علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۷، شماره ۱، ۱۳۹۸
صفحات ۷۳-۶۸
مقاله عادی

پیش‌بینی پیوند در شبکه‌های اجتماعی به‌وسیله تخصیص درجه‌ی همسایگی به رئوس در گراف‌های بدون جهت

حدیث بشیری*[۱]، غلامحسین دستغیبی فرد[۲]

*نویسنده مسئول، دریافت: ۹۷/۰۷/۱۷، بازنگری: ۹۷/۰۹/۱۸، پذیرش: ۹۸/۰۵/۲۵

[۱] دانشجوی کارشناسی ارشد، مهندسی کامپیوتر،دانشکده برق و کامپیوتر، دانشگاه شیراز، شیراز، ایران

[۲] دانشیار، مهندسی کامپیوتر، دانشکده برق و کامپیوتر، دانشگاه شیراز، شیراز، ایران

چکیده

امروزه شبکه‌های اجتماعی مثل فیس‌بوک، گوگل‌پلاس، اینستاگرام و غیره در زندگی افراد تأثیر بسزایی دارند. در این شبکه‌ها برای پیشنهاد افراد به یکدیگر از الگوریتم‌های پیش‌بینی پیوند استفاده می‌شود و یکی از موضوعات چالش‌برانگیز و پرکاربرد می‌باشد. محققین الگوریتم‌های مختلفی برای پیش‌بینی پیوند ارائه کرده‌اند، اما مشکل عمده الگوریتم‌های موجود، دقت پایین آن‌هاست. با توجه به اینکه درصد ارتباطات در شبکه‌های اجتماعی متفاوت است، در این مقاله با استفاده از وزن‌دار کردن یال‌ها و تخصیص درجه‌ی همسایگی، الگوریتمی برای تشخیص دوستان صمیمی ارائه داده‌ایم. آزمایش الگوریتم پیشنهادی بر روی سه مجموعه داده Facebook و Hamster و Email صورت گرفته است و در مقایسه با الگوریتم‌های جاری به ترتیب ۰,۴ ، ۲,۴ و ۶,۹ درصد بهبود دقت داشته‌ایم.

کلمات کلیدی: شبکه‌های اجتماعی، پیش‌بینی پیوند، آدامیک آدار، وزن دهی به یال‌ها، تخصیص درجه‌ی همسایگی، تشخیص دوستان صمیمی

۱- مقدمه

در شبکه‌های اجتماعی افرادی با اهداف خاص گرد هم می‌آیند و با یکدیگر به تعامل می‌پردازند و تصاویر و اطلاعاتی را باهم به اشتراک می‌گذارند. فیس‌بوک، اینستاگرام و گوگل‌پلاس و غیره، نمونه‌هایی از این شبکه‌ها می‌باشند. تعداد کاربران در شبکه‌های اجتماعی به‌صورت روزافزون در حال افزایش است، به‌عنوان‌مثال کاربران فعال در فیس‌بوک در سال ۲۰۱۷ برابر با ۲ میلیارد نفر است [۱].

شبکه‌های اجتماعی را معمولاً با یک گراف G=(V,E) نمایش می‌دهند، که V نشان‌دهنده گره‌ها است و E یال‌های گراف را نشان می‌دهد. گره‌ها و یال‌ها در شبکه‌های متفاوت ممکن است معناهای متفاوتی داشته باشند، به‌عنوان‌مثال در شبکه‌های اجتماعی، گره‌ها نشان‌دهنده افراد و یال‌ها ارتباط بین آن‌ها را نشان می‌دهند. یال‌ها می‌توانند جهت‌دار (یک‌طرفه) و یا بدون جهت (دوطرفه) باشند. شبکه‌ی اجتماعی فیس‌بوک یک شبکه با گراف بدون جهت است و وجود یال u به vدر شبکه به معنی وجود یال v به u نیز می‌باشد. مجموعه داده‌های استفاده شده در این مقاله همگی بدون جهت می‌باشند.

هدف پیش‌بینی پیوند این است که با استفاده از وضعیت شبکه در لحظه t، بتوانیم وضعیت شبکه در لحظه t+1 را تعیین کنیم. به‌عبارت‌دیگر می‌خواهیم یال‌هایی که در لحظه t+1 با احتمال بالا ممکن است، اضافه شوند را پیش‌بینی کنیم [۲].

به دلیل پویا بودن شبکه‌های اجتماعی، پیش‌بینی پیوند چالش‌های زیادی دارد که یکی از آن‌ها ساختمان داده مناسب است. زیرا در زمان‌های مختلف تعداد افراد شبکه و ارتباط بین آن‌ها ممکن است تغییر کند. به‌عبارت‌دیگر ممکن است، افراد و ارتباط بین این افراد اضافه و یا حذف (به دلیل حذف حساب کاربری یک شخص یا لغو دوستی بین دو فرد در شبکه) شود. برای انجام این کار باید ساختمان داده مناسبی برای شبکه داشته باشیم که در هر لحظه بتوانیم اطلاعات گراف را استخراج کنیم [۳].

پیش‌بینی پیوند کاربردهای متعددی دارد: پیدا کردن دوستان و آشنایان، پیدا کردن افراد با سلایق یکسان در شبکه‌های اجتماعی، کشف گروه‌های مجرم و تروریست [۴]، سیستم‌های پیشنهاددهنده در خریدهای اینترنتی معرفی کالا به دوستان و آشنایان [۵]، پیدا کردن رابطه بین پروتئین‌ها در بیوانفورماتیک[۶]، و غیره نمونه‌هایی از کاربردهای آن می‌باشند.

هدف این مقاله پیش‌بینی پیوند در شبکه‌های اجتماعی است. مجموعه داده‌های شبکه‌های اجتماعی می‌توانند همگن و یا ناهمگن باشند. در مجموعه داده‌های همگن فقط اطلاعات ساختار شبکه در دسترس است؛ اما مجموعه داده‌های ناهمگن ممکن است حاوی اطلاعات دیگری مثل نوع دوستی یا ارتباط، تعداد پست، چت و یا تگ (tag)های بین دو کاربر و غیره نیز باشد [۷]. در این مقاله از شبکه‌های همگن استفاده کرده‌ایم.

یکی از چالش‌های پیش‌بینی پیوند "شروع سرد" [۶] نام دارد. وقتی یک فرد جدید وارد یک شبکه اجتماعی می‌شود، هیچ اطلاعاتی از او نداریم تا بتوانیم ارتباطات آینده وی را پیش‌بینی کنیم. در چنین وضعیتی در شبکه‌ی فیس‌بوک، اگر افراد با ایمیل ثبت‌نام کرده باشد، برای پیشنهاد دوست، مخاطب‌های ایمیل او که در فیس‌بوک هستند را معرفی می‌کند و اگر از طریق موبایل ثبت‌نام کرده باشد مخاطب‌های تلفن او که در فیس‌بوک ثبت‌نام کرده‌اند را پیشنهاد می‌دهد [۸]. در گوگل‌پلاس افراد جدید موظف‌اند که در ابتدا ده نفر را به‌عنوان دوست انتخاب و معرفی کند [۹].

بقیه مقاله به شکل زیر سازماندهی شده است: در بخش ۲ ادبیات موضوع و در بخش ۳ پیشینه پژوهشی را به‌اختصار شرح می‌دهیم، در بخش ۴ روش پیشنهادی را به‌تفصیل توضیح می‌دهیم. در بخش ۵ به شرح مجموعه داده‌ها، متریک‌های ارزیابی می‌پردازیم و درنهایت نتایج ارزیابی روش پیشنهادی را با روش‌های جاری آمده است.

۲- ادبیات موضوع

در این بخش به شرح اهمیت، کاربرد، انواع و روش‌های پیش‌بینی پیوند می‌پردازیم.

۲-۱- اهمیت و کاربرد پیش‌بینی پیوند [۶]

۲-۱-۱- شبکه‌های اجتماعی و تجارت الکترونیک

با استفاده از پیش‌بینی پیوند می‌توان با پیشنهاد کردن کاربران جدید به‌عنوان دوست به کاربر و یا پیشنهاد کردن مطالبی که مورد پسند کاربر واقع می‌شوند، وی را ترغیب به حضور در شبکه و فعالیت نمایند. همچنین سیستم‌های توصیه گر، در سامانه‌های تجاری از قبیل سامانه‌های خرید و فروش نیز مورد استفاده قرار می‌گیرند تا بهترین محصولاتی که مورد پسند کاربر واقع می‌شوند را پیش‌بینی کرده و برای خرید به وی پیشنهاد دهد.

۲-۱-۲- بیوانفورماتیک و شبکه‌های بیولوژیکی

پیش‌بینی پیوند در این نوع شبکه‌ها می‌تواند به کشف آنتولوژی[۱] ژن‌های جدید، کشف ارتباطات ناشناخته میان ژن‌ها، کشف ژن‌هایی که حامل یک بیماری خاص هستند، کشف تأثیر داروها بر ژن‌ها، کشف داروهای جدید و کاربردهای دیگر کمک بسیاری کند.

۲-۱-۳- سامانه‌های امنیتی

مثالی از این قسمت می‌تواند شبکه‌ی تروریست‌ها باشد. پیش‌بینی پیوند در این شبکه‌ها می‌تواند کاربرد زیادی داشته باشد. کشف روابط مخفی میان سازمان‌ها و گروه‌های تروریستی می‌تواند به شناسایی و دستگیری تروریست‌ها و همچنین جلوگیری از هرگونه عملیات تروریستی کمک کند.

۲-۱-۴- امنیت اطلاعات

در امنیت اطلاعات نیز پیش‌بینی پیوند می‌تواند در پیدا کردن کامپیوتر بعدی که ممکن است در شبکه‌ی آلوده مورد حمله قرار گیرد، باعث ایمن سازی کامپیوترها در برابر ویروس‌های کامپیوتری و جلوگیری از انتشار ویروس شود.

۲-۲- انواع پیش‌بینی پیوند

مسئله پیش‌بینی پیوند به دو شکل ساختاری[۲] و زمانی مطرح می‌شود، ما در این مقاله از پیش‌بینی پیوند ساختاری استفاده می‌کنیم. که در ادامه به‌تفصیل شرح داده می‌شوند.

۲-۲-۱- پیش‌بینی پیوند ساختاری

در این حالت ما تصویری از گراف داریم که در آن تعدادی از یال‌ها وجود دارند و به دنبال آن هستیم که در مورد یال‌هایی که وجود ندارند پیش‌بینی انجام دهیم. عدم اطلاع ما در مورد یال‌های ناموجود به این دلیل است که این یال‌ها هنوز تشکیل نشده‌اند، به‌عبارت‌دیگر در شبکه‌های اجتماعی هنوز بین دو فرد دوستی انجام نگرفته است و یا به این دلیل است که فهمیدن اینکه بین دو گره ارتباط برقرار است یا خیر، هزینه‌بر یا غیرممکن است.

۲-۲-۲- پیش‌بینی پیوند زمانی

در این نوع پیش‌بینی پیوند ما تعدادی تصویر[۳] از کل گراف در زمان‌های مختلف به دست می‌آوریم و به بررسی مدل گسترش گراف در زمان می‌پردازیم. درنهایت هدف یافتن ساختار کل گراف در آینده است. ورودی ممکن است به دو صورت باشد: یا مجموعه‌ی یال‌ها به همراه زمان تشکیل آن‌هاست و یا تصویرهای مختلف گراف در زمان‌های $t_1, t_2, t_3, \dots, t_k$ است و هدف تشخیص t_n است که $t_n > t_k$ [۱۰].

معیار اولیه برای پیش‌بینی پیوند محاسبه میزان شباهت بین دو گره u و v است که با S_{uv} نشان داده می‌شود. در گراف‌های بدون جهت $S_{vu} = S_{uv}$. به‌طورمعمول روش‌های پیش‌بینی پیوند به سه دسته محلی، سراسری و نیمه محلی تقسیم می‌شوند [۱۱] که به شرح آن‌ها می‌پردازیم:

۲-۳- روش محلی

روش‌های محلی برای پیش‌بینی پیوند از مسیرهای به طول یک برای پیشنهاد ارتباط استفاده می‌کنند. به‌عبارت‌دیگر فقط دوستی‌های مستقیم بین دو شخص را در نظر می‌گیرد. این روش‌ها پیچیدگی زمانی کمتری دارند و معمولاً دقت پایین‌تری هم دارند. برخی از روش‌های محلی در ادامه آمده است.

۲-۳-۱- همسایه مشترک (CN)[۴]

یکی از ساده‌ترین شاخص‌ها برای پیش‌بینی پیوند شاخص همسایه‌های مشترک است. هر چه دو گره، دوستان مشترک بیشتری داشته باشند یعنی دو گره خیلی به هم شبیه هستند و در آینده با احتمال بالا باهم دوست می‌شوند. شاخص همسایه مشترک برای دو گره u ,v طبق رابطه (۱) محاسبه می‌شود [۱۲]:

$$S_{u,v} = |\Gamma(u) \cap \Gamma(v)| \qquad (۱)$$

در رابطه (۱)، $u, v \in V$، $\Gamma(u)$ و $\Gamma(v)$ به ترتیب مجموع همسایه‌های گره u و v هستند.

۲-۳-۲- ضریب جاکارد (JC)[۵]

در ضریب جاکارد برای ارزیابی، تعداد دوستان مشترک نسبت به کل دوستان را در نظر می‌گیرد. اگر تعداد همسایه‌های مشترک دو گره u وv برابر با $commom_{uv}$ باشد و تعداد کل همسایه‌های این دو گره برابر با $union_{uv}$ باشد همچنین تعداد همسایه‌های مشترک دو گره x و y برابر با $commom_{xy}$ باشد و تعداد کل همسایه‌های این دو گره برابر با $union_{xy}$ باشد و $commom_{xy} = commom_{uv}$ و $union_{uv} > union_{xy}$ باشد دو گره x,y با احتمال بیشتری در آینده دوست می‌شوند. مقدار شباهت به‌دست‌آمده از روش جاکارد همیشه مقداری بین ۰ و ۱ است. شاخص جاکارد برای دو گره u ,v طبق رابطه (۲) محاسبه می‌شود [۱۱]:

$$S_{u,v} = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \qquad (۲)$$

۲-۳-۳- شاخص آدامیک آدار (AA)[۶]

در اکثر مواقع استفاده تنها از تعداد_همسایه‌های مشترک دو گره کافی نیست، بلکه ویژگی‌های همسایه‌های مشترک را هم باید در نظر بگیریم. یکی از این ویژگی‌ها درجه است. این شاخص از درجه همه همسایه‌های مشترک بین دو گره u,v که با z نشان داده می‌شود، استفاده می‌کند. در این روش گره‌هایی که همسایه‌های کمتری دارند، امتیاز بالاتری می‌گیرند و همچنین هرچقدر که درجه همسایه‌های مشترک دو گره کمتر باشد، احتمال بالاتری می‌دهد که در آینده دوست شوند. شاخص آدامیک آدار برای دو گره u,v با استفاده از رابطه (۳) محاسبه می‌شود [۱۳]:

$$S_{u,v} = \sum_{z \in (\Gamma(u) \cap \Gamma(v))} \frac{1}{\log |\Gamma(Z)|} \qquad (۳)$$

در رابطه (۳)، z همسایه مشترک بین دو گره u , v را نشان می‌دهد.

۲-۴- روش‌های سراسری

روش‌های سراسری مسیرهای با طول بیشتر از یک را در نظر می‌گیرند و برای پیدا کردن شباهت بین دو گره کل گراف را در نظر می‌گیرند. از همین رو معمولاً دقت بالایی دارند، همچنین این روش‌ها پیچیدگی بالایی دارند و زمان‌بر هستند، به همین دلیل معمولاً برای شبکه‌های آنلاین استفاده نمی‌شوند. برخی روش‌های سراسری در ادامه آمده است.

۲-۴-۱- شاخص کاتز (K)[۷]

شاخص کاتز [۱۴] مسیرهای با طول بزرگ‌تر یا مساوی ۲ را در نظر می‌گیرد و همه مسیرهای موجود در گراف بین جفت گره‌ها را بررسی می‌کند. نکته مهم این است که این مسیر ممکن است طولانی باشد بنابراین باید تأثیر آن‌ها کمتر شود، یعنی مسیر با طول ۲ امتیاز بیشتری نسبت به مسیرهای با طول ۳ دارد. ازاین‌رو، پارامتر β برای کنترل این مسئله وجود دارد. شاخص کاتز برای دو گره u ,v طبق رابطه (۴) محاسبه می‌شود:

$$S_{u,v} = \sum_{L=1}^{\infty} \beta^L . |Paths_{uv}^L| \qquad (۴)$$

در رابطه (۴)، A ماتریس مجاورت گراف است و $|Paths_{uv}^L|$ تعداد مسیرهای با طول L بین دو گره u ,v است. β همان‌طور که توضیح داده شد یک پارامتر کنترلی است که مقدار خیلی کمی دارد.

۲-۴-۲- روش Prop Flow

این روش به این صورت عمل می‌کند که از یک گره مشخص یک جریان به سایر گره‌ها فرستاده می‌شود، اگر گراف وزن‌دار باشد، وزن یال نشان‌دهنده‌ی حداکثر ظرفیت آن یال برای عبور جریان است. بدیهی است هر یالی که وزن بیشتری داشته باشد احتمال بیشتری دارد که جریان از آن عبور کند. همچنین برای گراف‌های بدون وزن، وزن همه یال‌ها برابر با یک در نظر گرفته می‌شود.

۲-۵- روش‌های نیمه محلی

۲-۵-۱- روش FriendLink

این روش همه مسیرهای بین دو گره که طولی کمتر از یک آستانه را دارند بررسی می‌کند. شاخص FriendLink برای دو گره u ,v طبق رابطه (۵) محاسبه می‌شود:

$$S_{u,v} = \sum_{i=2}^{L} \frac{1}{i-1} . \frac{|Paths_{uv}^i|}{\prod_{j=2}^{i}(n-j)} \qquad (۵)$$

در رابطه (۵)، n تعداد کل گره‌های گراف است و $\prod_{j=2}^{i}(n-j)$ تعداد مسیرهای ممکن بین دو گره u ,v است [۱۵].

۳- پژوهش‌های پیشین

قبل از ایجاد و گسترش شبکه‌های اجتماعی روش‌های مختلفی برای اندازه‌گیری شباهت بین دو عنصر در زمینه‌های مختلف اعم از ریاضی، علوم گیاهی، علوم اجتماعی و... معرفی شدند. بعدها با گسترش شبکه‌های اجتماعی از این معیارها برای اندازه‌گیری شباهت بین دو گره استفاده شد. اولین روشی که برای این منظور ارائه شد شاخص جاکارد [۱۱] بود، این روش یک معیار برای مقایسه شباهت‌ها یا تفاوت‌های مجموعه نمونه‌های آماری بود. از مزایای آن محلی بودن آن است ولی این روش برای شبکه‌های بزرگ مانند فیس‌بوک مناسب نیست و دقتی برابر ۰,۴۳ به ما می‌دهد.

بعدازآن روش همسایه‌های مشترک [۱۲] مطرح شد، این روش بر پایه تعداد دوستان بیشتر عمل می‌کند و مناسب برای شبکه‌های مانند فیس‌بوک است. از مزایای آن می‌توان به‌سرعت آن اشاره کرد و نقص این روش این است که ازآنجایی‌که برای پیشنهاد افراد به یکدیگر فقط تعداد همسایه‌های مشترک را بدون در نظر گرفتن کل همسایه‌ها در نظر می‌گیرد، بنابراین برای مواردی که تعداد دوستان مشترک یکسان باشد عملکرد خوبی از خود نشان نمی‌دهد. مشکل دیگر این روش این است که همه همسایه‌ها را یکسان در نظر می‌گیرد. دقت این روش بر روی مجموعه داده فیس‌بوک برابر با ۰,۹۵۲ می‌باشد.

بعدازآن در سال ۲۰۰۳ آدامیک و آدار یک روش محلی به نام آدامیک آدار را ارائه کردند [۱۳]. این معیار باوجود محلی بودن یکی از بهترین معیارها برای پیش‌بینی پیوند به شمار می‌رود و توانسته است بر روی اکثر مجموعه داده‌ها دقت خوبی به ثبت برساند. مزایای آن سرعت بالای آن و در نظر گرفتن درجه هرکدام از همسایه‌های مشترک بین دو گره است. نقص این روش این است که اگر تعداد همسایه‌های گره مشترک زیاد باشد با احتمال کمتری آن را پیشنهاد می‌دهد، حتی اگر همسایه‌های مشترک آن همان گره‌هایی مشترک بین گره‌ها اصلی ما باشند.

در سال ۲۰۱۰ لیتن والتر و همکاران یک روش سراسری با نام prop flow ارائه کردند [۱۶]. این روش علاوه بر اینکه زمان‌بر بود نتوانست دقت خوبی برای شبکه فیس‌بوک به دست آورد، دقت آن برابر با ۰,۴ است. این روش برای شبکه‌های کوچک کاربرد بهتری دارد.

در سال ۲۰۱۲ روش نیمه محلی به نام friendlink [۱۵] ارائه شد، هدف از این روش استفاده از مزایای روش‌های محلی و سراسری بود. صرف‌نظر از زمان‌بر بودن این روش توانست دقت نسبتاً مناسبی برای فیس‌بوک ثبت کند.

در مقاله [۱۷] اکبری و همکاران روش‌های cn*cc, aa*cc, k*cc را ارائه کردند که در آن علاوه بر روش‌های سنتی ذکر شده از اطلاعات اجتماعات نیز استفاده شده است. دقت مسئله اصلی در پیش‌بینی پیوند است این روش‌ها به ازای افزایش کمی پیچیدگی دقت را بهبود می‌بخشد، بااین‌وجود نتوانستند بر روی مجموعه داده فیس‌بوک دقت خوبی به دست آورند.

همان‌طور که گفته شد یکی از معایب روش همسایه‌های مشترک، یکسان در نظر گرفتن همه دوستان می‌باشد. ما با استفاده از وزن دادن به یال‌ها سعی در تشخیص دوستان صمیمی داریم. ضمن اینکه روش آدامیک آدار در آزمایشات توانسته است دقت خوبی به‌جای گذارد، روش ما ترکیبی از وزن دادن به یال‌ها با یک معیار و آدامیک آدار است.

۴- روش پیشنهادی

در گراف‌های بدون وزن، یال‌ها هیچ وزنی ندارند، یعنی همه یال‌ها مشابه و برابر هستند و هیچ‌کدام نسبت به دیگری برتری ندارد. اما در شبکه‌های واقعی ممکن است یک یال دارای اهمیت بیشتری نسبت به دیگری باشد و لزوماً اهمیت همه یال‌ها یکسان نیست. به‌عنوان‌مثال در فیس‌بوک که یک گراف بدون جهت و بدون وزن است، اهمیت دو دوست یک شخص ممکن است یکسان نباشد. دوست یک شخص می‌تواند از اعضای خانواده و یا اینکه یکی از دوستان صمیمی وی باشد. با در نظر گرفتن این موضوع ما به همه‌ی یال‌های موجود در گراف یک درجه همسایگی یا وزن نسبت می‌دهیم. با استفاده از درجه همسایگی آن‌ها می‌توانیم دوستان صمیمی هر گره را مشخص کنیم و ازآنجایی‌که روش آدامیک آدار توانسته است دقت خوبی برای بیشتر مجموعه داده‌ها داشته باشد چون علاوه بر تعداد همسایه‌های مشترک درجه‌ی آن‌ها را هم در نظر می‌گیرد، ما از این شاخص در روش خود استفاده می‌کنیم. پایه روش ما تشخیص دوستان قوی و صمیمی در گراف بدون وزن و بدون جهت با استفاده از وزن گذاری یال‌ها است. روش پیشنهادی به‌صورت گام‌به‌گام به شرح زیر است:

گام ۱: به همه‌ی یال‌های در زمان t با استفاده از رابطه ۶ یک وزن اختصاص می‌دهیم. به‌وسیله این رابطه برای هر رابطه دوستی یک ارزش تعیین می‌کنیم.

$$\text{weights}(u,v)=\sum_{z\in(\Gamma(u)\cap\Gamma(v))}\min\left(\frac{|\Gamma(u)\cap\Gamma(z)|}{|\Gamma(u)\cup\Gamma(z)|},\frac{|\Gamma(z)\cap\Gamma(v)|}{|\Gamma(z)\cup\Gamma(v)|}\right) \qquad (۶)$$

در رابطه (۶) $\Gamma(u)$ همسایه‌های گره u رانشان می‌دهد و $|\Gamma(u)|$ تعداد همسایه‌های u را نشان می‌دهد. این رابطه ترکیب روش‌های آدامیک آدار و جاکارد است؛ دلیل استفاده از آدامیک آدار استفاده از خصوصیات همسایه‌های مشترک دو گره است و دلیل استفاده از روش جاکارد به دست آوردن مقداری است که میزان همسایگی دو گره را نشان می‌دهد. درواقع ما برای همه همسایه‌های مشترک دو گره، نسبت همسایگی گره اول با گره میانی و گره دوم با گره میانی را محاسبه می‌کنیم و در هر تکرار کمترین مقدار بین این دو مقدار را انتخاب می‌کنیم، وزن نهایی مجموع مقدار مینیمم‌ها به ازای همه همسایه‌های مشترک است.

گام ۲: دوستان صمیمی همه گره‌ها به‌وسیله‌ی مقایسه وزن آن‌ها با میانگین وزن‌ها مشخص می‌شوند.

گام ۳: با استفاده از روش آدامیک آدار شروع به پیش‌بینی پیوند می‌کنیم و یک لیست کاندید از یال‌هایی با احتمال بالا برای تشکیل شدن در لحظه t+1 می‌سازیم.

گام ۴: برای دو گره"a", "b" که در لیست کاندید هستند امتیاز آن‌ها را با استفاده از رابطه (۷) محاسبه می‌کنیم.

$$score(a,b)=\sum_{z\in a\cap b}\min(w_{az},w_{zb}) \qquad (۷)$$

به‌طوری‌که w_{az} وزن گره"z" در گره"a" است.

گام ۵: نتایج نهایی یال‌هایی که بزرگ‌تر یا مساوی با رابطه (۸) هستند به لیست پیشنهادها منتقل می‌شوند.

$$Avg * |\Gamma(a)\cap\Gamma(b)| \qquad (۸)$$

یعنی به ازای هر همسایه مشترک یک دوست صمیمی باید وجود داشته باشد.

گام ۶: به ازای هر یالی که پیش‌بینی می‌کنیم وزن‌های مرتبط با یال جدید را به‌روز می‌کنیم.

گام ۷: صد تا از یال‌های موجود در لیست پیشنهادها که بالاترین احتمال برای ایجاد در آینده را دارند انتخاب می‌کنیم (precision @100).

۵- ارزیابی نتایج

۵-۱- مجموعه داده‌ها

برای ارزیابی روش پیشنهادی و مقایسه آن با روش‌های دیگر، از سه مجموعه داده همگن استفاده کرده‌ایم:

الف- Facebook[۸]: که یک شبکه اجتماعی محبوب است و در این مجموعه داده گره‌ها افراد رانشان می‌دهند و یال‌ها رابطه دوستی بین آن‌ها را نشان می‌دهد.

ب- Petster hamster[۹]: که شامل دوستان و خانواده افراد در وب‌سایت hamsterster.com است.

پ- Email Enron[۱۰]: این مجموعه داده شامل حدود نیم میلیون ایمیل است. ابن داده‌ها در ابتدا به‌صورت عمومی و توسط کمیسیون تنظیم مقررات انرژی فدرال در تحقیقات خود، به وب فرستاده شد. گره‌های شبکه آدرس‌های ایمیل هستند و اگر از آدرس i یک ایمیل به آدرس j فرستاده شده باشد یک یال از گره i به گره j وجود دارد. خصوصیات این دو مجموعه داده در جدول ۱ نشان داده شده است:

جدول ۱- خصوصیات مجموعه داده‌ها

Dataset	#Nodes	#Edges
Facebook	4039	88234
Hamster	2426	16631
Email	36691	183830

۵-۲- ارزیابی

برای انجام پیش‌بینی پیوند ما نیاز به دو مجموعه داده داریم که عبارت‌اند از : مجموعه داده آزمایشی و مجموعه داده آموزشی[۱۱]. مجموعه داده آزمایشی مجموعه‌ای از یال‌هایی است که در زمان t+1 به گراف اضافه شده‌اند. مجموعه داده آموزشی مجموعه‌ای از همه یال‌های گراف در زمان t است. در مجموعه داده‌های مورد آزمایش، مجموعه داده‌های آزمایشی و آموزشی به‌صورت جداگانه وجود ندارد، بلکه فقط یک مجموعه وجود دارد که شامل همه یال‌ها در زمان t+1 است (۹). بنابراین ما به‌صورت تصادفی ۱۰٪ از کل یال‌ها را حذف می‌کنیم. مجموعه یال‌های حذف شده مجموعه داده آزمایشی را تشکیل می‌دهند و باقیمانده آن‌ها مجموعه داده آموزشی است. درنهایت ما از مجموعه داده آموزشی برای پیش‌بینی و برای اعتبارسنجی از مجموعه داده آزمایشی استفاده می‌کنیم. به این صورت که اگر یالی که پیش‌بینی کرده‌ایم جز این لیست باشد دقت افزایش پیدا می‌کند.

$$E = E_{train} \cup E_{test} \quad (۹)$$

قبل از شرح نتایج لازم است چند عبارت را به‌اختصار تعریف کنیم [۱۸]:

- True Positive (TP): مجموع تعداد یال‌هایی که روش موردنظر آن‌ها را پیش‌بینی کرده و آن‌ها در مجموعه داده آزمایشی وجود دارند. به‌عبارت‌دیگر در آینده رخ خواهند داد.
- False Positive (FP): مجموع تعداد یال‌هایی که روش موردنظر آن‌ها را پیش‌بینی کرده اما آن‌ها در مجموعه داده آزمایشی وجود ندارند و به‌عبارت‌دیگر در آینده رخ نخواهند داد.
- Precision: دقت پیش‌بینی را محاسبه می‌کند. ما از این معیار برای مقایسه روش‌ها استفاده کرده‌ایم. دقت هر روش به شکل رابطه (۱۰) محاسبه می‌شود.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (۱۰)$$

۵-۳- نتایج

در این قسمت ما نتایج روش پیشنهادی را با چند روش که قبلاً توضیح داده شده‌اند مقایسه می‌کنیم. جدول ۲ دقت این روش‌ها را نمایش می‌دهد. نتایج از پنج بار اجرای برنامه و سپس گرفتن میانگین از دقت‌ها، به‌دست‌آمده است.

جدول ۲- نتایج محاسبه دقت ۱۰۰ عنصر بالا

Name	CN*CC	AA*CC	K*cc	Proposed Method
Facebook	0.91	0.90	0.95	0.954
Hamster	0.774	0.84	0.732	0.866
Email	0.813	0.833	0.70	0.902

همان‌طور که در این جدول مشاهده می‌شود، با استفاده از روش پیشنهادی در هر سه مجموعه داده دقت افزایش پیدا کرده است. بهترین دقت برای مجموعه داده فیس‌بوک با استفاده از روش katz که یک روش سراسری و زمان‌بر است به‌دست‌آمده و برابر با ۰,۹۵ می‌باشد و نتایج روش پیشنهادی برای این مجموعه داده برابر با ۰,۹۵۴ است. برای مجموعه داده همستر بهترین دقت با استفاده از روش AA*CC به‌دست‌آمده که برابر با ۰,۸۴ و دقت روش پیشنهادی برابر با ۰,۸۶۶ است. همچنین برای مجموعه داده Email بهترین دقت به‌دست‌آمده برابر ۰,۸۳۳ است و دقت روش پیشنهادی ما برابر ۰,۹۰۲ است. در جدول ۳ نتایج بهبودهای به‌دست‌آمده با استفاده از هر قسمت از روش پیشنهادی نشان داده شده است.

جدول ۳- جزئیات بهبودهای به‌دست‌آمده

Hamster	Email	Facebook	میزان بهبود
1.09	2.947	0.205	دوستان سازگارتر
0.83	1.03	0.157	دوستان کمتر
0.68	2.923	0.038	دوستان صمیمی‌تر
2.01	3.48	0.309	دوستان سازگارتر و کمتر
1.65	2.11	0.181	دوستان سازگارتر و صمیمی‌تر
0.93	2.97	0.105	دوستان کمتر و صمیمی‌تر
2.4	6.9	0.4	جمع کل بهبودها

ما در روش جدید سه مؤلفه را باهم و درمجموع، به‌عنوان یک روش معرفی کرده‌ایم . دوستان کمتر، دوستان سازگارتر و دوستان صمیمی‌تر . همان‌طور که در جدول ۳ مشاهده می‌کنید، آزمایش‌هایی برای بررسی نتایج این معیارها انجام داده‌ایم و می‌خواهیم نتایج تغییر دقت را، با استفاده از هرکدام از آن‌ها به‌تنهایی نشان دهیم.

اگر برای پیش‌بینی لینک فقط دوستان سازگارتر را در نظر بگیریم، بیشترین تأثیر روی دیتاست همستر صورت می‌گیرد با بهبود ۱,۰۹ درصدی. که میزان بهبود به‌دست‌آمده روی همستر با روش ما برابر با ۲,۴ است. اگر معیار برای پیشنهاد افراد به یکدیگر فقط تعداد دوستان کمتر باشد، بیشترین تأثیر روی شبکه اجتماعی فیس‌بوک است و با بهبود نهایی ما فاصله دارد. به‌علاوه، نسبت به دوستان سازگارتر، روی هیچ‌کدام از دیتاست‌ها نتیجه خوبی نخواهیم گرفت. اگر معیار پیشنهاد دوستی را برابر با دوستان صمیمی‌تر بگیریم، نتایج به‌دست‌آمده نسبت به انتخاب دوستان کمتر بهبود می‌یابد.

در ادامه، ترکیب دوتایی از معیارهای دوستان کمتر و سازگارتر و ترکیب دو معیار دوستان کمتر و صمیمی‌تر و ترکیب دو شاخص دوستان کمتر و صمیمی‌تر را مورد آزمایش قرار دادیم، که نتایج به‌دست‌آمده نشان داد از بین این سه ترکیب، ترکیب شاخص‌های دوستان سازگارتر و دوستان صمیمی‌تر، نتایج بهتری بر روی دقت سه دیتاست به دست می‌دهند. که با نتایج نهایی به‌دست‌آمده از روش نهایی ما فاصله زیادی دارد. بهبود در این ترکیب برای دیتاست فیس‌بوک برابر با ۰,۳۰۹ است که روش نهایی برابر با ۰,۴ است، بهبود دقت بر روی دیتاست همستر برابر با ۳,۴۸ است که بهبود روش نهایی برابر با ۶,۹ و همچنین بر روی دیتاست ایمیل ۲,۰۱ بهبود دقت داشته، که دقت با استفاده از روش نهایی ۲,۴ بهبود داشته است.

مراجع

[1] M. Zuckerberg, "facebook," 2017. [Online]. Available: https://www.facebook.com/zuck?fref=ts. [Accessed: 01-Jan-2017].

[2] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Assoc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.

[3] Y. Dhote, N. Mishra, and S. Sharma, "Survey and analysis of temporal link prediction in online social networks," in *Advances in Computing, Communications and Informatics (ICACCI),* International Conference on, pp. 1178–1183, 2013.

[4] A. K. Menon and C. Elkan, "Link prediction via matrix factorization," in *Joint european conference on machine learning and knowledge discovery in databases*, pp. 437–452, 2011.

[5] G. Kossinets, "Effects of missing data in social networks," *Soc. Networks*, vol. 28, no. 3, pp. 247–268, 2006.

[6] J. Zhang, X. Kong, and S. Y. Philip, "Predicting social links for new users across aligned heterogeneous social networks," in *Data Mining (ICDM)*, IEEE 13th International Conference on, pp. 1289–1294, 2013.

[7] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," *Phys. Rev. E*, vol. 73, no. 2, p. 26120, 2006.

[8] "facebook," 2017. [Online]. Available: www.facebook.com.

[9] "google plus," 2017.

[10] W. Sherchan, S. Nepal, and C. Paris, "A survey of trust in social networks," *ACM Comput. Surv.*, vol. 45, no. 4, p. 47, 2013.

[11] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

[12] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E*, vol. 64, no. 2, p. 25102, 2001.

[13] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Soc. Networks*, vol. 25, no. 3, pp. 211–230, 2003.

[14] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[15] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos, "Fast and accurate link prediction in social networking systems," *J. Syst. Softw.*, vol. 85, no. 9, pp. 2119–2132, 2012.

[16] R. N. Lichtenwalter, J. T. Lussier, and N. V Chawla, "New perspectives and methods in link prediction," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 243–252, 2010.

[17] H. A. Deylami and M. Asadpour, "Link prediction in social networks using hierarchical community detection," in *Information and Knowledge Technology (IKT)*, 7th Conference on, pp. 1–5, 2015.

[18] J. Pei, X. Liu, P. M. Pardalos, W. Fan, S. Yang, and L. Wang, "Application of an effective modified gravitational search algorithm for the coordinated scheduling problem in a two-stage supply chain," *Int. J. Adv. Manuf. Technol.*, vol. 70, no. 1–4, pp. 335–348, 2014.

حدیث بشیری، در سال ۱۳۹۰ مدرک کاردانی خود را از آموزشکده فنی و حرفه‌ای دختران خرم‌آباد، در سال ۱۳۹۲، مدرک کارشناسی ناپیوسته خود را از دانشگاه غیرانتفاعی صفاهان اصفهان و در سال ۱۳۹۶ مدرک ارشد خود را در رشته نرم‌افزار کامپیوتر از دانشگاه شیراز دریافت کرد. زمینه تحقیقاتی موردعلاقه وی شبکه‌های اجتماعی، گراف‌ها می‌باشد.

آدرس پست الکترونیکی ایشان عبارت است از:

hadis.bashiri1@gmail.com

غلامحسین دستغیبی فرد، ارشد و دکترای خود را در رشته علوم کامپیوتر به ترتیب در سال‌های ۱۹۷۹ و ۱۹۹۰ از دانشگاه اکلاهما – نورمن – امریکا اخذ نموده است و در حال حاضر دانشیار بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات، دانشکده مهندسی برق و کامپیوتر دانشگاه شیراز می‌باشد. زمینه موردعلاقه وی عبارت‌اند از: پردازش موازی، محاسبات ابری و شبکه‌های اجتماعی می‌باشد.

آدرس پست الکترونیکی ایشان عبارت است از:

dstghaib@shirazu.ac.ir

[1] Anthology
[2] Structural
[3] Snapshot
[4] Common Neighbor
[5] Jaccard Coefficient
[6] Adamic and Adar
[7] Katz
[8] http://snap.stanford.edu/data/egonets-Facebook.html
[9] http://konect.uni-koblenz.de/networks/petster-hamster
[10] http://snap.stanford.edu/data/email-Enron.html
[11] Training dataset

The CSI Journal on Computing Science and Information Technology

The CSI Journal on
Computing Science and
Information Technology
Vol. 17, No.1, 2019

Link Prediction at Social Network by Assignment Degree of Neighborhood to Edges at No Direction Graphs

Hadis Bashiri, GholamHossein DastghaibiFard

Faculty of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran

Abstract

Nowadays, the existing social networks such as Facebook, Google Plus and Instagram have shown a great influence on human connections. In these social networks, in order to suggest a new connection to the clients, always a link prediction algorithm has been employed in network development. Many studies have been conducted to develop a robust link prediction algorithm; however, the applicability of the proposed algorithms may be crucial due to their low accuracy. Due to the different rate of definable connection in the existing social network, this paper presents a weighted edge and neighborhood degree based algorithm to recognize the sincere connections in social networks. The applicability of the defined algorithm has been examined in three sets of datasets gathered from Facebook, Hamster, and Email. The results respectively show 0.4%, 2.4% and 6.9% improvement accuracy for Facebook, Hamster and Email in comparison with the existing link prediction algorithms.

Keywords: Social Networks, Link Prediction, Adamic Adar, Assign Weight to Edges, Degree of Neighborhood Assignment, Recognizing Close Friends.