ارائه رهیافت جدید برای استخراج لینک های مفهومی پرتکرار از شبکه های اجتماعی

سامان پورسیاه*[۲] حسین سلامی [۱] حمید طباطبایی[۳] محبوب فریمانی[۴]

[۱] گروه مهندسی کامپیوتر، واحد قوچان، دانشگاه آزاد اسلامی، قوچان، ایران
[۲و۴] گروه مهندسی کامپیوتر، موسسه آموزش عالی فردوس، مشهد، ایران
[۳] باشگاه پژوهشگران جوان و نخبگان، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

*سامان پورسیاه

ارائه رهیافت جدید برای استخراج لینک های مفهومی پرتکرار از شبکه های اجتماعی

چکیده: لینک های مفهومی یکی از رهیافت های جدید برای توصیف شبکه های اجتماعی است که دانش نهفته در شبکه های اجتماعی را از طریق یک ساختار خلاصه شده با عنوان دیدگاه مفهومی ارائه می نماید. مهمترین چالش در بدست آوردن دیدگاه مفهومی از یک شبکه اجتماعی استخراج لینک های مفهومی پرتکرار است که این امر برای شبکه های با مقیاس بزرگ بسیار زمانبر است. در این نوشتار، روش جدیدی برای استخراج لینک های مفهومی پرتکرار از شبکه های اجتماعی ارائه شده است که با بکارگیری مفهوم وابستگی ها، سعی در تسریع فرآیند استخراج لینک های مفهومی دارد. روش پیشنهادی قادر خواهد بود در صورت وجود وابستگی ها در بین داده ها، این فرآیند را تسریع بخشد.

کلمات کلیدی: تحلیل شبکه های اجتماعی، لینک های مفهومی پرتکرار، داده کاوی، گراف کاوی، کاوش شبکه های اجتماعی

A New Approach to Extract Frequent Conceptual Links from Social Networks

Abstract: Conceptual link is a new approach for describing social networks. In this approach, the concealed knowledge in social network is presented using a concise structure called Conceptual view. Main challenge to achieve conceptual view from a social network is extracting frequent conceptual links, which is very time consuming for large networks. In this paper, a new method for extracting frequent conceptual links from social networks is provided where by using the concept of dependency, it is tried to accelerate the process of extracting conceptual links. The proposed method will be able to accelerate this process if there are dependencies between data.

Keywords: social network analysis, frequent conceptual link, data mining, graph mining, exploring social networks

۱-مقدمه

شبکهٔ اجتماعی ساختاری اجتماعی است که از عاملهایی (عموماً افراد یا سازمان ها) که توسط یک یا چند نوع خاص از وابستگی ها- مانند ایده‌ها و تبادلات مالی، دوستان، خویشاوندی، لینک‌های وب، سرایت بیماری‌ها (اپیدمولوژی)- به هم متصل اند، تشکیل شده است. شبکه های اجتماعی در دسته های متفاوتی وجود دارند که میتوان برخی از آنها را در [۱] یافت. نتایج تحقیقات مختلف بیانگر آن است که می‌توان از ظرفیت شبکه‌های اجتماعی در بسیاری از سطوح فردی و اجتماعی به منظور شناسایی مسائل و تعیین راه حل آنها، برقراری روابط اجتماعی، اداره امور تشکیلاتی، سیاستگذاری و رهنمون سازی افراد در مسیر دستیابی به اهداف استفاده نمود.

تحلیل شبکه اجتماعی روش تشخیصی قدرتمندی برای تحلیل طبیعت و الگوی ارتباطات میان اعضای یک گروه خاص می باشد. تحلیل شبکه اجتماعی کمک می کند مجموعه های پیچیده ای از روابط بین عاملهای مرتبط را به مثابه ی نقشه هایی(گراف یا نگاره های گروهی) از سمبلهای متصل تجسم کرده، و الگوهای موجود درون این مجموعه ها را تجسم و بررسی نموده و سنجه های دقیق اندازه، شکل و تراکم شبکه را به مثابه ی یک کل و موقعیت هر عنصر را درون آن محاسبه کنیم. برای نمونه در علم بیماریهای مسری[۱] تحلیل شبکه اجتماعی برای کمک در فهم اینکه چگونه الگوهای مخاطبین انسانی در گسترش بیماریهایی مانند HIV در جمعیت ها کمک کرده و یا مانع میشوند استفاده شده است. علاوه بر این تحلیل شبکه های اجتماعی ابزاری مفید برای تجسس در حجم بالاست- برای مثال برنامه ی آگاهی اطلاعات کلی[۲] تحقیقات مفصلی را در مورد استراتژیهایی که برای تحلیل شبکه های اجتماعی برای تعیین اینکه آیا شهروندان، تهدیدهای سیاسی محسوب می شوند یا خیر انجام داده است.

از میان انواع شبکه های اجتماعی، شبکه های اجتماعی آنلاین در بین محققین بیشتر مورد توجه بوده است. یک جنبه کلیدی از بسیاری از شبکه های اجتماعی آنلاین این است که غنی از داده می باشند، و لذا چالشها و فرصتهای بی سابقه ای را از دیدگاه کشف دانش و داده کاوی فراهم می آورند. یکی از مهمترین حوزه های مطالعاتی داده کاوی سنتی، مساله کاوش الگوی تکراری است. در زمینه ی ساختار داده ای پیچیده مانند شبکه ها، مسئله کاوش مجموعه اقلام تکراری بصورت یافتن زیرمجموعه ای از گره ها (زیرگراف هایی) که بصورت مکرر در یک شبکه رخ می دهند مطرح می شود و با عنوان گراف کاوی از آن یاد می شود. اگرچه روشهای ابتدایی در این حوزه معیارهایی که از نظریه ی گراف ناشی میشد را بکار می گرفته اند [۲] رهیافتهای جدید که با عنوان کاوش شبکه های اجتماعی یا بعنوان ساده تر کاوش لینک شناخته می شوند تلاش می نمایند علاوه بر ساختار شبکه، ویژگیهای گره ها را نیز برای استخراج نوع جدیدی از الگوها بررسی نمایند [۳] [۴] [۵].

در [۶]، رهیافت جدیدی با عنوان لینک های مفهومی برای توصیف شبکه های اجتماعی معرفی گردید. لینک های مفهومی، دانشی را در مورد گروههایی از گره هایی که در یک شبکه اجتماعی بصورت متراکمی به یکدیگر متصل شده اند فراهم آورده و از طریق یک ساختار کاهش داده شده که با عنوان دیدگاه مفهومی[۳] نامیده می شود منجر به یک نمایش معنایی از شبکه اجتماعی می گردد. با این حال، مسئله ی استخراج لینک های مفهومی پرتکرار حداکثر، همانند مسئله استخراج مجموعه آیتم های مکرر حداکثر [۷] دارای پیچیدگی NP-hard [۸] است. در این پژوهش هدف ما ارائه راهکاری جدید برای تسریع هرچه بیشتر در استخراج این لینکها می باشد. برای این منظور، الگوریتم D-MFCLMin ارائه شده است که با بکارگیری مفهوم وابستگی بین مجموعه آیتم ها، و با هرس فضای جستجو سبب کاهش زمان لازم برای استخراج لینکهای مفهومی پرتکرار می گردد.

ساختار این نوشتار به صورت زیر خواهد بود. در بخش بعد، ابتدا مفهوم لینکهای مفهومی ارائه شده و سپس در بخش سوم، روشهای پیشنهاد شده برای استخراج لینکهای مفهومی پرتکرار معرفی می شوند. در بخش چهارم به معرفی روش پیشنهادی می پردازیم. بخش پنجم، نتایج ازمایشات ارائه شده است و نهایتا در بخش ششم، نتیجه گیری و کارهای آتی آمده است.

۲-تشریح مسئله و تعاریف[۶]

در حوزه جستجوی لینکهای مفهومی پرتکرار، یک الگو بصورت "مجموعه ای از لینکهای بین دو گروه از گره ها تعریف میشود که گره ها در هر گروه، ویژگی های مشترکی را به اشتراک میگذارند". هنگامی که این الگوهای با تکرار کافی در شبکه یافت شوند، بعنوان

[۱] epidemiology

[۲] Total Information Awareness program

[۳] Conceptual View

الگوهای تکراری فرض شده و آنها را لینکهای مفهومی پرتکرار می نامیم.

به شکل رسمی تر فرض نمایید $G = (V,E)$ یک شبکه باشد که V مجموعه گره ها و E مجموعه یال ها با $E \subseteq V \times V$ است. V بصورت رابطه $R(A_1,\dots,A_N)$ تعریف می شود که هر A_i یک صفت است. بنابراین هر راس $v \in V$ بوسیله ی تاپل $(a_1,\dots,a_N)$ تعریف می شود که در آن $\forall k \in [1..N],$ $v[A_k] = a_k$، مقدار صفت A_k در v است. یک آیتم، یک عبارت منطقی بصورت $A = x$ است که A یک صفت و x یک مقدار می باشد. آیتم خالی بصورت $\emptyset$ نشان داده می شود. یک آیتم ست، ترکیبی از آیتم هاست، برای نمونه $A_1 = x$ و $A_2 = y$ و $A_3 = z$. یک آیتم ست، m، که ترکیبی از k آیتم غیرخالی باشد، یک مجموعه آیتم kتایی نامیده میشود و به صورت m^k نشان داده می شود$(|m^k| = k)$.

فرض کنید m و sm دو آیتم ست باشند. اگر $sm \subseteq m$، می گوییم که sm یک زیرآیتم ست و m یک ابرآیتم ست از sm است. برای مثال $sm = xy$ یک زیرآیتم ست از $m = xyz$ است.

مجموعه تمام آیتم ست های tتایی ساخته شده از V را با I^t نشان می دهیم. همچنین UI^t را بصورت زیر تعریف می نماییم(مجموعه تمام آیتم ست های حداکثر tتایی):

$$UI^t = \bigcup_{k=1}^{t} I^k \qquad (۱)$$

فرض کنید G یک گراف جهت دار باشد. بنابراین برای هر مجموعه آیتم m در UI^N، V_m را بصورت مجموعه گره هایی در V که مطابق با الگوی m می باشند (اصطلاحا m را برآورده می نمایند) نشان داده و تعریف می نماییم:

- مجموعه لینک سمت چپ m (LE_m): مجموعه لینک هایی از E که از گره هایی که m را برآورده می نمایند شروع می شوند.

$$LE_m = \{e \in E; e = (a,b), a \in V_m\}$$

- مجموعه لینک سمت راست m (RE_m): مجموعه لینک هایی از E که به گره هایی که m را برآورده می نمایند وارد می شوند.

$$RE_m = \{e \in E; e = (a,b), b \in V_m\}$$

تعریف ۱: **لینک مفهومی**. فرض کنید $m1$ و $m2$ دو آیتم ست و V_{m1} و V_{m2} به ترتیب مجموعه گره ها در V باشند که $m1$ و $m2$ را برآورده می نمایند. $E_{(m_1,m_2)}$ مجموعه لینک های متصل کننده ی گره ها در V_{m1} به گره ها در V_{m2} می باشد،

$$E_{(m_1,m_2)} = LE_{m1} \cap RE_{m2} = \{e \in E; e = (a,b)\ a \in V_{m1} \text{ and } b \in V_{m2}\} \qquad (۲)$$

تعریف ۲: پشتیبانی $E_{(m_1,m_2)}$ را نسبتی از لینکها در E که به $E_{(m_1,m_2)}$ تعلق دارد می نامیم.

$$supp(E_{(m_1,m_2)}) = \frac{|E_{(m_1,m_2)}|}{|E|} \qquad (۳)$$

تعریف ۳. گفته می شود که یک لینک مفهومی پرتکرار وجود دارد و می نویسیم $(m1,m2)$ اگر پشتیبانی $E_{(m_1,m_2)}$ بیشتر از حدآستانه ی حداقل β باشد، $(supp(E_{(m_1,m_2)}) > \beta)$.

تعریف ۴. فرض نمایید UI^t مجموعه ی تمام آیتم ست های حداکثر tتایی در V باشد، FL^t بعنوان مجموعه لینک های مفهومی پرتکرار که از این آیتم ست ها استخراج شده است، تعریف می شود.

$$FL^t = \bigcup_{m_1 \in UI^t, m_2 \in UI^t} \{E_{(m_1,m_2)}; \frac{|E_{(m_1,m_2)}|}{|E|} > \beta\} \qquad (۴)$$

ویژگی ۱. ویژگی پرتکرار بودن. بنابر تعریف ۳، اگر لینک $(m1,m2)$ پرتکرار باشد، مجموعه های LE_{m1} و RE_{m2} شرط زیر را براورده می نمایند:

$$|LE_{m1}| > \beta \times |E| \text{ و } |RE_{m2}| > \beta \times |E| \qquad (۵)$$

تعریف ۵. زیرلینک مفهومی. فرض نمایید دو آیتم ست $sm1$ و $sm2$ به ترتیب زیرآیتم ست های $m1$ و $m2$ در UI باشند. لینک مفهومی $(sm1, sm2)$ زیرلینک (m_1, m_2) نامیده می شود، بطور مشابه (m_1, m_2) ابرلینک $(sm1, sm2)$ نامیده میشود و آنرا بصورت $(sm1, sm2) \subseteq (m_1, m_2)$ می نویسیم.

ویژگی ۲. ویژگی بستار زیرین. اگر یک لینک مفهومی l پرتکرار باشد تمامی زیرلینکهای آن نیز پرتکرار می باشند. بنابراین اگر یک لینک پرتکرار نباشد، تمامی ابرلینک های آن نیز پرتکرار نمی باشند.

تعریف ۶. لینک مفهومی پرتکرار حداکثر. فرض نمایید β یک حد آستانه پشتیبانی داده شده باشد، گوییم لینک مفهومی پرتکرار حداکثر (MFCL)، هر لینک مفهومی پرتکرار l بگونه ای که هیچ ابرلینک $\acute{l}$ از l که پرتکرار باشد وجود نداشته باشد. بصورت رسمی تر:

$$\nexists \acute{l} \in FL^N \text{ بگونه ای که } l \subset \acute{l} \qquad (۶)$$

۳-کارهای پیشین

روشهای مشهوری از کاوش شبکه اجتماعی برای استخراج اشکال مختلف دانش از این شبکه ها پیشنهاد شده اند. مشابه با حوزه ی داده کاوی سنتی، دامنه ی کاوش شبکه های اجتماعی گستره ی وسیعی از وظایف مانند دسته بندی[۴]، خوشه بندی[۵]، جستجو برای الگوهای تکراری یا پیشگویی لینک[۶] را آدرس دهی می نماید. به ذاته این روشها را می توان به دو دسته تقسیم نمود [۸].

- رهیافت های مبتنی بر مدلسازی پیشگویی[۷] که دربرگیرنده تکنیک هایی است که حقایق جاری و گذشته را برای ایجاد فرضیات پیشگویانه در مورد رویدادهای آینده یا ناشناخته تحلیل می نمایند.
- رهیافت های مبتنی بر مدلسازی توصیفی[۸] که مجموعه ای از تکنیکها را پوشش می دهند که هدفشان خلاصه سازی داده ها بوسیله ی شناسایی برخی ویژگی های مرتبط بمنظور توصیف چگونگی سازماندهی چیزها و نحوه واقعی کارکردن آنهاست.

در این پژوهش تمرکز بر رهیافت توصیفی از شبکه های اجتماعی است. این رهیافتها را میتوان به ۴ دسته تقسیم نمود. ۱) خوشه بندی مبتنی بر اتصال[۹] (بعنوان مسئله ی شناسایی جوامع[۱۰] نیز شناخته می شود) که گروههای متراکمی از گره ها را جستجو می نماید و هدفشان، تجزیه شبکه به چندین مولفه متصل (جوامع) است، بگونه ای که گره ها در هر مولفه دارای اتصالاتی با تراکم بالا باشند درحالی که گره ها در مولفه های مختلف دارای کمترین تراکم اتصال باشند. از جمله روشهای پیشنهادی در این دسته می توان به الگوریتم SLPA [۹]، TopGC [10]، [۱۱]، MCD [۱۲]، CGGC [۱۳]، CONCLUDE [۱۴]، DSE [۱۵] و SPICi [۱۶] اشاره نمود. ۲) خوشه بندی ترکیبی[۱۱] که همزمان ساختار و صفات گره ها را برای شناسایی خوشه ها در نظر می گیرد. هدف این شکل جدید رهیافت ها، افراز شبکه برای جستجوی توازنی بین شباهت های ساختاری و صفات است بگونه ای که گره های با صفات مشترک در یک افراز گروهبندی شوند و گره های داخل یک افراز بصورت متراکمی متصل باشند. این نوع رهیافت ها یک افراز مفهومی تر از شبکه را که الزاما متناسب به زمینه باشد فراهم می آورند. از جمله روشهای خوشه بندی ترکیبی می توان به SA-Cluster [۴] و CESNA [۱۷] اشاره نمود. ۳) کاوش زیرگراف مکرر[۱۲] که بر استخراج زیرساختارهایی که مکررا در شبکه رخ می دهند تمرکز می نماید. پراستفاده ترین تعریف از یک الگو بصورت یک زیرگراف متصل است [۱۸]. بنابراین تکنیک هایی که برروی جستجوی الگوهای تکراری در شبکه های اجتماعی تمرکز می نمایند، هدفشان شناسایی زیرگراف هایی است که بصورت مکرر در یک شبکه خیلی بزرگ یا در یک پایگاهی از شبکه ها، براساس یک حدآستانه حداقل رخ می دهند. از جمله روشهای شاخص در این دسته میتوان به الگوریتمهای مبتنی بر Apriori [۱۹] و رشد الگو [۲۰] اشاره نمود. ۴) لینکهای مفهومی پرتکرار، اطلاعات ساختار شبکه و نیز ویژگی های گره را برای فراهم آوردن دانشی در مورد گروه هایی از گره ها که در یک شبکه اجتماعی دارای اتصال بیشتری می باشند را ترکیب می نماید. استخراج اتصالات مفهومی پرتکرار حداکثر پیچیدگی مشابه با مجموعه آیتم های مکرر حداکثر ایجاد می نماید زیرا ثابت شده است که این پیچیدگی NP-hard است. استخراج تمام MFCLها از یک شبکه اجتماعی ممکن است مسئله ای چالش برانگیز و از لحاظ محاسباتی پیچیده باشد. با توجه به تعاریف مربوط به مفهوم لینکهای مفهومی، به بررسی روشهای ارائه شده برای استخراج این لینک ها می پردازیم.

اگر فضای جستجو بسیار گسترده باشد، کشف تمامی لینکهای پرتکرار در یک شبکه بسیار هزینه بر است. در یک رهیافت ساده، لازم است تمامی مجموعه آیتم های ممکن را تولید کرده و سپس تکرار هر جفت از آنها را بررسی نمود. بمنظور کاهش این زمان، در ابتدا الگوریتم FLMIN [۲۱] پیشنهاد شد. این الگوریتم یک رهیافت پایین به بالا را با اعمال ویژگی ۲ برای کاهش تدریجی فضای جستجو برای دربرگرفتن ابرآیتم ستهایی که بصورت بالقوه ای در لینکهای مفهومی پرتکرار قرار خواهند گرفت بکار برد.

---

[۴] Classification

[۵] Clustering

[۶] Link Prediction

[۷] Predictive

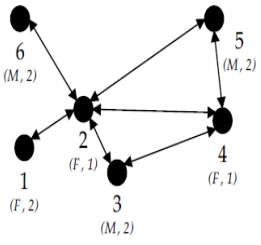
[۸] Descriptive

[۹] Link Based Clustering

[۱۰] Community Detection

[۱۱] Hybrid Clustering

[۱۲] Frequent Sub-graph Mining



(الف)

L_1	Support
((M, *), (F, *))	0.357
((M, *), (*, 1))	0.357
((F, *), (*, 1))	0.214
((F, *), (*, 2))	0.428
...	...

L_2	Support
((M, 2), (F, *))	0.357
((M, 2), (F, 1))	0.357
((F, 1), (*, 2))	0.428
((F, 1), (M, *))	0.357
...	...

(ب)

شکل ۱.نمونه ای از استخراج لینکهای مفهومی توسط الگوریتم FLMIN [21]

در شکل (۱)، مراحل عملکرد پایین به بالای الگوریتم FLMIN نشان داده شده است. علامت * به معنی این است که صفت می تواند هر مقداری را بپذیرد. در ابتدا الگوریتم با جستجوی لینکهای پرتکرار که شامل مجموعه آیتم ستهای 1تایی می باشند آغاز میگردد (L1 در تصویر بخش ب). سپس، با بکارگیری ویژگی ۲، می دانیم که ابر آیتم ستهایی از مجموعه آیتم ستهای 1تایی که در لینک های مفهومی قرار ندارند، در لینکهای مفهومی قرار نخواهند گرفت (L2 را مشاهده نمایید). به طور کلی در مرحله t، می توان فضای جستجو را تنها به زیرآیتم ستهایی که در لینک های مفهومی مرحله ی t-1 قرار دارند محدود نمود.

در [۲۲]، الگوریتم MAX-FLMin ارائه شد. در این الگوریتم، هدف یافتن لینکهای پرتکرار حداکثر می باشد. در مقایسه با الگوریتم قبل، در این الگوریتم تنها مجموعه آیتم ستهایی که ویژگی ۱ را برآورده می نمایند برای ایجاد لینکها مورد استفاده قرار می گیرند و سپس به بررسی پرتکرار بودن آنها پرداخته می شود. علاوه بر این، این الگوریتم در فرآیند بررسی لینک ایجاد شده بمنظور افزودن آن به لینک های پرتکرار، علاوه بر شرط پرتکرار بودن، عدم وجود لینک پرتکرار حداکثرتری نسبت به لینک فعلی را نیز بررسی می نماید. همچنین در صورتی که یک لینک پرتکرار به لیست نتیجه (لیست لینک های پرتکرار حداکثر) افزوده گردد، تمامی زیرلینکهای پرتکرار آن از لیست حذف خواهند شد.

در [۵]، الگوریتم H-MFCLMin ارائه گردید. در این الگوریتم، برای تسریع در استخراج لینک های پرتکرار مفهومی حداکثر، برخی از آیتم ستها فیلتر می شوند. مجموعه آیتم ستهای حذف شده، شامل آیتم ستهایی می شوند که تعداد گره های مربوطه آنها در شبکه کمتر از حد آستانه α باشد. α یک پارامتر ورودی برای الگوریتم می باشد. نویسندگان فرض نموده اند که لینک های مفهومی پرتکرار بین آیتم ستهای پرتکرار وجود دارد، در واقع این فیلتر سازی با این استدلال انجام می گیرد که احتمال کمی وجود دارد که آیتم ستهایی با فرکانس کم بتوانند نسبت بالایی از لینک ها را در سطح شبکه به خود جذب نمایند و لذا با فیلتر نمودن این نوع از آیتم ستها علیرغم کاهش فضای جستجو، اطلاعات خاصی از شبکه ی مفهومی نهایی از دست نخواهد رفت.

۴-الگوریتم پیشنهادی

در این نوشتار الگوریتم D-MFCLMin برای استخراج لینک های مفهومی پیشنهاد شده است. این الگوریتم با هرس نمودن فضای جستجو با استفاده از بکارگیری مفهوم وابستگی، باعث تسریع در استخراج لینک های مفهومی می شود. در ادامه ابتدا به معرفی مفهوم وابستگی پرداخته و سپس الگوریتم پیشنهادی معرفی و مورد بحث قرار خواهد گرفت.

تعریف ۷: وابستگی. فرض کنید m^t و n^t دو آیتم ست tتایی باشند. گوییم m^t به n^t وابسته است و بصورت $n^t \rightarrow m^t$ نشان می دهیم اگر به ازای هر $v \in V_{m^t}$ داشته باشیم $v \in V_{n^t}$. مجموعه تمام وابستگی های یک آیتم ست مانند m^t را بصورت $D(m^t)$ نشان می دهیم.

$$D(m^t) = \{n^t | \ n^t \rightarrow m^t\} \qquad (۷)$$

تعریف ۸: مجموعه آیتم ستهای انتخابی. فرض نمایید FL^t مجموعه لینک های مفهومی پرتکرار استخراج شده از مجموعه آیتم ستهای حداکثر tتایی باشد. LI^t_{sel} (RI^t_{sel}) مجموعه ای از مجموعه آیتم ستهایی است که در ایجاد این لینک ها بکار رفته اند.

$$LI^t_{sel} = \{m;\ E_{(m,n)} \in FL^t\}$$
$$RI^t_{sel} = \{m;\ E_{(n,m)} \in FL^t\} \qquad (۸)$$

ویژگی ۳: اگر آیتم ست n^t، در هیچ یک از لینک های مفهومی پرتکرار استخراج شده در FL^t قرار نداشته باشد ($n^t \notin LI^t_{sel}(RI^t_{sel})$)، آنگاه هیچ یک از آیتم ستهایی که به آن وابسته است ($\{m^t | n^t \in D(m^t)\}$) نیز در FL^t قرار نخواهند داشت.

اثبات: فرض نمایید که m^t یکی از آیتم ستهایی باشد که به آیتم ست n^t وابسته است، و نیز فرض نمایید n^t در هیچ لینک مفهومی پرتکراری قرار نداشته باشد ($n^t \notin LI^t_{sel}(RI^t_{sel})$)، لذا بنا به تعریف لینک مفهومی پرتکرار، به ازای تمامی آیتم ستهایی مانند n_j داریم:

$$(|RE_{n^t} \cap LE_{m_j}| < \beta \times |E|)\ |LE_{n^t} \cap RE_{n_j}| < \beta \times |E|$$

از طرفی با توجه به تعریف ۶ (وابستگی) میدانیم که $V_{m^t} \subseteq V_{n^t}$ و لذا داریم:

$$|RE_{m^t}| \leq |RE_{n^t}| \text{ و } |LE_{m^t}| \leq |LE_{n^t}|,$$

در نتیجه:

$$(|RE_{m^t} \cap LE_{n_j}| < \beta \times |E|)\ |LE_{m^t} \cap RE_{n_j}| < \beta \times |E|$$

و لذا ویژگی فوق اثبات می شود.

تعریف ۹: والدهای آیتم ست. برای هر آیتم ست m^t، $(t > 1)$ دو والد آنرا با $parent1\ (m^t)$ و $parent2\ (m^t)$ ($parent1\ (m^t), parent2\ (m^t) \in I^{t-1}$) نشان می دهیم بگونه ای که $m^t = parent1\ (m^t) . parent2\ (m^t)$.

تعریف ۱۰: سطح وابستگی[۱۳]. برای هر آیتم ست m، سطح وابستگی آن را با DL(m) نشان داده و بصورت زیر تعریف می نماییم:

$$DL(m) = \begin{cases} 0 & \text{if } D(m) = \emptyset \\ \max\limits_{n \in D(m)} DL(n) + 1 & \text{else} \end{cases} \quad (۹)$$

برای درک مفهوم وابستگی و نحوه استفاده از آن در الگوریتم D-MFCLMin، در ادامه مثالی آورده شده است، و سپس شبه کد آن را ارائه خواهیم نمود.

۴-۱- تشریح مفهوم وابستگی با استفاده از مثال

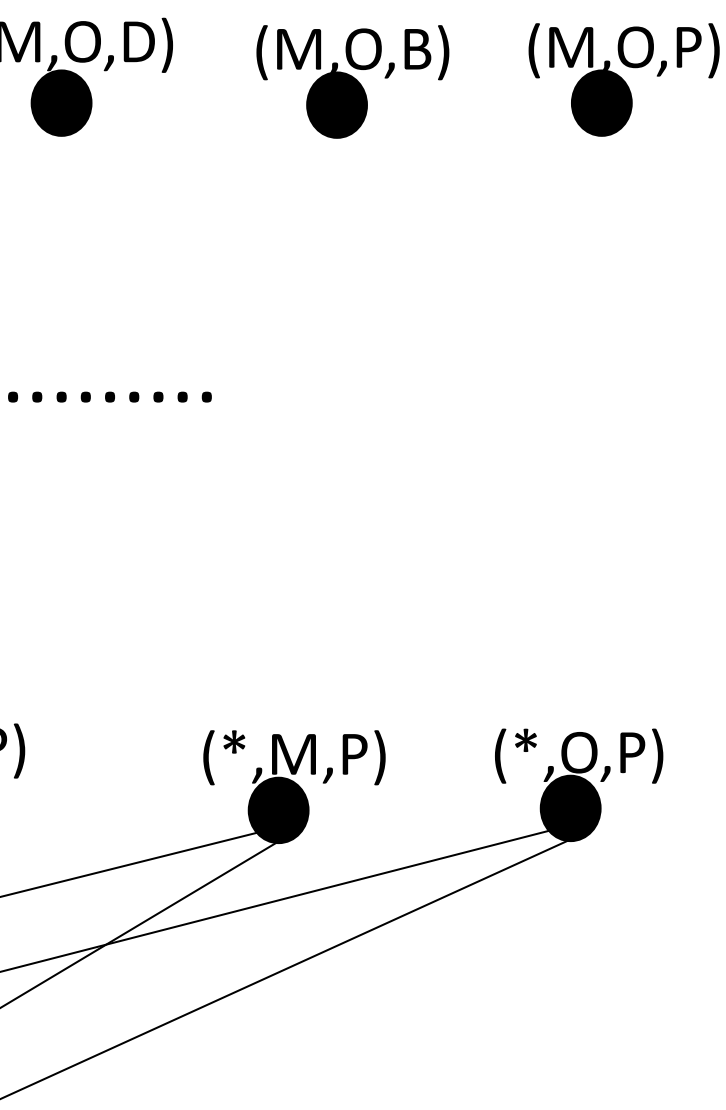
یک شبکه اجتماعی $G = (V, E)$ از افراد را در نظر بگیرید. هر فرد در این شبکه اجتماعی بصورت یک گره $v \in V$ و در قالب رابطه $R(Gender, AgeClass, Degree)$ تعریف شده است. وجود یک یال $e_{v1,v2} \in E$ نشاندهنده ی وجود ارتباط دوستی بین فرد $v1, v2$ می باشد. دامنه هر یک از صفات رابطه R بصورت زیر در نظر گرفته شده است.

$D_{Gender} = \{F, M\}$ (F: مرد, M: زن)

$D_{AgeClass} = \{T, M, O\}$(T: جوان, M: میانسال, O: مسن)

$D_{Degree} = \{D, B, P\}$(D: دیپلم, B: لیسانس, P: دکترا)

در شکل زیر، شبکه آیتم ستهای ممکن قابل ایجاد برای این رابطه نشان داده شده است.



شکل ۲: شبکه آیتم ست های ممکن برای رابطه R

[۱۳] Dependency Level

در طی اجرای الگوریتم، این شبکه بصورت سطح به سطح ایجاد می گردد. هر یک از آیتم ستهای این شبکه (بغیر از آیتم ست (*,*,*)) ممکن است در مجموعه LI (مجموعه آیتم ستهای سمت چپ) و یا RI (مجموعه آیتم ستهای سمت راست) و یا هر دو قرار گیرند. در هر مرحله از الگوریتم، سطح بعدی از این شبکه ایجاد و سپس آیتم ستهای واجدالشرایط به LI و RI افزوده شده و پس از آن، به ازای هر زوج n و m $(n \in LI, m \in RI)$ وجود لینک مفهومی (n, m) بررسی می گردد. همانگونه که مشاهده می شود تعداد آیتم ستهای ممکن با افزایش تعداد صفات رابطه و نیز تعداد مقادیر دامنه مربوط به هر صفت از رابطه، بشدت افزایش می یابد. از طرفی، افزایش تعداد گره ها در شبکه اجتماعی، سبب افزایش زمان لازم برای تعیین وجود لینک پرتکرار مفهومی بین دو آیتم ست خواهد شد.

همانگونه که پیشتر اشاره گردید، ایده الگوریتم پیشنهادی بکارگیری مفهوم وابستگی بمنظور کاهش فضای جستجوست. برای درک مفهوم وابستگی به مثال زیر توجه شود. سه آیتم ست M_1، M_2 و M_3 را در نظر بگیرید. M_1 و M_2 اعضای مجموعه آیتم ست LI و M_3 عضو مجموعه آیتم ست RI می باشد (یا بعکس). فرض کنید آیتم ست M_1 بصورت (*,M,*) – افراد میانسال، M_2 بصورت (*,F,*) –مردان، و M_3 بصورت (*,O,*)–افراد مسن باشند. همچنین فرض نمایید که وابستگی بصورت $M_2 \rightarrow M_1$ وجود داشته باشد. با توجه به مفهوم وابستگی، وجود این رابطه بدین معناست که در شبکه اجتماعی G، تمام افراد میانسال، مرد می باشند.

در الگوریتم پیشنهادی، زمانی که قصد بررسی وجود لینک مفهومی بین M_1 و M_3 را داریم، به سبب وجود وابستگی بین M_2 و M_1، ابتدا وجود لینک مفهومی بین M_2 و M_3 بررسی میشود. با توجه به ویژگی ۱، مطمئنا اگر بین M_2 و M_3 لینک مفهومی پرتکرار وجود نداشته باشد، بین M_1 و M_3 نیز وجود نخواهد داشت. بعبارت دیگر، هنگامی که مردان (M_2) با افراد مسن (M_3) رابطه ای ندارند، لذا با توجه به اینکه همه افراد میانسال، مرد می باشند $(M_2 \rightarrow M_1)$، افراد میانسال (M_1) نیز با افراد مسن رابطه ای نخواهد داشت.

همانگونه که مشاهده شد لازمه این فرآیند، این است که پیش از بررسی وجود لینک مفهومی پرتکرار بین M_1 و M_3، وجود لینک مفهومی پرتکرار بین M_2 و M_3 بررسی شده باشد. بمنظور برآورده نمودن این شرایط در الگوریتم پیشنهادی از مفهوم سطح وابستگی (تعریف ۴) استفاده گردیده است. با استفاده از سطح وابستگی، آیتم ستهایی که دارای وابستگی بیشتری می باشند، دیرتر مورد بررسی قرار خواهند گرفت. در بخش بعد، شبه کد الگوریتم پیشنهادی ارائه شده است.

۴-۲- شبه کد الگوریتم D-MFCLMin

الگوریتم D-MFCLMin با بسط الگوریتم ارائه شده در [۵] از طریق اعمال ویژگی ۳، فضای جستجو برای استخراج لینک های مفهومی پرتکرار را کاهش می دهد. شبه کد این الگوریتم در زیر آمده است. مشابه با الگوریتم H-MFCLMin، پارامترهای ورودی شامل α و β که بترتیب حد آستانه های مربوط به مجموعه آیتم ها و پشتیبانی لینک ها می باشد.

مشابه با الگوریتم H-MFCLMin در اولین گام (t=1)، آیتم ستهای تک آیتمی LI^1_{cand} (RI^1_{cand}) بنابر ویژگی های ۱ و ۲ (ویژگی های مربوط به آیتم ستهای واجدالشرایط) ایجاد می شوند (خطوط ۶ و ۷). پس از ایجاد این لیستها، آیتم ستهای آنها برحسب مقدار پشتیبانی و بصورت صعودی مرتب می شوند. برخلاف H-MFCLMin، پیش از جستجو برای یافتن لینک های مفهومی پرتکرار، وابستگی های بین آیتم ستهای موجود در $LI^t_{cand}(RI^t_{cand})$ بدست می آیند. برای این منظور، آیتم ستهایی tتایی از $LI^t_{cand}(RI^t_{cand})$ دو به دو الحاق شده و سپس براساس مقدار پشتیبانی آیتم ست حاصل، وجود وابستگی بین دو آیتم ست الحاق شده بررسی می گردد. در صورت عدم وجود وابستگی، آیتم ست الحاقی حاصل بعنوان یکی از آیتم ستهای کاندید مرحله بعد در لیست $LI^{t+1}_{cand}(RI^{t+1}_{cand})$ درج می گردد (خطوط ۲۵-۱۱). این درج به شکلی صورت می گیرد تا لیست بصورت صعودی برحسب مقدار پشتیبانی آیتم ستها مرتب باقی بماند.

پس از تعیین وابستگی ها در بین آیتم ستهای مرحله t، سطح وابستگی آنها (رابطه ۹) محاسبه و سپس در $LI^t_{cand}(RI^t_{cand})$ برحسب صعودی سطح وابستگی بدست آمده مرتب می شوند (خط ۲۶). پس از مرتب سازی، جستجو برای یافتن لینک های مفهومی پرتکرار انجام می گیرد. لینک های مفهومی پرتکرار یافت شده به لیست FL^t افزوده شده و سپس با حذف زیرلینک های مفهومی پرتکرار قرار گرفته در FL_{Vmax}، به عنوان لینک های مفهومی پرتکرار حداکثر به FL_{Vmax} افزوده می شوند (خطوط ۴۴-۲۷).

بصورت دقیقتر، این جستجو بدین صورت انجام می گیرد که به ازای هر آیتم ست $m_i \in LI_{cand}$ و $m_j \in RI_{cand}$، با این شرط که $|m_i| = t$ یا $|m_j| = t$، بررسی می شود که آیا لینک (m_i, m_j) پرتکرار است یا خیر. در الگوریتم پیشنهادی، پیش از انجام این بررسی، آیتم ستهای وابسته m_i و m_j بررسی می شوند. در صورتی که هیچ یک از آیتم ستهای وابسته آنها هنوز در FL^t افزوده نشده باشند، از بررسی پرتکرار بودن لینک بین این زوج چشم پوشی می شود (خط ۳۳). یادآوری می شود که آیتم ستها در LI^t_{cand} و RI^t_{cand} به ترتیب صعودی سطح وابستگی مرتب شده اند، لذا در هنگام بررسی یک آیتم ست، تمامی آیتم ستهایی که به آن وابسته می باشند، از پیش در این مرحله بررسی شده اند. پس از این گام، مشابه با الگوریتم H-MFCLMin بررسی پرتکرار بودن لینک مورد بررسی انجام می گیرد (خط ۳۴). در صورت پرتکرار بودن لینک (m_i, m_j)، m_i به LI^t_{sel} و m_j به RI^t_{sel} افزوده می شوند.

پس از پایان بررسی آیتم ستها در LI^t_{cand} و RI^t_{cand} برای استخراج لینک های مفهومی پرتکرار در مرحله t، آیتم ستهای LI^{t+1}_{cand} و RI^{t+1}_{cand} اصلاح می شوند. در این مرحله، هر آیتم ست m^{t+1} ($m^{t+1} \in LI^{t+1}_{cand}(RI^{t+1}_{cand})$) که هر دو آیتم ست والدش (تعریف ۸) در LI^t_{sel} (RI^t_{sel}) قرار نداشته باشند، از لیست مذکور حذف می شود(۴۹-۴۵). این حذف با توجه به ویژگی بستار زیرین (ویژگی ۲) انجام می گیرد.

Algorithm 1: D-MFCLMin Algorithm

Require: G = (V;E): Network, $\beta \in [0..1]$: Link support threshold and $\alpha \in [0..1]$: Itemset filtering threshold

1. FL_{Vmax}: Set of MFCLs $\leftarrow \emptyset$
2. LIcand: Stack of left-hand itemset candidates $\leftarrow \emptyset$
3. RIcand: Stack of right-hand itemset candidates $\leftarrow \emptyset$
4. FL^t: List of frequent conceptual links $\leftarrow \emptyset$
5. t: Iteration $\leftarrow$ 1

{Generation of the 1-itemsets}

6. $LI^1_{cand} \leftarrow$ Generate 1-itemsets m from V such as $|V_m| > \alpha$ and $|LE_m| > \beta \times |E|$
7. $RI^1_{cand} \leftarrow$ Generate 1-itemsets m from V such as $|V_m| > \alpha$ and $|RE_m| > \beta \times |E|$
8. Sort LI^1_{cand} , RI^1_{cand} itemsets by their Supports
9. $t \leftarrow 1$
10. do

{Determining Dependencies between $LI^t_{cand}(RI^t_{cand})$ itemsets}

11. for all itemset $m_i^t \in LI^t_{cand}(RI^t_{cand})$ do
12. for all itemset $m_j^t \in LI^t_{cand}(RI^t_{cand})$ do
13. if (m_i^t and m_j^t share $t-1$ item)
14. $m_k^{t+1} \leftarrow$ join m_i^t and m_j^t
15. if ($sup(m_k^{t+1}) = sup(m_i^t)$)
16. add m_j^t to $D(m_i^t)$
17. else
18. if ($|V_{m_k^{t+1}}| > \alpha$ and $\left|LE_{m_k^{t+1}}\right| > \beta \times |E|$ ($\left|RE_{m_k^{t+1}}\right| > \beta \times |E|$))
19. add m_k^{t+1} to $LI^{t+1}_{cand}(RI^{t+1}_{cand})$
20. parent1 (m_k^{t+1}) $\leftarrow m_i^t$
21. parent2 (m_k^{t+1}) $\leftarrow m_j^t$
22. end if
23. end if
24. end for
25. end for
26. Sort $LI^t_{cand}(RI^t_{cand})$ itemsets by their calculated dependency level

{Generation of frequent conceptual links}

27. $FL^t \leftarrow \emptyset$
28. $L^t_{sel} \leftarrow \emptyset$
29. $R^t_{sel} \leftarrow \emptyset$
30. for all itemset $m_i \in LI_{cand}$ do
31. for all itemset $m_j \in RI_{cand}$ do
32. if ($|m_i| = t$ or $|m_j| = t$)
33. if ($\exists(m_k, m_j) \in FL^t, \forall m_k \in D(m_i)$ and $\exists(m_i, m_k) \in FL^t, \forall m_k \in D(m_j)$)
34. if ($\exists l \in FL^t$ such as $(m_i, m_j) \subset l$ and $|(m_i, m_j)| > \beta \times |E|$)
35. add (m_i, m_j) to FL^t
36. remove all $q \in FL_{Vmax}$ such as $q \subset (m_i, m_j)$
37. add (m_i, m_j) to FL_{Vmax}
38. add m_i to L^t_{sel}
39. add m_j to R^t_{sel}
40. end if
41 end if
42. end if
43. end for
44. end for
45. for all itemset $m_i \in LI^{t+1}_{cand}(RI^{t+1}_{cand})$ do
46. if ($parent1(m_i) \notin L^t_{sel}(R^t_{sel})$ and $parent2(m_i) \notin L^t_{sel}(R^t_{sel})$)
47. remove m_i from $LI^{t+1}_{cand}(RI^{t+1}_{cand})$
48. end if

49. end for
50. $t \leftarrow t + 1$
51. while $FL^t \neq \emptyset$ and allCombinations() = false
52. return FLVmax

۴-۳- تحلیل روش پیشنهادی

ابتدا به بررسی هزینه ی مربوط به الگوریتم H-MFCLMin می پردازیم. فرض کنید بخواهیم وجود لینک مفهومی بین دو آیتم ست m_1^i و m_2^j ($i = t$ or $j = t$) را در مرحله t بررسی نماییم ($m_1^i \in LI_{cand}^t, m_2^j \in RI_{cand}^t$). برای این منظور، یال هایی از شبکه که گره مبداء آن متعلق به m_1^i و گره مقصد آن متعلق به m_2^j باشند شمارش خواهند شد، هزینه این بررسی بصورت زیر بدست می آید:

$$C(m_1^i, m_2^j) = 2.N.|E| \qquad (۱۰)$$

در رابطه فوق، N تعداد ویژگی های هر آیتم ست (که برابر با تعداد صفات رابطه ی R می باشد) است. برای بررسی تعلق یک گره به یک آیتم ست کافیست مقادیر ویژگیهای گره با آیتم ست مقایسه گردد که هزینه ی N را دربرخواهد داشت و از طرفی از آنجا که این عمل بایستی برای گره مبداء و نیز گره مقصد هر یک از یالها انجام گیرد، ۲ برابر این هزینه تحمیل خواهد شد.

در الگوریتم D-MFCLMin با در نظر گرفتن وابستگی ها، هزینه فوق بصورت زیر تغییر می نماید:

$$C(m_1^i, m_2^j) = C_d + (1 - p)(2.N.|E|) \qquad (۱۱)$$

در رابطه فوق، C_d هزینه بررسی وابستگیهای مربوط به دو مجموعه آیتم m_1^i و m_2^j، و p احتمال این است که وابستگی های مربوط به این دو آیتم ست، سبب جلوگیری از شمارش یالهای شبکه اجتماعی برای بررسی وجود لینک مفهومی بین آنها گردند.

مقدار C_d، به تعداد وابستگی های آیتم ستهای در حال بررسی و نیز تعداد لینکهای مفهومی یافت شده در مرحله مورد نظر بستگی دارد. در الگوریتم D-MFCLMin، به ازای هر زوج آیتم ست در حال بررسی، مشارکت آیتم ستهای وابسته آنها در لینک های مفهومی ای که تاکنون در مرحله جاری یافت شده اند بررسی میگردد، لذا این هزینه بصورت زیر می باشد.

$$C_d = \left(|D(m_1^i)| + |D(m_2^j)|\right)|FL^t| \qquad (۱۲)$$

لذا در ادامه مقدار دو عامل وابستگی های آیتم ست و لینکهای مفهومی یافت شده را بررسی می نماییم.

۴-۳-۱. تعداد وابستگی های یک آیتم ست

امکان تعیین دقیق تعداد وابستگی های یک آیتم ست وجود ندارد لذا حداکثر تعداد آنها را در نظر خواهیم گرفت. برای سادگی، فرض می نماییم که تعداد آیتم ستها در مرحله ی t، در LI_{cand}^t و RI_{cand}^t برابر باشد. با توجه به این فرض، در ادامه این متن تفاوتی بین این دو مجموعه قائل نشده و لذا با هدف کوته نویسی از I^t بجای این دو مجموعه استفاده می نماییم. همانگونه که قبلا اشاره گردید، آیتم ستها در هر مرحله بصورت صعودی براساس پشتیبانی مرتب شده اند. با توجه به فرض وجود حداکثر ممکن وابستگی ها در مجموعه I^t، اولین آیتم ست به هیچ آیتم ست دیگری وابسته نخواهد بود، دومین آیتم ست، تنها ممکن است به اولین آیتم ست وابسته باشد، سومین وابستگی حداکثر به دو آیتم ست قبلی وابسته خواهد بود، و به همین صورت؛ لذا حداکثر تعداد وابستگی های موجود بین تمامی آیتم ستها در مجموعه I^t برابر با

$$\frac{|I^t|(|I^t| - 1)}{2} \qquad (۱۳)$$

خواهد بود که با در نظرگرفتن توزیع یکنواخت این وابستگی در بین آیتم ستهای این مجموعه، حداکثر تعداد وابستگی ها برای هر آیتم ست بصورت

$$|D(m^t)| = \frac{|I^t| - 1}{2} \qquad (۱۴)$$

بدست می آید. لازم به ذکرست که حداکثر تعداد آیتم ستها در یک مرحله از رابطه بازگشتی زیر بدست می آید:

$$|I^M| = T(N, M) = \begin{cases} \sum_{i=1}^{N} K_i & M = 1 \\ \prod_{i=1}^{N} K_i & N = M \\ \sum_{i=M}^{N} K_i . T(i-1, M-1) & \text{else} \end{cases} \qquad (۱۵)$$

در رابطه فوق K_i، تعداد مقادیر ممکن برای ویژگی iام را نشان میدهد. برای مثال در مورد ویژگی جنسیت، تعداد مقادیر ممکن برابر با ۲ است.

۴-۳-۲. تعداد لینک های مفهومی یافت شده

عامل دوم موثر در هزینه بررسی وابستگی ها، تعداد لینک های مفهومی یافت شده در یک مرحله می باشد ($|FL^t|$). با فرض رشد یکنواخت تعداد لینک های مفهومی یافت شده، حداکثر تعداد لینک های مفهومی مورد بررسی قرار گرفته به ازای هر زوج آیتم ست برابر با:

$$\frac{(2|UI^t||I^t| - |I^t|^2)^2}{2} \qquad (۱۶)$$

خواهد بود. با توجه به مقادیر فوق، تعداد لینک های مفهومی که به ازای هر زوج آیتم ست به طور میانگین بررسی می شوند برابر با

$$\frac{2|UI^t||I^t| - |I^t|^2}{2} \qquad (۱۷)$$

می باشد. با توجه به روابط (۱۴) و (۱۷)، به طور کلی مقدار C_d بصورت زیر بدست می آید:

$$C_d = \frac{2|UI^t||I^t|^2 - |I^t|^3}{2} \qquad (۱۸)$$

حال با توجه به تعیین مقدار هزینه ی وابستگی ها به تحلیل و بررسی رفتار الگوریتم پیشنهادی خواهیم پرداخت.

بدترین شرایط در الگوریتم پیشنهادی زمانی رخ می دهد که علیرغم زیاد بودن میزان وابستگی ها، هرسی نیز رخ ندهد. میزان هرس شدن بستگی به

تعداد لینک های مفهومی یافت شده دارد؛ هر قدر میزان لینک های مفهومی یافت شده کمتر باشد، افزایش میزان وابستگی ها، احتمال بیشتری در هرس نمودن آیتم ستها خواهد داشت. از طرفی، تعداد لینک های مفهومی پرتکرار به مقدار β بستگی دارد، هر چقدر مقدار این پارامتر کمتر باشد، تعداد لینکهای مفهومی بیشتری استخراج خواهد شد. لذا انتظار خواهیم داشت که الگوریتم پیشنهادی در شرایطی که β مقدار کمی داشته باشد، عملکرد ضعیفتری از خود نشان دهد.

۵-آزمایشات و نتایج

در این بخش نتایج مربوط به ارزیابی روش پیشنهادی (D-MFCLMin) ارائه می شود. روش H-MFCLMin بعنوان روش مورد استفاده برای مقایسه در نظر گرفته شده است. ابتدا در زیربخش بعد مجموعه داده مورد استفاده معرفی میشود، سپس به بررسی نتایج حاصل خواهیم پرداخت.

۵-۱- دیتاست مورد استفاده

اغلب دیتاستهای موجود در زمینه شبکه های اجتماعی تنها حاوی اطلاعات اتصالات در شبکه می باشند، از آنجا که رهیافت لینکهای مفهومی پرتکرار جزء دسته ای از روشهاست که با بکارگیری همزمان اطلاعات لینکها و صفات گره های شبکه مبادرت به استخراج دانش می نمایند، لذا نیازمند به دیتاستی می باشیم که علاوه بر اطلاعات لینکها، دارای صفات گره ها نیز باشند. علاوه براین از آنجا که در این روش، صفات گره ها بایستی همگن باشند، استفاده از دیتاستهایی که مربوط به چندین نوع موجودیت می باشند نیز ممکن نمی باشند. با توجه به این محدودیتها علیرغم وجود تعداد زیادی از دیتاستهای مربوط به شبکه های اجتماعی، محدوده ی دیتاستهای قابل استفاده بسیار محدود می گردد.

در این پژوهش دیتاست مربوط به یک شبکه اجتماعی بنام Pokec مورد استفاده قرار گرفته است [۲۳]. این دیتاست یکی از مجموعه داده های مربوط به پروژه تحلیل شبکه دانشگاه استانفورد می باشد [۲۴]. این مجموعه داده ها حاوی بیش از ۵۰ مجموعه داده شبکه ای بزرگ از ده ها هزار گره و یال تا ده ها میلیون گره و یال می باشد. این شبکه ها شامل شبکه های اجتماعی، گراف های وب، شبکه های جاده ای، شبکه های اینترنتی، شبکه های همکاری و شبکه های ارتباطی می باشد. Pokec مشهورترین شبکه اجتماعی آنلاین در کشور اسلواکی میباشد. این دیتاست شامل پروفایل تغییریافته کاربران این شبکه اجتماعی بهمراه ارتباطات دوستی بین آنها می باشد. لازم به ذکر است در شبکه اجتماعی Pokec رابطه های دوستی جهت دار می باشد. مشخصات پروفایل کاربران شامل ۵۹ فیلد می باشد که تنها ۸ فیلد آن اجباری است. در جدول (۱) مشخصات این ۸ فیلد نشان داده شده است.

پس از بررسی و انجام پالایشهای لازم، از میان این ۸ فیلد، ۵ فیلد بعنوان صفات هر کاربر در نظر گرفته شد که عبارتند از عمومی یا خصوصی بودن پروفایل، جنسیت، منطقه، سال ثبت نام و رده سنی. منظور از رده سنی، حاصل تقسیم سن اظهار شده فرد بر عدد ۱۰ می باشد که ۱۰ رده ی مختلف حاصل خواهد شد ($\frac{age}{10}$). تعداد کاربران تا سال ۲۰۱۲ (زمان استخراج داده ها) ۱۶۳۲۸۰۳و تعداد لینکها برابر با ۳۰۶۲۲۵۶۴ می باشد. با توجه به تعداد زیاد گره ها و لینکها، و از طرفی ناقص بودن بخش عمده ای از اطلاعات سایر فیلدها، در این پژوهش تنها گره هایی که بیش از ۸۰ درصد از مشخصات پروفایل آنها دارای مقدار بود در نظر گرفته شد که تعداد آنها برابر با ۳۱۲۱۱ گره و تعداد اتصالات بین این گره ها برابر با ۲۶۱۹۴۵ می باشد.

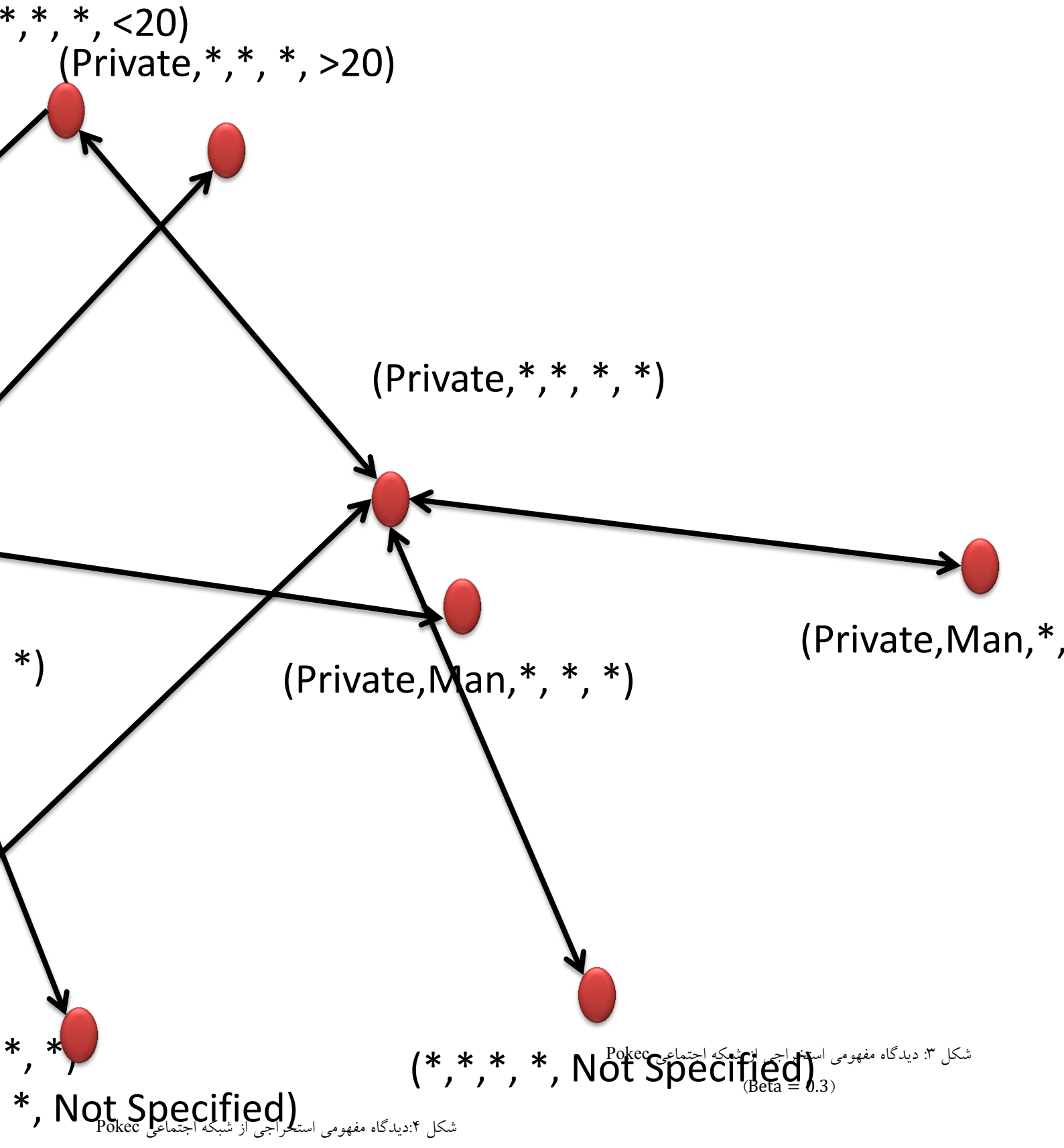
جدول (۱): مشخصات فیلدهای اجباری شبکه اجتماعی

عنوان فیلد	نوع فیلد	دامنه	توضیحات
user_id	عددصحیح	[تعداد کاربران-1]	عددی صحیح که نگاشت نام انتخابی کاربر
Public	بولین	True .. False	عمومی بودن پروفایل
Completion percentage	عدد صحیح	[1-100]	میزان دارای مقدار بودن فیلدها
Gender	بولین	True .. False	جنسیت
Region	متنی	[1-183]	منطقه ی زندگی کاربر*
last_login	datetime	۱۹۹۹ الی ۲۰۱۲	آخرین زمان لاگین کاربر
Registration	datetime	۱۹۹۹ الی ۲۰۱۲	زمان ثبت نام کاربر در سامانه
Age	عددصحیح	[1-100]	سن کاربر

*غالبا مناطقی در کشور اسلواکی اما برخی مناطق در کشورهای چک و آلمان نیز شامل میشوند

۵-۲- آزمایشات و نتایج

همانگونه که عنوان شد، بمنظور ارزیابی عملکرد روش پیشنهادی، نتایج حاصل از آنرا با نتایج حاصل از الگوریتم H-MFCLMin مقایسه نمودیم. لازم به ذکر است که خروجی های حاصل از هر دو روش مشابه می باشد بدین معنا که تفاوتی در لینکهای مفهومی پرتکرار استخراجی دو روش وجود ندارد. در شکل های زیر دیدگاه های مفهومی استخراجی از شبکه اجتماعی Pokec نمایش داده شده است. مقدار β برابر با 0.3 و 0.25 انتخاب شده است.



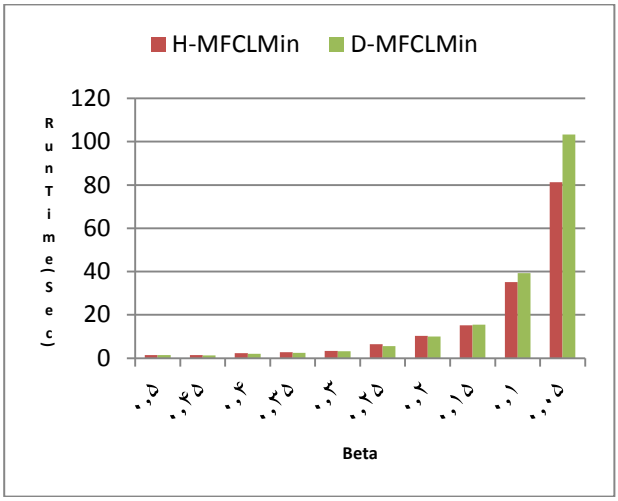
شکل ۳: دیدگاه مفهومی استخراجی از شبکه اجتماعی Pokec
(Beta = 0.3)

شکل ۴:دیدگاه مفهومی استخراجی از شبکه اجتماعی Pokec
(Beta = 0.25)

ویژگی جالب قابل مشاهده در شکل (۳) و (۴) دو طرفه بودن اغلب ارتباطات بین مجموعه آیتم هاست. در واقع اگر لینک های مفهومی بین آیتم ست A به B وجود داشته باشد، لینک مفهومی بین آیتم ست B به A نیز وجود دارد. همانگونه که قبلا اشاره شده بود، شبکه اجتماعی مذکور، جهت

دار است بدین معنی که ارتباطات دوستی یکطرفه می باشند. با این حال با خروجی بدست آمده مشخص میشود که کاربران این شبکه اجتماعی روابط دوستی دو طرفه دارند. با کاهش دادن مقدار پارمتر β تعداد مجموعه آیتم ها و لینکهای مفهومی پرتکراری بیشتری حاصل میشود و دیدگاه مفهومی بدست آمده بزرگتر می شود، لذا تنها به این ۲ مقدار بسنده خواهیم نمود.

علیرغم اینکه الگوریتم پیشنهادی (D-MFCLMin) و الگوریتم H-MFCLMin دیدگاه مفهومی یکسانی را از شبکه اجتماعی استخراج می نمایند، اما زمان صرف شده برای این کار در دو الگوریتم کمی متفاوت می باشد. در شکل (۵)، مدت زمان اجرای هر یک از این دو الگوریتم برای استخراج لینکهای مفهومی پرتکرار حداکثر از شبکه اجتماعی Pokec در مقادیر مختلف از پارامتر β نشان داده شده است. لازم به ذکرست، مقدار پارامتر α برابر با صفر در نظر گرفته شده است. هر دو الگوریتم ۱۰ بار اجرا شده و میانگین زمان اجرای بدست آمده بعنوان زمان اجرای آنها لحاظ گردیده است.



شکل ۵:زمان اجرای دو الگوریتم D-MFCLMin و H-MFCLMin در مقادیر مختلف Beta

همانگونه که مشاهده می شود، در مقادیر بالای β هر دو الگوریتم تقریبا عملکرد مشابهی را به نمایش می گذارند اما با کاهش مقدار این پارامتر، تفاوت در زمان اجرای دو الگوریتم بیشتر میشود. این تفاوت به سبب بررسی های روش پیشنهادی برای تعیین وابستگی های بین آیتم ستها می باشد.

نکته قابل ذکر این است که متاسفانه مقدار وابستگی در بین آیتم ستهای دیتاست مورد نظر صفر می باشد، لذا در واقعیت هرسی به موجب وجود وابستگی در این آزمایش صورت نگرفته است. اما در صورتی که وابستگی در بین آیتم ستها موجود باشد احتمال هرس فضای جستجو و در نتیجه تسریع در استخراج لینک های مفهومی پرتکرار حاصل خواهد گشت، و لذا میزان تفاوت در عملکرد دو الگوریتم افزایش بیشتری خواهد داشت.

۶-نتیجه گیری و کارهای آتی

فراگیر شدن استفاده از شبکه های اجتماعی، سبب حاصل آمدن حجم بسیار بالایی از اطلاعات گشته است و استخراج دانش از این اطلاعات به یکی از زمینه های مورد علاقه محققین بدل شده است. لینک های مفهومی پرتکرار یکی از رهیافتها برای استخراج دانش از این شبکه ها می باشد که علاوه بر داده های مربوط به ارتباطات، به اطلاعات مربوط به موجودیتهای این شبکه ها نیز تکیه می کند. در این نوشتار با معرفی و استفاده از مفهوم وابستگی، الگوریتم جدیدی برای تسریع در استخراج لینکهای مفهومی پرتکرار ارائه شده است. وجود وابستگی ها در بین داده ها، سبب می شود بتوان بخشی از فضای جستجو را هرس نمود و بدین ترتیب در فرآیند استخراج لینکهای مفهومی تسریع ایجاد کرد. با توجه به نبود وابستگی در دیتاست مورد استفاده، میزان این تسریع مشاهد نشد، با اینحال نتایج آزمایشات نشان داد که علیرغم عدم وجود وابستگی ها، الگوریتم پیشنهادی در مقایسه با الگوریتم H-MFCLMin عملکرد تقریبا یکسانی دارد. در نوشتار حاضر، مفهوم وابستگی به صورت قطعی استفاده گردید، در حالیکه میتوان با بسط این مفهوم به شکل وابستگی های تقریبی، هرس بیشتری از فضای جستجو را انجام داد که این امر در کارهای آتی انجام خواهد شد.

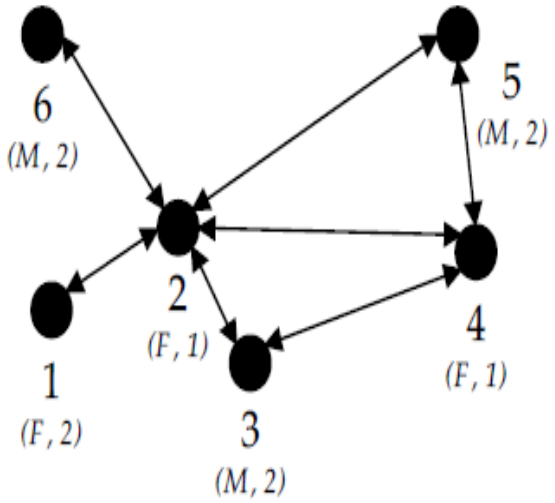
تقدیر وتشکر:

این مقاله مستخرج از طرح پژوهشی با حمایت مالی دانشگاه آزاد اسلامی واحد قوچان می باشد.

۷-منابع

1. Aggarwal, C.C., *An introduction to social network data analytics*, in *Social network data analytics*. 2011, Springer. p. 1-15.
2. West, D.B., *Introduction to graph theory*. Vol. 2. 2001: Prentice hall Upper Saddle River.
3. Tian, Y., R.A. Hankins, and J.M. Patel. *Efficient aggregation for graph summarization*. in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 2008. ACM.
4. Zhou, Y., H. Cheng, and J.X. Yu, *Graph clustering based on structural/attribute similarities.* Proceedings of the VLDB Endowment, 2009. **2**(1): p. 718-729.
5. Stattner, E. and M. Collard. *Towards a hybrid algorithm for extracting maximal frequent conceptual links in social networks*. in *IEEE 7th International Conference on Research Challenges in Information Science (RCIS)*. 2013. IEEE.

6. Stattner, E. and M. Collard. *Social-based conceptual links: Conceptual analysis applied to social networks*. in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. 2012. IEEE Computer Society.
7. Yang, G. *The complexity of mining maximal frequent itemsets and maximal frequent patterns*. in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004. ACM.
8. Stattner, E. and M. Collard, *Descriptive Modeling of Social Networks.* Procedia Computer Science, 2015. **52**: p. 226-233.
9. Xie, J. and B.K. Szymanski. *Towards linear time overlapping community detection in social networks*. in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2012. Springer.
10. Macropol, K. and A. Singh, *Scalable discovery of best clusters on large graphs.* Proceedings of the VLDB Endowment, 2010. **3**(1-2): p. 693-702.
11. Gopalan, P.K. and D.M. Blei, *Efficient discovery of overlapping communities in massive networks.* Proceedings of the National Academy of Sciences, 2013. **110**(36): p. 14534-14539.
12. Riedy, J., D.A. Bader, and H. Meyerhenke. *Scalable multi-threaded community detection in social networks*. in *Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012 IEEE 26th International*. 2012. IEEE.
13. Ovelgönne, M. and A. Geyer-Schulz, *An ensemble learning strategy for graph clustering.* Graph Partitioning and Graph Clustering, 2012. **588**: p. 187.
14. De Meo, P., et al., *Mixing local and global information for community detection in large networks.* Journal of Computer and System Sciences, 2014. **80**(1): p. 72-87.
15. Chen, J. and Y. Saad, *Dense subgraph extraction with application to community detection.* IEEE Transactions on Knowledge and Data Engineering, 2012. **24**(7): p. 1216-1230.
16. Jiang, P. and M. Singh, *SPICi: a fast clustering algorithm for large biological networks.* Bioinformatics, 2010. **26**(8): p. 1105-1111.
17. Yang, J., J. McAuley, and J. Leskovec. *Community detection in networks with node attributes*. in *2013 IEEE 13th International Conference on Data Mining*. 2013. IEEE.
18. Getoor, L. and C.P. Diehl, *Link mining: a survey.* ACM SIGKDD Explorations Newsletter, 2005. **7**(2): p. 3-12.
19. Agrawal, R. and R. Srikant. *Fast algorithms for mining association rules*. in *Proc. 20th int. conf. very large data bases, VLDB*. 1994.
20. Han, J., et al., *Mining frequent patterns without candidate generation: A frequent-pattern tree approach.* Data mining and knowledge discovery, 2004. **8**(1): p. 53-87.
21. Stattner, E. and M. Collard. *Flmin: An approach for mining frequent links in social networks*. in *International conference on networked digital technologies*. 2012. Springer.
22. Stattner, E. and M. Collard. *Max-flmin: An approach for mining maximal frequent links and generating semantical structures from social networks*. in *International Conference on Database and Expert Systems Applications*. 2012. Springer.
23. Takac, L. and M. Zabovsky. *Data analysis in public social networks*. in *International Scientific Conference and International Workshop Present Day Trends of Innovations*. 2012.
24. Leskovec, J. and A. Krevl, *{SNAP Datasets}:{Stanford} Large Network Dataset Collection.* 2015.



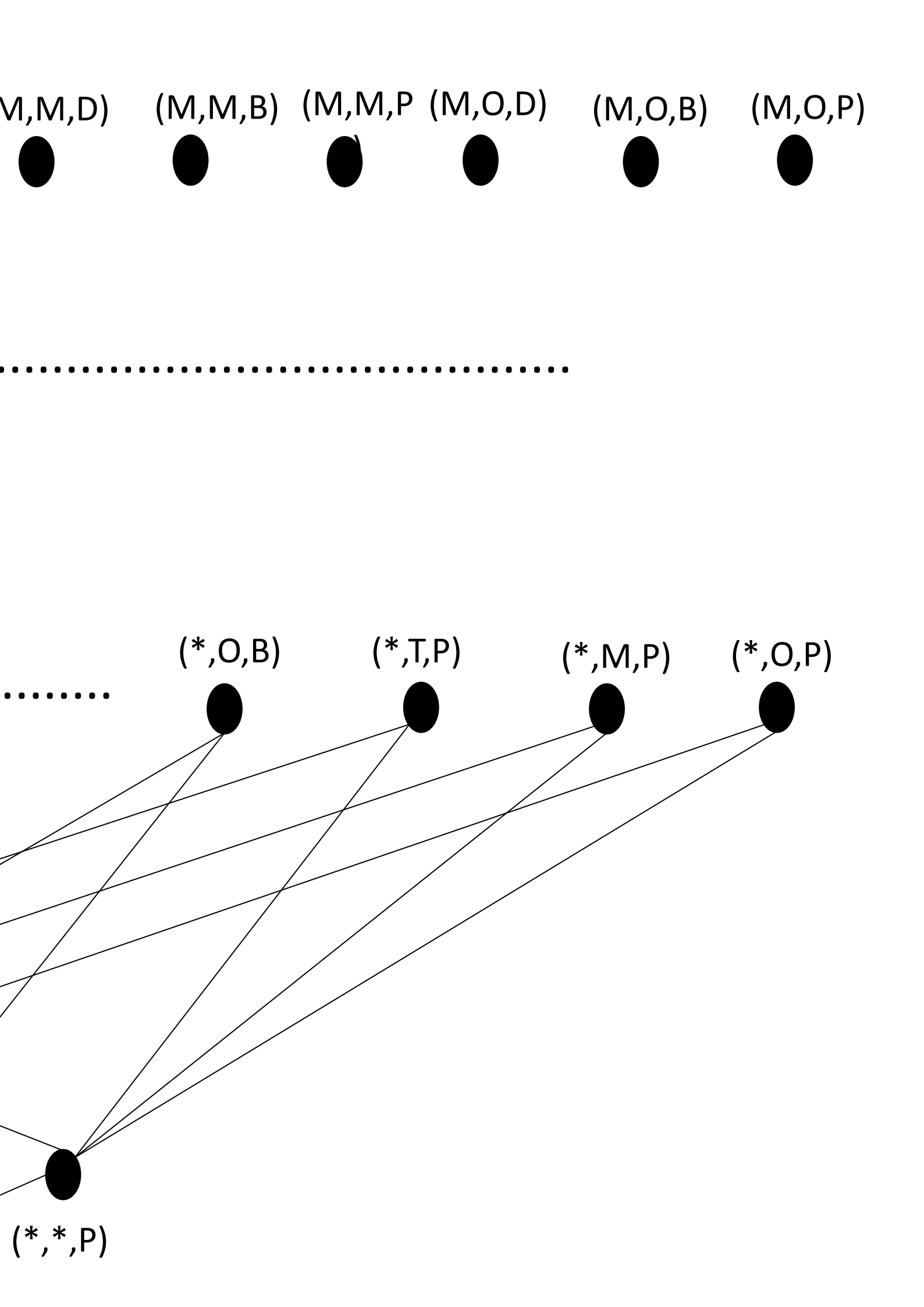
(الف)

L_1	Support
((M , *) , (F , *))	0.357
((M , *) , (* , 1))	0.357
((F , *) , (* , 1))	0.214
((F , *) , (* , 2))	0.428
…	…

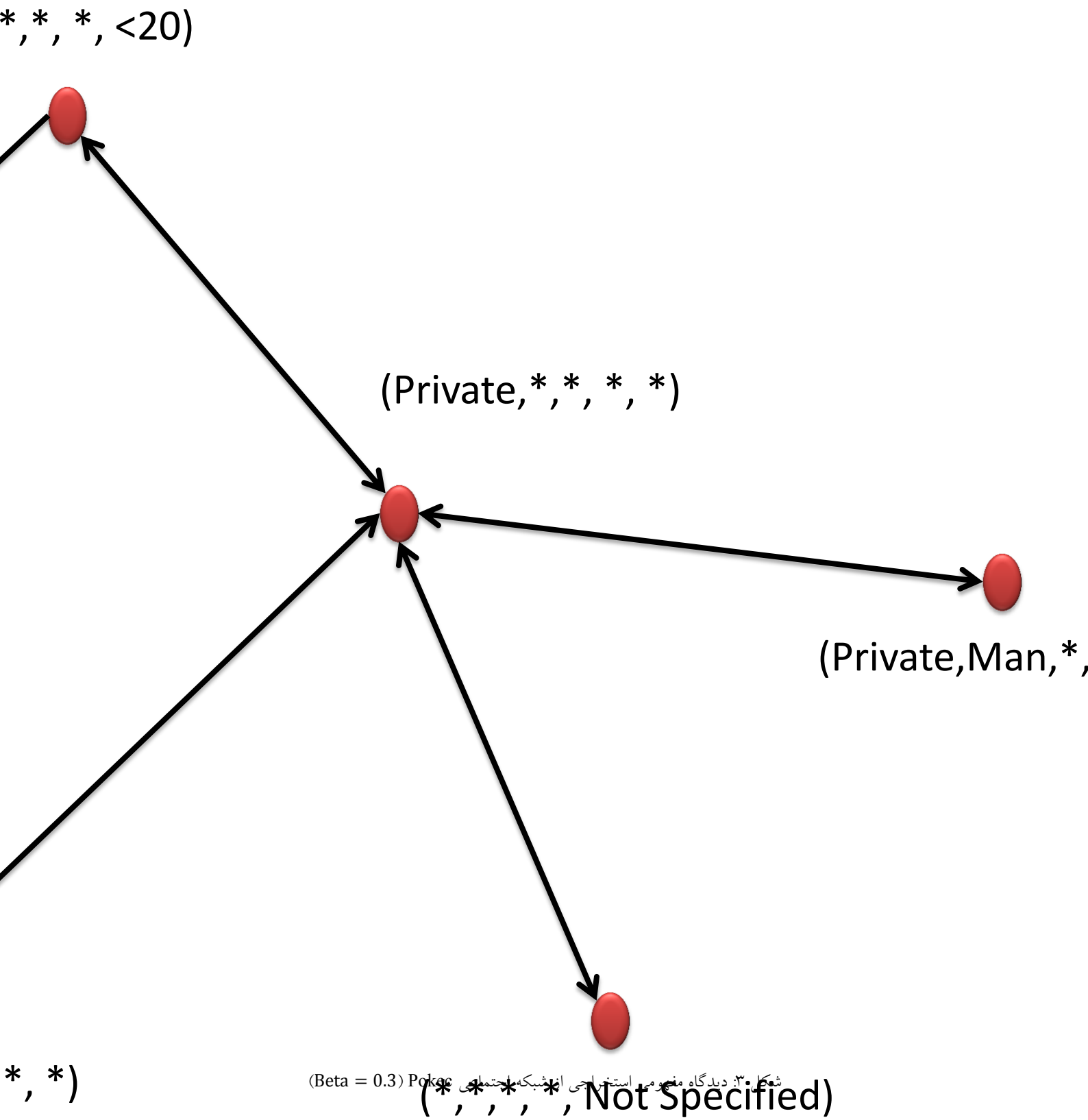
L_2	Support
((M , 2) , (F , *))	0.357
((M , 2) , (F , 1))	0.357
((F , 1) , (* , 2))	0.428
((F , 1) , (M , *))	0.357
…	…

(ب)

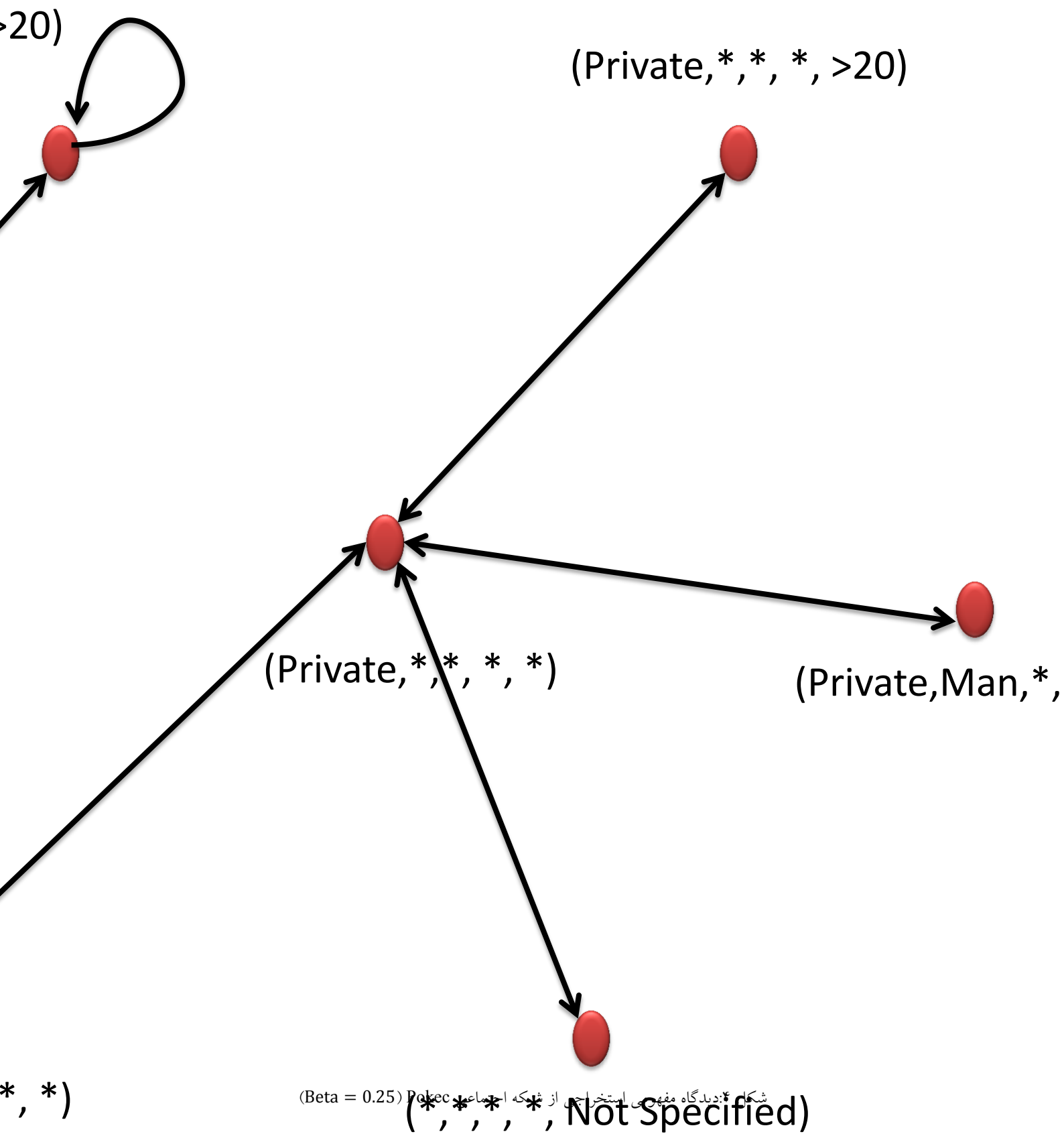
شکل ۱. نمونه ای از استخراج لینکهای مفهومی توسط الگوریتم FLMIN [21]


M,M,D)
(M,M,B)
(M,M,P
(M,O,D)
(M,O,B)
(M,O,P)
(*,O,B)
(*,T,P)
(*,M,P)
(*,O,P)
(*,*,P)

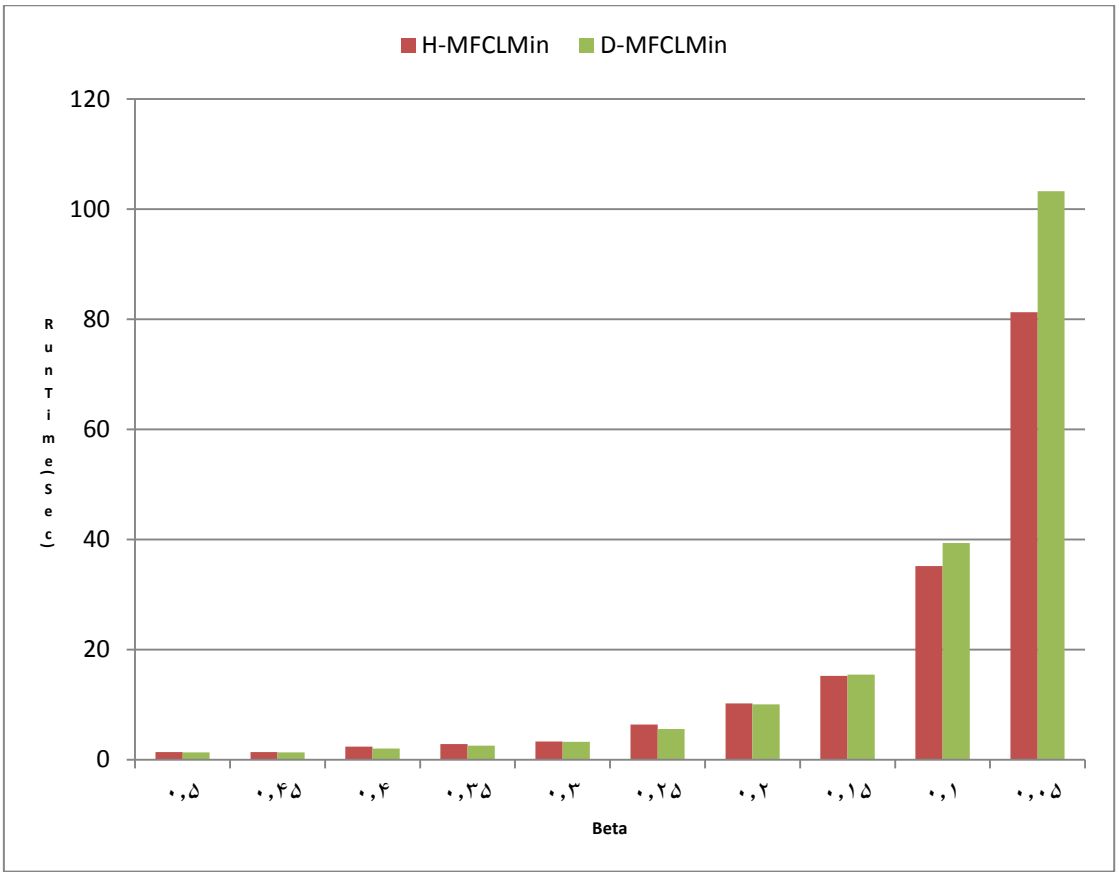
شکل ۲: شبکه آیتم ست های ممکن برای رابطه R



شکل ۳۰: دیدگاه منحوس استخراجی از شبکه اجتماعی (Beta = 0.3) Pokec



شکل ۴:دیدگاه مفهومی استخراجی از شبکه اجتماعی Pokec (Beta = 0.25)



شکل ۵:زمان اجرای دو الگوریتم D-MFCLMin و H-MFCLMin در مقادیر مختلف Beta