بررسی و تحلیل دادگان تشخیص صفحات اسپم درمحیط وب بر اساس نظریه گراف

مهدیه رعیتی[۱]، امیرحسین کیهانی‌پور[۲]

[۱] کارشناسی مهندسی کامپیوتر، دانشکده مهندسی دانشکدگان فارابی دانشگاه تهران
mahdieh.raeyati@ut.ac.ir
[۲] استادیار گروه مهندسی کامپیوتر، دانشکده مهندسی دانشکدگان فارابی دانشگاه تهران
keyhanipour@ut.ac.ir

چکیده

نظریه گراف که به مدل‌سازی روابط موجود بین عناصر مختلف مسئله مورد بررسی می‌پردازد، ابزار مفیدی را برای ساده‌سازی بخش‌های یک سیستم فراهم می‌کند. پیچیده‌تر شدن مسائل دنیای پیرامونی، به‌کارگیری نظریه گراف را به یک ضرورت تبدیل نموده است. این مقاله قصد دارد مجموعه دادگان عرضه شده به‌منظور شناسایی تشخیص صفحات اسپم در محیط وب را از منظر گراف مورد بررسی قرار دهد. برای این منظور، ابتدا گراف شباهت ویژگی‌های مجموعه داده، ایجاد می‌شود و سپس، گراف حاصل به‌لحاظ شاخص‌های ساختاری مختلف، مورد بررسی قرار خواهد گرفت. برای ارزیابی روش پیشنهادی، گراف شباهت به‌ازای دو دسته از ویژگی‌های متنی و پیوندی به‌ازای مجموعه داده WEBSPAM-UK2007 ایجاد گردید و بر اساس شاخص‌های فوق مورد مقایسه تحلیلی قرار گرفت. نتایج به‌دست آمده نشان می‌دهد که علیرغم بزرگ‌تر بودن و تراکم نسبی بالاتر گراف شباهت ویژگی‌های مبتنی بر متن، نسبت به گراف شباهت ویژگی‌های مبتنی بر پیوند، بر اساس شاخص‌های ضریب خوشه‌بندی، گراف شباهت ویژگی‌های مبتنی بر پیوند، انسجام نسبی بیشتری را دارا می‌باشد. این موضوع با توجه به اندازه نسبی بزرگ‌ترین مؤلفه همبند نیز تأیید می‌شود. این رویکرد، امکان مقایسه تحلیلی دادگان مختلف را فراهم می‌آورد. علاوه بر آن، می توان از نتایج این پژوهش به‌منظور طراحی دادگان جدید نیز استفاده نمود.

کلمات کلیدی: نظریه گراف، دادگان تشخیص صفحات اسپم، ویژگی‌های گراف، شاخص کندال.

۱ مقدمه

بازیابی اطلاعات وب با چالش‌های متعددی مواجه است. یکی از این چالش‌ها که بعضاً عدم توجه به آن، منجر به کاهش کیفیت نتایج جستجو و در نتیجه عدم رضایت کاربر می‌شود، عدم توجه به موضوع تشخیص و حذف

صفحات اسپم[1] از فرآیند بازیابی اطلاعات می‌باشد. بطور خلاصه، صفحات اسپم، شامل محتوای غیر مفید یا هرزی هستند که یا اصولا امکان تامین نیاز اطلاعاتی کاربر از طریق آنها فراهم نیست و یا اینکه تامین نیاز کاربر را با دشواری‌های مختلف، مواجه می‌کنند. این قبیل سایت‌ها و صفحات، غالبا با هدف دستکاری رتبه‌بندی نتایج جستجو و جلب توجه کار به صفحات هدف، طراحی می‌شوند و ممکن است محتوای قابل مشاهده توسط کاربر با محتوای واقعی این صفحات، متفاوت باشد. به منظور شناسایی صفحات اسپم، مجموعه‌های داده مختلفی طراحی شده است که از نظر ویژگی‌ها و مشخصات، بسیار متفاوت می‌باشند. در این پژوهش در نظر است با مدل‌سازی این دادگان با استفاده از نظریه گراف، چارچوبی برای مقایسه تحلیلی این دادگان فراهم آید. نظریه گراف بواسطه امکان مدل‌سازی روابط پیچیده بین عناصر مساله مورد بررسی، مورد اقبال وسیع محققان قرار گرفته است. بصورت خلاصه، یک گراف (شبکه) در ساده‌ترین شکل خود مجموعه‌ای از گره‌ها است که از طریق یالهای ارتباط دهنده، به یکدیگر متصل شده‌اند. این یال‌های عملا بیانگر روابط بین گره‌ها. در این مقاله به بررسی مجموعه دادگان مربوط به تشخیص صفحات اسپم با استفاده از این روش مدل‌سازی، می‌پردازیم. گره‌های این گراف، معادل ویژگی‌های عرضه شده در مجموعه دادگان مورد بررسی هستند و یالهای آن، میزان شباهت ویژگی‌ها را نشان می می‌دهند. در واقع با تبدیل مجموعه داده به گراف ویژگی، ویژگی‌های مهم و تعیین کننده و همچنین ارتباط بین ویژگی‌ها، به طور دقیق و مؤثر، قابل مشاهده خواهند بود. مراحل کلی روش پیشنهادی به شرح زیر خواهد بود:

- تعیین گره‌ها و یال‌های گراف
- تعیین معیار وزن‌دهی یال‌ها (شاخص کندال)
- ایجاد گراف ویژگی با استفاده از ثابت sigma
- تحلیل گراف و بررسی ویژگی‌های آن (اعم از توزیع درجه، معیارهای مرکزیت، قطر شبکه و ضرایب خوشه‌بندی محلی و سراسری)

در ادامه این نوشتار، ابتدا به مرور کاربردهای نظریه گراف در حوزه بازیابی اطلاعات وب، خواهیم پرداخت و سپس، چارچوب نظری روش پیشنهادی، بیان خواهد شد. پس از آن، نتایج بدست آمده از اجرای روش پیشنهادی روی یک مجموعه دادگان شناسایی صفحات اسپم و نیز تحلیل این نتایج ارائه می‌شود. در پایان نیز جمع‌بندی این پژوهش و ارائه رویکردهای توسعه آن، مطرح می‌گردد.

۲ مروری بر کارهای دیگران

نظریه گراف بدلیل غنای نظری و نیز قابلیت مدلسازی روابط پیچیده بین عناصر مساله، بصورت گسترده در حوزه بازیابی اطلاعات مورد استفاده قرار گرفته است. از جمله کاربردهای جالب توجه این نظریه در بازیابی اطلاعات وب، می‌توان به موضوع تعبیه گراف[2] به منظور طراحی الگوریتم‌های رتبه‌بندی کارآمد اشاره

[1] Spam
[2] Graph Embedding

نمود [۱]-[۴]. رویکرد دیگری که بر اساس کاربرد نظریه گراف، توسعه یافته است، مربوط به بکارگیری شبکه پیچشی گراف[۳] در طراحی روش‌های کارآمد در زمینه اطلاعات و به خصوص در مورد رتبه‌بندی بازیابی نتایج جستجوهای کاربران می‌باشد [۵]-[۹]. در زمینه بازیابی معنایی اطلاعات[۴] نیز از توانمندی‌های نظریه گراف برای بررسی روابط پیچیده بین عناصر داده، در قالب تولید گراف دانش[۵] استفاده شده است. بر این اساس، امکان ارتقای کیفیت بازیابی اطلاعات به ویژه در زمینه رتبه‌بندی نتایج جستجو، نسبت به روش‌های کلاسیک، فراهم خواهد آمد. از جمله این پژوهش‌ها می‌توان به [۱۰]-[۱۳] اشاره نمود. در [۱۴]، [۱۵] ایده تولید گراف شباهت ویژگی‌های دادگان به منظور بدست آوردن ویژگی‌های گرافی مطرح شده است و از این ویژگی‌ها به منظور ارائه یک روش یادگیری رتبه‌بندی[۶] استفاده شده است. از سوی دیگر در [۱۶] از نظریه گراف به منظور مقایسه تحلیلی دادگان موجود در حوزه یادگیری رتبه‌بندی استفاده شده است. در این روش نیز گراف شباهت ویژگی‌های موجود در مجموعه داده، به عنوان بازنمایی ثانویه از مجموعه دادگان مورد بررسی، تولید شده و به ازای کل گراف مجموعه‌ای از ویژگی‌های ساختاری استخراج گردیده است. بدین ترتیب، چارچوبی برای مقایسه مشخصات این دادگان فراهم آمده است.

در پژوهش فعلی با الهام از سه مقاله اخیر، از همین رویکرد برای تهیه چارچوبی برای مقایسه دادگان تشخیص صفحات اسپم در محیط استفاده شده است که بر اساس نظریه گراف، طراحی شده است. مشابه پژوهش‌های قبلی، با استفاده از روش پیشنهادی، امکان مقایسه تحلیلی انواع دادگان حوزه شناسایی صفحات اسپم، فراهم می‌شود.

۳ معرفی مجموعه داده

مجموعه داده WEBSPAM-UK2007[۷] بر اساس خزش[۸] صورت گرفته از دامنه uk. بدست آمده است و شامل فهرستی از سایت‌ها است که بصورت اسپم یا غیر اسپم، برچسب‌گذاری شده‌اند. مجموعه ویژگی uk-2007-05.link_based_features، شامل ویژگی‌های مبتنی بر لینک برای میزبان‌ها است که هم در صفحه اصلی و هم در صفحه با حداکثر رتبه صفحه[۹] در هر میزبان اندازه‌گیری می‌شود. ویژگی‌هایی نظیر درجه، درجه خروجی، رتبه صفحه، ضریب طبقه‌بندی، رتبه اعتماد ، رتبه صفحه کوتاه شده و ... .

[3] Graph Convolution Network
[4] Semantic Information Retrieval
[5] Knowledge Graph
[6] Learning to Rank
[7] https://chato.cl/webspam/datasets
[8] Crawl
[9] maximum page rank

۴ ایجاد گراف ویژگی

۱.۴ تعیین گره‌ها و یال‌ها

گراف ویژگی بر اساس ویژگی‌های موجود در مجموعه داده (ستون های مجموعه داده) ایجاد می‌شود؛ بنابراین گره‌ها، نام ستون‌های مجموعه داده هستند. پس از حذف دو ستون اول شامل شناسه و نام میزبان، - که ویژگی‌های غیر عددی و غیر مرتبط با سایر ویژگی‌ها هستند - بین هر دو جفت گره، یک یال ایجاد می‌کنیم. هر یال نشان دهنده ارتباط ویژگی‌های نقاط انتهایی خودش است. بنابراین در پایان این مرحله، یک گراف کامل خواهیم داشت.

۲.۴ تعیین معیار وزن‌دهی یال‌ها (شاخص کندال)

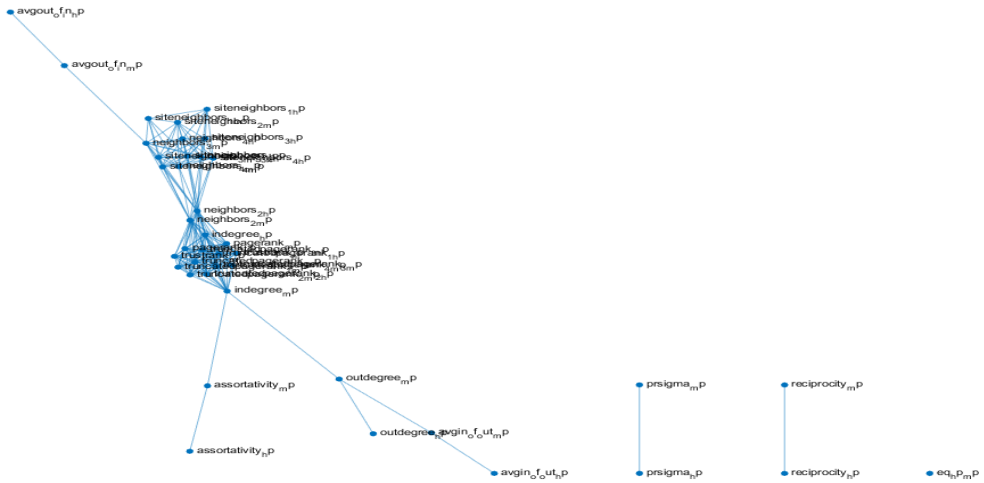
شاخص کندال، یک شاخص مناسب برای سنجش میزان ارتباط و همبستگی دو متغیر می‌باشد. لذا می‌توانیم برای ایجاد گراف ویژگی، از این آماره استفاده کنیم. شاخص کندال، عددی در بازه $[-۱, ۱]$ است. هرچه این عدد به یک یا منفی یک نزدیکتر باشد، میزان هماهنگی یا ناهماهنگی دو متغیر، بیشتر است؛ یعنی دو متغیر، تاثیر زیادی روی هم می‌گذارند و موجب تشدید یا تضعیف یکدیگر می‌شوند. بنابراین برای مشخص کردن میزان اهمیت و سنگین بودن یک یال، کافیست شاخص کندال را برای دو ستون متناظر دو گره در مجموعه داده ، محاسبه کنیم و قدر مطلق این مقدار را به عنوان وزن آن در نظر بگیریم . لذا هرچه وزن یالی به یک نزدیکتر باشد به معنی ارتباط قوی گره‌های انتهایی آن و هرچه به صفر نزدیکتر باشد به معنای ارتباط ضعیف‌تر است.

۳.۴ حذف یال‌های اضافی با استفاده از ثابت سیگما

گراف ایجاد شده در مرحله قبل، حاوی یال‌های اضافی است. ثابت سیگما، محدوده لازم برای حذف یا عدم حذف یال های گراف را مشخص می‌کند. اگر مقدار سیگما برابر با n باشد، تمام یال‌هایی که وزنشان از n کم‌تر باشد حذف خواهند شد. به همین ترتیب، هرچقدر ثابت سیگما بزرگ‌تر باشد، فیلتر سنگین‌تری روی گراف اعمال خواهد شد و فقط اتصال گره‌هایی که ارتباط بسیار قوی‌تری دارند حفظ خواهد شد. در این مقاله، مقدار سیگما، ۰٫۵ در نظر گرفته شده است. زیرا مقادیر کم‌تر از ۰٫۵، همبستگی اندکی دارند که بررسی آن ها، ضرورتی ندارد.

۵ تحلیل گراف

این گراف، شامل ۴۱ گره و ۲۰۲ یال است. میانگین درجه آن ۹٫۸ و میانگین وزن یال‌ها ۰٫۶۷ می‌باشد. همان‌طور که در شکل ۱ قابل مشاهده است، گراف از ۴ مؤلفه عمده تشکیل شده است به‌طوری‌که بزرگترین مؤلفه، ۰٫۸۷ = ۳۶/۴۱ گره‌ها را در بر دارد.



شکل ۱: گراف ویژگی استخراج شده از مجموعه ویژگی

۱.۵ توزیع درجه

در نمودار ۱، هیستوگرام توزیع درجه گراف نشان داده شده است. همانطور که در نمودار ۱ پیداست، توزیع درجه، از الگوی قانون توان پیروی نمی‌کند؛ زیرا اکثریت گره‌ها، درجه پایین ندارند و تعداد گره‌های درجه بالا، نسبتا زیاد است و به طور کلی، رفتار نمودار، نزولی (یا صعودی) نیست.

۲.۵ ضرایب خوشه‌بندی

فرمول محاسبه ضریب خوشه‌بندی برای کل شبکه به صورت زیر است [۱۷]:

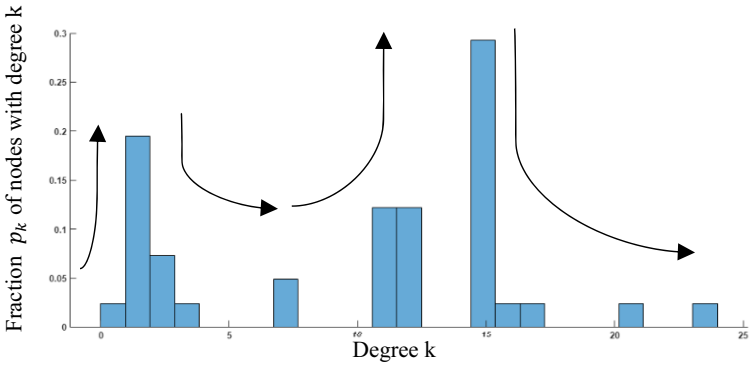
$$c = \frac{\text{۳ × تعداد مثلث‌ها در شبکه}}{\text{تعداد رئوس سه‌گانه متصل}} \quad (۱)$$

که در آن «سه‌گانه متصل» به‌معنای یک رأس منفرد با دو یال متصل به خودش است به‌گونه‌ای که اگر یال سومی به آنها اضافه شود، مثلث تشکیل خواهد شد [۱۸].

روش دوم: یک تعریف جایگزین از ضریب خوشه‌بندی، که به طور گسترده نیز استفاده می‌شود ، تعریف ضریب خوشه‌بندی محلی به صورت زیر است [۱۸]:

$$c_i = \frac{\text{تعداد مثلث‌های متصل به رأس } i}{\text{تعداد سه‌گانه با مرکز رأس } i} \quad (۲)$$

برای رئوس با درجه ۰ یا ۱ که صورت و مخرج هر دو صفر هستند، $c_i = ۰$ خواهد بود. سپس ضریب



نمودار ۱. هیستوگرام توزیع درجه گراف

خوشه‌بندی برای کل شبکه، میانگین مقادیر c_i خواهد بود [۱۸]:

$$C = \frac{۱}{n} \sum_i c_i \qquad (۳)$$

در این گراف، ضریب خوشه‌بندی سراسری با استفاده از معادله ۱ مقدار ۰٫۲۹۰۲ و با استفاده از معادله ۳ مقدار ۰٫۵۹۰۷۴ است.

همان‌طور که مشخص است، میانگین ضریب خوشه‌بندی محلی برای درجات ۱۵ تا ۲۴ اکیداً نزولی است. به‌طور کلی در اکثر شبکه‌های دنیای واقعی انتظار بر این است که گره‌های متعلق به مؤلفه‌های کوچک، درجات پایینی داشته باشند؛ زیرا گره‌های آن‌ها، محدوده انتخاب کوچک‌تری برای اتصال به گره‌های دیگر دارند؛ در حالی که مؤلفه‌های بزرگتر می‌توانند درجه بالاتری داشته باشند. با این وجود، معمولاً ضریب خوشه‌بندی محلی گره‌ها در مؤلفه‌های کوچک، بزرگتر است، زیرا هر مؤلفه ، به‌صورت جدا از بقیه شبکه و به‌عنوان یک شبکه کوچک‌تر عمل می‌کند [۱۹].

۳.۵ معیارهای مرکزیت

۱.۳.۵ مرکزیت درجه

گره‌های neighbors_2_hp و neighbors_2_mp در این گراف، بیشترین درجه‌ها را دارند (به ترتیب ۲۴ و ۲۱). بنابراین بیشترین مرکزیت درجه را خواهند داشت. بدیهی است که هیستوگرام مربوط به این مرکزیت، دقیقاً مشابه هیستوگرام توزیع درجه آن است.

۲.۳.۵ مرکزیت بینابینی

در این پژوهش، برای محاسبه مرکزیت بینابینی، از معکوس وزن یال‌ها استفاده شده است؛ زیرا وزن هر یال نشان دهنده میزان وابستگی و هماهنگی گره‌های دو سر آن است و هر چقدر معکوس این مقدار کم‌تر باشد، احتمال بالا رفتن مرکزیت بینابینی بیشتر است.

همانطور که انتظار می‌رفت، گره‌هایی که بخش‌های جدای گراف را به هم متصل می‌کنند، مرکزیت بینابینی بیشتری دارند. بیشترین مرکزیت بینابینی، مربوط به گره ۱ (indegree_mp) است. این گره، تنها راه اتصال ۶ گره پایین بزرگ‌ترین مؤلفه، به تمام گره‌های قسمت بالایی آن است. رتبه‌های بعدی متعلق به گره‌های ۲ و ۳ (neighbors_2_hp و neighbors_2_mp) است. این دو گره گلوگاه‌های ارتباطی قسمت بالا و پایین خودشان هستند. با اینکه گره ۱، نسبت به گره های ۲ و ۳ درجه به مراتب کمتری دارد، اما مرکزیت بینابینی آن بسیار بیشتر است. علت این است که گره ۱، تنها راه ارتباط بین تمام گره‌های بخش های بالا و پایین خودش است؛ در حالی که گره های ۲ و ۳، هر دو واسط ارتباطی بخش های بالا و پایین خودشان هستند و اگر هر کدام از آن‌ها از گراف حذف شود، دیگری می‌تواند نبودش را جبران کند.

گره‌های ۴ و ۵ (neighbors_3_mp و outdegree_mp) در جایگاه بعدی بینابینی قرار دارند. این دو نیز مثل موارد قبلی، گلوگاه ارتباطی هستند اما از آنجایی که واسط رسیدن به گره‌های کم‌تری از گراف هستند ، نسبت به سه گره قبلی، اهمیت کم‌تری دارند.

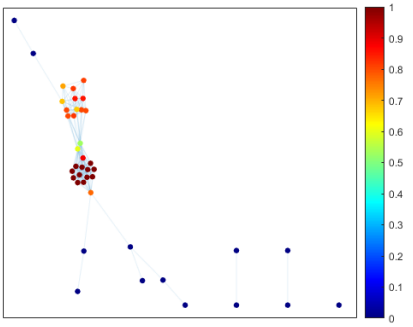
در نمودار ۴، مرکزیت در اکثر گره‌های گراف، مقدار کمی است. اما در انتهای توزیع، چند گره با مرکزیت بالا وجود دارند. در واقع نمودار مرکزیت بینابینی این گراف، مستعد پیروی از قانون توان (power law) است.

۳.۳.۵ مرکزیت بردار ویژه

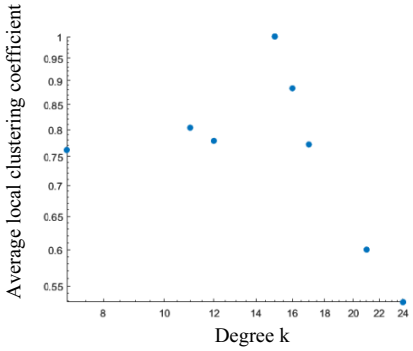
به‌منظور محاسبه مرکزیت برداروِیژه، می‌بایست وزن یال‌ها را به عنوان شاخصی برای اهمیت ارتباطات گره‌ها در نظر بگیریم. در واقع، نتیجه مرکزیت بردار ویژه، در صورتی که وزن یال‌ها را در نظر بگیریم، با زمانی که وزن‌ها را لحاظ نکنیم، متفاوت خواهد بود. در حالت اول، ماتریس مجاورت برای محاسبه مرکزیت بردار ویژه، با مقادیر وزن‌ها پر می‌شود و ارزش یک ارتباط را وزن آن تعیین خواهد کرد اما زمانی که گراف را بدون وزن در نظر بگیریم، صرفا اتصال به گره‌های درجه بالا، اهمیت خواهد داشت.

زمانی‌که تمام یال‌ها، به یک اندازه ارزش داشته باشند، عنصر تعیین کننده‌ی میزان اهمیت گره و همسایگانش، تنها «درجه» خواهد بود، لذا گره‌هایی که خودشان یا همسایگان نزدیکشان بیشترین درجه را داشته باشند، بیشترین مرکزیت را خواهند داشت؛ اما زمانی‌که وزن یال‌ها متفاوت است، برخی از اتصالات مهم‌تر هستند؛ بنابراین علاوه بر درجه، وزن اتصالات با همسایگان نیز تعیین کننده خواهند شد و همانطور که در شکل اول نمایان است، گره‌هایی که با اتصالات محکم‌تری به گره های مهم‌تر متصل هستند، مرکزیت بردار ویژه بالاتری خواهند داشت.

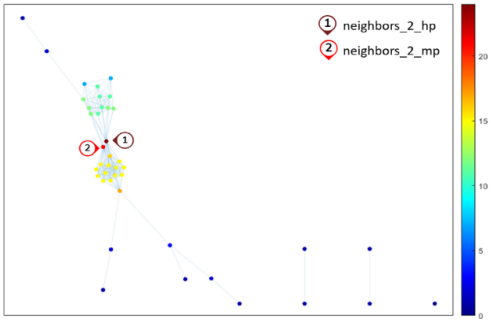
اما چرا بخش میانی گراف، نسبت به hub های شبکه که بیشترین درجات را دارند، مرکزیت بیشتری



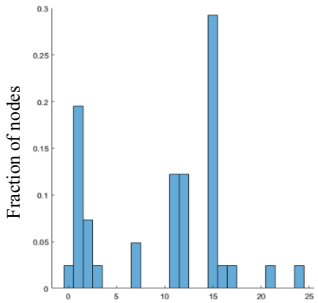
شکل ۲. ضرایب خوشه‌بندی محلی



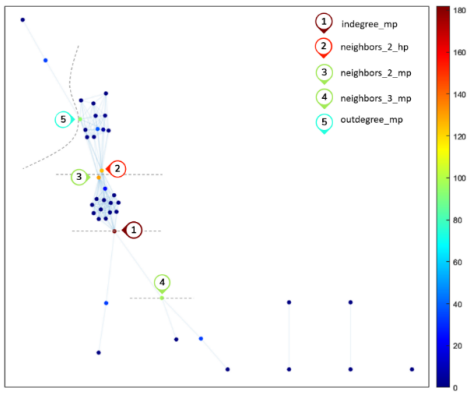
نمودار ۲. خوشه بندی محلی به عنوان تابعی از درجه



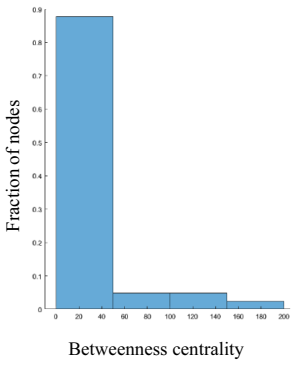
شکل ۳. مرکزیت درجه گراف



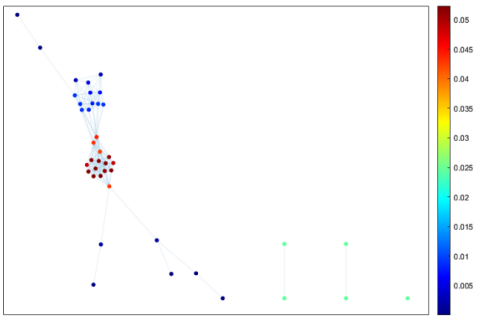
نمودار ۳. هیستوگرام مرکزیت درجه گراف



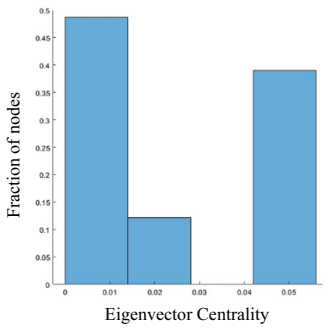
شکل ۴. مرکزیت بینابینی گراف



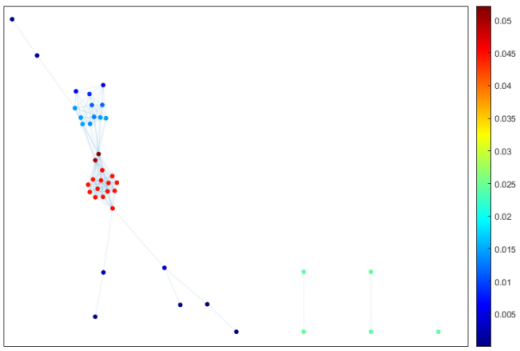
نمودار ۴. هیستوگرام مرکزیت



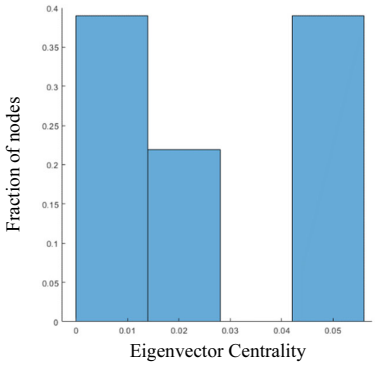
شکل۵. مرکزیت بردار ویژه گراف **با** در نظر گرفتن وزن یال‌ها



نمودار۵. هیستوگرام مرکزیت برداروِیژه گراف **با** در نظر گرفتن وزن یال‌ها



شکل۶. مرکزیت بردار ویژه گراف، **بدون** در نظر گرفتن وزن یال‌ها



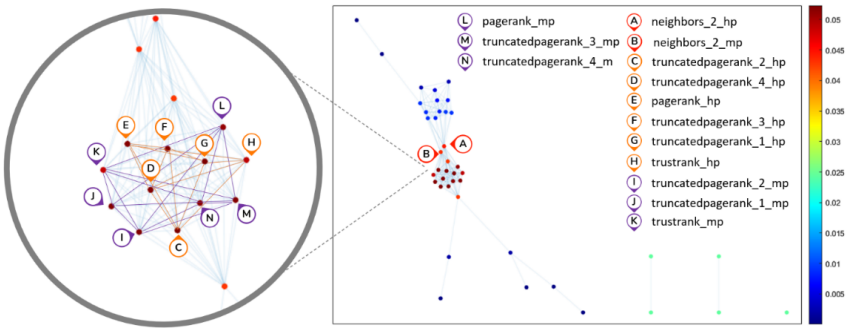
نمودار۶ . هیستوگرام مرکزیت بردار ویژه گراف

دارند؟ برای پاسخ به این سؤال، لازم است کیفیت اتصال گره ها را بررسی کنیم.

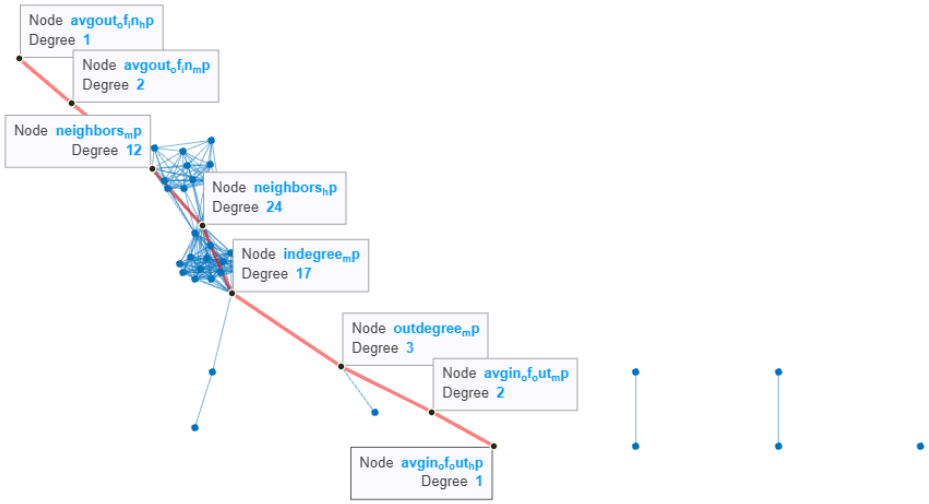
۴.۵ قطر شبکه

همان‌طور که در شکل ۸ پیداست، طولانی‌ترین فاصله، متعلق به دو گره avgin_of_out_hp و avgout_of_in_hp می‌باشد که با ۶ واسطه به یکدیگر متصل شده اند. مجموع این مسیر (اندازه قطر گراف) ۴٫۰۹۰۵ می‌باشد.

مرکزیت بینابینی، نسبت تعداد دفعاتی است که یک گره یا یک یال بر روی کوتاهترین مسیر میان گره‌های مختلف یک گراف قرار می‌گیرد. بررسی شکل ۸ نیز نشان می‌دهد که قطر گراف از نقاطی که بیشترین مرکزیت بینابینی را دارند (neighbors_2_hp, neighbors_3_mp indegree_mp)، عبور کرده است.



شکل ۷. نمایی نزدیک‌تر از مرکزیت بردار ویژه گراف با در نظر گرفتن وزن یال‌ها



شکل ۸. قطر گراف

جدول ۱: خلاصه پژوهش. برخی از ویژگی‌های اندازه‌گیری شده عبارتند از: تعداد گره‌ها (n)، تعداد یال‌ها (m)، میانگین درجه (avg1)، میانگین وزن یال‌ها (avg2)، اندازه بزرگ‌ترین مؤلفه (ns)، کسری از گره‌ها که در بزرگ‌ترین مؤلفه قرار دارند (s)، ضرایب خوشه‌بندی (C1,C2).

ویژگی‌های مورد استفاده	نوع گراف	n	m	avg1	avg2	n_s	S	L	C1	C2
ویژگی‌های مبتنی بر پیوند	بدون جهت	41	202	9.8	0.67	36	0.87	1.2	0.29	0.59
ویژگی‌های مبتنی بر متن	بدون جهت	96	854	17.79	0.64	75	0.78	1.63	0.2	0.7

۶ نتیجه‌گیری

در این پژوهش، گراف‌های شباهت ویژگی‌ها متناظر با دو دسته از ویژگی‌های مبتنی بر متن و نیز مبتنی بر پیوند از مجموعه داده مجموعه WEBSPAM-UK2007 ایجاد گردید و از منظر خصوصیات ساختاری، مورد مطالعه و مورد بررسی قرار گرفت. نتایج کلی به دست آمده در جدول ۱ آمده است.

همان‌طور که مشاهده می‌شود، علیرغم اینکه گراف شباهت ویژگی‌های مبتنی بر متن، نسبت به گراف شباهت ویژگی‌های مبتنی بر پیوند، بزرگ‌تر و به لحاظ تعداد یالها، متراکم‌تر است، با این حال، بر اساس شاخص‌های ضریب خوشه‌بندی، گراف شباهت ویژگی‌های مبتنی بر پیوند، انسجام نسبی بیشتری را دارا می‌باشد. این موضوع با توجه به اندازه نسبی بزرگ‌ترین مولفه همبند نیز تأیید می‌شود.

از این اطلاعات می‌توان به‌منظور کاهش فضای ابعاد داده مسئله شناسایی صفحات اسپم و نیز طراحی الگوریتم‌های جدید در این حوزه، استفاده نمود. از سوی دیگر، استفاده از رویکرد پیشنهای به‌منظور بررسی مجموعه‌های داده دیگر در حوزه شناسایی صفحات اسپم و نیز تحلیل تطبیقی این مجموعه‌های داده بسیار مفید خواهد بود. ضمن اینکه از این نتایج می‌توان به منظور تهیه یک مجموعه داده شناسایی صفحات اسپم در محیط وب فارسی نیز بهره گرفت.

مراجع

[1] Y. Zhang, D. Wang, and Y. Zhang, “Neural IR meets graph embedding: A ranking model for product search,” Web Conf. 2019 - Proc. World Wide Web Conf. WWW 2019, pp. 2390–2400, May 2019, doi: 10.1145/3308558.3313468.

[2] S. Liu, W. Gu, G. Cong, and F. Zhang, “Structural Relationship Representation Learning with Graph Embedding for Personalized Product Search,” Int. Conf. Inf. Knowl. Manag. Proc., pp. 915–924, Oct. 2020, doi: 10.1145/3340531.3411936.

[3] S. Bin Yang and B. Yang, “Learning to rank paths in spatial networks,” Proc. - Int. Conf. Data Eng., vol. 2020-April, pp. 2006–2009, Apr. 2020, doi: 10.1109/ICDE48307.2020.00225.

[4] Q. Xu, M. Li, and M. Yu, “Learning to rank with relational graph and pointwise constraint for cross-modal retrieval,” Soft Comput., vol. 23, no. 19, pp. 9413–9427, Oct. 2019, doi: 10.1007/S00500-018-3608-9/METRICS.

[5] Y. Qi, J. Zhang, Y. Liu, W. Xu, and J. Guo, "CGTR: Convolution Graph Topology Representation for Document Ranking," Int. Conf. Inf. Knowl. Manag. Proc., pp. 2173–2176, Oct. 2020, doi: 10.1145/3340531.3412073.

[6] R. Sawhney, S. Agarwal, A. Wadhwa, and R. Shah, "Exploring the scale-free nature of stock markets: Hyperbolic graph learning for algorithmic trading," Web Conf. 2021 - Proc. World Wide Web Conf. WWW 2021, pp. 11–22, Apr. 2021, doi: 10.1145/3442381.3450095.

[7] Y. Zhang et al., "Learning to Rank Ace Neural Architectures via Normalized Discounted Cumulative Gain," Aug. 2021, doi: 10.48550/arxiv.2108.03001.

[8] T. Formal, S. Clinchant, J. M. Renders, S. Lee, and G. H. Cho, "Learning to Rank Images with Cross-Modal Graph Convolutions," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12035 LNCS, pp. 589–604, 2020, doi: 10.1007/978-3-030-45439-5_39/FIGURES/2.

[9] F. Feng, X. He, X. Wang, C. Luo, Y. Liu, and T. S. Chua, "Temporal Relational Ranking for Stock Prediction," ACM Trans. Inf. Syst., vol. 37, no. 2, Mar. 2019, doi: 10.1145/3309547.

[10] F. Bianchi, M. Palmonari, M. Cremaschi, and E. Fersini, "Actively learning to rank semantic associations for personalized contextual exploration of knowledge graphs," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10249 LNCS, pp. 120–135, 2017, doi: 10.1007/978-3-319-58068-5_8/TABLES/4.

[11] I. Muhammad, D. Bollegala, F. Coenen, C. Gamble, A. Kearney, and P. Williamson, "Document Ranking for Curated Document Databases Using BERT and Knowledge Graph Embeddings: Introducing GRAB-Rank," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12925 LNCS, pp. 116–127, 2021, doi: 10.1007/978-3-030-86534-4_10/COVER.

[12] C. C. Ni, K. Sum Liu, and N. Torzec, "Layered Graph Embedding for Entity Recommendation using Wikipedia in the Yahoo! Knowledge Graph," Web Conf. 2020 - Companion World Wide Web Conf. WWW 2020, pp. 811–818, Apr. 2020, doi: 10.1145/3366424.3383570.

[13] G. Maheshwari, P. Trivedi, D. Lukovnikov, N. Chakraborty, A. Fischer, and J. Lehmann, "Learning to Rank Query Graphs for Complex Question Answering over Knowledge Graphs," in The 18th International Semantic Web Conference (ISWC 2019), Springer, 2019, pp. 487–504. doi: 10.1007/978-3-030-30793-6_28.

[14] J. Y. Yeh and C. J. Tsai, "Graph-based Feature Selection Method for Learning to Rank," ACM Int. Conf. Proceeding Ser., pp. 70–73, Nov. 2020, doi: 10.1145/3442555.3442567.

[15] J. Y. Yeh and C. J. Tsai, "A Graph-based Feature Selection Method for Learning to Rank Using Spectral Clustering for Redundancy Minimization and Biased PageRank for Relevance Analysis," Comput. Sci. Inf. Syst., vol. 19, no. 1, pp. 141–164, Jan. 2022, doi: 10.2298/CSIS201220042Y.

[16] A. H. Keyhanipour, "Graph-based comparative analysis of learning to rank datasets," Int. J. Data Sci. Anal., pp. 1–23, Jun. 2023, doi: 10.1007/S41060-023-00406-8/METRICS.

[17] A.-L. Barabási and M. Pósfai, Network Science, First edit. Cambridge University Press, 2016.

[18] M. Newman, Networks: A Introduction, Second edi. Oxford University Press, 2018.

[19] M. Newman, “The structure and function of complex networks,” SIAM Rev., vol. 45, no. 2, pp. 167–256, 2003, doi: 10.1137/S003614450342480.