به نام خدا

گزارش مسئله دوم

مسابقه مهارت سنجی فن‌آورد

بخش داده کاوی

پاییز ۹۵

مختصری از مطالب پیش رو:

- طرح مساله
- پیش پردازش
- نمودار گراف اولیه
- روش پیشنهادی
- معیار های ارزیابی
- نتایج
- گراف اجتماعات حاصل

طرح مسئله:

در این مسئله، تعداد ۲۸۶۵۷ ردیف داده از ارتباطات (تراکنش) افراد در یک شبکه اجتماعی داده شده است. در این دادگان هر ردیف داده شامل id دو فرد و وزن ارتباط آنها میباشد، میتوان برای این دادگان یک گراف وزن دار بی جهت (با توجه به غیر تکراری بودن داده ها) به اصطلاح گراف شبکه پیچیده (Complex Network) در نظر گرفت به این صورت که افراد رئوس و ارتباطات یال های گراف باشند.یا توجه به مجموع وزن های یال های گراف میتوان این گراف را به مجموعه های گسسته از راس ها خوشه‌بندی کرد به نحوی که مجموع این رئوس گراف اصلی را تشکیل میدهد و به این ترتیب میتوان اجتماعات کاربران این شبکه را مدل سازی کرد.

پس با یک مساله خوشه‌بندی گراف(Graph Clustering) از نوع درون گراف (Within-graph) و مشخصا مساله Community Structure Detection روبرو هستیم.

از روش های مرسوم برای حل این مساله میتوان به روش های زیر اشاره کرد:

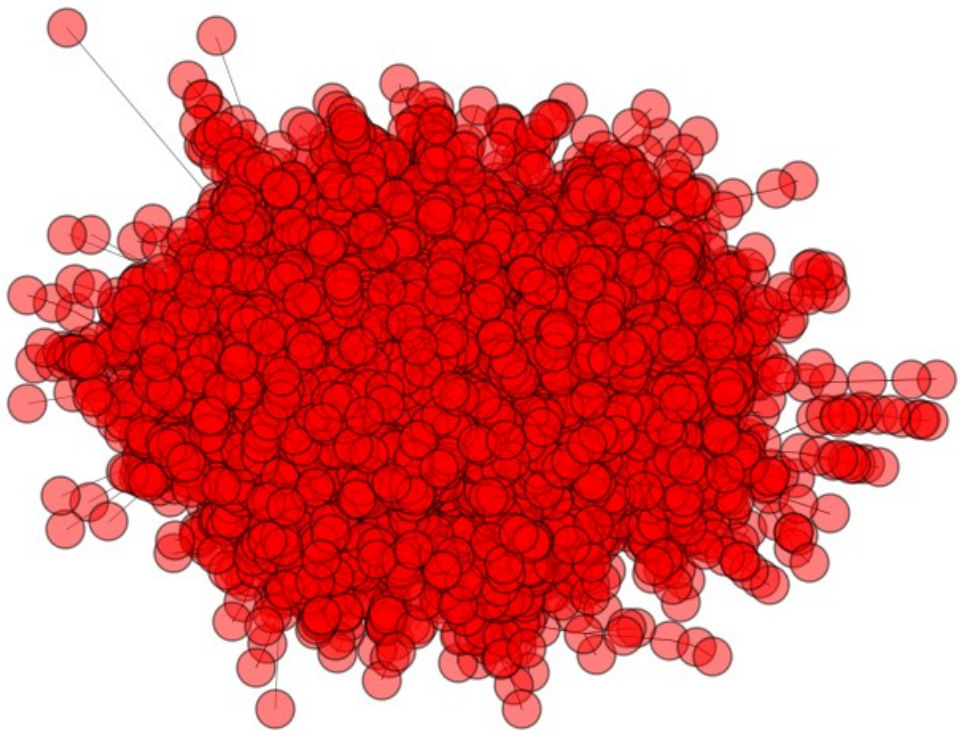
- k-Spanning Tree
- Shared Nearest Neighbor
- Betweenness Centrality Based
- Highly Connected Components یا Highly Connected Subgraph
- Maximal Clique Enumeration
- Kernel k-means

در این مساله خاص از روش louvain community detection که از روش های Modularity Optimization میباشد، استفاده شده است که در ادامه توضیح داده خواهد شد.

پیش پردازش:

در این مساله از آنجایی که تمام داده ها Numeric یا عددی می‌باشند، داده ها نیاز به پیش پردازش زیادی ندارند. پس از تبدیل فایل دادگان به فرمت مناسب استفاده، ستون های مبدا(Source)، مقصد(Destination) و وزن(weight) به ترتیب برای تشکیل گراف استفاده شده است.
این کار توسط کتابخانه NetworkX و به زبان Python انجام شده است.

نمودار گراف اولیه:



روش پیشنهادی:

در روش های Modularity Optimization برای افراز یک گراف وزن‌دار به اجتماعی از رئوس، یک تابع هزینه‌ی مناسب تعریف میشود که بر اساس آن میتوان گراف را خوشه‌بندی کرد.
برای این کار گراف با ماکزیمم ماژولاریتی توسط متد best_partition از کتابخانه Community محاسبه میشود. گراف حاصل بیشترین ماژولاریتی را مطابق الگوریتم Louvain دارد.

معیار های ارزیابی:

معیارهای گوناگونی برای ارزیابی خوشه‌بندی مناسب یک گراف وزن‌دار وجود دارد که در ادامه، به معیار انتخاب شده و دلیل انتخاب آن می‌پردازیم.

دو معیار مهم که برای ارزیابی مناسب بودن یک خوشه استفاده می‌شوند، **چگالی درون-خوشه‌ای و بین-خوشه‌ای** هستند. در یک خوشه‌بندی مناسب سعی بر این است که چگالی درون-خوشه‌ای بیشینه شود و چگالی بین-خوشه‌ای کمینه. این دو معیار طبق فرمول زیر در یک گراف قابل محاسبه هستند:

$$\delta_{\text{int}}(C) = \frac{\#internal\ edges\ of\ C}{n_c(n_c-1)/2}$$

$$\delta_{ext}(C) = \frac{\#inter-cluster\ edges\ of\ C}{n_c(n-n_c)}$$

- **چگالی درون-خوشه‌ای** :

نشان‌دهنده‌ی نسبت تعداد یال‌های در یک خوشه به تعداد کل یال‌های ممکن در آن خوشه است. همچنین چگالی بین-خوشه‌ای نشان‌دهنده‌ی تعداد یال‌های بین یک خوشه با سایر رئوس گراف، نسبت به کل تعداد یال‌های ممکن در گراف است. یک معیار مناسب که می‌توان تابع هزینه را بر اساس آن تعریف کرد تفاضل میان فاصله‌ی درون-خوشه‌ای و بین-خوشه‌ای است.
دو معیار بالا همان‌طور که مشخص است برای گراف‌های ساده و غیروزن‌دار کاربرد دارند. در صورت وزن‌دار بودن گراف این معیارها، وزن‌ها را در نظر نمی‌گیرند و تنها وجود یا عدم وجود یال بین رأس‌ها را مورد بررسی قرار می‌دهند. بنابراین نیاز به تغییری در معیار بالا برای قابل استفاده بودن در مسئله‌ی خوشه‌بندی گراف‌های وزن‌دار است.

معیاری که برای یک خوشه‌بندی در یک گراف وزن‌دار می‌تواند مناسب باشد آن است که مجموع اوزان یال‌ها درون یک خوشه یا اجتماع بیشینه و مجموع اوزان یک خوشه یا اجتماع با خوشه‌ها و اجتماعات دیگر کمینه شود. بنابراین معیار شهودی، معیار **ماژولاریتی**[1] که در بخش بعد به آن می‌پردازیم به عنوان معیار برای این مسئله انتخاب شده است.

- **معیار خوشه‌بندی ماژولاریتی:**

یک خوشه‌بندی مناسب در یک گراف، لزوما خوشه‌بندی نیست که در آن تعداد یال‌ها بین خوشه‌ها کمترین مقدار ممکن باشد، چرا که اگر به طور مثال یک رأس تنها در یک دسته باشد و تمام رئوس دیگر در دسته‌ای دیگر، آن‌گاه بر طبق این معیار یک جواب بهینه داریم که آشکارا نامناسب است. یک خوشه‌بندی مناسب آن خوشه‌بندی است که **تعداد یال‌ها میان دو خوشه‌ی مختلف، کمتر از مقدار مورد انتظار باشد**. معیار ماژولاریتی نیز بر این اساس تعریف می‌شود. در واقع ماژولاریتی در یک خوشه از تفریق تعداد یال‌های میان رئوس از تعداد یال‌های مورد انتظار در یک حالت تصادفی حاصل می‌شود. روش دقیق به دست آوردن ماژولاریتی در زیر توضیح داده شده است.

با در نظر گرفتن ماتریس مجاورت گراف مورد نظر، منظور از A_{ij} وزن یال بین دو رأس i و j است. همچنین در صورتی که حالت تصادفی را در نظر بگیریم، همچنین K_i را درجه‌ی رأس K در نظر بگیرید. m نیز تعداد کل یال‌های گراف است. بنابراین ماژولاریتی برای یک خوشه می‌تواند به صورت زیر تعریف شود:

$$A_{vw} - \frac{k_v k_w}{2m}$$

با جمع‌بستن تمامی این مقادیر به ازای هر دو رأس، به معادله‌ی زیر می‌رسیم:

$$Q = \frac{1}{2m} \sum_{uv} [A_{uv} - \frac{k_v k_u}{2m}] \delta(c_v, c_u)$$

مقدار تابع دلتا یک است اگر دو رأس در یک خوشه یا اجتماع باشند وگرنه صفر است. حال هدف ماکسیموم کردن مقدار Q است. مقدار Q عددیست در بازه‌ی [۱و ۱] و هرچه‌قدر به ۱ نزدیک‌تر باشد، نشان‌دهنده‌ی دسته‌بندی بهتر است.

بنابراین مسئله‌ی اولیه تبدیل به بیشینه‌کردن ماژولاریتی می‌شود. برای این کار الگوریتم‌های گوناگونی وجود دارد که در این مسئله از الگوریتم **Louvain** استفاده شده است.

1Modularity

نتایج:

تعداد اجتماعات به دست آمده به طوری که ماژولاریتی حداکثر شود، با استفاده از الگوریتم **Louvain** برابر ۱۶۸ خوشه بود. در فایل **output.csv** تعلق هر یک از منابع به هر کدام از خوشه‌ها به عنوان پاسخ نهایی آمده است. این فایل دارای دو ستون است که یکی شماره‌ی منبع و دیگری شماره‌ی خوشه‌ای است که آن منبع به آن تعلق دارد.

برای ارزیابی پاسخ، دو روش انتخاب شده‌اند. یکی محاسبه‌ی میزان ماژولاریتی برای تقسیم‌بندی مذکور است که مقدار آن برابر با **Modularity = 0.971198961478** است و نشان‌دهنده‌ی آن است که اجتماعات یافته شده، دارای بیشترین ارتباط درون خود، و کمترین ارتباط ممکن نسبت به حالت مورد انتظار هستند. این عدد به این جهت که به یک نزدیک است، نشان‌دهنده‌ی خوشه‌بندی مناسب بر اساس معیار ماژولاریتی است.

همچنین گراف القایی دیگری نیز روی اجتماعات به دست آمده تعریف شده است. رأس‌های این گراف القایی، اجتماعات به دست آمده هستند و یال‌های بین هر دو رأس نشان‌دهنده‌ی مجموع اوزانیست که بین آن دو اجتماع در گراف اصلی موجود است. بدیهی است که برای هر سه تایی (i, j, k) اگر $i = j$ باشد، میزان k که نشان‌دهنده‌ی مجموع اوزان بین دو اجتماع است، باید مقدار بزرگی باشد. چرا که در صورت برابری i و j در واقع اوزان بین منابع یک اجتماع مدنظر است که باید عدد بزرگی باشد. در صورتی که i و j برابر نباشند مقدار k باید تا حد امکان کوچک باشد، چرا که مجموع اوزان میان دو خوشه‌ی متمایز هر چه قدر کمتر باشد، خوشه‌بندی بهتر انجام گرفته است. مجموع اوزان میان خوشه‌ها در فایل **induced_graph.csv** قابل مشاهده است. به طور مثال مجموع اوزان بین منابع در میان اعضای خوشه‌ی اول، مطابق اطلاعات فایل، برابر ۴۰۳۰ است، و مجموع اوزان میان منابع خوشه‌ی اول با منابع خوشه‌ی شصتم، مطابق فایل برابر با ۱۰ است.

یکی دیگر از مزایای الگوریتم Louvain که بر اساس معیار ماژولاریتی عمل می‌کند آن است که در **هر خوشه، هر تعداد از رئوس می‌توانند قرار بگیرند** و اندازه‌ی هر کدام از خوشه‌ها(تعداد رئوس در یک اجتماع) لزوماً نباید با یکدیگر برابر باشند. بنابراین خوشه‌های گوناگون اندازه‌های گوناگونی دارند که در فایل **number_of_nodes.csv** اندازه‌ی هر کدام از خوشه‌ها قابل مشاهده است.

گراف اجتماعات حاصل: