هشتمین کنفرانس بین المللی فناوری اطلاعات و دانش
The Eighth International Conference on Information and Knowledge Technology

دسته‌بندی موضوعی پرس‌وجوهای زبان فارسی

محمد صادق زاهدی[۱،۲]، بهروز منصوری[۲]، مژگان فرهودی[۱] ، شیوا مرادخانی[۱]

[۱]مرکز تحقیقات مخابرات ایران،

{s.zahedi, farhoodi, sh.moradkhani }@itrc.ac.ir

[۲]دانشگاه تهران، b.mansouri@ut.ac.ir

چکیده

مسئله‌ی دسته‌بندی موضوعی پرس‌وجوها یکی از مسائل چالش برانگیز و مهم در زمینه داده‌کاوی است که در بسیاری از زمینه‌ها نظیر موتورهای جستجو، سیستم پرسش و پاسخ و سیستم‌های آنلاین تبلیغات کاربرد دارد. با وجود اهمیت بسیار بالای دسته‌بندی موضوعی پرس‌وجوها، هنوز پژوهشی در این زمینه برای زبان فارسی صورت نگرفته است لذا در این مقاله سعی شده است، در ابتدا راه‌کاری برای دسته‌بندی موضوعی پرس‌وجوهای زبان فارسی ارائه گردد و سپس بر اساس آن پرس‌وجوهای لاگ موتور جستجوی بومی تحلیل شود. دو روش مبتنی بر مدل زبانی و اسناد مرتبط با پرس‌وجو ارائه شده است که هر دو روش از نسخه‌ی توسعه یافته‌ی مجموعه داده‌ی همشهری به عنوان داده آموزشی استفاده می‌نمایند. برای ارزیابی روش‌های ارائه شده از یک مجموعه داده استاندارد برای دسته‌بندی موضوعی پرس‌وجوهای فارسی شامل ۷۰۰۰ پرس‌وجو و دسته‌ی موضوعی آن‌ها، استفاده شده است. نتایج حاصل شده حاکی از عملکرد مناسب روش‌های پیشنهادی برای دسته‌بندی موضوعی پرس‌وجوهای زبان فارسی دارند.

واژه‌های کلیدی

دسته بندی پرس‌وجو، دسته‌بندی موضوعی پرس‌وجو، لاگ موتور جستجو، Query Classification.

۱- مقدمه

دسته‌بندی موضوعی پرس‌وجوها یک مسئله‌ی تعریف شده در داده کاوی است که به این صورت تعریف می‌شود: سیستم دسته‌بندی موضوعی پرس‌وجوها، پرس‌وجو کاربر را به عنوان ورودی گرفته و دسته‌ی موضوعی مربوط به آن را از میان چند دسته‌ی از پیش تعریف شده استخراج می‌کند. دسته‌بندی موضوعی پرس‌وجوها در بسیاری از کاربردها مبتنی بر اینترنت مانند موتورهای جستجو، سیستم‌های پرسش و پاسخ و سرویس‌های آنلاین تبلیغات[1] کاربرد دارد. یکی از مهمترین کاربردهای سیستم دسته‌بندی خودکار پرس‌وجوها در موتورهای جستجو است. به عنوان مثال موتورهای جستجو در صورتی که دسته‌ی موضوعی پرس-وجوی کاربر را تشخیص دهند می‌توانند تبلیغات متناسب با آن دسته‌ی موضوعی را به کاربر نمایش دهند علاوه بر این تشخیص خودکار دسته‌ی موضوعی پرس‌وجو در دقت بازیابی اسناد مرتبط با پرس‌وجو نیز تاثیر گذار خواهد بود.

اهمیت موتورهای جستجو به عنوان دروازه ورودی اینترنت بسیار مهم است زیرا این موتورهای جستجو هستند که کاربران به وسیله آن‌ها در اینترنت گشت و گذار نموده و حتی گاهی اوقات، مخاطب را به سمت و سویی هدایت می‌کنند. لذا تحلیل رفتار کاربران در موتورهای جستجو از اهمیت بالایی برخوردار است. یکی از روش‌های اصلی تحلیل رفتار کاربران موتور جستجو از طریق تحلیل لاگ موتور جستجو است که همواره مورد توجه بسیاری از پژوهشگران بوده است[۱-۴]. اما چالش بزرگ برای تحلیل لاگ حجم بسیار بالای اطلاعات در لاگ موتور جستجو است. دسته‌بندی موضوعی خودکار پرس‌وجوها یکی از راه‌های کاهش حجم داده‌ها برای تحلیل رفتار کاربران در موتور جستجو است.

اما چالش اصلی دسته‌بندی موضوعی پرس‌وجوها طول کوتاه پرس‌وجوها و مبهم بودن آن است [۵]. آزمایشات ما نشان می‌دهد که میانگین طول پرس‌وجوهای کاربران در لاگ موتور جستجوی بومی پارسی‌جو برابر ۲,۹ کلمه بوده است همچنین در [۵] نشان داده‌شده است که ۸۰ درصد از پرس‌وجوهای کاربران کمتر از ۳ کلمه بوده است. طول کوتاه پرس‌وجوها باعث می‌شود که بردار ویژگی مربوط به پرس‌وجو بسیار تنک شود.

چالش دیگر مبهم بودن پرس‌وجوها است. بسیاری از پرس‌وجوها مبهم می‌باشند و ممکن است در چندین دسته‌ی موضوعی قرار گیرند لذا پیدا کردن دسته‌ی موضوعی درست دشوار خواهد بود.

[1] Online advertisement services

در این مقاله هدف ما ارائه‌ی راه‌کاری برای دسته‌بندی موضوعی پرس‌وجوهای زبان فارسی است. برای این کار دو روش پیشنهاد شده است: روش اول مبتنی بر بازیابی اسناد مرتبط با پرس‌وجو است. به این صورت که در ابتدا لیستی از اسناد مرتبط با پرس‌وجو از یک مجموعه داده که شامل تعداد زیادی اسناد به همراه برچسپ موضوعی آنها است، بازیابی شده سپس بر اساس دسته‌ی موضوعی این اسناد، موضوع پرس‌وجو تعیین می‌گردد. روش دوم مبتنی بر مدل زبانی است به این صورت که برای دسته‌های موضوعی مختلف بر اساس همان مجموعه اسناد برچسپ خورده یک مدل زبانی ایجاد شده و در نهایت به کمک این مدل‌های زبانی دسته‌ی موضوعی پرس‌وجو تعیین می‌گردد. برای مجموعه اسناد برچسپ خورده از نسخه‌ی توسعه داده شده مجموعه داده همشهری [۶] استفاده شده است.

ادامه‌ی مقاله به صورت زیر سازمان‌دهی شده است: در بخش دوم پژوهش‌های صورت گرفته در حوزه دسته‌بندی موضوعی پرس‌وجوها بررسی می‌شود. در بخش سوم مجموعه داده‌ای که برای دسته‌بندی و بررسی عملکرد روش‌های پیشنهادی استفاده شده معرفی می‌شود. سپس در بخش چهارم به معرفی الگوریتم‌های پیشنهادی پرداخته می‌شود. نتایج حاصل شده و تحلیل آن‌ها در بخش پنجم ارائه شده و در آخر به جمع‌بندی و کارهای آینده پرداخته می‌شود.

۲- کارهای مرتبط

دسته‌بندی پرس‌وجوها از دیدگاه‌های مختلف مورد توجه بسیاری از محققین بوده است. به عنوان مثال، دسته‌بندی زمانی[2] پرس‌وجوها در پژوهش‌هایی مانند [۹-۷] بررسی شده است درحالی که مقالاتی مانند [۱۰،۱۱] پرس‌وجوهای فضایی را دسته‌بندی می‌کنند. در [۱۲] راهکاری برای دسته بندی پرس‌وجوهای پرسشی ارائه شده است. به صورت کلی روش‌های دسته‌بندی پرس‌وجوها در یکی از دسته‌های زیر قرار می‌گیرند:

- دسته‌ی اول روش‌هایی هستند که با استفاده از منابع خارجی(منابعی به غیر از خود پرس‌وجو) پرس‌وجوی کاربر را توسعه داده‌اند منابع خارجی استفاده‌شده شامل: صفحات بازگرداننده شده برای پرس‌وجو توسط موتور جستجو، مجموعه داده خارجی موجود و استفاده از دسته‌بندی‌های موجود به عنوان یک دسته‌بندی واسط برای دسته‌بندی نهایی پرس‌وجو [۱۳، ۱۴].
- دسته دوم روش‌هایی هستند از یادگیری با ناظر[3] استفاده کرده‌اند و از داده‌های برچسب نخورده برای بهبود یادگیری استفاده نموده‌اند[۵، ۱۵].
- دسته سوم روش‌هایی هستند که اقدام به توسعه داده‌های آموزش[4] با استفاده از دسته‌بندی کردن خودکار تعدادی پرس وجو به کمک داده های مانند کلیک کاربران، نموده اند[۱۶].
- دسته‌ی چهارم روش‌های آگاه از زمینه که در این حالت رفتار کاربر را نیز در نظر می‌گیرند [۱۷].

دسته‌بندی موضوعی پرس‌وجوها با برگزاری رقابت KDDCCUP [۱۸] در ۲۰۰۵ بیش‌تر از قبل مورد توجه قرار گرفت. در این رقابت از شرکت‌کنندگان خواسته شده بود ۸۰۰۰۰۰ پرس‌وجو را به ۶۷ دسته از پیش تعیین شده دسته‌بندی کنند. [۱۹] با ارائه‌ی یک روش دسته‌بندی جدید مبتنی بر ساخت پل، راهکاری ارائه کرد که از تمامی روش‌های پیشنهادی در رقابت KDDCUP بهتر عمل کند. [۲۰] با استفاده از پرس‌وجوها از پیش دسته‌بندی شده به صورت دستی، روشی برای دسته‌بندی موضوعی پرس‌وجوها ارائه داد. راهکار ارائه شده در [۲۱] فقط از ویژگی‌های کلمات موجود در پرس‌وجو استفاده می‌کند و هیچ اطلاعات بیرونی و اضافه‌ای در روش دیده نمی‌شود. در یکی از پژوهش‌های اخیر، [۲۲] روشی سریع و مقیاس‌پذیر برای دسته‌بندی موضوعی پرس‌وجو ارائه داده که مبتنی بر بازیابی صفحات مرتبط ویکیپدیا است.

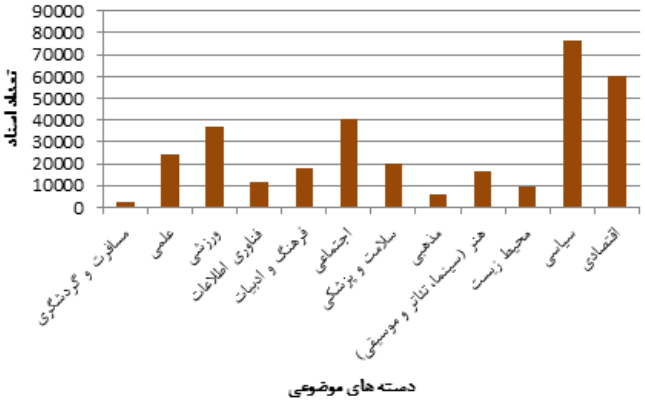
۳- مجموعه داده

یکی از پیش‌نیازهای اصلی مسئله‌ی دسته‌بندی موضوعی پرس‌وجوها، داده‌های آموزش است. در این راستا مجموعه داده‌ی همشهری[۶] توسعه داده شده است. مجموعه داده همشهری خبرهای تا سال ۱۳۸۵ را شامل می‌شود. در این مقاله این مجموعه داده گسترش داده شده است طوری که شامل حدود ۵۰۰ هزار خبر به همراه برچسب موضوعی آن‌ها است. این مجموعه شامل خبرهای همشهری از سال ۱۳۷۵ تا ۱۳۹۵ می‌باشد. مجموعه اسناد همشهری با خزش (Crawl) وب سایت همشهری و چندین مرحله پیش‌پردازش و برچسب‌گذاری حاصل آمده است. در نمودار شکل ۱ تعداد اسناد مربوط به هر دسته نشان داده شده است. همان طور که مشخص است تعداد اسناد مربوط به برخی از دسته‌ها کم می‌باشد که در آینده سعی خواهد شد که این مجموعه داده از طریق خزش سایت‌های تخصصی مربوط به هر دسته‌ی موضوعی، توسعه داده شود.

[2] Temporal

[3] Supervised learning

[4] training data



شکل ۱: توزیع خبرهای مجموعه داده همشهری بر اساس موضوع

اما برای ارزیابی روش‌های پیشـــنهادی به یک مجموعه داده استاندارد نیاز هست. در این راستا به صورت تصادفی ۷۰۰۰ پرس‌وجو از لاگ موتور جستجوی بومی پارسی‌جو[5] انتخاب شده است و هر پرس‌وجو به ۳ کاربر انسانی مختلف داده شده است که هر کابر دسته‌ی موضوعی پرس‌وجو را مشخص نموده است. لازم به ذکر است این مجموعه داده در آدرس (webazma.itrc.ac.ir) قابل دسترس است.

۴- روش‌های پیشنهادی

در این بخش به بررسی روش‌های پیشنهادی برای دسته‌بندی موضوعی پرس‌وجوها خواهیم پرداخت. ابتدا روش مبتنی بر بازیابی اطلاعات و سپس روش مدل زبانی بررسی می‌گردد.

۴-۱- روش مبتنی بر بازیابی اسناد مرتبط

روش پیشنهادی بر اساس روش‌های مختلف بازیابی اطلاعات و همچنین تکنیک رای‌دهی عمل می‌کند. مراحل انجام روش پیشنهادی برای دسته بندی یک پرس‌وجو به صورت زیر است:

- نمایه کردن مجموعه داده آموزش تهیه شده، توسط نرم افزار متن باز Terrier[6](یک نرم افزار متن باز برای بازیابی اطلاعات)
- بازیابی N سند مرتبط اول با پرس‌وجوی ورودی از مجموعه داده همشهری نمایه شده با استفاده از چهار روش مختلف بازیابی اطلاعات (BM25, BB2, LM , LemurTF_IDF)
- ترکیب اسناد بازیابی شده با استفاده از روش‌های مختلف ترکیب اطلاعات و استخراج M سند خیلی مرتبط با پرس-وجوی ورودی
- دسته‌بندی موضوعی پرس‌وجوها بر اساس دسته‌ی موضوعی M سند مرتبط استخراج شده

یکی از مراحل کلیدی روش ارائه شده برای دسته‌بندی پرس‌وجوها انتخاب M سند مرتبط با پرس‌وجو از مجموعه داده آموزش است. انتخاب درست این M سند در دقت روش دسته‌بندی بسیار تاثیر گذار است زیرا در صورتی که یک سند مرتبط وجود داشته باشد ولی جزو این M سند نباشد و یا اینکه سند نامرتبطی باشد که به صورت نادرست جزو این M سند قرار گرفته باشد دقت روش دسته‌بندی پایین می‌آید. لذا در این بخش سعی شده است این M سند با دقت مناسب‌تری بازیابی شود به صورتی که M سند خیلی مرتبط بازیابی شود. برای این کار از روش‌های مختلف ترکیب اطلاعات استفاده شده است. برای این کار ابتدا N سند مرتبط با پرس‌وجوی کاربر را به کمک چهار روش مختلف بازیابی اطلاعات بازیابی می‌کنیم و سپس به هر کدام از این اسناد یک امتیاز جدید نسبت می‌دهیم و در نهایت این اسناد بر اساس امتیاز جدید مرتبط شده و M سند مرتبط را از میان این لیست جدید بازیابی می‌کنیم. روش‌های زیر برای محاسبه‌ی امتیاز جدید اسناد و ترکیب اسناد بازیابی شده توسط روش‌های مختلف ارائه شده است:

روش اول : ترکیب بر اساس تعداد: امتیاز جدید در این روش بر اساس تعداد عمل می‌کند یعنی امتیاز یک سند برابر است با تعداد دفعاتی که یک سند توسط روش‌های بازیابی اطلاعات، بازیابی شده باشد لذا بر این اساس بیشترین امتیاز یک سند برابر ۴ است برای حالتی که یک سند توسط هر چهار روش بازیابی شده باشد.

روش دوم : ترکیب بر اساس تعداد و رتبه: مشابه حالت قبلی فقط در این حالت امتیاز جدید یک سند بر اساس رتبه و تعداد نسبت داده می‌شود. یعنی امتیاز جدید یک سند برابر است با مجموع عکس رتبه‌ی این سند در لیست مربوط به هر کدام از روش‌ها. در این حالت یک سند در صورتی امتیاز بالایی می‌گیرد که هم توسط روش‌های بیشتری از چهار روش بازیابی شده باشد و هم اینکه در هر کدام از این لیست‌ها رتبه‌ی بهتری داشته باشد.

روش سوم: ترکیب بر اساس تعداد و امتیاز: در این حالت امتیاز جدید یک سند بر اساس امتیاز و تعداد نسبت داده می‌شود. یعنی امتیاز جدید یک سند برابر است با مجموع امتیاز این سند در لیست مربوط به هر کدام از روش‌ها در تعداد دفعاتی که این سند توسط روش‌های مختلف بازیابی شده باشد. در این حالت یک سند در صورتی امتیاز بالایی می‌گیرد که هم توسط روش‌های بیشتری از چهار روش بازیابی شده باشد و هم اینکه در هر کدام از این لیست‌ها امتیاز بیشتری کسب کرده باشد.

حال به ازای هر پرس‌وجو، لیست M سند خیلی مرتبط، دسته‌ی موضوعی آنها و امتیاز مرتبط بودن هر سند، استخراج شده است. فرض کنید D مجموعه‌ی M سند مرتبط اول در رابطه با پرس‌وجو ورودی

[5] www.parsijoo.ir

[6] http://terrier.org/

باشد. برای هر عضو این مجموعه، d_n بیانگر شناسه سند، R_n رتبه‌ی سند، S_n امتیاز سند و C_n دسته‌ی مربوط به این سند می‌باشد. $C=\{C1,C2,...,Ck\}$ مجموعه‌ی دسته‌های موجود برای دسته‌بندی پرس‌وجوها باشد.

دسته‌بندی پرس‌وجو در واقع احتمال $P(Cj|Q)$ را محاسبه می‌کند. مطابق فرمول ۱ در واقع کلاسی که بیشترین مقدار این احتمال را داشته باشد به عنوان کلاس پرس‌وجو در نظر گرفته می‌شود.

$$C_{max} = \arg Max_{C_j \in C}\ P(Cj|Q) \qquad (1$$

برای محاسبه‌ی احتمال $P(Cj|Q)$ از فرمول ۲ استفاده شده است.

$$P(Cj|Q) = \sum_{m=1}^{|D|} P(Cj|d_m).P(d_m|Q) \qquad (2$$

که در این فرمول d بیانگر سند بازیابی شده می‌باشد. برای محاسبه‌ی $P(Cj \mid d)$ به صورت باینری عمل شده است و این احتمال در صورتی که دسته‌ی موضوعی سند C_j باشد برابر ۱ و در غیر این صورت صفر خواهد بود. اما برای محاسبه‌ی $P(d_m \mid Q)$ توابع مختلف در نظر گرفته شده است:

روش اول، تابع یکنواخت: در این حالت امتیاز $P(d\_m \mid Q)$ برای تمامی اسناد برابر ۱ در نظر گرفته شده است و تمامی اسناد بازیابی شده یک ارزش خواهند داشت. در این حالت در نهایت امتیاز $P(Cj \mid Q)$ برای دسته‌ی موضوعی C_j برابر تعداد اسناد با دسته‌ی موضوعی C_j در مجموعه اسناد مرتبط بازیابی شده D، خواهد بود.

روش دوم، تابع معکوس رتبه: در این حالت امتیاز $P(d\_m \mid Q)$ برای سند d برابر معکوس رتبه‌ی سند d در مجموعه اسناد D خواهد بود. در این حالت هر چه یک سند رتبه‌ی بهتری در لیست D داشته باشد تاثیر بیشتری در امتیاز نهایی $P(C_j|Q)$ خواهد داشت.

$$P(d_m|Q) = \frac{1}{R_m} \qquad (3$$

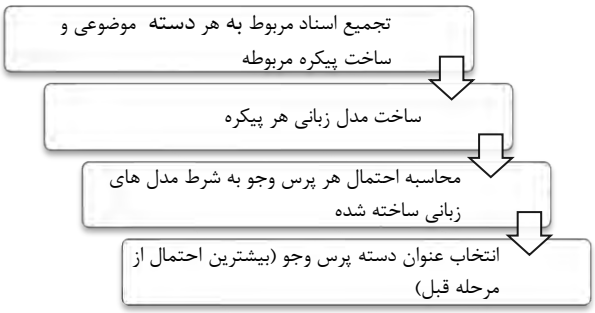
روش سوم، امتیاز مرتبط بودن: در این حالت امتیاز $P(d\_m \mid Q)$ برای سند d برابر امتیاز سند d در مجموعه اسناد D خواهد بود. در این حالت هر چه یک سند امتیاز بهتری در لیست D داشته باشد تاثیر بیشتری در امتیاز نهایی $P(Cj|Q)$ خواهد داشت.

$$P(d_m|Q) = s_m \qquad (4$$

۲-۴- روش مبتنی بر مدل زبانی

روش پیشنهادی دوم برای دسته‌بندی پرس‌وجوها استفاده از مدل زبانی n-gram است. به این صورت که سعی می‌شود برای اسناد مربوط به هر دسته در مجموعه داده‌ی آموزش تهیه شده یک مدل زبانی محاسبه گردد سپس احتمال پرس‌وجوی ورودی کاربر را برای هر مدل زبانی مربوط به هر دسته محاسبه نموده و از طریق آن دسته‌ی پرس‌وجو محاسبه گردد.

مراحل مختلف روش پیشنهادی در شکل ۲ نشان داده شده است. در ابتدا تمام اسناد واقع در مجموعه داده همشهری مربوط به یک دسته‌ی موضوعی، تجمیع شده و پیکره مربوط به آن دسته‌ی موضوعی ساخته می‌شود. سپس برای هر یک از پیکره‌ها که در واقع داده‌های آموزش برای آن دسته‌ی موضوعی هستند مدل زبانی ساخته می‌شود. برای ایجاد مدل زبانی، از مدل n-gram که با شمارش دنباله کلمات در پیکره بدست می‌آیند استفاده شده است. در روش استفاده شده برای هر پیکره مدلهای یک‌گرام(monogram)، دوگرام (bigram) و سه‌گرام(trigram) ساخته شده است.



شکل ۲: مراحل دسته بندی پرس و جو بر اساس مدل زبانی

در نهایت برای محاسبه‌ی دسته‌ی موضوعی مربوط به یک پرس‌وجو به صورت زیر عمل شده است:

فرض کنید پرس‌وجوی ورودی به صورت $Q=w_1w_2...w_m$ تعریف شده باشد که هر یک از w_i ، کلمات مربوط به پرس‌وجو را نشان می‌دهند. در این صورت احتمال $P(Q|Ci)$ برای هر یک از مدل‌های زبانی ساخته شده به صورت زیر محاسبه می‌شود

$$P(w_1\ w_2...w_m\ |C_i) = \prod_{i=1}^{m} P(w_i|w_{i-2}\ w_{i-1}) \qquad (5$$

وجود احتمال‌های صفر در مدل n-gram محاسبات را در عمل با مشکل مواجه می‌کند، زیرا باعث می‌شود احتمال بسیاری از پرس‌وجوها برابر با صفر گردد. برای حل این مشکل از روش‌های هموارسازی (smoothing) استفاده می‌شوند. در این مقاله ما از روش هموارسازی مبتنی بر Kneser-Ney [۲۳] که جز موثرترین روش‌های موجود برای هموارسازی است، استفاده شده‌است. با استفاده از تعاریف مدل زبانی و ساخت مدل Arpa برای هر دسته، برای هر پرس‌وجو بررسی می‌شود که آن پرس‌وجو با چه احتمالی توسط مدل زبانی مربوطه ساخته می‌شود. هر دسته‌ای که برای پرس‌وجو بیش‌ترین احتمال را داشته باشد، دسته آن پرس‌وجو را مشخص می‌کند.

۵- نتایج

در این بخش به بررسی عملکرد راهکارهای پیشنهادی برای دسته‌بندی خودکار پرس‌وجوها بر اساس موضوع خواهیم پرداخت. برای ارزیابی روش-های پیشنهادی از معیارهای استاندارد برای دسته‌بندی پرس‌وجوها که

شامل صحت[7]، دقت[8] و فراخوانی[9] است، استفاده شده است. معیار صحت نشان می‌دهد که چند درصد از پرس‌وجوها به درستی دسته‌بندی شده‌اند. معیار دقت، نشان می‌دهد که از بین پرس‌وجوهایی که الگوریتم در یک دسته قرار داده است، چه درصدی از آن‌ها به درستی متعلق به همان دسته می‌باشند و در نهایت فراخوانی که در واقع یک سنجه‌ی تمامیت است. این معیار برای یک دسته‌ی موضوعی خاص نشان می‌دهد از میان پرس‌وجوهایی که به آن دسته تعلق دارند، چه تعداد توسط روش پیشنهادی به درستی دسته‌بندی شده است.

روش پیشنهادی دو پارامتر روش ترکیب اطلاعات و نوع تابع امتیاز داشت که به ازای انتخاب توابع مختلف برای این دو پارامتر مختلف نتایج مختلفی حاصل گردید اما روشی که نتایج را بر اساس تعداد ترکیب کرده و سپس بر اساس امتیاز M سند مرتبط را بررسی می‌کند بهترین عملکرد را داشته است لذا در این بخش نتایج مربوط به آن گزارش می‌شود. جدول ۱ مقادیر معیارهای ارزیابی برای روش مبتنی بر روش بازیابی اسناد را نشان می‌دهد. همان طور که مشخص است روش پیشنهادی صحت بالایی دارد و میانگین صحبت برابر ۰,۹۲۹ بوده است و این به دلیل توزیع نامتعادل پرس‌وجوهای مربوط به دسته‌های موضوعی مختلف می‌باشد لذا معیار صحت، معیار مناسب برای ارزیابی نخواهد بود. برای حل این مشکل از معیار دقت و فراخوانی استفاده شده است. نتایج حاصل شده حاکی از این دارد که دقت روش پیشنهادی برای دسته‌های موضوعی مختلف، متفاوت می‌باشد که در جدول ۱ بر اساس رنگ تفکیک شده است. روش پیشنهادی برای دسته‌های مذهبی و سلامت دقت بسیار خوبی داشته و میانگین دقت برابر ۰,۸۷۸۵ بوده است و برای دسته‌های هنر، ورزشی، اقتصاد، علمی و فناوری اطلاعات دقت نسبتا خوبی داشته و میانگین دقت برابر ۰,۶۱۸۲ بوده است. اما میانگین دقت برای دسته‌های سیاسی، فرهنگ و محیط زیست مناسب نبوده است و برابر ۰,۳۶۹ بوده است. دسته‌های اجتماعی و گردشگری نیز بسیار بد عمل نموده است. در ادامه به تحلیل عمیق‌تر نتایج پرداخته شده است.

روش پیشنهادی برای دسته‌های موضوعی سلامت، سیاسی، فناوری اطلاعات و ورزشی فراخوانی خوبی داشته است و میانگین فراخوانی برای این دسته‌ها برابر ۰,۷۷۴۲ بوده است و برای دسته‌های محیط زیست، اقتصاد، فرهنگ، مذهبی، اجتماعی و هنر فراخوانی نسبتا خوبی داشته و میانگین ۰,۵۷ بوده است اما برای دسته‌ی گردشگری و علمی عملکرد مناسبی نداشته است.

جدول ۱: نتایج معیارهای ارزیابی برای روش مبتنی بر بازیابی اسناد مرتبط

دسته‌ی موضوعی	Accuracy	Precision	Recall
مذهبی	0.897	0.898	0.512
سلامت	0.959	0.859	0.825
هنر	0.888	0.651	0.527
ورزشی	0.972	0.619	0.701
اقتصاد	0.939	0.610	0.635
علمی	0.896	0.610	0.139
فناوری اطلاعات	0.938	0.601	0.763
سیاسی	0.877	0.464	0.808
فرهنگ	0.939	0.310	0.563
محیط زیست	0.983	0.333	0.656
اجتماعی	0.888	0.176	0.525
گردشگری	0.972	0	0.000

برای بررسی دقیق‌تر نتایج از ماتریس درهم‌ریختگی[10] استفاده شده است. نتایج حاصل شده نشان می‌دهد که کلاس‌های سیاسی، اجتماعی، فرهنگ و فناوری اطلاعات بیشترین FP را داشته‌اند به عبارتی یعنی برای این کلاس‌ها طبقه بند به اشتباه تعداد زیادی پرس‌وجو را مربوط به این کلاس‌ها برچسپ زده در حالی که کلاس واقعی آنها این کلاس‌ها نبوده است. دسته سیاسی بیشترین FP را داشته است با بررسی نتایج مشخص گردید که از انجایی که در مجموعه داده آموزش اخبار سیاسی موضوعات مختلفی داشته‌اند به عنوان مثال ممکن است یک خبر اقتصادی که جنبه‌ی سیاسی نیز داشته است منتشر شده باشد لذا روش پیشنهادی در یافتن دسته‌ی موضوعی سیاسی دچار مشکل شده است. دسته‌ی بعدی اجتماعی بوده است که بعد از سیاسی بیشترین تعداد FP را داشته است دلیل این نیز مشابه دسته‌ی سیاسی است. اسناد مربوط به دسته‌ی اجتماعی معمولا موضوعات مختلفی دارند و از کلمات مختلفی در آنها استفاده می‌شود که معمولا در دسته‌های دیگر نیز استفاده می‌شود لذا روش پیشنهادی ما نیز برای پرس‌وجوهای این دسته دچار مشکل شده است. از طرفی دسته‌های مذهبی، علمی، ورزشی، محیط زیست و سلامت کمترین FP را داشته‌اند و دلیل این نیز به خاطر خاص بودن این دسته‌های موضوعی است که معمولا اشتراک کمتری با دسته‌های دیگر دارند لذا منطقی است که FP کمتری داشته باشند. لذا یک نتیجه‌گیری کلی می‌تواند به این صورت باشد که روش پیشنهادی برای دسته‌های موضوعی خاص که اشتراک موضوعاتی کمتری با سایر دسته‌ها دارند FP کمتری دارد ولی برای دسته‌هایی مانند سیاسی و یا اجتماعی FP بیشتری داشته است.

اما روش پیشنهادی برای دسته‌ی گردشگری نتایج نا امید کننده‌ای داشته است لذا برای بررسی دلیل این امر به تحلیل پرس‌وجوهای مربوط به این دسته پرداخته شده است. علل عدم توانایی روش پیشنهادی در شناسایی پرس‌وجوهای دسته "گردشگری" را می‌توان در موارد زیر خلاصه کرد:

[7] Accuracy
[8] Precision
[9] Recall
[10] Confusion Matrix

- کم بودن مجموعه داده ی آموزش: در مجموعه داده‌ی آموزش گردهآوری شده تعداد اسناد مربوط به این دسته بسیار کم بوده است و این باعث شده است که در پاسخ به پرس‌وجوهای مربوط به این دسته اسناد کمتری بازیابی شوند و این نتایج را تحت تاثیر قرار داده است.
- متفاوت بودن محتوای همشهری با پرس‌وجوهای کاربر: با توجه به اینکه مجموعه داده همشهری یک مجموعه داده خبری است لذا بسیاری از کلمات مربوط به پرس‌وجوهای گردشگری ممکن است که در مجموعه داده همشهری کمتر استفاده شده باشد به عنوان مثال به دلیل خاص بودن پرس‌وجوهای گردشگری که اکثرا اسم یک منطقه یا یک شهر یا یک جای دیدنی است لذا این اسامی کمتر در اسناد همشهری تکرار شده است و این دقت دسته‌بندی را تحت تاثیر قرار می‌دهد.
- اشتراک داشتن برخی اسامی مانند نام شهرها با دسته‌ی سیاسی و اقتصادی: با توجه به اینکه در اسناد سیاسی و یا اقتصادی از اسامی شهرها و استان‌ها زیاد استفاده می‌شود لذا در بسیاری از حالات روش پیشنهادی به اشتباه دسته‌ی یک پرس‌وجوی گردشگری را سیاسی یا اقتصادی برچسپ زده است.

به عنوان مثال برای پرس‌وجوی "معنی قارلوق" روش پیشنهادی به دلیل وجود کلمه‌ی "معنی" در پرس‌وجو دسته‌ی آن را "فرهنگ" در نظرگرفته است و کاربر انسانی چون قارلوق نام یک روستا از توابع ساوه است آن را گردشگری برچسپ زده است و یا پرس‌وجوی عکس هایی از پاناما روش پیشنهادی به دلیل وجود کلمه‌ی پاناما که بیشتر در اسناد سیاسی وجود دارد دسته‌ی سیاسی را انتخاب کرده است.

یکی دیگر از دسته‌هایی که روش پیشنهادی نتایج بدی برای آن داشته است دسته‌ی اجتماعی بوده است. که در این بخش به تحلیل آن پرداخته شده است. یکی از دلایل اصلی پایین بودن دقت برای پرس-وجوهای دسته "اجتماعی"، این است که دسته مذکور با سایر دسته‌ها اشتراک دارد و نمی‌توان مرز مشخصی بین آنها تعیین نمود و حتی برچسب‌های تعیین شده برای برخی از پرس‌وجوهای این دسته به اشتباه توسط کاربر برچسپ اجتماعی خورده شده است. در جدول۲ مثال‌هایی از پرس‌وجوهای این دسته که روش پیشنهادی به اشتباه در دسته‌ای غیر از اجتماعی قرار داده است.

جدول ۲: مثال‌هایی از پرس‌وجوهای دسته اجتماعی

پرس‌وجو	دسته‌ی پیش‌بینی شده
راه های تغییر مسیر زندگی	محیط زیست
ستایش درنمایشگاه کتاب	هنر
کسری خدمت به مناسبت دهه فجر	سیاسی
افزایش زاد و ولد و ازدواج	سلامت
نیروی انتظامی سیستان وبلوچستان	سیاسی
نمایشگاه دستاردهای عشایر کشور	هنر
سیل چالوس ۹۴	محیط زیست
بیانات رهبری درباره کرامت خانواده	سیاسی

همانطور که در این جدول قابل مشاهده است، پرس‌وجویی مانند "نمایشگاه دستاوردهای عشایر کشور" مربوط به شاخه‌ی "هنر" می‌شود اما کاربران آن را به اشتباه "اجتماعی" برچسپ زده‌اند. اما پرس‌وجویی مانند"راه های تغییر مسیر زندگی" به دلیل وجود دو کلمه "راه" و "مسیر" که جز کلمات خاص دسته "محیط زیست" هستند به اشتباه توسط روش پیشنهادی در این دسته قرار گرفته‌اند. به طور کلی عوامل تشخیص اشتباه پرس‌وجوها توسط روش پیشنهادی را می‌توان در سه مورد خلاصه نمود:

- مشکلات مربوط به داده آموزشی:
  - رسمی بودن داده و نبود اطلاعات برای پرس‌وجوهای غیر رسمی
  - کمبود داده‌های مربوط به برخی از دسته‌ها مانند گردشگری
- اشتباهات مربوط به برچسپ زدن کاربران ارزیاب
- همپوشانی بین دسته‌های مختلف برای پرس‌وجو

روش دوم مبتنی بر مدل زبانی بود و از مدل‌های زبانی یک‌گرام، دوگرام و سه گرام استفاده شد که در این بین بهترین عملکرد مربوط به مدل زبانی یک گرام بود و نتایج مربوط به آن در جدول ۳ گزارش شده است.

جدول ۳ : نتایج معیارهای ارزیابی برای روش مبتنی بر مدل زبانی

موضوع	Accuracy	Precision	Recall
هنر	0.879	0.661	0.440
اقتصاد	0.944	0.627	0.619
سلامت	0.94	0.811	0.773
اجتماعی	0.943	0.305	0.423
مذهبی	0.878	0.737	0.557
فناوری اطلاعات	0.902	0.408	0.697
سیاسی	0.929	0.743	0.556
فرهنگ	0.941	0.331	0.564
محیط زیست	0.952	0.097	0.384
علمی	0.918	0.67	0.3
گردشگری	0.944	0.21	0.295
ورزشی	0.933	0.331	0.747

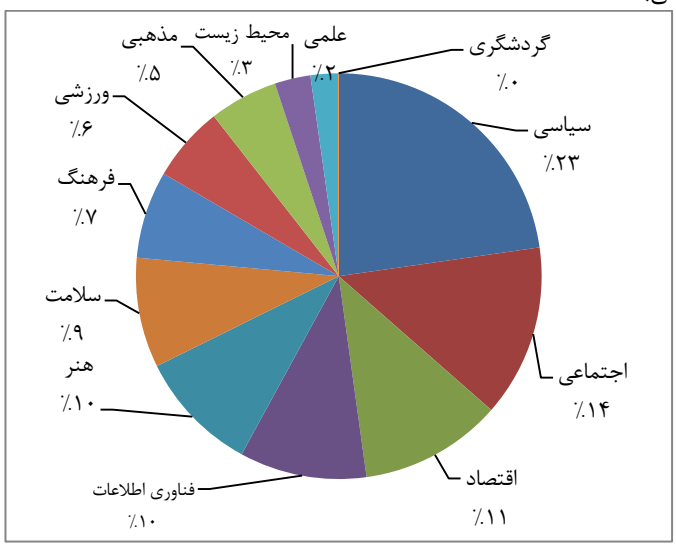
همانطور که در این جدول قابل مشاهده است روش مبتنی بر مدل زبانی برای دسته‌های "سلامت"، "هنر" و "مذهبی" عملکرد بسیار قابل قبولی داشته است. اگر چه کمترین دقت در این روش مربوط به دسته "گردشگری" است اما برای این دسته نسبت به روش قبلی بهتر عمل کرده

است. در این روش متوسط صحت برابر ۰,۹۲۵ بوده که نزدیک به روش قبل است اما میانگین دقت آن با ۰,۴۹۴ کمتر بوده است. نکته‌ی قابل توجه این است که روش مدل زبانی از انجایی که بر اساس فرکانس تکرار کلمات عمل می‌کند لذا توزیع مجموعه داده آموزشی باید برای کلاس‌های مختلف یکنواخت باشد. اما در مجموعه داده‌ی تهیه شده برای برخی از دسته‌ها مانند گردشگری و یا محیط زیست تعداد اسناد کمتری وجود دارد لذا روش مبتنی بر مدل زبانی برای این موارد خوب عمل نمی‌کند. اما همان طور که مشخص است برای دسته‌های سلامت، هنر، مذهبی و سیاسی عملکرد قابل قبولی داشته است.

۲-۵- تحلیل لاگ بر اساس دسته‌ی موضوعی پرس‌وجوها

برای تحلیل های این بخش یک نمونه ۱۰۰ هزارتایی به صورت کاملا تصادفی از پرس وجوهای لاگ موتور جستجوی بومی پارسی‌جو استخراج شده و با استفاده از روش پیشنهادی مبتنی بر بازیابی اطلاعات این پرس-وجوها دسته بندی شده است.

شکل ۳ توزیع پرس‌وجوهای کاربران را بر اساس موضوعات تعیین شده نشان می‌دهد. نتایج حاصل شده حاکی از این دارد که کاربران فارسی زبان بیشتر علاقه‌مند به موضوعات سیاسی، اجتماعی و اقتصاد بوده‌اند و این سه دسته سهم اعظمی از پرس‌وجوها را به خود اختصاص داده‌اند. موضوعات فناوری اطلاعات، سلامت و هنر در جایگاه بعدی قرار گرفته‌اند. همانطور که در این شکل مشخص است بیشترین سهم در پرس‌وجوهای کاربران مربوط به دسته‌ی "سیاسی" (۲۳٪) و سپس دسته‌ی "اجتماعی" (۱۴٪) می‌باشد. در این بین کمترین سهم پرس‌وجوها مربوط به دسته "گردشگری" (نزدیک به ۰٪) و دسته "علمی" (۲٪) می‌باشد



شکل ۳: توزیع موضوعی پرس‌وجوهای لاگ موتور جستجو

۶- نتیجه گیری و پیشنهادات

در این مقاله دو روش برای دسته‌بندی موضوعی پرس‌وجوی زبان فارسی ارائه شد. در روش اول از تکنیک‌های بازیابی اطلاعات استفاده شده است در حالی که در روش دوم مدل‌های زبانی برای دسته‌بندی بکار گرفته شده است. نتایج حاصل شده حاکی از این داشت که روش مبتنی بر بازیابی اسناد مرتبط بهتر از مدل زبانی عمل نموده است.

دقت روش‌های پیشنهادی به دقت مجموعه داده‌ی آموزش وابسته است لذا برای کارهای آینده مدنظر است که این مجموعه داده بهبود یابد و وب‌سایت‌های تخصصی برای هر دسته‌ی موضوعی خزش و تهیه گردد. همچنین در این پروژه یک نمونه ۱۰۰ هزار تایی از پرس-وجوهای لاگ دسته‌بندی و تحلیل گردید در آینده سعی خواهد شد که تحلیل‌های جامع‌تر و دقیق‌تری در رابطه با یک نمونه بزرگ‌تر از پرس-وجوها انجام شود. همچنین تحلیل پرس‌وجوهای مربوط به یک دسته‌ی موضوعی خاص مانند پرس‌وجوهای مذهبی، سیاسی و یا سلامت می‌تواند موضوع پژوهشی مناسبی باشد.

مراجع

[1] Jiang, Di, and Lingxiao Yang. "Query intent inference via search engine log." Knowledge and Information Systems (2016): 1-25.
[2] Palotti, João, et al. "How users search and what they search for in the medical domain." Information Retrieval Journal (2016): 1-36.
[3] Dumais, Susan, et al. "Understanding user behavior through log data and analysis." Ways of Knowing in HCI. Springer New York, 2014. 349-372.
[4] Lucchese, Claudio, et al. "Discovering tasks from search engine query logs." ACM Transactions on Information Systems (TOIS) 31.3 (2013): 14.
[5] S. M. Beitzel, E. C. Jensen, O. Frieder, D. Grossman, D. D. Lewis, A. Chowdhury, et al., "Automatic web query classification using labeled and unlabeled training data," in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005, pp. 581-582.
[6] AleAhmad, Abolfazl, et al. "Hamshahri: A standard Persian text collection." Knowledge-Based Systems 22.5 (2009): 382-387.
[7] Ren, Pengjie, et al. "Understanding temporal intent of user query based on time-based query classification." Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg, 2013. 334-345.
[8] Gupta, Dhruv, and Klaus Berberich. "Temporal query classification at different granularities." International Symposium on String Processing and Information Retrieval. Springer International Publishing, 2015.
[9] Joho, Hideo, et al. "Overview of NTCIR-11 Temporal Information Access (Temporalia) Task." NTCIR. 2014.
[10] Simion, Bogdan, Suprio Ray, and Angela Demke Brown. "Surveying the landscape: an in-depth analysis of spatial database workloads." Proceedings of the 20th International Conference on Advances in Geographic Information Systems. ACM, 2012.
[11] Simion, Bogdan. Analyzing and Improving the Performance of Spatial Database Processing. Diss. University of Toronto, 2015.
[12] Völske, Michael, et al. "What Users Ask a Search Engine: Analyzing One Billion Russian Question Queries." Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015.
[13] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen, "Building bridges for web query classification," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006, pp. 131-138.
[14] A. Z. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang, "Robust classification of rare queries using web

[19] Shen, Dou, et al. "Building bridges for web query classification." Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006.

[20] Jansen, Bernard J., and Danielle Booth. "Classifying web queries by topic and user intent." CHI'10 Extended Abstracts on Human Factors in Computing Systems. ACM, 2010.

[21] Hernández, I., et al. "A simple model for classifying web queries by user intent." 2nd Spanish Conference on Information Retrieval, CERI-2012. 2012.

[22] Laclavík, Michal, et al. "Search query categorization at scale." Proceedings of the 24th International Conference on World Wide Web. ACM, 2015.

[23] James, Frankie. "Modified kneser-ney smoothing of n-gram models." Research Institute for Advanced Computer Science, Tech. Rep. 00.07 (2000).

knowledge," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 231-238.

[15] S. M. Beitzel, E. C. Jensen, O. Frieder, D. D. Lewis, A. Chowdhury, and A. Kolcz, "Improving automatic query classification via semi-supervised learning," in Data Mining, Fifth IEEE international Conference on, 2005, p. 8 pp.

[16] X. Li, Y.-Y. Wang, and A. Acero, "Learning query intent from regularized click graphs," in Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008, pp. 339-346.

[17] H. Cao, D. H. Hu, D. Shen, D. Jiang, J.-T. Sun, E. Chen, et al., "Context-aware query classification," in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, pp. 3-10.

[18] Li, Ying, Zijian Zheng, and Honghua Kathy Dai. "KDD CUP-2005 report: Facing a great challenge." ACM SIGKDD Explorations Newsletter 7.2 (2005): 91-99.