تعامل انسان و اطلاعات، جلد۴ شماره۳
http://hii.khu.ac.ir
مقاله اصیل پژوهشی

ارزیابی همپوشانی و پوشش چهار موتور جستجوی بومی پارسی جو، یوز، پارسیک و ریسمون

محسن نوکاریزی: دانشیار دانشگاه فردوسی مشهد (نویسنده مسؤول). mnowkarizi@um.ac.ir
مهدی زینالی تازه کندی: دانشجوی کارشناسی ارشد علم اطلاعات و دانش‌شناسی دانشگاه فردوسی. ma.zeynali@mail.um.ac.ir

چکیده

دریافت: ۱۳۹۶/۰۷/۱۹
پذیرش: ۱۳۹۶/۰۹/۰۷

زمینه و هدف: پژوهش حاضر با هدف سنجش همپوشانی موتورهای جستجوی بومی پارسی جو، یوز، پارسیک، و ریسمون و مقایسه توانمندی‌های این موتورها در پوشش دادن وب نمایه‌پذیر انجام گرفت.

روش پژوهش: پژوهش از نوع کاربردی ارزیابانه بود. برای گردآوری اطلاعات از روش مبتنی بر کلیدواژه بهره گرفته شد، بدین ترتیب ابتدا کلیدواژه‌های انتخاب شده به موتورهای جستجو ارائه و از رکوردهای بازیابی شده نمونه‌گیری و با توجه به وجود یا نبود این رکوردها در موتورهای جستجو، داده های لازم گرداوری شد، بر این اساس برای تجزیه و تحلیل داده‌ها از آمار استنباطی استفاده شده است.

یافته‌ها: همپوشانی نسبی موتور جستجو پارسیک نسبت به پارسی‌جو و همپوشانی نسبی پارسی‌جو نسبت به یوز به طور متوسط ۲۶ درصد بود و موتور جستجوی پارسیک بیش‌ترین بازیابی را داشت. موتور جستجو ریسمون هیچ رکورد مشترکی با موتورهای جستجو مورد آزمون نداشت. سه موتور جستجو پارسیک، پارسی‌جو و یوز از بین ۲۲۵ رکورد، ۲۷ رکورد مشترک را بازیابی کردند؛ تفاوت معنی‌داری بین همپوشانی نسبی موتورهای جستجو وجود داشت. همچنین به طور متوسط موتورهای جستجو پارسیک، پارسی‌جو، یوز و ریسمون به ترتیب ۳۸ درصد، ۳۱ درصد، ۲۶ درصد و ۶ درصد از وب قابل نمایه را پوشش می‌دادند. میان میزان پوشش پایگاه موتورهای جستجو نیز تفاوت معنی‌داری مشاهده شد.

نتیجه‌گیری: ظاهراً هر کدام از موتورهای جستجو سیاست نمایه‌سازی متفاوتی داشتند و کاربران برای رسیدن به اطلاعات جامع درباره یک موضوع به جستجو در بیش از یک موتور جستجو نیاز دارند و می‌توان پیش‌بینی نمود که با جستجو در دو موتور جستجو پارسیک و پارسی‌جو، به حدود ۷۰ درصد وب نمایه‌پذیر دسترسی داشت.

کلیدواژه‌ها: ارزیابی بازیابی اطلاعات، وب فارسی، موتورجستجو، پوشش، همپوشانی، پارسی جو، یوز، پارسیک، ریسمون

مقدمه

بشر همواره به دنبال کشف پدیده‌های اطراف و افزایش آگاهی خویش نسبت به آن‌ها بوده است. هر فرد در زندگی روزمره خود، هر لحظه نیازمند دانستن درباره نحوه یافتن، کیفیت و قیمت بسیاری چیزها، مثل خدمات بهداشتی، تأمین اجتماعی، تسهیلات آموزشی و امکانات تربیتی است. افراد حتی در انجام کارهای خانه نیز نیازمند اطلاعات عملی مثل آشپزی، باغبانی و مانند آن هستند. آنان در انجام فعالیت‌ها و کارهای اداری نیز نیازمند اطلاعات فنی و گاه حرفه‌ای در مورد محیط کار، فرایند انجام کار، مدیریت و موارد مشابه هستند. به‌طورکلی، هر فعالیت انسانی دارای یک ورودی اطلاعاتی است و در این راستا، تمام فعالیت‌ها در سازمان‌های خدماتی و تولیدی نیازمند اطلاعات هستند (ویکری و ویکری، ۱۳۸۰، ۲۵– ۲۷)

نیاز به اطلاعات در کلیه تمدن‌های بشری قابل‌مشاهده و بررسی است. در تمدن‌های کهن، دانشمندان و فیلسوفان تنها قشر خاصی از جامعه بودند که کاربران اصلی اطلاعات به شمار می‌آمدند؛ اما امروزه این حدومرز به‌طور کامل برداشته‌شده است و همه نیازمند و کاربر اطلاعات‌اند (کوشا، ۱۳۸۱، ۲)

درگذشته‌ای نه‌چندان دور، کتابخانه‌ها و مراکز اطلاع‌رسانی تنها مکان‌هایی محسوب می‌شدند که منابع اطلاعاتی برای پاسخگویی به نیازهای اطلاعاتی استفاده‌کنندگان در آنجا قابل‌دسترس بود. با ظهور فناوری‌های نوین اطلاعاتی و ارتباطی به‌خصوص شبکه جهانی وب، تغییرات شگرفی در تولید، توزیع، انتشار، اشاعه و دسترسی به منابع اطلاعاتی به وجود آمد و وب به یکی از مهم‌ترین منابع اطلاعاتی تبدیل شد. از یک‌سو، تعداد استفاده‌کنندگان و از سوی دیگر حجم اطلاعات قابل‌دسترس از طریق وب به‌صورت شگفت‌آوری در حال افزایش است؛ به طوری که بیش از ۴ میلیارد صفحه وبی فارسی وجود دارد و روزبه‌روز بر این حجم افزوده می‌شود (جهانگرد، ۱۳۹۶). با وجود این افزایش، نبود ساختار و ناهمگنی وب موجب شده است که هیچ‌کس نتواند تمام

اطلاعات مورد نیازش را به طور کامل از این صفحات وبی به دست آورد؛ اما هر کس به ابزارهایی نیاز دارد که با پشتیبانی و کمک آن‌ها به نیاز اطلاعاتی خود پاسخ دهد (اندرسون[1]، ۲۰۰۶).

موتورهای جستجو در راستای پاسخگویی به چنین نیاز در بازیابی اطلاعات از وب توسعه یافتند. موتورهای جستجو، پایگاه‌های وبی هستند که به‌صورت خودکار نمایه‌های وب را ایجاد می‌کنند؛ به بیانی دیگر، آن‌ها با برنامه‌های توزیعی به نام روبات یا خزنده کار می‌کنند تا صفحات وبی را گردآوری نمایند (کلارک[2]، ۲۰۰۰). موتورهای جستجوی بومی از قبیل پارسیک[3]، پارسی‌جو[4]، یوز[5]، و ریسمون[6] ،هرکدام راهبرد و سیاست مجموعه‌سازی ویژه‌ای جهت نمایه‌سازی مدارک وبی دارند که موجب تفاوت در اندازه پایگاه و مدارک نمایه‌شده در آن‌ها می‌شود. البته هیچ‌کدام از موتورهای جستجو نمی‌توانند کل وب را نمایه‌سازی کنند.

کاربران موتورهای جستجو علاقه‌مند هستند بدانند اندازه وب چقدر است و هرکدام از موتورها چه اندازه از وب را پوشش می‌دهند و کدام‌یک از موتورها بیشترین میزان پوشش صفحات وبی را دارد و چقدر از صفحات وبی به‌صورت مشترک در موتورها نمایه‌سازی می‌شوند (باهارات و برودر[7]، ۱۹۹۸). این پرسش‌ها بنیه علمی داشته و مورد علاقه عامه هستند. در این راستا، جهت پاسخگویی به آن‌ها، ارزیابی میزان پوشش و همپوشانی موتورها به یکی از حوزه‌های مهم بازیابی اطلاعات تبدیل‌شده است. این عامل سبب شده است تا طراحان موتورهای جستجو، اندازه پایگاه خود را ارتقا دهند؛ چرا که اندازه پایگاه موتورهای جستجو به‌عنوان شاهدی بر کارآمدی آن‌ها تلقی می‌شود (کلارک، ۲۰۰۰). با توجه به اینکه با جستجو در پایگاه پژوهشگاه علوم و فناوری اطلاعات (ایران‌داک)، بانک اطلاعات نشریات کشور (مگیران)، سیویلیکا و جهاد دانشگاهی با کلیدواژه‌های «اندازه پایگاه موتورجستجو»، «اندازه پایگاه موتورکاوش» «اندازه نمایه موتورجستجو»، « اندازه نمایه موتورجستجو»، «پوشش موتور جستجو»،«وب فارسی»، «همپوشانی موتورجستجو» پژوهشی درزمینه‌ی پوشش و همپوشانی موتورهای جستجوی بومی مشاهده نشد و با توجه به اینکه هنوز مشخص نیست که هرکدام از این موتورهای جستجو چه میزان از وب را نمایه می‌کنند و کدام‌یک از موتورها بیش‌ترین پوشش را دارند و نیز میزان همپوشانی آن‌ها در چه حدی است، نیاز به پژوهشی که بر اساس اصول علمی به شناسایی موتور جستجوی کارآمد از این میان بپردازد، احساس شد تا از این طریق کاربران با شناخت و استفاده از موتور جستجوی کارآمد در کم‌ترین زمان نیاز اطلاعاتی خود را برطرف نمایند. دراین راستا، در پژوهش حاضر دو سؤال مطرح و دو فرضیه زیر آزمون شد:

سوال اول: میزان همپوشانی هر کدام از موتورهای مورد بررسی تا چه حد است؟

سؤال دوم: میزان پوشش هر کدام از موتورهای مورد بررسی تا چه حد است؟

فرضیه اول: میان میزان همپوشانی چهار موتورجستجوی پارسیک، پارسی جو، یوز و ریسمون تفاوت معنی‌داری وجود دارد.

فرضیه دوم: میان میزان پوشش چهار موتورجستجوی پارسیک، پارسی جو، یوز و ریسمون تفاوت معنی‌داری وجود دارد.

مبانی نظری و پیشینه پژوهش

قابلیت‌های موتورهای جستجو به‌طور مرتب در منابع مختلفی مورد مطالعه و مقایسه قرار می‌گیرد. در این منابع معیارهایی نظیر روش‌های مجموعه‌سازی، شیوه نمایه‌سازی، خط‌مشی چکیده‌نویسی و تسهیلات نمایش تحت بررسی قرار می‌گیرند. برخی از موتورهای جستجو بخشی از منابع اطلاعاتی اینترنت مانند صفحات وب یا گروه‌های خبری را جستجو می‌کنند، برخی دیگر صرفاً گروه خاصی از منابع را جستجو و در پایگاه خود نمایه می‌کنند (داورپناه، ۱۳۸۷، ۹۴). آنچه در نظام اطلاعاتی ذخیره می‌شود و محتوای آن را تشکیل می‌دهد، موضوع موردتوجه کاربران است که بر اساس خط‌مشی خاصی از پیش گزینش و فراهم‌شده است (فتاحی، ۱۳۸۳).

موتور جستجویی که از قابلیت‌های جستجویی پیشرفته‌ای برخوردار باشد، اگر نتواند مدارک کافی را پوشش دهد، نمی‌تواند برای کاربران رضایت‌بخش باشد (کوپر[8]، ۱۹۶۸؛ کلارک، ۲۰۰۰). به همین سبب اندازه نمایه موتورهای جستجو به‌عنوان شاهدی بر کارآمدی موتورها در نظر گرفته می‌شود. همچنین همپوشانی موتورهای جستجو یکی دیگر از

1. Anderson
2. Clarke
3. www.parseek.ir
4. www.parsijoo.ir
5. www.yooz.ir
6. www.rismoon.com
7. Bharat & Broder
8. Cooper

موضوعات مرتبط به پوشش موتورهای جستجو است که اهمیت فراوان دارد.

از لحاظ ساخت ادبی، واژه همپوشانی اسم مصدری است که از ترکیب پیشوند «هم-» در کنار صفت حالیه پوشان به همراه یای پسوند مصدرساز ساخته شده است (خلیلی، ۱۳۷۲). پوشان یعنی کسی یا چیزی که پوششی یا جامه‌ای بر تن دارد. پیشوند «هم» وقتی به کلمه‌ای می‌چسبد، ذاتاً به بیشتر از یک‌چیز دلالت می‌کند. از این رو، موجود بودن یک منبع اطلاعاتی در دو یا چند مجموعه مختلف را همپوشانی گویند (باکلند، هندلی و وال کر[1]، ۱۹۷۵؛ پویر[2]، ۱۹۸۴). به بیانی دیگر، اگر دو مجموعه از اشیا داشته باشیم، به‌طوری‌که برخی یا کل این اشیا متعلق به هر دو مجموعه باشد، بین دو مجموعه همپوشانی وجود دارد (اگه[3]، ۲۰۰۶). در این تعریف «مجموعه‌ای از اشیا می‌تواند اشاره به کتابخانه‌ها، پایگاه‌های اطلاعاتی، و موتورهای جستجو باشد و منظور از «اشیا»، منابع اطلاعاتی مختلف از قبیل کتاب، مقاله، صفحه‌ی وبی و مانند آن است. در پژوهش‌های همپوشانی در اصل تعداد اشیا مشترک بین دو مجموعه مطالعه می‌شود و در پژوهش‌های مختلف به پوشش و همپوشانی کتابخانه‌ها، پایگاه‌های اطلاعاتی و موتورهای جستجو پرداخته شده است.

پژوهش‌های اولیه، همپوشانی و پوشش کتابخانه‌ها را مورد بررسی قرار داده‌اند در این راستا، تاج‌الدین (۱۳۷۹)، همپوشانی نشریات ادواری کتابخانه های تخصصی و دانشگاهی تهران، باکلند، هندلی و وال‌کر (۱۹۷۵) نیز همپوشانی ۲۳ کتابخانه دانشگاهی و ملی بریتانیا را تعیین نمودند.

با پیدایش پایگاه‌های اطلاعاتی و درج مجلات در این پایگاه‌ها، پژوهش‌ها به سمت‌وسوی جدیدی معطوف شد و پژوهشگران به بررسی همپوشانی پایگاه‌های اطلاعاتی پرداختند که پژوهش‌های وود[4] و همکاران (۱۹۷۲، ۱۹۷۳)، و هود و ویلسون[5] (۲۰۰۸) نمونه‌ای از این پژوهش‌هاست.

پوشش و همپوشانی موتورهای جستجو در طی ۲۰ سال اخیر مطرح‌شده است و در این زمینه سایت‌هایی مختلفی وجود دارند که به‌طور روزانه میزان وب نمایه‌پذیر و اندازه پایگاه هرکدام از موتورها را تخمین می‌زنند. سرچ‌اینجین‌واچ[6] یکی از این وب‌سایت‌هاست که به‌طور جامع به موتورهای جستجو اختصاص یافته و توسط دنی سالیون[7] ایجاد شده است (باهارات و برودر، ۱۹۹۸). یکی دیگر از این سایت‌ها ورلدوایدوب‌سایز[8] است که میزان وب و اندازه نمایه موتورهای جستجو اصلی نظیر گوگل، بینگ، یاهو و اسک را بر اساس برآورد پژوهش‌های مشاور اینترنتی موریس دی‌کاندر[9] منتشر می‌کند (پاپیس[10]، ۲۰۱۶).

روش دیگر برای محاسبه اندازه وب و نمایه موتورهای جستجو رویکرد مستقیمی است که سیاهه آدرس‌های الکترونیکی از نمایه موتورهای جستجو تهیه می‌شود و اندازه پایگاه موتورها و اشتراکات آن‌ها محاسبه می‌شود. این روش تقریباً غیرممکن است؛ چون ارائه چنین سیاهه‌ای از موتورهای جستجو که اطلاعات حیاتی برای آن‌ها محسوب می‌شود، به‌سختی امکان‌پذیر است و معمولاً موتورهای جستجو چنین سیاهه‌ای را در اختیار پژوهشگران قرار نمی‌دهند.

برای حل مشکل عدم دسترسی به پایگاه موتورهای جستجو، برخی پژوهشگران روش‌های نمونه‌گیری را جهت محاسبه اندازه و همپوشانی موتورهای جستجو ارائه نموده‌اند. یکی از این‌ها محاسبه اندازه همپوشانی موتورها مبتنی بر نتایج بازیابی شده برای تعدادی پرسش است. باهارات و برودر (۱۹۹۸) از این روش در پژوهش خودشان استفاده کردند. در این روش تعدادی سؤال به‌صورت تصادفی انتخاب و در موتورها، جستجو شد و از نتایج بازیابی شده نمونه‌گیری به عمل آمد و در نهایت، وجود این رکوردها، در سایر موتورهای جستجو بررسی شد. این پژوهشگران در پژوهش خود اندازه و همپوشانی موتورهای جستجو هات‌بات، آلتاویستا، اکسایت و اینفوسیک[11] را در بازه زمانی اواسط ۱۹۹۷ تا نوامبر ۱۹۹۷ در ۱۰۰ نتیجه بازیابی‌شده برآورد کردند و دریافتند اندازه پایگاه هرکدام از موتورهای جستجو به ترتیب ۴۷ درصد، ۳۹ درصد، ۳۲ درصد ، ۱۸ درصد و همپوشانی بین چهار موتور جستجو کمتر از ۱/۴ درصد بود.

گیل و سیگنورینی[12] (۲۰۰۵) با استفاده از روش باهارات و برودر، اندازه وب نمایه‌پذیر و همچنین اندازه و همپوشانی موتورهای جستجو «گوگل»، «ام اس ان[13]»، «آسک» و «یاهو» را بررسی کردند. بر اساس نتایج این پژوهش، اندازه وب نمایه‌پذیر بیش از ۱۱/۵ بیلیون صفحه

1. Buckland, Hindle & Walker
2. Poyer
3. Egghe
4. Wood
5. Hood & Wilson
6. https://searchenginewatch.com/
7. Danny Sullivan
8. http://www.worldwidewebsize.com
9. Maurice de Kunder
10. *Pappas*
11. https://www. Hotbot.com, http://www.Altavista.com, www.Excite.com and www. Infoseek.com
12 . Gulli & Signorini
13. www.msn.com

برآورد شد. همچنین اندازه پایگاه گوگل ۶۸/۲ درصد، اماس‌ان ۴۹/۲ درصد، اسک ۴۳/۵ درصد و یاهو ۵۹/۱ درصد تخمین زده شد.

برخی پژوهشگران علت همپوشانی کم‌تر بین موتورهای جستجو را ناشی از بررسی تعداد پرسش‌های کم‌تر در پژوهش‌ها مطرح کرده‌اند. در این راستا، اسپینک، جانسن، بلکلی و کوشمن[1] (۲۰۰۶) همپوشانی چهار موتور جستجوی اماس‌ان، گوگل، یاهو و آسک را با استفاده از ۱۰،۰۰۰ پرسش بررسی کردند که این پرسش‌ها به‌صورت تصادفی از فایل تراکنش موتور ابرجستجوی داگ‌پایل[2] انتخاب شده بود؛ اما آنان فقط نخستین نتیجه بازیابی شده را بررسی کردند. نتایج پژوهش نشان داد که بازهم چهار موتور جستجو همپوشانی کم‌تری (۱/۱ درصد) داشتند.

اگه و گاورتس[3] (۲۰۰۷) به روش محاسبه همپوشانی پرداختند و یادداشت مفیدی با عنوان «نکته‌ای در موردسنجش همپوشانی» ارائه نمودند و خاطرنشان کردند که وقتی کل جامعه موردنظر بررسی نمی‌شود؛ از آمار استنباطی استفاده‌شده است و باید در تجزیه تحلیل داده‌ها نیز از آمار استنباطی استفاده شود که در اکثر پژوهش‌های این حوزه فراموش‌شده است.

در پژوهش دیگری، همپوشانی سه موتور جستجوی عمومی آلتاویستا، گوگل و هات‌بوت و دو موتور جستجوی تخصصی سایروس و بایووب[4] در حوزه زیست‌فناوری بررسی گردید. در این پژوهش، راثر، لون و جیلانی شاه[5] (۲۰۰۸) با استفاده از *سرعنوان‌های موضوعی کتابخانه کنگره*، ۲۰ پرسش جهت جستجو در موتورهای جستجو انتخاب و ۱۰ نتیجه بازیابی‌شده را بررسی کردند. نتایج پژوهش نشان داد که ۹۲/۵۳ درصد از نشانی‌های اینترنتی منحصر به یک موتور جستجو، ۵/۲۲ درصد مشترک بین دو موتور جستجو، ۰/۲۱ درصد مشترک بین سه موتور جستجو بود و هیچ همپوشانی بین چهار موتور جستجو وجود نداشت. همچنین همه نتایج موتور جستجوی بیووب منحصربه‌فرد بود.

میترا و آوه‌کار[6] (۲۰۱۷) همپوشانی گوگل اسکالر، سمانیتیک اسکالر، ماکروسافت آکادمیک و اسکاپوس[7] را درحوزه موضوعی علوم رایانه بررسی کردند. این پژوهشگران در پژوهش خود به ترتیب ۲۳۰۰ پرسش از سامانه رده‌بندی رایانش ای‌سی‌ام[8] و ۲۰۰ پرسش را از کلیدواژه های مقالات ارائه شده در کنفرانس اس‌آی‌جی‌کادی‌دی ای‌سی‌ام[9] گردآوری کردند و درنهایت ۲۵۰۰ پرسش را به این چهار سامانه بازیابی اطلاعات ارائه و هشت نتیجه نخست را بررسی کردند. نتایج پژوهش نشان داد که هر کدام از سامانه‌های بررسی شده، متون متفاوتی را نمایه سازی می‌کرد و همپوشانی آنها به صورت معنی داری درحد پایین بود و از این رو، پژوهشگران نباید به یکی از سامانه‌های بازیابی اطلاعات متکی باشند.

همپوشانی موتورهای ابرجستجوی آی‌سیزپی‌سی، ای‌زد توفایند، وان‌سیکیند، اینفوگرید و ویداو[10] با موتورهای جستجوی تحت پوشش آن‌ها در پژوهش اسفندیاری مقدم (۱۳۸۴) بررسی شد. در این پژوهش، پنج کلیدواژه در موتورهای ابرجستجو و موتورهای جستجوی تحت پوشش آن‌ها کاوش شد و سپس اشترک مدارک بازیابی‌شده در موتورهای ابرجستجو و موتورهای تحت پوشش آن‌ها بررسی گردید. میزان همپوشانی موتور ابرجستجوی ویداو،۵۶ درصد؛ ای زد توفایند، ۷۵ درصد؛ اینفو گرید، ۳۱ درصد، آی سیز پی سی، ۴۷ درصد و وان سیکیند، ۶۱ درصد به دست آمد و درنهایت فرا موتور ای‌زد توفایند را موفق‌ترین موتور ابرجستجو معرفی نمودند.

محمداسماعیل و قائمی (۱۳۸۸) به مقایسه همپوشانی پنج موتور جستجوی گوگل، یاهو، آلتا ویستا، ای‌اوال[11] و آسک و پنج موتورابرجستجو ماما، داگ پایل، متاکراولر،کلاستی و اینفو[12] پرداختند. این پژوهشگران پنج کلیدواژه از اصطلاحنامه کشاورزی انتخاب و به موتورهای جستجو و موتورهای ابرجستجو ارائه و ده نتیجه نخست بازیابی شده را بررسی نمودند. نتایج پژوهش نشان داد که موتورجستجوی یاهو با ۴۴٪ بیش‌ترین همپوشانی و موتورجستجوی آسک با ۲۲٪ کم‌ترین همپوشانی را داشتند. در بین موتورهای ابرجستجو نیز

1. Spink, Jansen, Blakely & Koshman
2.www. Dogpile.com
3. Egghe & Goovaerts
4. http:// www.Scirus.com andhttp:// www.Bioweb.com
5. Rather, Lone, Jeelanishah
6. Mitra & Awekar
7. http:// Scholar.google.com,http:// www.Semantic Scholar.org , https:// Academic .Microsoft.com ,and https://www. Scopus.com
8. ACM Computing Classication System
9 . ACM SIGKDD
10 . http://www.IcySpicy.com, http://www.Ez2find.com, http://www.1Second.com, http://www.InfoGrid.com andhttp://www. Widow.com
11. https://www.aol.com
12. http://www.Mamma.com, http://www.Dogpile.com, http://www.metacravler.com, http://www.Clusty.com &http://www. Info.com

متاکراولر با ۵۰٪ بیش‌ترین همپوشانی و موتور ابرجستجوی ماما با ۲۳٪ کمترین همپوشانی را داشتند.

اسفندیاری و بهاری موفق (۱۳۹۱) همپوشانی چهار موتورجستجو یاهو، لیوسرچ، گوگل و آسک را بررسی کردند. آن‌ها به منظور تعیین همپوشانی موتورهای جستجو، ده کلیدواژه از *سرعنوان‌های موضوعی پزشکی* را انتخاب و میزان اشتراکات رکوردهای بازیابی شده را محاسبه کردند. نتایج این پژوهش نشان داد که یاهو دارای بیش‌ترین نتیجه منحصربه‌فرد و گوگل داری کم‌ترین نتیجه منحصربه‌فرد بود. با وجود این، موتورجستجو گوگل بیش‌ترین همپوشانی را نشان داد و همپوشانی بین چهار موتورجستجو در حدود ۱۱ درصد بود.

گوهری، مکتبی فرد و جمالی مهموئی (۱۳۹۴) همپوشانی سه موتورجستجوی گوگل، یاهو و بینگ را در حوزه علوم انسانی بررسی کردند. آن‌ها به منظور تعیین کلیدواژه در حوزه علوم انسانی، در هر زمینه موضوعی نشریه‌ای را انتخاب نمودند که دارای رتبه علمی – پژوهشی بود و در پایگاه استنادی جهان اسلام نمایه شده و به زبان فارسی بود؛ سپس برای انتخاب کلیدواژه‌ها، چکیده‌های مقالات مربوط به آخرین دوره (۱۳۹۱) هریک از آن نشریات بررسی شد. در نهایت، ۷۵ کلیدواژه برای کاوش در موتورهای جستجو انتخاب شد. این پژوهشگران بیست نتیجه بازیابی شده نخست را بررسی کردند. نتایج پژوهش نشان داد موتورهای جستجوی گوگل، یاهو و بینگ به ترتیب، ۴۲٪، ۴۸٪ و ۵۸٪ همپوشانی داشتند.

در پژوهش دیگری، رجبی و نوروزی (۱۳۹۴) پنج موتورجستجوی جاماسب، گوگلر، کاوشگر، ریسمون و پارسیک را ارزیابی نمودند.آنها با استفاده از نظرات استادان، ۷ کلیدواژه جهت کاوش در موتورهای جستجو انتخاب و میزان اشتراک پنج موتور یادشده را محاسبه کردند. نتایج این پژوهش نشان داد همپوشانی موتورهای جستجوی جاماسب، ریسمون، گوگلر، کاوشگر و پارسیک به ترتیب ۶۰٪، ۰٪، ۴۰٪، ۴۰٪ و ۰٪ بود.

به‌طور خلاصه، پیشینه پژوهش نشان داد موتورهای جستجو سیاست‌های متفاوتی برای گزینش و فراهم‌آوری مدارک اینترنتی دارند که موجب تفاوت عملکرد، به‌ویژه تفاوت در پوشش و همپوشانی آن‌ها شده است. موضوع مهم دیگر روش پژوهش هرکدام از پژوهش‌های ذکرشده است. برخی پژوهش‌ها در انتخاب پرسش (کلیدواژه برای جستجو در موتورها) از نمونه‌گیری احتمالاتی استفاده کرده و برخی دیگر در انتخاب پرسش، روش خاصی نداشته‌اند. به بیانی دیگر، در برخی پژوهش‌ها تلاش شده تا روش آماری و دقیقی برای پژوهش انتخاب شود و به منظور رسیدن به هدف در انتخاب پرسش و انتخاب نشانی الکترونیکی (رکورد بازیابی شده) از نمونه‌گیری احتمالاتی استفاده شده است؛ با جود این، در تخمین و برآورد نتایج پژوهش از تحلیل آماری استفاده نشده است. اما هیچ پژوهش در این میان مشاهده نشد که به موتورهای جستجوی فارسی و میزان پوشش و همپوشانی آن‌ها پرداخته باشد. لذا ضروری است این مسأله مورد بررسی قرار گیرد.

روش پژوهش

روش، نمونه و ابزار: این پژوهش از نوع ارزیابانه بود. پژوهش ارزیابانه یا ارزشیابی نوعی از پژوهش‌های کاربردی است که هدف اصلی آن آزمایش کاربرد دانش در یک طرح یا برنامه ویژه است. به همین سبب، این نوع پژوهش دارای ماهیت کاربردی و جویای فایده عملی است و فرضیات این نوع پژوهش‌ها مبتنی بر متغیری است که عبارت از یک ارزش یا هدف یا اثر مطلوب است (پاول[1]، ۱۳۸۹، ۶۱).

ارزیابی به مفهوم تعیین عملکرد و ارزش سامانه، محصول و راهبرد است که به‌عنوان یک ضرورت مهم در علم، فناوری و بسیاری از حوزه‌های دیگر پذیرفته شده است (ساراسویک[2]، ۱۹۹۶). درواقع ارزیابی، قضاوت درباره یک کار، فکر، راه‌حل، روش و بسیاری از چیزهای دیگر در راستای اهداف است. در ارزیابی از معیارهایی استاندارد استفاده می‌شود تا میزان دقت، تأثیر، مقرون‌به‌صرفگی و رضایت‌بخش بودن موارد ذکرشده معین گردد (فیتزجرالد[3]، ۲۰۰۱). میزان پوشش و همپوشانی موتورهای جستجو از مهم‌ترین معیارهای ارزیابی‌های بازیابی اطلاعات است. چون در این پژوهش میزان پوشش و همپوشانی موتورهای جستجو مشخص و مورد مقایسه قرار گرفت، این پژوهش از نوع کاربردی ارزیابانه بود و در زمره پژوهش‌های ارزیابی بازیابی اطلاعات قرار گرفت. برای رسیدن به اهداف پژوهش، از روش‌های آمار استنباطی استفاده شده است. با توجه به استنتاج از پیشینه پژوهش، در این پژوهش ترکیبی از روش باهارات و برودر (۱۹۹۸) و روش اگه و گاورتس[4] (۲۰۰۷) استفاده شد.

با توجه به پیشینه پژوهش و به ویژه دیدگاه لواندوفسکی[5] (۲۰۱۲) مشاهده می‌شود که در مطالعات اولیه و برخی از مطالعات کنونی از تعداد پرسش‌های کم‌تری (۵ یا ۱۰ پرسش) استفاده شده است؛ اما اکثر مطالعات اخیر از بیش از ۲۵

1. Pavel
2. Sarasevic
3. Fitzgerald
4. Egghe & Goovaerts
5. Levandoweski

پرسش استفاده کرده بودند که گاهی این تعداد به ۵۰ مورد پرسش نیز رسیده است. از این رو، استفاده از ۳۰ تا ۳۵ پرسش به منظور ارزیابی بازیابی اطلاعات، تعداد مناسبی به نظر می‌رسد. به همین سبب، تعداد ۳۲ کلیدواژه به‌صورت تصادفی طبقه‌ای نسبتی از *سرعنوان‌های موضوعی فارسی* انتخاب شد تا مبنای انتخاب کلیدواژه‌ها سلیقه‌ای نباشند.

پژوهش‌های اندکی در زمینه موتورهای جستجوی بومی انجام‌شده است که البته در همین پژوهش‌های اندک به پوشش و همپوشانی آن‌ها کمترپرداخته شده است و نیز اطلاعات چندانی درزمینه‌ی پوشش آن‌ها وجود ندارد. به همین سبب موتورهای جستجوی بومی به منظور بررسی انتخاب شدند. تقریباً نزدیک به ۱۰ مورد موتور جستجوی بومی وجود دارد که با مرور پیشینه پژوهش و مشاهده رتبه موتورهای جستجو در سایت الکسا[1]، و نیز بررسی اولیه، از این موتورهای جستجو در نهایت، چهار موتور پارسی‌جو، یوز، پارسیک و ریسمون برای ارزیابی انتخاب شدند.

هر کدام از این پرسش‌ها در موتورهای جستجو بررسی شد و سپس از نتایج بازیابی شده نمونه‌گیری به عمل آمد. برای تعیین حجم نمونه از جدول کوکران استفاده شد. با توجه به این‌که بیش‌تر از ۷۰ درصد کاربران فقط از ۱۰ نتیجه اول بازدید می‌کنند (اسپینگ و جانسون[2]، ۲۰۰۴؛ جانسون و اسپینگ، ۲۰۰۶) و نیز ارائه مرتبط‌ترین موارد در۱۰ نتیجه نخست وجود دارد (اسفندیاری مقدم و بهاری موفق، ۱۳۹۱)، در این پژوهش، فقط ۱۰ نتیجه نخست بررسی شد. بر این اساس جامعه آماری پژوهش متشکل از ۱۲۸۰ رکورد بازیابی‌شده بود که تعداد ۳۰۰ رکورد بر اساس فرمول کوکران به عنوان حجم نمونه مناسب انتخاب شد.

شیوهٔ سنجش میزان همپوشانی: طبق نظر اگه و گاورتس (۲۰۰۷) جهت سنجش همپوشانی موتور جستجوی A نسبت به B ، کسری از B را محاسبه می‌کنیم. سپس نمونه از موتور جستجوی B انتخاب و تعداد عناصری از این نمونه که به موتور جستجوی A نیز تعلق دارد، مشخص می‌شود.

$$O(A|B) = \frac{|A \cap B|}{|B|}$$

$$O(B|A) = \frac{|A \cap B|}{|A|}$$

در این فرمول‌ها مفهوم علائم به شرح ذیل است:

علامت $O(A|B)$ نشان‌دهنده‌ی همپوشانی A نسبت به B، علامت $O(B|A)$ نشان‌دهنده‌ی همپوشانی B نسبت به A، $|A \cap B|$ نشان‌دهنده‌ی تعداد مدارک مشترک بین دو مجموعه A و B، $|A|$ نشان‌دهنده‌ی تعداد مدارک مجموعه A، و علامت $|B|$ نشان‌دهنده‌ی تعداد مدارک مجموعه B است.

شیوهٔ سنجش میزان پوشش (اندازه پایگاه): با توجه به نظر باهارات و برودر (۱۹۹۸) و گیل و سیگنورینی (۲۰۰۵)، به منظور سنجش پوشش موتور جستجوی A نسبت به موتور جستجوی B (همپوشانی A و B) از فرمول زیر استفاده می‌شود.

$$\frac{size\ A}{size\ B} = \frac{O\ (A\ |\ B)}{O\ (B\ |\ A)}$$

نمونه‌گیری و برآورد همپوشانی و پوشش: اگر تعداد عناصر مشترک دو مجموعه و اندازه مجموعه‌ها مشخص باشند، با فرمول‌های موجود می‌توان میزان همپوشانی نسبی را محاسبه کرد و معمولاً اندازه مجموعه‌ها مشخص هستند؛ چون اندازه کتابخانه ها معلوم است و اندازه نظام‌های بازیابی اطلاعات را نیز می‌توان با فنون بازیابی اطلاعات به دست آورد. اما مسئله اصلی، تعیین تعداد عناصر مشترک میان دو مجموعه است؛ چون واقعاً غیرممکن است که تمام مدارک موجود را در کتابخانه‌ها و به خصوص نظام‌های بازیابی اطلاعات بررسی کرد و ناگزیر باید نمونه‌گیری کرد. بدین ترتیب، استفاده از آمار استنباطی الزامی است.

مطابق نظر اگه و گاورتس (۲۰۰۷) ابتدا باید کسری از B یعنی تعداد عناصر مشترک دو مجموعه را معین کرد. لذا نمونه (اندازه N_B) وجود دارد و باید تعیین کرد که چه تعداد از عناصر این نمونه به مجموعه A تعلق دارد و سپس مقادیر حاصل را در فرمول اخیراً یاد شده قرار داد و عدد حاصل را $\bar{X}_{(A|B)}$ نامید. به دلیل این که تقسیم یک نوع میانگین است، می‌توان از نظریه کلاسیک سطح اطمینان برای میانگین که مبتنی بر نظریه حد مرکزی است، استفاده کرد؛ پس همپوشانی A نسبت به B با احتمال ۹۵ درصد در فاصله زیر قرار دارد:

$$O(A|B) \in [\bar{X}_{(A|B)} - 1.96 \sqrt{\frac{\bar{X}_{(A|B)}(1-\bar{X}_{A|B})}{N_B - 1}}$$

$$\text{و } \bar{X}_{(A|B)} + 1.96 \sqrt{\frac{\bar{X}_{(A|B)}(1-\bar{X}_{A|B})}{N_B - 1}} \ ]$$

برآورد روایی و پایایی آزمون: روایی پژوهش با استفاده از مطالعات پژوهشگران و به ویژه متون مرتبط نظیر باهارات و برودر(۱۹۹۷)، گیل و سیگنورینی (۲۰۰۵)، اگه و گاورتس (۲۰۰۷) و نظرخواهی از استادان و پایایی آن از طریق آزمون-

1. https://www.Alexa.com
2. Jansen

بازآزمون با همبستگی۰/۷۲ مورد تایید قرار گرفت. در مطالعه مقدماتی، پنج کلیدواژه به صورت تصادفی انتخاب و دو بار در فاصله زمانی یک هفته‌ای به موتورهای جستجو ارائه، و همبستگی نتایج بازیابی‌شده در دو زمان متفاوت با استفاده از آزمون پیرسون محاسبه گردید. نتیجه آزمون پیرسون نشان داد که همبستگی معنی‌داری بین نتایج حاصل از دو آزمون وجود داشت (جدول۱).

جدول ۱. آزمون پیرسون جهت سنجش پایایی پژوهش

متغیر	تعداد	آماره آزمون پیرسون	سطح معنی‌داری
مدارک بازیابی شده	۶۰	۰/۷۲۱	۰/۰۰

همان گونه که در جدول ۱ مشاهده می‌شود، میزان همبستگی در دو دوره زمانی متفاوت، بیش از ۰/۷۲ بود و این میزان نشان داد که پژوهش از پایایی مناسبی برخوردار بود.

یافته‌ها

در اینجا نخست به پرسش اول پژوهش پاسخ داده شد و سپس فرضیه ۱ مورد آزمون قرار گرفت. به منظور پاسخ به پرسش ۱، سطح همپوشانی موتورهای جستجو با اطمینان ۹۵٪ اندازه‌گیری شد و نتایج آن در جدول ۲ و نمودار ۱ آمده است.

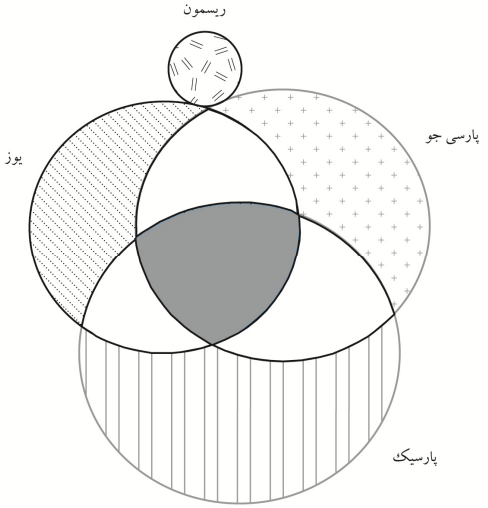
همان طور که در جدول ۲ مشاهده می‌شود، همپوشانی پارسیک نسبت به پارسی جو و همپوشانی پارسی جو نسبت به یوز بیشترین میزان را دارد و جالب این که موتورجستجوی ریسمون با هیچ کدام از موتورهای جستجو همپوشانی ندارد.

برای آزمون فرضیه نخست در باب معنی‌داری تفاوت در میزان همپوشانی، ابتدا بررسی نرمال بودن توزیع داده‌ها الزامی است. در این راستا، به منظور سنجش نرمال بودن توزیع داده‌ها از آزمون کولموگروف اسمیرنوف استفاده شد که نتیجه آزمون کولموگروف اسمیرنوف در جدول ۳ ارائه شده است.

همان طور که در جدول ۳ مشاهده می‌شود، میزان معنی‌داری

جدول ۲: برآورد همپوشانی موتورهای جستجو پارسیک، پارسی جو، یوز و ریسمون با سطح اطمینان ۹۵ درصد

میزان همپوشانی موتورها	در سطح نمونه	برآورد جامعه با سطح اطمینان ۹۵ درصد
همپوشانی یوز نسبت به ریسمون	۰	{۰ — ۰}
همپوشانی پارسی‌جو نسبت به ریسمون	۰	{۰— ۰}
همپوشانی پارسیک نسبت به ریسمون	۰	{۰— ۰}
همپوشانی ریسمون نسبت به یوز	۰	{۰— ۰}
همپوشانی پارسی‌جو نسبت به یوز	۰/۲۶۶	{۰/۳۶۵— -۰/۱۶۷}
همپوشانی پارسیک نسبت به یوز	۰/۱۷۳	{۰/۲۵۹— ۰/۰۸۷}
همپوشانی ریسمون نسبت به پارسی جو	۰	{۰— ۰}
همپوشانی یوز نسبت به پارسی‌جو	۰/۲۵۳	{۰/۳۵۲— ۰/۱۵۴}
همپوشانی پارسیک نسبت به پارسی‌جو	۰/۲۶۶	{۰/۳۶۵— ۰/۱۶۷}
همپوشانی ریسمون نسبت به پارسیک	۰	{۰— ۰}
همپوشانی یوز نسبت به پارسیک	۰/۱۶	{۰/۲۴۳— ۰/۰۷۷}
همپوشانی پارسی‌جو نسبت به پارسیک	۰/۱۴۶	{۰/۲۲۶ — ۰/۰۶۶}



نمودار ۱. همپوشانی چهار موتور جستجوی پارسیک، پارسی جو، یوز و ریسمون

جدول۳. آزمون کولموگروف اسمیرنوف برای سنجش نرمال بودن همپوشانی موتورهای جستجو

متغیر	تعداد	میانگین	انحراف معیار	آماره آزمون کولموگروف اسمیرنوف	سطح معنی‌داری
هم	۳۸۴	۰/۱۱	۰/۲۳	۹/۳۵	۰/۰۰

آزمون کولموگروف اسمیرنوف کم تر از ۰/۰۵ بود، لذا توزیع داده ها نرمال نبود؛ بدین ترتیب برای مقایسه همپوشانی موتورهای جستجو از آزمون کروسکال والیس استفاده شد که نتیجه این آزمون به‌صورت خلاصه در جدول ۴ ارائه شده است.

همان طور که جدول ۴ نشان می‌دهد، سطح معنی‌داری (۰/۰۰) آزمون کروسکال والیس کم‌تر از۰/۰۵ بود و این میزان نشان می‌دهد که تفاوت معنی‌داری بین همپوشانی موتورهای جستجو وجود داشت.

در ادامه به پرسش دوم پژوهش پاسخ داده شد و سپس فرضیه ۲ مورد آزمون قرار گرفت. به منظور پاسخ به پرسش ۲، سطح پوشش موتورهای جستجو با اطمینان ۹۵٪ اندازه‌گیری شد و نتایج آن در جدول ۵ و نمودار ۲ آمده است.

مشاهدات (جدول ۵) نشان داد که موتور جستجو پارسیک، بیش‌ترین پوشش و موتور جستجو ریسمون کم‌ترین پوشش را داشتند.

برای آزمون فرضیه دوم، ابتدا نرمال بودن توزیع داده‌ها بررسی گردید که نتایج آزمون کولموگروف اسمیرنوف درجدول ۶ ارائه شده است.

همان طور که در جدول۶ مشاهده می‌شود، سطح معنی‌داری آزمون کولموگروف اسمیرنوف۰/۰۰ کم تر از ۰/۰۵ است و این نشان می‌دهد که توزیع داده‌ها نرمال نیست. بدین ترتیب جهت سنجش معناداری تفاوت در پوشش موتورهای جستجو از آزمون کروسکال والیس استفاده شد که نتایج این آزمون در جدول ۷ ارائه شده است.

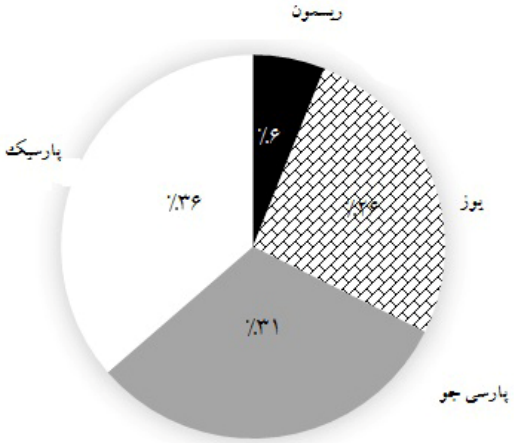
با توجه به جدول ۷ مشخص می‌شود که میزان معنی‌داری آزمون کروسکال والیس کم‌تر از ۰/۰۵ بود، لذا تفاوت معنی‌داری بین اندازه پایگاه موتورهای جستجو وجود داشت.

جدول ۴. آزمون کروسکال والیس برای سنجش معنی داری همپوشانی موتورهای جستجو

متغیر	تعداد	آماره آزمون کروسکال والیس	درجه آزادی	سطح معنی‌داری
هم پوشانی	۳۸۴	۱۰۹/۴۲	۱۱	۰/۰۰

جدول ۵. برآورد میزان پوشش چهار موتورجستجو پارسیک، پارسی جو، یوز و ریسمون با سطح اطمینان ۹۵ ٪

موتور جستجو	در سطح نمونه	برآورد در جامعه با سطح اطمینان ۹۵ درصد
پوشش پارسیک	۰/۳۶	{۰/۴۶۹–۰/۲۵۱}
پوشش پارسی جو	۰/۳۱	{۰/۴۱۵ – ۰/۲۰۵}
پوشش یوز	۰/۲۶	{۰/۳۵۹ – ۰/۱۶۱}
پوشش ریسمون	۰/۰۶	{۰/۰۸۷– ۰/۰۳۳}



نمودار ۲. میزان پوشش چهار موتور جستجوی پارسیک، پارسی جو، یوز و ریسمون

جدول ۶ آزمون کولموگروف اسمیرنوف برای سنجش نرمال بودن پوشش موتورهای جستجو

متغیر	تعداد	میانگین	انحراف معیار	آماره آزمون کولموگروف اسمیرنوف	سطح معنی داری
پوشش	۳۸۴	۰/۲۶	۰/۴۶	۸/۹۵	۰/۰۰

جدول ۷. آزمون کروسکال والیس برای سنجش معنی داری پوشش موتورهای جستجو

متغیر	تعداد	آماره آزمون کروسکال والیس	درجه آزادی	سطح معنی‌داری
پوشش	۳۸۴	۲۳۴/۲۳	۱۱	۰/۰۰

نتیجه‌گیری

نتایج پژوهش نشان داد همپوشانی موتورهای جستجو بومی در مقایسه با موتورهای عمومی در حد پایینی قرار داشت. همان طور که در جدول ۴ اشاره شده است، همپوشانی موتورهای بومی کم‌تر از ۳۷ درصد بود؛ در حالی که مطابق پژوهش گوهری و همکاران (۱۳۹۴) موتورهای جستجوی گوگل، یاهو و بینگ به ترتیب، ۴۲ ٪، ۴۸ ٪ و ۵۸ ٪

همپوشانی داشتند. این تفاوت حاکی از آن است که موتورهای جستجوی بومی در مقایسه با موتورهای جستجوی عمومی، سیاست نمایه‌سازی مستقل‌تر و متفاوت‌تری دارند. نکته جالب توجه این است که موتور جستجوی ریسمون هیچ همپوشانی با سایر موتورهای جستجو نداشت و سیاست نمایه‌سازی کاملاً متفاوتی از سایر موتورها را دنبال می‌کرد. این نتایج همراستا با نتایج پژوهش رجبی و نوروزی (۱۳۹۴) است. نتایج پژوهش نشان داد که موتور پارسیک بیش‌ترین میزان همپوشانی را با دیگر موتورهای جستجو داشت؛ درحالی که طبق پژوهش رجبی و نوروزی (۱۳۹۴)، موتور پارسیک هیچ گونه همپوشانی با سایر موتورها نداشت. شاید این تفاوت را بتوان به نوع نمونه و زمان متفاوت انجام این دو پژوهش نسبت داد. از سوی دیگر، موضوع مورد جستجو ممکن است در کسب نتایج متفاوت در این زمینه مؤثر باشد.

با کاوش در موتورجستجوی ریسمون، در برخی موضوعات هیچ رکوردی بازیابی نمی‌شد که نشان از محدودیت پوشش این موتور دارد و لزوم بازنگری در سیاست نمایه‌سازی این موتور جستجو را نشان می‌دهد. نتایج نشان داد که بین چهار موتور پارسیک، پارسی‌جو، یوز و ریسمون هیچ همپوشانی وجود نداشت. این نتیجه همسو با نتایج پژوهش راثر و همکاران (۲۰۰۸) بود. در پژوهش باهارت و برودر (۱۹۹۸) و اسپینگ و همکاران (۲۰۰۶) نیز همپوشانی بین چهار موتورجستجوی مورد بررسی درحد خیلی پایین (کمتر از ۱/۵ درصد) گزارش شد.

نتایج پژوهش درباره میزان پوشش موتورهای جستجو نشان داد که پارسیک، پارسی‌جو، یوز و ریسمون به ترتیب۳۶ درصد، ۳۱ درصد، ۲۶ درصد و ۶ درصد از وب نمایه‌پذیر را پوشش می‌دادند که این میزان اختلاف در پوشش موتورهای جستجو معنی‌دار بود. پوشش موتورهای جستجوی بومی در مقایسه با موتورهای جستجوی عمومی در حد پایینی قرار داشت. براساس نتایج پژوهش باهارات و برودر (۱۹۹۸)، موتور جستجو هات‌بوت در سال ۱۹۹۷ با پوشش ۴۷ درصد از وب نمایه‌پذیر، بیش‌ترین پوشش و اینفوسیک با پوشش ۱۸ درصد از وب نمایه‌پذیر کم‌ترین پوشش را داشتند. در مقایسه با نتایج به‌دست‌آمده در پژوهش حاضر که پارسیک با پوشش ۳۶ درصدی، بیش‌ترین، و ریسمون با پوشش ۶ درصدی، کمترین پوشش را داشتند.پس مشخص می‌شود که موتورهای جستجوی بومی پوشش بسیار کم‌تری در مقایسه با موتورهای جستجوی خارجی داشتند. با توجه به ارتقا و تقویت مداوم پایگاه نمایه موتورهای کاوش بین المللی، به نظر می‌رسد اختلاف فاصله پوشش موتورهای جستجوی بین‌المللی با موتورهای جستجو داخلی در زمان فعلی به مراتب بسیار بیش‌تر باشد. مطابق پژوهش گیل و سیگنورینی (۲۰۰۵)، موتور جستجو گوگل می‌تواند بیش از ۶۸ درصد وب نمایه‌پذیر را پوشش دهد که در مقایسه با اندازه پایگاه موتورهای جستجو بومی رقم چشمگیری است. حتی موتورجستجوی اسک که در آن پژوهش کم‌ترین میزان پوشش (۴۳ درصد) را در مقایسه با موتورهای بررسی شده نشان داد، نسبت به پارسیک میزان بیش‌تری از وب را پوشش می‌داد. لذا هنوز جای کار زیادی برای موتورهای جستجوی بومی وجود دارد.

براساس نتایج پژوهش‌های ارزشیابی میزان بازیابی اطلاعات می‌توان اطلاعات مفیدی را هم به کاربران و هم به مسؤولان و دست‌اندرکاران امور نظام‌های بازیابی اطلاعات کشور ارائه کرد. بدین ترتیب، کاربران با استفاده از نتایج این پژوهش‌ها، موتور جستجوهای کارآمد را برای جستجو خواهند شناخت و برای بازیابی اطلاعات مورد نظر از موتور کارآمد استفاده خواهند کرد و این موجب صرفه‌جویی در زمان آن‌ها خواهد شد. همچنین این نوع پژوهش‌ها برای مسؤولان امور نظام‌های بازیابی اطلاعات مفید است؛ ازآنجایی‌که در طی این پژوهش‌ها نقاط قوت و ضعف هر کدام از موتورهای جستجو مشخص می‌شود و متولیان هر کدام از موتورها با توجه به این نتایج می‌توانند نقاط قوت خود را تقویت و همچنین نقاط ضعف خود را برطرف کنند. ازاین‌رو پژوهش‌های ارزیابانه بازیابی اطلاعات از اهمیت فزاینده‌ای برخوردار هستند.

با توجه به نتایج، به تمامی پژوهشگران، کتابداران و دانشجویان پیشنهاد می‌شود که به هنگام کاوش و بازیابی اطلاعات از وب، جستجوی خود را در چندین موتور جستجو انجام دهند تا به اطلاعات جامعی در موضوع مورد نیاز خود دست یابند. در صورتی که کاربران به ناگزیر زمان کافی برای جستجو در همه موتورهای جستجو را ندارند، پیشنهاد می شود که موضوع خود را در موتور پارسیک و سپس در موتور پارسی‌جو کاوش نمایند؛ چرا که با کاوش در این دو موتور بیش از ۶۵ درصد وب دسترس‌پذیر است.

به متولیان امور موتورهای جستجو نیز پیشنهاد می شود، با استفاده از تجهیزات فنی لازم، میزان اندازه پایگاه موتورها را تقویت نمایند؛ چرا که اندازه پایگاه موتورهای بومی در مقایسه با موتورهای عمومی در سطح پایینی قرار دارند و درعین‌حال از توجه به معیارهایی از قبیل قابلیت‌ها و امکانات جستجویی، طراحی رابط کاربری مورد پسند، پیشنهاد کلیدواژه مناسب و مشابه آن مغفول نمانند.

search engine transaction logs. Information processing & management, 42(1), 248-263.
Khalili K(1993). Farhange Moshtagate Masader Farsi. Tehran: Moseseye Motaleat va Tahgigate Farhangi.
Kosha k (2002). Abzarhaye Jostejo Internet: Osul, maharatha va emkanate jostejo. Tehran: Ketabdar. (Persian)
Lewandowski D (2012). A Framework for Evaluating the Retrieval Effectiveness of Search Engines In Jouis, Christophe, Next Generation Search Engine: Advanced Models for Information Retrieval .Hershey, PA: IGI Global, retrieved 20 Decembers 2016 from http:// www.igi-global.com/book/next-generation-search-engines/59723
Mitra A, Awekar A (2017). On Low Overlap Among Search Results of Academic Search Engines. arXiv preprint arXiv. 823-824.
Mohammad Esmaeel S, gaemi M (2009). Mogayeseye Mezane Hamposhaneye Natayaje Bazyabi Shode dar Motorhaye Kavosh , Abarmotorhaye Kavoshe dar Bazyabeye Ettlaat Keshavarzi. Mahnameye Ettlae yabi , Ettlae Rasani, (21),55-61.
Pappas S (2016). How Big Is the Internet, Really? Retrievted 8 march 2017 http://www.livescience.com/54094-how-big-is-the-internet.html
Powell R (2000). Basic research methods for librarians. Translated by Najla Hriry.Tehran: Asar Nafes.
Poyer R (1984). Journal Article Overlap among Index Medicus Science Citation Index, Biological Abstracts, and Chemical Abstracts. Bull. Med. Libr. Assoc. 72(4).
Rajabi M, Norozi Y (2015). Motorhaye Jostojoye Farsi: Arzyabeye Emkanat Jostejo, Bazyabeye Etlaat,M0ezane Jameeyat va Maneeeyat va Taeen Hamposhaneye anha. Motalaat Meleye Katabdary va Sazmandeheye Etlaat, 26(3) 133-150.
Rather RA, Lone FA, SHah GJ (2008). Overlap in web search results: A study of five searches Engines. Library philosophy and practice. 226
Saracevic, T. (1996). Relevance reconsidered. In Proceedings of the second conference on conceptions of library and information science (CoLIS 2) (pp. 201-218).
Spink A, Janson B J (2004). A study of web search trends. Webology,1(2). Retrieved 10 December 2016 from www.webology.org.
Spink A, Jansen B J, Blakely C, Koshman S (2006). A study of results overlap and uniqueness among major web search engines. Information Processing & Management, 42(5), 1379-1391.
Vickrey B (2001). Information science in theory and practice. Transleted by Abdolhosien Faraj Pahlo. Mashhad: Daneshgahe Ferdowsiye Mashhad. (Persian)
Wood J, Flanagan C, Kenned H Edward (1972). Overlap in the Lists of Journals Monitored by Bios

References

Anderson B (2006). Indexing the Internet. Behavioral & Social Sciences Librarian, 25(1), 135-139.
Bharat K, Broder A (1998). A Technique for Measuring the Relative Size and Systems, 30(1), 379-388.
Buckland MK, Hindle A, Walker P. M (1975). Methodological problems in assessing the overlap between bibliographical files and library holdings. Information Processing and Management 11(3–4), 89–105.
Clarke SJ (2000) .Search Engines for the World Wide Web, Journal of Internet Cataloging, 2(3-4), 81-93.
Davarpanah M (2008). Hostejoye Ettlaate Elmi va Pazoheshi dar Manabee CHapi va Elektroniki. Tehran: Dabezesh.
Egghe L (2006). Properties of the n-overlap vector and n-overlap similarity theory. Journal of the American Society for Information Science and Technology57 (9)1165-1177.
Egghe L, Goovaerts M (2007). A note on measuring overlap. Journal of Information Science.33 (2), 189-195.
Fattahi R(2004). Tahlele Avamele Moather bar Nesbi Bodane Rabt dar Nezam Bazyabeye Etlaat. Ettelae SHenasi, 2 (1),7-22.
Gohari S, maktabifard L, Jamaleye Mehmamoee H (2015). Sanjeshe mezane Tekrar Bazyabeye Etlaat Farsi az Web ba Mogayeseye Motorhaye Kavoshe Emomi. Tahgigate Ettlae Ketabdary va Ettlae Resani Daneshgahi, 49 (2),239-254.
Gulli A, Signorini A (2005). The indexable web is more than 11.5 billion pages. In Special interest tracks and posters of the 14th international conference on World Wide Web (pp. 902-903). ACM.
Hood W W, Wilson C S (2003). Overlap in bibliographic databases. Journal of the American Society for Information Science and Technology, 54(12), 1091-1103.
Isfandyari Moghaddam A, Bahari Movaffagh, Z(2012).The Overlap Rate of Searching Medical Keywords in General Search Engines.Modereyat Etelaat Salamat,9(2).203-214.
Isfandyari Moghaddam A (2005).Barasiye Natayeje jostejo dar Abarmotorhaye Kavosh va Motorhaye Tahte Poshesh anha az janbeye Hamposhani va Rotbe Bandeye .Payannameye Karshenasiye Arshad.Daneshgahe Ferdowsiye Mashhad.
Isfandyari Moghaddam A, Parirokh M (2006). A comparative study on overlapping of search results in metasearch engines and their common underlying search engines. Library Review, 55(5), 301-306.
Jahangard N (2017). Chahar Melyard Safheye Farsi dar Web Retrevited in 20 Desamber 2017 from www.irna.ir/fa/News/82498650.
Jansen B J, Spink A. (2006). How are we searching the World Wide Web? A comparison of nine

is, CAS, and E. Journal of the American Society for Information Science.

Wood J, Flanagan C, Kenned (1973). Overlap Among the Journal Articles Selected for Coverage b BIO-SIS, CAS, and Ei. Journal of the American Society for Information Science.

Human Information Interaction. Vol. 4, No. 3
http://hii.khu.ac.ir
Original Article

The Overlap and Coverage of 4 Local Search Engines of Parsijoo, Yooz, Parseek and Rismoun

Mohsen Nowkarizi: Associate professor of Knowledge and Information Science, Ferdowsi University of Mashhad, Iran. (Corresponding author) mnowkarizi@um.ac.ir
Mahdi Zeynali Tazehkandi: MA Student of Knowledge and Information Science, Ferdowsi University of Mashhad, Iran.

Abstract

Background and Aim: The aim of this study was to measure the overlap of 4 local Persian search engines of Parsijoo, Yooz, Parseek, and Rismoun and to compare the capabilities of these engines in covering indexable web.

Methods: This was an applied and evaluative research. To collect data, a keyword-based method was used. First, the selected keywords were entered into the search engines and then a sample was extracted of the retrieved records. Finally, based on the existence or absence of these records in the search engines, the necessary data were gathered. Accordingly to analyze the data, inferential statistical methods were used.

Results: The relative overlap of the Parseek compared to that of Parsijoo and Parsijoo's one compared to Yooz was 26 percent on average and Parseek had the most recall. Rismoun had not any common records with the other investigated search engines. Three search engines (Parseekc, Parsijoo and Yooz retrieved 27 common records out of 225 recalled records; there was a significant difference between the relative overlap of the 4 search engines. Also, on average, Parseel, Parsigoo, Yooz and Rismoun covered respectively 38, 31, 26, and 6 percent of the indexable web. There was a significant difference between the coverage of the 4 search engines.

Conclusion: It seems that each search engine has a different indexing policy, and users need to search for more than one search engine to get comprehensive information about an issue. It can be predicted that by foraging in two search engines, Parseek and Parsijoo, one may access 70 percent of the indexable web.

Keywords: Information retrieval evaluation, Persian Web, Search engine, Coverage, Overlap, Parseek, Parsijoo, Yooz, Rismoun