

کارآمدی فیلد "زبان" در موتورهای کاوش بین‌المللی برای بازیابی نتایج مرتبط در زبان فارسی و عربی (مطالعه مقایسه‌ای: گوگل، بینگ، و یاهو)

مریم یقین | عبدالرسول جوکار

چکیده

هدف: کارآمدی فیلد "محدودیت زبانی" در سه موتور کاوش گوگل، بینگ، و یاهو و تعیین کارآمدترین موتور کاوش در ارائه نتایج مرتبط در زبان فارسی و عربی. **روش‌شناسی:** پیمایش بر روی ۳۸ واژه هم‌نویسه در زبان فارسی و عربی انجام شد. **یافته‌ها:** موتور کاوش گوگل در بازیابی و رتبه‌بندی نتایج مرتبط با زبان هدف، عملکرد ضعیف‌تری نسبت به موتور کاوش بینگ و یاهو داشت. سه موتور کاوش در بازیابی نتایج نامرتب با زبان هدف، تفاوت معناداری با یکدیگر نداشتند؛ اما در رتبه‌بندی نتایج نامرتب با زبان هدف، تفاوت معناداری در عملکرد موتور کاوش گوگل مشاهده شد. همچنین مشخص شد تأثیری در بازیابی نتایج نامرتب با زبان هدف ندارد.

نتیجه‌گیری: کاربران در بینگ و یاهو با اقبال بیشتری برای بازیابی نتایج مرتبط با زبان فارسی و عربی مواجه هستند.

کلیدواژه‌ها

موتور کاوش، محدودیت زبان، کارآمدی، بازیابی اطلاعات، رتبه‌بندی نتایج

کارآمدی فیلد "زبان" در موتورهای کاوش بین‌المللی برای بازیابی نتایج مرتبط در زبان فارسی و عربی (مطالعه مقایسه‌ای: گوگل، بینگ، و یاهو)

مریم یقظین^۱

عبدالرسول جوکار^۲

تاریخ دریافت: ۹۴/۱۲/۲۵

تاریخ پذیرش: ۹۵/۰۳/۰۸

مقدمه

بیشتر کاربران از موتورهای کاوش برای جستجو و دستیابی به اطلاعات مرتبط با نیازهای اطلاعاتی خود استفاده می‌کنند. با اینکه موتورهای کاوش بسیاری در دسترس است، گاه کاربران نمی‌توانند به سرعت به اطلاعات مطلوب دست یابند. از سوی دیگر، هر موتور کاوش نتایج متفاوتی برای پرس‌وجویی^۳ یکسان ارائه می‌دهد (کاشیک^۴، ۲۰۱۲). بنابراین، ارزیابی موتورهای کاوش برای دریافت مرتبط‌ترین اطلاعات از اهمیت بسیاری برخوردار است.

ارزیابی کیفیت موتورهای کاوش از نظر ویژگی‌های جستجو موارد مختلفی همچون عملکرد صحیح در جستجوی عنوان، عبارت‌های خاص، یا فایل‌های خاص؛ استفاده از عملگر بولی؛ توانایی جستجو در یک زبان، یک کشور، یا ناحیه خاص؛ یافتن صفحات مشابه؛ یا جستجو در نشانی وبی و دوره تاریخی خاص است. این ویژگی‌ها سبب می‌شود که دستیابی به اطلاعات مرتبط با نیازهای اطلاعاتی کاربران تسهیل شود و بازیابی بهتری را تجربه کنند (لواندوسکی و هشتاتر^۵، ۲۰۰۸). از میان ویژگی‌های مورد اشاره، امکان استفاده از فیلد زبان برای محدود کردن نتایج و بازیابی اطلاعات مرتبط از اهمیت ویژه‌ای برخوردار است. کاربرانی که اطلاعات را در محدوده زبانی خاص جستجو می‌کنند ممکن است نتوانند صفحات دیگر زبان‌ها را بخوانند یا چنانچه با کلمات هم‌نویسه در زبان‌های مختلف مواجه شوند نتایج

۱. دانشجوی دکترای علم اطلاعات و دانش‌شناسی، دانشگاه شیراز (نویسنده مسئول)

۲. استاد گروه علم اطلاعات و دانش‌شناسی، دانشگاه شیراز
ajowkar2003@yahoo.com

3. Query

4. Kaushik

5. Lewandowski & Höchstötter

بازیابی شده از دیگر زبان‌ها، حتی در صورت فهم کاربران، نامرتب تلقی می‌شوند؛ اگر این فیلد عملکرد درستی داشته باشد اطلاعات نامرتب از مجموعه نتایج حذف می‌شود. از دهه ۹۰ میلادی بر روی شناسایی خودکار زبان متون فعالیت‌هایی صورت گرفته است (کاوناو و ترنکل^۱، ۱۹۹۴؛ دانیگ^۲، ۱۹۹۴) و پژوهشگران به دنبال شناسایی چنین امکانی در وب بوده‌اند (مارتینز و سیلوا^۳، ۲۰۰۵) البته عواملی مانند مکان سرور نیز در تشخیص زبان صفحات وب بی‌تأثیر نبوده است (لواندوسکی، ۲۰۰۸).

با افزایش روزافزون کاربران و صفحات غیرانگلیسی زبان، عملکرد صحیح موتورهای کاوش برای دسترسی به اطلاعات مرتبط با زبان کاربر اهمیت بیشتری یافته است. از سوی دیگر، به سبب مهجوری زبان فارسی در مقایسه با زبان‌هایی چون انگلیسی، پایین بودن شمار گویشوران فارسی در سراسر جهان، کمبود موتور کاوش خاص زبان فارسی، و شباهت ظاهری زبان فارسی و عربی، پژوهش در این باره اهمیتی ویژه می‌یابد تا مشخص شود موتورهای کاوش بین‌المللی در بازیابی صحیح نتایج فارسی و عربی چقدر کارآمد هستند. به همین منظور، تلاش شد تا این مسئله در سه موتور کاوش گوگل، بینگ، و یاهو بررسی و مقایسه شود.

از مهم‌ترین تلاش‌های صورت گرفته در این زمینه می‌توان به پژوهش رجبی و نوروزی (۱۳۹۴) اشاره کرد. آنان با شناسایی و ارزیابی امکانات جستجو در موتورهای جستجوی فارسی دریافتند که موتورهای جستجوی فارسی از امکانات جستجوی مناسبی برخوردارند و عملکرد آنها قابل اطمینان است؛ اما به لحاظ بازیابی نتایج مرتبط و میزان جامعیت، بین موتورهای جستجو فاصله و تفاوت معناداری وجود دارد.

نتایج پژوهش دری (۱۳۹۳) نشان داد که موتورهای کاوش معنایی هاکیا^۴، داک‌داک‌گو^۵، کلاز^۶، لکس^۷، و فکت‌بایتز^۸ در شاخص ویژگی‌های عادی و معنایی، عملکرد مطلوب و کارایی مورد انتظار را نداشتند. از نظر جستجو در زبان خاص نیز هیچ‌یک دارای این ویژگی نبودند.

شریف (۱۳۹۱) نیز با استفاده از روش تجربی، تغییرات عملکرد یاهو و گوگل را از نظر پوشش کمی - زمانی نمایه‌سازی و نقش عناصر فراداده‌ای در رتبه‌بندی صفحه‌های وب بررسی کرد و نشان داد که این دو موتور کاوش در بُعد کمی - زمانی و بُعد کیفی عناصر ابرداده‌ای تفاوت چندانی در عملکرد خود نداشته‌اند.

محمداسمعیل و کیایی (۱۳۹۰) نیز دریافتند که موتور کاوش یاهو و ابرموتور کاوش کری‌گاید^۹ بیشترین مدارک فیزیک را بازیابی کردند و موتور کاوش عمومی

1. Cavnar & Trenkle
2. Dunning
3. Martins & Silva
4. Hakia
5. DuckDuckGo
6. Cluuz
7. Lexxe
8. Factbites
9. Curry Guide

آ-اوال^۱ و ابرموتور کاوش اینفو بیشترین درصد همپوشانی را با سایر موتورهای کاوش عمومی و ابرموتورها دارند.

ژانگ، فی، و لی^۲ (۲۰۱۳) در پژوهشی با مقایسه کارآمدی ویژگی‌های جستجو در موتورهای کاوش گوگل، گوگل چینی، و بایدو^۳ اعلام کردند که گوگل بهترین عملکرد را در بازیابی و رتبه‌بندی صفحات دارد.

کیم، فیلد، و کارترایت^۴ (۲۰۱۲) با مطالعه رفتار کاربران در جستجوی کتاب دیجیتال در کتابخانه دیجیتالی و موتورهای کاوش وب و استفاده از "محدودیت زبانی" در جستجوی پیشرفته نشان دادند که کاربران اغلب برای محدود کردن نتایج جستجو به ترتیب از فیلدهای عنوان، نویسندگان، ناشران، و زبان استفاده می‌کنند. لواندوسکی (۲۰۰۸) نیز دریافت که هیچ‌یک از موتورهای کاوش در ارائه نتایج با زبان رابط کاربری مورد استفاده مشکلی ندارند، اما گوگل و ام.اس.ان. هنگام محدود کردن نتایج به زبان خارجی به درستی عمل نمی‌کنند؛ و صفحاتی با زبان‌های متفاوت از زبان رابط کاربری رتبه‌های پایین‌تری را به خود اختصاص می‌دهند.

فتاحی^۵ و همکاران (۲۰۰۸) با انتخاب دو ویژگی جستجوی پیشرفته (جستجوی عنوان دقیق و جستجوی نشانی وبی) به مطالعه تأثیر اصطلاحات غیرموضوعی^۶ و نیمه‌موضوعی^۷ در نتایج جستجو پرداختند و نشان دادند که در صورت محدود کردن کاوش به جستجوی عنوان دقیق و نشانی وب نتایج بهبود می‌یابد. ژانگ و لین^۸ (۲۰۰۷) نیز به این نتیجه رسیدند که موتورهای کاوش گوگل، ای‌زدتوفایند^۹ و آنالین‌لینک^{۱۰} به ترتیب، بهترین موتورهای کاوش در پشتیبانی از زبان‌های متفاوت هستند. با اینکه این موتورهای کاوش از زبان‌های مختلف پشتیبانی می‌کنند، اما فقط تعداد اندکی از آنها هستند که امکان ترجمه بین زبانی دارند. علاوه بر آن، پشتیبانی این موتورها از زبان‌های مختلف در سطح نحوی - و نه معنایی - باقی مانده است. غانی، جونز، و لدنیک^{۱۱} (۲۰۰۵) ۱۰۰ صفحه وب اسلوونیایی زبان را که با فیلتر زبان محدود شده بود به‌طور تصادفی انتخاب و بررسی کردند. نتایج نشان داد که ۹۹ درصد صفحات به زبان اسلوونیایی و ۹۰ تا ۹۵ درصد صفحاتی که در دسته غیراسلوونیایی زبان دسته‌بندی شده بودند واقعاً غیراسلوونیایی بودند.

یافته‌های لواندوسکی (۲۰۰۴) در زمینه جستجو با محدودیت تاریخی در موتورهای کاوش گوگل، تئوما^{۱۲}، و یا هو نشان داد که جستجو با محدودیت تاریخی در این موتورهای کاوش به درستی کار نمی‌کند. اسروکا^{۱۳} (۲۰۰۰) نیز با مطالعه توانایی موتورهای کاوش وب برای بازیابی اطلاعات لهستانی دو موتور کاوش

1. AOL
2. Zhang, Fei, & Le
3. Baidu
4. Kim, Field, & Cartright
5. Fattahi
6. Non-topical terms
منظور از اصطلاحات موضوعی، اصطلاحاتی است که بار موضوعی دارند و موضوع مدرک را نشان می‌دهند. مانند سرعنوان‌های موضوعی و یا واژه Globalization. اما اصطلاحات غیرموضوعی، اصطلاحاتی مانند Introduction عبارت از عبارت globalization و اصطلاحات نیمه‌موضوعی اصطلاحاتی مانند Prevention ... Risk of
7. Semi-topical terms
8. Zhang & Lin
9. EZ2Find
10. Onlinelink
11. Ghani, Jones, & Mladenic
12. Teoma
13. Sroka

عمومی اینفوسیک^۱ و آلتاویستا^۲ و سه موتور کاوش بومی را مقایسه کرد و نشان داد که موتورهای کاوش اینفوسیک و اونت.پی.ال^۳ بیشترین میزان دقت در بازیابی اطلاعات مرتبط را داشتند. علاوه بر این، موتور کاوش اینفوسیک، بیشترین سرعت بازیابی را به خود اختصاص داد.

با توجه به پژوهش‌های پیشین ملاحظه می‌شود که تاکنون پژوهشی کارآمدی ویژگی "محدودیت زبانی" برای تشخیص نتایج فارسی و عربی در موتورهای کاوش بین‌المللی را بررسی نکرده است. به همین دلیل، پژوهش حاضر به این مهم پرداخت و در این راستا مؤلفه‌های زیر مدنظر قرار گرفت:

وضعیت میانگین نتایج نامرتب با زبان هدف در موتورهای کاوش مورد بررسی، تفاوت معناداری آنها، رتبه‌بندی نتایج نامرتب، و تأثیر ترافیک صفحات نامرتب با زبان هدف بر رتبه‌بندی آنان.

روش‌شناسی

در این پیمایش توصیفی در گام نخست، سه موتور کاوش گوگل، بینگ، و یاهو انتخاب شد. دلیل استفاده از این موتورهای کاوش در وهله اول، پوشش بین‌المللی و پشتیبانی آنها از زبان فارسی و عربی بود. به علاوه، این موتورها در سال ۲۰۱۵ به ترتیب در صدر موتورهای کاوش محبوب قرار داشتند^۴. فرض پژوهش حاضر این است که این گرایش جهانی برای ایران نیز صادق است.

در گام بعد، سیاهه‌ای از ۳۸ کلمه هم‌نویسه^۵ در زبان فارسی و عربی انتخاب شدند. به طور مثال، کلمه "بهار" در دو زبان فارسی و عربی به کار می‌رود. در زبان فارسی اغلب به معنای "فصل بهار" است، اما در زبان عربی به معنای "ادویه" یا "چاشنی" به کار می‌رود. دلیل انتخاب کلمات هم‌نویسه میان زبان فارسی و عربی به عنوان پرس و جو این بود که نتایجی در هر محدودیت زبانی (چه به زبان فارسی و چه به زبان عربی) بازیابی شوند تا امکان ارزیابی ویژگی محدودیت زبانی موتورهای کاوش فراهم شود. سیاهه واژگان منتخب در جدول ۱ آمده است.

جدول ۱. سیاهه واژگان منتخب

| مخابرات | بهار | زبون | حوصله | تقلیدی | صورت | خسیس |
|---------|------|-------|-------|--------|---------|------|
| منظور | ناظر | اداره | شراب | مجتمع | مکتب | نفر |
| دغدغه | برق | ملت | ادویه | تخلف | عکس | نشاط |
| مجرم | موظف | مصرف | تورم | اسباب | جمعیت | ملی |
| ملی | رقیب | غرور | لحیم | وجه | معلومات | ساری |
| میمون | رسوم | شوکت | | | | |

1. Infoseek
2. AltaVista
3. Onet.pl
4. <http://www.ebizmba.com>,
www.alexa.com
۵. کلماتی که به ظاهر یکسان هستند، اما با معانی متفاوت در دو زبان به کار می‌روند

در گام بعد، هر واژه دوبار در صفحه جستجوی پیشرفته هر موتور کاوش، یکبار با محدودیت زبان فارسی و بار دیگر با محدودیت زبان عربی جستجو شد. سپس، ۲۰ نتیجه اول هر جستجو، به منظور تعیین فارسی یا عربی بودن نتایج بررسی شد.^۱ نتایج نامرتب به آن دسته از نتایج به زبان عربی اطلاق می‌شود که پس از محدودیت زبان به فارسی بازبایی شدند. به علاوه، آن دسته از نتایج به زبان فارسی پس از محدودیت زبان به عربی نیز نامرتب با پرس و جو بودند.

دلیل بررسی ۲۰ نتیجه اول نیز این بود که صفحات اول بازبایی شده بهترین و مرتبط‌ترین نتایجی است که هر موتور کاوش ارائه می‌دهد. مطالعه رفتار کاربران در موتورهای کاوش هم نشان داده است که کاربران فقط چند نتیجه نخست را بررسی می‌کنند و به سراغ بقیه نتایج جستجو نمی‌روند (لواندوسکی، ۲۰۰۸).

از آنجا که تعیین زبان فارسی یا عربی نتایج دارای اطلاعات غیرمتنی امکان‌پذیر نبود، ۲۰ نتیجه اول که اطلاعات متنی داشتند بررسی شد. ارزیابی عملکرد هر موتور کاوش برای هر پرس و جو و مقایسه آنان با تعیین فراوانی و رتبه نتایج نامرتب بازبایی شده بررسی شد. از آنجا که وب‌سایت الکسا^۲، براساس ترافیک صفحات وب، رتبه‌ای به هر صفحه اختصاص می‌دهد، نشانی وبی هر نتیجه نامرتب در وب‌سایت الکسا بررسی و رتبه جهانی^۳ آن ثبت شد تا مشخص شود که آیا ترافیک صفحه یا تعداد زیاد بازدیدکنندگان صفحه نامرتب بر بازبایی آن در ۲۰ نتیجه اول تأثیرگذار بوده است یا خیر.

به منظور تأمین اعتبار، تمامی شرایط به دقت کنترل شد. متخصصان، سیاهه واژگان را تأیید کردند و برای اجتناب از تأثیر زبان رابط کاربری بر نتایج، در تمامی موارد جستجو، زبان رابط کاربری موتور کاوش، یکسان (عربی) در نظر گرفته شد. تمامی مراحل جستجو و گردآوری داده‌ها از آی.پی. یکسان انجام شد تا از تأثیر جغرافیا بر نتایج پیشگیری شود. از آنجا که هر روز بر تعداد منابع موجود در وب افزوده می‌شود، به تناسب حجم نمایه و بانک اطلاعاتی موتور کاوش نیز افزوده می‌شود. به همین دلیل، برای یکسان‌سازی شرایط برای سه موتور کاوش، هریک از کلیدواژه‌ها، در بازه زمانی سه روزه، در تاریخ ۱۵ تا ۱۷ مرداد ۱۳۹۴ در هر سه موتور کاوش جستجو شد. گردآوری داده‌ها با مشاهده فراوانی و رتبه نتایج نامرتب بازبایی شده در هر دوبار جستجو در هر موتور کاوش به تفکیک ثبت شد. تجزیه و تحلیل‌ها به کمک نرم‌افزار SPSS و با استفاده از فنون آمار توصیفی (فراوانی و درصد) و استنباطی (آزمون‌های تحلیل واریانس، کروسکال والیس^۴، یومان-ویتنی^۵، و همبستگی اسپیرمن^۶) صورت گرفت.

۱. در مجموع ۴۵۶۰ نتیجه (۲۰×۲۸×۲۸) بررسی شد.

2. www.alexa.com

3. Global rank

4. Kruskal-Wallis

5. Mann-Whitney U

6. Spearman

یافته‌ها

بازیابی نتایج نامرتبط با زبان هدف

در دو موتور کاوش بینگ و یاهو پس از محدود کردن به زبان فارسی، هیچ نتیجه نامرتبطی بازیابی نشد و تمامی نتایج (۱۰۰ درصد) به زبان فارسی بود؛ اما، در موتور کاوش گوگل ۱۴ نتیجه بازیابی شده (تقریباً ۲ درصد) به زبان عربی بازیابی شد. از سوی دیگر، با محدود کردن به زبان عربی، در دو موتور کاوش بینگ و یاهو، حدود ۴ درصد از نتایج به زبان عربی بازیابی نشد.^۱ اما موتور کاوش گوگل با بازیابی نشانی ۶۸ مدرک (تقریباً ۹ درصد) به زبان فارسی از دو موتور کاوش دیگر عملکرد ضعیف‌تری داشت (جدول ۲).

جدول ۲. وضعیت موتورهای کاوش به‌لحاظ فراوانی و درصد نتایج نامرتبط با زبان هدف

| موتور کاوش | نتایج نامرتبط با زبان هدف فارسی | | نتایج نامرتبط با زبان هدف عربی | |
|------------|---------------------------------|------|--------------------------------|------|
| | فراوانی | درصد | فراوانی | درصد |
| گوگل | ۱۴ | ۱/۸۴ | ۶۸ | ۸/۹۵ |
| بینگ | ۰ | ۱۰۰ | ۲۷ | ۳/۵۵ |
| ياهو | ۰ | ۱۰۰ | ۲۹ | ۳/۸۲ |

رتبه‌بندی نتایج نامرتبط با زبان هدف

جدول ۳ توزیع فراوانی رتبه‌های نتایج نامرتبط را در موتور کاوش گوگل پس از محدود کردن به زبان فارسی نشان می‌دهد.

جدول ۳. توزیع فراوانی رتبه‌های نتایج نامرتبط در موتور کاوش گوگل با زبان فارسی

| رتبه | فراوانی رتبه | درصد فراوانی رتبه | رتبه | فراوانی رتبه | درصد فراوانی رتبه |
|------|--------------|-------------------|------|--------------|-------------------|
| ۱ | ۰ | ۰ | ۱۱ | ۱ | ۷/۱ |
| ۲ | ۰ | ۰ | ۱۲ | ۰ | ۰ |
| ۳ | ۰ | ۰ | ۱۳ | ۰ | ۰ |
| ۴ | ۰ | ۰ | ۱۴ | ۱ | ۷/۱ |
| ۵ | ۰ | ۰ | ۱۵ | ۲ | ۱۴/۳ |
| ۶ | ۲ | ۱۴/۳ | ۱۶ | ۱ | ۷/۱ |
| ۷ | ۰ | ۰ | ۱۷ | ۰ | ۰ |
| ۸ | ۳ | ۲۱/۴ | ۱۸ | ۰ | ۰ |
| ۹ | ۱ | ۷/۱ | ۱۹ | ۲ | ۱۴/۳ |
| ۱۰ | ۱ | ۷/۱ | ۲۰ | ۰ | ۰ |

۱. با توجه به اینکه کلمات پرس‌وجو صرفاً در زبان فارسی و عربی (نه زبان دیگر) هم‌نویسه بوده‌اند، نتایج نامرتبط فقط به زبان مقابل بوده است.

داده‌ها حاکی از آن است که در موتور کاوش گوگل، رتبه ۸ بیشترین فراوانی نتایج نامرتبط را به خود اختصاص داده است (تقریباً ۲۱ درصد مجموع فراوانی رتبه‌های نامرتبط). پس از آن، رتبه‌های ۶، ۱۵، و ۱۹ بیشترین فراوانی را داشتند که هرکدام حدود ۱۴ درصد مجموعه فراوانی رتبه‌های نامرتبط را به خود اختصاص دادند. همچنین، رتبه‌های ۱ تا ۵، ۷، ۱۲، ۱۳، ۱۷، ۱۸، و ۲۰ با کمترین فراوانی، در هیچ‌یک از پرس‌وجوها نتایج نامرتبط نداشتند.

جدول ۴. توزیع فراوانی رتبه‌های نتایج نامرتبط در موتورهای کاوش با زبان عربی

| موتور کاوش | رتبه | فراوانی رتبه | درصد فراوانی رتبه | رتبه | فراوانی رتبه | درصد فراوانی رتبه |
|------------|------|--------------|-------------------|------|--------------|-------------------|
| گوگل | ۱ | ۲ | ۲/۹ | ۱۱ | ۳ | ۴/۴ |
| | ۲ | ۱ | ۱/۵ | ۱۲ | ۶ | ۸/۸ |
| | ۳ | ۱ | ۱/۵ | ۱۳ | ۳ | ۴/۴ |
| | ۴ | ۱ | ۱/۵ | ۱۴ | ۵ | ۷/۴ |
| | ۵ | ۵ | ۷/۴ | ۱۵ | ۳ | ۴/۴ |
| | ۶ | ۲ | ۲/۹ | ۱۶ | ۴ | ۵/۹ |
| | ۷ | ۳ | ۴/۴ | ۱۷ | ۴ | ۵/۹ |
| | ۸ | ۲ | ۲/۹ | ۱۸ | ۴ | ۵/۹ |
| | ۹ | ۳ | ۴/۴ | ۱۹ | ۴ | ۵/۹ |
| | ۱۰ | ۴ | ۵/۹ | ۲۰ | ۸ | ۱۱/۸ |
| بینگ | ۱ | ۰ | ۰ | ۱۱ | ۱ | ۳/۷ |
| | ۲ | ۲ | ۷/۴ | ۱۲ | ۲ | ۷/۴ |
| | ۳ | ۲ | ۷/۴ | ۱۳ | ۱ | ۳/۷ |
| | ۴ | ۱ | ۳/۷ | ۱۴ | ۲ | ۷/۴ |
| | ۵ | ۱ | ۳/۷ | ۱۵ | ۲ | ۷/۴ |
| | ۶ | ۳ | ۱۱/۱ | ۱۶ | ۱ | ۳/۷ |
| | ۷ | ۰ | ۰ | ۱۷ | ۰ | ۰ |
| | ۸ | ۱ | ۳/۷ | ۱۸ | ۳ | ۳/۷ |
| | ۹ | ۱ | ۳/۷ | ۱۹ | ۱ | ۳/۷ |
| | ۱۰ | ۱ | ۳/۷ | ۲۰ | ۲ | ۷/۴ |
| ياهو | ۱ | ۰ | ۰ | ۱۱ | ۱ | ۳/۴ |
| | ۲ | ۲ | ۶/۹ | ۱۲ | ۱ | ۳/۴ |
| | ۳ | ۲ | ۶/۹ | ۱۳ | ۱ | ۳/۴ |
| | ۴ | ۳ | ۱۰/۳ | ۱۴ | ۱ | ۳/۴ |
| | ۵ | ۲ | ۶/۹ | ۱۵ | ۴ | ۱۳/۸ |
| | ۶ | ۲ | ۶/۹ | ۱۶ | ۰ | ۰ |
| | ۷ | ۰ | ۰ | ۱۷ | ۱ | ۳/۴ |
| | ۸ | ۳ | ۱۰/۳ | ۱۸ | ۱ | ۳/۴ |
| | ۹ | ۰ | ۰ | ۱۹ | ۲ | ۶/۹ |
| | ۱۰ | ۰ | ۰ | ۲۰ | ۳ | ۱۰/۳ |

جدول ۴، توزیع فراوانی رتبه‌های نتایج نامرتب را در موتورهای کاوش گوگل، بینگ، و یاهو پس از محدود کردن به زبان عربی نشان می‌دهد. در گوگل، رتبه ۲۰ بیشترین فراوانی نتایج نامرتب را به خود اختصاص داد که تقریباً ۱۲ درصد مجموع فراوانی رتبه‌های نامرتب را شامل می‌شود. پس از آن، رتبه ۱۲ بیشترین فراوانی است که حدود ۹ درصد رتبه‌های نامرتب را تشکیل می‌دهد. همچنین، رتبه‌های ۲، ۳، و ۴ کمترین فراوانی نتایج نامرتب، یعنی حدود ۳ درصد رتبه‌های نامرتب را به خود اختصاص داده است. لازم به اشاره است، تمامی رتبه‌ها حداقل یک‌بار نتایج نامرتب داشته‌اند و رتبه ۱ به‌طور خاص، در هر دو پرس‌وجو، نتیجه نامرتب در برداشت. در موتور کاوش بینگ، رتبه ۶ بیشترین فراوانی و تقریباً ۱۱ درصد رتبه‌های نامرتب بود. کمترین فراوانی به رتبه ۱، ۷، و ۱۷ با صفر درصد از مجموع فراوانی رتبه‌های نامرتب اختصاص داشت.

همچنین، در موتور کاوش یاهو، بیشترین فراوانی به رتبه ۱۵ (با حدود ۱۴ درصد مجموع فراوانی رتبه‌های نامرتب) اختصاص داشت. رتبه‌های ۱، ۷، ۹، ۱۰، و ۱۶ کمترین فراوانی (صفر درصد) را از مجموع رتبه‌های نامرتب به خود اختصاص دادند.

آزمون معناداری میان موتورهای کاوش به‌لحاظ میانگین نتایج نامرتب با زبان هدف برای مقایسه میانگین نتایج نامرتب با زبان هدف در موتورهای کاوش مختلف از آزمون تحلیل واریانس استفاده شد (جدول ۵). همان‌گونه که مشاهده می‌شود پس از محدودیت زبان به فارسی و عربی، بین سه موتور کاوش گوگل، بینگ، و یاهو به‌لحاظ میانگین نتایج نامرتب با زبان هدف، اختلاف معناداری وجود نداشت.

جدول ۵. تحلیل واریانس برای سنجش معناداری اختلاف نتایج نامرتب با زبان هدف در موتورهای کاوش

| زبان هدف | مجموع مجذورات | درجه آزادی | میانگین مربعات | مقدار F | سطح معناداری |
|----------|---------------|------------|----------------|---------|--------------|
| فارسی | ۳/۴۴ | ۲ | ۱/۷۲ | ۲/۷۷ | ۰/۰۶۷ |
| | ۶۸/۸۴ | ۱۱۱ | ۰/۶۲ | | |
| | ۷۲/۲۸ | ۱۱۳ | | | |
| عربی | ۲۸/۱۲ | ۲ | ۱۴/۰۶ | ۲/۴۵ | ۰/۰۹۱ |
| | ۶۳۷/۰۰ | ۱۱۱ | ۵/۷۴ | | |
| | ۶۶۵/۱۲ | ۱۱۳ | | | |

آزمون معناداری میان موتورهای کاوش به لحاظ میانگین رتبه‌های نتایج نامرتبط با زبان هدف برای مقایسه رتبه‌های نتایج نامرتبط با زبان هدف در موتورهای کاوش گوگل، بینگ، و یاهو از آزمون کروسکال والیس استفاده شد. پس از محدود کردن به زبان فارسی، نتایج آزمون نشان از معناداری اختلاف میان رتبه‌های نتایج نامرتبط در سه موتور کاوش مختلف در سطح اطمینان ۹۹ درصد دارد ($\chi^2=16/71$ ، $P=0/000$) (جدول ۶). برای شناسایی موتور کاوشی که اختلاف معنادار را ایجاد کرده است، از آزمون یومان-ویتنی استفاده شد. نتایج نشان داد که اختلاف رتبه‌های نتایج نامرتبط میان سه موتور کاوش، به گوگل بازمی‌گردد، به نحوی که میانگین رتبه‌های نتایج نامرتبط در گوگل به طور معناداری نسبت به بینگ و یاهو بیشتر است. گرچه دو موتور کاوش بینگ و یاهو به لحاظ میانگین رتبه‌های نتایج نامرتبط با یکدیگر اختلاف معناداری نداشتند (جدول ۷).

جدول ۶. آزمون کروسکال-والیس برای مقایسه رتبه‌بندی نتایج نامرتبط با زبان هدف در موتورهای کاوش مختلف

| زبان هدف | خی دو | درجه آزادی | سطح معناداری |
|----------|-------|------------|--------------|
| فارسی | ۱۶/۷۱ | ۲ | ۰/۰۰۰ |
| عربی | ۹/۳۶ | ۲ | ۰/۰۰۹ |

پس از محدود کردن به زبان عربی، مشخص شد که بین سه موتور کاوش از نظر میانگین رتبه‌های نتایج نامرتبط اختلاف معناداری در سطح اطمینان ۹۹ درصد وجود داشت ($\chi^2=36/9$ ، $P=0/009$) (جدول ۶). برای شناسایی موتور کاوشی که این اختلاف را ایجاد کرده است، از آزمون یومان-ویتنی استفاده شد. نتایج نشان داد که اختلاف در میانگین رتبه‌های نتایج نامرتبط به موتور کاوش گوگل بازمی‌گردد، به نحوی که میانگین رتبه‌های نتایج نامرتبط در موتور کاوش گوگل به طور معناداری از دو موتور کاوش دیگر بیشتر بود. اختلاف میان میانگین رتبه‌های نتایج نامرتبط دو موتور کاوش بینگ و یاهو معنادار نبود (جدول ۷).

جدول ۷. موتورهای کاوش دارای رتبه‌بندی نتایج نامرتب متفاوت معنادار در آزمون یومان- ویتنی

| زبان هدف | موتور کاوش الف | موتور کاوش ب | مقدار یومان - ویتنی | مقدار Z | سطح معناداری |
|----------|----------------|--------------|---------------------|---------|--------------|
| فارسی | گوگل | بینگ | ۲۳۳۷/۵۰ | -۲/۹۱ | ۰/۰۰۴ |
| | | ياهو | ۲۳۸۰/۰۰ | -۲/۹۴ | ۰/۰۰۳ |
| عربی | گوگل | بینگ | ۲۰۶۰/۵۰ | -۲/۵۷ | ۰/۰۱۰ |
| | | ياهو | ۲۱۳۸/۵۰ | -۲/۴۳ | ۰/۰۱۵ |

تأثیر ترافیک صفحات نامرتب با زبان هدف بر رتبه‌بندی آنها

در این راستا، به منظور دریافت ترافیک هر صفحه به طور مجزا، از رتبه جهانی ترافیک صفحات توسط وبسایت الکسا استفاده شد. برای بررسی رابطه میان رتبه ترافیک صفحات نامرتب و رتبه آنان پس از محدود کردن به زبان فارسی و عربی از آزمون همبستگی اسپیرمن استفاده شد. با توجه به جدول ۸، رابطه معناداری میان رتبه ترافیک یا بازدید صفحات نامرتب و رتبه آنان در موتور کاوش گوگل پس از محدودیت زبان به فارسی وجود نداشت. همچنین، نتایج نشان از نبود رابطه معناداری میان رتبه ترافیک یا بازدید صفحات نامرتب و رتبه آنان در سه موتور کاوش پس از محدودیت زبان به عربی است.^۱

جدول ۸. نتایج آزمون همبستگی اسپیرمن میان رتبه ترافیک صفحات نامرتب و رتبه آنها در موتورهای کاوش

| زبان هدف | موتور کاوش | ضریب همبستگی | تعداد نتایج نامرتب | سطح معناداری |
|----------|------------|--------------|--------------------|--------------|
| فارسی | گوگل | ۰/۳۴ | ۱۳ | ۰/۲۵ |
| عربی | گوگل | ۰/۱۶ | ۶۰ | ۰/۲۳ |
| | بینگ | -۰/۱۲ | ۲۳ | ۰/۵۹ |
| | ياهو | -۰/۱۷ | ۲۰ | ۰/۴۷ |

۱. در وبسایت الکسا، رتبه ترافیک برخی صفحات نامرتب به دلیل فیلتر بودن صفحات دریافت نشد و این موارد از محاسبه حذف شد. در محدودیت زبان به فارسی، یک مورد از نتایج نامرتب در موتور کاوش گوگل، و در محدودیت زبان به عربی ۸ مورد نتایج نامرتب در موتور کاوش گوگل، ۴ مورد نتایج نامرتب در موتور کاوش بینگ، و ۹ مورد نتایج نامرتب در موتور کاوش یاهو در محاسبه لحاظ نشد.

نتیجه‌گیری

با جستجوی ۳۸ هم‌نویسه در زبان فارسی و عربی در سه موتور کاوش گوگل، بینگ، و یاهو و پس از محدود کردن زبان فارسی فقط دو موتور بینگ و یاهو عملکرد صددرصد موفق‌تری ارائه دادند. موتور کاوش گوگل به لحاظ بازیابی تعداد نتایج مرتبط

با زبان هدف، عملکرد ضعیف‌تری نسبت به موتور کاوش بینگ و یاهو داشت. پس از محدود کردن پرس‌وجوها به زبان عربی هیچ‌یک از موتورهای کاوش عملکرد صددرصد موفق‌تری ارائه ندادند، اما باز هم موتور کاوش بینگ و یاهو به ترتیب عملکرد بهتری نسبت به موتور کاوش گوگل داشتند. بنابراین، دو موتور کاوش بینگ و یاهو می‌توانند نتایج دقیق‌تری نسبت به موتور کاوش گوگل در زبان عربی به کاربران ارائه دهند. این یافته همسو با نتایج لواندوسکی (۲۰۰۸) است که نشان داد موتور کاوش گوگل نسبت به موتورهای کاوش اسک و یاهو عملکرد ضعیف‌تری به لحاظ بازیابی اطلاعات به زبان دیگر دارد.

از آنجا که انتظار می‌رود پس از محدودیت زبان، تمامی نتایج به زبان هدف بازیابی شود، رتبه‌بندی نتایج نامرتب نیز از اهمیت بسیاری برخوردار است. در واقع، انتظار می‌رود نتایج نامرتب در رتبه‌های اول ظاهر نشود، اما نتایج نشان داد که پس از محدود کردن زبان فارسی، بیشترین نتایج نامرتب در رتبه ۸ موتور کاوش گوگل ظاهر شد و در رتبه‌های ۱ تا ۵ در هیچ‌یک از پرس‌وجوها نتایج نامرتب ظاهر نشد. از سوی دیگر، در محدوده زبان عربی، موتور کاوش گوگل در تمامی رتبه‌ها حتی رتبه‌های ۱ تا ۵ حداقل یک‌بار نتایج نامرتب نشان داد. در موتور کاوش بینگ و یاهو رتبه ۱، در هیچ‌یک از پرس‌وجوها نتایج نامرتب نشان نداد، اما بیشتر رتبه‌ها حداقل یک‌بار نتایج نامرتب داشتند. به نظر می‌رسد که علت عملکرد بهتر موتورهای کاوش در زبان فارسی نسبت به زبان عربی نیز این است که جستجو از آی‌پی ایران صورت گرفته است، بنابراین نتایج زبان فارسی در رتبه‌های بهتری نسبت به دیگر زبان‌ها قرار می‌گیرند.

همچنین، سه موتور کاوش در بازیابی نتایج نامرتب با زبان هدف تفاوت معناداری با یکدیگر نداشتند. در واقع، هیچ‌یک از این سه موتور به لحاظ بازیابی نتایج مرتب با زبان هدف بر یکدیگر برتری ندارند. اما این سه موتور کاوش در رتبه‌بندی نتایج نامرتب با زبان هدف تفاوت معناداری با یکدیگر نشان دادند و این تفاوت نیز به موتور کاوش گوگل بازمی‌گردد. به بیان دیگر، گوگل با بازیابی نتایج نامرتب در رتبه‌های بهتر، کارآمدی کمتری نسبت به دیگر رقبای خود نشان داد. یافته‌های این پژوهش با نتایج پژوهش ژانگ، فی، و لی^۱ (۲۰۱۳) و لواندوسکی (۲۰۰۴) به لحاظ برتری موتور کاوش گوگل در زمینه عملکرد ویژگی‌های جستجو هم‌راستا نیست.

از سوی دیگر، با توجه به اینکه میان رتبه ترافیک و رتبه صفحات نامرتب در موتورهای کاوش مختلف ارتباط معناداری مشاهده نشد، به نظر می‌رسد که عملکرد

1. Zhang, Fei, & Le

موتورهای کاوش در بازیابی نتایج نامرتبط با زبان هدف به ترافیک و یا محبوبیت صفحات نامرتبط بازنمی‌گردد و به عواملی دیگر مانند ضعف ابزارهای جستجو و الگوریتم‌های تشخیص زبان در موتورهای کاوش مربوط می‌شود.

به نظر می‌رسد موتور کاوش گوگل از تشخیص زبان ایستا^۱ استفاده نکرده و به جای آن از تشخیص زبان درجه‌ای^۲ استفاده می‌کند. در تشخیص زبان ایستا، یک مدرک به یک زبان خاص تعلق می‌گیرد؛ اما، در تشخیص زبان درجه‌ای، مدرکی که شامل زبان‌های مختلف است به بیش از یک زبان اما با درجه‌های مختلف اختصاص می‌یابد. از سوی دیگر، زبان‌ها به برخی صفحات نمی‌توانند اختصاص یابند، چون هم زبان فارسی و هم زبان عربی دارند و یا اینکه گزینه انتخاب زبان در صفحه اول قرار گرفته است. همچنین، اگر تعداد کلمات موجود در صفحه بسیار اندک باشد (به‌طور مثال، فقط نام، شماره تماس، و نشانی) امکان شناسایی زبان مهیا نمی‌شود (لواندوسکی، ۲۰۰۸).

به‌طور کلی، پژوهش حاضر نشان داد کاربرانی که از فیلد زبان برای محدود کردن نتایج جستجو در موتورهای کاوش استفاده می‌کنند، در بینگ و یاهو از اقبال بالاتری برای دستیابی به نتایج مرتبط برخوردارند.

مآخذ

دری، راحله (۱۳۹۳). مقایسه و ارزیابی موتورهای جستجوی معنایی. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۰ (۲)، ۴۶۷-۴۸۷.

رجبی، منصور؛ نروزی، یعقوب (۱۳۹۴). موتورهای جستجوی فارسی: ارزیابی امکانات جستجو، بازیابی اطلاعات، میزان جامعیت و مانعیت و تعیین همپوشانی میان آنها. *مطالعات ملی کتابداری و سازماندهی اطلاعات*، ۲۶ (۳)، ۱۳۳-۱۵۰.

شریف، عاطفه (۱۳۹۱). بررسی تغییرات عملکرد دو موتور کاوش عمومی یاهو و گوگل از نظر پوشش کمی-زمانی نمایه‌سازی و توجه به عناصر ابر داده‌ای در رتبه‌بندی صفحه‌های وب. *پژوهش‌نامه کتابداری و اطلاع‌رسانی*، ۲ (۱)، ۱۷۵-۱۹۴.

محمداسمعیل، صدیقه؛ منصور کیایی، ربابه (۱۳۹۰). مقایسه موتورها و ابرموتورهای کاوش عمومی در بازیابی اطلاعات علم فیزیک و میزان همپوشانی آنها. *مطالعات ملی کتابداری و سازماندهی اطلاعات*، ۲۲ (۳)، ۱۳۰-۱۴۰.

1. Static-language detection
2. Graded-language detection

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. Paper presented at *the third Annual Symposium on Document Analysis and*

Information Retrieval, Las Vegas, Nevada, USA, April 11-13. Available July 5, 2015, from https://s3.amazonaws.com/academia.edu.documents/6397498/10.1.1.21.3248.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1502139164&Signature=qGsY8DhiWbcpjMZsgZmhs3ygaHs%3D&response-content-disposition=inline%3B%20filename%3DN-gram-based_text_categorization.pdf

Dunning, T. (1994). *Statistical identification of language, Technical Report* (MCCS 94-273). New Mexico State University: New Mexico.

Kaushik, A. (2012). Judging the capability of search engines and search terms. *International Journal of Information Dissemination and Technology*, 2 (1), 6-11.

Fattahi, R., Wilson, C. S., & Cole, F. (2008). An alternative approach to natural language query expansion in search engines: Text analysis of non-topical terms in web documents. *Information Processing & Management*, 44 (4), 1503-1516.

Ghani, R., Jones, R., & Mladenic, D. (2005). Building minority language corpora by learning to generate web search queries. *Knowledge and Information Systems*, 7 (1), 56-83.

Kim, J. Y., Feild, H., & Cartright, M. (2012, October). Understanding book search behavior on the web. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 744-753). ACM. October 29 - November 02.

Lewandowski, D. (2004). Date-restricted queries in web search engines. *Online Information Review*, 28 (6), 420-427.

Lewandowski, D. (2008). Problems with the use of web search engines to find results in foreign languages. *Online Information Review*, 32 (5), 668 - 672.

Lewandowski, D., & Höchstötter, N. (2008). *Web searching: a quality measurement perspective* (pp. 309-340). Springer Berlin Heidelberg. Available July 5, 2015, from https://link.springer.com/chapter/10.1007/978-3-540-75829-7_16

Martins, B., & Silva, M. J. (2005, March). Language identification in web pages. In *Proceedings of the 2005 ACM symposium on Applied computing* (pp. 764-768). ACM. March 13 – 17. Available July 5, 2015, from <http://dl.acm.org/citation.cfm?id=1066852>

Sroka, M. (2000). Web search engines for polish information retrieval: Questions of search capabilities and retrieval performance. *International Information & Library Research*, 32 (2), 87-98.

Zhang, J., & Lin, S. (2007). Multiple language supports in search engines. *Online Information Review*, 31 (4), 516-532.

Zhang, J., Fei, W., & Le, T. (2013). A comparative analysis of the search feature effectiveness of the major English and Chinese search engines. *Online Information Review*, 37 (2), 217-230.

استناد به این مقاله:

یقظین، مریم؛ جوکار، عبدالرسول (۱۳۹۶). کارآمدی فیلد "زبان" در موتورهای کاوش بین‌المللی برای بازیابی نتایج مرتبط به زبان فارسی و عربی (مطالعه مقایسه‌ای: گوگل، بینگ، و یاهو). *مطالعات ملی کتابداری و سازماندهی اطلاعات*، ۲۸ (۲)، ۱۴۱-۱۷۵.