الگوهای تعامل و راهبردهای بازآرایی پرس‌وجو توسط کاربران در یک موتور جستجوی فارسی

فاطمه کاوه یزدی[1]
دانشجوی دکتری مهندسی کامپیوتر دانشگاه یزد
علی محمد زارع بیدکی[2]
دانشگاه یزد
محمدرضا پژوهان[3]
دانشگاه یزد

تاریخ دریافت: ۱۳۹۵/۰۹/۱۹ تاریخ پذیرش: ۱۳۹۷/۰۳/۲۲

چکیده

فرایند جستجو در وب با زندگیِ برخط امروزی عجین شده است. موتورهای جستجو، با دریافت پرس‌وجوهای کاربران، تعداد محدودی از اسناد مرتبط را از میان چندین میلیارد صفحهٔ وب بازیابی می‌کنند. بنابراین موتورهای جستجو با ثبت مجموعهٔ پرس‌وجوهای کاربران در درازمدت می‌توانند به مجموعه‌ای از اطلاعات دربارهٔ الگوهای رفتاری کاربران دست یابند. این الگوها می‌توانند در فرایندهایی مانند گسترش پرس‌وجو، پیشنهاد پرس‌وجو و تصحیح املایی مورد استفاده قرار گیرند. این مقاله به بررسی الگوهای تعامل کاربران و بازآرایی پرس‌وجوهای ارسالی آنها به موتور جستجوی فارسی پارسی‌جو می‌پردازد و به‌طور اخص، پرس‌وجوهای کاربران از نظر واژگانی و زمانی و همچنین الگوهای بازآرایی را مورد بررسی قرار می‌دهد. توزیع آماری کلمات پرس‌جو، تغییرات طول پرس‌وجو در زمان‌های مختلف از شبانه‌روز و رفتارهای بازآرایی پرس‌وجو توسط کاربران از مهم‌ترین بررسی‌های آماری در این مقاله بوده‌اند. در ادامه تحلیل‌هایی دربارهٔ الگوهای تعامل کاربران مبتنی بر نتایج بررسی‌های مذکور ارائه شده و نتایج آن با بررسی‌های بین‌المللی مقایسه شده است. این مطالعه نشان‌دهنده همخوانی رفتاری کاربران فارسی زبان با کاربران موتورهای جستجو در سراسر جهان است.

کلیدواژه‌ها: موتور جستجوی فارسی، پرس‌وجو، بازآرایی، کلیک، توزیع قانون توان.

1. fkavehy@parsijoo.ir
2. alizareh@yazd.ac.ir
3. pajoohan@yazd.ac.ir

۱- مقدمه

موتورهای جستجو از مهم‌ترین ابزارها در دنیای امروز هستند که کاربران برای استفاده از این خدمات باید پرس‌وجوهای[1] خود را به آنها ارسال کنند؛ سپس موتور جستجو با استفاده از اسناد موجود در نمایه[2] اقدام به بازیابی اسناد مرتبط می‌کند. بدیهی است که بازیابی اسناد از وب و نمایش آنها برای کاربرانِ درخواست‌کننده بسیار پیچیده است؛ زیرا در بسیاری از موارد کاربران به‌دلیلِ نداشتنِ اطلاع دقیق از محتوای صفحات مورد نظر خود، ممکن است از عباراتی استفاده کنند که در محتوای صفحات دیده نشده باشند. در این حالت ممکن است کاربر نتیجهٔ مطلوبی دریافت نکند. از این مشکل با عنوان عدم هم‌خوانی عبارتی[3] یاد می‌شود و در زمرهٔ یکی از مشکلات جدی موتورهای جستجو در کسب رضایت کاربران است.

این موضوع را می‌توان از دو منظر مورد توجه قرار داد؛ (یک) از دید کاربران و (دو) از دید موتورهای جستجو. از دید کاربران، راهکار حل این مشکل می‌تواند بازآرایی پرس‌وجو[4] و تغییر پرس‌وجو باشد. از دید موتورهای جستجو نیز اعمال تغییراتی در پرس‌وجو تحت عنوان گسترش پرس‌وجو[5] می‌تواند همان اثر را داشته باشد. بررسی الگوهای تعامل کاربران در قالب بازآرایی پرس‌وجو می‌تواند به‌عنوان نشانه‌ای از میزان رضایت/عدم‌رضایت کاربران از کیفیت نتایج جستجو قلمداد شود. از سوی دیگر، بررسی آماری بازآراییِ صورت‌گرفته از سوی کاربران که منجر به حصول نتیجهٔ دلخواه کاربران و کلیک بر روی نتیجهٔ مذکور شده است، می‌تواند راهگشای موتورهای جستجو در فرایند گسترش پرس‌وجو باشد.

موتورهای جستجوی فارسی‌زبان در سال‌های اخیر توسعه یافته و مورد توجه قرار گرفته‌اند. کمبود داده و نبودِ تحقیقات سابقه‌دار بر روی دادگان این موتورها از یک سو و نیاز به استفاده از نتایج الگوهای بازآرایی توسط کاربران از سوی دیگر، ما را بر آن داشته است تا پژوهشی در این حوزه به انجام برسانیم و نتایج آن را منتشر کنیم.

۱-۲ بازآرایی پرس‌وجو

پیش از پرداختن به مسئلهٔ بازآرایی پرس‌وجو، به معرفی اجمالی مفاهیم اولیه در این حوزه خواهیم پرداخت و در ادامه، تعاملات کاربران فارسی زبان با یک موتور جستجوی فارسی را ازنظر آماری بررسی خواهیم کرد.

1. query
2. index
3. term mismatch
4. query reformulation
5. query expansion

فرایند جستجو یک فرایند کاربر-آغاز[1] است، که در آن کاربر با ارسال یک پرس‌وجو به موتور جستجو، فرایند جستجو را آغاز می‌کند. در این صورت برای تعریف یک تراکنش جستجو[2] باید گفت، اقدام کاربری مانند u که پرس‌وجوی q را در زمان t ارسال می‌کند، یک تراکنش جستجو را به‌وجود می‌آورد. به بیان بهتر، هر تراکنش به‌صورت یک سه‌تایی مرتّب[3] به فرم $\langle u,q,t \rangle$ قابل تعریف است. هر پرس‌وجوی q دارای یک یا چند عبارت (Term) است و به‌صورت مندرج در رابطه (۱) تعریف می‌شود.

$$q = \langle T_1 T_2 ... T_n \rangle \quad (۱)$$

در صورتی که کاربر u پس از ارسال پرس‌وجوی q، آن را تغییر دهد و نسخهٔ تغییریافتهٔ آن، یعنی q'، را در زمان $t+\Delta t$ ارسال کند، به پرس‌وجوی q' نسخهٔ بازآرایی‌شدهٔ q[4] گفته می‌شود و به این دو پرس‌جو، جفت پرس‌وجوی بازآرایی‌شده[5] اطلاق می‌شود (بیتزل[6] و همکاران، ۲۰۰۴؛ هوآنگ[7] و افثیمیادیس[8]، ۲۰۰۹؛ ژیانگ[9] و نی[10]، ۲۰۱۶). استفادهٔ عملی از این تعریف نیازمند ارائهٔ جزئیاتی معین برای هر بخش آن است که در ادامه به آنها پرداخته خواهد شد.

از آنجا که کاربران انتخاب‌های زیادی برای اعمال تغییر در پرس‌وجوی خود دارند، در این تعریف تنها به استفاده از واژهٔ تغییر بسنده شده و در عوض در هریک از تحقیقات مورد نظر، براساس نوع کاربردِ مورد نیاز، فهرستی از تغییراتِ قابل قبول ارائه شده است. الگوهای بازآرایی در تحقیقات متنوعی مانند (گوئو[11] و همکاران، ۲۰۰۸؛ هوآنگ و افثیمیادیس، ۲۰۰۹؛ و همکاران، ۲۰۰۷؛ تیوان[12] و همکاران، ۲۰۰۷؛ وایتل[13] و همکاران، ۲۰۰۷) مورد توجه قرار گرفته‌اند و یکی از کامل‌ترین فهرست‌ها در این میان توسط هوآنگ و افثیمیادیس (هوآنگ و افثیمیادیس، ۲۰۰۹) معرفی شده است. این فهرست شامل تغییر در ترتیب کلمات، تغییرات فاصله‌گذاری، حذف کلمه، افزودن کلمه، کوتاه‌کردنِ URLها، ریشه‌یابی، سرواژه‌سازی، زیررشته‌سازی، مخفف‌سازی، تصحیح املایی و انواعی از تغییرات جزئی‌تر است که در هیچ‌یک از دسته‌بندی‌های بالا قرار نمی‌گیرند. به‌علاوه، مواردی مانند عمومی‌سازی[14]، اختصاصی‌سازی[15]، تعیین توصیفگر[16] و مغایرت دستوری نیز در سایر تحقیقات مشابه ازقبیل

1. user-initiative
2. search interaction
3. ordered triple
4. reformulated query
5. reformulated query pair
6. S. M. Beitzel
7. J. Huang
8. E. N. Efthimiadis
9. J. Jiang
10. C. Ni
11. J. Guo
12. J. Teevan
13. M. Whittle
14. generalization
15. specialization
16. modifier

(جنسن و همکاران، ۲۰۰۷؛ انیک[1]، ۲۰۰۳) مورد ارجاع قرار گرفته‌اند. موارد ذکرشده را می‌توان براساس نوع تحلیل‌های مورد استفاده به دو دسته تقسیم کرد؛ دسته اوّل، مواردی مانند تغییر در ترتیب کلمات، تغییرات فاصله‌گذاری، حذف کلمه و افزودن کلمه که بدون استفاده از تحلیل‌های نحوی و معنایی نیز قابل تشخیص هستند و دسته دوم مواردی مانند تشخیص عمومی‌سازی، اختصاصی‌سازی، تعیین توصیفگر و مغایرت دستوری که تنها با استفاده از منابع اطلاعات بیرونی و تحلیل‌های نحوی و معنایی قابل تشخیص هستند.

نخستین گام پس از تعیین تغییرات مورد نظر در تشخیص جفت‌های بازآرایی‌شده، تخصیص پرس‌وجوها به کاربران و تعیین این حقیقت است که از دو پرس‌وجوی دریافتی توسط موتور جستجو در زمان‌های t و $t+\Delta t$ آیا هر دو پرس‌وجو متعلق به یک کاربر بوده‌اند یا خیر؟

ساده‌ترین راهکار برای تحقق این هدف معمولاً از تعریف نشست[2] ارائه‌شده توسط مرورگر بهره می‌گیرد و در عین حال طول نشست را چنان محدود درنظر می‌گیرد که بتوان فرض کرد در این بازهٔ زمانی نظرات کاربر تغییرات اساسی نخواهند کرد. راهکار دیگر، تعریف نشست براساس وابستگی موضوعی پرس‌وجوهاست که معمولاً خالی از خطا نیست و نیازمند استفاده از تحلیل‌های پیچیدهٔ زبانی و مفهومی است.

راهکار اوّل در زمان اجرا با دو سیاست متداول در این عرصه مورد استفاده قرار می‌گیرد. سیاست اوّل با تعریف بیشینهٔ طول بازهٔ زمانی فاقد فعّالیّت، مرز بین نشست‌ها را تعیین می‌کند که با استفاده از این تعریف می‌توان نشست‌های با طول متفاوت داشت و از مرتبط‌بودن پرس‌وجوها اطمینان حاصل کرد. سیاست دوم استفاده از نشست‌های با طول معین و ثابت است. در این حالت نشست‌ها دارای حداکثر طول ثابت هستند اما ممکن است پرس‌وجوی دوم از یک جفت به‌دلیلِ اتمام طول عمر نشست به ابتدای نشست بعدی انتقال یابد و در پردازش‌ها لحاظ نشود.

براساس نتایج تحقیقاتی که در پژوهش‌هایی مانند (حسن، ۲۰۱۳؛ پارک[3] و همکاران، ۲۰۱۵؛ اسلُن[4]، ینگ[5] و ونگ[6]، ۲۰۱۵؛ جیانگ و همکاران، ۲۰۱۴) منتشر شده است، بهترین راهکار استفاده از محدودیت بیشینه بازهٔ زمانی عدم فعالیت با طول ۳۰ دقیقه است. البته روش‌های دیگری، مانند محدودکردن تعداد جفت بازآرایی‌های متوالی، نیز در پژوهش‌ها مورد

1. P. Anick
2. Session
3. J. Y. Park
4. M. Sloan
5. H. Yang
6. J. Wang

استفاده قرار گرفته‌اند که به اندازهٔ روش بالا متداول نیستند و در عین حال ممکن است برخی از جفت‌های بازآرایی را، به‌دلیل انتخاب حدنصاب نامناسب، حذف کنند.

۳- مروری بر پژوهش‌ها در حوزهٔ بازآراییِ پرس‌وجو

یکی از مهم‌ترین روش‌ها در بررسی الگوهای رفتاری و تعامل کاربرانِ موتورهای جستجو، با بررسی فایل‌های تاریخچهٔ جستجو[1] صورت می‌پذیرد. در این روش، اطلاعاتِ جمع‌آوری‌شده توسط موتورهای جستجو از موارد متعدد پرس‌وجوهای ارسالی توسط کاربران مورد ارزیابی قرار می‌گیرد. در این مقاله، این تحقیقاتِ مبتنی بر اطلاعات تاریخچه‌ای را، براساس اینکه به دادگان پرس‌وجو و یا کلیک معطوف شده باشند، به دو دسته تقسیم می‌کنیم. در دستهٔ اوّل تحقیقاتی جای می‌گیرند که با بررسی پرس‌وجوها به‌تنهایی و بدون لحاظ کلیک به مطالعه رفتار کاربران پرداخته‌اند. یکی از باسابقه‌ترین تحقیقات در این زمینه توسط جَنسِن[2] و همکاران (۱۹۹۸) صورت گرفته است. در ادامهٔ این تحقیق، افراد و گروه‌های مختلفی برای بررسی پرس‌وجوهای کاربران اقدام کرده‌اند که از آن‌جمله می‌توان به هوآنگ و افثیمیادیس (۲۰۰۹) اشاره کرد. این تحقیق، غالب تحقیقات پیش از خود را به‌صورت کلی مورد بررسی قرار داده و مجموعه‌ای کامل از دسته‌بندی‌ها را برای بازآرایی‌های ممکن توسط کاربران ارائه کرده و سپس آمار مربوط به میزان کلیک کاربران را برای تخمین میزان اثربخشی آنها مورد استفاده قرار داده است.

تحقیقات با استفاده از پرس‌وجوها، در لایه‌ای با جزئیات بیشتر بر روی عبارت‌های تشکیل‌دهندهٔ پرس‌وجوها نیز صورت گرفته است و محققان مواردی مانند تغییرات طول پرس‌وجوها و توزیع عبارت‌های پرس‌وجوها را مورد بررسی قرار داده‌اند. به‌صورت نمونه می‌توان به نتایج تحقیقات جالب مانند (ایرون[3] و مک‌کرلی[4]، ۲۰۰۳؛ لِمپل[5] و مُران[6]، ۲۰۰۳؛ ونگ[7]، بِری[8] و یَنگ[9]، ۲۰۰۳) اشاره کرد که شواهد محکمی در تبعیت توزیع پرس‌وجوها و عبارات آنها از توزیع مبتنی بر قانون ارائه کرده‌اند و حتی سارایوا[10] و همکاران (۲۰۰۱) صحت وجود این پدیده را در یک موتور جستجوی بومی در برزیل نیز بررسی کرده و به نتایج مشابه رسیده‌اند.

در کنار تحقیقات مبتنی بر پرس‌وجوها، مجموعهٔ دیگری از پژوهش‌ها وجود دارند که از اطلاعات کلیک پرس‌وجوها نیز بهره می‌گیرند. از این دسته می‌توان به مواردی مانند (حسن و

---

1. log files
2. B. J. Jansen
3. N. Eiron
4. K. S. McCurley
5. R. Lempel
6. S. Moran
7. P. Wang
8. M. W. Berry
9. Y. Yang
10. P. C. Saraiva

همکاران، ۲۰۱۳؛ جیانگ، هی[1] و آلن[2]، ۲۰۱۴؛ وِبر[3] و جیمز[4]، ۲۰۱۱) اشاره کرد. براساس شواهد ارائه‌شده در حسن و همکاران (۲۰۱۳) و جیانگ و همکاران (۲۰۱۵) کلیک کاربران می‌تواند یکی از علائم بسیار مهم در تعیین میزان رضایتمندی کاربران از نتایج جستجو باشد. از طرف دیگر می‌توان از کلیک‌ها بر روی نتایج بازیابی‌شده در پاسخ به پرس‌وجوها به‌عنوان یک منبع اطلاعاتی برای بررسی بازآرایی پرس‌وجوها بهره گرفت.

در کنار تحقیقات پیشین مواردی نیز مانند کیزلوا[5] و همکاران (۲۰۱۴) وجود دارند که بر وجوه دیگری از این تعامل تمرکز کرده‌اند و رضایتمندی کاربران از نتایج پرس‌وجوهایی که پاسخ آنها با گذشت زمان تغییر می‌کند را با استفاده از داده‌های تاریخچهٔ جستجوی کاربران بررسی کرده‌اند. مجموعهٔ یافته‌های این تحقیق نشان می‌دهد می‌توان با بررسی واکنش کاربران نسبت به چنین پرس‌وجوهایی، تغییر پاسخ را پیش‌بینی کرد و از پرس‌وجوهای جدید به‌صورت خودکار در سرویس پیشنهاد پرس‌وجوی فوری[6] بهره گرفت. پژوهش دیگری که توسط بیتزِل و همکاران (۲۰۰۴) به‌اجرا درآمده است، توزیع پرس‌وجوها و موضوعات مختلف را در مقاطع زمانی مختلف در شبانه‌روز بررسی کرده است. نتایج این تحقیق، نشان‌دهندهٔ جهش جستجو در برخی از موضوعات در بازه‌های زمانی خاصی از شبانه‌روز است. به‌عنوان مثال، جستجوهای کاربران در حوزه‌های موسیقی و سرگرمی در ساعات پایانی روز و بامداد بیش از سایر زمان‌ها در شبانه‌روز است.

۴- بررسی آماری الگوهای تعامل کاربران

با توجه به توضیحات مندرج در قسمت‌های قبل، انواع تحلیل‌های ممکن بر روی دادگان پرس‌وجوی استخراج‌شده از موتور جستجوی فارسی، با ترکیب انواع پارامترها و دادگانِ دردسترس، به‌اجرا درآمده‌اند. با توجه به تنوع تحلیل‌ها، در نخستین گام به ارائهٔ طرح اجمالی تحلیل‌ها خواهیم پرداخت. تحلیل‌ها در این مقاله در دو سطح عبارت و پرس‌وجو به‌اجرا درآمده و پارامترهایی مانند تغییرات واژگانی، زمان ارسال پرس‌وجوها و نوع الگوهای بازآرایی پرس‌وجوها و کلیک پرس‌وجوها در بررسی‌ها لحاظ شده‌اند. جداول سه‌گانه‌ی یک تا سه طرح کلی ترکیب تحلیل‌ها را نشان می‌دهند.

1. D. He
2. J. Allan
3. I. Weber
4. A. Jaimes
5. J. Kiseleva
6. instant query suggestion

جدول ۱- فهرست بررسی‌های واژگانی انجام‌گرفته بر روی پرس‌وجوهای خام و بازآرایی‌شده

مشخصات واژگانی		نوع سطح مورد بررسی	
		سطح پرس‌وجو	سطح عبارت
نوع دادگان	پرس‌وجوی خام	مشخصات واژگانی پرس‌وجوهای خام	مشخصات واژگانی عبارت‌های پرس‌وجوهای خام
	جفت‌های بازآرایی‌شده	مشخصات واژگانی پرس‌وجوهای بازآرایی	مشخصات واژگانی عبارت‌های پرس‌وجوهای بازآرایی

جدول ۲- فهرست بررسی‌های مربوط به تغییرات بازه‌های زمانی بر روی پرس‌وجوهای خام و بازآرایی‌شده

تغییرات زمانی		نوع سطح مورد بررسی	
		سطح پرس‌وجو	سطح عبارت
نوع دادگان	پرس‌وجوی خام	تغییرات زمانی پرس‌وجوهای خام	تغییرات زمانی عبارت‌های پرس‌وجوهای خام
	جفت‌های بازآرایی‌شده	تغییرات زمانی پرس‌وجوهای بازآرایی	تغییرات زمانی عبارت‌های پرس‌وجوهای بازآرایی

جدول ۳- فهرست بررسی‌های تکمیلی بر روی الگوهای بازآرایی

		نوع پارامتر مورد بررسی	
		زمان	کلیک
نوع دادگان	الگوهای بازآرایی	توزیع زمانی الگوهای بازآرایی	تغییرات کلیک در الگوهای بازآرایی

۴-۱- آماده‌سازی دادگان

فاز مطالعاتی موتور جستجوی پارسی‌جو از سال ۱۳۸۰ آغاز شده، در سال ۱۳۸۸ وارد مرحلهٔ پیاده‌سازی شده و اوّلین نسخهٔ آن از سال ۱۳۸۸ با پوشش یک میلیون صفحه وب فارسی فعّالیّت رسمی خود را آغاز کرده است. این موتور جستجو در حال حاضر، با پانزده سرویس فعال، خدمات جستجوی متن، تصویر، ویدئو و آوا را در اختیار کاربران فارسی زبان قرار می‌دهد.

پرس‌وجوهای خام فارسی مورد استفاده در این پژوهش به‌صورت مستقیم از دادگان تاریخچه‌ای موتور جستجوی پارسی‌جو در بازهٔ شش ماههٔ دوم سال ۱۳۹۴ (از یکم مهرماه تا بیست و نهم اسفندماه) استخراج شده‌اند. تمام پرس‌وجوهای ارسال‌شده توسط کاربران روباتی[1] با استفاده از روش پیشنهادی توسط فلاح و ظریف‌زاده (۲۰۱۶) در این دادگان مشخص شده و

1. Bot users

از مجموعه حذف شده‌اند. پرس‌وجوهای خام باقیمانده بالغ بر هیجده میلیون و نهصد و سی و شش هزار و دویست و هشتاد و یک پرس‌وجو هستند.

از آنجا که در سیستم تاریخچهٔ موتور جستجوی پارسی‌جو از شناسهٔ نشست با بیشینهٔ فاصلهٔ مجاز عدم‌فعّالیت برابر سی دقیقه بهره گرفته می‌شود، اگر از آخرین پرس‌وجوی فرستاده شده توسط کاربر بیش از سی دقیقه زمان بگذرد، یک نشست جدید برای وی اختصاص می‌یابد. با احتساب این شرط، حداکثر فاصلهٔ مجاز بین دو پرس‌وجوی بازآرایی‌شده در حدود سی دقیقه منهای چند میلی‌ثانیه خواهد بود. پس از تخصیص پرس‌وجوها به هر کاربر و تعیین مرز نشست‌ها، در داخل هر نشست باید پرس‌وجو جفت شود و دسته‌بندی هر جفت نیز به آن اختصاص داده شود.

برای تولید جفت پرس‌وجوهای بازآرایی‌شده از ترکیب معیارهای فاصله در سطح نویسه[1] و سطح کلمه[2]، مشابه جیانگ و همکاران (۲۰۱۵)، بر روی پرس‌وجوهای ارسال‌شده توسط یک کاربر بهره گرفته شده است. در این روش فاصلهٔ ویرایشیِ دو پرس‌وجو در سطح نویسه و در سطح کلمه محاسبه شده که از آنها به‌ترتیب با عناوین ChEditDistance و WEditDistance نام برده می‌شود. علاوه بر این موارد، معیار ژاکارد[3] زیررشته‌های با طول حداکثر چهار بین دو پرس‌وجو نیز محاسبه شده است. معیار ژاکارد از معیارهای متداول برای شباهت‌سنجی بین دو مجموعه است. این معیار از حاصل تقسیم تعداد اعضای مجموعه‌ی حاصل از اشتراک دو مجموعه بر تعداد اعضای مجموعه‌ی اجتماع آنها به دست می‌آید. حدِّ بالای پارامتر ChEditDistance برابر ۰/۳۳ طول پرس‌وجو برحسب نویسه و حدِّ بالای پارامتر WEditDistance برابر دو درنظر گرفته شده است و ژاکاردِ زیررشته‌های با طول حداکثر چهار بین دو پرس‌وجو باید بزرگتر یا مساوی مقدار ۰/۲۵ باشد. درنهایت، آن دسته از پرس‌وجوها که در یک نشست زمانی قرار گرفته و با احتساب شروط مربوط به ChEditDistance و WEditDistance حدنصاب لازم را کسب کنند و یا در یک نشست واحد قرار گرفته و شرط مربوط به ژاکارد را برآورده کنند، جفت شده و برای تحلیل‌های آتی مورد استفاده قرار گرفته‌اند.

یکی از مواردی که در این تحقیق مورد توجه قرار گرفته است، جمع‌آوری اطلاعات کلیک هر کاربر است. زیرا یکی از اساسی‌ترین دادگان برای تحلیل میزان رضایت کاربران، از بررسی الگوهای کلیک آنها به‌دست می‌آید. دربارهٔ کلیک هیچ محدودیتی بر روی حالات مختلف درنظر گرفته نشده است، زیرا در پژوهش‌های متداول از وجود یا غیبت کلیک برای اوّلین پرس‌وجو در

1. character level
2. word level
3. Jaccard

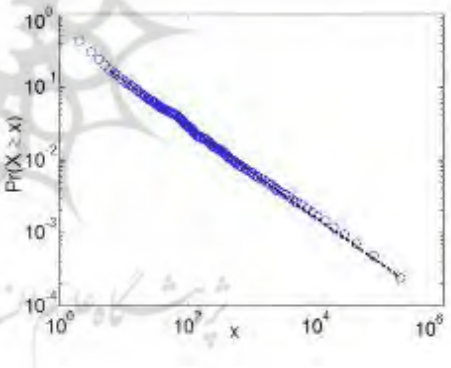
هر جفت بازآرایی‌شده به‌عنوان علائم متفاوتی در تعیین سطح رضایت کاربران بهره گرفته شده است.

۴-۲- بررسی مشخصات واژگانی

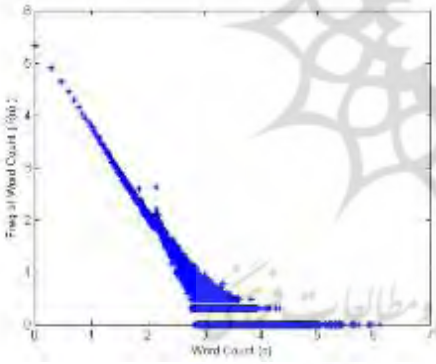
ساده‌ترین نوع تحلیل قابل انجام بر روی پرس‌وجوها و به‌صورت کلی دادگان واژگانی، استخراج مشخصات واژگانی[1] و توزیع آماری عبارت‌ها و یا بخش‌های متن است. به همین دلیل در اوّلین گام به استخراج مشخصات واژگانی پرس‌وجوها پرداخته‌ایم و در گام دوم جفت‌های بازآرایی‌شده را مورد توجه قرار داده‌ایم که نتایج این بررسی‌ها در بخش‌های زیر ارائه خواهد شد.

۴-۲-۱- مشخصات واژگانی در سطح عبارت‌ها

برای بررسی مشخصات واژگانی دو نوع از دادگان، یعنی پرس‌وجوهای خام و جفت‌های بازآرایی شده در دو سطح پرس‌وجو و عبارت، مورد مطالعه قرار گرفته‌اند. در سطح عبارت، نوع توزیع عبارت‌ها در پرس‌وجوها به‌عنوان اولین گزینه انتخاب شده است. نمودار شماره یک تابع توزیع تجمعی تعداد تکرار کلمات و نمودار شماره دو نمودار لگاریتم-لگاریتم تعداد تکرار عبارت‌ها در این دادگان را نشان می‌دهند. این نمودارها شاهدی مبنی بر تبعیت توزیع کلمات از توزیع قانون توان هستند که با یافته‌های (ایرون و مک‌کرلی، ۲۰۰۳؛ لِمپِل و مُران، ۲۰۰۳؛ ونگ، بِری و یَنگ، ۲۰۰۳؛ سارایوا و همکاران، ۲۰۰۱) نیز سازگار هستند. لازم به ذکر است در تمام این مطالعات توزیع کلی کلمات در تمام تاریخچهٔ پرس‌وجو مورد توجه قرار گرفته است.



شکل ۱: نمودار تابع توزیع تجمعی تعداد تکرار عبارت‌ها در پرس‌وجوهای خام

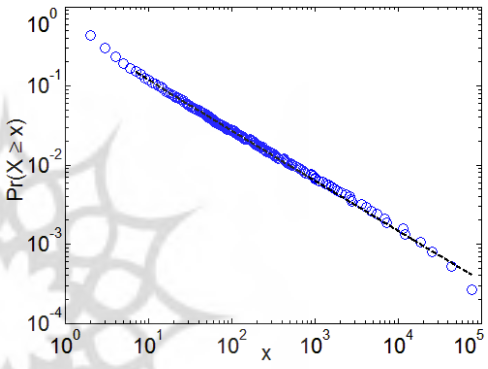


شکل ۲: نمودار لگاریتم-لگاریتم تعداد تکرار عبارت‌ها در پرس‌وجوهای خام

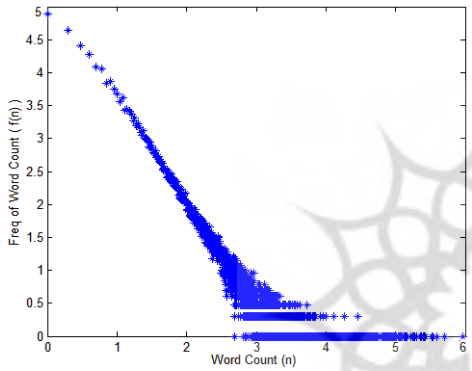
1. lexical specification

امّا شکل ظاهری این توزیع برای اظهارنظر دربارهٔ نحوهٔ بروز عبارت‌ها کافی نیست و پارامترهای توزیع قانون توان مذکور نیز باید مورد بررسی قرار گیرند تا بتوان دربارهٔ عبارت‌ها اظهارنظر کرد. به همین دلیل پیش از پرداختن به مقادیر پارامترها و تحلیل آنها باید این پارامترها را معرفی کرد.

به‌صورت کلی، برای هر مجموعه داده مانند x که از توزیع قانون توان تبعیت کند، تابع چگالی احتمال آن، یعنی $p(x)$ به‌صورت $p(x)d(x) = \Pr(x \le X < x + dx) = Cx^{-\alpha}dx$ تعریف می‌شود، که در آن X مقدار توصیف‌شده[1] و C یک ثابت نرمال‌سازی است. بدیهی است که این تابع نمی‌تواند برای همهٔ مقادیر $x \geq 0$ برقرار باشد و با کاهش x به سمت صفر، واگرا خواهد شد. برای حل این مشکل مقدار X_{min} به‌عنوان کمینهٔ مقادیر x تعیین می‌شود که به ازای xهای بزرگتر از این مقدار می‌توان از همگرایی این تابع اطمینان قابل قبول داشت (کلاوزت[2]، شالیزی و نیومن[3]، ۲۰۰۹).



نمودار ۳- تابع توزیع تجمعی تعداد تکرار عبارت‌ها در پرس‌وجوهای بازآرایی‌شده



نمودار ۴- نمودار لگاریتم-لگاریتم تعداد تکرار عبارت‌ها در پرس‌وجوهای بازآرایی‌شده

با استفاده از روش برازشِ معرفی‌شده در مقالهٔ کلاوزت، شالیزی و نیومن (۲۰۰۹) مقادیر α و $Xmin$ برای این توزیع‌ها محاسبه شدند که مقادیر آنها به‌ترتیب برابر با ۱/۶۲ و ۶ است. در گام بعدیِ بررسی مشخصات واژگانی در سطح عبارت‌ها، بررسی توزیع عبارت‌ها در پرس‌وجوهای بازآرایی‌شده را در دستور قرار می‌دهیم و توزیع عبارت‌ها در پرس‌وجوهای بازآرایی‌شده را

1. observed
2. A. Clauset
3. M. E. J. Newman

بررسی می‌کنیم. نمودار شماره سه تابع توزیع تجمعی توزیع تعداد تکرار کلمات در پرس‌وجوهای بازآرایی‌شده و نمودار شماره چهار لگاریتم-لگاریتم این دادگان را نمایش می‌دهد. مقادیر α و $Xmin$ برای این توزیع‌ها به‌ترتیب ۱/۶۳ و ۷ است. با مقایسهٔ مقادیر α و $Xmin$ در هر دو حالت می‌توان دید، توزیع عبارت‌ها در دادگان تاریخچهٔ پرس‌وجو و دادگان بازآرایی پرس‌وجو بسیار شبیه هستند.

برای مقایسهٔ این دو توزیع و با توجه به تبعیت آنها از توزیع توانی از آزمون ناپارامتری کولموگروف-اسمیرنوف[1] بهره گرفته شده است. مقدار p-value به‌دست‌آمده از این آزمون برابر ۰/۹۹ است بدین معنا که داده‌ها در سطح معناداری پنج درصد شواهد کافی برای ردّ فرضیه H_0 ارائه نمی‌کنند و بنابراین می‌توان ادعا کرد که این دو توزیع بسیار به هم شبیه هستند. به علاوه، از آزمون یویِ مَن-ویتنی[2] بهره گرفته شده است که برطبقِ نتایج این تست با مقدار p-value برابر ۰/۵۸ نیز نمی‌توان گفت اختلاف بین مقادیر میانگین دو توزیع معنادار است.

۴-۲-۲- مشخصات واژگانی در سطح پرس‌وجوها

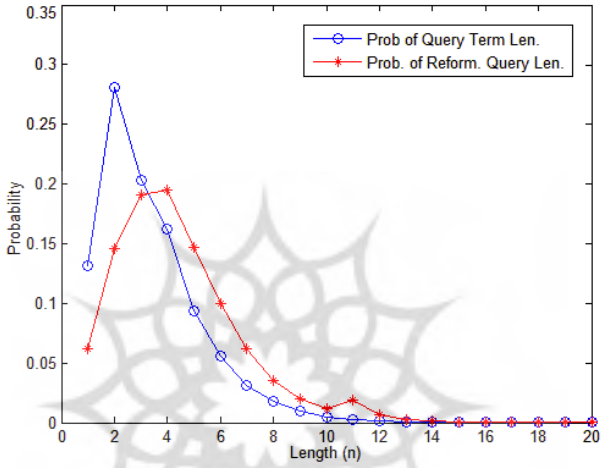
ساده‌ترین نوع تحلیل واژگانی در سطح پرس‌وجوها، بررسی تغییرات طول پرس‌وجو در دادگان است. اسپینک[3] و همکاران (۲۰۰۱) متوسط تعداد عبارات در پرس‌وجوهای موتور جستجوی Alta Vista در سال ۲۰۰۱ را برابر ۲/۳۵ اعلام کرده‌اند. ازطرفی، مطالعاتی مانند تقوی و همکاران (۲۰۱۲) نشان داده‌اند که طول پرس‌وجوهای کاربران به‌مرور زمان در حال افزایش است و تا زمان انجام تحقیق تقوی و همکاران (۲۰۱۲) به مرز ۳/۰۸ عبارت رسیده بوده است. تحقیقاتِ انجام‌شده توسط مؤسسات بهینه‌سازیِ جستجو مانند (ملُنی[4]، ۲۰۱۲) این مقدار را برای موتورهای جستجوی Ask.com و Google به‌ترتیب برابر ۴/۸۱ و ۴/۲۹ اعلام کرده‌اند.

در بررسی حاضر، متوسط تعداد عبارت‌ها در پرس‌وجوهای خام برابر ۴/۴۳ است و همچنین میانگین تعداد عبارت‌های مورد استفاده در پرس‌وجوی اول از جفت‌های بازآرایی‌شده برابر ۳/۰۱ عبارت است و میانگین تعداد عبارات در جفتِ بازآرایی‌شدهٔ آنها ۳/۳۸ است. میانگین پایین‌تر در جفت‌های بازآرایی‌شده، نشان‌دهندهٔ این حقیقت است که کاربران در پرس‌وجوهای با طولِ کوتاه‌تر از متوسطِ طولِ پرس‌وجوها نتوانسته‌اند نتایج مناسب را به‌دست آورند و اقدام به بازآرایی آنها کرده‌اند. این نتایج با یافته‌های مندرج در تحقیقاتی مانند هوآنگ و افثیمیادیس

1. Two-Sample Kolmogorov-Smirnov test
2. Mann-Whitney U-test
3. A. Spink
4. J. Meloni

(۲۰۰۹) و تقوی و همکاران (۲۰۱۲) کاملاً همخوانی دارد و این باور را که پرس‌وجوهای کوتاه‌تر ابهام بیشتری دارند و موتورهای جستجو در پاسخ‌گویی به آنها با مشکلات بیشتری مواجه می‌شوند تأیید می‌کند. برای مقایسهٔ جزئی‌ترِ آمار مربوط به تغییرات طول پرس‌وجوهای موجود در دادگان تاریخچهٔ جستجو و پرس‌وجوهای بازآرایی‌شده، نمودار شماره پنج تعداد تکرار پرس‌وجوها با طول‌های مختلف را نشان می‌دهد.

مقایسهٔ این نمودار با نمودارهایی نظیر آنچه در تقوی و همکاران (۲۰۱۲) و اسپینک و همکاران (۲۰۰۱) گزارش شده‌اند نشان می‌دهد، بر خلاف آمارهای جهانی که در آنها پرس‌وجوهای با طول یک کلمه بیش از سایر انواع پرس‌وجوها مورد استفاده قرار می‌گیرند، کاربران فارسی زبان از پرس‌وجوهای با طول دو (۲) عبارت، بیش از سایر انواع پرس‌وجوها بهره می‌گیرند. تغییر طول پرس‌وجوها در مطالعات جهانی، پس از طولِ یک، روندی کاهشی دارد که



نمودار ۵- احتمال پرس‌وجوهای خام و بازآرایی‌شده با طول‌های مختلف

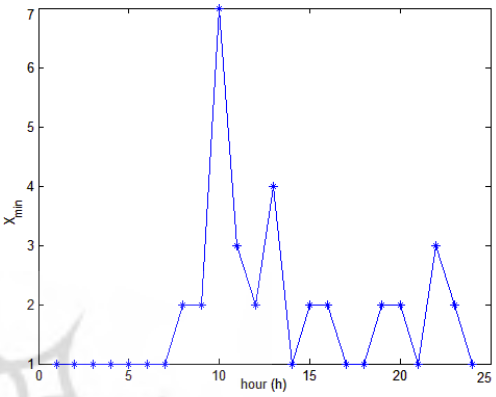
این روند نیز در پرس‌وجوهای فارسی به‌صورت مشابه قابل ملاحظه است؛ با این تفاوت که این روند از عدد دو آغاز شده است.

۴-۳- بررسی تغییرات زمانی

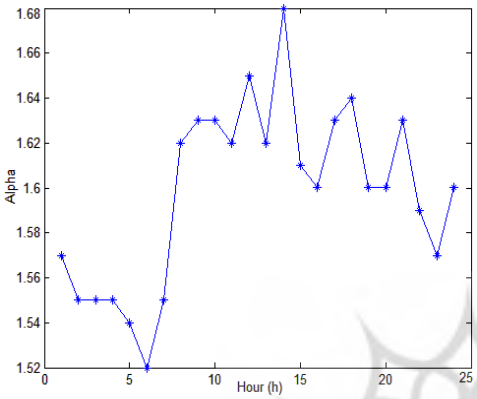
به‌طور کلی تغییرات الگوهای تعامل کاربران در بازه‌های زمانی مختلف به اندازهٔ سایر انواع تحلیل‌های صورت‌گرفته بر روی پرس‌وجوها مورد توجه قرار نگرفته‌اند. در زبان فارسی نیز با توجه به کم‌سابقه‌بودنِ تحقیقات بازیابی اطلاعات و دسترسی به اطلاعاتِ تاریخچهٔ پرس‌وجو، این تلاش‌ها بسیار نادر بوده است. تنها تحقیق منتشرشده در حوزهٔ تحلیلِ زمانی بر روی الگوهای جستجوی کاربران فارسی‌زبان در پژوهشِ کاوه‌یزدی، زارع‌بیدکی و زارع‌میرک‌آباد

(۲۰۱۴) ارائه شده و به بررسیِ وابستگیِ بین الگوهای تعامل کاربران به اخبار پرطرفدار در بازه‌های زمانی مختلف و تأثیرپذیری جستجوی کاربران از این اخبار محدود بوده است.

به همین دلیل، در این تحقیق، بر آن شدیم تا مشخصات کلی تعاملات کاربران در بازه‌های زمانی مختلف را ارزیابی کنیم. این بررسی‌ها در چهار دسته تقسیم‌بندی می‌شوند که این تقسیم‌بندی با لحاظ پارامتر سطح، اعم از عبارت و پرس‌وجو و همچنین نوع دادگان، یعنی پرس‌وجوهای خام و پرس‌وجوهای بازآرایی‌شده به‌دست آمده است و جزئیات و نتایج آن در بخش آتی مورد بررسی قرار می‌گیرد.



نمودار ۶- تغییرات مقادیر X_{min} برای پرس‌وجوهای خام در بازه‌های مختلف زمانی



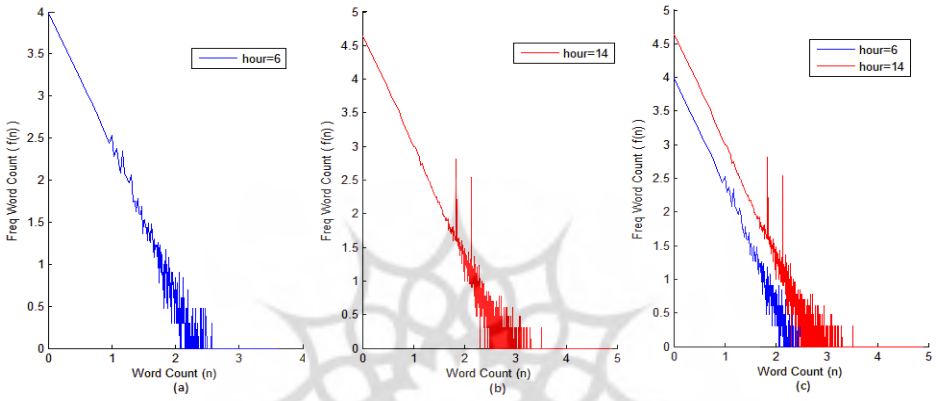
نمودار ۷- تغییرات مقادیر α برای پرس‌وجوهای خام در بازه‌های مختلف زمانی

۴-۳-۱- تغییرات زمانی در سطح عبارت‌ها

برای انجام این بررسی، دادگانِ تعداد تکرار عبارت‌ها در پرس‌وجوهای خام در بازه‌های یک‌ساعته بین ساعت صفر تا '۲۳:۵۹ جمع‌آوری شده و پارامترهای توزیع قانون توان با استفاده از روشِ معرفی‌شده در کلاوزت و همکاران (۲۰۰۹) محاسبه شده‌اند. نمودارهای شش و هفت، به‌ترتیب تغییرات پارامترهای X_{min} و α قانون توان پرس‌وجوهای خام در ساعات مختلف شبانه‌روز را مشخص می‌کنند. با توجه به اینکه پارامترهای توزیع دادگانِ پرس‌وجوهای خام و بازآرایی‌شده چندان تفاوتی نداشته‌اند، نتایج این آزمایش برای پرس‌وجوهای بازآرایی‌شده، ترسیم نشده است.

نتایج این بررسی‌ها نشان‌دهندهٔ تغییراتی جزئی در مقادیر پارامترهای توزیعِ برازش یافته بر توزیع کلمات در ساعات مختلف شبانه‌روز است. هرچند مقادیر تغییرات توان بسیار کوچک

هستند، امّا باید توجه کرد که این مقادیر به‌عنوان مقدار توان در یک تابع نمایی می‌توانند تأثیر قابل توجهی بر تعداد کلمات داشته باشند. برای واضح‌ترشدنِ اثر تغییراتِ پارامتر α بر توزیع کلمات، نمودار لگاریتم-لگاریتمِ توزیع عبارت‌ها در بازه‌های یک‌ساعتهٔ (۷-۶[۱]) و (۱۵-۱۴] که به‌ترتیب دارای کمترین و بیشترین مقدار α بوده‌اند، در نمودار هشت رسم شده است. این نمودار از سه بخش تشکیل شده است که به‌ترتیب نمودار لگاریتم-لگاریتم توزیع کلمات در بازهٔ (۷-۶]، بازهٔ (۱۵-۱۴] و نمایش توأم این دو توزیع در یک نمودار را نمایش می‌دهند. باید خاطرنشان کرد که با توجه به مقادیر X_{min} در نمودار شش**نمودار ۶**، می‌توان دید این مقادیر برای هر دو بازهٔ نمونه با هم برابر و مساویِ مقدار یک است؛ این برابری بدان معنا است که رفتار توزیع در بالای مقدارِ یک با دقت مناسبی بر توزیعِ تخمینی برازش می‌یابد. بنابراین رفتار کلمات با حداقل تعداد تکرار یک در هر دو بازهٔ زمانی مشابه و قابل مقایسه است.



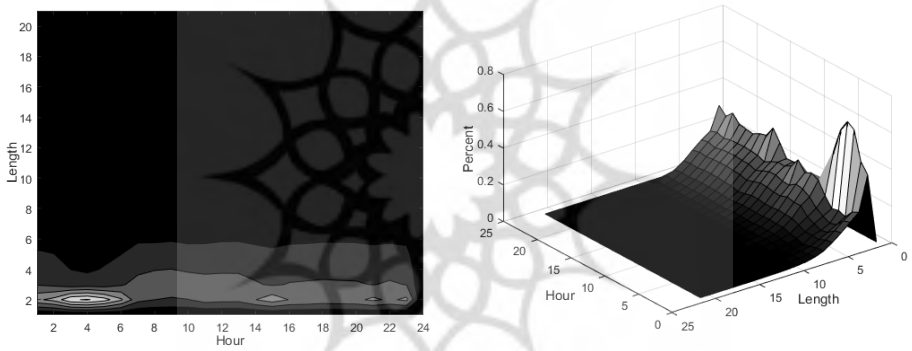
نمودار ۸- نمودار لگاریتم-لگاریتم تعداد تکرار عبارت‌ها در پرس‌وجوهای خام مربوط به بازهٔ (۷-۶] به رنگ آبی(a)، بازهٔ (۱۵-۱۴] به رنگ قرمز (b) و نمودار مقایسهٔ هر دو بازهٔ زمانی (c)

چنان‌که از نمودار هشت برمی‌آید، کلمات با بیشترین تعداد تکرار در ساعاتی با بیشترین α، تعداد تکراری بیش از کلمات مشابه در بازه‌های با αی کمتر دارند. از طرف دیگر، با دقت در دم سنگین توزیع می‌توان دریافت در بازهٔ زمانی (۱۵-۱۴] کلمات بیشتری با تعداد تکرار کم ظاهر شده‌اند که به‌معنای وجود تنوعِ بیشتر در کلمات است. همچنین تعدادی کلمه با تعداد تکرار غیرعادی که منجر به حضور نقاط پرت در نمودار مربوط به بازهٔ (۱۵-۱۴] شده‌اند در این دادگان وجود دارند.

۱ وجود [به معنای این است که بازه‌ی زمانی با در نظر گرفتن ۶ صبح به عنوان مبدا آغاز شده است و وجود) به معنای این است که ساعت ۷ صبح در بازه‌ی زمانی در نظر گرفته نمی‌شود.

به‌صورت کلی می‌توان گفت، در بازه‌های زمانی با αی بیشتر، بیشترین تعداد تکرار کلمات در بازه‌های بالاتری قرار می‌گیرند. از طرف دیگر، بخش دم سنگین توزیع قانون توان در این ساعات معمولاً متراکم‌تر هستند که منجر به افزایش شیب نمودار و در نتیجه افزایش α می‌شود.

به‌صورت شهودی می‌توان گفت، با توجه به موقعیت زمانیِ دو بازه که یکی در میانهٔ روز و دیگری در ساعات ابتدایی روز واقع هستند، تغییرات توزیع کلمات طبیعی است؛ زیرا در ساعات میانی روز معمولاً کاربرانِ زیادتری نسبت به ساعات ابتدایی صبح از خدمات جستجو استفاده می‌کنند که در نتیجه نیازهای متنوع‌تری دارند و از کلمات بیشتری برای بیان مقاصد خود بهره می‌گیرند. تعدد کاربران در ساعات میانی روز می‌تواند بر مقدار بیشینهٔ تعداد تکرار و شیب تابع مؤثر باشد. به‌صورت مشابه، تنوع نیازهای کاربران به تنوع بیشتر در پرس‌وجوها منجر می‌شود که نتیجه‌ای جز افزایش تعداد کلمات با تعداد تکرار کمینه ندارد. برای اطمینان از نتیجه‌گیری مزبور نیز به‌صورت مشابه از آزمون‌های ناپارامتری کولموگروف-اسمیرنوف و یوی مَن-ویتنی بهره گرفته شده است. مقدار p-value به‌دست‌آمده از آزمون اوّل برابر ۰/۰۹ است که با سطح معناداری ۱۰٪ شواهد کافی برای ردّ H_0 را فراهم نمی‌کند ولی با سطح معناداری پنج درصد آن را رد می‌کند.



نمودار ۹- کانتور تغییرات طول پرس‌وجوهای خام در بازه‌های زمانی مختلف به انضمام نمایش سه‌بعدی

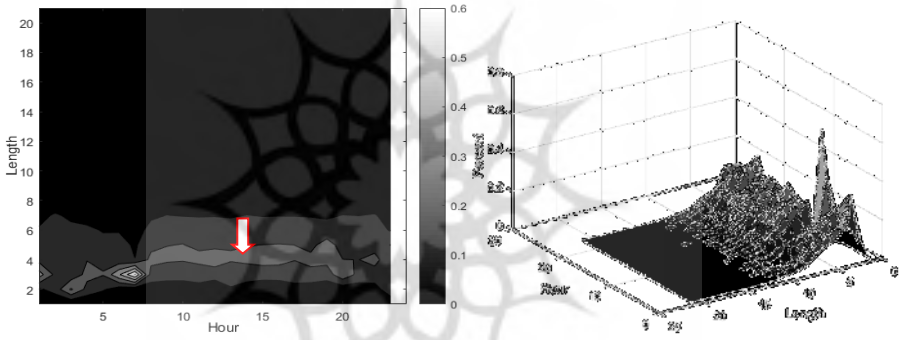
با توجه به سطح معناداریِ بالاتر از پنج درصد، از آزمون دوم نیز بهره می‌بریم. مقدار p-value آزمون من-ویتنی برابر ۰/۰۰۴۵ است که نشان‌دهندهٔ اختلاف معنادار بین مقادیر میانگین دو توزیع است. درنهایت، می‌توان گفت نتایج این دو آزمون نتایج شهودیِ بیان‌شده دربارهٔ تبعیت از توزیع تقریباً مشابه با شیفت در میانگین را تایید می‌کند.

۴-۳-۱- تغییرات زمانی در سطح پرس‌وجوها

تغییرات زمانیِ مورد بررسی در سطح پرس‌وجوها به سه دسته تقسیم شده‌اند که عبارتند از:

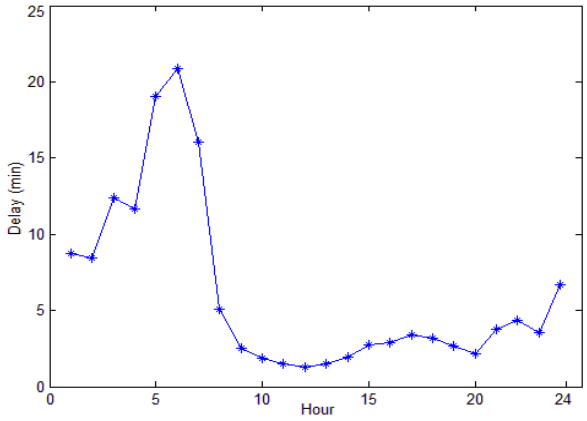
- بررسی تغییرات طول پرس‌وجوها در طول زمان
- بررسی تغییرات طول پرس‌وجوهای بازآرایی‌شده در طول زمان
- بررسی متوسط زمانی بین پرس‌وجوهای بازآرایی‌شده در طول زمان

برای بررسی تغییرات طول پرس‌وجو در گذر زمان در طول یک شبانه‌روز، از کانتور تغییرات طول در زمان بهره می‌گیریم که در نمودار نه قابل مشاهده است. در این نمودار فاصلهٔ مابین کمینه و بیشینهٔ فراوانی بر روی طیفی مابین سیاه خالص تا سفید خالص نگاشت می‌یابد. از طرفی، ماتریسی شامل فراوانی نرمال تعداد پرس‌وجوهای با طول متفاوت در هر ساعت از شبانه‌روز تولید شده است که در نمایش کانتور به هریک از مقادیر مندرج در این ماتریس براساس محل قرارگیری آن در فاصلهٔ بین کمینه تا بیشینه، رنگ مناسبی اختصاص داده می‌شود. درواقع، برای یک بازهٔ زمانیِ مشخص، تعداد پرس‌وجوهای با طول‌های مختلف شمارش شده و بر مجموعِ تعداد پرس‌وجوها در همان بازهٔ زمانی تقسیم شده است. با این رفتار می‌توان درصدی از پرس‌وجوها با طول مشخص را در هر بازهٔ زمانی تعیین کرد.



نمودار ۱۰- کانتور تغییرات طول پرس‌وجوهای بازآرایی‌شده در بازه‌های زمانی مختلف به انضمام نمایش سه‌بعدی

چنان‌که در نمودار نه قابل مشاهده است، در بیشتر بازه‌های زمانی، پرس‌وجوهای باطول کمتر از هشت عبارت، اکثریت موارد را تشکیل می‌دهند. از سوی دیگر می‌توان دید بیشترین درصد از پرس‌وجوها در بازهٔ زمانی بین یک بامداد تا هفت صبح را مواردی با طول بسیار کوتاه (بین یک تا سه) به خود اختصاص می‌دهند. این نسبت در ساعات دیگرِ شبانه‌روز رعایت نشده است و در ساعات کاری صبح تا قبل از ساعت پانزده، درصد پرس‌وجوهای با طول بین سه تا شش نسبت به بازهٔ زمانی بین ساعات یک بامداد تا هفت صبح، فراوانی بیشتری داشته‌اند.



نمودار ۱۱- تغییرات میانگین تأخیر بین دو پرس‌وجوی بازآرایی‌شده در بازه‌های زمانی مختلف

در بازه‌های زمانیِ مربوط به ساعات پانزده، بیست و یک و بیست و سه نیز پرس‌وجوهای با طول دو با درصد بیشتری به سیستم وارد شده‌اند، ولی در بقیهٔ بخش‌های بازهٔ بین هفت صبح تا بیست و سه، پرس‌وجوهای با طول دو تا چهار تقریباً به‌صورت یکنواخت با نرخ نزدیک به هم و البته بیشتر از سایر انواع پرس‌وجوها با طول‌های مختلف به سیستم وارد شده‌اند.

اما ممکن است این سؤال مطرح شود که تغییراتِ زمانیِ پرس‌وجوهای بازآرایی‌شده از چه الگویی تبعیت می‌کند؟ در بخش قبلی با بررسیِ زمانیِ مجموعهٔ پرس‌وجوهای بازآرایی‌شده مشخص شد پرس‌وجوهای با طول‌های کمتر، بیشتر در معرض فرایند بازآرایی قرار گرفته‌اند زیرا این پرس‌وجوها اصولاً ابهام بیشتری نسبت به پرس‌وجوهای با طول بیشتر دارند. برای بررسیِ صحت این ادعا و بررسی تغییرات نرخِ بازآرایی در بازه‌های زمانیِ مختلف کانتوری مشابه نمودار نه اما این بار برای پرس‌وجوهای بازآرایی‌شده در نمودار ده رسم شده است.

در این نمودار احتمال بروز پرس‌وجوها به‌ازای هر طول مشخص در بازه‌های زمانیِ مختلف، محاسبه شده است. چنان‌که در نمودار ده به‌چشم می‌خورد، بیشترین تجمع پرس‌وجوها در ناحیهٔ مربوط به پرس‌وجوهای با طول یک تا شش است. بدین معنا که این پرس‌وجوها معمولاً بیشترین درخواست برای بازآرایی را داشته‌اند. با دقت در نمودار نه مشخص می‌شود که حد پایینِ پرس‌وجوهای محتمل با احتمالی در حدود ۰/۲ (مشخص شده با پیکان به رنگ روشن) از پرس‌وجوهای با طول یک آغاز می‌شود، در حالی که در پرس‌وجوهای بازآرایی‌شده (نمودار ده**نمودار**) این نوار از پرس‌وجوهای با طول سه آغاز می‌شود. از طرفی، مرکز ناحیهٔ با احتمال ۰/۲ (مشخص شده با پیکان به رنگ روشن) در نمودار ده نسبت به نمودار نه از عرض کمتری

برخوردار است. بدین معنا که تجمع پرس‌وجوهای قابل بازآرایی بیشتر حول مواردی با طول سه رخ می‌دهد و طول‌های کمتر یا بیشتر از سه، کمتر در معرض تغییر هستند.

تا اینجا، الگوی تغییرات طول و مشخصات واژگانی ساده در پرس‌وجوهای خام و پرس‌وجوهای بازآرایی‌شده مورد بررسی قرار گرفتند ولی ممکن است این سؤال نیز مطرح شود که کاربران با چه سرعتی نظر خود را تغییر می‌دهند و برای جستجوی مجدد اقدام می‌کنند. برای پاسخ به این پرسش باید فاصلهٔ زمانیِ مابینِ تغییرات نظرات کاربران را اندازه گرفت. بررسی بر روی مجموعهٔ دادگان مشخص کرد، میانگین زمان تأخیر بین دو پرس‌وجوی بازآرایی‌شده برابر ۴/۸۳ دقیقه است. نظر به نتایج قبلی که نشان‌دهندهٔ تغییرات الگوی جستجوی کاربران در بازه‌های زمانی مختلف بوده است، باید اذعان کرد که گمان می‌رود فاصلهٔ زمانیِ مابین تغییر دو پرس‌وجو نیز با زمان متغیر باشد. نمودار یازده تغییرات میزان تأخیر کاربران در بازآرایی را در بازه‌های زمانیِ مختلف نمایش می‌دهد. چنان‌که از این تصویر برمی‌آید، در ساعاتی با بیشترین ترافیک که همزمان با ساعات میانی روز است، کاربران در بازهٔ بسیار کوتاه‌تری نسبت به بازآرایی پرسش‌های خود اقدام می‌کنند که به‌معنای صرف زمان کمتر برای بررسی نتایج پرس‌وجوی اول و سپس تلاش برای بازآرایی آن است. فاصلهٔ زمانیِ بازگشت کاربر در ساعات ابتداییِ صبح بسیار بیشتر است، بدین معنا که کاربران احتمالاً با بررسی نتایج، حتی بر روی برخی از آنها کلیک می‌کنند و سپس تصمیم می‌گیرند.

۴-۴- بررسی الگوهای بازآرایی

با توجه به تنوع فهرست تغییرات مجاز و متداول برای کاربران و برای افزایش سطح مقایسه‌پذیریِ نتایج با موارد مشابه برای زبان انگلیسی، در این تحقیق از مجموعهٔ بازآرایی‌های پیشنهادی هوآنگ و افثیمیادیس (۲۰۰۹) به‌عنوان فهرست پایهٔ تحلیل‌ها بهره گرفته شده است.

۴-۴-۱- بررسی توزیع زمانی الگوهای بازآرایی

جدول چهار نشان‌دهندهٔ آن دسته از بازآرایی‌هایی است که در این تحقیق مورد توجه قرار گرفته‌اند. لازم به‌ذکر است که برای تشخیص هر دسته از بازآرایی‌های مورد استفاده در این پژوهش، به‌جز تصحیح املایی، از قوانین پیشنهادی هوآنگ و افثیمیادیس (۲۰۰۹) بهره گرفته شده است و فقط موارد تصحیح املایی با استفاده از سیستم تصحیح‌گرِ املاییِ موتور جستجوی پارسی‌جو برچسب خود را دریافت کرده‌اند.

براساسِ شروط تعیین‌شده در بخش ۴-۱ و مشابه با جیانگ و همکاران (۲۰۱۵) و هوآنگ و افثیمیادیس (۲۰۰۹)، در صورتی که یک جفت پرس‌وجو شرط حدنصاب زمانی را کسب کند

ولی شروط مربوط به فاصلهٔ ویرایشی و ژاکارد را کسب نکند، به آن برچسبی با عنوان بازنویسی[1] و درصورتِ کسبِ این شرایط، برچسب بازآرایی اختصاص داده می‌شود. از مجموعِ هشت میلیون و دویست و چهل و سه هزار و هشتصد و چهل و شش جفتِ کشف‌شده با تعریف بالا، تعداد سه میلیون و چهارصد و چهار هزار و هشتصد و پنج جفت، معادل ۴۱/۳۰ ٪ از جفت‌ها، شامل تعریف بازنویسی و تعداد چهارمیلیون و هشتصد و سی و نه هزار و چهل و یک جفت یعنی ۵۸/۷۰ ٪ از جفت‌ها در حوزهٔ بازآرایی قرار می‌گیرند.

در بین انواع الگوهای بازآرایی، سررشته‌سازی، زیررشته‌سازی و تصحیح املایی بیشترین کسر از بازآرایی‌ها را به خود اختصاص داده‌اند. بالابودنِ نرخ سررشته‌سازی و زیررشته‌سازی به این دلیل اتفاق می‌افتد که گاهی کاربران ایرانی با تغییر یک وند و یا افزودن یک ترکیب یا صفت به ابتدا/ انتهای پرس‌وجو یا حذف آنها به بازآرایی می‌پردازند و از آنجا که افزایش و یا کاهش وندها و علائم جمع می‌توانند وضعیت زیرا سر رشته‌سازی را تداعی کنند، به همین دلیل این دو حالت بیش از سایر حالات در بین الگوها دیده می‌شوند.

نمودار دوازده درصد دو دستهٔ بازآرایی و بازنویسی را در کنار یکدیگر و درصد هریک از دسته‌های بازآرایی را با همان ترتیبِ معرفی‌شده در جدول چهار به نمایش می‌گذارد. نکتهٔ جالبی که در اینجا به چشم می‌خورد، نوع توزیع جفت‌های مربوط به بازنویسی در بازه‌های زمانی مختلف است. بدین معنا که کاربران در ساعات بامدادی و ابتدای شب، بیشتر از ساعات کاریِ روزانه به بازنویسی پرس‌وجوی خود می‌پردازند.

کِی[2] و هورویتز[3] (۱۹۹۹) در اواخر دههٔ نود میلادی پژوهشی بر روی الگوهای رفتاری کاربران به اجرا گذاشته بودند و نشان داده بودند که میزان پایبندی کاربران به هدف پرس‌وجوی اولیه در جریان بازآرایی و ارسال پرس‌وجوی دوم، با تأخیر زمانی بین این ارسال دو پرس‌وجو رابطهٔ عکس دارد.

با مقایسهٔ نمودارهای یازده و دوازده مشخص می‌شود که در این پژوهش نیز، بازه‌های زمانی که کاربران بیشترین تمایل به بازنویسی پرس‌وجوی خود را به نمایش گذاشته‌اند با بازه‌های زمانی مربوط به بیشترین زمان تأخیر در ارسال بازآرایی همخوانی دارد.

1. rrewriting
2. J. Kay
3. E. Horvitz

جدول ۴- فهرست انواع بازآرایی‌های تشخیص داده شده در پرس‌وجوهای کاربران به انضمام احتمال بروز آنها در دسته‌های بازآرایی/بازنویسی و در کل دادگان

شماره	عنوان دسته	نوع بازآرایی	درصد در کل دادگان	درصد در دسته
۱	بازنویسی	بازنویسی	۴۱/۳۰۱	۱۰۰
۲	بازآرایی	تغییرات فاصله‌گذاری	۱/۷۶۶	۳/۰۱
۳		زیررشته‌سازی[1]	۹/۴۴۴	۱۶/۰۹
۴		سررشته‌سازی[2]	۳۳/۰۰۱	۵۶/۲۲
۵		تصحیح املایی	۷/۲۷۵	۱۲/۳۹
۶		حذف عبارت	۲/۸۴۶	۴/۸۵
۷		اضافه‌کردن عبارت	۴/۱۲۲	۷/۰۲
۸		تغییر ترتیب کلمات	۰/۲۴۰	۰/۴

۴-۴-۲ بررسی تغییرات کلیک الگوهای بازآرایی

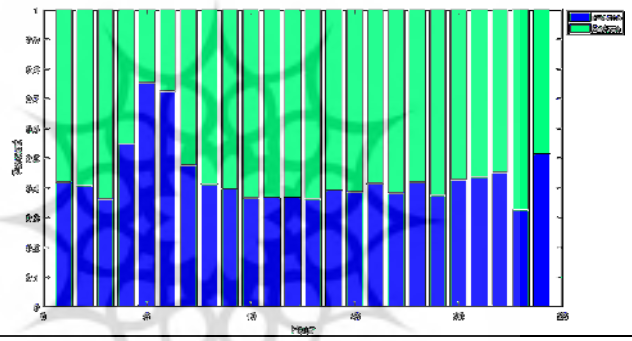
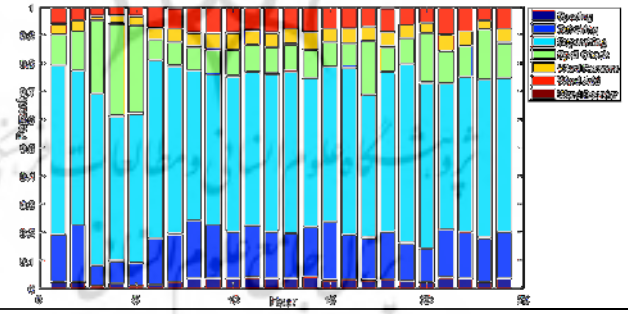
کلیک بر روی نتایج موتورهای جستجو، از مهم‌ترین فاکتورهای تعیین میزان رضایتمندیِ کاربران از نتایج موتورهای جستجو است. از آنجا که یک کاربر به‌ازای نتایج بازیابی‌شده می‌تواند بر روی نتایج کلیک کند و یا آنها را رها کند، بنابراین به‌ازای هر پرس‌وجو رفتار کاربر به دو دستهٔ کلیک[3] و رهاکردن[4] تقسیم می‌شود. در صورتی که کاربری پرس‌وجویی را بازآرایی کند، برای پرس‌وجوی دوم نیز می‌توان همین دو حالت را تعریف کرد و در نتیجه، تعداد حالات مربوط به رفتار کاربران به چهار دسته‌ی {Skip-Skip, Skip-Click, Click-Skip, Click-Click} تقسیم می‌شوند (مشابه هوآنگ و افشیمیادیس، ۲۰۰۹).

جدول ۵- فهرست انواع حالات کلیک در جفت‌های بازآرایی‌شده و احتمال بروز هریک از این حالات

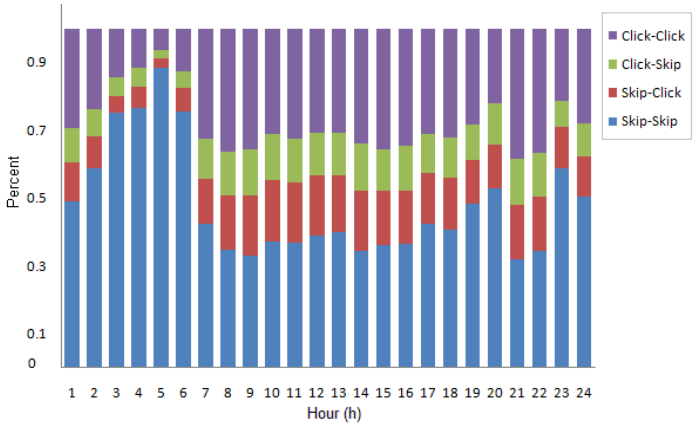
شمارهٔ الگو	الگو	درصد
۱	رهاکردن-رهاکردن (Skip-Skip)	۴۵/۳۸
۲	رهاکردن- کلیک (Skip-Click)	۱۴/۳۲
۳	کلیک-رهاکردن (Click-Skip)	۱۱/۲۵
۴	کلیک-کلیک (Click-Click)	۲۹/۰۴

1. substring
2. superstring
3. click
4. skip

نتایج مندرج در جدول پنج مشابه با یافته‌های نتایج هوآنگ و افشیمیادیس (۲۰۰۹) است. این نتایج نشان‌دهنده‌ی این حقیقت است که کاربران در مجموع در تعداد بیشتری از حالات بر روی نتایج کلیک کرده‌اند و در کمتر از نیمی از موارد، از نتایج هر دو پرس‌وجوی خود ناراضی بوده‌اند. نمودار سیزده نتایج تحلیل تغییرات زمانی الگوهای کلیک را نشان می‌دهد. براین اساس، مجموع نرخ پرس‌وجوها با حداقل یک کلیک، صرف‌نظر از نتایج بازیابی‌شده در ساعات کم‌ترافیک، بیشتر از سایر بازه‌های زمانی است. ممکن است در نگاه اوّل این یافته با تحلیل‌های بخش قبلی متضاد به نظر برسد ولی باید به یک نکته توجه کرد و آن اینکه در بازه‌های زمانیِ شبانگاهی و بامدادی، نسبت تعداد پرس‌وجوهای کلیک شده کمتر از ساعات دیگر شبانه‌روز هستند ولی کسر کوچکی از کاربران که بر روی نتایج کلیک کرده‌اند با بررسی نتایجِ بازیابی‌شده در فواصل زمانیِ بیشتری اقدام به ارسال پرس‌وجوی مجدد کرده‌اند و در صورتی که پرس‌وجوی بازآرایی‌شده‌ای ارسال کرده‌اند، پرس‌وجوی خود را با تغییرات بیشتری نسبت به ساعات دیگر شبانه روز ارسال کرده‌اند.

الف.	
ب	

نمودار ۱۲- درصد پرس‌وجوهای بازآرایی و بازنویسی‌شده در ساعات مختلف شبانه‌روز (الف) و درصد هر یک از روش‌های بازآرایی در این دسته (ب).



نمودار ۱۳- تغییرات زمانی احتمال هر یک از چهاردسته الگوی معرفی شده در (هوآنگ و افیثیمیادیس،۲۰۰۹)

۵- کاربردهای تحلیل الگوهای تعامل

در بخش‌های قبلی، آمارهای مربوط به انواع تعامل‌های کاربران، اطلاعات آماریِ مربوطه و توزیع زمانیِ آنها ارائه شد ولی مشخص نشد مجموعهٔ این اطلاعات با چه هدف و کاربردی استخراج شده‌اند یا لزوم بررسی الگوهای تعاملات کاربران با موتورهای جستجو چیست. به همین دلیل، در این بخش به معرفی کاربردهای نتایج به‌دست‌آمده از هریک از این تحلیل‌ها در حوزه‌های مختلف اشاره می‌شود.

اطلاع از مشخصات واژگانی و آماری پرس‌وجوها، به‌خصوص توزیع آماریِ مربوط به کلمات و اجزای آنها، می‌تواند راهگشای طراحیِ ساختارهایی برای ذخیره‌سازیِ داده‌های حجیم تاریخچهٔ پرس‌وجو باشد و یا راهکارهایی برای تنظیم مشخصات حافظه‌های نهان ارائه دهد. از سوی دیگر، می‌توان با مطالعهٔ میزان موفقیت‌آمیزبودنِ انواع الگوهای بازآرایی ازطریق کلیک‌ها و مدل‌های آماریِ حاکم بر آنها سیستم‌های گسترش پرس‌وجویی ارائه کرد که در عین بالابردنِ احتمالِ بازیابیِ اسناد مرتبط، تعداد تلاش‌های لازم از سوی کاربران برای نیل به هدف را کاهش دهند.

در حوزهٔ بررسی‌های زمانی می‌توان به نمونه‌هایی نظیر بیتزل و همکاران (۲۰۰۴) و ژانگ[1]، جنسن و اسپینک (۲۰۰۹) اشاره کرد و باید دانست که این نتایج در فرایندهای نگهداری موتورهای جستجو نقش به‌سزایی دارند. از آنجا که موتورهای جستجو برای حفظ پویایی خود باید دائماً فرایندهای به‌روزرسانی را به اجرا گذارند و در حین این فرایندها معمولاً عملکرد سیستم برای بازه‌ای هرچند کوتاه مختل می‌شود، باید بتوان زمان و طول مدت مناسب برای

1. Y. Zhang

این عملیات را از قبل تعیین کرد. برای این هدف می‌توان سیستم‌ها را براساس میزان بارِ کاریِ آنها در بازه‌های زمانیِ مختلف دسته‌بندی کرد و فرایند به‌روزرسانیِ آنها را در ساعاتی با کمترین ترافیک به اجرا گذارد.

۶- نتیجه‌گیری

در این تحقیق، الگوهای تعامل کاربران با یک موتور جستجوی فارسی ازطریقِ بررسیِ دادگانِ تاریخچهٔ جستجو مورد توجه قرار گرفته است. برای این منظور، پرس‌وجوهای خام از تاریخچهٔ جستجوی موتور جستجوی پارسی‌جو استخراج می‌شوند و سپس با اعمال تحلیل‌های واژگانی و زمانی بر روی آنها، آن دسته از پرس‌وجوهایی که شرایطِ تعیین‌شده برای بازآرایی یا بازنویسی را برآورده کنند، با همین عناوین برچسب‌گذاری می‌شوند.

در ادامه، بررسی‌های متنوعی بر روی پرس‌وجوهای خام و بازآرایی‌شده صورت گرفته است. این بررسی‌ها شامل ویژگی‌های واژگانی و زمانی و بررسی‌های الگوهای بازآرایی است. در قالب بررسی‌های واژگانی، مواردی ازجمله توزیعِ آماریِ عبارت‌ها در پرس‌وجوها و فراوانی پرس‌وجوها با طول‌های مختلف مورد بررسی قرار گرفتند. دستهٔ دومِ بررسی‌ها به تغییرات پرس‌وجوها، توزیع عبارت‌ها و فاصلهٔ زمانیِ بین ارسال پرس‌وجوهای بازآرایی‌شده در ساعات مختلفِ شبانه‌روز پرداخته‌اند. نهایتاً آخرین دسته با استفاده از رویکردِ الگوشناختی، پرس‌وجوهای کاربران را مورد توجه قرار داده‌اند. تحقیقات مشابه معمولاً تلاش خود را در یکی از رویکردهای سه‌گانهٔ بالا معطوف می‌کنند ولی با توجه به تازگیِ این حوزه در زبان فارسی و همچنین نیاز گروه‌های تحقیقاتی به دسترسی به اطلاعات مذکور، در این تحقیق تلاش‌ها به ارائهٔ تصویری کلی از تعامل کاربرانِ فارسی‌زبان با موتور جستجوی فارسی معطوف بوده است. به‌علاوه، در کنارِ نتایج هر بررسی، تحلیل‌هایی از رفتارهای کاربران در مقایسه با کاربران موتورهای جستجو ارائه شده است که نشان‌دهندهٔ همخوانی این رفتارها با رفتارهای کاربران موتورهای جستجوی بین‌المللی است. درنهایت، مجموعه‌ای از کاربردها در حوزهٔ طراحی و بهبود موتورهای جستجو که از نتایج این تحلیل‌ها بهره می‌گیرند، معرفی شده‌اند.

سپاسگزاری

نویسندگان مراتب قدرانیِ ویژهٔ خود را از همکاری خانم‌ها صدیقه طباطبایی و مهدیه فلاح و آقایان محمدصادق طاهرزاده و فرزاد نیازمند در مسیر جمع‌آوری و آماده‌سازیِ دادگان ابراز می

کنند. در پایان، از زحمات آقای دکتر سیدمحسن میرحسینی، استادیار گروه آمار در دانشکدهٔ ریاضی دانشگاه یزد برای ارائهٔ مشورت در اجرای تحلیل‌های آماری تشکر می‌کنیم.

منابع

Anick, P. (2003). "Using terminological feedback for web search refinement". *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. New York: ACM Press, 88–95. Available in: http://doi.org/10.1145/860435.860453

Beitzel, S. M., et al. (2004). "Hourly Analysis of a Very Large Topically Categorized Web Query Log". *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM Press, 321–328. Available in: http://doi.org/10.1145/1008992.1009048

Clauset, A., C. R. Shalizi, & M. E. J. Newman (2009). "Power-Law Distributions in Empirical Data". SIAM Review, 51(4), 661–703. Available in: http://doi.org/10.1137/070710111

Eiron, N., & K. S. McCurley (2003). "Analysis of Anchor Text for Web Search". *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval.* SIGIR '03, New York: ACM Press, 459-460. http://doi.org/10.1145/860435.860550

Fallah, M., & S. Zarifzadeh (2016). "Click Spam Detection Based on User Session Classification". *Proceedings of the 21th Annual Conference of Computer Society of Iran (CSICC2016)*. Tehran: CSI Conference Publications, 646–651.

Guo, J., et al. (2008). "A Unified and Discriminative Model for Query Refinement". *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* - SIGIR '08. New York: ACM Press, 379-386. Available in: http://doi.org/10.1145/1390334.1390400

Hassan, A., et al. (2013). "Beyond Clicks: Query Reformulation as a Predictor of Search Satisfaction". *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management*. New York: ACM Press, 2019–2028. Available in: http://doi.org/10.1145/2505515.2505682

Hassan A. A. (2013). "Identifying Web Search Query Reformulation Using Concept based Matching". *Empirical Methods in Natural Language Processing (EMNLP)*. Seattle, Washington: Association for Computational Linguistics, 1000–1010. Retrieved from

https://www.microsoft.com/en-us/research/publication/identifying-web-search-query-reformulation-using-concept-based-matching/

Huang, J., & E. N. Efthimiadis (2009). "Analyzing and Evaluating Query Reformulation Strategies in Web Search Logs". *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. New York: ACM Press, 77–86. Available in: http://doi.org/10.1145/1645953.1645966

Jansen, B. J., et al. (1998). "Real Life Information Retrieval: A Study of User Queries on the Web". *ACM SIGIR Forum*. 32(1), 5–17. Available in: http://doi.org/10.1145/281250.281253

Jansen, B. J., et al. (2007). "Defining a Session on Web Search Engines: Research Articles". *Journal of the American Society for Information Science and Technology*. 58(6), 862–871. Available in: http://doi.org/10.1002/ASI.V58:6

Jiang, J., et al. (2015). "Understanding and Predicting Graded Search Satisfaction". *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. New York: ACM Press, 57–66. Available in: http://doi.org/10.1145/2684822.2685319

Jiang, J., D. He, & J. Allan (2014). "Searching, Browsing, and Clicking in a Search Session: Changes in User Behavior by Task and over Time". *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. New York: ACM Press, 607–616. Available in: http://doi.org/10.1145/2600428.2609633

Jiang, J., & C. Ni (2016). "What Affects Word Changes in Query Reformulation During a Task-based Search Session?". *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. New York: ACM Press, 111–120. Available in: http://doi.org/10.1145/2854946.2854978

Jiang, J.-Y., et al. (2014). "Learning User Reformulation Behavior for Query Auto-completion". *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. New York: ACM Press, 445–454. Available in: http://doi.org/10.1145/2600428.2609614

Kaveh-Yazdy, F., A. M. Zareh-Bidoki, & M. R. Zare-Mirakabad (2014). "Social Event Detection Via Search Engine User Queries". *The 3rd Conference on Computational Linguistics (CLC '14)*. Tehran: Sharif University of Technology.

Kay, J., & E. Horvitz (1999). "UM99: User Modeling". *Proceedings of the seventh international conference on User modeling*. Springer, 392–397.

Kiseleva, J., et al. (2014). "Modelling and Detecting Changes in User Satisfaction". *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14*. New York: ACM Press, 1449–1458. Available in: http://doi.org/10.1145/2661829.2661960

Lempel, R., & S. Moran (2003). "Predictive Caching and Prefetching of Query Results in Search Engines". *Proceedings of the twelfth international conference on World Wide Web- WWW '03*. New York: ACM Press, 19-28. available in: http://doi.org/10.1145/775152.775156

Meloni, J. (2012). *SEO Keyword Alert: Long-Tail Search Most Common on Ask.com*. Retrieved from http://www.brafton.com/news/seo-keyword-alert-long-tail-search-most-common-on-ask-com/

Park, J. Y., et al. (2015). "A Large-Scale Study of User Image Search Behavior on the Web". *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. New York: ACM Press, 985–994. Available in: http://doi.org/10.1145/2702123.2702527

Saraiva, P. C., et al. (2001). "Rank-Preserving Two-Level Caching for Scalable Search Engines". *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval- SIGIR '01*. New York: ACM Press, 51–58. Available in: http://doi.org/10.1145/383952.383959

Sloan, M., H. Yang, & J. Wang (2015). "A Term-Based Methodology for Query Reformulation Understanding". *Information Retrieval Journal*. 18(2), 145–165. Available in: http://doi.org/10.1007/s10791-015-9251-5

Spink, A., et al. (2001). "Searching the Web: The Public and Their Queries". *Journal of the American Society for Information Science and Technology*. 52(3), 226–234. Available in: http://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.3.CO;2-I

Taghavi, M., et al. (2012). "An Analysis of Web Proxy Logs with Query Distribution Pattern Approach for Search Engines". *Computer Standards & Interfaces*. 34(1), 162–170. Available in: http://doi.org/10.1016/j.csi.2011.07.001

Teevan, J., et al. (2007). "Information Re-retrieval". *Proceedings of the 30th annual international ACM SIGIR conference on Research and*

development in information retrieval - SIGIR '07. New York: ACM Press, 151-158. Available in: http://doi.org/10.1145/1277741.1277770

Wang, P., M. W. Berry, & Y. Yang (2003). "Mining Longitudinal Web Queries: Trends and Patterns". *Journal of the American Society for Information Science and Technology*. 54(8), 743–758. Available in: http://doi.org/10.1002/asi.10262

Weber, I., & A. Jaimes (2011). "Who Uses Web Search for What: And How". *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. New York: ACM Press, 15–24. Available in: http://doi.org/10.1145/1935826.1935839

Whittle, M., et al. (2007). "Data Mining of Search Engine Logs". *Journal of the American Society for Information Science and Technology*. 58(14), 2382–2400. Available in: http://doi.org/10.1002/ASI.V58:14

Zhang, Y., B. J. Jansen, & A. Spink (2009). „Time Series Analysis of a Web Search Engine Transaction Log". *Information Processing & Management*. 45(2), 230–245. Available in: http://doi.org/10.1016/j.ipm.2008.07.003