دانشکده ریاضی، آمار و علوم کامپیوتر

ارزیابی الگوریتم‌های یافتن موتیف در شبکه‌های زیست‌شناسی

نگارش

الناز صابری انصاری

استاد راهنما

دکتر هایده اهرابیان

استاد مشاور

دکتر عباس نوذری دالینی

پایان نامه جهت دریافت درجهٔ کارشناسی ارشد
در رشتهٔ علوم کامپیوتر

اسفند ۱۳۸۷

تقدیم به

پدر و مادر عزیزم و همسر مهربانم
به خاطر محبت‌ها و حمایت‌هایشان.

چکیده

شبکه‌های زیست‌شناسی، عموماً شبکه‌های پیچیده و وسیعی هستند که حاوی اطلاعات مهمی می‌باشند. تاکنون تلاش‌های بسیاری به منظور درک ساختار و عملکرد این شبکه‌ها انجام شده است. در سال‌های اخیر مشخص شده است که شبکه‌های زیست‌شناسی در برگیرنده موتیف‌هایی هستند که به نظر می‌رسد، دارای عملکردهای خاصی در شبکه می‌باشند. موتیف‌ها زیرشبکه‌های کوچکِ همبندی هستند که در شبکه مورد بررسی با فراوانی بالاتری نسبت به شبکه‌های تصادفی مشاهده می‌شوند. اخیراً توجه زیادی بر روی موتیف‌ها، به منظور درک بهتر ساختار شبکه‌های پیچیده جلب شده است. الگوریتم‌های موجود برای یافتن موتیف‌های شبکه، دارای پیچیدگی محاسباتی زیادی هستند و قادر به پیدا کردن موتیف‌ها در شبکه‌های بزرگ و با اندازه بیشتر از هشت رأس نیستند. در این پایان‌نامه الگوریتمی به منظور یافتن موتیف‌هایی با هر اندازه، در شبکه‌ها ارائه شده است. آزمایش‌هایی که بر روی مجموعه‌ای از شبکه‌های زیستی و غیر زیستی انجام شده است، نشان می‌دهد که این الگوریتم دارای پیچیدگی محاسباتی کمتری نسبت به الگوریتم‌های قبلی می‌باشد. این امر منجر به تشخیص موتیف‌هایی با تعداد رئوس بیشتر و در شبکه‌های بزرگتر می‌شود که پیشرفت قابل توجهی در این زمینه است.

پیش‌گفتار

شبکه‌های پیچیده مانند شبکه‌های اجتماعی، شبکه‌های کامپیوتری و شبکه‌های زیستی که شامل تعداد رئوس فراوانی هستند، به تازگی توجه زیادی را به خود جلب کرده‌اند [۱]. تاکنون مطالعات بسیاری بر روی شبکه‌های زیست‌شناسی مانند شبکه‌های تأثیر متقابل پروتئین-پروتئین، شبکه‌های تنظیم ژن و شبکه‌های متابولیکی انجام شده است [۲,۳]. ابزارهای محاسباتی قدرتمندی به منظور استخراج اطلاعات از میان حجم انبوه داده‌های ذخیره شده در این شبکه‌ها، مورد نیاز هستند. دست یافتن به اطلاعاتی فراتر از مشخصه‌های کلی در شبکه‌ها، ملزم به بدست آوردن شناختی از اجزای ساختاری اصلی مربوط به کلاس‌های مختلف شبکه‌ها می‌باشد.

براساس این نظریه که "تکامل مدل‌هایی را که نمایان‌گر عملکردهای خاصی هستند، حفظ می‌کند" [۴]، در سال ۲۰۰۲ میلو[1] ادعا کرد که احتمال تکرار زیرساختارهای خاصی در یک شبکه و یا حتی در میان شبکه‌های مختلف وجود دارد [۵]. این زیرساختارها که نشان‌دهنده یک الگوی مشخص از تعاملات میان رأس‌های شبکه هستند، دارای عملکرد ویژه‌ای در شبکه می‌باشند. این زیرساختارهای تکرارشونده و معنی‌دار موتیف نامیده می‌شوند. در سال‌های اخیر، به منظور درک اصول کلیِ طراحی ساختار شبکه‌های پیچیده، موتیف‌ها توجه زیادی را به خود جلب کرده‌اند. مطالعه بر روی موتیف‌ها، همچنین برای درک بهتر عملکرد شبکه‌ها مفید می‌باشد. تحقیقات نشان داده است در شبکه‌هایی که فرایندهای مشابهی انجام می‌دهند، موتیف‌های مشابهی مشاهده شده است، هر چند این شبکه‌ها مربوط به سیستم‌های مختلف زیست‌شناسی باشند [۴]. به همین دلیل می‌توان از موتیف‌ها به عنوان یک مولفه اصلی برای کلاس‌بندی شبکه‌ها استفاده نمود. به علت اینکه عملکرد موتیف‌ها ممکن است از ابتدا مشخص نباشد، موتیف‌ها را به صورت مستقل از عملکرد آن‌ها و بر اساس فراوانی آن‌ها در شبکه تعریف می‌کنند. بنابراین می‌توان موتیف‌ها را به صورت زیرشبکه‌های کوچک همبندی در نظر گرفت که در شبکه حقیقی دارای فراوانی بالاتری نسبت به شبکه‌های تصادفی باشند. عمل تشخیص موتیف‌های شبکه، در حقیقت مسئله پیدا کردن زیرساختارهایی از شبکه است که دارای اهمیت آماری هستند.

تاکنون الگوریتم‌های مختلفی برای تشخیص موتیف‌های شبکه ارائه شده‌اند. این الگوریتم‌ها، عمل تشخیص موتیف‌های شبکه را بر اساس شمارش کامل زیرگراف‌ها و یا تخمین تراکم آن‌ها انجام می‌دهند. از الگوریتم‌های موجود در این زمینه می‌توان الگوریتم ارائه شده توسط میلو [۵]، الگوریتم‌های اِم‌فایندِر[2] [۶,۷]، مَویستو[3] [۸,۹] و فن‌مُد[4] [۱۰,۱۱] را نام برد. در میان این الگوریتم‌ها،

---

[1] Milo
[2] Mfinder
[3] MAVisto

الگوریتم فَن‌مُد سریعتر و بهتر از بقیه عمل می‌کند. این الگوریتم‌ها دارای پیچیدگی محاسباتی بالایی هستند و قادر به پیدا کردن موتیف‌ها در شبکه‌های بزرگ (با بیشتر از ۱۰۰،۰۰۰ یال) نمی‌باشند. این الگوریتم‌ها همچنین نمی‌توانند موتیف‌های با سایز بزرگ (بیشتر از هشت رأس) را پیدا کنند.

در این پایان‌نامه، الگوریتم جدیدی برای یافتن موتیف‌های شبکه ارائه شده است. این الگوریتم محدودیت‌های الگوریتم‌های قبلی را در زمینه اندازه شبکه‌ها یا سایز موتیف‌ها ندارد و پیچیدگی محاسباتی آن کمتر از الگوریتم‌های موجود می‌باشد.

برای تشخیص موتیف‌های شبکه، دانستن مفاهیم اولیه مربوط به گراف، شبکه و شبکه‌های زیستی ضروری است. در فصل اول از این پایان‌نامه این مفاهیم اولیه شرح داده شده‌اند و مسئله یافتن موتیف‌های شبکه، تعریف شده است. همچنین یک الگوریتم کلی برای پیدا کردن موتیف‌های شبکه ارائه شده است.

در فصل دوم و سوم به بررسی الگوریتم‌های موجود برای یافتن موتیف‌های شبکه، پرداخته شده است. در فصل دوم دو نمونه از الگوریتم‌های تشخیص موتیف‌های شبکه بررسی شده‌اند. یکی از آن‌ها الگوریتمی است که توسط میلو[۵] و همکارانش در سال ۲۰۰۲ ارائه شده است [۵] و موتیف‌های شبکه را بر اساس شمارش کامل زیرگراف‌ها تشخیص می‌دهد. به دلیل اینکه تعداد زیرگراف‌های شبکه‌ها حتی برای سایزهای کوچک بسیار زیاد می‌باشد، در الگوریتم دوم که در سال ۲۰۰۴ توسط کاشتان[۶] و همکارانش ارائه شده است [۶]، به جای شمارش کامل زیرگراف‌ها، یک روش نمونه‌گیری برای تخمین تراکم نسبی آن‌ها ارائه شده است. در فصل سوم نیز دو نمونه‌ی دیگر از الگوریتم‌های یافتن موتیف‌های شبکه بررسی می‌شوند. در یکی از الگوریتم‌ها که در سال ۲۰۰۵، توسط شرایبِر[۷] و شوبِرمایِر[۸] ارائه شده است [۸]، ابتدا سه دیدگاه مختلف برای تعیین فراوانی زیرگراف‌های شبکه مطرح شده و سپس یک الگوریتم جستجوی انعطاف‌پذیر برای محاسبه این فراوانی‌ها ارائه شده است. الگوریتم دیگر در سال ۲۰۰۶، توسط سِباستیَن وِرنیک[۹] ارائه شده [۱۰] که می‌تواند موتیف‌های شبکه را بر اساس شمارش کامل زیرگراف‌های شبکه و یا تخمین تراکم آن‌ها تشخیص بدهد. این الگوریتم بهتر و سریعتر از الگوریتم‌های قبلی عمل می‌کند. ولی قادر به پیدا کردن موتیف‌های با سایز بزرگتر از هشت نمی‌باشد.

و نهایتاً در فصل چهارم این پایان‌نامه، الگوریتم جدید ارائه شده برای یافتن موتیف‌های شبکه مورد بحث قرار می‌گیرد.

---

4 FANMOD
5 Milo
6 Kashtan
7 Schreiber
8 Schobbermeyer
9 Sebastian Wernicke

تقدیر و تشکر

اکنون که به یاری خداوند گامی دیگر را در زندگی خود پشت سر نهاده‌ام، از تمام عزیزانی که مرا در این راه یاری نمودند تشکر می‌کنم.

با تشکر از اساتید محترم، دکتر اهرابیان و دکتر نوذزی به خاطر تمامی کمک‌ها و راهنمایی‌های بی‌دریغشان.

با تشکر از پدر و مادر عزیز و خواهر مهربانم که همواره پشتیبان من بوده‌اند، به خاطر همه زحمت‌ها و محبت‌هایشان.

با تشکر از همسر مهربانم که همیشه در کنارم بود، به خاطر تمامی کمک‌ها و دلگرمی‌هایش.

با تشکر از پدربزرگ و مادربزرگ دوست‌داشتنی‌ام به خاطر تمام همراهی‌ها و مهربانیشان.

همچنین از دوستان عزیزم خانم مونا رزاقی، سحر اسدی و آقای شاهین محمدی و کلیه دوستانی که در گردآوری این پایان نامه مرا همراهی کردند تشکر می‌کنم.

فهرست مطالب

فصل اول - مفاهیم اولیه ۱

۱-۱ نظریه گراف ۱

۱-۱-۱ کاربرد گراف ها در زمینه های گوناگون ۲

۱-۱-۲ مفاهیم اولیه نظریه گراف ۳

۱-۲ شبکه‌ها ۸

۱-۲-۱ معیارهای بررسی شبکه‌ها ۹

۱-۲-۱-۱ توزیع درجات ۹

۱-۲-۱-۲ کوتاهترین مسیر ، میانگین طول مسیر و قطر شبکه‌ها ۱۰

۱-۲-۱-۳ ضریب دسته‌بندی ۱۰

۱-۲-۲ شبکه‌های پیچیده ۱۱

۱-۲-۳ مدل‌های شبکه‌های پیچیده ۱۲

۱-۲-۳-۱ شبکه‌های تصادفی ۱۳

۱-۲-۳-۲ شبکه های بدون اندازه ۱۵

۱-۲-۳-۳ شبکه‌های سلسله مراتبی ۱۶

۱-۳ سیستم‌های زیستی و شبکه‌ها ۱۷

۱-۳-۱ شبکه تنظیم/رونویسی ژن ۱۷

۱-۳-۲ شبکه متابولیک ۱۸

۱-۳-۳ شبکه عصبی ۱۹

۱-۳-۴ شبکه غذایی ۱۹

۱-۳-۵ شبکه تاثیر متقابل پروتئین-پروتئین ۱۹

۱-۴ موتیف‌ها و الگوریتم کلی یافتن آن‌ها ۲۰

۱-۴-۱ دسته‌بندی با فراوانی بالا در شبکه‌های زیستی ۲۱

۱-۴-۲ موتیف‌ها واحدهای اصلی شبکه‌های سلولی ۲۱

۱-۴-۳ الگوریتم کلی یافتن موتیف‌های شبکه ۲۲

فصل دوم - معرفی الگوریتم‌های یافتن موتیف‌های شبکه ۲۵
۲-۱ الگوریتم تشخیص موتیف‌های شبکه به وسیله شمارش کامل زیرگراف‌ها ۲۵
۲-۱-۱ الگوریتم شمارش تمام زیرگراف‌ها ۲۶
۲-۱-۲ بررسی شبکه‌های تصادفی ۲۷
۲-۱-۲-۱ مدلی مناسب برای تولید شبکه‌های تصادفی ۲۷
۲-۱-۲-۲ روش‌های تولید شبکه‌های تصادفی ۲۸
۲-۱-۳ تشخیص موتیف‌های شبکه توسط الگوریتم شمارش کامل ۲۹
۲-۱-۴ معیارهای آماری جهت تشخیص موتیف‌های شبکه ۳۰
۲-۱-۵ پیچیدگی زمانی الگوریتم ۳۱
۲-۱-۶ نتایج ۳۱
۲-۱-۷ بررسی اهمیت آماری موتیف‌ها با تغییر سایز شبکه ۳۴
۲-۱-۸ مزایا و کاستی‌ها ۳۴
۲-۲ الگوریتم تشخیص موتیف‌های شبکه به وسیله روش نمونه‌گیری زیرگراف ها ۳۶
۲-۲-۱ تخمین فراوانی زیرگراف‌ها با استفاده از روش نمونه‌گیری ۳۷
۲-۲-۱-۱ تراکم زیرگراف‌ها ۳۷
۲-۲-۱-۲ روش نمونه گیری زیرگراف‌ها ۳۷
۲-۲-۱-۳ تصحیحاتی بر روی الگوریتم نمونه‌گیری ۳۸
۲-۲-۱-۴ محاسبه فراوانی زیرگراف‌های با سایز n ۳۹
۲-۲-۲ بررسی شبکه‌های تصادفی ۳۹
۲-۲-۳ تشخیص موتیف‌های شبکه توسط روش نمونه‌گیری ۴۰
۲-۲-۴ معیارهای آماری جهت تشخیص موتیف‌های شبکه ۴۰
۲-۲-۵ پیاده‌سازی الگوریتم ۴۱
۲-۲-۶ بررسی تعمیم موتیف‌ها در شبکه‌ها ۴۲
۲-۲-۶-۱ تعمیم موتیف‌ها در شبکه تنظیم ژن ای.کُلای ۴۲
۲-۲-۶-۲ تعمیم موتیف‌ها در شبکه عصبی کرم سی.اِلگانس ۴۲

۲-۲-۷ پیچیدگی زمان الگوریتم نمونه‌گیری ۴۵
۲-۲-۸ مقایسه زمان اجرای روش نمونه‌گیری و الگوریتم جستجوی کامل ۴۶
۲-۲-۹ همگرایی الگوریتم نمونه‌گیری برای تشخیص موتیف‌ها ۴۷
۲-۲-۱۰ تخمین دقیق فراوانی زیرگراف‌ها و موتیف‌ها توسط روش نمونه‌گیری ۵۰
۲-۲-۱۱ مزایا و کاستی‌ها ۵۰

فصل سوم - معرفی سایر الگوریتم‌های یافتن موتیف‌های شبکه ۵۲
۳-۱ ارائه دیدگاهای مختلف برای تعیین فراوانی زیرگراف‌ها و یک الگوریتم جستجوی انعطاف‌پذیر در راستای تشخیص موتیف‌های شبکه ۵۲
۳-۱-۱ مفاهیم اولیه ۵۳
۳-۱-۲ دیدگاه‌های مختلف برای تشخیص فراوانی یک الگو ۵۳
۳-۱-۳ الگوریتم یافتن الگوهای رایج ۵۵
۳-۱-۳-۱ چگونگی پیمایش الگوها ۵۷
۳-۱-۳-۲ بالا بردن کارایی الگوریتم جستجو ۵۷
۳-۱-۴ پیاده‌سازی الگوریتم ۵۸
۳-۱-۵ پیچیدگی زمانی الگوریتم ۵۸
۳-۱-۶ مزایا و کاستی‌ها ۶۰
۳-۲ الگوریتمی سریع و موثر برای تشخیص موتیف‌های شبکه ۶۱
۳-۲-۱ نمادهای لازم ۶۲
۳-۲-۲ روش نمونه‌گیری زیرگراف‌ها بر اساس انتخاب تصادفی یال‌ها ۶۳
۳-۲-۳ تخمین تراکم زیرگراف‌ها براساس شمارش تصادفی ۶۴
۳-۲-۳-۱ شمارش تمام زیرگراف‌های با سایز k ۶۴
۳-۲-۳-۲ نمونه‌گیری یکنواخت زیرگراف‌های با سایز k ۶۸
۳-۲-۴ پیچیدگی زمانی الگوریتم ۷۰
۳-۲-۵ پیاده‌سازی الگوریتم ۷۱
۳-۲-۶ مزایا و کاستی‌ها ۷۲

فصل چهارم - الگوریتم یافتن موتیف‌های شبکه بدون محدودیت بر روی سایز آن‌ها ۷۴
۴-۱ الگوریتم تشخیص موتیف‌های شبکه بوسیله شمارش کامل زیرگراف‌ها ۷۴
۴-۱-۱ شمارش کامل زیرگراف‌های شبکه ۷۵
۴-۱-۲ کلاس‌بندی زیرگراف‌ها ۷۸
۴-۱-۳ تولید گراف‌های تصادفی ۸۰
۴-۱-۴ مشخص کردن موتیف‌های شبکه ۸۰
۴-۲ پیچیدگی زمانی الگوریتم ۸۱
۴-۳ نتایج ۸۱
۴-۴ مقایسه الگوریتم ارائه شده با الگوریتم‌های قبلی ۸۲
۴-۵ پیشنهادات ۸۳

مراجع ۸۴
واژه نامه فارسی به انگلیسی ۸۹
پیوست اول: ارگانیسم‌های معروف در مطالعات شبکه‌های زیستی ۱۰۵

فصل اول

مفاهیم اولیه

در این فصل مفاهیم اولیه مورد نیاز در این پایان‌نامه ارائه می‌گردد. در بخش اول، تعاریف مقدماتی نظریه گراف که در این پایان‌نامه مورد استفاده قرار گرفته‌اند، مطرح می‌شوند [۱۲,...,۲۲]. بخش دوم، مروری است بر مبانی شبکه‌ها و انواع آن‌ها [۲۳,...,۳۴]. در بخش سوم، به طور خاص شبکه‌های زیستی مورد بررسی قرار خواهند گرفت [۵,۲۴,۳۵,۳۶,۳۷] و نهایتاً در بخش چهارم، به معرفی موتیف‌ها و الگوریتم کلی یافتن آن‌ها پرداخته می‌شود [۴,...,۱۰,۲۴,۳۸].

۱-۱ نظریه گراف

امروزه گراف‌ها ساختمان‌داده‌های جهانی هستند. در حالت کلی، یک گراف می‌تواند یک شبکه از ارتباطات بین اشیاء را مدل کند. می‌توان ادعا کرد که تمام اشیایی که در جهان حقیقی وجود دارند یا یک شبکه از ارتباطات میان اجزایشان هستند و یا خود اجزای یک شبکه بزرگتر می‌باشند.

بوسیله گراف می‌توان هر یک از این دو مورد را تشریح کرد. در حالت اول یک گراف می‌تواند یک سیستم و ارتباطات بین اجزای آن سیستم را مدل کند و در حالت دوم تمام روابطی را که یک شی با بقیه اجزای سیستم دارد، نشان دهد [۱۲]. باید توجه کرد به وسیله گراف می‌توان نشان داد که دو جزء در یک سیستم بهم مرتبط هستند یا نه، ولی چگونگی این ارتباط را نمی‌توان مشخص کرد. بنابراین گراف یک ساختار ریاضی است که بوسیله آن می‌توان به بهترین نحو بسیاری از مسائل جهان را تشریح نمود.

معمولا گراف را با یک نمودار نمایش می‌دهند که در آن رئوس به صورت نقاط و یال‌ها بصورت خطوطی که این نقاط را بهم وصل می‌کنند و نشان‌دهنده ارتباط بین این نقاط هستند مشخص می‌شوند. برای مثال رئوس می‌توانند نشان‌دهنده افراد باشند و یال‌ها رابطه دوستی بین هر دو فرد [۱۳].

در این بخش با فرض اینکه خواننده با تعاریف مقدماتی گراف آشنایی دارد، تنها مرور کوتاهی روی این مفاهیم خواهد شد. پیشنهاد می‌شود جهت آشنایی بیشتر با مفاهیم مقدماتی گراف به [۱۴] رجوع شود.

۱-۱-۱ کاربرد گراف در زمینه‌های گوناگون

گراف‌ها ساختمان‌داده‌های رایجی برای مدل کردن شبکه‌ها می‌باشند. در دهه اخیر، استفاده از داده‌هایی با ساختار گراف در زمینه‌های کاربردی گوناگون افزایش یافته است. به عنوان مثال در زمینه‌های بیوانفورماتیک، بررسی شبکه‌های اجتماعی، بیولوژی و اینترنت و ... . دلیل آن این است که کلیه این زمینه‌ها، سیستم‌هایی را تشریح می‌کنند که شامل ارتباطاتی میان زیرساختارهایشان هستند. برای مثال یک شبکه اجتماعی، یک گروه از ارتباطات میان افراد است، یک شبکه تاثیر متقابل پروتئین- پروتئین، یک گروه از ارتباطات میان مولکول‌هاست، یا یک ساختار مولکولی، یک گروه از ارتباطات میان اتم‌ها می‌باشد و اینترنت یک شبکه از وب‌سایت‌های مرتبط بهم است. در نتیجه با انتخاب ساختمان‌داده گراف می‌توان به راحتی زیرساختارهای سیستم و ارتباطات میان این زیرساختارها را ذخیره و نگهداری نمود. در زیر خلاصه‌ای از کاربردهای گراف در این زمینه‌ها آورده شده است [۱۲]:

بیوشیمی: از زمان‌های گذشته، گراف‌ها در مدل کردن ترکیبات مولکولی شیمیایی کاربرد داشته‌اند [۱۵]. هدف بیوشیمی پیش‌بینی خصوصیت مولکول‌ها از روی ساختار آن‌ها است. مبدا بیشتر پایگاه‌داده‌های قدیمی مانند میوتاج[1] [۱۶] و پی.تی.سی[2] [۱۷] که برای الگوریتم‌های مربوط به استخراج گراف استفاده می‌شدند، از زمینه بیوشیمی می‌باشد.

بیوانفورماتیک: یکی از دلایل اصلی گرایش به ذخیره اطلاعات با ساختار گراف، حجم انبوه داده‌ها در زیست‌شناسی مولکولی می‌باشد. این داده‌ها شامل ساختارهای مولکولی، از آر.اِن.اِی[3] گرفته تا پروتئین‌ها و یا شبکه‌هایی مانند شبکه تاثیر متقابل پروتئین-پروتئین، شبکه‌های متابولیکی و شبکه‌های تنظیم ژن می‌باشند که می‌توان آن‌ها را بوسیله گراف مدل کرد. بیوانفورماتیک در حال جستجو برای یافتن عملکرد این شبکه‌ها و ساختارها است. امروزه، یک روش موفق برای پیش‌بینی طرزکار ساختارها، جستجو برای پیدا کردن شباهت میان ساختار مورد نظر با ساختارهایی است که عملکرد آن‌ها شناخته شده‌اند. برای مثال برای پیش‌بینی طرزکار یک ساختار پروتئینی جدید، می‌توان ساختار آن را با پایگاه‌داده‌ای از پروتئین‌هایی که کاربرد آن‌ها قبلا مشخص شده است، مقایسه نمود و پیش‌بینی کرد که طرزکار این پروتئین مانند طرز کار کدامیک از پروتئین‌هایی است که شباهت ساختاری زیادی به آن دارند. مدل‌های تکامل نیز این مفهوم را پشتیبانی می‌کنند، این مدل‌ها بیان کننده این امر هستند که پروتئین‌هایی که دارای ساختارهای توپولوژیکی مشابهی می‌باشند به احتمال زیادی دارای جد مشترک و در نتیجه عملکرد مشابهی هستند [۱۸].

---

[1] MUTAG
[2] PTC
[3] RNA

شبکه‌های تحلیل اجتماعی: این شبکه‌ها نیز نمونه‌ای از داده‌های با ساختار گراف می‌باشند. در این شبکه‌ها رئوس گراف نشان‌دهنده افراد هستند و یال‌ها نشان‌دهنده روابط میان آن‌ها می‌باشند. بررسی این شبکه‌ها هم از دیدگاه علمی و هم از دیدگاه تجاری مورد توجه است [۱۹].

اینترنت: چهارمین زمینه کاربرد گراف مدل کردن شبکه اینترنت می‌باشد. صفحات اِچ.تی.اِم.اِل[۴] رأس‌های این شبکه هستند و اَبَرپیوندها این رأس‌ها را بهم متصل می‌کنند. گوگِل[۵] نیز در الگوریتم معروف خود (پِیج‌رَنک[۶] [۲۰])، که برای رتبه‌بندی صفحات وب طراحی شده از این ساختار ارتباطی استفاده کرده است.

۱-۱-۲ مفاهیم اولیه نظریه گراف

گراف

یک گراف عبارتست از یک زوج $G=(V,E)$، که در آن V یک مجموعه ناتهی و با پایان از رأس‌ها یا گره‌ها و $E \subseteq V.V$، مجموعه‌ای از یال‌ها یا کمان‌ها است. به طوریکه هر یال $e=(u,v)$، دو رأس u و v از گراف را بهم متصل می‌کند. در حالت کلی، $V(G)$ مجموعه رئوس گراف G را نمایش می‌دهد و $E(G)$ مشخص کننده مجموعه یال‌های گراف G می‌باشد.

اندازه گراف

تعداد رئوس گراف $G=(V,E)$، نشان‌دهنده اندازه این گراف است که آن را با $|V|$ یا N نمایش می‌دهند. تعداد یال‌های گراف G را نیز با $|E|$ نشان می‌دهند.

گراف G' بزرگتر از گراف G است هرگاه $|V(G')|>|V(G)|$.

به گراف G اسپارس[۷] (خالی) گفته می‌شود هرگاه $|E| \approx N^2$.

گراف برچسب‌دار

یک گراف برچسب‌دار، یک سه تایی $G(V,E,L)$ می‌باشد. به طوریکه (V,E) نمایش‌دهنده یک گراف است و $L: V \cup E \to Z$، یک نگاشت از مجموعه رئوس V و یال‌های E به مجموعه رئوس و یال‌های برچسب خورده با اعداد حقیقی Z، می‌باشد. به گرافی که رئوس آن دارای برچسب باشند، گراف برچسب‌دار رأسی و به گرافی که یال‌های آن دارای برچسب باشند، گراف برچسب‌دار یالی می‌گویند. گراف وزن‌دار یک گراف برچسب‌دار یالی است.

---

[4] HTML
[5] Google
[6] Page Rank
[7] sparse

گراف جهت‌دار و بدون جهت

در گراف داده شده $G=(V,E)$، هرگاه به یال‌ها جهت داده شود، گراف جهت‌دار نامیده می‌شود. در یک گراف جهت‌دار، رابطه $(v_i,v_j)\neq(v_j,v_i)$ به ازای $\forall(v_i,v_j)\in V$ برقرار است. گراف G بدون جهت نامیده می‌شود، هرگاه به ازای $\forall(v_i,v_j)\in V$ داشته باشیم $(V_i,V_j)\in E\Leftrightarrow(V_j,V_i)\in E$.

همسایگی درگراف

دو رأس v_i و v_j از مجموعه رئوس گراف G را مجاور یا همسایه می‌نامند، هرگاه (v_i,v_j) یک یال از G باشد. مجموعه تمام رئوس مجاور رأس v_i با $N_G(v_i)$ نمایش داده و آن را مجموعه همسایگان رأس v_i می‌نامیم.

دو یال متمایز e_i و e_j در گراف G مجاور هستند اگر دارای یک رأس مشترک باشند.

اگر هر دو رأس دلخواه G با هم مجاور باشند، گراف G را کامل می‌نامیم.

درجه یک رأس در گراف

در گراف بدونِ جهت G، تعداد یال‌های گذرنده از یک رأس را، درجه آن رأس می‌گویند. درجه رأس v_i را با $d(v_i)$ نمایش می‌دهند. رأسی که هیچ همسایه‌ای نداشته باشد، یک رأس پایانی نامیده می‌شود.

کمترین درجه گراف G با رابطه زیر تعریف می‌کنند:

$$\delta(G):=\min\{d_G(v)\mid v\in V\} \qquad (۱-۱)$$

بیشترین درجه گراف G نیز با رابطه زیر تعریف می‌شود:

$$\Delta(G):=\max\{d_G(v)\mid v\in V\} \qquad (۱-۲)$$

میانگین درجات گراف G از رابطه زیر به دست می‌آید:

$$d_G(G):=\frac{1}{|V|}\sum_{v\in V}d_G(v) \qquad (۱-۳)$$

در گراف جهت‌دار G درجه ورودی هر رأس با $in\deg_G(v)$ و درجه خروجی هر رأس با $out\deg_G(v)$ نشان داده می‌شوند، و به ترتیب برابر با تعداد یال‌هایی هستند که به رأس v وارد و از رأس v خارج شده و به رئوس دیگر می‌روند.

گشت، مسیر، دور

یک گشت w به طول $\ell-1$ در گراف G عبارتست از یک دنباله غیر تهی $(v_1,e_1,v_2,e_2,...,v_{\ell-1},e_{\ell-1},v_\ell)$ از رئوس و یال‌های گراف، بطوری که در آن $e_i=(v_i,v_{i+1})$ و $1\le i\le\ell-1$. اگر در این گشت $v_1=v_\ell$ باشد، آنگاه گشت بسته است.

اگر همه رئوس گشت w از هم متمایز باشند، یک مسیر $p=(v_1,...,v_\ell)$ در گراف G تعریف می‌شود.

اگر در مسیر p ، $v_1 = v_\ell$ باشد، یک دور به دست می‌آید. دور به طول k را معمولا با C_k نمایش می‌دهند.

گراف همبند

گراف G را همبند می‌نامند، هرگاه بین هر دو رأس دلخواه آن حداقل یک مسیر وجود داشته باشد. در غیر این صورت گراف ناهمبند نامیده می‌شود. هر گراف ناهمبند از تعدادی گراف همبند که مولفه‌های همبندی آن نام دارند، تشکیل می‌شود.

ماتریس مجاورت

ماتریس مجاورت $A=(A_{ij})_{n\times n}$ در گراف $G=(V,E)$ به صورت زیر تعریف می‌شود، که در آن v_i و v_j رأس‌های گراف هستند.

$$A_{ij} := \begin{cases} 1 & if\,(v_i,v_j)\in E \\ 0 & ow. \end{cases} \qquad (۴-۱)$$

زیرگراف، زیرگراف القایی، کلیک

گراف $G'=(V',E')$ یک زیرگراف از گراف $G=(V,E)$ است اگر $V'\subseteq V$ و $E'\subseteq((V'.V')\bigcap E)$ باشد و با نماد $G'\subseteq G$ نشان داده می‌شود. در این صورت به گراف G فوق‌گراف G' گفته می‌شود.

اگر زیرگراف G' یک زیرگراف کامل باشد، به آن کلیک گفته می‌شود.

اگر در G'، $E'=((V'.V')\bigcap E)$ باشد، یعنی شامل تمام یال‌هایی از G باشد که هر دو رأس در V' را به هم وصل می‌کنند، آنگاه G' زیرگراف القایی G خوانده می‌شود.

هم‌ریختی گراف‌ها

برای چک کردن اینکه دو گراف شبیه هم هستند یا نه، نمی‌توان ماتریس مجاورت آن‌ها را با هم مقایسه کرد. زیرا ماتریس مجاورت هنگامی که ترتیب دو رأس عوض شود، تغییر می‌کند. در نتیجه برای فهمیدن شباهت دو گراف از مفهوم هم‌ریختی استفاده می‌شود. دو گراف $G=(V,E)$ و $G'=(V',E')$ هم‌ریخت نامیده می‌شوند، هرگاه یک نگاشت دو سویه $f:V\to V'$ وجود داشته باشد بطوری که رابطه

$$\forall v,v'\in V \;:\; (v,v')\in E \Leftrightarrow (f(v),f(v'))\in E'$$

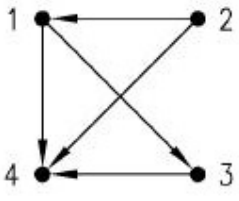
برقرار باشد و به صورت $G\approx G'$ نشان داده می‌شود. نگاشت f یک هم‌ریختی نامیده می‌شود. هم‌ریختی یک گراف با خودش، یک‌ریختی نامیده می‌شود.

برای بررسی این که آیا دو گراف G و G' هم‌ریخت هستند، باید تمام جایگشت‌های رئوس گراف G' در نظر گرفته شده و بررسی شود که آیا هیچ کدام از آن‌ها شبیه به G است یا خیر. برای مسئله هم‌ریختی دو گراف تاکنون هیچ الگوریتمی که قابل اجرا در زمان چند جمله‌ای باشد ارائه

نشده است [۲۱]. مسئله هم‌ریختی گراف‌ها یک مسئله NP است ولی هنوز ثابت نشده که این مسئله از کلاس P می‌باشد یا $NP-Complete$ [۱۰].

برای اینکه دوگراف هم‌ریخت باشند، باید تمام خصوصیات نامتغیر آن‌ها یکسان باشند. بنابراین برای اثبات اینکه دو گراف ناهم‌ریخت هستند گاهی تنها کافی است یک سری آزمون ساده انجام شود، مانند بررسی اندازه گراف‌ها و تعداد یال‌هایشان. اما اگر این دو معیار در دو گراف یکسان باشند، باید از مقایسه دنباله درجات رئوس و یا مقایسه طول کوتاهترین مسیر در گراف‌ها استفاده نمود، که این معیارها دارای پیچیدگی محاسباتی بالاتری هستند. بنابراین هر چه دو گراف بیشتر به هم شبیه باشند این آزمون‌ها ناکارآمدتر می‌شوند. یک روش کارآمد برای اثبات اینکه دوگراف ناهم‌ریخت هستند، بررسی و مقایسه خصوصیات نامتغیر آن‌ها به ترتیب افزایش پیچیدگی محاسباتی این خصوصیات می‌باشد. اگر گراف‌ها تنها در یکی از این خصوصیات متفاوت باشند، دیگر نمی‌توانند هم‌ریخت باشند.

سریع‌ترین برنامه‌ای که امروزه برای تست هم‌ریختی گراف‌ها وجود دارد، ناتی[۸] است که توسط مَک‌کِی[۹] ارائه شده است [۲۲]. دراین الگوریتم ابتدا یک سری آزمون از پیش تعریف شده برای رد کردن هم‌ریختی دو گراف انجام می‌شود. اگر تمام این آزمون ها با شکست مواجه شوند، آنگاه جایگشت‌های رئوس بررسی می‌شوند. در این الگوریتم یک شناسه منحصر به فرد، به هر کلاس هم‌ریختی اختصاص داده می‌شود. این شناسه، برچسبِ استاندارد نامیده می‌شود. برچسبِ استاندارد، رشته‌ای است که با کنار هم قرار دادن سطرهای ماتریس مجاورت بدست می‌آید. بدلیل اینکه ترتیب مختلف رئوس باعث تولید رشته‌های مختلفی می‌شوند، از میان تمام جایگشت‌های ممکن رئوس، ترتیبی به عنوان برچسبِ استاندارد انتخاب می‌شود که در آن رشته تولید شده در بزرگترین یا کوچکترین ترتیب قاموسی باشد. به عنوان مثال در شکل ۱-۱، یک گراف به همراه ماتریس مجاورتش نشان داده شده است، شناسه استاندارد مربوط به آن، رشته ۰۱۰۱۰۰۱۱۰۰۰۱۰۰۰۰ می‌باشد که با ترتیب رئوس به صورت (۴،۳،۱،۲) بدست آمده و از نظر ترتیب قاموسی بزرگترین رشته در میان تمام رشته‌های ممکن بدست آمده در جایگشت‌های مختلف رئوس است.



	1	2	3	4
1	0	0	1	1
2	1	0	0	1
3	0	0	0	1
4	0	0	0	0

شکل ۱-۱. نمایش یک گراف به همراه ماتریس مجاورتش.

[8] NAUTY
[9] McKay

ماکزیمم مجموعه مستقل

مجموعه‌هایی از رئوس و یال‌ها را مستقل می‌نامند، در صورتیکه هیچ کدام از اعضای آن‌ها با هم مجاور نباشند. ماکزیمم مجموعه مستقل S از گراف $G=(V,E)$، به صورت بزرگترین زیرمجموعه از رئوس V تعریف می‌شود، به طوریکه هیچ دو رأسی از S یک یال را در E تعریف نکنند. از آنجاییکه مسئله یافتن ماکزیمم مجموعه مستقل، یک مسئله $NP-Complete$ است، برای حل آن به منظور کارایی بهتر از یک الگوریتم اکتشافی برای تخمین جواب دقیق مسئله استفاده می‌شود. یک الگوریتم اکتشافی رایج برای این حل این مسئله، استفاده از الگوریتم حریصانه است که در هر مرحله یک رأس با کمترین درجه را از گراف انتخاب کرده و آن را به مجموعه مستقل اضافه می‌کند، سپس آن رأس و تمام همسایگانش را از گراف اصلی حذف می‌نماید و این روند را تا جایی ادامه می‌دهد که گراف تهی شود، سپس مجموعه بدست آمده را به عنوان خروجی برمی‌گرداند.

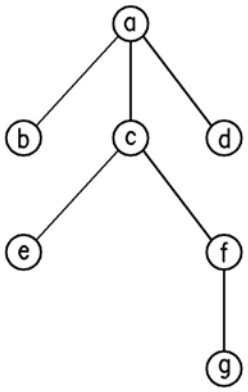
درخت

هرگراف بدون جهت هم‌بند که فاقد دور باشد، درخت نامیده می‌شود، که آن را با نماد T نمایش می‌دهند. هر درخت با n گره، $n-1$ یال دارد و دارای حداقل دو گره پایانی می‌باشد. هر دو گره یک درخت دقیقا با یک مسیر بهم متصل می‌شوند.

در هر درخت، معمولا یک گره خاص به عنوان ریشه در نظر گرفته می‌شود که آن را با نماد r نشان می دهند. در صورتیکه ریشه از پیش تعیین نشده باشد، یک گره دلخواه به عنوان ریشه انتخاب می‌گردد. از آنجاییکه در یک درخت از ریشه به هر گره تنها یک مسیر وجود دارد، می‌توان یک جهت فرضی برای هر یال، از سمت ریشه به سایر گره‌ها در نظر گرفت. این درخت یک گراف جهت‌دار را می‌سازد که به آن درختِ ریشه‌دار گفته می‌شود.

اگر $v\neq r$ گرهی از درخت T باشد، در این صورت پدر v، گره منحصر به فرد u است که از آن یالی به سمت v وجود دارد. به گره v فرزند u گفته می‌شود. گره‌ای که هیچ فرزندی نداشته باشد یا به عبارت دیگر درجه آن یک باشد، یک گره پایانی است که برگ نامیده می‌شود. به گره‌هایی از درخت با درجه بیشتر از یک، گره‌های میانی گفته می‌شود. گره‌هایی از درخت که دارای پدر مشترک باشند، خواهر خوانده می‌شوند. اجداد v، گره‌هایی متمایز از v هستند که در مسیر ریشه به این گره قرار دارند و شامل ریشه هم می‌شوند. به گره‌هایی از درخت که v از اجداد آن‌هاست، نوادگان v گفته می‌شود.

شکل ۱-۲ یک درخت را نشان می‌دهد. ریشه این درخت گره a است و برگ‌های آن گره‌های b,d,e,g و سایر گره‌ها میانی هستند. گره f یک گره میانی است که فرزند گره c و خواهر گره e و پدر گره g می‌باشد. اجداد گره f عبارتند از a و c.



شکل ۱-۲. یک درخت ریشه‌دار.

زیردرخت

اگر v گره‌ای از درخت T باشد، آنگاه به زیرگرافی از درخت T که شامل گره v، نوادگان این گره و تمام یال‌های واقع بر آن باشد، زیردرخت T با ریشه v گفته می‌شود و با نماد $T_v \subseteq T$ نشان داده می‌شود.

عمق، ارتفاع و سطح درخت

عمق گره v در درخت برابر است با طول مسیر ریشه تا v، و با $d_T(v)$ نمایش داده می‌شود. عمق ریشه درخت صفر است.

سطح در درخت از روی عمق گره‌ها تعریف می‌شود. گره‌هایی که دارای عمق یکسان هستند، هم سطح‌اند.

ارتفاع یک درخت طول بلندترین مسیر موجود از ریشه تا برگ‌هاست که با $h(T)$ نمایش داده می‌شود.

در شکل ۱-۲، عمق گره f برابر دو است و ارتفاع این درخت سه می‌باشد. در این درخت دو گره e و f هم سطح‌اند.

۱-۲ شبکه‌ها

شبکه‌ها سیستم‌هایی هستند که در آن‌ها اجزا و نحوه ارتباط آن اجزا با یکدیگر نشان داده شده‌اند. همانطور که گفته شد، می‌توان ادعا کرد که تمام اشیایی که در دنیای حقیقی وجود دارند یا یک شبکه از ارتباطات میان اجزایشان هستند و یا خود اجزای یک شبکه بزرگتر می‌باشند. در یک شبکه هر جزء را با یک رأس نمایش می‌دهند و چنانچه بین دو جز رابطه‌ای وجود داشته باشد، یالی میان رأس‌های متناظر آن‌ها در نظر گرفته می‌شود. این یال‌ها بر اساس ساختار شبکه، می‌توانند جهت‌دار یا بدون جهت در نظر گرفته شوند. در نتیجه گراف‌های جهت‌دار یا بدون جهت ساختمان‌داده مناسبی برای نمایش شبکه‌ها می‌باشند. نمایش دادن تصویری فرضی از نحوه ارتباطات اجزای یک سیستم

بدون استفاده از شبکه‌ها به‌راحتی قابل درک نیست. همچنین استفاده از شبکه‌ها، امکان پیش‌گویی رفتارهای مختلف شبکه را نیز فراهم می‌کند و اگر این پیش‌گویی‌ها قابل مقایسه با نتایج تجربی باشند می‌توان به صحت مدل پی برد و یا بر اساس نتایج مقایسه، مدل را تصحیح نمود.

پس از مطرح شدن بحث شبکه‌ها در اواخر قرن بیستم، تحقیقاتی در زمینه شبکه‌های پیچیده آغاز شد. ایده اصلی مطالعه روی این شبکه‌ها، به دست آوردن توانایی برای تحلیل سیستم‌هایی با هزاران و میلیون‌ها رأس به جای تحلیل شبکه‌های کوچک و مورد توجه قرار دادن خواص شبکه‌های دارای اجزای پویا بود [۲۳]. در ادامه ابتدا به بررسی معیارهای شبکه‌ها می‌پردازیم و سپس توضیحاتی راجع به شبکه‌های پیچیده ارائه می‌دهیم.

۱-۲-۱ معیارهای بررسی شبکه‌ها

خصوصیات یک شبکه عموما براساس دو نوع معیار عمومی و محلی مورد بررسی قرار می‌گیرند. خصوصیات عمومی شبکه به کسب یک بینش کلی در مورد شبکه کمک می‌کنند و وابسته به ساختار شبکه هستند. از مهمترین معیارهای عمومی شبکه‌ها می‌توان تابع توزیع درجات، کوتاهترین مسیر و ضریب دسته‌بندی را نام برد [۲۴]، که در ادامه به توضیح این معیارها پرداخته می‌شود. خصوصیات محلی، به بررسی خصوصیات شبکه در سطوح جزئی‌تر می‌پردازد. معیارهای محلی عموما زیر ساختارهای کوچک و محلی را می‌سنجد که به آن‌ها موتیف[۱۰] گفته می‌شود. در این پایان‌نامه نیز به بررسی موتیف‌ها و الگوریتم‌های یافتن آن‌ها می‌پردازیم.

۱-۲-۱-۱ توزیع درجات

اگر یک شبکه با گراف $G=(V,E)$ مدل شود و در آن درجه هر رأس با نماد k نشان داده شود، در این صورت میانگین درجات شبکه با نماد $<k>$ مشخص می‌شود و به صورت زیر محاسبه می‌گردد:

$$<k>=2.|E|/|V| \qquad (۵-۱)$$

میانگین درجات رئوس، یک معیار اندازه‌گیری برای بررسی خصوصیات عمومی شبکه می‌باشد و نشان می‌دهد در آن شبکه هر رأس به‌طور متوسط با چند رأس دیگر ارتباط دارد.

در شبکه‌ها به رأسی که درجه آن نسبت به سایر رأس‌ها به اندازه قابل توجهی بیشتر باشد، هاب[۱۱] گفته می‌شود.

توزیع درجات رئوس را با $P(k)$ نشان می‌دهند که بیان‌کننده این است که یک رأس دلخواه در شبکه، با چه احتمالی دارای درجه k می‌باشد. $P(k)$ با بدست آوردن تعداد رئوس با درجه k یعنی $N(k)$ و تقسیم آن به تعداد کل رأس‌های شبکه محاسبه می‌شود:

$$P(k)=N(k)/|V| \qquad (۶-۱)$$

---

[10] Motif
[11] Hub

توزیع درجات برای کلاس‌بندی شبکه‌ها به کار می‌رود. به عنوان مثال توزیع درجات نرمال در شبکه‌های تصادفی که در ادامه به بررسی آن‌ها می‌پردازیم، نشان‌دهنده توزیع یکنواخت درجات رأس‌های این شبکه‌ها می‌باشد و مشخص می‌کند که این شبکه‌ها دارای هاب نیستند. در مقابل اگر توزیع درجات شبکه‌ای از قانون توان تبعیت کند، نشان‌دهنده این است که در آن شبکه تعداد کمی هاب وجود دارد که به یکدیگر و به تعداد زیادی از رأس‌های با درجه کم وصل هستند [۲۴].

۱-۲-۱-۲ کوتاهترین مسیر ، میانگین طول مسیر

در شبکه‌ها برای حرکت بین دو رأس ممکن است مسیرهای زیادی وجود داشته باشد. در میان این مسیرها، طول کوتاهترین مسیر به عنوان فاصله دو رأس در نظر گرفته شده و با ℓ نمایش داده می‌شود. در شبکه‌های جهت‌دار فاصله بین دو رأس دلخواه، بسته به جهت حرکت بین آن دو رأس می‌تواند تفاوت داشته باشد. به بیشترین فاصله میان رأس‌ها در شبکه، قطر شبکه گفته می‌شود.

میانگین طول مسیر با $< \ell >$ نشان داده می‌شود که مشخص کننده میانگین طول کوتاهترین مسیرهای بین هر دو رأس در شبکه است و برای اندازه‌گیری قابلیت جابجایی در شبکه‌ها بکار می‌رود [۲۴].

پدیده دنیای کوچک

طبق مطالعاتی که روی شبکه‌ها انجام شده است، این نتیجه به دست آمده که میان بخش‌های مختلف شبکه‌ها، مسیرهای میان‌بُری وجود دارد. این مسیرها، اتصالاتی از شبکه هستند که دو سر آن‌ها هاب یا رأس‌های با درجه زیاد وجود دارد، مانند پل‌ها. این مسیرهای میان‌بُر باعث ایجاد راه‌های ارتباطی سریع بین رئوس می‌شوند. یک خصیصه مشترک میان اکثر شبکه‌ها صرف‌نظر از سایز آن‌ها، کوتاه بودن طول مسیرهای میان رأس‌های آن‌ها است. از این خصوصیت مشترک تحت عنوان پدیده دنیای کوچک نام برده می‌شود. این پدیده ابتدا در شبکه‌های اجتماعی دیده شد و سپس وجود آن در بسیاری شبکه‌های دیگر از جمله شبکه‌های زیستی گزارش شد. این پدیده یک خصوصیت بارز برخی از مدل‌های شبکه، از جمله شبکه‌های تصادفی است [۲۴]. طبق این خصوصیت گفته می‌شود در شبکه‌های اجتماعی از هر رأس می‌توان با حداکثر شِش اتصال، به هر رأس دیگری در شبکه متصل شد. در شبکه‌های وب این عدد به ۱۳ می‌رسد. در سلول برای اولین بار پدیده دنیای بسیار کوچک در مسیرهای متابولیسم گزارش شد که طبق آن بیشتر رأس‌ها با حداکثر سه یا چهار اتصال به‌هم مرتبط هستند.

۱-۲-۱-۳ ضریب دسته‌بندی

در بسیاری از شبکه‌ها، اگر رأس u به رأس v و رأس v به رأس z وصل باشد، با احتمال بالایی رأس u به طور مستقیم به رأس z متصل است. به این رابطه، رابطه مثلثی گفته می‌شود. این رابطه