به نام خداوند بخشنده مهربان

شبکه های اجتماعی و اقتصادی

دانشکده مهندسی کامپیوتر

مریم رمضانی
بهار ۱۴۰۴

سیستم های توصیه گر، گراف، شبکه | **تمرین چهارم** | تاریخ انتشار: ۱۰ خرداد ۱۴۰۴

۱. سوالات خود در مورد این تمرین را در کوئرا مطرح کنید.

۲. لطفا پاسخ خود را با توضیحات کافی ارائه دهید.

سوالات تئوری (۸۰ نمره)

تاریخ تحویل: ۲۳ خرداد ۱۴۰۴

پرسش ۱ (۲۰ نمره)

سوال: تحلیل گراف‌های ناهمگن و کاربرد Meta-Path در خوشه‌بندی

مدل‌سازی گراف‌های ناهمگن:

گرافی ناهمگن تعریف کنید که شامل نودهای متفاوت مانند مقاله‌ها، ژورنال‌ها و نویسندگان باشد. همچنین گراف دیگری شامل فیلم‌ها، کارگردان‌ها و بازیگران ایجاد نمایید. در هر دو گراف، نوع نودها و نوع یال‌ها (روابط میان نودها) را مشخص کنید.

تعریف Meta-Path:

مفهوم Meta-Path را در گراف‌های ناهمگن توضیح دهید. چرا این مفهوم برای تحلیل رابطه‌های غیر مستقیم بین نودها مهم است؟

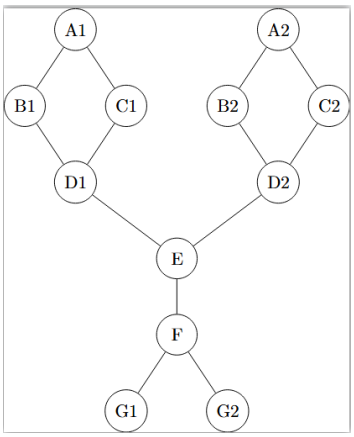
کاربرد Meta-Path در مدل‌سازی گراف:

نحوه‌ی استفاده از Meta-Pathها برای مدل‌سازی روابط میان نودهای هم‌نوع (مثلاً نویسنده با نویسنده یا فیلم با فیلم) را توضیح دهید. چگونه می‌توان از این مدل‌سازی برای اعمال الگوریتم‌های Community Detection استفاده کرد؟

طراحی Meta-Path برای گراف فیلم‌ها: حداقل سه Meta-Path مختلف پیشنهاد دهید که برای خوشه‌بندی (Clustering) فیلم‌ها مفید باشند. برای هر Meta-Path توضیح دهید چه نوع ارتباطی میان فیلم‌ها برقرار می‌شود و چرا برای کشف جامعه‌های معنایی مناسب است.

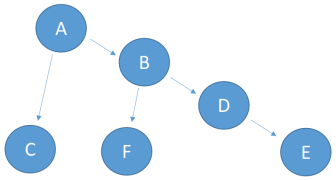
تعریف Meta-Path برای گراف مقالات جهت خوشه‌بندی نویسندگان: چگونه می‌توان از Meta-Pathها برای تعریف شباهت میان نویسندگان استفاده کرد؟ چند نمونه Meta-Path ارائه دهید که خوشه‌بندی نویسندگان بر اساس آن‌ها معنادار باشد (مثلاً همکاری مشترک، انتشار در ژورنال مشابه و ...).

پرسش ۲ (۲۰ نمره)



اجرای الگوریتم‌های Node2Vec، GraphSAGE و GAT را برای گراف داده شده به صورت گام‌به‌گام تا سه مرحله توضیح دهید. بر اساس پیچیدگی روابط بین رئوس، اندازه گراف و زمانی بودن توضیح دهید در هر حالت کدام الگوریتم مناسب‌تر است و چرا.

پرسش ۳ (۲۰ نمره)



یال	احتمال
$A \to B$	۰٫۹
$A \to C$	۰٫۴
$B \to D$	۰٫۶
$B \to F$	۰٫۳
$D \to E$	۰٫۵

مرکزیت در فضای تعبیه‌شده (Embedding Space) در برابر مرکزیت ساختاری

هر گره در یک فضای دوبعدی نهفته تعبیه شده که از طریق راهپیمایی‌های بایاس‌دار (Node2Vec) به دست آمده است. مختصات گره‌ها به صورت زیر است:

گره	(x, y)
A	(9.0, 1.0)
B	(8.0, 2.0)
C	(7.0, 0.0)
D	(6.0, 3.0)
E	(5.0, 4.0)
F	(7.5, 2.5)

آ) فاصله اقلیدسی گره F تا سایر گره‌ها را محاسبه کرده و میانگین فاصله‌ها را گزارش دهید.

ب) آیا مرکزیت در فضای تعبیه‌شده لزوماً با پتانسیل تأثیرگذاری در گراف ساختاری همبستگی دارد؟ با ذکر مثال مشخص از این گراف توضیح دهید.

تحلیل انتشار تأثیر (مدل IC)

آ) با استفاده از مدل انتشار مستقل (IC) تعداد مورد انتظار گره‌های فعال شده با شروع از گره A را محاسبه کنید. فرض کنید فرایند انتشار در حداکثر سه مرحله انجام می‌شود. فرمول $\text{Expected}(X \longrightarrow Y) = P(X \text{ activates } Y) \times P(X \text{ is active})$ همچنین موارد زیر را نیز محاسبه و تحلیل کنید:

- احتمال فعال شدن گره E
- تفاوت در انتظار تأثیر زمانی که گره B به جای A به‌عنوان seed انتخاب می‌شود

ب) توضیح دهید که چرا از الگوریتم‌های حریصانه در بیشینه‌سازی تأثیر استفاده می‌شود و به صورت خلاصه نشان دهید که تابع تأثیر زیرمجموعه‌گرا (submodular) است.

دینامیک – GRL همرخدادی متنی در Node2Vec

فرض کنید تعبیه‌های فوق با استفاده از الگوریتم Node2Vec به‌دست آمده‌اند با تنظیمات زیر: طول راهپیمایی: ۳، دو راهپیمایی برای هر گره، پارامتر بازگشت $p = ۱$، پارامتر in-out برابر $q = ۰/۲۵$ (تمایل به اکتشاف)

آ) دو راهپیمایی نمونه از گره A را شبیه‌سازی کنید. وزن انتخاب‌ها و مسیرها را مشخص کنید. مثال: $A \to B \to F \to B$

ب) بر اساس راهپیمایی‌ها، ماتریس همرخدادی بین گره‌ها را بسازید. این ماتریس چه تأثیری بر شباهت نقطه‌ای بین A و سایر گره‌ها در آموزش skip-gram دارد؟

ج) توضیح دهید که چگونه تغییر پارامتر q در گراف‌های پراکنده و متراکم تعبیه‌ها را تحت تأثیر قرار می‌دهد. این مسئله چه ارتباطی با سیستم‌های توصیه‌گر حساس به تأثیر دارد؟

ملاحظات راهبردی و نمایش چندوجهی

آ) محدودیت‌های استفاده از تعبیه‌های ایستا (مانند Node2Vec یا DeepWalk) در فرآیندهای پویا مثل انتشار تأثیر را توضیح دهید. یک روش جایگزین پیشنهاد دهید (مانند GRL زمانی یا مدل‌های مبتنی بر توجه.)

ب) در گراف‌های چندوجهی که تأثیر تنها ساختاری نیست (مثلاً وابسته به محتوا یا زمان یا اعتماد اجتماعی است)، چگونه می‌توان معماری‌های GRL را گسترش داد؟ یک اسکچ از نحوه انتقال پیام در GNN را که شامل احتمال تأثیر است، ارائه دهید.

به تمامی پرسش‌ها با استدلال دقیق پاسخ دهید. در صورت نیاز به مفاهیم نظری مانند زیرمجموعه‌گرایی، توزیع پایدار راهپیمایی تصادفی یا گرادیان‌های تابع هزینه skip-gram ارجاع دهید.

پرسش ۹ (۲۰ نمره) سیستم‌های پیشنهاددهنده

یک گراف دو بخشی user-item را در نظر بگیرید که در آن هر یال بین کاربر U و آیتم I نشان‌دهنده‌ی این است که کاربر U آیتم I را پسندیده است. همچنین ماتریس امتیازدهی را برای این مجموعه از کاربران و آیتم‌ها با R نمایش می‌دهیم، که در آن هر سطر از R مربوط به یک کاربر و هر ستون مربوط به یک آیتم است. اگر کاربر i آیتم j را پسندیده باشد، آنگاه $R_{i,j} = 1$، در غیر این صورت $R_{i,j} = 0$. همچنین فرض می‌کنیم m کاربر و n آیتم داریم، بنابراین ماتریس R ابعادی برابر با $m \times n$ خواهد داشت.

ماتریس P با ابعاد $m \times m$ را تعریف می‌کنیم به عنوان یک ماتریس قطری که در آن عنصر قطری iام برابر است با درجه‌ی رأس کاربر i، یعنی تعداد آیتم‌هایی که کاربر i آن‌ها را پسندیده است. به‌صورت مشابه، ماتریس Q با ابعاد $n \times n$ یک ماتریس قطری است که در آن عنصر قطری iام برابر است با درجه‌ی رأس آیتم i، یا به عبارتی تعداد کاربرانی که آیتم i را پسندیده‌اند. برای مثال به شکل زیر توجه کنید.

الف)

ماتریس شباهت کاربر به صورت غیر نرمال شده را به صورت $T = RR^T$ تعریف می‌کنیم (یعنی حاصل‌ضرب ماتریس R و ترانهاده‌ی آن). مفهوم عناصر T_{ij} (برای $i \neq j$) را از منظر ساختارهای گراف دو بخشی (مانند درجه‌ی رأس‌ها، مسیر میان رأس‌ها و ...) توضیح دهید (به شکل ۱ مراجعه کنید).

شباهت کسینوسی:

یادآوری می‌کنیم که شباهت کسینوسی بین دو بردار $\mathbf{u}$ و $\mathbf{v}$ به صورت زیر تعریف می‌شود:

$$\cos\text{-sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}$$

ب)

ماتریس **شباهت آیتم** را به صورت S_I با ابعاد $n \times n$ تعریف می‌کنیم، به‌طوری‌که عنصر سطر i و ستون j برابر است با شباهت کسینوسی بین **آیتم** i و **آیتم** j که متناظر با ستون‌های i و j از ماتریس R هستند. نشان دهید که:

$$S_I = Q^{-۱/۲} R^T R Q^{-۱/۲}$$

که در آن $Q^{-1/2}$ به صورت $1/\sqrt{Q_{rc}} = Q^{-1/2}$ برای تمام درایه‌های غیر صفر تعریف شده است و برای سایر درایه‌ها برابر صفر است.

همین سؤال را برای **ماتریس شباهت کاربران**، یعنی S_U تکرار کنید؛ به‌طوری‌که عنصر سطر i و ستون j برابر است با شباهت کسینوسی بین **کاربر** i و **کاربر** j که متناظر با سطرهای i و j از ماتریس R هستند. عبارت S_U را نیز به صورت یک عملیات ماتریسی بر حسب R، P و Q بیان کنید؛ توجه داشته باشید که نباید به صورت درایه به درایه S_U را تعریف نمایید.

$$\mathbf{R} = \begin{bmatrix} ۱ & ۰ & ۱ \\ ۱ & ۰ & ۰ \\ ۱ & ۱ & ۱ \\ ۰ & ۰ & ۱ \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} ۲ & ۰ & ۰ & ۰ \\ ۰ & ۱ & ۰ & ۰ \\ ۰ & ۰ & ۳ & ۰ \\ ۰ & ۰ & ۰ & ۱ \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} ۳ & ۰ & ۰ \\ ۰ & ۱ & ۰ \\ ۰ & ۰ & ۳ \end{bmatrix}$$

شکل ۱: گراف دو بخشی کاربر–آیتم.

پاسخ شما باید نحوه‌ی استخراج این رابطه‌ها را نشان دهد. (نکته: برای جذر درایه‌به‌درایه‌ی یک ماتریس، می‌توان آن را به توان ۱/۲ نوشت.)

ج)

روش پیشنهادی بر اساس فیلترینگ مشارکتی کاربر–کاربر برای کاربر u به این صورت است: برای تمام آیتم‌ها s محاسبه کنید

$$r_{u,s} = \sum_{x \in \text{کاربران}} \cos(\text{sim}(x, u)) \cdot R_{xs}$$

و سپس k آیتمی را پیشنهاد دهید که مقدار $r_{u,s}$ در آن‌ها بیشینه است.

به‌صورت مشابه، روش پیشنهادی بر اساس فیلترینگ مشارکتی آیتم–آیتم برای کاربر u به این صورت است: برای تمام آیتم‌ها s محاسبه کنید

$$r_{u,s} = \sum_{x \in \text{آیتم‌ها}} R_{ux} \cdot \cos(\text{sim}(x, s))$$

و سپس k آیتمی را پیشنهاد دهید که مقدار $r_{u,s}$ در آن‌ها بیشینه است.

ماتریس پیشنهادی Γ را تعریف می‌کنیم که ابعادی برابر با $m \times n$ دارد، به‌طوری‌که $\Gamma(i, j) = r_{i,j}$. ماتریس Γ را برای هر دو روش **فیلترینگ کاربر–کاربر** و **فیلترینگ آیتم–آیتم** برحسب R، P و Q بیابید. پاسخ نهایی شما باید شامل عملیات در سطح ماتریسی باشد و نه تعریف درایه به درایه.

راهنمایی: در حالت آیتم–آیتم داریم:

$$\Gamma = RQ^{-۱/۲} R^T.$$

پاسخ شما باید روند رسیدن به این عبارت‌ها را (حتی در مورد آیتم–آیتم که فرمول نهایی داده شده) توضیح دهد.

د)

جدول امتیازدهی زیر را بین پنج کاربر و شش آیتم در نظر بگیرید:

شناسه آیتم	۱	۲	۳	۴	۵	۶
۱	۵	۶	۷	۴	۳	؟
۲	۴	؟	۳	؟	۵	۴
۳	؟	۳	۴	۱	۱	؟
۴	۷	۴	۳	۶	؟	۴
۵	۱	؟	۳	۲	۲	۵

جدول ۱: امتیازدهی پنج کاربر (ردیف‌ها) به شش آیتم (ستون‌ها).

(د_۱) مقادیر نامشخص امتیازدهی کاربر ۲ را با استفاده از الگوریتم فیلترینگ مشارکتی مبتنی بر کاربر پیش‌بینی کنید. از ضریب همبستگی پیرسون با میانگین‌گیری استفاده کنید. فرض کنید اندازه گروه همتایان حداکثر ۲ است و همبستگی‌های منفی را حذف نمایید.

(د_۲) مقادیر نامشخص امتیازدهی کاربر ۲ را با استفاده از الگوریتم فیلترینگ مشارکتی مبتنی بر آیتم پیش‌بینی کنید. از شباهت کسینوسی تعدیل‌شده استفاده کنید. فرض کنید اندازه گروه همتایان حداکثر ۲ است و همبستگی‌های منفی را حذف نمایید.

اکنون، یک سامانه پیشنهاددهنده فیلم را در نظر بگیرید که در آن فیلم‌ها با ژانرها مرتبط هستند و امتیازدهی یک کاربر خاص مشخص است:

شناسه فیلم	کمدی	درام	عاشقانه	هیجانی	اکشن	ترسناک	دوست داشتن یا نداشتن
۱	۱	۰	۱	۰	۰	۰	دوست ندارد
۲	۱	۱	۱	۰	۱	۰	دوست ندارد
۳	۱	۱	۰	۰	۰	۰	دوست ندارد
۴	۰	۰	۰	۱	۱	۰	دوست دارد
۵	۰	۱	۰	۱	۱	۱	دوست دارد
۶	۰	۰	۰	۰	۱	۱	دوست دارد
آزمون_۱	۰	۰	۰	۱	۰	۱	؟
آزمون_۲	۰	۱	۱	۰	۰	۰	؟

جدول ۲: رابطه ژانر_فیلم و بازخورد کاربر در قالب دوست داشتن یا نداشتن.

تمام قوانین انجمنی را با حداقل پشتیبانی ۳۳٪ و اطمینان ۷۵٪ استخراج کنید. بر اساس این قوانین، آیا آیتم **آزمون_۱** یا **آزمون_۲** را به کاربر پیشنهاد می‌کنید؟

موارد مورد نیاز برای تحویل:

(i) تفسیر T_{ii} و T_{ij}.

(ii) بیان S_I و S_U برحسب ماتریس‌های R، P و Q به همراه توضیح.

(iii) بیان Γ برحسب R، P و Q به همراه توضیح.

(iv) محاسبات مبتنی بر تکنیک‌های فیلترینگ مشارکتی.

(v) پاسخ به سوالات.

سوالات عملی (۴۰ نمره)

تاریخ تحویل:

پرسش ۱ (۲۰ نمره) به فایل جوپیتر Q1.ipynb مراجعه کنید.

پرسش ۲ (۲۰ نمره) به فایل جوپیتر Q2.ipynb مراجعه کنید.