جستجوی شباهت در شبکه‌های ناهمگن بر مبنای فرامسیرهای وزن‌دار

سهیلا مولائی[1]، سما بابایی[2]، مصطفی صالحی[3]

[1] دانشجوی کارشناسی ارشد، دانشگاه تهران، تهران
soheila.molaei@ut.ac.ir

[2] دانشجوی کارشناسی ارشد، دانشگاه تهران، تهران
babaei.sama@ut.ac.ir

[3] عضو هیات علمی، دانشگاه تهران، تهران
mostafa_salehi@ut.ac.ir

چکیده

بسیاری از سیستم‌های اطلاعاتی را می‌توان به شکل شبکه‌ای ناهمگن، شامل گره‌ها و یال‌ها از انواع مختلف، مدل کرد. برای مثال در پایگاهی حاوی اطلاعات مقالات چاپ‌شده، انواع گره‌ها نظیر نویسنده، مقاله و کنفرانس و ارتباطات مابین آن‌ها قابل تعریف است. جستجوی شباهت گره‌ها در شبکه‌های ناهمگن از موضوعاتی است که در سال‌های اخیر مورد توجه محققین در حوزه‌ی علوم شبکه قرار گرفته است. برای این منظور شباهت گره‌ها با در نظر گرفتن مسیرهای مختلف مابین آن‌ها در شبکه تعریف شده است. به‌طور مشخص با استفاده از مفهوم فرامسیر –مسیرهایی که گره‌ها را از طریق چند نوع رابطه به یکدیگر متصل می‌کنند– معانی مختلفی از شباهت را خواهیم داشت. تاکنون چندین معیار شباهت بر پایه چارچوب فرامسیر مطرح شده است. با این حال میزان اهمیت هر فرامسیر در این روش‌ها در نظر گرفته نشده است که باعث می‌شود بخشی از اطلاعات تاثیرگذار در شناسایی گره‌های شبیه از بین برود. در این مقاله، یک معیار شباهت مبتنی بر فرامسیر، به نام SimSim، پیشنهاد داده شده است که با توجه به شباهت گره‌های درون یک مسیر، اهمیت آن مسیر را در نظر می‌گیرد. آزمایش‌ها بر روی مجموعه داده‌ی واقعی، میزان موثر بودن و کارایی این روش را به خوبی نشان می‌دهد.

کلمات کلیدی

جستجوی شباهت، شبکه‌ی اطلاعات ناهمگن، فرامسیر، شبکه‌ی اجتماعی

۱- مقدمه

در سال‌های اخیر، تحقیقات در حوزه علوم شبکه رشد زیادی داشته است و بسیاری از سیستم‌ها با دید شبکه‌ای مورد بررسی قرار گرفته‌اند. شبکه متشکل از تعدادی گره می‌باشد که توسط یال‌هایی به هم متصل شده‌اند. به عنوان نمونه‌هایی از شبکه می‌توان شبکه‌های اجتماعی، شبکه‌ی تحقیقات منتشر شده، شبکه‌ی زیستی و شبکه‌ی بزرگراه‌ها را نام برد [1].

در اغلب شبکه‌ها تنها روابط مابین گره‌ها مورد بررسی قرار گرفته است و اطلاعات اضافه در مورد ویژگی‌های زمانی و زمینه‌ای تعاملات نادیده گرفته شده است. اما اخیراً ویژگی ناهمگنی در سیستم‌های اطلاعاتی و در نظرداشتن آن در تحلیل‌ها، مورد توجه محققان این حوزه قرارگرفته است [2]. در واقع اکثر شبکه‌های جهان واقعی، ناهمگن[1] هستند. شبکه‌های ناهمگن، شبکه‌هایی شامل گره‌ها و یال‌ها از انواع مختلف بوده، که این یال‌ها می‌توانند نشانگر انواع مختلفی از روابط باشند [2,3]. از یک طرف، فرض اینکه همه‌ی گره‌ها از یک نوع هستند، ممکن است باعث از بین رفتن اطلاعات معنایی مهمی شود و از طرف دیگر باعث می‌شود که درک اطلاعات ارزشمندی چون

شکل کلی[۲] اطلاعات را از دست بدهیم. لذا لازم است که برای مطالعه‌ی این نوع شبکه‌ها چارچوب جدیدی تعریف شود و ابزارهای توسعه داده شده در حوزه‌ی علوم شبکه در سالیان گذشته، برای این نوع شبکه‌ها تعمیم داده شوند [4].

انواع مختلفی از تحلیل‌ها را می‌توان بر روی این شبکه‌ها استخراج کرد. یکی از این موارد، که از اهمیت ویژه‌ای برخوردار بوده، جستجوی شباهت بین اشیاء (گره‌ها) می‌باشد. در بسیاری از حالات، پیدا کردن اشیاء مشابه در شبکه به معنی پیدا کردن زوج‌ها، مانند پیدا کردن نویسنده‌های مقالات که زمینه‌های کاریشان مشترک است و از اعتبار مشابهی نیز برخوردارند، و یا پیدا کردن بازیگرانی که سبک فیلم‌های آن‌ها با هم مشابه است و یا پیدا کردن محصولاتی که کاربرد و محبوبیتشان شبیه یکدیگر است، می‌باشد.

در کارهای اخیر جستجوی شباهت در بین اشیاء در شبکه ناهمگن با در نظر گرفتن مسیرهای مختلف در شبکه، تعریف شده است. با استفاده از مفهوم مسیرهایی که اشیاء را از طریق چند نوع رابطه به یکدیگر متصل می‌کنند، که به آن‌ها فرامسیر[۳] گویند، معانی مختلفی از شباهت را خواهیم داشت. چندین معیار شباهت بر پایه‌ی چارچوب فرامسیر مطرح شده است که می‌توانند اشیاء شبیه در شبکه‌های ناهمگن را شناسایی کنند. مهم‌ترین این روش‌ها PathSim می‌باشد که در واقع روش پایه‌ی محاسبه‌ی شباهت بر اساس فرامسیرهاست [6]. با این حال این روش‌ها میزان اهمیت هر فرامسیر را در نظر نمی‌گیرند و باعث می‌شوند بخشی از اطلاعات به این صورت از بین برود.

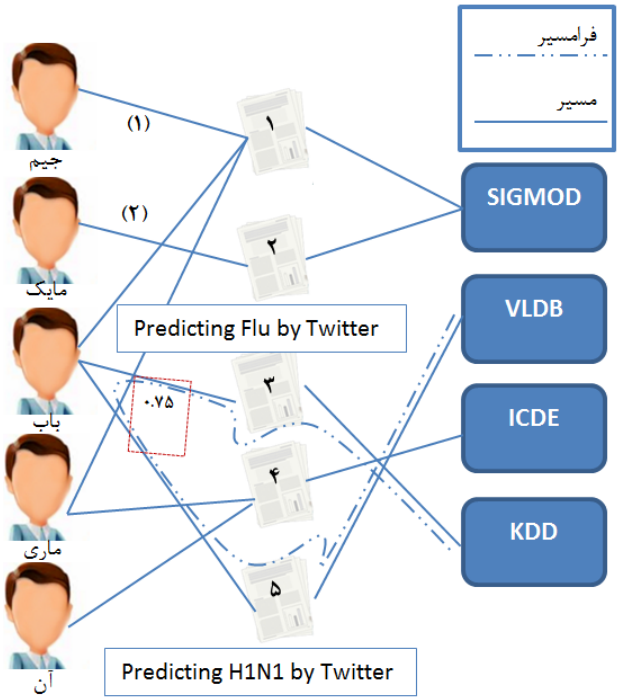
در این مقاله، یک معیار شباهت مبتنی بر فرامسیر پیشنهاد داده شده است که میزان اهمیتی برای هر نمونه مسیر در نظر می‌گیرد و بدین ترتیب، نتایج قابل استنادی به دست می‌آورد. میزان اهمیت هر نمونه مسیر، با توجه به شباهت اشیاء درون مسیر، محاسبه می‌شود. آزمایش‌ها بر روی مجموعه داده واقعی، میزان موثر بودن و کارایی این روش را به خوبی نشان می‌دهد.

در حقیقت، روش PathSim در به دست آوردن شباهت، به محتوای فرامسیرها توجه نمی‌کند و محتوای نمونه مسیرها را در حالتی ایده‌آل در نظر می‌گیرد که باعث می‌شود میزان شباهت بیشتری از واقعیت به دست آورد. حال آن‌که نوآوری روش پیشنهادی این مقاله این است که میزان اهمیت هر نمونه مسیر به صورت جداگانه در نظر گرفته شود و به صورت ضریبی بین ۰ و ۱، در هر نمونه مسیر دخیل باشد. بدین ترتیب، میزان شباهت اشیاء به یکدیگر، بیشتر به واقعیت نزدیک خواهد شد و این نتیجه مهم را در ادامه نشان خواهیم داد.

در ادامه‌ی مقاله، در بخش دوم، تعاریف و مفاهیم اولیه را مطرح خواهیم کرد. در بخش سوم مروری بر کارهای پیشین خواهیم داشت و با مزایا و معایب هر روش بیان خواهیم کرد. در بخش چهارم، راهکار پیشنهادی مطرح خواهد شد. در بخش پنجم ارزیابی روش پیشنهادی و مقایسه‌ی آن با رویکردهای پیشین عنوان می‌شود. در بخش ششم نیز به جمع‌بندی و پیشنهاداتی برای کارهای آتی پرداخته خواهد شد.

۲- مفاهیم اولیه

در این بخش، پیش از آن‌که به توضیح بعضی از مفاهیم اولیه‌ی این بحث پرداخته شود، ابتدا مثالی کاربردی عنوان می‌شود و مفاهیم اولیه روی این مثال توضیح داده خواهد شد. استفاده از داده‌ی کتابشناسی[۴] DBLP در اکثر مقالات مربوط به پیشنهاد معیار مشابهت در شبکه‌ی ناهمگن، مورد توجه بوده است. در این مقاله نیز برای نشان دادن مفاهیم اولیه و سپس کارایی روش پیشنهادی از شبکه‌ی کتابشناسی استفاده می‌کنیم. شبکه‌ی کتابشناسی DBLP، دارای مفاهیمی مانند نویسنده، مقاله، کنفرانس و لغات می‌باشد که در شکل (۱) می‌توان نحوه‌ی ارتباط این مفاهیم را مشاهده نمود.



شکل (۱) : شبکه‌ی کتابشناسی DBLP، شامل انواع گره‌ها نظیر نویسنده، مقاله و کنفرانس که می‌توانند توسط فرامسیرهایی به یکدیگر متصل باشند. برای مثال کنفرانس VLDB می‌تواند تحت فرامسیر VPAPV با نمونه مسیر $VLDB\text{-}P_5\text{-}Bob\text{-}P_3\text{-}KDD$ به کنفرانس KDD متصل شود.

تعریف ۱) شبکه‌ی اطلاعات ناهمگن[۵]: یک شبکه‌ی اطلاعات، انتزاعی از جهان واقعی است که بر روی اشیاء و تعاملات بین آن‌ها تمرکز می‌کند. این به این معنی است که این سطح از انتزاع، هم قدرت زیادی در ارائه و مرتب‌سازی اطلاعات ضروری جهان واقعی دارد و هم ابزاری کاربردی برای به دست آوردن دانش است. شبکه‌ی اطلاعات را به شکل زیر تعریف می‌کنیم [1,5]: یک گراف جهت‌دار $G=(V,E)$ است با یک تابع نگاشت نوع شیء $t:V\to A$ و تابع نگاشت نوع یال $\phi:E\to R$ که در آن هر گره $v\in V$ متعلق به یک شیء خاص $t(V)\in A$ است و هر یال $e\in E$ متعلق به یک رابطه‌ی خاص $\phi(e)\in R$ است که A نشان‌دهنده‌ی مجموعه گره‌ها و R نشان‌دهنده‌ی روابط بین گره‌ها است. تعریف می‌کنیم دو یال متعلق به یک رابطه هستند در صورتی که نقاط ابتدایی هر دو یال از یک نوع شیء باشد و نقاط انتهایی آن‌ها نیز از یک نوع شیء باشند. اگر $|A|>1$ و یا $|R|>1$ آن‌گاه شبکه‌ی اطلاعات ناهمگن داریم و در غیر این صورت شبکه اطلاعات همگن داریم. در شکل (۱)، دو یال (۱) و (۲)، هر دو متعلق به یک رابطه (رابطه‌ی نویسنده–مقاله) هستند، اما از آنجا که سه نوع شیء (نویسنده– مقاله–کنفرانس) و همچنین دو نوع یال (نویسنده–مقاله و مقاله–کنفرانس) داریم، این شبکه، یک شبکه‌ی اطلاعات ناهمگن است.

تعریف۲) فرامسیر: یک فرامسیر P مسیری است که بر طرح کلی شبکه $T_G=(A,R)$ تعریف می‌شود و به شکل $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} A_l$ نمایش داده می‌شود که یک رابطه ترکیب $R_1 o R_2 o \dots R_l$ بین انواع A_1 و A_{l+1} را نشان می‌دهد و o نشان‌دهنده عملگر ترکیب بر ارتباطات است [6]. برای راحتی کار، تنها نوع اشیاء را در مسیر نام می‌بریم: $P=(A_1 A_2 \dots A_{l+1})$. برای مثال در شبکه‌ی کتابشناسی شکل (۱)، رابطه‌ی شرکت کردن یک نویسنده در دو کنفرانس، به صورت فرامسیر VAV (کنفرانس، نویسنده، کنفرانس[۶]) عنوان می‌شود. به هر نمونه از یک فرامسیر، برای مثال شرکت کردن نویسنده باب در کنفرانس VLDB، به عنوان یک نمونه مسیر[۷] در نظر گرفته می‌شود.

مفهوم فرامسیر در شبکه‌های اطلاعات ناهمگن نقش مهمی دارد. براساس فرامسیرهای مختلف، اشیاء می‌توانند توسط مسیرهای مختلفی، روابط متفاوتی داشته باشند. در زمینه‌ی شباهت نیز فرامسیر نقش مهمی ایفا می‌کند، برای مثال، شباهت بین کنفرانس‌ها بر اساس فرامسیرهای مختلف، با هم متفاوت خواهد بود [6]. مثلاً تحت فرامسیر VPAPV، کنفرانس‌هایی که یک فرد واحد در هر دوی آن‌ها مقالاتی به چاپ رسانده است، بیشتر به یکدیگر شباهت دارند.

در جدول (۱) نشان‌های استفاده شده در تعاریف فوق، به طور خلاصه معرفی شده‌اند.

جدول (۱) : نشان‌گذاری

V	مجموعه گره‌های شبکه ناهمگن
E	مجموعه یال‌های شبکه ناهمگن
A	مجموعه انواع اشیاء
R	مجموعه انواع روابط
P	فرامسیر
P_i	نمونه مسیر

۳- کارهای پیشین

اخیراً، کارهای زیادی در زمینه‌ی تحلیل شبکه‌های ناهمگن صورت گرفته است که از بین آن‌ها معیار شباهت، تابعی مهم، پایه‌ای و کاربردی است. در شبکه‌های اطلاعات ناهمگن، تاکنون چندین معیار مختلف شباهت پیشنهاد شده است که در ادامه به مرور مهم‌ترین آن‌ها پرداخته شده است.

بعضی از معیارهای شباهت مانند شباهت کسینوسی، ضریب جاکارد، فاصله‌ی اقلیدسی و k-نزدیک‌ترین همسایه [7]، در شبکه‌ها به کار گرفته شده‌اند و بر اساس مقادیر ویژگی‌های گره‌ها در شبکه، میزان شباهت را محاسبه می‌کنند. این معیارها، روابط یالی بین اشیاء را در نظر نمی‌گیرند، به همین دلیل نمی‌توان این روش‌ها را به داده‌های شبکه‌ای بسط داد.

دسته‌ای دیگر از معیارهای شباهت وجود دارند که بر اساس ساختار یال‌ها و ارتباطات در شبکه، شباهت اشیاء را مشخص می‌سازند، از جمله این‌ها می‌توان به معیارهای شباهت نامتقارن[۸]، PageRank شخصی سازی شده[۹] و و ارزیابی احتمال شروع از یک گره و رسیدن به گره‌ی دیگر تحت الگوریتم قدم زدن تصادفی اشاره کرد [8]. در ادامه برخی از این روش‌ها را توضیح خواهیم داد.

معیار شباهت متقارن SimRank میزان شباهت دو گره را از طریق شباهت همسایه‌های دو گره تعیین می‌کند [9]. معیار شباهت SCAN شباهت دو گره را بر اساس مقایسه‌ی مجموعه همسایه‌های اولیه‌ی یک گره تعیین می‌کند [10]. این معیارها مختص شبکه‌های همگن هستند.

معیار شباهت PathSim روش پایه‌ی محاسبه‌ی شباهت بر اساس فرامسیرهاست [6]. این روش در شبکه‌ای ناهمگن که همه‌ی گره‌ها در آن از یک نوع هستند اما روابط متعددی بین آن‌ها برقرار است تعریف می‌شود. همچنین تمامی مسیرها در آن متقارن و دوطرفه در نظر گرفته می‌شود. PathSim از طریق فرامسیرها شباهت بین هر جفت گره از گره‌های شبکه را به دست می‌آورد و نتایج آن در بسیاری از حالات از سایر روش‌های به دست آوردن شباهت بر اساس قدم زدن تصادفی[۱۰]، با واقعیت انطباق بیشتری دارد.

این روش روابط زوج اشیاء را به صورت متقارن در نظر می‌گیرد و با فرامسیرهای متقارن کار می‌کند. بطور مشخص، شباهت دو شیء x و y که با $s(x,y)$ نشان داده می‌شود، برای فرامسیر متقارن ρ تحت رابطه‌ی (۱) محاسبه می‌شود:

$$s(x,y)=\frac{2\times\left|\{p_{x\to y}: p_{x\to y}\in\rho\}\right|}{\left|\{p_{x\to x}: p_{x\to x}\in\rho\}\right|+\left|\{p_{y\to y}: p_{y\to y}\in\rho\}\right|} \quad (۱)$$

که در آن $p_{x\to y}$ نمونه مسیری بین دو شیء x و y می‌باشد و $p_{x\to x}$ نمونه مسیری بین x و x است و $p_{y\to y}$ بین y و y می‌باشد.

معیار شباهت دیگری که مبتنی بر فرامسیرها معرفی شده است، AvgSim است که می‌تواند شباهت بین اشیاء از یک نوع و یا با نوع متفاوت را در یک چهارچوب یکنواخت به دست بیاورد [11]. این روش برای کار با شبکه‌های با مقیاس بزرگ نیز کارایی مناسبی دارد. مشکل این روش این است که تمامی فرامسیرها با هم یکسان فرض می‌شوند.

معیارهای شباهت مبتنی بر فرامسیر می‌تواند شباهت زوج اشیاء را برای مسیرهایی که دارای یال‌های متقارن هستند، به دست آورد، ولی داده‌ی واقعی حاوی ارتباطات غیرمتقارن بوده و نقشی اساسی در مفاهیم شباهت زوج گره‌ها بازی می‌کند. AsymSim روشی مبتنی بر فرامسیر برای اندازه‌گیری شباهت زوج گره‌ها است [8]، که هم مفاهیم شباهت زوج گره‌ها را به دست می‌آورد و هم به روابط نامتقارن در شبکه می‌پردازد. مشکل این روش این است که تنها ساختار محلی شبکه را مدنظر قرار می‌دهد و برای ساختار کلی شبکه مناسب نمی‌باشد.

روش دیگری برای به دست آوردن شباهت در شبکه‌های ناهمگن با عنوان HeteSim معرفی شده است [12]، که شامل ویژگی‌های زیر می‌باشد: ۱) معیاری یکنواخت: این روش می‌تواند ارتباط بین اشیای از یک نوع و یا اشیاء با نوع متفاوت را در یک چهارچوب یکسان اندازه‌گیری کند. ۲) معیاری محدود به مسیر: ارتباط بین جفت‌های اشیا، بر اساس مسیر جستجویی که دو شیء را به هم متصل می‌کند تعریف می‌شود. ۳) HeteSim دارای یک سری ویژگی‌های مناسب می‌باشد (مثلا تقارن و خود-بیشینگی[۱۱]) که این ویژگی‌ها برای بسیاری از وظایف داده‌کاوی، حیاتی هستند.

مشکل تمامی روش‌هایی که در بالا به آن‌ها اشاره شد، این است که تمامی فرامسیرها با هم یکسان فرض می‌شوند و بدین ترتیب، هیچ اولویتی برای هیچ کدام از نمونه مسیرها در نظر گرفته نمی‌شود و این موضوع باعث می‌شود بخشی از اطلاعات نادیده گرفته شود.

٤- راهکار پیشنهادی

همان طور که در بخش‌های پیشین اشاره شد، راهکارهای مختلفی برای به دست آوردن شباهت وجود دارد که از بین آن‌ها PathSim روش پایه‌ای و مهمی است که از فرامسیرها برای به دست آوردن شباهت استفاده می‌کند که از طریق رابطه‌ی (۱) محاسبه می‌شود. در رابطه‌ی $s(x,y)$ صورت کسر نشان‌دهنده مسیرهای عبوری از بین x و y است و مخرج کسر مسیرهای عبوری هرکدام بین خودشان، یعنی $x-x, y-y$ است، که برای توازن آمده است.

مشکل این روش این است که برای تمام مسیرها اگر مقاله‌های داده شده در کنفرانس‌ها به همدیگر شباهت داشته باشند یا خیر هیچ گونه تفاوتی در نظر گرفته نشده است. در حقیقت میزان شباهت مقالاتی که توسط یک نویسنده در دو کنفرانس به چاپ رسیده است، در میزان شباهت این دو کنفرانس به یکدیگر تاثیر به سزایی دارد. برای مثال در فرامسیر VPAPV یا همان VAV موضوع مقاله‌ها در نظر گرفته نشده است و فقط همین که یک نویسنده مشترک در هر دو کنفرانس مقاله داده است، در رابطه‌ی شباهت تاثیرگذار است. حال آن‌که، اگر شباهت موضوع مقاله‌های دو کنفرانس به یکدیگر بیشتر باشد، میزان شباهت این دو کنفرانس بیشتر از حالتی می‌شود که دو مقاله‌ی کاملاً بی‌شباهت را در این دو کنفرانس به چاپ برسانند.

در این مقاله، روشی در نظر گرفته شده است که شباهت مقاله‌ها را نیز در نظر بگیرد، یعنی مسیری که مقاله‌های شبیه‌تر دارند دارای ضریب بیشتر و مقاله‌هایی که شباهت کمتری به هم دارند دارای ضریب پایین‌تری در رابطه‌ی PathSim باشند. همان‌طور که در رابطه‌ی (۲) مشاهده می‌کنید، هر کدام از نمونه مسیرها دارای ضریبی می‌باشند. این رابطه را SimSim نامگذاری کرده‌ایم:

$$s(x,y)=\frac{2\times\left|\left\{\alpha_i p_{i_{x\to y}} : p_{i_{x\to y}}\in\rho\right\}\right|}{\left|\left\{\alpha_i p_{i_{x\to x}} : p_{x\to x}\in\rho\right\}\right|+\left|\left\{\alpha_i p_{i_{y\to y}} : p_{y\to y}\in\rho\right\}\right|} \quad (۲)$$

در رابطه‌ی (۲)، α_i ضریبی است که برای هر نمونه مسیر $p_{i_{x\to y}}$ در نظر گرفته می‌شود تا به هر نمونه مسیر وزنی برای تاثیرگذاری بدهد. این وزن از طریق میزان شباهت نام مقاله‌ها توسط روش n-gram به دست می‌آید [13]. این روش بیان می‌کند کلماتی که درجه‌ی ساختاری بالاتری دارند از نظر معنی بیشتر به یکدیگر شبیه هستند. در این روش هر کلمه متشکل از n-gramها است که n تعداد کاراکترهای مجاور در زیررشته هستند. با استفاده از این لیست معیار شباهت بین دو کلمه بر اساس تعداد n-gramهای مشترک و تمام n-gramهای موجود برای هر کلمه به دست می آید. برای bigram، n=۲ است و برای trigram، n=۳ است.

برای وضوح بیشتر مثالی برای فرامسیر VPAPV را در ادامه توضیح می‌دهیم. با توجه به شکل (۱)، باب مقالاتی در کنفرانس‌های KDD و VLDB با عنوان‌هایی که در شکل آمده است به چاپ رسانده است. بر طبق فرامسیر VPAPV، نمونه مسیر KDD-p3-Bob-p5-VLDB را داریم. میزان شباهت این دو کنفرانس در رابطه PathSim تحت فرامسیر فوق، ۱ در نظر گرفته شده است، یعنی کاملا به یکدیگر شباهت دارند، زیرا هر دو، تنها یک مقاله توسط یک فرد مشترک داشته‌اند و هیچ نویسنده‌ی دیگری نیز در این کنفرانس‌ها شرکت نکرده است. در این روش امتیاز هر مسیر ۱ در نظر گرفته شده است. با این فرض مقادیر را در رابطه‌ی (۱) قرار می‌دهیم که به این صورت است: $\frac{2\times(1\times1)}{1\times1+1\times1}=1$. اما در روش پیشنهادی، میزان شباهت مقاله‌ی سوم با مقاله‌ی پنجم توسط روش n-gram محاسبه می‌شود که در این مثال با n=۱، برابر ۰,۷۵۸ است و این عدد به عنوان ضریب (α_i) این نمونه مسیر به رابطه‌ی (۲) اضافه می‌شود. نحوه‌ی محاسبه‌ی این ضریب، به این صورت است که با استفاده از روش padded n-gram، کاراکترهای نام مقالات به صورت nتایی جدا شده و سپس با استفاده از رابطه‌ی جاکارد که در رابطه (۳) نشان داده شده است، میزان شباهت عنوان مقالات به دست می‌آید:

$$sim_{jaccard}(x,y)=\frac{|tok(x)\cap tok(y)|}{|tok(x)|+|tok(y)|-|tok(x)\cap tok(y)|}=\frac{|tok(x)\cap tok(y)|}{|tok(x)\cup tok(y)|} \quad (۳)$$

که با فرض n=۱، $tok(x)$ مجموعه‌ای شامل کاراکترهای عنوان مقاله است. نتیجه‌ی رابطه‌ی (۳) برای این مثال به این صورت خواهد بود:

$$\alpha=sim_{jaccard}(\text{predicting Flu by Twitter, predicting H1N1 by Twitter})=\frac{22}{29}=0.758$$

که در حقیقت، این دو عنوان دارای ۲۲ حرف مشترک (با احتساب کاراکتر فاصله) هستند و اجتماع کاراکترها نیز برابر ۲۹ است. در نتیجه شباهت این دو گره از طریق این فرامسیر $\frac{2\times(0.758\times1\times1)}{1\times1\times1+1\times1\times1}=0.758$ در نظر گرفته شده است. این ضرایب برای تمامی نمونه مسیرها محاسبه می‌شود و در رابطه (۲) قرار می‌گیرند. در واقع هر چه دو مقاله در هر کدام از نمونه مسیرهای فرامسیر VPAPV، به یکدیگر شباهت بیشتری داشته باشند، میزان شباهت دو کنفرانس ابتدا و انتهای این نمونه مسیر را بیشتر خواهند کرد و هر چه این شباهت بین مقالات کمتر باشد، شباهت بین کنفرانس‌های چاپ‌کننده آن مقالات نیز کمتر خواهد شد.

شبه‌کد الگوریتم روش پیشنهادی در شکل (۲) نشان داده شده است. همان‌طور که نشان داده شده است در ابتدا باید از مجموعه داده مورد نظر داده‌های مربوط به نویسنده‌ها جمع‌آوری شود. بعد از جمع‌آوری با توجه به فرامسیر انتخابی مثلا VPAPV، تمامی مسیرهای داده شناسایی می‌شود. برای هر مسیر توسط n-gram شباهت مقاله‌ها با هم پیدا شده و این ضریب به این مسیر داده می‌شود و در نهایت در رابطه‌ی شباهت قرار داده شده و بدین صورت شباهت دو کنفرانس با هم محاسبه می‌گردد.

SimSim Algorithm:

```
Choose a Database (in this case: DBLP):
for any metpath(VAV,AVA,etc.) do :
   Find all path instances for any metapath
    Calculate P(x,y) of each path instance with a similarity measure
(in this case: n-gram)
   if x==Source and y==Destination:
         Numerator=Sum P (source, Destination)
   else if x==Source and y==Source:
          SourceDenumerator=Sum P (Source, Source)
   else if x==Destination and y==Destination:
          DestinationDenumerator=Sum P (Destination,
Destination)
 Calculate SimSim formula:
    SimSim(Source, Destination)=Numerator/ (SourceDenumerator
+ DestinationDenumerator)
```

شکل (۲) : شبه‌کد روش پیشنهادی

۵- ارزیابی روش پیشنهادی

در این بخش به معرفی مجموعه داده‌ای که برای تست و ارزیابی از آن استفاده کرده‌ایم می‌پردازیم و نتایج حاصل از رابطه‌ی معرفی شده را در مقایسه با نتایج رابطه PathSim نشان می‌دهیم. برای مقایسه‌ی روش پیشنهاد شده با روش PathSim، شباهت کنفرانس‌ها را با هم تحت فرامسیر VPAPV محاسبه می‌کنیم و سپس توسط معیار ارزیابی تجمعی مبتنی بر بهره[۱۲] (nDCG) نتایج دو روش را با هم مقایسه می‌کنیم [14].

مجموعه داده‌ی مورد استفاده، شامل ۱۴۹۱۵ نویسنده، ۱۲۴۰۹ مقاله و ۱۵ کنفرانس می‌باشد که از مجموعه داده‌ی DBLP جمع‌آوری شده است. برای آزمایش، درخواستی[۱۳] برای به دست آوردن شبیه‌ترین کنفرانس‌ها به کنفرانس "PKDD" داده می‌شود و به این ترتیب، تمامی کنفرانس‌ها به ترتیب میزان شباهت‌شان با کنفرانس فوق، رتبه‌دهی می‌شوند که می‌توان نتایج این رتبه‌بندی‌ها را با هر کدام از روش‌های PathSim و SimSim در جدول (۲) مشاهده کرد. همان‌طور که مشاهده می‌شود، ترتیب کنفرانس‌های رتبه‌های ۳ تا ۶ در دو روش با یکدیگر متفاوتند. یعنی سومین کنفرانس شبیه به PKDD، تحت روش SimSim، کنفرانس ICDM است، اما رتبه‌ی همین کنفرانس تحت روش PathSim، برابر ۵ می‌باشد.

برای ارزیابی کیفیت رتبه‌بندی کنفرانس‌های شبیه تحت هر دو روش، با توجه به میزان شباهت کلی این کنفرانس‌ها به PKDD، هر کدام برچسب می‌خورند تا بتوانیم معیار nDCG را برای آن‌ها محاسبه کنیم. فرآیند برچسب‌گذاری به این صورت است که مقدار ۳ برای بیشترین میزان شباهت، مقدار ۲ برای شباهت تقریباً زیاد، مقدار ۱ برای شباهت تقریباً کم و مقدار ۰، برای کمترین میزان شباهت در نظر گرفته می‌شود و سپس از معیار nDCG (مقدار این معیار عددی بین ۰ و ۱ است و هرچه این میزان بالاتر باشد، نشان‌دهنده‌ی جواب‌های بهتر و مرتبط می‌باشد.) برای ارزیابی کیفیت رتبه‌دهی این روش‌ها از مقایسه‌ی نتایج رتبه‌دهی هر کدام با رتبه‌بندی برچسب‌گذاری شده، استفاده می‌کنیم.

سطر پایانی جدول (۲) نشان‌دهنده‌ی میزان nDCG برای دو روش PathSim و SimSim با nهای متفاوت است که همان‌طور که مشاهده می‌شود، این میزان برای رابطه SimSim با روش unigram (یعنی n-gram با n=۱) برابر ۰,۹۶۲۶ می‌باشد و این موضوع نشان‌دهنده‌ی این است که این روش، رتبه‌دهی را با بیشترین شباهت به روش پایه انجام داده است و در نتیجه به واقعیت بیشتر شبیه است. البته هرچه مقدار n در رابطه n-gram بیشتر شده، شباهت بین عنوان مقالات، دقیق‌تر لحاظ شده و با n=۲ بالاترین مقدار nDCG، یعنی ۱ به دست می‌آید. یعنی با افزایش مقدار n در رابطه n-gram میزان nDCG بیشتر شده و نسبت به PathSim حدود ۰,۰۸ بهبود داشته است.

٦- نتیجه گیری و کارهای آتی

در این مقاله، روش محاسبه‌ی شباهت بین اشیاء در شبکه‌ی اطلاعات ناهمگن، بر اساس مفهوم فرامسیر، پیشنهاد شد که این روش بهبودی بر معیار شباهت PathSim است که به عنوان روشی کارا و مناسب برای محاسبه‌ی شباهت اشیاء شبکه‌های ناهمگن در سال‌های اخیر مطرح بوده است.

نشان داده شد که روش PathSim تمامی نکات لازم در بررسی شباهت اشیاء را مدنظر قرار نمی‌دهد. برای مثال ماهیت فرامسیرهای مورد استفاده باعث می‌شوند هر نمونه مسیر از آن‌ها دارای ضریب اهمیت متفاوتی باشد و تمامی نمونه مسیرها تاثیر یکسانی در شباهت بین اشیاء نداشته باشند. بنابراین روش جدیدی مطرح شد که با در نظر گرفتن اهمیت هر فرامسیر در شباهت، معیار PathSim بهبود داده شود و نتایج قابل اعتمادی به دست آید.

برای بهبود این روش، می‌توان با در نظر گرفتن مفهوم شباهت نمونه مسیرها، معیار اندازه‌گیری شباهت برای فرامسیرهای نامتقارن را به دست آورد.

جدول (۲): رتبه‌بندی کنفرانس‌ها بر اساس میزان شباهتشان به کنفرانس PKDD تحت دو روش PathSim و SimSim

رتبه	*Base*	*بر اساس PathSim*	*بر اساس SimSim برای unigram*	*بر اساس SimSim برای bigram*
۱	PKDD	PKDD	PKDD	PKDD
۲	ICDM	SDM	SDM	ICDM
۳	SDM	PAKDD	ICDM	SDM
۴	PAKDD	Data Min. Knowl. Discov.	PAKDD	KDD
۵	KDD	ICDM	KDD	PAKDD
۶	Data Min. Knowl. Discov.	KDD	Data Min. Knowl. Discov.	Data Min. Knowl. Discov.
۷	SIGKDD Explorations	SIGKDD Explorations	SIGKDD Explorations	Knowl. Inf. Syst.
۸	Knowl. Inf. Syst.	Knowl. Inf. Syst.	Knowl. Inf. Syst.	SIGKDD Explorations
۹	J. Intell. Inf. Syst.	J. Intell. Inf. Syst.	J. Intell. Inf. Syst.	KDID
۱۰	KDID	KDID	KDID	J. Intell. Inf. Syst.
۱۱	TKDD	TKDD	TKDD	ICDE
۱۲	ICDE	ICDE	ICDE	TKDD
۱۳	VLDB	VLDB	VLDB	VLDB
۱۴	SIGMOD	SIGMOD	SIGMOD	SIGMOD
۱۵	TKDE	TKDE	TKDE	TKDE
nDCG	**۱**	**۰,۹۲۲۱**	**۰,۹۶۲۶**	**۱**

مراجع

[1] Sun, Y. and Han, J. "*Mining heterogeneous information networks: A structural analysis approach."* SIGKDD Explor. Newsl. 14(2), pp. 20–28, 2013.

[2] Han, J. *"Mining heterogeneous information networks by exploring the power of links,"* Discovery Science, pp. 13–30, 2009.

[3] Sun, Y. Han, J. Zhao, P. Yin, Z. Cheng, H. Wu, T. "*RankClus: integrating clustering with ranking for heterogeneous information network analysis,"* EDBT, pp. 565–576, 2009.

[4] Kivela, M. Arenas, A. Barthelemy, M. Gleeson, J. P. Moreno, Y. Porter, M. A. "*Multilayer Networks."* Journal of Complex Networks, pp. 1–69, 2014.

[5] Sun, Y. Han, J. Yu, Y. *"Ranking-based clustering of heterogeneous information networks with star network schema,"* KDD, pp. 797–806, 2009.

[6] Sun, Y. Han, J. Yan, X. Yu, P. S. Wu, T. "*PathSim : Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks.*" In Proc. 2011 Int. Conf. Very Large Data Bases (VLDB'11), Seattle, WA, 2011.

[7] Kolahdouzan, M. Shahabi, C. *"Voronoi-based K nearest neighbor search for spatial network databases,"* VLDB, pp. 840–851, 2004.

[8] Tedesco, J. *AsymSim: Meta path-based similarity with asymmetric relations.* Master's thesis, University of Illinois at Urbana-Champaign, 2013.

[9] Jeh, G. Widom, J. "*Simrank: a measure of structural-context similarity*," in KDD, pp. 538–543, 2002.

[10] Xiaowei, X. Nurcan, Y. Zhidan, F. Thomas, A. J. S. *"SCAN: an structural clustering algorithm for networks,"* KDD, pp. 824–833, 2007.

[11] Meng, X. Shi, C. Li, Y. Zhang, L. Wu, B. *"Relevance measure in large-scale heterogeneous networks."* Web Technologies and Applications -16th Asia-Pacific Web Conference, APWeb 2014, Changsha, China, September 5-7, 2014. Proceedings,Vol. 8709 of Lecture Notes in Computer Science, pp. 636-643, 2014.

[12] Shi, C. Kong, X. Huang, Y. Yu, P. S. Wu, B. "*HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks."* IEEE Transactions on Knowledge and Data Engineering, 26(10), pp. 2479-2492, 2014.

[13] Tengku, M. Sembok. Z. T. Bakar, A. "*Effectiveness of Stemming and n-grams String Similarity Matching on Malay Documents*," vol. 5, no. 3, pp. 208–215, 2011.

[14] Jarvelin, K. and Kekalainen, J. *"Cumulated gain-based evaluation of IR techniques."* ACM TOIS 20(4), pp. 422–446, 2002.

پانویس‌ها

[1] Heterogeneous
[2] Schema-level
[3] Meta-path
[4] Bibliographic
[5] Heterogeneous Information Network
[6] Venue-Author-Venue
[7] Path instance
[8] Asymmetrical similarity
[9] Personalized PageRank
[10] Random-walk
[11] Self-maximum
[12] Normalized Discounted Cumulative Gain
[13] Query