تأثیر الگوی موضوعی رفتار جستجوی کاربران نوجوان بر پیشنهاد پرس‌وجو

حیدر قاسم‌زاده، محمد قاسم‌زاده و علی‌محمد زارع بیدکی

چکیده: کاربران نوجوان هنگام جستجوی موضوع‌های مورد نظرشان، دایره لغات محدودی را در فرمول‌بندی پرس‌وجو به کار می‌برند. مسئله مهم دیگر آن است که کاربران نوجوان غالباً بر روی اقلام اولیه ارائه‌شده در لیست نتایج جستجو کلیک می‌کنند. در این پژوهش برای ترمیم و جبران این ویژگی‌ها، پیشنهاد می‌شود که الگوی موضوعی از روی رفتار کاربر نوجوان بر اساس جستجوهای قبلی کشف شوند و با تکیه بر الگوهای یافت‌شده، پرس‌وجوی مناسب استخراج و به کاربر نوجوان پیشنهاد گردد. در روش پیشنهادی، الگوهای موضوعی بر اساس ویژگی محبوبیت کلیک‌ها و مرتبط‌ترین موضوع‌ها از روی لاگ‌های جستجو که عموماً حجیم هستند استخراج می‌گردند. در ادامه با استفاده از کلاسه‌بندی دودویی، نزدیک‌ترین پرس‌وجو به پرس‌وجوی مورد نظر کاربر نوجوان مشخص می‌شود. در نتیجه با فیلترنمودن نویز ناوبری موضوعی بر اساس استخراج الگوهای موضوعی کلیک‌های کاربران نوجوان یک مدل کاربر با دقت بالاتری برای پیشنهاد پرس‌وجو حاصل می‌گردد. روش پیشنهادی با استفاده از ابزارهای Alteryx و weka پیاده‌سازی و عملکرد آن بر روی لاگ جستجوی AOL که شامل حدود ۲۰ میلیون نمونه تراکنش جستجو مربوط به ۶۵۰ هزار کاربر می‌باشد ارزیابی گردید. نتایج آزمایش‌ها نشان می‌دهند که پرس‌وجوهای ارائه‌شده توسط سیستم پیشنهادی به پرس‌وجوی مورد نظر کاربر نوجوان نزدیک‌تر است و به تبع آن موجب بهبود دستیابی به نتایج مرتبط می‌گردد.

کلیدواژه: **الگوی موضوعی، پیشنهاد پرس‌وجو، رفتار جستجو، کاربر نوجوان، لاگ جستجو.**

۱- مقدمه

در سال‌های اخیر کوشش‌های متعددی در راستای ارتقای عملکرد موتورهای جستجو در ارائه نتایج مرتبط به پرس‌وجوی کاربران انجام شده [۱] و در این راستا لحاظ‌کردن سن کاربران نیز مورد توجه بوده است. تعداد نوجوانان استفاده‌کننده از وب و مقدار زمان استفاده آنها از وب در سال‌های اخیر افزایش یافته است. بر طبق یک بررسی که در سال ۲۰۱۲ از ۸۰۲ والدین و نوجوانان ۱۲ تا ۱۷ سال آنها انجام شد، مشخص گردید که ۹۵٪ از نوجوانان به طور مرتب اینترنت را استفاده می‌کنند، ۷۸٪ دارای یک تلفن همراه و حدود ۴۷٪ دارای یک تلفن همراه هوشمند هستند [۲]. دو عامل اصلی در فرایند جستجوی کاربران توسط موتورهای جستجو مؤثر هستند: ۱) اطلاعات مورد نیازی که کاربر می‌خواهد توسط موتور جستجو برآورده شود باید توسط اضافه‌نمودن کلمه‌های کلیدی صحیح در پرس‌وجو بیان گردد و ۲) برای کسب اطلاعات مورد نیاز توسط کاربر باید بر اساس انتخاب صحیح از لیست نتایج ارائه‌شده توسط موتور جستجو کلیک‌کردن را انجام دهد. کاربران نوجوان به دلیل تجربه کمتری که در فرایند جستجو نسبت به کاربران بزرگ‌سال دارند، این دو عامل را به طور صحیح انجام نمی‌دهند [۳].

اغلب برای ارائه نتایج جستجوی با کیفیت بالا توسط موتور جستجو به کاربران نوجوان نیاز است که پرس‌وجوهای کاربران تصحیح شوند. پرس‌وجوهای مطلوب طولانی‌تر از آنهایی است که توسط کاربر وب ارسال می‌شود که شامل تعیین کلمات کلیدی صحیح برای کاربران نوجوان است. به هر حال، کاربران نوجوان پرس‌وجوهایی را به موتور جستجو وارد می‌کنند که فاقد ویژگی‌های لازم برای بازیابی مطالب مرتبط با اطلاعات مورد نیازشان است که در نهایت منجر به رهاکردن فرایند جستجو توسط آنها می‌گردد. همچنین تفاوت قابل توجهی در اندازه لغت مابین پرس‌وجوهای ارسال‌شده توسط کاربران نوجوان و بزرگ‌سال وجود دارد. بنابراین اهمیت و فوریت فراهم‌سازی ابزارهای کمکی برای پیشنهادکردن پرس‌وجوی صحیح به کاربران نوجوان مشخص می‌شود. فرمول‌بندی پرس‌وجو به عنوان اولین گام در جستجوی فرایند اطلاعات شناخته می‌شود. با توجه به عامل دوم در فرایند جستجو، کاربران نوجوان یک انحراف کلیک بیشتری نسبت به بزرگ‌سالان دارند که منجر به کلیک کمتر بر روی نتایج لیست‌شده در رتبه پایین‌تر می‌شود، رفتاری که مهارت‌های ناوبری و استخراج نتایج را مختل می‌سازد. در پژوهش‌های قبلی نشان داده شده که اطلاعات مورد نیاز و راهکارهای جستجوی نوجوانان و بزرگ‌سالان اساساً متفاوت هستند [۴].

نوجوانان نسبت به بزرگ‌سالان در هنگام جستجو تمرکز کمتری دارند و یک سبک پیشبرد غیر خطی را دنبال می‌کنند که منابعی که قبلاً بررسی نموده‌اند را دو مرتبه استفاده می‌کنند. این رفتار نشان می‌دهد که آنها هنگام جستجو در وب با مشکلاتی در تصمیم‌گیری در مورد اطلاعات مرتبط با موضوع مورد نظرشان مواجه هستند [۵].

به دلیل این که موتورهای جستجوی فعلی همه نوع اطلاعات را فراهم می‌سازند و با توجه به این که کاربران نوجوان پرس‌وجوهای نامناسب به موتور جستجو ارسال می‌کنند و بر روی نتایج در موقعیت‌های با رتبه بالا کلیک می‌کنند، بنابراین توسط موتور جستجو در معرض مطالبی که هدف نیستند و در بعضی مواقع می‌تواند برای آنها مضر باشند قرار می‌گیرند. یکی از راهکارهای دستیابی به پرس‌وجوی مناسب برای کاربران نوجوان، روش پیشنهاد پرس‌وجو است. امروزه انتظاراتی که از موتورهای جستجو است چیزی فراتر از یافتن تعدادی صفحه وب بر اساس کلمات پرس‌وجوی کاربر می‌باشد. مشکل آن است که موتور جستجو یک تعداد بسیار زیادی از صفحات وب مرتبط با کلمات پرس‌وجوی کاربر برمی‌گرداند و کاربران زمان زیادی را برای یافتن مطالب مورد نظرشان صرف می‌کنند.

این مقاله در تاریخ ۱۰ خرداد ماه ۱۳۹۶ دریافت و در تاریخ ۱۳ آبان ۱۳۹۶ بازنگری شد.

حیدر قاسم‌زاده، پردیس فنی و مهندسی، گروه مهندسی کامپیوتر، دانشگاه یزد، یزد، (email: h.ghasemzadeh@stu.yazd.ac.ir).

محمد قاسم‌زاده، پردیس فنی و مهندسی، گروه مهندسی کامپیوتر، دانشگاه یزد، یزد، (email: m.ghasemzadeh@yazd.ac.ir).

علی‌محمد زارع بیدکی، پردیس فنی و مهندسی، گروه مهندسی کامپیوتر، دانشگاه یزد، یزد، (email: alizareh@yazd.ac.ir).

این مشکل به مسئله سربار اطلاعات زاید معروف است. موتورهای جستجو علایق کاربران را که همان پرس‌وجوهایشان است در لاگ‌های جستجو ذخیره می‌کنند. کاربردهای مختلف وب با استفاده از تجزیه و تحلیل لاگ‌های جستجو و تکنیک‌های کاوش در وب، فعالیت‌های ناوبری کاربران را پیش‌بینی می‌کنند اما برای بهینه‌سازی نتایج موتور جستجو کمتر استفاده می‌شوند [۶] تا [۸].

از طرف دیگر، لاگ‌های جستجو یک منبع داده‌ای عظیم هستند و بنابراین برای تجزیه و تحلیل لاگ‌های جستجو نیاز به روش‌های تجزیه و تحلیل داده‌های عظیم است. در سال‌های اخیر، جستجوی حجم زیاد داده‌ها و تجزیه و تحلیل آنها در حال رشد هستند و بنابراین یک زمینه تحقیقاتی جدید به نام کاوش داده‌های عظیم به وجود آمده است. محققان روش‌ها و الگوریتم‌ها را برای استخراج دانش از این نوع داده‌ها مورد بررسی قرار می‌دهند [۹].

کاربران برای جستجو در وب، پرس‌وجوهای کوتاه و مبهم را استفاده می‌کنند. این نقص می‌تواند به کمک پیشنهاد یا بازنویسی پرس‌وجو کاهش یابد. پیشنهاد پرس‌وجو به کاربران می‌تواند توسط آنالیز لاگ پرس‌وجو و با استفاده از مکانیزم‌های کلاسه‌بندی و خوشه‌بندی انجام شود. به طور معمول لاگ پرس‌وجوی یک موتور جستجو حجیم است و آنالیز آن زمان زیادی طول می‌کشد [۱۰].

مسئله اصلی این پژوهش، پیشنهاد پرس‌وجوی مناسب به کاربر نوجوان است تا بر اساس آن بتواند به مطالب با کیفیت بالا توسط موتور جستجو دسترسی پیدا کند. به دلیل این که نویز در ناوبری موضوعی کلیک‌های کاربران نوجوان نسبت به بزرگ‌سالان بیشتر است، برای ساخت مدل رفتار کاربر تا نویز ناوبری موضوعی کلیک‌های کاربران نوجوان را با دقت بیشتر فیلتر کند، الگوهای موضوعی کلیک‌های کاربران نوجوان از لاگ جستجوی AOL استخراج می‌گردد. نویز در ناوبری موضوعی کلیک‌های کاربران نوجوان به این دلیل است که کاربران نوجوان اغلب موقع واردکردن یک پرس‌وجو، ویژگی‌های مورد نیاز خود را به طور ناقص وارد می‌کنند. این موضوع اغلب منجر به دریافت اطلاعات نامربوط می‌شود که در بعضی مواقع رهاسازی فرایند جستجو توسط آنها را به دنبال دارد. روش‌های پیشنهاد پرس‌وجو برای کمک به کاربران، پرس‌وجوهای کامل‌تر و مرتبط با پرس‌وجوی اولیه را مهیا می‌سازند تا آنها فرایند جستجوی خود را بر اساس این پیشنهادها تصحیح نمایند. روش پیشنهاد پرس‌وجو موجب می‌گردد که شانس کاربر برای دسترسی به مطالب مرتبط با پرس‌وجو بهبود یابد. در این رابطه یک روش پیشنهاد پرس‌وجو بر اساس الگوهای موضوعی رفتار جستجوی کاربران نوجوان مبتنی بر لاگ موتور جستجوی AOL پیشنهاد می‌گردد.

در این مقاله از روش آنالیز و کاوش رفتار جستجوی کاربر مبتنی بر لاگ جستجو AOL استفاده می‌گردد. با تطابق URLهای کلیک‌شده در لاگ جستجوی AOL با دامنه‌های در بخش "Kids and Teen" مربوط به Dmoz، جستجوهای مربوط به نوجوانان از لاگ جستجوی AOL استخراج می‌گردد. در Dmoz، URLهای مربوط به نوجوانان برای رده سنی ۱۳ تا ۱۵ سال هستند. سپس از رسانه اجتماعی Delicious تگ‌های مرتبط با موضوع‌های جستجوی کاربران نوجوان استخراج شده و الگوی موضوعی از URLهای کلیک‌شده در تراکنش‌های استخراج‌شده از لاگ جستجوی AOL کشف می‌شود. سرانجام از آنها برای پیشنهاد پرس‌وجو به کاربران نوجوان و بهبود نتایج جستجو استفاده می‌گردد. برای این کار با استفاده از کلاسه‌بندی دودویی نزدیک‌ترین پرس‌وجو به پرس‌وجوی مورد نظر کاربر نوجوان مشخص می‌گردد و به کاربر نوجوان پیشنهاد می‌شود.

۲- سابقه تحقیق

تحقیقات گسترده‌ای در رابطه با ارائه مناسب‌ترین پاسخ به پرس‌وجوی کاربران در موتورهای جستجو صورت گرفته که در اینجا چند مورد که ارتباط بیشتری با این پژوهش دارند معرفی می‌گردد.

تعدادی از پژوهش‌های انجام‌شده، روش‌هایی را به کمک لاگ‌های پرس‌وجوی موتور جستجوی وب برای کاوش پرس‌وجوهای مرتبط پیشنهاد کرده‌اند. خوشه‌بندی پرس‌وجوهای مشابه بر اساس URLهای در لاگ جستجوی یک موتور جستجو روش پیشنهادشده‌ای است که از چهار محاسبه فاصله مختلف استفاده می‌کند. این چهار محاسبه مبتنی بر ۱) کلمات کلیدی، ۲) تطابق رشته‌ای کلمات کلیدی، ۳) URLهای کلیک‌شده مشترک و ۴) فاصله سندهای کلیک انجام شده‌اند. مسئله اصلی این است که کدام نشست از پرس‌وجوها متعلق به یک فرایند جستجو است. مسئله بعدی کشف مورد علاقه‌ترین پرس‌وجوهایی است که توسط کاربران مختلف ارسال شده‌اند [۱۱].

برای تجزیه و تحلیل لاگ‌های پرس‌وجوی موتورهای جستجوی تجاری در مقیاس بزرگ، چند پژوهش انجام شده است. اسیلورستین و همکارانش در سال ۱۹۹۹ یک تجزیه و تحلیل بر روی لاگ پرس‌وجوی موتور جستجوی آلتاویستا شامل حدود یک میلیون مدخل انجام دادند. این تجزیه و تحلیل بر اساس نشست‌های پرس‌وجو و وابستگی عبارت‌های پرس‌وجو مبتنی بر یک مجموعه اندازه‌گیری‌های توصیفی مانند طول پرس‌وجو، تکرار پرس‌وجو، طول نشست و تکرار واژه انجام شده است. طبق نتایج این پژوهش مشخص گردید که کاربران از پرس‌وجوهای کوتاه (میانگین ۲٫۳ کلمه در پرس‌وجو) استفاده می‌کنند و نشست‌های کاربر نیز کوتاهند (میانگین ۲ پرس‌وجو در هر نشست). اغلب کاربران پرس‌وجوها را تغییر نمی‌دهند و ۷۷٫۵٪ پرس‌وجوها منحصر به فرد هستند که اشاره به تنوع گسترده‌ای از نیازهای اطلاعاتی دارند [۱۲].

همچنین در پژوهش دیگری اسپینک و همکارانش در سال ۲۰۰۱ نتایج مشابهی راجع به طول پرس‌وجو و مشخصات پرس‌وجو بر اساس لاگ پرس‌وجوی موتور جستجوی Exite گزارش دادند [۱۳].

جنبه‌های مختلفی از یک لاگ پرس‌وجوی AOL مانند الگوهای فرمول‌بندی پرس‌وجو، کارامدی موتور جستجو، مشخصات جمعیتی کاربر و تعاملات کاربر طی پژوهشی توسط پاس و همکارانش در سال ۲۰۰۶ مورد تجزیه و تحلیل قرار گرفت. آنها فضای پرس‌وجوی وسیع، نوعاً متنوع و ثابت را استفاده کردند. به عنوان نتیجه نشان دادند که ۲۰٪ کاربران تقریباً ۷۰٪ پرس‌وجوها را انجام می‌دهند و کمتر از ۱٪ حساب دامنه‌های وب برای ۵۰٪ کلیک‌های کاربران است [۱۴].

تجزیه و تحلیل‌های دیگری نیز روی همین لاگ پرس‌وجو بر اساس دسته‌بندی پرس‌وجوها و نشست‌ها مبتنی بر محبوبیت پرس‌وجوها انجام شده و رفتارهای مختلفی (مثل ضریب ناوبری، طول پرس‌وجو و طول زمان) مورد بررسی قرار گرفتند. تعریف مختلفی از نشست کاربر بر روی لاگ‌های پرس‌وجو در این پژوهش‌ها ارائه داده‌اند. در حالت کلی، یک نشست، دنباله‌ای از پرس‌وجوهای صادرشده برای ارضای یک نیاز اطلاعاتی است. همچنین نشست‌های جستجو با استفاده از یک زمان توقف پرس‌وجوها نیز تعریف شده‌اند. بر طبق این تعریف، دو پرس‌وجو در یک نشست هستند اگر اختلاف زمانی در آنها کوچک‌تر از یک مقدار آستانه داده‌شده باشد [۱۵] و [۱۶].

رفتار کاربر بر اساس داده‌های نوار ابزار و جستجوهای مبتنی بر سیستم یاهو توسط کومار و همکارانش و همچنین توسط چنگ و همکارانش در سال ۲۰۱۰ مورد بررسی قرار گرفته است. در این بررسی‌ها از نشست‌های

جستجوی کاربران بزرگ‌سال استفاده کرده‌اند و به طور کلی بر اساس ویژگی سن، تجزیه و تحلیل را انجام داده‌اند. همچنین نماهای صفحه در نشست‌های جستجو را بر اساس نوع محتوا (مانند بازی‌های خبری، پورتال)، نوع ارتباط (مانند ایمیل، شبکه‌بندی اجتماعی) و نوع جستجو (مانند وب، چندرسانه‌ای) طبقه‌بندی کردند. به عنوان نتیجه مشخص نمودند که حدود نیمی از نماهای صفحه متعلق به طبقه محتوایی، یک‌سوم متعلق به طبقه ارتباطی و یک‌ششم متعلق به طبقه جستجویی هستند [۱۷] و [۱۸].

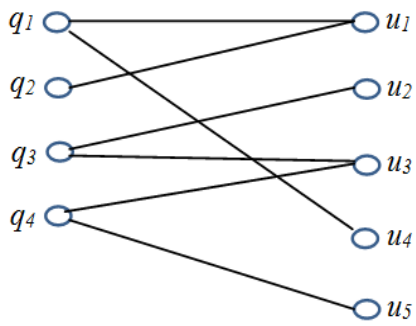
تورِس و همکارانش در سال ۲۰۱۴ در زمینه پیشنهاد پرس‌وجو به کاربران کودک پژوهش کردند. مسئله مورد نظرشان این بود که کودکان از مجموعه واژگان محدود برای کلمات کلیدی پرس‌وجوی خود استفاده می‌کنند. همچنین کودکان مشکل انتخاب کلمات کلیدی صحیح برای پرس‌وجو را نیز دارند. برای ایجاد پیشنهاد پرس‌وجو از تگ‌های موضوعی رسانه اجتماعی مرتبط با کودکان و انتخاب کلمات کلیدی مرتبط و صحیح استفاده کردند. همچنین یک قدم‌زدن تصادفی منحرف‌شده بر اساس گراف دوبخشی از منابع وب و تگ‌ها را پیشنهاد دادند. به عنوان نتیجه نشان دادند که روش پیشنهادی آنها برای پرس‌وجوهای رده سنی مابین ۱۰ الی ۱۲ سال عملکرد بهتری نسبت به موتورهای جستجوی فعلی دارد و رسانه اجتماعی منبع باارزشی برای تولید پیشنهادهای پرس‌وجو است. همچنین پیشنهادهای پرس‌وجو را برای رده‌های سنی مختلفی از کاربران با استفاده از منابع مطمئن وب و یک قدم‌زدن تصادفی منحرف‌شده ایجاد کردند. نشان دادند که روش پیشنهادی آنها موجب بهبود جستجوی کودکان می‌شود [۱۹] و [۲۰].

در تحقیق دیگری که تورِس و همکارانش در سال ۲۰۱۴ انجام دادند، مشابه با پژوهش قبلی برای ایجاد پیشنهاد پرس‌وجو از تگ‌های موضوعی رسانه اجتماعی مرتبط با کودکان و انتخاب کلمات کلیدی مرتبط و صحیح استفاده کردند. همچنین یک قدم‌زدن تصادفی منحرف‌شده بر اساس گراف دوبخشی از منابع وب و تگ‌ها را پیشنهاد دادند. کار دیگری که انجام دادند این بود که کیفیت رتبه‌بندی تگ‌ها را مورد توجه قرار دادند و توسط ترکیب آن با ویژگی‌های موضوعی و مدل‌سازی زبان استفاده‌شده در پرس‌وجوی کودکان، رتبه‌بندی تگ‌ها را بهبود بخشیدند. به عنوان نتیجه نشان دادند که روش پیشنهادی آنها برای پرس‌وجوهای رده سنی ۸ الی ۹ سال عملکرد بهتری نسبت به موتورهای جستجوی فعلی دارد [۲۱].

وانگ و همکارانش برای متمایزنمودن پرس‌وجوهای اطلاعاتی و مبهم از میانگین آنتروپی کلیک‌ها استفاده کردند. همچنین میانگین آنتروپی کلیک‌ها را بر روی توزیع کلیک‌های هر کاربر محاسبه نمودند. در شکل ۱ گراف دوبخشی کلیک‌ها و پرس‌وجوهای کاربران نشان داده شده است. بر اساس این گراف دوبخشی، دو پرس‌وجو با میانگین آنتروپی مشابه می‌تواند محاسبه گردد [۲۲].

هیوژانگ دوان و همکارانش، یک الگوی کلیک برای مدل‌سازی رفتار جستجوی کاربر معرفی کردند و نشان دادند که روش پیشنهادی آنها موجب بهبود پیشنهاد پرس‌وجو به کاربران می‌شود [۲۳].

برای مثال پرس‌وجوی "cars" را در نظر بگیرید. سیستم پیشنهاد پرس‌وجوی موتور جستجوی Google پرس‌وجوهای "car rentals"، "cars for sals"، "used cars"، "new car"، "Disney cars" و "car pictures" را به کاربر پیشنهاد می‌دهد. در حالی که پیشنهادهای پرس‌وجویی که نیازهای اطلاعاتی کودکان و نوجوانان را برآورده می‌سازند شامل "car games"، "car toys"، "car movies"، "car images" و "car coloring pages" هستند [۲۱].



شکل ۱: گراف دوبخشی از پرس‌وجوها و کلیک‌ها.

پژوهش‌هایی که بررسی شدند در مورد استفاده از محتوای پرس‌وجوهای گذشته کاربران برای پیشنهاد پرس‌وجو به کاربران رده سنی کودکان هستند. نقطه ضعف آن پژوهش‌ها در مورد عدم استفاده از بازخورد کاربر برای پیشنهاد پرس‌وجو به کاربران کودک است بدین معنا که می‌توان از الگوی ناوبری کلیک‌های کاربران برای پیشنهاد پرس‌وجو به کاربران رده سنی کودکان و نوجوانان استفاده نمود. همچنین برای پیشنهاد پرس‌وجوی با دقت بالاتر به کاربران نوجوان می‌توان از الگوی موضوعی کلیک‌های کاربران بهره برد.

فرضیه در این پژوهش آن است که تگ‌های با تکرار بیشتر و مربوط به URLهای مرتبط با موضوع‌های مورد جستجوی کاربران نوجوان، نامزد بهتر برای تولید پیشنهاد پرس‌وجو برای آنها هستند.

در روش پیشنهادی بر اساس الگوی موضوعی رفتار جستجوی کاربران نوجوان پیشنهاد پرس‌وجو صورت می‌گیرد. الگوی پیشنهادی مربوط به رده سنی خاصی از کاربران یعنی نوجوانان و مبتنی بر موضوع کلیک‌های کاربران نوجوان است که انحراف بیشتری در فرایند ناوبری کلیک‌ها بر اساس موضوع‌ها نسبت به بزرگ‌سالان دارند. سپس از این الگو برای پیشنهاد پرس‌وجو به کاربران نوجوان استفاده می‌شود.

۳- روش پیشنهادی

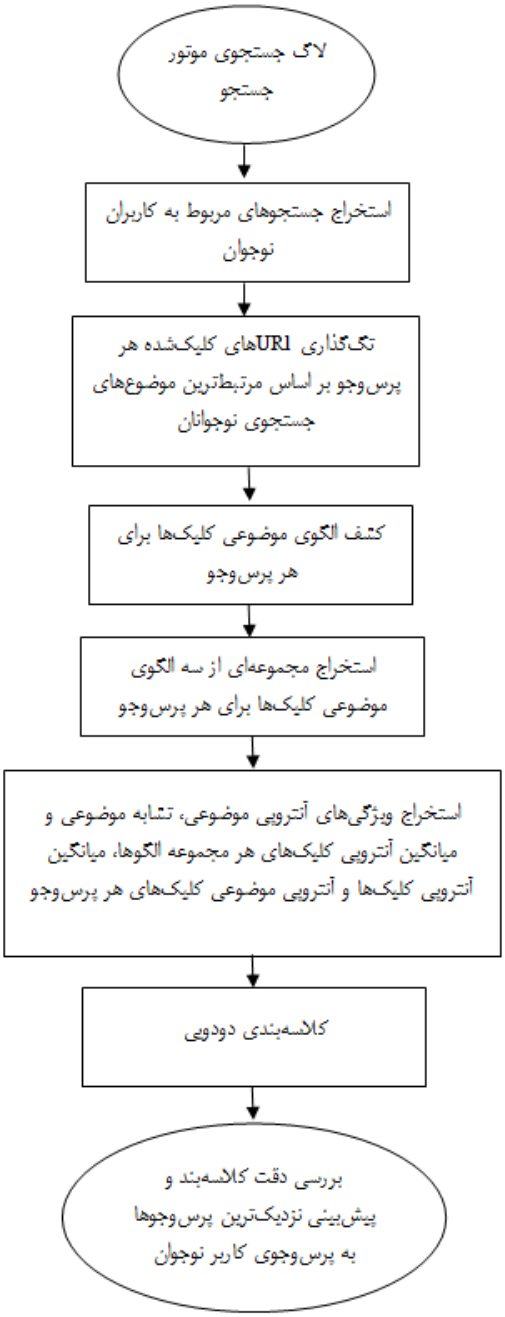
در این پژوهش استفاده از کشف الگوی موضوعیِ رفتار جستجوی کاربر نوجوان برای پیشنهاد پرس‌وجو به کاربر نوجوان که نزدیک‌ترین پرس‌وجو به پرس‌وجوی کاربر است، پیشنهاد می‌شود. به طور کلی از دو طریق می‌توان برای یافتن پرس‌وجوهای مشابه استفاده نمود:

- **استفاده از محتویات پرس‌وجو:** اگر دو پرس‌وجویی که شامل ترم‌های یکسان یا مشابه هستند دلالت به نیازهای اطلاعاتی یکسان یا مشابه دارند.

این روش برای پرس‌وجوهای طولانی قابل اعتماد است در حالی که کاربران اغلب پرس‌وجوهای کوتاه به موتورهای جستجو ارسال می‌کنند. در این حالت اطلاعات کافی برای نیازهای اطلاعاتی کاربر تأمین نمی‌شود. بنابراین معیار دوم به عنوان تکمیل‌کننده معیار اول استفاده شده است. معیار دوم مشابه با بینش تحت خوشه‌بندی اسناد در IR است. اعتقاد بر آن است که نزدیکی ارتباط اسناد منطبق با یکسانی پرس‌وجوها است. این بینش به صورت روش معکوس بینش اول استفاده می‌شود.

- **استفاده از بازخورد کاربر:** اگر دو پرس‌وجو بر اساس کلیک کاربر بر روی اسناد در نهایت به انتخاب سند یکسان منجر شود آن گاه آن دو پرس‌وجو مشابه هستند.

کلیک‌های روی سند قابل مقایسه با بازخورد کاربر در محیط IR سنتی هستند. این دو معیار مزایای خودشان را دارند. با استفاده از معیار اولی می‌توان پرس‌وجوهای با اجزای مشابه با همدیگر را گروه‌بندی نمود. با



شکل ۲: روندنمای روش پیشنهادی.

استفاده از معیار دوم می‌توان از داوری کاربر سود برد. همچنین معیار دوم برای خوشه‌بندی پرس‌وجوهای کاربر نیز استفاده شده است [۲۴].

به طور کلی اندازه‌گیری‌های مبتنی بر محتوا تمایل به تشخیص ترم‌های یکسان یا مشابه دارند. اندازه‌گیری‌های مبتنی بر بازخورد کاربر تمایل به تشخیص پرس‌وجوهای مرتبط با موضوع‌های یکسان یا مشابه دارند [۱۱].

معمولاً پژوهشگران از وابستگی ذاتی اسناد و پرس‌وجوها در یک پایگاه داده تراکنشی موتور جستجو بهره‌برداری نکرده‌اند در حالی که فاصله دو سند می‌تواند بدون آزمایش محتوای اسناد ارزیابی شود. این ویژگی content-ignorance نامیده شده است [۲۴]. مسئله‌ای که در معیار دوم وجود دارد آن است که رفتار هر کاربر حتی با نتایج جستجوی یکسان، متفاوت است. بنابراین می‌توان نتیجه گرفت که رفتار موضوعی کلیک‌های هر کاربر برای هر پرس‌وجو دارای نویز ولی منطبق با یک مجموعه الگوهای رفتاری مشترک است. راه حل پیشنهادی این است که می‌توان به وسیله شناسایی الگوهای مشترک در رفتارهای کاربر، نویز رفتارهای کاربر را فیلتر نمود و یک مدل کاربر با دقت بالاتری را به دست آورد. سپس در کاربردهای عملی مانند شناسایی پرس‌وجوهای مشابه برای پیشنهاد پرس‌وجو استفاده نمود.

شکل ۲ روندنمای کلی روش پیشنهادی را برای کشف الگوی موضوعی رفتار کاربران نوجوان و استفاده آن برای پیشنهاد پرس‌وجو نشان می‌دهد.

در مرحله اول، جستجوهای مربوط به کاربران نوجوان از لاگ موتور جستجوی AOL استخراج می‌گردد و سپس آنها تگ‌گذاری موضوعی می‌شوند. برای این کار از رسانه اجتماعی Delicious تگ‌های موضوعی مرتبط با URLهای کاربران نوجوان استخراج می‌شود. برای این منظور، منطبق بر پژوهش تورس و همکارانش [۲۱] با استفاده از کلیک‌های مربوط به دایرکتوری نوجوانان Dmoz و تطابق آن با URLهای مجموعه داده Del.icio.us تگ‌های موضوعی مربوط به نوجوانان استخراج می‌گردد. همچنین با تطابق کلیک‌های مربوط به دایرکتوری نوجوانان Dmoz با کلیک‌های لاگ جستجوی AOL تراکنش‌های جستجوی مربوط به کاربران نوجوان استخراج شده و هر کلیک مربوط به یک پرس‌وجو تگ‌گذاری موضوعی می‌شود.

در مرحله بعدی با استفاده از تگ‌های موضوعی استخراج‌شده از مجموعه داده Del.icio.us و لاگ جستجوی AOL الگوی موضوعی کلیک‌ها از URLهای کلیک‌شده در تراکنش‌های استخراج‌شده از لاگ جستجوی AOL کشف می‌شود.

در مرحله آخر با استفاده از کلاسه‌بندی دودویی نزدیک‌ترین پرس‌وجو به پرس‌وجوی مورد نظر کاربر نوجوان مشخص می‌گردد. برای این کار از جستجوهای استخراج‌شده از لاگ جستجوی AOL، ویژگی‌های آنتروپی موضوعی، تشابه موضوعی و میانگین آنتروپی کلیک‌های هر مجموعه الگوها، کلیک‌های هر پرس‌وجو استخراج می‌گردد.

در نهایت کلاسه‌بندی دودویی به روش نزدیک‌ترین همسایه (KNN) بر اساس این ویژگی‌های استخراج‌شده و ویژگی‌های میانگین آنتروپی کلیک‌ها، آنتروپی موضوعی کلیک‌ها، محبوبیت و طول پرس‌وجوها انجام و بر این اساس یک کلاسه‌بند دودویی با دقت بالاتر پیشنهاد می‌گردد.

۳-۱ الگوی موضوعی کلیک‌های کاربر

هر کاربر با نتایج جستجوی یکسان رفتار جستجوی متفاوتی را دارد. فرضیه بر آن است که رفتارهای نویزی بر اساس موضوع کلیک‌های هر جستجو منطبق بر یک مدل رفتاری پایه‌ای است که کاربران از یک مجموعه الگوهای رفتاری مشترک تبعیت می‌کنند. با شناسایی الگوهای مشترک در رفتارهای کلیک کاربر بر اساس موضوع مورد جستجو می‌توان نویزها را در رفتارهای کاربر فیلتر نمود و یک مدل کاربر دقیق به دست آورد و سپس برای شناسایی پرس‌وجوهای مشابه و پیشنهاد پرس‌وجو استفاده شود.

برای استخراج تگ‌های موضوعی مرتبط URLهای نوجوانان از یک گراف دوبخشی مربوط به مجموعه داده Del.icio.us از رسانه اجتماعی Delicious استفاده می‌شود.

- **تعریف ۱ (گراف دوبخشی از URLها و تگ‌ها):** گراف $G=(U,T,E)$ یک گراف دوبخشی از مجموعه داده Del.icio.us است که $U=\{u_1,u_2,\ldots,u_n\}$ مجموعه‌ای از URLها، $T=\{t_1,t_2,\ldots,t_m\}$ مجموعه‌ای از تگ‌ها و مجموعه $E=\{(u,t)\,|\,(u,t)\in U\times T\}$ مجموعه‌ای از یال‌های گراف هستند.

مجموعه U_t مجموعه URLهای استخراج‌شده مرتبط با نوجوانان است که به صورت زیر تعریف می‌شود:

- **تعریف ۲ (URLهای مرتبط با نوجوانان):** $U_t = \{u_1, u_2, \ldots, u_k\}$، مجموعه URLهای استخراج‌شده از دایرکتوری نوجوانان Dmozاست.

و T_t مجموعه تگ‌های موضوعی مرتبط با URLهای نوجوان را نشان می‌دهد که به صورت زیر تعریف می‌شود:

- **تعریف ۳ (تگ‌های موضوعی مرتبط با URLهای نوجوانان):** مجموعه $T_t = \{t_1, t_2, \ldots, t_k \,|\, URL(t_i) \in U_t\}$، مجموعه تگ‌های موضوعی مرتبط با URLهای نوجوانان است به طوری که U_t مجموعه URLهای مرتبط با نوجوان را شامل می‌شود و $URL(t_i)$ نشان‌دهنده یک URL است که توسط t_i در گراف دوبخشی G تگ شده است.

و سرانجام الگوی موضوعی پیشنهادی که الگوی موضوعی از کلیک‌ها را مشخص می‌سازد به صورت زیر تعریف می‌گردد:

- **تعریف ٤ (الگوی موضوعی کلیک‌ها):** یک پرس‌وجوی q، مجموعه اسناد کلیک‌شده D_q و مجموعه موضوع‌های T داده شده است آن گاه الگوی موضوعی کلیک‌های P_q، یک توزیع کلیک‌ها بر اساس موضوع $t \in T$ بر روی D_q است. این الگو یک توزیع احتمالی کلیک‌ها بر اساس مرتبط‌ترین موضوع‌ها از مجموعه T در مجموعه اسناد کلیک‌شده D_q به واسطه پرس‌وجوی q را نشان می‌دهد

$$P_q = \{(p(d_q), rel(d_q,t)) \,|\, t \in T,\ d_q \in D_q,\ \sum_{d_q \in D_q} p(d_q) = 1, rel(d_q,t) = TF.IDF(d_q,t)\} \quad (1)$$

که $p(d_q)$ احتمال کلیک بر روی سند d به واسطه پرس‌وجوی q است

$$p(d_q) = c_d^{(q)} . (\sum_{d \in D_q} c_d^{(q)})^{-1} \quad (2)$$

$c_d^{(q)}$ تعداد کلیک‌های بر روی سند d بر اساس پرس‌وجوی q است.

یک الگوی موضوعی کلیک‌های کاربران نوجوان، توسط سه عنصر چهارتایی نمایش داده می‌شود که نشان‌دهنده کلیک‌های بر روی سند با مرتبط‌ترین موضوع‌ها است. بنابراین هر الگوی موضوعی با یک لیست مرتبی از سه عنصر چهارتایی نوشته می‌شود

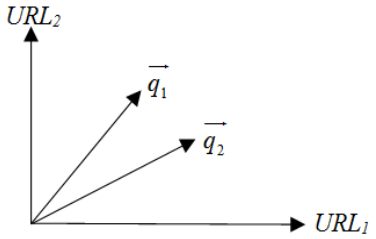
$$P_q = \{(u_1, t_1, p(u_1), rel(u_1,t_1)), (u_2,t_2, p(u_2), rel(u_2,t_2)), (u_3,t_3, p(u_3), rel(u_3,t_3))\} \quad (3)$$

که $p(u_i)$ نشان‌دهنده احتمال URLی است که بیشترین مرتبه کلیک را به واسطه پرس‌وجوی q در تراکنش‌های استخراج‌شده از لاگ AOL داشته است به طوری که $p(u_1) > p(u_2) > p(u_3)$ است.

$rel(u_i,t_i)$ نشان‌دهنده مرتبط‌ترین موضوع $t_i \in T$ از نظر امتیاز TF.IDF در سند $d_i \in D_q$ مربوط به URL کلیک‌شده u_i است.

- **تعریف ۵ (مجموعه الگوهای موضوعی):** مجموعه PS_q شامل سه الگو از مرتبط‌ترین موضوع‌های متفاوت نسبت به اسناد کلیک‌شده است که رفتار جستجوی موضوعی کاربر بر اساس پرس‌وجوی q را مشخص می‌سازد

$$PS_q = \{P_q, P'_q, P''_q\} \quad (4)$$



شکل ۳: نمایش برداری پرس‌وجوها در فضایی از URLها.

۳-۲ تشابه الگوهای موضوعی کلیک‌ها

یکی از ویژگی‌هایی که برای پیشنهاد پرس‌وجو استفاده می‌شود تشابه موضوعی مجموعه الگوهای کشف‌شده PS_q است. استفاده از این ویژگی موجب می‌شود پیشنهاد پرس‌وجو به پرس‌وجوی اصلی کاربر نوجوان از نظر تشابه موضوعی نزدیک‌تر شود. این تشابه بر اساس تشابه Jaccard و بر طبق (۵) محاسبه می‌گردد

$$Sim_T(PS_q) = \frac{\bigcap_T (P_q, P'_q, P''_q)}{\bigcup_T (P_q, P'_q, P''_q)} \quad (5)$$

$\bigcap_T$ و $\bigcup_T$ عملیات اشتراک و اجتماع موضوعی بر اساس موضوع‌های در مجموعه T را نشان می‌دهند. ویژگی دیگری که استخراج می‌شود تشابه دو مجموعه از الگوهای موضوعی کلیک‌ها است که تشابه دو پرس‌وجو از نظر الگوهای موضوعی کشف‌شده محاسبه می‌گردد.

برای استخراج این ویژگی از یک گراف دوبخشی پرس‌وجوها و URLها مشابه شکل ۱ استفاده می‌شود. گراف $G(V,E)$ یک گراف دوبخشی در میان گره‌های Q و U است به طوری که $Q \cup U = V$ و $Q \cap U = \varnothing$ و هر یال در E یک گره در Q و یک گره در U است.

هر پرس‌وجو به عنوان یک بردار نمایش داده می‌شود که عنصر iام به ارتباط مابین پرس‌وجو و URL i دلالت دارد. شکل ۳ چگونگی نمایش برداری پرس‌وجوها را نشان می‌دهد.

بردار پرس‌وجو در (۶) نشان داده شده که $rel(q_k,u_i)$ به میزان ارتباط مابین پرس‌وجو q_k با URL u_i دلالت دارد [۲۵]

$$\overrightarrow{q_k} = [rel(q_k,u_1), rel(q_k,u_2), \ldots, rel(q_k,u_m)] \quad (6)$$

در این پژوهش، بردار پرس‌وجوی (۷) با جایگزینی الگوی موضوعی کلیک‌های (۳) و مجموعه الگوهای موضوعی (۴) در (۶) نمایش داده می‌شود و برای محاسبه تشابه مابین پرس‌وجوها مورد استفاده قرار می‌گیرد

$$\overrightarrow{q_k} = [rel(u_1,t_1), rel(u_2,t_2), rel(u_2,t_2), rel(u_1,t_4), rel(u_2,t_5), rel(u_3,t_6), rel(u_1,t_7), rel(u_2,t_8), rel(u_3,t_9)] \quad (7)$$

$rel(u_i,t_j)$ میزان ارتباط تگ t_j نسبت به URL کلیک‌شده u_i به واسطه پرس‌وجوی q را مشخص می‌سازد.

حال مجموعه S_q را به عنوان مجموعه الگوهای موضوعی کلیک‌های مشترک مربوط به پرس‌وجوی q طبق (۸) تعریف می‌کنیم

$$S_q = \{(PS_{q'}, Sim(PS_{q'}, PS_q)) \,|\, PS_{q'}, PS_q \text{ are Topic Click Patterns of } q' \text{ and } q\} \quad (8)$$

S_q یک مجموعه از تمام مجموعه الگوهای موضوعی مانند $PS_{q'}$ مربوط به پرس‌وجوی q' است که مشابه با مجموعه الگوهای موضوعی PS_q مربوط به پرس‌وجوی q می‌باشد. این تشابه نشان می‌دهد که یک

مجموعه الگوهای موضوعی $PS_{q'}$ به چه میزان یک مجموعه الگوهای موضوعی PS_q در مجموعه D_q را تبعیت می‌کند. در حالت کلی تابع تشابه در (۹) تعریف می‌شود

$$Sim(PS_{q'}, PS_q) = ۱ - Dis(PS_{q'}, PS_q) \quad (۹)$$

تابع Dis فاصله دو مجموعه الگوهای موضوعی را محاسبه می‌کند که هر تابع فاصله‌ای می‌تواند استفاده شود. در روش پیشنهادی برای محاسبه تشابه موضوعی دو مجموعه الگوهای موضوعی از اندازه‌گیری تشابه cosine دو بردار بر طبق (۱۰) استفاده می‌شود. هر مجموعه الگوهای موضوعی به عنوان برداری از موضوع‌های در آن الگو نمایش داده می‌شود

$$Sim_T(PS_{q'}, PS_q) = cosine(PS_{q'}, PS_q) = \frac{\overrightarrow{PS_{q'}}.\overrightarrow{PS_q}}{\left|\overrightarrow{PS_{q'}}\right|.\left|\overrightarrow{PS_q}\right|} \quad (۱۰)$$

۳-۳ آنتروپی الگوهای موضوعی کلیک‌ها

در پژوهش‌های انجام‌شده، آنتروپی کلیک به عنوان آنتروپی اطلاعاتی توزیع کلیک‌های کاربر محاسبه می‌گردد. به طور رسمی، پرس‌وجوی q و مجموعه اسناد کلیک‌شده D_q داده شده‌اند، آنتروپی کلیک مربوط به پرس‌وجوی q توسط (۱۱) محاسبه می‌شود

$$ClickEntropy(q) = -\sum_{d \in D_q} p(d|q) \log p(d|q) \quad (۱۱)$$

که $p(d|q)$ احتمال کلیک‌های روی سند d در میان تمام کلیک‌های انجام‌شده بر اساس q است.

آنتروپی کلیک، میزان ابهام پرس‌وجوی کاربر را محاسبه می‌کند. برای پرس‌وجوهای ناوبری که مقصود جستجو آشکار می‌باشد آنتروپی کلیک مقدار پایینی را شامل می‌شود در حالی که برای پرس‌وجوهای اطلاعاتی با مقصودهای مبهم، آنتروپی کلیک مقدار بالایی را تحویل می‌دهد. پرس‌وجوی اطلاعاتی به پرس‌وجویی گفته می‌شود که شامل تعدادی کلمات کلیدی است و بر اساس آن ناوبری کلیک‌های کاربر روی تعدادی از اسناد برای دستیابی به اطلاعات مورد نیاز خود انجام می‌شود. در نتیجه هرچه تعداد کلیک‌های کاربر بیشتر باشد به معنای ابهام بیشتر پرس‌وجو است در حالی که پرس‌وجوی ناوبری به پرس‌وجویی گفته می‌شود که شامل یک URL است و سندی دارای ارتباط بسیار بالا با پرس‌وجوی ناوبری در اختیار کاربر قرار می‌گیرد. در نتیجه کاربر، ناوبری کلیک‌های بسیار کمی را برای دستیابی به اطلاعات مورد نیاز خود انجام می‌دهد و آنتروپی کلیک مقدار پایینی را شامل می‌شود.

حال آنتروپی الگوی موضوعی کلیک را به عنوان یک ویژگی کلاسه‌بندی دودویی پیشنهاد می‌کنیم. در این پژوهش، آنتروپی الگوی موضوعی کلیک، آنتروپی اطلاعاتی از توزیع الگوی موضوع‌های کلیک کاربران نوجوان را نشان می‌دهد. هرچه این آنتروپی کمتر باشد ابهام موضوعی پرس‌وجوی پیشنهادی به کاربر نوجوان نیز کمتر است. به طور رسمی، پرس‌وجوی q و مجموعه الگوهای موضوعی کلیک PS_q داده شده‌اند، آنتروپی مجموعه الگوهای موضوعی کلیک‌ها توسط (۱۲) محاسبه می‌گردد. معادله (۱۲) با جایگزینی الگوهای موضوعی کلیک‌های (۳) و (۴) در (۱۱) به دست می‌آید

$$TP\_Entropy(q) = -\sum_{P_q \in PS_q} rel(u,t).\log(rel(u,t)) \quad (۱۲)$$

$rel(u,t)$ امتیاز TF.IDF موضوع t در سند d مربوط به یک الگو P_q در مجموعه PS_q را مشخص می‌سازد. این آنتروپی اطلاعاتی توزیع الگوی موضوعی کلیک‌های کاربران نوجوان برای پرس‌وجوها را نشان می‌دهد. هدف اصلی کاهش ابهام موضوعی پرس‌وجوهای کاربران نوجوان است که مقدار کمتر آنتروپی الگو نشان‌دهنده ابهام موضوعی کمتر پرس‌وجو است.

برای پیشنهاد پرس‌وجو از نزدیک‌ترین پرس‌وجو در لاگ نسبت به پرس‌وجوی کاربر استفاده می‌شود. برای این منظور کوتاه‌ترین فاصله مابین پرس‌وجوی کاربر و پرس‌وجوهای در لاگ محاسبه می‌گردد. میزان پشتیبانی یک پرس‌وجو توسط محبوبیت پرس‌وجو در لاگ پرس‌وجو اندازه‌گیری می‌شود. از الگوی موضوعی کلیک‌های کاربر برای اندازه‌گیری تشابه مابین پرس‌وجوها بر طبق (۱۰) استفاده می‌شود و ماکسیمم تشابه مابین دو الگو در نظر گرفته می‌شود.

٤- پیاده‌سازی و اجرا

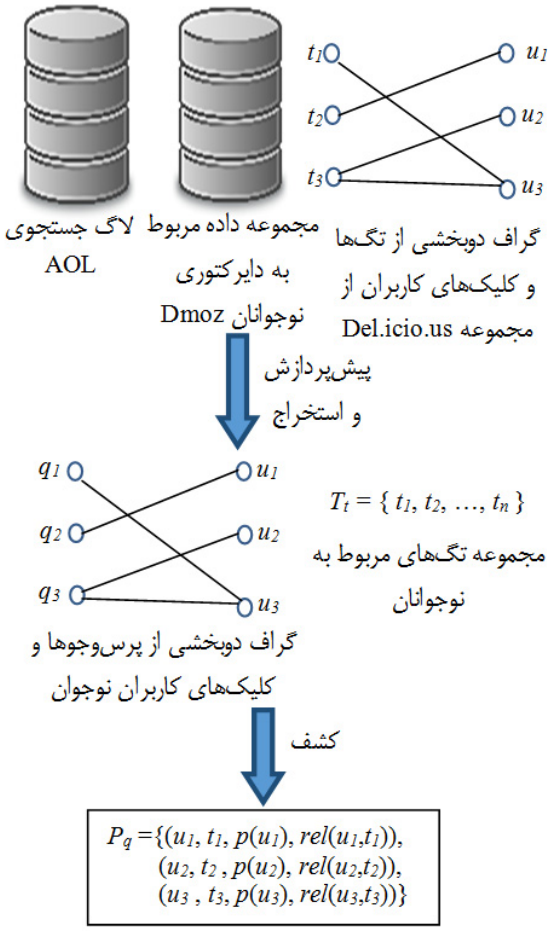
روش پیشنهادی بر روی یک میکروکامپیوتر با ۴ گیگابایت حافظه اصلی و پردازنده اینتل ۱٫۸ گیگاهرتز با به کارگیری ابزار Alteryx [۲۶]، برای کشف الگوی موضوعی کلیک کاربران نوجوان و ابزار Weka [۲۷]، برای کلاسه‌بندی دودویی پیاده‌سازی گردید. برای کشف الگوی موضوعی رفتار جستجوی کاربر نوجوان از مجموعه داده‌ای Del.icio.us مربوط به رسانه اجتماعی Delicious که مجموعه داده ایجادشده توسط وتزکر [۲۸] است استفاده شد. این مجموعه داده بزرگ‌ترین مجموعه داده اجتماعی تگ‌شده است که قابل دسترس برای امور پژوهشی مرتبط است و شامل چهارصد و پنجاه میلیون تگ می‌باشد. از این مجموعه داده، تگ‌های موضوعی مرتبط با URLهای کاربران نوجوان استخراج گردید. سپس از بخش مربوط به کلیک‌های لاگ جستجوی AOL که شامل حدود بیست میلیون نمونه پرس‌وجوی مربوط به ۶۵۰ هزار کاربر می‌باشد بهره گرفتیم. هر رکورد در این لاگ جستجو شامل صفت‌های زیر است:

- **صفت AnonID:** یک شماره شناسه کاربر بی‌نام است.
- **صفت Query:** شامل پرس‌وجوی ارسال‌شده توسط کاربر است.
- **صفت queryTime:** شامل زمان ارسال پرس‌وجو است.
- **صفت ItemRank:** شامل رتبه موقعیت نتیجه بازدیدشده توسط کاربر بر روی هر کدام از رکوردها است.
- **صفت ClickURL:** شامل URL کلیک‌شده بر روی یک نتیجه جستجو است.

شکل ۴ روند کلی کشف الگوی موضوعی کلیک‌ها را برای هر پرس‌وجو در ابزار Ateryx نشان می‌دهد. به دلیل این که لاگ جستجو AOL، مجموعه داده دایرکتوری نوجوانان مربوط به Dmoz و مجموعه داده Del.icio.us حجیم هستند، بنابراین برای اکتشاف الگوی موضوعی کلیک‌ها از ابزار Alteryx استفاده می‌شود که می‌توان بر اساس آن پردازش داده‌های عظیم را انجام داد.

در این مدل ابتدا عمل استخراج جستجوهای کاربران نوجوان از لاگ جستجوی AOL و تگ‌های مربوط به URLهای نوجوانان بر اساس انطباق URLهای در دایرکتوری نوجوانان مربوط به Dmoz با URLهای در لاگ جستجوی AOL و مجموعه داده Del.icio.us انجام می‌شود. بعد از فیلترینگ بر روی جستجوهای استخراج‌شده، یک گراف دوبخشی از پرس‌وجوها و کلیک‌های کاربران نوجوان و یک مجموعه داده تگ‌های مربوط به نوجوانان حاصل می‌گردد. سپس سه تگی که مرتبط‌ترین تگ‌ها به اسنادی که منطبق با هر URL مربوط به نوجوانان است، بر اساس محاسبه امتیاز TF.IDF مشخص می‌شود. با استفاده از عمل گروه‌بندی بر روی پرس‌وجوها و پرتکرارترین موضوع کلیک‌ها، یک الگوی موضوعی به

دست می‌آید.

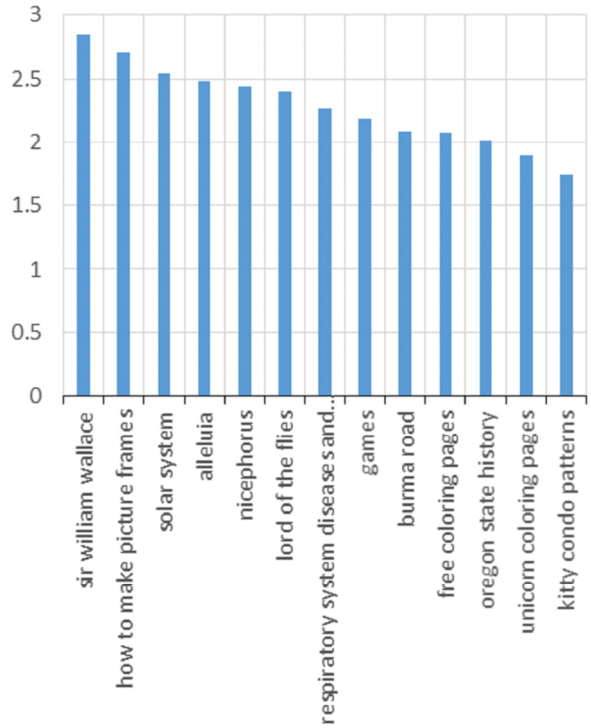


شکل ۴: روند کشف الگوی موضوعی کلیک‌های کاربران نوجوان به ازای هر پرس‌وجو.

الگوی (۱۳) یک الگوی موضوعی کلیک‌های مربوط به پرس‌وجوی free coloring pages به صورت مجموعه‌ای از چهارتایی‌ها را نشان می‌دهد. این الگو شامل سه URL است که اولین مورد آن یعنی http://www.coloringcastle.com با احتمال ۰٫۰۰۴۵ بیشترین مرتبه کلیک بر اساس مرتبط‌ترین تگ "lessons" با مقدار ارتباط ۰٫۵۲۹۰ را در میان تراکنش‌های استخراج‌شده از لاگ AOL داشته است. بعد از آن آنتروپی هر الگوی موضوعی بر اساس (۱۲) و تشابه مجموعه الگوهای موضوعی بر اساس (۱۰) به ازای هر پرس‌وجوی استخراج‌شده از لاگ AOL محاسبه می‌گردد. شکل ۵ آنتروپی الگوی موضوعی کلیک‌های کاربران نوجوان به ازای هر پرس‌وجوی استخراج‌شده از لاگ جستجوی AOL را نشان می‌دهد.

هرچه مقدار آنتروپی الگوی موضوعی کلیک‌ها کمتر باشد به معنای این است که ابهام در پرس‌وجو کمتر است و همچنین به معنای ابهام کمتر در ناوبری موضوعی کلیک‌ها نیز می‌باشد. در این پژوهش، پیشنهاد پرس‌وجو به عنوان یک وظیفه کلاسه‌بندی در نظر گرفته می‌شود که با اضافه‌کردن یک واژه به پرس‌وجوی اصلی فضای جستجو محدود می‌گردد. از لاگ پرس‌وجوی AOL برای این منظور استفاده می‌شود. برای انجام کلاسه‌بندی دو کلاس YES و NO به ترتیب برای پیشنهاد پرس‌وجو یا عدم پیشنهاد پرس‌وجوهای کاندید در نظر گرفته می‌شود. برای این منظور پرس‌وجوهایی که بیش از ۱۰ مرتبه در لاگ تکرار شده باشند به عنوان

$$TopicClickPattern = \{(\text{http://www.coloringcastle.com}, lessons, ۰٫۰۰۴۵, ۰٫۵۲۹۰), (\text{http://www.activityvillage.co.uk}, color, ۰٫۰۰۱۴, ۰٫۰۹۵۶), (\text{http://familycrafts.about.com}, Books, ۰٫۰۰۱۳, ۰٫۳۰۴۵)\} \quad (۱۳)$$



شکل ۵: آنتروپی مجموعه الگوهای موضوعی به ازای هر پرس‌وجوی کاربران نوجوان.

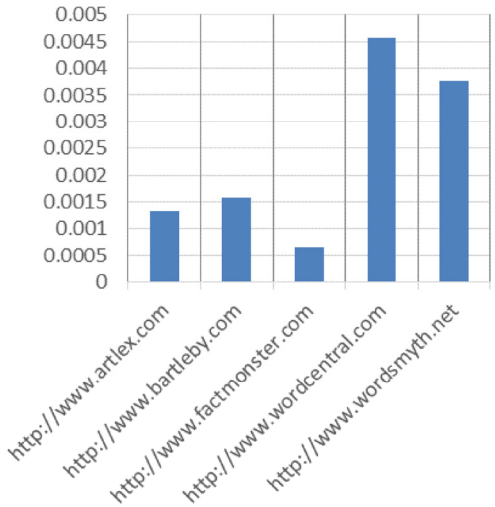
پیشنهاد پرس‌وجوی کاندید انتخاب می‌شوند. بنابراین پرس‌وجوهای کاندید اگر در لاگ بیش از ۱۰ مرتبه تکرار شده باشد در کلاس YES وگرنه در کلاس NO برچسب‌گذاری می‌شوند.

در کل ۱۵۶ پرس‌وجو به عنوان پیشنهاد پرس‌وجوی کاندید برچسب‌گذاری شده وجود خواهد داشت. سپس کلاسه‌بندی دودویی برای هر پرس‌وجوی کاندید صورت می‌گیرد. آزمایش، استفاده از الگوی موضوعی کلیک‌ها و آنتروپی الگوی موضوعی در یک کاربرد واقعی را روشن می‌سازد. یک کلاسه‌بندی دودویی برای هر کدام از پرس‌وجوهای کاندید استفاده شده است. در این روش، الگوی موضوعی کلیک‌های کاربران نوجوان از لاگ جستجو، توسط ابزار Alteryx کشف گردیده و تشابه الگوی موضوعی کلیک‌ها محاسبه شده است. سپس با استفاده از کلاسه‌بندی دودویی نزدیک‌ترین پرس‌وجو به پرس‌وجوی مورد نظر کاربر نوجوان مشخص گردیده است.

ویژگی‌های استفاده‌شده در مرتبه اول اجرای کلاسه‌بندی دودویی برای هر پرس‌وجو در جدول ۱ معرفی شده است. در مرتبه دوم بر اساس ویژگی سنتی محبوبیت که در پژوهش‌های قبلی استفاده گردیده، کلاسه‌بندی دودویی انجام شده است.

با استفاده از ابزار WEKA، کلاسه‌بندی دودویی انجام گرفته است. بعد از عملیات پیش‌پردازش عمل کلاسه‌بندی به روش نزدیک‌ترین همسایه (KNN) دو مرتبه اجرا می‌گردد. کارایی آن در جدول ۲ آورده شده است.

در حقیقت پیشنهاد پرس‌وجو با کاهش ابهام رابطه مستقیم دارد و برای پرس‌وجوهای اطلاعاتی هدف اصلی کاهش ابهام است. برای پرس‌وجوهای ناوبری که مقصود به طور کامل روشن است هدف اصلی استخراج پرس‌وجوهای با تشابه بیشتر است. در نتیجه ویژگی‌های آنتروپی کلیک و میانگین آنتروپی که در پژوهش‌های قبلی توسط پژوهشگران



شکل ۶: احتمال کلیک URLها برای پرس‌وجوی "dictionary".

جدول ۱: ویژگی‌های استفاده‌شده در کلاسه‌بندی دودویی.

	ویژگی
۱	آنتروپی موضوع‌های مجموعه الگوهای موضوعی
۲	تشابه مجموعه الگوهای موضوعی با یکدیگر
۳	میانگین آنتروپی موضوعی کلیک‌ها
۴	تشابه موضوعی مجموعه الگوهای موضوعی
۵	طول پرس‌وجو
۶	آنتروپی کلیک‌های مجموعه الگوهای موضوعی
۷	میانگین آنتروپی کلیک‌ها

جدول ۲: نتایج کلاسه‌بندی دودویی برای پیشنهاد پرس‌وجو.

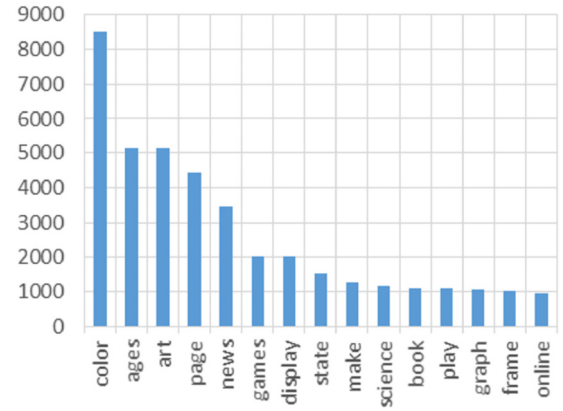
ویژگی	Accuracy	Precission	Recall
محبوبیت	٪۷۴٫۳۵۹	۰٫۸۰۴	۰٫۸۳۳
تشابه و آنتروپی موضوعی الگوها	٪۹۵٫۵۱۲	۰٫۹۹	۰٫۹۴۴

استفاده شده است برای پرس‌وجوهای اطلاعاتی مناسب نیستند زیرا مقدار بالایی را شامل می‌شوند. بنابراین یک سیستم پیشنهاد با استفاده از ویژگی‌های آنتروپی کلیک و میانگین آنتروپی به سمت فقط پیشنهاددادن پرس‌وجوهای ناوبری منحرف می‌شود و در نتیجه، استفاده از ویژگی استخراج‌شده آنتروپی مجموعه الگوهای موضوعی کلیک‌ها و تشابه آنها برای کاربران نوجوان موجب جلوگیری از این انحراف و انحراف موضوعی سیستمِ پیشنهاد می‌شود.

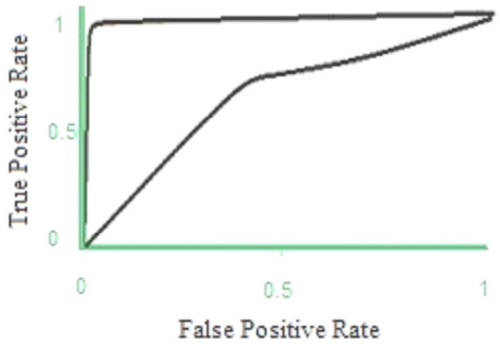
۵- نتایج آزمایش

آزمایش کارایی سیستم بر اساس بخش پرس‌وجوهای لاگ جستجوی AOL صورت گرفته است. در ابتدا با تطابق URLهای کلیک‌شده در لاگ AOL با دامنه‌های لیست‌شده در بخش "Kids and Teens" مربوط به Dmoz پرس‌وجوهای مربوط به نوجوانان از لاگ جستجوی AOL استخراج گردیده است. شکل ۶ احتمال کلیک URLها را بر اساس پرس‌وجوی "dictionary" نشان می‌دهد.

در میان URLهای کلیک‌شده بالاترین احتمال کلیک را http://www.wordcentral.com و بعد از آن احتمال بالاتر را http://www.wordsmyth.net داشته است. توزیع احتمالی URLهای کلیک‌شده هر سند بر طبق (۲) برای هر کدام از پرس‌وجوهای لاگ جستجوی AOL محاسبه شده است.



شکل ۷: بیشترین فراوانی تگ‌های موضوعی کاربران نوجوان در محتوای اسناد استخراج‌شده از لاگ جستجوی AOL.



شکل ۸: منحنی ROC کلاسه‌بندی دودویی.

شکل ۷ بیشترین فراوانی تگ‌های موضوعی در محتوای اسناد مربوط به URLهای نوجوانان استخراج‌شده از لاگ AOL را نشان می‌دهد. تگ‌های مربوط به URLهای نوجوانان از مجموعه داده Del.icio.us استخراج شده است. پرس‌وجوی "color" بالاترین فراوانی در میان اسناد کلیک‌شده کاربران نوجوان در تراکنش‌های استخراج‌شده از لاگ جستجوی AOL را دارد.

جدول ۲ نتیجه کلاسه‌بندی برای پیشنهاد پرس‌وجو را نشان می‌دهد. کلاسه‌بندی دو مرتبه انجام می‌شود. سطر اول جدول نتایج کلاسه‌بندی بر اساس پژوهش‌های قبلی پژوهشگران مانند پژوهش هیوژانگ دوان و همکارانش [۲۳] را نشان می‌دهد که از ویژگی محبوبیت پرس‌وجوها استفاده شده است. دقت کلاسه‌بندی دودویی برابر با ٪۷۴٫۳۵۹ به دست آمده است. سطر دوم جدول، نتایج کلاسه‌بندی را با استفاده از ویژگی‌های پیشنهادی شامل تشابه و آنتروپی موضوعی نشان می‌دهد که دقت کلاسه‌بندی ٪۹۵٫۵۱۲ به دست آمده است.

با توجه به این که نویز در ناوبری موضوعی کلیک‌های کاربران نوجوان نسبت به بزرگسالان بیشتر است روش پیشنهادی الگوی موضوعی کلیک‌ها یک مدل با دقت بالاتر برای فیلترنمودن نویز ناوبری موضوعی کلیک‌های کاربران نوجوان پیشنهاد می‌کند و بر اساس آن کلاسه‌بندی دودویی دقت بیشتری برای کلاسه‌بندی پرس‌وجوها را نشان می‌دهد. همچنان که در جدول ۲ آمده است با افزودن ویژگی تشابه و آنتروپی موضوعی مجموعه الگوها، دقت کلاسه‌بندی نیز افزایش می‌یابد.

شکل ۸ منحنی ROC مربوط به نتیجه کلاسه‌بندی دودویی را نشان می‌دهد. این منحنی حساسیت مابین پرس‌وجوهای پیشنهادی که به درستی کلاسه‌بندی شده و موارد نادرست را در هر دو مرتبه کلاسه‌بندی نشان می‌دهد. منحنی ROC مربوط به کلاسه‌بندی با استفاده از دو ویژگی آنتروپی و تشابه موضوعی مجموعه الگوهای موضوعی کلیک‌ها،

جدول ۳: پیشنهاد پرس‌وجو به کاربران نوجوان برای "FREE COLORING PAGES".

پیشنهاد پرس‌وجو	تشابه موضوعی	آنتروپی موضوعی	پیشنهاد پرس‌وجو
unicorn_coloring_pages	۰٫۹۶	۰٫۲۱۷۳	YES
free online games	۰٫۹۲	۰٫۲۷۶۰	YES
free_online_lessons	۰٫۴۴	۰٫۷۱۳۵	NO

بالاتر از کلاسه‌بندی با ویژگی سنتی محبوبیت پرس‌وجوها است و نشان‌دهنده این است که با دقت بالاتر پیشنهادهای پرس‌وجوی کاندید را کلاسه‌بندی می‌کند.

در جدول ۳ نتایج کلاسه‌بندی دودویی برای پیشنهاد پرس‌وجو بر اساس پرس‌وجوی "free coloring pages" نشان داده شده است. دو پرس‌وجوی پیشنهادشده ابتدایی که در کلاس YES پیش‌بینی شده است، مقدار آنتروپی موضوعی مجموعه الگوهای آن پایین و تشابه موضوعی مجموعه الگوهای آن با پرس‌وجوی اصلی بیشتر است. در حالی که پیشنهاد پرس‌وجوی "free_online_lessons" که در کلاس NO پیش‌بینی شده است مقدار آنتروپی موضوعی مجموعه الگوهای آن بالاست و تشابه موضوعی مجموعه الگوهای آن نسبت به پرس‌وجوی اصلی کمتر است. این نشان از صحت آزمایش دارد.

در حالت کلی آزمایش نشان می‌دهد که شباهت موضوعی الگوها برای پیشنهاد پرس‌وجو به کاربر نوجوان مفید می‌باشد. نتایج آزمایش حاکی از آن است که ویژگی الگوی موضوعی کلیک‌ها، سبب نزدیک‌شدن پیشنهاد پرس‌وجو به پرس‌وجوی مورد نظر کاربر می‌گردد.

٦- نتیجه‌گیری و پیشنهادها

در این پژوهش به دلیل این که نویز در ناوبری موضوعی کلیک‌های کاربران نوجوان نسبت به بزرگ‌سالان بیشتر است، با استفاده از روش پیشنهادی مجموعه الگوهای موضوعی کلیک‌ها، یک مدل با دقت بیشتر برای فیلترنمودن نویز ناوبری موضوعی کلیک‌های کاربران نوجوان حاصل شد. الگوهای موضوعی کلیک‌های استخراج‌شده از لاگ جستجو، موجب نزدیکی بیشتر پیشنهاد پرس‌وجو نسبت به پرس‌وجوی اصلی کاربر نوجوان می‌گردد. یک سیستم پیشنهاد با استفاده از ویژگی‌های آنتروپی کلیک و میانگین آنتروپی به سمت فقط پیشنهاددادن پرس‌وجوهای ناوبری منحرف می‌شود. در نتیجه استفاده از ویژگی استخراج‌شده آنتروپی مجموعه الگوهای موضوعی کلیک‌ها و تشابه آنها برای کاربران نوجوان موجب جلوگیری از انحراف موضوعی سیستمِ پیشنهاد می‌شود. به هر حال استفاده از الگوهای موضوعی کلیک‌ها برای کلاسه‌بندی دودویی پرس‌وجوها بالقوه بر روی مرتبط‌بودن نتایج بازیابی اطلاعات برای کاربران نوجوان تأثیر بسزایی دارد.

به عنوان پیشنهاد برای پژوهش آینده می‌توان تگ‌های وابسته به الگوی موضوعی کلیک‌های کاربران نوجوان را رتبه‌بندی نمود. سپس از تگ‌های با رتبه بالا برای فرمول‌بندی مجدد پرس‌وجوی اصلی کاربر نوجوان بهره برد.

مراجع

[1] A. T. Mulik and H. Palkar, "A survey on development of search engine," *Int. Advanced Research J. in Science, Engineering and Technology*, vol. 4, no. 4, pp. 116-117, Jan. 2017.

[2] M. Madden, A. Lenhart, M. Duggan, S. Cortesi, and U. Gasser, *Teens and Technology 2013*, Washington, DC: Pew Research Center's Internet & American Life Project, 2013.

[3] A. Druin, E. Foss, H. Hutchinson, E. Golub, and L. Hatley, "Children's roles using keyword search interfaces at home," in *Proc. of the 28th Int. Conf. on Human Factors in Computing Systems-CHI'10*, pp. 413-422, Atlanta, GA, USA, 10-15 Apr. 2010.

[4] D. Bilal, "Children's use of the Yahooligans! web search engine. III. cognitive and physical behaviors on fully self-generated search tasks," *J. of the American Society for Information Science and Technology*, vol. 53, no. 13, pp. 1170-1183, Nov. 2002.

[5] D. Bilal, "Children's use of the Yahooligans! web search engine: II. cognitive, physical, and affective behaviors on fact-based search tasks," *J. of the American Society for Information Science and Technology*, vol. 52, no. 2, pp. 118-136, Oct. 2001.

[6] M. Caramia, G. Felici, and A. Pezzoli, "Improving search results with data mining in a thematic search engine," *Computers and Operations Research*, vol. 31, no. 14, pp. 2387-2404, Dec. 2004.

[7] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query recommendation using query logs in search engines," in *Proc. of the Int. Conf. on Current Trends in Database Technology*, *EDBT'04*, pp. 588-596, Heraklion, Greece, 14-18 Mar. 2004.

[8] E. Foss, *et al.*, "Children's search roles at home: implications for designers, researchers, educators, and parents," *J. of the American Society for Information Science and Technology*, vol. 63, no. 3, pp. 558-573, Mar. 2012.

[9] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: a review," *SIGMOD Rec.*, vol. 34, no. 2, pp. 18-26, Jun. 2005.

[10] Y. Liu, J. Miao, M. Zhang, S. Ma, and L. Ru, "How do users describe their information need: query recommendation based on snippet click model," *Expert Systems with Applications*, vol. 38, no. 11, pp. 13847-13856, Oct. 2011.

[11] J. Wen, J. Nie, and H. Zhang, "Clustering user queries of a search engine," in *Proc. 10th Int. Conf. on World Wide Web, WWW'01,* pp. 162-168, Hong Kong, China, 1-5 May. 2001.

[12] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, "Analysis of a very large web search engine query log," *ACM SIGIR Forum*, vol. 33, no. 1, pp. 6-12, Fall 1999.

[13] A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic, "Searching the web: the public and their queries," *J. of the American Society for Information Science and Technology*, vol. 52, no. 3, pp. 226-234, Feb. 2001.

[14] G. Pass, A. Chowdhury, and C. Torgeson, "A picture of search," in *Proc. of the 1st Int. Conf. on Scalable Information Systems, InfoScale'06*, vol. 152, 7 pp., Hong Kong, 30 May- 1 Jun. 2006.

[15] D. J. Brenes and D. Gayo-Avello, "Stratified analysis of AOL query log," *Information Sciences*, vol. 179, no. 12, pp. 1844-1858, May 2009.

[16] R. Jones and K. L. Klinkner, "Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs," in *Proc. of the 17th ACM Conf. on Information and Knowledge Management*, pp. 699-708, Napa Valley, CA, USA, 26-30 Oct. 2008.

[17] R. Kumar and A. Tomkins, "A characterization of online browsing behavior," in *Proc. of the 19th Int. Conf. on World Wide Web*, *WWW'10,* pp. 561-570, Raleigh, NC, USA, 26-30 Apr. 2010.

[18] Z. Cheng, B. Gao, and T. Liu, "Actively predicting diverse search intent from user browsing behaviors," in *Proc. of the 19th Int. Conf. on World Wide Web*, *WWW'10,* pp. 221-230, Raleigh, NC, USA, 26-30 Apr. 2010.

[19] S. D. Torres, D. Hiemstra, I. Weber, and P. Serdyukov, "Query recommendation for children," in *Proc. of the 21th ACM Int. Conf. on Information and Knowledge Management, CIKM'12*, pp. 2010-2014, Maui, HI, USA, 29 Oct. 2- Nov. 2012.

[20] S. D. Torres, D. Hiemstra, and T. Huibers, "Vertical selection in the information domain of children," in *Proc. of the 13th ACM/IEEE-CS Joint Conf. on Digital Libraries*, *JCDL'13,* pp. 57-66, 22-26 Jul. 2013.

[21] S. D. Torres, D. Hiemstra, I. Weber, and P. Serdyukov, "Query recommendation in the information domain of children," *J. of the Association for Information Science and Technology*, vol. 65, no. 7, pp. 1368-1384, Jul. 2014.

[22] Y. Wang and E. Agichtein, "Query ambiguity revisited: clickthrough measures for distinguishing informational and ambiguous queries," in *the Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, pp. 361-364, 2-4 Jun. 2010.

[23] H. Duan, E. Kiciman, and C. Zhai, "Click patterns: an empirical representation of complex query intents," in *Proc. of the 21st ACM Int. Conf. on Information and Knowledge Management, CIKM'12*, pp. 1035-1044, Maui, HI, USA, 29 Oct.- 2 Nov. 2012.

[24] D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log," in *Proc. of the 6th ACM SIGKDD Int. Conf. on*

Knowledge Discovery and Data Mining, KDD'00, pp. 407-416, Boston, MA, USA, 20-23 Aug. 2000.
[25] M. Hosseini and H. Abolhassani, "Clustering search engine log for query recommendation," in *Proc.-Advances in Computer Science and Engineering*, vol. 6, pp. 380-387, Kish Island, Iran, 9-11 Mar. 2008.
[26] Alteyx Inc., "Ateryx Designer x64," Boulder, Colorado, 2015. Availabe: http://www.alteryx.com.
[27] "Weka 3: Data Mining Software in Java," Machine Learning Group at the University of Waikato, Hamilton, New Zealand, 2015. Availabe: http://www.cs.waikato.ac.nz/~ml/weka/.
[28] R. Wetzker, C. Zimmermann, and C. Bauckhage, "Analyzing social bookmarking systems: a del.icio.us cookbook," in *Proc. of the ECAI Mining Social Data Workshop*, pp. 26-30, Jul. 2008.

حیدر قاسم‌زاده در سال ۱۳۷۸ مدرک کارشناسی مهندسی کامپیوتر خود را از دانشگاه آزاد اسلامی واحد میبد و در سال ۱۳۸۱ مدرک کارشناسی ارشد مهندسی کامپیوتر خود را از دانشگاه علوم و تحقیقات تهران دریافت نمود. از بهمن ماه ۱۳۹۱ الی بهمن ماه ۱۳۹۶ نام‌برده جهت انجام دوره دکتری در دانشگاه یزد مشغول به تحصیل و پژوهش بودند. زمینه‌های علمی مورد علاقه ایشان عبارتند از: بازیابی هوشمند اطلاعات، داده‌های حجیم، داده کاوی و شبکه‌های عصبی مصنوعی.

محمد قاسم‌زاده در سال ۱۳۶۸ مدرک کارشناسی مهندسی کامپیوتر خود را از دانشگاه شیراز و در سال ۱۳۷۴ مدرک کارشناسی ارشد مهندسی کامپیوتر (هوش ماشین و رباتیک) خود را از دانشگاه صنعتی امیرکبیر دریافت نمود. از بهمن ماه ۱۳۸۰ الی بهمن ماه ۱۳۸۴ نام‌برده جهت انجام دوره دکتری (علوم کامپیوتر) در دانشگاه تربیر و دانشگاه پتسدام هر دو در کشور آلمان مشغول به تحصیل و پژوهش بودند. ایشان در سال ۱۳۸۴ موفق به اخذ درجه دکتری علوم کامپیوتر گردید. محمد قاسم‌زاده از سال ۱۳۷۵ تاکنون به عنوان عضو هیأت علمی در دانشگاه یزد مشغول به تدریس و تحقیق می‌باشند. زمینه‌های علمی مورد علاقه نام‌برده عبارتند از: طراحی و تحلیل الگوریتم‌ها، پردازش زبان طبیعی، سیستم‌های هوشمند و محاسبات نرم.

علی‌محمد زارع بیدکی در سال ۱۳۷۸ مدرک کارشناسی مهندسی کامپیوتر خود را از دانشگاه صنعتی اصفهان و در سال ۱۳۸۱ مدرک کارشناسی ارشد مهندسی کامپیوتر (معماری کامپیوتر) خود را از دانشگاه تهران دریافت نمود. از سال ۱۳۸۴ الی ۱۳۸۸ نام‌برده جهت انجام دوره دکتری مهندسی کامپیوتر (نرم افزار) در دانشگاه تهران مشغول تحصیل و پژوهش بودند. ایشان در سال ۱۳۸۸ موفق به اخذ درجه دکتری مهندسی کامپیوتر گردید. در حال حاضر دکتر زارع بیدکی عضو هیأت علمی دانشگاه یزد می‌باشند. زمینه‌های علمی مورد علاقه نام‌برده عبارتند از: بازیابی اطلاعات در وب ، وب کاوی و داده کاوی، پردازش و مدیریت گرافهای حجیم، نمایه سازی و جستجو، پردازش زبان‌های طبیعی، مدل‌سازی کاربر و وب معنایی.