ارزیابی خودکار جویشگرهای متنی مبتنی بر تجمیع آرا در حوزه وب فارسی

فرزانه شعله[۱]، معصومه عظیم‌زاده[۱]، محمدمهدی یدالهی[۱]، اکبر میرزایی[۱]، مژگان فرهودی[۱]

[۱]گروه سکوهای فناوری اطالعات، پژوهشکده فناوری اطالعات، مرکز تحقیقات مخابرات ایران، تهران
{f.shoeleh, azim_ma, mm.yadollahi, ak.mirzaei , farhoodi }@itrc.ac.ir

چکیده

امروزه با توجه به رشد روز افزون صفحات وب و استفاده فراوان کاربران از جویشگرها به منظور بازیابی اطلاعات از وب، ارزیابی جویشگرها به ویژه در حوزه‌های بومی بسیار مورد توجه قرار گرفته است. از این رو، در هر کشوری از جمله ایران، جویشگرهایی با تمرکز بر روی حوزه های خاصی از وب توسط محققین این عرصه به وجود آمده اند که همگی در تلاش اند عملکرد مناسبی در رقابت با جویشگرهای همه منظوره مانند گوگل داشته باشند. از این رو، بحث ارزیابی جویشگرها به یکی از مباحث مطرح و با اهمیت در حوزه بازیابی اطلاعات تبدیل شده است. در این مقاله، روشی مبتنی بر تجمیع آرا به منظور ارزیابی خودکار جویشگرهای متنی با نام اختصاری VAWSEE ارائه گردیده است. تمرکز اصلی سیستم ارزیابی VAWSEE بر روی حوزه وب فارسی بوده و با توسعه روشی نوین برای شباهت سنجی مبتنی بر محتوا با الهام از راهکارهای تشخیص تقلب سعی در ارزیابی جویشگرهای متنی در این حوزه را دارد. روش پیشنهادی با ارزیابی انسانی بر روی مجموعه پرس‌وجوهای جمع آوری شده از کاربران محک زده شده است و میزان همبستگی دو روش خودکار و انسانی مورد بررسی قرار گرفته است. نتایج بدست آمده از این آزمایشات حاکی از مناسب و قابل اتکا بودن روش پیشنهادی است.

کلمات کلیدی

بازیابی اطلاعات، جویشگر، ارزیابی جویشگرهای متنی، تجمیع آرا.

۱- مقدمه

با توجه به اهمیت جویشگرها به عنوان درگاه ورود ۸۰ درصد کاربران به وب و همچنین نقش آنها در بازیابی اطلاعات از وب، مبحث ارزیابی جویشگرها از اهمیت بالایی برخوردار است. مهمترین فاکتور رضایت کاربران از یک موتور جستجو، مرتبط بودن نتایج ارائه شده توسط آن می‌باشد. بنابراین ارزیابی این معیار از اهمیت بالایی برخوردار است که با استفاده از شاخص دقت سنجیده می‌شود. روش‌های مطرح در حوزه تخمین دقت از دیدگاه‌های مختلفی قابل طرح هستند. وظیفه‌ی موتور جستجو پاسخ‌گویی به پرس‌وجو-هایی می‌باشد که کاربران به آن ارسال می‌کنند. در پاسخ به هر پرس‌وجو، موتور جستجو لیستی از اسناد را به صورت رتبه‌بندی‌شده، در اختیار کاربر قرار می‌دهد. هر چقدر که اسناد موجود در لیست بازگردانده شده دارای کیفیت مناسب‌تری باشند گفته می‌شود که موتور جستجو از دقت بالاتری برخوردار می‌باشد.

روشهای مختلفی برای ارزیابی جویشگرها وجود دارد که با توجه به عامل ایجاد مجموعه قضاوت به دو دسته مهم روشهای ارزیابی خودکار [۲۱-۳] و انسانی [۲] [۲۲] [۲۳] [۲۴] قابل تقسیم می‌باشند. در واقع یک سیستم ارزیابی جویشگر شامل سه بخش اصلی است که عبارتند از واحد تولید پرس و جوها، واحد قضاوت و واحد محاسبه معیارها. ایجاد واحد قضاوت مهمترین و هزینه‌برترین بخش در یک سیستم ارزیابی است. مبنای خودکار یا انسانی بودن روش ارزیابی نیز به نحوه پیاده‌سازی بخش قضاوتهای مرتبط برمی-گردد.

موضوع مهمی که باید توجه داشت آن است که عمدتا روشهای ارزیابی انسانی از دقت بالاتری برخوردار است. اما به دلیل هزینه‌بر بودن و زمان‌بر بودن بکارگیری آنها، استفاده از روشهای خودکار نقش مهمی در خودارزیابی

جویشگرها دارد به نحوی که توسعه‌دهندگان آنها به منظور امکان ارزیابی تاثیر تغییرات اعمال شده و نسخه‌های جدید نسبت به نسخه‌های پیشین از این روشها به صورت گسترده‌ای استفاده می‌کنند.

روش‌های ارزیابی خودکار برای هر دو دسته پرس و جوهای پیمایشی [۳] [۵] [۱۲] [۱۹] [۲۰] [۲۱] و اطلاعاتی [۴]، [۱۰]، [۱۱] [۱۵] کاربرد دارند. از منظر مکانیزم ایجاد مجموعه قضاوت و داده‌های ورودی مورد استفاده آنها روش‌های ارزیابی خودکار جویشگرها، در دو دسته مهم روشهایی مبتنی بر بازخورد کاربر [۱۳] [۱۴] [۱۵] [۱۶] [۱۸] و روش‌های مبتنی بر تجمیع آرا و رتبه بندی مجدد نتایج موتورهای جستجو[۱] [۳] [۵] [۸] [۱۲] [۱۹] [۲۰] [۲۱] قابل ارائه هستند. روش‌های مبتنی بر بازخورد کاربران از اطلاعاتی نظیر کلیک کاربر یا رفتار ضمنی کاربر حین جستجو استفاده می‌کنند. این دسته از روشها با توجه به بهره‌برداری از مشارکت غیرمستقیم کاربران از دقت بالایی برخوردار هستند اما محدودیتهایی برای استفاده و پیاده‌سازی آنها وجود دارد. به عنوان مثال جویشگر باید دارای بازار قابل توجهی باشد تا بتوان حجم داده قابل توجهی جمع‌آوری نمود و همچنین امکان دسترسی به داده‌های مورد نیاز نیز وجود داشته باشد. دسته دوم روشها تحت عنوان " روش‌های مبتنی بر تجمیع آرا و رتبه بندی مجدد نتایج موتورهای جستجو" از نتایج رتبه‌بندی شده توسط جویشگرها به عنوان داده ورودی اولیه استفاده می‌کند. دو سیاست متداول استفاده شده در این زمینه عبارتند از استفاده از اجماع جویشگرها و توافق آنها روی نتایج مشابه و یا رتبه‌بندی مجدد نتایج جویشگرها(مجتمع شده در یک مخزن) با استفاده از شاخصها و مکانیزمهای متنوع. برای پرس و جوهای پیمایشی نیز علاوه بر روشهای مذکور دسته دیگری از روش‌ها تحت عنوان روش‌های مبتنی بر آیتم‌های شناخته شده [۴]، [۱۰]، [۱۱] نیز مطرح هستند. با توجه به اینکه به ازای هر پرس و جوی پیمایشی دقیقا یک جواب مشخص وجود دارد از برخی منابع موجود در وب نظیر دانش سازماندهی شده در فهرست‌ها و دایره‌المعارف‌ها[۱] نیز می‌توان برای ایجاد مجموعه قضاوت استفاده نمود.

هدفی که در این مقاله دنبال می‌شود پیاده‌سازی یک روش ارزیابی خودکار برای ارزیابی جویشگرهای بومی و مقایسه آنها با چندجویشگر بین-المللی نظیر گوگل و بینگ می‌باشد. واضح است که به دلیل عدم دسترسی به جزئیات پیاده‌سازی این سامانه‌ها، ارزیابی آنها به روش جعبه سیاه انجام می-شود. همچنین به منظور اطمینان از صحت عملکرد روش پیاده‌سازی شده همبستگی روش با روش ارزیابی انسانی نیز مقایسه می‌شود. روش پیشنهادی در دسته دوم روشها یعنی رتبه‌بندی مبتنی بر تجمیع آرا قرار می‌گیرد. واضح است که دسته اول روشها که مبتنی بر بازخورد کاربران می‌باشد کاربردی برای هدف این مقاله نخواهد داشت چرا که در این حوزه نیاز به داده‌های واقعی کاربران در سطح وسیع وجود دارد که با توجه به اینکه جویشگرهای بومی در نقطه آغازین راه هستند و هنوز به صورت گسترده مورد استفاده قرار نگرفته‌اند، برای آنها کاربرد نخواهد داشت.

هدف از این پژوهش ارائه راهکاری نوین مبتنی بر تجمیع آرا برای ارزیابی خودکار جویشگرهای متنی در حوزه وب فارسی است که با نام اختصاری VAWSEE معرفی می‌گردد. مهمترین ویژگی روش پیشنهادی که آن را با سایر روشهای ارائه شده تا کنون متمایز می‌کند، عدم بایاس شدن آن به روش رتبه‌بندی مورد استفاده جویشگرها می‌باشد. چرا که از اجماع جویشگرها براساس شباهت نتایج URL ها یا محتوای اصلی نتایج ارائه شده توسط آنها استفاده می‌کند.

به همین منظور در ادامه ابتدا مروری بر روشهای ارزیابی خودکار جویشگر متنی خواهیم داشت. سپس راهکار پیشنهادی ارائه شده و نهایتا نتایج و تحلیل‌های مرتبط ارائه خواهد شد.

۲- کارهای مرتبط

در این بخش ابتدا روشهای ارزیابی خودکار مبتنی بر بازخورد کاربر را مرور می‌کنیم، سپس به بررسی روشهای مبتنی بر تجمیع آرا و رتبه‌بندی مجدد خواهیم پرداخت. در مقاله‌ی [۱۳] اطلاعات موجود در فایل کلیک کاربران به عنوان نشانه‌هایی دال بر ترجیحات نسبی اسناد نسبت به یکدیگر در نظر گرفته شده‌اند که متفاوت با کارهای قبلی است که کلیک شدن اسناد را به عنوان مرتبط بودن مطلق آن‌ها در نظر می‌گیرد. روش مطرح در [۱۴] روی پرس و جوهای پیمایشی تمرکز نموده است. در این مقاله برای ایجاد مجموعه قضاوت از ترکیبی از رفتار ضمنی کاربر و کلیک روی نتایج استفاده گردیده است. پاسخ متناظر با یک پرس‌وجوی پیمایشی، سندی در لیست نتایج متناظر با آن است که بیشتر از سایر اسناد موجود در این لیست، توسط کاربران کلیک شده است. یعنی کاربران چنین نتیجه‌ای را، نتیجه‌ی مناسبی برای پرس‌وجوی خود یافته‌اند و علاقه‌مند به مشاهده آن بوده‌اند. برای شناسایی پرس و جوهای پیمایشی روش مقاله‌ی [۱۵] بکاررفته و به این منظور از توزیع کلیک‌های کاربران بر روی اسناد مختلف موجود در نتایج بازگشتی متناظر با هر پرس‌وجو بهره گرفته شده است. در مقاله‌ی [۱۵]، روشی برای دسته‌بندی پرس‌وجوها به سه دسته‌ی اطلاعاتی، تراکنشی[۲] و پیمایشی[۳] است. در این مقاله فرض شده است که در حالتی که اکثریت کلیک‌های کاربران، متمرکز بر روی یک سند باشند، آن‌گاه چنین پرس‌وجویی از نوع پیمایشی است. روش مطرح شده در مقاله‌ی [۱۶] یک روش نیمه خودکار است. در این روش با ارائه لیستی از ترکیب نتایج جویشگرهای مختلف، تنها از نیروی انسانی خواسته می‌شود مطابق با رفتار معمولی که در استفاده از یک جویشگر دارد، بر روی اسنادی که در لیست نتایج هستند و آن‌ها را پسندیده است کلیک نماید. بنابراین نحوه‌ی به دست آوردن داده‌های کلیک به‌صورت بی طرفانه و از طریق واسط یکسان است.

در مقاله [۱۸] روشی دیگر برای ارزیابی خودکار در نظر گرفته شده است که اصطلاحا از بازخوردهای ضمنی برای این موضوع استفاده می‌کند. در این مقاله ارزیابی سه سطحی برای تعیین مرتبط بودن نتایج استفاده شده است. معیارهای مرتبط بودن عبارتند از رونوشت گرفتن از صفحه[۴]، اضافه کردن به صفحات مورد علاقه[۵]، نشانه‌گذاری صفحه[۶]، چاپ کردن[۷]، ذخیره کردن[۸] و Scroll کردن صفحه[۹]. انجام هر یک از این اعمال در حقیقت می‌تواند یک بازخورد ضمنی از سمت کاربر تلقی گردد در غیر این‌صورت سند بازگشتی می‌تواند به عنوان نامرتبط در نظر گرفته شود. روش Reference Count [۸] مبتنی بر تجمیع آرا است. ابتدا پرس‌وجو به هریک از جویشگرها تحت بررسی داده می‌شود. از نتایج بازگشتی پرس‌وجوی فعلی، n سند دارای بالاترین رتبه در نظر گرفته می‌شوند. سپس، بررسی می‌شود که هر سند در نظر گرفته شده برای هر جویشگر، چند بار در مجموعه اسناد در نظر گرفته-شده از سایر جویشگرها تکرار شده است. مجموع تکرارهای برای هر سامانه‌ی بازیابی اطلاعات به عنوان امتیاز آن سامانه لحاظ می‌گردد. جویشگرها بر

اساس امتیازات محاسبه شده مورد رتبه‌بندی قرار می‌گیرند. در روش ارزیابی خودکار انتساب تصادفی مرتبط [۵]، نویسنده ایده انتخاب تصادفی اسناد به عنوان اسناد مرتبط را برای هر پرس و جو مطرح ساخته و از آن در ارزیابی اسناد و پرس و جوهای موجود در TREC استفاده کرده است.

روش AWSEEM از جمله روشهای تجمیع رتبه در مقاله‌ی [۱۲] معرفی گردیده است. در این روش پرس‌وجوها به هشت جویشگر از جمله یاهو و امام‌ان ارسال شده‌اند. سپس ۲۰۰ نتیجه‌ی اول هر یک از این جویشگرها به یک مجموعه‌ی نهایی افزوده شده و با توجه به میزان شباهت‌شان با پرس و جوی ارسالی مورد رتبه‌بندی قرار گرفته‌اند. از جمله نقاط ضعف این روش این است که موتورهای جستجو در هنگام رتبه‌بندی اسناد پارامترهای متعددی را مدنظر قرار می‌دهند، در حالی‌که روش AWSEEM تنها به محتوای متنی صفحات وب توجه می‌کند. این روش به عنوان مبنای کار مقاله [۳] قرار گرفته است. برای بهبود این روش پارامترهای دیگری نظیر PageRank و AlexaRank برای انجام ارزیابی به این روش افزوده شدند. روش مطرح شده در مقاله‌ی [۱۹] روشی بر اساس تجمیع رتبه‌ها است. تجمیع رتبه‌ها روشی است که در آن رتبه‌بندی‌های صورت گرفته توسط چند روش مختلف با یکدیگر به نحوی ترکیب می‌گردند تا رتبه‌بندی که حاصل می‌شود دارای کیفیت بالاتری باشد. در روش مطرح شده در این مقاله مهمترین مرحله گام یادگیری است که در آن قواعد رتبه‌بندی به کمک چهار روش ۱) الگوریتم PageRank، ۲) معیار شباهت دودویی (بازیابی دودویی)، ۳) مدل فضای برداری و ۴) بازخوردهای ضمنی کاربران استخراج می‌شود. روش [۲۰] از سه روش مختلف امتزاج داده‌ها برای رتبه‌بندی سامانه‌های بازیابی اطلاعات استفاده می‌کند. از روش‌های امتزاج داده‌ها برای ایجاد مجموعه اسناد شبه مرتبط استفاده می‌شود. در این روش‌ها نتایج بازگشتی جویشگرهای مختلف، با استفاده از تکنیک‌های مختلف با یکدیگر ادغام می‌گردند و تعدادی از اسناد با بالاترین رتبه‌ها در رتبه‌بندی نهایی به عنوان اسناد شبه مرتبط در نظر گرفته می‌شوند. سپس این اسناد شبه مرتبط برای ارزیابی موثر بودن سامانه‌های بازیابی مورد استفاده قرار می‌گیرند.

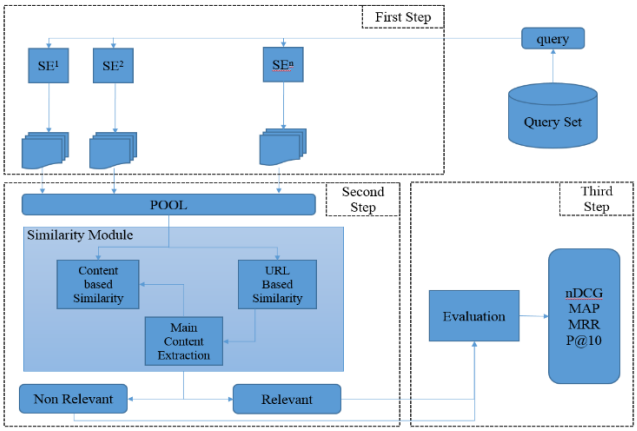
۳- راهکار پیشنهادی

هدف از این پژوهش ارائه راهکاری نوین مبتنی بر تجمیع آرا برای ارزیابی خودکار جویشگرهای متنی در حوزه وب فارسی است. در **Error! Reference source not found.** شمای کلی روش پیشنهادی با نام اختصاری VAWSEE ارائه شده است. بطور کلی روش پیشنهادی سه گام اساسی دارد:

۱) گام اول: بازیابی نتایج از n جویشگر متنی متفاوت و ساخت مجموعه نتایج بازگشتی به ازای هر پرس و جو

۲) گام دوم: مشخص نمودن سطح مرتبط بودن هر یک از نتایج بازگشتی از جویشگرها

۳) گام سوم: محاسبه معیارهای ارزیابی متفاوت



شکل (۱) شمای کلی سیستم ارزیابی خودکار جویشگرهای متنی مبتنی بر تجمیع آرا

واضح است که گام اساسی هر سیستم ارزیابی با هدف ارزیابی دقت جویشگرها، گام دوم است که در آن به ایجاد مجموعه قضاوت می‌پردازیم. در این مقاله، سعی شده است راهکاری نوین مبتنی بر تجمیع آرا برای مشخص نمودن مرتبط بودن یا نبودن یک نتیجه و همچنین سطح مرتبط بودن آن ارائه گردد. برای این منظور، به ازای هر پرس‌وجو، نتایج بازگشتی از جویشگرهای متفاوت با یکدیگر مقایسه شده و نتیجه‌ای که توسط بیشتر از m جویشگر (m<n) بازیابی شده است، به عنوان نتیجه مرتبط در نظر گرفته خواهد شد.لازم به ذکر است که مقایسه نتایج از طریق شباهت سنجی در دو سطح آدرس صفحه وب و همچنین محتوای آن صفحه صورت خواهد گرفت.

در مولفه شباهت‌سنجی، نتایج در سطح اول از منظر آدرس URL مورد مقایسه قرار گرفته و نتیجه‌ای که URL های آن بعد از نرمالیزه شدن مشابه با URL نرمال شده از m-1 جویشگر دیگر باشد، به عنوان نتیجه مرتبط در استخر نتایج مرتبط به ازای پرس‌وجو مربوطه در نظر گرفته خواهد شد. منظور از نرمال کردن آدرس یک صفحه وب، حذف عناصری است که اغلب در اکثر آدرس‌ها تکرار خواهند شد. از جمله این عناصر می‌توان به www، https، http، index.htm و غیره اشاره نمود. پس از این گام، آدرس نرمال شده را به دو قسمت دامنه-آدرس و پارامترها تقسیم کرده و برای مقایسه هر قسمت همانند دو رشته با آن‌ها برخورد می‌شود. لازم به ذکر است که در قسمت پارامترها امکان جابجا بودن پارامترها نیز لحاظ شده است، بدین صورت که آدرس‌های مشابه ممکن است در ترتیب پارامتر ارسالی به یک وب سرویس متفاوت باشند، در این حالت این آدرس‌ها یکسان تشخیص داده خواهند شد.

در سطح دوم، نتایج از منظر محتوا مورد شباهت سنجی قرار خواهند گرفت. بدین صورت که پس از محاسبه شباهت مبتنی بر آدرس اگر نتیجه‌ای دارای مورد مشابهی نباشد به مولفه استخراج محتوا اصلی ارسال شده تا صفحه وب معادل با آن دانلود و محتوای اصلی آن استخراج و ذخیره گردد. برای این منظور از بسته کد باز گوس[۱۰] استفاده شده و متن اصلی یک صفحه وب که حاوی اطلاعات مفید است به عنوان فایل محتوای اصلی آن استخراج و ذخیره می‌گردد. از این رو، فایل محتوای اصلی یک صفحه وب شامل بخشی از آن صفحه وب خواهد بود که بار اطلاعاتی داشته باشد.

پس از استخراج فایل محتوای اصلی صفحات وب، با استفاده از روش های تشخیص تقلب برای هر یک از فایل‌های محتوای اصلی یک مشخصه بصورت «اثر انگشت»[۱۱] استخراج می گردد. با مقایسه اثر انگشت‌های

استخراج شده به ازای هر یک از فایل‌های محتوای اصلی مرتبط با نتایج بازگشتی از جویشگرهای متنی می‌توان نتایج را از منظر محتوایی مورد مقایسه قرار داد. اگر اثر انگشت مربوط به محتوای اصلی دو نتیجه بازگشتی از یک مقدار پیش فرض مانند θ بیشتر باشد، دو نتیجه مزبور در روش پیشنهادی به عنوان دو نتیجه مشابه در نظر گرفته می‌شود و چنانچه آن نتیجه حاوی m-1 نتیجه مشابه در لیست نتایج بازگشتی از جویشگرهای دیگر باشد، راهکار پیشنهادی ما آن نتیجه را به عنوان یک نتیجه مرتبط در نظر خواهد گرفت. به منظور درک بهتر از روند کارکرد الگوریتم مطرح شده، در شکل (۱) الگوریتم کلی ارزیابی را نشان داده شده است.

۱. ابتدا فایل پیکربندی که قبلا توسط کاربر سیستم ارزیابی خودکار ایجاد شده است خوانده و تنظیمات لازم برای پارامترهای موجود در سیستم صورت می‌گیرد.
۲. به ازای هر پرس‌وجو ، نتایج جویشگرهای مختلف جمع آوری شده و یک مخزن که حاوی تمامی نتایج است ایجاد می‌شود.
۴. به ازای نتایج هر جویشگر مراحل زیر دنبال می‌شود:
۴-۱. به ازای هر نتیجه مراحل زیر دنبال می‌شود:
۴-۱-۱. در مولفه شباهت‌سنج، هر نتیجه با نتایج بازگشتی از سایر جویشگرها براساس آدرس URL آن‌ها مورد مقایسه و شباهت‌سنجی قرار می‌گیرد.
۴-۱-۱-۱. آدرس URL ها نرمال شده و موارد معمول آن حذف می‌گردد.
۴-۱-۱-۲. آدرس URL ها به دو قسمت دامنه-آدرس و پارامترها تقسیم می‌شود. و هر قسمت به صورت جداگانه مورد مقایسه قرار می‌گیرد.
۴-۱-۱-۳. در صورت وجود تشابه بین آدرس URL ها، به مقدار آرای نتایج مربوطه یکی اضافه خواهد شد.
۴-۱-۱-۴. نتایجی که مقدار آرای آنها کمتر از m است، برای دانلود به مولفه استخراج محتوای اصلی ارسال می‌شود. در غیر اینصورت به گام شش خواهیم رفت.
۴-۱-۲. در مولفه استخراج محتوای اصلی ، عمل دانلود و استخراج محتوای اصلی توسط بسته گوس بصورت موازی انجام می‌شود. از این رو، به ازای هر نتیجه، یک فایل متنی حاوی محتوای اصلی ذخیره خواهد شد.
۵. در مولفه شباهت‌سنج مبتنی بر محتوا، به ازای هر فایل متنی حاوی محتوای اصلی مربوط به یک نتیجه، مراحل زیر انجام می‌گردد.
۵-۱. الگوریتم غربال کردن بر روی فایل متنی اعمال شده و به ازای هر فایل برداری از اثر انگشت‌های مربوط به آن صفحه بدست می‌آید.
۵-۲. به ازای هر دو فایل ضریب شباهت جاکارد مابین بردار اثر انگشت‌ها آنها به عنوان میزان شباهت نتایج مرتبط به آن دو فایل، محاسبه می‌شود.
۵-۳. در صورتی که شباهت یک نتیجه با نتیجه دیگر از حد آستانه تعیین شده θ بالاتر باشد، به مقدار آرای نتایج مربوطه یکی اضافه خواهد شد.
۶. مرتبط بودن هر یک از نتایج محاسبه می‌گردد.
۶-۱. در صورتی که مقدار آرای یک نتیجه بزرگتر یا مساوی m باشد، نتیجه مورد نظر به عنوان یک نتیجه مرتبط با پرس‌وجو علامت‌گذاری می‌شود.
۷. معیارهای ارزیابی نظیر فراخوانی، دقت و nDCG برای هر جویشگر به ازای این پرس‌وجو محاسبه می‌گردد.

شکل (۱) الگوریتم سیستم ارزیابی خودکار جویشگرهای متنی

در راهکار پیشنهادی، به منظور شباهت سنجی مبتنی بر محتوای دو نتیجه از الگوریتم تشخیص تقلب بنام الگوریتم «غربال کردن»[۱۲] معرفی شده در [۲۵] الهام گرفته شده است. به منظور استفاده از این الگوریتم، محتوای اصلی دو نتیجه به عنوان دو فایل متنی مجزا در نظر گرفته شده و در واقع به عنوان ورودی‌های الگوریتم تشخیص تقلب خواهند بود. در این الگوریتم، ابتدا، هر فایل متنی پالایش شده و به اصطلاح نرمال می‌گردد. بصورتیکه کارکترهای خاص نظیر فاصله، خط جدید، کاما، نقطه، و غیره از متن حذف می‌شوند. سپس، برای هر کدام از فایل‌ها رشته‌های n-gram ای استخراج می‌گردد. هر رشته n-gram را با یک الگوریتم معروف درهم سازی کرده و تبدیل به عدد خواهیم کرد. از این رو، به ازای هر فایل محتوای اصلی مربوط به یک نتیجه، آرایه‌ای از اعداد خواهیم داشت. با در نظر گرفتن یک پنجره با سایز w بر روی این آرایه حرکت خواهیم کرد و در هر حرکت مینیمم عدد موجود در آن پنجره را به عنوان نماینده در نظر خواهیم گرفت. با این راهکار، آرایه اعداد متناظر با درهم سازی n-gram ها تبدیل به آرایه‌ای از اعداد متناظر با کمترین مقدار در هم سازی می‌شوند. آرایه به دست آمده در واقع به نوعی معرف شباهت سنجی محلی (به اندازه سایز w) در فایل متنی خواهد بود. به منظور درک بهتر، خوانندگان می‌تواند به مثال‌های ذکر شده در [۲۵] رجوع نمایند.

به منظور محاسبه میزان شباهت مبتنی بر محتوا مابین دو نتیجه، طبق فرمول (۱) ضریب شباهت جاکارد بین دو بردار اثر انگشت حاصل شده از اعمال الگوریتم غربال کردن بر روی فایل محتوای اصلی مربوط به دو نتیجه مورد نظر، محاسبه می‌شود.

$$similarity\ (d_1, d_2) = \frac{|fingerprints(d_1) \cap fingerprints(d_1)|}{|fingerprints(d_1) \cup fingerprints(d_1)|} \quad (۱)$$

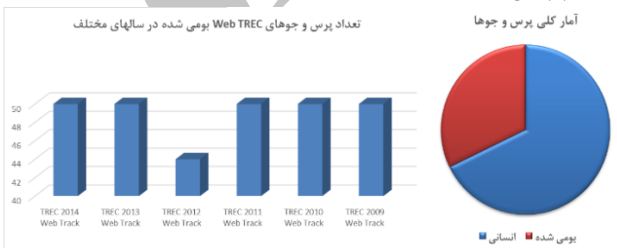
در روش پیشنهادی، پس از مشخص شدن شباهت بین نتایج بازگشتی از جویشگرهای متفاوت سطح ارتباط آن‌ها در سه سطح مختلف: «نامرتبط» (۰)، «کمی مرتبط» (۱) و «مرتبط» (۲) سطح‌بندی می‌شود. از آنجایی که جویشگر گوگل در حال حاضر بزرگترین و بهترین جویشگر موجود در دنیا می‌باشد و اکثر جویشگرها با آن مقایسه شده و همچنین سعی در رسیدن به کارکرد مشابه به آن را دارند، ۵ نتیجه اول این جویشگر بصورت پیش فرض «کمی مرتبط» در نظر می‌گیریم و باقی نتایج به همراه نتایج سایر جویشگرهای دیگر، اگر طبق مولفه شباهت سنج روش پیشنهادی شبیه بودند به عنوان «مرتبط» و اگر شبیه نبودند به عنوان «نامرتبط» در نظر گرفته خواهند شد.

٤- نتایج

مجموعه دادگان: یکی از عوامل بسیار تاثیرگذار بر کیفیت ارزیابی جویشگرها مجموعه پرس‌وجوها می‌باشد. به این منظور در این پروژه ایجاد مجموعه پرس‌وجوهای کنترل شده به صورت انسانی و هم ایجاد مجموعه پرس و جوها بر اساس رفتار کاربر یا لاگ جویشگرها مورد توجه قرار گرفته است. در حالت اول علاوه بر اینکه پرس‌وجوها کنترل شده و پالایش می‌شوند به صورت انسانی نیز برچسب می‌خورند بنابراین می‌توان تحلیل‌های دقیقتر و بیشتری را ارائه داد. منتها با توجه به اینکه ایجاد تعداد زیادی پرس و جوی

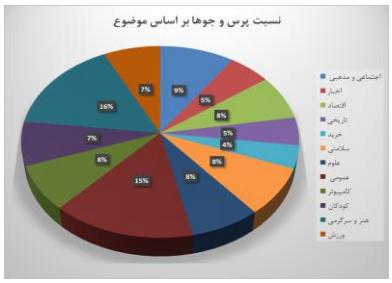
کنترل شده هزینه بر و زمانبر است همچنین کاربر پرس و جوها را به صورت هدفمند تولید می‌کند بنابراین استفاده از پرس‌وجوهای لاگ جویشگر می‌تواند مشکلات ذکر شده در بالا را کاهش دهد. در این پژوهش هر دو این روش‌ها جهت ایجاد مجموعه پرس‌وجوها مورد نظر می‌باشد.

در این مرحله لازم بود روش خودکار ارزیابی جویشگر متنی مورد ارزیابی قرار گیرد و میزان همبستگی آن با روش انسانی ارزیابی شود. بنابراین نیاز به استفاده از پرس‌وجوهایی بود که در ارزیابی انسانی مورد استفاده قرار گرفته است. بنابراین حدود ۴۰ عدد از پرس‌وجوها ایجاد شده انتخاب گشته و در ارزیابی انسانی جهت ارزیابی روش خودکار مورد استفاده قرار گرفت. این پرس‌وجوها به صورت جمع‌سپاری توسط خود کاربران ایجاد شده بودند. پرس‌وجوهای انسانی کنترل شده در دو دسته ایجاد شدند. یک بخش پرس‌وجوهای پیشنهادی کاربران فارسی زبان بود و بخش دوم پرس‌وجوهای بومی شده هستند. پرس‌وجوهای بومی شده پرس‌وجوهایی هستند که از دادگان استاندارد TREC از سال ۲۰۰۹ تا ۲۰۱۴ اخذ شده و برای استفاده داخلی بومی سازی شده‌اند. در مجموع، ۱۱۲۲ پرس‌وجو ایجاد شد که ۸۲۹ عدد از آنها پرس‌وجوهای انسانی و مابقی مبتنی بر TREC ایجاد شده‌اند. در شکل (۳) این آمار ارائه شده است.

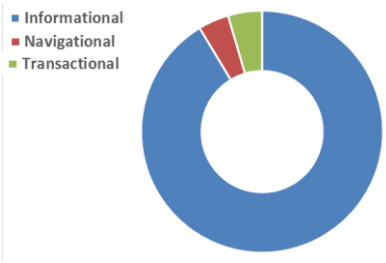


شکل (۳) آمار پرس‌وجوهای انسانی مبتنی بر منبع

پرس‌وجوهای تهیه شده از هفت نقطه نظر برچسب خوردند: ۱) نیاز اطلاعاتی یا هدف کاربر از پرس‌وجو. ۲) نوع پرس‌وجو. ۳) دسته موضوعی پرس و جو به این منظور ۱۲ دسته در نظر گرفته شد. ۴) نوع پرس و جو که می‌تواند هر یک از سه نوع اطلاعاتی، پیمایشی و تراکنشی باشد. ۵) منبع پرس‌وجو که نشان می‌دهد انسانی است یا بومی شده. ۶) زبان پرس‌وجو نشان دهنده فارسی، انگلیسی یا دوزبانه بودن است. ۷) تاریخ اضافه شدن به مجموعه پرس‌وجوها. شکل (۴) آمار دسته‌بندی موضوعی پرس‌وجوها را نشان می‌دهد. همانطور که این شکل نشان می‌دهد بیشتر پرس و جوی کاربران در حوزه هنر و سرگرمی و عمومی بوده است. بعد از آن بیشتر نیاز اطلاعاتی کاربران در حوزه اجتماعی و مذهبی بوده است. همچنین در شکل (۵) آمار پرس و جوها به تفکیک نوع دیده می‌شود. همانگونه که این شکل نشان می‌دهد آمار پرس‌وجوهای اطلاعاتی حدود ۹۰ درصد پرس‌وجوهای این مجموعه را در برمی‌گیرد. لازم به ذکر است گرچه روش پیشنهادی به منظور ارزیابی پرس‌وجوهای اطلاعاتی طراحی شده است، اما بدلیل تجمیع آرا بین جویشگرها می‌تواند بر روی ارزیابی جویشگرها به ازای پرس‌وجوهای پیمایشی نیز مفید باشد.



شکل (٤)آمار پرس‌وجوها به تفکیک موضوعی



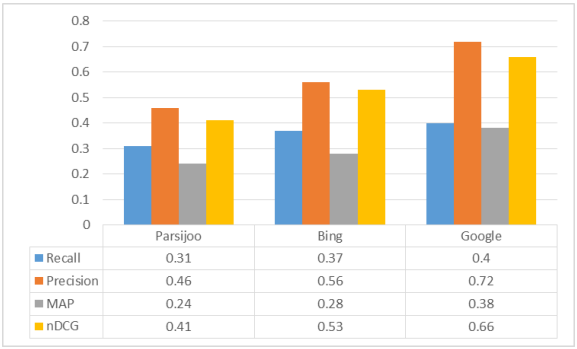
شکل (٥)آمار پرس و جوها به تفکیک نوع

آزمایشات و نتایج : به منظور ارزیابی روش ارائه شده، در این بخش آزمایشی ترتیب داده شده که در آن نتایج حاصل از ارزیابی خودکار دقت جویشگرهای متنی با نتایج حاصل از ارزیابی انسانی همین جویشگرها مورد مقایسه قرار گرفته است. بدین منظور، مجموعه‌ای متشکل از ۱۰ کاربر انسانی در نظر گرفته شده و به ازای هریک از آنها ۴ پرس‌وجوی مختلف برگرفته شده از مجموعه دادگان ذکر شده استخراج گردیده است. سپس این پرس‌وجوها در اختیار ده کاربر انسانی قرار گرفت تا به وسیله آنها، سه جویشگر گوگل، بینگ و پارسی‌جو را مورد ارزیابی قرار دهند. برای هر پرس‌وجو، ده نتیجه اول هر موتور جستجو دریافت شده و برای ارزیابی در اختیار کاربران قرار گرفته است. به منظور فراهم کردن امکان مقایسه نتایج ارزیابی خودکار و انسانی جویشگرها، پرس‌وجوهای انتخابی نیز به عنوان ورودی سامانه ارزیابی خودکار جویشگرهای متنی مبتنی بر تجمیع آرا مورد استفاده قرار گرفته شده است. به منظور تعیین میزان دقت نتایج بازگشتی از جویشگرها، در مقایسه‌ها جویشگرها از معیارهای اصلی Recall، Precision، MAP و همچنین nDCG محاسبه شده طبق فرمول (۲) استفاده شده است.

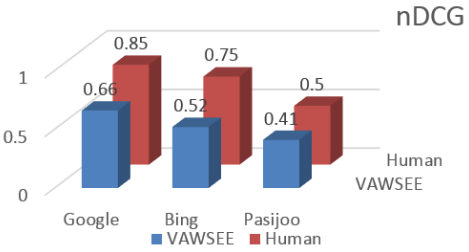
$$nDCG_p = \frac{DCG_p}{IDCG_p}$$
$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2^{i+1}} \qquad (۲)$$

در جدول (۱) نتایج حاصل از این مقایسه به ترتیب بر اساس معیارهای ذکر شده در بالا نمایش داده شده است. لازم به ذکر است که بدست اوردن مقدار دقیق Recall برای جویشگرها امری غیرممکن است زیرا به ازای یک پرس‌وجو اطلاعاتی مجموعه نتایج مرتبط موجود در دنیای وب بصورت کامل در اختیار ما نخواهد بود. از این رو، باید توجه داشت که مقدار محاسبه شده برای این معیار، مقداری تخمینی است. بدین صورت که مجموعه کل نتایج مرتبط برای یک پرس‌وجو را برابر با اجتماع تمامی صفحات مرتبط تشخیص داده شده از جویشگرهای مورد ارزیابی در نظر گرفته‌ایم.

جدول (۱) : نتایج حاصل از ارزیابی خودکار جویشگرهای متنی گوگل، بینگ و پارسی‌جو



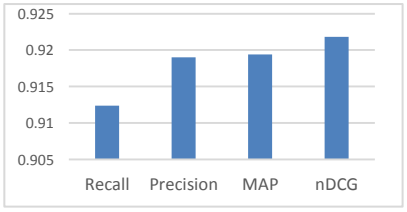
در شکل (۶) نتایج بدست آمده از مقایسه جویشگرهای متنی توسط سیستم ارزیابی خودکار در مقایسه با ارزیابی انسانی بر مبنای معیار nDCG به نمایش در آمده است. تفاوت عمده این معیار با سایر معیارها در این است که علاوه بر در نظر گرفتن سطوح مختلف ارتباط برای نتایج بازگشتی، جایگاه یک نتیجه را نیز مد نظر قرار میدهد. از این رو، این معیار به عنوان یکی از مهمترین معیارهای شناخته شده در ارزیابی سیستم‌های بازیابی اطلاعات بخصوص جویشگرها تبدیل شده است. با بررسی نتایج حاصل از ارزیابی جویشگرهای متفاوت، در خواهیم یافت که ارزیابی خودکار جویشگرهای متنی در اغلب موارد محافظه کارانه تر از نتایج به دست آمده از ارزیابی انسانی است. به عبارت دیگر، میزان دقت به دست آمده برای هر جویشگر متنی در ارزیابی خودکار عموماً پائین‌تر از میزان دقت به دست آمده در ارزیابی انسانی است. با این حال، نکته جالب توجهی که وجود دارد آن است که رتبه‌بندی جویشگرها از منظر دقت در هر دو ارزیابی خودکار و انسانی یکسان است.



شکل (٦) نتایج حاصل از ارزیابی خودکار و ارزیابی انسانی جویشگرهای متنی بر اساس معیار nDCG

به منظور تعیین میزان همبستگی مابین نتایج به دست آمده از ارزیابی خودکار و ارزیابی انسانی، از معیار همبستگی پیرسن در سطح کلی استفاده شده است. بصورتیکه، میزان همبستگی مابین مقادیر به دست آمده برای معیار nDCG به عنوان مهمترین معیار دقت در ارزیابی دقت جویشگر، در نظر گرفته شده است. نتایج به دست آمده برای همبستگی در شکل (۷) نشان داده شده است. این نتایج نشان می‌دهند که میزان همبستگی در سطح کلی مابین دو روش ارزیابی خودکار و انسانی برای هر ۴ معیار بیش از ۰.۹۱ است. از این‌رو، بدلیل وجود همبستگی بالا مابین دو روش ارزیابی، می‌توان ادعا نمود که نتایج به دست آمده از روش ارزیابی خودکار پیشنهاد شده در این مقاله مناسب و قابل اتکا است.

در روش پیشنهادی، جویشگرها به صورت جعبه سیاه در نظر گرفته شده‌اند و ارزیابی صرفا بر اساس خروجی آنها که همان صفحات بازگشتی به ازای یک پرس‌وجو و ترتیب آن‌ها صورت می‌گیرد. از این رو، انتظار می‌رود، محاسبه معیارها در ارزیابی خودکار از ارزیابی انسانی کمتر باشد. زیرا در ارزیابی انسانی، یک انسان با هوشمندی خود تشخیص خواهد داد که آیا یک نتیجه مرتبط با پرس‌وجوی آن است یا خیر. بخصوص در برخی موارد، مرتبط بودن را نیز ممکن از بصورت نسبی با سایر نتایج در نظر بگیرد. هرچند سعی شده است درآزمایشات طراحی شده این استقلال حفظ شود، اما نتایج خالی از ارتباط نیز نخواهد بود. در صحت آزمایی روش ارایه شده، مهم شباهت روش پیشنهادی در رتبه بندی جویشگرها با ارزیابی انسانی است. که این شباهت به وضوح در شکل‌های (۶) نشان داده شده است و همچنین بصورت آماری در شکل (۷) محاسبه و نمایش داده شده است. از این رو، میتوان ادعا نمود که روش پیشنهادی توانایی و عملکرد خوبی در نحوه رتبه بندی جویشگرهای متنی بخصوص در حوزه وب فارسی را دارد.



شکل (۷) میزان همبستگی نتایج حاصل از سیستم ارزیابی انسانی و ارزیابی خودکار ارائه شده

۵- نتیجه‌گیری

در این مقاله، با توجه به اهمیت موضوع مقایسه جویشگرهای متنی بومی روش ارزیابی خودکار جویشگرهای متنی مبتنی بر تجمیع آرا با نام اختصاری VAWSEE ارائه گردید. روش پیشنهادی در این مقاله شامل سه گام اساسی: ۱) بازیابی نتایج از n جویشگر متفاوت و ساخت مجموعه نتایج بازگشتی به ازای هر پرس‌وجو، ۲) مشخص نمودن سطح مرتبط بودن هر یک از نتایج بازگشتی از جویشگرها و ۳) محاسبه معیارهای ارزیابی است. در گام دوم، به منظور تشخیص سطح مرتبط بودن یا نبودن یک نتیجه راهکاری مبتنی بر تجمیع آرا مابین جویشگرها پیشنهاد شده است. در تجمیع آرا، نیاز است بتوان شباهت بین دو صفحه وب متناظر با دو نتیجه از دو جویشگر متنی متفاوت بدست آید. از این رو، در روش VAWSEE به منظور تشخیص شباهت مابین صفحات وب بازگردانده شده از جویشگرها، علاوه بر شباهت‌سنجی در سطح آدرس صفحات از شباهت سنجی در سطح محتوای اصلی صفحات نیز استفاده شده است. به منظور شباهت‌سنجی محتوایی دو صفحه وب، از یکی از معروفترین الگوریتم‌های تشخیص تقلب استفاده شده است. نتایج حاصل شده از روش پیشنهادی تحت یک آزمایش که حاوی مجموعه‌ای از پرس‌وجوهای گردآوری شده از کاربران است با ارزیابی انسانی که بر روی همین مجموعه پرس‌وجوها انجام شده است، محک زده شد. نتایج میزان همبستگی این روش با ارزیابی انسانی همگی بیانگر مناسب بودن روش پیشنهادی و همچنین قابل اتکا بودن نتایج آن جهت در مقایسه و ارزیابی جویشگرهای متنی است.

در ادامه می‌توان به بهینه‌سازی زمان اجرای روش پرداخت. یکی از مشکلات روش ارائه شده، زمان اجرای بالای آن است، زیرا تمامی صفحات به منظور شباهت سنجی محتوایی بایستی دانلود و ذخیره گردند. برای حل این مشکل علاوه بر استفاده از راهکارهای پیاده‌سازی الگوریتم بصورت موازی می‌توان راهکار پیشنهادی را در مولفه شباهت سنجی بهبود بخشید. بدین صورت که

بسته به نوع و دسته پرس‌وجو کنترل‌های متفاوتی در شباهت سنجی انجام داد. به عنوان مثال، در پرس‌و جوهای پیمایشی شباهت در سطح محتوا می‌تواند صورت نگیرد.

مراجع

[۱] س. موسوی ، م. عظیم زاده ، م. محمودی، ع. یاری ، **ارائه چارچوبی جامع و کارا برای ارزیابی موتورهای جستجوی فارسی**، هجدهمین کنفرانس ملی سالیانه انجمن کامپیوتر،تهران، اسفند ۱۳۹۱.

[۲] م. عظیم زاده، ش. سموری، ع. یاری، **بررسی و مقایسه کیفی موتورهای جستجو در حوزه وب فارسی**، هجدهمین کنفرانس ملی سالیانه انجمن کامپیوتر،تهران، اسفند ۱۳۹۱.

[3] R. Badie, M. Azimzadeh, M.S. Zahedi, S. Samuri, "Automatic evaluation of search engines: Using webpages' content, web graph link structure and websites' popularity" Seventh International Symposium on Telecommunications (IST2014), September 09-11, 2014.

[4] M. Mahmoudy, M.S. Zahedi, M. Azimzadeh, "Evaluating the retrieval effectiveness of search engines using Persian navigational queries", Seventh International Symposium on Telecommunications, September 09-11, 2014.

[5] I. Soboroff, C. Nicholas, and P. Cahan, "Ranking retrieval systems without relevance judgments," in 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001, pp. 66-73.

[6] S. P. Harter, "Variations in relevance assessments and the measurement of retrieval effectiveness," JASIS, vol. 47, pp. 37-49, 1996.

[7] A. Spink and H .Greisdorf, "Regions and levels: measuring and mapping users' relevance judgments," Journal of the American Society for Information science and Technology, vol. 52, pp. 161-173, 2001.

[8] S. Wu and F. Crestani, "Methods for ranking information retrieval systems without relevance judgments," in Proceedings of the 2003 ACM symposium on Applied computing, 2003, pp. 811-816.

[9] J. Callan, M. Connell, and A. Du, "Automatic discovery of language models for text databases," in ACM SIGMOD Record, 1999, pp. 479-4.۹۰

[10] A. Chowdhury and I. Soboroff, "Automatic evaluation of world wide web search services," in 25th annual international ACM SIGIR conference on Research and development in information retrieval, 2002, pp. 421-422.

[11] S. M. Beitzel, E. C. Jensen, A. Chowdhury, and D. Grossman, "Using titles and category names from editor-driven taxonomies for automatic evaluation," in Proceedings of the twelfth international conference on Information and knowledge management, 2003, pp. 17-23.

[12] F. Can, R. Nuray, and A.B. Sevdik, "Automatic performance evaluation of Web search engines," Information processing & management, vol.40, pp.495-514, 2004.

[13] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005, pp. 154-161.

[14] Y. Liu, Y. Fu, M. Zhang, S. Ma, and L. Ru, "Automatic search engine performance evaluation with click-through data analysis," in Proceedings of the 16th international conference on World Wide Web, 2007, pp. 1133-1134.

[15] Y. Liu, M. Zhang, L. Ru, and S. Ma, "Automatic query type identification based on click through information," in Information Retrieval Technology:2006, pp. 593-600.

[16] T. Joachims, "Evaluating Retrieval Performance Using Clickthrough Data," ed: Citeseer, 2003.

[17] G. Mood, "Boes, Introduction to the theory of statistics," McCraw-Hill Statistics Series, 1974.

[18] H. Sharma and B. J. Jansen, "Automated evaluation of search engine performance via implicit user feedback," in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval ,۲۰۰۵ ,pp. 649-650.

[19] R. Ali and M. S. Beg, "Automatic performance evaluation of web search systems using rough set based rank aggregation," in International Conference on Intelligent Human Computer Interaction, 2009,pp.344-358.

[20] R. Nuray and F. Can, "Automatic ranking of information retrieval systems using data fusion," Information processing & management, vol. 42, pp. 595-614, 2006.

[21] H. Sadeghi.,(2011). "Automatic Performance Evaluation of Web search Engines using judgements of Meta search Engines", Online Information Review,ISSN:1468-4527,Emerald Publishing Limited, pp.957-971.

[22] Tawileh W, Griesbaum J, Mandl T. Evaluation of five web search engines in Arabic language. Proceedings of LWA. (2010).

[23] Lewandowski, Dirk. Evaluating the retrieval effectiveness of Web search engines using a representative query sample. Journal of the Association for Information Science and Technology (2015).

[24] Bar-Ilan J, Levene M. A method to assess search engine results. Online Information Review 35(6),854-868. (2011).

[25] Schleimer, Saul, Daniel S. Wilkerson, and Alex Aiken. "Winnowing: local algorithms for document fingerprinting." international conference on Management of data. ACM, 2003.

زیرنویس‌ها

[1] Encyclopedia
[2] Transactional
[3] Navigational
[4] Copying
[5] Adding To Favorite
[6] Bookmarking
[7] Printing
[8] Saving
[9] Scrolling
[10] Goose
[11] Fingerprint
[12] Winnowing