فصل دهم

موضوع : Mining Social-Network Graphs

به نام خدا

مقدار زیادی داده با تجزیه و تحلیل از شبکه های اجتماعی به دست می آید.

شناخته شده ترین نمونه یک شبکه اجتماعی ، رابطه "دوستان" است که در سایت هایی مانند فیس بوک یافت می شود.

با این حال ، همانطور که خواهیم دید بسیاری از منابع داده دیگر وجود دارند که افراد یا نهادهای دیگر را به هم وصل می کنند.

در این فصل تکنیک های تحلیل داده های موجود در شبکه ها را بررسی می کنیم.

یک سوال مهم در مورد یک شبکه اجتماعی ، چگونگی شناسایی "اجتماعات" ، یعنی زیر مجموعه گره ها (افراد یا اشخاص دیگری که شبکه را تشکیل می دهند) با اتصالات غیرمعمول قوی است.

برخی از تکنیک های مورد استفاده برای شناسایی جوامع مشابه الگوریتم های خوشه بندی است که در فصل 7 درباره آنها گفتیم.

با این حال ، مجموعه ها تقریباً هرگز قسمت هایی از گره ها را در یک شبكه تقسیم نمی كنند.

در عوض ، مجموعه ها معمولاً با هم همپوشانی دارند.

به عنوان مثال ، شما ممکن است به چندین گروه مختلف از دوستان یا همکلاسی های خود تعلق داشته باشید.

مردم یک جامعه تمایل دارند یکدیگر را بشناسند ، اما مردم دو جامعه مختلف به ندرت یکدیگر را می شناسند.

شما نمی خواهید فقط در یکی از گروه ها قرار بگیرید و همچنین معقول نیست که همه ی افراد از همه ی جوامع خود را حتما در یک گروه قرار دهید.

همچنین در این فصل الگوریتم های کارآمد را برای کشف سایر خصوصیات گراف ها بررسی می کنیم.

ما به یک مفهومی تحت عنوان simrank نگاه می کنیم که این مفهوم به معنی کشف شباهت بین گره های یک گراف است.

ما شمارش مثلثی را به عنوان روشی برای سنجش ارتباط در یک جامعه استفاده می کنیم.

ما برای اندازه گیری دقیق و تقریبی اندازه مجموعه ها و گره های در یک گراف ، الگوریتم های مناسبی را ارائه می دهیم.

سرانجام ، برای محاسبه بستار انتقالی ، به الگوریتم های ویژه ای را بررسی می کنیم.

10.1 شبکه های اجتماعی به صورت گراف

بحث خود را در مورد شبکه های اجتماعی با معرفی یک مدل گرافی آغاز می کنیم.

هر نمودار گرافی مناسب به نمایش یک اجتماع در شبکه های اجتماعی نیست.

بنابراین، ما درباره ی ایده ی اصل محلیت که جز ویژگی های اصلی شبکه های اجتماعی است با کمک نودها و یال ها در گراف ها صحبت می کنیم. با کمک گره ها و یال ها تمایل خوشه بندی در شبکه ها بررسی می کنیم.

در این بخش همچنین برخی از انواع مختلف شبکه های اجتماعی که در عمل مورد استفاده قرار می گیرند را بررسی می کنیم.

10.1.1 شبکه اجتماعی چیست؟

وقتی به یک مفهوم شبکه اجتماعی فکر می کنیم ، به فیس بوک ، توییتر ، Google+ یا وب سایت دیگری فکر می کنیم که "شبکه اجتماعی" نامیده می شود و در واقع این شبکه ها نماینده ای از مفهوم شبکه های اجتماعی هستند.

ویژگی های اساسی یک شبکه اجتماعی عبارتند از:

1. مجموعه ای از اشخاص در شبکه های اجتماعی در شبکه وجود دارند که به طور معمول این موجودات مردم هستند اما می توانند چیزهای دیگری نیز باشند؛ در بخش 10.1.3 به مثالهای بیشتری در این مورد می پردازیم.
2. حداقل یک رابطه بین موجودیت های موجود (کاربران) در شبکه های اجتماعی وچود دارد. به رابطه ی بین موجودیت ها در فیس بوک Relationship می گویند. ارتباط ها یا وجود داشته یا ندارد پس دو نفر یا دوست هستند یا نیستند؛  
   گرچه در گراف های شبکه های اجتماعی رابطه ها دارای یک درجه هستند.  
   این درجه ها می تواند مقدار گسسته داشته باشد. به عنوان مثال در گوگل پلاس این درجه ها با عنوان های دوستان، خانواده، آشنایان و ... شناخته می شود.  
   این درجه می تواند یک عدد صحیح باشد یا یک عدد کسری باشد که از میزان صحبت بین دونفر بدست می آید.
3. محلیت در این شبکه ها به صورت غیرتصادفی است. این شرط برای نرمال سازی سخت ترین شرط است. اما ارتباطش به گرایش خوشه ها و روابط آنها می پردازد. یعنی اگر موجودیت A به هر دو موجودیت B و C مربوط باشد احتمال رخداد آن از میانگین B و C بدست می آید.

10.1.2 شبکه های اجتماعی به صورت گراف

شبکه های اجتماعی به طور معمول به عنوان گراف هایی مدل سازی می شوند که بعضا از آنها به عنوان یک گراف اجتماعی یاد می کنیم.

در این نمودار گراف گره ها موجودیت ها هستند و یال ها اتصال و ارتباط بین دو موجودیت (کاربر) هستند. اگر در این گراف درجه ای موجود باشد روی یال ها برچسب زده می شود.

غالبا گراف های موجود در شبکه های اجتماعی بدون جهت هستند؛ مثل گراف دوستان در شبکه اجتماعی فیس بوک.

اما می توان گراف های جهت دار هم داشت؛ مانند نمودار فالورهای توییتر یا گوگل پلاس.

مثال 10.1 : شکل 10.1 نمونه ای از یک گراف کوچک شبکه اجتماعی است؛ گره های این گراف از A تا G نامگذاری شده است. رابطه ی موجود در این گراف به عنوان دوست شناخته می شود که یال ها را تشکیل می دهد.

به عنوان مثال موجودیت B با A و C و D دوست می باشد.

آیا این گراف واقعا نشان دهنده ی یک شبکه ی اجتماعی و نمایش روابط آنها است؟

ابتدا توجه داشته باشید که گراف زیر دارای 9 یال می باشد.



شکل 10.1 : نمونه ای از شبکه اجتماعی کوچک

. در واقع 21 جفت گره می تواند در این شبکه دارای لبه باشد یا حداکثر یال های موجود در این شبکه 21 باشد.

فرض کنید X و Y و Z نودهای شکل 10.1 هستند که بین x و y و همچنین بین x و z دارای یال می باشد.

احتمال وجود یال بین y و Z چقدر است؟

اگر این نمودار بزرگ باشد این احتمالا به صورت کسری می باشد. یعنی در واقع 9 تقسیم بر 21 که برابر 0.429 می شود.

اما ، از آنجا که نمودار كوچك است ، بین احتمال واقعی و نسبت تعداد لبه ها به تعداد جفت گره ها تفاوت قابل ملاحظه ای وجود دارد.

از آنجایی که می دانیم یال های (X، Y) و (X، Z) وجود دارد ، تنها 7 یال باقی مانده است. این 7 یال می توانند بین هر 19 گره باقی مانده از گره ها قرار بگیرند.

بنابراین احتمال یک لبه بین z و y برابر 7 تقسیم بر 19 که برابر 0.368 می شود است.

حال باید این احتمال را محاسبه کنیم که لبه (Y ، Z) در شکل 10.1 وجود داشته باشد ، با توجه به اینکه لبه ها (X ، Y) و (X ، Z) وجود دارند. آنچه در واقع باید حساب کنیم ، جفت گره هایی است که می توانند Y و Z باشند ، بدون اینکه نگرانی در مورد کدام گره Y باشد و کدام Z باشد. اگر X A باشد ، باید Y و Z به ترتیب B و C باشند. از آنجا که لبه (B ، C) وجود دارد ، A یک مثال مثبت (که در آن لبه وجود دارد) است و هیچ نمونه منفی (جایی که لبه وجود ندارد) در آن نیست.

در مواردی که X به جای C ، E یا G است نتایج یکسان است.

در هر حالت ، X فقط دو همسایه دارد و لبه بین همسایگان وجود دارد. بنابراین ، ما تاکنون چهار نمونه مثبت و صفر مثال منفی را دیده ایم.

حال ، در نظر بگیرید که X = F. F دارای سه همسایه ، D ، E و G است. لبه هایی بین دو سه جفت همسایه وجود دارد ، اما هیچ لبه ای بین G و E وجود ندارد. بنابراین ، ما دو نمونه مثبت دیگر را می بینیم و اولین نمونه منفی خود را می بینیم.

اگر X = B ، دوباره سه همسایه وجود دارد ، اما فقط یک جفت همسایه ، A و C ، یک لبه دارند. بنابراین ، ما دو مثال منفی دیگر ، و یک مثال مثبت ، برای کل هفت مثبت و سه منفی داریم.سرانجام ، وقتی X = D ، چهار همسایه وجود دارد. از شش جفت همسایه ، فقط دو نفر بین آنها لبه دارند.

بنابراین ، تعداد کل نمونه های مثبت 9 و تعداد کل نمونه های منفی 7 است. در شکل 10.1 می بینیم که کسر ما برابر 9 تقسیم بر 16 است که در واقع برابر 0.563 می شود. این کسر خیلی بیشتر از مقدار قابل انتظار ما که 0.368 است می باشد.

در نتیجه شکل 10.1 واقعا اصل locality در شبکه های اجتماعی را نشان می دهد.

10.1.3 انواع شبکه های اجتماعی

نمونه های زیادی از شبکه های اجتماعی وجود دارد که ماهیت دوستان ندارد. در اینجا ، اجازه دهید تعدادی از نمونه های دیگر شبکه های اجتماعی که با اصل locality روابط را نشان می دهند ذکر کنیم.

شبکه های تلفن

در اینجا گره ها شماره تلفن ها را نشان می دهند ، که در واقع افراد هستند. اگر در طی مدت زمان مشخصی مانند ماه گذشته یا از ابتدا تا به حال بین این تلفن ها تماس برقرار شده باشد بین دو گره وجود دارد. یال ها را می توان با تعداد تماس های انجام شده بین این تلفن ها در طول دوره ی مشخص وزن داد. جوامع در یک شبکه تلفنی از گروههایی تشکیل می شوند که مرتباً ارتباط برقرار می کنند: برای مثال گروه هایی از دوستان ، اعضای یک باشگاه یا افرادی که در همان شرکت کار می کنند.

شبکه های ایمیلی

در شبکه های ایمیلی گره ها آدرس ایمیل افراد را نشان می دهند. یال ها بیانگر وجود حداقل یک ایمیل بین دو آدرس ایمیل می باشد. از طرف دیگر ممکن است لبه ها در این نوع شبکه ها به صورت یک طرفه یا دوطرفه باشند. از نمایش هرزنامه به عنوان دوست در این شبکه ها خودداری می شود یا با رویکردی آدرس های ایمیل هرزنامه را با لبه های ضعیف و لبه های دیگر را با لبه های قوی تر نشان می دهیم.

اجتماعاتی که در شبکه های ایمیلی مشاهده می شود از همان گروه بندی هایی هستند که در ارتباط با شبکه های تلفنی از آنها یاد کردیم. یک نوع دیگر برای مرتب سازی شبکه های ایمیلی از افرادی که از طریق تلفن های همراه متن های خود را می نویسند است.

شبکه های همکاری

گره ها افرادی را نشان می دهند که مقالات تحقیقاتی را منتشر کرده اند. بین دو فرد که یک یا چند مقاله را به طور مشترک منتشر کرده اند ، لبه وجود دارد. به صورت اختیاری ، می توان لبه ها را با تعداد انتشارات مشترک برچسب گذاری کرد. گروه های این نوع شبکه نویسنده هایی هستند که روی یک موضوع خاص کار می کنند. نمای جایگزین از همان داده ها به عنوان گرافی است که در آن گره ها مقالات هستند. اگر حداقل یک نویسنده مشترک داشته باشند ، دو مقاله به یک لبه متصل می شوند. اکنون ، گروه هایی را تشکیل می دهیم که مجموعه ای از مقالات در همین موضوع را شامل هستند.

چندین نوع داده دیگر وجود دارد که دو شبکه را به روشی مشابه ایجاد می کند.

به عنوان مثال ، می توانیم به افرادی که مقالات ویکی پدیا را برای بار اول منتشر می کنند و افرادی که مقاله هایی را ویرایش می کنند تقسیم کنیم. اگر ویرایش یک مقاله به صورت مشترک باشد ، دو گره ویرایشگر به یکدیگر متصل هستند. گروه هایی از ویراستاران که به صورت مشترک کار کرده اند را در یک دسته قرار می دهیم. بطور مضاعف ، می توانیم شبکه ای از مقالات بسازیم و در صورت ویرایش آنها توسط همان شخص ، مقاله ها را وصل کنیم.

در اینجا ، ما مقالاتی را درمورد موضوعات مشابه یا مرتبط با هم جمع می کنیم.

در واقع ، داده های مربوط به همکاری ، همانطور که در فصل 9 مورد بحث قرار گرفت ، اغلب می توانند به عنوان تشکیل یک جفت شبکه ، یکی برای مشتریان و دیگری برای محصولات مشاهده شوند.

مشتریانی که کالاهای مشابهی را خریداری می کنند ، به عنوان مثال ، کتاب های علمی تأیید می کنند ، جوامع تشکیل می دهند و بصورت دوگانه ، کالاهایی که توسط همان مشتریان خریداری می شوند ، جوامع را تشکیل می دهند؛

به عنوان مثال ، تمام کتاب های علمی تولیدی.

مثال های دیگر از گراف های شبکه های اجتماعی

بسیاری از پدیده های دیگر گراف هایی ا ایجاد می کنند که چیزی شبیه به گراف های شبکه های اجتماعی است ، به خصوص نمایش محلیت ها.

مثالهای این بخش عبارتند از: شبکه های اطلاعاتی (اسناد ، گراف های در بستر وب ، ثبت اختراعات) ، شبکه های زیرساختی (جاده ها ، هواپیماها ، لوله های آب ، نیروگاهها) ، شبکه های بیولوژیکی (ژن ها ، پروتئین ها ، شبکه های غذایی حیوانات که یکدیگر را می خورند) و همچنین انواع دیگر ، مانند شبکه های خرید محصول (به عنوان مثال ، Groupon).

10.1.4 گراف ها دارای چندین نوع گره

پدیده های اجتماعی دیگری وجود دارند که موجودیت های مختلفی را درگیر می کنند. تحت عنوان "شبکه های همکاری" ، انواع مختلفی از گراف ها که از دو نوع گره واقعاً تشکیل شده اند را دیده ایم.

شبکه های نویسندگی می توانند گره های نویسنده و گره های کاغذ را تشکیل دهند. در بحث بالا ، ما با از بین بردن گره های یکی از این دو نوع ، دو شبکه اجتماعی ایجاد کردیم اما لازم نیست این کار را انجام دهیم. ما می توانیم به طور کلی به ساختار فکر کنیم.

به عنوان مثال پیچیده تر ، کاربران در سایتی مانند del.icio.us برچسب ها را در صفحات وب قرار می دهند. بنابراین سه نوع مختلف موجودیت وجود دارد: کاربران ، برچسب ها و صفحات. ممکن است فکر کنیم اگر تمایل به استفاده از همان برچسب ها به طور مکرر یا اگر تمایل به برچسب زدن به همان صفحات دارند ، کاربران به نوعی به یکدیگر متصل شده اند. به طور مشابه ، برچسب ها اگر در همان صفحات ظاهر شوند می توانند مرتبط باشند؛ یا اگر برچسب ها توسط کاربران زیادی استفاده شود ممکن است صفحات مشابه به نظر برسند.

یک روش طبیعی برای نمایش چنین اطلاعاتی است که به عنوان گراف k-partite شناخته می شود؛ که k همیشه بزرگتر از 1 است. در بهش 8.3 گرافی دارای k=2 را مشاهده کرده ایم. به طور کلی یک نمودار k-partite از مجموعه گره هایی با k جداکننده تشکیل شده است که بین گره های یک گروه هیچ یالی وجود ندارد.

مثال 10.2 : شکل 10.2 نمونه ای از یک گراف k-partite است که مقدار k = 3 است. سه نوع گره وجود دارد: گره کاربران را در این گراف با {U1, U2} و گره برچسب ها را با {T1, T2, T3, T4} و گره صفخات وب را با {W1, W2, W3} نمایش می دهیم. توجه کنید تمام یال ها بین دو مجموعه مختلف متصل هستند.

ممکن است حس کنید این نمودار اطلاعاتی در مورد سه نوع موجودیت را نشان می دهد. به عنوان مثال لبه ی (U1, T2) به این معنی است که کاربر U1 برچسب T2 را حداقل در یک صفحه ی وب قرار داده است.

توجه کنید که گراف جزئیاتی را که می تواند مهم باشد را به ما نمی گوید. به عنوان مثال برای نمایش اینکه چه کسی چنین برچسبی را در این صفحه قرار داده است نیاز به نمایش پیچیده تری مانند روابط موجود در بانک های اطلاعاتی سه ستونه نیاز داریم.



شکل 10.2 : نمودار k-partite با مقدار k=3 برای نمایش کاربران، برچسب ها و صفحات وب

10.1.5 تمرینات مربوط به بخش 10.1

تمرین 10.1. : لبه های گراف G را به عنوان گره های G’ در نظر می گیریم.

1. اگر (X ، Y) یالی از گراف G باشد ، XY ، نمایانگر مجموعه ای هماهنگ از X و Y گره ای از گراف G’ است. توجه داشته باشید که XY و Y X یک گره G’ را نشان می دهند ، و نه دو گره مختلف.
2. اگر (X ، Y) و (X ، Z) لبه های گراف G باشند ، در گراف G’ یک لبه بین XY و XZ وجود دارد. اگر گره های G’ که این گره ها نمایان هستند دارای یک گره (از G) مشترک هستند ، گره های G’ بین آنها یال دارد.
3. اگر ساخت گراف را به صورت دوتایی را در شبکه ای از دوستان بکار گیریم ، تعبیر لبه های نمودار نتیجه چیست؟
4. ساخت دوتایی را در گراف شکل 10.1 اعمال کنید.
5. درجه گره XY در گراف G’ چگونه با درجه X و Y در گراف G ارتباط دارد؟
6. تعداد لبه های گراف G’ مربوط به درجه گره های گراف G توسط یک فرمول خاص است. آن فرمول را کشف کنید.
7. آنچه ما آنرا به عنوان دوگان نامیدیم در واقع یک دوتایی واقعی نیست؛ زیرا استفاده از این روش ساختن گراف در گراف G’ لزوماً یک ایزومورف گراف G می دهد. یک نمونه از گراف G را بدهید که در آن دوتایی از G’ از نظر گراف G ایزومورف است و مثال دیگری هم بدهید که در آن گرافG’ از نظر G ایزومورف نیست.

10.2 خوشه بندی گراف شبکه های اجتماعی

جنبه مهم شبکه های اجتماعی این است که آنها حاوی جوامع موجوداتی هستند که توسط بسیاری از لبه ها به هم وصل می شوند. به عنوان مثال ، اینها با گروهی از دوستان در مدرسه یا گروههایی از محققان علاقمند به همان موضوع مطابقت دارد. در این بخش ، خوشه بندی نمودار را به عنوان راهی برای شناسایی جوامع در نظر می گیریم. به نظر می رسد که تکنیک هایی که در فصل 7 آموخته ایم ، معمولاً برای مشکل خوشه بندی نمودارهای شبکه های اجتماعی نامناسب است.

10.2.1 معیار فاصله در گراف شبکه های اجتماعی

اگر بخواهیم از تکنیک های خوشه بندی استاندارد در یک گراف شبکه های اجتماعی استفاده کنیم ، اولین گام ما تعیین یک روش اندازه گیری فاصله است. هنگامی که یال های نمودار دارای برچسب هستند ، این برچسب ها بسته به آنچه که آنها نشان می دهند ، می توانند به عنوان اندازه گیری فاصله قابل استفاده باشند. اما هنگامی که لبه ها بدون برچسب هستند ، مانند نمودار «دوستان» ، برای تعیین فاصله ی مناسب ، کار زیادی نمی توان انجام داد.

اولین فرض ما این است که در نظر بگیریم گره ها نزدیک هستند و اگر یالی بین آنها باید یا نباشد دارای یک فاصله ی معین است. بنابراین ، می توان گفت که فاصله d(x,y) در صورت وجود لبه (x ، y) 1 است و در صورت عدم وجود چنین لبه ای 0 است. ما می توانیم از دو مقدار دیگری مثل 1 و ∞ استفاده کنیم تا زمانی که فاصله لبه ها به هم نزدیک تر باشد.

هیچ یک از معیارهای بررسی فاصله با دو ارزش 0 و 1 یا 0 و ∞ یک معیار واقعی و درست برای فاصله نیست. دلیل این امر این است که در هنگام اتصال مثلثی بین گره ها این مقدار فاصله بین دو گره نقض می شود. یعنی اگر لبه های (A ، B) و (B ، C) وجود داشته باشد ، اما هیچ لبه ای (A ، C) وجود ندارد ، در این صورت فاصله از A تا C از مجموع مسافت های A تا B از C بیشتر می شود. می توانیم برای این مشکل فاصله ی تا یک لبه را به طور مستقیم 1 در نظر بگیریم و برای missing edge ها از فاصله ی 1.5 استفاده کنیم که اینکار مشکل بالا را حل می کند اما مشکل توابع محاسبه ی فاصله به نابرابری مثلثی محدود نمی شود که در بخش بعدی این موضوع را خواهیم دید.