کیفیت‌سنجی جویشگرهای متنی در بستر جمع‌سپاری

معصومه عظیم‌زاده[۱]، محمد مهدی اثنی‌عشری[۲]، مژگان فرهودی[۳]

[۱] گروه سکوهای فناوری اطلاعات، پژوهشکده فناوری اطلاعات، مرکز تحقیقات مخابرات ایران، تهران، ایران
azim_ma@itrc.ac.ir

[۲] دانشکده مهندسی کامپیوتر، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران
esnaashari@kntu.ac.ir

[۳] گروه سکوهای فناوری اطلاعات، پژوهشکده فناوری اطلاعات، مرکز تحقیقات مخابرات ایران، تهران، ایران
farhoodi@itrc.ac.ir

چکیده

جویشگرها یکی از مهمترین ابزارهای بازیابی اطلاعات هستند. اهمیت جویشگرها از آن جا نشات می‌گیرد که از مهمترین درگاههای دسترسی کاربران به وب بوده و باید دستیابی کاربران به اطلاعات مورد نیاز را در حجم انبوهی از اطلاعات تسهیل نمایند. با افـزایش روزافزون حجم اطلاعات وب و نیاز کاربران به دسترسی به آنها در کوتاهترین زمان ممکن، فرایند بازیابی و استخراج اطلاعات اهمیت ویژه‌ای یافته است. با توجه به نیاز به دسترسی کاربران به وب از طریق یک درگاه امن و بومی در سال‌های اخیر موضوع جویشگرهای بومی در کشور ما مورد توجه زیادی قرار گرفته است. به نحوی که چندین جویشگر بومی متکی به دانش و تجربه متخصصان داخلی و بر اساس فرهنگ و بوم ایرانی اسلامی گسترش پیدا کرده است. در راستای بهبود کیفیت و افزایش مخاطبین این جویشگرهای نوپا یکی از نیازمندیهای اساسی مانیتورینگ پیوسته و ارزیابی کیفیت نتایج آنها می‌باشد. در این راستا در این مقاله کیفیت جویشـگرهای بومی مبتنی بر بستر جمع‌سپاری ارزیابی شده که در آن محوریت ارزیابی متکی به روش ارزیابی انسانی است. نتایج ارزیابی برای دو جویشگر بومی پارسی‌جو و یوز در مقایسه با جویشگرهای بین‌المللی گوگل و بینگ به دست آمده است. برآیند نتایج ارزیـابی صـورت پذیرفته، رتبه‌بندی کیفی جویشگرها را به ترتیب به صورت گوگل، بینگ، پارسی‌جو و یوز نشان داده است.

کلمات کلیدی

جمع‌سپاری، جویشگر، ارزیابی کیفی.

۱- مقدمه

با رشد و ظهور اینترنـت و ارائـه خـدمات و اطلاعـات بـر بسـتر وب، امـروزه جویشگرها به عنوان شاهراه اطلاعاتی از اهمیت زیادی برخوردار هستند. یـک جویشگر خوب باید بتواند نیازمنـدیهای اطلاعـاتی کـاربران را بـه روش‌هـای هوشمندانه از میان انبوهی از اطلاعات موجود در وب که با ساختارها و فرمت-های متنوعی هستند تامین کند.

با توجه به آنکه کاربران معمولاً تنها به نتایج اول ارائه شده از جویشگرها توجه دارند [1]، بنابراین هر چه نتایج ارائه شده در این رتبه‌هـا از کیفیـت بـالاتری برخوردار باشند، میزان اقبال کاربران به ایـن جویشـگرها بیشـتر خواهـد بـود. فرآیند کیفیت‌سنجی جویشگرها می‌تواند نشان دهد که این بخش از نتایج هر جویشگر، تا چه حد توانسته نیازهای اطلاعاتی کاربران را مرتفع نماید. هـم از این روست که این فرایند، از اهمیت بسیار بالائی برخوردار است، چرا که نقاط ضعف و قوت جویشگرها را مشخص نموده و اجازه تصـحیح عملکـرد آنهـا را فراهم می‌آورد.

از سوی دیگر، امروزه در برخی کشورها مانند کشور ما به دلیل حساسیت روی حفاظت اطلاعات و استقلال در فضای سایبری، پرداختن به موضوع جویشگرهای بومی از اولویت بالایی برخوردار است. بنا به همین دلیل، مرکز تحقیقات مخابرات ایران، به عنوان یکی از متولیان دولتی توسعه جویشگرهای بومی، با برگزاری ارزیابی‌های مداوم، به پایش وضعیت جویشگرهای بومی از منظرهای مختلف پرداخته است. در این مقاله، به ارائه و تحلیل نتایج حاصل از ارزیابی کیفی این جویشگرها در مقایسه با جویشگرهای بین‌المللی بینگ و گوگل خواهیم پرداخت. روش مورد استفاده جهت دستیابی به این نتایج، ارزیابی انسانی مبتنی بر بستر جمع‌سپاری بوده است.

ادامه این مقاله به صورت زیر سازماندهی شده است. در بخش دوم، به مروری بر فعالیت‌های تحقیقاتی صورت پذیرفته در زمینه ارزیابی جویشگرهای متنی پرداخته شده است. بخش سوم به معرفی بستر جمع‌سپاری مورد استفاده اختصاص دارد. روال ارزیابی صورت پذیرفته و مجموعه دادگان مورد استفاده به منظور ارزیابی در بخش چهارم تشریح گردیده است. نتایج به دست آمده به همراه تحلیل آنها در بخش پنجم بیان شده است. بخش ششم به جمع‌بندی مقاله اختصاص دارد.

۲- کارهای مرتبط

به صورت کلی، روش‌های ارزیابی و کیفیت‌سنجی نتایج حاصل از جویشگرهای متنی را می‌توان به سه دسته تقسیم‌بندی نمود [2]:

- **ارزیابی‌های انسانی:** در این دسته از روش‌ها، که دقیق‌ترین و مناسب‌ترین روش برای ارزیابی نتایج جویشگرها محسوب می‌شوند، تعدادی از نیروهای انسانی به کار گرفته می‌شوند تا میزان ارتباط نتایج حاصل از جویشگرها به پرس‌وجوهای ارسالی را تعیین نمایند. روش‌های ارائه شده در [3-10,12,13] در این دسته جای می‌گیرند.
- **ارزیابی‌های خودکار:** با وجود آنکه ارزیابی انسانی از دقت بالایی برخوردار است، اما هزینه انجام بالایی دارد و به همین دلیل، انجام مداوم آن در بسیاری از موارد امکان‌پذیر نیست. در این گونه موارد، راهکار جایگزین استفاده از روش‌های ارزیابی خودکار است. در این روش‌ها، تعیین میزان ارتباط نتایج به پرس‌وجوها به صورت خودکار و به کمک روش‌هایی نظیر شباهت‌سنجی نتایج جویشگرهای مختلف صورت می‌پذیرد. برخی از روش‌های ارزیابی خودکار جویشگرهای متنی در [14,15,16,17,18,19,20] معرفی شده‌اند.
- **ارزیابی‌های ضمنی:** راهکار دیگری که می‌توان برای کیفیت‌سنجی نتایج جویشگرها مورد استفاده قرار داد، انجام ارزیابی به صورت ضمنی است. در این روش، رفتار کاربر در حین کار با جویشگرهای مختلف به کمک افزونه‌های مرورگر مورد بررسی قرار گرفته و بر اساس میزان توجه کاربر به هر نتیجه، میزان ارتباطی برای آن نتیجه لحاظ می‌شود. روش‌های ارائه شده در [21,22,23] در این دسته جای می‌گیرند. مشکل عمده این دسته از روش‌ها آن است که باید به نوعی کاربران را مجاب نمود که افزونه مرورگر مورد نظر را نصب نمایند. در عین حال، کاربری که افزونه را نصب نموده است، دیگر رفتار معمول خود را در مواجهه با جویشگرها نخواهد داشت و تا حدی به صورت تصنعی رفتار خواهد نمود.

با توجه به آنکه تمرکز این مقاله، بر ارزیابی انسانی جویشگرهای متنی است، در ادامه این بخش، تنها به معرفی پژوهش‌های پیشین انجام شده در این حوزه خواهیم پرداخت. روش کلی ارزیابی انسانی به این صورت است که تعدادی پرس و جو و نیاز اطلاعاتی[1] متناظر با آن در اختیار کاربران قرار می‌گیرد و از آنها خواسته می‌شود تا به ارزیابی نتایج بازگردانده شده توسط جویشگرها بپردازند. پارامترهای مهمی که در ارزیابی باید مورد توجه قرار گیرند عبارتند از: تعداد پرس و جوها، نیروی انسانی، تعیین جویشگرهای تحت آزمون و مشخص نمودن سطوح ارتباط (مابین نتایج و پرس‌وجوها). تعداد پرس‌وجوهایی که در پژوهش‌های مختلف به کار رفته بسیار متغیر است: از حدود ۱۰ تا ۲۰ پرس و جو در [3,4] تا ۲۵، ۵۰ و مواردی حتی بیشتر در [5,6]. منبع انتخاب پرس‌وجوها می‌تواند گستره‌ای از لاگ‌های موجود، افراد خبره و کاربران شرکت کننده در آزمون باشد [7]. نیروی انسانی مورد استفاده در ارزیابی‌ها اکثراً دانشجویان هستند. برای انتخاب جویشگرها معمولاً از بزرگترین و معروفترین آنها و گاهی اوقات جویشگرهای جدید یا خاص زبانه [8] استفاده می‌شود. در بیشتر کارها برای ارزیابی میزان مرتبط بودن سند از قضاوت دو یا سه سطحی و در برخی موارد از قضاوتهایی با تعداد سطوح مرتبط بودن بیشتر نیز استفاده شده است. در کارهای مرتبط تمهیداتی برای عدم قضاوت جانبدارانه کاربران در نظر گرفته شده است. به عنوان مثال در برخی از موارد واسط کاربری جویشگر به کاربر نمایش داده نشده و تنها نتایج بازگردانده شده برای قضاوت در اختیار کاربران قرار می‌گیرد [1]. همچنین به منظور جلوگیری از تأثیر رتبه‌بندی جویشگرها بر قضاوت کاربران، نتایج بازیابی شده به صورت تصادفی به آنها ارائه می‌شود.

ویسم تویل و همکاران [12] در سال ۲۰۱۰ به ارزیابی جویشگرهای بومی عربی زبان پرداختند. در این ارزیابی، جویشگرهای بومی عربی در مقایسه با گوگل امتیاز کمتری کسب نمودند. در پژوهش دیگری که در سال ۲۰۱۱ صورت پذیرفته است، در ارزیابی تنها از پرس‌وجوهای پیمایشی[2] استفاده شده است [4]. نتیجه حاصل از این پژوهش نشانگر برتری جویشگرهای گوگل و یاهو بوده است.

در زبان فارسی فعالیت‌های انجام شده در زمینه ارزیابی جویشگرها بسیار محدود است و اغلب آنها به ارزیابی جویشگرهایی در یک حوزه‌ی خاص پرداخته‌اند. به عنوان مثال در خصوص جستجو در یک دامنه خاص در مقاله [5]، توانایی شش جویشگر و شش فراجویشگر در پاسخ به پرس و جوهای حوزه داروشناسی مورد بررسی قرار گرفت. در این مقاله جویشگرهای یاهو و گوگل بهترین نتیجه را در بازیابی مستندات داروشناسی به دست آورده‌اند. از جمله فعالیت‌هائی که در آن ارزیابی جویشگرها مبتنی بر ویژگی‌ها صورت گرفته، مقاله [13] است که در آن، ۱۶ جویشگر فارسی بر اساس ویژگی‌های به دست آمده از سایت الکسا مورد ارزیابی قرار گرفتند که در آن سایت قطره بهترین نتیجه را در بر داشته است.

۳- معرفی بستر جمع‌سپاری مورد استفاده

جمع‌سپاری به ترکیبی از دو کلمه جمعیت و برونسپاری اطلاق می‌شود و به مفهوم انجام کاری با کمک جمعیتی از نیروهای انسانی است. این کار معمولاً از طریق فراخوان عمومی در اینترنت انجام می‌شود. ابزار مورد استفاده به منظور راه‌اندازی بستر جمع‌سپاری در این مقاله Bossa نام دارد که بستری متن‌باز است و در سال ۲۰۰۷ توسط دانشگاه برکلی توسعه داده شده است [11]. این بستر امکان تعریف تنوعی از فعالیت‌های جمع‌سپاری را فراهم می‌کند که هریک در قالب مجموعه‌ای از وظایف در اختیار کاربران قرار می‌گیرد. البته واسط کاربری چندان مناسبی ندارد و برای استفاده از آن، لازم است که واسط کاربری تقریباً از ابتدا توسعه داده شود. زبان توسعه در این بستر PHP است و این بستر در سامانه عامل Debianقابل استفاده است.

٤- روال ارزیابی انسانی جویشگرهای متنی در بستر جمع سپاری

در این بخش به معرفی مراحل مختلف انجام فعالیت کیفیت‌سنجی پرداخته شده است. روال انجام این فعالیت در شکل ۱ نشان داده شده است. به این منظور ابتدا لیستی از افراد مشارکت‌کننده در فعالیت ارزیابی تعیین شد. سپس مجموعه پرس وجوهای ارزیابی از منابع مختلف استخراج شد. به منظور امکان ارزیابی کیفیت جویشگر مبتنی بر دسته‌های مختلف پرس‌وجوها، فعالیت دیگری در بستر جمع‌سپاری برای برچسب‌گذاری پرس وجوها تعریف شد. در ادامه مجموعه پرس‌وجوهای برچسب‌خورده به جویشگرها ارسال و ۵ نتیجه اول هر جویشگر برای ارزیابی در اختیار کاربران قرار گرفت. در نهایت دقت جویشگرها مبتنی بر معیار nDCG محاسبه گردید.

در ادامه، ابتدا مجموعه دادگان مورد استفاده در این ارزیابی معرفی شده و سپس به فعالیتهای برچسب‌گذاری دادگان و جزئیات روال انجام شده جهت ارزیابی جویشگرهای متنی خواهیم پرداخت.

شکل(۱):روال انجام ارزیابی انسانی

٤-۱- مجموعه دادگان

بدین منظور، پرس‌وجوها از چهار منبع دریافت شدند:

• منبع اول، لاگ جویشگر پارسی‌جو بود. پرس و جوهای ارسالی به این جویشگر در بازه زمانی آوریل تا جولای سال ۲۰۱۶ به همراه تعداد تکرار هریک جمع‌آوری شدند. سپس به صورت تصادفی و با روش Random Weighted Sampling، ۳۰۰ پرس و جو از میان این پرس و جوها انتخاب گردیدند.

• منبع دوم دریافت پرس‌وجوها، لاگ جویشگر یوز بود. به طریقه مشابه با پارسی‌جو، تیم توسعه دهنده یوز نیز مجموعه‌ای متشکل از ۳۰۰ پرس و جو را در اختیار قرار دادند.

• پرس‌وجوهای مورد استفاده در کنفرانس‌های سالانه TREC منبع سوم را تشکیل دادند. از میان Trackهای مختلف این کنفرانس، دو Track وب (Web Track) و پرسش و پاسخ (Question Answering Track) انتخاب شده و از میان آنها، ۲۵۰ پرس و جو به تصادف انتخاب گردید. همچنین، از میان همین مجموعه، ۲۵۰ پرس‌وجوی دیگر نیز به تصادف انتخاب شده و به فارسی بازگردانده شد. در فرایند ترجمه به فارسی، بومی‌سازی نیز انجام شد. مثلاً پرس‌وجوی «history of orcas island» به «تاریخ جزیره کیش» تغییر یافت و یا به جای «تام کروز» از «رضا عطاران» استفاده شد. بدین ترتیب، در مجموع ۵۰۰ پرس‌وجو آماده شده و به لیست پرس‌وجوهای مورد استفاده جهت ارزیابی افزوده شد.

• در نهایت، آخرین منبع مورد استفاده جهت تولید مجموعه دادگان ورودی وب‌سایت الکسا و رتبه‌های بالای این وب‌سایت در میان بازدیدکنندگان ایرانی بود. لیست ۵۰۰ وب‌سایت اول الکسا در ایران دریافت شد. در نهایت ۵۰ پرس و جوی پیمایشی فارسی و ۵۰ پرس و جوی پیمایشی انگلیسی به روش تصادفی به دست آمد. این دو مجموعه به عنوان پرس‌وجوهای پیمایشی به مجموعه پرس‌وجوهای مورد استفاده در ارزیابی افزوده شدند. البته لازم به ذکر است که برخی از پرس‌وجوها در این مجموعه، نظیر «باما» یا «Blog» مبهم بودند که برای رفع ابهام از آنها، یا واژه وب‌سایت در ابتدای آنها افزوده گردید («وب‌سایت باما») و یا نام دامنه وب‌سایت در انتهای آن قرار گرفت («Blog.ir»).

بدین ترتیب، مجموعه‌ای متشکل از ۱۲۰۰ پرس‌وجو جهت ارزیابی جویشگرهای متنی به دست آمد.

٤-۲- فعالیت برچسب‌گذاری پرس و جوها

به منظور انجام تحلیل‌های دقیق‌تر از عملکرد جویشگرهای متنی لازم بود، پرس‌وجوهایی که به منظور ارزیابی آنها مورد استفاده قرار می‌گیرد در دسته‌های مختلف برچسب‌گذاری شود. با توجه به آنکه پرس و جوهای مذکور برای ارزیابی جویشگر متنی تهیه گردیده است، مبنای ایجاد برچسبها جنبه‌های مختلف ارزیابی کیفی این جویشگرها می‌باشد.

- صحت نگارش: صحیح، خطادار
- انتظار صریح فایل یا وب‌سایت: وب‌سایت، فایل
- وابستگی زمانی: مبتنی بر فصل، مبتنی بر رویداد، عدم وابستگی زمانی
- طول پرس و جو: یک کلمه، دوکلمه، سه کلمه، چهارکلمه، پنج کلمه و بیشتر
- نوع پایه: اطلاعاتی، پیمایشی و تراکنشی
- حساس به آخرین نسخه: حساس به آخرین نسخه، غیرحساس به آخرین نسخه
- دسته‌بندی موضوعی: فناوری اطلاعات، مذهبی، علمی، عمومی، سلامت و پزشکی، هنر و سرگرمی، ورزش، اقتصادی، سیاسی، مسافرت و گردشگری، اجتماعی، فرهنگ و ادبیات، محیط زیست و تاریخی

- اخلاقی بودن: اخلاقی و غیراخلاقی
- مبهم بودن: مبهم، غیرمبهم

۴-۳- فعالیت ارزیابی جویشگرهای متنی

روال انجام ارزیابی جویشگرهای متنی، به این صورت است که مجموعه پرس‌وجوها روی جویشگرهای متنی مختلف ارسال شده و نتایج دریافتی از این جویشگرها، به همراه پرس‌وجوها جهت تعیین میزان مرتبط بودن آنها در اختیار کاربران قرار گیرد. به این منظور فعالیت ارزیابی جویشگر متنی روی بستر جمع‌سپاری تعریف گردید. ویژگیهای این بستر عبارتند از اینکه در این بستر ارزیابی به صورت کاملا بی‌طرفانه انجام می‌شود، چرا که کاربر از اینکه نتیجه ارائه شده از سوی کدام جویشگر ارائه شده اطلاعی ندارد. همچنین انعطاف‌پذیری بالایی برای افزایش یا کاهش تعداد وظایف و تعداد مشارکت-کنندگان در فعالیت وجود دارد. در ارزیابی جویشگر متنی هر وظیفه عبارت است از یک پرس و جو به همراه نتیجه ارائه شده از سوی یک جویشگر همراه با سطوح ارتباطی تعریف شده برای ارزیابی. کاربر برای انجام هر وظیفه باید میزان ارتباط پرس وجو با نتیجه را مبتنی بر سطوح تعریف شده مشخص نماید.

۴-۳-۱- زیرساخت جمع‌سپاری جهت کیفیت‌سنجی

آنچه که قرار است در این فعالیت صورت پذیرد آن است که کاربر بتواند با دریافت یک زوج «پرس‌وجو و لینک یک وب‌سایت»، مشخص نماید که لینک مزبور تا چه حد به پرس‌وجوی ارائه شده مرتبط است. شکل ۲ واسط کاربری طراحی شده بدین منظور را نشان می‌دهد.

شکل(۲): واسط کاربری مربوط به انجام فعالیت «ارزیابی جویشگرهای متنی»

هر چقدر که اطلاعات ارائه شده در لینک وب‌سایت، مفیدتر بوده و با پرس و جو ارتباط بیشتری داشته باشند، میزان ارتباط بالاتر خواهد بود. مقادیری که کاربر می‌تواند به عنوان میزان ارتباط انتخاب نماید به شرح زیر هستند:

- زمانی که باز شدن صفحه وب‌سایت با مشکل مواجه می‌شود، که ممکن است صفحه فیلتر باشد یا کاربر به لینک خراب ارجاع داده شده باشد.
- در حالتی که صفحه وب مورد نظر، بدون هرگونه مشکلی باز می‌شود، میزان ارتباط آن با پرس و جوی ارائه شده یکی از موارد «عالی»، «خیلی خوب»، «خوب»، «بد»، «خیلی بد»، «اسپم» یا «غیراخلاقی» به شرح زیر خواهد بود:
  - عالی: میزان ارتباط فقط و فقط در صورتی «عالی» است که پرس‌وجوی ارائه شده به منظور یافتن آدرس یک وب‌سایت باشد و صفحه وب‌سایت ارائه شده نیز دقیقاً همان وب‌سایت را ارائه نماید.
  - خیلی خوب: اگر پرس و جوی ارائه شده با هدف دریافت اطلاعات در مورد یک موضوع مشخص مطرح شده باشد، و صفحه وب‌سایت ارائه شده نیز اطلاعات بسیار مناسب و مطلوبی در زمینه موضوع مورد نظر ارائه نماید، میزان ارتباط «خیلی خوب» خواهد بود.
  - خوب: در صورتی که لینک صفحه مورد نظر دقیقا جواب مورد انتظار برای پرس‌وجو نباشد، اما حداقلی از اطلاعات را در مورد آن پرس و جو ارائه کند، میزان ارتباط «خوب» خواهد بود.
  - بد: گزینه «بد» باید در صورتی انتخاب شود که محتوای موجود در لینک ارائه شده، پاسخی برای پرس و جوی مورد نظر نباشد، اما بی ارتباط به آن هم نباشد.
  - خیلی بد: این گزینه در صورتی انتخاب می‌شود که محتویات لینک ارائه شده، مطلقاً بی ارتباط با موضوع پرس و جو باشند.
  - اسپم: در برخی موارد، وب‌سایت‌هایی مشاهده می‌شوند که اگر چه در خصوص موضوع پرس و جوی ارائه شده، اطلاعات حداقلی دارند، اما عمده اطلاعات ارائه شده در آنها تبلیغاتی و فاقد ارتباط با موضوع است.
  - غیراخلاقی: این گزینه در صورتی انتخاب می‌شود که لینک ارائه شده باز شود، اما حاوی محتوای غیراخلاقی باشد.

۴-۳-۲- نهایی سازی اطلاعات

با توجه به تفاوت سلیقه‌های انسانی، به منظور نهایی‌سازی اطلاعات و محاسبه معیار nDCG، هر زوج (پرس‌وجو و لینک) ابتدا در اختیار دو کاربر قرار می‌گیرد. در صورتی که دو کاربر پاسخ یکسان ارائه دهند، نتیجه تأیید می‌شود. اما در صورتی که پاسخ دو کاربر متفاوت باشد، زوج مزبور در اختیار کاربر سوم قرار می‌گیرد. اگر از بین این سه کاربر، دو کاربر پاسخ‌های یکسان داشته باشند، نتیجه دریافتی از این دو کاربر تأیید می‌شود، اما در غیر این صورت، از بین سه میزان ارتباط مختلف دریافتی، میزان ارتباط میانه به عنوان نتیجه تأیید می‌گردد. به عنوان مثال، اگر یک کاربر گزینه «خیلی خوب»، کاربر دوم گزینه «بد» و کاربر سوم گزینه «خیلی بد» را انتخاب کرده باشند، نتیجه «بد» در نظر گرفته می‌شود.

۴-۳-۳- نحوه محاسبه معیارها

به منظور تعیین میزان دقت جویشگرها، از معیار nDCG محاسبه شده طبق فرمول (۱) استفاده شده است.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \qquad (۱)$$

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2^{i+1}}$$

با توجه به آنکه در پایان فعالیت ارزیابی جویشگرهای متنی، نتایج باید در قالب معیار nDCG ارائه گردد، برای هر میزان ارتباط، یک Gain در نظر گرفته شده است که به شرح جدول ۱ است.

جدول ۱: مقدار Gain در نظر گرفته شده برای هر میزان از ارتباط

میزان ارتباط	Gain
عالی	۳۱
خیلی خوب	۱۵
خوب	۷
بد	۳
خیلی بد/فیلتر/لینک خراب	۰
اسپم/غیراخلاقی	-۱

ذکر این نکته در اینجا حائز اهمیت است که در نظر گرفتن یک سطح مجزا برای پرس‌وجوهای پیمایشی، برخلاف روشی که به صورت معمول در کنفرانس‌هایی نظیر TREC و CLEF مورد استفاده قرار می‌گیرد، با توجه به آن لحاظ شده است که به صورت صنعتی، جویشگرهای مطرح نظیر بینگ برای ارزیابی نتایج خود، بدین شکل عمل می‌نمایند. شاید در ابتدا اینگونه به نظر برسد که بدین شکل، جویشگرهایی که به پرس‌وجوهای پیمایشی بهتر پاسخ می‌دهند، نتایج بهتری را نسبت به جویشگرهایی که به پرس‌وجوهای اطلاعاتی بهتر پاسخ می‌دهند کسب خواهند نمود، اما واقعیت آن است که با توجه به نرمال شدن نتایج نسبت به نتایج ارائه شده توسط تمامی جویشگرها، چنین اتفاقی عملاً رخ نمی‌دهد.

۵- نتایج فعالیت ارزیابی جویشگرهای متنی

با تعیین میزان ارتباط نتایج دریافتی از جویشگر به پرس‌وجوها و بدست آوردن Gain برای هر میزان ارتباط، معیار nDCG برای چهار جویشگر گوگل، بینگ،پارسی‌جو و یوز محاسبه گردیده است که نتایج نهایی در شکل ۳ آورده شده است. لازم به ذکر است که ارزیابی‌ها و نتایج آنها مربوط به آذرماه ۱۳۹۵ هستند.

طبق نتایج بدست آمده بر مبنای معیار nDCG، جویشگر متنی گوگل بهترین رتبه را با nDCG برابر با مقدار ۰/۶۹ در بین سایر جویشگرها کسب نموده است. جویشگرهای بینگ، پارسی‌جو و یوز با مقدارnDCG متفاوت و کمتر از گوگل به ترتیب در مقام‌های دوم تا چهارم قرار دارند.

به منظور بررسی دقیقتر عملکرد جویشگر در ارزیابی انسانی، نتایج کیفیت سنجی بدست آمده به تفکیک برچسب‌های مختلف، از جمله دسته‌بندی پرس‌وجو، نوع پایه پرس‌وجو، صحیح یا خطادار بودن پرس‌وجوها و غیره قابل ارائه است که در ادامه چند نمونه از این نتایج ارائه گردیده است.

شکل(۳): نتایج کلی بر مبنای معیار nDCG برای جویشگرهای متنی گوگل،بینگ، پارسی‌جو و یوز

شکل ۴ میزان nDCG به دست آمده از جویشگرهای مختلف برای پرس‌وجوهای خطادار و صحیح را نشان می‌دهد. نتایج نشان می‌دهد که تمامی جویشگرها در مواجهه با پرس‌وجوهای خطادار عملکرد ضعیف‌تری داشته‌اند، اما این افت کیفیت برای گوگل و پارسی‌جو کمتر از دو جویشگر دیگر است و به عبارت دیگر، می‌توان گفت که زیربخش «تصحیح خطای پرس‌وجوی ورودی» در این دو جویشگر از کیفیت بالاتری برخوردار است.

در شکل ۵ نتایج nDCG حاصل از جویشگرهای مختلف به صورت دسته‌بندی شده برای زبان‌های «فارسی»، «انگلیسی»، «فارسی/انگلیسی» و «سایر زبان‌ها» ارائه گردیده است. گوگل تقریباً برای تمامی حالت‌ها نتایج یکسانی ارائه کرده است که نشان دهنده قدرت این جویشگر در بازیابی چندزبانه اطلاعات است. البته به نظر می‌رسد که برای حالتی که زبان پرس و جو غیر از فارسی یا انگلیسی بوده است، نتایج این جویشگر نیز دچار افت شده است، اما دراین خصوص توجه به این نکته حائز اهمیت است که تعداد پرس‌وجوهای موجود در دسته «سایر زبان‌ها» در حدی نبوده که بتوان به نتایج به دست آمده از آن اعتماد نمود. دو جویشگر پارسی‌جو و یوز برای زبان فارسی نتایج بهتری ارائه کرده‌اند و در مقابل، نتایج حاصل از بینگ برای پرس‌وجوهای انگلیسی بهتر بوده است. همین نکته نشان دهنده تمرکز بیشتر این جویشگرها روی یک زبان و توجه کمتر به سایر زبان‌هاست.

	صحیح	خطادار
Bing	0.591503058	0.357756363
Google	0.698120556	0.664920821
Parsijoo	0.49864127	0.444454396
Yooz	0.405564419	0.256344684

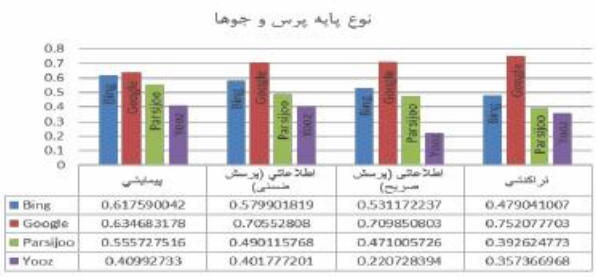
شکل (٤): نتایج nDCG جویشگرهای متنی به تفکیک صحیح یا خطادار بودن پرس و جوها

	فارسی	انگلیسی	فارسی-انگلیسی	سایر زبان ها
Bing	0.565535548	0.629506481	0.509952992	0.217474145
Google	0.702844001	0.682233358	0.703715286	0.531604862
Parsijoo	0.521059723	0.440641644	0.483624898	0.12617882
Yooz	0.416120335	0.363470465	0.360106544	0.207186397

شکل (٥): نتایج nDCG جویشگرهای متنی به تفکیک زبان پرس‌وجوها

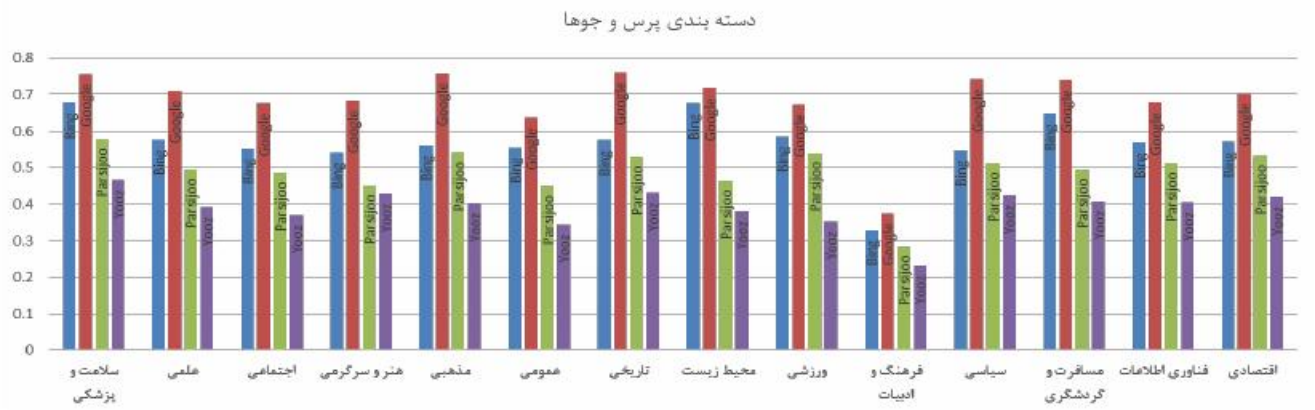
نتایج nDCG به تفکیک نوع پایه پرس‌وجوها در شکل ۶ مشاهده می‌شود. به جز گوگل، تمامی جویشگرها در پرس‌وجوهای تراکنشی افت کیفیت دارند. همچنین، جویشگر یوز برای پرس‌وجوهای اطلاعاتی که حاوی پرسش صریح هستند نسبت به سایر جویشگرها افت کیفیت محسوس‌تری را تجربه کرده است.

رتبه هر یک از چهار جویشگر بر اساس معیار nDCG به تفکیک دسته‌بندی موضوعی پرس‌وجو در شکل ۷ آورده شده است. همانطور که مشاهده می‌شود، تقریباً در تمامی دسته‌ها، ترتیب امتیاز جویشگرها یکسان است. به عبارت دیگر، تقریباً در تمامی دسته‌ها، گوگل در رتبه اول و بینگ، پارسی‌جو و یوز به ترتیب در رتبه‌های دوم تا چهارم قرار دارند. تنها دسته متفاوت از این لحاظ، دسته «پرس و جوهای نامعتبر» است که در آن ترتیب پارسی‌جو و بینگ تغییر کرده است. بدیهی است که با توجه به نامعتبر بودن پرس‌وجوها در این دسته، عملاً نتایج مناسبی از جویشگرها در پاسخ به آنها دریافت نمی‌شود و لذا ترتیب به دست آمده نکته خاصی را نشان نمی‌دهد.



	پیمایشی	اطلاعاتی (پرسش مشخص)	اطلاعاتی (پرسش صریح)	ناوبری
Bing	0.617590042	0.579901819	0.531172237	0.479041007
Google	0.634683178	0.70552808	0.709850803	0.752077703
Parsijoo	0.555727516	0.490115768	0.471005726	0.392624773
Yooz	0.40992733	0.401777201	0.220728394	0.357366968

شکل (٦): نتایج nDCG جویشگرهای متنی به تفکیک نوع پایه پرس‌وجوها.



شکل (۷): نتایج nDCG جویشگرهای متنی به تفکیک دسته بندی موضوعی پرس‌وجوها

٦- جمع‌بندی

این مقاله به ارائه نتایج حاصل از انجام ارزیابی انسانی جویشگرهای بومی یوز و پارسی‌جو در مقایسه با جویشگرهای بین‌المللی بینگ و گوگل اختصاص داشت. ارزیابی‌ها در بستر جمع‌سپاری Bossa و با استفاده از حدود ۵۰ ارزیاب انسانی صورت پذیرفتند. ۱۲۰۰ پرس‌وجو که متشکل از لاگ جویشگرهای پارسی‌جو و یوز، برخی پرس‌وجوهای کنفرانس TREC و برخی از وب‌سایت‌های دارای رتبه مناسب الکسا بودند برای ارزیابی‌ها مورد استفاده قرار گرفتند و برای هر پرس‌وجو، ۵ نتیجه اول هر جویشگر مورد ارزیابی قرار گرفتند. نتایج حاصل از این ارزیابی‌ها بر اساس معیار nDCG نشان داد که ترتیب کیفی جویشگرها گوگل، بینگ، پارسی‌جو و یوز است. به منظور تحلیل بیشتر نتایج، ۱۲۰۰ پرس‌وجوی مورد استفاده برچسب‌گذاری شدند و نتایج nDCG برای هر دسته به صورت مجزا تعیین گردید. نتایج این فعالیت نشان دادند که جویشگرهای پارسی‌جو و یوز برای پرس‌وجوهای زبان فارسی دقت بالاتری را از خود نشان می‌دهند. همچنین، پرس‌وجوهای تراکنشی نقطه ضعف جویشگرهای بومی هستند و باید در این زمینه خود را تقویت نمایند. نکته قابل توجه دیگر در این زمینه توانمندی مناسب جویشگر پارسی‌جو در مواجهه با پرس‌وجوهای خطادار است.

مراجع

[1] Griesbaum, J. "Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de." Information Research, 2004.
[2] R. Ali and M. M. Sufyan, "An Overview of Web Search Evaluation Methods," *Computer and Electrical Engineering Journal*, Vol. 37, Issue 6, pp. 835-848, 2011.
[3] Leighton, H. V. and J. Srivastava, "First 20 precision among World Wide Web search services (search engines)." Journal of the American Society for Information Science 50(10): 870-881, 1999.
[4] Gordon, M. and P. Pathak, "Finding information on the World Wide Web: the retrieval effectiveness of search engines." Information processing and management 35(2): 141-180, 1999.
[5] Lewandowski, D. "The retrieval effectiveness of search engines on navigational queries", Emerald Group Publishing Limited, 2011.
[6] Tawileh, W., J. Griesbaum, et al., "Evaluation of five web search engines in Arabic language, 2010.
[7] Jason Morrison, P. "Tagging and searching: Search retrieval effectiveness of folksonomies on the World Wide Web." Information Processing & Management 44(4): 1562-1579, 2008.
[8] Bitirim, Y., Y. Tonta, et al. "Information retrieval effectiveness of Turkish search engines." Advances in information systems: 93-103, 2002.

[۹] محمداسماعیل، ص.، ق.، ا. لفظی، et al. "مقایسه موتورها و ابر موتورهای کاوش در بازیابی اطلاعات داروشناسی."
[10] Erfanmanesh, M. and F. Didegah, "Evaluating Function of Persian Search Engines on the Web Using Correspondence Analysis." International Journal of Information Science and Management (IJISM) 8(2): 75-5, 2012.
[11] http://boinc.berkeley.edu/trac/wiki/BossaIntro
[12] Tague-Sutcliffe, J. "The pragmatics of information retrieval experimentation, revisited." Information Processing & Management 28(4): 467-490, 1992.
[13] Soboroff, I., C. Nicholas, et al. "Ranking retrieval systems without relevance judgments". Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2001.
[14] F. Can, R. Nuray, and A. B. Sevdik, "Automatic Performance Evaluation of Web Search Engines," Information Processing and Management, Vol. 40, Issue 3, pp. 495-514, 2004.
[15] Y. Shang and L. Li, "Precision Evaluation of Search Engines," World Wide Web, Vol. 5, Issue 2, pp. 159-173, 2002.
[16] Y. Liu, Y. Fu, M. Zhang, Sh. Ma, and L. Ru, "Automatic Search Engine Performance Evaluation with Click-through Data Analysis," in Proc. of the 16th Intl. Conf. on World Wide Web, New York, USA, 2007, pp. 1133-1134.
[17] M. Mahmoudi, R. Badie, M. S. Zahedi, and M. Azimzadeh, "Evaluating the Retrieval Effectiveness of Search Engines using Persian Navigational Queries," in Proc. of the 7th Intl. Symposium on Telecommunications, Tehran, Iran, 2014, pp. 563-568.
[18] R. Cen, Y. Liu, M. Zhang, and Sh. Ma, "Automatic Search Engine Performance Evaluation with the Wisdom of Crowds," Lecture Notes in Computer Science, Vol. 5839, pp. 351-362, 2009.
[19] B. Carterette and R. Jones, "Evaluating Search Engines by Modeling the Relationship Between Relevance and Clicks," in Proc. of the 21st Advances in Neural Information Processing Systems, Vancouver, British Columbia, Canada, 2008, pp. 217-224.
[20] G. Dupret, V. Murdock, and B. Piwowarski, "Web Search Engine Evaluation Using Clickthrough Data and a User Model," in Proc. of the Workshop on Query Log Analysis: Social and Technological Challenges, Banff, Alberta, Canada, 2007.
[21] Zh. Liu, Y. Feng, and H. Wang, "Automatic Deep Web Query Results User Satisfaction Evaluation with Clickthrough Data Analysis," Intl. Journal of Smart Home, Vol. 8, No. 5, pp. 25-32, 2014.
[22] G. Smith and H. Ashman, "Evaluating Implicit Judgements from Image Search Interactions," in Proc. of the Web Science: Society On-Line, Athens, Greece, 2009.
[23] G. Smith and H. Ashman, "Evaluating Implicit Judgements from Image Search Clickthrough Data," Journal of the American Society for Information Science and Technology, Vol. 63, Issue 12, pp. 2451-2462, 2012.

زیرنویس‌ها

[1] Information Need

[2] Navigational Queries