علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
مجلد ۱۲، شماره ۱ (الف)، ۱۳۹۳
صفحات ۱-۹
مقاله عادی

پیش‌بینی لینک در شبکه‌های اجتماعی با استفاده از الگوریتم بهینه‌سازی مورچگان

احسان شرکت مسعود رهگذر مسعود اسدپور

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران، تهران، ایران

چکیده

با گسترش روزافزون شبکه‌های اجتماعی در دنیا، لزوم استفاده از روش‌های مقیاس‌پذیر برای پیش‌بینی لینک در شبکه‌های اجتماعی بیش از پیش احساس می‌شود. در این مقاله به معرفی یک روش بر اساس الگوریتم بهینه‌سازی کلونی مورچگان[1](ACO) با رویکرد ساختاری بدون ناظر، برای پیش‌بینی لینک در شبکه‌های اجتماعی پرداخته شده است. روش پیشنهادی بر روی گراف‌های با سایز بزرگ دارای زمان اجرای قابل قبول‌تری نسبت به سایر روش‌های ساختاری پیش‌بینی لینک دارا می‌باشد. برای ارزیابی عملکرد الگوریتم پیشنهادی از روش‌های مساحت زیر نمودار ROC، دقت- فراخوانی[2] و همچنین از روش دقت nتای اول[3] استفاده شده است. در برخی از مجموعه‌های دادگان مورد استفاده، روش پیشنهادی دارای بهترین عملکرد نسبت به سایر روش‌های ساختاری پیش‌بینی لینک می‌باشد. بر اساس بررسی‌های انجام شده، عملاً روش مشابه دیگری وجود ندارد که در آن از الگوریتم بهینه‌سازی کلونی مورچگان برای پیش‌بینی لینک در شبکه‌های اجتماعی استفاده شده باشد.

کلمات کلیدی: پیش‌بینی لینک، شبکه‌های اجتماعی، الگوریتم کلونی مورچگان.

۱- مقدمه

پیش‌بینی امکان وجود ارتباط میان دو موجودیت بر اساس ویژگی‌های موجودیت‌ها و دیگر لینک‌های مشاهده شده در گراف را پیش‌بینی لینک می‌گویند [۱]. به عبارت دیگر اگر در زمان t یک تصویر لحظه‌ای[4] از مجموعه لینک‌ها داشته باشیم، هدف پیش‌بینی لینک‌ها در زمان t+1 می‌باشد [۲]. در سال ۲۰۰۵ اولین بار به طور رسمی در مقاله [۳] مبحثی به نام لینک‌کاوی[5] مطرح شد. در این مقاله پیش‌بینی لینک به عنوان یکی از زیر شاخه‌های لینک‌کاوی مطرح شده است. در شبکه‌های زیستی مثل شبکه‌های تعامل پروتئین- پروتئین[6] و شبکه‌های متابولیکی کشف ارتباط میان پروتئین‌ها نیازمند صرف وقت و هزینه بسیار زیاد برای انجام تحقیقات مربوطه در محیط آزمایشگاه می‌باشد. می‌توان به کمک پیش‌بینی ارتباط میان پروتئین‌ها در این شبکه‌ها ضمن کاهش هزینه‌ها تعداد آزمایشات لازم را کاهش داد [۱]. گاهی در برخی از شبکه‌ها تعدادی لینک به خاطر وجود خطا در شبکه به صورت اتفاقی به وجود می‌آیند. این لینک‌های نادرست می‌توانند ساختار شبکه را مختل کرده و مطالعه آن را با مشکل مواجه کنند. به کمک پیش‌بینی لینک می‌توان این لینک‌ها را شناسایی و از شبکه حذف کرد [۶]. می‌توان مدل تکاملی شبکه را با پیش‌بینی لینک‌ها در آینده بدست آورده و به کمک آن شناخت بهتری از شبکه مورد نظر بدست آورد [۲]. از طرف دیگر، در حوزه پزشکی پیش‌بینی نحوه‌ی شیوع یک بیماری خاص، توسط پیش‌بینی لینک امکان‌پذیر می‌باشد [۱]. در شبکه‌های اجتماعی می‌توان دوستان جدیدی را به کاربران شبکه پیشنهاد داده و باعث افزایش وفاداری کاربران به شبکه اجتماعی مربوطه شد [۴]. پیشنهاد یک کالا یا خدمات به کاربران یک فروشگاه مجازی نیز از دیگر کاربردهای پیش‌بینی لینک می‌باشد [۵].

روش‌های مختلف پیش‌بینی لینک را می‌توان به سه دسته تقسیم‌بندی نمود [۱].

دسته اول راهکارهای مبتنی بر معیارهای شباهت[7] می‌باشند [۲]. این روش‌ها از ویژگی‌های ساختاری گراف شبکه برای پیش‌بینی‌های خود استفاده می‌کنند. برای مثال طول کمترین مسیر میان دو گره یا تعداد همسایه‌های مشترک میان دو گره از معیارهای شباهت می‌باشند. روش ضرایب جاکارد شبیه به روش تعداد همسایه‌های مشترک می‌باشد با این تفاوت که در این روش زوج گره‌هایی که

دارای تعداد همسایه‌های مشترک زیاد و همسایه‌های غیرمشترک کمی باشند، دارای شباهت بیشتری هستند. در روش دیگر بر اساس درجه گره‌ها به پیش‌بینی لینک پرداخته می‌شود. در این روش دو گره با درجه بیشتر، احتمال بیشتری دارد که در آینده با یکدیگر ارتباط برقرار کنند [۲].

دسته دوم، روش‌های مبتنی بر بیشترین احتمال می‌باشند [۱]. در این روش‌ها ضمن بررسی ساختاری شبکه، به استخراج قواعد و ویژگی‌هایی که باعث افزایش احتمال وجود لینک‌ها می‌شود، پرداخته می‌شود. هر چند در این دسته از روش‌ها ویژگی‌های خاص مربوط به شبکه را در نظر می‌گیرند؛ ولی این گونه روش‌ها زمان‌بر بوده و تنها بر روی شبکه‌های دارای گره‌های کم قابل اجرا هستند.

دسته سوم از روش‌ها، راهکارهای مبتنی بر آمار می‌باشند [۱]. در این گونه از روش‌ها از مدل‌های آماری و توزیع‌های مربوطه برای پیش‌بینی لینک استفاده می‌شود.

اکثر روش‌های کنونی بر روی شبکه‌های بدون وزن و جهت تعریف شده است. می‌توان با در نظر گرفتن وزن و جهت در کیفیت خروجی‌ها تأثیرگذار بود [۷]. همچنین این امکان وجود دارد که به جای استفاده از یک تصویر لحظه‌ای از دنباله‌ای از تصاویر لحظه‌ای برای پیش‌بینی لینک نیز استفاده کرد [۸].

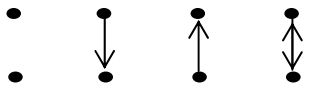
در این مقاله به کمک الگوریتم بهینه‌سازی کلونی مورچگان یک روش جدید برای پیش‌بینی لینک در شبکه‌های اجتماعی معرفی شده است. روش پیشنهادی ضمن بهبود کیفیت پیش‌بینی‌ها در برخی از مجموعه‌های دادگان، باعث بهبود کیفیت پیش‌بینی‌ها شده و به‌کارگیری آن برای شبکه‌های اجتماعی بزرگ نیز امکان‌پذیر می‌باشد. بخش‌های بعدی مقاله شامل معرفی برخی از روش‌های ساختاری که بر اساس پیدا کردن زیر گراف‌های خاص می‌باشند، پرداخته شده و سپس در ادامه به شرح روش پیشنهادی و معرفی مجموعه‌های دادگان مورد استفاده و نحوه ارزیابی روش ارائه شده پرداخته شده است.

۲- کارهای مشابه

در سال‌های اخیر، برخی از روش‌های پیش‌بینی لینک بر اساس بررسی زیرگراف‌های درون گراف شبکه مورد نظر، معرفی شده‌اند. در ادامه به شرح برخی از آنها پرداخته شده است.

مقاله [۹] بر اساس بدست آوردن الگوی تکامل سه‌تایی‌ها[۸] در طی تصاویر لحظه‌ای متوالی از شبکه به پیش‌بینی لینک‌ها می‌پردازد. این روش از نوع ساختاری با کمک ناظر[۹] می‌باشد. اگر گراف را جهت‌دار در نظر بگیریم، در مجموع ۶۴ گروه سه‌تایی مختلف در یک گراف وجود خواهند داشت. زیرا طبق شکل ۱ بین هر جفت گره سه‌تایی می‌تواند ۴ نوع ارتباط شامل عدم اتصال، اتصال اولی به دومی، اتصال دومی به اولی و اتصال متقابل وجود داشته باشد. با شمارش این ۶۴ گروه سه‌تایی در طی دو تصویر لحظه‌ای متوالی از گراف شبکه می‌توان به ماتریس تغییرات سه‌تایی‌ها[۱۰] دست پیدا کرد. بررسی‌ها نشان می‌دهد که ماتریس تغییرات سه‌تایی‌ها برای مجموعه دادگان مختلف دارای تفاوت‌هایی می‌باشد؛ به طوری که این تفاوت‌ها نشان دهنده ویژگی‌های خاص هر مجموعه داده‌ای است. اکنون به کمک این ماتریس و شمردن تعداد مشارکت زوج گره‌هایی از گراف (که با یکدیگر لینک ندارند) در هر یک از این ۶۴ سه‌تایی مختلف، می‌توان احتمال وقوع لینک میان این زوج گره را پیش‌بینی کرد.

مقاله [۱۰] ضمن معرفی یک زیر ساختار از گراف به نام VCP[۱۱] به ارائه یک روش ساختاری با ناظر برای پیش‌بینی لینک و تحلیل نتایج پیش‌بینی‌ها پرداخت هاست. در این مقاله، ساختار $VCP_{s,t}^{n,r}$ به معنای تمامی زیر گراف‌هایی است که دو گره s و t در آن‌ها حضور دارند و دارای n گره و بین هر دو گره حداکثر r رابطه وجود دارد (بدون آنکه بین s و t رابطه‌ای وجود داشته باشد). بدیهی است که اگر مقدار r برابر با یک باشد، یعنی گراف جهت دار نمی‌باشد و اگر r برابر با دو باشد، یعنی گراف جهت دار می‌باشد.



شکل ۱- انواع مختلف ارتباط میان دو گره در یک سه‌تایی

با افزایش مقدار n و r از مرز ۴، تعداد زیر گراف‌های صادق طبق تعریف VCP بسیار زیاد می‌شود. در این مقاله تنها به بررسی VCPs با مقدار n برابر با ۳ و ۴ و مقدار r برابر با ۱ و ۲ پرداخته شده است. اگر مقدار n برابر با ۳ و مقدار r برابر با ۲ باشد، عملاً VCP همان سه‌تایی تعریف شده در مقاله [۹] می‌باشد. با تقسیم خروجی‌های یک روش پیش‌بینی لینک به دو بخش (یکی شامل m لینک پیش‌بینی شده با امتیاز بالا و m لینک پیش‌بینی شده با امتیاز پایین) و مقایسه توزیع آماری VCPs این دو دسته با یکدیگر، می‌توان به نقاط ضعف و قوت روش پیش‌بینی لینک به کار رفته پی‌برد. به همین شکل نیز می‌توان به کمک یک فاز یادگیری[۱۲] به پیش‌بینی لینک پرداخت. از مشکلات این روش زمان بر بودن محاسبات مربوط به VCPs با مقدار n و r به ترتیب بیش از ۴ و ۳ در گراف شبکه‌های بزرگ می‌باشد.

مقاله [۱۱] با گسترش تعریف ضریب خوشه‌بندی[۱۳]، به ارائه روش جدید احتمالاتی مبتنی بر ساختار برای پیش‌بینی لینک پرداخته است. ضریب خوشه‌بندی، میزان تمایل سه‌تایی‌ها به مثلث شدن را در یک گراف نشان می‌دهد. رابطه (۱) نحوه محاسبه ضریب خوشه‌بندی را در یک گراف نشان می‌دهد.

$$\text{ضریب خوشه‌بندی} = \text{(تعداد سه‌تایی‌ها)}/(۳ \times \text{تعداد مثلث‌ها}) \qquad (۱)$$

مقاله [۱۱] ضریب خوشه‌بندی را به صورت تقسیم تعداد حلقه‌های به طول k بر تعداد مسیرهای به طول k در گراف تعریف کرده است. به این ضریب خوشه‌بندی، ضریب خوشه‌بندی عمومی[۱۴] گفته می‌شود. اکنون می‌توان پیش‌بینی لینک را به صورت احتمال آنکه مسیرهای به طول ۳ تبدیل به دورهای به طول ۳ شوند و احتمال اینکه مسیرهای به طول ۴ تبدیل به دورهای به طول ۴ و به همین ترتیب برای مسیر های با طول بیشتر در نظر گرفت. با محاسبه ضریب خوشه‌بندی می‌توان پارامترهای این مدل احتمالاتی را محاسبه کرد.

در اوایل دهه ۹۰ میلادی الگوریتم کلونی مورچگان (ACO) برای اولین بار توسط دوریگو و همکارانش معرفی شد [۱۲]. هدف اصلی از معرفی این روش، حل تقریبی مسائل پیچیده ترکیبی بهینه‌سازی می‌باشد. اساس این روش الهام گرفتن از ویژگی رفتار مورچه‌ها هنگام جستجوی غذا می‌باشد. هنگام جستجوی غذا، در ابتدا مورچه‌ها به طور تصادفی محیط اطراف خود را مورد بررسی قرار می‌دهند. به محض پیدا کردن غذا، هر مورچه میزان کیفیت و کمیت غذا را مورد توجه قرار می‌دهد و در راه برگشت به لانه مقداری از آن را با خود حمل می‌کند. هنگام بازگشت مورچه مذکور ماده‌ای شیمیایی به نام فرمون را نیز در مسیر بازگشت خود می‌ریزد. میزان فرمون ریخته شده متناسب با کیفیت و کمیت منبع غذایی پیدا شده است. این فرمون باعث ایجادیک ارتباط غیر مستقیم با دیگرمورچه‌ها می‌شود. به طوری که مورچه‌ها را قادر می‌سازد که کمترین فاصله میان لانه و منبع غذایی را بدست آورند [۱۲]. در این مقاله احتمال حرکت مورچه kام از نود i به نود j از رابطه (۲) استفاده شده است.در این رابطه τ_{ij} میزان فرمون روی یال i و j است و η_{ij} میزان هزینه یال i و j می‌باشد. منظور از N_i همسایه‌های گره i می‌باشد. مقادیر α و β ضرایب کنترلی این رابطه محسوب می‌شوند. با تغییر این مقادیر می‌توان اثر

میزان فرمون یا هزینه پیمایش یک مسیر را در انتخاب مسیر گره بعدی، کاهش یا افزایش داد.

$$p_{ij}^{k} = \frac{{\tau_{ij}}^{\alpha} {\eta_{ij}}^{\beta}}{\sum_{L \in N_i^k} \left({\tau_{iL}}^{\alpha} {\eta_{iL}}^{\beta} \right)} \quad (۲)$$

یکی از مسائل نزدیک به مسئله پیش‌بینی لینک، مسئله سیستم‌های پیشنهاد دهنده می‌باشد. در این گونه از سیستم‌ها بر اساس رفتار قبلی مشتری یک کالا یا خدمات را به وی پیشنهاد می‌دهند. در اینجا به جای لینک دادن دو فرد به یکدیگر، یک فرد به یک کالا ارتباط داده می‌شود. برای سیستم‌های پیشنهاد دهنده نیز، روش‌هایی مبتنی بر ACO وجود دارد [۱۳]. برخی از روش‌ها نیز وجود دارند که به کمک ACO به استخراج زیرگراف‌ها (از جمله گراف کامل[۱۵]) می‌پردازند [۱۴].

در مقاله حاضر ضمن بهره‌بردن از ویژگی مورچه‌ها در زمان جستجوی غذا، به ارائه یک روش جدید برای پیش‌بینی لینک پرداخته شده است.از آنجایی که از الگوریتم کلونی مورچگان استفاده شده است نام روش پیشنهادی را ACOLP قرار داده‌ایم.

۳- پیش‌بینی لینک به کمک الگوریتم ACO

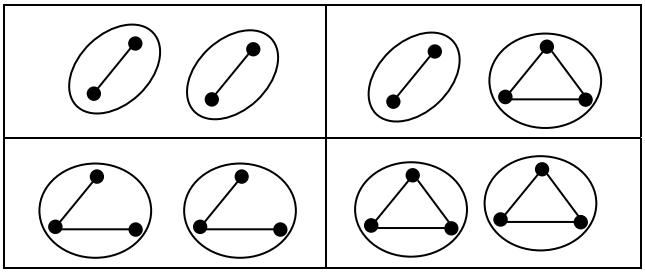
می‌توان با بررسی زیر ساختارهای (زیر گراف) یک گراف شبکه اجتماعی ویژگی‌های ساختاری مهمی از آن گراف را بدست آورد. با افزایش مرتبه زیر ساختار مورد نظر، اطلاعات ساختاری مهم‌تری بدست می‌آید. به طور کلی می‌توان پیش‌بینی لینک را در سطوح مختلفی از زیر ساختارها مورد بررسی قرار داد. برخی از این سطوح در جدول ۱ آورده شده است. در پایین‌ترین سطح (سطح اول) تنها دو گره وجود دارد و از آنجایی که میان دو زیر ساختار حداقل باید یک لینک وجود داشته باشد تا بتوان به کمک آن ارتباطی میان دو زیر ساختار برقرار کرد، پس در پایین‌ترین سطح این ارتباط همان یال‌های موجود در گراف می‌باشند. پس در پایین‌ترین سطح یالی برای پیش‌بینی وجود ندارد (اگر گراف جهت‌دار نباشد). در سطح بعدی ارتباط میان یک گره با یک خط (دو گره که با یک یال به یکدیگر متصل شده‌اند) می‌باشد. عملاً حالت مورد پیش‌بینی در این سطح حالت متعددی[۱۶] می‌باشد. بسیاری از روش‌های پیش‌بینی لینک ساختاری در این سطح، ساختار گراف را مورد بررسی قرار می‌دهند. برای مثال روش همسایه‌های مشترک یا روش ضرایب جاکارد از این جمله موارد می‌باشند [۲].

از آنجایی که با افزایش سطوح از ویژگی‌های ساختاری بیشتری از گراف استفاده می‌شود، لینک پیش‌بینی شده در سطوح بالاتر از ارزشمندی بیشتری نسبت به سطوح پایینی خود برخوردار است. در جدول ۱ در تمامی سطوح یک نقطه با یک زیر ساختار دیگر مورد بررسی قرار گرفته است. می‌توان میزان سطوح را بالاتر برد، به طوری که ارتباط میان دو خط با یکدیگر، دو سه‌تایی با یکدیگر و یا دو مثلث با یکدیگر نیز داشته باشیم. در شکل ۲ برخی از این سطوح آورده شده است. علاوه بر این موارد می‌توان جهت را نیز به سطوح مختلف اضافه کرد. با این کار حالات بیشتری برای پیش‌بینی به وجود می‌آیند.

روش پیشنهادی بر روی سطح چهارم تعریف شده است. در این سطح ارتباط یک گره با یک مثلث مورد بررسی قرار می‌گیرد. همان طوری که از جدول ۱ مشخص است، در این سطح ۲ حالت برای پیش‌بینی وجود دارد. با وجود آنکه با افزایش سطح زیر ساختارها، تعداد لینک‌های مورد پیش‌بینی کمتر می‌شوند، ولی لینک‌های پیش‌بینی شده از آنجایی که از ویژگی‌های ساختاری بیشتری از گراف استفاده نموده‌اند، از ارزشمندی بیشتری برخوردار هستند. این موضوع نیز در درون سطوح صادق است. برای مثال در زیر ساختار سطح چهارم در جدول ۱، دو زیر ساختار a و b قرار دارند. از آنجایی که زیر ساختار b از ویژگی‌های ساختاری بیشتری نسبت به زیر ساختار a بهره برده است، این زیر ساختار ارزشمندتر است. برای اثبات این موضوع، میزان دقت لینک‌های پیش‌بینی شده در دو زیر ساختار a و b در جدول ۱ را مورد بررسی قرار دادیم. نتایج حاصله ادعای ما را تایید می‌کند. راه دیگر اثبات این ادعا استفاده از روش معرفی شده در مقاله [۱۱] می‌باشد. برای تنها یال موجود برای پیش‌بینی در جدول ۱ قسمت b، دو مسیر به طول ۲ برای تبدیل شدن به دور به طول ۲ و یک مسیر به طول ۳ برای تبدیل شدن به دور به طول ۳ وجود دارد. این در حالی است که برای شکل a در همین جدول، برای دو لینک ممکن برای پیش‌بینی، برای یکی تنها یک مسیر به طول ۲ برای تبدیل شدن به دور به طول ۲ و برای دیگری یک مسیر به طول ۲ و یک مسیر به طول ۳ وجود دارد. پس طبق مدل احتمالاتی ارائه شده در مقاله [۱۱]، احتمال وقوع تنها یال موجود در زیرگراف b در جدول ۱ نسبت به دو یال موجود برای پیش‌بینی زیرگراف a در همان جدول بیشتر می‌باشد.

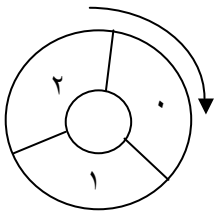
جدول ۱- برخی از سطوح مختلف زیر ساختارها. در این جدول خطوط خط چین یال‌های مورد نظر برای پیش‌بینی می‌باشند

سطح	زیرساختار	لینک‌های موجود برای پیش‌بینی لینک مابین بخش سمت راست و چپ (خط‌چین‌ها)
سطح ۱		
سطح ۲		
سطح ۳		
سطح ۴		



شکل ۲- سایر سطوح قابل بررسی

اکنون به کمک الگوریتم بهینه‌سازی کلونی مورچگان به استخراج زیر ساختارهای سطح چهارم پرداخته می‌شود. مورچه‌ها به کار گرفته شده دارای ویژگی‌های خاصی می‌باشند که در ادامه به شرح این ویژگی‌ها پرداخته می‌شود. برای استخراج زیر ساختاری‌های سطح چهارم، تنها کافی است مثلث‌ها را پیدا کنیم. به این صورت که در هر بار حرکت مورچه بر روی گراف، نام گره عبوری در یکی از خانه‌های حافظه آن مورچه ذخیره می‌شود (شکل ۳).



شکل ۳- حافظه مورچه‌ها

در صورتی که حافظه پر شود، بر روی خانه قبلی اطلاعات نوشته می‌شود. ولی قبل از آن بررسی می‌شود که آیا مقداری که قرار است در یک خانه از حافظه نوشته شود، با مقدار قبلی آن خانه برابر است یا خیر، اگر برابر بود یک مثلث پیدا شده است. میزان فرمون روی تمامی یال‌ها در ابتدا برابر با یک واحد می‌باشد. در هر بار پیدا کردن یک مثلث میزان فرمون تمامی یال‌های آن مثلث یک واحد افزایش پیدا می‌کند. برای آنکه به سرعت تمامی بخش‌های گراف توسط مورچه‌ها پیمایش شود، مورچه‌ها دارای این ویژگی هستند که بر خلاف روش‌های معمول، به یال‌های دارای فرمون کمتر علاقه‌مند می‌باشند. احتمال حرکت مورچه kام از گره iام به گره jام از رابطه (۳) بدست می‌آید.

$$P_{ij}^{k} = \frac{(\frac{1}{\tau_{ij}})^{\alpha}}{\sum_{l \in N_i^k} (\frac{1}{\tau_{ij}})^{\alpha}} \qquad (۳)$$

در رابطه (۳)، τ میزان فرمون را در آن یال نشان می‌دهد. مقدار α یک ضریب ثابت می‌باشد که در این مقاله آن را برابر با یک در نظر گرفته‌ایم. تفاوت رابطه (۳) با رابطه (۲) در این است که مقدار β برابر با صفر در نظر گرفته شده است. علاوه بر این، رابطه (۳) با مقدار فرمون رابطه عکس دارد. می‌توان با دادن ضریب منفی به فرمون‌های آزاد شده توسط مورچه‌ها این رابطه را به صورت رابطه مستقیم (مانند رابطه (۲)) تبدیل کرد.

در این الگوریتم، فرمون رها شده بر روی یال‌ها تبخیر نمی‌شود و مورچه‌ها نیز از این ویژگی برخوردارند که اگر پس از مدتی یالی بدون فرمون اولیه پیدا نکردند، می‌میرند. همچنین اگر ۳ یال قبلی عبوری یک مورچه دارای فرمون بیشتر از یک واحد باشد، آن مورچه می‌میرد. این ویژگی باعث شده است که مورچه‌ها بخش‌های مورد بررسی شده قبلی را دوباره مورد بررسی قرار ندهند. علت مرگ مورچه‌ها می‌تواند به خاطر گرفتار شدن در جزایر دو گره‌ای و یا گرفتاری بر روی گره‌های برگ و یا پر شدن محیط اطراف آن‌ها از فرمون باشد. چون مورچه‌ها می‌میرند، در ابتدا با مورچه‌هایی در حدود ۱۰ تا ۲۰ برابر تعداد نودهای گراف، الگوریتم اجرا می‌شود و پس از در حدود ۲ یا ۳ بار حرکت مورچه‌ها، در حدود نیمی از آن‌ها به علت پر شدن محیط اطرافشان از فرمون، می‌میرند. الگوریتم زمانی پایان می‌پذیرد که یا تمامی مورچه‌ها مرده باشند و یا آنکه به تعداد مورد نظر مورچه‌ها در گراف حرکت کرده باشند.

برای بدست آوردن تمامی مثلث‌ها در گراف روش‌های مختلفی وجود دارد که ساده‌ترین آنها از مرتبه زمانی $O(n^3)$ می‌باشد. که n تعداد گره‌های گراف می‌باشد. بهترین الگوریتم که تا کنون برای بدست آوردن تمامی مثلث‌ها ارائه شده است از مرتبه زمانی $O(n^{2.376})$ و پیچیدگی مکانی[۱۷] از مرتبه $O(n^2)$ می‌باشد [۱۹]. در زمانی که گراف شبکه مورد بررسی خلوت باشد می‌توان با پیچیدگی زمانی کمتری مسئله را حل نمود به طوری که، برای گراف از نوع قانون نمایی[۱۸]، می‌توان با پیچیدگی زمانی $O(mn^{1/\alpha})$ و پیچیدگی مکانی $O(n^2)$ مسئله را حل نمود [۱۹]. منظور از m، تعداد یال‌ها و مقدار α ما بین ۲ و ۳ می‌باشد. با این وجود برای گراف‌های بزرگ و یا غیرخلوت، استفاده از این روش‌ها از لحاظ زمان اجرا یا فضای مصرفی مقرون به صرفه نمی‌باشد. الگوریتم ارائه شده از آنجایی که یک روش تخمینی می‌باشد، پیچیدگی زمانی از مرتبه $O(n)$ و فضای مصرفی از مرتبه $O(m)$ می‌باشد.

پس از پیدا کردن مثلث‌ها، رابطه سایر گره‌ها با این مثلث‌ها طبق روابط موجود در سطح چهارم جدول ۱ را محاسبه می‌کنیم. از آنجایی که شکل b در سطح چهارم جدول ۱ از اهمیت بیشتری نسبت به شکل a در همان جدول برخوردار است، اگر گراف شبکه مورد بررسی خلوت[۱۹] نباشد، تنها کافی است که زیر ساختار b را مورد بررسی قرار دهیم. از آنجایی که ممکن است یک لینک (یال) به علت قرار گرفتن در زیر ساختارهای متعددی از نوع سطح چهارم، چندین بار مورد پیش‌بینی قرار گیرد، در هر بار پیش‌بینی امتیاز این لینک افزایش می‌یابد. لازم به ذکر است که در هر بار پیش‌بینی هر دو یال موجود شکل a در سطح چهارم جدول ۱ مورد پیش‌بینی قرار می‌گیرند.

مقدار فرمون رها شده بر روی یال‌ها تنها نقش تعیین کننده برای پیدا کردن بهتر و سریع‌تر مثلث‌ها دارند و در تعیین مقدار امتیاز[۲۰] یک لینک پیش‌بینی شده نقشی ندارند. می‌توان برای آنکه از فرمون‌ها نیز برای تعیین مقدار امتیاز لینک‌ها استفاده کنیم به این صورت عمل کرد که پس از آنکه لینک‌های مورد نظر پیش‌بینی شدند، یک بار دیگر مورچه‌ها بر روی گراف، البته با لینک‌های پیش‌بینی شده در مرحله قبل حرکت کنند. سپس هر لینکی که دارای فرمون بیشتری باشد از امتیاز بیشتری برخوردار می‌شود. آزمایشات نشان می‌دهد که این کار در برخی از مجموعه‌های دادگان باعث بهبود نسبی پیش‌بینی‌ها شده و در برخی اثر مثبتی ندارد. با توجه به سربار زمانی، این کار عملاً از این قسمت در این مقاله صرفه‌نظر شده است. می‌توان از فرمون‌های رها شده بر روی یال‌ها برای بدست آوردن مرکزیت[۲۱] گره‌ها نیز استفاده کرد. به این صورت که گره‌ای که به یال‌های دارای فرمون بیشتری متصل است دارای مرکزیت بیشتری نسبت به سایر گره‌ها است. در ادامه به معرفی مجموعه‌های دادگان مورد استفاده و ارزیابی روش ارائه شده، پرداخته می‌شود.

۴- نتایج و آزمایشات

در این بخش به معرفی مجموعه‌های دادگان به کار رفته به همراه نتایج ارزیابی عملکرد الگوریتم پیشنهادی و مقایسه آن با سایر روش‌های پیش‌بینی لینک مبتنی بر ساختار بر روی این مجموعه‌های دادگان پرداخته می‌شود. نخست در بخش ۴-۱ به معرفی مجموعه‌های دادگان به کار رفته اشاره می‌شود و سپس در بخش ۴-۲ انواع مختلفی از ارزیابی‌ها بر روی نتایج حاصل از الگوریتم پیشنهادی و سایر الگوریتم‌ها صورت می‌گیرد. در بخش ۴-۳ نیز نحوه محاسبه مرکزیت داده‌ها در کنار پیش‌بینی لینک توسط الگوریتم پیشنهادی بیان شده است.

۴-۱- مجموعه‌های دادگان

مجموعه‌های دادگان hep-ph و astro-ph مربوط به مقاله‌های نوشته شده توسط نویسندگان موضوعات فیزیک انرژی بالا- پدیده‌شناسی[۲۲] و اخترفیزیک[۲۳] می‌باشد [۲]. این مجموعه‌های دادگان شامل دو بخش مقالات نوشته شده در طی سال‌های ۹۴ تا ۹۶ و سال‌های ۹۷ تا ۹۹ است. بر اساس تعریف استفاده شده در مقاله [۲]، تنها برای این دو مجموعه داده‌ای، از گراف هسته برای پیش‌بینی‌ها استفاده شده است. گراف هسته گرافی است که تمامی گره‌های آن حداقل ۳ بار در زمان t و حداقل ۳ بار در زمان t+1 دارای تعامل با سایر گره‌های گراف باشند. برای این منظور در جدول ۲ در ستون تعداد یال‌های اضافه شده به گره‌های زمان t، یال‌های اضافه شده به گراف هسته نوشته شده است. برای سایر مجموعه‌های

دادگان که در ادامه معرفی می‌شوند، گراف هسته شامل تمامی گره‌های گراف زمان t می‌باشد.

مفهوم گراف هسته در مقاله‌ی [۲] مطرح شده است. منظور از گراف هسته به شماره ۳ آن است که تمامی گره‌هایی از گراف زمان t انتخاب شوند، که حداقل ۳ بار با سایر گره‌ها لینک داشته باشند، و در زمان t+1 نیز همین گره‌ها، حداقل ۳ لینک با سایر گره‌ها داشته باشند. علت تعریف، گراف هسته به سه دلیل بوده است. دلیل اول: از یک الگوریتم پیش‌بینی لینک تنها انتظار می‌رود ارتباط جدید یک گره با سایر گره‌های موجود را پیش‌بینی نماید، نه آنکه ارتباط یک گره با گره‌هایی که قرار است در آینده بوجود بیایند. دلیل دوم: از آنجایی که در برخی از شبکه‌ها ممکن است برخی از گره‌ها در زمان بعدی از بین بروند، بنابراین انتظار پیش‌بینی لینک برای آنها از تعریف پیش‌بینی لینک خارج خواهد بود. دلیل سوم: علت آنکه در مقاله [۲] گراف هسته به شماره ۳ بیان شده است، کاهش اثر خلوتی ماتریس در نتایج پیش‌بینی‌ها می‌باشد. به عبارت دیگر هر چه قدر مقدار درجه گراف هسته بالا رود، گره‌هایی انتخاب می‌شوند که فعال‌تر از سایر گره‌ها بوده و بنابراین احتمال اینکه الگوریتم پیش‌بینی، پیش‌بینی‌های صحیح‌تری داشته باشد، بیشتر است. برای مثال اگر گراف هسته به شماره ۱۰۰ انتخاب شود، گره‌هایی انتخاب می‌شوند که بسیار در زمینه‌ی ایجاد ارتباط فعال‌تر از سایرین هستند و بنابراین احتمال رخ داد لینک جدید میان آنها و در نتیجه بهبود عملکرد الگوریتم‌های پیش‌بینی لینک در این موارد بیشتر خواهد شد. علت عدم‌استفاده از تعریف گراف هسته در سایر مجموعه‌های دادگان به دو دلیل است. دلیل اول: تنها در مقاله‌ی [۲] از تعریف گراف هسته استفاده شده است و در سایر مقاله‌ها در رابطه با پیش‌بینی لینک این مورد به چشم نمی‌خورد. دلیل دوم: باز در این مجموعه‌های دادگان انتظار نمی‌رود که لینک بین یک گره موجود و گره‌ای که بعداً بوجود می‌آید، پیش‌بینی شود. بلکه به نوعی تنها انتظار می‌رود لینک‌هایی که بعداً ما بین گره‌های موجود بوجود می‌آیند، پیش‌بینی شود. به عبارت دیگر، مقدار درجه در گراف هسته در این مجموعه‌های دادگان برابر با یک فرض شده است. یعنی تنها گره‌هایی از گراف زمان t انتخاب شده‌اند که حداقل یکبار در زمان t و حداقل یکبار در زمان t+1 با سایر گره‌ها ارتباط داشته‌اند.

جدول ۲- اطلاعات مربوط به مجموعه های داده ای مورد استفاده

نام مجموعه داده ای	گراف زمان t		گراف زمان t+1		
	تعداد گره	تعداد یال	تعداد گره	تعداد یال	تعداد یال‌های اضافه شده به گره‌های گراف زمان t
astro-ph	۵۳۴۳	۲۰۹۲۴	۱۱۹۳۸	۱۸۰۰۰	۵۷۵۱
hep-ph	۵۴۱۴	۲۳۹۰۰	۷۱۴۴	۶۳۷۷	۳۲۹۸
disease-p	۴۳۷	۱۰۲۷۸	۴۳۷	۱۰۲۷۸	۱۰۲۷۸
disease-g	۳۹۹	۳۰۳۰۱	۳۹۹	۱۹۶۶	۱۹۶۶
huddle	۱۴۷۸	۶۲۱۰۳	۱۷۸۷	۸۳۹۱۹	۵۱۳۰۰
irBlogs	۱۶۹۵۸۷	۲۱۶۵۹۳	۶۵۰۸۴۸	۵۱۷۷۸۸۲	۷۷۵۱۹۸

مجموعه داده‌ای disease-g و disease-p شامل شبکه‌ای است که در آن گره‌ها نشان دهنده یک بیماری و یال‌ها نشان دهنده وجود یک ویژگی ژنتیکی مشترک میان دو بیماری است [۱۰]. این شبکه بدون جهت می‌باشد. از آنجایی که این شبکه تقریباً در طول زمان ثابت است، به ناچار برای انجام پیش‌بینی‌ها، ۲۰ درصد از لینک‌ها را به طور تصادفی از گراف حذف کرده و سپس به پیش‌بینی این یال‌های حذف شده پرداختیم.

مجموعه داده‌ای huddle اطلاعات مربوط به یک فروشگاه در دانشگاه نوتردام می‌باشد، به طوری که گره‌ها نشان دهنده محصولات فروشگاه و یال‌ها نشان دهنده خرید دو کالا در یک خرید به طور همزمان می‌باشد [۱۰]. گراف زمان t اطلاعات خرید مشتریان در طی ماه‌های ژوئن الی آگوست ۲۰۰۴ و در گراف زمان t+1 اطلاعات مشتریان در طی ماه‌های سپتامبر تا نوامبر سال ۲۰۰۴ میلادی است.

مجموعه داده‌ای irBlogs شامل وبلاگ‌های فارسی زبان و لینک‌های میان آن‌ها است [۲۰]. این مجموعه داده‌ای در آزمایشگاه تحقیقات پایگاه داده دانشگاه تهران ایجاد شده است. این مجموعه شامل بیش از ۶۵۰۰۰۰ وبلاگ به زبان فارسی است.این مجموعه از انواع مختلفی از ارائه دهندگان وبلاگ در ایران گردآوری شده است. گراف زمان t در سال ۲۰۱۰ و گراف زمان t+1 در سال ۲۰۱۱ ایجاد شده‌است. اطلاعات آماری مربوط به این مجموعه داده‌ای در سال ۲۰۱۱ در جدول ۳ آورده شده است.

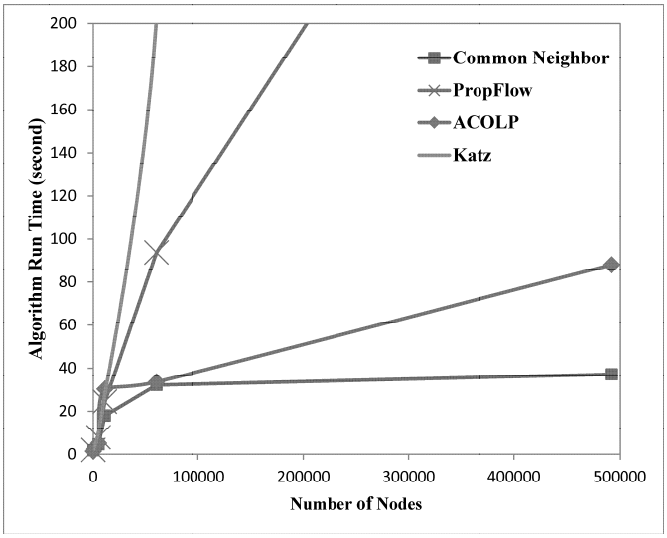
جدول ۳- اطلاعات آماری مربوط به مجموعه داده‌ای irBlogs در سال ۲۰۱۱

نام ارائه دهنده وبلاگ	تعداد (کل)	درصد (کل)
blogfa.com	۴۶۸۴۰۵	۷۱.۹۶
persianblog.ir	۵۹۲۲۷	۹.۰۹
mihanblog.com	۴۸۰۳۲	۷.۳۷
persiangig.com	۱۹۷۹۸	۳.۰۴
blogspot.com	۱۵۷۱۹	۲.۴۱
blogsky.com	۱۴۹۸۳	۲.۳۰
parsiblog.com	۱۳۷۱۷	۲.۱۰
rozblog.com	۳۴۲۱	۰.۵۲
iranblog.com	۲۵۹۱	۰.۳۹
bloghaa.com	۱۱۸۸	۰.۱۸
javanblog.com	۱۰۶۷	۰.۱۶
googlepages.com	۱۰۰۲	۰.۱۵
blogdoon.com	۴۲۴	۰.۰۶
mahblog.com	۴۱۹	۰.۰۶
shakheh.com	۲۶۳	۰.۰۴
other	۵۹۰	۰.۰۹

۴-۲- آزمایشات

برای ارزیابی عملکرد الگوریتم پیشنهادی، این الگوریتم را با تمامی روش‌های ساختاری پیش‌بینی لینک ارائه شده در [۲] و روش PropFlow ارائه شده در [۱۵] مقایسه کردیم. تمامی این روش‌ها بدون ناظر[۲۴] می‌باشند. روش PropFlow یک الگوریتم بر مبنای گام تصادفی محدود شده می‌باشد. از نرم‌افزار معرفی شده در [۱۶] برای آزمون برخی از الگوریتم‌ها استفاده شده است.

در راستای مقایسه‌ی زمان اجرای الگوریتم‌ها و به تبع آن تحلیل قابلیت مقیاس‌پذیر یا لگوریتم‌ها، زمان اجرای الگوریتم پیشنهادی را با یکی از سریع‌ترین روش‌های موجود یعنی روش همسایه‌های مشترک مقایسه نمودیم. از طرفی برای نشان دادن عدم مقیاس‌پذیری برخی از الگوریتم‌ها نتایج زمان اجرای دو الگوریتم Katz و PropFlow را نیز با روش پیشنهادی مقایسه نمودیم. در شکل ۴ نمودار مقایسه زمان اجرای الگوریتم‌های منتخب نشان داده شده است. برای بررسی مقیاس‌پذیری الگوریتم‌ها زمان اجرای الگوریتم‌ها را با رشد تعداد گره‌های گراف بررسی نمودایم. نتایج حاصل از این آزمایش نشان می‌دهد که الگوریتم پیشنهادی مقیاس‌پذیر بوده و برای گراف‌های بزرگ قابل استفاده می‌باشد. از طرفی نتایج خروجی پیش‌بینی‌های الگوریتم نیز نسبت به سایر روش‌ها قابل ملاحضه می‌باشد.



شکل ۴- مقایسه زمان اجرای الگوریتم‌ها با رشد شبکه

اولین روش ارزیابی بر اساس تعریف به کار رفته در مقاله [۲] است. در این روش به تعداد لینک‌های جدید به وجود آمده دقت لیست پیش‌بینی‌ها را بدست می‌آوریم. در مقاله [۲] میزان دقت الگوریتم، نسبت به حالت یک پیش‌بینی کننده تصادفی مقایسه می‌شود. برای آنکه میزان دقت یک پیش‌بینی کننده تصادفی محاسبه شود، تنها کافی است تعداد لینک‌های جدید به وجود آمده در زمان t+1 را بر تعداد لینک‌هاینا‌موجود در گراف زمان t تقسیم کنیم. در جدول ۴، درصد دقت پیش‌بینی کننده تصادفی برای مجموعه‌های دادگان به کار رفته، آورده شده است.

جدول ۴- درصد دقت پیش بینی کننده تصادفی

نام مجموعه داده‌ای	درصد دقت پیش بینی کننده تصادفی
astro-ph	۰.۴۷۵ درصد
hep-ph	۰.۲۰۷ درصد
disease-p	۱۵.۸۱۸ درصد
disease-g	۲.۶۷۲ درصد
huddle	۴.۹۷۹ درصد
irBlogs	۰.۰۰۳ درصد

در جدول ۵ نتایج دقت الگوریتم‌های پیش‌بینی کننده با الگوریتم پیشنهادی (ACOLP) مقایسه شده است. در این جدول میزان بهبود نسبت به حالت تصادفی نشان داده شده است. لازم به ذکر است که از آنجایی که در روش پیشنهادی در هر بار لیست پیش‌بینی‌ها تا حدودی متفاوت باشد، در جدول ۵ میانگین دقت الگوریتم پیشنهادی آورده شده است. در جدول ۵ به علت زمان بر بودن الگوریتم SimRank و Graph Distance از محاسبه آن‌ها برای مجموعه داده‌ای irBlogs صرفه نظر شده است.

در تمامی بخش‌های این مقاله، برای الگوریتم Katz مقدار β برابر با ۰.۰۰۵ و تا حداکثر فاصله ۵ از هر گره محاسبات صورت می‌گیرد. برای الگوریتم RootedPageRank احتمال برگشت به گره قبلی برابر با ۰.۵ ، برای SimRank پارامتر ورودی برابر با ۰.۸ ، برای PropFlow برابر با ۵ و برای الگوریتمACOLP هر دو حالت شکل‌های a و b در جدول ۱ در سطح چهارم، در نظر گرفته شده است.

روش‌های دیگر مقایسه میان الگوریتم‌های پیش‌بینی لینک با روش پیشنهادی، استفاده از معیار مساحت زیر نمودار ROC [۱۷] و مساحت زیر نمودار دقت - فراخوانی [۱۸] می‌باشد. در جدول ۶ نتایج ارزیابی مساحت زیر نمودار ROC و در جدول ۷ نتایج ارزیابی مساحت زیر نمودار دقت - فراخوانی آورده شده است. در شکل‌های ۵، ۶ و ۷ نمودار مقایسه ROC و در شکل‌های ۸، ۹ و ۱۰ نمودار مقایسه‌ای دقت - فراخوانی میان چند روش برتر با روش پیشنهادی نشان داده شده است. منحنی ROC معمولاً برای مسائل تصمیم‌گیری باینری در یادگیری ماشین استفاده می‌شود. این در حالی است که اگر مجموعه داده‌ای نا اریب باشد، منحنی دقت - فراخوانی تصویر با مفهوم‌تری از مقایسه الگوریتم‌ها نشان می‌دهد [۱۸]. لازم به ذکر است که برای گراف irBlogs در تمامی الگوریتم‌ها جهت را در نظر نگرفته‌ایم. پس از بررسی نتایج به این جمع‌بندی رسیدیم که روش پیشنهادی از نظر کیفیت خروجی‌ها در برخی از مجموعه‌های داده نسبت به سایر روش‌ها بهترین می‌باشد.

۵- نتیجه‌گیری

در این مقاله یک روش جدید پیش‌بینی لینک از نوع ساختاری بدون ناظر معرفی شد. این روش بر اساس رفتار مورچه‌ها در زمان پیدا کردن غذا می‌باشد. ارزیابی‌های مختلفی برروی الگوریتم پیشنهادی صورت پذیرفت. از این جمله موارد می‌توان به روش دقت n جواب اول، مساحت زیر نمودار ROC و مساحت زیر نمودار دقت - فراخوانی اشاره کرد. انواع مختلفی از مجموعه‌های دادگان برای آزمون الگوریتم نیز استفاده شد. نتایج حاکی از آن است که در برخی از مجموعه‌های دادگان الگوریتم پیشنهادی دارای بهترین عملکرد نسبت به سایر روش‌ها برخوردار است. در سایر مجموعه‌های دادگان نتایج حاصله نزدیک به بهترین الگوریتم‌ها می‌باشد. هیچ روش ساختاری در تمامی مجموعه‌های دادگان دارای بهترین جواب‌ها نمی‌باشد، بنابراین نوع ساختار گراف هر مجموعه داده‌ای مناسب برای یک سری از روش‌ها برای پیش‌بینی لینک می‌باشد. روش ارائه شده به نوعی یک روش تقریبی می‌باشد و بنابراین هدف اصلی آن بهینه‌سازی زمان اجرا می‌باشد. در مقایسه با سایر روش‌ها روش پیشنهادی دارای زمان اجرا مناسبی در مجموعه دادگان بزرگ می‌باشد.

جدول ۵- مقایسه دقت الگوریتم‌ها با یکدیگر نسبت به پیش‌بینی کننده تصادفی

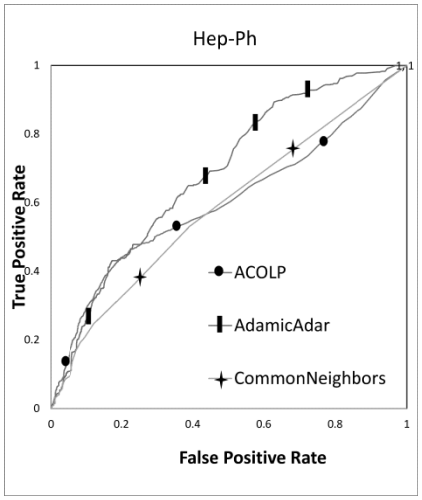
ACOLP	PropFlow	SimRank	Rooted PageRank	Graph Distance	Katz	Preferential Attachment	Jaccard Coefficient	Adamic Adar	Common Neighbor	
۱۹.۰۳	۱۶.۳۶	۱۵.۰۷	۱۶.۴۲	۹.۴۰	۱۶.۷۶	۴.۶۸	۱۶.۴۳	۱۶.۸۰	۱۸.۰۰	astro-ph
۳۰.۴۸	۲۷.۸۷	۲۶.۲۳	۲۸.۴۵	۱۱.۷۴	۲۴.۰۶	۱۴.۸۳	۲۷.۷۳	۲۸.۷۴	۲۴.۶۴	hep-ph
۴.۱۶	۲.۹۳	۰.۰۳	۳.۷۸	۱.۷۸	۴.۳۲	**۴.۵۰**	۲.۶۴	۴.۱۱	۴.۰۹	disease-p
۱۸.۸۲	۱۴.۹۰	۱.۴۱	۱۷.۴۳	۲.۸۲	۱۶.۵۹	۷.۸۱	۹.۶۴	۱۷.۳۴	۱۶.۹۴	disease-g
۵.۷۰	۳.۹۱	۰.۳۸	۴.۱۲	۳.۱۳	۵.۵۳	۵.۰۱	۴.۶۸	۵.۶۴	۵.۶۷	huddle
۲۹۲.۳۳	۶۵۰.۶۷	-	**۶۵۹.۳۳**	-	۱۵۸.۰۰	۴۷.۰۰	۴۲۱.۰۰	۶۴۷.۶۶	۲۳۵.۰۰	irBlogs

جدول ۶- مساحت زیر نمودار ROC

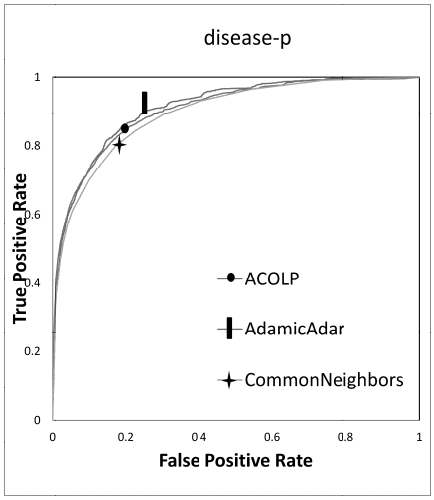
ACOLP	PropFlow	SimRank	Rooted PageRank	Graph Distance	Katz	Preferential Attachment	Jaccard Coefficient	Adamic Adar	Common Neighbor	
۰.۶۰۲۶	۰.۵۱۷۷	۰.۳۸۰۷	۰.۵۰۴۷	۰.۵۰۲۹	۰.۶۰۳۰	**۰.۶۶۲۹**	۰.۴۶۴۷	۰.۶۳۳۸	۰.۶۰۵۳	astro-ph
۰.۶۰۰۷	۰.۶۴۶۱	۰.۵۶۷۱	۰.۶۴۵۴	۰.۵۰۰۰	۰.۵۰۶۸	۰.۴۷۸۳	۰.۶۴۹۲	**۰.۶۸۴۳**	۰.۵۸۶۰	hep-ph
۰.۸۵۷۱	۰.۷۴۶۶	۰.۱۹۹۴	۰.۷۸۳۳	۰.۵۰۰۰	۰.۹۱۵۰	**۰.۹۲۹۵**	۰.۷۴۶۵	۰.۸۹۲۲	۰.۸۹۰۴	disease-p
۰.۹۰۵۴	۰.۸۱۷۰	۰.۵۸۵۳	۰.۸۴۵۱	۰.۵۰۰۰	۰.۸۹۳۷	۰.۸۱۳۸	۰.۷۹۱۷	**۰.۹۱۰۸**	۰.۸۹۴۷	disease-g
۰.۷۵۶۴	۰.۶۷۰۰	۰.۵۲۱۲	۰.۶۸۶۰	۰.۵۰۰۰	۰.۷۶۱۱	۰.۷۴۲۹	۰.۷۲۳۷	**۰.۷۶۸۸**	۰.۷۶۷۷	huddle
۰.۶۸۸۳	۰.۵۸۰۵	-	۰.۵۸۸۰	-	**۰.۸۰۰۱**	۰.۵۴۵۶	۰.۵۰۰۰	۰.۵۴۶۲	۰.۶۶۵۳	irBlogs

جدول ۷- مساحت زیر نمودار دقت- فراخوانی

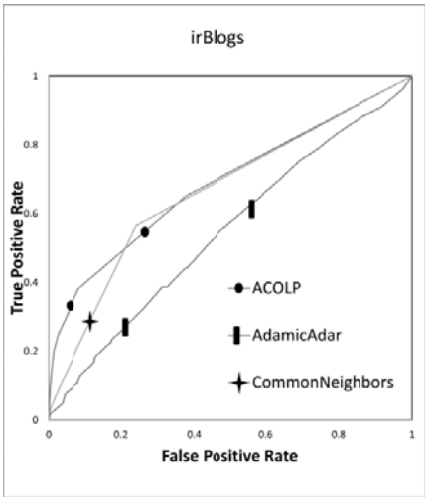
ACOLP	PropFlow	SimRank	Rooted PageRank	Graph Distance	Katz	Preferential Attachment	Jaccard Coefficient	Adamic Adar	Common Neighbor	
۰.۰۲۴۳۶۴	۰.۰۱۴۳۸۸	۰.۰۱۰۶۵۸	۰.۰۱۳۹۹۶	۰.۰۰۷۷۰۶	۰.۰۲۳۷۱۷	**۰.۰۲۸۴۶۴**	۰.۰۱۳۰۵۰	۰۰۲۷۲۷۲	۰۰۲۴۷۶۵	astro-ph
۰.۰۱۲۸۲۵	۰.۰۰۹۸۳۷	۰.۰۰۷۶۴۹	۰.۰۰۹۸۶۶	۰.۰۰۶۶۲۵	۰.۰۰۸۰۸۸	۰.۰۰۸۹۵۲	۰.۰۰۹۷۷۷	۰.۰۱۲۷۶۱	۰.۰۰۹۸۰۳	hep-ph
۰.۶۹۰۴۶۹	۰.۳۵۱۸۳۳	۰.۰۹۵۹۴۵	۰.۴۹۸۸۸۵	۰.۱۵۷۸۳۷	۰.۷۷۱۳۵۴	**۰.۷۸۵۰۲۲**	۰.۴۶۰۸۶۵	۰.۷۳۲۴۷۲	۰.۷۳۱۳۸۳	disease-p
۰.۴۶۵۶۰۹	۰.۲۳۳۱۸۲	۰.۰۴۵۲۷۲	۰.۲۹۶۹۲۷	۰.۰۳۵۵۹۴	۰.۳۶۷۹۳۷	۰.۱۵۲۸۷۱	۰.۲۰۷۷۱۰	۰.۳۹۵۰۷۹	۰.۳۸۲۲۸۶	disease-g
۰.۲۱۷۴۷۵	۰.۱۲۱۵۰۴	۰.۰۵۶۷۷۲	۰.۱۲۹۲۹۵	۰.۰۵۹۵۷۵	۰.۲۲۲۲۷۴	۰.۱۹۵۴۵۹	۰.۱۷۵۴۵۳	**۰.۲۳۰۳۷۲**	۰.۲۳۰۱۹۲	huddle
۰.۰۳۸۱۰۰	۰.۰۲۱۹۴۹	-	۰.۰۲۳۰۵۶	-	۰.۰۰۹۱۹۲	۰.۰۰۱۳۶۴	۰.۰۰۷۰۰۰	۰.۰۲۰۹۶۷	۰.۰۰۹۷۱۹	irBlogs



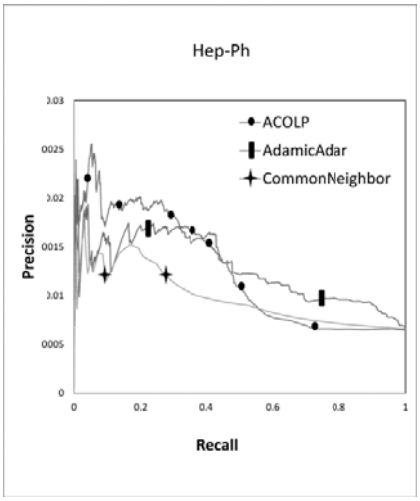
شکل ۵- نمودار مقایسه‌ای ROC برای مجموعه داده‌ای Hep-Ph



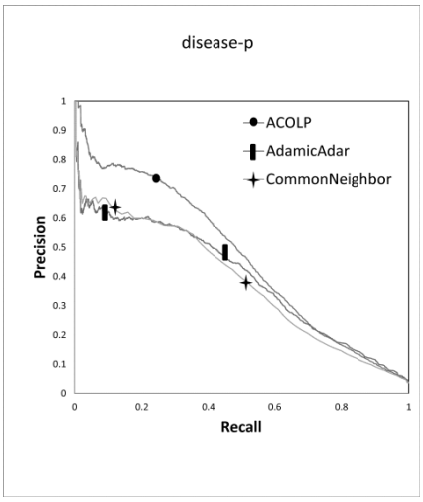
شکل ۶- نمودار مقایسه‌ای ROC برای مجموعه داده‌ای disease-p



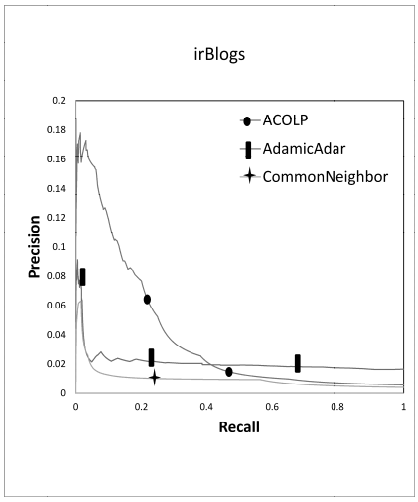
شکل ۷- نمودار مقایسه‌ای ROC برای مجموعه داده‌ای irBlogs



شکل ۸- نمودار مقایسه‌ای Precision-Recall برای مجموعه داده‌ای Hep-Ph



شکل ۹- نمودار Precision-Recall برای مجموعه داده‌ای disease-p



شکل ۱۰- نمودار Precision-Recall برای مجموعه داده‌ای irBlogs

مراجع

[1] L. Linyuan, and T. Zhou, "Link Prediction in Complex Networks: A Survey," *Journal of Statistical Mechanics and Its Applications*, vol. 390, no. 6, pp. 1150-1170, 2011.

[2] D. Liben-Nowell, and J. Kleinberg, "The Link-prediction Problem for Social Networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[3] L. Getoor, and C. P. Diehl, "Link Mining: a Survey," *ACM Trans. SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 3-12, 2005.

[4] D. Yin, L. Hong, and B. D. Davison, "Structural Link Analysis and Prediction in Microblogs," *Proc, ACM Int'l Conf. Information and Knowledge Management*, pp. 1163-1168, 2011.

[5] Z. Huang, X. Li, and H. Chen, "Link Prediction Approach to Collaborative Filtering," *Proc, ACM/IEEE-CS Joint Int'l Conf. Digital Libraries*, pp. 141-142, 2005.

[6] Z. Huang, and D. K. Lin, "The Time-series Link Prediction Problem with Applications in Communication Surveillance," *INFORMS Journal on Computing,* vol. 21, no. 2, pp. 286–303, 2009.

[7] M. Tsuyoshi, and S. Moriyasu, "Link Prediction of Social Networks Based on Weighted Proximity Measures," *Proc, IEEE/WIC/ACM Int'l Conf. Web Intelligence,* pp. 85-88, 2007.

[8] T. Tylenda, R. Angelova, and S. Bedathur, "Towards Time-aware Link Prediction in Evolving Social Networks," *Proc, ACM Int'l Workshop. Social Network Mining and Analysis*, pp. 9, 2009.

[9] K. Juszczyszyn, K. Musial, and M. Budka, "Link Prediction Based on Subgraph Evolution in Dynamic Social Networks," *Proc, IEEE Int'l Conf. Social Computing*, pp. 27-34, 2011.

[10] R. N. Lichtenwalter, and N. V. Chawla, "Vertex Collocation Profiles: Subgraph Counting for Link Analysis and Prediction," *Proc, Int'l Conf. World Wide Web*, pp. 1019-1028, 2012.

[11] Z. Huang, "Link Prediction Based on Graph Topology: The Predictive Value of the Generalized Clustering Coefficient," *Proc, Int'l Workshop. Link Analysis: Dynamics and Static of Large Networks*, pp. 1-9, 2006.

[12] M. Dorigo, and C. Blum, "Ant Colony Optimization Theory: A Survey," *Journal of Theoretical Computer Science*, vol. 344, no. 2, pp. 243-278, 2005.

[13] J. Sobecki, and J. M. Tomczak, "Student Courses Recommendation Using Ant Colony Optimization," *Springer Journal of Intelligent Information and Database Systems*, pp. 124-133, 2010.

[14] S. Fenet, and C. Solnon, "Searching for Maximum Cliques with Ant Colony Optimization," *Springer Journal of Applications of Evolutionary Computing*, pp. 236-245, 2003.

[15] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New Perspectives and Methods in Link Prediction," *Proc, ACM Int'l Conf. Knowledge Discovery and Data Mining*, pp. 243-252, 2010.

[16] R. N. Lichtenwalter, and N. V. Chawla, "Lpmade: Link Prediction Made Easy," *Journal of Machine Learning Research,* vol. 12, no. 2, pp. 2489-2492, 2011.

[17] D. J. Hand, "Measuring Classifier Performance: a Coherent Alternative to the Area under the ROC Curve," *Journal of Machine Learning,* vol. 77, no. 1, pp. 103-123, 2009.

[18] J. Davis, and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," *Proc, ACM Int'l Conf. Machine Learning*, pp. 233-240, 2006.

[19] M. Latapy, "Main-memory Triangle Computations for Very Large (sparse (power-law)) Graphs," *Journal of Theoretical Computer Science*, vol. 407, no. 1, pp. 458-473, 2008.

[20] A. AleAhmad, and A. Habibian, "A standard Persian weblog collection, Database Research Group," irBlogs of ECE Department, University of Tehran, May 2010.

[21] T. Zhou, L. Lü, and Y. C. Zhang, "Predicting Missing Links via Local Information," *Journal of European Physical*, vol. 71, no. 4, pp. 623-630, 2009.

احسان شرکت در سال ۱۳۶۵ در شهر تاریخی اصفهان دیده به جهان گشود. وی مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار در سال ۱۳۸۹ از دانشگاه اصفهان و مدرک کارشناسی‌ارشد خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار در سال ۱۳۹۱ از دانشگاه تهران اخذ نمود. ایشان در حال حاضر پژوهشگر در مرکز ملی فضای مجازی (مرکز تحقیقات مخابرات) می‌باشد.

آدرس پست‌الکترونیکی ایشان عبارت است از:

e.sherkat@ut.ac.ir

مسعود رهگذر دوره کارشناسی خود را از دانشگاه صنعتی شریف در گرایش مهندسی برق در سال ۱۳۵۸ به اتمام رسانید. وی در سال ۱۳۶۲ موفق به اخذ مدرک کارشناسی‌ارشد در زمینه سیستم‌های پایگاه داده از دانشگاه پاریس ۶ شده و سپس با ادامه تحصیل در همین دانشگاه، موفق به اخذ مدرک دکتری در زمینه پایگاه‌های داده توزیع شده در سال ۱۳۶۶ شد. ایشان دارای حدود ۲۰ سال سوابق کاری در صنایع خارج از کشور در فرانسه و اروپا بوده و طی سال‌های ۱۳۵۹ تا ۱۳۷۹ در بیش از ۷۰ پروژه انفورماتیک در کشورهای اروپایی، فرانسه و آمریکا و نیز طی سال‌های ۱۳۷۹ تا ۱۳۹۲ در بیش از ۱۵ پروژه پژوهشی کاربردی در صنایع و سازمان‌های دولتی و خصوصی در ایران مشارکت داشته است. احراز بیش از ۱۰ مورد مسئولیت‌های اجرایی در دانشگاه تهران طی سال‌های ۱۳۷۹ تا ۱۳۹۲ بخشی دیگر از فعالیت‌های ایشان می‌باشد. وی تا کنون بیش از ۱۱۰ مقاله در کنفرانس‌ها و مجلات علمی پژوهشی بین‌المللی به چاپ رسانیده است. در حال حاضر عضو هیأت علمی و دانشیار دانشکده مهندسی برق و کامپیوتر پردیس دانشکده‌های فنی دانشگاه تهران بوده و زمینه‌های مورد علاقه ایشان شامل معماری سیستم‌ها و پایگاه‌های داده توزیع یافته، رایانش ابری، جستجوی اطلاعات، داده‌کاوی، شبکه‌های اجتماعی و بیوانفورماتیک می‌باشد.

آدرس پست‌الکترونیکی ایشان عبارت است از:

rahgozar@ut.ac.ir

مسعود اسدپور در سال ۱۳۵۴ در شهرستان لامرد فارس متولد شد. وی مدرک کارشناسی خود را از دانشگاه صنعتی شریف تهران در رشته مهندسی کامپیوتر- نرم‌افزار در سال ۱۳۷۶ اخذ کرد. ایشان در سال ۱۳۷۸کارشناسی‌ارشد رشته هوش ماشین و رباتیک دانشگاه تهران را به اتمام رسانید. وی دوره دکترا و فوق دکترای خود را در زمینه های یادگیری ماشینی و رباتیک گروهی در دانشگاه EPFL سوئیس در سال‌های ۲۰۰۲ تا ۲۰۰۷ گذراند. وی سپس به عنوان هیات علمی دانشکده برق و کامپیوتر دانشگاه تهران به گروه هوش ماشین و رباتیک این دانشگاه پیوست و آزمایشگاه شبکه‌های اجتماعی را بنا نهاد. او عضو وابسته گروه‌های نرم‌افزار و فناوری اطلاعات این دانشکده نیز می‌باشد. وی از سال ۲۰۰۵ عضو IEEE بوده است و در سال‌های ۱۹۹۸تا ۲۰۰۱ و سپس از ۲۰۱۰ تا ۲۰۱۲ با IPM همکاری داشته است. موضوعات پژوهشی مورد علاقه وی عبارتند از: شبکه‌های اجتماعی، رباتیک گروهی، یادگیری ماشینی و الهام از طبیعت در رباتیک.

آدرس پست الکترونیکی ایشان عبارت است از:

asadpour@ut.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۹۱/۷/۱۰
تاریخ اصلاح: ۹۲/۵/۲۴
تاریخ قبول شدن: ۹۲/۶/۱۱
نویسنده مرتبط: احسان شرکت، دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران، تهران، ایران.

---

[1]Ant Colony Optimization
[2]Precision-Recall
[3]Top-N Precision
[4]Snapshot
[5]Link Mining
[6]Protein-Protein Interaction (PPI) Network
[7]Similarity Measures
[8]Triad
[9]Supervised
[10]Triad Transition Matrix (TTM)
[11]Vertex Collocation Profiles
[12]Learning
[13]Clustering Coefficient
[14]Generalized Clustering Coefficient
[15]Clique
[16]Transitivity
[17]Space Complexity
[18]Power-Law graphs
[19]Sparse
[20]Score
[21]Centrality
[22]High Energy Physics-Phenomenology
[23]Astrophysics
[24]Unsupervised