Rank$ برای نشان دادن شباهت غیر صفر به SRR استفاده می شود حتی زمانی که T هیچ عبارت پرس و جو را شامل نمی شود تا نشان دهد که سند کامل نتیجه باید دارای برخی از اصطلاحات پرس و جو باشد زیرا در بین ۱۰ نتیجه برتر رتبه بندی شده است .در ادامه برخی از مهم ترین الگوریتم های ادغام را شرح میدهیم: .(Rasolofo et al, ۲۰۰۳)

الگوریتم TS (Title Scoring): این الگوریتم شباهت بین درخواست کاربر و عنوان هر SRR را با استفاده از فرمول محاسبه می کند) یعنی T عنوان است (و نتایج را به ترتیب نزولی این شباهت ها رتبه بندی می کند. .(Rasolofo et al, ۲۰۰۳)

الگوریتم(Snippet Scoring) SS : این الگوریتم مشابه الگوریتم TS است با این تفاوت که عنوان با قطعه جایگزین می شود. .(Rasolofo et al, ۲۰۰۳)

الگوریتم **TSS۱**(روش ۱ برای ترکیب عنوان بندی و امتیاز بندی قطعه): این الگوریتم به شرح زیر عمل می کند .برای هر SRR ، اگر عنوان آن شامل حداقل یک عبارت از پرس و جو باشد ، شباهت آن با Sim(Title, q)محاسبه می شود؛ در غیر این صورت ، اگر قطعه آن حداقل دارای یک عبارت از پرس و جو باشد ، شباهت آن توسط Sim(Snippet, q) محاسبه می شود ، سپس نتایج به ترتیب نزولی این شباهت ها رتبه بندی می شوند .این الگوریتم بر اساس تجربیاتی که نشان می دهد الگوریتم TS عملکرد بهتری نسبت به الگوریتم SS دارد ، عنوان را بر قطعه ترجیح می دهد.(Rasolofo et al, ۲۰۰۳).

الگوریتم TSS۲(رویکرد ۲ برای ترکیب عنوان بندی و امتیازدهی قطعه.): در این الگوریتم ، شباهت SRR مجموع وزنی امتیاز عنوان و قطعه قطعه آن است .وزن بیشتری به اولی (۰,۹) نسبت به آخری (۰,۱) داده می شود. .(Rasolofo et al, ۲۰۰۳)

چندین تغییر در الگوریتم های اساسی فوق نیز معرفی شده است. زمانی که چندین SRR دارای شباهت های مشابه هستند .واریاسیون دوم محاسبه می شود. .(Rasolofo et al, ۲۰۰۳)

ادغام بر اساس رتبه های محلی نتایج

در این بخش ، چندین الگوریتم ادغام نتایج را معرفی می کنیم که اساساً بر اساس رتبه های محلی نتایج بازیابی شده است .این الگوریتم ها را می توان به چهار دسته زیر طبقه بندی کرد: (Rasolofo et al., ۲۰۰۳).

.۱- روشهای مبتنی بر .Round-Robin این روشها از لیست نتایج هر موتور جستجوی جزیی در هر دور به ترتیب خاصی یک نتیجه را می گیرند. .(Rasolofo et al., ۲۰۰۳)

۲- روشهای تبدیل شباهت این روش ها رتبه های محلی را به شباهت ها تبدیل می کند تا بتوان از تکنیک های ادغام مبتنی بر شباهت استفاده کرد. .(Rasolofo et al., ۲۰۰۳)

۳-روش های رای گیری این روشها هر موتور جستجوی جزیی را به عنوان یک رای دهنده و هر نتیجه را به عنوان یک نامزد در انتخابات در نظر می گیرند .تکنیک های رأی گیری بیشتر برای موتورهای فرا جویشگر مناسب است که موتورهای جستجوی جزیی آنها در مجموعه اسناد خود همپوشانی قابل ملاحظه ای دارند. .(Rasolofo et al., ۲۰۰۳)

۴-روشهای مبتنی بر یادگیری ماشین این نوع روش بر اساس داده های آموزشی یک رتبه کلی برای هر نتیجه در لیست نتایج ادغام شده می آموزد. .(Rasolofo et al., ۲۰۰۳)

فرض میکنیم که N تا موتور جستجوی جزیی بصورت {S۱…SN} داریم که برای ارزیابی یک پرس و جوی q مشخص شده استفاده میشوند و همچنین RLi = (Ri۱, Ri۲,…) لیستی از نتایج بازگشتی از Si برای پرس و جوی q است.(Rasolofo et al, ۲۰۰۳)

روشهای مبتنی بر ROUND-ROBIN

چندین نوع استراتژی ادغام مبتنی بر ROUND-ROBIN وجود دارد که می توان آنها را در دو الگوریتم زیر خلاصه کرد. .(Rasolofo et al, ۲۰۰۳)

۱-الگوریتم SimpleRR: این روش ساده شامل دو مرحله است .در مرحله اول ، موتورهای جستجوی انتخاب شده را به صورت دلخواه انتخاب کنید. در مرحله دوم ، نتایج را از لیست نتایج این موتورهای جستجو در تعدادی تکرار یا دور بردارید و نتایج را به همان ترتیب که نتایج گرفته شده مرتب کنید. در هر دور ، بر اساس ترتیب موتورهای جستجو در مرحله ۱ ، نتیجه بعدی که هنوز از هر RLi گرفته نشده است را بگیرید. .(Rasolofo et al, ۲۰۰۳)

این روند تا زمانی که تمام لیست نتایج خالی شود ، تکرار می شود .در صورت تمام شدن لیست نتایج ، روند ROUND-ROBIN با لیست نتایج باقی مانده ادامه می یابد. این روش ساده ادغام بعید است عملکرد خوبی داشته باشد زیرا همه نتایج را با رتبه محلی یکسان احتمال مشابهی را مرتبط می داند و این واقعیت را نادیده می گیرد که مفید بودن موتورهای جستجوی مختلف انتخاب شده معمولاً برای یک پرس و جو متفاوت است. .(Rasolofo et al, ۲۰۰۳)

۲-الگوریتم PriorityRR: این روش الگوریتم SimpleRR را با اولویت دادن به موتورهای جستجو که دارای رتبه های بالاتر در مرحله انتخاب موتور جستجو هستند ، بهبود می بخشد به عبارت دیگر ، الگوریتم PriorityRR با الگوریتم SimpleRR تنها در نحوه انتخاب موتورهای جستجو متفاوت است ، یعنی اولی موتورها را به ترتیب نزولی نمرات رتبه بندی خود انتخاب میکند در حالی که دومی از ترتیب تصادفی استفاده می کند. .(Rasolofo et al, ۲۰۰۳)

توجه داشته باشید که الگوریتم PriorityRR تفاوت بین نمرات موتور جستجو را در نظر نمی گیرد (یعنی فقط از اطلاعات سفارش استفاده می شود). در زیر یک نسخه تصادفی از الگوریتم PriorityRR است. ما این روش را الگوریتم RandomRR می نامیم. .(Rasolofo et al, ۲۰۰۳)

۳- الگوریتم RandomRR: به یاد بیاورید که روش انتخاب موتور جستجوی MRDD ابتدا تعیین می کند که از هر موتور جستجوی جزیی چند نتیجه برای یک پرس و جو مشخص شده به دست آورد تا دقت بازیابی را به حداکثر برساند. فرض کنید تعداد نتایج مورد نظر از هر موتور جستجوی اجزای انتخاب شده بازیابی شده و لیست نتایجRLN ، ...، RL۱بدست آمده است. .(Rasolofo et al, ۲۰۰۳)

فرض کنید n تعداد کل نتایجی است که هنوز انتخاب نشده اند وni نتیجه هنوز در RLi هستند.

روشهای مبتنی بر شباهت

میتوان از تابع زیر برای تبدیل یک رتبه محلی به یک مقدار شباهت استفاده کرد:

$$Rank\_Sim(rank) = ۱ - \frac{rank - ۱}{num\_of\_retrieved\_docs}$$

این تابع شباهت رتبه ۱ را به نتایج برتر و با مقدار بالاتر نسبت میدهد؛ شباهت نتایج دیگر به رتبه های محلی نتایج و تعداد کل نتایج بازیابی شده بستگی دارد. مشاهده می شود که این تابع شباهت بالاتری را به همان نتیجه رتبه بندی شده از موتور جستجو که نتایج بیشتری را بازیابی کرده است ، اختصاص می دهد. (Lee, J,۱۹۹۷)

تابع شباهت در DWISE به شکل زیر است:(Lee, J,۱۹۹۷)

$$sim(rank) = ۱ - (rank - ۱) * \frac{rs_{min}}{m * rs_i}$$

برای پرس و جوی q داده شده، rs_i نمره رتبه بندی موتور جستجویSi است. rs_{min}در بین تمام موتورهای جستجو که برای q انتخاب شده اند ، رتبه موتور جستجو با کمترین رتبه است و rank رتبه محلی نتیجه R از Si است. (Lee, J,۱۹۹۷)

برای پرس و جو q داده شده ، روش SAFE شامل سه مرحله زیر است: (Lee, J,۱۹۹۷)

۱-هنگام بررسی اسناد موجود در SD (Si) به عنوان بخشی از CSD مجموعه اسناد ، شباهت جهانی بین qو هر سند در SD (Si) را با استفاده از یک تابع شباهت جهانی محاسبه کنید. به عبارت دیگر ، آمار مجموعه مانند فرکانس سند برای CSD برای محاسبه وزن اصطلاحات اسناد در SD (Si) استفاده می شود(. Lee, J,۱۹۹۷)

۲- تعیین کنید که اسناد موجود در SD (Si) باید در بین همه اسناد موجود در Si رتبه بندی شوند.در این قسمت دو مورد وجود دارد:۱) هیچ اسنادی در SD (Si) در لیست نتایج رتبه بندی شده RLi که توسط Si برای q بازگردانده می شود ، یعنی $SD(Si) \cap RLi \neq \emptyset$ ۲) SD (Si) ∩ RLi =∅ ظاهر نمی شود. .
(Lee, J,۱۹۹۷)

در مورد اول ، فرض بر این است که اسناد برگشتی در RLi قبل از همه اسناد در SD (Si) رتبه بندی می شوند و اسناد در SD (Si) به طور یکنواخت بین رده های همه اسناد در Si نسبت به q توزیع می شوند. در مورد دوم ، اسناد در RLi SD (Si) مانند RLi و اسناد باقی مانده در SD (Si) به طور یکنواخت رتبه بندی می شوند(. Lee, J,۱۹۹۷)

۳-شباهت های جهانی اسناد در RLi را با برازش منحنی برآورد کنید. به طور خاص ، SAFEرابطه بین شباهت اسناد نمونه و رتبه های تخمینی آنها را با استفاده از رگرسیون خطی در موارد زیر تعیین می کند: (Lee, J,۱۹۹۷)

$$sim(d) = m * f(rank(d)) + e$$

جایی که sim (d) شباهت جهانی یک سند نمونه برداری شده را در SD (Si) نشان می دهد و rank(d) رتبه تخمینی آن در بین اسناد موجود در Si است m .و e دو ثابت هستند و () f یک تابع برای ترسیم رتبه بندی سند در توزیع های مختلف است. (Lee, J,۱۹۹۷)

۳-۵-مشکلات فراجویشگرها:

همانطور که در بخشهای قبلی مورد بحث قرار گرفت ، پیشرفتهای زیادی در جهت یافتن راه حلهای کارآمد و دقیق برای مشکل پردازش پرس و جوها در محیط موتورهای فرا تحقیق انجام شده است .با این حال ، به عنوان یک مسئله ر حال ظهور ، بسیاری از مشکلات برجسته هنوز باید حل شوند .در این بخش ، ما چند چالش ارزشمند در این زمینه را لیست می کنیم. (Meng et al., ۱۹۹۹b, ۲۰۰۲)

۱-یکپارچه سازی سیستم های محلی با استفاده از تکنیک های مختلف نمایه سازی : استفاده از تکنیک های نمایه سازی مختلف در سیستم های محلی مختلف می تواند تأثیر جدی بر سازگاری شباهت های محلی داشته باشد .مشاهده دقیق می تواند نشان دهد که استفاده از تکنیک های مختلف نمایه سازی در واقع می تواند بر دقت برآورد هر یک از سه جزء نرم افزار (یعنی انتخاب پایگاه داده ، انتخاب اسناد و ادغام نتایج) تأثیر بگذارد.(Meng et al., ۱۹۹۹b, ۲۰۰۲)

مطالعات جدیدی باید انجام شود تا دقیقتر بررسی شود که چه تاثیری بر جای می گذارد و چگونه می توان بر این تأثیر غلبه یا کاهش داد. مطالعات قبلی تا حد زیادی بر عملکردهای مختلف شباهت محلی و طرح های وزن دهی محلی متمرکز شده است.(Meng et al., ۱۹۹۹b, ۲۰۰۲)

۲-یکپارچه سازی سیستم های محلی که از انواع مختلف پرس و جوها پشتیبانی می کند (به عنوان مثال ، پرس و جوهای بولی در مقابل پرس و جوهای فضای بردار): بیشتر بحث های ما در اینجا بر اساس پرس و جوهایی در مدل فضای بردار است . از آنجا که ممکن است از روشهای بسیار متفاوتی برای رتبه بندی اسناد برای پرس و جوهای بولی و برداری استفاده شود ، هنگام ادغام سیستمهای محلی که از پرس و جوهای بولی و فضای بردار پشتیبانی می کند ، احتمالاً با مشکلات جدی روبرو خواهیم شد (Meng et al., ۱۹۹۹b, ۲۰۰۲).

۳-کشف دانش در مورد موتورهای جستجوی جزیی: بسیاری از سیستم های محلی مایل به ارائه طراحی و اطلاعات آماری کافی در مورد سیستم های خود نیستند. آنها چنین اطلاعاتی را اخصوصی می دانند .با این حال ، بدون اطلاعات کافی در مورد یک سیستم محلی ، برآورد مفید بودن سیستم محلی با توجه به یک پرس و جو مشخص ممکن است به طور دقیق انجام نشود .(Meng et al., ۱۹۹۹b, ۲۰۰۲)

یک راه حل احتمالی برای این معضل ایجاد ابزارهایی است که بتوانند در مورد یک سیستم محلی در مورد اصطلاحات نمایه سازی استفاده شده و اطلاعات آماری خاصی در مورد این اصطلاحات و همچنین عملکرد شباهت مورد استفاده از پرس و جوها اطلاعات کسب کنند؛ از این ابزارهای یادگیری یا کشف دانش می توان نه تنها برای افزودن موتورهای جستجوی جزیی جدید به یک موتور فرا جویشگر، بلکه تشخیص ارتقاء عمده یا تغییرات سیستم های جزیی موجود نیز استفاده کرد. (Meng et al., ۱۹۹۹b, ۲۰۰۲).

۴-روشهای موثرتر ادغام نتایج را توسعه دهید : تا کنون ، اکثر روشهای ادغام نتیجه ای که تحت ارزیابی تجربی گسترده ای قرار گرفته اند ، روشهایی هستند که برای ادغام داده ها پیشنهاد شده اند. این روشها ممکن است در محیط موتور فراجویشگر که پایگاههای داده موتورهای جستجوی جزیی مختلف یکسان نیستند نامناسب باشد .(Meng et al., ۱۹۹۹b, ۲۰۰۲)

روشهای جدیدی که ویژگیهای ویژه محیط موتورهای فرا جویشگر را در نظر می گیرند باید طراحی و ارزیابی شوند. یکی از ویژگیهای خاص این است که وقتی یک سند توسط یک موتور جستجو بازیابی نمی شود ، ممکن است به این دلیل باشد که آن سند توسط موتور جستجو نمایه نمی شود (Meng et al., ۱۹۹۹b, ۲۰۰۲).

۵-همکاری مناسب بین یک موتور فرا جویشگر و سیستم های محلی را مطالعه کنید :دو روش فوق العاده برای ساختن موتور فراجویشگر وجود دارد. یکی این است که یک رابط کاربری را بر روی موتورهای جستجوی جزیی خودمختار اعمال کنید. در این مورد ، نمی توان انتظار همکاری از این سیستم های محلی را داشت. مورد دیگر این است که از سیستم های محلی برای پیوستن به یک موتور فرا جویشگر دعوت شود. (Meng et al., ۱۹۹۹b, ۲۰۰۲)

در این مورد ، توسعه دهنده موتور فراجویشگر ممکن است شرایطی را تعیین کند ، مانند اینکه چه تابع های شباهتی باید استفاده شود و چه اطلاعاتی در مورد پایگاه داده های جزیی باید ارائه شود ، که باید برای پیوستن یک سیستم محلی به موتور فراجویشگر کافی باشد. یک مسأله جالب این است که دستورالعمل هایی در مورد اینکه چه اطلاعاتی از سیستم های محلی برای تسهیل ساخت یک موتور فرا جویشگر مفید است ارائه شود. توسعه دهندگان موتورهای جستجو ممکن است از چنین دستورالعمل هایی برای طراحی یا ارتقاء موتورهای جستجوی خود استفاده کنند . (Meng et al., ۱۹۹۹b, ۲۰۰۲)

۶-ترکیب تکنیک های نمایه سازی و وزن دهی جدید برای ساخت موتورهای جستجوی بهتر: برخی از تکنیک های جدید نمایه سازی و وزن دهی اصطلاحات برای موتورهای جستجو برای اسناد HTML ایجاد شده است. به عنوان مثال ، برخی از موتورهای جستجو (به عنوان مثال ، Googleو Webor از اصطلاحات لنگر در یک صفحه وب برای فهرست بندی صفحه وب استفاده می کنند که توسط URL مرتبط با لنگر پیوند داده می شود. منطق این است که وقتی نویسندگان صفحات وب یک پیوند به صفحه وب دیگر p اضافه می کنند ، در برچسب لنگر توضیح صفحه p را به URL آن اضافه می کنند. این توضیحات که برای آن لینک و صفحه نوشته میشود خیلی به ما برای فهمیدن محتویات آن صفحه کمک میکند. (Meng et al., ۱۹۹۹b, ۲۰۰۲)

به عنوان مثال دیگر ، برخی از موتورهای جستجو وزن یک عبارت را با توجه به موقعیت آن در صفحه وب و نوع فونت آن محاسبه می کنند. مثلا اگر یک عبارت در عنوان صفحه ظاهر شود وزن بیشتری دارد. روش مشابهی نیز در AltaVista ، HotBotو Yahoo استفاده می شود. گوگل وزن بیشتری را به واژه هایی با فونت بزرگتر یا پررنگ اختصاص می دهد. مشخص است که همزمانی و مجاورت اصطلاحات تأثیر قابل توجهی بر ارتباط اسناد دارد. یک مشکل جالب این است که چگونه می توان این تکنیک های جدید

را در کل فرایند بازیابی و نمایندگان پایگاه داده ادغام کرد تا موتورهای جستجوی بهتر ساخته شوند (Meng et al., ۱۹۹۹b, ۲۰۰۲).

۷-بهبود اثربخشی فراجویشگرها: اکثر تکنیک های موجود پایگاه داده ها و اسناد را بر اساس شباهت های بین پرس و جو و اسناد موجود در هر پایگاه داده رتبه بندی می کنند. شباهت ها بر اساس تطابق شرایط در پرس و جو و اسناد محاسبه می شود. (Meng et al., ۱۹۹۹b, ۲۰۰۲)

مطالعات در زمینه بازیابی اطلاعات نشان می دهد که وقتی پرس و جوها تعداد زیادی اصطلاح دارند ، همبستگی بین اسناد بسیار مشابه و اسناد مربوطه وجود دارد به شرطی که توابع شباهت مناسب و طرح های وزن دهی اصطلاحاتی مانند تابع کسینوس و فرمول وزن tfw* idfw استفاده شود. (Meng et al., ۱۹۹۹b, ۲۰۰۲)

با این حال ، برای پرسش های کوتاه ، معمولی در محیط اینترنت ، همبستگی فوق ضعیف است .دلیل این است که برای یک پرس و جو طولانی ، اصطلاحات در پرس و جو زمینه ای را برای یکدیگر فراهم می کند تا به معنای واضح کردن معانی اصطلاحات مختلف کمک کند .در یک پرس و جو کوتاه ، معنای خاص یک اصطلاح اغلب نمی تواند به درستی مشخص شود؛ به طور خلاصه ، یک سند مشابه برای یک پرس و جو کوتاه ممکن است برای کاربری که پرس و جو را ارسال کرده است مفید نباشد ، زیرا ممکن است واژه های منطبق معانی متفاوتی داشته باشند .واضح است که همین مشکل برای موتورهای جستجو نیز وجود دارد. (Meng et al., ۱۹۹۹b, ۲۰۰۲)

برای رفع این مشکل باید روش هایی تدوین شود .در زیر ایده های امیدوار کننده ای وجود دارد .اول ، اهمیت یک سند را که با پیوندهای بین اسناد به عنوان مثال ، (PageRankو Authority) تعیین شده است ، با شباهت سند با یک پرس و جو تعیین کنید .دوم ، پایگاه های داده را با مفاهیم مرتبط کنید؛ هنگامی که یک پرس و جو توسط موتور فراجویشگر دریافت می شود ، ابتدا به تعدادی از مفاهیم مناسب نگاشت می شود و سپس پایگاه های داده مرتبط با مفاهیم ترسیم شده برای انتخاب پایگاه داده استفاده می شود .مفاهیم مرتبط با پایگاه داده/پرس و جو برای ارائه برخی زمینه ها برای اصطلاحات در پایگاه داده/پرس و جو استفاده می شود.(Meng et al., ۱۹۹۹b, ۲۰۰۲)

۸- تصمیم بگیرید که اجزای نرم افزاری موتور فراجویشگر را در کجا قرار دهید :یکی از موضوعاتی که ما در مورد آن بحث نکرده ایم این است که اجزای جزیی فراجویشگر باید در کجا قرار بگیرند. به عنوان مثال ، به جای داشتن انتخابگر پایگاه داده در سایت جهانی ، می توانیم آن را در همه سایت های محلی توزیع کنیم .نماینده هر پایگاه داده محلی نیز می تواند به صورت محلی ذخیره شود. در این سناریو ، هر درخواست کاربر برای انتخاب پایگاه داده به تمام سایتهای محلی ارسال می شود (Meng et al., ۱۹۹۹b, ۲۰۰۲).

سپس هر سایت مفید بودن پایگاه داده خود را در رابطه با پرس و جو تعیین می کند تا مشخص شود آیا موتور جستجوی محلی آن باید برای پرس و جو فراخوانی شود یا خیر. اگرچه این انتخابگر پایگاه داده هزینه ارتباطی بالاتری را متحمل می شود ، اما مزایای جذابی نیز دارد .(Meng et al., ۱۹۹۹b, ۲۰۰۲)

اول ، برآورد مفید بودن پایگاه داده اکنون می تواند به صورت موازی انجام شود. دوم: با ذخیره نمایندگان پایگاه داده به صورت محلی ، مسئله مقیاس پذیری بسیار کمتر از زمانی که از انتخاب پایگاه داده متمرکز استفاده می شود ، اهمیت کمتری پیدا می کند.اجزای دیگر مانند انتخاب کننده سند نیز ممکن است مکان های جایگزین داشته باشند. .(Meng et al., ۱۹۹۹b, ۲۰۰۲)

منابع فارسی

[۱] علیرضا یاری ، بررسی جویشگرهای متنی،تهران: پژوهشکده ارتباطات و فناوری اطلاعات(مرکز تحقیقات مخابرات ایران)، ۱۳۹۴

[۲] علیرضا یاری ، تحلیل و ارائه موضوعات راهبردی موتور جستجوی بومی،تهران: پژوهشکده ارتباطات و فناوری اطلاعات(مرکز تحقیقات مخابرات ایران)، ۱۳۹۶

منابع لاتین

[۱] Arzanian, B., Akhlaghian, F., & Moradi, P. (۲۰۱۰, January). A multi-agent based personalized meta-search engine using automatic fuzzy concept networks. In *۲۰۱۰ Third International Conference on Knowledge Discovery and Data Mining* (pp. ۲۰۸-۲۱۱). IEEE.

[۲] Arguello, J. (۲۰۱۱). *Federated search for heterogeneous environments* (Doctoral dissertation, Yahoo! Research).

[۳] Arguello, J. (۲۰۱۷). Aggregated search. *Foundations and Trends in Information Retrieval*, *۱۰*(۵), ۳٦۵-۵۰۲.

[۴] Erfanmanesh, M. A., & Didegah, F. (۲۰۱۲). Evaluating Function of Persian Search Engines on the Web Using Correspondence Analysis. *International Journal of Information Science and Management (IJISM)*, *۱۰*(۲), ۷۷-۸۷.

[۵] Feng, J., Gu, J., & Zhou, Z. (۲۰۱۳). An Improved Ranking Aggregation Method for Meta-Search Engine. In *Multimedia and Ubiquitous Engineering* (pp. ۱۹۳-۲۰۰). Springer, Dordrecht.

[۶] Gulli, A., & Signorini, A. (۲۰۰۵, May). Building an open source meta-search engine. In *Special interest tracks and posters of the ۱٤th international conference on World Wide Web* (pp. ۱۰۰٤-۱۰۰۵).

[۷] Gupta, D., & Singh, D. (۲۰۱٦, August). MetaFusion: An efficient metasearch engine using genetic algorithm. In *۲۰۱٦ Ninth International Conference on Contemporary Computing (IC۳)* (pp. ۱-٦). IEEE.

[۸] Hassanpour, H., & Zahmatkesh, F. (۲۰۱۲). An adaptive meta-search engine considering the user's field of interest. *Journal of King Saud University-Computer and Information Sciences*, *۲٤*(۱), ۷۱-۸۱.

[۹] Jansen, B. J., Spink, A., & Koshman, S. (۲۰۰۷). Web searcher interaction with the Dogpile. com metasearch engine. *Journal of the American Society for Information Science and Technology*, *۵۸*(۵), ۷٤٤-۷۵۵.

[۱۰] Jadidoleslamy, H. (۲۰۱۲). Search result merging and ranking strategies in meta-search engines: a survey. *Int. J. Comput. Sci*, *۹*, ۲۳۹-۲۵۱.

[۱۱] Jamshidi, M., Haji, M., Kamankesh, M. R., Daghineh, M., & Shaltooki, A. A. (۲۰۱۹). A Multi-Criteria Ranking Algorithm Based on the VIKOR Method for Meta-Search Engines. *JOIV: International Journal on Informatics Visualization*, *۳*(۳), ۲٤۸-۲۵٤.

[١٢] Keyhanipour, A. H., Moshiri, B., Piroozmand, M., & Lucas, C. (٢٠٠٦, July). Webfusion: Fundamentals and principals of a novel meta search engine. In *The ٢٠٠٦ IEEE International Joint Conference on Neural Network Proceedings* (pp. ٤١٢٦-٤١٣١). IEEE.

[١٣] Kumar, N. (٢٠١٧, August). Document Clustering Approach for Meta Search Engine. In *IOP Conference Series: Materials Science and Engineering* (Vol. ٢٢٥, No. ١, p. ٠١٢٢٩١). IOP Publishing.

[١٤] Meng, W., Yu, C., & Liu, K. L. (٢٠٠٢). Building efficient and effective metasearch engines. *ACM Computing Surveys (CSUR)*, *٣٤*(١), ٤٨-٨٩.

[١٥] Raval, V., & Kumar, P. (٢٠١٢, March). SEReleC (Search Engine Result Refinement and Classification)-a Meta search engine based on combinatorial search and search keyword based link classification. In *IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM-٢٠١٢)* (pp. ٦٢٧-٦٣١). IEEE.

[١٦] Shu, B., & Kak, S. (١٩٩٩). A neural network-based intelligent metasearch engine. *Information Sciences*, *١٢٠*(١-٤), ١-١١.

[١٧] Shokouhi, M., & Si, L. (٢٠١١). Federated search. *Foundations and Trends in Information Retrieval*, *٥*(١), ١-١٠٢.

[١٨] Tayal, D. K., Jain, A., Dimri, N., & Gupta, S. (٢٠١٥). MetaSurfer: a new metasearch engine based on FAHP and modified EOWA operator. *International Journal of System Assurance Engineering and Management*, *٦*(٤), ٤٨٧-٤٩٩.

[١٩] Tazehkandi, M. Z., & Nowkarizi, M. (٢٠٢٠). Evaluating the effectiveness of Google, Parsijoo, Rismoon, and Yooz to retrieve Persian documents. *Library Hi Tech*.

[٢٠] Vijaya, P., & Chander, S. (٢٠١٨). Metasearch Engine: A Technology for Information Extraction in Knowledge Computing. In *Knowledge Computing and its Applications* (pp. ٢٠٩-٢٣٣). Springer, Singapore

[٢١] Wang, M., Li, Q., & Lin, Y. IM Search: An Agent-based Personalized Metasearch Engine

[٢٢] N. Craswell. (٢٠٠٠) *Methods in Distributed Information Retrieval*. Ph.D. thesis, School of Computer Science, The Australian National University, Canberra, Australia, ٢٠٠٠.

[٢٣] M. Shokouhi and L. Si. Federated Search. *Foundations andTrends in Information Retrieval*, ٢٠١١ (in press).

[٢٤] Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Y. Halevy. (٢٠٠٨) Google's deep Web crawl. In *Proc. ٣٤th Int. Conf. on Very Large Data Bases*, pages ١٢٤١–١٢٥٢, ٢٠٠٨.

[٢٥] Wang and D. DeWitt. (٢٠٠٤) Computing PageRank in a distributed internet search engine system. In *Proc. ٣٠th Int. Conf. on Very Large Data Bases.*

[٢٦] A.Laender,B. Ribeiro-Neto,A.da Silva, and J.Teixeira. (٢٠٠٢)Abrief survey ofWeb data extraction tools. *ACMSIGMOD Rec.*, ٣١(٢):٨٤–٩٣, ٢٠٠٢.

[٢٧] S. Raghavan and H. Garcia-Molina. (٢٠٠١) Crawling the hidden Web. In *Proc. ٢٧th Int. Conf. on Very Large Data Bases*, pages ١٢٩–١٣٨, ٢٠٠١.

[٢٨] W Meng, CT Yu – (٢٠١٠) Advance Metasearch Engine

[٢٩] Y. Lu, W. Meng, L. Shu, C. Yu, and K. Liu. (٢٠٠٥) Evaluation of result merging strategies for metasearch engines. In *Proc. ٦th Int. Conf. on Web Information Systems Eng.*, pages ٥٣–٦٦, ٢٠٠٥.

[٣٠] Y. Rasolofo, D. Hawking, and J. Savoy. (٢٠٠٣) Result merging strategies for a

current news metasearcher. *Information Proc. & Man.*,

[٣١] J. Lee. (١٩٩٧). ١٩٩٧ Analyses of multiple evidence combination. In *Proc. ٢٠th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages ٢٦٧–٢٧٦, ١٩٩٧.

[٣٢] G.Salton. (١٩٨٩) *AutomaticText Processing:TheTransformation,Analysis, and Retrieval of Information by Computer*. AddisonWesley, ١٩٨٩.

[٣٣] W. B. Croft, D. Metzler, and T. Strohman. (٢٠٠٩) *Search Engines: Information Retrieval in Practice*. Addison-Wesley, ٢٠٠٩.

[٣٤] R. A. Baeza-Yates and B. Ribeiro-Neto. (١٩٩٩) *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., ١٩٩٩.

[٣٥] G.Salton. (١٩٨٩) *AutomaticText Processing:TheTransformation,Analysis, and Retrieval of Information by Computer*. AddisonWesley, ١٩٨٩.

[٣٦] S. E. Robertson and S.Walker. (١٩٩٩) Okapi/Keenbow at TREC-٨. In *Proc. The ٨th Text Retrieval Conf.*, pages ١٥١–١٦١, ١٩٩٩.

[٣٧] H.Turtle and J. Flood. (١٩٩٥) Query evaluation: Strategies and optimizations. *Information Proc.&Man*, ٣١:٨٣١–٨٥٠, ١٩٩٥.

[٣٨] E. M. Voorhees and D. K. Harman. (٢٠٠٥) *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, ٢٠٠٥.

[٣٩]Lawrence, S. & Lee Giles, C. Inquirus - The NECIMetasearch Engine. In the Proceedings of the SeventhInternational World Wide Web Conference, Brisbane,Australia, Elsevier Sience, ١٩٩٨.