ترکیب شباهت ساختاری و غیرساختاری کاربران در شبکه‌های اجتماعی*

اعظم ملک محمد[1]، هادی خسروی فارسانی[2]

[1] دانشکده کامپیوتر، مؤسسه آموزش عالی صفاهان ، اصفهان،
malekmohammad@safahan.ac.ir

[2] دانشکده مهندسی کامپیوتر، دانشگاه شهرکرد ، شهرکرد
khosravi@eng.sku.ac.ir

Archive of SID

چکیده

مسئله تخمین شباهت در بسیاری از علوم و زمینه‌ها کاربرد دارد. برای مثال علوم داده کاوی، وب کاوی، خوشه بندی، موتورهای جستجو و شبکه‌های اجتماعی نیاز به تعریف و پیاده‌سازی شباهت دارند. در شبکه‌های اجتماعی تخمین شباهت بین کاربران یکی از مسائل مطرح در این زمینه است و کاربردهای زیادی دارد. در این تحقیق با هدف افزایش دقت در تخمین شباهت بین کاربران شبکه‌های اجتماعی، الگویی جدید جهت ترکیب الگوریتم‌های ساختاری و غیرساختاری ارائه شده است. در بخش ارزیابی، الگوریتم‌های ساختاری SRank، P-Rank و SimRank توسط الگوی پیشنهادی با الگوریتم غیرساختاری OF[1] ترکیب شده و بر روی بخشی از مجموعه داده شبکه اجتماعی توئیتر پیاده‌سازی می‌شوند. نتایج ارزیابی نشان دهنده‌ی دقت بالای الگوریتم حاصل از ترکیب SRank و OF توسط الگوی پیشنهادی در تخمین شباهت بین کاربران است.

کلمات کلیدی:
شبکه‌های اجتماعی، شباهت ساختاری،شباهت غیرساختاری.

* این مقاله برگرفته از پایان نامه تحصیلات تکمیلی در موسسه آموزش عالی صفاهان است.

www.SID.ir

۱- مقدمه

امروزه شبکه‌های اجتماعی جایگاه و نقش مهمی در زندگی روزمره افراد پیدا کرده است. اولین شبکه اجتماعی در سال ۱۹۹۷ در ایالات متحده امریکا ایجاد گردید که کاربران آن قادر به ساختن پروفایل شخصی و برقراری ارتباط با یکدیگر بودند [۱]. پس از آن شبکه‌های اجتماعی دیگری با موضوعات متفاوتی در سراسر جهان ایجاد شدند. شبکه‌های اجتماعی امکان اشتراک اطلاعات، ایده‌ها، علایق، فعالیت‌ها و وقایع زندگی را برای کاربران خود فراهم می‌کنند. برخی از شبکه‌های اجتماعی عبارتند از Facebook، Twitter ، LinkedIn، Google Plus+ و غیره است.

ساختار شبکه‌های اجتماعی به صورت گراف است که از گره و پیوندهای تشکیل شده است. گره‌ها نشان دهنده افراد، سازمان و یا سایر موجودیت‌هایی[۲] هستند که تعاملاتی[۳] را آغاز و یا دریافت می‌کنند و پیوندها نشان دهنده‌ی تعاملات، همکاری[۴] و یا نفوذ[۵] گره‌ها بر یکدیگر هستند. نمونه‌ای از این پیوندها عبارتند از قیمت‌ها، ایده‌ها، تبادلات مالی، رابطه دوستی و خویشاوندی، تجارت، تبلیغات، پیوندهای وب، سرایت بیماری و یا مسیرهای هواپیمایی و غیره است. گراف شبکه‌های اجتماعی ساختار کاملاً پویا دارند و با گذشت زمان با اضافه و حذف شدن گره‌ها و پیوندها تغییر می‌یابد. شناسایی مکانیزمی که این تغییرات را پیش‌بینی و استنتاج نماید، یکی از مسائل مطرح در این زمینه است.

تخمین شباهت بین کاربران و پیش‌بینی پیوندهایی که ممکن است در آینده نزدیک بین آن‌ها ایجاد شوند یکی از موضوعات مطرح در زمینه شبکه‌های اجتماعی است. شباهت در شبکه‌های اجتماعی بر اساس معیارهای از پیش تعیین شده تعریف و محاسبه می‌گردد. در این مقاله، شباهت به این صورت تعریف می‌شود .«کاربرانی که از نظر معیار از پیش تعیین شده مشابه هستند را یک رتبه یکسان اختصاص داده می‌شود و یا آن‌ها را درون یک خوشه قرار داده می‌شود». این معیار می تواند یکی از خصوصیات پروفایل کاربران هم‌چون سن، تحصیلات، محل تحصیل، محل زندگی و غیره است. موقعیت شبکه‌ای کاربران در گراف شبکه اجتماعی نیز نوعی معیار برای تعیین شباهت کاربران است. در سیستم‌های پیشنهاد دهنده[۶] مجموعه ای از پرس-وجوهای از پیش تعیین شده توسط کاربران پاسخ داده می شود؛ پاسخ کاربران برداری از اولویت‌ها برای هر کاربر ایجاد می‌کند و برای معیار شباهت کاربران از این بردار اولویت استفاده می‌گردد.

در شبکه‌های اجتماعی، پیدا کردن میزان شباهت افراد کاربرد بسیاری دارد. برای مثال پیدا کردن مسیر برای گسترش اخبار، عقاید، نظریه‌ها و دیدگاه‌های سیاسی با یافتن کاربران مشابه امکان‌پذیر است. یافتن انجمن‌ها و صفحات مورد علاقه کاربر، دوستان هم فکر و عقیده نیز نیاز به یافتن کاربران مشابه دارد. در سیستم‌های پیشنهاد دهنده برای یافتن آیتم‌های مورد علاقه کاربر نیاز به یافتن شباهت بین علایق کاربر و آیتم‌های موجود است. امروزه از شبکه‌های اجتماعی برای تبلیغ محصولات و بازاریابی استفاده می‌شود. تبلیغ محصولات نیز نیاز به یافتن کاربرانی است که علایق و سلیقه‌هایشان مشابه با محصول مورد نظر است [۲]. تشخیص رهبران[۷] و پیروان[۸] در شبکه‌های اجتماعی یکی دیگر از کاربردهای تخمین شباهت در شبکه‌های اجتماعی است [۳].

۲- روش‌های تخمین شباهت کاربران شبکه‌های اجتماعی

کاربران شبکه‌های اجتماعی به صورت کلی دارای دو خصوصیت ساختاری و غیرساختاری هستند. خصوصیات ساختاری و یا شبکه‌ای به معنای موقعیت قرارگیری کاربر در گراف شبکه است. از خصوصیات ساختاری می‌توان به درجه رأس[۹]، تعداد مثلثات[۱۰] ، ضریب خوشه بندی[۱۱]، مرکزیت بردار[۱۲] و طول متوسط کوتاه ترین مسیر[۱۳] اشاره نمود [۳]. خصوصیات ساختاری گراف شبکه اجتماعی به صورت ماتریس مجاورت نمایش داده می‌شود.

خصوصیات غیرساختاری و یا پروفایلی به معنای اطلاعات دسته‌بندی شده[۱۴] موجود در پروفایل کاربر است. پروفایل کاربر در شبکه‌های اجتماعی شامل اولویت‌ها، سلایق، رویدادهای زندگی و غیره است. یکی از ویژگی‌های بارز اطلاعات دسته‌بندی شده مرتب نبودن مقادیر فیلدهای آن است. از دیگر خصوصیت بارز پروفایل کاربران شبکه‌های اجتماعی، کامل نبودن آن است زیرا تکمیل اطلاعات پروفایل کاربران شبکه‌های اجتماعی در اختیار کاربر است. بسیاری از کاربران اطلاعات پروفایل را کامل وارد نمی‌کنند و یا اشتباه وارد می‌کنند.

الگوریتم‌های تخمین شباهت، با توجه به خصوصیات ساختاری و یا غیرساختاری، سعی در پیدا کردن میزان شباهت بین کاربران می‌نماید و از این دیدگاه به سه دسته اصلی تقسیم می‌گردند:

- الگوریتم‌هایی که تنها با در نظر گرفتن خصوصیات ساختاری کاربران سعی در تخمین میزان شباهت بین کاربران می نماید.
- الگوریتم‌هایی که تنها با در نظر گرفتن خصوصیات غیرساختاری کاربران و اطلاعات دسته‌بندی شده موجود در پروفایل کاربران عمل تخمین شباهت را انجام می دهند.
- الگوریتم‌هایی که با در نظر گرفتن خصوصیات ساختاری و غیرساختاری سعی در تخمین میزان شباهت کاربران نموده است.

با گسترش شبکه‌های اجتماعی و اهمیت یافتن مسئله شباهت در آن‌ها، ابتدا الگوریتم های دسته اول پا به عرصه نهادند و تعداد زیادی از این الگوریتم‌ها معرفی گردید. اما تنها استفاده از این دسته از الگوریتم‌ها برای تعیین شباهت کاربران کافی نیست. زیرا بسیاری از کاربران از نظر شبکه‌ای هیچ شباهتی ندارند، لیکن از نظر خصوصیات معنایی شباهت دارند. بنابراین الگوریتم‌های دسته دوم با ایده گرفتن از مسئله شباهت در داده‌های پیوسته و اعمال آن بر داده‌های دسته‌بندی شده موجود در پروفایل کاربران ایجاد گردید. بر اساس اصل هوموفیلی، در شبکه‌های اجتماعی بیشتر افرادی که خصوصیات پروفایلی مشابه دارند، از نظر ساختاری نیز به هم متصل‌تر هستند [۴]. به همین دلیل در سال‌های اخیر سعی بر ترکیب این دو نوع الگوریتم برای افزایش میزان دقت تخمین شباهت شده است و چالش اصلی در این نوع الگوریتم‌ها ترکیب موثر این دو نوع الگوریتم به منظور بهبود نتایج است. زیرا شباهت ساختاری و غیرساختاری دو مسئله کاملاً مستقلی هستند و هر کدام سهم متمایزی در تعیین شباهت دارند. در این مقاله بررسی شباهت کاربران شبکه‌های اجتماعی با استفاده از ترکیب موثر شباهت ساختاری و غیرساختاری کاربران انجام خواهد گرفت.

Archive of SID

www.SID.ir

۳- کارهای مرتبط

در زمینه تخمین شباهت بین کاربران شبکه‌های اجتماعی با استفاده از خصوصیات ساختاری، تحقیقات زیادی انجام شده است [۲و۵و۶و۷و۸ و۹]. هم‌چنین در زمینه شباهت غیرساختاری کاربران نیز تحقیقاتی انجام شده است. [۱۰و۱۱] اما در نظر گرفتن هم خصوصیات ساختاری و هم خصوصیات غیرساختاری کاربران در تخمین شباهت بین آنان و پیش‌بینی پیوندهای جدید در بین آن‌ها مسئله‌ای جدید است و در این زمینه تحقیقات کمی انجام شده است.

در [۱۲] روشی برای تعیین شباهت کاربران با استفاده از خصوصیات ساختاری و غیر ساختاری آن‌ها ارائه شده است. در قسمت شباهت غیرساختاری از روش OF استفاده شده است. مزیت مقاله [۱۲] نسبت به سایر مقاله‌ها این است که ساختار پروفایل کاربران را به صورت مجموعه‌ای از آیتم‌ها در نظر می‌گیرد و آیتم‌های چندگانه همانند تحصیلات را نیز پشتیبانی می‌نماید. در قسمت شباهت ساختاری از روش NS استفاده شده است. در مقاله [۱۲] برای استنباط مقادیر پروفایل وارد نشده از روش رای‌گیری اکثریت[۱۵] استفاده شده است. در پایان، نتایج بر روی گراف شبکه‌های اجتماعی جهت‌دار Youtube و بدون جهت Facebook ارزیابی و مقایسه شده است. در مقاله [۱۲] یک الگوریتم واحد جهت اندازه‌گیری شباهت ساختاری و غیرساختاری ارائه نشده است و تنها هم‌بستگی بین شباهت ساختاری و غیرساختاری را بیان می‌کند.

در [۱۳] الگوریتم Random Walk with restart معرفی شده است. در این مقاله با استفاده از خصوصیات گره‌ها و پیوندهای موجود شروع به یادگیری یک دسته‌بندی می‌کند تا گره‌های موجودی که پیوندی با آن‌ها ندارد را در گروه گره‌های مثبت و یا منفی قرار دهد. گره‌های مثبت گره‌هایی هستند که ممکن است در آینده، پیوندی بین گره مبدأ و یکی از گره‌های مثبت برقرار گردد. ابتدا یک الگوریتم یادگیرنده با توجه به خصوصیات دو گره به پیوند بین آن‌ها یک رتبه اختصاص می‌دهد. سپس الگوریتم Random Walk با توجه به رتبه‌های ایجاد شده شروع به حرکت در گراف می‌کند. به بیان دیگر الگوریتم Random Walk توسط رتبه پیوندها تحت تاثیر قرار می‌گیرد و بیشتر تمایل به حرکت در پیوندهایی دارد که رتبه آن‌ها بیشتر است. در این روش، خصوصیات گره‌ها و پیوندها بایستی از قبل مشخص باشد بنابراین به‌کارگیری این روش در شبکه‌های واقعی امکان‌پذیر نیست.

یکی از کاربردهای مهم تخمین شباهت خوشه‌بندی کاربران است. در [۱۴] الگوریتم SA-cluster معرفی شده است که پس از تخمین شباهت با استفاده از خصوصیات ساختاری و غیرساختاری کاربران، آن‌ها را درون خوشه مناسب قرار می‌دهد. کاربران درون یک خوشه رتبه یکسانی دارند و نمی‌توان رتبه مجزا به کاربران درون یک خوشه اختصاص داد. در مقاله [۱۴] اطلاعات پروفایل کاربران را به صورت اطلاعات دسته بندی شده در نظر نمی‌گیرد و تنها یک خصوصیت برای پروفایل کاربران در نظر گرفته است.

٤- شباهت ساختاری

در بخش شباهت ساختاری الگوی پیشنهادی این تحقیق از الگوریتم‌های SRank، SimRank و P-Rank استفاده شده است. الگوریتم‌های SimRank و P-Rank بر پایه‌ی همسایگان گره عمل می‌کنند و الگوریتم SRank بر پایه‌ی مسیرهای کوتاه بین دو گره عمل می‌کند. در ادامه الگوریتم‌های فوق شرح داده خواهند شد.

٤-۱- SRank [۱۵]

ایده اصلی الگوریتم SRank به این صورت بیان می‌شود: «دو گره در یک گراف جهت‌دار در صورتی با یکدیگر مشابه‌اند که از طریق چندین مسیر با طول کوتاه به یکدیگر متصل باشند». به طور دقیق‌تر شباهت بین دو گره a و b در یک گراف توسط دو شرط زیر تحت تاثیر قرار می‌گیرد:

- تعداد مسیرهای کوتاه از a به b
- طول مسیرهای کوتاه از a به b

برای محاسبه شباهت بر اساس SRank، ابتدا بایستی مقدار دسترسی تعریف و محاسبه شود. P_p ماتریس احتمال انتقال با ابعاد N*N در گراف G است. طول این ماتریس p است. مقدار دسترسی از گره a به گره b به صورت رابطه (۱) تعریف می‌شود.

$$H(a,b) = w_1 * P^1_{a,b} + \ldots + w_p * P^p_{a,b} + \ldots + w_{n-2} * P^{n-2}_{a,b} \qquad (۱)$$

w_i وزن تعریف شده برای تمام مسیرها با طول i است و $P^p_{a,b}$ احتمال انتقال از گره a به گره b توسط مسیرهایی به طول p است. $P^p_{a,b}$ برابر با تعداد مسیرهایی با طول p از گره a به b تقسیم بر تعداد مسیرهایی با طول p از گره a به تمام گره‌های موجود در گراف است.

$$P^p_{a,b} = \frac{k_p(a,b)}{\sum_{\forall x \in G-\{a\}} k_p(a,x)} \qquad (۲)$$

به دست آوردن مقدار دسترسی با در نظر گرفتن تمام مسیرها با طول‌های متفاوت در یک گراف، زمان زیادی لازم دارد. بنابراین $H_s(a,b)$ به صورت زیر تعریف می‌شود و جایگزین $H(a,b)$ خواهد شد. در مقاله [۱۵] مسیرهایی با طول‌های متفاوت بررسی شده است و به این نتیجه رسیده است که حداکثر طول مسیر سه، بیشترین میزان دقت در تخمین شباهت بین گره‌های موجود در گراف دارد.

$$H_s(a,b) = w_1 * P^1_{a,b} + \ldots + w_s * P^s_{a,b} \qquad 1 \le s \le 3 \qquad (۳)$$

برای بدست آوردن نتایج صحیح بایستی وزنی که به مسیرهایی کوتاه اختصاص داده می‌شود بالاتر از وزنی باشد که به مسیرهای طولانی اختصاص داده می‌شود. وزن اختصاص داده شده به مسیرها از طریق فرمول (۴) بدست می‌آید.

$$w_p = 2^{s-p} \qquad (۴)$$

یک روش ساده برای تخمین شباهت بین گره a و گره b نرمال سازی $H_s(a, b)$ با توجه به بزرگترین و کوچکترین H در کل گراف است. بنابراین از فرمول (۵) به منظور تخمین شباهت بین گره‌ها استفاده می‌شود.

$$SRank_s(a,b) = \frac{H_s(a,b) - H_{Min}}{H_{Max} - H_{Min}} \qquad (۵)$$

Archive of SID

www.SID.ir

۴-۲- SimRank [۵]

ایده اصلی الگوریتم SimRank به این صورت بیان می‌شود: «دو گره در یک گراف جهت‌دار در صورتی با یکدیگر مشابه‌اند که به همسایگان مشابه بیشتری متصل و مرتبط باشند». الگوریتم SimRank تنها همسایگان ورودی دو گره را برای تخمین شباهت بین آن‌ها در نظر می‌گیرد. شباهت بین گره a و b را با s(a,b) نشان داده شده است و با استفاده از رابطه بازگشتی زیر قابل محاسبه است. اگر a=b باشد شباهت بین این گره ۱ است و در غیر اینصورت:

$$s(a,b)=\frac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)||I(b)|}\sum_{j=1}^{}s(I_i(a),I_j(b)) \qquad (۶)$$

C سطح اطمینان[۱۶] الگوریتم و مقداری ثابت بین صفر و یک است. در مقاله [۵] ثابت شده است که بهترین مقدار برای C، ۰٫۸ است. در آزمایشات این مقاله نیز به منظور افزایش حداکثری دقت الگوریتم SimRank مقدار ۰٫۸ برای سطح اطمینان این الگوریتم در نظر گرفته می‌شود. I(a) نشان دهنده همسایگان ورودی گره a و I(b) نشان دهنده همسایگان ورودی گره b است. اگر گره a یا b هیچ همسایه ورودی نداشته باشد، مقدار صفر به عنوان شباهت بین دو گره فوق در نظر گرفته خواهد شد.

۴-۳- P-Rank [۹]

ایده اصلی الگوریتم P-Rank به این صورت بیان می‌شود: «دو گره در یک گراف جهت‌دار در صورتی با یکدیگر مشابه‌اند که به موجودیت‌ها و گره‌های مشابه بیشتری متصل و مرتبط باشند». الگوریتم P-Rank از همسایگان ورودی و خروجی هر گره برای تخمین شباهت بین استفاده می‌کند. شباهت بین گره a و b را با s(a,b) نشان داده شده است و با استفاده از رابطه بازگشتی زیر قابل محاسبه است. اگر a=b باشد شباهت بین این گره ۱ است و در غیر اینصورت:

$$\begin{aligned}s(a,b)=&\lambda*\frac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)||I(b)|}\sum_{j=1}^{}s(I_i(a),I_j(b))\\&+(1-\lambda)*\frac{C}{|O(a)||O(b)|}\sum_{i=1}^{|O(a)||O(b)|}\sum_{j=1}^{}s(O_i(a),O_j(b))\end{aligned} \qquad (۷)$$

در الگوریتم P-Rank اگر تعداد همسایگان ورودی گره a یا گره b، صفر باشد از همسایگان خروجی استفاده می‌کند. چنان‌چه تعداد همسایگان خروجی گره a یا گره b، صفر باشد از همسایگان ورودی استفاده می‌کند. اگر تعداد همسایگان ورودی و خروجی گره a یا گره b صفر باشد، میزان شباهت بین این دو گره صفر خواهد بود. مقدار ثابت C در الگوریتم P-Rank نیز نشان‌دهنده‌ی سطح اطمینان الگوریتم بوده و بهترین میزان آن مقدار ۰٫۸ است. مقدار λ وزن نسبی است که به یال‌های ورودی و خروجی اختصاص داده می‌شود. در حالت نرمال وزن اختصاصی به یال‌های ورودی و خروجی یکسان و برابر با ۰٫۵ است. در بخش آزمایشات این مقاله نیز مقدار ۰٫۸ به عنوان سطح اطمینان الگوریتم و مقدار ۰٫۵ به عنوان وزن اختصاصی به یال‌های ورودی و خروجی در نظر گرفته می‌شود.

۵- شباهت غیرساختاری

در بخش شباهت غیرساختاری الگوی پیشنهادی این تحقیق از الگوریتم OF استفاده شده است. این الگوریتم بر پایه‌ی فراوانی رخداد است. برای محاسبه شباهت غیرساختاری دو کاربر در شبکه اجتماعی یکی از فیلدهای پروفایل کاربر را در نظر گرفته و شباهت دو کاربر بر اساس مقدار فیلد مورد نظر محاسبه می‌شود. در محاسبه شباهت خصوصیات X و Y، اگر مقدار این دو خصوصیت مانند یکدیگر باشند، مقدار شباهت متناظر یک خواهد بود و در غیراینصورت مقدار شباهت این دو برابر با فرمول (۸) است.

$$S(X,Y)=\frac{1}{1+\log(\frac{N}{f(X)})*\log(\frac{N}{f(Y)})} \qquad (۸)$$

مقدار f(X) و f(Y) نشان دهنده‌ی تعداد فراوانی و تکرار مقدار X در پروفایل کاربران کل گراف است

۶- الگوی پیشنهادی در ترکیب الگوریتم شباهت ساختاری و غیرساختاری

الگوی پیشنهادی جدید این تحقیق جهت ترکیب الگوریتم شباهت ساختاری و غیرساختاری در شکل ۱ نشان داده شده است. ترکیب الگوریتم‌های ساختاری و غیرساختاری با استفاده از الگوی فوق قبلا انجام نشده است. این الگو برای هر الگوریتم شباهت ساختاری و غیرساختاری قابل پیاده‌سازی است. با توجه به شکل ۱، ابتدا بیشترین میزان شباهتی که کاربر x با سایر گره‌های موجود درگراف که جزء دوستان کاربر x نیستند، تعیین می‌شود. برای مثال این مقدار n است؛ اگر مقدار n از صفر بزرگتر باشد به این معناست که حتما گره x حداقل با یکی از گره های موجود در گراف اتصال دارد. در صورتی که گره x با هیچ یک از گره‌های موجود در گراف اتصال نداشته باشد مقدار n برابر با صفر خواهد شد. در مواردی هم‌چون کاربران تازه وارد و یا کاربری که کمترین اتصال با سایرکاربران دارد، معمولاً مقدار n صفر است.

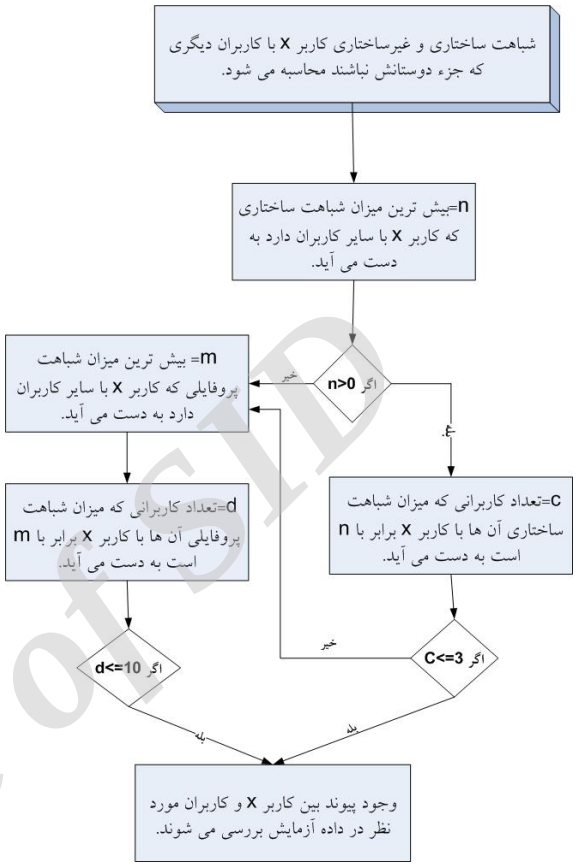
اگر مقدار n بزرگتر از صفر باشد تعداد کاربرانی که جزء دوستان گره x نیستند و میزان شباهت ساختاری آن‌ها با گره x برابر با بیشترین مقدار یا همان n باشند مشخص می‌شود. اگر تعداد این گره‌ها بیش از اندازه‌ی معقول باشند در نظر گرفتن و بررسی تمام آن‌ها در داده آزمایش کار عاقلانه‌ای نیست زیرا برای مثال کاربری از نظر خصوصیات ساختاری با ۱۰۰ نفر بیشترین شباهت ساختاری را دارد. پیشنهاد دادن این ۱۰۰ نفر به کاربر فوق درست نیست. بنابراین در تعداد کاربرانی که ممکن است با کاربر x بیشترین شباهت ساختاری را داشته باشند محدودیت قائل می‌شود و اگر تعداد آن‌ها از حد مجاز بیشتر باشد از پیشنهاد دادن آن‌ها چشم‌پوشی می‌کند و به سراغ شباهت غیرساختاری کاربر x با سایر کاربران می‌رود. در صورتی که تعداد کاربرانی که از نظر شباهت ساختاری بیشترین مقدار را دارند کمتر از حد مجاز باشد آن‌ها را به کاربر مورد نظر پیشنهاد می دهد و برای صحت پیشنهادات خود به سراغ داده آزمایش می‌رود.

اگر بین کاربر x و یکی از کاربرانی که بیشترین شباهت ساختاری را با کاربر x دارند، رابطه‌ای در داده آزمایش وجود داشته باشد به این معناست که الگوریتم با موفقیت برای کاربر x پیوندهای جدید پیش‌بینی کرده است. در طرح پیشنهادی این تحقیق حد استفاده شده در تعداد کاربرانی که بیشترین

Archive of SID

www.SID.ir

شباهت ساختاری با کاربر x دارند مقدار ۳ در نظر گرفته شده است. در صورتی تعداد این کاربران کمتر از ۳ باشد تمام آن‌ها به کاربر پیشنهاد داده می‌شوند و در صورتی که تعداد آن‌ها بیشتر از ۳ باشد از شباهت پرفایلی کاربر در پیش‌بینی پیوندهای جدید برای آن کاربر استفاده می‌کند.



شکل (۱) : الگوی پیشنهادی در ترکیب شباهت ساختاری و غیرساختاری

در صورتی که مقدار n برابر با صفر باشد از شباهت ساختاری کاربر در پیشنهاد پیوندهای جدید به وی نمی‌تواند استفاده کند زیرا گره x در چنین حالتی کمترین اتصال را با سایر گره‌ها دارد و یا هیچ اتصالی ندارد و شباهت غیرساختاری کاربر در پیش‌بینی پیوندهای جدید برای وی موثرتر است. بنابراین در این‌حالت بیشترین میزان شباهت پروفایلی که کاربر x با سایر گره‌های موجود در گراف دارد تعیین می‌شود. برای مثال این مقدار m است اگر مقدار m از صفر بزرگتر باشد به این معناست که حتما گره x حداقل با یکی از گره های موجود در گراف شباهت معنایی دارد. سپس تعداد کاربرانی که جزء دوستان گره x نیستند و میزان شباهت غیرساختاری آن‌ها با گره x برابر با بیشترین مقدار یا همان m باشند مشخص می‌شود. در این قسمت نیز یک مقدار حدی در تعداد این گره‌ها در نظر گرفته می‌شود که در این تحقیق مقدار ۱۰ برای آن مشخص شده است. اگر تعداد کاربرانی که شباهت معنایی آن‌ها با کاربر x برابر با m است از حد انتخاب شده یا همان عدد ۱۰ کمتر باشد؛ تمام آن‌ها به کاربر پیشنهاد داده می‌شود. در صورتی که تعداد این کاربران از حد مجاز تعیین شده بیشتر باشد به کاربر پیشنهاد داده نمی‌شود.

انتخاب اعداد ۳ و ۱۰ از اصل خاصی را پیروی نمی‌کند اما با توجه به پراکندگی داده‌های موجود در مجموعه داده و تک‌فیلدی بودن مقادیر پروفایل کاربران و پس از آزمایشات این نتیجه استنباط شد که در مجموعه داده استفاده شده در این تحقیق، میزان حد استفاده شده برای قسمت شباهت غیرساختاری بایستی بیشتر از حد استفاده شده برای قسمت شباهت ساختاری باشد. در این تحقیق هم‌چنین اعداد متفاوتی برای میزان حد شباهت ساختاری و غیرساختاری مشروط به بزرگتر بودن حد شباهت پروفایلی نسبت به حد شباهت ساختاری، در نظرگرفته شد و پس از بررسی نتایج آن مشخص شد که با انتخاب هر عددی نتایج الگوریتم‌ها تقریباً ثبات خاصی دارند و ترتیب دقت و بازیابی الگوریتم‌ها در تمام موارد یکسان است. حدود انتخاب شده ممکن است در مجموعه داده‌های دیگر متفاوت باشد. شبه کد مربوط به الگوی ترکیب شباهت ساختاری و غیرساختاری در انتهای مقاله در شکل ۸ نشان داده شده است.

۷- ارزیابی

در ارزیابی این تحقیق از مجموعه داده شبکه اجتماعی توئیتر استفاده شده است [۱۶]. این مجموعه داده شامل ۲۱ میلیون کاربر و حدود ۲۰۰ میلیون روابط دوستی بین کاربران است. به طور میانگین، هر کاربری ۱۰ دوست دارد. در این مجموعه داده سه فایل وجود دارد؛ فایل اول شامل روابط دوستی بین کاربران است و محتوای آن به صورت زیر است: [USER ID1]\tab [USER ID 2]

این جمله به این معناست که کاربر۱ رابطه دوستی با کاربر۲ دارد. فایل دوم شامل اطلاعات پروفایل کاربران است و فایل سوم شامل tweet های مربوط به کاربران است. ابتدا تعدادی از کاربران به صورت تصادفی انتخاب می‌شوند و یال‌های بین کاربران فوق از مجموعه داده استخراج شده و تحت یک گراف کوچک ذخیره می‌شوند. این نمونه گراف کوچک نیز به صورت تصادفی به دو قسمت داده آزمایش[۱۷] و داده آموزش[۱۸] تقسیم می‌شود. داده آموزش به منظور آزمایش روش و داده آزمایش به منظور ارزیابی روش مورد استفاده قرار می‌گیرد.

در گراف ایجاد شده، از هر گره یک یال به صورت تصادفی حذف می‌گردد. یال‌های حذف شده در قسمت داده آزمایش قرار می‌گیرند و یال‌های باقی مانده در داده آموزش قرار می‌گیرند. در این مقاله، دو نمونه گراف زیر به روش ذکر شده ایجاد شده‌اند:

- گراف اول شامل ۳۰۱ کاربر و ۹۳۷ یال بین کاربران است.
- گراف دوم شامل ۱۷۶۶ کاربر و ۶۶۵۲ یال بین کاربران است.

الگوریتم‌ها در قسمت داده آموزش اجرا می‌شوند و سعی می‌کنند برای هر کاربر یک دوست جدید با بیشترین شباهت ساختاری پیدا کرده و به وی پیشنهاد دهند. اگر بین کاربر مورد نظر و دوست پیشنهاد داده شده به وی، یک رابطه از قبل وجود داشته است و این رابطه در داده آزمایش وجود دارد، الگوریتم فوق قادر به پیش‌بینی درست یال بوده است؛ در غیراینصورت نتیجه پیش‌بینی آن نادرست است. این فرایند برای تمام گره‌های موجود در گراف اتفاق می‌افتد.

۷-۱- معیار ارزیابی

در این مقاله، برای ارزیابی کارایی الگوریتم‌ها، از معیارهای زیر استفاده می‌شود:

Archive of SID

www.SID.ir

- Precision: این معیار، مقدار پیش‌گویی صحیح الگوریتم را نشان می‌دهد و برابر است با کسری از موارد بازیابی شده که صحیح هستند. در زمینه شبکه‌های اجتماعی، به معنای درصد از دوستان پیشنهادی به تمام کاربرانی که صحیح باشند و در داده آزمایش وجود داشته باشند [۱۷].

$$\text{Precision} = 100 * \frac{\text{number of correct found}}{\text{number of nodes}} \quad (۹)$$

- Recall: این معیار حساسیت نتایج الگوریتم را نشان می‌دهد و برابر است با کسری از موارد مرتبط که بازیابی می‌شوند. در زمینه شبکه‌های اجتماعی، به معنای درصد از دوستان حذف شده‌ای که در داده آزمایش قرار دارند و به کاربران به عنوان دوستان جدید پیشنهاد داده شده‌اند. بالا بودن این معیار به معنای این است که الگوریتم بیشتر نتایج مرتبط را برمی‌گرداند [۱۷].

$$\text{Recall} = 100 * \frac{\text{number of correct found}}{\text{number of edges in test data}} \quad (۱۰)$$

- F-Measure: از این معیار برای تست دقت الگوریتم استفاده می‌شود. F-Measure از میانگین هارمونیک یا میانگین وزن‌دار Precision و Recall به‌دست می‌آید [۱۷].

$$\text{F-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{recall}} \quad (۱۱)$$

۷-۲- نتایج ارزیابی

در این قسمت سه روش شباهت ساختاری SRank، SimRank و P-Rank با استفاده از الگوی پیشنهادی با الگوریتم شباهت غیرساختاری OF ترکیب می‌شوند و به روی دو گراف معرفی‌شده پیاده‌سازی می‌شوند. تمام آزمایش‌ها بر روی یک سیستم کامپیوتر با پردازنده ۲٫۵ گیگاهرتز و Core™5 با ۴ گیگابایت حافظه اصلی، تحت سیستم عامل ویندوز ۷ و به زبان جاوا پیاده‌سازی و اجرا شده‌اند.

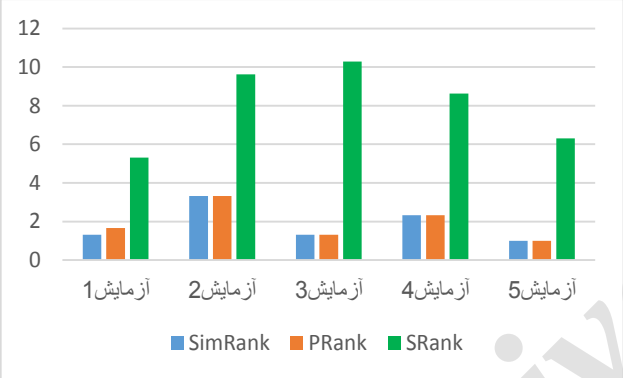
۷-۲-۱- آزمایش شماره۱

نتایج حاصل از ۵ آزمایش بر روی گراف اول در جدول ۱ نشان داده شده است.

جدول (۱) : نتایج پیش‌بینی الگوریتم‌ها بر روی گراف اول

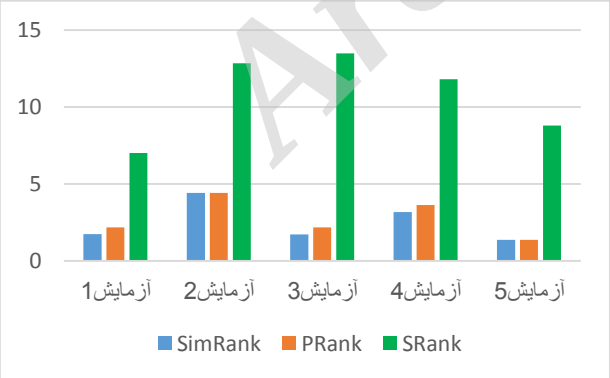
آزمایش	تعداد یال‌ها در داده آزمایش	تعداد یال‌ها در داده آموزش	تعداد پیش‌بینی‌های صحیح		
			SimRank & OF	P-Rank & OF	SRank & OF
۱	۲۲۸	۷۰۹	۴	۵	۱۶
۲	۲۲۶	۷۱۱	۱۰	۱۰	۲۹
۳	۲۳۰	۷۰۷	۴	۵	۳۱
۴	۲۲۰	۷۱۷	۷	۸	۲۶
۵	۲۱۶	۷۲۱	۳	۳	۱۹

در هر دوره آزمایش سه الگوریتم برروی گراف آموزش مربوطه اجرا می‌شوند. مشخصات گراف آموزش و آزمایش هر آزمایش در ستون دوم و سوم آزمایش مربوطه مشخص شده است. برای مثال در آزمایش شماره۱ بعد از تقسیم‌بندی تصادفی پیوندهای موجود در گراف اول، ۲۲۸ پیوند در داده آزمایش و ۷۰۹ پیوند در داده آزمایش قرار گرفته است. الگوریتم‌ها برای تمام کاربران موجود در گراف تخمین شباهت انجام می‌دهند و یک پیوند جدید به آنان پیشنهاد می‌دهند. به منظور تعیین صحت پیوند پیشنهادی توسط الگوریتم، از داده آزمایش استفاده می‌شود. در صورتی‌که پیوند پیشنهادی در داده آزمایش وجود داشته باشد پیشنهاد الگوریتم صحیح است و در غیر اینصورت پیشنهاد الگوریتم نادرست است. در ستون ۴ تعداد پیشنهادات صحیح الگوریتم حاصل از ترکیب SimRank با OF توسط الگوی پیشنهادی نشان داده شده است و به ترتیب در ستون‌های ۵ و ۶ تعداد پیشنهادات صحیح الگوریتم ترکیبی P-Rank با OF و SRank با OF نشان داده شده است. با توجه به تعداد پیشنهادات صحیح هرالگوریتم نتایج دقت و بازیابی هر الگوریتم به‌دست می‌آید. نتایج Precision مربوط به آزمایش‌های انجام شده بر روی گراف اول در شکل ۲ نشان داده شده است.



شکل (۲) : نتایج Precision حاصل از آزمایش شماره ۱

نتایج Recall مربوط به آزمایش‌های انجام شده بر روی گراف اول در شکل ۳ نشان داده شده است.



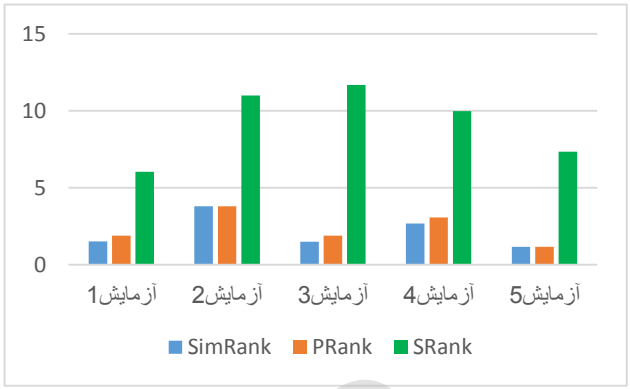
شکل (۳) : نتایج Recall حاصل از آزمایش شماره ۱

نتایج F-Measure مربوط به آزمایش‌های انجام شده بر روی گراف اول در شکل ۴ نشان داده شده است. با توجه به نتایج F-Measure مشخص است که الگوریتم SRank که بر مبنای مسیرهای کوتاه بین دو گره

Archive of SID

www.SID.ir

است؛ در مقایسه با سایر روش‌ها که بر مبنای همسایگان گره هستند، از دقت بالاتری برخوردار است.



شکل (٤) : نتایج F-Measure حاصل از آزمایش شماره ۱

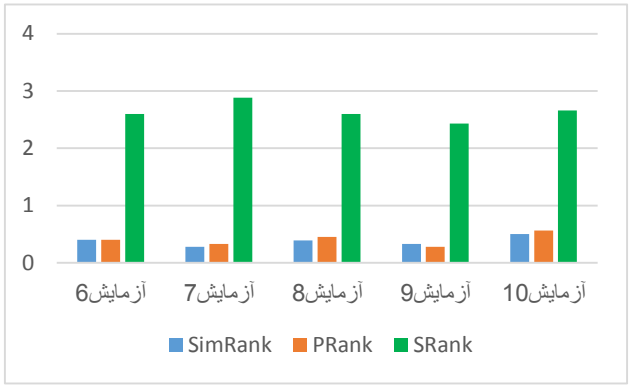
۷-۲-۲- آزمایش شماره۲

نتایج حاصل از ۵ آزمایش بعدی بر روی گراف دوم در جدول ۲ نشان داده شده است.

جدول (۲) : نتایج پیش‌بینی الگوریتم‌ها بر روی گراف دوم

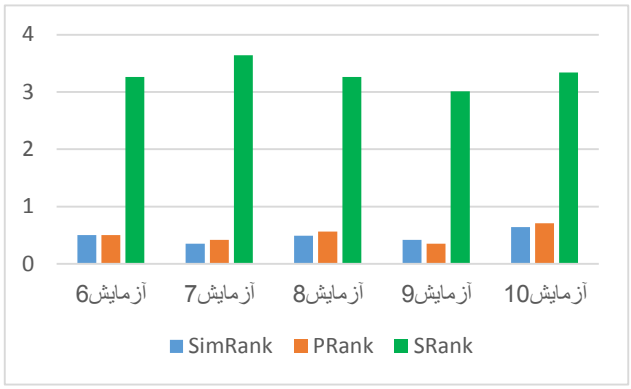
آزمایش	تعداد یال‌ها در داده آزمایش	تعداد یال‌ها در داده آموزش	تعداد پیش‌بینی‌های صحیح		
			SimRank & OF	P-Rank & OF	SRank & OF
۶	۱۴۰۸	۵۲۴۴	۷	۷	۴۶
۷	۱۳۹۸	۵۲۵۴	۵	۶	۵۱
۸	۱۴۱۰	۵۲۴۲	۷	۸	۴۶
۹	۱۴۲۴	۵۲۲۸	۶	۵	۴۳
۱۰	۱۴۰۷	۵۲۴۸	۹	۱۰	۴۷

با توجه به جدول ۲، نتایج Precision مربوط به آزمایش‌های انجام شده بر روی گراف دوم در شکل ۵ نشان داده شده است.



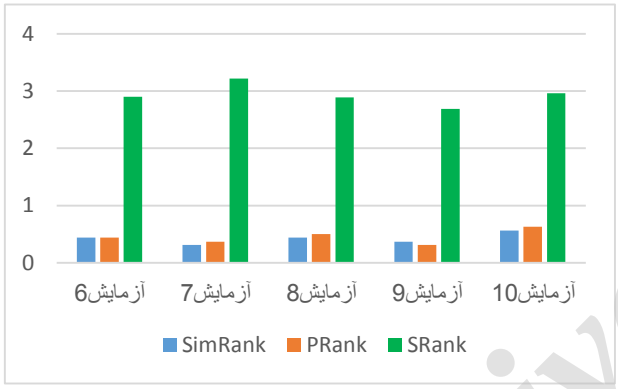
شکل (۵) : نتایج Precision حاصل از آزمایش شماره ۲

نتایج Recall مربوط به آزمایش‌های انجام شده بر روی گراف دوم در شکل ۶ نشان داده شده است.



شکل (٦) : نتایج Recall حاصل از آزمایش شماره ۲

نتایج F-Measure مربوط به آزمایش‌های انجام شده بر روی گراف دوم در شکل ۷ نشان داده شده است.



شکل (۷) : نتایج F-Measure حاصل از آزمایش شماره۲

در نتایج آزمایش شماره ۲ نیز مشخص است که الگوریتم SRank که بر مبنای مسیرهای کوتاه بین دو گره است؛ در مقایسه با سایر روش‌ها که بر مبنای همسایگان گره هستند، از دقت بالاتری برخوردار است.

۸- نتیجه

در این مقاله الگویی جدید جهت ترکیب الگوریتم‌های شباهت ساختاری و غیرساختاری ارائه شده است. جهت پیاده‌سازی الگوی ارائه شده از الگوریتم‌های SRank، SimRank و P-Rank در بخش شباهت ساختاری و الگوریتم OF در بخش شباهت غیرساختاری استفاده شده است. آزمایشات بر روی بخشی از مجموعه داده شبکه اجتماعی توئیتر به منظور تخمین شباهت کاربران شبکه‌های اجتماعی و پیش‌بینی پیوندهای جدید بین آنان پیاده‌سازی شده است. دو گراف کوچک از مجموعه داده اصلی استخراج شده و پیوندهایی از گراف به صورت کاملاً تصادفی حذف شده‌اند. هدف الگوریتم‌ها پیش‌بینی پیوندهای حذف شده است. با توجه به نتایج آزمایشات مشخص شده است که الگوریتم‌های شباهت ساختاری بر اساس مسیرهای کوتاه در تخمین شباهت و پیش‌بینی پیوندهای جدید در مقایسه با الگوریتم‌های شباهت ساختاری بر اساس همسایگان گره دقیق‌تر هستند. همچنین مشخص گردید

Archive of SID

www.SID.ir

که طبق الگوی ارائه شده، الگوریتم‌های شباهت ساختاری بر اساس مسیرهای کوتاه در مقایسه با الگوریتم‌های شباهت ساختاری بر اساس همسایگان گره بهتر می‌توانند با الگوریتم‌های شباهت غیرساختاری ترکیب شوند.

Algorithm 1

```
1: procedure SIMILARITY(Input TrainData:Graph, TestData: Graph; Out put CorrectPrediction:Int)
2:     for all (a) ∈ G do
3:         MaxStructuralSimilarity ←n
4:         if n > 0 then
5:             ArrayMaxStructuralSimilarNode ← y|StructuralSimilarity(a, y) == n
6:             if (Count(ArrayMaxStructuralSimilarNode)) ≤ 3 then
7:                 TestPredictionCorrect(x, ArrayMaxStructuralSimilarNode)
8:             else
9:                 MaxUnstructuralSimilarity ←m
10:                if m > 0 then
11:                    ArrayMaxUnstructuralSimilarNode ← y|UnStructuralSimilarity(a, y) == m
12:                    if (Count(ArrayMaxUnstructuralSimilarNode) ≤ 10) then
13:                        TestPredictionCorrect(a, ArrayMaxUnstructuralSimilarNode)
14:        if (n) = 0 then
15:            MaxUnstructuralSimilarity ←m
16:            if m > 0 then
17:                ArrayMaxUnstructuralSimilarNode ← y|UnStructuralSimilarity(a, y) == m
18:                if (Count(ArrayMaxUnstructuralSimilarNode)) ≤ 10 then
19:                    TestPredictionCorrect(a, ArrayMaxUnstructuralSimilarNode)
```

شکل (۸) : شبه کد مربوط به الگوی جدید پیشنهادی جهت ترکیب الگوریتم شباهت ساختاری و غیرساختاری

مراجع

[1] W. Kim, O.-R. Jeong, and S.-W. Lee, "On social Web sites," Information Systems, vol. 35, pp. 215-236, 2010.

[2] M. Zhang, Z. He, H. Hu, and W. Wang, "E-rank: A Structural-Based Similarity Measure in Social Networks," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, 2012, pp. 415-422.

[3] M. Z. Shafiq, M. U. Ilyas, A. X. Liu, and H. Radha, "Identifying Leaders and Followers in Online Social Networks," Selected Areas in Communications, IEEE Journal on, vol. 31, pp. 618-628, 2013.

[4] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," Annual review of sociology, pp. 415-444, 2001.

[5] G. Jeh and J. Widom, "SimRank: a measure of structural-context similarity," in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, pp. 538-543.

[6] L. A. Adamic and E. Adar, "Friends and neighbors on the web," Social networks, vol. 25, pp. 211-230, 2003.

[7] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," Physical review E, vol. 69, p. 026113, 2004.

[8] E. Spertus, M. Sahami, and O. Buyukkokten, "Evaluating similarity measures: a large-scale study in the orkut social network," in Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005, pp. 678-684.

[9] P. Zhao, J. Han, and Y. Sun, "P-Rank: a comprehensive structural similarity measure over information networks," in Proceedings of the 18th ACM conference on Information and knowledge management, 2009, pp. 553-562.

[10] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," red, vol. 30, p. 3, 2008.

[11] P. Bhattacharyya, A. Garg, and S. F. Wu, "Analysis of user keyword similarity in online social networks," Social network analysis and mining, vol. 1, pp. 143-158, 2011.

[12] C. G. Akcora, B. Carminati, and E. Ferrari, "Network and profile based measures for user similarities on social networks," in Information Reuse and Integration (IRI), 2011 IEEE International Conference on, 2011, pp. 292-298.

[13] L. Backstrom and J. Leskovec, "Supervised random walks: predicting and recommending links in social networks," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 635-644.

[14] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," *Proceedings of the VLDB Endowment*, vol. 2, pp. 718-729, 2009.

Archive of SID

www.SID.ir

[15] H. Khosravi-Farsani, M. Nematbakhsh, and G. Lausen, "SRank: Shortest paths as distance between nodes of a graph with application to RDF clustering," Journal of Information Science, vol. 39, pp. 198-210, 2013.
[16] Dataset-UDI-TwitterCrawl-Aug2012, Sep 2014, https://wiki.cites.illinois.edu/wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012.
[17] S. Robertson, "Evaluation in information retrieval," in Lectures on information retrieval, ed: Springer, 2001, pp. 81-92.

زیرنویس‌ها

[1] Occurrence frequency
[2] Entity
[3] Interaction
[4] Collaboration
[5] Influence
[6] Recommender systems
[7] Leaders
[8] Followers
[9] Vertex degree
[10] Number of triangles
[11] Clustering coefficient
[12] Eigenvector centrality
[13] Average shortest path length
[14] Categorical data
[15] Majority voting
[16] Confidence level
[17] Test data
[18] Train data

Archive of SID

www.SID.ir