سیستم‌های پیشنهاددهنده در شبکه‌های اجتماعی

نویسندگان

مهندس آتوسا مهران‌فر

دکتر مریم رستگارپور

ویراستار

مهندس عباس مسلمانی

سرشناسه	:	مهران‌فر، آتوسا، ۱۳٦۷ –
عنوان و نام پدیدآور	:	سیستم‌های پیشنهاددهنده در شبکه‌های اجتماعی/ نویسنده آتوسا مهران‌فر؛ ویراستار عباس مسلمانی.
مشخصات نشر	:	اهواز: انتشارات علوم و فنون پزشکی اهواز، ۱۳۹۵.
مشخصات ظاهری	:	۱۲۰ص.
شابک	:	978-600-7937-86-0
وضعیت فهرست نویسی	:	فیپا
موضوع	:	نظام‌های توصیه‌گر
موضوع	:	Recommender systems (Information filtering)
موضوع	:	سامانه‌های پالایش اطلاعات
موضوع	:	Information filtering systems
موضوع	:	شبکه‌های اجتماعی -- الگوهای ریاضی
موضوع	:	Social networks -- Mathematical models
شناسه افزوده:		مریم، رستگارپور،۱۳٦۰
شناسه افزوده	:	مسلمانی، عباس، ۱۳٦٤ –، ویراستار
رده بندی کنگره	:	QA۹/۷٦/ن۵۷م۹ ۱۳۹۵
رده بندی دیویی	:	۰۰۵/٤۳۷
شماره کتابشناسی ملی	:	٤٤۸۷۲۸٦

نویسندگان: آتوسا مهران فر،مریم رستگارپور

انتشارات:علوم و فنون پزشکی اهواز

ویراستار:عباس مسلمانی

طراح جلد: مروارید مهران فر

قیمت: ۲۵۰۰۰ تومان

نوبت چاپ: اول ۱۳۹۵

تیراژ: ۱۰۰۰نسخه

آدرس انتشارات:اهواز-امانیه- خیابان سقراط غربی جنب مجتمع تصادفات دادگستری-کوچه خاکسار- پلاک ۱۲۰

اخطار!

تقدیم به

پدر و مادر مهربانم

که در تمام مراحل زندگی کنارم بوده‌اند.

چکیده

امروزه ما در عصر شبکه‌ها زندگی می‌کنیم؛ عصری که در آن شکل‌گیری شبکه‌های گوناگون اجتماعی، شیوه‌های ارتباطی و اطلاع‌رسانی نوینی به عرصه‌ی گسترده‌ی ارتباطات اجتماعی معرفی کرده است. شبکه‌های اجتماعی به‌طور فزاینده‌ای محبوب شده‌اند، زیرا کانال‌های ارتباطی سریعی هستند که افراد می‌توانند از طریق وسایل الکترونیکی شخصی در هر مکان و در هر زمان با دوستان، آشنایان و همکاران و ... ارتباط برقرار کنند. با ظهور web2 و گسترش شبکه‌های اجتماعی در سال‌های اخیر محققین به منبع اطلاعاتی دیگری برای بهبود کیفیت پیشنهادها پی بردند که همان اطلاعات موجود در شبکه‌های اجتماعی بود و بر همین اساس کارهای تحقیقاتی زیادی در این حوزه شکل ریزی گردید. برای مدل‌سازی این شبکه‌ها، از گراف استفاده شد که در آن افراد گره‌ها را تشکیل می‌دهند و روابط بین افراد با یال‌ها نمایش داده می‌شود. در این میان مطالعه‌ی ساختارها، شیوه‌ها، مزیت‌ها و تأثیرهای ارتباط در شبکه‌های اجتماعی اهمیتی فراوان یافته است و همچنین تحلیل شبکه‌های اجتماعی حاصل از این ارتباطات روزبه‌روز از اهمیت بیشتری برخوردار می‌شود .تحلیل شبکه اجتماعی عبارت است از نگاشت و اندازه‌گیری روابط و همکاری در بین افراد، گروه‌ها، سازمان‌ها و هر موجودیتی که قابلیت پردازش اطلاعات و دانش داشته باشد. یکی از مهم‌ترین جنبه‌ها در تحلیل شبکه‌های اجتماعی استخراج جوامع در این شبکه‌ها است. تاکنون روش‌های گوناگونی برای استخراج جوامع از روی ساختار گراف اجتماعی ارائه‌شده است و در برخی از روش‌ها محتوای ارتباطی بین بازیگران نیز در استخراج جوامع موردتوجه قرار گرفته است. هدف آن شناسایی یال‌هایی است که در پیوند بین جوامع شبکه‌های اجتماعی مهم‌تر و مؤثرتر هستند .برای شناسایی و تشخیص جوامع در شبکه‌های اجتماعی با استفاده از الگوریتم میانگی یال، روشی پیشنهاد می‌دهیم به‌طوری‌که هر یال بین دو گره، با دریافت یک درجه عضویت در بازه‌ی $[0,1]$ ، میزان تأثیر و نفوذش در تشخیص جوامع با بقیه متفاوت است. میانگی یال، یالی است که در محاسبه کوتاه‌ترین مسیر بین جفت گره‌ها در شبکه، بیشترین دفعات از آن یال عبور می‌شود. همچنین پس از آن خواهیم توانست به کمک الگوریتم فازی میانگی یال، جوامع را با دقت بالاتری کشف و بررسی کنیم و میزان چگالی هر خوشه را با توجه به ساختار کشف‌شده بهبود دهیم. پس از آشنایی با شبکه‌های اجتماعی و بررسی

ویژگی‌های رفتاری و ساختاری این شبکه‌ها، به بررسی و مطالعه مرکزیت در شبکه‌های اجتماعی خواهیم پرداخت.

فهرست مطالب

مقدمه

با توجه به رشد روزافزون اطلاعات، توانایی یک سایت در پاسخگویی به بازدیدکنندگان و هدایت موفقیت‌آمیز آن‌ها به سمت اطلاعات مفید و مناسب، عاملی کلیدی در موفقیت نهایی سایت‌ها و کسب‌وکارهایشان به‌حساب می‌آید؛ بنابراین داشتن یک سیستم هوشمند که قادر باشد علایق کاربران را یاد بگیرد و براساس این علایق به‌طور خودکار علایق غیرمرتبط را فیلتر کند یا اطلاعات مرتبط را در زمان کم به کاربر پیشنهاد دهد، ضروری است. سیستم‌های شخصی‌سازی وب با ایجاد پیشنهادهایی که با سلایق کاربران مرتبط باشند (تولید یکسری پیشنهادهای شخصی)، به کاربران کمک می‌کنند تا بدون درخواست صریح آن‌ها، اطلاعات موردنیاز خود را بیابند. کاوش استفاده از وب از تکنیک‌های مفید برای ساخت سیستم‌های شخصی‌سازی وب است که از فایل‌های ثبت وقایع سرویس‌دهنده‌ی وب جهت ارائه پیشنهادهای آتی به کاربر جاری استفاده می‌کند. سیستم‌های شخصی‌سازی وب متفاوتی جهت پیش‌بینی صفحات درخواستی آینده‌ی کاربر با بهره‌گیری از کاوش استفاده از وب پیشنهاد شده است، اما این سیستم‌ها دارای دقت و پوشش مناسبی در ارائه‌ی پیشنهادها به کاربران نیستند.

سیستم‌های پیشنهاددهنده سیستم‌هایی هستند که در پیدا کردن و انتخاب نمودن آیتم‌های موردنظر کاربران به آن‌ها کمک می‌کنند. طبیعی است که این سیستم‌ها بدون در اختیار داشتن اطلاعات کافی و صحیح در مورد کاربران و آیتم‌های موردنظر آن‌ها قادر به پیشنهاد دادن نمی‌باشند؛ بنابراین یکی از اساسی‌ترین اهداف آن‌ها جمع‌آوری اطلاعات گوناگون در رابطه با سلایق کاربران و آیتم‌های موجود در سیستم است. منابع و روش‌های گوناگونی برای جمع‌آوری چنین اطلاعاتی وجود دارد. یک روش جمع‌آوری اطلاعات به‌صورت صریح که در آن کاربر صراحتاً اعلام می‌کند که به چه چیزهایی علاقه دارد. روش دیگر روش ضمنی است که کمی دشوارتر است و در آن سیستم باید سلایق کاربر را با کنترل و دنبال کردن رفتارها و فعالیت‌های او بیابد علاوه بر اطلاعات ضمنی و صریح برخی از سیستم‌ها نیز هستند که از اطلاعات شخصی کاربران استفاده می‌کنند. به این‌گونه از اطلاعات Demographic Information گفته می‌شود که گروهی از سیستم‌های پیشنهاددهنده مبتنی بر همین اطلاعات بناشده‌اند.

انسان‌ها با افراد غریبه بیشتر از سایر موجودات دیگر همکاری کرده و اطلاعات مبادله می‌کنند و این امر همراه با ورود به جوامع است و از جمع این جوامع یک شبکه جهان‌گستر تشکیل می‌شود. جامعه‌گرایی افراد موجب می‌شود که آن‌ها بتوانند در فعالیت‌های معدودی متخصص شوند و آن‌ها را به دلیل ارضای نیازها و باقی خواسته‌هایشان به دیگران وابسته می‌نماید. در نتیجه افراد را می‌توان اتم‌های شناوری در بازارهایی که در آن کالا و خدمات مبادله می‌نمایند، به‌حساب آورد؛ اما در بازبینی‌های دقیق‌تر می‌توان گفت که انسان‌ها تا حدودی ترجیح می‌دهند خودشان را به سایر افراد در گروه‌ها و اجتماع‌هایی از همه نوع نظیر خانواده‌ها، سکونتگاه‌ها، مذهب‌ها، سازمان‌ها و گاهی نیز اجتماع‌های مجازی پیوند بزنند. شبکه‌ی اجتماعی، یک ساختار اجتماعی ساخته‌شده از افراد یل سازمان‌هایی است که گره‌ها را تشکیل می‌دهند .این گره‌ها به‌وسیله‌ی انواع خاصی از وابستگی‌های متقابل مانند دوستی، خویشاوندی، منافع مشترک ... به هم متصل شده‌اند. تحلیل شبکه اجتماعی عبارت است از نگاشت و اندازه‌گیری روابط و همکاری در بین افراد، گروه‌ها، سازمان‌ها و هر موجودیتی که قابلیت پردازش اطلاعات و دانش داشته باشد. مسائل مربوط به حوزه‌ی شبکه‌های اجتماعی تحت تأثیر علومی چون جامعه‌شناسی، انسان‌شناسی، روانشناسی، ارتباطات، فناوری اطلاعات و اقتصاد قرار دارند .اخیراً مدل‌سازی ریاضی رفتارهای افراد در شبکه‌های اجتماعی به دلیل کاربردهای موجود در بازاریابی و سیاست نیز از اهمیت ویژه‌ای برخوردار گشته است. به‌طورمعمول شبکه‌های اجتماعی را می‌توان در قالب گراف نمایش داد که در این گراف گره‌ها معادل با بازیگران شبکه‌های اجتماعی بوده و یال‌های گراف نشان‌دهنده ارتباط بین بازیگران هست. با توجه به ساختار شبکه اجتماعی و یک‌طرفه یا دوطرفه بودن ارتباط، گراف متناظر می‌تواند گراف جهت‌دار یا بدون جهت باشد .همچنین درصورتی‌که قوت ارتباط بین افراد در شبکه اجتماعی یکسان نباشد، گراف متناظر با شبکه یک گراف وزن‌دار خواهد بود که در آن وزن هر یال متناظر باقوت ارتباط است. گراف شبکه‌های واقعی، به لحاظ وارد شدن افراد جدید در فضای شبکه و ایجاد ارتباطات جدید، به‌سرعت در حال رشد هستند .در واقع باید گفت شبکه‌های اجتماعی، محیط‌های بسیار پویایی هستند.

امروزه به علت رشد روزافزون اینترنت و حجم عظیمی از اطلاعات نیاز به سیستم‌هایی داریم تا بتوانند مناسب‌ترین صفحات و محصولات را به کاربر توصیه کنند سیستم‌هایی که این وظیفه را انجام می‌دهند سیستم‌های پیشنهاددهنده نامیده می‌شوند سیستم‌های پیشنهاددهنده با یک‌سری

الگوریتم‌ها و روش‌های خاصی سعی می‌کنند که مناسب‌ترین اقلام از قبیل داده اطلاعات و کالا را شناسایی کرده و نزدیک‌ترین کالا به سلیقه کاربر را به وی پیشنهاد دهند.

سامانه‌های توصیه‌گر به‌طورکلی به سه دسته تقسیم می‌شوند؛ در رایج‌ترین تقسیم‌بندی، آن‌ها را به سه گروه محتوا محور[1] دانش‌محور[2] و صافی سازی تجمعی[3]، تقسیم می‌کنند که البته گونه چهارمی تحت عنوان سیستم‌های پیشنهاددهنده ترکیبی[4] هم برای آن‌ها قائل می‌شوند.

وب، یک ابزار ارتباطی مهم و منبع بازیابی اطلاعات است که طی یک فرایند آشفته و غیرمتمرکز در حال رشد است. این روند منجر به تولید حجم وسیعی از مستندات متصل به یکدیگر گشته است که از هیچ‌گونه سازمان‌دهی منطقی برخوردار نیستند. یافتن اطلاعات و دانش مرتبط و مورد نیاز از میان تعداد زیادی از صفحات موجود، می‌تواند بسیار وقت‌گیر و دشوار باشد. از سوی دیگر ظهور سرویس‌های مبتنی بر وب مانند تجارت الکترونیکی و یادگیری تحت وب موجب ایجاد تغییرات اساسی در روش استفاده از اینترنت شده است و وب‌سایت‌ها را به محیطی برای تجارت تبدیل کرده و موجب افزایش رقابت بین آن‌ها شده است؛ بنابراین نیاز به افزودن خدمات اضافی به سرویس‌های وب به عنوان لازمه‌ی ایجاد مشتری پابرجا به وضوح احساس می‌شود. این خدمات اضافی تنها با تمرکز بر نیازها و علایق فردی مشتریان و فراهم کردن سرویس‌ها و محصولات موردنظر آن‌ها امکان‌پذیر است.

داشتن یک سیستم هوشمند که قادر باشد علایق کاربران را یاد بگیرد و براساس این علایق به‌طور خودکار علایق غیرمرتبط را فیلتر کند یا اطلاعات مرتبط را در زمان کم به کاربر پیشنهاد دهد، ضروری است. برای حل این مشکل، شخصی‌سازی وب[5] به یک پدیده‌ی محبوب به منظور سفارشی کردن محیط‌های وب تبدیل شده است. براساس تکنولوژی شخصی‌سازی، سرویس‌های موردنیاز کاربران براساس علایق و اولویت‌های آنان و بدون درخواست صریح، منطبق می‌شوند. شخصی‌سازی وب مجموعه‌ای از عملیات است که تجربه وب را برای یک کاربر خاص یا مجموعه‌ای از کاربران سازمان‌دهی می‌کند و پیشنهادهای پویا براساس الگوهای رفتاری کاربران ارائه می‌دهد. به این معنی

[1] Content-Based
[2] Knowledge Base
[3] Filtration Cumulative
[4] Hybrid RS
[5] Web Personalization

که سیستم خود را با هر کاربر سازگار می‌نماید و مطابق علایق و نیازهای او به درخواست‌های وی پاسخ می‌دهد. فواید یک سیستم شخصی‌سازی وب عبارت‌اند از:

- شخصی‌سازی سرویس‌های ارائه شده توسط یک وب‌سایت نقش مهمی در کاهش گرانبار شدن اطلاعات ایفا می‌کند و وب‌سایت را به یک محیط کاربرپسندتر برای افراد تبدیل می‌کند.
- با فراهم کردن اطلاعات دلخواه کاربر به روش مناسب و در زمان مناسب باعث بهبود گردش کاربر در وب‌سایت می‌شود.
- در تجارت الکترونیکی مکانیسمی برای درک بهتر نیازهای مشتری، شناسایی تمایلات آینده‌ی او و درنهایت افزایش پابرجایی مشتری به سرویس ارائه شده فراهم می‌کند.

سیستم‌های شخصی‌سازی وب به‌طور وسیع در حوزه تجارت الکترونیک، تبلیغات هدفمند و موتورهای جستجو استفاده می‌شوند. سیستم‌های توصیه‌گر یک نمونه از سیستم‌های شخصی‌سازی وب هستند.

امروزه تکنیک‌های وب‌کاوی[1] به‌طور گسترده‌ای برای شخصی‌سازی وب به‌کار گرفته شده‌اند. یکی از انواع وب‌کاوی، کاوش استفاده از وب[2] می‌باشد که از تکنیک‌های ساخت سیستم‌های توصیه‌گر وب است. یک سیستم توصیه‌گر وب ساخته شده توسط کاوش استفاده از وب، فایل‌های ثبت وقایع[3] سرویس‌دهنده‌ی وب را به عنوان ورودی می‌گیرد و از تکنیک‌های داده‌کاوی از جمله کاوش قوانین انجمنی، استخراج الگوهای ترتیبی و خوشه‌بندی برای استخراج الگوهای پیمایشی کاربران جهت ارائه‌ی پیشنهادهای آتی به کاربر جاری استفاده می‌کند. روش‌های متفاوتی جهت ایجاد سیستم‌های شخصی‌سازی وب پیشنهاد شده است، اما این روش‌ها دقت و پوشش کافی را دارا نیستند، هدف این پژوهش ارائه یک رویکرد جدید براساس خوشه‌بندی فازی و قوانین انجمنی وزن‌دار برای افزایش دقت و پوشش سیستم‌های شخصی‌سازی وب است.

[1] Web Mining
[2] Web Usage Mining
[3] Web Log Files

فصل اول

اصول سیستم‌های پیشنهاددهنده

برای پایه‌ریزی و ایجاد یک سیستم پیشنهاددهنده کارا مواردی وجود دارند که باید به آن‌ها توجه نمود و در پروسه طراحی و پیاده‌سازی سیستم آن‌ها را باید در نظر گرفت. این موارد به شرح ذیل می‌باشند:

- **نوع داده‌های موجود در بستر سیستم**: بنا به کاربرد سیستم ممکن است انواع مختلفی از منابع اطلاعاتی در سیستم وجود داشته باشد. این اطلاعات می‌توانند امتیازهای کاربران به آیتم‌ها[1]، اطلاعات شخصی کاربران، محتوای مربوط به آیتم‌های سیستم، ارتباطات موجود در شبکه‌های اجتماعی و اطلاعات مربوط به موقعیت کاربر[2] باشند. طبیعی است که در پروسه طراحی یک سیستم پیشنهاددهنده باید به نوع داده‌های در اختیار توجه بسیار نمود.
- **الگوریتم فیلترینگ مورد استفاده**: هدف سیستم‌های پیشنهاددهنده در واقع رتبه‌بندی آیتم‌های سیستم به لحاظ نزدیک بودن به علایق کاربران می‌باشد تا در هنگام ارائه پیشنهاد آیتم‌هایی با رتبه بالاتر را به کاربر پیشنهاد دهند. برای این منظور الگوریتم‌های متعددی پیشنهاد شده‌اند که مهم‌ترین آن‌ها عبارت‌اند از:
  - Collaborative Filtering
  - Content-based Filtering
  - Social-based Filtering
  - Knowledge-based Filtering
  - Context-aware Filtering
  - Hybrid Filtering

[1] Ratings
[2] Location-Aware Information

- **مدل انتخابی برای سیستم:** در حال حاضر برای پیاده‌سازی سیستم‌های پیشنهاددهنده دو راهکار استفاده می‌شود. یکی استفاده مستقیم از داده‌های موجود سیستم است که به این راهکار، روش مبتنی بر حافظه[1] گفته می‌شود و دیگری که کمی هوشمندتر است استفاده از یک مدل در سیستم است که بدان راهکار مبتنی بر مدل[2] گفته می‌شود (استفاده از الگوریتم‌های ژنتیک، شبکه‌های عصبی، فازی).

- **تکنیک مورداستفاده در پیشنهاد دادن**: راهکارها و تکنیک‌های مختلفی برای پیاده‌سازی هسته یک سیستم پیشنهاددهنده وجود دارد. به‌عنوان‌مثال می‌توان از الگوریتم‌های ژنتیک، شبکه‌های عصبی، شبکه‌های Bayesian، روش‌های احتمالی و یا الگوریتم‌های هماسیگی برای پیاده‌سازی هسته چنین سیستم‌هایی استفاده می‌شوند.

- **مقیاس‌پذیری مورد انتظار سیستم.**

- **کارایی مطلوب سیستم** (ازنظر میزان حافظ مصرفی و زمان).
- **کیفیت نتایج قابل‌ارائه**: نتایجی که یک سیستم پیشنهاددهنده ارائه می‌دهد می‌تواند با توجه به کاربرد آن و اینکه در چه حوزه‌ای از بازار کسب‌وکار در حال استفاده است متفاوت باشد. به‌عنوان‌مثال در برخی از سیستم‌ها ممکن است هدف ارائه جدیدترین‌ها باشد درحالی‌که در برخی دیگر هدف می‌تواند پوشش دادن به‌تمامی آیتم‌ها باشد. به‌طورکلی در سیستم‌های پیشنهاددهنده هدف فراهم آوردن سه فاکتور تازگی[3]، دقت[4] و پوشش[5] در نتایج پیشنهادی است.

[1] Memory-Based
[2] Model-Based
[3] Novelty
[4] Precision
[5] Coverage

روش‌ها و الگوریتم‌های فیلترینگ

بدون شک اساسی‌ترین جزء در سیستم‌های پیشنهاددهنده الگوریتم و راهکار فیلترینگ آن می‌باشد. در ادامه به مهم‌ترین راهکارهای مورداستفاده در این زمینه خواهیم پرداخت:

Collaborative Filtering: یکی از مهم‌ترین و پراستفاده‌ترین روش‌های فیلترینگ در سیستم‌های پیشنهاددهنده روش Collaborative Filtering می‌باشد که پایه و مبنای کار در بسیاری از راهکارهای دیگر نیز بشمار می‌رود. روش کار این الگوریتم درواقع به همان صورتی است که ما در تصمیم‌گیری‌های روزمره‌مان عمل می‌نماییم. به‌عنوان‌مثال کالایی را می‌خریم که بیشتر موردپسند دیگران واقع‌شده باشد (مثلاً در دنیای تجارت الکترونیک امتیازهای[1] بیشتری به آن داده‌شده باشد)؛ بنابراین می‌توان گفت درروش Collaborative Filtering آنچه مدنظر قرار داده می‌شود، بیشتر تجربه دیگران است تا خود فرد.

در این روش ابتدا باید اجازه داد تا کاربران در سیستم مشارکت نمایند و به آیتم‌های مختلف موجود در سیستم امتیاز دهند. البته این امتیاز دادن‌ها می‌توان به‌صورت ضمنی نیز اتفاق بیفتد و توسط سیستم تشخیص داده شود. به‌عنوان‌مثال یک نوع امتیاز دادن ضمنی می‌توان بدین شکل باشد که آیتم‌هایی که بیشتر دانلود شده‌اند احتمالاً از محبوبیت بیشتری برخوردار بوده‌اند و درنتیجه امتیاز بیشتری نسبت به بقیه به آن‌ها داده می‌شود.

توجه داشته باشید که در این راهکار سیستم بر مبنای امتیازات، آیتم‌ها را رتبه‌بندی می‌کند و آیتم‌هایی با بیشترین امتیاز را به کاربر پیشنهاد می‌دهد. به همین دلیل درصورتی‌که سیستم تازه شروع به کارکرده باشد و یا آیتم جدیدی به سیستم اضافه شود، اطلاعات کافی از آیتم‌ها در دسترس نخواهد بود و درنتیجه نمی‌توان به‌درستی امتیازدهی و رتبه‌بندی را انجام داد. این یکی از مشکلات اساسی و مهم در این‌گونه سیستم‌هاست که با عنوان شروع سرد[2] شناخته می‌شود. البته این سیستم‌ها

[1] Rate
[2] Cold Start

از مشکل دیگری نیز رنج می‌برند که پراکندگی داده‌ها[1] می‌باشد. بدین معنی که اطلاعات در سیستم وجود دارد اما پراکنده هستند و نمی‌توان به‌درستی و با قطعیت گفت که چه آیتمی مقبولیت بیشتری دارد.

مشکل شروع سرد و پراکندگی داده‌ها

همان‌طور که گفته شد یکی از مهم‌ترین مشکلاتی که سیستم‌های پیشنهاددهنده مبتنی بر CF با آن مواجه هستند، مسئله شروع سرد می‌باشد. این مشکل زمانی ایجاد می‌شود که به‌منظور ارائه پیشنهاد، اطلاعات لازم و کافی در سیستم وجود نداشته باشد. این حالت ممکن است به یکی از دلایل زیر رخ دهد.

۱. **شروع کار سیستم پیشنهاددهنده**: راهکاری که در چنین حالاتی پیشنهاد می‌شود این است که با استفاده از روش‌های مناسب کاربران را تشویق به دادن روی به آیتم‌ها نماییم و زمانی اقدام به پیشنهاد به کاربر کنیم که به‌اندازه کافی اطلاعات جمع‌آوری‌شده باشد.
۲. **ورود کاربر جدید به سیستم**: مهم‌ترین مشکل برای سیستم‌های پیشنهاددهنده مبتنی بر CF زمانی است که کاربر جدیدی وارد سیستم می‌شود. در این صورت اطلاعات کافی در مورد آیتم‌ها وجود دارد اما ازآنجاکه کاربر جدیدالورود هنوز به آیتمی رأی نداده است نمی‌توان از روش‌های معمول مورد استفاده در CF استفاده نمود. برای حل چنین مشکلی در سیستم، عموماً CF را با دیگر روش‌های رایج در سیستم‌های پیشنهاددهنده ترکیب می‌کنند و یک سیستم ترکیبی[2] را می‌سازند (مثلاً CF با Content-based Filtering).
۳. **درج آیتم جدید در سیستم**: عموماً آیتم‌های جدید دارای هیچ امتیازی نمی‌باشند. بر همین اساس در لیست پیشنهادها هرگز آورده نمی‌شوند و از دیدگاه کاربران نیز پنهان می‌مانند. این مسئله باعث می‌شود که در آینده نیز به آن‌ها هیچ امتیازی داده نشود. البته این مسئله در سیستم‌های پیشنهاددهنده از اهمیت بالایی برخوردار نیست و می‌توان بر اساس روش‌ها و ابزارهای دیگری این آیتم‌ها را به کاربران نشان داد تا به آن‌ها رأی دهند.

[1] Data Sparsity
[2] Hybrid

به دلیل مشکلات شروع سرد و نیز پراکندگی داده‌ها عموماً سیستم‌های CF را به‌صورت ترکیبی با سایر راهکارها بکار می‌برند تا از مزایای آن‌ها بهره‌مند شده و درعین‌حال معایب آن را نیز برطرف نمایند. به‌عنوان‌مثال یکی از راهکارهایی که پیشنهاد شده است پیاده‌سازی Collaborative Tagging در یک سیستم مبتنی بر Collaborative Filtering است تا بتوان سلایق کاربران را شناخت و آیتم‌ها را بر اساس تمایلات کاربران دسته‌بندی نمود.

یکی دیگر از راه‌های مقابله با مشکلات ذکر شده استفاده از تکنیک‌های خوشه‌بندی می‌باشد که عموماً برای حل مشکل شروع سرد بکار گرفته می‌شود. در این روش می‌توان آیتم‌ها یا کاربران و یا هر دوی آن‌ها[1] را خوشه‌بندی کرد. این تکنیک‌ها علاوه بر برطرف کردن مشکل ذکر شده باعث بهبود کارایی سیستم پیشنهاددهنده نیز می‌گردند.

برای برطرف نمودن مشکل پراکندگی داده‌ها نیز عموماً از تکنیک‌های dimensionality reduction استفاده می‌شود. در کنار این روش، تکنیک‌های Latent Semantic Index و Singular Value Decomposition نیز وجود دارند. در مورد تکنیک SVD باید گفت که علیرغم نتایج بسیار خوب، این تکنیک سربار پردازشی بالایی دارد و از آن‌ها بهتر است در کاربردهای آفلاینی استفاده نمود که تغییرات زیادی در اطلاعات آن‌ها وجود ندارند.

الگوریتم KNN[2]

رایج‌ترین الگوریتم مورداستفاده در روش CF، الگوریتم KNN می‌باشد. در این الگوریتم دو رویکرد وجود دارد. رویکرد اول نگاهی کاربربه‌کاربر[3] به سیستم دارد. الگوریتم‌هایی که بر مبنای این نگاه پیاده‌سازی می‌شوند شامل سه گام هستند:

[1] Bi-Clustering
[2] K Nearest Neighbors
[3] User-To-User

گام اول: در این گام بر اساس یک معیار شباهت (cosine، Pearson Correlation، Mean square difference) برای کاربر a تعداد k همسایه انتخاب می‌شود. این همسایگان، آن‌هایی هستند که بیشترین شباهت را به کاربرa دارند.

گام دوم: در گام دوم به ازای تمامی آیتم‌های موجود در سیستم معیاری کمی برای پیش‌بینی[1] آنکه آیا آیتم i موردپسند کاربر a قرار خواهد گرفت یا خیر محاسبه می‌شود. محاسبه این معیار کمی با استفاده از راهکارهای مختلفی (میانگین امتیازات[2]، حاصل جمع وزن‌دار[3] و ...) از روی امتیازهایی که همسایگان کاربر a به آیتم i داده‌اند حاصل می‌شود.

گام سوم: بر اساس گام دوم، از بین تمامی آیتم‌ها N آیتمی که بیشتری مقدار پیش‌بینی را دارند به کاربر پیشنهاد داده می‌شوند.

از جمله مزیت‌های این الگوریتم سادگی و درعین‌حال دقت نتایج حاصل از آن است. البته دو مشکل اساسی نیز دارد که عبارت‌اند از مقیاس‌پذیری کم[4] و آسیب‌پذیری در مقابل پراکندگی داده‌ها در پایگاه داده. با افزوده شدن کاربر جدید به سیستم معیارهای شباهت و مقادیر پیش‌بینی‌ها باید مجدداً حساب شوند که با افزایش تعداد کاربران و بزرگ شدن سیستم این مسئله سربار محاسباتی زیادی را به سیستم وارد می‌آورد و مشکل‌ساز می‌شود.

برای حل مشکل مقیاس‌پذیری در این الگوریتم، نسخه دیگری از آن با رویکردی متفاوت ارائه شده است. این رویکرد نگاه آیتم به آیتم[5] دارد و توانسته است مشکل مقیاس‌پذیری را تا حد قابل‌توجهی کاهش دهد. در این نسخه نیز سه گام اصلی دیده می‌شود که به شرح زیر می‌باشند:

---

[1] Prediction
[2] Average
[3] Weighted Sum
[4] Scalability
[5] Item-To-Item

گام اول: ابتدا بر اساس معیارهای شباهت برای هر آیتم i تعداد q همسایه را تعیین می‌کنیم.

گام دوم: درصورتی‌که کاربر a به آیتم i تاکنون امتیازی نداده باشد، بر اساس امتیازاتی که این کاربر به آیتم‌های همسایه i داده است مقدار پیش‌بینی را محاسبه می‌کنیم.

گام سوم: بر اساس مقادیر پیش‌بینی‌ها[1] آیتم‌هایی را که بیشترین مقدار پیش‌بینی را دارند به کاربر a پیشنهاد می‌کنیم.

تعیین میزان شباهت

اساس کار در سیستم‌های پیشنهاددهنده مبتنی بر CF تعیین میزان شباهت بین کاربران و آیتم‌هاست تا بر اساس آن بتوان همانسیگی را به دست آورد. برای این منظور نیاز به یک معیار شباهت وجود دارد. در روش‌های سنتی شباهت بر اساس امتیازدهی تعیین می‌شد. بدین گونه که اگر دو کاربر به یک آیتم رأی می‌دادند سیستم نتیجه می‌گرفت که شباهتی بین این دو آیتم وجود دارد. در مورد آیتم‌ها نیز این مسئله به‌گونه‌ای دیگر قابل‌تعریف است. بدین‌صورت که درصورتی‌که به دو آیتم توسط کاربران یکسانی رأی می‌دادند سیستم احتمال می‌داد که شباهتی بین این دو آیتم وجود دارد.

در کنار این معیار، موارد دیگری نیز مطرح هستند که از آن‌ها در تعیین شباهت بین کاربران و آیتم‌ها استفاده می‌شود. برخی از این معیارها عبارت‌اند از:

Pearson Correlation، Cosine، Adjusted Cosine، Mean Square Difference و Constraint Correlation

با ظهور Web 2 و گسترش شبکه‌های اجتماعی این روزها گرایش به سمت استفاده از پتانسیل موجود در این شبکه در امر تعیین میزان شباهت و ارائه پیشنهادهای دقیق‌تر به وجود آمده است. بر

[1] Recommendations

همین مسائل مسائلی همچون trust، Credibility و reputation نیز در تعیین میزان شباهت استفاده می‌شود.

Demographic Filtering: اطلاعاتی نظیر سن، جنسیت، ملیت و ... در گروه اطلاعات دموگرافیک[1] قرار می‌گیرند. سیستم‌هایی که از این روش استفاده می‌کنند بر این اساس عمل می‌کنند که کاربرانی که صفات دموگرافیک مشابهی دارند (مثلاً در یک بازه سنی قرار می‌گیرند) احتمالاً سلایق و خواسته‌ها مشابهی نیز دارند.

Content-based Filtering: برخلاف روش قبلی که پیشنهادها بر اساس تجربیات دیگران داده می‌شد، در Content-based Filtering تکیه بر اطلاعات و سلایق کاربر جاری است، بدین شکل که برای ارائه پیشنهاد، به انتخاب‌ها و تجربیات وی در گذشته توجه می‌شود. عموماً در چنین روش‌هایی نیاز به تحلیل و آنالیز اطلاعات و محتویات[2] مربوط به کاربر و آیتم‌های موجود در سیستم است تا بتوان میزان شباهت بین کاربر و آیتم‌های سیستم (و نیز شباهت بین آیتم‌ها با یکدیگر) را تعیین نمود. در نتایج پیشنهادی به کاربر، آیتم‌هایی آورده می‌شوند که شباهت بیشتری به آیتم‌هایی دارند که قبلاً کاربر آن‌ها را انتخاب نموده است. نکته مهم و اساسی در چنین سیستم‌هایی انتخاب یک معیار شباهت[3] می‌باشد.

روش Content-based Filtering دارای دشواری‌ها و معایبی می‌باشند که عبارت‌اند از:

- **دشوار بودن استخراج داده‌ها و اطلاعات در مورد آیتم‌ها و کاربران**: در این روش هدف پیدا نمودن شباهت بین آیتم‌های مختلف بر اساس صفات و محتویات آن‌هاست که در برخی

[1] Demographic
[2] Contents
[3] Similarity Measure

از کاربردها و حوزه‌ها (مانند موزیک، ویدئو وبلاگ‌ها) این کار بسیار پیچیده و دشوار است. در این روش نیاز به راهکارهایی برای استخراج صفات به‌صورت خودکار می‌باشد.

- **مشکل overspecialization**: در طول حیات سیستم، تلاش بر این است که آیتم‌هایی به کاربر پیشنهاد داده شوند که شباهت بیشتری به آیتم‌های انتخاب شده توسط او در گذشته داشته باشند. این باعث می‌شود که آیتم‌هایی که ممکن است موردپسند کاربر باشند ولی شباهتی به آیتم‌های انتخاب شده در گذشته ندارند، به کاربر هر گز پیشنهاد داده نشوند و از دید وی مخفی بمانند.
- **عدم امکان گرفتن بازخورد کاربران**: معمولاً در سیستم‌هایی که از این راهکار استفاده می‌کنند امکان گرفتن بازخورد از کاربران وجود ندارد. به‌عنوان‌مثال در چنین سیستم‌هایی معمولاً کاربران به آیتم‌ها امتیاز نمی‌دهند (برخلاف آنچه در مورد سیستم‌های CF داشتیم). این مسئله باعث می‌شود که نتوان دریافت که آیا پیشنهاد داده شده به کاربر صحیح بوده است یا خیر.

به دلیل وجود چنین معضلاتی معمولاً این راهکار را به‌صورت ترکیبی با راهکارهای دیگر مورد استفاده قرار می‌دهند. به‌عنوان‌مثال یکی از کارهای خوب در زمینه ایجاد سیستم‌های ترکیبی، ترکیب Content-based Filtering با شبکه‌های اجتماعی بوده است. در چنین سیستم‌هایی علاوه بر ratings از اطلاعات موجود در شبکه‌های اجتماعی مانند کامنت‌ها، بلاگ‌ها، ارتباطات بین دوستان، like ها و followers استفاده می‌شود تا بتوان کیفیت و دقت نتایج را بهبود داد.

گام‌های روش CBF[1]

در روش Content-based Filtering سه گام اصلی وجود دارد:

- **استخراج صفات[1] مربوط به آیتم‌ها**: برای آنکه یک سیستم مبتنی بر CBF به‌خوبی عمل نماید، ابتدا می‌بایست صفات مربوط به آیتم‌ها استخراج شوند. عموماً بیشتر صفات به‌طور

[1] Content-Based Filtering

صریح همراه با آیتم‌ها در سیستم درج می‌شوند؛ بنابراین استخراج این‌گونه صفات با مشکل خاصی مواجه نیست؛ اما گروهی دیگر از صفات هستند که بر اساس دامنه[۲] سیستم، برای استخراج آن‌ها باید از تکنیک‌های خاصی استفاده نمود. به‌عنوان‌مثال در سیستم‌هایی که آیتم‌ها اسناد متنی هستند، می‌بایست از روش‌های کلاسیک بازیابی اطلاعات[۳] استفاده نمود تا بتوان به صفاتی از قبیل term frequency، inverse document frequency و document length دست پیدا کرد.

- **مقایسه صفات آیتم‌ها با سلایق کاربر**: پس از مشخص شدن صفات آیتم‌ها باید تحلیل‌هایی صورت پذیرد که نشان دهد آیتم‌های موجود در سیستم تا چه اندازه با علایق کاربر همخوانی دارند که این کار عموماً با استفاده از روش‌هایی از قبیل روش‌های اکتشافی[۴] و یا الگوریتم‌های خوشه‌بندی انجام می‌شود.
- **پیشنهاد دادن آیتم‌هایی که شباهت بیشتری به سلایق کاربر دارند**.

گرایش موجود در مورد سیستم‌های CBF ترکیب آن‌ها با شبکه‌های اجتماعی و استفاده از اطلاعاتی نظیر tag، comment و social network sharing است. معروف‌ترین این سیستم‌ها، سیستم‌های پیشنهاددهنده تگ[۵] هستند. البته حوزه کار بر روی تگ‌ها به دو دسته تولید سیستم‌های پیشنهاددهنده تگ و استفاده از تگ‌ها در پیشنهادها تقسیم می‌شود.

گروه‌بندی کاربران

سیستم‌های CBF به دلیل نیاز به اعمال پردازشی و تحلیلی فراوان کارایی کمتری نسبت به بقیه سیستم‌ها دارند. یکی از راهکارهای مؤثری که برای بهبود کارایی آن‌ها پیشنهاد شده است گروه‌بندی کاربران و ارائه پیشنهاد به کل گروه است (بجای یک کاربر). اگر چه در این روش بهبودی در دقت و کیفیت نتایج حاصل نمی‌شود اما در کارایی و کم شدن سربار پردازشی تأثیر بسیاری دارد.

---

[1] Attributes
[2] Domain
[3] Information Retrieval
[4] Heuristic
[5] Tag Recommendation Systems

Social-based Filtering: همان‌طور که گفته شد با گسترش شبکه‌های اجتماعی گروهی از محققان به سمت استفاده از اطلاعات موجود در این شبکه‌ها (نظیر trust، followed، Followers، friends، comments، blog و tags) در سیستم‌های پیشنهاددهنده رفتند. توجه داشته باشید که این اطلاعات ممکن است به‌صورت صریح و یا ضمنی جمع‌آوری شوند. بر اساس نتایج حاصل از به‌کارگیری این اطلاعات مشخص‌شده است که این کار باعث بهبود نتایج پیشنهادی و همچنین کاهش مشکل پراکندگی داده‌ها شده است.

در حوزه استفاده از شبکه‌های اجتماعی در سیستم‌های RS، مطالعات و تحقیقات علمی به دو دسته تقسیم می‌شوند. گروهی به دنبال استفاده از اطلاعات موجود در این شبکه‌ها در جهت بهبود کارایی سیستم‌های موجود رفتند که نتایج حاصل از کارهای آن‌ها گواه بر تأثیر مثبت این اطلاعات در سیستم‌های RS دارد. در مقابل گروهی دیگر از محققین به سمت ایجاد یک سیستم پیشنهاددهنده جدید مبتنی بر Social Filtering رفتند. این گروه دیگر به دنبال ترکیب شبکه‌های اجتماعی با سایر سیستم‌های پیشنهاددهنده نیستند. بلکه قصد دارند از پتانسیل‌های موجود در چنین شبکه‌هایی برای ایجاد یک سیستم مستقل استفاده نمایند.

اعتماد و اعتبار

در میان اطلاعات موجود در شبکه‌های اجتماعی، اعتماد[1] و اعتبار[2] نسبت به بقیه توجهی بیشتری به خود جذب کرده‌اند. معیار اعتماد در واقع میزان اعتبار یک کاربر در بین سایر کاربران است که نقش مؤثری در پیشنهادها می‌تواند داشته باشد. به‌عنوان‌مثال هرچه میزان اعتماد یک کاربر بیشتر باشد، امتیازهایی که او به آیتم‌ها می‌دهد از درجه اهمیت و وزن بیشتری نسبت به سایرین برخوردار است. دو راه‌کار برای تعیین اعتماد کاربران وجود دارد. روش اول از طریق اطلاعاتی است که صریحاً از خود کاربر کسب می‌شود. روش دوم نیز از طریق اطلاعات ضمنی و روابط بین کاربران که در شبکه‌های اجتماعی موجود است می‌باشد. برخی از راهکارهایی که تا کنون در مقالات پیشنهاد

[1] Trust
[2] Reputation

شده‌اند عبارت‌اند از: مکانیزم‌های انتشار اعتماد (trust propagation mechanism)، روش Follow The Leader، شبکه اعتماد[۱]، معیارهای شباهت مبتنی بر خصوصیات فردی[۲]، Distrust Analysis.

در مورد آیتم‌ها، معیار اعتبار مورد استفاده قرار می‌گیرد. می‌توان این معیار را برای آیتم‌ها از روی تعداد امتیازاتی که کاربران به یک آیتم می‌دهند (صریح) و یا با بررسی نحوه کار کاربران با آیتم‌ها (ضمنی) تعیین نمود.

استفاده از برچسب

یکی از امکاناتی که در بیشتر سیستم‌های پیشنهاددهنده اجتماعی وجود دارد این است که کاربران امکان اختصاص برچسب[۳] را به آیتم‌ها دارند؛ بنابراین در چنین سیستم‌هایی ما مجموعه از سه‌تایی‌های <user, item, tag> را داریم که این مجموعه‌ها باعث ایجاد فولکسونومی‌ها[۴] می‌گردد. پتانسیل بالایی که توسط تگ‌ها، به‌عنوان فراداده[۵]، در فواکسونومی‌ها به وجود آمده است منجر شده است به اینکه گروهی از سیستم‌های پیشنهاددهنده کاملاً بر مبنای تگ‌ها[۶] شکل بگیرند. علاوه بر این‌گونه از سیستم‌ها، به دلیل مزایای بسیار استفاده از برچسب‌ها، در برخی از سیستم‌های ترکیبی از این برچسب‌ها برای تقویت و بهبود کیفیت نتایج پیشنهادی به‌کارگیری شده است.

Context-aware Filtering: حرکت به سمت Web3 یا همان Internet Of Things باعث ظهور نسل جدیدی از سیستم‌های پیشنهاددهنده شده است. در چنین محیطی دستگاه‌ها و حس‌گرهای گوناگونی وجود دارند که اطلاعاتی از شرایط کاربر[۷] را جمع‌آوری می‌کنند. چنین اطلاعاتی را می‌توان در سیستم‌های پیشنهاددهنده مورداستفاده قرار داد تا نسل جدیدی از سیستم‌ها بنام سیستم‌های مبتنی بر Context-aware information شکل بگیرند. تأکید این سیستم‌ها بر

[1] Trust Network
[2] Personality-Based Similarity Measures
[3] Tag
[4] Folksonomies
[5] Metadata
[6] Tag Recommendation Systems
[7] Context

اطلاعاتی از قبیل زمان، مکان، اطلاعات حاصل از دوربین‌های امنیتی، RFID ها و شبکه‌های حسگر بیسیم و نیز پارامترهای سلامت، عادات خرید و غذا خوردن فرد می‌باشد. این اطلاعات را می‌توان به‌صورت صریح و یا با استفاده از روش‌های داده‌کاوی[1] کسب نمود.

سیستم‌های پیشنهاددهنده مبتنی بر موقعیت

یکی از موفق‌ترین این سیستم‌ها که در سال‌های اخیر کارهای خوبی نیز بر روی آن انجام شده است، سیستم‌های پیشنهاددهنده مبتنی بر موقعیت[2] می‌باشند. این‌گونه از سیستم‌ها که عموماً در برنامه‌های تلفن همراه نمود دارند، بر اساس موقعیت فعلی کاربر پیشنهادهایی را در حوزه خاصی به وی می‌دهند.

دسته‌ای از این سیستم‌ها، به‌صورت ترکیبی با سیستم‌های CF استفاده می‌شوند. راهکار رایج در این نوع سیستم‌های پیشنهاددهنده بدین‌صورت است که امتیازدهی به‌صورت سنتی در قالب یک روش CF به آیتم‌ها داده می‌شود؛ اما هنگامی‌که پیشنهاد به کاربر داده می‌شود اطلاعات جغرافیایی او نیز در پروسه ارائه پیشنهاد دخیل می‌گردند. این مسئله باعث ایجاد پیچیدگی در سیستم می‌شود. چرا که با تغییر مکان کاربر آیتمی که قبلاً مطلوب کاربر بوده و باید به وی پیشنهاد داده می‌شد، ممکن است دیگر موردپسند کاربر نباشد. به‌عنوان‌مثال فرض کنید فردی که در محل کار است به دنبال رستورانی برای سفارش ناهار می‌گردد. رستورانی که در نزدیکی محل کارش است به او پیشنهاد داده می‌شود و فرد به این رستوران امتیاز مثبت می‌دهد. حال اگر فرد به خانه خود بازگردد که در فاصله دوری نسبت به رستوران قرار گرفته است، آیا در هنگام سفارش غذا بازهم این رستوران باید به وی پیشنهاد داده شود یا خیر؟ اگرچه سیستم بر مبنای الگوریتم CF تشخیص می‌دهد که این رستوران موردپسند کاربر است اما به دلیل دوری از محل فعلی کاربر نباید به وی نشان داده شود.

[1] Data Mining
[2] Location-Aware Recommendation Systems

سیستم‌های پیشنهاددهنده مبتنی بر دانش

سیستم‌های پیشنهاددهنده مبتنی بر دانش[1]، نسل جدیدی از سیستم‌های پیشنهاددهنده هستند که مبتنی بر دانش موجود در رابطه با کاربران و آیتم‌ها هستند. چنین سیستم‌هایی، پیشنهادها خود را بر پایه تفسیر و استنباط خود از سلایق و نیازهای کاربر ارائه می‌دهند و از دیدگاه تئوری نسبت به سایر روش‌های ذکر شده از دقت و کیفیت بیشتری برخوردار هستند. طبیعی است که برای پیاده‌سازی چنین سیستم‌هایی نیاز به یک بستر و ساختار مبتنی دانش وجود دارد (آنتولوژی‌ها، case-based reasoning، constraint-based reasoning، knowledge vectors و social knowledge).

یکی از فیلدهای کاری در زمینه سیستم‌های پیشنهاددهنده مبتنی بر دانش، Workflow می‌باشد که مبتنی بر مدل users-roles-tasks است. در این مدل تشریح می‌شود که یک هر کدام از کاربران در چه نقش‌هایی چه وظایفی را انجام می‌دهند. شبکه‌های نظیر به نظیر[2] نیز حوزه دیگری از تحقیقات را به خود اختصاص می‌دهند که در آن دانش سیستم در مورد آیتم‌ها و کاربران در بین peer ها توزیع شده است.

رویکردهای موجود در سیستم‌های پیشنهاددهنده

سیستم‌های توصیه گر دوست را می‌توان با توجه به نوع داده‌هایی که استفاده می‌کنند و روشی که آن داده‌ها را تحلیل می‌کنند، به دسته‌های متفاوتی تقسیم‌بندی کرد. بر اساس [دسته‌بندی]چهار رویکرد متفاوت

در سیستم‌های پیشنهاددهنده مرتبط به روش ما وجود دارد:

- سیستم‌های پیشنهاددهنده مبتنی بر توپولوژی گراف
- سیستم‌های پیشنهاددهنده مبتنی بر پروفایل کاربران
- سیستم‌های پیشنهاددهنده مبتنی بر فیلترینگ مشارکتی

[1] Knowledge-Based Filtering
[2] Peer-To-Peer

- سیستم‌های پیشنهاددهنده با رویکرد ترکیبی

- **سیستم‌های پیشنهاددهنده مبتنی بر توپولوژی گراف**

در رویکرد سیستم‌های پیشنهاددهنده مبتنی بر توپولوژی گراف، ویژگی‌های ذاتی ساختار شبکه برای تعیین شباهت میان گره‌ها مورد استفاده قرار می‌گیرد. هدف این سیستم‌ها پیدا کردن افراد فعال و مؤثر در شبکه‌های اجتماعی از طریق تجزیه‌وتحلیل گراف اجتماعی می‌باشد.

پیش‌بینی لینک (ارتباط-پیوند) یکی از مسائل شناخته‌شده در شبکه‌های اجتماعی است که کاربردهای زیادی دارد. یکی از مهم‌ترین کاربردهای آن طراحی سیستم‌های پیشنهاددهنده بر مبنای ساختار توپولوژیکی گراف در شبکه‌های اجتماعی می‌باشد. هدف پیش‌بینی لینک این است که با در نظر گرفته تصویری از گراف در زمان t لینک‌های جدیدی در زمان t+1 پیش‌بینی شوند که تصویر لحظه‌ای در زمان t+1 می‌تواند یک هفته، یک ماه، یک سال و حتی سال‌ها بعد از تصویر لحظه‌ای t گرفته شود. الگوریتم‌های پیش‌بینی لینک که به الگوریتم‌های مبتنی بر شباهت گره‌ها شناخته شده‌اند به دو دسته‌ی کلی الگوریتم‌های محلی و سراسری تقسیم می‌گردند.

- معیارهای شباهت الگوریتم‌های محلی (مانند FAOF، ضریب جاکارد و ضریب Adamic/Ader) به‌طور اساسی بر ساختار گره‌های محلی تمرکز دارند و مسیرها با طول‌های مختلف از شبکه را موردبررسی قرار نمی‌دهند و در عوض تنها مسیرهای با طول حداکثر دو را بین کاربر و دوستان کاندیدش در نظر می‌گیرند، در نتیجه دقت پائینی در تشخیص دوستان همسایه دارند.
- الگوریتم‌های سراسری (مانند شاخص وضعیت کاتز، الگوریتم SimRank و الگوریتمPageRank) ساختار کلی مسیر در شبکه را تشخیص می‌دهند که برای شبکه‌های بزرگ محاسبات سنگینی را در بر دارند. همچنین به این دلیل که گراف را به‌صورت سراسری پیمایش می‌کنند، ضبط (تصرف[1]) ویژگی‌های محلی گراف را از دست می‌دهند.
- برخی دیگر از الگوریتم‌ها (مانند الگوریتم FriendLink) از ترکیبی از این دو الگوریتم و در جهت استفاده از مزایای هر کدام طراحی شده‌اند و برخلاف روش‌های محلی که تنها

[1] Capture

مسیرهای با طول حداکثر دو را در نظر می‌گیرند، فرض می‌کنند که هر فرد می‌تواند به دیگران از طریق مسیرهایی با طول‌های مختلف متصل گردد. در نتیجه افق همسایه‌های محلی کاربر را با بهره‌گیری از مسیرهایی با طول بیشتر گسترش می‌دهد و این کار باعث بالاتر رفتن دقت و همچنین سرعت پیش‌بینی ارتباطات بین افراد می‌شود.

- **سیستم‌های پیشنهاددهنده مبتنی بر محتوا**

به‌طورکلی الگوریتم‌های مبتنی بر محتوا، آیتم‌هایی (اشیاء یا اشخاص) با اولویت بالاتر را بر اساس شرح آیتم‌ها و مشخصات پروفایلی کاربران (مانند سن، جنسیت و غیره) توصیه می‌کند.

در روش‌های سیستم‌های پیشنهاددهنده مبتنی بر محتوا، برخی از ویژگی‌های کاربر برای محاسبه شباهت میان کاربران مختلف مورد استفاده قرار می‌گیرد. ویژگی‌های کاربر می‌تواند از منابع مختلفی مانند اطلاعات ثبت‌نام کاربر، موقعیت مکان کاربر، علائق کاربر و غیره استخراج شود. چندین معیار مختلف برای اندازه‌گیری شباهت مشخصات پروفایلی کاربر که به‌طور عمده شامل داده‌های طبقه‌بندی می‌باشد وجود که در این میان اساسی‌ترین شاخص (معیار- اندازه‌گیری) هم‌پوشانی شباهت است که در صورت برابری مقدار دو آیتم ۱ و در غیر این صورت ۰ را برمی‌گرداند. معیار (شاخص) پیچیده‌تر یک شباهت غیر صفر را برای مقادیر آیتم‌های غیر یکسان و در حالت برابر بودن ارزش آیتم‌ها شباهت ۱ را برمی‌گرداند.

- **سیستم‌های پیشنهاددهنده مبتنی بر فیلترینگ مشارکتی**

در این سیستم‌ها به هر کاربر، آیتم‌های جدیدی بر اساس تراکنش‌های قبلی صورت گرفته و با در نظر گرفتن اولویت کاربران مشابه، پیشنهاد داده می‌شود. فیلترینگ مشارکتی مبتنی بر کاربر، فرآیند اجتماعی پیشنهاد دوست در شبکه‌های اجتماعی را مدل می‌کند. در این روش حجم زیادی از اطلاعات درباره‌ی رفتارها، فعالیت‌ها و یا ترجیحات (اولویت‌های) کاربر جمع‌آوری‌شده و برای پیش‌بینی کاربران مشابه تجزیه‌وتحلیل می‌شوند.

الگوریتم‌های فیلترینگ مشارکتی به‌طورکلی به دو دسته تقسیم می‌شوند: الگوریتم‌های مشارکتی مبتنی بر حافظه و الگوریتم‌های مشارکتی مبتنی بر مدل. در الگوریتم‌های مبتنی بر حافظه کل پایگاه

داده کاربر-محصول مورد استفاده قرار می‌گیرد، درحالی‌که در الگوریتم‌های مبتنی بر مدل، ابتدا یک مدل برای رتبه‌بندی ایجاد شده و سپس پیش‌بینی‌ها انجام می‌گیرد.

- **سیستم‌های پیشنهاددهنده با رویکرد ترکیبی**

در این روش‌ها معمولاً بیش از یک رویکرد موجود، برای ایجاد سیستم‌های پیشنهاددهنده با یکدیگر ترکیب می‌شوند. به‌طور مثال ترکیب الگوریتم‌های مبتنی بر ساختار گراف با الگوریتم‌های مبتنی بر محتوا و یا ترکیب الگوریتم‌های مبتنی بر فیلترینگ مشارکتی با الگوریتم‌های مبتنی بر محتوا. هدف از این ترکیب استفاده از مزیت‌های هر روش جهت بالا بردن کارایی سیستم پیشنهاد شده می‌باشد. در بعضی از پژوهش‌ها نیز محققان، از ترکیب الگوریتم‌های تکاملی (مانند الگوریتم ژنتیک، کلونی مورچگان و غیره) با الگوریتم‌های دیگر برای ایجاد یک سیستم‌های پیشنهاددهنده استفاده نموده‌اند.

طبقه‌بندی عملکردهای اصلی شخصی‌سازی وب

یک سیستم شخصی‌سازی وب می‌تواند عملکردهای گوناگونی از یک خوشامدگویی ساده گرفته تا عملکردهای پیچیده‌تری مانند تحویل یک محتوای شخصی‌سازی‌شده در اختیار قرار دهد. این عملکردها به چند دسته تقسیم می‌شوند که در ادامه بیان خواهند شد.

- **به‌خاطرسپاری**

ساده‌ترین عملکرد است که در آن سیستم، اطلاعات مربوط به کاربر از قبیل نام و سابقه‌ی مرور او را ذخیره می‌کند. هنگامی‌که کاربر به وب‌سایت برمی‌گردد این اطلاعات بدون هیچ پردازش دیگری مورداستفاده قرار می‌گیرند. ازجمله‌ی این عملکردها عبارت‌اند از:

  - خوشامدگویی به کاربر
  - نشان کردن صفحات برای کاربر
  - حقوق دسترسی شخصی‌سازی‌شده

- **راهنمایی**

آن دسته از عملیات شخصی‌سازی که به‌منظور کمک‌رسانی به کاربر برای دریافت سریع اطلاعات موردنیاز خود در وب‌سایت و نیز فراهم کردن مرورهای جایگزین برای او انجام می‌شود، در این دسته قرار می‌گیرند. نمونه‌هایی از این عملکردها عبارت‌اند از:

- توصیه‌ی لینک
- آموزش کاربر

- **تطبیق**

تغییر محتوا، ساختار و طرح کلی صفحات با در نظر گرفتن دانش، ترجیحات و علایق کاربر. مثال‌هایی از این عملکردها عبارت‌اند از:

- طرح کلی شخصی‌سازی‌شده
- تطبیق محتوا
- تطبیق لینک‌ها
- قیمت‌گذاری شخصی‌سازی‌شده
- تمایز شخصی‌سازی‌شده بین محصولات

- **پشتیبانی اجرای وظیفه**

این طبقه پیشرفته‌ترین طبقه از اعمال شخصی‌سازی است که شامل اجرای یک عمل خاص از طرف کاربر (و بدون دخالت او) هست. ازجمله این عملکردها می‌توان به موارد زیر اشاره کرد:

- پیغام‌رسانی شخصی‌سازی‌شده
- تکمیل شخصی‌سازی‌شده‌ی پرسش
- مذاکره‌ی شخصی‌سازی‌شده

در این پژوهش از سیستم شخصی‌سازی وب برای راهنمایی کاربران (توصیه صفحات به آن‌ها) استفاده می‌شود.

رویکردهای موجود در شخصی‌سازی وب

سیستم‌های شخصی‌سازی وب خودکار را می‌توان براساس نوع داده‌هایی که استفاده می‌کنند و روشی که آن داده‌ها را تحلیل می‌کنند، به دسته‌های متفاوتی تقسیم‌بندی کرد. براساس چهار رویکرد

فیلترینگ برای شخصی‌سازی وب با استفاده از سیستم‌های توصیه‌گر وجود دارد که عبارت‌اند از: فیلترینگ مبتنی بر قانون، فیلترینگ مبتنی بر محتوا، فیلترینگ مشارکتی و فیلترینگ ترکیبی. هرکدام از این سیستم‌ها دارای محدودیت‌های خاصی است که این محدودیت‌ها در جدول (۱) بیان شده‌اند. در ادامه هر یک از این سیستم‌ها به‌اختصار شرح داده خواهند شد.

- **سیستم‌های فیلترینگ مبتنی بر قانون**[1]

این سیستم نیاز به دخالت دستی طراحان وب‌سایت و عملیات کاربران جهت شخصی‌سازی هستند. در این رویکرد مجموعه‌ای از پرسشنامه‌ها به‌صورت یک درخت تصمیم به کاربران ارائه می‌شوند. طبق جواب‌های داده‌شده توسط کاربر، یک مجموعه از قوانین به‌صورت دستی تعریف می‌شوند و مدل ایستای کاربر ایجاد می‌گردد. براساس قوانین اساسی و مدل کاربر، محتوای صفحات وب مطابق نیازهای کاربر، سفارشی می‌شود. موتور شخصی‌سازی Yahoo! و Websphere Personalization شرکت IBM، نمونه‌هایی از این سیستم‌ها می‌باشند.

- **سیستم‌های فیلترینگ مبتنی بر محتوا**[2]

این سیستم‌ها رفتار هر کاربر را براساس علایق گذشته‌ی وی و سلیقه‌های شخصی او مدل می‌کنند و پروفایل کاربر را ایجاد می‌کنند. سپس سیستم صفحات یا اقلام جدیدی را براساس شباهت محتوایی آن‌ها با صفحات و اقلامی که در پروفایل کاربر موجود است به کاربر توصیه می‌کنند. مکانیسم معمول در این سیستم‌ها معمولاً مقایسه‌ی کلمات کلیدی نشان‌دهنده‌ی صفحات یا توصیف اقلام است. نمونه‌هایی از این سیستم‌ها عبارت‌اند از؛ Letizia و WebWatcher.

- **سیستم‌های فیلترینگ مشارکتی**[3]

این سیستم‌ها یکی از موفق‌ترین رویکردها در ساخت سیستم‌های توصیه‌گر هستند. هدف این سیستم‌ها، شخصی‌سازی بدون تحلیل محتوای وب است. تمرکز اصلی این سیستم‌ها به‌جای شباهت بر مبنای اقلام، بیشتر بر شباهت بین کاربران است. این سیستم‌ها سابقه‌ی ترجیحات کاربر موردنظر را

[1] Rule Based Filtering Systems
[2] Content Based Filtering System
[3] Collaborative Filtering System

با سابقه‌های تمامی کاربران دیگر به‌منظور یافتن کاربران دارای علایق مشابه با کاربر موردنظر مقایسه می‌کنند. به این مجموعه کاربران دارای علایق مشابه، همسایگی کاربر جاری گفته می‌شود. نگاشت بین سابقه‌ی یک کاربر به همسایگانش می‌تواند بر مبنای شباهت رتبه‌بندی اقلام، دسترسی به صفحات با محتوای مشابه و یا خرید اقلام مشابه انجام شود. همسایگی به‌دست‌آمده سپس برای توصیه‌ی اقلامی که توسط کاربر جاری دسترسی و یا خریداری شده‌اند مورداستفاده قرار می‌گیرد.

- **سیستم‌های فیلترینگ ترکیبی**[1]

هدف این رویکرد، ترکیب دو یا چند رویکرد فیلترینگ جهت برطرف کردن محدودیت‌های آن‌ها است. روش‌های مختلفی برای این ترکیب وجود دارد. به‌این‌ترتیب که می‌توان هرکدام از رویکردها را به‌تنهایی به‌عنوان یک سیستم توصیه‌گر اجرا نمود و سپس نتایج پیش‌بینی آن‌ها را با هم ترکیب کرد، یا اینکه روش‌ها را در ابتدا با یکدیگر ترکیب نمود و یک سیستم توصیه‌گر از آن‌ها ایجاد کرد.

جدول (۱): محدودیت‌های رویکردهای مختلف شخصی‌سازی وب

نام رویکرد	محدودیت‌ها
فیلترینگ مبتنی بر قانون	- نیاز به حجم قابل‌توجهی فعالیت در هنگام ساخت و نگهداری - نیاز به همکاری طراح وب و کاربر
فیلترینگ مبتنی بر محتوا	- مشکل بودن استخراج اطلاعات موردنیاز از آیتم‌های متفاوت - محدود بودن پیشنهادها به آیتم‌های مشابه آیتم‌های مشاهده‌شده توسط کاربر (عدم در نظر گرفتن آیتم‌هایی که ممکن است موردعلاقه کاربر باشند، ولی متوجه آن‌ها در سایت نشود)
فیلترینگ مشارکتی	- عدم توسعه‌پذیری خوب با افزایش تعداد کاربران - شروع سرد (برای آیتم‌های جدید)

[1] Hybrid Filtering System

نقش کاوش استفاده از وب در شخصی‌سازی وب

به‌طورکلی وب‌کاوی را می‌توان داده‌کاوی بر روی داده‌های محتوا، ساختار و استفاده وب به‌حساب آورد. هدف وب‌کاوی کشف مدل‌ها و الگوهای نهفته در منابع وب هست. کشف چنین الگوهایی دارای کاربردهای مهمی است که ازجمله‌ی آن‌ها می‌توان به سیستم‌هایی که میزان مؤثر بودن یک سایت را در برآوردن انتظارات کاربر ارزیابی می‌کنند، تکنیک‌هایی برای متعادل کردن پویای بار و بهینه‌سازی وب‌سرورها برای دستیابی مؤثرتر کاربران و کاربردهای مربوط به ساختاردهی مجدد و تطبیق یک سایت براساس نیازهای پیش‌بینی‌شده‌ی کاربر اشاره کرد.

در سال‌های اخیر تکنیک‌های کاوش استفاده از وب به‌عنوان رویکردی که مبتنی بر کاربر است، در شخصی‌سازی وب ارائه شده‌اند که برخی از مشکلات مربوط به فیلترینگ مشارکتی را کاهش می‌دهند. به‌طور خاص کاوش استفاده از وب برای افزایش گسترش‌پذیری سیستم‌های شخصی‌سازی‌شده‌ی سنتی که برمبنای تکنیک‌های فیلترینگ مشارکتی می‌باشند، استفاده شده است.

کاوش استفاده از وب که کاوش ثبت‌های وب نیز خوانده می‌شود، رویکردی جهت جمع‌آوری و پیش‌پردازش داده‌های استفاده وب است. سپس الگوهایی می‌سازد که رفتار و علایق کاربران را نشان می‌دهند. این الگوها می‌توانند به‌صورت خودکار توسط سیستم‌های شخصی‌سازی وب برای پیش‌بینی علایق شخصی کاربران استفاده شوند. هدف شخصی‌سازی وب براساس کاوش استفاده از وب توصیه کردن یک مجموعه از اشیا به کاربر جاری شامل لینک، آگهی، متن، محصول و غیره با جهت‌گیری به سمت ترجیحات و علایق کاربر می‌باشد. این عمل با تطابق جلسه جاری کاربر با الگوهای کاربردی کشف شده از طریق کاوش استفاده از وب صورت می‌گیرد. به این الگوهای کاربردی، پروفایل‌های تجمعی کاربرد گفته می‌شود، زیرا یک نمایش تجمعی از فعالیت‌ها و علایق مشترک گروهی از کاربران فراهم می‌کنند. اشیای توصیه شده به آخرین صفحه در جلسه جاری کاربر پیش از فرستاده شدن آن به مرورگر کاربر، افزوده می‌شوند.

فرآیند کلی شخصی‌سازی وب براساس کاوش استفاده از وب شامل سه مرحله است:

- **آماده‌سازی و پیش‌پردازش داده‌ها:** در این مرحله داده‌های کاربرد جمع‌آوری شده و سپس جهت تشخیص نشست‌های کاربران، پیش‌پردازش می‌شوند. این مرحله همچنین شامل یکپارچه‌سازی داده از منابع مختلف مانند پایگاه‌های داده و سرورهای خدمات کاربردی می‌باشد.
- **کاوش الگوهای پیمایشی کاربران:** در این مرحله الگوهای کاربرد مفید و مدل‌های توصیه توسط الگوریتم‌های داده‌کاوی از داده‌های پیش‌پردازش شده، کشف می‌شوند.
- **تحلیل الگوهای کشف شده:** در این مرحله موتور توصیه، جلسه‌ی جاری کاربر را همراه با الگوهای کشف شده برای فراهم کردن محتوای شخصی‌سازی‌شده مورد استفاده قرار می‌دهد تا توصیه‌های هوشمند ارائه دهد.

از بین این مراحل تنها مرحله‌ی سوم به‌صورت برخط انجام می‌شود. در ادامه دو مرحله‌ی کاوش الگوهای پیمایشی کاربران و تحلیل الگوهای کشف شده به‌طور مختصر بررسی خواهند شد.

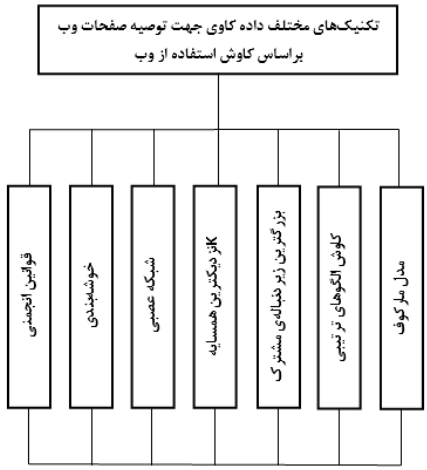
کاوش الگوهای پیمایشی کاربران

در این مرحله تکنیک‌های متنوع کشف دانش بدون ناظر می‌توانند برای استخراج الگوها بکار روند. بعضی از این تکنیک‌ها در مرحله‌ی توصیه نیز بکار گرفته می‌شوند. تکنیک‌هایی که به‌طور عمده استفاده می‌گردند، در شکل (۱) نشان داده شده‌اند. تکنیک‌هایی مانند خوشه‌بندی نشست‌ها می‌توانند منجر به کشف خوشه‌های مهمی از کاربران شوند. سایر تکنیک‌ها مانند خوشه‌بندی اقلام (مثلاً مشاهده صفحه‌ها)، کاوش قوانین انجمنی یا کشف الگوهای ترتیبی نیز می‌توانند برای یافتن ارتباطات مهم بین اقلام بر مبنای الگوهای پیمایشی کاربران سایت مورد استفاده قرار گیرند. در مورد خوشه-بندی و کشف قوانین انجمنی عموماً ترتیب بین مشاهده صفحه‌ها در نظر گرفته نمی‌شود. در مورد کشف الگوهای ترتیبی و مدل مارکوف، نیاز به در نظر گرفتن ترتیب بین مشاهده صفحات در نشست می‌باشد.

تحلیل الگوهای کشف‌شده

ماژول اصلی در این مرحله، موتور توصیه است. هدف موتور توصیه، تطبیق جلسه‌ی جاری کاربر با الگوهای کشف‌شده از طریق کاوش استفاده از وب و توصیه‌ی یک مجموعه از صفحات به کاربر است.

به مجموعه‌ی صفحات توصیه‌شده، مجموعه‌ی توصیه گفته می‌شود. هر یک از رویکردهای خوشه‌بندی، کاوش قوانین انجمنی و کاوش الگوهای ترتیبی، شبکه عصبی، طولانی‌ترین زیردنباله‌ی مشترک و k نزدیک‌ترین همسایه می‌توانند برای ایجاد توصیه استفاده شوند.



شکل (۱): تکنیک‌های مختلف داده‌کاوی جهت شخصی‌سازی صفحات وب براساس کاوش استفاده از وب

- **روش مبتنی بر خوشه‌بندی:** در این روش می‌توان جلسه‌ی جاری را به‌صورت برداری از صفحات نشان داد و خوشه‌ای که بیشترین شباهت را با آن دارد به‌عنوان مبنا برای توصیه استفاده کرد.
- **رویکرد مبتنی بر قوانین انجمنی:** در این روش می‌توان جلسه‌ی جاری کاربر را با سمت چپ مجموعه‌ی قوانین مقایسه کرد تا قوانین مفید برای آن کاربر را پیدا کرد. تمامی عبارات سمت راست این قوانین را می‌توان برحسب اطمینان مرتب کرد و n قلم اول آن را به‌عنوان مجموعه‌ی توصیه برای کاربر انتخاب کرد.

- **روش مبتنی بر الگوهای ترتیبی:** این روش نیز با در نظر گرفتن ترتیب و تغییر روش مبتنی بر قوانین انجمنی عمل می‌کند.
- **شبکه عصبی:** یک شبکه عصبی یک سیستم پردازش اطلاعات توزیع‌شده‌ی موازی است که توانایی یادگیری و خودسازماندهی دارد. این سیستم از تعدادی موجودیت پردازشی که به‌صورت شبکه به هم متصل هستند، تشکیل شده است. شبکه عصبی یک مدل قابل انعطافی را ایجاد می‌کند که برای کاربردهای مختلفی از جمله پیش‌بینی، رگرسیون غیرخطی و یا کلاس بندی قابل استفاده است. در این روش ابتدا شبکه با خوشه‌های به دست آمده از مرحله‌ی قبل آموزش داده می‌شود، سپس نشست جاری کاربر به شبکه وارد می‌شود و مناسب‌ترین خوشه جهت توصیه دادن توسط شبکه عصبی به کاربر معرفی می‌گردد.
- **طولانی‌ترین زیردنباله‌ی مشترک:** مسئله‌ی مقایسه‌ی دو زیررشته‌ی a و b در تعیین شباهت آن‌ها، یک مسئله‌ی اساسی در تطابق الگو است. یکی از شکل‌های پایه‌ی این مسئله، تعیین طولانی‌ترین زیردنباله مشترک دو زیررشته‌ی a و b است. این الگوریتم برای یافتن منطبق‌ترین الگوی پیمایشی با نشست جاری کاربر، جهت پیش‌بینی درخواست‌های آتی بکار گرفته می‌شود.
- **k نزدیک‌ترین همسایه:** مدل k نزدیک‌ترین همسایه، یک سیستم توصیه‌گر مبتنی بر فیلترینگ مشارکتی را طبق سه گام ایجاد می‌کند: گام ۱) ماتریس علاقه‌مندی‌های کاربران با استفاده از اطلاعات استفاده، ایجاد می‌شود. گام ۲) موتور توصیه تکنیک‌های آماری و یا تکنیک‌های یادگیری ماشین را برای یافتن k کاربر (همسایه نیز خوانده می‌شوند) که رفتاری مشابه با رفتار کاربر جاری دارند را بکار می‌گیرد. همسایه‌ها طبق درجه شباهت کاربر جاری با سایر کاربران انتخاب می‌شوند. گام ۳) پس از شکل‌گیری همسایه‌ها، موتور توصیه n آیتم را که توسط کاربران همسایه مشاهده شده‌اند و بیشترین علاقه به آن‌ها را دارند و در نشست جاری کاربر نیز وجود ندارند، به کاربر توصیه می‌کند.

پژوهش‌های انجام شده در زمینه شخصی‌سازی وب

تحقیقات زیادی در زمینه‌ی شخصی‌سازی وب از سال ۲۰۰۰ تاکنون انجام گرفته است. در این قسمت پژوهش‌هایی که در ۵ سال گذشته تاکنون در زمینه‌ی شخصی‌سازی وب با بهره‌گیری از کاوش

استفاده از وب انجام شده‌اند، معرفی شده و به‌طور مختصر مورد بررسی قرار می‌گیرند. پژوهش‌های انجام شده در این زمینه به دو دسته‌ی پژوهش‌های مبتنی بر یک تکنیک کاوش استفاده از وب و پژوهش‌های مبتنی بر ترکیب چند تکنیک کاوش استفاده از وب، تقسیم شده‌اند.

تمامی پژوهش‌ها دارای یک فاز برون‌خط[۱] و یک فاز برخط[۲] هستند. در فاز برون‌خط تمامی پژوهش‌های انجام شده، ابتدا داده‌های ثبت خام دریافت می‌شوند، سپس این داده‌ها پیش‌پردازش می‌گردند. در مرحله‌ی پیش‌پردازش، داده‌های اضافی حذف شده و نشست‌های کاربران ایجاد می‌شوند. عملیات این مرحله در تمامی پژوهش‌ها تقریباً یکسان است و فقط در مواردی مانند حذف بعضی از نشست‌ها با طول‌های خاص و یا نحوه‌ی تشخیص کاربران و ایجاد نشست‌ها با هم تفاوت دارند. در این فاز پس از پیش‌پردازش داده‌ها، در بعضی از پژوهش‌ها برای بیان تأثیر هر کدام از صفحات در نشست‌ها، وزنی به ازای هر صفحه براساس پارامترهای متفاوت مانند زمان مشاهده صفحه، فرکانس مشاهده و غیره، محاسبه شده است. سپس الگوهای پیمایشی کاربران از داده‌های پیش‌پردازش شده، استخراج می‌گردند. در فاز برخط، موتور توصیه نشست جاری کاربر را دریافت می‌کند و با استفاده از تکنیک‌های متفاوت، الگوی مناسب برای پیشنهاد صفحات به کاربر را یافته و پیشنهادهای مناسب برای وی را تولید می‌کند. تفاوت پژوهش‌های مختلف در نحوه‌ی استخراج الگوهای پیمایشی کاربران و تکنیک موتور توصیه در ارائه‌ی پیشنهادها به کاربران است.

روش‌های الهام گرفته شده از طبیعت[۳]

بسیاری از روش‌های ارائه شده در زمینه سیستم‌های پیشنهاددهنده از الگوریتم‌های مبتنی بر طبیعت الهام گرفته‌اند. به‌طورکلی می‌توان این راهکارها را به دو گروه راهکارهای الگوریتم تکاملی (GA) و راهکارهای شبکه‌های عصبی (NN) تقسیم‌بندی نمود. البته مدل‌هایی نیز هستند که بر پایه شبکه‌های ایمنی بدن[۴] پیشنهادشده‌اند.

کارهایی که در حوزه الگوریتم‌های ژنتیک صورت گرفته است عموماً در جهت ایجاد مدل‌های ترکیبی و یا خوشه‌بندی بوده است. در بحث خوشه‌بندی این راهکارها بدین‌صورت عمل می‌نمایند که

[1] Offline Phase
[2] Online Phase
[3] Bio-Inspired
[4] Artificial Immune Network

کاربران را به گروه‌هایی دسته‌بندی می‌کنند طوری که کاربران مشابه در گروه‌های یکسانی قرار بگیرند. این کار باعث می‌شود که بجای فرد پیشنهادها به گروه داده شود و بدین ترتیب از زمان و سربار محاسبات کاسته شود. مدل‌های ترکیبی کاربران[1]، عموماً ترکیبی از CF با روش‌هایی از قبیل Demographic Filtering و یا Content-based Filtering هستند که به‌عنوان‌مثال در اینجا کروموزوم‌های الگوریتم ژنتیک را می‌توان به‌عنوان Demographic Information کاربران در نظر گرفت.

الگوریتم‌های شبکه‌های عصبی بر پایه عملکرد و رفتار سیستم عصبی بدن شکل‌گرفته‌اند که قدرت یادگیری را بر اساس ورودی‌ها و الگوهای پیشین برای سیستم‌های پیشنهاددهنده فراهم می‌کنند. در راهکارهای مبتنی بر شبکه‌های عصبی هدف برخوردار نمودن سیستم از این قوه یادگیری است؛ بنابراین پروسه‌های یادگیری نظری SOM و Case Base Reasoning در کارهای انجام شده پیشین دیده می‌شوند.

راهکارهای مبتنی بر شبکه ایمنی بدن نیز به‌منظور رفع مشکل پراکندگی داده‌ها و افزایش مقیاس‌پذیری سیستم ارائه‌شده‌اند. البته همانند راهکارهای دیگر این روش نیز به‌صورت ترکیبی با روش‌های CF بیشتر مورد استفاده قرار گرفته است.

متدولوژی ارزیابی

برای ارزیابی روش پیشنهادی از سه معیار استاندارد مختلف به نام‌های دقت، پوشش و F1 استفاده شده است. همان‌طور که در فصل دوم مشاهده شد، این معیارها برای ارزیابی بسیاری از سیستم‌های شخصی‌سازی وب بکار گرفته شده‌اند. هر نشست تست ts به طول |ts| در مجموعه داده-های تست به دو قسمت تقسیم می‌شود. w صفحه از نشست ts را جدا کرده و آن را نشست جاری کاربر (پنجره‌ی پیشنهاد) می‌نامیم؛ بنابراین فرض می‌شود که نشست w با طول |w| نشست جاری کاربر فعال بوده و از آن برای تولید پیشنهادها استفاده می‌شود. rp = |ts| - w صفحه‌ی باقیمانده که

[1] Hybrid User Models

معادل با صفحات بازدید نشده توسط کاربر جاری هستند، با خروجی سیستم پیشنهادی مقایسه می‌گردند. نتیجه‌ی فاز پیشنهاد، ارائه‌ی برداری با rs صفحه به کاربر جاری است.

دقت پیشنهاد برابر است با نسبت پیشنهادهای درست به کل پیشنهادها. دقت پیشنهادها با استفاده از رابطه (۱) محاسبه می‌گردد:

$$\text{precision}(rs, rp) = \frac{|rs \cap rp|}{|rs|}$$ رابطه (۱)

پوشش پیشنهاد برابر است با نسبت پیشنهادهای درست تشخیص داده شده به کل صفحات باقیمانده در ادامه‌ی همان نشست. پوشش پیشنهادهای ارائه شده با استفاده از رابطه (۲) اندازه‌گیری می‌شوند:

$$\text{coverage}(rs, rp) = \frac{|rs \cap rp|}{|rp|}$$ رابطه (۲)

معیار F1 زمانی بیشترین مقدار را خواهد داشت که هر دو معیار دقت و پوشش بیشترین مقدار را داشته باشند. معیار F1 پیشنهادهای ارائه شده با استفاده از رابطه (۳) اندازه‌گیری می‌شوند:

$$\text{F1}(rs, rp) = \frac{2 * (precision\ (rs, rp) * \text{cov}\ erage\ (rs, rp))}{precision\ (rs, rp) + \text{cov}\ erage\ (rs, rp)}$$ رابطه (۳)

اهمیت و ضرورت سیستم‌های پیشنهاددهنده

بیش از یک دهه، سیستم‌های پیشنهاددهنده به‌عنوان یک ابزار کلیدی برای غلبه بر حجم عظیم اطلاعات ارائه شده است و تعداد زیادی از الگوریتم‌ها و سیستم‌ها توسعه یافته‌اند. با وجود همه این تلاش‌ها، سیستم‌های پیشنهاددهنده هنوز با چالش‌های بسیاری از جمله، بهبود دقت پیش‌بینی[1]، پراکندگی داده‌ها[2] و عدم وجود منابع کافی[3] روبرو هستند. درواقع سیستم‌های پیشنهاددهنده یکی از

[1] Improving Prediction Accuracy
[2] Data Sparsity
[3] Cold-Start Issues

ابزارهای اصلی برای غلبه بر مشکل انباشت اطلاعات[1] بوده است. با توجه اینکه کاربران علایق خود را به‌صورت صریح بیان نمی‌کنند نیاز به یک سیستم پیشنهاددهنده با دقت بالا احساس می‌شود. با توسعه‌ی سریع محیط جهانی وب، افراد می‌توانند دانش و اطلاعات خود را از طریق مجموعه‌ای از ابزارهای انتشار برخط نظیر سیستم‌های اشتراک‌گذاری برخط و یا سایت‌های اجتماعی، به اشتراک بگذارند. تاکنون ابزارهای زیادی جهت کنترل و سازمان‌دهی این اطلاعات ارائه شده‌اند. سیستم‌های پیشنهاددهنده، نمونه‌ای از موفق‌ترین ابزارهای شخصی‌سازی وب[2] هستند. مهم‌ترین وظیفه‌ی یک سیستم پیشنهاددهنده، شناسایی و معرفی آیتم‌های موردعلاقه‌ی کاربر در یک فضای بسیار بزرگ از آیتم‌های قابل انتخاب (مثل موسیقی، فیلم، کتاب، صفحه وب و ...) است.

تحقیقات نشان می‌دهد، میان بسیاری از شبکه‌های واقعی اعم از سیستم‌های بیولوژیکی و خصوصیات مشترکی وجود دارد که در میان، سیستم‌های همکاری علمی توجه زیادی را در سال‌های اخیر به خود جلب نموده است. ساختار انجمنی خصوصیات این سیستم‌ها ساختار انجمنی برای توضیح خوشه‌بندی یا گروه‌بندی ویژگی‌های شبکه‌های اجتماعی تعریف شده است. یک شبکه با ساختار انجمنی می‌تواند به‌سادگی به انجمن‌های مختلف تقسیم شود. پیوندهای بین گره‌های موجود درون انجمن‌ها، نسبت به پیوند این گره‌ها با سایر گره‌های شبکه متراکم‌تر هستند .اگرچه، تاکنون تعریفی جامع برای انجمن ارائه نشده است، اما بااین‌حال، برخی پارامترهای قابل‌اندازه‌گیری پذیرفته شده برای تشخیص وجود این نوع ساختارها و میزان اهمیت آن‌ها در شبکه‌های معین، ارائه گردیده است .الگوریتم‌های تشخیص ساختار انجمنی به منظور بهبود کارایی، بهبود زمان اجرا استفاده می‌شوند، ولی برای واقعی سازی کشف جوامع در شبکه‌های اجتماعی، زمان اجرا اولویت اول نباشد. البته با توجه به این که خوشه‌بندی یکاست، الگوریتمی باید ارائه شود که در زمان چندجمله‌ای بتواند مسئله را حل کند. الگوریتم NP-HARD بااین‌حال اولویت دادن و توجه بیش‌ازحد به پارامتر زمان می‌تواند دقت الگوریتم کشف جوامع را پایین بیاورد و باعث شود که تعدادی از جوامع که واقعاً در شبکه‌های اجتماعی وجود دارند به دلیل بهبود پارامتر زمان، از دست بروند. در واقع منظور یال‌هایی است که در محاسبه کوتاه‌ترین مسیر بین جفت گره‌ها در شبکه، بیش‌ترین دفعات از آن‌ها عبور می‌شود .با توجه

[1] Information Overload
[2] Web Personalization

به این‌که در شبکه‌های اجتماعی یال نشان‌دهنده ارتباط بین دو فرد است، می‌توان نتیجه گرفت یال‌هایی که در محاسبه کوتاه‌ترین مسیر از آن‌ها به‌دفعات زیادی عبور می‌شود، پل‌هایی هستند که در دو طرف این پل یک زیر گراف بسیار چگال وجود دارد. اگر کشف این یال‌های مابینی به کمک الگوریتم‌های کوتاه‌ترین مسیر در زیر گراف‌های ایجادشده به‌صورت فازی و تعیین یک درجه عضویت برای این یال‌ها ادامه پیدا کند، می‌توانیم قریب به اکثریت جوامع موجود در گراف را کشف کنیم و با بهینه شدن چگالی جوامع، میانگین چگالی گراف را درنهایت بهینه‌سازی کنیم.

فصل دوم

تعریف شبکه‌های اجتماعی

شبکه‌ی اجتماعی، به هر ساختار اجتماعی متشکل از مجموعه‌ای از عامل‌ها و روابط میان آن‌ها گفته می‌شود .مجموعه عامل‌ها می‌تواند نماینده‌ی مجموعه‌ای از افراد، گروه‌ها یا سازمان‌های یک جامعه باشد. مجموعه‌ای عامل‌ها را گره‌های (رأس‌های) شبکه می‌نامند. روابط میان گره‌ها بیانگر نوعی وابستگی بین اعضای شبکه است. این نوع وابستگی می‌تواند رابطه‌ی دوستی یا هم‌گروهی، ارتباط تجاری یا سیاسی و مثال‌های دیگری ازاین‌دست باشد. روابط میان اعضای شبکه و تأثیرپذیری و تأثیرگذاری آن‌ها بر یکدیگر، باعث بروز پدیده‌هایی در حوزه مسائل اقتصادی، سیاسی، اجتماعی، فرهنگی می‌شود؛ به‌عنوان‌مثال، درزمینه‌ی نوع پوشش افراد در یک جامعه، اگر عده‌ای از دوستان و آشنایان یک فرد که تأثیرگذاری بالایی روی او دارند، از یک نوع پوشش به‌خصوص استفاده کنند، آن فرد هم بعد از مدتی، از مجموعه‌ی دوستانش پیروی خواهد کرد ونیز روی گروه دیگری از دوستانش تأثیر می‌گذارد؛ کم‌کم این رفتار انتشار می‌یابد و بخشی از جامعه از یک نوع پوشش خاص پیروی می‌کنند و پدیده‌ی اجتماعی پیروی از مد بروز می‌یابد .از طرفی بروز پدیده‌هایی در حوزه‌های یادشده، باعث به وجود آمدن روابط جدید یا کم‌رنگ شدن یا از بین رفتن روابط قبلی می‌شود. درگیری میان دو کشور، نمونه‌هایی از چنین پدیدهای در حوزه‌ی مسائل سیاسی است که باعث می‌شود دوستان و متحدان هر یک از دو کشور با طرف مقابل روابط کم‌رنگ و سردتری داشته باشند و حتی روابط خود را قطع کنند .از طرفی ممکن است دو کشور که با یکدیگر روابط خوبی ندارند، باهم متحد شوند تا در مقابل دشمن مشترک خود، پیروز شوند؛ بنابراین به نظر عاقلانه می‌رسد که سیاست‌گذاران و نیز به‌طورکلی افراد ذینفع در این حوزه‌ها، به دنبال بررسی علل پیدایش این پدیده‌ها و چگونگی شکل‌گیری روابط در میان اعضای یک شبکه باشند. از این طریق هم می‌توانند از وقوع یک پدیده جلوگیری کنند و یا بستر پیدایش یا گسترش آن را فراهم کنند. یک شرکت تولیدکننده‌ی تلفن همراه برای در اختیار داشتن کل بازار و افزایش سود اقتصادی‌اش تصمیم می‌گیرد به گروهی از افراد در جامعه، مانند دانشجویان یک دانشگاه به‌خصوص، تخفیف‌های ویژه‌ای بدهد. دولت حاکم بر یک کشور برای مهار انتقادات و اعتراضات علیه خود، گروهی از رهبران کلیدی

مخالفش را تحت نظر می‌گیرد تا اعتراضات میان سایرین فراگیر نشود و به‌این‌ترتیب رخدادهای نامطلوب بعدی مهار شود .این نمونه‌ها و موارد بسیار دیگری، مثال‌هایی هستند که اهمیت مطالعه و تحلیل شبکه‌های اجتماعی را مشخص می‌سازد تحلیل شبکه‌های اجتماعی، در ابتدا، به‌عنوان فن کلیدی در جامعه‌شناسی مدرن مطرح شد؛ اما به‌سرعت در رشته‌هایی نظیر مطالعات ارتباطات، علوم اطلاعات، مطالعات سازمانی، اقتصاد، زیست-شناسی راه یافت و پیروان فراوانی پیدا کرد .در ابتدا تحلیل این شبکه‌ها به شیوه‌های سنتی‌ای که مربوط به همان رشته بود، انجام می‌شد که دشواری‌های فراوانی به همراه داشت .مطالعه‌ی شبکه‌های اجتماعی به‌طور عمده به دلایل زیر دشوار است:

- ابعاد این شبکه‌ها عموماً بسیار بزرگ و دائماً در حال گسترش است و غالباً ساختار پیچیده‌ای دارند
- جمع‌آوری اطلاعات و داده‌ها مشکل است .پیدا کردن تمام روابط میان اعضای یک شبکه آسان نیست. حتی در برخی از شبکه‌ها، شناسایی تمام اعضای شبکه دشوار است.
- در بروز یک پدیده و شکل‌گیری ارتباطات، عوامل مختلفی در حوزه‌های گوناگون تأثیرگذار است. آشنایی با یک حوزه به‌تنهایی کافی نیست.

پیدایش شبکه‌های اجتماعی برخطی نظیر فیسبوک و توییتر در اینترنت و نیز به‌کارگیری مفاهیمی از علوم نظری، همچون گراف و شبکه، مطالعه و تحلیل این‌گونه شبکه‌ها را سهولت بخشید و شاخه‌ی جدیدی به نام تحلیل شبکه‌های اجتماعی را به وجود آورد .تحلیل شبکه‌های اجتماعی را می‌توان به دو شاخه‌ی عمده تقسیم کرد:

۱. **تحلیل ساختاری:** در این کتاب، بیشتر ویژگی‌های شبکه مورد بحث است و جنبه‌های تأثیرگذاری افراد، در نظر گرفته نمی‌شود .برخی ویژگی‌های ساختاری از قبیل اندازه‌ی قطر، توزیع درجات، تراکم یال‌ها، میانگین فاصله رئوس از یکدیگر، همبندی شبکه و اندازه‌ی مؤلفه‌ها هستند
۲. **تحلیل رفتاری:** در این نوع تحلیل، جنبه‌های تأثیرگذاری گره‌ها در نظر گرفته می‌شود و موضوع مورد بحث علت و چگونگی شکل‌گیری ساختارها و پدیده‌ها در شبکه است.

برای مثال، در یک شبکه‌ی دوستی وقتی صحبت از وجود رابطه میان دو فرد می‌کنیم، در مورد ساختار شبکه صحبت می‌کنیم؛ درحالی‌که وقتی به یال‌های شبکه وزن یا جهت می‌دهیم و میزان تأثیرگذاری افراد بر یکدیگر را بررسی می‌کنیم، در مورد رفتار شبکه بحث می‌کنیم در سال‌های اخیر مطالعات متنوعی از جنبه‌های مختلف روی این شبکه‌ها انجام شده است که نمونه‌هایی را در زیر آورده‌ایم:

- بررسی افراد یا نقاط کلیدی در انتشار یک خبر در شبکه
- بررسی شکل‌گیری گروه‌های سیاسی و احزاب در جامعه
- بررسی دوره‌ی زمانی همه‌گیر شدن یک بیماری و توقف آن در میان افراد جامعه
- جستجو و بررسی شبکه‌های مجرمانه و مخفی، نظیر شبکه تروریستی یا شبکه‌های اقتصادی فاسد

در بسیاری از این مطالعات، مسئله به بهینه‌سازی یک کمیت، روی آن شبکه تبدیل می‌شود: کمینه کردن زمان انتشار یک خبر در سطح شبکه، بیشینه کردن افرادی که یک کالای خاص را می‌خرند، کمینه کردن افرادی که باید مقاوم شوند تا از گسترش یک بیماری واگیردار جلوگیری شود، مثال‌هایی از بهینه‌سازی یک کمیت است .برای حل این مسائل، شبکه‌ی موردنظر را با یک گراف مدل می‌کنند که اعضای شبکه، همان گره‌های گراف و ارتباط میان اعضا، همان یال‌های شبکه هستند وسعی می‌کنند از ویژگی‌های ساختاری و رفتاری خود شبکه‌ها استفاده کنند تا روش‌ها و الگوریتم‌های مناسب‌تری ارائه دهند یکی از اولین گام‌های مطالعه و تحلیل شبکه‌های اجتماعی، بررسی خواص و ساختار مشترک میان این نوع شبکه‌ها و شناسایی تفاوت‌های میان آن‌ها با سایر شبکه‌ها است .در ادامه برخی از ویژگی‌های مهم این شبکه‌ها را معرفی می‌کنیم و مورد بررسی قرار می‌دهیم.

شبکه‌های اجتماعی در سال‌های اخیر به دلیل گسترش روزافزون و ایجاد شدن دستگاه‌های فعال اینترنتی مانند کامپیوترهای شخصی، دستگاه‌های موبایل و دیگر نوآوری‌های اخیر سخت‌افزاری مانند تبلیت‌های اینترنتی بسیار محبوب شده‌اند. این محبوبیت رو به رشد بسیاری از شبکه‌های اجتماعی مانند فیسبوک و تویتر را اثبات می‌کند. چنین شبکه‌های اجتماعی منجر به یک انفجار عظیم از شبکه

داده محور در طیف گسترده‌ای از زمینه‌ها شده‌اند. شبکه‌های اجتماعی را می‌توان چه درزمینه سیستم‌هایی مانند فیسبوک که به‌صراحت برای تعاملات اجتماعی طراحی‌شده است و یا برحسب سایت‌های دیگر مانند که فیلیکر[1] برای سرویس‌های مختلفی مانند اشتراک‌گذاری محتوا طراحی شده است تعریف کرد، اما علاوه بر این همچنین سطح گسترده‌ای از تعاملات اجتماعی را نیز در برمی‌گیرد. واضح است که مفهوم شبکه‌های اجتماعی به‌صورت خاص به یک شبکه اجتماعی مبتنی بر اینترنت مانند فیسبوک محدود نمی‌شود. مشکل شبکه‌های اجتماعی اغلب در زمینه جامعه‌شناسی ازنظر تعاملات عمومی بین گروهی از موجودیت‌ها موردبررسی قرار می‌گیرد. این تعاملات ممکن است در هر شکل متعارف و یا غیرمتعارف باشد تجزیه‌وتحلیل شبکه‌های اجتماعی (مربوط به نظریه شبکه‌ها) به‌عنوان یک تکنیک کلیدی در جامعه‌شناسی مدرن پدید آمده است. اهمیت این حوزه تحقیقاتی در علوم انسانی، زیست‌شناسی، مطالعات و ارتباطات، اقتصاد، جغرافیا، علوم اطلاعات و ... می‌باشد و به‌همین دلیل به‌عنوان یک موضوع محبوب، مورد مطالعه قرار می‌گیرد. به‌طورمعمول شبکه‌های اجتماعی را می‌توان در قالب گراف نمایش داد که در این گراف گره‌ها معادل با بازیگران شبکه‌های اجتماعی بوده و یال‌های گراف نشان‌دهنده ارتباط بین بازیگران هست. با توجه به ساختار شبکه اجتماعی و یک‌طرفه یا دوطرفه بودن ارتباط، گراف متناظر می‌تواند گراف جهت‌دار یا بدون جهت باشد. همچنین درصورتی‌که قوت ارتباط بین افراد در شبکه اجتماعی یکسان نباشد، گراف متناظر با شبکه یک گراف وزن‌دار خواهد بود که در آن وزن هر یال متناظر با قوت ارتباط است. گراف شبکه‌های واقعی، به لحاظ وارد شدن افراد جدید در فضای شبکه و ایجاد ارتباطات جدید، به‌سرعت در حال رشد هستند. درواقع باید گفت شبکه‌های اجتماعی، محیط‌های بسیار پویایی می‌باشند یک شبکه‌اجتماعی، یک ساختاراجتماعی ساخته‌شده از افراد (سازمان‌ها) است که گره‌ها را تشکیل می‌دهند. این گره‌ها به‌وسیله‌ی انواع خاصی از وابستگی‌های متقابل مانند دوستی، خویشاوندی، منافع مشترک، روابط جنسی، باورهای مشترک و غیره به‌هم متصل شده‌اند. تجزیه‌وتحلیل شبکه‌های اجتماعی، بیانگر روابط اجتماعی است که به کمک تئوری گراف مدل می‌شوند شبکه‌های اجتماعی شامل گره‌ها (افراد یا سازمان‌ها) و روابط (یال‌ها، پیوندها یا اتصالات) می‌باشد. گره‌ها بازیگران فردی درون شبکه‌ها هستند

[1] Flickr

و علاقه‌مندی‌ها همان روابط بین گره‌های گراف هستند، بنابراین ساختارهای مبتنی بر گراف خیلی پیچیده هستند.

ویژگی‌های ساختاری

شبکه‌های اجتماعی ویژگی‌های ساختاری‌ای دارند که آن‌ها را متمایز از سایر گراف‌ها می‌کند . برای مطالعه‌ی این ویژگی‌های، ابتدا کمیتی قابل‌اندازه‌گیری و مناسب تعریف می‌شود .این کمیت روی شبکه‌های اجتماعی و سایر گراف‌ها اندازه‌گیری می‌شود و به‌عنوان مبنایی برای مقایسه‌ی آن‌ها قرار می‌گیرد در این بخش، برخی خصوصیات و کمیت‌های مهم در تحلیل شبکه‌های اجتماعی را بیان می‌کنیم

پدیده جهان کوچک

شبکه‌ای با ویژگی جهان کوچک[1]، گرافی است که با وجود آن که تعداد زیادی از گره‌ها، همسایه‌ی یکدیگر نیستند، ولی هر دو گره دلخواه توسط زنجیره‌ی کوتاهی از گره‌های دیگر به یکدیگر متصل در Lهستند .به‌طور دقیق‌تر شبکه‌ای با این ویژگی، گرافی است که فاصله‌ی بین هر دو گره تصادفی L در شبکه با لگاریتم تعداد گره‌های شبکه N متناسب است:

$$L \propto \log(N)$$ رابطه (۴)

شبکه‌های اجتماعی نظیر فیسبوک، شبکه‌ی اینترنت، ویکی‌پدیا، شبکه‌های ژنتیکی، نمونه‌هایی هستند که ویژگی جهان کوچک را دارا می‌باشند سرآغاز تحقیقات برای بررسی مسئله‌ی جهان کوچک را، شاید بتوان معمای معروف شش گام جدایی[2] دانست که توسط نویسنده‌ی مجارستانی، کارینتی[3]، مطرح شد .معمای کارینتی این بود که دو نفر با در نظر گرفتن زنجیره‌های حداکثر به طول پنج از افراد دیگر، یکدیگر را می‌شناسند. کارینتی به دلیل پیشرفت فن‌آوری در جوامع، شبکه‌های دوستی گسترش‌یافته و فراتر از فاصله‌های جغرافیایی رفته و افزایش دائمی ارتباطات انسانی باعث کوچک شدن دنیای مدرن شده است به‌طوری‌که هر دو نفر با زنجیرهای کوتاه از آشنایانشان به یکدیگر متصل هستند .به دنبال طرح چنین معمایی، به مدت نزدیک به نیم‌قرن، ریاضیدانان و جامعه‌شناسان، مطالعات و آزمایش‌هایی به منظور بررسی این نظریه انجام دادند. یکی از

[1] Small World
[2] Six Degree Of Separation
[3] Frigyes Karinthy

معروف‌ترین این مطالعات، آزمایش مشهور میلگرم[1] روانشناس آمریکایی در سال ۱۹۶۷ بود. میلگرم قصد داشت میانگین فاصله‌ی بین هر دو نفر را به دست آورد .میلگرم تعدادی نامه را به ۱۶۰ فرد به‌طور تصادفی در ایالت نبراسکای آمریکا فرستاد و از آن‌ها خواست نامه را به یکی از دوستان یا آشنایانشان که حدس می‌زنند به گیرنده‌ی نامه نزدیک‌تر باشد بدهند. گیرنده‌ی نامه‌ها سهامدارانی در شهر بوستون بودند. هدف آزمایش این بود که نامه‌ها به دوستان و آشنایان دست‌به‌دست شود تا به گیرنده‌ی نامه برسد .بسیاری از نامه‌ها در این مسیر گم شدند و اصلاً به مقصد نرسیدند .به ازای هر دست‌به‌دست شدن هم روی نامه یک علامت گذاشته می‌شد تا وقتی به مقصد رسید فاصله‌ی شخص اولیه و گیرنده از روی تعداد علامت‌ها تعیین شود .در بین نامه‌هایی که به مقصد رسیدند میانگین علامت‌های روی نامه‌ها ۶ بود .ولی باز به دلیل عدم رسیدن بسیاری از نامه‌ها به مقصد، محققان متقاعد به وجود فاصله‌ی کم بین افراد جامعه نشدند.

تراکم یال‌ها

در شبکه‌ی اجتماعی اگرچه فاصله‌ی هر دو رأس دلخواه چندان زیاد نیست و شبکه همبندی بالایی دارد اما تعداد یال‌ها و به عبارتی تراکم آن‌ها بالا نیست. تعداد یال‌های این شبکه عموماً در حدود $O(n)$و اصطلاحاً از نوع گراف‌های خلوت[2] هست

مقیاس آزاد

شبکه‌ای مقیاس آزاد[3] است که توزیع درجه‌ی گره‌هایش، از توزیع قانون توان پیروی کند.

رابطه (۵) $$P(K) \propto C.K^{-\gamma} \quad 2 < \gamma < 3$$

که $p(k)$ عدد ثابتی برای نرمال کردن است. C عدد ثابتی برای نرمال کردن است. γ پارامتری است که مقدار آن، عموماً در شبکه‌های اجتماعی بین 2 و 3 است .طبق این توزیع، درجه‌ی بسیاری از گره‌ها در یک شبکه‌ی اجتماعی کم است و نیز، تعداد اندکی از گره‌ها درجه‌ی بسیار بالایی دارند. شبکه‌ای اینترنت، شبکه‌های اجتماعی نظیر فیسبوک، شبکه‌های زیستی، شبکه‌ای خطوط هوایی،

[1] Stanely Milgram
[2] Sparse
[3] Scale- Free

مثال‌هایی از شبکه‌هایی با ویژگی مقیاس آزاد هستند. اولین بار پرایس[۱] در مطالعه‌ی گراف ارجاعات در مقاله‌های علمی[۲] متوجه شد که تعداد لینک‌هایی که به یک مقاله اشاره می‌کنند، توزیع قانون توان را دارد .وی همچنین مکانیزمی برای توضیح علت وقوع چنین پدیده‌ای در شبکه‌ی ارجاعات ارائه داد که بعدها به مکانیزم اتصال ترجیحی[۳] مشهور شد. در این مکانیزم بیان شد که هر گره‌ای تمایل دارد به گره‌ای با اعتبار بیشتر متصل شود .درجه‌ی یک گره نیز یکی از معیارهایی می‌تواند باشد که اعتبار آن گره را مشخص می‌کند

بستار سه‌تایی

بستار سه‌تایی [۴] مفهومی است که توسط جامعه‌شناس آلمانی، سیمل[۵] در سال ۱۹۰۰ مطرح شد و بعدها با انتشار مقاله‌ای از جامعه‌شناس آمریکایی، گرنوتر[۶] در سال ۱۹۷۳، به طور وسیعی در شبکه‌های اجتماعی به کار رفت .اصل بستار سه‌تایی می‌گوید اگر در یک شبکه‌ی اجتماعی، دو گره غیر مجاور، همسایه‌ی مشترکی داشته باشند، احتمال ایجاد یال بین این دو گره، از احتمال ایجاد یال بین دو گره تصادفی و غیر مجاور در آن شبکه، بیشتر است. اگر شبکه را در بازه‌های زمانی متعدد بررسی کنیم، به‌خوبی افزایش تعداد سه‌تایی‌ها (مثلث‌ها) را می‌توانیم مشاهده کنیم .چنین ساختاری، با این چگالی، در گراف‌های کاملاً تصادفی دیده نمی‌شود .این ساختار ناشی از درستی اصل بستار سه‌تایی است.

ضریب خوشه‌بندی

به دنبال مطرح شدن اصل بستار سه‌تایی و تراکم بالای مثلث‌ها، معیارهایی برای اندازه‌گیری چگالی این سه‌تایی‌ها در شبکه‌های اجتماعی معرفی شد. ضریب خوشه‌بندی، یکی از معروف‌ترین این معیارهاست. همچنین ضریب خوشه‌بندی معیاری است که مشخص می‌کند گره‌های یک شبکه تا چه حد تمایل دارند باهم تشکیل یک خوشه دهند. شواهد نشان می‌دهد که در شبکه‌های واقعی و

[1] Price
[2] Citation Between Scientific Papers
[3] Preferential Attachment Preferential Attachment
[4] Closure Triad
[5] George Simmel
[6] Mark Granovetter

پیچیده و به‌خصوص در شبکه‌های اجتماعی گره‌ها تمایل دارند با برخی گره‌های دیگر، گروه‌هایی تشکیل دهند، به‌طوری‌که گره‌های داخل یک گروه شباهت زیادی به یک گراف کامل دارند.

تعریف ضریب خوشه‌بندی

این خصوصیت بیان می‌کند که دو گره در شبکه اجتماعی که همسایه یک گره دیگر هستند نسبت به دو گره که به‌طور تصادفی از کل گره‌های موجود در شبکه اجتماعی انتخاب شوند با احتمال بیشتری با یکدیگر در ارتباط هستند .این خصوصیت بر اساس ضریب خوشه‌بندی قابل‌اندازه‌گیری است.

هدف از خوشه‌بندی

هدف خوشه‌بندی یافتن خوشه‌های مشابه می‌باشد معیار تعیین بهترین خوشه‌بندی را می‌توان نشان داد که هیچ معیار مطلقی برای بهترین خوشه‌بندی وجود ندارد بلکه این بستگی به مسئله و نظر کاربر دارد که باید تصمیم بگیرد که آیا نمونه‌ها به‌درستی خوشه‌بندی شده‌اند یا خیر. بااین‌حال معیارهای مختلفی برای خوب بودن یک خوشه‌بندی ارائه می‌شود که می‌تواند کاربر را برای رسیدن به یک خوشه‌بندی مناسب راهنمایی کند یکی از مسائل مهم در خوشه‌بندی انتخاب تعداد خوشه‌ها می‌باشد. در بعضی از الگوریتم‌ها تعداد خوشه‌ها از قبل مشخص می‌شود و در بعضی دیگر خود الگوریتم تصمیم می‌گیرد که داده‌ها به چند خوشه تقسیم شوند.

یادگیری با نظارت در مقابل یادگیری بدون نظارت

در یادگیری با نظارت از ابتدا دسته‌ها مشخص هستند و هر یک از داده‌های آموزشی به دسته‌ای خاص نسبت داده می‌شود و اصطلاحاً گفته می‌شود ناظری وجود دارد که در هنگام آموزش اطلاعاتی علاوه بر داده‌های آموزش در اختیار یادگیرنده[1] قرار می‌دهد. ولی در یادگیری بدون نظارت هیچ اطلاعاتی به‌جز داده‌های آموزشی در اختیار یادگیرنده قرار ندارد و این یادگیرنده است که بایستی در داده‌ها به دنبال ساختاری خاص بگردد.

[1] Learner

مسائل درگیر با روش‌های خوشه‌بندی موجود

متأسفانه چندین مسئله در خصوص روش‌های خوشه‌بندی مطرح است که هنوز به شکل کامل پاسخ داده نشده‌اند؛ و همچنان تلاش‌های بسیاری به‌منظور حل آن‌ها انجام می‌گیرد.

روش‌های خوشه‌بندی قادر نیستند تمامی نیازهای مسائل را به‌طور هم‌زمان برآورده کنند.

به دلیل پیچیدگی محاسباتی زیاد در برخورد با مجموعه داده‌های بزرگ با تعداد داده زیاد و تعداد ویژگی‌های زیاد برای هر داده عملی نیستند.

به دلیل وابستگی شدید به تعریف معیار شباهت بین داده‌ها در مسائلی که تعریف معیار شباهت مشکل باشد نتایج مطلوبی تولید نمی‌کنند. (در داده‌ها با تعداد ویژگی زیاد)

برای نتایج آن‌ها می‌توان تفسیرهای مختلفی بیان کرد.

خوشه‌بندی در مقابل چندی‌سازی برداری

همان‌گونه که بحث شد، خوشه‌بندی نوعی سازمان‌دهی داده‌ها است، بر اساس شباهتی که بین آن‌ها تعریف می‌شود به‌گونه‌ای که شباهت بین داده‌هایی که درون یک خوشه قرار می‌گیرند، نسبت به داده‌هایی که درون خوشه‌های متفاوت قرار می‌گیرند، بیشتر باشد.

در کاربردهای ارتباطی و فشرده‌سازی داده‌ها از روش‌هایی به نام چندی‌سازی برداری استفاده می‌شود که از بعضی جنبه‌ها می‌توان آن‌ها را معادل خوشه‌بندی در نظر گرفت. در چندی‌سازی برداری نیز داده‌ها بر اساس میزان شباهت بین آن‌ها به دسته‌هایی تقسیم می‌شوند و هر دسته به‌وسیله یک بردار که به آن کلمه کد[1] گفته می‌شود جایگزین می‌گردد. به مجموعهٔ این کلماتِ کد اصطلاحاً کتابِ کد[2] گفته می‌شود.

در بعضی از بحث‌های علمی بین خوشه‌بندی و چندی‌سازی برداری تفاوت‌هایی قائل می‌شوند. زیرا خوشه‌بندی را یک رهیافت بدون نظارت برای تحلیل داده‌ها در نظر می‌گیرند ولی چندی‌سازی برداری را روشی برای کشف خوشه‌ها نمی‌شناسند بلکه آن را راهی برای نمایش داده‌ها با تعداد عناصر کمتر

[1] CodeWord
[2] CodeBook

به‌گونه‌ای که اطلاعات ازدست‌رفته حداقل شود، می‌شناسند. علی‌رغم تفاوت بیان‌شده می‌توان روش‌های بکار رفته در هر یک آن‌ها را در دیگر نیز بکار برد در اینجا بین خوشه‌بندی و چندی‌سازی برداری تفاوتی قائل نمی‌شویم.

روش‌های خوشه‌بندی

روش‌های خوشه‌بندی را می‌توان از چندین جنبه تقسیم‌بندی کرد:

۱- خوشه‌بندی انحصاری[1] و خوشه‌بندی با هم‌پوشی[2]

در روش خوشه‌بندی انحصاری پس از خوشه‌بندی هر داده دقیقاً به یک خوشه تعلق می‌گیرد مانند روش خوشه‌بندی K-Means. ولی در خوشه‌بندی با همپوشی پس از خوشه‌بندی به هر داده یک درجه تعلق به ازای هر خوشه نسبت داده می‌شود. به عبارتی یک داده می‌تواند با نسبت‌های متفاوتی به چندین خوشه تعلق داشته باشد. نمونه‌ای از آن خوشه‌بندی فازی است.

۲- خوشه‌بندی سلسله مراتبی[3] و خوشه‌بندی مسطح[4]

در روش خوشه‌بندی سلسله مراتبی، به خوشه‌های نهایی بر اساس میزان عمومیت آن‌ها ساختاری سلسله مراتبی نسبت داده می‌شود؛ مانند روش Single Link. ولی در خوشه‌بندی مسطح تمامی خوشه‌های نهایی دارای یک میزان عمومیت هستند مانند K-Means. به ساختار سلسله مراتبی حاصل از روش‌های خوشه‌بندی سلسله مراتبی دندوگرام[5] گفته می‌شود.

با توجه با اینکه روش‌های خوشه‌بندی سلسله مراتبی اطلاعات بیشتر و دقیق‌تری تولید می‌کنند برای تحلیل داده‌های با جزئیات پیشنهاد می‌شوند ولی از طرفی چون پیچیدگی محاسباتی بالایی دارند برای مجموعه داده‌های بزرگ روش‌های خوشه‌بندی مسطح پیشنهاد می‌شوند.

الگوریتم k-means استاندارد

در این الگویتم داده‌ها را به k خوشه مجزا تقسیم کنیم. این الگوریتم به دو فاز مجزا تقسیم می‌شود:

---

[1] Exclusive or Hard Clustering
[2] Overlapping or Soft Clustering
[3] Hierarchical
[4] Flat
[5] Dendogram

در فاز اول برای هر خوشه یک نقطه را به‌عنوان نقطه ثقل خوشه یا نقطه مرکزی خوشه به دست می‌آوریم و در فاز بعدی به دست می‌آوریم که هر نقطه از مجموعه به کدام مرکز خوشه نزدیک‌تر است و آن نقطه را به خوشه مربوطه نسبت می‌دهیم. در حالت کلی برای به دست آوردن فاصله بین نقاط داده و مراکز خوشه‌ها از فاصله اقلیدسی استفاده می‌شود. زمانی که تمام نقاط در خوشه‌ها قرار گرفتند مرحله اول به اتمام رسیده و خوشه‌بندی اولیه انجام می‌شود. سپس دوباره برای خوشه‌ها مراکز جدیدی به دست می‌آوریم و فاصله هر نقطه را نسبت به این نقاط مرکزی اندازه می‌گیریم تا خوشه‌ها به‌روز شوند و این کار تا زمانی ادامه پیدا می‌کند که خوشه‌ها همگرا شوند؛ اما اشکال عمده این الگوریتم این است که با توجه به مقدار مراکز اولیه خوشه‌های متفاوتی تولید می‌شود و درنتیجه کیفیت خوشه‌های نهایی شدیداً به انتخاب مراکز اولیه خوشه‌ها وابسته است. این الگوریتم ازلحاظ محاسباتی گران و متناسب با تعداد نقاط، تعداد خوشه‌ها و تعداد تکرارها نیاز به زمان دارد. در قسمت بعد الگوریتم اصلاح‌شده k-means را بررسی می‌کنیم که این اشکالات را رفع کرده است.

الگوریتم k-means بهبودیافته

همان‌طور که بیان شد، الگوریتم خوشه‌بندی k-means استاندارد از لحاظ محاسباتی سنگین و کیفیت نتایج خوشه‌های آن، شدیداً به انتخاب مراکز اولیه خوشه‌ها وابسته است. به همین دلیل محققان سعی کرده‌اند تا با ارائه روش‌هایی این نقایص را برطرف کنند و الگوریتم k-means را بهبود ببخشند.

در روش k-means بهبودیافته هر دو فاز الگوریتم k-means برای بهبود دقت و کارایی آن اصلاح می‌شود. در مرحله اول مراکز اولیه خوشه‌ها برای تولید خوشه‌ها با دقت بالاتر از یک روش سیستماتیک به‌جای انتخاب تصادفی استفاده می‌کنند.

مرحله دوم با تشکیل خوشه‌های اولیه بر اساس فاصله نسبی هر نقطه از مراکز اولیه خوشه‌ها شروع می‌شود. این خوشه‌ها متعاقباً به‌وسیله یک روش اکتشافی تنظیم می‌شوند، بنابراین کارایی بهبود می‌یابد.

الگوریتم‌های k-means توزیع‌شده

همان‌طور که اشاره شد با توجه به توزیع‌شدگی داده‌ها در مجموعه سایت‌های مختلف و عدم متمرکز کردن آن‌ها در یک مجموعه نیاز به الگوریتم‌هایی می‌باشد که در محیط‌های توزیع‌شده به‌خوبی عمل کنند. الگوریتم k-means یکی از الگوریتم‌هایی است که می‌تواند در این موارد نتایج مورد قبولی را ارائه دهد، این الگوریتم به‌صورت استاندارد قادر خواهد بود در یک مجموعه متمرکز کار کند، لکن باکمی تغییر در ساختار این الگوریتم و ترکیب آن با دیگر الگوریتم‌ها می‌توان آن را به‌صورت توزیع‌شده استفاده کرد. در زیر به ذکر نمونه‌هایی از این الگوریتم که در محیط‌های توزیع‌شده کار می‌کنند، می‌پردازیم.

خوشه‌بندی گراف

در مبحث خوشه‌بندی گراف، گروه‌بندی رأس‌های گراف به خوشه‌ها با در نظر گرفتن ساختار یال‌های گراف، به‌گونه‌ای است که باید یال‌های درون هر خوشه ماکزیمم و بین خوشه‌ها حداقل تعداد یال موجود باشد. خوشه‌بندی گراف به معنای گروه‌بندی گره‌های گراف ورودی به خوشه‌ها است.

در محیط (فضای) گراف، هر خوشه باید مستقیماً متصل باشد، باید حداقل یک مسیر چندگانه متصل برای هر جفت از گره‌ها درون یک خوشه وجود داشته باشد. اگر رأس از رأس قابل‌دسترس نباشد، آن‌ها نباید در خوشه‌های یکسان (مشابه) قرار بگیرند.

مقاومت شبکه

یک شبکه در صورتی مقاوم است که در برابر حذف درصدی از گره‌ها و یال‌هایش از کار نیفتد و کارایی‌اش را در حد قابل‌توجهی از دست ندهد. حذف گره‌ها و یال‌ها معمولاً در اثر از بین رفتن تصادفی یا تخریب عمدی رخ می‌دهد .بدیهی است که برای اندازه‌گیری کارایی یک شبکه، نیاز به تعریف کارایی و انتخاب یک معیار مناسب داریم. البته این تعریف بستگی به کاربرد شبکه دارد .در بسیاری از این موارد میانگین کوتاه‌ترین فاصله‌ی بین رئوس در شبکه را معیار اندازه‌گیری کارایی در نظر می‌گیرند. در آزمایش‌هایی که برای بررسی مقاومت شبکه انجام‌شده دو روش وجود دارد:

- **از بین رفتن تصادفی:** در این روش در چندین مرحله به‌صورت تصادفی درصدی از رئوس یا یال‌های شبکه را انتخاب و از گراف حذف می‌کنند. (برای مثال در هر مرحله ۵ درصد حذف می‌کنند.) در هر مرحله کارایی شبکه را با توجه به معیار انتخابی اندازه‌گیری و با مقدار آن در مرحله اولیه (پیش از حذف هرگونه رأس یا یالی) مقایسه می‌کنند. در این روش به نحوه‌ی مقاومت شبکه را در برابر اشکالات احتمالی که در شبکه پیش می‌آید بررسی می‌کنند به همین خاطر باید انتخاب‌ها تصادفی باشد.
- **تخریب عمدی:** تنها تفاوت این روش با روش قبلی، انتخاب رئوس یا یال‌هاست که نباید تصادفی باشد .در این حالت فرض بر وجود یک دشمن است که قصد از کار انداختن شبکه را دارد و برای ایجاد اختلال در شبکه، گره یا لینک‌های کلیدی‌تر[1] را حذف می‌کند. معیارهایی نیز برای کلیدی بودن رئوس و یال‌ها وجود دارد، از قبیل درجه یا مرکزیت یک رأس در شبکه، باری که یک یال تحمل می‌کند، سهمی که یک یال در کوتاه‌ترین فاصله بین رئوس دیگر دارد. بدیهی است که مقاومت شبکه ارتباط تنگاتنگی باکیفیت همبندی و شکل اتصالات شبکه دارد. شبکه‌ای که یال یا رأس برشی دارد، مقاومت بسیار پایینی دارد. برعکس مقاومت شبکه‌ای با ضریب خوشه‌بندی زیاد به نظر می‌رسد که بالاست.

مرکزیت

بدیهی است که میزان اهمیت افراد مختلف در یک جامعه یکسان نیست. برخی از آن‌ها به دلیل جایگاه اجتماعی، روابط و یا دوستان بانفوذشان از اهمیت بیشتری برخوردارند. این اهمیت امکان دسترسی بیشتر به اطلاعات و یا نقش پررنگ‌تر در انتقال آن را برای افراد فراهم می‌کند .ازاین‌روست که افراد خاصی را در اجتماع تأثیرگذار و تعیین‌کننده میدانیم. اهمیت و محبوبیت این افراد در زمینه‌های مختلف با معیارهای متفاوتی تعیین می‌شود، به‌عنوان‌مثال سیاستمداران، مدیران، نویسندگان و هنرمند اندر یک جامعه ازجمله این افراد مرکزی به‌حساب می‌آیند .مفهوم مرکزیت در شبکه‌های اجتماعی نیز صادق است ازاین‌رو برخی از رأس‌ها مهم‌تر و مرکزی‌تر از دیگر رأس‌ها به نظر می‌رسند .در دهه‌ی ۱۹۵۰ تلاش برای تعریف معیارهای مرکزیت در گراف آغاز شد .در طی سال‌ها،

[1] Critical

معیارهای مرکزیت متفاوتی معرفی شده است، این امر بدین دلیل است که هیچ معیار واحدی برای مرکزیت وجود ندارد. این مفهوم همچنان از محبوبیت خاصی برخوردار است و از مفاهیم پایه‌ای تحلیل شبکه‌ها به‌حساب می‌رود. برای هر مسئله، این معیار متفاوت و متناسب با محیط مسئله و رابطه‌ی بین رأس‌ها تعریف می‌شود. در این جا چند مثال ساده از لزوم وجود معیارهای متفاوت برای مفهوم مرکزیت آورده شده است.

مثال خود را با مسئله‌ی انتخاب یک نماینده در کلاس درس آغاز می‌کنیم .برای این شبکه اجتماعی روابط مختلفی می‌توان در نظر گرفت، یکی از این روابط می‌تواند برقراری یال جهت‌دار بین دو رأس (الف) و (ب) در شرایطی باشد که رأس (الف) بخواهد به رأس (ب) رأی بدهد. در چنین محیطی، رأسی به‌عنوان نماینده انتخاب خواهد شد که بیش‌ترین رأی را داشته باشد، پس کافی است عضو مرکزی را رأسی با بیش‌ترین درجه ورودی معرفی کنیم .این نوع از معیارهای مرکزیت را تحت عنوان معیارهای مرکزیت مبتنی بر درجه می‌نامیم.

شبکه‌ی اجتماعی انتخاب نماینده در کلاس را می‌توان از منظر دیگری نیز مورد مطالعه قرار داد، کافی است رابطه‌ای که یال (الف) را به (ب) متصل می‌کند، توانایی (الف) برای متقاعد کردن (ب) به تغییر رأی باشد .یک شبکه با چنین رابطه‌ای را شبکه تأثیرات می‌نامند. بی‌آنکه به کلیت مسئله خدشه وارد شود می‌توان فرض کرد که رقابت اصلی بین دو کاندیدای مختلف x و y است که موجب تقسیم کلاس به دو گروهX وY می‌شود. گروه X به فرد x و گروه Y به فرد y رأی خواهند داد. برخی از افراد در شبکه، دوستانی از هر دو گروه دارند .فرض کنید یکی از این افراد خود بخواهد به فرد x رأی دهد و درعین‌حال دوستان زیادی در Y داشته باشد که توانایی قانع کردن آن‌ها را دارد. فرد موردنظر در سیستم معرفی شده فردی مرکزی به‌حساب می‌آید چرا که توانایی متقاعد کردن تعداد زیادی از افراد را داراست و در نتیجه در رأی نهایی انتخابات تأثیرگذار است. در چنین سبک‌های مرکزیت را متعلق به افرادی می‌دانیم که در انتقال عقاید در شبکه نقش بیشتری دارند. این نوع از مرکزیت که به خاصیت در میان واقع بودن می‌پردازد را مرکزیت میانگی می‌نامیم که اساس این تحقیق را در برمی‌گیرد. از زاویه‌ی دیگری نیز می‌توان شبکه اجتماعی کلاس را بررسی کرد. می‌توانیم رابطه‌ی حاکم در برقراری یال بین دو رأس را رابطه‌ی دوستی در نظر بگیریم .در این صورت مرکزیت

یک فرد که دوستان مهم‌تری دارد نسبت به افرادی که با افراد کم‌اهمیت‌تری دوست هستند، بیشتر خواهد بود .در چنین محیط‌هایی مرکزیت هر فرد تحت تأثیر مرکزیت همسایگانش است ازاین‌رو این‌گونه مرکزیت‌ها را معمولاً بازخوردی می‌نامند. باآنکه در طی سال‌ها معیارهای متفاوتی از مفهوم مرکزیت پیشنهاد شده است، این مفهوم فرای توصیف‌هایی چون برتری رأسی و یا اهمیت ساختاری، خوش‌تعریف نیست .به‌علاوه افراد مختلف تفسیرهای متفاوتی از جمله خودمختاری، کنترل، ریسک، آشکاری، تأثیرپذیری، تعلق، دلالی، استقلال یا قدرت از این معیار ارائه داده‌اند.

در طی سال‌ها، تلاش برای ارائه تعریف ریاضی دقیق برای مرکزیت انجام شده است. به‌عنوان‌مثال در مجموعه‌ای از ضابطه‌ها را برای این معیار معرفی نموده است .ازجمله این ضوابط آن است که اضافه کردن قید به هر رأس باید منجر به افزایش درجه مرکزیت شود .به‌بیان‌دیگر مرکزیت هیچ رأسی نباید به دلیل افزودن قید در شبکه کاهش پیدا کند .این شرط اگرچه جالب به نظر می‌رسد و می‌تواند گامی مهم در خوش‌تعریف نمودن مفهوم مرکزیت باشد اما دارای مشکلاتی نیز است. یکی از این مشکلات حذف بسیاری از معیارهای مطرح مرکزیت ازجمله معیارهای مبتنی بر مفاهیم واسطه‌ای است.گرچه این قید خصوصیات مطلوبی از مرکزیت را ارائه می‌دهد اما تلاشی برای تعریف مفهوم نمی‌کند. در رویکرد دیگری را برای پاسخ به سؤال «مرکزیت چیست» در پیش گرفته است. وی معیارهای مختلف منتشرشده مرکزیت را بررسی کرده و در نهایت آن‌ها را در قالب سه گروه مفهومی کلی درجه‌ای و نزدیکی و میانگی طبقه‌بندی کرده است. وی برای هر کدام از این سه دسته فرمول‌های استانداردی نیز معرفی کرده است .هر سه این دسته‌ها بیش‌ترین مرکزیت را به رأس میانی در شبکه ستارهای شکل نسبت می‌دهند که توجیه درستی از مفهوم است.

به‌تازگی دید جدیدی مبتنی بر مدل را از مرکزیت معرفی نموده است که بیشتر بر خروجی رأس‌ها زمانی که شبکه شامل جریان گذرنده از یال‌هاست تمرکز دارد .وی معتقد است که مسائل بنیادی مطرح در رابطه با جریان‌ها در شبکه، عموماً دو سؤال زیر است:

- جریان ترافیک هر از چند گاه از رأس‌ها عبور می‌کند؟
- برای آنکه یک رأس جریان فرستاده شده در شبکه را دریافت کند چقدر زمان لازم است؟

وی بر اساس پاسخ این دو سؤال، مرکزیت مبتنی بر جریان را معرفی می‌کند.

برای بررسی مفهوم مرکزیت دو دیدگاه متفاوت موجود است .دیدگاه مبتنی بر تئوری گراف که مفاهیم مرکزیت را با توجه به نوع خصوصیات محاسبه شدنشان طبقه‌بندی می‌کند و دیدگاه مبتنی بر مدل که بیشتر بر خروجی مرکزیت تمرکز دارد. تمرکز غالب ما بر معیارهای مبتنی بر تئوری گراف است. ازاین‌رو معمولاً شبکه‌های اجتماعی را به‌صورت یک گراف بدون جهت $G = (V, E)$ در نظر می‌گیریم که همان‌طور که گفتیم V مجموعه رأس‌ها و E مجموعه یال‌هاست. از نمایش‌های رایج برای گراف استفاده از مفهوم ماتریس مجاورت است، درایه‌های این ماتریس از یک و صفر تشکیل شده است که نمایانگر وجود و یا عدم وجود یال بین رأس‌ها است.

تعاریف اولیه در گراف

شبکه‌های اجتماعی ترکیبی از جنبه‌های متفاوتی از رفتار و تعاملات پیچیده در میان افراد مختلف می‌باشد، بنابراین تجزیه‌وتحلیل این شبکه‌ها با ابزارهای سنتی کار آسانی نیست. درنتیجه برای رسیدن به این هدف از ساختار مبتنی بر گراف استفاده می‌شود که اغلب بسیار پیچیده می‌باشند.

نظریه گراف

نظریه گراف شاخه‌ای از ریاضیات است که درباره اشیاء خاصی در ریاضیات به نام گراف بحث می‌کند. این مبحث درواقع شاخه‌ای از توپولوژی است که با جبر و نظریه ماتریس‌ها پیوند مستحکم و تنگاتنگی دارد. نظریه گراف برخلاف شاخه‌های دیگر ریاضیات نقطه آغاز مشخصی دارد و آن انتشار مقاله‌ای از لئونارد اویلر، ریاضیدان سوئیسی، برای حل مسئله پل‌های کونیگسبرگ در سال ۱۷۳۶ است.

تعریف دقیق‌تر گراف به این صورت است، که یک گراف$G=(V, E)$ شامل مجموعه‌ای از گره‌های V و مجموعه‌ای از یال‌های E می‌باشد که هر یال دو گره را به یکدیگر متصل می‌کند.

- همیشه گراف G یک گراف بدون جهت و بدون وزن است؛ بنابراین(v_i , v_j) و (v_j , v_i) نشان دهنده یک لبه یکسان در گراف می باشند.
- گراف G دارای لبه های چند گانه نیست؛ بنابراین دو گره v_i و v_j تنها توسط یک لبه در E به یکدیگر متصلند و هیچ لبه ی دیگری در E آن‌ها را به یکدیگر متصل نمی کند.

- گراف G فاقد حلقه می‌باشد (یک گره نمی‌تواند به خودش متصل شود).
- گراف G نشان دهنده روابط دوستی در میان کاربران شبکه اجتماعی آنلاین می‌باشد. در این گراف گره ها نشان دهنده کاربران و هر یال متصل کننده دو کاربر بیانگر ارتباط دوستی موجود در میان این دوکاربر می‌باشد.

همچنین تعاریف اولیه زیر را برای یک گراف بدون جهت خواهیم داشت:

- **درجه راس:** مجموع تعداد لبه های متصل به یک راس درجه آن راس نامیده می‌شود و با $\deg(v_i)$نشان داده می‌شود.
- **ماتریس مجاورت:** ماتریس مجاورت A مربوط به گراف G ماتریسی مربع با سطرها و ستون هایی برچسب خورده با رئوس گراف است که در خانه‌های (v_i , v_j) مربوط به آن، با توجه به اینکه دو کاربر v_i و v_j دوست هستند یا خیر مقادیر ۰ و ۱ قرارمی‌گیرد. برای یک گراف بدون جهت ماتریس مجاورت، متقارن است.
- **مسیر:**یک مسیر $p(v_0,v_x)$ از راس مبدا v_0 به راس مقصد v_x دنباله ای از یالها به صورت $(v_0,v_1), (v_1,v_2), \dots , (v_{(x-1)},v_x), \text{where } e_i=(v_i,v_{(i+1)}) \in E \ (0\leq i<x)$می‌باشد. برای دو راس v_i و v_j کوتاهترین مسیر میان آن‌ها، مسیری با کمترین تعداد یال می‌باشد.

نمادهای پر استفاده در این پایان نامه در جدول (۲) نمایش داده شده است.

جدول (۲): نمادهای پر تکرار و توضیح مختصر آن ها

نماد	توضیح
G	گراف بدون جهت و بدون وزن
V	مجموعه ای از گره های گراف
E	مجموعه ای از یال های گراف
N	تعداد رئوس در گراف G
v_i	گره گراف G
e_i	یال گراف G
$P(v_o,v_x)$	مسیری از گره مبدا v_o به گره مقصد v_x
$Deg(v_i)$	درجه ی گره v_i
A	ماتریس مجاورت
K	تعداد خوشه های درنظر گرفته شده در گراف G
a_i	سن کاربر i
d_a	فاصله سنی دو کاربر
g_i	جنسیت کاربر i
$Sim_{location}$	شباهت بر اساس مکان جغرافیایی
Sim_{gender}	شباهت بر اساس جنسیت دو کاربر
Sim_{music}	شباهت براساس هنرمندان مورد علاقه
Sim_{friend}	شباهت براساس دوستان مشترک
Sim_{age}	شباهت سنی
$Sim_{profile}$	شباهت پروفایلی دو کاربر
Sim_{struct}	شباهت ساختاری
SIM (u,v)	شباهت کلی دو کاربر uوv

در ادامه هر کدام از شاخص‌های مرکزیت را توضیح می‌دهیم.

مرکزیت درجه[1]

مرکزیت درجه‌ای، ساده‌ترین و معروف‌ترین معیار برای تعیین مرکزیت است .در مرکزیت درجه‌ای هر رأس را تعداد یال‌هایی تعریف می‌کند که با آن رأس مجاور هستند. شاخص مرکزیت درجه‌ای یکی از شاخص‌های شبکه‌ای است که در تحلیل ساختار کل شبکه‌ها و موقعیت‌های افراد در شبکه مفید است. این معیار توسط ماتریس مجاورت A صورت زیر محاسبه می‌شود.

$$C_i^{DEG} = \sum j \; a_{ij}$$ رابطه (۶)

درجه‌ی گره، تعداد پیوندهایی است که آن گره دارد .در توییتر تعداد دنبال کنندگان و در فیسبوک تعداد دوستان گره، درجه‌ی آن گره هستند .در شبکه‌های جهت‌دار، دو شاخص درجه وجود دارد؛ درون درجه[2]، تعداد پیوندهایی است که به یک گره اشاره دارند؛ برون درجه[3]، تعداد پیوندهایی است که از گره منشأ می‌گیرند و به گره‌های دیگر اشاره دارند .در یک ساختار ستارهای، گرهای که در مرکز قرار گرفته است، به نظر می‌رسد به دلیل جایگاه استثنایی‌ای که دارد، اهمیت فراوانی نیز داشته باشد، به این دلیل که همه‌ی یال‌ها از آن رد می‌شوند؛ اما در شبکه‌های دنیای واقعی، چنین ساختارهای ستارهای شکلی بسیار نادرند .درنهایت اهمیت درجه‌ای، بسیار به طبیعت اتصال‌ها و نیز طبیعت روابط بستگی دارد. به‌هرحال، درجه، یک شاخص مفید در داده‌های شلوغ است و یک آغاز آسان محسوب می‌شود

موجودیتی با مرکزیت درجه‌ای بالا، دارای ویژگی‌های زیر است:

- به‌طورکلی یک بازیگر فعال در شبکه است؛
- اغلب یک متصل‌کننده یا قطب در شبکه است؛
- ضرورتاً مرتبط‌ترین موجودیت در شبکه نیست (یک موجودیت ممکن است روابط زیادی داشته باشد که اکثریت آن‌ها دارای موجودیت‌های سطح پایین باشند)؛
- ممکن است در موقعیت ممتازی در شبکه قرار داشته باشد؛
- اغلب می‌تواند به‌عنوان شخص سوم یا واسطه شناخته شود

[1] Degree Centrality
[2] In-Degree
[3] Out-Degree

مرکزیت نزدیکی[1]، یافتن شایعه‌پراکنان[2]

با توجه به اینکه قدرت را می‌توان به‌وسیله مذاکرات و مبادلات مستقیم نشان داد، اما قدرت همچنین از طریق عمل کردن، به‌عنوان یک (نقطه مرجع) که سایر افراد به‌وسیله آن خودشان را مورد قضاوت قرار می‌دهند، رخ می‌دهد .همچنین قدرت به‌وسیله مرکز توجه شدن توسط افرادی که دیدگاه‌هایشان به‌وسیله تعداد زیادی از افراد دیگر شنیده می‌شود، به وجود می‌آید .افرادی که قادرند در کوتاه‌ترین طول مسیر به دیگر افراد برسند، یا کسی که توسط دیگران در کوتاه‌ترین طول مسیر در دسترس، است، موقعیت‌های مطلوبی دارند .این مزیت ساختاری می‌تواند به قدرت ترجمه شود

مرکزیت نزدیکی، فاصله یک فرد با کلیه افراد دیگر در شبکه را می‌سنجد .افراد با میزان نزدیکی بالا، احتمالاً اطلاعات را خیلی سریع‌تر از دیگران دریافت می‌کنند، به خاطر اینکه میانجی‌های کمتری بین آن‌ها وجود دارد .در سنجش مرکزیت نزدیکی، ارزیابی از طریق قضاوت کردن درباره نزدیکی یک فرد به افراد دیگر صورت می‌گیرد .این نوع مرکزیت از طریق طول مسیرها یا گام‌هایی که برای یک فرد موردنیاز است تا به دیگر افراد در شبکه برسد، اندازه‌گیری می‌شود .افرادی که قادرند به دیگر افراد با طول مسیر کوتاه‌تری برسند یا آن‌هایی که با طول مسیرهای کوتاه‌تر توسط دیگران دسترس‌پذیر ترند، به طورکلی قدرت و نفوذ بیشتری در درون شبکه دارند.

شاخص مرکزیت نزدیکی، بر اساس فاصله ژئودسیک محاسبه می‌شود .در برای محاسبه مرکزیت نزدیکی به‌صورت مجموع کمترین مسیرها از یک رأس به دیگر رأس‌های موجود عمل نمود .چنانچه در فرمول نیز نشان داده شده است در محاسبه این معیار می‌توان از ماتریس کوتاه‌ترین مسیر D کمک گرفت.

رابطه (۷)
$$c_c(v) = \frac{1}{\sum_{t \in V} d(v,t)}$$

اینکه یک موجودیت در شبکه چقدر سریع می‌تواند به موجودیت‌های بیشتری در آن شبکه دسترسی پیدا کند، شاخصی است که مرکزیت نزدیکی را سنجش می‌کند .موجودیتی با مرکزیت نزدیکی بیشتر به‌طورکلی دارای ویژگی‌های زیر است:

- دسترسی سریعی به سایر موجودیت‌ها در شبکه دارد؛

[1] Closeness Centrality
[2] Gossipmongers

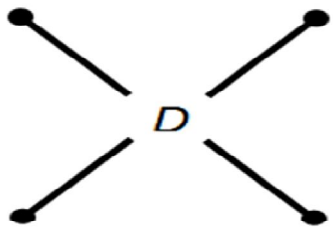
- مسیر کوتاهی به سایر موجودیت‌ها دارد؛
- به سایر موجودیت‌ها نزدیک است؛
- رؤیت پذیری بالایی درباره آنچه در شبکه در حال اتفاق افتادن است، دارد؛

نکته قابل‌توجه این است که اگر شبکه دارای موجودیتی باشد که هیچ پیوندی دریافت نکرده است (به هیچ موجودیت دیگر پیوند نداده باشد). مقدار نزدیکی برای موجودیت در شبکه صفر خواهد بود.

از مسائلی که در این حیطه بسیار مطرح است مسئله مکان‌یابی امکانات رفاهی است .مسئله به مکان‌یابی یک مرکز تجاری در شهر می‌پردازد .برای جذب مشتری بیشتر و فراهم کردن آسایش آن‌ها بهتر است مرکز تجاری در جایی احداث شود که به همه‌ی ساکنان شهر نزدیک باشد .به بیانی دیگر، هدف یافتن رأسی چون $v \in V$ با مینیمم فاصله‌ی کلی تا باقی رأس‌ها در گراف شهر است.

مرکزیت میانگی[1]، یافتن تنگناهای ارتباطی[2] یا پل‌های اجتماع[3]

فاصله میان دو فرد توسط شمارش کمترین تعداد جست‌های یکی به دیگری همسایه به همسایه اندازه‌گیری می‌شود. برای مثال افرادی که همسایه‌ی ما نیستند اما با همسایه‌ی ما همسایه‌اند، با ما دو (جست) فاصله دارند. همان طور که گفتیم کوتاه‌ترین مسیر میان دو نفر (فاصله مستقیم) یا (فاصله‌ی ژنودیسک) نام دارد. مرکزیت میانگی بر این فرض استوار است که فرد می‌تواند با قرار گرفتن بر یک تنگ راه ارتباطی (مسیر ژئودسیک) قدرت کسب کند .در شکل زیر گره D در جایگاه قدرت دارد؛ تمام ارتباطات میان سایر گره‌ها باید از D بگذرند.



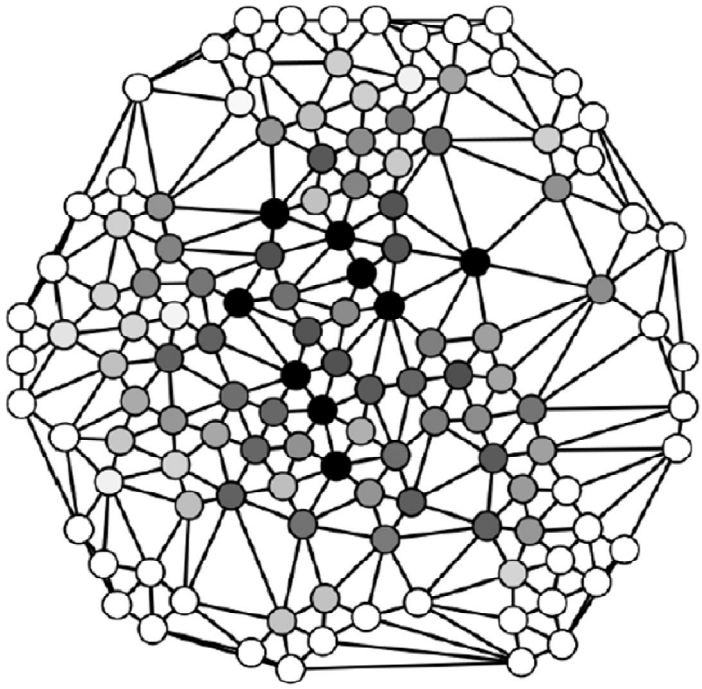
شکل(۲): D در جایگاه قدرت قرار دارد

[1] Betweenness Centrality
[2] Communication Bottlenecks
[3] Community Bridges

جایگاه تنگ راهی می‌تواند پرمخاطره باشد، چرا که فشار قابل‌توجهی به همراه دارد. بهترین معیار میانه‌ای شناخته‌شده برای مرکزیت، معیار میانگی معرفی شده توسط فریمن است. وی مرکزیت میانگینی رأسی چون v تعداد دفعاتی که دیگر رأس‌های گراف برای رسیدن به یکدیگر به عبور از رأس چون v نیاز دارند بیان می‌کند.

رابطه (۸)
$$C_k^{DEG} = \sum_i \sum_j \frac{g_{ikj}}{g_{ij}}$$

در واقع این معیار سهم رأس k را از ترافیک تمام کوتاه‌ترین مسیرهای موجود از i به j به ازای تمام iها و jها بیان می‌دارد. درصورتی‌که تنها یک کوتاه‌ترین مسیر از هر رأس به رأس‌های دیگر موجود باشد، معیار فوق برابر تعداد کوتاه‌ترین مسیرهایی است که از رأس داده‌شده‌ی k عبور می‌کند. در شکل زیر، رأس‌های بارنگ تیره‌تر معرف مرکزیت میانگی بیشتر هستند.



شکل (۳): نمایش مفهوم مرکزیت واسطه‌ای در گراف

شاخص مرکزیت میانگی، موقعیت یک موجودیت را درون یک شبکه برحسب توانایی‌اش برای ایجاد ارتباط با سایر زوج‌ها یا گروه‌ها در شبکه شناسایی می‌کند. موجودیتی با بالاترین مرکزیت میانگی به‌طورکلی در شبکه دارای ویژگی‌های زیر است:

- موقعیت مطلوب و مستحکمی در شبکه به دست آورده است؛
- نقطه مجزایی از گسیختی را به نمایش می‌گذارد؛
- تأثیر خیلی زیادی بر آنچه در شبکه اتفاق میافتد، دارد.

به‌طورکلی مرکزیت میانگی، نقطه‌ای است که بینابین بسیاری از جفت نقاط دیگر باشد؛ درواقع نقاطی هستند که راهه‌ای ارتباطی نقاط دیگر از آن‌ها می‌گذرد. این نقاط دارای قدرت ایزوله کردن یا افزایش ارتباطات هستند. افرادی که به‌عنوان واسطه برای جریان اطلاعات عمل می‌کنند ارزش میانگی بالایی خواهند داشت.

انواع مختلفی از معیارهای میانگی موجود است، برخی در نظر گرفتن تنها کوتاه‌ترین مسیرها را برای این معیار کافی نمی‌دانند. باآنکه در حالت‌های خاصی چون انتقالات بین‌شهری و تجارت، این کوتاه‌ترین مسیرهای ممکن هستند که با قطعیت مورد استفاده قرار می‌گیرند اما این امر در همه‌ی محیط‌ها برقرار نیست، جریان یافتن اطلاعات و یا انتشار تأثیرات در شبکه، از این قسم‌اند. اطلاعات برای انتشار یافتن در شبکه تمام مسیرهای ممکن را با احتمالی برابر در نظر می‌گیرد. بدین ترتیب منطقی به نظر می‌رسد که به جای استفاده از g_{ikj} در رابطه (۸) تمام مسیرهای موجود از i به j را که از k عبور می‌کنند در نظر بگیریم .در چندین معیار میانگی را با در نظر گرفتن تمام مسیرهای موجود، دنباله‌ها و همچنین گشته‌ها (به‌صورت وزن‌دار و یا وزن برابر معکوس طول) بررسی کرده است .دیدگاه دیگری نیز وجود دارد که در آن، با این توجیه که مسیرهای خیلی طولانی شانس کمی در انتشار اطلاعات در شبکه دارند محدودیت‌هایی را بر روی طول مسیرهای محاسبه شده اعمال می‌کند. این کار سبب می‌شود مسیرهای خیلی طولانی در محاسبات مرکزیت میانگی دخیل نشوند. این قسم از معیارها، مرکزیت میانگی kتایی نامیده می‌شوند. k نشان‌دهنده‌ی ماکسیمم طول مسیرهای محاسبه شده است. درحالت خاصی از معیار میانگی را با $k = 2$ بر ر کرده است. همچنین حالت خاصی از معیار واسطه‌ای را مطرح نموده‌اند که آن را معیار دلالی‌ای می‌نامند. نوع بسیار

هوشمندانه‌ای از معیار ۲ واسطه‌ای می‌تواند حالتی باشد که تمام مسیرها با طول‌های متفاوت در محاسبات، با وزنی مرتبط با معکوس طولشان محاسبه شوند.

محاسبه‌ی تمام مسیرها یا گشت‌های موجود، موجب شمارش چندباره‌ی مسیرها می‌شود چرا که همواره مسیرهای مختلف در تعدادی یال مختلف مشترک‌اند. برای حل این مسئله پیشنهاد در نظر گرفتن مسیرهای یال مجزا را داده است .معیار وی مرکزیت واسطه‌ای جریان نامیده می‌شود، دلیل این نام‌گذاری ارتباط بسیار مشخص موجود بین مسیرهای یال مجزا و حداکثر مقدار جریان موجود بین دو رأس است .ا ز آنجا که معیار واسطه‌ای جریان تنها مسیرهای یال مجزا را در نظر می‌گیرد، می‌توان آن را معیار اندازه‌گیری میزان جریان که در صورت غیبت یک رأس انتقال پیدا نمی‌کند، دانست.

مرکزیت میانگی، نقش مهم دیگری نیز دارد؛ به ما امکان می‌دهد تا اتصال گران کران‌های را تشخیص دهیم؛ یعنی کسانی که به‌مثابه‌ی پل‌هایی میان دو یا چند اجتماع که راه دیگری برای ارتباط با هم ندارند (حفره‌ی ساختاری میانشان هست)، عمل می‌کنند. برای نشان داد افرادی که میان حفره‌های ساختاری پل می‌زنند، سریع‌تر از دیگران ترقی می‌کنند.

مرکزیت میانگی عددی بین صفر و یک است .در حالت صفر، با برداشتن گره هیچ اتفاق خاصی در شبکه نمی‌افتد و همه‌ی گره‌ها به هم متصل باقی می‌مانند و حتی فواصل ژئودسیک میانشان از میان نمی‌رود، اما در حالت یک، اگر گره پل را بردارید، رابطه‌ی اطرافیان او به‌کلی باهم قطع می‌شود.

مرکزیت بردار ویژه[1]

مرکزیت بردار ویژه یک معیار اندازه‌گیری است که نفوذ یک گره در یک شبکه را تعیین می‌کند. اختصاص رتبه نسبی به تمام گره‌ها در شبکه مبتنی بر این مفهوم است که اتصال به گره‌های با رتبه بالاتر کمک بیشتری به رتبه گره نسبت به اتصال به گره‌های با رتبه کمتر می‌کند.

استفاده از ماتریس مجاورت برای یافتن مرکزیت بردار ویژه برای یک گراف $G = (V, E)$ با $|V|$ گره، یک ماتریس مجاورت درنظرمی‌گیریم. درایه $A_{ij} = 1$ است، اگر گره i به گره j پیوند خورده

[1] Eigenvector Centrality

باشد و درغیر این صورت $A_{ij} = 0$ است. امتیاز مرکزیت گره iام که با x_i نشان داده شده است، با رابطه (۹) قابل محاسبه است.

رابطه (۹) $$x_i = \frac{1}{\lambda}\sum_{j\in M(i)} x_j = \frac{1}{\lambda}\sum_j^N A_{i,j}x_j$$

که $M(i)$ مجموعه‌ای از همسایه‌های i و λ یک ثابت است که مقدار ویژه نام دارد.

مرکزیت مابینیت[1]

مابینیت، یک معیار اندازه‌گیری گره‌های درون یک شبکه اجتماعی می‌باشد. این معیار یک معیار اندازه‌گیری برای کمی کردن کنترل یک بازیگر در ارتباطات با انسان‌های دیگر در یک شبکه‌اجتماعی می‌باشد. به‌طور مفهومی، گره‌هایی که با احتمال بالا به‌صورت تصادفی در کوتاه‌ترین مسیر بین دو گره‌های که به‌صورت تصادفی انتخاب شده‌اند، مابینیت بالایی دارند، بیشتر به‌عنوان واسط بین اعضای شبکه اجتماعی مطرح هستند.

مرکزیت گره v در یک گراف $G = (V, E)$ با V گره به‌صورت زیر محاسبه می‌گردد.

- برای هر جفت از گره‌ها (s,t)، کوتاه‌ترین مسیر بین آن‌ها را محاسبه می‌کنیم.
- برای هر جفت از گره‌های (s,t)، کسری از کوتاه‌ترین مسیرهای که از میان گره‌ها (در اینجا v) عبور می‌کنند را تعیین می‌کنیم.
- این کسرها را بین هر جفت رئوس (s,t)، جمع می‌کنیم. از رابطه‌ی زیر مرکزیت مابینیت به‌دست می‌آید:

رابطه (۱۰) $$C_B(V) = \sum_{S\neq v\neq t\in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

در معادله بالا پارامترها به‌صورت زیر تعریف می‌شوند:

σ_{st}تعداد کوتاه‌ترین مسیرها از گره s تا گره t و $\sigma_{st}(v)$ تعداد مسیرهای که از v عبور می‌کنند.

[1] Betweenness Centrality

مرکزیت نزدیکی[1]

در درون حوزه نظریه گراف و تحلیل شبکه‌ها، انواع مختلفی از معیارهای اندازه‌گیری مرکزیت وجود دارد که اهمیت نسبی یک رأس درون شبکه اجتماعی را تعیین می‌کند (به‌عنوان‌مثال درون شبکه اجتماعی چقدر نفوذ دارد). بسیاری از مفاهیم مرکزیت، اولین بار در تحلیل شبکه‌های اجتماعی توسعه یافتند و بسیاری از اصطلاحات مورداستفاده برای اندازه‌گیری مرکزیت در شبکه‌های اجتماعی مجازی منشأ جامعه‌شناختی دارند.

در گراف‌ها، یک معیار اندازه‌گیری و فاصله واقعی بین تمام جهت گره‌ها، به‌وسیله‌ی طول کوتاه‌ترین مسیر بین گره‌ها تعریف می‌شود. دور بودن یک گره به‌صورت جمع فاصله‌های آن با تمام گره‌های دیگر، تعریف می‌شود و نزدیکی به‌صورت معکوس دور بودن یک گره تعریف می‌گردد؛ بنابراین یک گره مرکزی‌تر، فاصله‌های آن با دیگر گره‌ها پایین‌تر است. نزدیکی می‌تواند به‌عنوان یک معیار اندازه‌گیری برای منتشر کردن اطلاعات از گره s به تمام دیگر گره‌ها به‌کار رود. در تعریف کلاسیک مرکزیت نزدیکی، منتشر کردن اطلاعات با استفاده از کوتاه‌ترین مسیرها، مدل شده‌است.

این مدل ممکن است برای همه نوع سناریوهای ارتباطی واقع‌بینانه نباشد. باید به این نکته توجه کرد که با استفاده از تئوری گراف، مرکزیت نزدیکی تمام گره‌ها در یک گراف غیر متصل خواهد بود. مرکزیت نزدیکی از رابطه‌ی زیر قابل‌محاسبه است.

رابطه (۱۱)
$$c_v = \frac{n-1}{\sum_{u \in V} d(v,u)}$$

که در آن n تعداد افراد موجود در شبکه (تعداد گره‌ها) و $d(v,u)$ طول کوتاه‌ترین مسیر از v به u است.

جریان داده

مرکزیت، یکی از موضوعات مطرح در شبکه‌های پیچیده است که کار زیادی روی آن انجام شده است؛ اما نکته‌ای که معمولاً در نظر گرفته نمی‌شود جریان‌های داده روی شبکه‌های مختلف است. برای نمونه، همان‌طور که گفته شد بعضی از معیارهای مرکزیت تنها کوتاه‌ترین مسیرهای بین گره‌ها

[1] Closeness Centrality

را در نظرمیگیرند. با اعمال چنین معیاری روی شبکه، به صورت ضمنی فرض می‌شود که همه‌ی ترافیک شبکه از طریق کوتاه‌ترین مسیرها منتقل می‌شود. اکثر معیارهای مرکزیت موجود برای جریان‌های معمول که موردعلاقه‌ی محققان است، جوابگو نیستند و استفاده از آن‌ها منجر به انتخاب گره‌هایی اشتباه به‌عنوان گره‌های با مرکزیت بالا می‌شود. در این قسمت به بررسی عوامل مؤثر روی نحوه‌ی جریان داده در یک شبکه می‌پردازیم.

انواع جریان داده

داده‌ها یا اشیاء در شبکه‌ها ممکن است از راه‌های مختلفی منتقل شوند.

- **کوتاه‌ترین مسیر**[1]: در بعضی از شبکه‌ها اشیاء و داده‌ها تنها از طریق کوتاه‌ترین مسیر بین دو گره بین آن دو منتقل می‌شوند. مثلاً در مورد بسته‌ی پستی بین دو گره در شبکه، بسته از کوتاه‌ترین مسیر بین گره‌ها عبور می‌کند.
- **مسیر**[2]: در بعضی از شبکه‌ها اشیاء و داده‌ها از طریق همه‌ی مسیرهای بین دو گره منتقل می‌شوند. مثلاً در مورد نحوه‌ی انتشار یک ویروس در شبکه‌های اجتماعی، یک فرد بعد از اولین ابتلا در برابر ویروس مصون می‌شود و هرگز دوباره مبتلابه آن ویروس نمی‌شود.
- **دنباله**[3]: توالی تعدادی یال همسایه‌ی تصادفی است که در آن هیچ یالی دو بار تکرار نشده است .نمونه‌ی جریان‌هایی که ازاین‌گونه انتقال داده استفاده می‌کنند، جریان شایعه‌پراکنی در شبکه‌های اجتماعی است. ممکن است یک فرد یک شایعه را به چند نفر بگوید و بعد از مدتی خودش آن‌ها بشنود؛ اما یک شایعه را دو بار برای شخص دیگری تعریف نخواهد کرد.
- **گشت**[4]: توالی تعدادی یال همسایه‌ی تصادفی است. ممکن است یک یال چند بار نیز تکرار شود. نمونه‌ی شبکه‌هایی که از این مدل برای انتقال داده‌ها استفاده می‌کنند، شبکه‌ای است که توسط ردوبدل کردن پول تشکیل می‌شود. ممکن است یک پول دوباره به یک فرد برگردد و این فرد آن پول را دوباره به همان فرد قبلی بدهد.

[1] Geodesics
[2] Path
[3] Trail
[4] Walk

انواع انتشار داده

انواع مختلف روش‌های انتشار داده به شرح زیر است:

- **نسخه‌برداری موازی** :شی در هر گره شبکه به تعداد همسایه‌ی آن کپی می‌شود و به آن‌ها فرستاده می‌شود؛ مانند نامه‌های الکترونیکی که به لیست همه‌ی دوستان فرستاده می‌شود
- **نسخه‌برداری ردیفی** :شی در هر گره کپی می‌شود و به یکی از همسایه‌های گره داده می‌شود؛ مانند سرایت ویروس.
- **انتقال** :شی کپی نمی‌شود بلکه به یکی از همسایه‌های گره داده می‌شود؛ مانند شبکه‌ای که از خریدوفروش یک جنس دست‌دوم تشکیل می‌شود.

مرکزیت گره‌ها در جریان‌های متفاوت

شبکه به‌خودی‌خود مفهومی ندارد و با جریان یافتن داده در آن مفهوم پیدا می‌کند .با توجه به ماهیت جریان و نحوه‌ی انتشار آن ممکن است معیارهای مرکزیت بیان‌شده، قابل‌اعمال کردن روی آن شبکه شوند.

جدول (۳): انواع مرکزیت‌های داده‌ی قابل‌استفاده در انواع مختلف راه و جریان داده

نوع جریان /نوع انتشار	نسخه‌برداری موازی	نسخه‌برداری ردیفی	انتقال
کوتاه‌ترین مسیر	–	نزدیکی	نزدیکی، میانگی
مسیر	نزدیکی، درجه	–	–
دنباله	نزدیکی، درجه	–	–
گشت	نزدیکی، درجه	–	–

همان‌طور که دیده می‌شود، معیارهای مرکزیت تنها در موارد خاص قابل‌استفاده‌اند. مثلاً معیار میانگی تنها در صورت انتقال داده شی از کوتاه‌ترین مسیر قابل‌استفاده است.

چالش‌های موجود در شبکه‌های اجتماعی

در ادامه چالش‌های موجود در شبکه‌های اجتماعی را موردبررسی قرار خواهیم داد. این چالش‌ها که عموماً ناشی از گسترده بودن شبکه‌های اجتماعی است، امروزه به موضوع اصلی تحقیقات دانشمندان و ارائه راه‌حل‌هایی برای این مشکلات شده است.

گسترده بودن شبکه‌های اجتماعی و نحوه ذخیره‌سازی ارتباطات

با توجه به اینکه شبکه‌های اجتماعی از گره‌ها که نماینده افراد و یال‌ها که نماینده ارتباط بین افراد است، تشکیل شده است، با بزرگ شدن شبکه، نحوه ذخیره‌سازی این ارتباطات سخت خواهد شد .با دو ساختار داده می‌توان ماتریس را در حافظه کامپیوتری ذخیره کرد:

- ساختار داده آرایه دوبعدی (ماتریس)
- ساختار داده لیست پیوندی

اگر شبکه اجتماعی نظیر فیسبوک را در نظر بگیریم، طبق آخرین آمار ارائه شده در مارس ۲۰۱۳ تعداد کاربران از مرز یک میلیارد نفر در سراسر کره زمین گذشت. با توجه به حجم عظیمی از تعداد کاربران، اگر بخواهیم به‌صورت ماتریس روابط بین افراد را مشخص کنیم به ماتریسی با ابعاد ۱۰۰۰۰۰۰۰۰*۱۰۰۰۰۰۰۰۰ نیاز داریم .علاوه بر حجم بالای موردنیاز برای ذخیره‌سازی، زمان موردنیاز برای جستجو و بازیابی برای ماتریسی با چنین ابعادی بسیار زمان‌بر خواهد بود. به‌طور واضح برای نمایش گراف شبکه‌های اجتماعی از ماتریس نباید استفاده کرد. ولی ساختار داده لیست پیوندی می‌تواند گزینه بهتری برای ذخیره‌سازی باشد، زیرا جستجو در لیست پیوندی به زمان خطی احتیاج دارد، ولی جستجو در ماتریس در زمان تابع درجه دوم صورت می‌پذیرد.

طبق نظریه جامعه‌شناسان هر فرد در کره زمین به‌طور میانگین ۷۵۰ نفر را می‌شناسد، بنابراین می‌توانیم نتیجه بگیریم که به ازای هر سطر که دارای ۱ میلیارد درایه است، فقط ۷۵۰ درایه آن دارای مقدار یک که نشان‌دهنده برقراری ارتباط بین افراد است، دارای مقدار است .این نشانه واضحی است از اینکه ما در اینجا با یک ماتریس اسپارس مواجه هستیم؛ بنابراین به‌صورت قابل‌توجهی می‌توانیم حجم ذخیره‌سازی را پایین بیاوریم.

یافتن گروه‌ها در گراف شبکه‌های اجتماعی

مسئله یافتن گروه‌ها در گراف شبکه‌های اجتماعی یکی از چالش‌های اساسی است .همان‌طور که گفتیم دسته را می‌توان به‌عنوان تعدادی گره (افراد) در نظر گرفت که به‌صورت کامل به هم متصل دسته داریم هستند (گراف کامل) اگر k فرد به‌صورت کامل به هم متصل باشند، گوییم یک kدسته داریم. در واقع یک k-دسته عبارت است ازk گره که باهم به صورت کامل وصل هستند و همان‌طور که گفته شد یافتن تمام این زیر دسته‌ها در یک گراف مسئله‌ایNP-HARD است، زیرا برای هر گره باید اتصال با تمام گره‌های دیگر بررسی شود. این عمل باید برای تمام گره‌ها انجام گیرد.

امنیت با استفاده از تجزیه‌وتحلیل شبکه‌های اجتماعی

ثابت شده است که اطلاعات استخراج‌شده از شبکه‌های اجتماعی ابزاری مفید در جهت برقراری امنیت است .یک مثال کاربردی مربوط به امنیت تجزیه‌وتحلیل تروریسم است. اگر در یک گراف شبکه‌های اجتماعی، فردی به‌عنوان مرکزیت، عامل پیوند چندین گروه تروریستی به هم باشد، جهت متلاشی کردن کارکرد و برنامه‌های این گروه تروریستی به‌جای مقابله با تمام افراد گروه، می‌توان با عوامل پیونددهنده مقابله کرد و کارکرد کل گروه را مختل کنیم. این مطالعات به وسیله جمع‌آوری اطلاعات روی شبکه‌های گسترده وب و تجزیه‌وتحلیل آن با استفاده از شبکه‌های اجتماعی است .یک چالش تحقیقاتی عمده روی تحلیل شبکه‌های اجتماعی نظارت اینترنتی برای فعالیت‌های غیرقانونی برای حفاظت از زیرساخت‌های حیاتی، است.

جامعه، موبایل، اشتراک‌گذاری محتوای فراگیر و توزیع رسانه زنده

اشتراک‌گذاری محتوا و نیازهای توزیعی به‌صورت فزاینده‌ای ادامه خواهد یافت .تلفن همراه، دوربین‌های دیجیتال و دیگر دستگاه‌های فراگیر، مقدار بسیار عظیمی از داده‌ها تولید می‌کنند که کاربران می‌خواهند در صورت امکان به‌صورت بلادرنگ توزیع شوند.

هرزنامه‌ها، نظرات و تعامل خصمانه در رسانه‌های اجتماعی

تشخیص هرزنامه و تشخیص آگهی‌ها، چالش‌های تحقیقاتی هستند که نیاز به توجه بیشتری از جانب جامعه پژوهشی دارند. ازآنجاکه کاربران و داده‌ها افزایش می‌یابند، هرزنامه‌ها و تبلیغات در حال

رشد هستند .علاوه بر این اهمیت شبکه‌های اجتماعی برای تحت تأثیر قرار دادن نظرات کاربران باید با مکانیزهای مناسب برای جلوگیری از نظرات مغرضانه و جعلی، با توجه به ارتباط برای کسب‌وکار، محافظت شود.

شخصی‌سازی برای تعاملات اجتماعی

به‌منظور بهبود تعاملات اجتماعی و افزایش ظرفیت‌های اجتماعی، موتورهای جستجو باید در جستجو حتماً مواردی که کاربر آن‌ها را دوست دارد و مواردی را که دوست ندارد را اولویت‌بندی کند. موتور جستجویی کاراتر خواهد بود که قادر به ارائه خدمت، فقط به کاربرانی با محتوای مربوط داشته باشد. در این زمان الگوریتم‌های شخصی‌سازی برای موتورهای جستجو موردمطالعه قرارگرفته‌اند

استفاده از شبکه‌های اجتماعی برای کسب‌وکار و بازاریابی

شبکه‌های اجتماعی الگوهای همکاری جدید بین کاربران شبکه را مصرف می‌کنند و مطالعات جدی در استفاده از تعدادی پلت فرم برای اهداف کسب‌وکار داخلی انجام گرفته است. بااین‌حال یکی از برجسته‌ترین چالش‌های تحقیقاتی، چگونگی استفاده از شبکه‌های اجتماعی برای ارتباطات خارجی، پشتیبانی از مشتری و دوره‌های بازاریابی هدفمند است.

مسائل اجتماعی و اخلاقی در یک جهان شبکه شده

در هر جامعه کوچک یا بزرگ، جوامع اجتماعی آنلاین با موضوعات اخلاقی و اجتماعی حیاتی روبرو هستند که نیاز به مراقبت‌های ویژه و رسیدگی حساس دارند .به اشتراک‌گذاری اطلاعات شخصی، حفاظت از بهره‌برداری از کودکان و مسائل یادگیری باید مورد مطالعه قرار گیرند و پاسخ آن‌ها به‌صورت مناسب داده شود.

جستجوی بلاگ‌ها، توییت‌ها و سایر رسانه‌های اجتماعی

جستجو در بلاگ‌ها، توییت‌ها و سایر رسانه‌های اجتماعی هنوز یک موضوع باز است. کاربران مختلف وقتی‌که به‌عنوان استفاده‌کننده از شبکه‌های اجتماعی می‌آیند، نیازهای مختلفی دارند. جستجوهای بلادرنگ برای تعادل بین کیفیت، قدرت، ارتباط و به هنگام بودن از محتوا داریم.

جوامع در شبکه‌های اجتماعی

در اغلب شبکه‌های اجتماعی مجموعه‌ای از بازیگران وجود دارند که ارتباطات قوی‌تری با یکدیگر داشته و موضوعات موردعلاقه آن‌ها نیز مشابه است که در ادبیات شبکه‌های اجتماعی به این مجموعه بازیگران مجمع یا انجمن گفته می‌شود. بنا بر تعریف ارائه شده در (یک انجمن مجموعه‌ای از گره‌ها است که گره‌های موجود در آن به سایر گره‌های موجود در انجمن نزدیک‌تر هستند تا سایر گره‌های خارج از آن انجمن) .معیار نزدیکی دو گره می‌تواند کوتاه‌ترین مسیر بین دو بازیگر یا مجموع تمامی مسیرهای مستقل بین دو بازیگر و توابع فاصله دیگری ازاین‌دست باشد. انجمن‌ها اطلاعات ارزشمندی

در مورد نوع ارتباط بازیگران، نحوه انتقال اطلاعات بین آن‌ها و نحوه توزیع بازیگران در شبکه اجتماعی ارائه می‌کنند و درواقع به‌عنوان جزء اصلی سازنده این شبکه‌ها محسوب می‌شوند.

بر اساس تعریف دیگری که در ارائه گردیده است (یک انجمن مجموعه‌ای از گره‌ها بوده که ارتباطات میان گره‌های آن انجمن قوی‌تر و چگال تر از ارتباطات میان گره‌های انجمن با سایر گره‌ها در شبکه اجتماعی است) .برای اینکه بتوان بر اساس تعاریف ارائه شده جوامع را استخراج کرد باید بتوان کیفیت جوامع استخراج شده را بر اساس این تعاریف تعریف نمود که دو نمونه از معیارهای کیفیت جوامع که در ارائه گردیده است عبارت‌اند از:

- **ضریب خوشه‌بندی** :همان‌طور که توضیح داده شد این خصوصیت بیان می‌کند که دو گره در شبکه اجتماعی که همسایه یک گره دیگر هستند، نسبت به دو گره که به‌طور تصادفی از کل گره‌های موجود در شبکه اجتماعی انتخاب می‌شوند با احتمال بیشتری با یکدیگر در ارتباط هستند .در قسمت خصوصیات شبکه‌های اجتماعی فرمول ضریب خوشه‌بندی را بیان نمودیم.
- **قطر انجمن** :یکی از معیارهای مورداستفاده برای اندازه‌گیری گراف قطر گراف است که از آن به‌عنوان معیار اندازه‌گیری کیفیت یک انجمن نیز استفاده می‌شود. قطر یک انجمن حداکثر طول کوتاه‌ترین مسیر بین هر دو گره دلخواه در این انجمن است و هر چه قطر یک انجمن کمتر باشد نشان از چگالی بیشتر ارتباطات و قوت آن‌ها بین بازیگران موجود در انجمن است. در فصل بعد به تشریح روش‌های استخراج جوامع خواهیم پرداخت.

مدل‌های شبکه

در ابتدای مطالعه‌ی شبکه‌های اجتماعی، فرض بر این بود که این گراف‌ها کاملاً تصادفی هستند و ساختار ویژه‌ای ندارد؛ اما درنهایت، محققین با انجام آزمایش‌های بسیاری، ازجمله بررسی توزیع درجات، اندازه‌گیری ضریب خوشه‌بندی، دریافتند که ساختار این شبکه‌ها با گراف‌های تصادفی تفاوت زیادی دارد. مطالعه بر روی شبکه‌های مختلف و مشاهده‌ی برخی ساختارها و رفتارهای مشترک در آن‌ها، موجب شد تا دانشمندان مکانیزها و مدل‌هایی برای توجیه نحوه‌ی رفتار و بروز برخی

ساختارهای خاص در این شبکه‌ها ارائه دهند .مطالعه‌ی مدل‌ها از چندین جهت می‌توان سودمند باشد:

- شبکه‌های واقعی می‌توانند بسیار بزرگ و پیچیده باشند و مطالعه‌ی یک موضوع خاص روی آن‌ها دشوار باشد .درصورتی‌که مدل‌ها کمک می‌کنند که شبکه‌ای با همان ویژگی‌ها ولی در ابعاد کوچک‌تر در اختیار داشته باشیم و بتوانیم به‌طور تقریبی درباره‌ی شبکه‌ی اصلی بحث کنیم.

- به دست آوردن اطلاعات برخی از شبکه‌ها در بسیاری از موارد یا امکان‌پذیر نیست و یا باید هزینه‌ی گزافی صرف شود. برخی از این شبکه‌ها، شبکه‌های مجرمانه و مخفی هستند که هم برخی از اعضای (گره‌ها) آن مشخص نیستند و هم در بسیاری از موارد روابط (یال‌ها) میان اعضا قابل‌مشاهده نیست .در این‌گونه از موارد استفاده از مدل‌ها می‌تواند بسیار کارگشا باشد.
- پیدا کردن مدل و مکانیزمی که گرافی با خواص نزدیک به شبکه‌های اجتماعی تولید کند، به ما کمک می‌کند که رفتار این شبکه‌ها را بهتر درک کنیم.

در ادامه برخی از مدل‌های معروفی را که از ابتدا برای شبکه‌های اجتماعی ارائه‌شده، معرفی می‌کنیم و برخی ویژگی‌های آن‌ها را مثال می‌زنیم.

گراف‌های تصادفیErdos_Renyi

گراف‌های تصادفی اولین بار توسط دو ریاضیدان مجارستانی به نام اردوش و رینی موردبررسی قرار گرفتند .این دو، برخی ویژگی‌های گراف تصادفی را به دست آوردند .در زیر، ابتدا مدل گراف تصادفی اردوش-رینی را تعریف می‌کنیم و سپس به ذکر برخی ویژگی‌های آن می‌پردازیم.

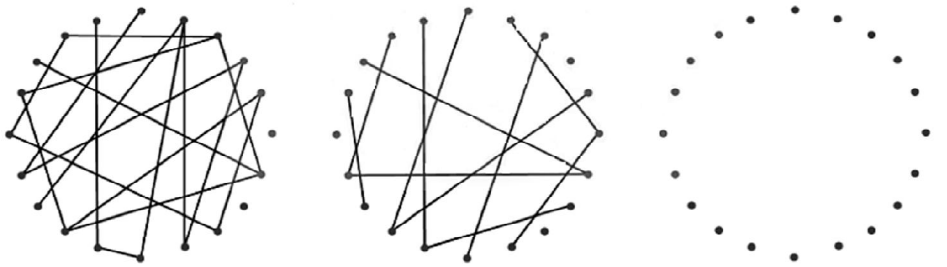
دو نسخه از این گراف‌های تصادفی وجود دارد.

تعریف گراف تصادفی ER(n,m)

در این مدل، به‌عنوان ورودی، دو پارامتر n ، تعداد رئوس وm ، تعداد یال‌ها داده و از بین تمام گراف‌هایی که n رأس و m یال دارند با احتمال مساوی و یکنواخت، یکی به‌عنوان خروجی انتخاب می‌شود. m می‌تواند در بازه$\left[\mathbf{0},\binom{\mathbf{n}}{\mathbf{2}}\right]$مقدار بگیرد.

تعریف گراف تصادفی ER(n,p)

در این نسخه از گراف‌های تصادفی، دو پارامتر n، تعداد رئوس و P ، احتمال حضور هر یال در گراف، به‌عنوان ورودی داده می‌شودو خروجی گرافی است که به ازای هر یک از $\binom{n}{2}$ یال ممکن گراف، هرکدام با احتمال p در گراف حضور و با احتمال 1-p حضور ندارد.



آ) p=0 ب) p=0.1 ج) p=0.2

شکل(۴): نمونه‌ای از گراف‌های تصادفی اردوش-رینی

به‌سادگی می‌توان دریافت که دو نسخه‌ی تولید گراف تصادفی اردوش -رینی، بسیار به هم شبیه هستند و با تبدیل $p=\frac{m}{\binom{n}{2}}$ می‌توان دو مدل را به یکدیگر تبدیل کرد .ازاین‌رو، تنها به ذکر ویژگی‌های یکی از آن‌ها (مدل دوم) بسنده می‌کنیم.

برخی ویژگی‌ها

ثابت شده است که فاصله‌ی بین هر دو گره در گراف تصادفی، متناسب با لگاریتم تعداد گره‌های گراف است، که این همان ویژگی جهان کوچک را بیان می‌دارد که شبکه‌های اجتماعی هم دارا هستند.

به‌راحتی می‌توان اثبات کرد که متوسط درجه‌ی هر رأس در این گراف $(n-1)p$ است و توزیع درجات، از توزیع نرمال پیروی می‌کند؛ بنابراین درجه‌ی رئوس مختلف بسیار همگون است، در حالی که شبکه‌های اجتماعی، این‌گونه نیست و درجه‌ی رئوس ناهمگون هستند و از توزیع قانون توان پیروی می‌کند .از طرفی می‌توان مشاهده کرد که ضریب خوشه‌بندی گراف‌های اردوش -رینی، بسیار نزدیک بهp است، درحالی‌که ضریب خوشه‌بندی در یک شبکه‌ی اجتماعی با همان تعداد رأس و یال، بسیار بیشتر است .گراف تصادفی اردوش -رینی، به خاطر تفاوتی که در توزیع درجات و ضریب خوشه‌بندی با شبکه‌های اجتماعی داشت، مدل مناسبی شناخته نشد.

گراف‌های تصادفی Watts-Strogatz

پس از مدل گراف تصادفی اردوش-رینی، گراف تصادفی واتس-استروگتس دومین مدل شناخته‌شده برای شبکه‌های اجتماعی است. این مدل که ترکیبی از یک گراف تصادفی و یک لاتیس منظم حلقوی است، توسط در مجله‌ی نیچر معرفی شد .این مدل به خاطر، ضعفی که در کمیت ضریب خوشه‌بندی در مدل گراف تصادفی اردوش – رینی بود، معرفی شد .ابتدا مدل را تعریف می‌کنیم و سپس برخی ویژگی‌هایش را بررسی می‌کنیم.

تعریف گراف تصادفی WS(n,2k, ß)

این مدل به‌عنوان ورودی، سه پارامتر n، تعداد رئوس،$2k$ ، تعداد همسایه‌های هر رأس و ß پارامتر تصادفی کردن $(0 \leq ß \leq 1)$ را می‌گیرد. در ابتدا هر رأس بهk نزدیک‌ترین همسایه‌ی چپ وk نزدیک‌ترین همسایه راست متصل است (رئوس گراف را روی یک دایره در نظر بگیرید) .در این صورت یک لاتیس منظم حلقوی به وجود می‌آید. رئوس را با همان ترتیبی که روی دایره قرار دارند $N_1,N_2,\ldots,N_n$در نظر بگیرید .هر یال n_i , n_j که $i<j$ را با احتمال ß جابه‌جا کنید و با یال ni ,nk تعویض کنید به‌نحوی‌که یال چندگانه یا طوقه درست نشود. رأس k باید با احتمال مساوی و یکنواخت از میان رئوس مجاز انتخاب شود.

همان‌طور که از تعریف گراف تصادفی واتس- استروگتس پیداست، هر چه قدر پارامتر بیشتر باشد، βباشد، گراف به حالت اولیه‌ی آن، یعنی لاتیس منظم نزدیک‌تر است و هر چه قدر پارامتر β بیشتر باشد گراف به یک گراف تصادفی نزدیک‌تر است.

شکل (۵): تأثیر پارامتر تصادفی در گراف واتس-استروگتس

ساختار گراف لاتیس منظم به‌گونه‌ای است که باعث می‌شود ضریب خوشه‌بندی گراف بالا باشد، اما قطر آن و نیز میانگین فاصله‌ی بین رئوس (برخلاف شبکه‌ی اجتماعی) زیاد است .از طرفی در گراف تصادفی ضریب خوشه‌بندی بسیار کم است و در عوض قطر و میانگین فاصله‌ی بین رئوس کم است. ترکیب این دو نوع گراف ایده‌ی اصلی گراف تصادفی واتس-استروگتس است. ویژگی‌های بینابینی از هردو مدل لاتیس منظم و گراف کاملاً تصادفی دارد.

این مدل خاصیت جهان کوچک را داراست .مشکلی که دارد توزیع درجاتش است که همانند گراف تصادفی اردوش -رینی، توزیع قانون توان نیست .همچنین مشکل دیگر این مدل، عدم رشد آن . است و تعداد گره هایش از ابتدا تعیین می‌شود و ثابت می ماند.

گراف های تصادفیBarabasi-Albert

مدل باراباسی -آلبرت، به دنبال جبران ضعف مدل‌های قبلی که عدم تطابق توزیع درجات آن مدل‌ها با شبکه‌های واقعی بود، مطرح شد .ازاین‌رو، مدل باراباسی-آلبرت، یکی از چندین مدلی است که برای تولید شبکه‌ای با ویژگی مقیاس-آزاد معرفی‌شده است .این مدل، دو مفهوم رشد شبکه و اتصال ترجیحی که در شبکه‌های واقعی رایج است را برآورده می‌کند.

رشد شبکه، به این معناست که تعداد گره‌های گراف لازم نیست از ابتدا مشخص و ثابت باشد، بلکه در طی زمان افزایش می‌یابد .اتصال ترجیحی، به این معناست که هر گره‌ای که تعداد اتصالات (یال‌های) بیشتری دارد، در آینده با احتمال بیشتری یال جدیدی به آن متصل می‌شود.

تعریف گراف تصادفی

در این مدل، از یک گراف اولیه با m_0 گره، شروع می‌کنیم. در نسخه‌های مختلف مدل باراباسی آلبرت وضعیت همبندی این m_0 گره متفاوت است. در بعضی از آن‌ها، این گراف اولیه یک دور است، در برخی دیگر یک گراف کامل است. هر گره جدید را که اضافه می‌کنیم، به m گره متمایز قبلی متصل می‌نماییم (یال چندگانه نداریم). توجه کنید که $m \leq m_0$ باید باشد .همسایه‌های گره جدید را با احتمالی متناسب با درجه‌شان انتخاب می‌کنیم.

رابطه (۱۲)
$$p(i) = \frac{k_i}{\sum_{j \in G} k_j}$$

که در آن درجه‌ی گره i است. به عبارتی هر گره‌ای که درجه‌اش بیشتر باشد احتمال انتخاب شدنش به‌عنوان همسایه بیشتر است. در مدل‌های تعمیم‌یافته، برای کاهش ناهمگونی شبکه، از فرمول زیر به‌جای فرمول بالا استفاده می‌کنند.

رابطه (۱۳)
$$p(i) = \frac{k_i + B}{\sum_{j \in G} k_j + B}$$

که در آن B عددی ثابت است. به این شکل احتمال انتخاب گره‌های با درجه‌ی پایین کمی زیاد می‌شود و درنهایت توزیع درجات شبکه‌ی به وجود آمده دارای انحراف معیار پایین‌تری خواهد بود. از پارامتر B برای تنظیم ناهمگونی این‌گونه شبکه‌ها استفاده می‌شود .این نحوه‌ی ساخت در بسیاری از شبکه‌های اجتماعی نیز دیده می‌شود .مثلاً در شبکه‌های اجتماعی بر خط، احتمال اینکه یک تازه‌وارد با فردی دوست شود که تعداد دوستانش زیاد است، بیشتر است و یا در شبکه‌ی حاصل از پیوندها بین صفحات وب، احتمال این‌که یک صفحه‌ی جدید پیوندی به صفحه‌ای محبوب ایجاد کند، بسیار زیاد است.

برخی ویژگی‌ها

مهم‌ترین ویژگی مدل باراباسی_آلبرت، مقیاس آزاد بودن آن است، یعنی توزیع درجاتش از قانون توان پیروی می‌کند و آن سه است:

$$p(k) \propto k^{-3}$$ رابطه (۱۴)

میانگین درجات گراف حاصل از این مدل با فرض $m=m_0$ برابر $2m$ میانگین فاصله‌ی بین هر دو رأس در این گراف، تقریباً به‌صورت لگاریتمی برحسب اندازه‌ی شبکه افزایش می‌یابد:

$$l \propto \frac{\ln N}{\ln \ln N}$$ رابطه (۱۵)

فرمول دقیقی برای ضریب خوشه‌بندی در گراف باراباسی-آلبرت، محاسبه نشده است، اما به‌صورت تجربی، ضریب خوشه‌بندی در این گراف از گراف اردوش-رینی بیشتر است.

فصل سوم

آشنایی با تئوری فازی

واژه (فازی) در فرهنگ لغت آکسفورد به‌صورت (مبهم، گنگ، نادقیق، مغشوش، درهم و نامشخص) تعریف‌شده است. تئوری فازی لطفی زاده استاد ایرانی‌الاصل دانشگاه برکلی در سال ۱۹۶۵ میلادی مطرح گردید. هرچند این نظریه در ابتدا با مخالفت‌هایی مواجه گشت ولی به‌مرورزمان ارزش آن مشخص گردید. به‌طوری‌که امروزه در تمامی زمینه‌ها کاربرد پیداکرده است. مدل کردن یک سیستم که شرایط اولیه و مرزی غیردقیق دارد و یا مسائلی که با عدم قطعیت همراه هستند، همواره سخت و مشکل بوده است .به همین علت زاده تئوری فازی را ارائه نمود. در واقع منطق فازی تکنولوژی جدیدی است که شیوه‌های مرسوم برای طراحی و مدل‌سازی یک سیستم را که نیازمند ریاضیات پیشرفته نسبتاً پیچیده است را با استفاده از مقادیر و شرایط زبانی و یا به عبارتی دانش فرد خبره و باهدف ساده، دقیق و کارآمدتر شدن طراحی تا اندازه زیادی تعدیل و تکمیل می‌نماید .علیرغم اینکه منطق فازی بر پایه ریاضیات پیشرفته و پیچیده قرار دارد اما یادگیری آن بسیار آسان است .اساساً اگرچه سیستم‌های فازی پدیده‌های غیرقطعی و نامشخص را توصیف می‌کند با این .حال خود تئوری فازی یک تئوری دقیق است.

تئوری احتمالات نیز یکی از مهم‌ترین تئوری‌های سنتی جهت توصیف پدیده‌های غیرقطعی و سنتی است. یکی از وجوه تمایز بین تئوری احتمالات و فازی این است که تئوری احتمالات، تئوری حوادث تصادفی است. این تئوری می‌تواند در مورد پرتاب یک سکه و پیش‌بینی نتیجه شیر یا خط بکارگرفته شود .در اینجا برداشت ما حول پیش‌بینی وقوع یک حادثه در آینده می‌چرخد. درحالی‌که تئوری فازی متمرکز بر رخدادهای تصادفی نیست و تمرکز این تئوری روی مفاهیم زبانی است. عدم قطعیت تئوری فازی ناشی از عدم شفافیت در معنی یک واژه‌ی زبانی و محاوره‌ای مانند واژه‌های بلند و گرم و غیره است. یکی از کاربردهای تئوری فازی نیز فرموله کردن دانش بشری است که در قالب واژگان زبانی و محاوره‌ای بیان می‌شود .تئوری فازی و احتمالات مکمل یکدیگرند و در بسیاری موارد قابل‌ترکیب هستند.

آشنایی با مجموعه‌های فازی

تئوری فازی بر مبنای نظریه‌ی مجموعه‌های فازی شکل‌گرفته است. در این بخش، بابیان یک مثال لزوم ایجاد مفهوم مجموعه‌های فازی نشان داده خواهد شد .مسئله‌ی جداسازی ماشین‌های داخلی و خارجی یک شهر مفروض است. بسیاری از ماشین‌ها وجود دارند که درصدی از قطعات آن‌ها داخلی و بقیه خارجی هستند .در این حالت این سؤال مطرح می‌شود که این ماشین‌ها در کدام دسته قرار می‌گیرند؟ اساساً مجموعه‌هایی وجود دارند که نمی‌توان آن‌ها را به فرم مجموعه‌های کلاسیک تعریف کرد. چون مرز مشخص و روشنی ندارند، در این حالت مجموعه‌های فازی مطرح می‌شوند.

قبل از بیان نحوه‌ی نمایش مجموعه‌های فازی، روش‌های نمایش مجموعه‌های کلاسیک توضیح داده می‌شود .یک مجموعه کلاسیک A یا به‌اختصار مجموعه A در فضای جهانیU را می‌توان با فهرست کردن اعضاء (روش فهرست[1]). یا با مشخص کردن ویژگی‌هایی که باید توسط اعضاء مجموعه ارضاء گردد (روش قاعده[2]) تعریف کرد .روش فهرست را می‌توان در مجموعه‌های متناهی بکار برد، بنابراین کاربرد محدودی دارد .روش قاعده کلی‌تر است .درروش قاعده یک مجموعه A را بدین‌صورت می‌توان تعریف نمود:

رابطه (۱۶) $A = \{x \in U | \text{ شروطی برآورده را می کند } x\}$

روش سومی نیز برای تعریف مجموعه A وجود دارد .روش تعلق که یک تابع تعلق دو مقدار را برای A معرفی می‌کند که با نشان داده می‌شود به‌نحوی‌که:

رابطه (۱۷) $$\mu_A(x) = \begin{cases} 1, x \in A & \text{اگر} \\ 0, x \notin A & \text{اگر} \end{cases}$$

بدین ترتیب با معلوم بودن تابع تعلق یک مجموعه برای تمام اعضای مجموعه‌ی مرجع، اعضای آن مجموعه نیز مشخص می‌شوند.

[1] List Method
[2] Rule Method

تعریف مجموعه فازی

یک مجموعه فازی A در فضای جهانی U به‌وسیله یک تابع $\mu_A(x)$ که مقادیری در بازه [0,1] اختیار می‌کند، مشخص می‌شود؛ بنابراین یک مجموعه فازی تعمیم یک مجموعه کلاسیک است که اجازه می‌دهد تابع تعلق هر مقداری را در بازه [0,1] اختیار کند .به‌عبارت‌دیگر یک مجموعه کلاسیک فقط می‌توانست دو مقدار 1,0 داشته باشد درحالی‌که تابع تعلق یک مجموعه فازی، یک تابع پیوسته در محدوده‌ی [0,1] است .درواقع می‌بینیم که هیچ‌چیز در مورد مجموعه فازی گنگ و مبهم نیست بلکه مجموعه فازی، مجموعه‌ای است با یک تابع تعلق پیوسته.

یک مجموعه فازی A را در U با یک مجموعه از زوج‌های مرتب x و مقدار آن نمایش داد، بدین ترتیب:

$$A = \{x, \mu_A(x) | x \epsilon U\}$$ رابطه (۱۸)

هنگامی‌که U پیوسته باشد به‌عنوان‌مثال را معمولاً بدین‌صورت مشخص می‌کنند:

$$A = \int_U \frac{\mu_A(x)}{x}$$ رابطه (۱۹)

که علامت $\int$ نشان‌دهنده انتگرال نیست، بلکه اجتماع تمامی نقاط و مقدار تابع تعلق متناظر را نشان می‌دهد. هنگامی‌که U گسسته باشد، A معمولاً بدین‌صورت نوشته می‌شود:

$$A = \sum_U \frac{\mu_A(x)}{x}$$ رابطه (۲۰)

که در اینجا نیز علامت Σ به معنای جمع ریاضی نیست بلکه اجتماع تمامی نقاط $x \epsilon U$ و تابع تعلق متناظر $\mu_A(x)$ را نشان می‌دهد.

گراف فازی

گراف در حقیقت نمونه رابطه‌ای ساده از تعاملات سیستم مدل شده است .یک گراف روشی مناسب به‌منظور ارائه اطلاعات بین اشیاء است .همان‌طور که گفته شد، این اشیاء خود به‌وسیله رأس‌ها و رابطه آن‌ها به‌وسیله یال‌های اتصال‌دهنده یا یال‌ها به نمایش درمی‌آیند .در هنگام بروز ابهام در توصیف اشیاء، رابطه‌ها یا هردوی آن‌ها، طبیعی است که نیازمند طراحی یک مدل گراف فازی هستیم. موارد استفاده از روابط فازی بسیار گسترده و پراهمیت است، به‌خصوص در زمینه تجزیه‌وتحلیل خوشه‌ای[1]، شبکه‌های عصبی، شبکه‌های کامپیوتری، شناسایی الگو، تصمیم‌گیری و سیستم‌های خبره[2] .در هر یک از این موضوعات ساختار ریاضی اساسی، گراف‌های فازی هستند.

پروفسور زاده گراف فازی را نماینده‌ای برای نمایش داده‌های مبهم و روابط آن‌ها تعریف می‌کند.

$$\tilde{G} = (V, \tilde{E})$$ رابطه (۲۱)

که طبق تعریف گراف V را مجموعه رأس‌ها و $\tilde{E}$ را مجموعه روابط فازی بین رأس‌های گراف تعریف می‌کنیم. ما می‌توانیم این فرض را داشته باشیم که V نیز یک مجموعه فازی باشد. در این مورد، می‌گوییم که گراف ارائه‌کننده روابط فازی از گره‌های فازی است.

خوشه‌بندی فازی

در خوشه‌بندی کلاسیک هر نمونه ورودی متعلق به یک و فقط یک خوشه می‌باشد و نمی‌تواند عضو دو خوشه و یا بیشتر باشد و به زبان دیگر خوشه‌ها همپوشانی ندارند. حال حالتی را در نظر بگیرید که میزان تشابه یک نمونه با دو خوشه و یا بیشتر یکسان باشد در خوشه‌بندی کلاسیک باید تصمیم‌گیری شود که این نمونه متعلق به کدام خوشه است. تفاوت اصلی خوشه‌بندی کلاسیک و خوشه‌بندی فازی در این است که یک نمونه می‌تواند متعلق به بیش از یک خوشه باشد.

[1] Clustering Analysis
[2] Expert Systems

الگوریتم خوشه‌بندی c میانگین

مشابه الگوریتم c میانگین کلاسیک در این الگوریتم نیز تعداد خوشه‌ها (c) از قبل مشخص‌شده است. تابع هدفی که برای این الگوریتم تعریف‌شده است به‌صورت زیر می‌باشد:

رابطه (۲۲)
$$J = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} d_{ik}^{2} = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \|x_k - v_i\|^2$$

در رابطه فوق m یک عدد حقیقی بزرگ‌تر از یک است که در اکثر موارد برای m عدد دو انتخاب می‌شود. X_k نمونه k ام است و V_i نماینده یا مرکز خوشه i ام است. U_{ik} میزان تعلق نمونه i ام در خوشه k ام را نشان می‌دهد. علامت ||*|| میزان تشابه (فاصله) نمونه با (از) مرکز خوشه می‌باشد که می‌توان از هر تابعی که بیانگر تشابه نمونه و مرکز خوشه باشد را استفاده کرد. از روی U_{ik} می‌توان یک ماتریس U تعریف کرد که دارای c سطر و n ستون می‌باشد و مؤلفه‌های آن هر مقداری بین صفرتا یک را می‌توانند اختیار کنند. اگر تمامی مؤلفه‌های ماتریس U به‌صورت صفر و یا یک باشند الگوریتم مشابه c میانگین کلاسیک خواهد بود. بااینکه مؤلفه‌های ماتریس U می‌توانند هر مقداری بین صفر تا یک را اختیار کنند اما مجموع مؤلفه‌های هر یک از ستون‌ها باید برابر یک باشد و داریم:

رابطه (۲۳)
$$\sum_{i=1}^{c} u_{ik} = 1, \forall k = 1,\ldots,n$$

معنای این شرط این است که مجموع تعلق هر نمونه به c خوشه باید برابر یک باشد. با استفاده از شرط فوق و مینیمم کردن تابع هدف خواهیم داشت:

رابطه (۲۴)
$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}}$$

رابطه (۲۵)
$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{2/(m-1)}}$$

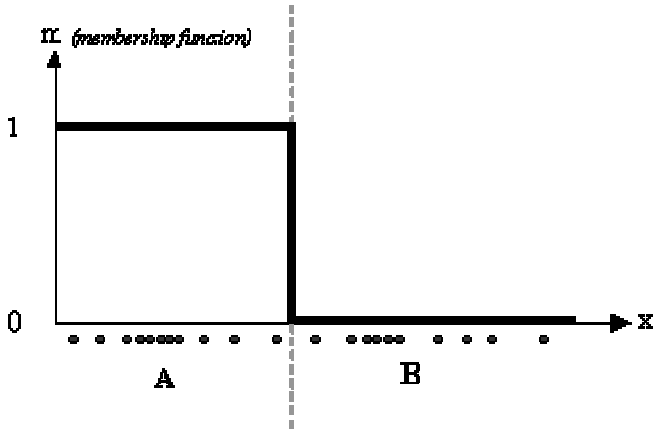
مراحل الگوریتم:

- مقداردهی اولیه برای c، m و U^0. خوشه‌های اولیه حدس زده شوند.
- مراکز خوشه‌ها محاسبه شوند (محاسبه v_i ها).
- محاسبه ماتریس تعلق از روی خوشه‌های محاسبه‌شده در مرحله دو.
- اگر $||Ul+1-Ul|| \leq \varepsilon$ الگوریتم خاتمه می‌یابد و در غیر این صورت برو به مرحله دو.

در شکل زیر یک توزیع یک‌بعدی از نمونه‌های ورودی را آورده شده است.

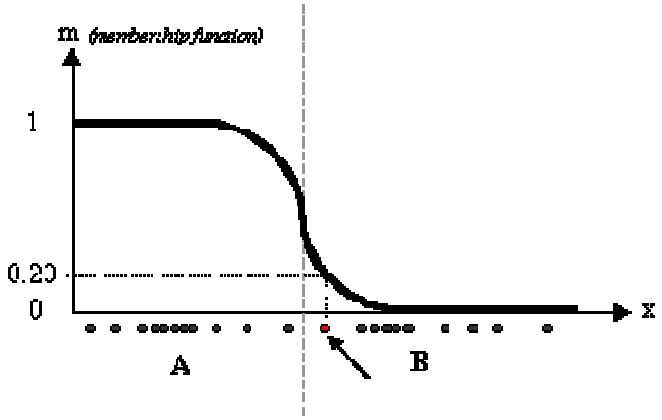
شکل (۶): توزیع یک‌بعدی نمونه‌ها

اگر از الگوریتم c میانگین کلاسیک استفاده کنیم داده‌های فوق به دو خوشه مجزا تقسیم خواهند شد و هر نمونه تنها متعلق به یکی از خوشه‌ها خواهد بود. به‌عبارت‌دیگر تابع تعلق هر نمونه مقدار صفر یا یک خواهد داشت. نتیجه خوشه‌بندی کلاسیک مطابق شکل زیر است:



شکل (۷): خوشه‌بندی کلاسیک نمونه‌های ورودی

شکل (۷) تابع تعلق مربوط به خوشه A را نشان می‌دهد. تابع تعلق خوشه B متمم تابع تعلق A می‌باشد. همان‌طور که مشاهده می‌کنید نمونه‌های ورودی تنها به یکی از خوشه‌ها تعلق دارند و به‌عبارت‌دیگر ماتریس U به‌صورت باینری می‌باشد. حال اگر از خوشه‌بندی فازی استفاده کنیم خواهیم داشت:



شکل (۸): خوشه‌بندی فازی نمونه‌ها

مشاهده می‌کنید که در این حالت منحنی تابع تعلق هموارتر است و مرز بین خوشه‌ها به‌طورقطع و یقین مشخص نشده است. به‌عنوان‌مثال نمونه‌ای که با رنگ قرمز مشخص شده است با درجه تعلق دو دهم به خوشه A و با درجه تعلق هشت دهم به خوشه B نسبت داده شده است.

نقاط قوت الگوریتم c میانگین فازی

- همیشه همگرا می‌شود.
- بدون نظارت بودن الگوریتم.

نقاط ضعف الگوریتم c میانگین فازی

- زمان محاسبات زیاد است.
- حساس به حدس‌های اولیه هست و ممکن در مینیمم‌های محلی متوقف شود.
- حساس به نویز می‌باشد.

اگر معیار تشابه در تابع هدف بر اساس فاصله تعریف شود می‌توان از تعاریف مختلفی که در مورد فاصله وجود دارد استفاده کرد که در زیر چند نمونه از این توابع آورده شده است:

روش خوشه‌بندی c میانگین فازی-امکانی (fpcm)

این روش از ترکیب خوشه‌بندی fcm با خوشه‌بندی pcm به‌دست‌آمده است تابع هدف در این نوع خوشه‌بندی توسط رابطه زیر تعریف می‌شود.

$$J_{fcm} = \sum_{i=1}^{N} \sum_{j=1}^{C} \left(u_{ij}^{m} + t_{ij}^{\eta} \right) \left\| x_i - c_j \right\|^2$$ رابطه (۲۶)

که در آن t_{ij} میزان درجه خصوصیت داده j ام به خوشه i ام و η کنترل‌کننده میزان تأثیر خصوصیت بر خوشه‌بندی می‌باشد و معمولاً در بازه [5 3] در نظر گرفته می‌شود برای به‌روزرسانی مقادیر u_{ij} از رابطه زیر به‌روزرسانی مقادیر عمومیت t و برای به‌روزرسانی بردارهای میانگین cj از رابطه زیر به دست می‌آید:

$$t_{ij} = \left[\sum_{k=1}^{N}\left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{\eta-1}}\right]^{-1}$$ رابطه (۲۷)

$$c_j = \frac{\sum_{i=1}^{N}\left(u_{ij}^{m}+t_{ij}^{\eta}\right)x_i}{\sum_{i=1}^{N}\left(u_{ij}^{m}+t_{ij}^{\eta}\right)}$$ رابطه (۲۸)

روش خوشه‌بندی pcm

رهیافت فازی pcm[1] برای مقاوم نمودن fcm در برابر داده‌های پرت و نویز دار می‌باشد محدودیت نرمال‌سازی رهیافت فازی احتمالی fcm به‌صورت $\forall j, \max u_{i,j} > 0$ ساده شد. سپس یک عبارت پنالتی به تابع هدف یا هزینه به‌صورت زیر اضافه شده است.

$$J_f\left(X, U_f, C\right) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2} + \sum_{i=1}^{c}\eta_i \sum_{j=1}^{n}\left(1-u_{ij}\right)^{m}$$ رابطه (۲۹)

افزوده شدن عبارت دوم به رابطه فوق مانع میل کردن تمام درجه عضویت‌ها به سمت صفر می‌شود زیرا با کم کردن مقدار ui,j در عبارت اول معکوس آن در عبارت دوم افزایش خواهد یافت. برای مشخص نمودن شکل خوشه باید مقدار η_iتخمین زده شود.

$$\eta_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} + d_{ij}^{2}}{\sum_{j=1}^{n} u_{ij}^{m}}$$ رابطه (۳۰)

$$u_{ij} = \frac{1}{1+\left[d_{ij}^{2}/\eta_i\right]^{1/(m-1)}}$$ رابطه (۳۱)

[1] Possibilistic C-Means

فصل چهارم

ساختار کلی سیستم پیشنهادی

بسیاری از سیستم‌های توصیه‌گر، معماری پیشنهاد شده از دو فاز تشکیل شده است. یک فاز برون خط و یک فاز بر خط. فاز برون خط از ماژول آماده‌سازی و پیش‌پردازش داده‌ها و ماژول کاوش الگوهای پیمایشی تشکیل شده است. در فاز برخط جهت انجام پیش‌بینی، از الگوهای پیمایشی کشف شده در فاز برون خط استفاده می‌شود. ماژول اصلی این فاز، موتور توصیه است. در ادامه هر کدام از ماژول‌ها توضیح داده خواهند شد.

آماده‌سازی و پیش‌پردازش داده‌ها

سرور وب تمام پیمایش‌های کاربران را در قالب فایل‌های ثبت دسترسی ذخیره می‌کند. داده‌های وب جمع‌آوری شده معمولا دارای حجم زیاد، بسیار ناهمگن و ساختارنیافته می‌باشند. وظیفه‌ی این مرحله آماده‌سازی داده‌های ثبت خام برای کاوش الگوهای پیمایشی است. این مرحله از چندین گام تشکیل شده است. در این پژوهش، این گام‌ها عبارت‌اند از؛ جمع آوری داده‌ها، پاکسازی داده‌ها، ساختاردهی داده‌ها و پیش پردازش نهایی داده‌ها. این مراحل برای هر مسئله‌ی کاوش استفاده از وب یکسان هستند. این ماژول در نهایت منجر به تولید مجموعه‌ای از m صفحه $P = \{p\_1, p\_2, \dots, p\_m\}$، یک مجموعه از n نشست کاربر $S = \{s\_1, s\_2, \dots, s\_n\}$، می‌شود که هر $s\_i \epsilon S$ زیرمجموعه‌ای از P می‌باشد. هر نشست یک دنباله به طول m از زوج‌های مرتب است.

که $s = \langle (p\_1\hat{}s, w(p\_1\hat{}s)), (p\_2\hat{}s, w(p\_2\hat{}s)), \dots ((p\_m\hat{}s, w(p\_m\hat{}s)) \rangle$ (در آن، $p_i^s = p_j$ برای یک $j \in \{1, \dots, m\}$ و $w(p_i^s)$ وزن صفحه p_i^s در نشست s است که یک عدد حقیقی غیرمنفی است. این وزن اهمیت صفحه‌ی p_i را در نشست s نشان می‌دهد. خروجی این مرحله یک ماتریس علاقه‌مندی‌ها خواهد بود که ستون‌های آن نمایانگر صفحات و هر کدام از سطرها، معادل یک نشست کاربر است و مقادیر موجود در ماتریس، وزن صفحات است که نمایانگر میزان علاقه‌مندی کاربر در آن نشست خاص به صفحه‌ی بازدید شده توسط او است. در ادامه هر یک از گام‌ها به

طور مختصر بررسی شده‌اند.

جمع آوری داده‌ها

در این مرحله فایل‌های ثبت دسترسی از سرورهای مورد نظر جمع آوری می‌گردند و در پایگاه داده جهت بررسی‌های آتی ذخیره می‌شوند. فایل‌های ثبت دسترسی متفاوتی با توجه به نوع تنظیمات سرور وجود دارند، اما تمامی این فایل‌ها دارای اطلاعات پایه‌ای از قبیل؛ تاریخ و زمان ارسال درخواست، نام منبع درخواست شده (URL)، آدرس IP سرویس گیرنده و متد درخواست (GET, POST, HEAD و غیره) هستند.

پاکسازی داده‌ها

در این مرحله داده‌های موجود بررسی می‌شوند و موارد نامربوط یا اضافی آن‌ها حذف می‌گردند. با حذف این موارد غیرمفید، اندازه‌ی داده‌ها نیز کاهش پیدا می‌کند. انتخاب داده‌هایی که باید حذف شوند، به هدف نهایی سیستم شخصی‌سازی وب بستگی دارد. درخواست‌هایی که در اغلب موارد از فایل‌های ثبت حذف می‌شوند، شامل موارد زیر هستند:

- **درخواست‌هایی با متد دسترسی متفاوت از 'GET'**: چنین درخواست‌هایی به درخواست‌های صریح کاربران اشاره نمی‌کنند. این درخواست‌ها بی اهمیت در نظر گرفته می‌شوند و در نتیجه باید از فایل ثبت حذف گردند.
- **درخواست‌های مربوط به اشیا چندرسانه‌ای:** بخش قابل توجهی از تکرارها در فایل‌های ثبت از مشخصات پروتکل HTTP ناشی می‌شود که برای هر فایل، تصویر، صوت، ویدیو و ... موجود در صفحات وب نیاز به یک درخواست جداگانه به سرور دارد. معمولا وارده‌هایی که به تصویر، صوت، فایل‌های ویدیویی و اسکریپت‌های CGI مراجعه می‌کنند، اضافی محسوب می‌شوند. این فایل‌ها بدون این که کاربر به طور صریح درخواست آن‌ها را بدهد، دانلود می‌شوند و از این رو بخشی از فعالیت حقیقی مرور کاربر محسوب نمی‌شوند. در نتیجه چنین داده‌هایی معمولا از فایل‌های ثبت حذف می‌شوند. البته حذف این گونه موارد از فایل‌های ثبت می‌تواند منجر به از دست رفتن اطلاعات ارزشمندی شود. یک مثال می‌تواند وب‌سایتی باشد که به‌طور عمده‌ای شامل

محتوای چندرسانه‌ای است. حذف و یا نگهداری این گونه درخواست‌ها به تحلیلگر بستگی دارد. در این پژوهش این گونه درخواست‌ها حذف شده‌اند.

- **درخواست‌های خراب و ناموفق:** چنین درخواست‌هایی توسط وارده‌های ثبتی که شامل یک کد خطای HTTP هستند، نمایش داده می‌شوند. یک درخواست موفق دارای کد وضعیت ۲۰۰ است و درخواست‌هایی که به غیر از این کد وضعیت را دارند، درخواست‌های ناموفق هستند و باید حذف گردند. علاوه بر این درخواست‌هایی که شامل مقدار گمشده در بعضی از فیلدها هستند نیز باید حذف گردند.
- **درخواست‌های مربوط به ربات‌های وب:** وارده‌های متناظر با اسپایدرها و خزش گرهای وب (مانند برنامه‌های دانلود کامل یک وب‌سایت و درخواست‌های موتورهای جستجو) نیز باید حذف شوند. دو راهکار برای تشخیص این درخواست‌ها وجود دارند: راهکار اول این است که تمامی وارده‌هایی که در قسمت URL آن‌ها ‘Robot.txt’ نوشته شده است، باید مشخص شده و حذف گردند. راهکار دوم مبتنی بر این است که ربات‌های وب صفحات را به‌صورت خودکار بازیابی می‌کنند، بنابراین داری سرعت پیمایش بسیار بالایی هستند. درنتیجه تمامی IP هایی که سرعت پیمایش در آن‌ها از آستانه‌ی مشخصی بیشتر است، باید حذف گردند. مقدار این آستانه با تحلیل رفتار پیمایشی فایل ثبت در نظر گرفته شده، تعیین می‌شود.

ساختاردهی داده‌ها

در این مرحله درخواست‌های ساختارنیافته در داده‌های ثبت به نشست‌های کاربران تبدیل می‌شوند. یک نشست کاربر مجموعه‌ای از صفحات بازدید شده توسط آن کاربر در یک بازدید خاص او از وب‌سایت است. تشخیص دقیق کاربران و نشست‌های آن‌ها در شخصی‌سازی وب اهمیت ویژه‌ای دارد، زیرا مدل‌های کاربران بر مبنای رفتار آن‌ها ساخته می‌شوند که آن‌ها هم به‌صورت نشست‌های کاربران وجود دارند. تشخیص کاربران از داده‌های ثبت یک وظیفه‌ی مشکل است، زیرا یک کاربر می‌تواند از کامپیوترهای مختلفی استفاده کند و نیز کاربران مختلف می‌توانند از یک کامپیوتر استفاده کنند.

راه‌های متفاوتی جهت تشخیص کاربران وجود دارند، اما بسیاری از آن‌ها مانند login، زحمتی را بر کاربر متحمل می‌کنند و یا مانند کوکی‌ها، تهدیدی برای امنیت و خصوصی بودن محسوب می‌شوند؛

بنابراین در این پژوهش از یک متد هیوریستیک استفاده شده است که در آن هر IP یکتا به عنوان یک کاربر در نظر گرفته می‌شود (قابل ذکر است که یک IP ممکن است توسط چندین کاربر استفاده شود). بعد از تشخیص کاربران، درخواست‌های مربوط به آن‌ها جدا می‌شوند، سپس باید نشست‌های آن‌ها استخراج گردند.

در این پژوهش جهت تشخیص نشست‌ها از یک متد مبتنی بر زمان استفاده شده است. به این ترتیب که اگر مدت زمان یک نشست از یک آستانه‌ی تعیین شده تجاوز کرد، به عنوان یک نشست جدید کاربر در نظر گرفته می‌شود. یافته‌های تجربی آستانه‌ی ۳۰ دقیقه را برای مدت زمان یک نشست پیشنهاد کرده‌اند.

پیش‌پردازش نهایی داده‌ها

علاوه بر مراحل پیش، چندین عمل تبدیل نیز می‌تواند بر روی داده‌های تراکنشی انجام شود. چنین اعمالی باعث بهبود در دقت توصیه‌های سیستم شخصی‌سازی بر مبنای کاوش استفاده از وب می‌شوند نشست‌های کوتاه به عنوان نشست‌های تصادفی در نظر گرفته می‌شوند و باید حذف گردند.

جهت نشان دادن تأثیر صفحات، به هر کدام از صفحات موجود در نشست‌ها براساس پارامترهای «مدت زمان مشاهده صفحه»، «فرکانس مشاهده صفحه» و «تاریخ» یک وزن تخصیص داده می‌شود که نشان دهنده‌ی میزان علاقه‌ی کاربر به آن صفحه در آن نشست خاص است.

مدت زمان مشاهده صفحه توسط کاربر، اهمیت آن صفحه را نشان می‌دهد، زیرا اگر صفحه‌ای برای وی جذاب نباشد، به سرعت از آن صفحه می‌گذرد و به صفحه‌ی دیگری می‌رود. البته باید در نظر داشت که اگر طول یک صفحه‌ی وب کم باشد، زمان مشاهده‌ی آن نیز متناسبا کمتر می‌شود. لذا زمان مشاهده‌ی صفحه متناسب با اندازه صفحه است و این تناسب باید در محاسبه میزان اهمیت صفحه در نظر گرفته شود. در این پژوهش جهت محاسبه میزان اهمیت یک صفحه براساس مدت زمان مشاهده‌ی آن، از رابطه (۳۲) استفاده شده است:

$$t(p) = \frac{\frac{\text{duration}(p)}{\text{size}(p)}}{\max_{q \in T}\left(\frac{\text{duration}(p)}{\text{size}(p)}\right)}$$

رابطه (۳۲)

در فرمول فوق، size(p)، اندازه‌ی صفحه را برحسب بایت و duration(p)، زمان مشاهده صفحه را برحسب ثانیه نشان می‌دهد.

فرکانس صفحه به این مفهوم است که در یک نشست امکان دارد کاربر به یک صفحه چندین بار مراجعه کند که هر چه تعداد این ارجاع‌ها به یک صفحه در یک نشست خاص بیشتر باشد، آن صفحه در نشست مذکور نسبت به سایر صفحات مهم‌تر است. جهت محاسبه میزان اهمیت یک صفحه براساس فرکانس آن از رابطه (۳۳) استفاده شده است:

$$f(p)=\frac{\text{NumberofVisit}(p)}{\sum_{q\in T}\text{NumberofVisit}(q)}$$ رابطه (۳۳)

تاریخ این گونه اهمیت خود را نشان می‌دهد که احتمال درخواست صفحات جدید نسبت به صفحات قدیمی برای کاربران بیشتر است و صفحات قدیمی احتمال بازدید کمتری دارند، زیرا کاربران به دنبال اطلاعات جدید هستند. در محاسبه میزان اهمیت براساس این معیار یک دوره‌ی زمانی در نظر گرفته می‌شود. با توجه به میزان تغییرپذیری صفحات، این دوره زمانی می‌تواند به صورت روزانه، هفتگی، ماهانه و غیره تعیین گردد. برای محاسبه میزان اهمیت یک صفحه براساس تاریخ آن از رابطه (۳۴) استفاده شده است:

$$d(p)=\begin{cases}1 & dif_d<\min\\ 1-\left(\frac{dif_d}{\max}\right) & \min<dif_d<\max\\ 0 & dif_d>\max\end{cases}$$ رابطه (۳۴)

که در فرمول فوق، $dif_d=date_{now}-date_p$ تفاضل بین تاریخ فعلی ($date_{now}$) با تاریخ مشاهده‌ی صفحه توسط کاربر ($date_p$) را نشان می‌دهد. در واقع به هر کدام از صفحات براساس میزان نزدیکی به تاریخ فعلی، وزنی اختصاص داده می‌شود. min حداقل آستانه برای تفاضل تعداد روزها را نشان می‌دهد. در صورتی که dif_d از حداقل آستانه‌ی تعیین شده، کمتر باشد، نشان می‌دهد که صفحه اخیرا توسط کاربر ملاقات شده است، بنابراین وزن بیشتری به آن اختصاص داده می‌شود. max حداکثر آستانه برای تفاضل تعداد روزها را نشان می‌دهد. در صورتی که dif_d از حداکثر آستانه‌ی تعیین شده بیشتر باشد، نشان دهنده‌ی این است که این صفحه در تاریخی بیشتر از

دوره‌ی زمانی در نظر گرفته، مشاهده شده است، بنابراین به این صفحه وزن صفر تعلق می‌گیرد و این صفحه در آن نشست خاص نادیده گرفته می‌شود. در این پژوهش با توجه به مجموعه داده‌های در دسترس، دوره‌ی زمانی به صورت ماهانه در نظر گرفته شد. بنابراین min برابر یک روز و max بر؛ ر ۳۰ روز خواهند بود.

در نهایت به هرکدام از صفحات با استفاده از رابطه (۳۶) وزنی اختصاص داده می‌شود:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}$$

رابطه (۳۵)

$$w(p) = \frac{2 * f(p) * t(p)}{f(p) + t(p)} * d(p)$$

رابطه (۳۶)

برای تعیین وزن صفحات آخر در نشست‌ها، میانگین وزن برای آن صفحه در سایر نشست‌هایی که در آن‌ها صفحه‌ی آخر نبوده است، محاسبه می‌گردد و به عنوان وزن به این صفحه اختصاص داده می‌شود.

کاوش الگوهای پیمایشی

در این مرحله الگوهای کاربرد مفید و مدل‌های توصیه توسط الگوریتم‌های داده‌کاوی از داده‌های پیش‌پردازش شده، کشف می‌شوند.

خوشه‌بندی فازی نشست‌ها

ماتریس علاقه‌مندی‌ها، محاسبه شده در گام قبلی، به عنوان ورودی برای الگوریتم خوشه‌بندی محسوب می‌شود. خوشه‌بندی، نشست‌های کاربران را براساس میزان علاقه‌مندی آن‌ها به صفحات، گروه‌بندی می‌کند. مطالعات اخیر مزایای استفاده از منطق فازی را برای بهبود تکنیک‌های شخصی‌سازی اثبات کرده است. بکارگیری منطق فازی در سیستم‌های توصیه‌گر در معرفی گردید. از آنجاییکه سلیقه‌ی کاربران دارای عدم قطعیت است، قرار گرفتن یک نشست در یک خوشه‌ی خاص نمی‌تواند سلیقه‌ی یک کاربر را بدرستی نشان دهد؛ بنابراین در این پژوهش جهت خوشه‌بندی نشست‌ها از خوشه‌بندی فازی استفاده شده است تا بتوان عدم قطعیت سلایق کاربران را به درستی پوشش داد.

برای انجام خوشه‌بندی فازی، از الگوریتم FCM[1] استفاده شده است. این متد خوشه‌بندی موجب می‌شود که یک داده بتواند به دو یا چند خوشه با درجه‌های تعلق متفاوت، تعلق گیرد. در سال ۱۹۹۶ روسپینی اولین مدل برای خوشه‌بندی با استفاده از تکنیک‌های فازی را معرفی کرد. این الگوریتم خوشه‌بندی مبتنی بر کمینه‌سازی تابع هدف رابطه (۳۷) است:

$$J_{FCM} = \sum_{j=1}^{C}\sum_{i=1}^{N} \mu_{ij}^{m} \left\| x_i - c_j \right\|^2$$ رابطه (۳۷)

در فرمول بالا، C تعداد خوشه‌ها، N تعداد داده‌ها و μ_{ij} درجه تعلق داده‌ی i به خوشه‌ی j است. m یک عدد حقیقی بزرگتر از یک است که پارامتر فازی نام دارد. در اکثر موارد مقدار دو برای آن اتخاذ می‌شود. x_i، iمین داده و c_j، مرکز خوشه‌ی jم است. ||*|| فاصله‌ی داده‌ی x_i از مرکز c_j است که می‌توان از هر تابع فاصله‌ای استفاده کرد و معمولا از فاصله‌ی اقلیدسی استفاده می‌شود. تابع فوق به طور مستقیم کمینه نمی‌شود، بنابراین باید از الگوریتم تکرارشونده استفاده شود. درجه تعلق داده‌ی i به خوشه‌ی j توسط رابطه (۳۸) و مراکز خوشه‌ها توسط رابطه (۳۹) بروز می‌شوند:

$$c_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{m} x_i}{\sum_{i=1}^{N} \mu_{ij}^{m}}$$ رابطه (۳۸)

این تکرار زمانی متوقف می‌شود که شرط رابطه (۳۹) برقرار شود:

$$\left\| \mu_{ij}^{(l)} - \mu_{ij}^{(l-1)} \right\| < \varepsilon$$ رابطه (۳۹)

از آنجاییکه در FCM مراکز اولیه خوشه‌بندی به‌صورت تصادفی انتخاب می‌شوند، در روش پیشنهادی این پژوهش ابتدا الگوریتم K-Means با معیار فاصله اقلیدسی مجموعه داده‌های ورودی را یک بار خوشه‌بندی می‌کند، زیرا K-Means سرعتی بیشتر نسبت به FCM دارد، پس با سرعت بالا و دقت پایینی می‌تواند مراکز خوشه‌ها را به دست آورد. مراکز به دست آمده در الگوریتم K-Means به عنوان مراکز اولیه در FCM استفاده می‌شوند. در الگوریتم FCM ابتدا باید درجه عضویت اولیه برای هر داده به مرکز هر خوشه مقداردهی شود که در FCM استاندارد این کار معمولا

[1] Fuzzy C-Means

به‌صورت تصادفی انجام می‌شود. در اینجا با توجه به استفاده از الگوریتم K-Means، فاصله هر داده تا هر یک از مراکز خوشه‌ها نرمالیزه را کرده و به عنوان درجه عضویت اولیه به الگوریتم FCM داده می‌شود.

کاوش قوانین انجمنی وزن‌دار خوشه‌ها

قوانین انجمنی ارتباطات میان اقلام را بر مبنای الگوهای وقوع آن‌ها با یکدیگر در تراکنش‌ها (بدون در نظر گرفتن ترتیب آن‌ها) نشان می‌دهند. در مورد نشست‌های وب، قوانین انجمنی ارتباطات بین مشاهده‌ی صفحه‌ها بر مبنای الگوهای پیمایشی کاربران را نشان می‌دهند. بیشتر رویکردهای کشف قوانین انجمنی بر مبنای الگوریتم Apriori می‌باشند که یک استراتژی تولید و آزمایش را بکار می‌برد. این الگوریتم دارای دو فاز است. در فاز اول، گروه‌های اقلام (مشاهده صفحات ظاهر شده در ثبت پیش‌پردازش شده) را که با یکدیگر به‌طور مکرر در تراکنش های زیادی ظاهر شده‌اند (و یک آستانه‌ی پشتیبانی تعیین شده توسط کاربر را ارضا می‌کنند) پیدا می‌کند. چنین گروه‌هایی از اقلام به مجموعه اقلام مکرر معروفند. یک ویژگی مهم پشتیبانی به نام ویژگی بستاری رو به پایین در الگوریتم Apriori به این شرح است: اگر یک مجموعه اقلام، معیار حداقل پشتیبانی را برآورده نسازد آن‌گاه هیچ یک از ابرمجموعه‌های آن نیز این معیار را برآورده نخواهند کرد. این ویژگی برای هرس کردن فضای حالت در حین هر تکرار الگوریتم Apriori اساسی است. در فاز دوم قوانین از مجموعه آیتم‌ست‌های تکرارشونده استخراج می‌گردند، سپس قوانینی که دارای ضریب اطمینان بالاتری از مینیمم ضریب اطمینان تعیین شده توسط کاربر هستند، به عنوان قوانین نهایی انتخاب می‌گردند. در این پژوهش نیز از این الگوریتم استفاده شده است.

در این مرحله، باید قوانین انجمنی وزن‌دار هر خوشه استخراج گردند. در این پژوهش از الگوریتم قوانین انجمنی وزن‌دار پیشنهاد شده در استفاده شده است. در الگوریتم پیشنهادی آن‌ها الگوریتم کاوش قوانین انجمنی معمول با اختصاص وزن به هر صفحه‌ی موجود در نشست برای نشان دادن میزان اهمیت آن صفحه در آن نشست خاص توسعه داده شده است و الگوریتم جدیدی به نام الگوریتم کاوش قوانین انجمنی وزن‌دار ارائه شده است.

وزن هرکدام از صفحات براساس سه پارامتر زمان سپری شده بر روی صفحه، فرکانس مشاهده

صفحه و تاریخ مشاهده صفحه محاسبه شد. وزن اقلام در هر نشست، براساس وزن صفحاتی که شامل آن‌ها است تعیین می‌گردد و با $w(X,s)$ نمایش داده می‌شود که در آن X مجموعه اقلام و s نشست را نشان می‌دهد. ساده‌ترین راه برای به دست آوردن وزن اقلام، در نظر گرفتن مینیمم وزن عضوهایی که شامل آن‌ها است، می‌باشد که در رابطه (۴۰) نشان داده شده است:

$$w(X,s)=\begin{cases}\min\left(w(p_1,p_2,...,p_k)\right) & X\subseteq s\\ 0 & X\not\subset s\end{cases}$$

رابطه (۴۰)

در رابطه‌ی بالا، k تعداد عضوهای اقلام است.

با اختصاص وزن به اقلام، می‌توان به هر کدام از نشست‌ها نیز وزنی نسبت داد. نسبت دادن وزن به هر کدام از نشست‌ها موجب می‌شود تا بتوان تفاوت نشست‌های مختلف را بهتر تشخیص داد. بدین ترتیب تراکنشی که وزن بالاتری دارد، در کاوش نتایج حاصل، نقش بیشتری ایفا می‌کند. ساده-ترین راه برای محاسبه وزن هر نشست، به دست آوردن میانگین وزن اقلامی می‌باشد که نشست شامل آن‌هاست. در این قسمت یک وزن جدید برای هر نشست ارائه شده است. از آنجاییکه هر نشست با درجه‌های تعلق متفاوتی به خوشه‌های فازی متفاوت، متعلق است، بنابراین وزن و ارزش آن نشست در خوشه‌های متفاوت، به درجه تعلق آن نشست در خوشه‌ها بستگی دارد. بدین ترتیب وزن نشست s_k در خوشه‌ی c، مطابق رابطه‌ی (۴۱) محاسبه می‌شود:

$$w(s_k)=\frac{\sum_{i=1}^{|s_k|}w(p_i)}{|s_k|}*\mu_{kc}$$

رابطه (۴۱)

در رابطه‌ی فوق μ_{kc} درجه تعلق نشست s_k به خوشه‌ی c را نشان می‌دهند.

ضریب پشتیبانی قوانین انجمنی وزن‌دار اقلام X در میان همه‌ی نشست‌ها، مطابق رابطه (۴۲) محاسبه می‌گردد:

$$wsp(X)=\frac{\sum_{s_i\in S}w(s_i)*w(X,s_i)}{\overline{w}*\sum_{k=1}^{|S|}w(s_k)}$$

رابطه (۴۲)

که در فرمول بالا، $\overline{w}$ میانگین وزن همه‌ی صفحات در کل نشست‌ها و S مجموعه‌ای از همه‌ی

نشست‌ها است.

ضریب اطمینان وزن‌دار برای قوانین انجمنی وزن‌دار نیز طبقه رابطه (۴۳) تعریف می‌شود:

$$wconf(X \Rightarrow Y) = \frac{wsp(X \cup Y)}{wsp(X)}$$ رابطه (۴۳)

در این مدل، در کنار دو پارامتر ضریب پشتیبانی و ضریب اطمینان وزن‌دار، وزن هر صفحه نیز در هر قانون نمایش داده می‌شود؛ بنابراین قوانین انجمنی وزن‌دار به شکل زیر تعریف می‌شوند:

به‌طوری‌که $(p_1, p_2, \dots, p_k), (q_{k+1}, q_{k+2}, \dots, q_{k+l})$ سرآمد و بدنه‌ی قوانین انجمنی وزن‌دار را به ترتیب نشان می‌دهند و δ بیانگر ضریب پشتیبانی وزن‌دار و α، ضریب اطمینان وزن‌دار و $(w_1, w_2, \dots, w_{k+l})$، بیانگر وزن متناظر هر یک از صفحات است.

خروجی این مرحله قوانین انجمنی وزن‌دار استخراج شده به ازای هر خوشه‌ی فازی است.

موتور توصیه

هنگامی که یک کاربر وارد سیستم می‌شود، موتور توصیه با تحلیل رفتار وی در مقایسه با نتایج حاصل از مدل کردن رفتار کاربران در فاز برون‌خط، به پیش‌بینی رفتار آتی وی می‌پردازد و لیستی از صفحاتی که به‌احتمال‌زیاد موردتوجه او قرار خواهند گرفت، در اختیار وی قرار می‌دهد. صفحات توصیه‌شده به آخرین صفحه از نشست جاری کاربر قبل از ارسال به مرورگر وی اضافه می‌شوند.

موتور توصیه پس از دریافت نشست جاری کاربر، ابتدا وزن صفحات را برای آن محاسبه می‌کند.

محاسبه درجه تعلق نشست جاری کاربر به خوشه‌ها

در این مرحله موتور توصیه درجه تعلق نشست جاری کاربر به تمامی خوشه‌های فازی را محاسبه می‌کند. محاسبه‌ی درجه تعلق به این صورت خواهد بود که میزان شباهت نشست جاری کاربر به مراکز هر کدام از خوشه‌ها تعیین می‌گردد و سپس با توجه به آن، درجه‌های تعلق برای آن نشست خاص به خوشه‌های متفاوت محاسبه می‌گردد.

تعیین تعداد صفحات انتخابی از هر خوشه

در این پژوهش یک راهکار جدید برای تعیین تعداد صفحات انتخابی جهت پیشنهاد از هر خوشه ارائه شده است. در این راهکار از یک متد غیرفازی‌سازی استفاده می‌شود. به‌این‌ترتیب که یک تناسب بین تعداد صفحات انتخابی از هر خوشه و درجه تعلق نشست جاری کاربر به آن خوشه برقرار می‌شود. برای مثال درصورتی‌که سه خوشه وجود داشته باشند و ۱۰ صفحه باید به کاربر توصیه شوند و درجه تعلق نشست جاری کاربر به خوشه‌های یک تا سه به ترتیب (۰.۷، ۰.۲ و ۰.۱) باشد، تعداد صفحاتی انتخابی جهت توصیه به کاربر از خوشه‌های یک تا سه به ترتیب هفت، دو و یک صفحه خواهند بود.

یافتن قوانین انجمنی وزن‌دار منطبق با نشست جاری کاربر

در این گام، دنباله‌ی صفحات مشاهده شده در نشست جاری کاربر با بخش سرآمد قوانین انجمنی وزن‌دار مربوط به خوشه‌ی فازی، منطبق می‌شوند و قوانینی که صفحات موجود در بخش سرآمد آن‌ها در نشست جاری کاربر وجود دارند، استخراج می‌گردند. سپس درجه‌ی تطابق نشست جاری کاربر با سرآمد هر یک از قوانین انجمنی وزن‌دار استخراج شده، توسط فرمول‌های (۴۴) و (۴۵) محاسبه می‌گردد:

$$\text{MatchScore}\left(s, r_L\right) = 1 - \frac{1}{4}\sqrt{\frac{\text{Dissimilarity}\left(s, L_r\right)}{\sum_{i:r_{Li}} 1}}$$

رابطه (۴۴)

$$\text{Dissimilarity}\left(s, r_L\right) = \sum_{i:r_{Li}>0}\left(\frac{2*\left(w\left(s_i\right) - w\left(r_{Li}\right)\right)}{w\left(s_i\right) + w\left(r_{Li}\right)}\right)^2$$

رابطه (۴۵)

در فرمول‌های بالا، s، نشست جاری کاربر و r_L، بخش سرآمد قانون انجمنی وزن‌دار را نشان می‌دهد. به دلیل اینکه توصیه‌ی هر کدام از صفحات مستقل از یکدیگر است، بخش بدنه‌ی هر قانون انجمنی وزن‌دار دارای یک صفحه است. این مرحله به ازای هر خوشه‌ی فازی به‌صورت جداگانه انجام می‌گردد.

تولید مجموعه توصیه

در این مرحله به ازای هر کدام از صفحات موجود در بخش بدنه‌ی قوانین انجمنی وزن‌دار استخراج شده در مرحله‌ی قبل که در نشست جاری کاربر وجود ندارند، امتیاز پیشنهاد طبق رابطه (۴۶) محاسبه می‌گردد:

$$\mathrm{Re}\,c(s, X \Rightarrow p) = \mathrm{MatchScore}\,(s, X) * \mathrm{wconf}\,(X \Rightarrow p) \qquad \text{رابطه (۴۶)}$$

در فرمول بالا مشاهده می‌شود که امتیاز به ازای هر صفحه براساس دو فاکتور درجه تطابق و ضریب اطمینان وزن‌دار آن قانون انجمنی وزن‌دار تعیین می‌گردد. پس از محاسبه‌ی امتیاز، صفحات براساس امتیازهای آن‌ها به‌صورت نزولی مرتب می‌گردند. سپس براساس تعداد صفحات انتخابی تعیین شده از هر خوشه در مرحله‌ی ۳-۴-۲، n تا از صفحاتی که بالاترین امتیاز را دارند، انتخاب می‌گردند و به‌عنوان صفحات توصیه به کاربر پیشنهاد می‌شوند.

فهرست منابع

منابع فارسی

۱ احمدیان، س. مرادی، ر. اخلاقیان طاب ف. (۱۳۹۳). ارائه راهکارهای مبتنی بر روابط اعتماد به‌منظور افزایش کارایی سیستم‌های توصیه‌گر.

۲ پروازه، ف.، هارون آبادی، ع؛ و نیزاری، م. (۱۳۹۴). مقایسه و ارزیابی تکنیکهای توصیه دوست در شبکه های اجتماعی . سومین کنفرانس بین المللی پژوهشهای کاربردی در مهندسی کامپیوتر و فن آوری اطلاعات، تهران، دانشگاه صنعتی مالک اشتر

۳ حیدری، ط. کارگر، م. (۱۳۹۱). سیستم‌های توصیه گر. دومین همایش ملی کامپیوتر برق و فن‌آوری اطلاعات، خمین، دانشگاه آزاد اسلامی واحد خمین.

۴ دیفن، ج؛ و شیری احمدآبادی، م. (۱۳۹۲). ارائه مدل مفهومی سیستم توصیه گر گردشگری مبتنی بر شبکه اجتماعی و با بهره‌گیری از فناوری وب معنایی. دوازدهمین کنفرانس ملی سیستم‌های هوشمند، بم، انجمن سیستم‌های هوشمند ایران، مجتمع آموزش عالی بم.

۵ ساعد، ن.، صادق زاده، م. (۱۳۹۳). ارائه الگوریتمی جهت خوشهبندی گراف شبکه های اجتماعی مبتنی بر مرکزیت گره ها. اولین همایش ملی مهندسی برق و کامپیوتر در شمال کشور، بندر انزلی، موسسه آموزش عالی موج.

۶ ساعد، ن.، صادق زاده، م. (۱۳۹۳). بهبود خوشه‌بندی در شبکه‌های اجتماعی مبتنی بر الگوریتم میانگی یال. اولین همایش ملی مهندسی برق و کامپیوتر در شمال کشور، بندر انزلی، موسسه آموزش عالی موج.

۷ سمیعی، ا.، میرزائی، ک. (۱۳۹۴). ارزیابی یک سیستم توصیه گر با استفاده از الگوریتم تکاملی ترکیبی GAPSO. اولین کنفرانس بین‌المللی وب پژوهشی، تهران، دانشگاه علم و فرهنگ.

۸ عادلی، ه.، نیک‌نفس، ع. (۱۳۸۷). طراحی سیستم توصیه گر تغذیه بیماران مبتنی بر

چرخ‌دستی‌های هوشمند و الگوریتم ژنتیک. پنجمین کنفرانس بین‌المللی مدیریت فناوری اطلاعات و ارتباطات، تهران، ندای اقتصاد بامداد (ناب).

۹ قیافه داودی، ف. (۱۳۹۱). ارائه یک سیستم توصیه گر در وب‌گاه‌های نشانه‌گذاری اجتماعی با استفاده از مدل کاربر. دانشگاه تهران (با حمایت مرکز تحقیقات مخابرات ایران).

۱۰ کی پور، ا.، برداری، م؛ و شیرازی، ح. (۱۳۹۳). ارائه روشی برای پیشگویی پیوند بین راس‌های موجود در شبکه‌های اجتماعی، فصلنامه علمی-پژوهشی مدیریت فناوری اطلاعات، ۶(۳), ۴۷۵-۴۸۶

۱۱ معینی، ط. (۱۳۹۳). ارائه روشی جهت بهبود صحت سیستم‌های پیشنهاددهنده در شبکه‌های اجتماعی با استفاده از تشخیص انجمن‌ها. همایش سیستم‌های هوشمند کامپیوتری دانشگاه پیام نور.

۱۲ معینی، ط. (۱۳۹۳). بهبود صحت سیستم‌های پیشنهاددهنده با استفاده از تشخیص اجتماعات در شبکه‌های اجتماعی. کنفرانس داده‌کاوی، پنجمین کنفرانس داده‌کاوی ایران، دانشگاه صنعتی امیرکبیر.

۱۳ نیک‌نفس، ا.، مقدم چرکری ن؛ و نیک‌نفس، ع. (۱۳۸۷). سیستم توصیه گر مبتنی بر روش PROMETHEE II برای دسته‌های مختلف اقلام با تکرار خرید پایین. چهاردهمین کنفرانس سالانه انجمن کامپیوتر ایران، تهران، انجمن کامپیوتر، دانشگاه صنعتی امیرکبیر.

منابع لاتین

14 Achananuparp, P., H. Han, O. Nasraoui and R. Johnson (2007). Semantically enhanced user modeling. Proceedings of the 2007 ACM symposium on Applied computing, ACM, pp. 1335-1339.

15 Agarwal, V., and Bharadwaj, K. K. (2013). A collaborative filtering framework for friends recommendation in social networks based on interaction intensity and adaptive user similarity. Social Network Analysis and Mining, vol. 3, no. 3, pp. 359-379.

16 Agrawal, R. and R. Srikant (1994). Fast algorithms for mining association rules. Proceedings of the 20th International Conference very large data bases, pp. 487-499.

17 Akbari, F., Tajfar, A. H., and Nejad, A. F. (2013). Graph-Based Friend Recommendation in Social Networks Using Artificial Bee Colony. Dependable, Autonomic and Secure Computing (DASC), 11th International Conference on, pp. 464-468.

18 AlMurtadha, Y., M. N. Sulaiman, N. Mustapha and N. I. Udzir (2011). Improved web page recommender System Based on Web Usage Mining. Proceedings of the 3rd International Conference on Computing and Informatics, pp. 32-36.

19 AlMurtadha, Y., M. N. Sulaiman, N. Mustapha and N. I. Udzir (2011). IPACT: Improved web page recommendation system using profile aggregation based on clustering of transactions. American Journal of Applied Sciences, vol. 8, no. 3, pp. 277-383.

20 Anitha, A. (2010). A new web usage mining approach for next page access prediction. International Journal of Computer Applications, vol. 8, no. 11, pp. 7-10.

21 Balasko, B., J. Abonyi and B. Feil (2005). Fuzzy clustering and data analysis toolbox. Department of Process Engineering, University of Veszprem, Veszprem.

22 Bobadilla, J., Ortega, F. and Hernando, A. (2013). Recommender systems survey. Knowledge-Based Systems, vol. 46, pp. 109-132.

23 Burke, R. (2002). Hybrid recommender systems: Survey and experiments. User modeling and user-adapted interaction, vol. 12, no. 4, pp. 331-370.

24 Cantador, I., Konstas I. and Jose, J. M. (2011). Categorising social tags to improve folksonomy-based recommendations. Science, Services and Agents on the World Wide Web, vol. 9, no. 1, pp. 1-15.

25 Castellano, G., A. M. Fanelli and M. A. Torsello (2011). NEWER: A system for NEuro-fuzzy WEb Recommendation. Applied Soft Computing, vol. 11, no. 1, pp.793-806.

26 Castellano, G., A. M. Fanelli and M. A. Torsello (2011). NEWER: A system for NEuro-fuzzy WEb Recommendation. Applied Soft Computing,

vol. 11, no. 1, pp. 793-806.

27 Chaudhuri, A. (2015). Intuitionistic fuzzy possibilistic c means clustering algorithms. Advances in Fuzzy Systems, p.1.

28 Chintalapudi, K. K,. Kam, M.(1998). A Noise-Resistant Fuzzy C Means Algorithm for Clustering, IEEE World Congress on Computational Intelligence Vol. 2, pp. 1458-1463.

29 Correa, C., Valero, C., Barreiro, P., Diago, M. P., and Tardáguila, J. (2011). A comparison of fuzzy clustering algorithms applied to feature extraction on vineyard. Proceedings of the XIV Conference of the Spanish Association for Artificial Intelligence.

30 Dixit, D. and J. Gadge (2010). Automatic Recommendation for Online Users Using Web Usage Mining. International Journal of Managing Information Technology, vol. 2, no. 3, pp. 33-42.

31 Döring, C., Lesot, M. J., and Kruse, R. (2006). Data analysis with fuzzy clustering methods. Computational Statistics & Data Analysis, vol. 51, no. 1, pp. 192-214.

32 Feitosa, R. M. Labidi, S. Silva dos Santos A. L and. Santos, N. (2013). Social Recommendation in Location-Based Social Network Using Text Mining. Intelligent Systems Modelling & Simulation (ISMS), vol. 4th International Conference on, pp. 67-72.

33 Felfernig, A., Gordea, S., Jannach, D., Teppan, E. and Zanker, M. (2007). A short survey of recommendation technologies in travel and tourism. OEGAI Journal, vol. 25, no. 7, pp. 17-22.

34 Forsati, R., M. Meybodi and A. G. Neiat (2009). Web page personalization based on weighted association rules. International Conference on Electronic Computer Technology, IEEE, pp. 130-135.

35 Forsati, R., Meybodi, M. R., Ghari Neiat, A. (2009). Web page personalization based on weighted association rules, In International Conference on Electronic Computer Technology, pp. 130-135,.

36 Gan, G., Ma C. and Wu, J. (2007). Data clustering: theory, algorithms, and applications, Vol. 20: Siam.

37 Gao, M., K. Liu and Z. Wu (2010). Personalisation in web computing and informatics: Theories, techniques, applications, and future research. Information Systems Frontiers, vol. 12, no. 5, pp. 607-629.

38 Göksedef, M. and Ş. Gündüz-Öğüdücü (2010). Combination of Web page recommender systems. Expert Systems with Applications, vol. 37, no.4, pp. 2911-2922.

39 Guerbas, A., O. Addam, O. Zaarour, M. Nagi, A. Elhajj, M. Ridley and R. Alhajj (2013). Effective web log mining and online navigational pattern prediction. Knowledge-Based Systems, 49, pp. 50-62.

40 Gupta, A., Jain, R., and Song, S. (2008). Movie Recommendations Using Social Networks.

41 Halkidi, M., Batistakis Y. and Vazirgiannis, M. (2001). On Clustering Validation Techniques. Journal of Intelligent Systems, vol. 17, no.2-3, pp 107-145.

42 Han, J., Kamber, M. and Pei, J. (2011). Data mining: concepts and techniques: concepts and techniques. Elsevier.

43 He, J., Chu, W. W. (2010). A social network-based recommender system (SNRS), Springer US.

44 Hun, X., Wang, L., Crecpi, N., Park, S. and Cuevas, A. (2015). Alike people, alike interests? Inferring interest similarity in online social networks. Decision Support Systems, vol. 69, pp. 92-106.

45 Jalali, M., N. Mustapha, M. N. Sulaiman and A. Mamat (2010). WebPUM: A Web-based recommendation system to predict user future movements. Expert Systems with Applications, vol. 37, no. 9, pp. 6201-6212.

46 Kangas, S. (2002). Collaborative Filtering and Recommendation Systems, VTT information technology.

47 Kazienko, P. Musiał, K. and Kajdanowicz, T. (2011). Multidimensional social network in the social recommender system. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 41, no. 4, pp. 746-759.

48 Kim, H. N., El Saddik, A., and Jung, J. G. (2012). Leveraging personal photos to inferring friendships in social network services. Expert Systems with Applications, vol. 39, no. 8, 6955-6966.

49 Lucas, J. P., N. Luz, M. N. Moreno, R. Anacleto, A. Almeida Figueiredo and C. Martins (2013). A hybrid recommendation approach for a tourism system. Expert Systems with Applications, vol. 40, no. 9, pp. 3532-3550.

50 Manesh, M. A., Harounabadi, A., & Golabpour, A. (2015). A Hybrid Approach for Web Personalization based on Fuzzy Clustering and Weighted Association vol, 5. no, 17., PP. 2468-248

51 Moghaddam, A., Nodoshan, J. (2010). Geometry optimization of triangle labyrinth spillway using ANFIS models and genetic algorithm. Journal of Modeling in Engineering, vol. 5, no. 19, pp. 57-68.

52 Naruchitparames, J., Gunes, M. H., and Louis, S. J. (2011). Friend recommendations in social networks using genetic algorithms and network topology. IEEE Congress on Evolutionary Computation ,pp. 2207-2214.

53 Nigam, B. and S. Jain (2010). Generating a new model for predicting the next accessed web page in web usage mining. 3rd International Conference on Emerging Trends in Engineering and Technology (ICETET), IEEE, pp. 485-490.

54 NING, L. J., and DUAN, H. Y. (2014). An algorithm for friend-recommendation of social networking sites based on SimRank and ant colony optimization. The Journal of China Universities of Posts and Telecommunications, 21, pp. 79-87.

55 Niranjan, U., R. Subramanyam and V. Khanaa (2010). An efficient system based on closed sequential patterns for web recommendations. International Journal of Computer Science Issues, vol. 7, no. 4, pp. 26-34.

56 Pamutha, T. Chimphlee, S. Kimpan, C. and Sanguansat, P. (2012). Data Preprocessing on Web Server Log Files for Mining Users Access Patterns. International Journal of Research and Reviews in Wireless Communications (IJRRWC), vol. 2, no. 2, pp. 92-98.

57 Papadimitriou, A., Symeonidis, P., and Manolopoulos, Y. (2012). Fast and accurate link prediction in social networking systems. Journal of Systems and Software, vol. 85, no. 9, pp. 2119-2132.

58 Passant, A., Raimond, Y. (2008). Combining Social Music and Semantic Web for music-related recommender systems. In The 7th International Semantic Web Conference, vol. 19.

59 Poornalatha, G. and P. S. Raghavendra (2012). Web Page Prediction by Clustering and Integrated Distance Measure. Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, pp. 1349-1354.

60 Pujahari, A. and Padmanabhan, V. (2015). A New Grouping Method Based on Social Choice Strategies for Group Recommender System. Computational Intelligence in Data Mining, vol. 1, pp. 325-332.

61 Rashidi, S., A. Harounabadi and M. Dezfouli (2012). Prediction of users' future requests using neural network. Management Science Letters, vol. 2, no. 6, pp. 2119-2124.

62 Rashidi, S., Harounabadi, A. and Dezfouli, M. (2012). Prediction of users' future requests using neural network. Management Science Letters, vol. 2, no. 6, pp. 2119-2124.

63 Ricci, F., Lior, R. and Bracha, S. (2011). Introduction to recommender systems handbook. Springer US, pp. 1-35.

64 Santra, A. K., Jayasudha, S. (2012). Classification of web log data to identify interested users using naïve Bayesian classification. International Journal of Computer Science Issues, vol. 9, no. 1, pp. 381-387.

65 Silva, N. B., Tsang, I. R., Cavalcanti, G. D., and Tsang, I. J. (2010). A graph-based friend recommendation system using genetic algorithm. IEEE Congress on Evolutionary Computation,pp. 1-7.

66 Spiliopoulou, M., B. Mobasher, B. Berendt and M. Nakagawa (2003). A framework for the evaluation of session reconstruction heuristics in web-usage analysis. Informs journal on computing, vol. 15, no. 2, pp. 171-190.

67 Suchanek, F. M., Kasneci, G. and Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. Web Semantics: Science, Services and Agents on the World Wide Web, vol. 6, no. 3, pp. 203-217.

68 Sudhamathy G., Venkateswaran, C. J. (2012). Matrix based Fuzzy Clustering for Categorization of Web Users and Web Pages. International Journal of Computer Applications, vol. 43, no. 14, pp. 43-47.

69 Sujatha, V., Punithavalli, A. (2010). An approach to user navigation pattern based on ant based clustering and classification using decision tress. International Journal of Advanced Engineering Scienced And Technologies, vol. 1, no. 2, pp. 112-117,.

70 Symeonidis, P., and Mantas, N. (2013). Spectral clustering for link prediction in social networks with positive and negative links. Social Network Analysis and Mining, vol. 3, no.4, pp. 1433-1447.

71 Symeonidis, P., and Perentis, C. (2014). Link Prediction in Multi-modal Social Networks. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 147-162. Springer Berlin Heidelberg.

72 Symeonidis, P., and Tiakas, E. (2014). Transitive node similarity: predicting and recommending links in signed social networks. World Wide Web, vol. 17, no. 4, pp. 743-776.

73 Tahmasebi, P., Hezarkhani, A. (2012). A hybrid neural networks-fuzzy logic-genetic algorithm for grade estimation. Computers & geosciences, vol. 42, pp. 18-27.

74 Talabeigi, M., R. Forsati and M. R. Meybodi (2010). A hybrid web recommender system based on cellular learning automata. IEEE International Conference on Granular Computing IEEE, pp. 453-458.

75 Thiyagarajan, R., K. Thangavel and R. Rathipriya (2014). Recommendation of Web Pages using Weighted K-Means Clustering. International Journal of Computer Applications, vol. 86, no. 14, pp. 44-48.

76 Thwe, P. (2014). WEB PAGE ACCESS PREDICTION BASED ON INTEGRATED APPROACH. International Journal of Computer Science and Business Informatics, vol. 12, no. 1, pp. 55-64.

77 Tian, X., Song, Y., Wang, X., and Gong, X. (2012). Shortest path based potential common friend recommendation in social networks. Cloud and Green Computing (CGC), 2012 Second International Conference on ,pp. 541-548.

78 Tsekouras, G. E., Sarimveis, H. (2004). A new approach for measuring the validity of the fuzzy c-means algorithm, Advances in Engineering Software 35 567–575.

79 Udzir, N. I, Mustapha, N., AlMurtadha, Y., Sulaiman, M. N. (2011) IPACT: Improved web page recommendation system using profile aggregation based on clustering of transactions, American Journal of Applied Sciences, vol. 8, no. 3, pp. 277-283.

80 Wan, S., Lan, Y., Guo, J., Fan, C., and Cheng, X. (2013). Informational friend recommendation in social media. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval,pp. 1045-1048.

81 Yamashita, A., Kawamura, H. and Suzuki, K. (2010). Similarity Computation Method for Collaborative Filtering Based on Optimization. JACIII, vol. 14, no. 6, pp. 654-660.

82 Zahid, N., Limouri, M. and Essaid, A. A new cluster-validity for fuzzy clustering, Pattern Recognition 32 (1999) 1089-1097.

83 Zbal, G., Karaman, H. (2008). Matchbook A Web Based Recommendation System For Matchmaking. 23rd International Symposium on Computer and Information Sciences, pp. 1 - 6.

84 Zhen, L., Huang G. Q. and Jiang, Z. (2009). Recommender system based on workflow. Decision Support Systems, vol. 48, no. 1, pp. 237-245.