

# علوم رایانش و فناوری اطلاعات

## نشریه علمی انجمن کامپیوتر ایران

صاحب امتیاز: انجمن کامپیوتر ایران

مدیر مسئول: دکتر جعفر حبیبی

سر دبیر: دکتر احمد خونساری

### شورای علمی

قاسم جابری پور، دانشیار دانشگاه شهید بهشتی  
جعفر حبیبی، دانشیار دانشگاه صنعتی شریف  
امیر حسین جهانگیر، دانشیار دانشگاه صنعتی شریف  
شاهین حسابی، دانشیار دانشگاه صنعتی شریف  
سید حمید حاجی سید جوادی، دانشیار دانشگاه شاهد  
مسعود رهگذر، دانشیار دانشگاه تهران  
مهدی صدیقی، دانشیار دانشگاه صنعتی امیرکبیر  
هشام فیلی، دانشیار دانشگاه تهران  
عبدالرسول قاسمی، دانشیار دانشگاه خواجه نصرالدین طوسی  
مقصود عباسپور، دانشیار دانشگاه شهید بهشتی  
محمد عبداللهی ازگمی، دانشیار دانشگاه علم و صنعت ایران  
مهدی کارگهی، دانشیار دانشگاه تهران  
مازیار گودرزی، دانشیار دانشگاه صنعتی شریف  
ناصر مزینی، دانشیار دانشگاه علم و صنعت ایران

حمیدرضا ربیعی، استاد دانشگاه صنعتی شریف  
حمید سربازی آزاد، استاد دانشگاه صنعتی شریف  
کریم فائز، استاد دانشگاه صنعتی امیرکبیر  
اکبر غفارپور رهبر، استاد دانشگاه صنعتی سهند  
احسان الله کبیر، استاد دانشگاه تربیت مدرس  
کیوان ناوی، استاد دانشگاه شهید بهشتی  
ناصر یزدانی، استاد دانشگاه تهران  
محمد حسین یغمایی مقدم، استاد دانشگاه فردوسی مشهد  
مرتضی آنالویی، دانشیار دانشگاه علم و صنعت ایران  
محسن ابراهیمی مقدم، دانشیار دانشگاه شهید بهشتی  
حسین اسدی، دانشیار دانشگاه صنعتی شریف  
احمد اکبری ازیرانی، دانشیار دانشگاه علم و صنعت ایران  
رضا برنگی، دانشیار دانشگاه علم و صنعت ایران  
حسین پدرام، دانشیار دانشگاه صنعتی امیرکبیر  
نصراله مقدم چرکری، دانشیار دانشگاه تربیت مدرس

### همکاران دفتر نشریه

لیلا نورانی  
عبدالله غفاری ششجوانی

### نشانی

تهران، خیابان آزادی، ضلع غربی دانشگاه صنعتی شریف، کوچه شهید ولی... صادقی، پلاک ۲۶، طبقه ۴، واحد ۱۶، دفتر انجمن کامپیوتر ایران، نشریه علوم رایانش و فناوری اطلاعات

تلفن: ۰۲۱-۶۶۰۳۲۰۰۰-۶۶۰۸۷۲۲۴

دورنگار: ۰۲۱-۶۶۰۲۱۱۴۹

پست الکترونیکی: csitjour@gmail.com

نشانی سایت: <http://csi.org.ir/fa/publication/archive/name/csit>

مقالات درج شده در این نشریه صرفاً بیانگر نظرات مؤلفین آنها است و مسئولیت صحت و سقم داده‌ها و نتایج بر عهده آنها است.

لیتوگرافی، چاپ و صحافی:

## فهرست مقالات

- یادگیری عمیق در سامانه‌های توصیه‌گر ..... ۱  
امید عباسی و مهدیه سلیمانی باغشاه
- ارائه روشی متنی بر محتوا و جذابیت برای خلاصه‌سازی مجموعه تصاویر اجتماعی ..... ۱۳  
زهرا ریاحی سامانی و محسن ابراهیمی مقدم
- استفاده از شبکه‌های اجتماعی برای بهبود مدیریت ارتباط با مشتری ..... ۲۷  
الناز مظاهری و سید ابوالقاسم میرروشندل
- مدل‌سازی تکامل شبکه‌های ارتباطات کلاسی نرم‌افزارهای شی‌گرا ..... ۳۶  
مجتبی صادقیان، محمد خوانساری و فرید دهقان
- سیستم پیشنهاددهنده قانون محور تکاملی ..... ۴۷  
بهزاد سلیمانی نیسیانی، نسیم سلطانی، رضا مفیدی و محمد حسین ندیمی شهرکی
- بهبود توان مصرفی و زمان اجرا در رایانش ابری رهیافت تخصیص منبع و زمانبندی وظایف ..... ۵۶  
عاطفه یکتا اول و محمود فتحی





## یادگیری عمیق در سامانه‌های توصیه‌گر

مهديه سلیمانی باغشا

امید عباسی

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، تهران، ایران

### چکیده

روش پالایش همکارانه یکی از کارآمدترین و پرکاربردترین روش‌های مورد استفاده در بسیاری از سامانه‌های توصیه‌گر است. روش‌های اولیه پالایش همکارانه ویژگی‌های نهان کاربران و گزینه‌ها را با استفاده از تجزیه ماتریس امتیازات به دست می‌آوردند، اما این روش‌ها با مشکل شروع سرد و تنک بودن ماتریس امتیازات مواجه می‌شوند. در سال‌های اخیر استفاده از اطلاعات اضافه موجود، در کنار ماتریس امتیازات برای به دست آوردن ویژگی‌های نهان مورد توجه قرار گرفته است. از طرف دیگر، مدل‌های یادگیری عمیق، توانایی بالایی در یادگیری بازنمایی مناسب، به خصوص در مواقعی که با داده‌های خام سروکار داریم از خود نشان داده‌اند. باتوجه به این قابلیت یادگیری عمیق، در این پژوهش از شبکه‌های عمیق برای به دست آوردن نمایش مناسب از گزینه‌ها استفاده شده است. به‌طور خاص، یک مدل ترکیبی از یادگیری عمیق در کنار تجزیه ماتریسی ارائه شده است که یک ارتباط دوطرفه بین ویژگی‌های به‌دست آمده از تجزیه ماتریسی و ویژگی‌های محتوایی استخراج شده با استفاده از یادگیری عمیق درباره گزینه‌ها ایجاد می‌کند. در این مدل به‌طور همزمان پارامترهای هر دو بخش یادگرفته می‌شود. مقایسه مدل پیشنهادی با برترین روش‌های ارائه شده در سال‌های اخیر بر روی مجموعه داده‌های مختلف دنیای واقعی، برتری روش پیشنهادی بر سایر روش‌های ارائه شده را نشان می‌دهد.

**کلمات کلیدی:** سامانه‌های توصیه‌گر، شبکه‌های عمیق، پالایش همکارانه، محتوای گزینه‌ها.

### ۱- مقدمه

از گزینه‌ها، براساس علاقه‌مندی شخصی‌سازی شده آن‌ها نمایش داده می‌شود که گزینه‌های این لیست به کمک سامانه توصیه‌گر انتخاب شده‌اند.

در سامانه‌های ارائه پیشنهاد معمولاً مجموعه‌ای از کاربران، گزینه‌ها و بازخوردهای کاربران نسبت به گزینه‌ها را در اختیار داریم که این بازخوردها می‌توانند به صورت صریح بیان شوند (همانند امتیازی که کاربران به گزینه‌ها می‌دهند) و یا به صورت ضمنی از فعالیت‌های کاربر در سیستم، استنباط شوند. معمولاً بازخوردهای کاربران در ماتریسی به نام ماتریس امتیازدهی ذخیره می‌شود که این ماتریس عموماً بسیار تنک است [۳]. تنک بودن ماتریس امتیازات یکی از بزرگترین چالش‌های سامانه‌های توصیه‌گری است که از روش‌های تجزیه ماتریسی برای ارائه پیشنهاد استفاده می‌کنند. از این رو از روش‌های مختلفی جهت بهبود عملکرد رویکرد تجزیه ماتریسی استفاده می‌شود که از جمله آن می‌توان به استفاده از اطلاعات اضافی موجود همانند محتوای گزینه‌ها، ارتباط کاربران با یکدیگر، نمایه کاربران و یا ارتباط بین گزینه‌ها اشاره کرد. اما با توجه به این که اطلاعات اضافه موجود عموماً به صورت متنی هستند، نمایش آن‌ها به کمک روش‌هایی همانند

امروزه حجم اطلاعاتی که در دسترس ما قرار دارد به صورتی فزاینده افزایش یافته است به طوری که، ما در اقیانوسی از داده‌ها غرق شده‌ایم اما همچنان تشنه‌ی دانش هستیم [۱]. با توجه به این که افراد به روش‌های گوناگون و به سادگی می‌توانند داده‌ها را در اینترنت قرار دهند، حجم داده‌های موجود بسیار زیاد شده است. یکی از راه‌های برخورد با مشکل سربار اطلاعات استفاده از سامانه‌های توصیه‌گر است. سامانه‌های توصیه‌گر برنامه‌هایی هستند که گزینه‌هایی را که احتمال می‌دهند مورد علاقه کاربر باشد به او پیشنهاد می‌دهند [۲]. این پیشنهاد می‌تواند دربرگیرنده‌ی انواع تصمیم‌گیری‌ها همانند انتخاب کالا، فیلم و یا کتاب از بین انبوهی از گزینه‌ها باشد. این سامانه‌ها به صورتی گسترده در بسیاری از وبسایت‌ها (همانند آمازون<sup>۱</sup>، اخبار گوگل<sup>۲</sup>، نتفلیکس<sup>۳</sup>) جهت ارائه پیشنهاد به کاربران استفاده می‌شوند. معمولاً در این وبسایت‌ها به کاربران لیستی مرتب شده

• بررسی نتایج عملی بر روی سه مجموعه داده از دنیای واقعی نشان می‌دهد که روش پیشنهادی بهتر از سایر روش‌های ارائه پیشنهاد عمل می‌کند و توانسته به نتایج بهتری دست یابد.

در ادامه ابتدا در بخش ۲، تعدادی از مهمترین کارهای پیشین در زمینه سامانه‌های توصیه‌گر که از شبکه‌های عمیق در ساختار خود استفاده کرده‌اند معرفی می‌شوند. سپس مدل پیشنهادی به تفصیل ارائه می‌شود. در بخش ارزیابی مدل پیشنهادی، ابتدا مجموعه داده‌های مورد استفاده و روش‌های مقایسه معرفی می‌شوند و پس از آن نتایج عملی روش پیشنهادی گزارش می‌شود. در نهایت نتیجه‌گیری و کارهای آتی ارائه خواهد شد.

## ۲- کارهای پیشین

سامانه‌های توصیه‌گر معمولاً براساس روشی که جهت انتخاب گزینه‌ها برای پیشنهاد اتخاذ می‌نمایند در سه دسته محتوا-محور، پالایش همکارانه و روش‌های ترکیبی قرار می‌گیرند. رویکرد محتوا-محور از اطلاعات اضافه موجود در نمایه کاربران و توضیحات موجود درباره گزینه‌ها استفاده می‌نماید [۱۰]. رویکرد پالایش همکارانه از فعالیت‌های گذشته کاربران با سیستم همانند امتیازاتی که به گزینه‌ها داده‌اند بهره می‌گیرد و از اطلاعاتی همانند محتوای گزینه‌ها استفاده نمی‌کند [۱۱]. هر کدام از روش‌های محتوا-محور و پالایش همکارانه برای ارائه پیشنهاد از منابع اطلاعاتی متفاوتی استفاده می‌کنند. به همین دلیل، محدودیت‌هایی دارند که سبب می‌شود نتوانند در بعضی حالت‌ها پیشنهاد‌های مناسبی را ارائه کنند. برای جلوگیری از بسیاری از این محدودیت‌ها و افزایش کیفیت پیشنهادها روش‌های ترکیبی ارائه شدند که از ترکیب روش‌های مبتنی بر محتوا و پالایش همکارانه به دست می‌آیند [۲].

روش‌های ترکیبی در سامانه‌های توصیه‌گر با توجه به این که یک ارتباط دوطرفه بین اطلاعات موجود در ماتریس امتیازات و اطلاعات اضافی موجود درباره گزینه‌ها و کاربران در نظر گرفته‌اند یا نه، در دو دسته ارتباط محکم<sup>۱</sup> و ارتباط ضعیف<sup>۲</sup> قرار می‌گیرند [۷]. در روش‌های ترکیبی ارتباط ضعیف، اطلاعات اضافه تنها یک بار مورد بررسی قرار می‌گیرند و ویژگی‌های آن‌ها استخراج می‌گردد و سپس از این ویژگی‌ها در روش‌های پالایش همکارانه استفاده می‌شود. چون ارتباط در این حالت یک طرفه است اطلاعات موجود در جدول امتیازات نمی‌توانند در جهت انتخاب بهتر ویژگی‌ها کمکی بکنند. اما در روش‌های ترکیبی ارتباط محکم، یک ارتباط دوطرفه بین انتخاب ویژگی‌ها از اطلاعات اضافی و جدول امتیازات وجود دارد. این ارتباط دوطرفه از یک سو با استفاده از جدول امتیازات سبب انتخاب ویژگی‌های بهتر از اطلاعات اضافی می‌گردد و از سوی دیگر با استخراج ویژگی‌های بهتر سبب افزایش دقت مدل می‌گردد.

CTR<sup>۱</sup> [۷] یک روش ترکیبی ارتباط محکم است که جهت پیشنهاد مقالات به کاربران ارائه شده است. این روش یک مدل گرافی احتمالی است که از ترکیب مدل عنوانی LDA<sup>۱</sup> [۱۳] و روش تجزیه ماتریسی [۱۴] به دست می‌آید.

روش CTR بر مبنای فرض اساسی موجود در رویکرد تجزیه ماتریسی بنا شده است که طبق آن فرض ویژگی‌های کاربران و گزینه‌ها را می‌توان به یک فضای برداری نهان با ابعاد کوچکتر نگاهت کرد به صورتی که، میزان علاقه یک کاربر به یک گزینه از ضرب داخلی این دو بردار در آن فضا به دست آید.

CTR کاربران را با استفاده از عنوان‌های مورد علاقه آن‌ها که با استفاده از روش LDA و از متن مقالات مورد علاقه آن‌ها به دست آمده است نمایش می‌دهد و فرض می‌کند مقالات با استفاده از یک مدل موضوعی تولید شده‌اند. اما جهت مدل کردن مقالات فرض می‌شود که در بازنمایی هر مقاله علاوه بر موضوع‌های آن که با استفاده از مدل‌های موضوعی به دست می‌آید، به امتیاز کاربران به آن مقاله

TF-IDF انجام می‌پذیرد که این دست از روش‌ها با مشکل ابعاد بالا و تنگ بودن نمایش به دست آمده از داده‌ها مواجه می‌شوند. در سال‌های اخیر و با توجه به پیشرفت چشمگیر در سرعت پردازنده‌ها و استفاده از پردازنده‌های گرافیکی رویکرد شبکه‌های عمیق بسیار مورد توجه قرار گرفته است [۴] و نشان داده شده است این رویکرد به خصوص در مواقعی که با داده‌های خام سروکار داریم توانایی بالایی در یادگیری بازنمایی مناسب داده‌ها دارد و توانسته پیشرفت‌های چشمگیری در زمینه‌هایی نظیر بینایی کامپیوتر [۵]، پردازش زبان طبیعی و بازشناسی گفتار [۶] حاصل نماید.

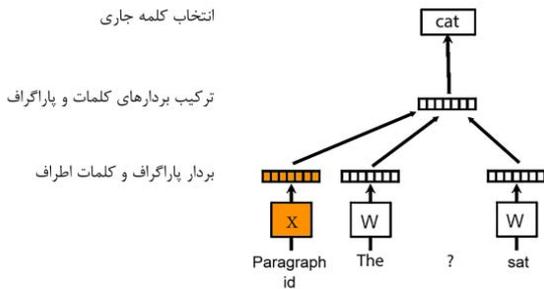
در مدل‌های یادگیری عمیق، ویژگی‌ها می‌توانند به کمک روش‌های ناظر<sup>۴</sup> و بدون ناظر به دست آیند. گرچه این روش‌ها در یادگیری خودکار ویژگی‌ها بیشتر از روش‌های کم عمق مورد توجه قرار دارند، اما نمی‌توانند همانند مدل پالایش همکارانه شباهت و ارتباط‌های ضمنی بین گزینه‌ها را به دست آورند [۷]. این امر نشان می‌دهد می‌توان از ترکیب روش‌های یادگیری عمیق و پالایش همکارانه در بهبود عملکرد سامانه‌های توصیه‌گر بهره برد.

در دست داشتن ارتباط‌های موجود بین گزینه‌های سیستم یکی دیگر از مواردی است که در پیش‌بینی امتیازات کاربران به گزینه‌ها می‌تواند بسیار مفید واقع شود. با استفاده از ارتباط‌های بین گزینه‌ها می‌توان گزینه‌های شبیه به یکدیگر را پیدا کرد که این امر به خصوص در حالت شروع سرد<sup>۵</sup> برای گزینه‌ها می‌تواند بسیار مفید باشد، چرا که در این حالت امتیازی برای گزینه‌ای که تازه وارد سیستم شده است ثبت نشده است و سیستم نمی‌تواند با استفاده از ماتریس امتیازات گزینه‌های شبیه به یکدیگر را بیابد.

در این مقاله، یک مدل ترکیبی ارائه می‌شود که از ماتریس امتیازات، ماتریس ارتباطات بین گزینه‌ها و اطلاعات موجود در مورد محتوای گزینه‌ها در کنار هم، جهت ارائه یک پیشنهاد به کاربر استفاده می‌کند. با توجه به مشکلات ذکر شده برای روش‌هایی نظیر TF-IDF، در روش پیشنهادی از مدل تبدیل پاراگراف به بردار [۸] جهت به دست آوردن نمایش مناسب برای گزینه‌ها استفاده شده است. در واقع شبکه‌ای مشابه skip-gram که در [۹] معرفی شده است جهت استخراج ویژگی‌های مناسب از محتوای گزینه‌ها مورد استفاده قرار می‌گیرد. در مدل ارائه شده با ایده گرفتن از روش مطرح شده در [۷]، یک ارتباط دوطرفه محکم بین ویژگی‌های استخراج شده از محتوای گزینه‌ها و ماتریس امتیازات ایجاد می‌گردد به صورتی که علاوه بر این که ویژگی‌های استخراج شده از محتوای گزینه‌ها با استفاده از اطلاعات اضافه موجود به پیش‌بینی بهتر امتیازات نامشخص کمک می‌کنند، ماتریس امتیازات نیز به شبکه استخراج ویژگی‌ها از محتوا کمک می‌کند تا ویژگی‌های مهمتری را از اطلاعات اضافه موجود استخراج نماید. به این ترتیب ویژگی‌های مهمتری از گزینه‌ها که دقت مدل نهایی را افزایش می‌دهد استخراج می‌گردد.

از مهمترین نوآوری‌های این پژوهش می‌توان به موارد زیر اشاره کرد:

- در این مقاله یک مدل ترکیبی ارائه شده است که از سه منبع متفاوت اطلاعاتی (ماتریس امتیازات، ماتریس ارتباطات بین گزینه‌ها و اطلاعات محتوای گزینه‌ها) در کنار یکدیگر جهت ارائه پیشنهاد استفاده می‌کند و یک ارتباط محکم بین تمام اجزای سیستم با یکدیگر برقرار می‌نماید.
- ارائه یک مدل جهت استفاده از ارتباطات بین گزینه‌ها به منظور ارائه پیشنهاد به کاربر.
- استفاده از یک مرحله پیش‌پردازش داده‌ها با استفاده از روش ارائه شده در [۸] که منجر به پیدا کردن یک فضای معنایی مناسب با ابعاد بسیار کمتر و متراکم‌تر جهت نمایش گزینه‌ها می‌شود. استفاده از این نحوه نمایش برای محتوای گزینه‌ها سبب کاهش بعد داده‌های ورودی می‌گردد که در نهایت سبب افزایش دقت و سرعت یادگیری در مدل نهایی می‌شود.



شکل ۱- ساختار یادگیری بردار پاراگراف. در این مثال از ترکیب بردار یک کلمه قبل و بعد از کلمه جاری به همراه بردار پاراگراف اطراف جهت پیش‌بینی کلمه جاری استفاده شده

### ۳- روش‌های پیش‌نیاز

در این بخش به‌طور مختصر روش‌های پیش‌نیاز مورد استفاده در روش پیشنهادی معرفی می‌شوند. ابتدا روش تبدیل پاراگراف به بردار [۸] معرفی می‌گردد که با استفاده از آن نمایش گزینه‌ها به دست می‌آید. سپس شبکه SDAE<sup>[۱۶]</sup> که جهت استخراج ویژگی‌ها مورد استفاده قرار گرفته است بررسی می‌شود.

#### ۳-۱- روش تبدیل پاراگراف به بردار

یکی از مهمترین بخش‌ها در طراحی سامانه‌های توصیه‌گری که از محتوای اطلاعات گزینه‌ها نیز استفاده می‌کنند، نحوه‌ی به دست آوردن بازنمایی گزینه‌ها با استفاده از متن خام موجود درباره آن گزینه، است. سامانه‌های توصیه‌گر عموماً نیاز دارند که متن مرتبط با هر گزینه ابتدا به یک بردار با طول ثابت و مشخص نگاشت شود تا بتوانند از این اطلاعات استفاده نمایند و پس از این نگاشت هر گزینه به کمک بردار خود شناخته شود. متداول‌ترین روش مورد استفاده جهت به دست آوردن بردار با طول مشخص، استفاده از روش کیسه کلمات [۱۷] است که به علت سادگی و کارایی به نسبت خوب، بسیار مورد توجه قرار گرفته است. اما این روش چندین مشکل دارد از جمله این که:

- در این روش ترتیب کلمات از دست می‌رود، در نتیجه دو متن تا زمانی که کلمات یکسانی داشته باشند بدون توجه به ترتیب کلمات مورد استفاده نمایش یکسانی نیز خواهند داشت.
- این روش تقریباً هیچ دیدی نسبت به معنای کلمات یا به‌طور دقیق‌تر فاصله معنایی کلمات از یکدیگر ندارد.
- بازنمایی به دست آمده برای هر کلمه در این روش بسیار تنک و دارای بعد بسیار بالایی است چرا که در این روش هر کلمه به کمک یک بردار S بعدی (که S نشان‌دهنده تعداد کلمات موجود در دایره واژگان مورد استفاده است) نمایش داده می‌شود که تنها یک عضو این بردار برابر یک است و سایر اعضا صفر هستند. این امر سبب می‌شود روش‌هایی که از روش کیسه کلمات جهت به دست آوردن بردار متن خود استفاده می‌کنند با مشکل ابعاد بالا مواجه شوند.

در مدل پیشنهادی از روش تبدیل پاراگراف به بردار [۸] جهت به‌دست آوردن بازنمایی گزینه‌ها استفاده شده است. بازنمایی به دست آمده از این طریق بسیار مورد توجه است چرا که این بازنمایی بسیاری از ساختارهای زبانی و الگوهای کلمات را حفظ کرده است به‌طوری که بسیاری از این الگوها را می‌توان به صورت ترکیب خطی آن‌ها نمایش داد [۱۸]. به عنوان مثال اگر نمایش به دست آمده از

نیز توجه می‌شود. به عبارت دیگر اگر بردار  $v_j$  بازنمایی گزینه  $j$  در فضای نهان باشد آن‌گاه  $v_j = \theta_j + \epsilon_j$  است که در این نمایش  $\theta_j$  موضوع‌های استخراج شده از مدل‌های موضوعی برای آن مقاله را نمایش می‌دهد و  $\epsilon_j$  با استفاده از تجزیه ماتریس امتیازات کاربران به گزینه‌ها به دست آمده است. یکی از مهمترین مزیت‌های CTR نسبت به روش‌های قدیمی‌تر (علاوه بر دقت بالاتر) این است که این مدل می‌تواند گزینه‌هایی که تازه وارد سیستم شده‌اند را نیز با استفاده از محتوای آن‌ها پیشنهاد نماید. در نتیجه مشکل شروع سرد برای گزینه‌ها برطرف می‌گردد.

علی‌رغم توانایی بالای یادگیری عمیق در به دست آوردن ویژگی‌های مناسب، متأسفانه تا همین اواخر پژوهش‌های اندکی در زمینه استفاده از شبکه‌های عمیق در سامانه‌های توصیه‌گر انجام شده است. CDL<sup>[۱۹]</sup> یکی از موفق‌ترین روش‌های ارائه شده در این زمینه می‌باشد. این روش همانند CTR هم از اطلاعات اضافه موجود درباره گزینه‌ها و هم از ماتریس امتیازات کاربران به گزینه‌ها جهت ارائه پیشنهاد استفاده می‌نماید.

این روش از دو بخش کلی تشکیل شده است: (۱) استخراج ویژگی‌ها و (۲) تجزیه ماتریس امتیازات. روش CDL جهت به دست آوردن بردار ویژگی‌های نهان کاربران از روش تجزیه ماتریس امتیازات استفاده می‌کند، اما این فرض را اضافه می‌کند که ویژگی‌های نهان گزینه‌ها علاوه بر مقادیر به دست آمده از تجزیه ماتریس امتیازات به محتوای گزینه‌ها نیز وابسته است. در نتیجه، ابتدا ویژگی‌های محتوایی گزینه‌ها را با استفاده از یک شبکه خودرمزگذار<sup>[۱۱]</sup> از روی توصیف موجود درباره هر گزینه به دست می‌آورد و سپس در آموزش مدل تلاش می‌کند ویژگی‌های نهان به دست آمده برای هر گزینه به صورتی یادگرفته شوند که علاوه بر تأثیر پذیرفتن از تجزیه ماتریس امتیازات به ویژگی‌های محتوایی به دست آمده از شبکه خودرمزگذار نیز نزدیک باشند. در این روش هم به دلیل استفاده از محتوای گزینه‌ها مشکل شروع سرد برای گزینه‌ها وجود ندارد اما همچنان مشکل شروع سرد برای کاربران وجود دارد.

MDA-CF<sup>[۱۵]</sup> یکی دیگر از روش‌هایی است که شبکه عمیق را با تجزیه ماتریسی ترکیب کرده است. در این مدل نیز همانند CDL از شبکه عمیق جهت استخراج ویژگی‌های مناسب از اطلاعات اضافی موجود استفاده شده است با این تفاوت که اطلاعات اضافی موجود درباره کاربران نیز مورد استفاده قرار گرفته است. MDA-CF با استفاده از اطلاعات اضافی موجود درباره کاربران این فرض را به مدل CDL اضافه می‌کند که ویژگی‌های نهان کاربران علاوه بر ماتریس امتیازات به محتوای اطلاعات اضافه موجود درباره کاربران نیز وابسته است.

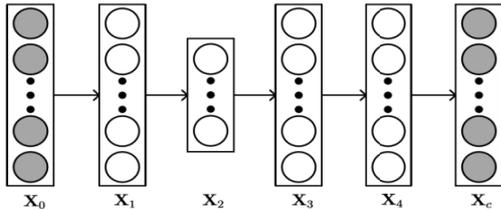
با توجه به این که هر دو مدل CDL و MDA-CF از ماتریس امتیازات در کنار محتوای اطلاعات اضافی موجود جهت ارائه پیشنهاد استفاده می‌نمایند در نتیجه با دقت بالایی قادر به تشخیص امتیازات ناموجود هستند و می‌توانند با مشکل شروع سرد به خوبی مقابله نمایند. در هر سه روش بالا، پس از به دست آوردن بازنمایی کاربران و گزینه‌ها، همانند روش تجزیه ماتریسی پس از ضرب داخلی نمایش‌های به دست آمده در یکدیگر میزان شباهت هر کاربر به هر گزینه محاسبه می‌گردد و گزینه‌هایی که بیشترین شباهت را به هر کاربر دارند به او پیشنهاد می‌گردند.

یکی از مشکلات اساسی در تمام این روش‌ها استفاده از یک بردار با ابعاد بالا و تنک، جهت ورود اطلاعات اضافه به سیستم است. در CDL و MDA-CF از روش TF-IDF جهت به دست آوردن نمایش اطلاعات اضافه موجود استفاده شده است که سبب می‌گردد ورودی با ابعاد بالا و تنک وارد مدل شود و منجر به کاهش دقت مدل و افزایش زمان پردازش داده‌ها گردد. در CTR نیز بازنمایی به دست آمده برای گزینه‌ها به‌خصوص در زمان‌هایی که اطلاعات اضافی تنک هستند، چندان موثر نمی‌باشد، در نتیجه این اطلاعات کمک چندانی به بهبود نتایج نمی‌کنند.

می‌کند با استفاده از این فضا نسخه اصلی داده‌ها را بازسازی نماید و به این صورت ویژگی‌های مهم داده‌های ورودی که همان فضای نهان کوچک است را می‌آموزد. یک نمونه از این شبکه در شکل ۲ آمده است. در نهایت این شبکه تلاش می‌کند مسئله بهینه‌سازی زیر را حل کند:

$$\min_{\{W_1\}, \{b_1\}} \|X_C - X_L\|_2^2 + \lambda \sum_1 \|W_1\|_2^2 \quad (2)$$

که در این نمایش  $X_C$  نسخه اصلی ورودی‌ها و  $X_L$  خروجی لایه آخر شبکه را نمایش می‌دهد.  $\{W_1\}, \{b_1\}$  وزن‌ها و بایاس شبکه را مشخص می‌کنند و  $\lambda$  پارامتر تنظیم است.



شکل ۲- نمایش ساختار شبکه SDAE. در این نمایش  $X_0$  داده نويز وارد شده و  $X_C$  داده بدون نويز را نمایش می‌دهند

## ۴- روش پیشنهادی

در این بخش روش پیشنهادی معرفی می‌گردد. در ابتدا نمادگذاری و تعریف مسئله ارائه می‌شود و سپس مدل پیشنهادی که از ترکیب روش تبدیل پاراگراف به بردار و شبکه SDAE در کنار روش تجزیه ماتریسی به دست می‌آید معرفی خواهد شد.

### ۴-۱- نمادگذاری و تعریف مساله

در این مقاله مجموعه کاربران را با  $I = \{i_1, i_2, \dots, i_N\}$  نمایش می‌دهیم که هر  $i_n$  نشان‌دهنده یک کاربر خاص است و  $N$  تعداد کل کاربران را نشان می‌دهد. مجموعه گزینه‌ها را نیز با  $J = \{j_1, j_2, \dots, j_M\}$  نمایش می‌دهیم که در این مجموعه هر  $j_m$  نشان‌دهنده یک گزینه خاص (مثلاً مقاله و یا فیلم) خواهد بود و  $M$  نیز تعداد کل گزینه‌ها است. به ازای هر گزینه همانند  $m$  یک مجموعه توالی<sup>۱۳</sup> که شامل  $T_m$  لغت است وجود دارد. این مجموعه شامل اطلاعات اضافی است که از هر گزینه در اختیار داریم. این اطلاعات اضافی برای مقالات، متن چکیده آن‌ها و برای فیلم‌ها داستان آن فیلم است که عموماً به صورت یک پاراگراف متن خام است که گزینه مورد نظر را به طور کلی توصیف می‌کند. در مجموعه توالی لغات  $W_{t_m}$  نشان‌دهنده لغت  $t$ -ام در توصیف موجود درباره گزینه  $m$  است. هر لغت به کمک بردار  $S$  بعدی نمایش داده می‌شود که فقط یک عنصر آن برابر ۱ است و باقی عناصر آن صفر هستند در این جا  $S$  تعداد کل کلمات مجزای موجود در دایره لغات است و  $W^S$  نشان‌دهنده لغت  $S$ -am در مجموعه لغات می‌باشد.

در این مقاله جهت ثبت تراکنش‌های کاربر با سیستم از یک ماتریس امتیازات دوتایی<sup>۱۴</sup> استفاده می‌شود و آن را با  $R = [R_{nm}]_{N \times M}$  نمایش می‌دهیم. این ماتریس می‌تواند بازخوردهای مستقیم و غیرمستقیم یک کاربر به یک گزینه را در خود ذخیره نماید. به عنوان مثال در مجموعه دادگان citeulike که شامل تعدادی مقاله است کاربران به گزینه‌ها به طور مستقیم امتیازی نمی‌دهند اما می‌توانند هر مقاله را در کتابخانه مقالات خود ذخیره نمایند.

این طریق برای هر کلمه همانند word را با  $\text{rep}(\text{'word'})$  نمایش دهیم، نتیجه محاسبه برداری  $\text{rep}(\text{'Tehran'}) - \text{rep}(\text{'Iran'}) + \text{rep}(\text{'France'})$  به  $\text{rep}(\text{'Paris'})$  نزدیک‌تر از نمایش سایر کلمات می‌باشد. روش تبدیل پاراگراف به بردار به دنبال به دست آوردن یک نمایش با ابعاد پایین برای هر کلمه و هر پاراگراف است به صورتی که با استفاده از نمایش به دست آمده برای کلمات اطراف هر کلمه و پاراگرافی که آن کلمه در آن قرار دارد بتوان کلمه جاری را پیش‌بینی کرد. این مدل در نهایت به دنبال یافتن یک تابع همانند  $f$  است که با دریافت کلمات اطراف هر کلمه و پاراگراف جاری، کلمه جاری را پیش‌بینی کند. در نتیجه تابع  $f$  را به صورت زیر خواهیم داشت:

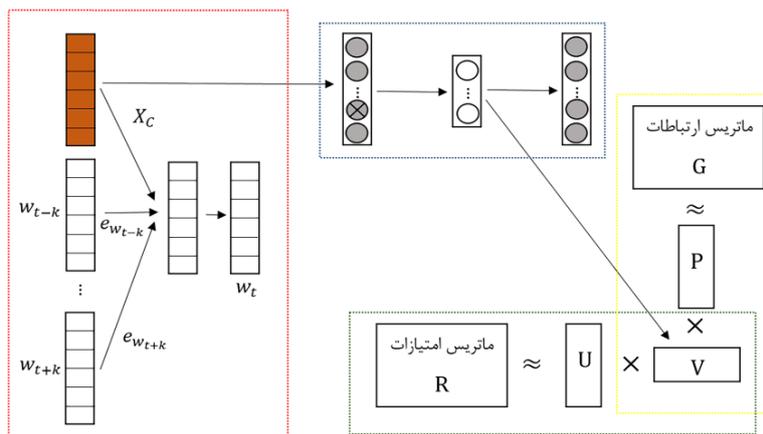
$$f(D_i, w_{t-k}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+k}) = P(w_t = i | \text{context}) \quad (1)$$

که در این رابطه  $w_t$  نشان‌دهنده کلمه  $t$ -ام در دنباله کلمات است،  $D_i$  نمایش پاراگراف جاری را مشخص می‌کند و  $k$  تعداد کلمات اطراف هر کلمه که برای پیش‌بینی کلمه جاری استفاده می‌شود را نشان می‌دهد. در عمل این تابع به کمک یک شبکه جلورو به دست می‌آید و به کمک روش پس‌انتشار خطا مقادیر بهینه برای نمایش کلمات و پاراگراف را به دست می‌آورد. در این روش جهت آموزش شبکه در هر مرحله یک نمونه با طول مشخص از محتوای یک پاراگراف انتخاب می‌شود و سپس با محاسبه خطا، پارامترهای شبکه در کنار ماتریس پاراگراف‌ها  $(D)$  و کلمات  $(w)$  به‌روز می‌شوند. در این روش ماتریس کلمات در تمام پاراگراف‌ها مشترک است و یک کلمه در پاراگراف‌های مختلف بازنمایی یکسانی خواهد داشت و ماتریس پاراگراف نیز در تمام کلمات آن پاراگراف مشترک است.

در مرحله بعد جهت محاسبه بردار یک پاراگراف جدید که تاکنون دیده نشده نیز کافی است یک سطر به ماتریس پاراگراف‌ها یا همان  $D$  افزوده شود. سپس دوباره شبکه با استفاده از این پاراگراف آموزش داده شود ولی تنها مقادیر بردار پاراگراف جدید به‌روز می‌شوند و سایر پارامترهای شبکه ثابت نگاه داشته می‌شوند. یک شمای کلی از این مدل در شکل ۱ آمده است. در نتیجه با استفاده از این روش مدل پیشنهادی قادر است به یک بردار یکتا برای هر کدام از گزینه‌ها دست یابد که ویژگی‌های اصلی هر گزینه را از متن آن گزینه استخراج کرده است و بسیاری از مشکلات روش‌های معمول مورد استفاده را برطرف کرده است. از جمله این که در این روش اطلاعات بسیاری درباره یک پاراگراف همچون ترتیب کلمات حفظ می‌شود و از این ترتیب در ساخت بردار پاراگراف استفاده می‌شود و نکته‌ی مهم‌تر این‌که با استفاده از این روش می‌توان هر گزینه را به کمک یک بردار با تراکم بالا و بعد کم نمایش داد  $[A]$  در نتیجه ورودی سامانه توصیه‌گر حجم کمتری خواهد داشت و پردازش آن سریع‌تر انجام می‌پذیرد.

### ۳-۲- شبکه خودرمزگذار

پس از بازنمایی گزینه‌ها به کمک روش معرفی شده سامانه توصیه‌گر باید ویژگی‌های مناسب برای هر گزینه را از بازنمایی آن‌ها استخراج کند. در این‌جا منظور از "استخراج ویژگی‌های مناسب" به دست آوردن ویژگی‌هایی از گزینه‌ها است که به سیستم کمک می‌کند که سریع‌تر به کارایی بالاتری که ممکن بود نتواند با استفاده از بازنمایی اولیه گزینه‌ها به آن کارایی برسد، دست یابد [۱۶]. در روش پیشنهادی همانند [۷] از شبکه SDAE جهت استخراج ویژگی‌های مناسب از نمایش اولیه گزینه‌ها استفاده می‌شود. SDAE یک شبکه جلورو جهت یادگیری ویژگی‌های مهم داده‌های ورودی است. این شبکه نسخه نويزی شده داده‌ها را در ورودی دریافت می‌کند و پس از نگاشت آن‌ها به یک فضای نهان کوچک‌تر تلاش



شکل ۳- ساختار کلی روش ارائه شده. ورودی‌ها ماتریس امتیازات، ماتریس ارتباطات گزینه‌ها و اطلاعات اضافی موجود درباره گزینه‌ها هستند. روش ارائه شده همزمان ماتریس‌های امتیازات و ارتباطات را تجزیه می‌کند و ویژگی‌های نهان گزینه‌ها را به کمک تجزیه این دو ماتریس در کنار اطلاعات اضافی موجود درباره گزینه‌ها به دست می‌آورد. بخش مشخص شده در حاشیه قرمز رنگ نمایش گزینه‌ها را به دست می‌آورد و سپس به کمک یک خودرمزگذار که در بخش با حاشیه با رنگ آبی مشخص شده ویژگی‌های مهم گزینه‌ها استخراج می‌شود. این ویژگی‌ها در کنار ویژگی‌های به دست آمده از تجزیه ماتریس امتیازات (حاشیه سبز) و ارتباطات (حاشیه زرد) جهت پیش‌بینی امتیازات نامشخص استفاده می‌شوند

وابسته بودن به ماتریس امتیازات به محتوای آن‌ها نیز وابسته است و محتوای گزینه‌ها نیز در ویژگی‌های نهان آن‌ها تاثیرگذار است.

در مدل پیشنهادی از یک شبکه عمیق جهت به دست آوردن بازنمایی گزینه‌ها و از یک خودرمزگذار جهت استخراج ویژگی‌های مهم از بازنمایی به دست آمده از گزینه‌ها استفاده شده است. با توجه به این که ویژگی‌های نهان به دست آمده برای هر گزینه باید با محتوای آن گزینه نیز ارتباط داشته باشد، در نتیجه ویژگی‌های استخراج شده از شبکه خودرمزگذار باید در ویژگی‌های نهان گزینه‌ها تاثیرگذار باشند. برای این منظور مدل تلاش می‌کند ویژگی‌های نهان گزینه‌ها را به نحوی آموزش دهد که به ویژگی‌های استخراج شده از محتوای گزینه‌ها نزدیک باشد. استفاده از مدل پیشنهادی سبب می‌شود برای گزینه‌هایی که امتیازی برای آن‌ها در دسترس نیست بتوان با استفاده از محتوای آن‌ها، ویژگی‌هایشان را به دست آورد و در نتیجه این گزینه‌ها را نیز به کاربران پیشنهاد داد و مشکل شروع سرد برای گزینه‌ها از این طریق برطرف می‌گردد. به علاوه در این مدل یک ارتباط محکم دوطرفه بین ویژگی‌های استخراج شده از گزینه‌ها و ویژگی‌های به دست آمده از ماتریس امتیازات وجود دارد. این ارتباط سبب می‌شود علاوه بر این که ویژگی‌های محتوای گزینه‌ها به پیش‌بینی بهتر امتیازات کمک می‌کند، ویژگی‌های مهم‌تری نیز از محتوای گزینه‌ها استخراج شده و بازنمایی گزینه‌ها نیز به نحوی به دست آید که ویژگی‌های مهم‌تری از گزینه‌ها را در بر داشته باشد و از این طریق دقت مدل را افزایش دهد.

فرض دیگری که در مدل پیشنهادی در نظر گرفته شده است این است که گزینه‌هایی که به طور صریح مشخص شده است که با یکدیگر ارتباط دارند (مثلاً وجود ارجاع یک مقاله به مقاله دیگر) نیز می‌توانند در یافتن ویژگی‌های نهان گزینه‌ها به ما کمک کنند، چرا که با توجه به ارتباط آن‌ها با یکدیگر ویژگی‌های نهان آن‌ها نیز باید به یکدیگر ارتباط داشته باشد. با توجه به این که در نهایت کلمات موجود در اطلاعات اضافی موجود درباره گزینه‌ها ویژگی‌های محتوایی آن گزینه را می‌سازند در برخی حالت‌ها ممکن است دو گزینه کلمات مشترک بسیاری داشته باشند در نتیجه بردار ویژگی‌های آن‌ها به یکدیگر شبیه باشد اما از لحاظ معنایی با یکدیگر ارتباط چندانی نداشته باشند. به عنوان مثال مطرح شده در [۱۲] در خصوص دو مقاله A و B را در نظر بگیرید که هر دو درباره استفاده از یادگیری ماشین در شبکه‌های اجتماعی ارائه شده‌اند.

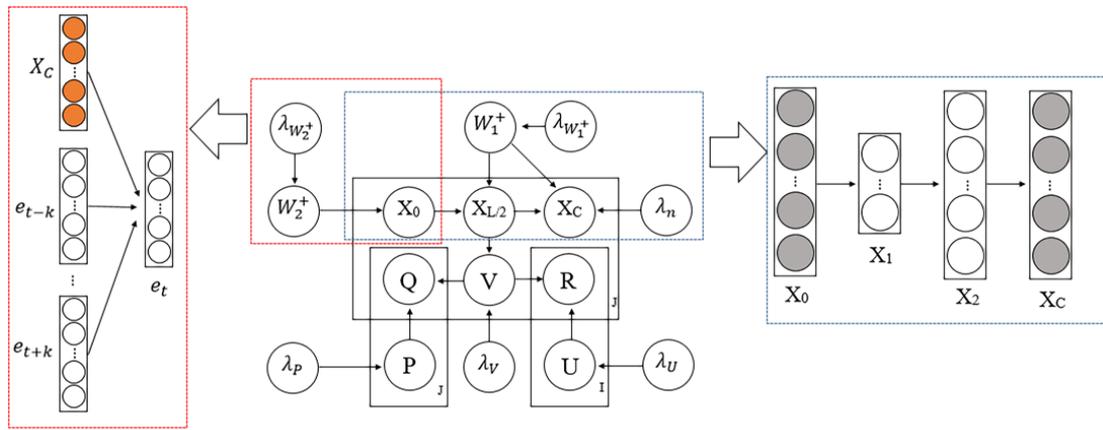
مقدار  $R_{nm} = 1$  است اگر کاربر  $m$  گزینه  $n$  را در کتابخانه شخصی خود ذخیره کرده باشد و در غیر این صورت  $R_{nm} = 0$  خواهد بود. در این‌جا در کنار ماتریس امتیازات و مجموعه توالی لغات، یک ماتریس ارتباطات بین گزینه‌ها هم در نظر می‌گیریم و آن را با  $G = [G_{mm}]_{M \times M}$  نمایش می‌دهیم. این ماتریس نیز یک ماتریس دوتایی است که نشان‌دهنده ارتباطات بین گزینه‌ها است. به عنوان مثال در مجموعه مقالات به‌زای دو مقاله  $m$  و  $m'$  مقدار  $G_{mm'} = 1$  است اگر یکی از این دو مقاله به دیگری ارجاع داده باشد و در غیر این صورت این مقدار برابر صفر خواهد بود.

در نهایت سیستم باید بتواند با دریافت برخی از امتیازات موجود در ماتریس  $R$  به همراه مجموعه توالی لغات  $w_{1m}, w_{2m}, \dots, w_{T_{mm}}$  که همان اطلاعات اضافی موجود درباره گزینه‌ها هستند در کنار ماتریس ارتباطات گزینه‌ها  $G$ ، مقادیر نامشخص  $R$  را پیش‌بینی نماید. خلاصه‌ای از تمام نمادهای استفاده شده در این مقاله در جدول ۱ قابل مشاهده است.

#### ۴-۲- روش پیشنهادی

مدل پیشنهادی یک روش ترکیبی جهت ارائه پیشنهاد است که هم از ماتریس امتیازات و هم از اطلاعات اضافه موجود جهت ارائه پیشنهاد استفاده می‌کند. ایده‌ی کلی این مدل در شکل ۳ آمده است. در مدل ارائه شده همانند روش تجزیه ماتریسی فرض می‌شود که یک فضای نهان برای کاربران و گزینه‌ها وجود دارد که میزان علاقه هر کاربر به هر گزینه از ضرب داخلی بازنمایی آن‌ها در این فضا به دست می‌آید. در نتیجه می‌توان با تجزیه ماتریس امتیازات به بازنمایی کاربران و گزینه‌ها دست یافت.

با توجه به این که در اکثر موارد اطلاعات چندانی درباره کاربران در دسترس نیست و تنها از ارتباطات کاربر با سیستم می‌توان به ویژگی‌های نهان کاربر دست یافت در روش ارائه شده نیز جهت به دست آوردن ویژگی‌های نهان کاربران تنها از ماتریس امتیازات کاربر به گزینه‌ها استفاده می‌شود و این ویژگی‌ها با استفاده از تجزیه ماتریس امتیازات به دست می‌آید. اما با توجه به در دست بودن اطلاعات اضافی درباره گزینه‌ها در سیستم فرض می‌شود که بردار نهان گزینه‌ها علاوه بر



شکل ۴- بخش مشخص شده در مستطیل با حاشیه نقطه‌چین آبی شبکه یک SDAE دو لایه را مشخص می‌کند و بخش مشخص شده در مستطیل با حاشیه نقطه‌چین قرمز نشان‌دهنده شبکه تبدیل پاراگراف به بردار است

تخمینی که از متغیرهای دیگر تاکنون به دست آمده، استفاده می‌شود و این عمل تا رسیدن به یک کمینه محلی برای تابع خطای تعریف شده ادامه پیدا می‌کند. تابع لگاریتم درست‌نمایی در مدل پیشنهادی به صورت زیر تعریف می‌شود:

$$\begin{aligned} \mathcal{L} = & -\frac{\lambda_q}{2} \sum_j \|p_j\|_2^2 - \frac{\lambda_u}{2} \sum_i \|u_i\|_2^2 - \frac{\lambda_w}{2} (\|W^{(2)}\|_2^2 + \|b^{(2)}\|_2^2) \\ & - \frac{\lambda_s}{2} \sum_j \sum_i \|\sigma(X_{i-1,j} W_1^{(2)} + b_1^{(2)}) - X_{i,j}\|_2^2 \\ & - \frac{\lambda_v}{2} \sum_j \|v_j - X_{L2,j}^T\|_2^2 - \frac{\lambda_n}{2} \sum_j \|X_{L2,j} - X_{C,j}\|_2^2 \\ & - \lambda_p \sum_{ij} \frac{B_{ij}}{2} (G_{ij} - v_j p_j^T)^2 - \sum_{ij} \frac{C_{ij}}{2} (R_{ij} - u_i v_j^T)^2 \end{aligned} \quad (3)$$

در تابع هدف (۳)  $\lambda_u, \lambda_w, \lambda_q, \lambda_s, \lambda_e, \lambda_n, \lambda_v, \lambda_p$  پارامترهای منظم‌ساز مدل هستند. خروجی لایه میانی شبکه SDAE را نمایش می‌دهد که نشان‌دهنده ویژگی‌های محتوایی به دست آمده از گزینه  $z$  است. در این تابع هدف  $X_{0,z}$  برابر  $A_j \times D$  است که  $A_j$  نیز یک بردار  $M$  بعدی می‌باشد که فقط عنصر  $z$  آن برابر یک است و سایر مقادیر آن صفر هستند در نتیجه دقیقاً برابر با سطر  $z$ -ام ماتریس  $D$  که همان بازنمایی برداری گزینه  $z$  است خواهد بود. باتوجه به این که ماتریس امتیازات به صورت غیرمستقیم از فعالیت‌های کاربر در سیستم به دست آمده است.

در نتیجه از مقدار صفر در ماتریس امتیازات می‌توان به دو صورت برداشت کرد، یا کاربر به این گزینه علاقه‌ای ندارد و یا کاربر این گزینه را تا به حال ندیده است و مشخص نیست که به این گزینه علاقه دارد یا نه. در نتیجه در روش پیشنهاد شده همانند [۱۲] و [۷] از پارامتر اطمینان  $C_{ij}$  در کنار ماتریس امتیازات استفاده می‌شود که مقدار آن در حالاتی که امتیاز  $R_{ij} = 1$  است بیشتر از مواقعی است که  $R_{ij} = 0$  است. ماتریس  $B$  نیز جهت سنجش مقدار ارتباط بین دو گزینه استفاده شده است و مقادیر آن با توجه به قدرت ارتباط بین دو گزینه تنظیم می‌شود. در این تابع هدف،  $X_{0,z}$  نسخه نویزی شده نمایش برداری گزینه  $z$  است که از روش تبدیل پاراگراف به بردار به دست آمده است. جمله چهارم و ششم در تابع هدف معرفی شده معادل تابع هدف شبکه SDAE است.

در نتیجه متن این مقالات بسیار به یکدیگر شبیه است و در نتیجه احتمالاً ویژگی‌های استخراج شده از متن آن‌ها نیز بسیار به یکدیگر شبیه خواهد بود. حال فرض کنید مقاله  $A$  یک الگوریتم یادگیری ماشین را ارائه می‌کند که در یک کاربرد از شبکه‌های اجتماعی مورد استفاده قرار می‌گیرد و مقاله  $B$  از روش‌های استاندارد معرفی شده در یادگیری ماشین جهت بررسی و تحلیل یک ویژگی خاص در شبکه‌های اجتماعی استفاده کرده است. کاربرانی که در زمینه یادگیری ماشین فعالیت می‌کنند مقاله  $A$  را ترجیح می‌دهد و کاربرانی که در زمینه شبکه‌های اجتماعی فعالیت می‌کنند مقاله  $B$  را ترجیح خواهند داد. اما با توجه به محتوا دو مقاله شبیه یکدیگر هستند و به هر دو دسته کاربر پیشنهاد می‌گردند. اما در صورتی که از ارتباط بین مقالات نیز استفاده گردد با توجه به ارجاع‌های مورد استفاده در مقاله این دو مقاله با توجه به این که در زمینه‌های متفاوتی قرار دارند ویژگی‌های به دست آمده برای آن‌ها نیز متفاوت خواهد بود و در نتیجه به کاربران متفاوتی پیشنهاد خواهند شد [۱۲].

علاوه بر این موارد، استفاده از ارتباطات بین گزینه‌ها برای گزینه‌هایی که به تازگی وارد سیستم شده‌اند می‌تواند بسیار مفید باشد. با توجه به این که استفاده از روش تجزیه ماتریس امتیازات در این حالت نمی‌تواند کمکی به پیدا کردن گزینه‌های مشابه کند و استفاده از محتوا نیز در برخی موارد مشکل اشاره شده در به دست آوردن دسته گزینه را دارد در نتیجه می‌توان از ارتباطات بین گزینه‌ها نیز جهت پیدا کردن گزینه‌های مشابه و پیشنهاد آن به کاربران استفاده کرد. ویژگی مهمی که سبب تمایز روش پیشنهادی نسبت به سایر روش‌ها می‌گردد این است که تمام بخش‌های مدل پیشنهادی با یکدیگر یک ارتباط محکم دارند. این امر سبب می‌گردد که ویژگی‌های به دست آمده در هر بخش بر روی ویژگی‌های نهان به دست آمده در قسمت‌های دیگر تاثیر بگذارد. بنابراین، در صورتی که در یکی از بخش‌ها اطلاعات چندانی وجود نداشته باشد سایر بخش‌ها به مدل کمک می‌کنند که از آن بخش نیز مهمترین ویژگی‌ها استخراج گردد تا مدل در نهایت به پیش‌بینی بهتری از ماتریس امتیازات برسد. نمایش مدل گرافکی روش پیشنهادی در شکل ۴ قابل مشاهده است. ویژگی‌های نهان نهایی به دست آمده برای گزینه‌ها تاثیر مستقیم می‌گذارند.

### ۴-۳- تخمین مقادیر متغیرها

در مدل پیشنهادی جهت به دست آوردن مقادیر بهینه متغیرها از روش تخمین بیشینه کردن امید استفاده می‌کنیم. به این صورت که در به‌روزرسانی هر متغیر

جدول ۱- خلاصه نمادها

نماد	توصیف
N	تعداد کاربران
M	تعداد گزینه‌ها
d	تعداد ویژگی‌های پنهان گزینه‌ها و کاربران
H	تعداد ویژگی‌های نمایش اولیه گزینه‌ها
T <sub>m</sub>	تعداد لغات مورد استفاده در توصیف گزینه m
k	تعداد لغات مجاور مورد استفاده در تبدیل کلمه به بردار
S	تعداد کل لغات موجود در دایره لغات
$W_{tm} \in \mathbb{R}^{1 \times S}$	نمایش لغت t-ام در دنباله لغات توصیف گزینه m
$W^t \in \mathbb{R}^{1 \times S}$	نمایش لغت t-ام در مجموعه تمام کلمات موجود در دایره واژگان
$e^t \in \mathbb{R}^{1 \times H}$	بردار ویژگی‌های کلمه t-ام در دایره واژگان
$e_w \in \mathbb{R}^{1 \times H}$	بردار ویژگی‌های کلمه w
$E \in \mathbb{R}^{S \times H}$	ماتریس نمایش تمام بردارهای کلمات موجود در دایره واژگان
$D \in \mathbb{R}^{M \times H}$	ماتریس نمایش تمام بردارهای گزینه‌های موجود در سامانه
$G \in \mathbb{R}^{M \times M}$	ماتریس ارتباطات بین گزینه‌ها
$X_c^m \in \mathbb{R}^{1 \times H}$	بردار ویژگی‌های اولیه به دست آمده از طریق شبکه تبدیل پاراگراف به بردار برای گزینه m
$R \in \mathbb{R}^{M \times N}$	ماتریس امتیازات
$U \in \mathbb{R}^{N \times d}$	ویژگی‌های پنهان کاربران
$V \in \mathbb{R}^{M \times d}$	ویژگی‌های پنهان گزینه‌ها
$P \in \mathbb{R}^{d \times M}$	ویژگی‌های پنهان ارتباطات گزینه‌ها
$W^{(1)} \in \mathbb{R}^{H \times S}$	ماتریس وزن‌های شبکه تبدیل پاراگراف به بردار
$b^{(1)} \in \mathbb{R}^{1 \times S}$	بردار بایاس شبکه تبدیل پاراگراف به بردار
$K_1$	تعداد گرهای لایه ۱-ام شبکه SDAE
$W_1^{(2)} \in \mathbb{R}^{K_1-1 \times K_1}$	ماتریس وزن‌های لایه ۱-ام شبکه SDAE
$b_1^{(2)} \in \mathbb{R}^{1 \times K_1}$	بردار بایاس لایه ۱-ام شبکه SDAE

$$U = U - \alpha[(C \odot (U \times V^T - R)) \times V + \lambda_u U] \quad (۴)$$

$$P = P - \alpha[(B \odot (V \times P^T - G)) \times V + \lambda_p P] \quad (۵)$$

$$V = V - \alpha[(C \odot (U \times V^T - R))^T \times U + \lambda_p (B \odot (V \times P^T - G))^T \times P + \lambda_v (V - X_{Lj}^T)] \quad (۶)$$

در این روابط  $\odot$  نشان‌دهنده ضرب عنصر به عنصر مقادیر دو ماتریس است و  $\times$  ضرب دو ماتریس را نمایش می‌دهد.  $\alpha$  نیز پارامتر گام یادگیری است. با در دست داشتن مقادیر به‌روز شده برای ماتریس‌های  $U, V$  و  $P$  سایر پارامترهای مدل تنها وزن‌های شبکه SDAE هستند که آن مقادیر را می‌توان به کمک روش پس‌انتشار خطا به‌روزرسانی کرد. در نتیجه با چندین بار به‌روزرسانی پی‌درپی پارامترهای مدل می‌توان به یک جواب بهینه محلی برای تابع خطای تعریف شده در معادله (۳) دست یافت.

#### الگوریتم ۱- الگوریتم پیش‌پردازش داده‌های ورودی

- ورودی:** اطلاعات اضافه موجود درباره گزینه‌ها (دنباله کلمات توصیف گزینه  $W_{1,m}, \dots, W_{T,m}$ )
- خروجی:** مقادیر بهینه اولیه برای وزن‌های شبکه SDAE و شبکه تبدیل پاراگراف به بردار.
- تا رسیدن به خطای اعتبار سنجی کمتر از  $\epsilon$
  - برای تمام گزینه‌های  $j = 1, \dots, M$
  - برای تمام کلمات موجود در دنباله کلمات، بردار احتمال این کلمه را با استفاده از شبکه تبدیل پاراگراف به بردار محاسبه کن.
  - با استفاده از خطای به دست آمده در گام قبل و روش پس انتشار خطا پارامترهای شبکه  $(W^{(1)}, b^{(1)}, D, E)$  را به‌روز کن.
  - مقدار  $X_{Cj}$  را برابر سطر  $j$  ماتریس  $D$  قرار بده.  $(X_{Cj} \leftarrow D[j,:])$
  - یک نویز گوسی با واریانس  $\lambda_j$  به  $X_{Cj}$  اضافه کن و حاصل را در  $X_{0j}$  قرار بده.  $(X_{Cj} \leftarrow D[j,:])$
  - مقدار  $X_{Lj}$  را با استفاده از شبکه SDAE محاسبه کن و خطای شبکه را برابر  $X_{Cj} - X_{Lj}$  قرار بده.
  - با استفاده از مجذور خطای به دست آمده در مرحله قبل و روش پس انتشار خطا پارامترهای شبکه  $(W^{(2)}, b^{(2)})$  را به‌روزرسانی کن.

الگوریتم نهایی جهت به دست آوردن مقادیر بهینه برای  $U, V$  و  $P$  از دو قسمت تشکیل شده است. در مرحله اول مقادیر اولیه بهینه برای وزن‌های شبکه SDAE و شبکه تبدیل پاراگراف به بردار محاسبه می‌گردند (که در الگوریتم ۱ آمده است). مدل تلاش می‌کند با استفاده کلمات موجود در توصیف یک گزینه، بردار ویژگی‌های آن گزینه را با استفاده از شبکه تبدیل پاراگراف به بردار محاسبه نماید و سپس از این بردار به عنوان ورودی اولیه شبکه SDAE استفاده می‌نماید و با تکرار این فرایند برای تمام گزینه‌ها وزن‌های اولیه این دو شبکه را محاسبه می‌نماید. این مرحله پیش‌پردازش داده‌ها کمک می‌کند تا برای مدل نهایی سریع‌تر و با دقت بالاتری مقادیر بهینه پیدا شود. در مرحله بعد با استفاده از مقادیر به دست آمده برای بردار توصیف گزینه‌ها روش تلاش می‌کند مقادیر بهینه برای پارامترهای نهایی مدل را محاسبه نماید. فرایند اجرای این مرحله در الگوریتم ۲ آمده است.

#### ۴-۴- پیش‌بینی امتیازات

پس از این که مقادیر بهینه محلی برای پارامترهای مدل محاسبه شدند، اگر داده‌های مشاهده شده را با  $D$  نمایش دهیم امتیازات نامشخص را می‌توان به کمک رابطه زیر محاسبه کرد:

مشابه روش CDL [۷] با استفاده از این جمله، شبکه تلاش می‌کند ویژگی‌هایی را از ورودی به دست آورد که به کمک آن‌ها بتوان نسخه اصلی ورودی را به کمک نسخه نویزی شده آن ساخت. در نتیجه به کارگیری شبکه SDAE سبب انتخاب ویژگی‌های مهم گزینه‌ها خواهد شد. جمله پنجم نیز معادل تابع خطای یک شبکه جلورو است که مقدار  $jz$  را به عنوان هدف در نظر گرفته و  $X_{Cj}$  ورودی آن شبکه است. با الهام از روش CDL، این جمله به‌عنوان پل بین محتوای گزینه‌ها و ماتریس امتیازات به کار گرفته شده است. این لایه میانی در کنار بردار ویژگی‌های به دست آمده از ماتریس امتیازات قرار می‌گیرد و یک ارتباط دوطرفه بین ویژگی‌های محتوایی و ویژگی‌های استخراج شده از ماتریس امتیازات ایجاد می‌کند که از یک طرف سبب انتخاب بهتر ویژگی‌های نهان گزینه‌ها می‌گردد و از طرف دیگر این انتخاب ویژگی‌های مناسب سبب افزایش دقت مدل در پیش‌بینی عناصر خالی امتیازات خواهد شد.

از جمله مهمترین تفاوت‌های روش پیشنهادی و روش CDL استفاده از ارتباطات بین گزینه‌ها با یکدیگر است که در جمله هفتم تابع هدف معرفی شده مورد استفاده قرار گرفته است و سبب نزدیک شدن نمایش گزینه‌های مرتبط با یکدیگر می‌گردد و تفاوت دیگر استفاده از روش تبدیل پاراگراف به بردار جهت به دست آوردن نمایش اولیه گزینه‌ها است که سبب کاهش بعد داده‌های ورودی و در نتیجه کاهش تعداد پارامترهای آزاد می‌شود و از این طریق باعث جلوگیری از بیش‌برازش و همچنین کاهش زمان رسیدن به حالت پایدار می‌گردد. جهت به دست آوردن مقادیر جدید بردارهای  $U_i, V_j$  و  $P_j$  نیز در هر مرحله با در نظر گرفتن مقادیر جاری برای وزن‌های شبکه، گرادینان  $L$  را نسبت به بردارهای  $U_i, V_j$  و  $P_j$  محاسبه کرده و مقدار آن را برابر صفر قرار می‌دهیم و در نتیجه به قوانین به‌روزرسانی زیر می‌رسیم:

عوامل فیلم استفاده شده است، به این صورت که بین فیلم‌هایی که کارگردان مشترک دارند یک ارتباط با قدرت ۱۰ و به ازای هر بازیگر مشترک ارتباط با قدرت ۱ در نظر گرفته شده است.

جدول ۲- خلاصه اطلاعات آماری مجموعه داده‌های مورد استفاده

MovieLens	citeulike-T	citeulike-A	مجموعه داده
۲.۱۰۳	۷.۹۴۷	۵.۵۵۱	تعداد کاربران
۹.۴۲۳	۲۵.۹۷۵	۱۶.۹۸۰	تعداد گزینه‌ها
۱۶۰.۳۳۲	۱۳۴.۸۶۰	۲۰۴.۹۸۷	تعداد امتیازات
٪۹۹.۳	٪۹۹.۹۳	٪۹۹.۸	تنکی ماتریس امتیازات
۱.۴۶۱.۹۷۲	۳۲.۵۶۵	۴۴.۷۰۹	تعداد ارتباط‌های گزینه‌ها
۷۶	۱۷	۳۷	میانگین تعداد امتیازات کاربر
۱۴۳.۴	۱.۳	۲.۶	میانگین ارتباط‌های گزینه‌ها
۲۹۶.۰۰	۲۰.۰۰۰	۸.۰۰۰	تعداد تگ
۱۵۷	۶۴	۶۷	میانگین تگ هر گزینه

خلاصه‌ای از اطلاعات آماری این مجموعه‌های داده در جدول ۲ آمده است. با توجه به این که روش‌های مورد نظر برای مقایسه با روش پیشنهادی از تگ برای نمایش گزینه‌ها استفاده می‌کنند، خلاصه‌ای از اطلاعات آماری تگ‌ها نیز در این جدول آمده است.

علاوه بر مجموعه داده‌های معرفی شده جهت بهبود عملکرد روش تبدیل پاراگراف به بردار یک مجموعه داده دیگر نیز از وبسایت IMDB استخراج شد. این مجموعه داده شامل متن تمام نظرات کاربران درباره فیلم‌های موجود در مجموعه داده movielens است که شامل ۱,۱۰۸,۵۷۹ پاراگراف متن است که نظر هر کاربر را درباره یک فیلم بیان می‌کند. این مجموعه داده در مرحله پیش‌آموزش جهت بهبود عملکرد روش تبدیل پاراگراف به بردار مورد استفاده قرار می‌گیرد و سبب می‌شود برای کلمات بردارهای بازنمایی بهتری به دست آید و در نتیجه ارتباط بین کلماتی که از لحاظ معنایی به یکدیگر نزدیک هستند بهتر مشخص گردد.

## ۵-۲- معیار ارزیابی

سامانه‌های توصیه‌گر را می‌توان از جنبه‌های مختلفی همانند کارایی، دقت و مقیاس‌پذیری الگوریتم ارائه شده مورد بررسی قرار داد. از این بین، دقت مهمترین ویژگی‌ای است که سامانه‌های توصیه‌گر را با توجه به آن با یکدیگر مقایسه می‌کنند، چرا که هدف در سامانه‌های توصیه‌گر فهمیدن کاربر و ارائه پیشنهادهایی به او است که به آن‌ها علاقه دارد تا از این طریق اعتماد کاربر به سیستم جلب شود [۱۹]. برای اندازه‌گیری دقت در روش ارائه شده، ابتدا کاربران را به صورت تصادفی به دو دسته آموزش و آزمون جدا می‌کنیم. در اینجا از ۹۰٪ کاربران برای آموزش و ۱۰٪ مابقی کاربران برای آزمون استفاده کرده‌ایم. از بین کاربران آزمون نیز Q گزینه مورد علاقه کاربر را که به صورت تصادفی انتخاب شده‌اند به داده‌های آموزش اضافه می‌کنیم و سایر گزینه‌های مورد علاقه کاربر را در داده‌های آموزش قرار نمی‌دهیم. برای ارزیابی و مقایسه روش ارائه شده با سایر روش‌ها در شرایط مختلف، دو حالت داده‌های تنک و متراکم را در نظر گرفته‌ایم. مقدار Q را در آزمایشات برای این دو حالت به ترتیب برابر ۱ و ۱۰ قرار داده‌ایم. همانند روش‌های پایه مورد مقایسه [۷] [۱۲] با توجه به این که امتیازات از بازخورد غیرمستقیم کاربر به دست آمده‌اند ما نیز از معیار فراخوان<sup>۱۷</sup> جهت ارزیابی و مقایسه روش پیشنهادی استفاده می‌کنیم. به‌خصوص با توجه به این که امتیاز صفر در ماتریس امتیازات می‌تواند ناشی از عدم علاقه کاربر به آن گزینه و یا عدم اطلاع کاربر از آن گزینه باشد، استفاده از معیارهایی همچون معیار دقت<sup>۱۸</sup> نمی‌تواند دقیق باشد چرا که این معیار

$$\mathbb{E}[R_{ij}|D] \approx \mathbb{E}[u_i|D](\mathbb{E}[\theta_j|D] + \mathbb{E}[\epsilon_j|D] + \mathbb{E}[\pi_j|D])^T \quad (7)$$

که در این رابطه  $\theta_j$  از ماتریس امتیازات،  $\epsilon_j$  از اطلاعات محتوای گزینه‌ها و  $\pi_j$  از ارتباطات بین گزینه‌ها به دست آمده‌اند. اگر هر کدام از این سه دسته اطلاعات برای یک گزینه در دسترس نباشند از دسته‌های اطلاعاتی دیگر برای پیش‌بینی استفاده می‌شود. به عنوان مثال اگر برای یک گزینه هیچ امتیازی ثبت نشده باشد در نتیجه از ویژگی‌های به دست آمده از ماتریس امتیازات نمی‌توان استفاده کرد. پس در پیش‌بینی امتیاز این گزینه فقط از اطلاعات محتوا و ارتباطات آن استفاده می‌شود. در نتیجه برای این گزینه خواهیم داشت:

$$R_{ij} = u_i(\epsilon_j + \pi_j)^T \quad (8)$$

با استفاده از این رابطه می‌توان برای هر گزینه با استفاده از اطلاعات موجود رای را پیش‌بینی کرد. به عنوان مثال برای گزینه‌هایی که تازه وارد سیستم شده‌اند می‌توان با استفاده از محتوای آن‌ها در کنار ارتباطات بین گزینه‌ها آن‌ها را پیشنهاد کرد و در نتیجه یکی از بزرگترین چالش‌های موجود در سامانه‌های توصیه‌گر را با استفاده از این روش برطرف نمود. از طرف دیگر در روش پیشنهادی با استفاده از منابع اطلاعاتی مختلف دقت پیش‌بینی سیستم افزایش می‌یابد.

## ۵- ارزیابی روش پیشنهادی

جهت ارزیابی روش پیشنهادی از ۳ مجموعه داده که از فعالیت‌های کاربران در دنیای واقعی به دست آمده استفاده شده است و میزان دقت روش پیشنهادی در این مجموعه داده‌ها با سایر روش‌ها مورد بررسی قرار گرفته است.

### ۵-۱- مجموعه داده‌ها

در این مقاله ما از ۳ مجموعه داده از دنیای واقعی استفاده کرده‌ایم. دو مجموعه داده از داده‌های citeulike<sup>۱۵</sup> و یک مجموعه داده از داده‌های movielens<sup>۱۶</sup> هستند. Citeulike یک وبسایت است که به کاربران خود کمک می‌کند از بین تعداد انبوهی از مقالات چاپ شده مقاله‌های مورد علاقه خود را پیدا کنند. در این وبسایت هر کاربر برای خود یک کتابخانه شخصی می‌سازد و می‌تواند مقالات مورد علاقه خود را در کتابخانه شخصی خود قرار دهد. برای هر مقاله هم معمولاً عنوان، چکیده و ارجاعات آن مقاله در دسترس قرار دارد. دو مجموعه داده اول از این وبسایت به وسیله نویسندگان مقالات CTR و CDL جمع‌آوری شده‌اند. اما جمع‌آوری این داده‌ها با رویکردهای مختلفی انجام شده است و در نتیجه این دو مجموعه داده تفاوت‌های بسیاری از لحاظ بعد داده‌ها، میزان تنک بودن داده‌ها و موارد دیگر دارند. مجموعه داده سوم، movielens است که از دو قسمت تشکیل شده است. قسمت اول که شامل عناوین فیلم‌ها، امتیاز کاربران به فیلم‌ها و عوامل فیلم است در همین مجموعه داده قرار دارد. با توجه به این که در این پژوهش به محتوای گزینه‌ها نیز جهت ارائه پیشنهاد نیاز داریم و در این مجموعه داده از محتوای گزینه‌ها اطلاعاتی وجود ندارد، قسمت دوم این مجموعه داده که شامل داستان و خلاصه فیلم‌ها است را ما از وبسایت IMDB استخراج کردیم. همانند [۷] به جهت این که این مجموعه داده نیز با دو مجموعه داده اول که از بازخورد غیرمستقیم کاربر به دست آمده همخوانی داشته باشد، تنها از بازخوردهای مثبت کاربران به فیلم‌ها (امتیاز ۵ از ۵) استفاده شده است و سایر امتیازات در نظر گرفته نشده‌اند. در این مجموعه داده نیز کاربرانی که کمتر از ۳ امتیاز دارند در نظر گرفته نشده‌اند. برای ایجاد ارتباط بین گزینه‌ها در این مجموعه داده که فیلم‌ها هستند، از

آمده‌اند و از یک شبکه ۲ لایه SDAE که ۶۰ نورون در لایه پنهان دارد استفاده شده است.

در شبکه تبدیل پاراگراف به بردار نیز از  $H = 300$  نورون برای توصیف هر کلمه و پاراگراف و از پنجره  $k = 10$  کلمه در هر طرف هر کلمه برای به دست آوردن بردار کلمه جاری استفاده شده است. تعداد ویژگی‌های نهان مورد استفاده جهت نمایش گزینه‌ها و کاربران نیز که در تمام روش‌ها مورد استفاده قرار می‌گیرد به صورت مشترک برابر  $D = 50$  در نظر گرفته شده است.

#### ۵-۴- ارزیابی روش‌ها

در این بخش کارایی روش پیشنهادی را با روش‌های CDL [۷] و تجزیه ماتریسی [۳] مقایسه می‌کنیم. جهت مقایسه این سه روش از معیار  $\text{recall}@M$  با مقادیر  $M = [50, 100, 150, 200, 250, 300]$  استفاده می‌شود. نتایج حاصل از این آزمایش در مجموعه داده‌های مختلف در شکل ۵ آمده است. نتایج به دست آمده از مقایسه این روش‌ها نشان می‌دهد استفاده از اطلاعات اضافی تأثیر چشم‌گیری بر روی نتایج دارد و این تأثیر با افزایش تعداد نمونه‌های بازگردانده شده به کاربر نمایان‌تر می‌گردد و در مقادیر بالاتر  $M$  فاصله نتایج از یکدیگر افزایش می‌یابد. دلیل این امر این است که زمانی که تعداد اندکی گزینه باید به کاربر بازگردانده شود گزینه‌های معروف به احتمال زیادی انتخاب می‌شوند چون که اکثر کاربران به آن‌ها علاقه دارند. در نتیجه فاصله در مقادیر پایین‌تر  $M$  کم است و با افزایش  $M$  با توجه به محدود بودن گزینه‌های مشهور روش‌هایی همانند تجزیه ماتریسی که بر این اساس عمل می‌کنند نمی‌توانند عملکرد خوب خود را در مقادیر بالای  $M$  حفظ نمایند.

بازخوردهای منفی کاربر را نیز در نظر می‌گیرد. اما چون معیار یادآوری تنها به بازخوردهای مثبت کاربر وابسته است می‌تواند به درستی محاسبه گردد.

برای محاسبه مقدار یادآوری برای یک کاربر ابتدا میزان امتیاز یک کاربر به تمامی گزینه‌ها را پیش‌بینی می‌کنیم و سپس آن‌ها را به صورت نزولی مرتب کرده و  $M$  گزینه اول را به کاربر باز می‌گردانیم. سپس مقدار یادآوری در  $M$  گزینه اول با استفاده از رابطه زیر محاسبه می‌شود:

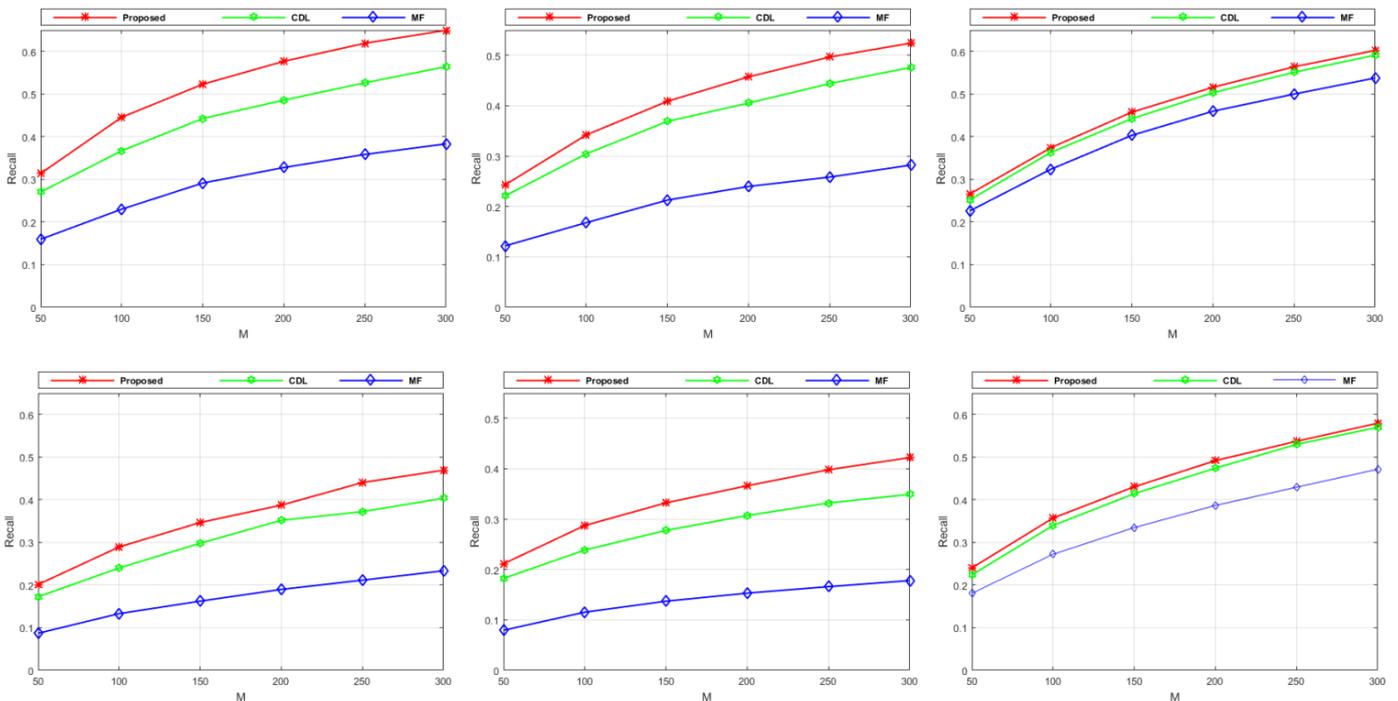
$$\text{recall}@M = \frac{\text{تعداد گزینه‌های مورد علاقه کاربر از بین } M \text{ گزینه اول انتخاب شده}}{\text{تعداد کل گزینه‌هایی که کاربر به آن‌ها علاقه دارد}} \quad (۹)$$

این مقدار برای تمام کاربران محاسبه می‌گردد و سپس میانگین یادآوری برای تمام کاربران مقدار یادآوری کل سیستم را مشخص می‌کند.

#### ۵-۳- مدل‌های مورد مقایسه و تنظیمات آزمایش

جهت بررسی عملکرد مدل پیشنهادی، این مدل را با روش CDL که در زمان نوشتن این مقاله بهترین نتایج در سامانه‌های توصیه‌گر روی هر دو مجموعه داده citeulike داشته‌است، مقایسه می‌کنیم. برای اجرای این روش از پیاده‌سازی ارائه‌دهندگان این روش [۷] استفاده شده است و پارامترهای کد نیز همان مقادیر پیش‌فرض مشخص شده در مقاله قرار داده شده است.

روش دیگری که مورد مقایسه قرار گرفته است روش تجزیه ماتریسی است. با توجه به این که این روش از اطلاعات محتوای گزینه‌ها استفاده نمی‌کند مقایسه این روش با روش پیشنهادی منصفانه نخواهد بود و نتایج تنها برای مشخص نمودن اهمیت اطلاعات اضافی مورد استفاده، گزارش شده‌اند. جهت به دست آوردن پارامترهای مدل در روش پیشنهادی داده‌های اعتبارسنجی مورد استفاده قرار گرفته است و با استفاده از این تکنیک اعتبارسنجی متقابل مقادیر



شکل ۵- کارایی روش‌های مورد بررسی در مجموعه داده‌های مختلف. در این نمودارها محور افقی نشان‌دهنده تعداد گزینه‌هایی است که برای کاربر بازگردانده می‌شود. سطر اول مربوط به مجموعه داده‌های متراکم است که در آن‌ها  $Q=10$  است و در سطر دوم مربوط به مجموعه داده‌های تنک است که در آن‌ها  $Q=1$  قرار داده شده است. ستون سمت چپ مربوط به مجموعه داده citeulike-A، ستون وسط citeulike-T و ستون سمت راست مربوط به داده‌های MovieLens می‌باشد.

مشهور در سیستم را نیز شناسایی کرد. به عنوان مثال در مجموعه داده citeulike که از ارجاعات بین مقالات ارتباطات بین آن‌ها ایجاد شده است می‌توان مقالاتی که پایه مباحث مختلف هستند را با استفاده از ارجاعاتی که به آن مقاله شده است شناسایی کرد چرا که مقالات پایه‌ای و اساسی هر مبحث عموماً بسیار مورد ارجاع قرار می‌گیرند و کاربرانی که به تازگی وارد یک مبحث شده‌اند به دنبال این مقالات هستند. به علاوه استفاده از ارتباطات بین گزینه‌ها به خصوص در مجموعه داده‌های تنک و حالت شروع سرد برای گزینه‌ها می‌تواند بسیار مفید باشد چرا که ماتریس امتیازات در این حالت نمی‌تواند کمک چندانی در پیدا کردن گزینه مشابه بکند ولی با استفاده از ارتباطات بین گزینه‌ها می‌توان گزینه‌های مشابه با یکدیگر را حتی در حالتی که تاکنون امتیازی برای یک گزینه ثبت نشده است پیدا کرد.

در این پژوهش از ارجاعات بین مقالات جهت ایجاد ارتباط بین گزینه‌ها استفاده شده است نمودارهای ستون سمت چپ شکل ۶ میزان تاثیر این ارتباطات را نمایش می‌دهند. نتایج حاصل اجرای آزمایش بر روی مجموعه داده citeulike-A در دو حالت تنک و متراکم را نشان می‌دهد.

با مقایسه نتایج به دست آمده در حالت متراکم و تنک مشخص می‌شود که در حالت داده‌های متراکم استفاده از ارتباطات بین مقالات تنها اندکی سبب بهبود نتایج شده است اما همانطور که مشخص است در حالت داده‌های تنک استفاده از ارتباطات بین گزینه‌ها تاثیر محسوس‌تری را بر روی نتایج نشان می‌دهد. همانطور که گفته شده، با توجه به این که در حالت تنک تنها ۱ گزینه به ازای هر کاربر تست، وجود دارد در نتیجه استفاده از ماتریس امتیازات کمک چندانی در پیدا کردن گزینه‌های مشابه نمی‌کند اما استفاده از ارتباطات همان یک گزینه به سیستم کمک می‌کند تا بتواند گزینه‌های مشابه را با استفاده از ارتباطات بین گزینه‌ها پیدا کند و در نتیجه سبب بهبود عملکرد سیستم می‌شود. اما در حالت متراکم با توجه به وجود امتیازات بیشتر برای هر کاربر سیستم می‌تواند گزینه‌های مشابه را با استفاده از همان ماتریس امتیازات به دست آورد و در نتیجه ارتباطات کمک چندانی در بهبود عملکرد سیستم در این حالت نمی‌کنند. یکی از دلایل دیگری که می‌تواند سبب کم بودن اثر ارتباطات در نتایج شود کم بودن تعداد ارتباطات بین گزینه‌ها است با توجه به این که در مجموعه داده مورد استفاده به ازای هر گزینه به طور میانگین تنها ۳ ارتباط با گزینه‌های دیگر وجود دارد در نتیجه ماتریس ارتباطات بسیار تنک خواهد بود که سبب کم شدن اثر ارتباطات در نتایج می‌گردد.

همانطور که گفته شد مدل ارائه شده یک رویکرد ترکیبی است که از محتوای گزینه‌ها در کنار ماتریس امتیازات جهت ارائه پیشنهاد استفاده می‌کند. در این پژوهش جهت به دست آوردن فضای نهان گزینه‌ها از روی محتوای آن‌ها ابتدا متن توصیف آن‌ها با روش تبدیل پاراگراف به بردار به یک بردار با ابعاد کوچکتر نگاشت می‌گردد سپس با استفاده از شبکه SDAE ویژگی‌های مهم این گزینه‌ها استخراج می‌گردد و میزان تاثیر محتوا در پیش‌بینی امتیازات نیز توسط پارامتر  $\lambda$  کنترل می‌گردد. میزان تاثیر این پارامتر در نمودارهای ستون سمت راست شکل ۶ مشخص شده است. جهت انجام آزمایش همانند بخش‌های قبلی از مجموعه داده‌ی citeulike-A در دو حالت تنک و متراکم استفاده شده است.

همانطور که در نمودارهای این بخش مشخص است استفاده از محتوای گزینه‌ها تاثیر قابل توجهی بر روی نتایج دارد چرا که کاربران عموماً به دنبال گزینه‌های مشابه با گزینه‌هایی که دوست دارند هستند و با استفاده از محتوای گزینه‌ها می‌توان گزینه‌های مشابه با یکدیگر را شناسایی کرد. از طرف دیگر تاثیر دو طرفه ویژگی‌های به دست آمده برای گزینه‌ها از ماتریس امتیازات و ویژگی‌های به دست آمده از محتوای گزینه‌ها بر یکدیگر سبب می‌شود ویژگی‌های مهم گزینه‌ها از متن محتوای آن‌ها استخراج شوند و انتخاب ویژگی‌های مناسب سبب بهبود عملکرد سیستم در حالت کلی می‌گردد.

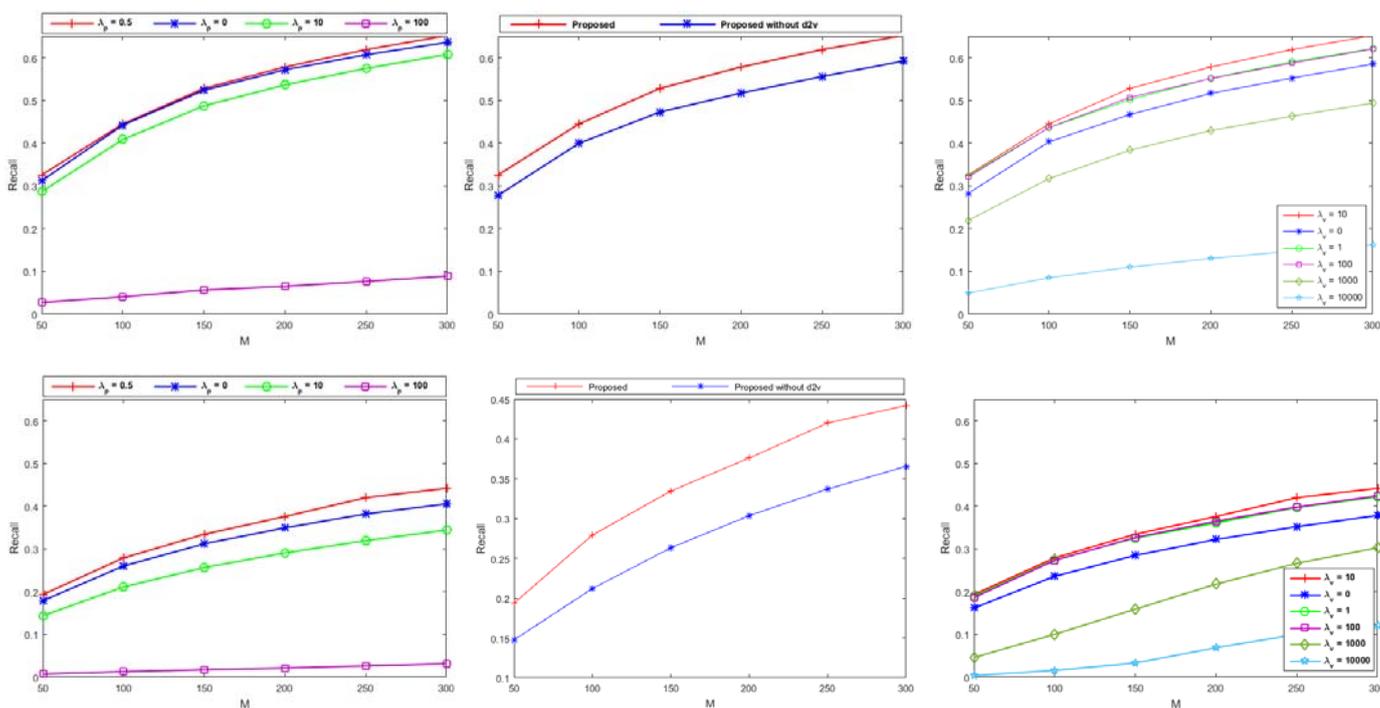
اما روش تجزیه ماتریسی غیرمغرم سادگی خود در حالت‌هایی که جدول امتیازات تنک نباشد بسیار خوب عمل می‌کند و همانطور که در نمودارهای مربوط به مجموعه داده movielens مشخص است این روش توانسته است به نتایجی نزدیک به روش CDL و روش پیشنهادی که مدل‌هایی پیچیده هستند دست یابد. اما عملکرد آن در مجموعه داده‌های تنک همانند citeulike-T به شدت افت می‌کند. یکی دیگر از دلایل نزدیک شدن نتایج مدل تجزیه ماتریسی به روش CDL و روش پیشنهادی می‌تواند به دلیل کم ارزش بودن اطلاعات متنی اضافی درباره فیلم‌ها نسبت به متن مقالات باشد در نتیجه این اطلاعات نمی‌توانند به خوبی فیلم مورد نظر را توصیف کنند که سبب کاهش کارایی مدل می‌شود. هر دو روش CDL و روش پیشنهادی از اطلاعات اضافی موجود درباره گزینه‌ها استفاده می‌کنند و از شبکه عمیق جهت استخراج ویژگی‌های نهان گزینه‌ها بهره می‌برند، اما نتایج ارائه شده نشان‌دهنده این مطلب است که روش پیشنهادی توانایی بهتری در به دست آوردن ویژگی‌های نهان گزینه‌ها دارد. یکی از مهمترین دلایل برتری روش ارائه شده نسبت به روش CDL استفاده از روش تبدیل پاراگراف به بردار جهت نمایش گزینه‌ها است. چرا که با این روش، ورودی با ابعاد بسیار کوچکتر و با حفظ ویژگی‌های اصلی گزینه‌ها را خواهیم داشت که هم سبب افزایش سرعت رسیدن به حالت پایدار می‌گردد و هم دقت مدل را در حالت کلی به دلیل کم کردن پارامترهای آزاد و جلوگیری از بیش‌برازش افزایش می‌دهد. یکی دیگر از دلایل برتری روش ارائه شده بر مدل CDL استفاده از اطلاعات ارتباطات بین گزینه‌ها است که سبب پیدا کردن گزینه‌های شبیه به هم در حالت شروع سرد برای گزینه‌ها می‌گردد.

## ۵-۵- بررسی میزان تاثیر بخش‌های مختلف مدل پیشنهادی

مدل ارائه شده از ۴ بخش کلی، شبکه تبدیل پاراگراف به بردار، شبکه SDAE، تجزیه ماتریس ارتباطات و تجزیه ماتریس امتیازات، تشکیل شده است و تمام این بخش‌ها در کنار یکدیگر به طور همزمان آموزش می‌بینند و میزان تاثیر آن‌ها بر مدل نهایی به کمک پارامترهایی که از بیرون تنظیم می‌گردد مشخص می‌شوند. در این بخش میزان تاثیر هر کدام از بخش‌ها بر روی نتایج کلی مدل مورد بررسی قرار گرفته است. میزان تاثیر تمام بخش در شکل ۶ آورده شده است (نتایج گزارش شده از اجرای آزمایش بر روی مجموعه داده citeulike-A به دست آمده‌اند).

همانطور که از نتایج گزارش شده در نمودارهای ستون وسط شکل ۶ مشخص است استفاده از روش تبدیل پاراگراف به بردار تاثیر قابل توجهی بر روی نتایج به دست آمده دارد و این تاثیر در حالت تنک نمایان‌تر است. بخش تبدیل پاراگراف به بردار از دو جهت تاثیر به‌سزایی بر عملکرد کلی سامانه می‌گذارد. اول این که با استفاده از این بخش تعداد ابعاد ورودی بسیار کاهش می‌یابد و ورودی از یک حالت تنک با بعد بالا به یک ورودی با ابعاد بسیار کمتر و به مقادیر حقیقی بین ۱- تا ۱ تبدیل می‌شود. در مثال مجموعه داده‌های citeulike-A که نتایج در آن بررسی شده است ورودی در حالت TF-IDF یک بردار ۵,۰۰۰ بعدی است که به طور میانگین تنها ۶۷ عنصر آن غیرصفر هستند ولی در صورت استفاده از روش تبدیل پاراگراف به بردار به ورودی ۳۰۰ بعدی که تمام عناصر آن بین ۱- تا ۱ هستند می‌رسیم. کم شدن تعداد ابعاد ورودی سبب می‌شود پارامترهای آزاد شبکه کمتر شوند و در نتیجه از بیش‌برازش جلوگیری شده و همچنین شبکه در مدت زمان کمتری به حالت پایدار با دقت بالاتری برسد.

در مدل پیشنهادی از ارتباطات بین گزینه‌ها نیز جهت ارائه پیشنهاد استفاده شده است. استفاده از ارتباطات بین گزینه‌ها سبب می‌شود گزینه‌های مشابه به یکدیگر را علاوه بر محتوا و شباهت بین آرای دریافتی از کاربران، از روی ارتباطات آن‌ها نیز بتوان پیدا کرد و همچنین با استفاده از این روش می‌توان گزینه‌های



شکل ۶- میزان تاثیر بخش‌های مختلف مدل. در این نمودارها محور افقی نشان‌دهنده تعداد گزینه‌هایی است که برای کاربر بازگردانده می‌شود. سطر اول مربوط به مجموعه داده‌های مترکم است که در آن‌ها  $Q=10$  است و سطر دوم مربوط به مجموعه داده‌های تنک است که در آن‌ها  $Q=1$  قرار داده شده است. ستون سمت چپ مربوط به میزان تاثیر بخش ارتباطات بین گزینه‌ها، ستون میانی میزان تاثیر بخش تبدیل پاراگراف به بردار و ستون سمت راست میزان تاثیر اطلاعات محتوای گزینه‌ها را نمایش می‌دهند. در تمام نمودارها از مجموعه داده‌های citeulike-A استفاده شده است

## مراجع

- [1] D. Almazro, G. Shahatah, L. Albdulkarim, M. Kherees, R. Martinez, and W. Nzoukou, "A Survey Paper on Recommender Systems," *CoRR*, 2010.
- [2] J. Tang, X. Hu, and H. Liu, "Social recommendation: a review," *Social Network Analysis and Mining*, pp. 1113-1133, 2013.
- [3] Y. Koren, R. Bell, and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," *Computer*, pp. 30-37, August 2009.
- [4] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," *Neural Networks*, pp. 85-117, 2015.
- [5] N. Wang, and D.-Y. Yeung, "Learning a Deep Compact Image Representation for Visual Tracking," *Advances in Advances in Neural Information Processing Systems 26*, pp. 809-817, 2013.
- [6] K. Nal, G. Edward, and B. Phil, "A Convolutional Neural Network for Modelling Sentences," *CoRR*, 2014.
- [7] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative Deep Learning for Recommender Systems," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1235-1244, 2015.

## ۶- نتیجه‌گیری و کارهای آتی

در این مقاله یک روش ترکیبی ارائه شد که از منابع اطلاعاتی مختلف جهت ارائه پیشنهاد به کاربر بهره می‌گیرد. مدل پیشنهادی با استفاده از اطلاعات اضافه موجود درباره گزینه‌ها و ارتباطات بین آن‌ها مشکل شروع سرد را برطرف کرده است و با ایجاد یک ارتباط محکم بین اجزای مدل توانسته نتایج بهتری نسبت به برترین روش موجود در سامانه‌های توصیه‌گر بر روی ۳ مجموعه داده دنیای واقعی کسب کند.

از جمله مواردی که به کمک آن‌ها می‌توان روش پیشنهادی را توسعه داد، استفاده از اطلاعات اضافه‌ای همچون ارتباطات بین کاربران با یکدیگر و اطلاعات موجود در نمایه کاربران است. در ادامه می‌توان از اطلاعات نمایه کاربران نیز استفاده کرد. در نتیجه می‌توان به کاربرانی که تازه وارد سیستم شده‌اند نیز گزینه‌هایی را با استفاده از ارتباطات آن‌ها و یا اطلاعات نمایه کاربران پیشنهاد کرد و به این ترتیب مشکل شروع سرد برای کاربران نیز تا حد زیادی برطرف شود.

در این مقاله محور اساسی آزمایشات بر روی مجموعه داده‌هایی بود که از بازخوردهای غیرمستقیم کاربر نسبت گزینه‌ها در سیستم به وجود آمده بودند و به همین دلیل برای هر گزینه تنها دو حالت مورد علاقه کاربر و یا عدم علاقه کاربر به آن گزینه وجود داشت اما در مجموعه داده‌هایی که کاربران به گزینه‌ها یک امتیاز (مثلاً بین ۱ تا ۱۰) انتساب می‌دهند سیستم باید بتواند میزان علاقه کاربر به هر گزینه را با یک مقدار عددی مشخص کند. روش پیشنهادی با تغییر اندکی قابلیت به دست آوردن امتیاز را نیز دارد که این موضوع نیز می‌تواند مورد بررسی قرار بگیرد.

آدرس پست‌الکترونیکی ایشان عبارت است از:

kiboabbasi@ce.sharif.edu

**مهديه سلیمانی باغشاه** دوره کارشناسی، کارشناسی‌ارشد و دکتری خود را به ترتیب در سال‌های ۸۲، ۸۴ و ۸۹ در دانشکده مهندسی کامپیوتر دانشگاه صنعتی شریف به اتمام رسانده است. همچنین وی از سال ۹۱ تاکنون به عنوان استادیار دانشکده مهندسی کامپیوتر این دانشگاه به آموزش و پژوهش پرداخته است. زمینه‌های تحقیقاتی مورد علاقه ایشان یادگیری ماشین، مدل‌سازی احتمالاتی، بازیابی اطلاعات و سیستم‌های هوشمند است.



آدرس پست‌الکترونیکی ایشان عبارت است از:

soleymani@sharif.edu

**اطلاعات بررسی مقاله:**

تاریخ ارسال: ۱۳۹۶/۰۴/۱۸

تاریخ اصلاح: ۱۳۹۶/۰۶/۰۶

تاریخ قبول شدن: ۱۳۹۷/۰۱/۲۲

نویسنده مرتبط: دکتر مهديه سلیمانی، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، تهران، ایران.

<sup>1</sup><https://www.amazon.com>

<sup>2</sup><https://news.google.com>

<sup>3</sup><https://www.netflix.com>

<sup>4</sup>Supervised

<sup>5</sup>Cold Start

<sup>6</sup>Tightly Coupled

<sup>7</sup>Loosely Coupled

<sup>8</sup>Collaborative Topic Regression

<sup>9</sup>Latent Dirichlet Allocation

<sup>10</sup>Collaborative Deep Learning

<sup>11</sup>Auto-Encoder

<sup>12</sup>Stacked Denoising Auto-Encoder

<sup>13</sup>Sequence

<sup>14</sup>Binary

<sup>15</sup><http://www.citeulike.org/faq/data.adp>

<sup>16</sup><http://ir.ii.uam.es/hetrec2011>

<sup>17</sup>Recall

<sup>18</sup>Precision

[8] Q. V. Le, and T. Mikolov, "Distributed Representations of Sentences and Documents," *International Conference on Machine Learning*, pp. 1188-1196, 2014.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *CoRR*, 2013.

[10] P. Lops, M. De Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender systems handbook*, Springer, 2011, pp. 73-105.

[11] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, pp. 77-84, 2012.

[12] C. Wang, and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 448-456, 2011.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, pp. 993-1022, 2003.

[14] R. Salakhutdinov, and A. Mnih, "Probabilistic matrix factorization," *Neural Information Processing Systems*, pp. 1-8, 2011.

[15] S. Li, J. Kawale, and Y. Fu, "Deep collaborative filtering via marginalized denoising auto-encoder," *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 811-820, 2015.

[16] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, pp. 3371-3408, 2010.

[17] Z. S. Harris, "Distributional structure," *Word*, pp. 146-162, 1954.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, pp. 3111-3119, 2013.

[19] M. A. Abbasi, J. Tang, and H. Liu, "Trust-aware recommender systems," *Machine Learning book on computational trust*, Chapman & Hall/CRC Press, 2014.

**امید عباسی** دوره کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار و در سال ۹۳ از دانشگاه شیراز و کارشناسی‌ارشد خود را در رشته مهندسی کامپیوتر گرایش هوش مصنوعی و در سال ۹۵ از دانشگاه صنعتی شریف اخذ نموده است. زمینه‌های مورد علاقه ایشان یادگیری عمیق، سامانه‌های توصیه‌گر و بازیابی اطلاعات است.



## ارائه روشی متنی بر محتوا و جذابیت برای خلاصه‌سازی مجموعه تصاویر اجتماعی

محسن ابراهیمی مقدم

زهرا ریاحی سامانی

دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران

### چکیده

روش‌های خلاصه‌سازی تصویر امروزه به ابزاری مهم برای استفاده از مجموعه‌های تصویری بزرگ تبدیل شده‌اند. این روش‌ها مجموعه کوچکی از تصاویر را به‌عنوان نماینده از یک مجموعه بزرگ‌تر انتخاب می‌کنند. مطالعات جدید نشان داده‌اند که تصاویری که در روش‌های خلاصه‌سازی تصویر انتخاب می‌شوند باید از یک سو پوشش اطلاعاتی مناسبی از مجموعه اولیه فراهم کنند و از سوی دیگر معیارهایی مانند کیفیت و زیبایی تصاویر را در انتخاب تصاویر خلاصه مد نظر قرار دهند. با علم به این موارد، در این مقاله روشی خودکار برای خلاصه‌سازی مجموعه‌های تصویری با به کارگیری و مطالعه دو دسته ویژگی با عنوان ویژگی‌های محتوایی و ویژگی‌های جذابیت ارائه شده است. در روش پیشنهادی دو دسته ویژگی معرفی شده به‌طور خودکار و با استفاده از بستر شبکه‌های اجتماعی و هستان‌شناسی دامنه استخراج شده و تاثیرگذاری آنها در خلاصه‌سازی تصاویر با استفاده از بازخورد کاربران مد نظر می‌شود. خلاصه نهایی با ترکیب خلاصه ساخته شده به‌وسیله دو دسته ویژگی به‌دست آورده می‌شود. نتایج روش پیشنهادی با تصاویر خلاصه‌ای که توسط گروهی از کاربران بر روی مجموعه‌ای از تصاویر انتخابی از فلیکر انجام شده است مقایسه شده است. آزمایش‌ها نشان می‌دهد که روش پیشنهادی در مقایسه با روش‌های جاری خلاصه‌سازی تصویر توانسته است به بهبود سی و پنج درصدی پوشش اطلاعاتی در خلاصه کردن مجموعه تصویری به میزان یک درصد حجم اولیه و بهبود چهار درصدی پوشش اطلاعاتی در خلاصه کردن مجموعه تصویری به میزان پنج درصد حجم اولیه دست یابد. همچنین روش پیشنهادی توانسته است به بهبود سی و نه درصدی در ساختن خلاصه‌های شبیه خلاصه‌های دستی دست یابد.

**کلمات کلیدی:** تصاویر اجتماعی، خلاصه‌سازی تصویر، بازخورد کاربر.

### ۱- مقدمه

این اطلاعات جانبی و همچنین اطلاعات جانبی دیگر که در دوربین‌های رقمی موجود هستند امروزه در سیستم‌های چندرسانه‌ای استفاده می‌شوند.

هدف یک سیستم خلاصه‌سازی مجموعه تصویر آن است که با دریافت یک مجموعه بزرگ از تصاویر بتواند مجموعه کوچک‌تری از تصاویر را از مجموعه اولیه به عنوان تصاویر نماینده یا تصاویر خلاصه انتخاب کند، به گونه‌ای که مجموعه انتخابی را بتوان جوهره و یا چکیده‌ای از مجموعه اولیه دانست [۲].

روش‌های اولیه در خلاصه‌سازی تصویر معیارهای پوشش<sup>۱</sup> و تکرار<sup>۲</sup> را دنبال می‌کردند، به این ترتیب که تصاویر مجموعه خلاصه باید تا حد ممکن شبیه مجموعه اولیه بوده و تا حد ممکن ناشبیه به همدیگر باشند [۳]. روش‌های جدیدتر این ایده را مطرح می‌کنند که معیارهای دیگری مانند کیفیت یا جذابیت هم برای انجام عمل خلاصه‌سازی مهم هستند. رادیناک و همکاران [۴] یک آزمایش بررسی جمعی انجام داده‌اند و در آن به تحلیل اینکه کاربر انسانی چگونه خلاصه‌سازی تصویر را انجام می‌دهد پرداخته‌اند. نتایج این مطالعه نشان می‌دهد که تصاویری که

با ظهور دوربین‌های دیجیتال و تلفن‌های هوشمند اندازه مجموعه‌های تصویری رو به افزایش است. مطالعات نشان می‌دهد که روزانه ۵۰۰ میلیون تصویر در اینترنت بارگذاری می‌شود (فقط با در نظر گرفتن Instagram, Facebook, Flickr و Snapchat) [۱]. این مسئله موجب افزایش تقاضا برای روش‌های خلاصه‌سازی تصویر شده است که به کاربرها در مرور و پیدا کردن مجموعه‌های تصویری کمک می‌کنند.

تصاویر اجتماعی یا تصاویر بارگذاری شده در شبکه‌های اجتماعی و سایت‌های اشتراک تصویر معمولاً با داده‌های جانبی همراه هستند. به عنوان مثال اگر کاربری آنها را جذاب ببیند می‌تواند آنها را نشانه کند و یا بر روی آنها یادداشت بنویسد.

برای ارزیابی روش پیشنهادی یک مجموعه خلاصه انسانی برای مجموعه‌ای از تصاویر استخراجی از فلیکر در دامنه شهرها و مناطق ایجاد شده است. نتایج خلاصه‌سازی با نتایج کاربر انسانی مقایسه شده است. نتایج نشان می‌دهد که روش پیشنهادی نتایج بهتری در مقایسه با روش‌های دیگر در حوزه خلاصه‌سازی تصویر و همچنین در مقایسه با روش‌های تک معیاره دارد.

نوآوری‌های این مقاله را به صورت زیر می‌توان نام برد:

- ۱- یک روش خودکار برای خلاصه‌سازی تصاویر اجتماعی ارائه شده است که محتوا و ویژگی‌های جذابیت را هم‌زمان در نظر می‌گیرد.
  - ۲- دو مجموعه قابل تفسیر از ویژگی‌های خلاصه‌سازی پیشنهاد شده و میزان اثر بخشی آنها در خلاصه‌سازی تصویر با استفاده از بازخورد کاربران به دست آورده شده است.
  - ۳- روش‌هایی خودکار برای استخراج ویژگی‌های معرفی شده به کار گرفته شده و روش‌های موجود در این حوزه بهبود داده شده‌اند.
- ساختار مقاله در ادامه به صورت زیر است. در قسمت بعدی در مورد کارهای مرتبط در حوزه خلاصه‌سازی تصویر صحبت خواهد شد. سپس به معرفی روش پیشنهادی پرداخته و پس از آن نتایج روش پیشنهادی ارائه شده و با کارهای مرتبط مقایسه می‌شود.

## ۲- کارهای مرتبط

در قسمت قبلی مقاله به خطوط اصلی تحقیقات در حوزه خلاصه‌سازی مجموعه تصاویر اشاره کردیم. در این بخش به بررسی جزئی‌تر روش‌های مطرح در حوزه خلاصه‌سازی تصویر می‌پردازیم.

عمدتاً روش‌های مطرح در خلاصه‌سازی مجموعه تصاویر رویکرد بدون نظارت دارند. به تازگی روش‌هایی با رویکرد با نظارت نیز مطرح شده‌اند [۱]. رویکردهای با نظارت معمولاً مبتنی بر روش‌های یادگیری هستند که تلاش می‌کنند شیوه خلاصه‌سازی دستی که به وسیله انسان انجام می‌شود را یاد بگیرند. این گونه روش‌ها معمولاً در مرتبه تصاویر چند ده تایی قابل به‌کارگیری هستند، اما در مقابل، روش‌های بدون نظارت بر روی تعداد چند صد و یا چند هزار تصویر تمرکز دارند. از بعد دیگر مقایسه، می‌توان روش‌های خلاصه‌سازی تصویر را به دو دسته عمده تقسیم کرد. روش‌های خلاصه‌سازی مبتنی بر ویژگی‌های تصویری و روش‌های خلاصه‌سازی چند حالتی<sup>۷</sup>.

روش‌های خلاصه‌سازی تصویری، تنها از ویژگی‌های تصویری استفاده می‌کنند. سیستم‌های خلاصه‌سازی چند حالتی، از اطلاعات جانبی تصاویر نیز برای انجام عمل خلاصه‌سازی استفاده می‌کنند. تصاویر موجود بر روی وب و یا تصاویر به اشتراک‌گذاری شده بر روی شبکه‌های اجتماعی دارای داده‌های جانبی هستند. به عنوان مثال، کاربران می‌توانند بر روی آنها داده‌های متنی یادداشت کنند و یا اگر کاربری تصویری را جذاب ببیند می‌تواند آن را به اشتراک بگذارد و یا می‌تواند تصویر را جزء تصاویر مورد علاقه‌اش قرار دهد. همچنین دوربین‌های دیجیتال، امروزه اطلاعاتی از زمان و مکان تصویر را ثبت می‌کنند. روش‌های متعددی در ادبیات موضوع مطرح شده است که از این اطلاعات جانبی استفاده می‌کنند [۴، ۷، ۸]. اکثر این روش‌ها در حضور داده‌های جانبی معمولاً موفق عمل می‌کنند اما این داده‌ها همیشه در دسترس نیستند. در ادامه در ابتدا به شرح روش‌های خلاصه‌سازی تصویری خواهیم پرداخت و پس از آن روش‌های خلاصه‌سازی چند حالتی مطرح می‌شوند.

یکی از مطرح‌ترین روش‌های خلاصه‌سازی تصویری روش [۳] است. نویسندگان در این مقاله یک نسخه حریصانه از الگوریتم k-means را برای خلاصه‌سازی مجموعه تصاویر ارائه داده‌اند. الگوریتم پیشنهادی به طور تکراری

به عنوان خلاصه انتخاب می‌شوند از نظر معنایی شبیه تصاویر دیگر هستند. از طرف دیگر، این نتایج نشان می‌دهد که معیارهای دیگری مانند کیفیت و زیبایی تصاویر نیز در خلاصه‌سازی مهم هستند [۴].

در این مقاله به معرفی و استخراج خودکار دو دسته ویژگی برای خلاصه‌سازی تصویر پرداخته‌ایم. دو دسته ویژگی مطرح شده شامل ویژگی‌های محتوایی و ویژگی‌های جذابیت هستند.

هدف ویژگی‌های محتوایی پیدا کردن تصاویری از مجموعه اولیه است که در آنها بتوانند پوشش اطلاعاتی مجموعه اولیه را در تعداد کمتر تصویر فراهم کنند. رویکردی که ما در این مقاله برای ویژگی‌های محتوایی به کار گرفته‌ایم رویکردی معنایی است و علاوه بر شباهت بصری تصاویر، مفاهیمی که در آنها وجود دارند نیز در نظر گرفته شده است. روش‌هایی که تا به امروز بر روی پوشش اطلاعاتی تمرکز داشتند به طور عمده بهینه‌سازی معیارهای تصویری یا عددی را دنبال می‌کردند [۵]. این روش‌ها معمولاً شباهت بصری را در نظر می‌گیرند. به تازگی روش‌هایی مطرح شده‌اند که به مدل کردن اطلاعات معنایی مجموعه تصویر پرداخته‌اند [۲]. در این مقاله بهبودی بر روی مدل معنایی [۲] ارائه شده است. روش ارائه شده در [۲] بر مبنای استفاده از هستان‌شناسی<sup>۸</sup> دامنه و استخراج ویژگی‌هایی تحت عنوان ویژگی‌های هستان‌شناسی است. جزئیات این ویژگی‌ها در قسمت ۲-۳ بحث خواهد شد. در این مقاله ویژگی‌های هستان‌شناسی معرفی شده در [۲] بهبود داده شده‌اند و همچنین دسته دومی از ویژگی‌ها تحت عنوان ویژگی‌های جذابیت برای خلاصه‌سازی تصویر معرفی و به کار گرفته شده‌اند.

ویژگی‌های جذابیت شامل کیفیت تصویر، احساس، اثر بخشی و زیبایی تصویر هستند. برای استخراج ویژگی‌های جذابیت، در این مقاله پیشنهاد شده است که از اطلاعات موجود بر روی تصاویر بارگذاری شده در شبکه‌های اجتماعی استفاده شود. شبکه‌های اجتماعی به کاربران خود اجازه می‌دهند که رفتارهای اجتماعی بر روی تصاویر انجام دهند. به عنوان مثال اگر یک تصویر برای یک کاربر جذاب باشد می‌تواند آن را پسند<sup>۹</sup> کند، بر روی آن یادداشت<sup>۱۰</sup> بنویسد و یا مجدداً به اشتراک<sup>۱۱</sup> بگذارد. مطالعات جدید نشان می‌دهد که این فراداده‌ها می‌توانند برای نشان دادن جذابیت تصویر به کار گرفته شوند [۶]. در این مقاله از این بستر برای پیدا کردن تصاویر جذاب استفاده شده است. جزئیات این ویژگی‌ها در قسمت ۱-۳ بحث خواهد شد.

همچنان که در بخش‌های آتی توضیح داده خواهد شد ویژگی‌های محتوایی و جذابیت دارای ابعاد بالا بوده و به محتوای تصاویر حساس هستند. ارائه یک روش یادگیری با حضور این بعد بالای ویژگی، نیازمند تعداد زیادی داده آموزشی است. بنابراین در این مقاله پیشنهاد شده است که خلاصه‌سازی را با استفاده از ویژگی‌های محتوا و جذابیت به طور جداگانه انجام داده و سپس خلاصه نهایی با ترکیب دو خلاصه ساخته شده به دست آورده شود. بدین ترتیب سه پیمانانه برای خلاصه‌سازی تصویر تعریف شده است. پیمانانه اول پیمانانه محتوایی نام دارد و امتیازی برای میزان پوشش محتوایی هر تصویر - از مجموعه اولیه - استخراج می‌کند. پیمانانه دوم پیمانانه جذابیت نام دارد و برای هر تصویر امتیازی برای میزان جذابیت آن تصویر استخراج می‌کند. سومین و آخرین پیمانانه، پیمانانه تجمیع و خلاصه‌سازی است که تصاویر خلاصه را با بررسی اطلاعات هر دوی پیمانانه‌های محتوایی و جذابیت تولید می‌کند.

مسئله بعدی انتخاب روش ترکیب نتایج پیمانانه‌های محتوایی و جذابیت است. در این مقاله از روش جمع وزن‌دار استفاده شده است. انتخاب وزن‌ها با استفاده از یک آزمایش بازخورد از کاربر و در یک مرحله آموزش انجام شده است. در واقع وزن‌ها نشان‌دهنده این مسئله هستند که هر کدام از دو گروه ویژگی معرفی شده چه میزان در تولید خلاصه تاثیرگذار هستند. جزئیات آزمایش بازخورد کاربر در قسمت ۳-۳ توضیح داده شده است.

تصاویر به مسئله پیدا کردن بردار ویژه ماتریس مجاورت تبدیل شده است [۱۷]. ریاحی و ابراهیمی مقدم در [۲] یک نسخه معنایی از گراف مشابهت را ساخته و از الگوریتم‌های مرکزیت گراف برای پیدا کردن تصاویر نماینده استفاده کرده‌اند. هم چنین در روش پیشنهادی در [۱۶] یک نسخه بهبود داده شده از الگوریتم PageRank پیشنهاد داده شده است. در این روش پس از استخراج هر کدام از تصاویر خلاصه، گراف اولیه براساس تفاوت با تصویر انتخابی وزن‌دهی مجدد می‌شود. به این ترتیب تصاویر خلاصه متنوع‌تر می‌شوند.

دسته دوم کارهای مرتبط، روش‌های خلاصه‌سازی تصویر چند حالت هستند. تصاویر اجتماعی و یا تصاویر بارگذاری شده بر روی وب دارای اطلاعات جانبی با محتوای متنی، زمانی، رفتاری یا مکانی هستند. روش‌های بسیاری در ادبیات موضوعی مطرح شده است که از یکی [۱۲] و یا ترکیبی [۴، ۷، ۱۸] از این اطلاعات جانبی برای خلاصه‌سازی تصویر استفاده می‌کنند. اکثر این روش‌ها در حضور این داده‌های جانبی موفق عمل می‌کنند.

اولین داده جانبی که در تحلیل‌ها استفاده شده است داده مکانی است. کندی و همکاران [۱۹] از داده‌های مکانی و متنی برای پیدا کردن تصاویر نماینده استفاده می‌کنند. به این صورت که ابتدا برچسب‌های متنی و اطلاعات مکانی را تحلیل کرده و برچسب‌های نماینده را به دست می‌آورند و سپس در مرحله بعد تصاویر مربوط به برچسب‌های نماینده انتخاب می‌شوند. جف و همکاران [۱۸] از ترکیبی از داده‌های متنی، مکانی، زمانی و اجتماعی و الگوریتم‌های خوشه‌بندی برای خلاصه‌سازی تصویر استفاده می‌کنند.

داده جانبی دیگری که به طور وسیعی به کار برده شده است داده متنی است. پانگ و همکاران روشی را برای خلاصه‌سازی تصاویر استفاده کرده‌اند که در آن با روش‌های تحلیل متن همراه تصویر ابتدا عنوان‌های متنی نماینده را انتخاب کرده و سپس تصاویر مرتبط با آن عنوان‌ها را انتخاب می‌کنند [۲۰]. کامارگو و همکاران از یک روش فاکتورگیری نامنفی ماتریس برای پیدا کردن عنوان‌های نهفته در فضای متنی استفاده می‌کنند و سپس نتایج این فاکتورگیری را برای انجام فاکتورگیری نامنفی ماتریس در فضای تصویر به کار می‌برند [۸].

رفتارهای اجتماعی کاربران منبع دیگری از اطلاعات است که می‌توان از آن استفاده کرد. به عنوان مثال جیانگ و همکاران از ویژگی‌های تصاویری که روی شبکه‌های اجتماعی دارای یادداشت هستند برای خلاصه‌سازی مجموعه تصویر استفاده کرده‌اند [۲۱].

به تازگی روش‌های چند معیاره و با نظارت نیز معرفی شده‌اند. این روش‌ها مبتنی بر یادگیری خلاصه‌سازی از روی خلاصه‌سازی دستی انسان هستند. به عنوان مثال شیاتشک و همکاران [۱] روشی با نظارت برای خلاصه‌سازی تصاویر ارائه داده‌اند که بر مبنای بهینه‌سازی معیاری با نام V-rouge است که به صورت فاصله بین خلاصه دستی و اتوماتیک تعریف می‌شود.

رودیناک و همکاران [۴] یکی دیگر از روش‌های با نظارت و چند حالت را ارائه داده‌اند. روش پیشنهادی آن‌ها چند حالت است به این دلیل که از منابع اطلاعاتی متفاوتی مانند داده‌های متنی، مکانی و زمانی استفاده می‌کند و چند معیاره است به این دلیل که هم محتوا و هم ویژگی‌های زیبایی تصویر را در نظر می‌گیرد. آن‌ها از ماشین بردار پشتیبان برای رتبه‌بندی تصاویر استفاده کرده‌اند.

شن و همکاران [۷] روش دیگری ارائه داده‌اند که از مجموعه‌ای از داده‌های مکانی، زمانی و معنایی برای انجام خلاصه‌سازی استفاده می‌کنند. آن‌ها از مدل مخلوط گاوسی برای انجام خوشه‌بندی استفاده می‌کنند و تصاویر را در تعدادی رویداد خوشه‌بندی می‌کنند. سپس از یک الگوریتم رتبه‌بندی برای رتبه‌بندی تصاویر استفاده می‌کنند.

جدول ۱ مقایسه‌ای بر روی روش‌های جدیدتر خلاصه‌سازی تصویر را نشان می‌دهد. این روش‌ها در نوع ورودی، میزان نظارت، رهیافت و روش خلاصه‌سازی متفاوت هستند در این مقاله روشی برای خلاصه‌سازی مجموعه تصویر ارائه می‌شود

تصاویری را برای خلاصه انتخاب می‌کند که شباهت مجموعه انتخابی با مجموعه اولیه را بیشینه و شباهت میان مجموعه انتخابی را کمینه کند. روش پیشنهادی در [۵] هم بهینه‌سازی مشابهی را دنبال می‌کند. این روش بر این مبنای استوار است که تصاویر مجموعه اولیه را می‌توان با ترکیب خطی و تنک ویژگی‌های محلی تصاویر خلاصه بازسازی کرد. تابع هدف [۵] به صورت زیر تعریف می‌شود:

$$\min_{D,A} \sum_i \|x_i - D\alpha_i\|^2 + \lambda \sum_i \|\alpha_i\| + \beta \sum_{j \neq k} \text{corr}(d_k, d_j) \quad (1)$$

که در آن مولفه اول اختلاف تصاویر اولیه با تصاویر بازسازی شده به وسیله ترکیب خطی تصاویر خلاصه است، مولفه دوم شرط تنک بودن را بر آورده می‌کند و مولفه سوم شباهت ویژگی‌های تصاویر خلاصه به یکدیگر را کمینه می‌کند. این روش از یک نسخه از الگوریتم شبیه‌سازی حرارت برای بهینه‌سازی تابع هدف استفاده می‌کند.

الگوریتم‌های خوشه‌بندی نیز برای خلاصه‌سازی مجموعه‌های تصویری به طور گسترده‌ای به کار گرفته شده‌اند. این روش‌ها مبتنی بر خوشه‌بندی مجموعه تصاویر و سپس انتخاب تصاویر نماینده از هر خوشه هستند. روش‌های خوشه‌بندی متفاوتی در ادبیات موضوعی برای این کار به کار گرفته شده‌اند: مانند الگوریتم انتشار وابستگی [۹] و خوشه‌بندی طیفی [۱۰]. روش‌های خلاصه‌سازی مبتنی بر خوشه‌بندی تصویر، ویژگی‌های مختلفی مانند ویژگی‌های محلی [۱۰]، سراسری [۹، ۱۱] و یا ترکیبی از هر دوی آنها [۱۲، ۱۳] از تصاویر استخراج می‌کنند.

روش‌های مرکزیت گراف نیز برای خلاصه‌سازی تصاویر استفاده شده‌اند [۱۴-۱۶]. یکی از مطرح‌ترین آن‌ها روش مطرح شده در [۱۴] است. در این روش از الگوریتم PageRank بر روی گراف مشابهت تصویر برای خلاصه‌سازی تصویر استفاده شده است. در گراف مشابهت نودها تصاویر هستند و شباهت بین آنها وزن یال گراف را مشخص می‌کند و الگوریتم PageRank مبتنی بر مرکزیت مقدار ویژه عمل می‌کند. مرکزیت مقدار ویژه مرکزیتی است که تعداد همسایه‌های یک راس و همچنین اهمیت همسایه‌های آن راس را به عنوان معیاری از اهمیت آن در نظر می‌گیرد و مبتنی بر این اصل است که راسی که به یک راس مهم نزدیک‌تر است اهمیت بالاتری در مقایسه با راس‌های دورتر دارد. این مرکزیت در مقابل مرکزیت درجه مطرح می‌شود که در آن اهمیت هر راس به عنوان تابعی از تعداد یال‌های آن در نظر گرفته می‌شود. اگر  $A_{ij}$  را به صورت شباهت بین دو راس  $i$  و  $j$  (معادله ۲) و  $e_i$  را به عنوان مرکزیت بردار ویژه‌ی مجموعه تصاویر تعریف شود  $e_i$ ، به صورت معادله‌ای تعریف می‌شود که مرکزیت درجه همسایه‌های  $v$  را اندازه می‌گیرد [۱۷].

$$A_{ij} = \text{Sim}_{ij} \quad (2)$$

$$e_i = \frac{1}{\lambda} \sum_j A_{ij} e_j \quad (3)$$

در این رابطه،  $\lambda$  فاکتور نرمال‌سازی است و باعث می‌شود که جمع  $e_i$ ‌ها یک شود ( $\sum_i e_i = 1$ ). با بازنویسی متغیرها به فرم برداری  $E = (e_1, e_2, \dots, e_n)^T$  معادله ۳ به صورت معادله ۴ بازنویسی می‌شود [۱۷]:

$$E = \frac{1}{\lambda} AE \quad (4)$$

$$AE = \lambda E \quad (5)$$

که در آن  $E$  به صورت بردار ویژگی ماتریس مجاورت  $A$  تعریف می‌شود. بنابراین مسأله‌ی پیدا کردن تعدادی تصویر به عنوان نماینده از مجموعه اولیه

به سیستم وارد می‌کند. در گام آموزش، طبقه‌بند با مفاهیم هستان‌شناسی دامنه آموزش داده می‌شود. در گام آزمایش از طبقه‌بند آموزش داده شده برای محاسبه ویژگی‌های هستان‌شناسی استفاده شده و سپس ویژگی‌های هستان‌شناسی برای محاسبه محتوای اطلاعاتی تصاویر استخراج می‌شود.

ایده پیمانه جذابیت پیدا کردن مدلی است که بتوان از روی آن ویژگی‌های سطح بالای جذابیت را از ویژگی‌های سطح پایین و تصویری به دست آورد. در این مقاله پیشنهاد شده است که از طبقه‌بندی استفاده شود که بتواند تصاویر جذاب را از غیرجذاب جدا کند. در گام آموزش، طبقه‌بند با تصاویر جذاب و غیرجذاب آموزش داده می‌شود. برای فراهم کردن داده آموزش برای طبقه‌بند، از رفتارهای کاربران در شبکه‌های اجتماعی استفاده شده است. به این ترتیب که دو مجموعه از تصاویر با برچسب‌های جذاب و غیرجذاب با استفاده از تصویری که در شبکه‌های اجتماعی نشانه<sup>۱۱</sup> دارند جمع‌آوری می‌شود. رفتارهای کاربران نیز با توجه به محتوا و زمینه‌ی تصاویر متفاوت هستند. به عنوان مثال در شهری مانند رم دامنه‌ی تعداد نشانه‌ها بیشتر از شهری با تعداد کمتر بازدیدکننده است. بنابراین جذابیت باید برای محتواهای متفاوت به طور جداگانه‌ای مدل شود. با مدل کردن جذابیت در شبکه‌های اجتماعی، تصاویر جذاب و غیرجذاب به عنوان داده آموزشی به طبقه‌بند جذابیت داده می‌شود. از این طبقه‌بند در گام آزمایش برای ساختن امتیاز جذابیت استفاده می‌شود.

پس از گام آموزش نوبت به گام آزمایش می‌رسد. مراحل این گام در شکل ۲ نشان داده شده است. روش پیشنهادی شامل سه پیمانه است. دو پیمانه اول بر روی ویژگی‌های محتوایی و ویژگی‌های جذابیت کار می‌کنند. پیمانه اول برای هر تصویر امتیازی استخراج می‌کند که نشان‌دهنده میزانی از اطلاعات است که یک تصویر از کل مجموعه تصاویر را پوشش می‌دهد. این پیمانه را از این پس پیمانه محتوا می‌نامیم. پیمانه دوم بر روی ویژگی‌های جذابیت کار می‌کند و یک امتیاز به هر تصویر نسبت می‌دهد که نشان‌دهنده میزان جذابیت هر تصویر است. از این پس این پیمانه را پیمانه جذابیت می‌نامیم.

پیمانه سوم پیمانه تجمیع و خلاصه‌سازی است. هدف این پیمانه این است که خروجی دو پیمانه دیگر را ترکیب کرده و یک عدد نهایی و نشان‌دهنده میزان نماینده بودن تصویر استخراج کند. برای انجام تجمیع از روش جمع وزن‌دار استفاده کرده‌ایم. به این ترتیب هر تصویر یک امتیاز خلاصه‌سازی دریافت می‌کند که برابر جمع وزن‌دار امتیاز جذابیت و امتیاز محتوا است.

لازم به ذکر است که هستان‌شناسی دامنه و بستر شبکه‌های اجتماعی در گام آموزش استفاده می‌شوند و هیچ اطلاعات جانبی‌ای در گام آزمایش استفاده نمی‌شود. این مسئله موجب می‌شود روش پیشنهادی در شرایطی که اطلاعات جانبی در دسترس نیستند قابل به‌کارگیری باشد.

در ادامه جزئیات هر کدام از پیمانه‌ها توضیح داده شده است.

### ۳-۱- پیمانه جذابیت

هدف این پیمانه این است که به هر تصویر امتیازی نشان‌دهنده میزان جذابیت نسبت داده شود. برای این هدف در این قسمت طبقه‌بندی قرار داده شده است که بین تصاویر جذاب و غیرجذاب طبقه‌بندی انجام می‌دهد. این طبقه‌بند با استفاده از ویژگی‌های سطح پایین تصویر در مورد جذاب بودن یا عدم‌جذاب بودن یک تصویر تصمیم می‌گیرد. در گام آموزش، مجموعه‌ای از تصاویر جذاب و غیرجذاب به طبقه‌بند داده شده و طبقه‌بند آموزش داده شده در گام آزمایش برای تصمیم در مورد جذابیت تصاویر به کار گرفته می‌شود. در انتها، اطمینان تعلق هر تصویر به رده<sup>۱۲</sup> تصاویر جذاب به عنوان امتیاز جذابیت در نظر گرفته می‌شود.

که هم ارتباطات معنایی را برای خلاصه‌سازی در نظر می‌گیرد و هم پارامترهای کیفیت و جذابیت را لحاظ می‌کند.

### ۳- روش پیشنهادی

همان‌طور که در قسمت قبل عنوان شد، بیشتر روش‌های خلاصه‌سازی تصویر بر روی پوشش اطلاعاتی تمرکز داشته‌اند و روش‌های اندکی در ادبیات موضوعی وجود دارند که معیارهای دیگری مانند کیفیت و جذابیت را نیز در نظر گرفته‌اند. در روش پیشنهادی از دو دسته ویژگی استفاده شده است. دسته اول ویژگی‌های محتوایی هستند که به صورت ویژگی‌های سطح بالا بر روی ویژگی‌های تصویری سطح پایین تصویر تعریف می‌شوند. هدف ویژگی‌های محتوایی این است که اطلاعات موجود در تصاویر را مدل کنند.

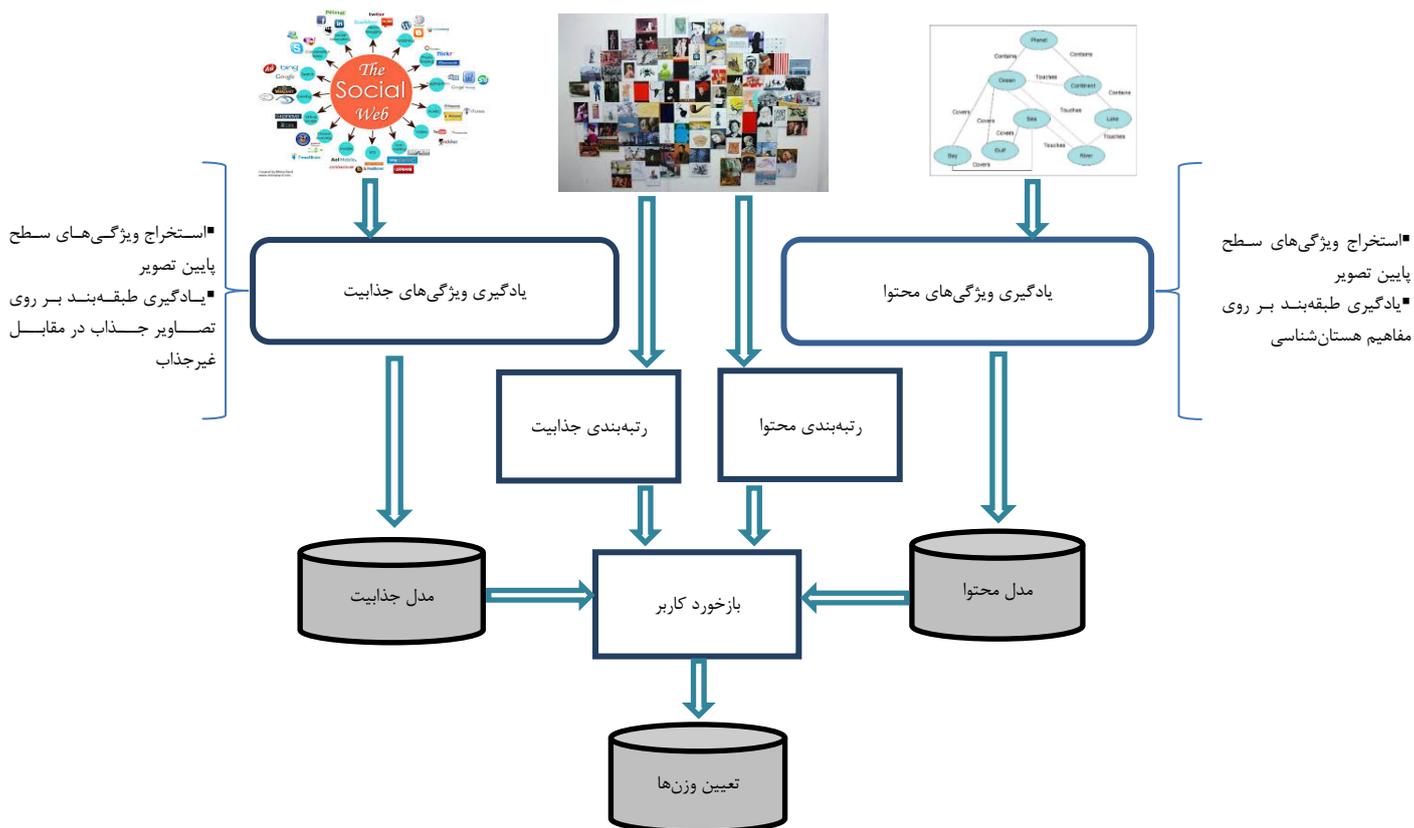
جدول ۱- مقایسه روش‌های خلاصه‌سازی تصویر

نام روش	حالت یا منبع اطلاعات	میزان نظارت	رهیافت
می [۲۲]	ویدئو	بدون نظارت	بهبودسازی معیار عددی
ناسی [۲۳]	ویدئو	بدون نظارت	آنالیز گراف
کامارو [۸]	تصویر + متن	بدون نظارت	بهبودسازی معیار عددی
شن [۷]	تصویر + مکان + زمان	بدون نظارت	خوشه‌بندی
ریاحی [۲]	تصویر	بدون نظارت	تحلیل گراف و روابط معنایی
فانگ [۲۴]	تصویر	با نظارت	یادگیری TOPIC ASPECT ها
شیاتشک [۱]	تصویر	با نظارت	یادگیری توابع SUB-MODULAR
روش پیشنهادی	تصویر	بدون نظارت	چند معیاره

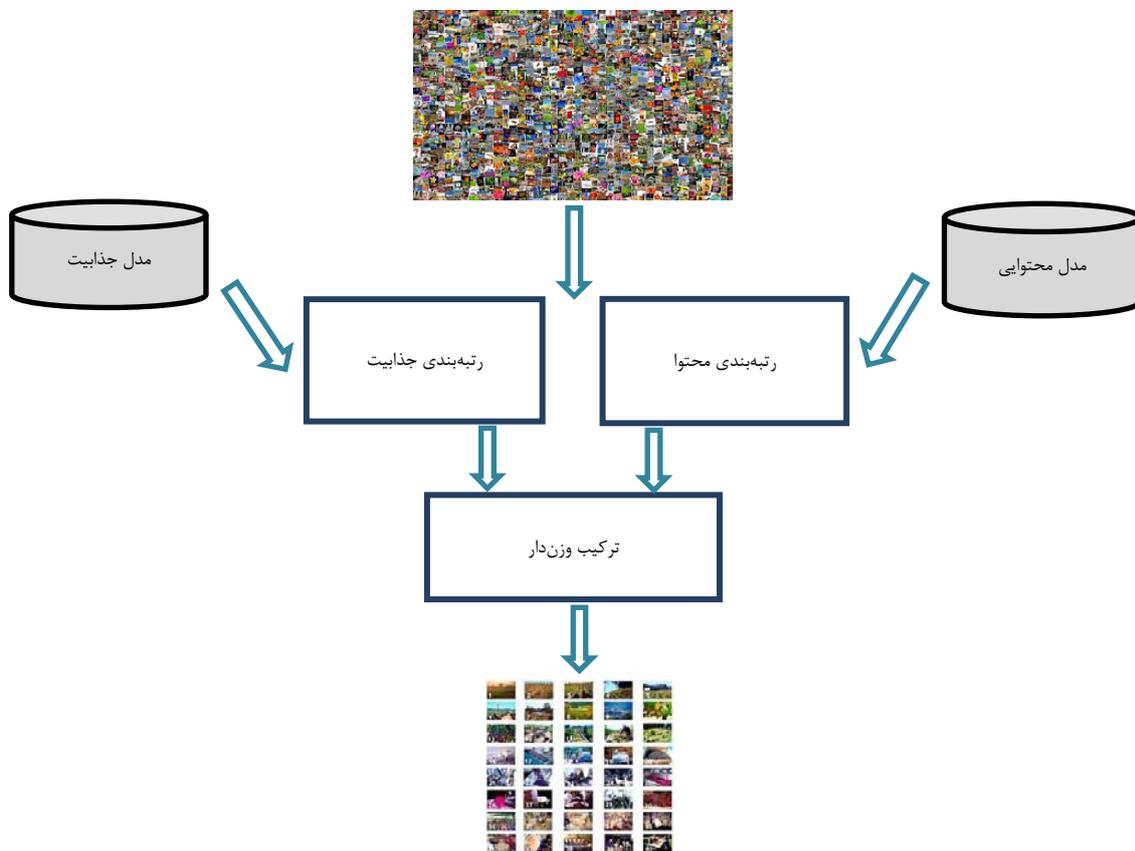
دسته دوم ویژگی‌ها، ویژگی‌های جذابیت هستند. این ویژگی‌ها به صورت ترکیبی از ویژگی‌های کیفیت، زیبایی و احساس بر روی ویژگی‌های سطح پایین تصویر تعریف می‌شوند.

روش پیشنهادی دارای یک گام<sup>۸</sup> آموزش است. در گام آموزش مدل‌های پیمانه محتوا و پیمانه جذابیت آموزش داده می‌شود و وزن‌دهی با استفاده از بازخورد کاربر انجام می‌شود. در گام آزمایش<sup>۹</sup> الگوریتم هر کدام از پیمانه‌ها بر روی تصاویر اعمال شده و نتایج آن‌ها با هم ترکیب می‌شود. گام آموزش در شکل ۱ نشان داده شده است.

ایده پیمانه محتوا بر مبنای مدل‌سازی معنایی اطلاعات است. هدف آن است که ویژگی‌های سطح بالای معنایی را از روی ویژگی‌های سطح پایین تصویری به دست آوریم. در این مقاله از یک بهبود انجام شده بر روی روش ارائه شده در [۲] استفاده کرده‌ایم. در این روش هستان‌شناسی دامنه به عنوان ورودی به سیستم داده می‌شود. سپس طبقه‌بندی طراحی می‌شود که تصاویر را به مفاهیم هستان‌شناسی دامنه طبقه‌بندی کند. هستان‌شناسی دامنه مفاهیم حائز اهمیت در یک دامنه و ارتباط بین آن‌ها را به سیستم خلاصه‌سازی تصویر معرفی می‌کند. فرض کنیم تصاویر اولیه مربوط به مکان‌های دیدنی یک شهر هستند. هستان‌شناسی دامنه، مفاهیم پراهمیت در یک دامنه مانند مکان‌های دیدنی مهم در یک شهر را به سامانه<sup>۱۰</sup> معرفی می‌کند. این مکان‌های دیدنی برای هر شهری متفاوت هستند به عنوان مثال مفاهیمی مانند amphitheater و fountain برای شهری مانند رم (مطابق با Colosseum و Trevi Fountain) و مفاهیمی مانند رودخانه و مسجد برای شهری مانند اصفهان در هستان‌شناسی دامنه وجود دارند. به بیان دیگر، هستان‌شناسی دامنه، دانش مورد نیاز درباره محتوا و زمینه تصاویر را



شکل ۱- روش پیشنهادی گام آموزش



شکل ۲- روش پیشنهادی گام آزمایش

تصویر [۲۸] به دست آورده می‌شود و وزن بصری میزان مجزا بودن موضوع از پس زمینه را مشخص می‌کند.

• میزان تاثیر شامل وضوح و اثربخشی بوده و بیان‌کننده‌ی این است که چه عواطف مشخصی از تصویر احساس می‌شود و یا به عبارت دیگر تصویر چه احساساتی را برانگیخته می‌کند. در این مقاله از ویژگی‌های خط و رنگ استفاده شده است. مطالعات نشان می‌دهد، خطوط یک تصویر باعث ایجاد تاثیر است عاطفی می‌شوند [۲۹]. خطوط افقی با یک افق ثابت، آرامش و احساس امنیت را القا می‌کنند، خطوط عمودی واضح و مستقیم هستند و باعث ایجاد حس کرامت و ابدیت می‌شوند. از طرف دیگر، خطوط کج ناپایدار بوده و احساس پویایی را ایجاد می‌کنند. خطوط با جهت‌های بسیار مختلف بیان‌کننده‌ی بی‌نظمی و سردرگمی هستند. هر چه خط بلندتر، ضخیم‌تر و غالب‌تر باشد اثر روانی القا شده‌ی آن قوی‌تر می‌شود. برای پیدا کردن خطوط ایستا و پویا شیب خطوط در تصاویر با استفاده از تبدیل هاف به دست آورده شده و خطوط براساس زاویه آنها به کلاس ایستا و پویا تقسیم می‌شوند [۳۰]. برای ویژگی‌های رنگ سه دسته ویژگی زیر استفاده شده‌اند:

**ویژگی اول احساسی:** ویژگی احساسی اول القاکننده‌ی جفت کلمات معنادار سبک- سنگین<sup>۲۹</sup>، سرزنده- خسته‌کننده<sup>۳۰</sup>، پرتراوت- ویران<sup>۳۱</sup> و سرد- گرم<sup>۳۲</sup> است. بر پایه‌ی تئوری روانشناسی رنگ‌ها، رنگ‌های سایه روشن احساسات سبک، سرزنده، پرتراوت و گرم را منتشر می‌کنند. رنگ‌های سرد و گرم تأثیر متفاوتی روی احساس می‌گذارند، رنگ‌های گرم القاکننده‌ی عواطف شادی، سرزندگی و طراوت بوده در حالی که رنگ‌های سرد القاکننده‌ی عواطف ناراحتی، کسل‌کنندگی و افسردگی هستند.

**ویژگی دوم احساسی:** ویژگی احساسی دوم ارتباط قوی‌ای با جفت کلمات معنادار سخت- نرم<sup>۳۳</sup>، عالی- معتدل<sup>۳۴</sup> و پرچلوه- ظریف<sup>۳۵</sup> دارد. براساس تئوری روانشناسی رنگ‌ها، تصاویر با اشباع بالا القاکننده‌ی حس اعتدال و ملایمت و ظرافت می‌باشد، در حالی که رنگ‌های با اشباع پایین القاکننده‌ی احساسات عالی و پرچلوه است؛ به‌طور مشابه احساسات سخت- نرم ارتباط نزدیکی با تباین دارند [۳۰].

**ویژگی سوم احساسی:** ویژگی احساسی سوم ارتباط قوی‌ای با کلمات معنادار تخیل- واقعیت و قطعی- غیرقطعی دارد. براساس تئوری دریافت بصری، تباین و روشنایی بالا و لبه‌های تیز القاکننده‌ی واقعیت و قطعی بوده در حالی که تباین و روشنایی پایین و لبه‌های پهن‌تر القاکننده‌ی عواطف غیرقطعی و تخیل است [۳۰]. جزئیات این ویژگی‌ها در جدول ۲ نمایش داده شده است.

بردار ویژگی به دست آمده براساس این جدول، یک بردار ۴۹ بعدی است. سپس از تجزیه و تحلیل مولفه‌های اصلی<sup>۳۶</sup> بر روی بردار ویژگی ساخته شده برای کاهش بعد استفاده شده و از ماشین بردار پشتیبان با کرنل خطی برای طبقه‌بندی بین رده‌های جذاب و غیرجذاب استفاده شده است. با استفاده از ماشین بردار پشتیبان تصاویر مجموعه داده با برجسب‌های جذاب و یا غیرجذاب طبقه‌بندی می‌شوند.

برجسب تصاویر با بررسی  $I_i \in [-1, 1]$  به وسیله طبقه‌بند پیش‌بینی می‌شود که در آن  $I_i$  فاصله نرمال شده خروجی ماشین بردار پشتیبان تا خط جداساز است. در این مقاله با استفاده از تابع سیگموئید بر روی  $I_i$  میزان جذابیت تصویر به صورت زیر محاسبه شده است:

$$A_{x_i} = \frac{1}{1+e^{-I_i}} \quad (۶)$$

که این مقدار به عنوان امتیاز جذابیت یک تصویر در نظر گرفته شده است.

در اینجا اولین مسئله فراهم کردن تصاویر جذاب و غیرجذاب برای آموزش است. در روش پیشنهادی برای حل این موضوع از بستر شبکه‌های اجتماعی استفاده شده است. کاربران در شبکه‌های اجتماعی می‌توانند بر روی تصاویر فعالیت انجام دهند، به عنوان مثال وقتی یک تصویر برای آنها جذاب است کاربر می‌تواند آن را نشانه کند (مانند پسند<sup>۳۷</sup> در فیس‌بوک و یا علاقه<sup>۳۸</sup> در فلیکر). تحقیقات جدید نشان می‌دهد که این اطلاعات می‌توانند برای تعیین اینکه یک تصویر دارای جذابیت<sup>۳۹</sup> هست یا خیر مورد استفاده قرار بگیرند [۶]. هم‌چنین در برخی تحقیقات گزارش شده است که این ویژگی‌ها بعضاً از ویژگی‌های متنی و تصویری مهم‌تر هستند [۲۵]. پدرو و همکاران هم چنین گزارش می‌کنند که بین تعداد نشانه‌ها و بقیه رفتارهای اجتماعی مانند به اشتراک‌گذاری یا یادداشت همبستگی<sup>۴۰</sup> وجود دارد [۲۶]. بنابراین در این مقاله تصمیم گرفتیم از نشانه‌ها برای پیدا کردن تصاویر جذاب استفاده کنیم. به این ترتیب تصاویر به دو مجموعه جذاب و غیرجذاب براساس تعداد نشانه‌ها تقسیم می‌شوند.

برای انجام این کار یک مقدار آستانه برای تعداد نشانه‌ها تعریف می‌شود. به این ترتیب تصاویری که تعداد نشانه‌هایشان بالاتر از آستانه است به عنوان تصاویر جذاب و تصاویری که تعداد نشانه‌هایشان پایین‌تر از آستانه است به عنوان تصاویر غیرجذاب معرفی می‌شوند.

مسئله بعد مقدار آستانه برای نشانه‌ها می‌باشد. کردانی و ابراهیمی مقدم در مطالعه‌ی نشان داده‌اند که مقدار آستانه به محتوای تصاویر مورد نظر وابسته است [۶]. آن‌ها نشان می‌دهند که انتخاب یک مقدار آستانه سراسری امکان‌پذیر نبوده و مقدار آستانه باید با توجه به محتوای تصاویر انتخاب شود. هم‌چنین در این مطالعه آنها نشان داده‌اند که میانگین تعداد نشانه‌ها را می‌توان به عنوان آستانه برای تصمیم‌گیری در مورد جذابیت تصاویر استفاده کرد. بنابراین در این مقاله نیز ما مقدار آستانه را برای هر دامنه از تصاویر به طور جداگانه برابر با میانگین تعداد نشانه‌ها انتخاب کرده‌ایم.

با داشتن دو دسته از تصاویر جذاب و غیرجذاب از ویژگی‌های سطح پایین زیر برای آموزش جذابیت استفاده شده است. این ویژگی‌ها در سه دسته کیفیت ادراکی<sup>۴۱</sup>، حساسیت زیبایی شناختی<sup>۴۲</sup> و میزان تاثیر<sup>۴۳</sup> جای می‌گیرند و برای اولین بار در [۲۷] معرفی شده‌اند:

• کیفیت ادراکی ویژگی‌هایی از تصویر را در نظر می‌گیرد که کیفیت را تحت تاثیر قرار می‌دهند مانند روشنایی<sup>۴۴</sup>، درخشندگی<sup>۴۵</sup>، تباین<sup>۴۶</sup>، رنگ، وضوح و ماتی<sup>۴۷</sup>. روشنایی در دو بعد، درخشندگی<sup>۴۸</sup> و تباین<sup>۴۹</sup> تعریف می‌شود. درخشندگی به صورت میانگین شدت روشنایی<sup>۵۰</sup> پیکسل‌های تصویر سیاه- سفید و تباین، واریانس شدت روشنایی پیکسل‌های تصویر سیاه- سفید می‌باشند. نقشه برجسته قسمت‌های مهم‌تر یک تصویر را نشان می‌دهد، برای استخراج برخی ویژگی‌ها از تصویر ضرب شده در نقشه برجسته استفاده می‌شود. منطقه موضوعی<sup>۵۱</sup> تصویر به ناحیه‌ای گفته می‌شود که چگالی اشیاء موجود در تصویر در آن منطقه بیشتر است، بنابراین کلیات تصویر از آن قابل برداشت است. لذا این منطقه از اهمیت بالایی برخوردار است. به همین دلیل برخی ویژگی‌ها از منطقه موضوع تصویر استخراج می‌شوند.

• حساسیت زیبایی به این مسئله می‌پردازد که آیا یک تصویر از نظر زیبایی جذابیت دارد یا خیر. برای این دسته از ویژگی‌ها قانون سوم‌ها<sup>۵۲</sup>، سادگی و وزن بصری در نظر گرفته شده است. قانون سوم‌ها در عکاسی تاکید می‌کند که قرار گرفتن عناصر مهم تصویر در محل برخورد خطوط افقی و عمودی‌ای که تصویر را به سه قسمت تقسیم می‌کنند، بیش‌ترین توجه را جلب می‌کند. برای سادگی توزیع لبه‌ها در تصویر اولیه و تصویر ضرب در نقشه برجسته تصویر و همچنین مقادیر متمایز خلوص در تصویر اولیه و منطقه موضوع اصلی

جدول ۲- ویژگی‌های جذابیت تصویر

ویژگی	دسته	نام ویژگی	تعداد بعد	توضیحات
کیفیت ادراکی		براقی <sup>۲۷</sup>	۲	میانگین (روشنایی) و واریانس (تباين) شدت تصاویر خاکستری [۲۶]
		رنگ	۷	میانگین و انحراف معیار اشباع <sup>۲۸</sup> و خلوص <sup>۲۹</sup> [۲۶، ۲۹]. تباين رنگها [۲۶] رنگارنگی <sup>۳۰</sup> (تفاوت رنگ تصویر در مقابل خاکستری) [۲۶] طبیعی بودن <sup>۳۱</sup> (ضریبی از حضور پوست، آسمان و چمن در تصویر که بعد از انتقال تصویر به فضای LUV به دست آورده می‌شود) ویژگی‌های رنگی [۳۱]
		وضوح و ماتی <sup>۳۲</sup>	۳	وضوح: میانگین و انحراف معیار لاپلاسین تصویر که با توجه به میانگین درخشندگی <sup>۳۳</sup> محلی نرمال شده است [۲۶]. تیری: فرکانس توزیع تبدیل FFT تصویر [۲۸]
		کیفیت موضوع	۱۲	برق، رنگ و تیزی تصویر ضرب شده در نقشه برجسته <sup>۳۴</sup> تصویر، و میزان تاری در منطقه موضوع تصویر
حساسیت زیبایی		قانون سومها	۱	قرار دادن موضوع تصویر در مکانی نزدیک نقطه قوی مرکزی [۳۲]
		سادگی	۴	توزیع لبه‌ها در تصویر اولیه و تصویر ضرب در نقشه برجسته تصویر، شمارش مقادیر متمایز خلوص در تصویر اولیه و منطقه موضوع اصلی تصویر [۲۸]
میزان تاثیر		وزن بصری	۱	میزان مجزا بودن موضوع اصلی از پس‌زمینه [۳۳]
		پویایی	۲	تعداد نسبی و طول خطوط ایستا و پویا [۳۴]
		احساس رنگ	۱۷	هیستوگرام طراحی شده برای بررسی تاثیر احساسی رنگ تصویر: میران سرد یا گرم بودن رنگ تصویر، سخت یا نرم بودن رنگ تصویر، واقعی یا تخیلی بودن رنگ تصویر [۳۵]

### ۳-۲- پیمانانه محتوا

برای مدل کردن اطلاعات محتوایی از هستان‌شناسی دامنه استفاده شده است. هستان‌شناسی دامنه حاوی مفاهیم مهم در یک دامنه و ارتباط بین آنها است. به بیان دیگر، هستان‌شناسی دامنه دانش عمومی در زمینه تصاویر را به سیستم می‌دهد. شکل ۳ یک مثال از هستان‌شناسی مربوط به تصاویر شهر پاریس را نشان می‌دهد. برای انجام خلاصه‌سازی محتوایی تصاویر، در ابتدا تصاویر مجموعه اولیه با استفاده از طبقه‌بند به مفاهیم مختلف هستان‌شناسی دامنه طبقه‌بندی شده و پس از آن ویژگی‌های هستان‌شناسی بر روی خروجی طبقه‌بند تعریف می‌شوند.

برای انجام طبقه‌بندی از یک روش طبقه‌بندی سلسله مراتبی استفاده شده است. این روش در [۳۶] پیشنهاد داده شده است. به این ترتیب که در هر نود در هستان‌شناسی یک طبقه‌بند قرار داده شده است. وظیفه هر طبقه‌بند آن است که بین رده‌هایی که زیر مفهوم مفهوم جاری هستند طبقه‌بندی انجام دهد. در هر نود در هستان‌شناسی، طبقه‌بندی بین زیرمفاهیم مفهوم جاری انجام شده و نودی که تصویر به آن طبقه‌بندی می‌شود برای کاوش مجدد انتخاب می‌شود. این کار به صورت تکراری انجام می‌شود تا به مفاهیم برگ برسیم. در انتها هر تصویر به دنباله‌ای از مفاهیم در هستان‌شناسی طبقه‌بندی شده و هر تصویر با سبب کلمات در فضای هستان‌شناسی نمایش داده می‌شود.

برای هر تصویر یک مجموعه از ویژگی‌های هستان‌شناسی به صورت زیر تعریف می‌شوند [۲].

$$F_i = (C_i^1, \dots, C_i^j, \dots, C_i^c) \quad (7)$$

که در آن  $C_i^j$  مفهومی در فضای هستان‌شناسی است و  $C_i^j$  ویژگی هستان‌شناسی تصویر  $i$  با توجه به مفهوم  $j$  است که به صورت زیر محاسبه می‌شود و  $c$  تعداد مفاهیم است [۲].

$$C_i^j = \alpha_j \times Conf_i^j \quad (8)$$

در این رابطه  $Conf_i^j$  اطمینان به تعلق تصویر  $i$  به مفهوم  $j$  است. که با استفاده از تابع سیگموئید بر روی خروجی طبقه‌بند به دست می‌آید. اگر  $L_i^j$  فاصله نرمال شده خروجی ماشین بردار پشتیبان تا خط جداساز برای تعلق تصویر  $i$  به کلاس متعلق به مفهوم  $j$  باشد در این مقاله با استفاده از تابع سیگموئید بر روی  $L_i^j$  اطمینان به تعلق تصویر  $i$  به کلاس متناظر با مفهوم  $j$  مطابق معادله زیر به دست آورده شده است.

$$Conf_i^j = \frac{1}{1 + e^{-L_i^j}} \quad (9)$$

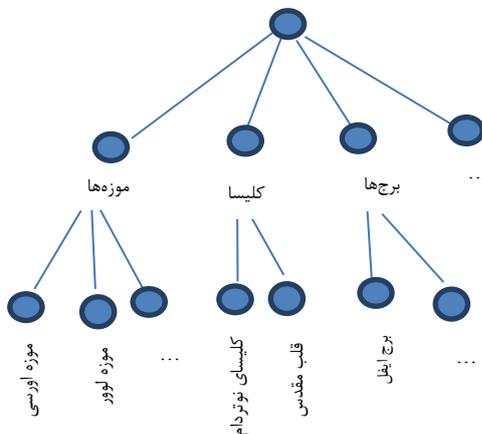
$\alpha_j$  به کمک رابطه ۹ محاسبه می‌شود و برابر است با جمع اطمینان به تعلق تصاویر به رده  $j$  تقسیم بر جمع اطمینان به تعلق تصاویر به همه رده‌ها [۲].

$$\alpha_j = \frac{N_j}{N_t} = \frac{\sum_{i=1}^n C_i^j}{\sum_{j=1}^c \sum_{i=1}^n C_i^j} \quad (10)$$

دلیل استفاده از  $\alpha_j$  این است که تعداد رخداد مفاهیم در محاسبه ویژگی‌های هستان‌شناسی در نظر گرفته شود. به این ترتیب مفاهیمی که بیشترین رخداد را دارند تاثیر بیشتری روی ویژگی‌های هستان‌شناسی دارند. تصاویر خلاصه باید نماینده مجموعه اولیه باشد پس به مفاهیم پررخداد شانس بیشتری برای انتخاب داده می‌شود [۲].

به‌عنوان مثال، بردار ویژگی تصاویر مرتبط با شهری مانند پاریس، شامل مفاهیمی که در هستان‌شناسی آن شهر هستند مانند (برج، کلیسا، موزه و ...) می‌باشد. برای هر تصویر میزان تعلق آن به مفاهیم مختلف ( $Conf_i^j$ ) به‌دست آورده شده و با استفاده از آن، بردار ویژگی برای هر تصویر ساخته می‌شود. در کنار آن پارامتر  $\alpha$  برای هر مفهوم محاسبه شده که برابر اطمینان به تعلق تصاویر به یک مفهوم تقسیم بر اطمینان به تعلق تصاویر به بقیه مفاهیم است. در انتها بردار ویژگی تصویر از ضرب  $Conf_i^j$  در  $\alpha_j$  به دست می‌آید (رابطه ۸).

در این مقاله از نودهای میانی طبقه‌بند برای ساختن ویژگی‌های هستان‌شناسی استفاده شده است.



شکل ۳- هستان‌شناسی دامنه مربوط به شهر پاریس

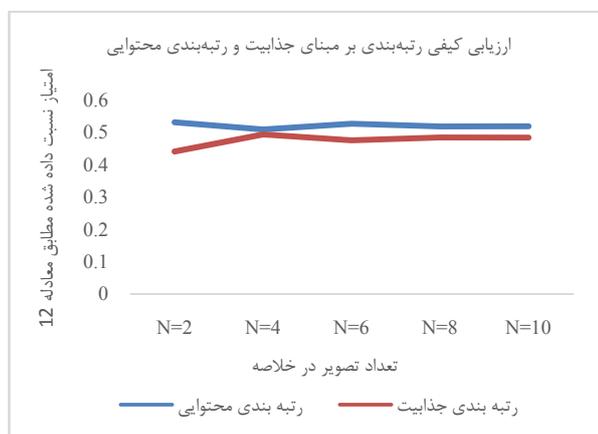
$$\text{Weight}_{\text{set}} = \frac{\sum_{j=1}^U (\text{grade}_j(1) + \sum_{i=2}^M \frac{\text{grade}_j(i)}{\log_2 i})}{U * M} \quad (13)$$

بعد از انجام ارزیابی توسط کاربران یک امتیاز کل برای میزان نماینده بودن مطابق معادله‌ی ۱۲ نسبت داده می‌شود. در معادله‌ی ۱۲ منظور از Set خلاصه محتوایی و یا خلاصه جذابیت است، U تعداد ارزیابی‌های موجود یا به بیان دیگر تعداد کاربرانی است که در آزمایش بازخورد کاربر حضور داشته‌اند، M تعداد کل تصاویر موجود در خلاصه است و  $\text{grade}_j(i)$  امتیازی است که کاربر j به تصویر iam نسبت داده است. این روش نمره‌دهی از روش تخفیف افزایش تجمعی<sup>۴۵</sup> در بازایی اطلاعات [۳۷] و با در نظر گرفتن این مسئله که روش‌های بازایی قوی‌تر باید تصاویر را در مرتبه‌ی بالاتری بازایی کنند به کار گرفته شده است. وزن نهایی نرمال شده به دست آمده از رابطه ۱۳ اهمیت هر مجموعه را نشان می‌دهد. این مقدار برای مقادیر متناظر آلفا و بتا در نظر گرفته می‌شود.

$$\alpha = \frac{\text{Weight}_{\text{ContentRankedSet}}}{\text{Weight}_{\text{AttractivenessRankedSet}} + \text{Weight}_{\text{ContentRankedSet}}}$$

$$\beta = \frac{\text{Weight}_{\text{AttractivenessRankedSet}}}{\text{Weight}_{\text{AttractivenessRankedSet}} + \text{Weight}_{\text{ContentRankedSet}}} \quad (14)$$

شکل ۴ نتایج ارزیابی‌های کاربران را نشان می‌دهد. محور عمودی وزن‌های تعریف شده در معادله ۱۲ را نشان داده و محور افقی اندازه‌های متفاوت خلاصه را نشان می‌دهد. از شکل این نتیجه‌گیری حاصل می‌شود که هر دوی این ویژگی‌ها دارای اهمیت یکسانی هستند ولی ویژگی‌های محتوایی را می‌توان حائز اندکی اهمیت بیشتر دانست.



شکل ۴- ارزیابی خلاصه‌سازی محتوایی و جذابیت

#### ۴- ارزیابی روش پیشنهادی

در این بخش به ارزیابی روش پیشنهادی خواهیم پرداخت. بدین ترتیب ابتدا در مورد پیش‌فرض‌های روش و مجموعه داده مورد استفاده شرح خواهیم داد. سپس به ارزیابی روش پیشنهادی خواهیم پرداخت. در قسمت اول نتایج روش پیشنهادی را با خلاصه‌های ساخته شده به وسیله‌ی انسان مقایسه کرده‌ایم و در بخش دوم پیمانه محتوا را با معیارهای کمی ارزیابی کرده‌ایم.

به عنوان معیار فاصله، از فاصله کسینوسی بین بردار ویژگی تصویر و بردار ویژگی مجموعه استفاده شده است.

$$\text{ContentScore}_{iC} = \frac{F_i \cdot F_C}{|F_i| |F_C|} \quad (11)$$

که در آن  $F_i$  و  $F_C$  بردارهای ویژگی مجموعه و تصویر هستند. بردار ویژگی مجموعه از میانگین‌گیری بردار ویژگی تصاویر به دست می‌آید و نقطه نشانگر ضرب داخلی بین بردارهای ویژگی است. این معیار فاصله نشان می‌دهد که یک تصویر چقدر از اطلاعات کل مجموعه تصاویر را در فضای هستان‌شناسی پوشش می‌دهد. هر چه مقدار ContentScore در فضای هستان‌شناسی بزرگتر باشد نشان‌دهنده آن است که آن تصویر پوشش اطلاعاتی بالاتر و نمایندگی بالاتری دارد.

#### ۳-۳- تجمیع و خلاصه‌سازی

هدف این پیمانه جمع‌بندی بین امتیازهای محتوایی و جذابیت و انجام خلاصه‌سازی است. ایده‌ی به کار گرفته شده در این مقاله، استفاده از جمع وزن‌دار امتیازهای جذابیت و معنایی است. ورودی این قسمت مجموعه‌ی تصاویر ورودی و امتیازهای جذابیت و معنایی و خروجی آن لیست رتبه‌بندی شده‌ی تصاویر براساس امتیاز کلی آنها است. به این ترتیب جمع وزن‌دار امتیازهای معنایی و جذابیت به دست می‌آید و به متغیر TotalScore مقدار دهی می‌شود.

$$\text{TotalScore}_i = \alpha \times \text{ContentScore}_i + \beta \times \text{AttractivenessScore}_i \quad (12)$$

$\alpha$  و  $\beta$  وزن‌هایی هستند که به امتیازهای محتوایی و جذابیت نسبت داده می‌شود و تصاویر خلاصه براساس مقدار TotalScore انتخاب می‌شوند. برای تصمیم‌گیری در مورد آلفا و بتا یک آزمایش بازخورد از کاربران انجام شده است. جزئیات این آزمایش در ادامه مطرح شده است. نتایج حاکی از آن است که هر دوی ویژگی‌های محتوایی و جذابیت تقریباً به یک اندازه در خلاصه‌سازی مجموعه تصاویر مهم هستند.

آزمایش انتخاب آلفا و بتا بدین صورت طراحی شده است که به ازای ۱۲ شهر و هر شهر ۱۰۰ تصویر (در مجموع ۱۲۰۰ تصویر) از میان تصاویر موجود برای آموزش انتخاب شدند. این تعداد تصویر با توجه به دو ملاحظه زیر انتخاب شده‌اند:

- در دسترس بودن کاربر انسانی برای ارزیابی
- مطابقت با روش‌های مشابه یادگیری در حوزه خلاصه‌سازی تصویر، روش‌های مشابه در حوزه یادگیری با نظارت نیز از حدود صد تصویر به ازای هر مفهوم استفاده می‌کنند. در نظر گرفتن بیشتر از صد تصویر در خلاصه‌سازی برای کاربر انسانی دشوار است [۱].

تصاویر به پیمانه‌های استخراج ویژگی‌های محتوایی و ویژگی‌های جذابیت داده شده و ده تصویر با امتیاز بالای محتوایی و جذابیت به دست آورده می‌شود. تصاویر بازایی شده به کاربران نشان داده شد و از آنها خواسته شده که به مجموعه‌ی اولیه نگاه کرده و سپس یک نمره‌ی میزان نماینده بودن به تصویر بازایی شده نسبت دهند. امتیاز میزان نماینده بودن با استفاده از جدول ۳ نسبت داده شده است.

جدول ۳- امتیازدهی به تصاویر براساس میزان نماینده بودن

غیر نماینده	نمایندگی پایین	نمایندگی متوسط	نمایندگی بالا
۰	۱	۲	۳

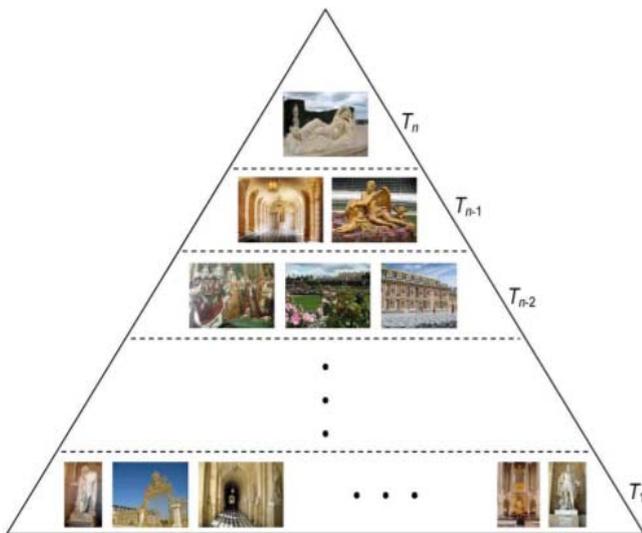
داده و از آنها خواستیم مجموعه‌ی کوچکی از تصاویر را به عنوان خلاصه انتخاب کنند. در انتها خلاصه‌ی تولید شده به وسیله‌ی روش پیشنهادی را با خلاصه‌های دستی تولید شده به وسیله‌ی کاربران مقایسه کرده‌ایم.

در این قسمت به روشی نیاز بود که بتواند نتایج خلاصه‌سازی را با مجموعه‌ای از خلاصه‌های انسانی مقایسه کند. رودیناک و همکاران [۴] نشان داده‌اند که معمولاً توافق پایینی بین خلاصه‌های تولید شده به وسیله کاربران انسانی وجود دارد. تعداد تصاویر مشترک بین دو خلاصه انسانی میانگینی نزدیک به یک دارد. این مسئله انتخاب رهیافت‌هایی [۳۹]- که معیارهایی برای خلاصه‌های VERT [۳۸] و ROUGE مثل دستی و اتوماتیک در دامنه‌های متن و ویدیو هستند- را با سختی روبرو می‌کند.

بر همین اساس رودیناک و همکاران روشی هرمی برای مقایسه‌ی مجموعه‌های از خلاصه‌های دستی و خلاصه‌ی اتوماتیک ارائه داده‌اند. در این روش تصاویر در یک هرم قرار داده می‌شوند- شکل ۵- خلاصه nم هرم هستند در n به این صورت که تصاویری که در لایه دستی حضور دارند. به این ترتیب تصاویر در لایه‌ی اول هرم تنها توسط یک کاربر انتخاب شده‌اند و هر چه در هرم بالاتر می‌رویم تصاویر توسط n کاربر وجود n کاربران بیشتری انتخاب شده‌اند، تا اینکه در لایه دارد که توسط همه‌ی کاربران انتخاب شده است. رودیناک و همکاران تصویر در آن که حداکثر s با تعداد S عنوان می‌کنند که یک خلاصه مقدار ارزیابی شده را می‌گیرد باید به ترتیب زیر ساخته شده باشد [۴] و معیار d را برای ارزیابی خلاصه دستی ساخته شده به کار می‌برند. این معیار تعداد تصاویر مشترک بین خلاصه‌ی اتوماتیک و تصاویر موجود در هر لایه هرم را می‌شمارد و در عین حال به تصاویر لایه‌های بالای هرم امتیاز بیشتری نسبت می‌دهد [۴]:

$$d_{\max} = \sum_{i=0}^n i \times |T_i| + \theta \times (N_R - \sum_{i=0}^n |T_i|) \quad (15)$$

$$\theta = \max_i \left( \sum_{j=1}^n |T_j| \right) \geq N_R$$



شکل ۶- روش هرم پیشنهادی در رودیناک و همکاران [۴]

$$d = \frac{1}{d_{\max}} \times \sum_{i=1}^n i \times |T_i \cap R| \quad (16)$$

ما نیز در این مقاله از این روش برای ارزیابی روش پیشنهادی استفاده کرده‌ایم. روش پیشنهادی در این مقاله با روش‌های زیر مقایسه شده است.

#### ۱-۴- پیش فرض‌ها

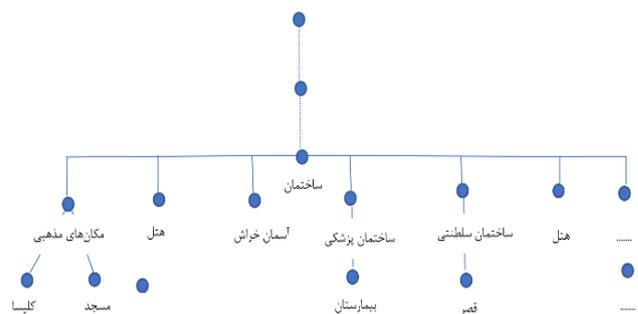
روش پیشنهادی تنها به خلاصه‌سازی می‌پردازد و برای از بین بردن داده‌های نویزی روشی پیشنهاد می‌کند. همچنین فرض بر این است که تصاویر حول موضوع - یا موضوعات مشخصی - هستند. بنابراین اگر داده‌ها نویزی باشند و یا تصاویر حول موضوع مشخصی نباشد روش به درستی عمل نخواهد کرد.

#### ۲-۴- مجموعه داده

متاسفانه مجموعه داده استاندارد با تعداد تصویر زیاد برای خلاصه‌سازی مجموعه تصویر تا به حال ارائه نشده است، گرچه اخیراً تلاش‌هایی برای ایجاد مجموعه داده استاندارد خلاصه برای تعداد صدها تصویر انجام شده است [۱]. به عنوان مثال [۱] از ۱۴۰۰ تصویر مرتبط با ۱۴ مکان مختلف استفاده کرده است که به وسیله کاربر انسانی خلاصه‌سازی شده است، اما نویسندگان مجموعه داده خود را منتشر نکرده‌اند.

تصاویری که برای آزمایش روش پیشنهادی به کار گرفته شده است ۱۲۰۰۰ تصویر مربوط به ۱۲ شهر مختلف است که از فلیکر انتخاب شده‌اند. سپس از FlickrAPI برای استخراج و تحلیل آماری ۱۰۰۰۰ تصویر بالاتر مربوط به هر شهر و انتخاب آستانه برای مقدار جذابیت استفاده شده است. برای هر شهر ۵۰۰ تصویر جذاب و ۵۰۰ تصویر غیرجذاب ارزیابی شده و با استفاده از آن طبقه‌بند جذابیت آموزش داده شده است.

هستان‌شناسی مربوط به هر شهر از روی DBpedia استخراج شده و سپس برای هر مفهوم، مفهوم منطبق با آن از وردنت ارزیابی شده است. نودهای برگ در هستان‌شناسی استخراج شده و به طور بازگشتی پدر آن نودها انتخاب شده است. در حالتی که برای یک نود بیش از یک مسیر به ریشه وجود داشته، مسیر دارای بیشترین همپوشانی با مسیرهای دیگر انتخاب شده است. در انتها از ImagNet برای آموزش طبقه‌بند استفاده کردیم. جزئیات بیشتر در مورد روش استخراج هستان‌شناسی و نگاشت آن به وردنت در [۲] توضیح داده شده است. به عنوان نمونه یک زیرگراف استخراج شده از وردنت برای مفهوم ساختمان در شکل ۵ نشان داده شده است.



شکل ۵- نمونه‌ای از زیر گراف استخراجی از وردنت برای مفهوم ساختمان

#### ۳-۴- مقایسه با خلاصه دستی

همان‌طور که ذکر شد در قسمت اول مقایسه، نتایج روش پیشنهادی را با خلاصه‌های ساخته شده به وسیله انسان مقایسه کرده‌ایم. برای انجام این کار ۱۲۰۰ تصویر مربوط به ۱۲ شهر مختلف انتخاب شده است (۱۰۰ تصویر برای هر شهر). تصاویر اولیه را به مجموعه‌ای از ۱۶ دانشجوی تحصیلات تکمیلی نمایش

شکل ۸ مجموعه‌ای از ۳۲ تصویر مربوط به شهر پاریس را نمایش می‌دهد و شکل ۹ نتایج خلاصه‌سازی با روش پیشنهادی را نشان می‌دهد. همان‌طور که از شکل مشخص است مجموعه‌ی اولیه شامل تصاویر تکراری زیادی است که طیف وسیعی از معیارهای جذابیت را پوشش می‌دهد و مجموعه‌ی انتخابی یک مجموعه ترکیبی است که هم پوشش اطلاعاتی و هم معیارهای جذابیت را در نظر گرفته است.

#### ۴-۴- ارزیابی پیمانانه محتوا

در این بخش به ارزیابی پیمانانه محتوا می‌پردازیم. داده‌های تست برای این قسمت از ۱۲۰۰۰ تصویر انتخابی از فلیکر انتخاب شده‌اند. برای ارزیابی پیمانانه محتوا از معیار معرفی شده در [۲] استفاده می‌کنیم. در روش پیشنهادی در [۲] از نوده‌های انتهایی در هستان‌شناسی برای ساخت ویژگی‌های هستان‌شناسی استفاده شده است، اما در این مقاله از همه نوده‌های میانی از ریشه تا برگ به عنوان ویژگی‌های تاثیرگذار در ساخت ویژگی‌های هستان‌شناسی استفاده شده است. معیار معرفی شده به صورت زیر است.

$$\gamma = \frac{\text{Coverage}}{\text{Redundancy}} \quad (17)$$

Coverage یا پوشش، این معیار شباهت مجموعه‌ی انتخابی را با مجموعه‌ی اولیه اندازه می‌گیرد و Redundancy یا تکرار، شباهت مجموعه‌ی اولیه را با خودش در نظر می‌گیرد. تصاویر نماینده‌ی انتخابی باید مشابه با مجموعه اولیه و غیرمشابه با همدیگر باشند و به عبارت دیگر پوشش بالا و تکرار پایین نشان‌دهنده‌ی سیستم خلاصه‌سازی بهتر است. در این مقاله از معیار  $\gamma$  برای ارزیابی روش خلاصه‌سازی استفاده شده است. هر چه معیار گاما بزرگتر باشد روش پیشنهادی کارا تر است.

محاسبه‌ی پوشش و تکرار به روش دستی دشوار است. بنابراین در [۲] پوشش و تکرار با شمارش تعداد انطباق‌های SIFT بین دو تصویر شمرده شده‌اند<sup>۴۶</sup>. شکل ۹ نتایج مقایسه روش پیشنهادی با روش به کار گرفته شده در [۲] را نمایش می‌دهد. نسبت  $\gamma$  را برای اندازه‌های مختلف خلاصه به دست آورده‌ایم. اندازه‌ی خلاصه بین یک درصد تا ۵ درصد اندازه مجموعه اولیه تغییر داده شده است. همان‌طور که از شکل پیداست روش پیشنهادی که از همه نوده‌های هستان‌شناسی برای ساخت ویژگی استفاده می‌کند در مقایسه با روش [۲] که تنها از نوده‌های برگ برای ساخت ویژگی‌های هستان‌شناسی استفاده می‌کند نتایج بهتری در خروجی‌ها دارد. هم‌چنین در مقایسه با سایر روش‌ها که تنها از معیار عددی برای خلاصه‌سازی استفاده می‌کنند کارا تر عمل کرده است. بدین ترتیب روش پیشنهادی در مقایسه با روش‌های جاری خلاصه‌سازی تصویر توانسته است به بهبود سی و پنج درصدی پوشش اطلاعاتی در خلاصه کردن مجموعه تصویری به میزان یک درصد حجم اولیه و بهبود چهار درصدی پوشش اطلاعاتی در خلاصه کردن مجموعه تصویری به میزان پنج درصد حجم اولیه دست یابد.

اکثر این روش‌ها به بهبود معیارهای عددی و تصویری پرداخته و از الگوریتم‌های حریمانه و یا مرکزیت گراف برای خلاصه‌سازی استفاده می‌کنند. روش پیشنهادی با استفاده از هستان‌شناسی دامنه، هم از معیارهای معنایی برای خلاصه‌سازی استفاده می‌کند و هم اطلاعاتی از فضای ویژگی تصاویر را با استفاده از هستان‌شناسی در اختیار دارد. به همین دلیل از روش‌های حریمانه و عددی بهتر عمل می‌کند.

• k-means حریمانه: این روش در [۳] مطرح شده است. این روش یک تابع هدف را بهینه می‌کند که شباهت مجموعه‌ی انتخابی با مجموعه‌ی اولیه را بیشینه و شباهت مجموعه انتخابی با خودش را کمینه می‌کند.

• رتبه‌بندی بصری: این روش در [۱۴] به کار برده شده است. در این روش از PageRank بر روی ویژگی‌های تصویری برای انجام خلاصه‌سازی استفاده می‌شود.

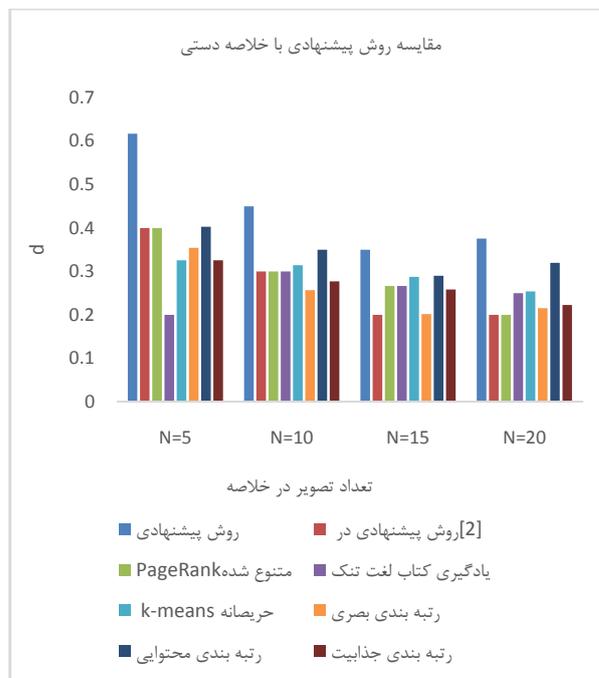
• شبیه‌سازی حرارت با یادگیری تنک: تابع هدف پیشنهادی در این روش بر این مبنا استوار است که تصاویر مجموعه اولیه را می‌توان با ترکیب خطی و تنک مجموعه خلاصه بازسازی کرد. سپس از شبیه‌سازی حرارت برای بهینه‌سازی این تابع هدف استفاده شده است [۵].

• PageRank متنوع شده: این روش بهبودی را بر روی الگوریتم PageRank ارائه می‌کند. در این روش پس از استخراج هر کدام از تصاویر خلاصه، گراف اولیه براساس تفاوت با تصویر انتخابی وزن دهی مجدد می‌شود. به این ترتیب تصاویر خلاصه متنوع‌تر می‌شوند [۱۶].

• رتبه‌بندی جذابیت: در این روش تصاویر براساس امتیاز جذابیت رتبه‌بندی می‌شوند. به عبارت دیگر این روش، خروجی پیمانانه جذابیت به تنهایی را در بر دارد.

• رتبه‌بندی محتوا: در این روش تصاویر براساس امتیاز محتوا رتبه‌بندی می‌شوند. به عبارت دیگر این روش، خروجی محتوا به تنهایی را در بر دارد.

شکل ۶ نتایج ارزیابی را نشان می‌دهد. محور عمودی پارامتر  $d$  است که با استفاده از معادله‌ی ۱۴ محاسبه شده است و محور افقی اندازه‌های مختلف خلاصه را نشان می‌دهد. همان‌طور که در شکل قابل ملاحظه است روش پیشنهادی نتیجه بهتری در مقایسه با روش‌های [۱۴]، [۳]، [۵] و [۱۶] که تنها معیارهایی عددی را بهینه می‌کردند داشته و هم‌چنین نتایج در مقایسه با استفاده از پیمانانه‌های محتوا و جذابیت بهبود یافته است. بدین ترتیب روش پیشنهادی توانسته است به بهبود سی و نه درصدی در ساختن خلاصه‌های شبیه خلاصه‌های دستی دست یابد.



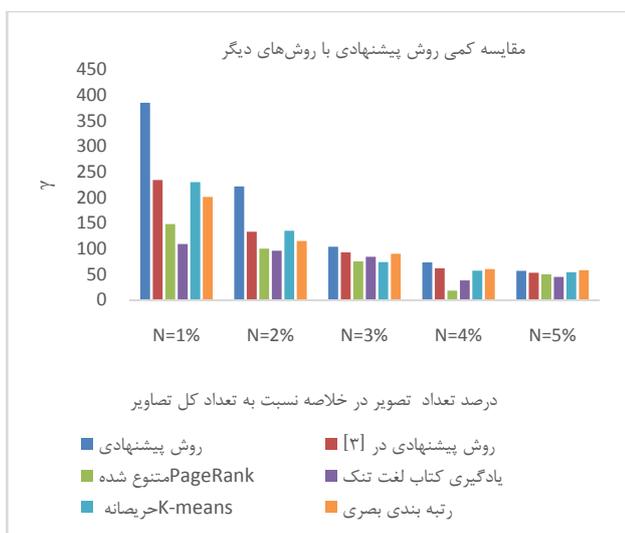
شکل ۷- ارزیابی روش پیشنهادی، مقایسه با خلاصه دستی



شکل ۸- مجموعه‌ای از ۳۲ تصویر مربوط به شهر پاریس



شکل ۹- نتیجه خلاصه‌سازی تصاویر شکل ۹ با روش پیشنهادی



شکل ۱۰- ارزیابی پیمانه محتوا

همان‌طور که مشاهده می‌کنیم در اندازه‌های کوچک‌تر خلاصه، روش پیشنهادی بهبود چشمگیرتری نسبت به بقیه روش‌ها دارد. از آنجایی که با افزایش تعداد تصاویر خلاصه، تصاویر بیشتری در خلاصه شرکت می‌کنند، شانس پوشش اطلاعاتی مجموعه اولیه افزایش می‌یابد. تا اینکه با انتخاب اندازه خلاصه برابر با اندازه مجموعه اولیه، معیار پیشنهادی برای همه روش‌ها یک مقدار به عنوان خروجی بر می‌گرداند - و این مقدار برابر یک است. به این ترتیب روشی کارا تر عمل می‌کند که در تعداد پایین خلاصه بتواند پوشش اطلاعاتی مناسبی از مجموعه اولیه ارائه دهد.

همچنین همان‌طور که در بخش‌های قبل توصیف شد، در این مقاله از طبقه‌بند سلسله مراتبی استفاده شده و از همه نودهای هستان‌شناسی برای ساخت ویژگی استفاده شده است. بدین ترتیب و تعداد ویژگی‌ها در مقایسه با [۲] افزایش می‌یابد. این مساله منجر به افزایش حافظه مورد نیاز برای نگهداری ویژگی‌ها می‌شود. با این حال اگر طبقه‌بندی را صورت موازی در همه نودها انجام دهیم می‌توان با زمان اجرایی مشابه با روش [۲] ولی با مصرف حافظه بیشتر به جواب مطلوب رسید.

جدول ۴- مقایسه زمان اجرای روش‌ها

روش	زمان اجرا برای صد تصویر بر حسب ثانیه
روش پیشنهادی	~۲۱
ریاحی و ابراهیمی مقدم [۲]	~۱۸
K-means حریمانه [۳]	~۹۵
رتبه‌بندی بصری [۱۴]	~۳۱
یادگیری افزونه و تکراری [۵]	~۲۱
PageRank متنوع شده [۱۶]	~۱۹

[2] Z. R. Samani, and M. E. Moghaddam, "Aknowledge-based semantic approach for image collection summarization," *Multimedia Tools and Applications*, pp. 1-23, 2016.

[3] I. Simon, N. Snavey, and S. M. Seitz, "Scene summarization for online image collections," in *Computer Vision, 2007. ICCV 2007, IEEE 11<sup>th</sup> International Conference on IEEE, 2007*, pp. 1-8.

[4] S. Rudinac, M. Larson, and A. Hanjalic, "Learning crowdsourced user preferences for visual summarization of image collections," *Multimedia, IEEE Transactions on*, vol. 15, pp. 1231-1243, 2013.

[5] C. Yang, J. Shen, J. Peng, and J. Fan, "Image collection summarization via dictionary learning for sparse representation," *Pattern Recognition*, vol. 46, pp. 948-961, 2013.

[6] M. Kardaani, and M. E. Moghadam, "Attractive social image extraction based on users' social behaviors," in *2015 9th Iranian Conference on Machine Vision and Image Processing (MVIP), 2015*, pp. 105-110.

[7] X. Shen, and X. Tian, "Multi-modal and multi-scale photo collection summarization," *Multimedia Tools and Applications*, pp. 1-15, 2015.

[8] J. E. Camargo, and F. A. González, "Multimodal latent topic analysis for image collection summarization," *Information Sciences*, vol. 328, pp. 270-287, 2016.

[9] Y. Jia, J. Wang, C. Zhang, and X.-S. Hua, "Finding image exemplars using fast sparse affinitypropagation," in *Proceedings of the 16th ACM international conference on Multimedia, 2008*, pp. 639-642.

[10] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proceedings of the 18th international conference on World wide web, 2009*, pp. 761-770.

[11] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol, "Visual diversification of image search results," in *Proceedings of the 18th international conference on World wide web, 2009*, pp. 341-350.

[12] J. a. G. Fan, Yuli, H. Luo, D. A. Keim, and Z. Li, "A novel approach to enable semantic and visual image summarization for exploratory image search," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval, 2008*, pp. 358-365

[13] Y. H. Yang, P. T. Wu, C. W. Lee, K. H. Lin, W. H. Hsu, and H. H. Chen, "ContextSeer: context search and recommendation at query time for shared consumer photos," in *Proceedings of the 16th ACM international conference on Multimedia, 2008*, pp. 199-208.

[14] Y. Jing, and S. Baluja, "Visualrank:Applying pagerank to large-scale image search," *Pattern Analysis and Machine*

#### ۴-۵- مقایسه زمان اجرا

در گام آخر به مقایسه زمان اجرای روش‌های مورد مقایسه پرداخته‌ایم. برای این کار روش‌های مورد بررسی، بر روی ششصد تصویر اجرا شده و زمان اجرای آنها را در جدول ۴ گزارش داده شده است. مجموعه داده مورد استفاده برای این آزمایش، ششصد تصویر برای شش شهر (پاریس، رم، بارسلونا، سن پترزبورگ، هونگ کنگ، لندن) است که زمان اجرا بین شش مجموعه صد تایی مربوط به هر شهر، میانگین گرفته شده و به عدد بالاتر گرد شده است. بستر سخت‌افزاری یک دستگاه با واحد پردازش مرکزی Intel 2.5 GHz Core i-7 و ۸ گیگا بایت حافظه رم است. همان‌طور که در جدول می‌بینیم کمترین زمان اجرا مربوط به روش [۲] است و پس از آن روش‌های [۵]، [۱۶] و روش پیشنهادی با اختلاف کمتری قرار می‌گیرند. بیشترین زمان اجرا مربوط به K-means حریمانه [۳] است. زمان ذکر شده تنها برای آزمایش مدل‌ها می‌باشد. لازم به ذکر است که زمانی نیز برای آموزش مدل پیشنهادی لازم است که این زمان بیش از چند ده ساعت به طول می‌انجامد.

#### ۵- نتیجه‌گیری

در این مقاله روشی چند معیاره و خودکار برای خلاصه‌سازی مجموعه‌ای از تصاویر اجتماعی مطرح شد. دو پیمانه‌ی محتوا و جذابیت در روش پیشنهادی معرفی شده است که هر کدام مجموعه‌ی متفاوتی از ویژگی‌ها را در نظر می‌گیرند. یک روش خودکار نیز برای استخراج ویژگی‌ها ارائه شد. خلاصه نهایی با ترکیب خروجی این دو پیمانه تولید می‌شود. نتایج آزمایش‌ها نشان می‌دهد روش پیشنهادی در مقایسه با روش‌های دیگر مطرح در حوزه خلاصه‌سازی تصویر که تک معیاره و عددی بودند به بهبود بالای سی درصد در ساخت خلاصه‌های شبیه خلاصه انسانی و بهینه‌سازی معیار عددی دست یافته است. در مقابل، روش پیشنهادی روشی مبتنی بر یک فار یادگیری بوده و میزان حافظه مصرفی این روش در مقابل برخی روش‌های دیگر بالاتر است.

این مقاله جهت‌های متفاوتی را برای کارهای آینده معرفی می‌کند. ترکیب نتایج پیمانه‌های جذابیت و محتوا به سادگی با استفاده از جمع وزن‌دار انجام شده است. روش‌های پیشرفته‌تری در ادبیات برای رتبه بندی چند معیاره وجود دارد. استفاده از این تکنیک‌های چند معیاره برای ترکیب نتایج دو پیمانه ایده‌ی جالبی برای دنبال کردن است.

#### مراجع

[1] S. Tschitschek, R. K. Iyer, H. Wei, and J. A. Bilmes, "Learning Mixtures of Submodular Functions for Image Collection Summarization," in *Advances in Neural Information Processing Systems, 2014*, pp. 1413-1421.

- [27] B. Geng, L. Yang, C. Xu, X.-S. Hua, and S. Li, "The role of attractiveness in web image search," in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 63-72.
- [28] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, pp. 419-426.
- [29] R. Datta, D. Joshi, J. Li, and J. Wang, "Studying aesthetics in photographic images using a computational approach," *Computer Vision–ECCV 2006*, pp. 288-301, 2006.
- [30] J. Itten, "The Art of Color: The Subjective Experience and Objective Rationale of Color (trans. Ernest van Haagen)," ed: New York: Reinhold Publishing Corporation, 1973.
- [31] K.-Q. Huang, Q. Wang, and Z.-Y. Wu, "Natural color image enhancement and evaluation algorithm based on human visual system," *Computer Vision and Image Understanding*, vol. 103, pp. 52-63, 2006.
- [32] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 271-280.
- [33] Y. Luo, and X. Tang, "Photo and video quality evaluation: Focusing on the subject," *Computer Vision–ECCV 2008*, pp. 386-399, 2008.
- [34] J. Machajdik, and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 83-92.
- [35] W. Wei-ning, Y. Ying-lin, and J. Sheng-ming, "Image retrieval by emotional semantics: A study of emotional space and feature extraction," in *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, 2006, pp. 3534-3539.
- [36] Z. Abdollahpour, Z. R. Samani, and M. E. Moghaddam, "Image classification using ontology based improved visual words," in *2015 23rd Iranian Conference on Electrical Engineering*, 2015, pp. 694-698.
- [37] K. Järvelin, and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, pp. 422-446, 2002.
- [38] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out: Proceedings of the ACL-04 workshop*, 2004.
- [39] Y. Li, and B. Merialdo, "VERT: automatic evaluation of video summaries," in *Proceedings of the international conference on Multimedia*, 2010, pp. 851-854.
- Intelligence, IEEE Transactions on*, vol. 30, pp. 1877-1890, 2008.
- [15] L. Yang, "Mining canonical views from internet image collections," Citeseer, 2011.
- [16] J. Wang, L. Jia, and X.-S. Hua, "Interactive browsing via diversified visual summarization for image search results," *Multimedia systems*, vol. 17, pp. 379-391, 2011.
- [17] D. Poole, *Linear algebra: A modern introduction*: Cengage Learning, 2014.
- [18] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating summaries and visualization for large collections of geo-referenced photographs," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, 2006, pp. 89-98.
- [19] L. S. Kennedy, and M. Naaman, "Generating diverse and representative image search results for landmarks," in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 297-306.
- [20] Y. Pang, Q. Hao, Y. Yuan, T. Hu, R. Cai, and L. Zhang, "Summarizing tourist destinations by mining user-generated travelogues and photos," *Computer Vision and Image Understanding*, vol. 115, pp. 352-363, 2011.
- [21] J.-W. Jeong, H.-K. Hong, J.-U. Heu, I. Qasim, and D.-H. Lee, "Visual Summarization of the Social Image Collection Using Image Attractiveness Learned from Social Behaviors," *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pp. 538-543, 2012.
- [22] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. D. Feng, "Video summarization via minimum sparse reconstruction," *Pattern Recognition*, vol. 48, pp. 522-533, 2015.
- [23] S. U. Naci, U. Damnjanovic, B. Mansencal, J. Benois-Pineau, C. Kaes, M. Corvaglia, E. Rossi, and N. Aginako, "The COST292 experimental framework for rushes summarization task in TRECVID 2008," in *Proceedings of the 2nd ACM TRECVID video summarization workshop*, 2008, pp. 40-44.
- [24] H. Fang, W. Lu, F. Wu, Y. Zhang, X. Shang, J. Shao, and Y. Zhuang, "Topic aspect-oriented summarization via group selection," *Neurocomputing*, vol. 149, pp. 1613-1619, 2015.
- [25] R. van Zwol, A. Rae, and L. Garcia Pueyo, "Prediction of favourite photos using social, visual, and textual signals," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1015-1018.
- [26] J. San Pedro, and S. Siersdorfer, "Ranking and classifying attractiveness of photos in folksonomies," in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 771-780.

- <sup>30</sup>Lively-Tedious
- <sup>31</sup>Vibrant-Desolate
- <sup>32</sup>Cool-Warm
- <sup>33</sup>Hard-Soft
- <sup>34</sup>Magnificent-Modest
- <sup>35</sup>Showy-Elegant
- <sup>36</sup>Principle Component Analysis
- <sup>37</sup>Lighting
- <sup>38</sup>Saturation
- <sup>39</sup>Hue
- <sup>40</sup>Colorfulness
- <sup>41</sup>Naturalness
- <sup>42</sup>Sharpness and Blur
- <sup>43</sup>Luminance
- <sup>44</sup>Saliency Map
- <sup>45</sup>Discounted Cumulative Gain (DCG)
- <sup>46</sup>Scale Invariant Feature Transform

**زهرا ریاحی سامانی** مدرک کارشناسی خود را از دانشگاه شریف و مدرک کارشناسی ارشد و دکتری خود را از دانشگاه شهید بهشتی اخذ نموده و اکنون محقق دوره پسا دکتری در دانشگاه پنسیلوانیا است. تحقیقات مورد علاقه ایشان به کارگیری تکنیک‌های محاسباتی، یادگیری و تحلیلی در راستای استخراج اطلاعات عمیق و پیچیده از داده‌های تصویری است.



آدرس پست‌الکترونیکی ایشان عبارت است از:

[z\\_rsamani@sbu.ac.ir](mailto:z_rsamani@sbu.ac.ir)

**محسن ابراهیمی مقدم** از سال ۱۳۸۵ به عنوان عضو هیات علمی دانشگاه شهید بهشتی فعالیت می‌کند. ایشان هر سه مقطع تحصیلی را در رشته مهندسی کامپیوتر در دانشگاه صنعتی شریف به پایان رسانده‌اند و هم‌اکنون به عنوان دانشیار دانشکده مهندسی و علوم کامپیوتر دانشگاه شهید بهشتی مشغول هستند. ایشان مسئول آزمایشگاه تحقیقات پردازش تصویر سیستم‌های توزیع شده هستند و تا کنون بالغ بر ۶۰ مقاله ژورنال و ده‌ها مقاله کنفرانس منتشر کرده‌اند. از علائق تحقیقاتی ایشان می‌توان به زمینه‌های پردازش تصویر و بینایی ماشین اشاره کرد.



آدرس پست‌الکترونیکی ایشان عبارت است از:

[m\\_moghadam@sbu.ac.ir](mailto:m_moghadam@sbu.ac.ir)

#### اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۶/۰۶/۰۶

تاریخ اصلاح: ۱۳۹۶/۰۷/۱۰

تاریخ قبول شدن: ۱۳۹۶/۱۰/۰۴

نویسنده مرتبط: دکتر محسن ابراهیمی مقدم، دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران.

- <sup>1</sup>Coverage
- <sup>2</sup>Redundancy
- <sup>3</sup>Ontology
- <sup>4</sup>Like
- <sup>5</sup>Comment
- <sup>6</sup>Share
- <sup>7</sup>Multi Modal
- <sup>8</sup>Phase
- <sup>9</sup>Test
- <sup>10</sup>System
- <sup>11</sup>Bookmark
- <sup>12</sup>Class
- <sup>13</sup>Like
- <sup>14</sup>Favorite
- <sup>15</sup>Appeal
- <sup>16</sup>Correlation
- <sup>17</sup>Perceptual Quality
- <sup>18</sup>Aesthetic Sensitivity
- <sup>19</sup>Affective Tune
- <sup>20</sup>Brightness
- <sup>21</sup>Luminance
- <sup>22</sup>Contrast
- <sup>23</sup>Sharpness and Blur
- <sup>24</sup>Brightness
- <sup>25</sup>Contasrt
- <sup>26</sup>Intensity
- <sup>27</sup>Subject Region
- <sup>28</sup>Rule of Thirds
- <sup>29</sup>Light-Heavy



## استفاده از شبکه‌های اجتماعی برای بهبود مدیریت ارتباط با مشتری

الناز مظاهری سید ابوالقاسم میرروشندل

دانشکده فنی، دانشگاه گیلان، رشت، ایران

### چکیده

در دنیای امروز شبکه‌های اجتماعی توجه افراد بسیاری را به خود جلب کرده‌اند. در رسانه‌های اجتماعی محصولات بسیاری توسط شرکت‌ها معرفی می‌شوند و انتخاب در این حجم عظیم برای مشتری دشوار خواهد بود. سیستم‌های پیشنهاددهنده بخش مهمی از تجارت الکترونیکی هستند، زیرا تعداد اندکی محصول را مطابق با سلیقه مشتری به او پیشنهاد می‌دهند. کارهای انجام شده در این زمینه راه‌حلهایی را ارائه می‌دهند که شرکت‌ها بتوانند وفاداری مصرف‌کنندگان خود را جلب کنند و از طریق شناسایی سلیقه مشتریان پیشنهادات صادقانه‌ای را به آن‌ها ارائه دهند. در این مقاله یک سیستم پیشنهاددهنده ترکیبی به منظور پیشنهاد فیلم ارائه می‌شود. این سیستم به منظور بهره‌مندی از بهترین ویژگی‌ها و غلبه بر محدودیت‌های موجود در تکنیک‌های فیلترسازی مشترک و فیلترسازی مبتنی بر محتوا آن‌ها را ترکیب می‌کند. هدف الگوریتم پیشنهادی پیش‌بینی پسندیدن یا نپسندیدن فیلم، توسط کاربر است. این الگوریتم به کمک اعتبارسنجی متقابل ارزیابی شده است و نتیجه‌ی قابل‌قبولی را به دست آورده است.

**کلمات کلیدی:** سیستم پیشنهاددهنده ترکیبی، شبکه‌ی اجتماعی، مدیریت ارتباط با مشتری.

### ۱- مقدمه

محصولات به افراد انجام شده است [۸]. هنگامی که کاربران نمی‌توانند آنچه در جستجوی آن هستند را پیدا کنند، حفظ ارتباط با مشتریان در سایت تجارت الکترونیک دچار مشکل می‌شود [۴]. با استفاده از سیستم پیشنهاددهنده<sup>۲</sup>، به‌عنوان بخشی از فعالیت‌های مدیریت ارتباط با مشتری، می‌توان مشکلات بسیاری را حل کرد. زیرا این سیستم‌ها می‌توانند به کاربران کمک کنند، محصول مناسب با سلیقه‌ی خود را انتخاب کنند [۹]. وظیفه‌ی سیستم پیشنهاددهنده، پیش‌بینی پتانسیل خرید یا ارزشی است که کاربر به یک محصول خواهد داد و ارائه فهرستی از محصولات مناسب به مشتری است [۱، ۴].

این مقاله از شبکه‌های اجتماعی مرتبط با فیلم برای بهبود مدیریت ارتباط با مشتری استفاده کرده است. در واقع این شبکه‌ها، سیستم‌های پیشنهاددهنده‌ی فیلم هستند که هر یک با در نظر گرفتن ویژگی‌ها و الگوریتم‌های مشخص، فیلم مناسب را به کاربران پیشنهاد می‌دهند. سودمندی این سیستم بدین شرح است: کاربری برای آخر هفته خود در نظر دارد از میان فیلم‌های بسیاری که وجود دارند، یک فیلم را برای تماشا انتخاب کند. اگر شرکت تولیدکننده فیلم بتواند با استفاده از وبسایت اختصاصی خود و دیگر شبکه‌های اجتماعی، فیلم مناسب را با

مدیریت ارتباط با مشتری<sup>۱</sup> موجب تعاملات یک سازمان با مشتریان جاری و آینده می‌شود. رقابت در زمینه‌ی تجارت الکترونیکی مستلزم مدیریت کارآمد داده‌های وب است [۱]. با ظهور وب ۲ رشد مهمی در جنبه‌های اجتماعی وب به وجود آمد، به‌گونه‌ای که شبکه‌های اجتماعی همانند توییتر، فیس‌بوک و لینکدین رشد قابل‌توجهی را در سال‌های اخیر تجربه کرده‌اند [۲، ۳]. توجه به موضوع مدیریت ارتباط با مشتری ضروری است، زیرا ارزش شرکت‌های بزرگ به مشتریان آن‌هاست. مطالعات نشان می‌دهد که سوددهی مشتری فعلی بیشتر از مشتریان جدید است [۴].

ادغام سیستم‌های مدیریت ارتباط با مشتری موجود با فناوری رسانه‌ی اجتماعی، مفهوم جدید مدیریت ارتباط با مشتری اجتماعی<sup>۲</sup> را مطرح می‌کند. مدیریت ارتباط با مشتری اجتماعی برای به دست آوردن مشتریان بیشتر و پردازش‌های خرید بیشتر به شرکت‌ها کمک می‌کند [۵، ۶، ۷].

در ۲۰ سال اخیر تحقیقات بسیاری در زمینه‌ی چگونگی پیشنهاد خودکار

تجربه کافی و شخصی درباره‌ی موضوعی خاص ندارد، انتخاب کنند. به کمک پیشنهادات ارائه شده توسط کارشناسان یا افراد دیگر در صورت عدم یا کمبود دانش نیز می‌توان انتخاب کرد. سیستم‌های پیشنهاددهنده نیز به افراد کمک می‌کنند، محصولات را با توجه به خواسته‌ها و نیازهای خود انتخاب کنند [۱۲].

در سال‌های اخیر شرکت‌ها در شبکه‌های اجتماعی حضور چشمگیری داشته‌اند. موفقیت یک برند به فراهم آوردن خواسته‌ها و نیازهای مشتریان وابسته است. شرکت‌ها در صورتی می‌توانند انتظار وفاداری از مشتریان خود داشته باشند که سود و منفعت آن‌ها را نیز در نظر بگیرند. به همین منظور در جدول ۱ عملکرد پنج شرکت معتبر فورد فیستا، شورلت تاهو، استارباکس، جت‌بلو و اسکن در شبکه‌های اجتماعی بررسی شده است. همان‌طور که در جدول ۱ می‌توان مشاهده نمود، سه شرکت فورد فیستا، شورلت تاهو، استارباکس برای معرفی محصولات خود اقدام به برگزاری کمپین نموده‌اند. موارد بررسی شده و جزئیات بیشتر را می‌توان در جدول ۱ مشاهده نمود.

در تحقیقی دیگر، تأثیر کشف جوامع در شبکه‌های اجتماعی به‌منظور بهبود مدیریت ارتباط با مشتری بررسی شده است. مشتریان شرکت می‌توانند جوامع مختلفی را تشکیل دهند. برای کشف جوامع باید به ارتباطات میان شرکت، منابع و کاربران توجه کرد. از پروفایل کاربران به‌منظور ایجاد جوامع استفاده می‌شود. کاربرانی که در یک جامعه هستند، خواسته‌ها و علاقه‌مندی‌های مشابه دارند. فرصت اندکی برای شرکت در گفتگوهایی که میان کاربران شکل می‌گیرد، وجود دارد. بنابراین فراهم آوردن عامل‌هایی که بتوانند به‌صورت خودکار و با سرعت بالا به سؤالات و شکایات مشتریان پاسخ دهند به افزایش اعتبار برند کمک بسیاری می‌کند. رسیدگی سریع به شکایات مشتریان بسیار ضروری است. در صورتی که مشتری در شبکه‌های اجتماعی ناراضی خود را درباره‌ی محصولی بیان کند و به این شکایت رسیدگی نشود یا با تأخیر رسیدگی شود، این موضوع موجب ناراضی سایر کاربران نیز می‌شود. نتیجه‌ی این جریان کاهش اعتبار برند است.

جدول ۲ راه‌کارهایی برای مدیریت مشتریان در شبکه‌های اجتماعی را بیان می‌کند. در سطر اول جدول ۲ تأثیر کشف جوامع در شبکه‌های اجتماعی به‌منظور بهبود مدیریت ارتباط با مشتری مورد بررسی قرار می‌گیرد. در سطر دوم این جدول سیستمی ارائه می‌شود که توانایی ارائه‌ی پاسخ سریع به سؤالات و شکایات مشتریان را دارد.

توجه به علاقه‌مندی‌هایی کاربر (کاربر در این وبسایت حساب کاربری دارد) به او پیشنهاد دهد و کاربر پس از تماشای فیلم آن را بپسندد، توانسته است نظر او را جلب کند و او را به مشتری وفادار برای خود تبدیل کند و در زمان و هزینه‌ای که مشتری بابت خرید یا اجاره فیلم‌های اشتباه صرف می‌کند، صرفه‌جویی کند. در واقع وبسایت اختصاصی تولیدکننده‌ی فیلم، همانند یک دوست، به کاربر فیلمی مناسب با سلیقه‌ی او را پیشنهاد می‌دهد که با این عملکرد توانسته است اهداف مدیریت ارتباط با مشتری را پیاده‌سازی کند.

در بخش بعد کارهای مرتبط با تحقیق بیان می‌شوند. سپس در بخش سوم به معرفی مجموعه داده و در بخش چهارم به ارائه گراف کاربر-فیلم پرداخته می‌شود. در بخش پنجم الگوریتم پیشنهادی مطرح و در بخش ششم الگوریتم پیشنهادی ارزیابی می‌گردد. در نهایت، در بخش هفتم نتیجه‌گیری این تحقیق بیان می‌شود.

## ۲- کارهای مرتبط

در این بخش، ابتدا توضیحات مختصری درباره‌ی مدیریت ارتباط با مشتری، شبکه‌ی اجتماعی و سیستم پیشنهاددهنده بیان می‌شود. در ادامه، به ارائه کارهای مرتبط با موضوع تحقیق پرداخته می‌شود.

**مدیریت ارتباط با مشتری:** مدیریت ارتباط با مشتری می‌تواند تمامی سازمان را تحت پوشش خود قرار دهد، روابط سازمان با مشتریان را بهبود دهد و همچنین اطلاعات مربوط به مشتری را ذخیره و تجزیه و تحلیل نماید. ویژگی‌های اصلی مدیریت ارتباط با مشتری بازاریابی، فروش و خدمات هستند [۱۰]. صاحبان کسب و کار معتقدند که یافتن راه‌کارهایی برای حفظ مشتریان فعلی نسبت به یافتن راه‌کارهایی برای جذب مشتریان جدید مقرون به صرفه‌تر است و به افزایش درآمد کمک بیشتری می‌کند [۵].

**شبکه‌ی اجتماعی:** به کمک اینترنت، شبکه‌های اجتماعی می‌توانند میلیون‌ها نفر از سراسر جهان را در یک مکان گرد هم بیاورند. سایت‌های شبکه‌های اجتماعی خدمت‌دهنده‌های مبتنی بر وب هستند که به افراد اجازه می‌دهند یک پروفایل عمومی یا نیمه عمومی در سیستم ایجاد کنند و فهرستی از کاربران دیگر که با آن‌ها در ارتباط هستند را ایجاد و به اشتراک بگذارند [۱۰، ۱۱].

**سیستم پیشنهاددهنده:** در بیشتر مواقع افراد مجبور هستند در شرایطی که

جدول ۱- عملکرد پنج شرکت معتبر در شبکه‌های اجتماعی

مرجع	نام شرکت	زمینه‌ی فعالیت	برگزاری کمپین	عملکرد مناسب	نتیجه
[۱۳]	فورد فیستا	تولید خودرو	✓	- انتخاب کاربران فعال در شبکه‌های اجتماعی برای بارگذاری مطالب در وبسایت اختصاصی کمپین و دیگر شبکه‌های اجتماعی - در نظر گرفتن سود متقابل برای شرکت‌کنندگان	- افزایش فروش - افزایش وفاداری
[۱۳]	شورلت تاهو	تولید خودرو	✓	استفاده از برنامه‌ی پربیننده کارآموز	- افزایش فروش - عدم در نظر گرفتن اقدامی برای تقویت برند و ایجاد وفاداری در میان مشتریان فعلی و جدید
[۱۳]	استارباکس	بزرگ‌ترین کافی‌شاپ زنجیره‌ای	✓	برگزاری مسابقات تبلیغاتی و در نظر گرفتن محصولات شرکت به‌عنوان جوایز	- افزایش فروش - افزایش وفاداری
[۱۴]	جت‌بلو	شرکت هواپیمایی	-	- اختصاص زمان بیشتر برای پاسخ به شکایات و سؤالات - ایجاد تصویری همانند یک دوست برای مشتریان	رضایت‌مندی بیشتر مشتریان شرکت از خدمات ارائه شده
[۱۵]	اسکن	صنعت غذایی	-	علاقه‌مند به گفتگو با افراد ذینفع	- شفاف بودن عملکرد شرکت - عدم حذف پیام‌ها: در صورت منطقی بودن و تکراری نبودن

مشتریان برای انتخاب کمک کنند، می‌توانند وفاداری و رضایتمندی مشتریان خود را جلب کنند. در جدول ۳ راه‌کارهایی ارائه شده است که به کمک آن‌ها می‌توان محصول موردنیاز مشتری را یافت. در این جدول چالش‌های موجود در زمینه‌ی سیستم‌های پیشنهاددهنده، راه‌حل این چالش‌ها، تکنیک‌ها و روش‌های موردنیاز برای پیاده‌سازی این راه‌حل‌ها و همچنین مزایای آن‌ها نیز ارائه شده است.

در شبکه‌های اجتماعی و در دنیای وب محصولات بسیاری توسط شرکت‌های مختلف معرفی می‌شوند. به دلیل افزایش دسترس‌پذیری محصولات، مشتریان یکتا بودن فروشندگان خود را از دست داده‌اند. برای شرکت‌ها یافتن محصول موردعلاقه‌ی مشتری بسیار دشوار شده است. به کمک سیستم‌های پیشنهاددهنده می‌توان محصولات جذاب برای مشتریان را پیدا کرد. اگر شرکت‌ها بتوانند به

جدول ۲- مدیریت مشتریان در شبکه‌های اجتماعی

مرجع	هدف	روش	سازوکار	ضرورت	عملکرد
[۱۶]	کشف جوامعی از مشتریان با استفاده از وب‌سایت اختصاصی شرکت و شبکه‌های اجتماعی عمومی	منابع داخلی: برچسب منابع خارجی: برچسب درهم‌سازی <sup>۴</sup> و پیوند <sup>۵</sup> بهوب‌سایت	امتیازدهی منابع: اظهارنظر درباره‌ی منبع در وب‌سایت اختصاصی: امتیاز ۱ اظهارنظر درباره‌ی منبع و اشتراک‌گذاری در یک شبکه‌ی اجتماعی: امتیاز ۲ اظهارنظر درباره‌ی منبع و اشتراک‌گذاری در دو شبکه‌ی اجتماعی: امتیاز ۳	درجه‌ی علاقه‌مندی کاربر به یک منبع وابسته است به: - تعداد منابع امتیاز داده شده توسط کاربر با برچسب مشابه و همچنین میانگین این امتیازها - رفتار کاربر در وب‌سایت اختصاصی شرکت و شبکه‌های اجتماعی. - علاقه‌مندی مخاطبین <sup>۶</sup> کاربر.	استفاده از یک الگوریتم خوشه‌بندی همانند k-means یا سلسله‌مراتبی برای ایجاد جوامعی از کاربران با علاقه‌مندی‌های مشابه
[۱۷]	- کاوش گفتگوها بر روی سکوها برای معرفی، اولویت‌بندی پست‌ها و پیام‌هایی مرتبط با شرکت به‌منظور ارائه پاسخ سریع به آن‌ها	استفاده از سیستم مدیریت ارتباط با مشتری SCION <sup>۷</sup>	استفاده از سه پودمان: - جمع‌آوری داده‌ها - تحلیل گر محتوا - رابط کاربر  - استفاده از موتور تفسیر SystemText به‌صورت کارآمد و با دقت بالا به‌منظور تعیین نیت هر پست	اولویت هر پست وابسته است به: - امتیاز اضطرابی بودن پست - امتیاز اعتبار نویسنده پست - اهمیت عملکرد تجاری	توانمند ساختن یک عامل یا یک نماینده به‌منظور: - ردیابی ارتباطات با مشتریان - نظارت بر ارتباطات - پاسخ به سوالات مشتریان - تشویق کاربران به مشارکت - پردازش ۹۵۰ پست رسانه‌ی اجتماعی در هر ثانیه توسط یک موتور تفسیر این سیستم

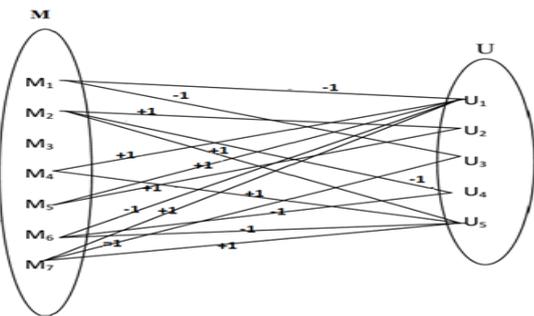
جدول ۳- یافتن محصول جذاب برای مشتری

مرجع	چالش	راه‌حل	تکنیک	روش	مزایا
[۱]	زیاد بودن و طبیعت پویا اطلاعات	استفاده از سیستم پیشنهاددهنده پویا WebBluegillRecomannealing	- پیشنهاد محصول براساس علاقه‌مندی‌های کاربر و علاقه‌مندی‌های کاربران هم‌سلیقه - پیشنهاد محصول با استفاده از الگوریتم تریید شبیه‌سازی شده <sup>۸</sup> و هوش گروهی ماهی بلوگیل	- استفاده از الگوریتم cooling به‌منظور قرار دادن عامل‌های مشابه در صفحه نمایش نزدیک به یکدیگر و عامل‌های غیرمشابه دورتر از یکدیگر - استفاده از الگوریتم cluster-creation برای خوشه‌بندی - ارائه پیشنهادات پویا با استفاده از الگوریتم Bluegill-BestPredictions	- مدیریت حجم بالا اطلاعات در وب - افزایش مقیاس‌پذیری - استفاده از عامل‌های هوشمند
[۴]	- مدیریت حجم عظیم داده‌ها - تولید اطلاعات جدید - ساخت‌نیافته یا نیمه ساخت‌یافته مانند اظهارنظرها در شبکه‌های اجتماعی	- استفاده از محاسبات ابری برای ذخیره‌ی داده‌ها با حجم عظیم - استفاده از ETL <sup>۹</sup> برای تبدیل داده‌ها و بارگذاری آن‌ها در جداول - استفاده از انبار داده‌ای نیمه ساخت‌یافته مانند انبار داده‌ای NoSQL	- استفاده از احساسات موجود در داده‌های عظیم برای پیشنهاد محصولات (مفهوم سیستم پیشنهاددهنده بر مبنای احساسات، پیشنهاد دادن بر مبنای ویژگی‌هاست).	- استخراج احساسات موجود در جمله‌ها برای هر ویژگی به کمک جستجو در واژه‌نامه احساسات <sup>۱۰</sup> - ایجاد یک الگوی سلیقه - ارزیابی تکرار الگوهای استخراج شده - انتخاب احساسات پرتکرار برای هر ویژگی	- استفاده از پتانسیل موجود در داده‌های عظیم برای سیستم‌های پیشنهاددهنده - بهبود عملکرد در کسب و کار، توسط مدیریت و جستجوی تغییرات در داده‌های عظیم
[۱۸]	حجم بالای اطلاعات در وب	استفاده از روش‌های مبتنی بر مورد در نظر گرفتن ۲۰ همسایه برای کاربر (پیش محاسبه)	- استفاده از اعتبارسنجی متقابل ۱۰-fold - استفاده از میانگین خطای مطلق <sup>۱۱</sup>	- کسینوس پایه <sup>۱۲</sup> - کسینوس تطبیق داده شده <sup>۱۳</sup> (انتخاب به‌عنوان بهترین الگوریتم) - همبستگی <sup>۱۴</sup>  - مجموع وزن <sup>۱۶</sup> (انتخاب به‌عنوان بهترین الگوریتم) - رگرسیون <sup>۱۷</sup>	عملکرد بهتر راه‌حل پیشنهادی در مقایسه با الگوریتم پنج-مارک <sup>۱۵</sup> که مبتنی بر کاربر است.
		مدل‌سازی برای یافتن موارد مشابه (پیش محاسبه) - کمتر شدن زمان اجرا بیش از ۲/۸ برابر - توان عملیاتی بیشتر - مقیاس‌پذیری بیشتر			

ادامه جدول ۳- یافتن محصول جذاب برای مشتری

مرجع	چالش	راه‌حل	تکنیک	روش	مزایا
[۱۹]	- الزام بازگشت بی‌درنگ به مجموعه نتایج در بعضی از برنامه‌های کاربردی - محدود بودن اطلاعات مشتریان جدید - فرار بودن داده‌های مشتریان	فیلترسازی مشترک سنتی	معایب نوع روش - الگوریتم فیلترسازی مشترک نزدیک‌ترین همسایه <sup>۱۸</sup> - فاصله کسینوسی <sup>۱۹</sup>	- آزمایش بر روی تعدادی از مشتریان - جزءبندی فضای موردها - نادیده گرفتن بسیاری از موارد برجسته یا غیربرجسته	
		مدل‌های خوشه‌بندی	انجام محاسبات قطعه‌های زیاد و نسبت دادن کاربران به قطعه‌ای با بیشترین تشابه	تقسیم مشتریان به قطعه‌های زیاد و نسبت دادن کاربران به قطعه‌ای با بیشترین تشابه	گروه‌ها با اندازه‌ی کوچک کارایی و مقیاس پذیری بهتری به صورت برخط دارند.
		روش‌های مبتنی بر جستجو	موارد رأی داده شده و خریداری شده توسط کاربر یک پرس‌وجو <sup>۲۰</sup> را برای یافتن موارد مشابه دیگر شکل می‌دهند.	- اگر کاربر موارد زیادی را خریده باشد، پرس‌وجو بر روی همه‌ی موارد غیرعملی خواهد بود. - خیلی کلی <sup>۲۱</sup> - خیلی محدود <sup>۲۲</sup>	اگر کاربری محصولات کمی را خریده باشد یا به محصولات کمی رأی داده باشد، الگوریتم‌های پیشنهاددهنده مبتنی بر جستجو مقیاس‌پذیر و کارآمد هستند.
		الگوریتم فیلترسازی مشترک مورد به مورد، استفاده شده توسط Amazon.com	- بیشتر از اینکه هدف سازگاری کاربر با مشتریان مشابه باشد، هدف ایجاد جدولی از موارد مشابه است.	- ایجاد جدولی از موارد مشابه به صورت غیربرخط (این اقدام توسط یافتن مواردی صورت می‌پذیرد که مشتری تمایل داشته است آن‌ها را با یکدیگر بخرد).	- افزایش کارایی و مقیاس پذیری - ایجاد جدولی از موارد مشابه به صورت غیربرخط - عالی بودن کیفیت پیشنهادات - اجرا قابل قبول با محدود شدن داده‌های کاربر

۳- مجموعه داده



شکل ۱- نمونه‌ای از گراف کاربر-فیلم

در این تحقیق از مجموعه داده‌ی hetrec2011-movieLens-2k استفاده شده است [۲۰]. این مجموعه داده شامل ۱۰,۱۰۹ فیلم است و ۲,۱۱۳ کاربر به این فیلم‌ها رأی داده‌اند. در این مقاله، از ۴۳ کاربر و ۳,۳۷۴ فیلم مختلف استفاده شده است. این مجموعه داده حاوی ۲۰ ژانر مختلف است. در واقع به‌طور میانگین هر فیلم دارای ۲/۴۰ ژانر است. رأی‌هایی که کاربران به هر فیلم داده‌اند، در محدوده‌ی ۱ تا ۵ قرار گرفته است.

۴- گراف کاربر-فیلم

برای درک بهتر مجموعه داده آن را به صورت گراف کاربر-فیلم مدل‌سازی می‌کنیم. شبکه کاربر-فیلم را می‌توان به عنوان یک گراف دوبخشی می‌دانیم.  $G = \langle U, M, E \rangle$  مدل‌سازی کرد.  $U$  و  $M$  دو مجموعه‌ی منحصراً به‌فرد از کاربران و فیلم‌ها هستند.  $E$  مجموعه‌ای از یال‌های  $G$  و همچنین زیرمجموعه‌ای از  $U \times M$  است.

$$\begin{aligned}
 U &= \{U_1, U_2, U_3, U_4, U_5\} \\
 M &= \{M_1, M_2, M_3, M_4, M_5, M_6, M_7\} \\
 E &= \{(U_1, M_1), (U_1, M_4), (U_1, M_5), (U_1, M_6), (U_1, M_7), \\
 &\quad (U_2, M_2), (U_2, M_5), (U_3, M_1), (U_3, M_7), (U_4, M_2), \\
 &\quad (U_4, M_6), (U_5, M_2), (U_5, M_4), (U_5, M_6), (U_5, M_7)\} \\
 G &= \{\langle U_1, M_1, -1 \rangle, \langle U_1, M_4, +1 \rangle, \langle U_1, M_5, +1 \rangle, \\
 &\quad \langle U_1, M_6, -1 \rangle, \langle U_1, M_7, +1 \rangle, \langle U_2, M_2, +1 \rangle, \\
 &\quad \langle U_2, M_5, +1 \rangle, \langle U_3, M_1, -1 \rangle, \langle U_3, M_7, -1 \rangle, \\
 &\quad \langle U_4, M_2, -1 \rangle, \langle U_4, M_6, -1 \rangle, \langle U_5, M_2, +1 \rangle, \\
 &\quad \langle U_5, M_4, +1 \rangle, \langle U_5, M_6, -1 \rangle, \langle U_5, M_7, +1 \rangle\}
 \end{aligned}$$

تماشا کرده‌اند و آن را پسندیده‌اند. کاربر  $U_1$  یک مرتبه با کاربر  $U_2$  و سه مرتبه با کاربر  $U_5$  هم‌سلیقه بوده است. دسته‌ی دوم، کاربرانی که هم‌سلیقه با کاربر موردنظر هستند و فیلم پیشنهادی را تماشا کرده‌اند و آن را نپسندیده‌اند. کاربر  $U_1$  یک‌مرتبه با کاربر  $U_4$  هم‌سلیقه بوده است.

در مرحله‌ی پنجم فیلترسازی مشترک را به کمک بیزین ساده و با استفاده از رابطه‌ی (۱) محاسبه می‌کنیم. دسته‌بندی بیزین ساده یکی از الگوریتم‌های موفق برای دسته‌بندی بسیاری از دامنه‌هاست. برخلاف سادگی، این الگوریتم قابلیت رقابت با روش‌های پیچیده‌ی دیگر به‌خصوص در دسته‌بندی متنی و فیلترسازی مبتنی بر محتوا را دارد. فرض می‌کنیم، ویژگی‌ها از یکدیگر مستقل هستند [۲۲].

$$\prod_{i=1}^n p(f_i | \text{class}_j) p(\text{class}_j) \quad (1)$$

در رابطه‌ی (۱)،  $p(\text{class}_j)$  احتمال رخداد  $\text{class}_j$  و  $p(f_i | \text{class}_j)$  احتمال تولید نمونه  $f_i$  در  $\text{class}_j$  هستند. یادگیری سریع (یک‌بار مرور اجمالی) را انجام می‌دهد، دسته‌بندی سریع، عدم حساسیت به ویژگی‌های غیرمرتبط، مدیریت داده‌های واقعی و گسسته، مدیریت داده‌های رشته‌ای از مزایای دسته‌بندی بیزین ساده هستند. فرض استقلال ویژگی‌ها محدودیت این روش است [۲۳]. فیلترسازی مشترک را به کمک بیزین ساده برای هر یک از دسته‌های "like" و "dislike" محاسبه می‌کنیم. در واقع به کمک روابط (۲) و (۳)،  $CF_{\text{like}}$  و  $CF_{\text{dislike}}$  را به دست می‌آوریم.

$$CF_{\text{like}} = p(\text{"like"}) \prod_{i=1}^n p(f_i | \text{"like"}) \quad (2)$$

$$CF_{\text{dislike}} = p(\text{"dislike"}) \prod_{i=1}^n p(f_i | \text{"dislike"}) \quad (3)$$

در مرحله‌ی ششم فیلترسازی مبتنی بر محتوا را به کمک بیزین ساده و با استفاده از رابطه‌ی (۴) به‌منظور تعیین اینکه آیا کاربر ژانرهای فیلم پیشنهادی را می‌پسندد و یا خیر محاسبه می‌کنیم. ژانرهای فیلم‌های موجود در گراف کاربر - فیلم در جدول ۴ آمده است.

$$p(\text{class}_j) p(\text{genres}_{\text{class}_j} | \text{class}_j) \quad (4)$$

جدول ۴- ژانرهای فیلم‌های گراف کاربر - فیلم

Movie	Genres	Movie	Genres
M <sub>1</sub>	Adventure	M <sub>5</sub>	Action
M <sub>2</sub>	Adventure	M <sub>5</sub>	Adventure
M <sub>2</sub>	Animation	M <sub>5</sub>	Thriller
M <sub>2</sub>	Children	M <sub>6</sub>	Crime
M <sub>2</sub>	Comedy	M <sub>6</sub>	Comedy
M <sub>2</sub>	Fantasy	M <sub>6</sub>	Romance
M <sub>3</sub>	Drama	M <sub>6</sub>	Children
M <sub>3</sub>	Fantasy	M <sub>7</sub>	Adventure
M <sub>3</sub>	Mystery	M <sub>7</sub>	Drama
M <sub>4</sub>	Adventure	M <sub>7</sub>	Fantasy
M <sub>4</sub>	Children	M <sub>7</sub>	Mystery
M <sub>4</sub>	Fantasy	M <sub>7</sub>	Sci-Fi

ژانر فیلم‌هایی که کاربر تماشا کرده است را به دو دسته تقسیم می‌کنیم: دسته‌ی اول تعداد ژانرهایی است که کاربر فیلم‌های متناظر با آن‌ها را پسندیده است؛ این تعداد برای کاربر  $U_1$  برابر با ۱۱ است. دسته‌ی دوم تعداد ژانرهایی است که کاربر فیلم‌های متناظر با آن‌ها را نپسندیده است؛ این تعداد برای کاربر  $U_1$  برابر با ۵ است.

سپس ژانرهای مطابق با ژانرهای فیلم پیشنهادی و تماشا شده توسط کاربر

هر  $e = (u, m) \in E$  نشان می‌دهد که کاربر  $u \in U$  به فیلم  $m \in M$  رأی داده است. مجموعه  $W$  شامل وزن‌های یال‌های موجود در مجموعه  $E$  است. این وزن‌ها دارای مقدار +۱ یا -۱ هستند. +۱ نمایانگر دسته‌ی "like" و -۱ نمایانگر دسته‌ی "dislike" است. از این گراف دوبخشی وزن‌دار برای سیستم پیشنهاددهنده‌ی فیلم استفاده می‌کنیم [۲۱]. ذکر این نکته ضروری است که اگر کاربر به یک فیلم رأی بالاتر از ۳ بدهد، به این معناست که او این فیلم را می‌پسندد و یال مرتبط با این کاربر و فیلم دارای وزن +۱ خواهد بود. اما اگر کاربر به این فیلم رأی کمتر از ۳ بدهد، یعنی او این فیلم را نمی‌پسندد و یال مرتبط به این کاربر و فیلم دارای وزن -۱ خواهد بود. اگر کاربر به این فیلم رأی ۳ بدهد، یعنی او نسبت به این فیلم بی‌تفاوت است؛ در نتیجه یالی برای این کاربر و فیلم وجود نخواهد داشت. شکل ۱ نمونه‌ای از گراف کاربر-فیلم را نشان می‌دهد، که دارای ویژگی‌های زیر است:

## ۵- الگوریتم پیشنهادی

در این بخش، الگوریتم پیشنهادی برای سیستم پیشنهاددهنده ارائه می‌شود. همان‌طور که در شکل ۱ مشاهده می‌شود، کاربر  $U_1$  فیلم  $M_2$  را ندیده است. آیا کاربر پس از تماشای این فیلم آن را خواهد پسندید؟ روندنما<sup>۳۳</sup> الگوریتم پیشنهادی را می‌توان در شکل ۲ مشاهده نمود. در ادامه مراحل الگوریتم نیز شرح داده می‌شوند.

در مرحله‌ی اول فیلم‌هایی که کاربر موردنظر تماشا کرده است را پیدا می‌کنیم و این فیلم‌ها را به دو دسته تقسیم می‌کنیم. دسته‌ی اول شامل فیلم‌هایی می‌شود که کاربر تماشا کرده و آن‌ها را پسندیده است. دسته‌ی دوم شامل فیلم‌هایی می‌شود که کاربر تماشا کرده و آن‌ها را نپسندیده است. فیلم‌های  $M_4$ ،  $M_5$  و  $M_7$  را کاربر  $U_1$  پسندیده است. فیلم‌های  $M_1$  و  $M_6$  را کاربر  $U_1$  نپسندیده است.

در مرحله‌ی دوم کاربران هم‌سلیقه جستجو می‌شوند. کاربران هم‌سلیقه کاربرانی هستند که فیلم‌هایی که کاربر موردنظر دیده است را تماشا کرده‌اند و همچنین با او هم‌نظر هستند. این کاربران به دو دسته تقسیم می‌شوند: دسته‌ی اول کاربرانی هستند که فیلم‌هایی که کاربر موردنظر دیده است را تماشا کرده‌اند و همچنین همانند این کاربر آن‌ها را پسندیده‌اند. کاربر  $U_1$  با کاربر  $U_2$  در زمینه‌ی فیلم  $M_5$  و با کاربر  $U_5$  در زمینه‌ی فیلم‌های  $M_4$  و  $M_7$  هم‌سلیقه است. دسته‌ی دوم کاربرانی هستند که فیلم‌هایی که کاربر موردنظر دیده است را تماشا کرده‌اند و همچنین همانند این کاربر آن‌ها را نپسندیده‌اند. کاربر  $U_1$  با کاربر  $U_3$  در زمینه‌ی فیلم  $M_1$ ، با کاربر  $U_4$  در زمینه‌ی فیلم  $M_6$  هم‌سلیقه است و همچنین با کاربر  $U_5$  در زمینه‌ی فیلم  $M_6$  هم‌سلیقه است.

در مرحله سوم کاربران هم‌سلیقه‌ای جستجو می‌شوند که فیلم پیشنهادی را مشاهده کرده‌اند. این کاربران نیز به دو دسته تقسیم می‌شوند: دسته‌ی اول کاربرانی هستند که با کاربر موردنظر هم‌سلیقه هستند، فیلم پیشنهادی را تماشا کرده‌اند و آنرا پسندیده‌اند. از میان کاربر هم‌سلیقه با کاربر  $U_1$  که این فیلم را تماشا کرده‌اند و همچنین آن را پسندیده‌اند، کاربران  $U_2$  و  $U_5$  انتخاب می‌شوند. دسته‌ی دوم کاربرانی هستند که با کاربر موردنظر هم‌سلیقه هستند، فیلم پیشنهادی را تماشا کرده‌اند اما آن را نپسندیده‌اند. از میان کاربر هم‌سلیقه با کاربر  $U_1$  که این فیلم را تماشا کرده‌اند و همچنین آن را نپسندیده‌اند، کاربر  $U_4$  انتخاب می‌شود.

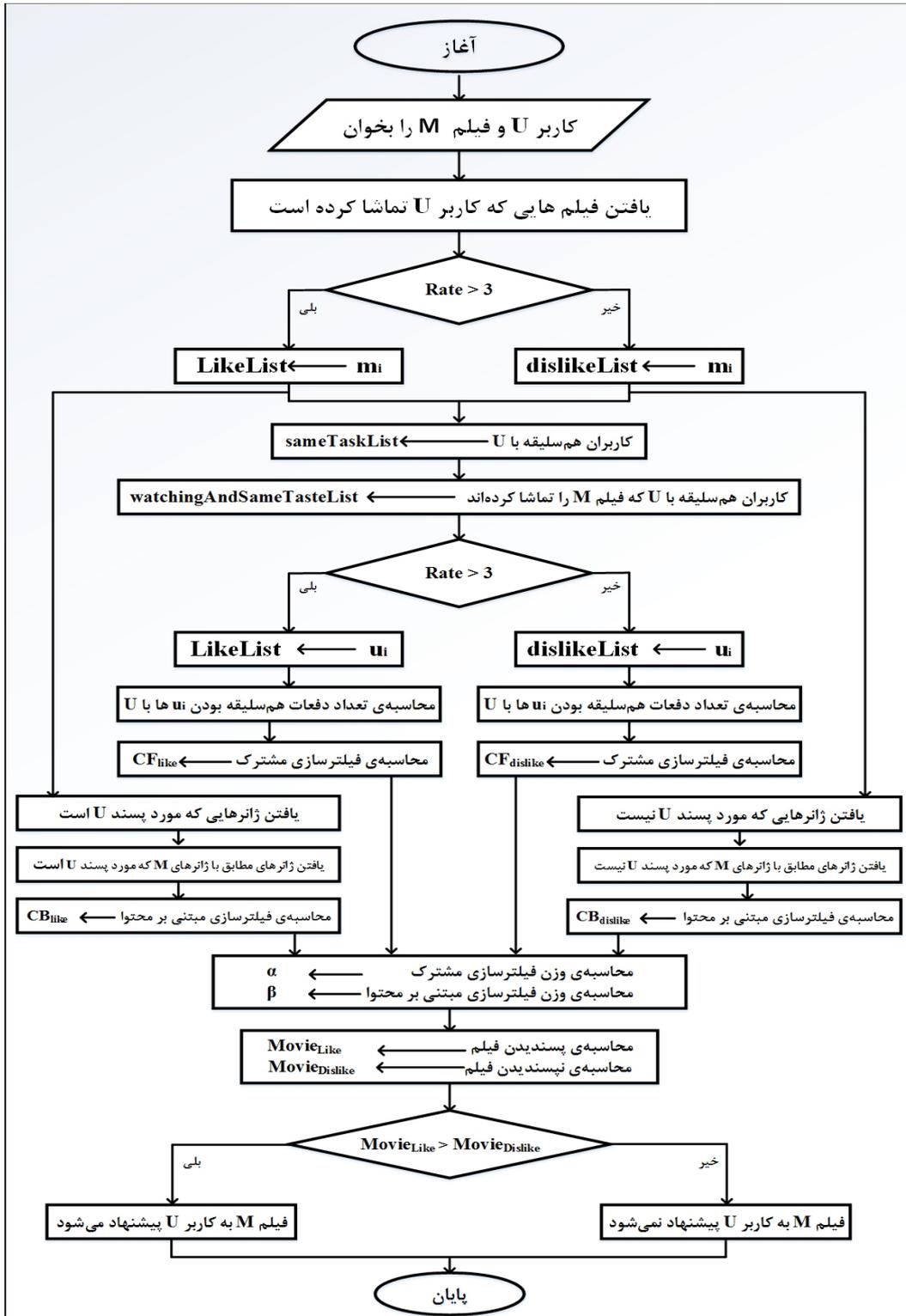
در مرحله‌ی چهارم تعداد دفعات هم‌سلیقه بودن کاربران با کاربر موردنظر محاسبه می‌شود. در این مرحله تعداد دفعاتی که هر کاربر با کاربر موردنظر هم‌سلیقه بوده است را پیدا می‌کنیم. کاربران در دو دسته متفاوت قرار می‌گیرند. دسته‌ی اول، کاربرانی که هم‌سلیقه با کاربر موردنظر هستند و فیلم پیشنهادی را

برای هریک از دسته‌های "like" و "dislike" محاسبه می‌کنیم. در واقع به کمک روابط (۵) و (۶)،  $CB_{like}$  و  $CB_{dislike}$  را محاسبه می‌کنیم.

$$CB_{like} = p(\text{"like"}) p(\text{genres}_{like} | \text{"like"}) \quad (5)$$

$$CB_{dislike} = p(\text{"dislike"}) p(\text{genres}_{dislike} | \text{"dislike"}) \quad (6)$$

جستجو می‌شوند. ژانر فیلم‌هایی که کاربر تماشا کرده است و مطابق با ژانر یا ژانرهای فیلم پیشنهادی هستند را به دو دسته تقسیم می‌کنیم: دسته اول مربوط به تعداد ژانرهایی می‌شود که کاربر فیلم‌های متناظر با آن‌ها را پسندیده است و همچنین این ژانرها مطابق با فیلم پیشنهادی هستند؛ این تعداد برای کاربر  $U_1$  برابر با ۶ است. دسته دوم مربوط به تعداد ژانرهایی می‌شود که کاربر فیلم متناظر با آن‌ها را نپسندیده است و همچنین این ژانرها مطابق با فیلم پیشنهادی هستند؛ این تعداد برای کاربر  $U_1$  برابر با ۳ است. فیلترسازی مبتنی بر محتوا را



شکل ۲- روندنما الگوریتم پیشنهادی

کاربر، یافتن کاربران مشابه یا ایجاد یک پروفایل مبتنی بر محتوا امکان پذیر نیست. به کمک فیلترسازی مشترک می توان به کاربر محصولاتی را پیشنهاد داد که در گذشته تجربه استفاده از آن ها را نداشته است. این موضوع در طول زمان موجب تغییر در روند علاقه مندی های کاربر می شود. نقاط قوت و ضعف الگوریتم را می توان در جدول ۶ مشاهده نمود.

جدول ۵- نتایج به دست آمده از الگوریتم پیشنهادی

دقت الگوریتم	پیش بینی نادرست	پیش بینی درست	تعداد رأی های مخالف ۳	تعداد رأی های برابر با ۲	کل رأی ها	تست
۷۷٪	۳۹۲	۱,۳۴۷	۱,۷۴۷	۳۵۵	۲,۱۰۲	۱
۷۷٪	۳۶۷	۱,۳۰۱	۱,۶۷۳	۳۲۶	۲,۰۰۰	۲
۷۸٪	۳۵۶	۱,۲۹۱	۱,۶۵۱	۳۴۹	۲,۰۰۰	۳
۷۸٪	۳۴۰	۱,۲۸۹	۱,۶۴۴	۳۵۶	۲,۰۰۰	۴
۷۷٪	۳۶۸	۱,۲۶۲	۱,۶۳۸	۳۶۰	۲,۰۰۰	۵
۷۷/۴٪	-	-	-	-	-	میانگین

جدول ۶- نقاط قوت و ضعف الگوریتم

نقاط ضعف الگوریتم	نقاط قوت الگوریتم
مشکل شروع سرد (مربوط به کاربران جدید)	یوبایی (توانایی یادگیری در صورت تغییر در علاقه مندی های کاربر در طول زمان)
-	عدم استفاده از نمونه برداری برای از دست ندادن داده های با اهمیت
-	قابلیت اجرا با سرعت قابل قبول در پردازش های پرخط
-	مقیاس پذیری (قابل اجرا بودن الگوریتم برای هر تعداد کاربر)
-	غلبه بر مشکل کم پشتی
-	غلبه بر مشکل شروع سرد (مربوط به محصولات جدید)

مقایسه الگوریتم پیشنهادی ارائه شده در این مقاله با الگوریتم های سیستم های پیشنهاددهنده ارائه شده در تحقیقات دیگر (که از مجموعه داده استاندارد hetrec2011-movielens-2k نیز استفاده کرده اند) امکان پذیر نیست؛ زیرا هدف الگوریتم های سیستم های پیشنهاددهنده دیگر، پیش بینی رأی دقیقی است که کاربر به فیلم خواهد داد، در حالی که الگوریتم پیشنهادی ارائه شده در این مقاله دسته "like" و "dislike" را برای فیلمی که کاربر به آن رأی خواهد داد را پیش بینی می کند.

## ۷- نتیجه گیری

بزرگ ترین چالش رسانه های اجتماعی، یافتن و استفاده از فرصت هاست. در حجم عظیم داده ها و اطلاعات موجود در وب، سیستم پیشنهاددهنده برای انتخاب محصول مناسب کمک بسیاری به مشتری می کند. بنابراین در این تحقیق، الگوریتمی برای سیستم پیشنهاددهنده ارائه شد که بتواند با وجود فیلم های

در مرحله ی هفتم احتمال پسندیدن یا نپسندیدن فیلم توسط کاربر را محاسبه می کنیم. در ابتدا برای فیلترسازی مشترک و فیلترسازی مبتنی بر محتوا، وزن تعیین می کنیم.  $\alpha$  را به عنوان وزن فیلترسازی مشترک و  $\beta$  را به عنوان وزن فیلترسازی مبتنی بر محتوا در نظر می گیریم. از ۱,۱۷۸ سطر مجموعه داده به عنوان مجموعه توسعه<sup>۲۵</sup> برای تنظیم وزن ها استفاده می کنیم. با آزمایش هایی که بر روی مجموعه توسعه انجام داده ایم، به صورت تجربی  $\alpha = ۰/۵۵$  و  $\beta = ۰/۴۵$  را انتخاب می کنیم. movie<sub>like</sub> و movie<sub>dislike</sub> را به کمک روابط (۷) و (۸) محاسبه می کنیم.

$$\text{movie}_{\text{like}} = \alpha \text{CF}_{\text{like}} + \beta \text{CB}_{\text{like}} \quad (۷)$$

$$\text{movie}_{\text{dislike}} = \alpha \text{CF}_{\text{dislike}} + \beta \text{CB}_{\text{dislike}} \quad (۸)$$

اگر احتمال اینکه کاربر فیلم پیشنهادی را نپسندد (movie<sub>like</sub>) از احتمال اینکه کاربر فیلم پیشنهادی را نپسندد (movie<sub>dislike</sub>) بیشتر باشد، فیلم مورد نظر به کاربر پیشنهاد داده می شود، در غیر این صورت از پیشنهاد فیلم به کاربر صرف نظر می شود. احتمال اینکه کاربر  $U_1$  فیلم  $M_2$  را نپسندد؛ برابر با ۰/۲۶ و احتمال اینکه کاربر  $U_1$  فیلم  $M_2$  را نپسندد؛ برابر با ۰/۲۱ است، در نتیجه فیلم  $M_2$  به کاربر  $U_1$  پیشنهاد می شود.

## ۶- ارزیابی الگوریتم پیشنهادی

روندنا الگوریتم پیشنهادی توسط ابزار اکلیس<sup>۲۶</sup>، مبتنی بر زبان برنامه نویسی جاوا پیاده سازی شده است. به منظور محاسبه ی دقت الگوریتم پیشنهادی از اعتبارسنجی متقابل<sup>۲۷</sup> ۵-fold برای مجموعه داده استفاده شده است. نتایج را می توان در جدول ۵ مشاهده نمود. میانگین دقت الگوریتم پیشنهادی پس از ۵ مرحله تست ۷۷/۴ است. نکته ای که در این مجموعه داده و به صورت مشابه در بسیاری از مجموعه داده های رأی دیگر وجود دارد این است که، اغلب امتیاز رأی ها بالاست. زیرا کاربران سعی می کنند محصولاتی را انتخاب کنند که احتمال زیادی می دهند آن ها را خواهند پسندید.

به منظور دستیابی به بالاترین کارایی و غلبه بر مشکلات تکنیک های پیشنهاددهنده سنتی، در این تحقیق تکنیک پیشنهاددهنده ترکیبی ارائه شد. به صورتی که بهترین ویژگی های دو تکنیک پیشنهاددهنده، در یک تکنیک پیشنهاددهنده ترکیب شدند. در این تکنیک پیشنهاددهنده ترکیبی، تکنیک فیلترسازی مشترک با تکنیک فیلترسازی مبتنی بر محتوا برای حل مشکلات کم پشتی<sup>۲۸</sup> و شروع سرد<sup>۲۹</sup> ترکیب شدند.

مشکل کم پشتی در فیلترسازی مشترک به کمک فیلترسازی مبتنی بر محتوا حل شد، زیرا مشکل کم پشتی زمانی اتفاق می افتد که کاربران بسیاری به بیشتر محصولات رأی نمی دهند. در این صورت ماتریس رأی کاربر کم پشت خواهد بود. در ماتریس کم پشت احتمال یافتن کاربران مشابه برای کاربر مورد نظر کاهش می یابد؛ در این شرایط از تاریخچه ی زانرهای مورد علاقه ی کاربر برای تکنیک فیلترسازی مبتنی بر محتوا استفاده کردیم.

مشکل شروع سرد مربوط به محصولات جدید و کاربران جدید می شود. در الگوریتم پیشنهادی مشکل شروع سرد مربوط به محصولات جدید توسط فیلترسازی مبتنی بر محتوا حل شد؛ زیرا تکنیک مبتنی بر محتوا براساس رأی های کاربران دیگر نیست. از این تکنیک می توان برای پیشنهاد همه ی محصولات (محصولات جدید و قدیم) استفاده کرد. اما الگوریتم پیشنهادی در ارائه پیشنهاد به کاربر جدید با مشکل روبه رو است. زیرا بدون دانستن اولویت های قبلی یک

[6] D. Liberona, M. Ruiz, and D. Fuenzalida, "Customer knowledge management in the age of social networks," In 7th International Conference on Knowledge Management in Organizations: Service and Cloud Computing (pp. 353-364). Springer Berlin Heidelberg, 2013.

[7] K. J. Trainor, J. M. Andzulis, A. Rapp, and R. Agnihotri, "Social media technology usage and customer relationship performance: A capabilities-based examination of social CRM," *Journal of Business Research*, 67(6), 1201-1208, 2014.

[8] M. D. Ekstrand, J. T. Riedl, and J. A. Konstan, "Collaborative filtering recommender systems," *Foundations and Trends in Human-Computer Interaction*, 4(2), 81-173, 2011.

[9] A. Demiriz, "Enhancing product recommender systems on sparse binary data," *Data Mining and Knowledge Discovery*, 9(2), 147-170, 2004.

[10] S. Mohan, E. Choi, and D. Min, "Conceptual modeling of enterprise application system using social networking and web 2.0 "social CRM system"," *Convergence and Hybrid Information Technology, 2008. ICHIT'08. International Conference on. IEEE*, 2008.

[11] Ellison, and B. Nicole "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication* 13.1 (2007): 210-230.

[12] P. Resnick, and H. R. Varian, "Recommender systems," *Communications of the ACM*, vol. 40, no. 3, pp. 56-58, 1997.

[13] E. Brown, "How implementing social media strategies (the right way) attracts customer loyalty," The Faculty of the Public Communication Graduate Program School of Communication American University Washington, DC, 1-39, 2010.

[14] D. Halova, and R. Menne, "Using Social CRM to influence Customer Service and Loyalty: A Perspective in the Airline Industry," Master's Thesis. BUSN39: Global Marketing. Lund University, 2013.

[15] K. Johnsson, "Strategies for social customer relationship management," Master's Thesis. Agricultural Programme – Economics and Management. Degree thesis No 874, 2014.

[16] I. Basaille-Gahitte, L. Abrouk, N. Cullot, and E. Leclercq, "Using social networks to enhance customer relationship management," In Proceedings of the Fifth International Conference on Management of Emergent Digital EcoSystems (pp. 169-176), ACM, October 2013.

[17] J. Ajmera, H. I. Ahn, M. Nagarajan, A. Verma, D. Contractor, S. Dill, and M. Denesuk, "A crm system for social media: challenges and experiences," In Proceedings of the 22nd international conference on World Wide Web (pp. 49-58), International World Wide Web Conferences Steering Committee, May 2013.

بسیاری که روزانه توسط شرکت‌های فیلم‌سازی تولید می‌شوند، فیلم مناسب باسلیقه‌ی کاربر را به او پیشنهاد دهد. در الگوریتم پیشنهادی تکنیک‌های فیلترسازی مشترک و فیلترسازی مبتنی بر محتوا با یکدیگر ترکیب شدند. در واقع برای پیشنهاد فیلم به کاربر از یک سیستم پیشنهاددهنده‌ی ترکیبی استفاده شد. در روش ارائه شده علاوه بر انتخاب‌های گذشته‌ی کاربر از انتخاب‌های کاربران هم‌نظر با او نیز استفاده شد. الگوریتم پیشنهادی پس از ارزیابی، نتیجه‌ی قابل قبول ۷۷/۴ را بر روی مجموعه داده استاندارد به دست آورد. در روش پیشنهادی برای افزایش سرعت به دلیل امکان از دست دادن داده‌های مهم، نمونه‌برداری انجام نشد. بنابراین برای افزایش سرعت در پردازش‌های برخط، از الگوریتم بی‌زین ساده و تقسیم کردن فیلم‌ها در دو دسته‌ی پسندیدن و نپسندیدن استفاده شد.

در این تحقیق تنها از رأی‌های کاربران و ژانر فیلم‌ها برای پیشنهاد فیلم به کاربر استفاده شد، بنابراین زمانی که یک کاربر جدید وارد سیستم می‌شود، سیستم باید به کاربر پیشنهادات تصادفی را ارائه دهد تا بتواند علاقه‌مندی‌های او را متوجه شود. در این حالت ممکن است که تعداد پیشنهادات اشتباه توسط سیستم پیشنهاددهنده بالا باشد، در این شرایط امکان دارد کاربر به سیستم اطمینان نکند و بار دیگر برای خرید مراجعه نکند. چهار راه‌حل برای مشکل شروع سرد مربوط به کاربران جدید در نظر گرفته شده است که در آینده به آن‌ها پرداخته خواهد شد. در راه‌حل اول، سیستم از بازخوردهای کاربر پس از پیشنهاد فیلم به او یاد می‌گیرد و پیشنهادات خود را بهبود می‌دهد. در این حالت علاوه بر رأی‌های کاربر به فیلم و ژانرهای موردعلاقه، از اظهار نظرهایی که درباره‌ی فیلم بیان می‌کند، برای کسب اطلاعات در زمینه‌ی علاقه‌مندی‌های او نیز استفاده می‌شود. در راه‌حل دوم زمانی که کاربر در سیستم پیشنهاددهنده فیلم عضو می‌شود، از او درخواست شود که ژانرهای موردعلاقه خود را بیان کند. در راه‌حل سوم هنگامی که کاربر در سیستم برای عضویت اقدام می‌کند، از او خواسته شود، آدرس دیگر شبکه‌های اجتماعی خود را نیز اعلام کند. راه‌حل چهارم ترکیب سه راه‌حل فوق است.

## مراجع

[1] A. Alphy, and S. Prabakaran, "A Dynamic Recommender System for Improved Web Usage Mining and CRM Using Swarm Intelligence," *The Scientific World Journal*, 2015.

[2] C. Giannakis-Bompolis, and C. Boutsouki, "Customer relationship management in the era of social web and social customer: an investigation of customer engagement in the Greek retail banking sector," *Procedia-Social and Behavioral Sciences*, 148, 67-78, 2014.

[3] H. Zwikstra, F. Hogenboom, D. Vandic, and F. Frasinca, "Connecting customer relationship management systems to social networks," In 7th International Conference on Knowledge Management in Organizations: Service and Cloud Computing (pp. 389-400), Springer Berlin Heidelberg, 2013.

[4] F. E. Jamiy, A. Daif, M. Azouazi, and A. Marzak, "The potential and challenges of big data-Recommendation systems next level application," *arXiv preprint arXiv: 1501.03424*, 2015.

[5] M. Das, "Social Customer Relationship Management: Proposal for SCRM for an IT Services Company," Master's Thesis. Industrial Management. Helsinki Metropolia University of Applied Sciences. Master's Degree, 2012.

- <sup>7</sup>Social Channel is Open Now  
<sup>8</sup>Simulated Annealing (SA)  
<sup>9</sup>Extract, Transform, Load  
<sup>10</sup>Sentiment Lexicon  
<sup>11</sup>Mean Absolute Error (MAE)  
<sup>12</sup>Basic Cosine  
<sup>13</sup>Adjusted Cosine  
<sup>14</sup>Correlation  
<sup>15</sup>Bench-Mark  
<sup>16</sup>Weighted Sum  
<sup>17</sup>Regression  
<sup>18</sup>Nearest-Neighbor CF Algorithm  
<sup>19</sup>Cosine Distance  
<sup>20</sup>Query  
<sup>21</sup>Too General  
<sup>22</sup>Too Narrow  
<sup>23</sup>Flowchart  
<sup>24</sup>Single Scan  
<sup>25</sup>Development Set  
<sup>26</sup>Eclipse  
<sup>27</sup>Cross Validation  
<sup>28</sup>Sparsity  
<sup>29</sup>Cold-Start

[18] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in Proceedings of the 10th international conference on World Wide Web. ACM, 2001, pp. 285–295.

[19] G. Linden, B. Smith, and J. York, "Amazon. Com recommendations: Item-to-item collaborative filtering," IEEE Internet computing , vol. 7, no. 1, pp. 76–80, 2003.

[20] H. Byström, "Movie Recommendations from User Ratings," Stanford University, Tech. Rep, 2013.

[21] M. Fu, "CS224W Project Report," 2011.

[22] K. Miyahara, and M. J. Pazzani, "Collaborative filtering with the simple Bayesian classifier," In PRICAI 2000 Topics in Artificial Intelligence (pp. 679-689), Springer Berlin Heidelberg, 2000.

[23] E. Keogh, "Naïve Bayes Classifier," Dr. Eamonn Keogh's lecture- UCR, 2011.

**الناز مظاهری** فارغ‌التحصیل کارشناسی‌ارشد رشته فناوری اطلاعات گرایش تجارت الکترونیکی از دانشگاه گیلان هستند و زمینه‌های مورد علاقه ایشان داده‌کاوی، متن‌کاوی و هوش تجاری هستند.  
 آدرس پست‌الکترونیکی ایشان عبارت است از:  
 elnaz.mazaheri27@yahoo.com



**سید ابوالقاسم میرروشندل** فارغ‌التحصیل از دانشکده فنی دانشگاه تهران در مقطع کارشناسی در رشته مهندسی کامپیوتر با گرایش نرم‌افزار و دانشگاه صنعتی شریف در مقاطع کارشناسی‌ارشد و دکتری در رشته مهندسی کامپیوتر گرایش هوش مصنوعی. از سال ۱۳۹۱ عضو هیات علمی گروه مهندسی کامپیوتر دانشگاه گیلان بوده و زمینه‌های مورد علاقه ایشان پردازش زبان‌های طبیعی، داده‌کاوی، یادگیری ماشینی و پردازش تصویر هستند.  
 آدرس پست‌الکترونیکی ایشان عبارت است از:  
 mirroshandel@guilan.ac.ir



#### اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۱۲/۲۵

تاریخ اصلاح: ۱۳۹۶/۰۷/۱۰

تاریخ قبول شدن: ۱۳۹۷/۰۲/۰۳

نویسنده مرتبط: دکتر سید ابوالقاسم میرروشندل، دانشکده فنی، دانشگاه گیلان، رشت، ایران.

<sup>1</sup>Customer Relationship Management (CRM)

<sup>2</sup>Social Customer Relationship Management (SCRM)

<sup>3</sup>Recommender System

<sup>4</sup>Hashtag

<sup>5</sup>Link

<sup>6</sup>Contacts



## مدل سازی تکامل شبکه های ارتباطات کلاسی نرم افزارهای شی گرا

مجتبی صادقیان محمد خوانساری فرید دهقان

دانشکده علوم و فنون نوین، دانشگاه تهران، تهران، ایران

### چکیده

شبکه ای ارتباطات کلاس های نرم افزارها با نگهداری و توسعه نرم افزار تکامل پیدا می کند. مطالعه تغییرات ویژگی های این شبکه ها به فهم روند طراحی و قواعد حاکم بر رشد نرم افزار کمک می کند. هدف ما در این مقاله، مدل سازی تکامل شبکه ساختار نرم افزارهای شی گرا، با بررسی خصوصیات شبکه در روند تغییرات مهندسی نرم افزار (ایجاد و حذف کلاس ها) است. سیستم های نرم افزاری شی گرا به صورت شبکه ای جهت دار که در آن گره ها متناظر با کلاس ها و یال ها متناظر با ارتباطات بین آن ها است (شامل انجمنی، ارث بری، تحقق و تجمع)، مدل سازی می شوند. مطالعات ما نشان داده شبکه نرم افزار دارای خصوصیات جهان کوچک بوده و گره های ایجاد شده و حذف شده دارای توزیع های درجه ورودی قانون توانی و درجه خروجی نرمال هستند. ارزیابی مدل پیشنهادی بر روی داده های واقعی پروژه متن باز EGit (با ۱۳ نسخه و میانگین ۷۲۰ کلاس) انجام می شود. نتایج این مدل سازی به پیش بینی روند توسعه نرم افزار و استخراج قواعد حاکم بر تکامل نرم افزار به توسعه دهندگان کمک می کند.

**کلمات کلیدی:** کاوش مخازن کد منبع نرم افزار، تحلیل شبکه، تکامل شبکه، شبکه ای ارتباطات کلاسی نرم افزارهای شی گرا.

### ۱- مقدمه

شکل شبکه امکان مطالعه رفتار، ویژگی های ساختاری و تغییرات ساختار کلاسی نرم افزار را فراهم می کند [۶]، [۷].

کاوش مخازن کد منبع سیستم های نرم افزاری<sup>۲</sup> از منابع اصلی شناسایی الگوهای تغییرات پویای سیستم های نرم افزاری است. کاوش فایل های نرم افزاری مخازن کد و استخراج نمودار کلاسی UML پروژه های نرم افزاری، امکان نمایش شبکه ای ساختار نرم افزار در سطح کلاس را فراهم می کند [۱]، [۸].

در سال های اخیر، مطالعات فراوانی بر روی شبکه های ساختاری نرم افزارها صورت گرفته است و خواص توزیع درجات توانی، جهان کوچکی، ساختار انجمنی را بررسی کرده است [۹]. برای نمونه لی و همکاران با مشاهده خاصیت پیمانهای شبکتهای نرم افزار واقعی، روش اتصال پیمانهای بجای پیوست تک گره برای شبیه سازی تکامل شبکه نرم افزار ارائه کردند. در این روش به منظور اتصال یال های ورودی و خروجی در شبکه جهت دار به هنگام اتصال گره های جدید به شبکه اولیه، از روش اتصال ترجیحی<sup>۴</sup> استفاده شده است [۷]. احتمال اتصال یال های جدید در روش اتصال پیمانهای با درجه های ورودی و خروجی گره ها (میزان وابستگی کلاس ها) متناسب است. تعداد یال های اتصالی متناسب با میزان همبستگی و

تکامل سیستم های نرم افزاری شی گرا، به تغییرات حاصل از فرایند نگهداری<sup>۱</sup> و حفظ عملکرد سیستم و یا اضافه کردن عملکردهای جدید با تغییر نیاز اطلاق می شود. مدل سازی تکامل نرم افزارها به فهم روند طراحی و توسعه نرم افزار، تضمین کیفیت روند توسعه، پیش گویی تغییرات آینده و یافتن الگوهای تغییر بهینه در هر سیستم نرم افزاری کمک می کند. نتایج مدل سازی می تواند بصورت مستقیم در فعالیت های مهندسی بر پایه مدل جهت تست، طراحی و توسعه سیستم های نرم افزاری استفاده شود [۱].

روش های مختلفی برای مدل سازی تکامل سیستم های نرم افزاری ارائه شده است. تحلیل آماری کدها از جمله روش های شناخته شده برای مطالعه تکامل نرم افزار هست [۲]، مدل های آماری با استفاده از تعداد خطوط کد<sup>۲</sup> نرم افزار [۳]، رشد اندازه بخشی از بسته ها و یا فایل کلاس های اساسی نرم افزار را مدل سازی می کنند. مدل های آماری قادر به مدل سازی ساختار کلی نرم افزار از جمله ارتباطات بین کلاس ها نیستند [۴]، [۵]. نمایش ساختار سیستم های نرم افزار شی گرا به

را آشکار ساخت. جمع‌بندی از روش‌های بکار برده شده برای مدل‌سازی تکامل سیستم‌های نرم‌افزار در جدول ۱ نمایش داده شده است. مدل‌های تکامل شبکه‌ای پیشین، عمدتاً بر روی رشد شبکه متمرکز شده [۷] و یا اضافه/حذف شدن یال‌ها را در شبکه نرم‌افزار مدل‌سازی می‌کنند [۱۰]. مطالعات بر روی تغییرات نسخه‌های متوالی پروژه‌های شی‌گرا نشان داده است که الگوهای تغییرات هر دسته از مجموعه کلاس‌های شبکه از یکدیگر متفاوت هستند. تحلیل‌های آماری معیارهای مهندسی نرم‌افزار الگوی این تغییرات را در کلاس‌های نرم‌افزاری آشکار می‌کند [۱۱].

با در نظر گرفتن ایده‌ها و محدودیت‌های مدل‌سازی‌های پیشین و استفاده از معیارها تحلیل رشد نرم‌افزار، در این مقاله مدلی برای پیشگویی تکامل شبکه‌های نرم‌افزار بر پایه شبکه ساختار کلاس‌های نرم‌افزار و تحلیل تغییرات آماری کلاس‌ها در روند توسعه سیستم‌های نرم‌افزاری ارائه می‌شود. هدف مدل‌سازی تکامل شبکه‌های ارتباطات کلاسی نرم‌افزارهای شی‌گرا با استفاده از ابزارها و مفاهیم تحلیل شبکه‌های اجتماعی [۱۲] و مطالعه تغییرات شبکه‌ای ساختار در روند توسعه و نگهداری سیستم‌های نرم‌افزاری است. ما از شبکه ارتباطات کلاس‌های نرم‌افزار که در آن گره‌ها متناظر با کلاس‌ها که دارای خصوصیت<sup>۶</sup> سن و یال‌ها نشان‌دهنده ارتباطات بین آنها (شامل انجمنی<sup>۷</sup>، اثربری<sup>۸</sup>، تحقق<sup>۹</sup> و تجمع<sup>۱۰</sup>) است، برای مدل‌سازی تکامل سیستم‌های نرم‌افزاری استفاده می‌کنیم.

نوآوری مدل پیشنهادی ارائه شده، استفاده از شبکه‌های جهت‌دار نرم‌افزار به‌منظور نمایش انتقال اطلاعات و فراخوانی‌های بین پارامترهای کلاس‌ها است. علاوه بر این، ما از چهار ارتباط اساسی بین کلاس‌های نرم‌افزارهای شی‌گرا به منظور نمایش یال‌های کلاس‌ها در شبکه نرم‌افزار استفاده می‌کنیم. مدل پیشنهادی تغییرات گره‌های شبکه (ایجاد و حذف گره‌ها) را به منظور پیشگویی روند تکامل شبکه شبیه‌سازی می‌کند. در این مقاله، روند تکامل پروژه متن باز<sup>۱۱</sup> از خانواده پروژه‌های eclipse مورد مطالعه و تحلیل قرار گرفته است. همچنین، صحت و دقت مدل ارائه شده در پیشگویی نسخه‌های متوالی پروژه EGit مورد ارزیابی قرار می‌گیرد.

در ادامه مقاله، مدل پیشنهادی در بخش ۲ شرح داده می‌شود، ارزیابی نتایج شبیه‌سازی مدل پیشنهادی بر روی شبکه‌های واقعی در بخش ۳ ارائه می‌شود. در نهایت بخش ۴، شامل نتیجه‌گیری و پیشنهادها برای کارها آتی است.

## ۲- مدل تکاملی شبکه ساختار نرم‌افزار

در این بخش نخست مجموعه داده‌های مورد استفاده (پروژه EGit) شرح داده می‌شود. شبکه ارتباطات کلاسی پروژه EGit توصیف و خصوصیات آماری این شبکه مورد مطالعه قرار می‌گیرد و تغییرات آماری کلاس‌ها در طول تکامل پروژه EGit ارزیابی می‌گردد. در نهایت براساس تحلیل‌های کسب شده، مدل تکامل پیشنهادی ارائه می‌گردد.

### ۲-۱- مجموعه دادگان

در این مقاله ما از پروژه‌ی متن باز eclipse جهت نمایش شبکه نرم‌افزار استفاده می‌کنیم. Eclipse یک محیط توسعه نرم‌افزاری چندزبانه برای محیط توسعه مجتمع با قابلیت اضافه کردن افزونه هست. این محیط توسعه در ابتدا با زبان جاوا و برای توسعه برنامه‌های این زبان استفاده می‌شده است. EGit یک سرویس از سوی تیم ارائه‌دهندگان eclipse برای سیستم کنترل نسخه Git است. Git یک<sup>۱۲</sup> SCM توزیع شده است که به توسعه‌دهندگان این امکان را می‌دهد تا یک کپی کامل از تمام گذشته هر نسخه از کد در اختیار داشته باشند. EGit انجام

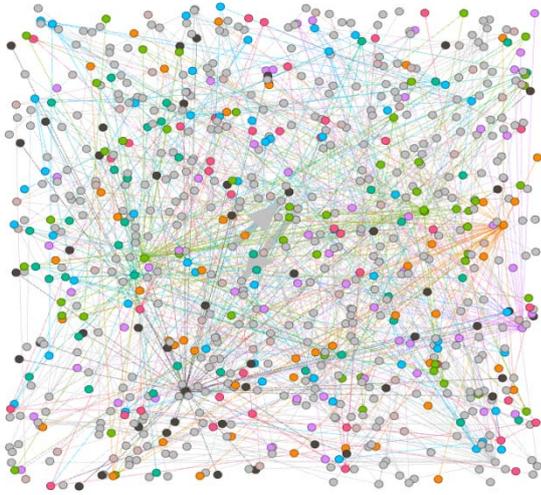
جفت‌شدگی در شبکه است. شبکه تولید شده توسط روش اتصال پیمان‌های خاصیت جهان کوچکی را از خود نشان می‌دهد. در این روش تنها رشد سیستم‌های نرم‌افزاری برای مدل‌سازی تکامل شبکه‌های نرم‌افزاری در نظر گرفته شده و تغییراتی چون حذف کلاس‌ها نادیده گرفته شده است [۷]. لذا قادر به شبیه‌سازی رفتار واقعی سیستم‌های نرم‌افزاری نیست. چیکلیز و همکاران مدل پیش‌گویی بر پایه شبکه و با ترکیب اطلاعات پیشین تکامل نرم‌افزار و قواعد مرتبط با حوزه شبکه (برای نمونه خصوصیات مقیاس آزادی و قوانین وراثت بین کلاس‌ها) برای تکامل نرم‌افزارها ارائه شده است [۱۰]. آنها فرایندهای رشد نرم‌افزار برای نمونه ایجاد و حذف ارتباط و مفهوم ساختار انجمنی را در مدل‌سازی لحاظ کرده‌اند. این مدل در یک حلقه تکراری با محاسبه تعداد نسخه‌ها جهت شبیه‌سازی، اتصال ترجیحی، رونوشت رفتار گذشته و تاثیر دادن سن یال‌ها و قواعد حوزه، اقدام به اضافه کردن گره‌های جدید می‌کند. توجه صرف به ارتباطات بین کلاس‌ها و نمونه‌برداری از توزیع‌های تجمعی حذف/ایجاد یال‌ها از جمله محدودیت‌های این مدل‌سازی است.

جدول ۱- روش‌های پیشگویی بکار برده شده برای مدل‌سازی تکامل ساختار نرم‌افزار

مرجع	نام مقاله	روش پیشنهادی	مزایا	معایب	مجموعه داده‌ها
[۱۱]	تکامل شبکه سیستم عامل لینوکس (۲۰۱۷)	کاوش مخازن کد نرم‌افزار با استفاده از معیارهای شبکه‌های پیچیده	تحلیل بر پایه رخداد	عدم در نظر گرفتن کلاس‌ها (اصلی‌ترین مولفه سیستم‌های شی‌گرا)	LOS
[۱۳]	چهارچوبی برای کسب، مدل‌سازی ایستا و تحلیل تکامل مدل‌های نرم‌افزاری (۲۰۱۶)	مدل‌سازی ایستا داده‌های کسب شده براساس مدل ترکیبی ARMA-GARCH	شبیه‌سازی رفتار ایستا پروژه‌های نرم‌افزاری	حجم بالای اطلاعات و محاسبه پیچیده ماتریس محاسبات	Java
[۱۴]	پیش‌بینی روند تکامل نرم‌افزار Java با به کار گیری مدل‌های شبکه‌ای (۲۰۱۵)	پیشگویی بر پایه شبکه و داده‌های تکامل نرم‌افزار	تاثیر دادن سن یال‌ها مدل‌سازی اضافه کردن گره‌های جدید و تغییر ارتباطات	توجه صرف به ارتباطات بین کلاس‌ها (یال‌های شبکه) نمونه‌برداری از توزیع‌های تجمعی حذف/ایجاد یال‌ها	aTunes FreeCol JEdit Jmol Weka
[۷]	روش اتصال پیمان‌های برای تکامل شبکه‌های نرم‌افزاری (۲۰۱۳)	روش اتصال ترجیحی بسته‌های کلاس‌ها	نمایش خاصیت پیمان‌های در شبکه‌های نرم‌افزار و مدل‌سازی اتصال ترجیحی	عدم در نظر گرفتن تغییرات انجمنی‌های موجود در نرم‌افزار (حذف کلاس‌ها)	Eclipse

گروه دیگری از تحقیقات به مطالعه و تحلیل روند، رفتار و خصوصیات تکامل نرم‌افزار از دید شبکه می‌پردازند. برای نمونه زایو و همکاران تکامل شبکه سیستم عامل لینوکس<sup>۵</sup> را در نسخه‌های انتشاری مورد مطالعه قرار دادند [۱۱]. آن‌ها نسخه‌های انتشاری سیستم عامل را بصورت شبکه‌ی جهت‌دار نمایش دادند، که در آن توابع با گره‌ها و فراخوانی بین آنها با یال‌ها نمایش داده شده است. نتایج تحلیل نشان‌دهنده رشد خطی اندازه سیستم عامل به همراه نزول یکنواخت ضریب خوشه‌بندی است. توزیع درجه‌های ورودی و خروجی شبکه ساختاری هسته لینوکس توانی است. همچنین مطالعات آنها نشان دهنده تکامل پیمان‌های توابع براساس هفت دنباله از تغییرات شامل مداومت، رشد، انقباض، تولد، تقسیم، حذف و ادغام است. تحلیل آماری این تغییرات، الگوی و نسبت تغییرات پیمان‌های توابع

کلاس مقصد  $v_j$  است. خواص و رفتار شبکه ساختار کلاسی نرم‌افزار در ادامه تحلیل می‌گردد. ویژگی اساسی شبکه‌های ارتباطات کلاسی نرم‌افزار EGit الگوی مشابه توزیع درجه توانی گره‌ها (مشاهده شده در شبکه‌های مقیاس آزاد<sup>۱۳</sup>) است. وجود خصوصیات شبکه‌های مقیاس آزاد و توزیع توانی معمولاً منسوب به حضور اتصال ترجیحی است. در زمینه نرم‌افزارهای شی‌گرا، این مدل تکامل نمایانگر آن است که کلاس‌های که نقش مرکزی در سیستم دارند (در نقش «خدا»<sup>۱۴</sup>) که سرویس‌هایی را به کلاس‌های مشتری ارائه می‌کند) نقش جذب‌کننده را برای کلاس‌های جدید دارند. با تکامل گام به گام نرم‌افزار این پدیده، موجب اتصال درصد زیادی از کلاس‌های جدید به کلاس‌های با درجه ورودی بالا می‌شود و در نتیجه گره‌های (کلاس‌های) مهم در شبکه مهم باقی می‌مانند که این امر نشانه پایداری طراحی سیستم‌های نرم‌افزاری هست، به این معنی است که رفتار بسیاری از اشیاء<sup>۱۵</sup> در طول زمان پایدار خواهند بود و نیازی به تغییر نیست [۷]، [۱۴].



شکل ۱- شبکه ساختار کلاس پروژه EGit نسخه v4.3.0 با ۷۹۹ گره و ۸۴۲ یال

تفاوت اساسی بین شبکه‌های ساختاری نرم‌افزار و دیگر شبکه‌های پیچیده همبستگی<sup>۱۶</sup> منفی بین درجه‌های ورودی و خروجی در شبکه‌های نرم‌افزاری است [۷]. عدم تقارن بین توزیع درجه‌های ورودی و خروجی نشان می‌دهد که گره‌های با درجه بزرگتر غالباً دارای درجه خروجی کوچک هستند و برعکس. به علاوه، گره‌های با درجه ورودی بزرگ معمولاً وظایف ساده و بنیادی را انجام می‌دهند. کلاس‌های متناظر با این گره‌ها مکرراً استفاده می‌شوند و ضرورتاً وابسته به دیگر کلاس‌ها نیستند. در مقابل، گره‌های با درجه خروجی بزرگتر معمولاً به گره‌های زیادی وابسته‌اند و دارای ساختار داخلی پیچیده‌ای هستند و بنابراین احتمال وابستگی به دیگر کلاس‌ها بسیار اندک است [۱۶].

مطالعه بر روی شبکه‌های نرم‌افزارهای شی‌گرا نشان‌دهنده ساختار انجمنی<sup>۱۷</sup> در این شبکه‌ها است [۹]. شبکه نرم‌افزار در طول توسعه و تکامل خود این ساختار انجمنی را حفظ می‌کند و این به دلیل توسعه نرم‌افزارهای شی‌گرا براساس فرضیه «لگو»<sup>۱۸</sup> است. این فرضیه بیانگر ساخت نرم‌افزار بر پایه تعداد زیادی از انجمن‌های کوچک و مستقل از هم است [۱].

## ۲-۳- تغییرات کلاس‌های نرم‌افزارهای شی‌گرا

در این مقاله ما توسعه و تکامل سیستم‌های نرم‌افزاری شی‌گرا را براساس دو تغییر اساسی ایجاد و حذف کلاس‌ها مطالعه می‌کنیم. استخراج و شناسایی این تغییرات در نسخه‌های متوالی پروژه‌های نرم‌افزار با استفاده از مخازن کد منبع و نگاشت آنها

جستجوها در گذشته کد را بسیار سریع و تطبیق‌پذیر می‌کند. این پشتیبانی از طریق مجموعه‌ای از افزونه‌های EGit\_project ارائه شده است. قابلیت EGit بر پایه کتابخانه JGit است. ما از ۱۳ نسخه پروژه EGit به شرح جدول ۲ استفاده می‌کنیم.

## ۲-۲- شبکه‌ی ارتباطات کلاسی نرم‌افزار

کلاس‌ها اجزای اصلی در سیستم‌های نرم‌افزاری شی‌گرا هستند. در این مقاله ما از شبکه جهت‌دار ارتباطات بین کلاس‌های نرم‌افزار استفاده می‌کنیم. گره‌ها در این شبکه متناظر با کلاس‌ها هستند که با نام کلاس‌ها برچسب خورده و حاوی خصوصیت سن هر کلاس هستند. سن هر گره متناسب با تعداد گام‌های است که کلاس مفروض در نسخه‌های متوالی توسعه پروژه حضور دارد و دچار تغییر نشده است. برای نمونه در پروژه EGit یک کلاس خاص در نسخه v2.0.0 دارای سن یک و با فرض عدم تغییر همان کلاس در روند توسعه در نسخه v4.3 دارای سن ۱۳ خواهد بود.

جدول ۲- نسخه‌های پروژه EGit

Num.	Tags	Release Date	Size
1	v2.0.0	2012-05-31 01:56:11	11.0 MB
2	v3.0.0	2013-05-08 17:04:52	11.8 MB
3	v3.1.0	2013-09-27 15:11:43	11.9 MB
4	v3.2.0	2013-11-13 18:42:22	12.2 MB
5	v3.3.0	2014-03-03 03:08:25	12.4 MB
6	v3.4	2014-05-06 02:32:29	13.7 MB
7	v3.5.0	2014-09-26 10:38:59	13.9 MB
8	v3.6.0	2014-11-12 19:41:59	14.0 MB
9	v3.7.0	2015-02-04 02:31:49	14.2 MB
10	v4.0.0	2015-03-23 21:36:36	14.3 MB
11	v4.1	2015-09-28 12:15:45	15.1 MB
12	v4.2	2015-11-11 01:32:52	15.2 MB
13	v4.3	2016-03-23 16:15:38	15.6 MB

یال‌های بین گره‌های شبکه ارتباطات جهت‌دار بین کلاس‌های سیستم نرم‌افزاری را نمایش می‌دهند. استفاده از یال‌های جهت‌دار در ساختار شبکه امکان نمایش دقیق‌تر جریان و انتقال اطلاعات (برای مثال ساختار فراخوانی‌ها، انتقال پیام و وراثت) را فراهم می‌کند [۱۵]. مدل پیشنهادی از چهار ارتباط اساسی بین کلاس‌های نرم‌افزار به نام‌های انجمنی، ارث بری، تحقق و تجمع (کسب شده از تجمیع ماتریس مجاورت ارتباطات در نمودار کلاسی UML) برای نمایش ارتباطات بین کلاس‌ها در شبکه نرم‌افزارهای شی‌گرا استفاده می‌کند. شکل ۱ نمونه‌ای از شبکه ساختار کلاسی نسخه v4.3.0 پروژه EGit را نمایش می‌دهد. شبکه ساختار کلاسی نرم‌افزار به صورت زیر نمایش داده می‌شود.

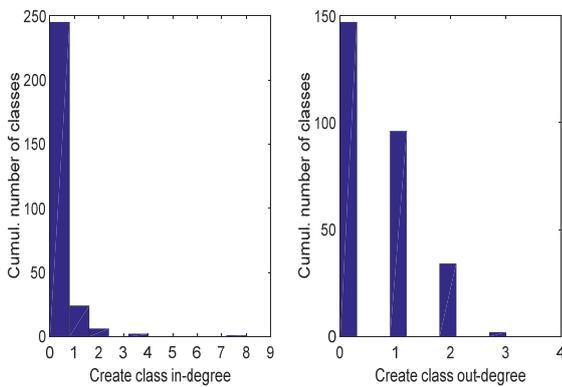
$$N = (V, E) \quad (1)$$

که در آن  $G$  گراف جهت‌دار ساختار کلاس نرم‌افزار،  $V = \{v_1, v_2, \dots\}$  مجموعه گره‌های شبکه که نمایانگر کلاس‌های نرم‌افزار و  $E = \{e_{i,j}, \dots\}$  مجموعه یال‌های شبکه که در آن  $v_i \rightarrow v_j$  نشان دهنده ارتباطات از کلاس مبدأ  $v_i$  به

می‌کند. برای مدل‌سازی اتصال کلاس‌های جدید به شبکه نرم‌افزار ما خصوصیات (توزیع درجات ورودی/خروجی) تمام کلاس‌های ایجاد شده (در ۱۳ نگارش متوالی) را مطالعه می‌کنیم. در هر گام تکامل شبکه نرم‌افزار، با شناسایی گره‌های متناظر با کلاس‌های ایجاد شده (با استفاده از برجسب هر گره) تعداد همسایه‌های ورودی و خروجی گره‌ها شمارش و نمودار فراوانی تجمعی برای درجه‌های ورودی و خروجی مجموع تمام کلاس‌های ایجاد شده در روند تکامل شبکه نرم‌افزار EGit ترسیم می‌شود.

شکل ۳ مجموع تعداد کلاس‌های ایجاد شده (اضافه شده به شبکه) از هر درجه برای درجات ورودی و خروجی را بصورت مجزا نمایش می‌دهد. نمودار حاصله توزیع درجه‌های ورودی/خروجی تمام کلاس‌های ایجاد شده در اتصال به شبکه‌های پیشین برای پروژه EGit را آشکار می‌کند.

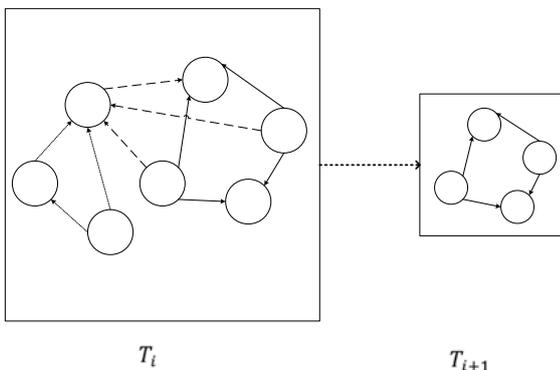
مدل تکامل پیشنهادی با بهره‌گیری از توزیع‌های درجات گره‌ها روند ایجاد گره‌های جدید در شبکه نرم‌افزار را نشان می‌دهد.



شکل ۳- فراوانی درجه‌های کلاس‌های تازه اضافه شده

### ۲-۳-۲- حذف کلاس‌های قدیمی

قدیمی شدن برخی از عملکردها در روند توسعه نرم‌افزار و یا عدم نگهداری طولانی مدت منجر به حذف کلاس‌های از ساختار نرم‌افزار می‌شود، برای نمونه حذف ارجاع دهنده HoverManager در نسخه ۷.3.6.0. شکل ۴ نشان‌دهنده مفهوم حذف کلاس‌ها در شبکه نرم‌افزار است.



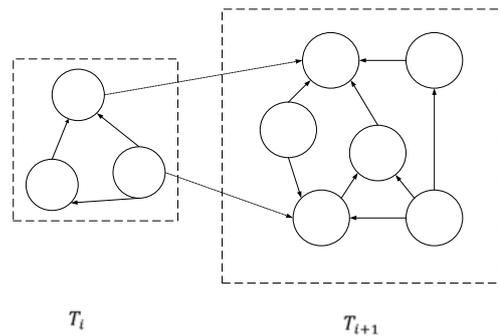
شکل ۴- حذف کلاس‌های قدیمی

کلاس‌های حذف شده به آسانی قابل شناسایی‌اند (با مقایسه نام کلاس‌ها در نسخه‌های متوالی). دنبال کردن لیست کلاس‌های حذف شده در نسخه‌های متوالی توسعه نرم‌افزارهای شی‌گرا امکان تحلیل الگوی حذف این کلاس‌ها در ساختار

به شبکه ساختاری نرم‌افزار به ما در فهم الگوی تکامل ساختار نرم‌افزارهای شی‌گرا کمک می‌کند.

### ۲-۳-۱- ایجاد کلاس‌های جدید

اساس رشد سیستم‌های نرم‌افزاری بر پایه اضافه شدن مجموعه کلاس‌های جدید به ساختار پروژه نرم‌افزار به‌منظور اضافه کردن عملکردهای جدید از سوی کاربران و توسعه‌دهندگان است، (شکل ۲ اضافه شدن کلاس‌ها به شبکه نرم‌افزار را نمایش می‌دهد). برای نمونه وابستگی maven به اجزاء نمایه‌ساز در نسخه ۷.4.2.0 پروژه EGit توسط توسعه‌دهندگان به عنوان ابزاری برای خودکارسازی ساخت اضافه گردیده است.



شکل ۲- ایجاد کلاس‌های جدید

مقایسه کلاس‌های نسخه‌های متوالی پروژه‌های نرم‌افزاری شی‌گرا می‌تواند به شناسایی کلاس‌های تازه اضافه شده کمک کند. بصورت عملی با مقایسه لیست کلاس‌های نرم‌افزار در دو نسخه متوالی  $T_i$  و  $T_{i+1}$  می‌توان تعداد کلاس‌های جدید اضافه شده به نسخه متوالی به همراه لیست کامل این کلاس‌ها را کسب کرد، (جدول ۳).

جدول ۳- دنباله تعداد کلاس‌های اضافه شده

Projects	Tags	Creation
EGit	v2.0.0.	-
	v3.0.0	49
	v3.1.0	4
	v3.2.0	34
	v3.3.0	16
	v3.4.0	15
	v3.5.0	21
	v3.6.0	6
	v3.7.0	9
	v4.0.0	8
	v4.1.0	89
	v4.2.0	5
	v4.3.0	31

مشاهد رفتار گره‌های متناظر با کلاس‌های جدید در شبکه نسخه متوالی ( $T_{i+1}$ ) پروژه نرم‌افزار، به استخراج الگوی اتصال گره‌های جدید به شبکه کمک

## ۲-۴- مدل تکامل ساختاری پیشنهادی

در این بخش مدل تکامل پیشنهادی بر پایه توزیع‌های کسب شده از دو نوع تغییر، ایجاد یا حذف گره‌های متناظر با کلاس‌ها ارائه می‌گردد. شبکه اولیه برای اعمال الگوهای تغییرات در روش پیشنهادی، نمودار UML ارتباطات کلاسی پروژه واقعی نرم‌افزار است. برخلاف مدل‌سازی‌های پیشین ما از ۴ ارتباط اساسی بین کلاس‌های نرم‌افزاری (شامل انجمنی، ارث بری، تحقق و تجمع) برای ترسیم ارتباطات بین گره‌های شبکه استفاده کردیم، که باعث شده شبکه ترسیم شده شباهت بیشتری به مدل واقعی سیستم نرم‌افزاری از جنبه نمایش ارتباطات کلاسی داشته باشد. فرایند کلی مدل‌سازی تغییرات ساختار نرم‌افزار در شکل ۶ نمایش داده شده است.



شکل ۶- مراحل مدل پیشنهادی تکامل ساختاری

در اینجا تغییرات اشاره به فرایند ایجاد و حذف کلاس‌های نرم‌افزاری دارد. در ادامه مدل‌سازی هر یک از دو تغییر اساسی در روند توسعه نرم‌افزار بر روی شبکه نرم‌افزار تشریح می‌شود.

## ۲-۴-۱- مدل‌سازی ایجاد گره‌های جدید

ایجاد کلاس‌های جدید در روند رشد نرم‌افزار شامل اضافه شدن مجموعه‌ای از کلاس‌ها به ساختار اصلی نرم‌افزار است. مجموعه کلاس‌های جدید می‌تواند تک کلاس و یا تعدادی از کلاس‌ها (بسته) باشند که با یکدیگر ارتباط دارند. تعداد گره‌های تازه اضافه شده به شبکه نرم‌افزار با محاسبه توزیع احتمالی گسسته بر روی مجموعه دنباله داده‌های تعداد کلاس‌های ایجاد شده در هر گام تکاملی (جدول ۳) بدست می‌آید. تابع جرم احتمال برای وقوع هر یک از اعداد دنباله تعداد کلاس‌های ایجاد شده توسط رابطه ۱ محاسبه می‌شود.

$$P(C) = \frac{\text{Num.ofCreatedClasses}}{\text{Sum(Num.ofCreatedClasses)}} \quad (2)$$

که در آن Num.of Created Classes تعداد کلاس‌های ایجاد شده در هر نسخه انتشاری پروژه نرم‌افزار و  $P(\text{Num.of Created Classes})$  فراوانی نسبی تعداد کلاس‌های ایجاد شده است.

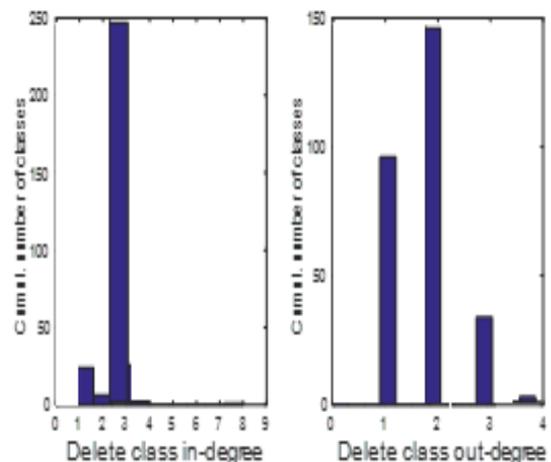
شبکه نرم‌افزار را برای ما فراهم می‌کند. همانند تحلیل ایجاد کلاس‌های جدید، به منظور مدل‌سازی حذف کلاس‌ها در روند تکامل شبکه نرم‌افزار، مجموع تمام کلاس‌های حذف شده پیشین شناسایی و تحلیل می‌شوند. جدول ۴ تعداد کلاس‌های حذف شده در گام‌های توسعه پروژه نرم‌افزار (۱۳ نگارش پروژه EGit) را به ما نمایش می‌دهد.

مطالعات ما بر روی شبکه کلاس‌های نرم‌افزار نشان‌داد احتمال حذف شدن کلاس‌ها وابسته به درجات گره‌های متناظر در شبکه سیستم‌افزاری است.

جدول ۴- دنباله تعداد کلاس‌های حذف شده

Projects	Tags	Deletion
EGit	v2.0.0	-
	v3.0.0	5
	v3.1.0	3
	v3.2.0	16
	v3.3.0	10
	v3.4.0	2
	v3.5.0	6
	v3.6.0	0
	v3.7.0	1
	v4.0.0	1
	v4.1.0	5
	v4.2.0	0
	v4.3.0	12

شکل ۵، مجموع تعداد کلاس‌های حذف شده (گره‌های حذف شده از شبکه) برای درجات ورودی و خروجی را بصورت مجزا نمایش می‌دهد. نمودار حاصله توزیع درجه‌های ورودی/خروجی تمام کلاس‌های حذف شده در اتصال به شبکه‌های پیشین برای پروژه EGit را آشکار می‌کند. الگوهای توزیع‌های بدست آمده به مدل‌سازی حذف گره‌ها در روند تکامل شبکه کمک می‌کند.



شکل ۵- توزیع درجه‌های کلاس‌های حذف شده

درجه خروجی مشخص در شبکه نرم‌افزار، فرمول ۷،  $P_{Network\_class\_age}$  احتمال شرطی انتخاب گره با سن خاص و  $P_{In\_neighbor}$  مقدار تابع جرم احتمال برای انتخاب گره‌ها به عنوان همسایه ورودی است.

$$P_{Network\_class\_outdegree} = \frac{P_{degree\ i\ between\ New\ class\ outdegree}}{P_{degree\ i\ in\ whole\ network}} \quad (7)$$

با در نظر گرفتن تعداد درجه‌های ورودی/خروجی هر گره تازه ایجاد شده و براساس مقادیر کسب شده برای تابع جرم احتمال همسایه‌ها و توزیع احتمال گسسته، همسایه‌های هریک از گره‌های اضافه شده به شبکه نرم‌افزار تعیین می‌شوند.

مطالعه مجموع کلاس‌های اضافه شده (بسته‌های نرم‌افزاری) در هر گام نشان داده است که مجموعه کلاس‌های تازه ایجاد شده زیر شبکه‌ای متصل از گره‌ها با توزیع درجات ورودی/خروجی توانی هستند. دنباله درجات ورودی/خروجی گره‌های شبکه مشابه درجه‌های اتصال گره‌ها به شبکه نرم‌افزار از رابطه ۳ بدست می‌آید. در ادامه با استفاده از الگوریتم [۱۷] زیرشبکه جهت‌دار گره‌های تازه ایجاد شده ساخته می‌شود. الگوریتم مذکور با استفاده از دنباله‌ی درجات گراف نخست دنباله را به صورت نزولی مرتب کرده و سپس در هر گام بزرگترین عدد دنباله را حذف و از بقیه رئوس یک واحد کم می‌کند (مشتمقات پیاپی دنباله) تا در مرحله آخر به دنباله‌ای می‌رسد که تمام جملاتش صفر است.

## ۲-۴-۲- مدل‌سازی حذف گره‌های قدیمی

حذف گره‌های شبکه نرم‌افزاری شامل دو مرحله است: (۱) تعیین تعداد گره‌های که باید حذف شوند. (۲) شناسایی و حذف گره‌ها از شبکه. مشابه مدل‌سازی ایجاد گره‌ها، تعداد گره‌های حذف شده در هر گام، با استفاده از توزیع احتمالی گسسته دنباله تعداد کلاس‌های حذف شده در پروژه‌ی واقعی نرم‌افزار (جدول ۴) محاسبه می‌شود.

مطالعات ما نشان‌دهنده وابستگی تابع الگوی حذف یک گره خاص به ۳ پارامتر متمایز است. پارامتر اول احتمال حذف مربوط به خود گره به شرط داشتن ویژگی‌های خاص (با معیارهای درجات و سن گره‌ها)، پارامتر دوم احتمال حذف گره به شرط داشتن همسایه‌های خروجی خاص و پارامتر سوم احتمال حذف گره به شرط داشتن همسایه‌های ورودی خاص است. تحلیل آماری دنباله درجات حذف شده نشان‌دهند توزیع درجه ورودی توانی و توزیع درجه خروجی نرمال است (شکل ۵). به‌صورت عملی احتمال حذف گره در شبکه توسط رابطه ۸ محاسبه می‌شود.

$$P_{Delete\_node} = P_{Delete\_Network\_node} + P_{Delete\_node\_outneighbor} + P_{Delete\_node\_inneighbor} \quad (8)$$

که در آن  $P_{Delete\_Network\_node}$  احتمال حذف گره با در نظر گرفتن ویژگی‌های شبکه‌ای،  $P_{Delete\_outneighbor}$  احتمال حذف گره با در نظر گرفتن احتمال همسایه‌های خروجی،  $P_{Delete\_inneighbor}$  احتمال حذف گره  $i$  با در نظر گرفتن احتمال همسایه‌های ورودی و  $P_{Delete\_node}$  مقدار تابع جرم احتمال برای انتخاب گره‌ها جهت حذف شدن است.

ویژگی‌های شبکه‌ای گره متناسب با درجه ورودی، خروجی و سن گره است، لذا سه صفت درجات ورودی و درجات خروجی و سن کلاس‌ها در کل باهم تعیین‌کننده احتمال حذف یک گره  $P_{Delete\_Network}$  هستند، رابطه ۹.

هر کلاس در مجموعه کلاس‌های تازه ایجاد شده با توجه به نقش و عملکردی که در سیستم دارد با مجموعه‌ای از کلاس‌های نرم‌افزار ارتباط خواهد داشت، بنابراین گام بعدی تشخیص تعداد و نوع ارتباط کلاس‌های تازه ایجاد شده (درجه ورودی/خروجی گره‌های متناظر در شبکه نرم‌افزار) است.

با فرض تک عضوی نبودن مجموعه گره‌های تازه ایجاد شده، مطالعات نشان می‌دهد که زیر شبکه ایجاد شده در روند توسعه نرم‌افزار توزیع درجه ورودی/خروجی توانی دارد (شکل ۳). لذا با کسب دنباله درجات ورودی/خروجی و نگاشت آن بر روی توزیع احتمال توانی، دنباله درجات ورودی/خروجی مجموعه گره‌های تازه ایجاد شده اضافه شده به ساختار شبکه اصلی نرم‌افزار محاسبه می‌شود.

فرض کنید  $P(x) = Cx^n$  توزیع توانی استخراج شده از دنباله درجات کلاس‌های تازه ایجاد شده است که در آن  $x \in [x_0, x_1]$  دنباله تعداد درجات ورودی/خروجی برای هریک از کلاس‌های جدید ایجاد شده توسط رابطه ۳ محاسبه می‌شود:

$$Rand\_PowerLaw(x) = [(x_1^{n+1} - x_0^{n+1})y + x_0^{n+1}]^{\frac{1}{(n+1)}} \quad (9)$$

که در آن  $y$  متغیر با توزیع یکنواخت با بازه  $[0,1]$ ،  $n$  توان توزیع،  $x_0, x_1$  محدوده توزیع و  $x$  متغیر توزیع شده توانی (درجه گره‌های تازه ایجاد شده) است. برای اتصال گره‌های تازه ایجاد شده به شبکه نرم‌افزار احتمال اتصال به گره‌های همسایه در شبکه نرم‌افزار محاسبه می‌شود. احتمال انتخاب هر گره متناسب با مقادیر تابع جرم احتمال برای انتخاب گره‌های شبکه نرم‌افزار به عنوان همسایه‌های گره‌های تازه ایجاد شده است. بررسی‌های ما نشان داده است که احتمال انتخاب هر گره همسایه متناسب با توزیع احتمال درجه‌های ورودی/خروجی و سن کلاس‌ها است. بنابراین برای همسایه‌های خروجی (کلاس‌هایی که به آنها وابسته‌اند) داریم:

$$P_{Out\_neighbor} = P_{Network\_class\_indegree} * P_{Network\_class\_age} \quad (4)$$

که در آن  $P_{Network\_class\_indegree}$  احتمال شرطی انتخاب گره با درجه ورودی خاص (احتمال رخداد گره  $i$  با درجه ورودی  $indegree$  در دنباله درجات ورودی کلاس‌های تازه ایجاد شده به شرط احتمال رخ داد همان گره با درجه ورودی مشخص در شبکه نرم‌افزار، فرمول ۵)،  $P_{Network\_class\_age}$  احتمال شرطی انتخاب گره با سن خاص (احتمال رخ داد گره با سن  $age$  در دنباله سن کلاس‌های تازه ایجاد شده به شرط احتمال رخ داد همان گره با سن مشخص در شبکه نرم‌افزار) و  $P_{Out\_neighbor}$  مقدار تابع جرم احتمال برای انتخاب گره‌ها به عنوان همسایه خروجی است.

$$P_{Network\_class\_indegree} = \frac{P_{degree\ i\ between\ New\ class\ indegree}}{P_{degree\ i\ in\ whole\ network}} \quad (5)$$

بطور مشابه برای همسایه‌های ورودی گره‌های تازه ایجاد شده (کلاس‌های که وابسته به کلاس‌های جدیداند) داریم:

$$P_{In\_neighbor} = P_{Network\_class\_outdegree} * P_{Network\_class\_age} \quad (6)$$

که در آن  $P_{Network\_class\_outdegree}$  احتمال شرطی انتخاب گره با درجه خروجی خاص (احتمال رخداد گره با درجه خروجی  $out\_degree$  در دنباله درجات خروجی کلاس‌های تازه ایجاد شده به شرط احتمال رخ داد همان گره با

شناسایی گره‌ها با استفاده از توزیع‌های بدست آمده براساس تعداد گام‌های شبیه‌سازی است.

**Algorithm: Evolution model  
(N, list of changing features)**

1. n.create, n.delete = Calculate\_Changes\_Number  
(list of all changes' number)
2. Delete\_classes(  
N, n.delete, list of deleted classes' degrees & ages)
3. Find\_neighbors(  
N, indegree, outdegree, list of classes' degrees & ages)
4. Create\_classes(  
N, n.create, list of created classes' degrees & ages)
5. Aging\_classes(N)
6. Return N

در گام اول تعداد کلاس‌های که باید حذف و ایجاد شوند محاسبه می‌شود. Calculate\_Changes\_Number تعداد کلاس‌های کاندیدای تغییرات ایجاد و حذف را با پارامتر ورودی لیست‌های تعداد کلاس‌های ایجاد شده و حذف شده مشخص می‌نماید (فرمول ۱). Delete\_classes تابع حذف کلاس‌های کاندید ورودی تابع شبکه نرم‌افزار تعداد کلاس‌های کاندید حذف و لیست درجات و سن کلاس‌های حذف شده تا کنون است. Create\_classes تابع ایجاد کلاس‌ها با پارامترهای ورودی شبکه نرم‌افزار، تعداد کلاس‌های کاندید برای ایجاد و لیست درجات و سن کلاس‌های ایجاد شده تا کنون است. Find\_neighbors همسایه‌های گره‌های جدید در شبکه ساختاری نرم‌افزار را مشخص می‌کند. به منظور محاسبه سن کلاس‌ها، تابع Aging\_classes تعریف شده است.

این تابع در هر گام شبیه‌سازی شبکه نرم‌افزار یک واحد به خصوصیت سن گره‌های شبکه اضافه می‌کند (برای کلاس‌های ایجاد شده در هر گام نخست سن واحد را تخصیص می‌دهد). گام‌های ۱ تا ۶ برای ایجاد شبکه نرم‌افزار در هر گام توسعه (نسخه‌های متوالی) تکرار می‌شوند و در هر گام ورودی شبکه نرم‌افزار شبکه ایجاد شده در مرحله قبلی است.

پیچیدگی زمانی الگوریتم پیشنهادی وابسته به مرحله ساخت زیرشبکه (بسته‌های نرم‌افزاری) است که برابر با پیچیدگی زمانی الگوریتم مرجع [۱۷] یا همان  $O(n \log n)$  است.

### ۳- نتایج شبیه‌سازی

ارزیابی مدل پیشنهادی با استفاده از ابزارها و سنجه‌های حوزه علوم شبکه با بکارگیری دادگان واقعی انجام می‌گیرد. عملکرد مدل پیشنهادی در مقایسه با شبکه‌های شبیه‌سازی شده توسط مدل اتصال پیمان‌های [۷] ارزیابی می‌شود.

با توجه به اینکه سایر روش‌های مدل‌سازی مطالعه شده بر پایه تحلیل توابع نرم‌افزار [۱۱]، تحلیل آماری تکامل نرم‌افزار [۱۳] و یا مدل‌سازی‌های وابسته به قواعد دامنه مجموعه داده‌ها [۱۴] هستند، مدل اتصال پیمان‌های [۷] نزدیک‌ترین مدل‌سازی به مدل پیشنهادی ما بر پایه شبیه‌سازی تکامل شبکه‌های ارتباطات کلاس‌های نرم‌افزار است.

مجموعه‌ای از نسخه‌های پیشین مجموعه دادگان (در حدود ۶۰٪ مجموعه نسخه‌های در دسترس نرم‌افزار برای هر پروژه) برای محاسبه توزیع خصوصیات شبکه (پارامترهای مدل پیشنهادی شامل توزیع درجات و سن) استفاده می‌شود. سایر نسخه‌ها به عنوان مجموعه داده تست استفاده می‌شوند. متغیرهای روش اتصال پیمان‌های به منظور مقایسه با مدل پیشنهادی به شرح جدول ۵ است.

$$P_{Delete\_Network\_node} = P_{Delete\_class\_indegree} * P_{Delete\_class\_outdegree} * P_{Delete\_class\_age} \quad (9)$$

که در آن  $P_{Delete\_class\_indegree}$  احتمال شرطی حذف گره با درجه ورودی خاص،  $P_{Delete\_class\_outdegree}$  احتمال شرطی حذف گره با درجه خروجی خاص،  $P_{Delete\_class\_age}$  احتمال شرطی حذف گره با سن خاص (احتمال حذف گره با سن age در دنباله سن کلاس‌های حذف شده به شرط احتمال رخ داد همان گره با سن مشخص در شبکه نرم‌افزار) است.

$$P_{Delete\_class\_indegree} = \frac{cdf\_power-law(indegree)}{p(indegree)} \quad (10)$$

$P_{Delete\_class\_indegree}$  احتمال حذف گره با درجه ورودی خاص یک احتمال شرطی برای انتخاب گره با درجه ورودی خاص از لیست دنباله درجات ورودی گره‌های حذف شده که توسط تابع توزیع تجمعی توانی به شکل،  $cdf\_power-law(indegree)$  (که از نگاشت بر روی دنباله درجات ورودی گره‌های حذف شده بدست می‌آید) محاسبه می‌شود به شرط انتخاب گره‌ای با درجه ورودی خاص در شبکه نرم‌افزار است.

$$P_{Delete\_class\_outdegree} = \frac{cdf\_Gaussian(outdegree)}{p(outdegree)} \quad (11)$$

$P_{Delete\_class\_outdegree}$  احتمال حذف گره با درجه خروجی خاص یک احتمال شرطی برای انتخاب گره با درجه خروجی خاص از لیست دنباله درجات خروجی گره‌های حذف شده که توسط تابع توزیع تجمعی نرمال،  $cdf\_Gaussian(outdegree)$  (که از نگاشت بر روی دنباله درجات خروجی گره‌های حذف شده بدست می‌آید) محاسبه می‌شود به شرط انتخاب گره‌ای با درجه خروجی خاص در شبکه نرم‌افزار است.

تحلیل روند حذف کلاس‌ها نشان‌دهنده همبستگی بین رفتار همسایه‌ها و گره حذف شده است. بنابراین، احتمال حذف یک گره وابسته به احتمال انتخاب همسایه‌های گره‌های حذف شده است، لذا داریم:

$$P_{Delete\_node\_inneighbor} = P_{node\_inneighbor\_outdegree} * P_{node\_inneighbor\_age} \quad (12)$$

$$P_{Delete\_node\_outneighbor} = P_{node\_outneighbor\_indegree} * P_{node\_outneighbor\_age} \quad (13)$$

که در آن  $P_{node\_outneighbor\_indegree}$  احتمال شرطی درجه‌های همسایه‌های گره‌های شبکه با در نظر گرفتن توزیع درجه‌های همسایه‌های گره‌های حذف شده در پروژه نرم‌افزاری واقعی (مجموعه دادگان) و  $P_{node\_outneighbor\_age}$  احتمال شرطی سن همسایه‌های گره حذف شده (کسب شده از احتمال شرطی سن همسایه‌های گره‌های حذف شده در پروژه نرم‌افزاری واقعی) است. در نهایت با محاسبه این احتمالات برای تمام گره‌های شبکه و استفاده از توزیع احتمال گسسته مدل قادر به شناسایی و حذف کلاس‌ها در روند توسعه نرم‌افزار خواهد بود.

### ۲-۵- تجمیع مدل‌سازی‌های تغییرات

فرایند شبیه‌سازی تکامل ساختار نرم‌افزار شامل تجمیع مدل‌سازی تغییرات حاصله بر روی شبکه اولیه (شبکه واقعی نسخه نرم‌افزار) و تکرار روند محاسبه احتمال‌ها و

جدول ۵- متغیرهای مدل اتصال پیمانهای

$n_0$	اندازه اولیه شبکه
$g$	نرخ رشد شبکه
$m$	شیب خط رشد گره ها و یال های شبکه (میانگین تعداد یال های اضافه شده با اضافه شدن تک گره)
$\Gamma$	نرخ جفت شدگی <sup>۱۹</sup> - درجه عمومی استفاده شده در نرم افزار (نسبت تعداد یال هایی که گره جدید را به شبکه اضافه می کنند بر تعداد کل یال های گره - مجموع یال های داخلی بسته گره و یال های که به شبکه نرم افزار متصل می کنند).

مقادیر این متغیرها برای پروژه EGIt در جدول ۶ نمایش داده شده است. معیارهای که در ادامه مورد مطالعه قرار می گیرند نماینده ساختار همبندی شبکه نرم افزار هستند، در نتیجه می توانند به عنوان شاخصی از میزان شباهت شبکه پیش بینی شده در مقایسه با نسخه نهایی سیستم نرم افزار واقعی در نظر گرفته شوند.

جدول ۶- مقادیر متغیرهای مربوط به مدل اتصال پیمانهای در مجموعه دادگان

Parameters	EGit project
$n_0$	10
$g = \frac{1}{n_0}$	0.1
$m$	9.021
$\Gamma$	0.4445

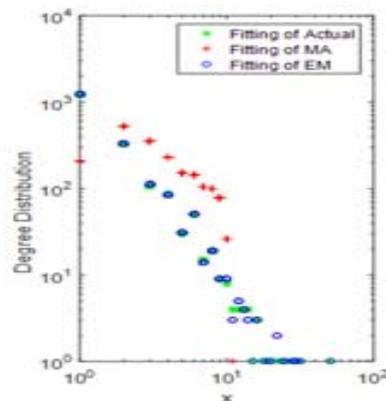
توجه: در این بخش ما از نمادهای زیر استفاده می کنیم.

- شبکه های واقعی (Actual Software Networks): Actual
- روش اتصال پیمانهای (Modular Attachment model): MA
- مدل پیشنهادی (The Proposed Evolution Model): EM

این نکته قابل توضیح است که اندازه شبکه برای پیش بینی تکامل پروژه در مدل پیشنهادی، شبکه اولیه نرم افزار EGIt (نسخه v2.0.0) است.

۳-۱- توزیع درجات

اولین معیار سنجش دقت توزیع درجه های شبکه های تولید شده توسط مدل پیشنهادی و روش اتصال پیمانهای در مقایسه با شبکه واقعی نرم افزار است. شکل ۷.



شکل ۷- توزیع درجات شبکه های واقعی (Actual)، شبکه های تولید شده توسط روش اتصال پیمانهای (MA) و مدل پیشنهادی (EM)

همانگونه که شکل ۷ نمایش می دهد، درجه شبکه های واقعی توزیع توانی دارند که مطابق با شبکه های تولید شده توسط مدل پیشنهادی است. در مقایسه با شبکه های تولید شده توسط مدل اتصال پیمانهای، شبکه های تولیدی مدل پیشنهادی دارای تطابق دقیق تری با توزیع درجات شبکه های واقعی eclipse است. از دید مهندسی نرم افزار کلاس های با قابلیت استفاده مجدد بیشتر محبوب هستند، که منجر به طراحی سیستم های نرم افزاری با درجه ورودی بزرگ می شوند. از سوی دیگر کلاس های با ساختار داخلی پیچیده اغلب به دلیل سختی نگهداری نرم افزار بسیار محبوب نیستند، بنابراین توزیع ارتباطات بین کلاس های نرم افزاری از توزیع توانی پیروی می کند (شکل ۷).

کلاس های پایه برای عملکرد ساختار نرم افزار حیاتی هستند. شکست برخی از گره های اساسی با درجه ورودی بالا در شبکه نرم افزار به شکنندگی اساسی عملکرد سیستم و نقص عملکرد منجر می شود، بنابراین حفظ عملکرد بدون خطای کلاس های اساسی بسیار مهم خواهد بود.

۳-۲- همبستگی بین درجات ورودی و خروجی

در شبکه های شبیه سازی و واقعی نرم افزار، ما از ضریب همبستگی برای سنجش رابطه بین درجه ورودی و خروجی استفاده می کنیم. نتایج ضریب همبستگی برای پروژه EGIt در جدول ۷ نمایش داده شده است. نتایج نشان دهنده عدم تقارن بین توزیع درجه ورودی و خروجی است.

جدول ۷- ضریب همبستگی در پروژه EGIt

Correlation coefficient	actual network	Modular attachment model	The Proposed Evolution Model
EGit v4.3.0	-0.2443	-0.1967	-0.2511

دلیل این عدم تقارن این است که گره های با درجه ورودی بالا نماینده کلاس های پایه و ساده در سیستم های نرم افزاری بوده، لذا به صورت مکرر استفاده شده و کلاس های دیگر به آنها وابسته هستند. در مقابل، گره ها با درجه خروجی بالا معمولاً نماینده کلاس های هستند که وابسته به گره های زیادی بوده و ساختار داخلی پیچیده ای دارند و در نتیجه احتمال وابستگی دیگر کلاس ها کمتر هست [۷].

۳-۳- خاصیت پیمانهای

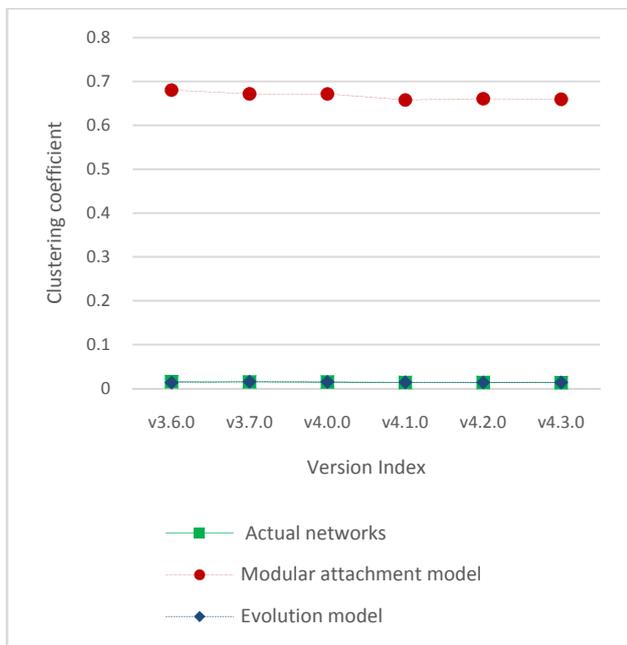
ساختار انجمنی شبکه های نرم افزار همواره در طول تکامل حفظ می شود. ما از معیار پیمانهای<sup>۲۰</sup> برای سنجش خاصیت پیمانهای شبکه های شبیه سازی شده و شبکه های واقعی استفاده می کنیم [۹]. پیمان های (Q) یک معیار استاندارد برای تعیین قدرت ساختار یک انجمن در شبکه است. مقدار پیمان های بال نشان می دهد که انجمن های به وضوح تعریف شده ای در شبکه وجود دارد. تعریف پیمان های به شکل زیر است، فرمول ۱۴.

$$Q = \sum (e_{ii} - a_i^2) \quad (14)$$

که در آن  $e_{ii}$  نسبت یال های بین گره های درون انجمن  $i$  به کل یال های گراف و  $a_i$  نشان دهنده نسبت تمام یال های است که از مرز انجمن عبور می کنند. شکل ۸ مقادیر پیمانهای محاسبه شده توسط الگوریتم لوین<sup>۲۱</sup> [۱۸] برای شبکه های شبیه سازی شده و شبکه های واقعی نمایش می دهد. مقادیر پیمانهای شبکه های

### ۳-۴- جهان کوچکی

خوشه‌بندی یکی از ویژگی‌های اساسی شبکه‌های جهان کوچک است. مقادیر ضریب خوشه‌بندی<sup>۲۲</sup> برای شبکه‌های شبیه‌سازی شده و شبکه‌های واقعی در شکل ۹ نمایش داده شده است.



شکل ۱۰- مقادیر ضریب خوشه‌بندی شبکه‌های واقعی و شبیه‌سازی شده

خوشه‌بندی معیاری برای سنجش همبستگی پیمانه‌های شبکه‌های نرم‌افزاری است. میانگین طول کوتاه‌ترین مسیر از دیگر معیار مهم در شبکه‌های جهان کوچک است. این معیار نشان‌دهنده سرعت ارتباطات در سیستم‌های نرم‌افزاری است. شکل ۱۰ نشان‌دهنده مقادیر میانگین طول مسیر<sup>۲۳</sup> برای شبکه‌های واقعی و شبیه‌سازی شده است.

همانطور که نتایج نشان دادند که شبکه‌های نرم‌افزاری با وجود داشتن اندازه بزرگ (تعداد گره‌ها و یال‌های زیاد) دارای میانگین طول مسیر کوتاه هستند، بنابراین انتقال داده‌های بین کلاس‌ها و پیمانه‌ها بسیار موثرتر است.

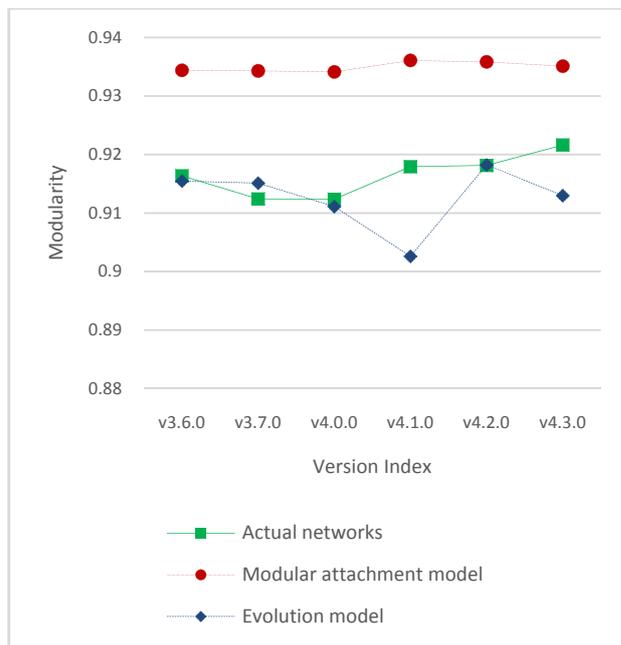
خوشه‌بندی بالا و میانگین طول مسیر کوتاه از ویژگی‌های شبکه‌های جهان کوچک هستند که در طول توسعه و تکامل شبکه نرم‌افزار و شبیه‌سازی مدل تکامل پیشنهادی حفظ شده‌اند.

مدل پیشنهادی در مقایسه با مدل اتصال پیمانه‌ای در شکل‌گیری سیستم‌های نرم‌افزاری واقعی بسیار موفق‌تر عمل کرده است. استفاده از تحلیل آماری تغییرات ساختار نرم‌افزار در مدل‌سازی شبکه نرم‌افزاری منجر به مدل‌سازی بهتری نسبت به مدل اتصال پیمانه‌ای شده است بگونه‌ای که شبکه پیش‌بینی شده دارای ساختاری بسیار مشابه به شبکه واقعی نرم‌افزار از لحاظ معیارهای مشخص‌کننده شبکه است.

### ۴- نتیجه‌گیری

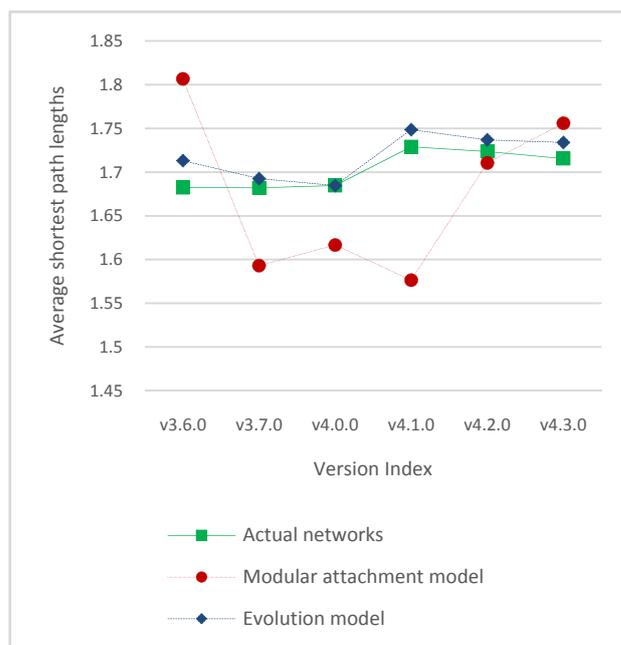
کاوش مخازن کد سیستم‌های نرم‌افزاری متن باز امکان مطالعه و تحلیل روند تکامل نرم‌افزار را فراهم کرده است. مدل‌سازی‌های تکامل سیستم‌های نرم‌افزار مبتنی بر شبکه، صرفاً بر پایه رشد و تغییر اتصالات شبکه‌های نرم‌افزاری بوده و سنجه‌های مهندسی نرم‌افزار را کمتر در نظر می‌گرفتند. در این مقاله، با کاوش

شبیه‌سازی شده توسط مدل تکاملی تفاوت اندکی با شبکه واقعی نرم‌افزارها دارد در حالی که مقادیر پیمانه‌ای در شبکه تولید شده توسط مدل اتصال پیمان‌های اختلاف قابل توجهی دارد.



شکل ۸- مقادیر پیمانه‌ای شبکه‌های واقعی و شبیه‌سازی شده

از دیدگاه مهندسی نرم‌افزار، نرم‌افزارهای با خاصیت پیمانه‌ای بالا نشان‌دهنده اصل حداقل جفت‌شدگی و حداکثر انسجام هستند، لذا بصورت ذاتی پیمانه‌هایی با ساختار قوی در شبکه نرم‌افزار ایجاد می‌گردند. علاوه بر این ساختار پیمانه‌های سیستم‌های نرم‌افزاری می‌تواند به تقویت همکاری داخل کلاسی با حداقل ارتباطات خارج پیمانه‌ای کمک کند [۹].



شکل ۹- میانگین طول کوتاه‌ترین مسیر در شبکه‌های واقعی و شبیه‌سازی شده

[4] G. Rasool, and N. Fazal, "Evolution Prediction and Process Support of OSS Studies: A Systematic Mapping," *Arab. J. Sci. Eng.*, May 2017.

[5] L. Madeyski, and M. Kawalerowicz, "Software engineering needs agile experimentation: a new practice and supporting tool," in *Software Engineering: Challenges and Solutions*. Springer, 2017, pp. 149–162.

[6] L.-Z. Zhu, B.-B. Yin, and K.-Y. Cai, "Generating Mechanisms for Evolving Software Mirror Graph," *J. Mod. Phys.*, vol. 03, no. 09, pp. 1050–1059, 2012.

[7] H. Li, H. Zhao, W. Cai, J.-Q. Xu, and J. Ai, "A modular attachment mechanism for software network evolution," *Phys. Stat. Mech. Its Appl.*, vol. 392, no. 9, pp. 2025–2037, May 2013.

[8] G. Canfora, M. Di Penta, and L. Cerulo, "Achievements and challenges in software reverse engineering," *Commun. ACM*, vol. 54, no. 4, p. 142, Apr. 2011.

[9] L. Šubelj, and M. Bajec, "Community structure of complex software systems: Analysis and applications," *Phys. Stat. Mech. Its Appl.*, vol. 390, no. 16, pp. 2968–2975, Aug. 2011.

[10] T. Chaikalis, and A. Chatzigeorgiou, "Forecasting Java Software Evolution Trends Employing Network Models," *IEEE Trans. Softw. Eng.*, vol. 41, no. 6, pp. 582–602, Jun. 2015.

[11] G. Xiao, Z. Zheng, and H. Wang, "Evolution of Linux operating system network," *Phys. Stat. Mech. Its Appl.*, vol. 466, pp. 249–258, Jan. 2017.

[12] T. Hellmann, and M. Staudigl, "Evolution of social networks," *Eur. J. Oper. Res.*, vol. 234, no. 3, pp. 583–596, May 2014.

[13] H. Shariat Yazdi, L. Angelis, T. Kehrer, and U. Kelter, "A framework for capturing, statistically modeling and analyzing the evolution of software models," *J. Syst. Softw.*, vol. 118, pp. 176–207, Aug. 2016.

[14] T. Chaikalis, and A. Chatzigeorgiou, "Forecasting Java Software Evolution Trends Employing Network Models," *IEEE Trans. Softw. Eng.*, vol. 41, no. 6, pp. 582–602, Jun. 2015.

[15] W. Pan, B. Li, Y. Ma, and J. Liu, "A Novel Software Evolution Model Based on Software Networks," in *Complex Sciences*, vol. 5, J. Zhou, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1281–1291.

[16] H. Li, L.-Y. Hao, R. Chen, X. Ge, and H. Zhao, "Symmetric Preferential Attachment for New Vertices Attaching to Software Networks," *New Gener. Comput.*, vol. 32, no. 3–4, pp. 271–296, Aug. 2014.

[17] H. Kim, C. I. Del Genio, K. E. Bassler, and Z. Toroczkai, "Constructing and sampling directed graphs with

مخازن کد نرم‌افزار و مطالعه تکامل سیستم‌های نرم‌افزاری شی‌گرا و استفاده از خصوصیات شبکه کلاسی نرم‌افزار و تغییرات مهندسی نرم‌افزار، مدلی برای پیشگویی تکامل شبکه‌های نرم‌افزاری ارائه گردید.

شبکه جهت‌دار کلاس‌های نرم‌افزار که گره‌ها متناظر با کلاس‌ها و یال‌ها نشان‌دهنده چهار ارتباط اساسی انجمنی، ارث بری، تحقق و تجمع بین آنهاست، به عنوان پایه مدل‌سازی استفاده شد. مطالعه تغییرات آماری تغییرات ساختاری نرم‌افزار (ایجاد، حذف کلاس‌ها) در نسخه‌های متوالی، الگوی تغییرات کلاس‌های نرم‌افزار را آشکار می‌کند. مدل پیشنهادی با تجمیع الگوهای تغییر اساسی بر روی شبکه کلاسی نرم‌افزار، نسخه‌های آتی نرم‌افزار را شبیه‌سازی می‌کند. به‌منظور نمایش کارایی و برتری مدل پیشنهادی، قدرت شبیه‌سازی مدل بر روی پروژه‌ی واقعی متن باز eclipse به نام EGit (۱۳ نگارش مختلف) و در مقایسه با مدل اتصال پیمان‌های ارزیابی شد. نتایج نشان‌دهنده توزیع درجه توانی زیر شبکه‌های اضافه شده (مجموعه کلاس‌های تازه اضافه شده) و درجات ورودی و توزیع نرمال درجات خروجی گره‌های حذف شده است.

شبکه تولید شده توسط مدل پیشنهادی نسبت به مدل رشد اتصال پیمان‌های توزیع درجه شباهت بیشتری با شبکه واقعی نرم‌افزار داشته و ضریب همبستگی بین درجه‌های منفی و مثبت در شبکه شبیه‌سازی شده توسط مدل پیشنهادی به مدل شبکه واقعی نزدیک‌تر است. شبکه تولید شده توسط مدل پیشنهادی از لحاظ معیارهای جهان کوچک (میانگین طول کوتاهترین مسیر و خوشه‌بندی) دارای ارزش مقادیر نزدیک‌تری به شبکه‌های واقعی در مقایسه با شبکه شبیه‌سازی شده مدل اتصال پیمان‌هایی است. ضریب پیمان‌های و اندازه‌های انجمن‌های شبکه مدل پیشنهادی در مقایسه با شبکه اتصال ترجیحی تطابق نزدیک‌تری به شبکه نرم‌افزار واقعی دارند. اندازه شبکه تولید شده توسط مدل پیشنهادی نسبت به شبکه ایجاد شده توسط اتصال ترجیحی تطابق بهتری با شبکه واقعی نرم‌افزار دارد.

غنی کردن ارتباطات شبکه نرم‌افزاری مورد استفاده با بهره‌گیری از سایر ارتباطات بین کلاسی بصورت مستقیم در بهبود دقت تحلیل رفتار شبکه‌های نرم‌افزاری تأثیرگذار است. مدل پیشنهادی از چهار ارتباط اساسی بین کلاس‌های نرم‌افزار برای نمایش ارتباطات گره‌ها و تحلیل شبکه نرم‌افزار استفاده کرده است. اضافه کردن سایر ارتباطات بین کلاس‌های نرم‌افزار (برای نمونه چندگانگی<sup>۲۴</sup> و ترکیب‌بندی<sup>۲۵</sup>) به نمایش واقعی‌تر روابط بین گره‌های شبکه نرم‌افزار کمک می‌کند. در مدل‌سازی الگوی دو دسته تغییرات سیستم‌های نرم‌افزاری در شبکه کلاسی توسط مدل پیشنهادی میزان تأثیر صفات گوناگون برای نمونه توزیع درجات و سن بصورت خطی در نظر گرفته شده است. سنجش میزان تأثیر و اهمیت هر یک از این صفات در مدل‌سازی تغییرات با استفاده از مدل‌های مارکوف<sup>۲۶</sup> و بازخوردی<sup>۲۷</sup> در افزایش دقت مدل‌سازی تغییرات تأثیرگذار خواهد بود.

## مراجع

[1] T. Mens, and S. Demeyer, *Software Evolution*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[2] S. W. Thomas, B. Adams, A. E. Hassan, and D. Blostein, "Studying software evolution using topic models," *Sci. Comput. Program*, vol. 80, pp. 457–479, Feb. 2014.

[3] J. Liebig, A. Von Rhein, C. Kastner, S. Apel, J. Dorre, and C. Lengauer, "Scalable analysis of variable software," in *Proceedings of the 2013 9<sup>th</sup> Joint Meeting on Foundations of Software Engineering*. ACM, 2013, pp. 81–91.

**اطلاعات بررسی مقاله:**

تاریخ ارسال: ۱۳۹۶/۰۶/۱۳

تاریخ اصلاح: ۱۳۹۶/۰۷/۱۰

تاریخ قبول شدن: ۱۳۹۷/۰۱/۲۲

نویسنده مرتبط: دکتر محمد خوانساری، دانشکده علوم و فنون نوین، دانشگاه تهران، تهران، ایران.

given degree sequences," *New J. Phys.*, vol. 14, no. 2, p. 023012, Feb. 2012.

[18] Y. Jiang, Y. Huang, P. Li, S. Gao, Y. Zhang, and Y. Yan, "How to detect communities in large networks," in *International Conference on Intelligent Computing*. Springer, 2015, pp. 76–84.

**مجتبی صادقیان** مدرک کارشناسی و کارشناسی ارشد خود را به ترتیب از دانشگاه تبریز در سال ۹۳ و دانشگاه تهران در سال ۹۵ دریافت نموده است. پایان‌نامه ایشان در مورد تحلیل و مدل‌سازی تکامل شبکه‌های نرم‌افزاری متن‌باز شی‌گرا با استفاده از مفاهیم شبکه‌های پیچیده و مهندسی



نرم‌افزار است.

آدرس پست‌الکترونیکی ایشان عبارت است از:

[m.sadeghian1991@ut.ac.ir](mailto:m.sadeghian1991@ut.ac.ir)

**محمد خوانساری** مدرک کارشناسی، کارشناسی ارشد و دکتری خود را از دانشگاه صنعتی شریف در مهندسی کامپیوتر به ترتیب در سال‌های ۱۳۷۵، ۱۳۷۷ و ۱۳۸۷ اخذ کرد. او یکی از مشارکت‌کنندگان اصلی راه‌اندازی مرکز پژوهشی فناوری اطلاعات و ارتباطات پیشرفته دانشگاه صنعتی شریف (AICT) و دارای یک بورس تحقیقاتی از موسسه DAAD آلمان است. از سوابق اجرایی - پژوهشی ایشان می‌توان به مدیریت طرح ملی نرم‌افزارهای آزاد/متن‌باز (لینوکس فارسی)، رئیس پژوهشکده فناوری اطلاعات و رئیس مرکز تحقیقات مخابرات ایران (پژوهشگاه ارتباطات و فناوری اطلاعات)، عضو هیأت علمی پردیس بین‌الملل دانشگاه صنعتی شریف (کیش) و رئیس مرکز فناوری اطلاعات و فضای مجازی دانشگاه تهران اشاره کرد. ایشان مؤسس گرایش بین‌رشته‌ای مهندسی سامانه‌های شبکه‌ای در کارشناسی ارشد مهندسی فناوری اطلاعات می‌باشند. از شهریور ۱۳۹۰ تاکنون، ایشان عضو هیئت علمی دانشکده علوم و فنون نوین دانشگاه تهران هستند. زمینه‌های پژوهشی ایشان عبارتند از: علوم شبکه و شبکه‌های پیچیده، شبکه‌های سنسور بی‌سیم چندرسانه‌ای و سلامت، نرم‌افزارهای آزاد/متن‌باز، سیستم‌های اطلاعات سلامت.



آدرس پست‌الکترونیکی ایشان عبارت است از:

[m.khansari@ut.ac.ir](mailto:m.khansari@ut.ac.ir)

**فرید دهقان** دانشجوی دکتری رشته مهندسی فناوری اطلاعات است و در زمینه تحلیل تعامل در توسعه نرم‌افزار بر بستر گیت هاب کار می‌کند. به تحلیل شبکه‌های اجتماعی تعاملی و داده‌های بزرگ و نیز تحلیل گراف‌های احتمالاتی و داده‌کاوی علاقه دارد.



آدرس پست‌الکترونیکی ایشان عبارت است از:

[farid.dehghan@ut.ac.ir](mailto:farid.dehghan@ut.ac.ir)

- <sup>1</sup>Maintenance
- <sup>2</sup>Number of Code Lines
- <sup>3</sup>Mining Software Code Repositories
- <sup>4</sup>Preferential Attachment
- <sup>5</sup>Linux Operating System (LOS)
- <sup>6</sup>Attribute
- <sup>7</sup>Association
- <sup>8</sup>Generalization
- <sup>9</sup>Realization
- <sup>10</sup>Aggregation
- <sup>11</sup><https://github.com/eclipse/egit>
- <sup>12</sup>Source Code Management
- <sup>13</sup>Scale-Free
- <sup>14</sup>God
- <sup>15</sup>Objects
- <sup>16</sup>Correlation
- <sup>17</sup>Community
- <sup>18</sup>Lego
- <sup>19</sup>Coupling
- <sup>20</sup>Modularity
- <sup>21</sup>Louvain
- <sup>22</sup>Clustering Coefficient
- <sup>23</sup>Average Shortest Path Lengths
- <sup>24</sup>Multiplicity
- <sup>25</sup>Composition
- <sup>26</sup>Markov Model
- <sup>27</sup>Feedback Model



## سیستم پیشنهاددهنده قانون محور تکاملی

بهزاد سلیمانی نیسیانی<sup>۱و۲</sup> نسیم سلطانی<sup>۳</sup> رضا مفیدی<sup>۳</sup> محمد حسین ندیمی شهرکی<sup>۴</sup>

<sup>۱</sup> دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد اصفهان (خوراسگان)، اصفهان، ایران

<sup>۲</sup> دانشکده برق و کامپیوتر، دانشگاه کاشان، کاشان، ایران

<sup>۳</sup> دانشکده فنی و مهندسی، موسسه آموزش عالی علامه نائینی، نائین، ایران

<sup>۴</sup> دانشکده کامپیوتر، دانشگاه آزاد اسلامی واحد نجف‌آباد، نجف‌آباد، ایران

### چکیده

سیستم‌های پیشنهاددهنده برای ارائه محصولات به مشتریان بالقوه طراحی شده و همین‌طور فیلتر مشارکتی روشی رایج در سیستم‌های پیشنهاددهنده است. این سیستم با شناسایی کاربران مشابه براساس زمان ورود و انجام تراکنش‌های قبلی، پیشنهادهایی را تولید خواهد نمود. دقت پایین پیشنهادها به دلیل تنگی یکی از نگرانی‌های عمده مربوط به روش‌های پیشنهاددهنده فیلتر مشارکتی است، برای رفع این مشکل، محققان متعددی استفاده از استخراج قوانین انجمنی را ارائه داده‌اند. تولید قوانین انجمنی با سرعت بالاتر و حفظ کیفیت، می‌تواند باعث ارائه پیشنهادهایی با کیفیت و سرعت بالاتر گردد. در این راستا، این مقاله سیستم پیشنهاددهنده مبتنی بر قوانین انجمنی تولید شده با استفاده از الگوریتم ژنتیک را ارائه می‌دهد که نیاز به قوانین با رأس ثابت دارد و در نهایت قوانینی با رأس ثابت را تولید می‌کند. آزمایش‌های انجام شده روی مجموعه داده موبیلنز، سرعت بالاتر تولید قوانین را نسبت به کارهای دیگر نشان می‌دهد و در مقایسه با الگوریتم ازدحام ذرات سرعت همگرایی بهتری در عین حفظ کیفیت داشته است.

**کلمات کلیدی:** سیستم پیشنهاددهنده، فیلتر مشارکتی، استخراج قوانین انجمنی، بهینه‌سازی ازدحام ذرات، بهینه‌سازی چند هدفه.

### ۱- مقدمه

۱-۲ مبتنی بر دانش<sup>۲</sup> [۵]: دانش در مورد کاربران، اقلام، چگونگی یک قلم خاص و نیازهای یک کاربر خاص را ثبت می‌کنند. پیشنهادات براساس این دانش صورت می‌گیرد.

۱-۳ فیلتر مشارکتی<sup>۴</sup> (CF) [۲، ۶]: از اطلاعات کاربران مشابه جهت پیشنهاد استفاده می‌کند. سیستم‌های پیشنهاددهنده مشارکتی<sup>۵</sup> اطلاعات متناسب با ترجیحات کاربر جمع‌آوری می‌نمایند. در فیلتر مشارکتی کاربران با ترجیحات مشابه در سایر موارد هم می‌توانند با هم شبیه باشند.

روش فیلتر مشارکتی دارای چالش‌های بسیاری است. از جمله برای مقابله با تنگی<sup>۶</sup> در پایگاه داده به منظور بهبود مقیاس‌پذیری و افزایش بهره‌وری پیشنهادات. تلاش‌های زیادی برای غلبه بر مشکلات فیلتر مشارکتی و در نتیجه تولید پیشنهادات با کیفیت شده است. این روش‌ها را می‌توان به این صورت دسته‌بندی نمود:

موفقیت کسب و کار آنلاین بستگی به توانایی ارائه محصولات شخصی، خدمات و اقلام<sup>۱</sup> اطلاعاتی به مشتریان بالقوه، دارد. این زمینه باعث تمایل به سیستم‌های پیشنهاددهنده شده است. این سیستم‌ها با اعمال روش‌های آماری و کشف دانش، خدماتی را به مشتریان ارائه می‌دهد [۱-۳]. سه نوع عمده از سیستم‌های پیشنهاددهنده عبارتند از:

۱-۱ مبتنی بر محتوا<sup>۲</sup> [۴]: محتوای اقلام را تحلیل نموده و سعی دارد نظم در میان آنها را درک نماید. به یک کاربر پیشنهادهایی خواهد شد در مورد شباهت زیاد آن اقلام به آن‌هایی که کاربر در گذشته ترجیح داده است.

جست‌وجوی سراسری ژنتیک و قانون ارتباط ترکیب شده تا اطلاعات مورد نیاز دریافت شوند.

۵. الگوریتم کشف قوانین انجمنی مبتنی بر بهینه‌سازی تجمعی ذرات چند هدفه (MOPSO<sup>۱۲</sup>) [۱۱]، که در این الگوریتم سعی شده تا ایرادات الگوریتم ASARM را برطرف نماید. این الگوریتم با ترکیب استخراج قوانین انجمنی (ARM) با الگوریتم بهینه‌سازی تجمعی ذرات چند هدفه (MOPSO) در چارچوب فیلتر مشارکتی به بهبود دقت و صحت پیشنهادات کمک نموده است.

در این مقاله شیوه به کار گرفته شده در الگوریتم MOPSO الگو برداری شده و با الگوریتم ژنتیک جایگزین شده است. سرعت همگرایی بالاتر الگوریتم ژنتیک نسبت به الگوریتم بهینه‌سازی ذرات تجمعی (PSO) می‌تواند سرعت تولید قوانین و پیشنهادها را بالا ببرد. بخش ادبیات تحقیق توضیحاتی در مورد استخراج قوانین انجمنی، الگوریتم ژنتیک و بهینه‌سازی چند هدفه آورده شده است. بخش سوم الگوریتم استخراج قوانین انجمنی مبتنی بر ژنتیک (GARM) ارائه شده است. بخش چهارم استراتژی پیشنهاد شرح داده شده است. بخش ۵ الگوریتم GARM با کارهای مشابه مقایسه شده و بخش پایانی در مورد کارهای آینده صحبت شده است.

## ۲- پیش‌زمینه

در این بخش، زمینه‌های کلی استخراج قوانین انجمنی و الگوریتم ژنتیک ارائه شده است. در قسمت سوم بهینه‌سازی چند هدفه به صورت خلاصه شرح داده شده است.

### ۲-۱- استخراج قوانین انجمنی (ARM)

استخراج قوانین انجمنی برای اولین بار توسط آگراوال<sup>۱۴</sup> معرفی شده است و یکی از مهم‌ترین تکنیک‌های داده‌کاوی است [۱۸]. هدف آن کشف قوانین انجمنی مطلوب از پایگاه داده‌های تراکنش تجاری است. بنابراین، ARM می‌تواند در تصمیم‌گیری کسب و کار به تصمیم‌گیرندگان در جهت خرید محصولات کمک کند.

از دیدگاه کلی کاوش قوانین انجمنی را می‌توان یک فرایند دو گامی در نظر گرفت. گام اول یافتن کلیه مجموعه اقلام مکرر و گام دوم تولید قوانین انجمنی قوی با کمک مجموعه اقلام مکرر. به دلیل هزینه بالای محاسباتی در گام اول، به طور معمول الگوریتم‌ها بر روی بهینه‌سازی عملیات این مرحله متمرکز می‌شوند، چون کارایی الگوریتم با توجه به پیچیدگی این مرحله سنجیده می‌شود.

قوانین به صورت  $A \rightarrow B$  بوده که A دنباله<sup>۱۵</sup> قانون و B راس<sup>۱۶</sup> قانون می‌باشد. به این معناست که اگر A اتفاق بیفتد به احتمال زیاد B هم اتفاق خواهد افتاد. معیار قوی بودن قوانین انجمنی که در گام دوم مورد بررسی قرار می‌گیرد، براساس دو پارامتر ضریب حمایت<sup>۱۷</sup> و ضریب اطمینان<sup>۱۸</sup> سنجیده خواهد شد. این دو پارامتر به صورت زیر محاسبه می‌شود:

$$\text{sup}(A \rightarrow B) = \frac{\text{تعداد تراکنش‌های شامل A و B}}{\text{تعداد تراکنش‌های مجموعه داده}} \quad (۱)$$

$$\text{Conf}(A \rightarrow B) = \frac{\text{تعداد تراکنش‌های شامل A و B}}{\text{تعداد تراکنش‌های شامل A}} \quad (۲)$$

مجموعه داده شامل m تراکنش روی n قلم داده می‌باشد. A و B مجموعه اقلام هستند، به طوری که  $A \cap B = \emptyset$  همواره صادق است. قوانینی که ضریب

• روش‌های سنتی مبتنی بر همسایگی [۷]: سیستم‌های پیشنهاددهنده فیلتر مشارکتی مبتنی بر قلم پیش‌بینی را از طریق سابقه همسایگان اقلام محاسبه می‌کنند و انتخاب درست نزدیک‌ترین همسایگان که بیشترین شباهت را با قلم مورد نظر داشته باشند تاثیر زیادی بر افزایش دقت پیش‌بینی‌ها خواهد داشت. برای یافتن نزدیک‌ترین همسایگان باید شباهت بین تمام اقلام موجود در سیستم محاسبه شود.

• روش خوشه‌بندی [۸]: الگوریتم‌های خوشه‌بندی می‌توانند با خوشه‌بندی اقلام فضای جستجو را کاهش دهند و مقیاس‌پذیری سیستم را افزایش دهند. در نتیجه با محاسبه شباهت بین اعضای که در یک خوشه قرار دارند، پیش‌بینی‌ها تولید خواهند شد.

• استدلال مبتنی بر مورد [۹]: استدلال مبتنی بر مورد، یک روش تصمیم‌گیری، مبتنی بر راه حل مسائل مشابه است. این روش توانایی یادگیری قدرتمندی دارد. می‌تواند از تجربیات گذشته، برای برخورد با مسائل جدید، استفاده کند. در وضعیت‌های متغیر و پویا، زمانی که شرایط ناشناخته و غیرواضح هستند، این روش استدلال مناسبی است.

• الگوریتم‌های ترکیبی [۱۰]: ترکیبی از CF با تکنیک‌های مبتنی بر حافظه و CF با تکنیک‌های مبتنی بر محتوا

• کاوش قوانین انجمنی (ARM)<sup>۱۹</sup> [۱۱، ۱۲]: برای تشخیص وابستگی‌های مهم میان اقلام موجود در یک پایگاه داده تراکنشی به کار می‌رود، به نحوی که حضور بعضی اقلام در تراکنش‌ها بر حضور برخی اقلام دیگر در همان تراکنش‌ها دلالت داشته باشد.

تنکی اثر بخشی روش‌های فیلتر مشارکتی را محدود می‌کند. هنگامی که تراکنش داده‌ها نادر است، این سیستم در شناسایی شباهت‌ها در منافع مصرف کننده ناکارآمد است. اساساً، مشکل تنکی را می‌توان در دو بعد تجزیه و تحلیل نمود. CF مبتنی بر کاربر و CF مبتنی بر قلم. به دلیل تنکی در مجموعه داده، دو کاربر که علائق مشابه دارند اما قلم مشترکی را رای نداده‌اند، علائق دو کاربر مشابه در نظر گرفته نمی‌شود. در این شرایط، CF مبتنی بر کاربر پیشنهادها با کیفیت پایین تولید می‌کند. مشابه این وضعیت برای سیستم‌های CF مبتنی بر قلم رخ می‌دهد، زمانی که دو قلم مشابه هستند ولی چون توسط یک کاربر امتیاز مشابه دریافت نکرده‌اند مشابه در نظر گرفته نمی‌شوند.

هانگ<sup>۸</sup> و همکاران یک روش برای برخورد با مشکل تنکی از طریق بررسی قوانین انجمنی در میان مصرف‌کنندگان از طریق تراکنش‌های گذشته و بازخورد، ارائه داده‌اند [۱۳]. بعدها، لئونگ<sup>۹</sup> و همکاران با معرفی یک چارچوب CF براساس قوانین انجمنی فازی و در مشابهت چند سطحی (FARAMS)<sup>۱</sup> به مشکل تنکی پرداخته‌اند [۱۴].

از جمله راه حل‌هایی که از طریق کشف قوانین انجمنی ارائه شده عبارتند از:

۱. الگوریتم کاوش قوانین انجمنی با ضریب حمایت و فقی (ASARM)<sup>۱۱</sup> [۱۲]، که در این الگوریتم تولید قانون با الگوریتم CBA-RG انجام شده که تکامل یافته الگوریتم اپریوری<sup>۱۲</sup> می‌باشد.

۲. الگوریتم کاوش قوانین انجمنی با استفاده از الگوریتم ژنتیک که از داده‌کاوی برای کشف الگوهای خرید مکرر مشتریان استفاده کرده است [۱۵]. در ابتدا مشتریان برگزیده و در نهایت قوانین انجمنی با استفاده از آنها اکتشاف می‌شود.

۳. الگوریتم اکتشاف قوانین انجمنی با استفاده از الگوریتم ژنتیک برای یافتن قوانین ارتباطی است که در این مقاله فرد به عنوان مجموعه‌ای از قوانین ارائه شده است. الگوریتم پیشنهادی یک جایگزین برای یافتن قوانین با پیش‌بینی دقیق و دشوار، قابل فهم و جالب است [۱۶].

۴. کاوش قوانین انجمنی با استفاده از ژنتیک و قانون ارتباط، که برای یافتن الگوهای مکرر استفاده شده است [۱۷]. برای کاهش پیچیدگی زمانی و

### ۳- الگوریتم استخراج قوانین انجمنی مبتنی بر ژنتیک<sup>۲۲</sup> (GARM)

قوانین انجمنی تولید شده برای سیستم پیشنهاددهنده فیلتر مشارکتی به صورت قوانین با رأس ثابت است، بدین معنی که اگر کاربری به راس ثابت  $f$  وابستگی داشته باشد سیستم پیشنهاددهنده باید بتواند تمام اقلام مرتب با  $f$  را بیابد و به مشتری پیشنهاد دهد.

به عنوان مثال چهار رأس  $a, b, c, d$  به رأس  $f$  منجر می‌شوند و اگر کاربر به رأس  $f$  وابستگی داشته سیستم پیشنهاددهنده باید بتواند سایر رأس‌هایی که منجر به  $f$  می‌شوند مثل  $a, b, c, d$  را به کاربر پیشنهاد دهند.

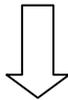
رأس ثابت می‌تواند قلم و یا کاربری باشد که فیلتر مشارکتی برای آن پیشنهاد تولید خواهد کرد. تا پایان مقاله به جای قلم/کاربر هدف که قرار است برای آنها پیشنهاد تولید شود از واژه شی هدف به جای هر دو استفاده شده است.

#### ۳-۱- طرح نمایش<sup>۲۳</sup>

قانون انجمنی  $A \rightarrow B$  دو مجموعه قلم غیرمتصل  $A$  و  $B$  است. در این مقاله سیستم پیشنهاددهنده، قوانین با رأس ثابت نیاز دارد، از این رو الگوریتم استخراج قوانین انجمنی مبتنی بر ژنتیک (GARM) به استخراج قوانین تنها برای یک شی هدف (کاربر/قلم) می‌پردازد.

برای کاربرد این الگوریتم، هر قانون به عنوان یک کروموزوم ارائه می‌گردد. از آنجا که رأس قانون برای یک شی هدف خاص بوده و ثابت است، روی دنباله قانون عملیات انجام می‌گردد. بنابراین هر یک از کروموزوم‌ها در فضای جستجو توسط  $(A = \{A_1, A_2, \dots, A_n\})$  نشان داده می‌شود. به عنوان مثال قانون  $I_5 \rightarrow I_8$  به صورت  $5 \rightarrow 8$  در نظر گرفته شده است. چون رأس قانون ( $I_8$ ) ثابت است، عملیات ژنتیک روی دنباله قانون ( $I_5$ ) انجام می‌گیرد. عدد ۵ به صورت دودویی به صورت (۰۱۰۱) خواهد بود. هر کدام از بیت‌های صفر و یک تشکیل ژن‌های کروموزوم را می‌دهند.

شماره کاربرها	شماره اقلام			
	$I_1$	$I_2$	$I_3$	$I_4$
User 1	3	0	1	2
User 2	0	1	4	0
User 3	0	0	0	2
User 4	1	5	2	4
User 5	4	0	1	2



شماره کاربرها	شماره اقلام			
	$I_1$	$I_2$	$I_3$	$I_4$
User 1	1	0	0	1
User 2	0	0	1	0
User 3	0	0	0	1
User 4	0	1	0	1
User 5	1	0	0	0

شکل ۱- طرح تبدیل

حمایت و ضریب اطمینان آنها بالاتر از حداقل ضریب حمایت و حداقل ضریب اطمینان تعیین شده توسط مساله باشد، قانون قوی نامیده می‌شوند.

الگوریتم اپریوری برای کشف قوانین انجمنی استفاده می‌شود که توسط آگراول و همکاران طراحی شده است. این الگوریتم در نظر دارد برای پیدا کردن مجموعه اقلام مکرر از حداقل ضریب حمایت استفاده نموده و پس از آن قوانین انجمنی با کیفیت را با استفاده از حداقل ضریب اطمینان معین کند. این الگوریتم به عنوان پایه‌ای برای تجزیه و تحلیل سبد بازار به وسیله پیدا کردن ارتباط بین اقلام مختلف در سبد خرید، استفاده می‌شود.

#### ۲-۲- الگوریتم ژنتیک

الگوریتم ژنتیک<sup>۱۹</sup> راه حل مناسبی برای جایگزینی روش‌های بهینه‌سازی سنتی، با استفاده از جستجوهای تصادفی در محیط‌های پیچیده است [۱۹، ۲۰]. الگوریتم ژنتیک نوع خاصی از الگوریتم‌های تکامل است که از تکنیک‌های زیست‌شناسی برگشتی مانند وراثت و جهش استفاده می‌کند، بدین صورت که هر نسل با پیشرفت دائم خود باعث بهبود و تکامل نسل بعدی می‌گردد که در الگوریتم ژنتیک با استفاده از جهش و وراثت اینکار صورت می‌گیرد. این الگوریتم برای اولین بار توسط جان هلند معرفی شد.

ایده اساسی این الگوریتم انتقال خصوصیات موروثی توسط ژن‌هاست. مجموعه خصوصیات انسان توسط کروموزوم‌های او به نسل بعدی منتقل می‌شوند [۲۱، ۲۲]. هر ژن در این کروموزوم‌ها نماینده یک خصوصیت است. به عنوان مثال ژن ۱ می‌تواند رنگ چشم باشد، ژن ۲ طول قد، ژن ۳ رنگ مو و الی آخر. کروموزوم‌ها توسط عملیات ترکیب<sup>۲۰</sup> و جهش<sup>۲۱</sup> باز تولید شده و نسل جدید را بوجود می‌آورند.

#### ۲-۳- بهینه‌سازی چند هدفه

الگوریتم‌های بهینه‌سازی چند هدفه به صورت بهینه‌سازی  $\{f_1(p), f_2(p), \dots, f_n(p)\}$  طراحی شده که  $p \in P$  (منطقه امکان‌پذیر) بوده و در آن  $f_1, f_2, \dots, f_n$  توابع هدف مختلف هستند. این توابع به طور همزمان بهینه‌سازی می‌شوند. اغلب توابع هدف متضاد هستند، بنابراین هیچ راه‌حل واحد به طور همزمان نمی‌تواند به بهینه‌سازی تمام توابع منجر شود. مجموعه‌ای از راه‌حل‌ها برای مشکل بهینه‌سازی چند هدفه وجود دارد [۲۳، ۲۴].

در اینجا مقادیر ضریب اطمینان و ضریب حمایت به عنوان دو تابع هدف برای جستجو راه‌حل بهینه (قوانین) به کار گرفته شده است. یک قانون بهینه گفته می‌شود اگر دارای کیفیت بالا باشد. کیفیت قوانین با در نظر گرفتن متوسط وزنی مقادیر ضریب اطمینان و ضریب حمایت قانون ارزیابی می‌گردد. بنابراین، برای یک قانون انجمنی  $A \rightarrow B$ ، کیفیت قانون به صورت زیر تعریف می‌گردد:

$$\text{quality}(A \rightarrow B) = W_1 \text{sup}(A \rightarrow B) + W_2 \text{conf}(A \rightarrow B) \quad (3)$$

در این رابطه،  $W_1$  و  $W_2$  وزن اختصاص یافته به مقادیر ضریب اطمینان و ضریب حمایت بوده به طوری که  $W_1 + W_2 = 1$  و  $W_1, W_2 \in [0, 1]$  است.  $W_1$  با ارزش بیشتر و  $W_2$  با ارزش کمتر نشان می‌دهد که کیفیت قانون بیشتر بستگی به مقدار ضریب حمایت قانون دارد. در مقابل،  $W_1$  با ارزش کمتر و  $W_2$  با ارزش بیشتر نشان می‌دهد که کیفیت قانون بیشتر بستگی به مقدار ضریب اطمینان قانون دارد. این وزن‌ها به صورت تجربی انتخاب می‌گردند [۱۱].

### ۳-۲- طرح تبدیل<sup>۲۴</sup>

- گام سوم: قوانین استخراج شده با کیفیت بالا در هر نسل به همراه کیفیت آنها در یک پایگاه داده ذخیره می‌شود.  
 - گام چهارم: از پایگاه داده،  $k$  قانون برتر از نظر کیفیت، انتخاب شده و به عنوان ورودی به سیستم پیشنهاددهنده ارسال می‌شوند.  
 همانطور که گفته شد، این سیستم پیشنهاددهنده قوانین با رأس ثابت را نیاز دارد. به همین علت الگوریتم فوق قوانین با رأس شی هدف را استخراج می‌کند. هنگام اعمال ترکیب و جهش نیز تنها دنباله قانون تغییر می‌کند.

به منظور استخراج قوانین انجمنی از مجموعه داده‌های تراکنشی، طرح انتقال مورد نیاز است که می‌تواند فرآیند مرور داده‌ها را سرعت بخشد [۲۵]. در شکل ۱ در جدول بالایی، پنج کاربر را نشان می‌دهد که به چهار قلم رای داده‌اند. آرا به صورت یک تا پنج داده شده است. در جدول پایینی طرح تبدیل یافته نمایش داده شده است. برای تبدیل، معدل رای‌های داده شده برای هر کاربر محاسبه شده و آرا کمتر از معدل صفر و بالاتر از معدل یک در نظر گرفته می‌شود.

### ۴- استراتژی پیشنهاد

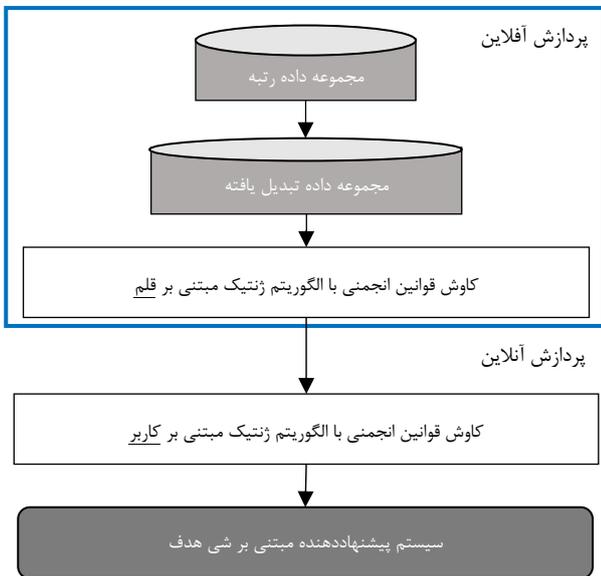
### ۳-۳- طرح الگوریتم GARM

سیستم پیشنهاددهنده براساس قوانین انجمنی استخراج شده، پیشنهاد را تولید می‌کند. دو گونه قوانین انجمنی در بخش قبل گفته شد. قوانین بر پایه قلم به صورت آفلاین و قوانین بر پایه کاربر به صورت آنلاین، هر دو برای دقیق‌تر شدن پیشنهادها به کار گرفته می‌شوند. مدل پیشنهاد شده این مقاله در شکل ۳ نشان داده شده است.

با توجه به طرح تبدیل گفته شده مجموعه داده مورد استفاده، به صورت تعدادی کاربر بوده که به اقلام رای داده‌اند. پس از تبدیل مجموعه داده به صورت پسند<sup>۲۵</sup> و یا ناپسند<sup>۲۶</sup> (صفر و یک) برای هر کاربر در می‌آید. سیستم پیشنهاددهنده برای چنین مجموعه داده‌ای دو حالت برای پیشنهاد دارد.

۱. پیشنهادها را براساس شباهت میان اقلام تولید می‌کند که به صورت آفلاین صورت می‌پذیرد. در این حالت قوانین انجمنی بر پایه قلم نیاز دارد. (مانند  $I_5 \rightarrow I_8$ )  
 ۲. پیشنهادها را براساس شباهت میان کاربرها تولید می‌کند که به صورت آنلاین صورت می‌پذیرد. برای این حالت قوانین انجمنی بر پایه کاربر نیاز دارد. (مانند  $U_7 \rightarrow U_8$ )

همان‌طور که در طرح نمایش گفته شد. کروموزم‌ها به صورت عدد دودویی شی هدف خواهد بود. اگر کاربر ۱ در شکل ۱، کاربر هدف باشد. اقلام  $I_1$  و  $I_4$  قلم هدف در نظر گرفته می‌شود. در این شرایط الگوریتم قوانین با رأس  $I_1$  و  $I_4$  را تولید می‌کند. سیستم پیشنهاددهنده براساس این قوانین برای کاربر ۱ پیشنهاد تولید می‌کند. شبه کد این الگوریتم در شکل ۲ نشان داده شده است که در آن مجموعه داده تبدیل شده (دودویی)  $D_t$  و شیء هدف  $t$  به عنوان ورودی دریافت شده و  $k$  مورد برتر قوانین انجمنی مستقیم  $R_t$  به عنوان خروجی بر می‌گرداند.



شکل ۳- مدل سیستم پیشنهاددهنده

#### GARM Algorithm:

**Input:** Trans Formed Dataset  $D_t$ , Target Object  $t$

**Output:**  $R_t \leftarrow$  Top-k Direct Association Rule in the Archive  $R_t$

**Step 1:** Initial Population Randomly and set  $R_t$

**Step 2:** for  $g=1$  to  $q // q$ : Maximum Generation Length

Select Some Rules in  $R_t$

Do Cross Over and Mutation

Evaluate Fitness by quality  $(A \rightarrow B) =$

$$W_1 \sup(A \rightarrow B) + W_2 \text{conf}(A \rightarrow B)$$

Update  $R_t$

End for

**Step 3:**  $R_t \leftarrow$  Find Top-k rules in  $R_t$

**Step 4:** send Top-k Rules to Recommendation System

شکل ۲- شبه کد الگوریتم GRAM

- سیستم پیشنهاددهنده بر پایه کاربر: ساختار قوانینی که برای سیستم پیشنهاددهنده بر پایه کاربر استفاده می‌شود به صورت  $U_x \rightarrow U_t$  است. برای این قانون، اقلامی که کاربر  $U_t$  در آینده خواهد پسندید، در گذشته  $U_x$  پسندیده است. در این شرایط اقلامی که  $U_x$  پسندیده به کاربر  $U_t$  پیشنهاد می‌گردد [۱۱].
- سیستم پیشنهاددهنده بر پایه قلم: ساختار قوانینی که برای سیستم پیشنهاددهنده بر پایه قلم استفاده می‌شود به صورت  $I_x \rightarrow I_t$  است. برای این قانون، کاربری که  $I_t$  را پسندیده در آینده  $I_x$  را خواهد پسندید. در این شرایط  $I_x$  پیشنهاد می‌گردد [۱۱].

### ۵- ارزیابی و تفسیر نتایج

در این قسمت سه آزمایش انجام شده است. آزمایش روی الگوریتم GARM انجام شده تا بهبود عملکرد سرعت این الگوریتم را نشان دهد. الگوریتم پیشنهادی با الگوریتم MOPSO-ARM مقایسه شده است [۱۱]. آزمایش‌ها روی مجموعه داده موویلنز 100k انجام شده است. در این بخش ابتدا مجموعه داده مورد استفاده شرح داده شده

- گام اول: برای نسل اول تعدادی کروموزوم به صورت تصادفی تولید می‌شوند.  
 - گام دوم: به تعداد مشخص  $q$  مرتبه این اعمال را انجام می‌دهد: کروموزم‌ها به صورت دو به دو با تابع انتخاب<sup>۲۷</sup> برای عملیات برش<sup>۲۸</sup> و جهش<sup>۲۹</sup> انتخاب می‌شوند؛ خروجی این عملیات، نسل جدید است که برای هر کدام از آنها تابع ارزیاب با توجه به مجموعه داده محاسبه می‌شود و بهترین آنها جایگزین نسل قبل خواهند شد. این روال ادامه می‌یابد تا زمانی که تعداد  $q$  مرتبه خاتمه یافته یا کیفیت قوانین در نسل‌های جدید بهبودی نیابد و در واقع الگوریتم به همگرایی برسد. کیفیت قوانین با رابطه ۳ سنجیده می‌شود.

استفاده می‌شود. اینکه چه تعداد از قوانین با کیفیت در نظر گرفته شود، مورد ارزیابی قرار می‌گیرد. ضرایب  $W_1$  و  $W_2$  پارامترهای کنترلی هستند که مورد ارزیابی قرار گرفته‌اند. این پارامترهای کنترلی توسط پارامترهای recall، precision، accuracy و F1-measure ارزیابی می‌شوند.

$$\text{accuracy} = \frac{\text{تعداد اقلامی که صحیح دسته‌بندی شده‌اند}}{\text{کل اقلام دسته‌بندی شده}} \quad (۴)$$

$$\text{precision} = \frac{\text{تعداد اقلام پیشنهاد شده صحیح}}{\text{کل اقلام پیشنهاد شده}} \quad (۵)$$

$$\text{recall} = \frac{\text{تعداد اقلام پیشنهاد شده صحیح}}{\text{کل اقلام پسندیده شده بوسیله کاربر}} \quad (۶)$$

$$\text{F1\_measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (۷)$$

#### ۴-۵- تنظیمات ارزیابی

در این بخش ابتدا پارامترهای کنترلی  $W_1$  و  $W_2$  مورد بررسی قرار می‌گیرد. سپس تابع انتخاب برای الگوریتم ژنتیک بررسی می‌شود. سایر پارامترهای ارزیابی در جدول ۲ نشان داده شده است.

جدول ۲- پارامترهای ارزیابی

نوع برش	تک نقطه‌ای
نوع جهش <td>تصادفی</td>	تصادفی
$W_1$	۰.۶
$W_2$	۰.۴

#### ۵-۵- پیش آزمایش روی پارامترهای کنترلی $W_1$ ، $W_2$ و

##### تعداد قوانین با کیفیت

در این قسمت ابتدا ضرایب  $W_1$  و  $W_2$  بررسی می‌شود. سپس تعداد قوانین با کیفیت بررسی می‌گردد. پارامترهای زیر برای سنجش کیفیت پیشنهادات به کار می‌رود. در شکل ۴ نشان می‌دهد رابطه ۳ با  $W_2$  های مختلف، پیشنهادات دقت متفاوتی را خواهند داشت (با توجه به آنکه مجموع  $W_1$  و  $W_2$  برابر ۱ است). در این شرایط هنگامی که  $W_2 = 0.4$  است،  $W_1 = 0.6$  خواهد بود. روی  $W_2 = 0.4$  مقادیر هر سه پارامتر به یک توازن می‌رسند. این شکل نشان می‌دهد به ازای مقادیر مختلف  $W_2$ ، کیفیت پارامترها خیلی تفاوت نمی‌کنند و حدود ۶۰ درصد هستند. اما روی مقدار  $0.4$  تقریباً متوازن تر هستند.

در شکل ۵ دقت پیشنهادات را به ازای تعداد قانون با کیفیت نشان می‌دهد. به ازای ۲۰ قانون با کیفیت هر سه پارامتر به توازن رسیده‌اند.

در پیش آزمایش اول تعداد جمعیت اولیه برای ۴، ۱۰ و ۱۶ مورد بررسی قرار گرفته است. الگوریتم GARM برای اجرای الگوریتم در تعداد تکرارهای ۳۰، ۴۰، ۵۰ و غیره در نظر گرفته شده است تا به همگرایی برسد. همان‌گونه که در شکل ۶ و جدول ۳ نشان داده شده است. با توجه به زمان اجرای حداقل الگوریتم در تعداد

است. سپس تاثیر پارامترهای مختلف روی الگوریتم GARM بررسی شده و در قسمت آخر مقایسه انجام شده است.

#### ۵-۱- محیط ارزیابی

برای ارزیابی الگوریتم‌های GARM و MOPSO از کامپیوتر با مشخصات زیر استفاده شده است.

RAM 2G -

CPU Doul-core E5400 @ 2.70 GHz -

- ویندوز ۷

- نرم‌افزار متلب ۳۰

از نرم‌افزار متلب به عنوان ابزار ارزیابی استفاده شده است. نرم‌افزار متلب به صورت تخصصی روی مجموعه داده‌های بزرگ کار می‌کند و کدهای خاص برای انجام عملیات روی ماتریس را دارد. این نرم‌افزار برای پیاده‌سازی انواع الگوریتم‌های تکاملی به کار گرفته می‌شود. به همین علت برای ارزیابی این پژوهش، مناسب تشخیص داده شده است.

#### ۵-۲- مجموعه داده

مجموعه داده استفاده شده در آزمایش‌ها مولینز<sup>۳۱</sup> 100k است. این مجموعه داده توسط پروژه تحقیقاتی گروه لنز از دانشگاه مینسوتا جمع‌آوری شده است [۲۶]. این مجموعه داده شامل ۱۰۰ هزار رکورد بوده که ۹۴۳ کاربر به ۱۶۸۲ فیلم رای داده‌اند. آرا در بازه ۱-۵ است. مشخصات دقیق مجموعه داده در جدول ۱ آمده است.

جدول ۱- مشخصات مجموعه داده

تعداد کاربر	۹۴۳
تعداد قلم	۱۶۸۲
بازه آرا	۵ تا ۱
حداقل تعداد رای	۲۰
حداکثر تعداد رای	۷۳۷
نوع اعداد	صحیح

برای آزمایش‌ها تعداد مختلف کاربر هدف به تصادف نمونه برداری شده است. برای هر آزمایش از میان کاربرهای با تعداد رای کم (حدود ۲۰ رای)، با تعداد رای زیاد (حدود ۶۰۰ رای) و با حدود ۳۰۰ رای، به طور مساوی نمونه‌برداری شده است.

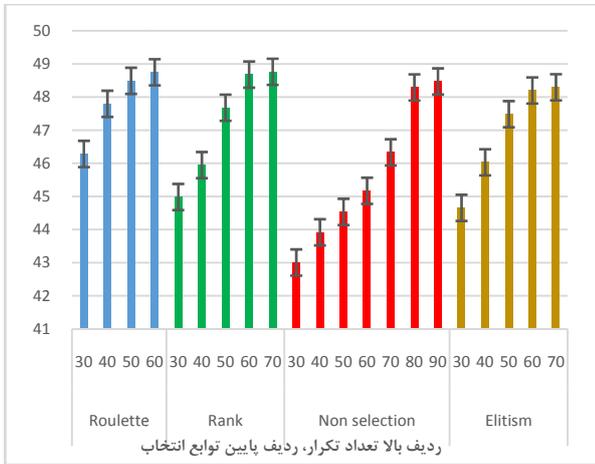
#### ۵-۳- معیار ارزیابی و بررسی اثر پارامترها

میانگین کیفیت قوانین در هر تکرار<sup>۳۲</sup>، معیار مقایسه میان پارامترهای مختلف است. زمان اجرای الگوریتم در لحظه همگرایی معیار مقایسه نهایی خواهد بود. برای بررسی اثر پارامترها سه کاربر هدف به تصادف انتخاب شده است.

کیفیت قوانین انجمنی تولید شده توسط رابطه ۳ سنجیده می‌شود. در این رابطه  $w_1 = 0.6$  و  $w_2 = 0.4$  در نظر گرفته شده است [۱۱]. تعداد جمعیت اولیه و تابع انتخاب با همین معیار سنجیده می‌شود. سیستم پیشنهاددهنده براساس قوانین قوی پیشنهاد تولید می‌کند.

پارامتر کیفیت که برای سنجش قوانین به کار می‌رود، در رابطه ۴ معرفی شده است. این رابطه هر دو پارامتر ضریب حمایت و ضریب اطمینان را برای سنجش کیفیت در نظر می‌گیرد. تعدادی از قوانین با کیفیت برای سیستم پیشنهاددهنده

می‌دهد. زمان اجرا در لحظه همگرایی برای تابع انتخاب چرخ رولت ۱۰۲ دقیقه، تابع انتخاب رتبه ۱۲۰ دقیقه، تابع انتخاب نخبه‌گرایی ۱۲۱ دقیقه و بدون تابع انتخاب ۱۶۳ دقیقه می‌باشد. با توجه به میانگین کیفیت برابر در همه‌ی توابع مختلف در نهایت از تابع انتخاب چرخ رولت با توجه به زمان اجرای حداقل، برای اجرای نهایی مورد نظر قرار خواهد گرفت.



شکل ۷- میانگین کیفیت در هر تکرار برای توابع انتخاب مختلف

#### ۵-۷- مقایسه

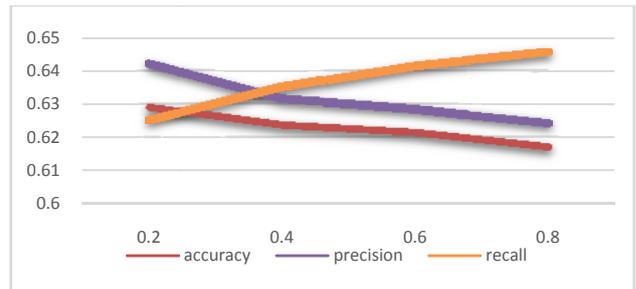
اجرای نهایی برای تعداد ۱۰۰، ۲۵۰ و ۵۰۰ قلم هدف انجام گرفته است. زمان اجرا و میانگین کیفیت الگوریتم GARM و الگوریتم MOPSO پس از همگرایی در جدول ۴ نشان داده شده است. زمان اجرای الگوریتم GARM تقریباً ۲۵ درصد نسبت به الگوریتم MOPSO بهبود یافته است. خروجی قوانین انجمنی به سیستم پیشنهاددهنده داده می‌شود. خروجی قوانین انجمنی الگوریتم MOPSO با سیستم پیشنهاددهنده آزمایش شده است و دقت قوانین استخراجی از آن ارزیابی شده که مطلوب و مورد قبول بوده است [۱۱]. سرعت بالاتر الگوریتم GARM نسبت به الگوریتم MOPSO در تولید قوانین انجمنی در کل باعث سرعت بالاتر در تولید پیشنهاد می‌گردد.

جدول ۴- مقایسه الگوریتم‌های GARM و MOPSO

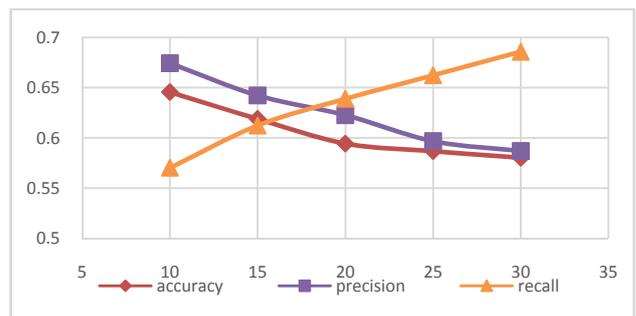
نام الگوریتم	تعداد قلم	زمان اجرا (ساعت)	میانگین کیفیت
GARM	۱۰۰	۱.۷	۴۸.۵۴
	۲۵۰	۴.۲۲	۴۸.۵۴
	۵۰۰	۸.۳۳	۴۸.۵۶
MOPSO	۱۰۰	۲.۲	۴۸.۴۸
	۲۵۰	۵.۴۷	۴۸.۵۳
	۵۰۰	۱۰.۹۷	۴۸.۵۰

الگوریتم MOPSO به دلیل سربار محاسباتی نسبت به الگوریتم GARM کندتر عمل می‌کند. علت این سربار محاسباتی، محاسبه روابط ۶ و ۷ در هر تکرار است. در حالی که الگوریتم GARM با انجام عملیات ترکیب و جهش، سربار بسیار

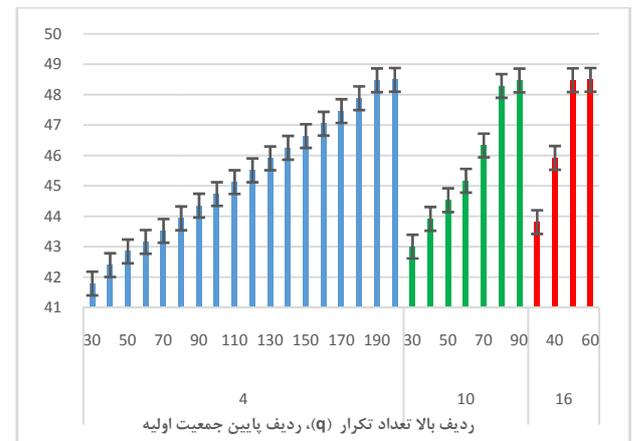
جمعیت ۱۰ نفر، برای اجرای نهایی جمعیت ۱۰ نفره مورد نظر قرار می‌گیرد.



شکل ۴- کیفیت پیشنهادها به ازای w2های مختلف (w1+w2=1)



شکل ۵- کیفیت پیشنهادها به ازای تعداد قانون مختلف (k)



شکل ۶- میانگین کیفیت در هر تکرار برای جمعیت‌های اولیه مختلف

جدول ۳- همگرایی الگوریتم به ازای جمعیت‌های مختلف

جمعیت (نفر)	تعداد تکرار الگوریتم	زمان اجرا (دقیقه)
۴	۱۹۰	۱۶۷
۱۰	۸۰	۱۶۳
۱۶	۵۰	۱۷۳

#### ۵-۶- پیش آزمایش برای توابع انتخاب الگوریتم ژنتیک

با توجه به تنوع موجود در توابع انتخاب الگوریتم ژنتیک، در پیش آزمایش دوم الگوریتم GARM به ازای توابع انتخاب چرخ رولت ۳۳، رتبه ۳۴ و نخبه‌گرایی ۳۵ و بدون انتخاب ۳۶ اجرا شده است. در حالت Non selection دقیقاً هر نسل جدید جایگزین نسل قبل می‌گردد. شکل ۴ میانگین همگرایی توابع انتخاب را برای تکرارهای مختلف نشان

[5] R. Burke, "Hybrid recommender systems: Survey and experiments," *User modeling and user-adapted interaction*, vol. 12, pp. 331-370, 2002.

[6] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," *Internet Computing, IEEE*, vol. 7, pp. 76-80, 2003.

[7] H. Ma, I. King, and M. R. Lyu, "Effective missing data prediction for collaborative filtering," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 39-46.

[8] G.-R. Xue, C. Lin, Q. Yang, W. Xi, H.-J. Zeng, Y. Yu, and Z. Chen, "Scalable collaborative filtering using cluster-based smoothing," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2005, pp. 114-121.

[9] S. Tyagi, and K. K. Bharadwaj, "A Collaborative Filtering Framework Based on Fuzzy Case-Based Reasoning," in *Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011*, 2012, pp. 279-288.

[10] D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles, "Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach," in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, 2000, pp. 473-480.

[11] S. Tyagi, and K. K. Bharadwaj, "Enhancing collaborative filtering recommendations by utilizing multi-objective particle swarm optimization embedded association rule mining," *Swarm and Evolutionary Computation*, vol. 13, pp. 1-12, 2013.

[12] W. Lin, S. A. Alvarez, and C. Ruiz, "Efficient adaptive-support association rule mining for recommender systems," *Data mining and knowledge discovery*, vol. 6, pp. 83-105, 2002.

[13] Z. Huang, H. Chen, and D. Zeng, "Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering," *ACM Transactions on Information Systems (TOIS)*, vol. 22, pp. 116-142, 2004.

[14] C. W.-k. Leung, S. C.-f. Chan, and F.-l. Chung, "A collaborative filtering framework based on fuzzy association rules and multiple-level similarity," *Knowledge and Information Systems*, vol. 10, pp. 357-381, 2006.

[15] M. Grami, R. Gheibi, and F. Rahimi, "A novel association rule mining using genetic algorithm," in *Information and Knowledge Technology (IKT), 2016 Eighth International Conference on*, 2016, pp. 200-204.

[16] W. Soto, and A. Olaya-Benavides, "A genetic algorithm for discovery of association rules," in *Chilean Computer Science Society (SCCC), 2011 30th International Conference of the*, 2011, pp. 289-293.

پایینی دارد و هدفمندتر و منطقی‌تر عمل می‌کند. همچنین لازم به ذکر است به عنوان مثال برای ۵۰۰ کاربر، اجرای الگوریتم GARM حدوداً ۸.۳۳ ساعت طول کشیده است یعنی به طور متوسط برای هر قلم ۱ دقیقه طول کشیده است تا قوانین انجمنی محاسبه شده و پیشنهادهای مشخص گردند که این مسئله با هدف اول الگوریتم‌های تکاملی سازگار است یعنی در زمان خیلی کم، بدون محاسبات پیچیده و زمان‌بر الگوریتم‌هایی مانند Apriori موفق به انجام پیشنهاد با کیفیت متوسط شده‌ایم که این زمان در الگوریتم MOPSO برای هر قلم ۱.۳ دقیقه طول کشیده است.

## ۶- نتیجه‌گیری و کارهای آینده

کاوش قوانین انجمنی یک مساله چند هدفه است که این مقاله با استفاده از یک روش اندازه‌گیری وزنی برای الگوریتم ژنتیک به صورت چند هدفه وزنی پرداخته است. الگوریتم ژنتیک با معرفی تابع ارزیاب خود به کاوش قوانین کاربردی در فضای جستجو می‌پردازد که این روش از روش‌های سنتی سریع‌تر و ضریب اطمینان بالاتری دارد. در نهایت الگوریتم GARM به دلیل سربار محاسباتی کمتر نسبت به الگوریتم MOPSO، سرعت بالاتری دارد و همین‌طور از لحاظ سرعت همگرایی با توجه به آزمایشات انجام شده می‌توان گفت الگوریتم GARM سرعت بالاتری نسبت به الگوریتم MOPSO دارا است که متعاقب آن باعث تولید پیشنهادهای بیشتر می‌شود.

دلیل اصلی بهبود پیشنهادات ذات الگوریتم می‌باشد که الگوریتم MOPSO ماهیتاً برای محیط‌های پیوسته مناسب است علی‌رغم اینکه الگوریتم GA سعی دارد با جایگشت‌های مختلف، نسل جدید به وجود آورد که برای سیستم گسسته این پژوهش مناسب‌تر می‌باشد، نتایج نیز این مساله را نشان می‌دهند.

قوانین قوی براساس همین معیار انتخاب و به سیستم پیشنهاددهنده داده می‌شوند. در حالی که معیارهای دیگری مانند لیفت<sup>۳۷</sup>، خی دو و غیره هستند که می‌تواند قوانین بهتری را انتخاب کند. سیستم پیشنهاددهنده‌ای که در این مقاله معرفی شد، قبلاً در مقاله الگوریتم MOPSO معرفی شده بود. بهبود سیستم پیشنهاددهنده برای تولید پیشنهادات با کیفیت‌تر می‌تواند برای کارهای بعدی مورد نظر باشد.

## مراجع

[1] G. Adomavicius, and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, pp. 734-749, 2005.

[2] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: an open architecture for collaborative filtering of netnews," in *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, 1994, pp. 175-186.

[3] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender systems: an introduction*: Cambridge University Press, 2010.

[4] A. Ansari, S. Essegai, and R. Kohli, "Internet recommendation systems," *Journal of Marketing research*, vol. 37, pp. 363-375, 2000.

آدرس پست الکترونیکی ایشان عبارت است از:

b.soleimani@grad.kashanu.ac.ir

**نسیم سلطانی** تحصیلات تکمیلی خود را در رشته مهندسی کامپیوتر گرایش نرم افزار به انجام رسانده و اکنون فارغ التحصیل مقطع کارشناسی ارشد و پژوهشگر در حوزه‌ی محاسبات ابری است. تحقیقات مورد علاقه‌ی نامبرده هوش مصنوعی، الگوریتم‌های فرا اکتشافی،



داده کاوی و بهبود زمانبندی وظایف می‌باشد.

آدرس پست الکترونیکی ایشان عبارت است از:

n.soltani@naeini.ac.ir

**رضا مفیدی** فارغ التحصیل رشته مهندسی کامپیوتر گرایش نرم افزار می‌باشد که اکنون فارغ التحصیل مقطع کارشناسی ارشد است که در حوزه‌ی سیستم‌های پیشنهاددهنده و قوانین انجمنی فعالیت دارد. تحقیقات مورد علاقه‌ی ایشان الگوریتم‌های فرا اکتشافی و



داده کاوی می‌باشد.

آدرس پست الکترونیکی ایشان عبارت است از:

rezamofidi66@naeini.ac.ir

**محمد حسین ندیمی شهرکی** استادیار و رئیس مرکز تحقیقات مه داده در دانشگاه آزاد اسلامی واحد نجف آباد است. موضوعات پژوهشی مورد علاقه‌ی ایشان داده کاوی، کاوش شبکه‌های اجتماعی، سیستم‌های هوشمند و کاوش داده‌های پزشکی



می‌باشد.

آدرس پست الکترونیکی ایشان عبارت است از:

nadimi@ieee.org, nadimi@iaun.ac.ir

#### اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۱۲/۰۵

تاریخ اصلاح: ۱۳۹۶/۱۱/۲۰

تاریخ قبول شدن: ۱۳۹۷/۰۲/۰۵

نویسنده مرتبط: بهزاد سلیمانی نیسانی، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد اصفهان (خوراسگان)، اصفهان، ایران.

[17] S. Ghosh, S. Biswas, D. Sarkar, and P. Sarkar, "Association rule mining algorithms and Genetic Algorithm: A comparative study," in *Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on*, 2012, pp. 202-205.

[18] T. I. R. Agrawal, and A. Swami, "Mining association rules between sets of items in large databases," *Proceeding of the ACM International Conference on Management of Data, ACM*, pp. 207-216, 1993.

[19] M. Srinivas, and L. M. Patnaik, "Genetic algorithms: A survey," *Computer*, vol. 27, pp. 17-26, 06 August 2002 1994.

[20] A. Ganjehkaviri, M. M. Jaafar, S. Hosseini, and H. Barzegaravval, "Genetic algorithm for optimization of energy systems: Solution uniqueness, accuracy, Pareto convergence and dimension reduction," *Energy*, vol. 119, pp. 167-177, 2017.

[21] T. İnkaya, and M. Akansel, "Coordinated scheduling of the transfer lots in an assembly-type supply chain: A genetic algorithm approach," *Journal of Intelligent Manufacturing*, vol. 28, pp. 1005-1015, 2017.

[22] M. M. J. Kabir, S. Xu, B. H. Kang, and Z. Zhao, "A New Multiple Seeds Based Genetic Algorithm for Discovering a Set of Interesting Boolean Association Rules," *Expert Systems with Applications*, 2017.

[23] F. Ramezani, J. Lu, J. Taheri, and F. K. Hussain, "Evolutionary algorithm-based multi-objective task scheduling optimization model in cloud environments," *World Wide Web*, vol. no. 6, pp. 1-21, 2015.

[24] F. Ramezani, J. Lu, and F. Hussain, "Task Scheduling Optimization in Cloud Computing Applying Multi-Objective Particle Swarm Optimization," in *Service-Oriented Computing*, ed: Springer, 2013, pp. 237-251.

[25] S.-Y. Wur, and Y. Leu, "An effective Boolean algorithm for mining association rules in large databases," in *Database Systems for Advanced Applications, 1999. Proceedings., 6th International Conference on*, 1999, pp. 179-186.

[26] GroupLens. (1998, 2016-11-17). *MovieLens Dataset*. Available: <https://grouplens.org/datasets/movielens/>

**بهزاد سلیمانی نیسانی** دانشجوی سال پنجم دکتری در

رشته مهندسی نرم افزار کامپیوتر است. حوزه تحقیقاتی که تا کنون در آنها پژوهش انجام داده و مقاله نیز به چاپ رسانده است بیشتر شامل الگوریتم‌های داده کاوی در زمینه قواعد انجمنی، طبقه بندی، سیستم‌های پیشنهاددهنده بوده است.



از جمله زمینه‌های تحقیقاتی دیگر ایشان در زمینه سیستم‌های توزیع شده مانند شبکه‌های نظیر به نظیر، محاسبات ابری، همچنین در زمینه‌های مهندسی نرم افزار و هوش مصنوعی می‌توان اشاره کرد. وی هم‌اکنون مدرس دانشگاه کاشان و دانشگاه آزاد اسلامی واحد اصفهان (خوراسگان) و دانشگاه شهید اشرفی اصفهانی است که بر روی الگوریتم‌های متن کاوی و کشف گزارش‌های متنی تکراری در پروژه دکتری خود کار می‌کند.

<sup>1</sup>Item

<sup>2</sup>Content-Based

<sup>3</sup>Knowledge-Based

<sup>4</sup>CF: Collaborative Filtering

<sup>5</sup>Collaborative Recommender Systems

<sup>6</sup>Sparse

<sup>7</sup>ARM: Association Rule Mining

<sup>8</sup>Huang

<sup>9</sup>Leung

<sup>10</sup>FARAMS: Fuzzy Association Rule And Multiple-Level Similarity

<sup>11</sup>ASARM: Adaptive-Support Association Rule Mining

<sup>12</sup>Apriori

<sup>13</sup>MOPSO: Multi-Objective Particle Swarm Optimization Association Rule Mining

<sup>14</sup>Agrawal

- 
- <sup>15</sup>Tail
  - <sup>16</sup>Head
  - <sup>17</sup>Support
  - <sup>18</sup>Confidence
  - <sup>19</sup>Genetic Algorithm
  - <sup>20</sup>Crossover
  - <sup>21</sup>Mutation
  - <sup>22</sup>Genetic Algorithm Association Rule Mining
  - <sup>23</sup>Representation Schema
  - <sup>24</sup>Transformation Schema
  - <sup>25</sup>Like
  - <sup>26</sup>Dislike
  - <sup>27</sup>Selection
  - <sup>28</sup>Cross Over
  - <sup>29</sup>Mutation
  - <sup>30</sup>Matlab
  - <sup>31</sup>Movielens
  - <sup>32</sup>Iteration
  - <sup>33</sup>Roulette
  - <sup>34</sup>Rank
  - <sup>35</sup>Elitism
  - <sup>36</sup>Non Selection
  - <sup>37</sup>Lift

## بهبود توان مصرفی و زمان اجرا در رایانش ابری رهیافت تخصیص منبع و زمانبندی وظایف

عاطفه یکتا اول محمود فتحی

دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران

### چکیده

جریان‌های کاری علمی یکی از راه‌های معروف به منظور مدلسازی برنامه‌های کاربردی ابر می‌باشند. یکی از چالشی‌ترین موضوعات تحقیق در این زمینه، زمانبندی وظایف این جریان کاری بر روی منابع در دسترس است به گونه‌ای که بتوانیم به حداقل زمان تکمیل اجرای وظایف دست یابیم. از طرفی، یکی از مسائل چالش‌برانگیز دیگر در رایانش ابر کاهش میزان توان مصرفی مراکز داده می‌باشد. مشکل موجود مغایرت این معیارهای کیفی است که الگوریتم‌های بسیاری سعی در بهینه‌سازی آنها دارند. در این مقاله به منظور کاهش توان مصرفی و زمان اجرا، تخصیص منبع را در دو سطح انجام داده‌ایم: ابتدا در سطح زیرساخت به گونه‌ای منابع فیزیکی را به ماشین‌های مجازی تخصیص می‌دهیم، که با افزایش بهره‌وری منابع کمترین میزان توان مصرفی را در مرکز داده داشته باشیم. در مرحله بعدی، برنامه‌های کاربردی را به گونه‌ای بر روی این ماشین‌های مجازی زمانبندی می‌کنیم که به حداقل زمان تکمیل وظایف دست یابیم. نتایج آزمایش‌ها بیانگر موفقیت این روش می‌باشد.

**کلمات کلیدی:** رایانش ابر، گراف وظایف، زمانبندی وظایف، تخصیص منابع، مدیریت توان.

### ۱- مقدمه

داده است. براساس [۱]، گزارشات نشان می‌دهند که حدود ۵۰ درصد از بودجه‌ی مدیریتی مرکز داده‌ی آمازون صرف هزینه‌هایی برای خنک‌سازی ماشین‌های فیزیکی می‌شود. در این میان با وجود بهبودها و بهینه‌سازی‌هایی که در بهره‌وری انرژی سخت‌افزار صورت گرفته است، مصرف انرژی مراکز داده به خاطر افزایش نیاز به منابع رایانشی روز به روز در حال افزایش است و این افزایش مصرف انرژی به دلیل تأثیری که بر روی هزینه‌های عملیاتی می‌گذارد یکی از نگرانی‌های عمده برای صاحبان مراکز داده محسوب می‌شود. مسلماً به همراه داشتن پارامترهای کیفیت سرویس مانند کاهش زمان اجرای برنامه‌های کاربردی می‌تواند باعث افزایش توان مصرفی شود. مشکل موجود مغایرت این معیارهای کیفی است که این الگوریتم‌ها سعی در بهینه‌سازی آنها دارند. آنچه مسلم است این است که تخصیص منابع قسمت مهمی از مدیریت منابع است و انتخاب متدهای مناسب برای اختصاص یافتن بار کاری<sup>۵</sup> یا ماشین مجازی به میزبان مناسب راه رسیدن به هدف تخصیص بهینه منابع است.

تاکنون الگوریتم‌های بسیاری در رابطه با نحوه‌ی تخصیص و یا زمانبندی منابع در رایانش ابر ارائه شده‌اند که هدف برخی از این الگوریتم‌ها بهینه‌سازی عاملی از کیفیت سرویس مانند کاهش زمان تکمیل وظایف و یا عواملی مانند کاهش توان مصرفی مراکز داده<sup>۱</sup> بوده است. جریان‌های کاری علمی<sup>۲</sup> یکی از راه‌های بسیار معروف و مرسوم به منظور مدلسازی برنامه‌های کاربردی<sup>۳</sup> می‌باشند به گونه‌ای که این برنامه‌ها بتوانند بر روی سیستم‌های توزیع شده نظیر ابر اجرا شوند. زمانی که یک جریان کاری تشکیل می‌شود، یکی از چالشی‌ترین موضوعات تحقیق در این زمینه این است که چگونه وظایف مختلف این جریان کاری را بر روی منابع در دسترس زمانبندی کنیم. زمانبندی موثر برنامه‌های کاربردی یک راه حیاتی برای دستیابی به کارایی بالا در محیط‌های محاسباتی ناهمگن می‌باشد. از طرفی یکی از مسائل عمده و چالش‌برانگیز دیگر در رایانش ابر کاهش میزان توان مصرفی مراکز



دسترس و ۳- گام مرتب کردن وظایف به منظور مرتب کردن وظایف نگاشت شده درون هر پردازنده.

### ۱۲- تکثیر وظیفه

ایده‌ی پشت این الگوریتم‌های زمانبندی این است که گرافی از وظایف را با نگاشت برخی از وظایف زمانبندی کند که این کار سربار ارتباط میان پردازنده‌ها را کاهش می‌دهد. الگوریتم‌های تکثیر وظیفه مطابق با استراتژی انتخاب وظیفه برای عمل تکثیر متفاوتند. الگوریتم‌های این دسته معمولاً برای تعدادی از پردازنده‌های همانند بوده و نسبت به الگوریتم‌های دیگر دسته‌ها پیچیدگی بیشتری دارند.

### ۲-۱-۱-۲- تکنیک‌های جستجوی تصادفی هدایت شده<sup>۱۳</sup> [۲]

این دسته از تکنیک‌ها از انتخاب تصادفی استفاده می‌کنند تا در میان فضای مسئله هدایت شوند. البته عملکرد این جستجو صرفاً با یک جستجوی تصادفی که از طریق متدهای جستجو انجام می‌گیرد یکسان نمی‌باشد. این تکنیک‌ها دانش به دست آمده از نتایج جستجوی قبلی را با برخی از ویژگی‌های تصادفی ترکیب می‌کنند تا نتایج جدید تولید کنند. الگوریتم‌های ژنتیک مشهورترین و گسترده‌ترین تکنیک‌های استفاده شده برای مسائل زمانبندی وظایف می‌باشند. الگوریتم‌های ژنتیک کیفیت خوبی از خروجی زمانبندی تولید می‌کنند. اگرچه که دفعات زمانبندی آنها معمولاً بسیار بیشتر از تکنیک‌های مبتنی بر مکاشفه است. علاوه بر این در یک الگوریتم ژنتیک، چندین پارامتر کنترلی باید به درستی تعیین شوند. مجموعه‌ی بهینه از پارامترهای کنترلی مورد استفاده برای زمانبندی یک گراف ممکن است برای گرافی دیگر جواب بهینه را ندهد. علاوه بر الگوریتم‌های ژنتیک، متدهای simulated annealing و جستجوی محلی، متدهای دیگر در این دسته هستند.

### ۲-۱-۲- زمانبندی پویای وظایف

در زمانبندی پویا، تعداد وظایف، مکان ماشین‌های مجازی و شیوه‌ی تخصیص منبع ثابت نیست و دفعات ورود وظایف قبل از ارسال آنها مشخص نیست [۲]. زمانبندی پویا، وظایف را بر روی منابع در زمان اجرا زمانبندی می‌کند تا به یک تعدیل بار میان المان‌های پردازشی دست باید [۴].

در مقاله‌ی [۵] یک ایده‌ی زمانبندی پویا به نام DGS<sup>۱۴</sup> معرفی شده که به صورت پویا منابع مجازی را به وظایف محاسباتی تخصیص می‌دهد و با استفاده از استراتژی حریصانه<sup>۱۵</sup> بهبود یافته فرآیند زمانبندی و اجرا را تکمیل می‌کند. در ابتدا اگر ناظر ماشین مجازی تشخیص داد که ماشین مجازی در مرکز داده وجود ندارد یا توان محاسباتی آن محدود است، ماشین‌های مجازی جدید ایجاد خواهند شد. بعد از اینکه اولین زمانبندی توسط استراتژی حریصانه صورت گرفت، اگر ناظر ماشین مجازی دید که برخی از ماشین‌های مجازی در حال اجرا در وضعیت بحرانی به سر می‌برند، ماشین‌های مجازی جدید ایجاد کرده و دومین زمانبندی را شکل خواهد داد. در حقیقت در بازه‌های زمانی از پیش تعیین شده مقدار بار هر ماشین مجازی محاسبه شده و این مقدار با مقدار بار آستانه مقایسه می‌شود. در صورتی که این مقدار از مقدار آستانه‌ی تعیین شده بیشتر باشد، ماشین مجازی دچار سربار شده و ناظر ماشین مجازی، ماشین مجازی دیگری دقیقاً مشابه با ماشین مجازی قبلی ایجاد خواهد کرد. سپس مازول زمانبندی وظایف، وظایفی را که در صف ماشین مجازی سربار شده قرار داشتند، به دو ماشین مجازی روشن تخصیص می‌دهد. در صورتی که باز هم سربار وجود داشت این روال تکرار خواهد شد.

### ۳- تکنیک‌های کاهش توان مصرفی

امروزه برنامه‌های کاربردی مدرن که کاربرد تجاری یا علمی دارند نیاز به زیرساخت‌های رایانشی با کارایی بالایی دارند. این منجر به مراکز داده رایانشی با مقیاس بسیار بالا می‌شود که مصرف توان الکتریکی بالایی دارند. با وجود بهبودها و بهسازی‌هایی که در بهره‌وری انرژی سخت‌افزار صورت گرفته است، مصرف انرژی کلی مرکز داده به خاطر افزایش نیاز به منابع رایانشی روز به روز در حال افزایش است. جدای از هزینه‌های سرسام‌آور کاربردی، ساختن مرکز داده‌ای که برای پاسخگویی به اوج بار یا حجم کاری درخواستی طراحی می‌شوند در حالی که بهره‌وری متوسط پایینی دارند از دلایل دیگر مصرف توان بالای مراکز داده می‌باشند [۶]. علاوه بر آن کمبود سیستم‌های خنک‌کننده یا خرابی آنها می‌تواند منجر به گرم شدن شدید منابع مرکز داده شده که باعث پایین آمدن قابلیت اطمینان سیستم و کاهش طول عمر دستگاه‌ها و تجهیزات داخل مرکز داده شود. علاوه بر آن مصرف توان بالای زیرساخت‌ها باعث انتشار قابل توجهی گاز کربن دی اکسید منجر به پدیده‌ی گلخانه‌ای می‌شود [۷].

**تکنیک‌های سخت‌افزاری:** از نقطه‌نظر سخت‌افزاری مدیریت ایستای انرژی شامل روش‌های بهینه‌سازی است که در زمان طراحی در سطوح مختلف سخت‌افزار شامل مدار، لاجیک و معماری سیستم به کار می‌روند. تکنیک‌های سخت‌افزاری مدیریت توان به صورت پویا متناسب با اجزای سخت‌افزاری فرق می‌کنند. اما این روش‌ها معمولاً به دو دسته‌ی مقیاس‌پذیری پویای ولتاژ و فرکانس و خاموش کردن اجزا به‌طور کامل یا جزئی در زمان‌های بیکاری یا عدم فعالیت سیستم تقسیم‌بندی می‌شوند. مقیاس‌پذیری پویای ولتاژ و فرکانس<sup>۱۶</sup> در بعضی از اجزای سخت‌افزاری نظیر پردازنده‌ها تعبیه شده است. با بهره‌گیری از این قابلیت پردازنده‌ها در زمان‌های کاهش بار کاری می‌توانند در وضعیت‌های عملیاتی با مصرف پایین توان قرار بگیرند. مصرف انرژی در مراکز داده با استفاده از این قابلیت فقط می‌تواند حدود ۱۰ تا ۲۰ درصد کاهش داده شود [۸].

**تکنیک‌های نرم‌افزاری:** روش‌های نرم‌افزاری مدیریت ایستای انرژی شامل استفاده از تکنیک‌های برنامه‌نویسی موازی، الگوریتم‌های چندبخشی<sup>۱۷</sup>، بهینه‌سازی کد برنامه و بهینه‌سازی فرآیند کامپایل کد هستند که به هنگام توسعه‌ی برنامه‌های کاربردی استفاده می‌شوند. تکنیک‌های نرم‌افزاری مدیریت توان به صورت پویا شامل مکانیزم‌ها و سیاست‌هایی هستند که از دانش درباره‌ی وضعیت جاری سیستم استفاده می‌کنند. این تکنیک‌های نرم‌افزاری مطابق با سیاست‌های خود از روش‌های سخت‌افزاری مدیریت پویای توان استفاده می‌کنند.

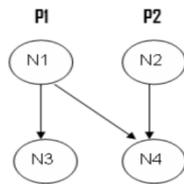
به دلیل محدودیت‌های موجود بسیاری از کارهای مرتبط برای کاهش انرژی مصرفی از رویکرد متمرکزسازی بارهای کاری بر روی تعداد کمتری از سرورها و خاموش کردن سرورهای بیکار یا قرار دادن آنها در مدهای خواب یا خواب زمستانی و یا از ترکیب رویکرد متمرکزسازی با قابلیت مقیاس‌پذیری ولتاژ و فرکانس استفاده کرده‌اند [۸].

**فناوری مجازی‌سازی:** در کنار تکنیک‌های نرم‌افزاری و سخت‌افزاری برای مدیریت توان، فناوری مجازی‌سازی می‌تواند با بهبود درصد استفاده از منابع، توان مصرفی سرورها را کاهش دهد. فناوری مجازی‌سازی این امکان را فراهم می‌سازد که یک یا بیشتر ماشین‌های مجازی بتوانند بر روی یک سرور فیزیکی واحد اجرا شوند. به این ترتیب ضمن افزایش میزان استفاده از منابع، مقدار سخت‌افزار مورد نیاز برای اجرای برنامه‌های کاربردی کاهش می‌یابد. با بهره‌گیری از قابلیت مهاجرت<sup>۱۸</sup> ماشین‌های مجازی، امکان مدیریت پویای منابع سخت‌افزاری فراهم می‌شود. ماشین‌های مجازی می‌توانند در زمان اجرا بر روی حداقل سرور فیزیکی یکپارچه‌سازی شوند. در این حالت سرورهای بیکار می‌توانند در وضعیت خاموشی و یا مدهای عملیاتی کم‌مصرف قرار بگیرند. از آنجایی که انرژی مصرفی ماشین‌های

صرف کلی توان افزایش پیدا نکند. این الگوریتم حداقل به میزان ۳۹٪ منجر به بهبود توان مصرفی می‌شود.

#### ▪ الگوریتم LPHEFT

این الگوریتم [۱۰] که نسخه‌ی کم‌مصرف الگوریتم HEFT می‌باشد زمان اجرای یکسان و میزان توان مصرفی کمتری نسبت به الگوریتم HEFT نتیجه می‌دهد (الگوریتم HEFT در بخش ۶-۱ به‌طور کامل توضیح داده خواهد شد). در این الگوریتم ولتاژ پردازنده در طول زمان‌های بیکاری پردازنده مقیاس‌بندی می‌شود. زمان بیکاری برای یک پردازنده زمانی ایجاد می‌شود که آن پردازنده منتظر اجرای وظیفه‌ای است که آن وظیفه وابسته به وظایف دیگری است که بر روی پردازنده‌های دیگر اجرا می‌شود. زمان بیکاری با کاهش سرعت پردازنده حذف می‌شود که این کار باعث می‌شود که مدت زمانی که پردازنده وظایف قبلی را انجام می‌دهد، افزایش یابد. توجه به این نکته ضروری است که تمام وابستگی‌های بین هر وظیفه با اجدادش می‌بایست برآورده شود.



با استفاده از این الگوریتم اگر  $EFT(N, P1) + C(P1, P2) > EFT(N2, P2)$  باشد،  $EFT(N2, P2)$  افزایش پیدا می‌کند بدلیل اینکه P2 مجبور است منتظر بماند تا اجرای N1 تمام شده و سپس N4 را اجرا کند. اگر زمان بیکاری روی P2 جستجو شود، در صورتی که  $EST(N4, P2)$  کمتر از  $EFT(N2, P2)$  باشد، زمان‌های بیکاری را بر روی این پردازنده یافت خواهیم کرد. به همین دلیل ما به دنبال یافتن زمان‌های بیکار در دسترس بر روی تمامی پردازنده‌ها هستیم.

### ۵- مدلسازی مسئله

#### ۵-۱- مدل جریان کاری

جریان کاری یک برنامه‌ی کاربردی معمولاً به صورت یک گراف بدون دور جهت‌دار (DAG) مدلسازی شده و به صورت  $W=(A, D)$  نشان داده می‌شود. این گراف شامل  $n$  وظیفه  $A = \bigcup_{i=1}^n A_i$  می‌باشد که این وظایف از طریق جریان‌های کنترلی<sup>۲۱</sup> و جریان‌های داده<sup>۲۲</sup> به همدیگر وابسته می‌باشند.  $D = \{(A_i, A_j, Data_{ij}) \mid (A_i, A_j) \in A \times A\}$  نشان‌دهنده‌ی سایز داده‌ی مورد نیاز برای انتقال از وظیفه‌ی  $A_i$  به وظیفه‌ی  $A_j$  می‌باشد. مجموعه‌ی وظایف والد<sup>۲۳</sup>  $A_i$  (وظایفی که قبل از اجرای وظیفه‌ی  $A_i$  می‌بایست اجرا شوند) را به صورت  $Pred(A_i) = \{A_k \mid (A_k, A_i, Data_{kj}) \in D\}$  نشان می‌دهیم. ما فرض می‌کنیم که بار کاری محاسباتی هر وظیفه به صورت طول هر وظیفه داده شده است که در واقع نشان‌دهنده‌ی تعداد دستورالعمل‌های ماشین مورد نیاز برای اجرای آن وظیفه می‌باشد.

#### ۵-۲- مدل زمان اجرای وظایف در گراف

برای محاسبه‌ی زمان اجرای یک جریان کاری ابتدا می‌بایست زمان اجرای وظیفه‌ی  $A_i$  بر روی ماشین مجازی  $VM_j$  را بدست آوریم که ما این زمان را به صورت

فیزیکی در حالت بیکاری، به عنوان مثال در حالت ۱۰ درصد استفاده از پردازنده، بیشتر از ۵۰ درصد انرژی مصرفی نسبت به حالت حداکثر بهره‌وری آنها می‌باشد [۹]. با استفاده از قابلیت متمرکزسازی ماشین‌های مجازی بر روی تعداد کمتری از ماشین‌های فیزیکی می‌توان تعداد ماشین‌های بیکار را کاهش و در نتیجه انرژی مصرفی را کمینه کرد.

### ۴- کارهای مرتبط با هدف کاهش زمان اجرا و کاهش توان مصرفی

الگوریتم‌های زمانبندی مبتنی بر لیست، مشهورترین الگوریتم‌ها در زمینه‌ی زمانبندی ایستای وظایف می‌باشند. این الگوریتم‌ها به هر وظیفه اولویتی نسبت داده و بر همین اساس وظایف را به ترتیب نزولی مرتب می‌کنند. سپس وظایف براساس این اولویت زمانبندی می‌شوند. سه مکاشفه‌ی زمانبندی مبتنی بر لیست که به منظور کاهش توان مصرفی ارائه شده‌اند عبارتند از:

- 1- Power aware list scheduling (PALS)
- 2- Power aware task clustering (PATC)
- 3- Low power heterogeneous-earliest-finish-time algorithm (LPHEFT)

#### ▪ الگوریتم PALS

الگوریتم PALS که توسط لیزا وانگا<sup>۱۹</sup> و همکارانش [۴] معرفی شد، یک الگوریتم زمانبندی مبتنی بر لیست برای یافتن بهترین زمان پاسخگویی و یک مکاشفه‌ی زمانبندی آگاه از توان برای وظایف موازی ارائه می‌کند. در این الگوریتم وظایف از طریق الگوریتم زمانبندی ETF<sup>۲۰</sup> زمانبندی می‌شوند. همچنین به منظور کاهش توان مصرفی، ولتاژ پردازنده برای وظایفی که حساس نیستند و اهمیت زیادی ندارند با استفاده از الگوریتم Non-critical time slot voltage scaling مقیاس‌بندی شده و کاهش می‌یابد. این الگوریتم حداقل به میزان ۴۴.۳٪ منجر به کاهش توان مصرفی می‌شود.

- الگوریتم زمانبندی ETF:

الگوریتم ETF یک الگوریتم زمانبندی مبتنی بر لیست برای یافتن بهترین زمان پاسخگویی برای هر وظیفه می‌باشد. این الگوریتم به هر وظیفه اولویتی نسبت داده سپس وظایف آماده با بالاترین اولویت را انتخاب کرده و آنها را بر روی پردازنده‌ای که زودترین زمان شروع را دارد زمانبندی می‌کند.

- الگوریتم Non-Critical Time slot Voltage scaling:

این الگوریتم نشان می‌دهد که چطور برای وظایف غیر حساس مقیاس‌بندی ولتاژ انجام می‌شود. برای هر پردازنده، تمامی بازه‌های زمانی جستجو می‌شود. زمانی که پردازنده بیکار است یا داده‌ای در یک بازه‌ی زمانی انتقال می‌یابد، الگوریتم فرکانس پردازنده را کاهش می‌دهد. هنگامی که در یک بازه‌ی زمانی یک وظیفه‌ی غیر حساس اجرا می‌شود، میزان زمان جای افتاده را محاسبه می‌کند، زمان اجرای آن وظیفه را به اندازه‌ی این زمان محاسبه شده گسترش می‌دهد و فرکانس پردازنده را به مقدار مناسب کاهش می‌دهد.

#### ▪ الگوریتم PATC

الگوریتم PATC [۴] یک مکاشفه‌ی زمانبندی برای وظایف موازی است که با هدف کاهش توان مصرفی توسط لیزا وانگا و همکارانش معرفی شد. این الگوریتم در ابتدا تمام یال‌ها را به عنوان امتحان نشده علامت‌گذاری کرده و هر وظیفه را به یک شاخه‌ی جداگانه تخصیص می‌دهد. بعد از مرتب کردن تمامی یال‌ها به ترتیب نزولی زمان انتقال، الگوریتم به صورت تکراری وظایفی را که هزینه‌ی انتقال میان آنها بالاست با همدیگر ترکیب می‌کند البته این کار را در صورتی انجام می‌دهد که

مجموعه‌ای از  $n$  ماشین مجازی داده شده است تا بر روی  $m$  ماشین فیزیکی قرار داده شود.

$$\{VM_i(pe_i, mips_i, ram_i, bw_i) \mid i = 1, 2, 3, \dots, n\}$$

$$\{M_j(PE_j, MIPS_j, RAM_j, BW_j) \mid j = 1, 2, 3, \dots, n\}$$

هر ماشین مجازی به  $pe_i$  المان پردازشی،  $mips_i$ ،  $ram_i$ ،  $bw_i$  (پهنای باند) بدون توقف یا بدون مهاجرت در طول زمان اجرا نیاز دارد. در اینجا ما سه نوع منبع پردازنده، حافظه و پهنای باند را مورد بررسی قرار می‌دهیم. هدف این زمانبندی مصالحه‌ای بین کاهش مصرف انرژی در تامین ساختن بیشترین نیازمندی‌های ماشین‌های مجازی است، بدین معنی که بتوانیم میزبانی را انتخاب نماییم که در عین حال که میزان توان مصرفی کمتری نسبت به سایر میزبان‌ها دارد، ظرفیت و کارایی بالاتری داشته و بتواند تعداد بیشتری ماشین مجازی را میزبانی کند، تا بدین ترتیب از تعداد کمتری ماشین فیزیکی استفاده شود و توان مصرفی مرکز داده کاهش یابد. الگوریتم تعدیل بار که الگوریتم پیش‌فرض استفاده شده برای تخصیص منابع فیزیکی به ماشین‌های مجازی در CloudSim و WorkflowSim می‌باشد به صورت زیر است:

```

Input: a VM (vm) and a set of Host (H)
Output: a host for vm
For each host in host list do
Select the host with less processor in use
CloudSim.allocationMap.put (vm.getId (), host.getId ());
    
```

شکل ۲- الگوریتم تعدیل بار برای تخصیص منابع فیزیکی به ماشین‌های مجازی

این الگوریتم هر ماشین مجازی را بر روی میزبانی قرار می‌دهد که نسبت به سایر میزبان‌ها تعداد کمتری پردازنده‌ی در حال استفاده داشته باشد به عبارت دیگر بهره‌وری کمتری داشته باشد.

### ۶-۱- الگوریتم تخصیص منبع پیشنهادی آگاه به توان

این الگوریتم هر ماشین مجازی را بر روی میزبانی قرار می‌دهد که بزرگترین مقدار  $G_H = \frac{h.TotalMips}{h.GetPower(100\%)}$  را داشته باشد. برای هر میزبان،  $G_H$  می‌تواند به عنوان نسبتی از مجموع MIPS‌های تمامی پردازنده‌های ماشین فیزیکی به بیشترین توان مصرفی پردازنده زمانی که میزان بهره‌وری ۱۰۰٪ است، در نظر گرفته شود. با این کار ما در واقع میزبانی را انتخاب می‌نماییم که در عین حال که میزان توان مصرفی کمتری نسبت به سایر میزبان‌ها دارد، سرعت، کارایی و ظرفیت بالاتری داشته، چرا که هر چقدر  $h.TotalMips$  بیشتر باشد، تعداد هسته‌های آن میزبان بیشتر بوده و برای تخصیص ماشین‌های مجازی به تعداد کمتری میزبان نیازمند می‌شویم و بدین ترتیب از تعداد کمتری ماشین فیزیکی استفاده شده و توان مصرفی مرکز داده کاهش می‌یابد.

در الگوریتم پیشنهادی ما شکل ۳، ورودی به صورت مجموعه‌ای از میزبان‌ها و ماشین مجازی، که قرار است بر روی میزبان مناسب قرار داده شود، داده شده است. در ابتدا متغیر MaxValue را برابر با یک مقدار کمینه قرار می‌دهیم سپس برای هر میزبان، میزان بهره‌وری پردازنده‌های آن را با استفاده از رابطه ۴ محاسبه می‌کنیم. در صورتی که این مقدار (میزان بهره‌وری پردازنده‌ها) کمتر از ۱ باشد، مقدار  $G_H$  برای آن میزبان و ماشین مجازی محاسبه می‌شود، در صورتی که مقدار  $G_H$  میزبان از MaxValue بیشتر باشد و سایر نیازمندی‌های دیگر آن ماشین مجازی را (مانند حافظه و پهنای باند) فراهم کند مقدار  $G_H$  به عنوان MaxValue و آن میزبان به عنوان مناسب‌ترین میزبان انتخاب می‌شود.

مجموع زمان مورد نیاز برای انتقال بزرگترین داده ورودی از هر  $A_p \in pred(A_p)$  و زمان مورد نیاز برای اجرای وظیفه‌ی  $A_i$  بر روی ماشین مجازی  $VM_j$  تعریف می‌کنیم:

$$t(A_i, VM_j) = \max_{A_p \in pred(A_i)} \left\{ \frac{Data_{pi}}{b_{pi}} \right\} + \frac{workload(A_i)}{s_j} \quad (1)$$

که  $Data_{pi}$  سایز داده‌ای است که باید بین  $A_p$  و  $A_i$  انتقال داده شود،  $b_{pi}$  پهنای باند بین ماشین مجازی که وظیفه‌ی  $A_p$  را اجرا کرده با  $VM_j$  که قرار است وظیفه‌ی  $A_i$  را اجرا کند، می‌باشد.  $workload(A_i)$ ، طول وظیفه‌ی  $A_i$  (تعداد دستورالعمل‌های ماشین برای اجرای  $A_i$ ) و  $s_j$  سرعت ماشین مجازی  $VM_j$  می‌باشد. بنابراین زمان تکمیل اجرای وظیفه‌ی  $A_i$  ( $T_{A_i}$ ) با بررسی زمان اجرای وظیفه‌ی  $A_i$  و اجداد این وظیفه به صورت زیر تعریف می‌شود:

$$T_{A_i} = \begin{cases} t(A_i, VM_j) \rightarrow pred(A_i) = \phi \\ \max_{A_p \in pred(A_i)} \{T_{A_p} + t(A_i, VM_j)\} \rightarrow pred(A_i) \neq \phi \end{cases} \quad (2)$$

بنابراین زمان اجرای یک جریان کاری به صورت زیر محاسبه می‌شود [۱۱]:

$$T_w = \max_{i \in [1, n]} \{T(A_i, VM_j)\} \quad (3)$$

### ۵-۳- مدل توان

در اینجا به جای استفاده از مدل توان مصرفی در پردازنده‌های چند هسته‌ای، از داده‌های واقعی برای مصرف توان استفاده می‌کنیم. این داده‌ها به وسیله‌ی نتایج SPEC power benchmark فراهم شده است که در جدول ۴ در بخش پیوست آمده است. مقادیر درون این جدول، میزان توان مصرفی هر یک از سرورها را به ازای افزایش ۱۰ درصدی میزان بهره‌وری نشان می‌دهد. براساس [۹] بهره‌وری پردازنده‌ی یک میزبان را به صورت مجموع بهره‌وری پردازنده‌های مختلف آن میزبان تعریف می‌کنیم.

$$U_{cpu}(t) = \sum_{c=1}^{PE} \sum_{i=1}^r \frac{MIPS_{i,c}}{MIPSC} \quad (4)$$

در رابطه‌ی بالا  $PE$  نشان‌دهنده‌ی تعداد پردازنده‌های میزبان و  $r$  نشان‌دهنده‌ی تعداد ماشین‌های مجازی هستند که بر روی میزبان قرار دارند. همچنین  $MIPS_{i,c}$  نشان‌دهنده‌ی MIPS تخصیص داده شده‌ی پردازنده‌ی  $c$  میزبان به ماشین مجازی  $i$  و  $MIPSC$  نشان‌دهنده‌ی مجموع MIPS مربوط به پردازنده‌ی  $c$  میزبان می‌باشد.

### ۶- تخصیص منبع در سطح زیرساخت با هدف کاهش توان مصرفی

در این قسمت شیوه‌ی تخصیص جدیدی را مطرح می‌کنیم تا ماشین فیزیکی که از نظر میزان توان مصرفی کارآمدتر است را برای نگاشت ماشین‌های مجازی انتخاب کنیم. مسئله‌ی زمانبندی ماشین‌های مجازی را به صورت زیر فرمول‌بندی می‌کنیم:

این روند برای تمامی میزبان‌های موجود ادامه یافته و هر بار پس از بررسی بهره‌وری پردازنده‌های میزبان، در صورتی که مقدار  $G_h$  از مقدار  $MaxValue$  قبلی بیشتر باشد، جایگزین  $MaxValue$  قبلی می‌شود و آن میزبان به عنوان مناسب‌ترین میزبان انتخاب می‌شود. برای تخصیص سایر ماشین‌های مجازی، میزبان مناسب با استفاده از این الگوریتم محاسبه می‌شود.

```

Find Host for VM by Green Metric
Energy aware VM Allocation Policy
Output: best Host (a best host for allocation of VM)
Input: a VM (vm) and a set of Host (H)
{
  MaxValue=Integer.MIN_VALUE
  For all (Power Host h in Host)
  {
    Calculate utilization of CPU:
    
$$U_{cpu}(t) = \frac{PE}{c} \sum_{r=1}^r \frac{MIPS_{i,c}}{MIPSC_c}$$

    If (utilization of CPU < 1)
    {
      
$$G_h = \frac{h.TotalMips}{h.GetPower(100\%)}$$

      If ( $G_h > MaxValue$  & &  $h.IsSutableForVM(vm)$ )
      {
        MaxValue =  $G_h$ 
        bestHost = h
      }
    }
  }
  Return best Host
CloudSim.AllocationMap.put (vm.getId, host.getId)
}

```

شکل ۳- الگوریتم تخصیص منبع پیشنهادی آگاه به توان

## ۷- زمانبندی برنامه‌ی کاربردی بر روی ماشین‌های مجازی

تابع هدف ما در این مسئله زمانبندی، نگاشت وظایف بر روی ماشین‌های مجازی و ترتیب بخشیدن به اجرای آنهاست به گونه‌ای که نیازمندی‌های اولویت بین وظایف برآورده شده و کمترین زمان اجرای کل گراف حاصل شود.

فرض می‌کنیم که محیط محاسباتی شامل مجموعه‌ای از  $q$  ماشین مجازی ناهمگن می‌باشد که به صورت کامل به یکدیگر متصل می‌باشند. در این مدل فرض شده است که وظایف یک برنامه‌ی کاربردی به صورت غیرانحصاری اجرا می‌شوند.  $W$  ماتریس هزینه‌ی محاسباتی با اندازه‌ی  $n * q$  است به گونه‌ای که هر  $w_{i,j}$  زمان تخمینی از زمان تکمیل وظیفه‌ی  $n_i$  بر روی ماشین  $q_j$  را نشان می‌دهد. قبل از زمانبندی، وظایف با میانگین زمان اجرای‌شان بر روی ماشین‌های مختلف برچسب‌گذاری می‌شوند. میانگین زمان اجرای یک وظیفه‌ی  $n_i$  به صورت زیر تعریف می‌شود:

(۵)

$$\bar{w}_i = \frac{\sum_{j=1}^q w_{i,j}}{q}$$

نرخ انتقال داده بین ماشین‌های مجازی نیز در ماتریس  $B$  با اندازه‌ی  $q * q$  ذخیره می‌شود. هزینه‌ی انتقال یک یال  $(I,k)$  که برای انتقال داده از وظیفه‌ی  $n_i$

$$c_{i,k} = \frac{data_{i,k}}{B_{m,n}} \quad (۶)$$

زمانی که هم  $n_i$  و هم  $n_k$  بر روی ماشین‌های مجازی یکسان زمانبندی شوند،  $c_{i,k}$  صفر خواهد بود. قبل از زمانبندی، میانگین هزینه‌ی انتقال محاسبه شده و بر روی هر یال برچسب‌گذاری می‌شود. بنابراین میانگین هزینه‌ی انتقال یک یال  $(I, k)$  که بین دو وظیفه‌ی  $n_i$  و  $n_k$  می‌باشد به صورت زیر محاسبه می‌شود:

$$\bar{c}_{i,k} = \frac{data_{i,k}}{B} \quad (۷)$$

که  $\bar{B}$  میانگین نرخ انتقال میان ماشین‌های مجازی (میانگین پهنای باند میان ماشین‌های مجازی) می‌باشد.

قبل از ارائه‌ی تابع هدف، ابتدا مقادیر  $EST$  و  $EFT$  را تعریف می‌کنیم.  $EST(n_i, q_j)$  و  $EFT(n_i, q_j)$  به ترتیب نشان‌دهنده‌ی زودترین زمان شروع اجرا و زودترین زمان پایان اجرای وظیفه‌ی  $n_i$  بر روی ماشین مجازی  $q_j$  می‌باشد. برای یک وظیفه‌ی ورودی  $EST(n_{entry}, q_j) = 0$  می‌باشد. برای وظایف دیگر موجود در گراف، مقادیر  $EST$  و  $EFT$  به صورت بازگشتی و با شروع از وظیفه‌ی ورودی محاسبه می‌شود. برای محاسبه‌ی  $EFT$  وظیفه‌ی  $n$ ، تمامی وظایف پیشین (والد) آن می‌بایست زمانبندی شود.

$$EST(n_i, q_j) = \max\{avail[j], \max_{n_m \in pred(n_i)} (AFT(n_m) + c_{m,i})\} \quad (۸)$$

$$EFT(n_i, q_j) = w_{i,j} + EST(n_i, q_j) \quad (۹)$$

که در روابط بالا  $pred(n_i)$  مجموعه‌ی وظایف پیشین (والد) وظیفه‌ی  $n_i$  و  $avail[j]$  زودترین زمانی‌ست که ماشین  $j$  آماده‌ی اجرای  $n_i$  است. اگر  $n_k$  وظیفه‌ی آخر زمانبندی شده بر روی ماشین  $q_j$  باشد،  $avail[j]$  زمانی است که ماشین  $q_j$  اجرای وظیفه‌ی  $n_k$  را تمام کرده است و آماده است تا وظیفه‌ی دیگری را اجرا کند. بلاک  $\max$  داخل تساوی  $EST$ ، زمان آماده را برمی‌گرداند یعنی زمانی که تمامی داده‌های مورد نیاز برای اجرای وظیفه‌ی  $n_i$  به ماشین مجازی  $q_j$  رسیده است. بعد از اینکه وظیفه‌ی  $n_i$  بر روی ماشین مجازی  $q_j$  زمانبندی شد، زودترین زمان شروع و زودترین زمان پایان وظیفه‌ی  $n_i$  بر روی ماشین  $q_j$  به ترتیب برابر با زمان شروع واقعی،  $AST(n_i)$ ، و زمان پایان واقعی،  $AFT(n_i)$ ، آن وظیفه نامیده می‌شود. بعد از اینکه تمامی وظایف در گراف زمانبندی شد، طول این زمانبندی (زمان اجرای کل وظایف گراف) زمان پایان واقعی وظیفه‌ی خروجی خواهد بود. اگر چندین وظیفه‌ی خروجی وجود داشته باشد این زمان به صورت زیر محاسبه می‌شود [2]:

$$makespan = \max\{AFT(n_{exit})\} \quad (۱۰)$$

### ۷-۱- الگوریتم زمانبندی HEFT

الگوریتم زمانبندی HEFT (Heterogeneous Earliest Finish Time Algorithm) یک الگوریتم معروف زمانبندی برای کمینه ساختن زمان اجرای وظایف در یک جریان کاری می‌باشد که توسط هالوک<sup>۲۴</sup> و همکارانش معرفی شد [۲]. این الگوریتم شامل دو فاز می‌باشد: فاز اولویت‌بندی وظایف و فاز انتخاب

خروجی (آخرین وظیفه) شروع می‌کنیم، این مرتبه‌بندی، مرتبه‌بندی رو به بالا نامیده می‌شود. برای وظیفه‌ی خروجی  $n_{exit}$  این مقدار برابر است با:

$$rank_u(n_{exit}) = \overline{w_{exit}} \quad (۱۲)$$

در فاز انتخاب ماشین مجازی برای هر وظیفه  $n_i$ ، پردازنده  $p_j$  انتخاب شده و مقدار  $EFT(n_i, p_j)$  (زودترین زمان پایان) براساس رابطه‌ی (۹) محاسبه می‌شود. در نهایت وظیفه  $n_i$  به پردازنده‌ای اختصاص داده می‌شود که  $EFT$  کمتری برای آن وظیفه داشته باشد.

## ۸- ارزیابی

در این بخش به جزئیات شبیه‌سازی الگوریتم‌های معرفی شده خواهیم پرداخت. سپس از طریق رسم نمودار به ارزیابی روش‌های مذکور می‌پردازیم. محیط مورد بررسی ما یک محیط زیرساخت به عنوان سرویس است و همچنین ابزار ما برای شبیه‌سازی نرم‌افزار Workflowsim [۱۲] و [۱۳] است که توسعه‌یافته‌ی شبیه‌ساز CloudSim می‌باشد. این شبیه‌ساز، به ما اجازه‌ی ایجاد یک محیط مجازی‌سازی شده را می‌دهد و از تخصیص منابع براساس تقاضا پشتیبانی می‌کند. همچنین امکان ارسال جریان‌های کاری از طرف کاربر را فراهم می‌آورد.

### ۸-۱- مشخصات مرکز داده و جریان‌های کاری مورد استفاده

مرکز داده‌ی ما از ۱۰۰ گره فیزیکی ناهمگون و ۲۲۰ ماشین مجازی ناهمگون تشکیل شده است. مشخصات سرورهای مورد استفاده در جدول ۱ آمده است:

جدول ۱- مشخصات و تنظیمات سرورهای مورد نظر

سرور	نوع پردازنده	کارایی پردازنده MIPS	تعداد هسته	حافظه GB	پهنای باند شبکه
HP ProLiant ML110 G3	PentiumD930	۳۰۰۰	۲	۴	۱GB/s
HP ProLiant ML110 G4	Intel Xeon 3040	۱۸۶۰	۲	۴	۱GB/s
HP ProLiant ML110 G5	Intel Xeon 3075	۲۶۶۰	۲	۴	۱GB/s
IbmX3250	XeonX3470	۲۹۳۳	۴	۸	۱GB/s
IbmX3250	XeonX3480	۳۰۶۷	۴	۸	۱GB/s

ماشین‌های مجازی مورد استفاده نیز در ۴ نوع مختلف ایجاد شده‌اند، مشخصات انواع ماشین‌های مجازی به نمونه‌های Amazon EC2 مربوط می‌شود که در جدول ۲ آمده است.

جدول ۲- مشخصات و تنظیمات ماشین‌های مجازی موردنظر

نوع ماشین مجازی	کارایی پردازنده MIPS	تعداد هسته	حافظه MB	پهنای باند MB/s
High-CPU Medium	۲۵۰۰	۱	۸۷۰	۱۰۰
Extra Large	۲۰۰۰	۱	۱۷۴۰	۱۰۰
Small	۱۰۰۰	۱	۱۷۴۰	۱۰۰
Micro	۵۰۰	۱	۶۱۳	۱۰۰

ماشین مجازی. در فاز اولویت‌بندی وظایف، اولویت هر وظیفه براساس معیار B-rank به آنها داده می‌شود که این معیار در واقع نشان‌دهنده‌ی فاصله‌ی آن وظیفه تا انتهای جریان کاری می‌باشد. بعد از اینکه اولویت و مرتبه‌ی هر وظیفه که در واقع نشان‌دهنده‌ی ترتیب اجرای وظایف هستند، مشخص شد، در فاز دوم وظایف به ترتیب اولویت‌شان برای اجرا بر روی منابع مجازی قرار می‌گیرند و هر وظیفه بر روی منبعی قرار می‌گیرد که زمان پایان کمتری را نسبت به منابع مجازی دیگر برای آن وظیفه فراهم کند.

```

Require: W=(A,D),      ▷ Workflow Application
Require:                ▷ Set of resources
Ensure: sched w={ (Ai, sched(Ai)) | ∀Ai ∈ A }
                        ▷ Workflow Schedule

1: Function HEFT (W, R)
2: Rank ← B - Rank (A)  ▷ Order the tasks
                        according to B - Rank
3: sched w ← φ        ▷ Initialize workflow schedule
                        with empty set
4: for i ← 1,n do      ▷ Iterate over the ranked tasks
5:   Tmin ← ∞
6:   for j ← 1,m do    ▷ Iterate over all resources
7:     TRanki ← maxAp ∈ pred(Ranki) { TAp + t(Ranki, Rj) }
                        ▷ Compute completion time of Ranki
8:     if TRanki < Tmin then ▷ Save the minimum
                        completion time
9:       Tmin ← TRanki
10:      Rmin ← Rj
11:    end if
12:  end for
13:  sched w ← sched w ∪ (Ranki, Rmin)
                        ▷ Schedule the task
14: end for
15: return sched w
16: end function
    
```

شکل ۴- الگوریتم HEFT [۱۱]

در این الگوریتم ابتدا هزینه‌های محاسباتی وظایف و هزینه‌های انتقال یال‌ها به ترتیب براساس روابط (۵) و (۷) محاسبه می‌شوند. سپس تمامی وظایف جریان کاری براساس معیار B-rank اولویت‌بندی می‌شوند. در این اولویت‌بندی، اولویت هر وظیفه براساس رابطه‌ی زیر محاسبه می‌شود که این رتبه‌بندی براساس هزینه‌های محاسباتی و انتقال آن وظیفه مشخص می‌شود. سپس با مرتب کردن وظایف براساس مرتبه‌یشان به صورت نزولی لیستی از وظایف تولید می‌شود.

$$rank_u(n_i) = \overline{w_i} + \max_{n_j \in succ(n_i)} (\overline{c_{i,j}} + rank_u(n_j)) \quad (۱۱)$$

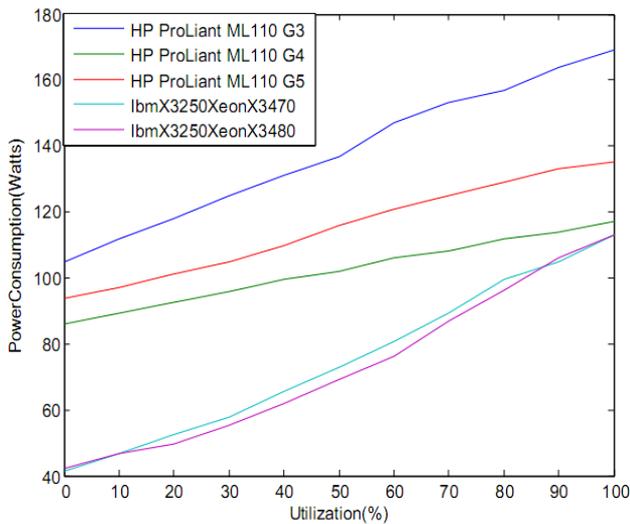
که در این رابطه  $succ(n_i)$  مجموعه‌ای از وظایفی هستند که بلافاصله بعد از وظیفه‌ی  $n_i$  اجرا می‌شوند.  $\overline{c_{i,j}}$  میانگین هزینه‌ی انتقال یال  $(i,j)$  و  $\overline{w_i}$  میانگین هزینه‌ی محاسباتی وظیفه‌ی  $n_i$  می‌باشد. به دلیل اینکه این اولویت به صورت بازگشتی و با پیمایش گراف به سمت بالا محاسبه می‌شود و ابتدا از وظیفه‌ی

برای جایگذاری هر ماشین مجازی، میزبانی را انتخاب می‌کند که کمترین افزایش توان مصرفی را داشته باشد. (در پیوست این الگوریتم معرفی شده است).

**الگوریتم تعدیل بار:** این الگوریتم که در بخش ۶ مقاله نیز معرفی شد، هر ماشین مجازی را بر روی میزبانی قرار می‌دهد که نسبت به سایر میزبان‌ها تعداد کمتری پردازنده‌ی در حال استفاده داشته باشد به عبارت دیگر بهره‌وری کمتری داشته باشد.

الگوریتم پیشنهادی آگاه به توان ما میزبانی را انتخاب می‌کند که بیشترین مقدار  $G_{ij}$  را داشته باشد و الگوریتم PABFD ترجیح می‌دهد که یک ماشین مجازی جدید را به میزبانی تخصیص دهد که کمترین افزایش در توان مصرفی را داشته باشد. برای مثال مجموعه‌ای از ۱۲ ماشین مجازی (که هر یک به ۱ هسته نیاز دارد) داده شده است. الگوریتم PABFD براساس شکل ۵ که روند افزایش توان مصرفی را نشان می‌دهد، ۶ عدد سرور نوع HP ProLiant ML110 G4 را استفاده می‌کند که هر یک دارای ۲ هسته است به دلیل اینکه با افزایش بهره‌وری، کمترین میزان افزایش توان را در بین دیگر میزبان‌ها دارد.

در مقابل الگوریتم پیشنهادی ما از ۳ سرور نوع IbmX3250XeonX3480 استفاده می‌کند که هر یک دارای ۴ هسته است چرا که بیشترین میزان MIPS و کمترین توان مصرفی در بهره‌وری ۱۰۰ درصد را دارد. بنابراین چنین انتخابی می‌تواند هم تعداد سرورهای مورد استفاده و هم توان مصرفی مرکز داده را کاهش دهد. الگوریتم تعدیل بار نیز یک ماشین مجازی جدید را به میزبانی تخصیص می‌دهد که بهره‌وری کمتری داشته باشد و این امر باعث می‌شود که تعداد میزبان‌های مورد استفاده افزایش یافته و توان مصرفی مرکز داده افزایش یابد.



شکل ۵- میزان توان مصرفی سرورها بر حسب وات در سطوح متفاوت بهره‌وری

براساس نتایج شبیه‌سازی شکل ۷ که بر روی ۴ نوع گراف متفاوت با تعداد وظایف مختلف صورت گرفته است مشاهده می‌کنیم که الگوریتم آگاه به توان معرفی شده نسبت به الگوریتم تعدیل بار، در تمامی حالات کاهش چشمگیری در میزان توان مصرفی مرکز داده داشته است. که البته این میزان کاهش بسته به نوع گراف و تعداد وظایف گراف متفاوت است. در حقیقت در تمامی الگوریتم‌های تخصیص منبع مورد استفاده، با به کار بردن مثلا ۵ ماشین مجازی زمان اجرا به کمترین مقدار خود می‌رسد ولی با این تفاوت که بسته به مکان ماشین مجازی، تعداد سرورهای مورد نیاز متفاوت هستند.

ما در این شبیه‌سازی پهنای باند میان ماشین‌های مجازی را یکسان در نظر گرفته‌ایم. به منظور محاسبه‌ی توان مصرفی سرورها، از داده‌های واقعی توان مصرفی که از نتایج استفاده از SPEC power benchmark فراهم شده است استفاده می‌کنیم. مقادیر درون جدول ۴ در قسمت پیوست، میزان توان مصرفی هر یک از سرورها را به ازای افزایش ۱۰ درصدی میزان بهره‌وری نشان می‌دهد.

ما از ۴ نوع گراف یا جریان کاری متفاوت برای نشان دادن برنامه‌ی کاربردی ارسال شده از طرف کاربر استفاده می‌کنیم. که هر یک از این گراف‌ها با یک جریان کاری از برنامه‌ی کاربردی در دنیای واقعی مطابقت می‌کند. این جریان‌های کاری عبارتند از جریان کاری Montage با تعداد ۲۵، ۵۰، ۱۰۰ و ۱۰۰۰ وظیفه، Sipht با تعداد ۳۰، ۶۰، ۱۰۰ و ۱۰۰۰ وظیفه، CyberShake با تعداد ۳۰، ۵۰، ۱۰۰ و ۱۰۰۰ وظیفه و Inspiral با تعداد ۳۰، ۵۰، ۱۰۰ و ۱۰۰۰ وظیفه [۱۴].

## ۸-۲- شبیه‌سازی و ارزیابی الگوریتم HEFT

در این قسمت به مقایسه و ارزیابی زمان تکمیل وظایف جریان‌های کاری در الگوریتم HEFT با الگوریتم‌های MCT، MAX-MIN و MIN-MIN [۱۵] می‌پردازیم. الگوریتم MCT [۱۵] جز الگوریتم‌های حریم‌های معروفی است که برای زمانبندی پویای وظایف به کار می‌رود. در این الگوریتم بر خلاف الگوریتم HEFT اولویت بین وظایف مورد بررسی قرار نمی‌گیرد و هر وظیفه با ترتیبی دلخواه، بر روی منبعی قرار می‌گیرد که نسبت به سایر منابع زمان تکمیل کمتری را برای آن وظیفه داشته باشد.

این امر باعث می‌شود برخی از وظایف بر روی ماشینی قرار بگیرند که الزاما کمترین زمان اجرا را به همراه نداشته باشند. الگوریتم HEFT یک الگوریتم معروف زمانبندی به منظور کمینه ساختن زمان اجرای وظایف در یک جریان کاری می‌باشد. همانطور که نتایج شبیه‌سازی شکل ۶ نشان می‌دهند الگوریتم HEFT به‌طور میانگین نسبت به سه الگوریتم زمانبندی دیگر کاهش زمان تکمیل جریان کاری را داشته است که این کاهش زمان در ۴ جریان کاری مورد استفاده ما متفاوت است. این میزان کاهش در جدول ۳ نشان داده شده است. به همین منظور ما برای زمانبندی جریان‌های کاری ورودی از الگوریتم HEFT استفاده نموده‌ایم.

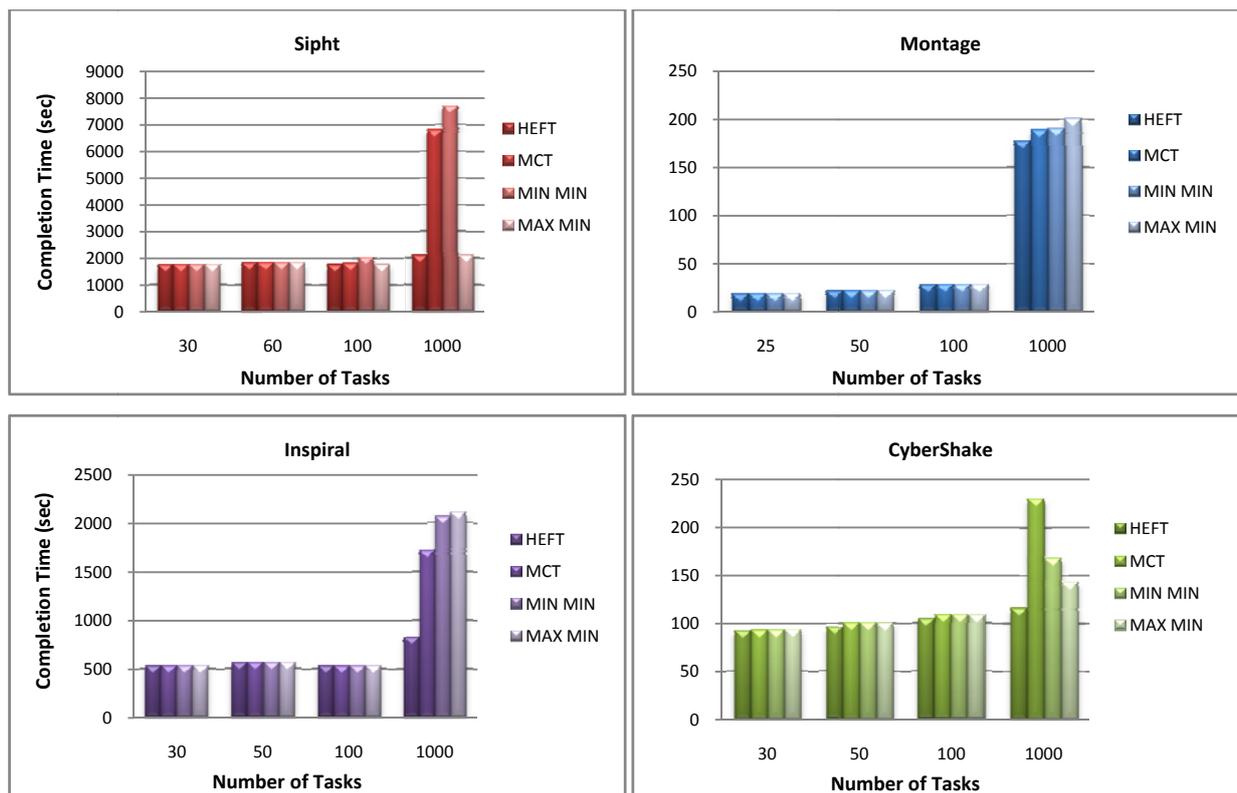
جدول ۳- درصد کاهش زمان اجرای جریان‌های کاری در الگوریتم HEFT نسبت به الگوریتم‌های زمانبندی دیگر

جریان کاری / الگوریتم زمانبندی	MCT	MAX MIN	MIN MIN
Montage	۰.۰۳	۶۰.۰	۵۰.۰۳
Sipht	۱۱.۸۲	۹۰.۰۰۲	۱۴.۶۷
Cybershake	۱.۲۳	۶۰.۳	۰.۶۱
Inspirial	۲.۲۳	۲۳.۲	۱۳.

## ۸-۳- شبیه‌سازی و ارزیابی الگوریتم پیشنهادی آگاه به توان

ما در این شبیه‌سازی از ورژن WorkflowSim-1.0-master برای مدل‌سازی و شبیه‌سازی مرکز دادیمان استفاده کردیم. در اینجا الگوریتم آگاه به توان معرفی شده خود را با الگوریتم‌های تخصیص منبع زیر مقایسه می‌کنیم:

**الگوریتم PABFD:** این الگوریتم که در مرجع [۱۶] معرفی شده است، تمام ماشین‌های مجازی را براساس EST (زودترین زمان شروع) آنها مرتب می‌کند و



شکل ۶- نمودار زمان تکمیل وظایف (ثانیه) در جریان‌های کاری مختلف

راست) را در سه الگوریتم تخصیص منبع تعدیل بار، تخصیص منبع PABFD و تخصیص منبع آگاه به توان، نشان می‌دهد. براساس این نتایج، میانگین بهره‌وری منابع فیزیکی در الگوریتم پیشنهادی آگاه به توان نسبت به دو الگوریتم دیگر افزایش چشمگیری داشته است و این مقدار در الگوریتم تعدیل بار از همه کمتر است. علاوه بر این در الگوریتم آگاه به توان تعداد منابع فیزیکی مورد استفاده برای میزبانی ماشین‌های مجازی به منظور رسیدن به حداقل زمان اجرای ممکن، بسیار کمتر شده است و همین امر موجب شده تا بهره‌وری منابع فیزیکی افزایش یابد. در نتیجه با بالا رفتن بهره‌وری منابع فیزیکی و کاهش تعداد منابع فیزیکی مورد استفاده برای اجرای برنامه‌های کاربردی، میزان توان مصرفی مرکز داده کاهش چشمگیری پیدا کرده است.

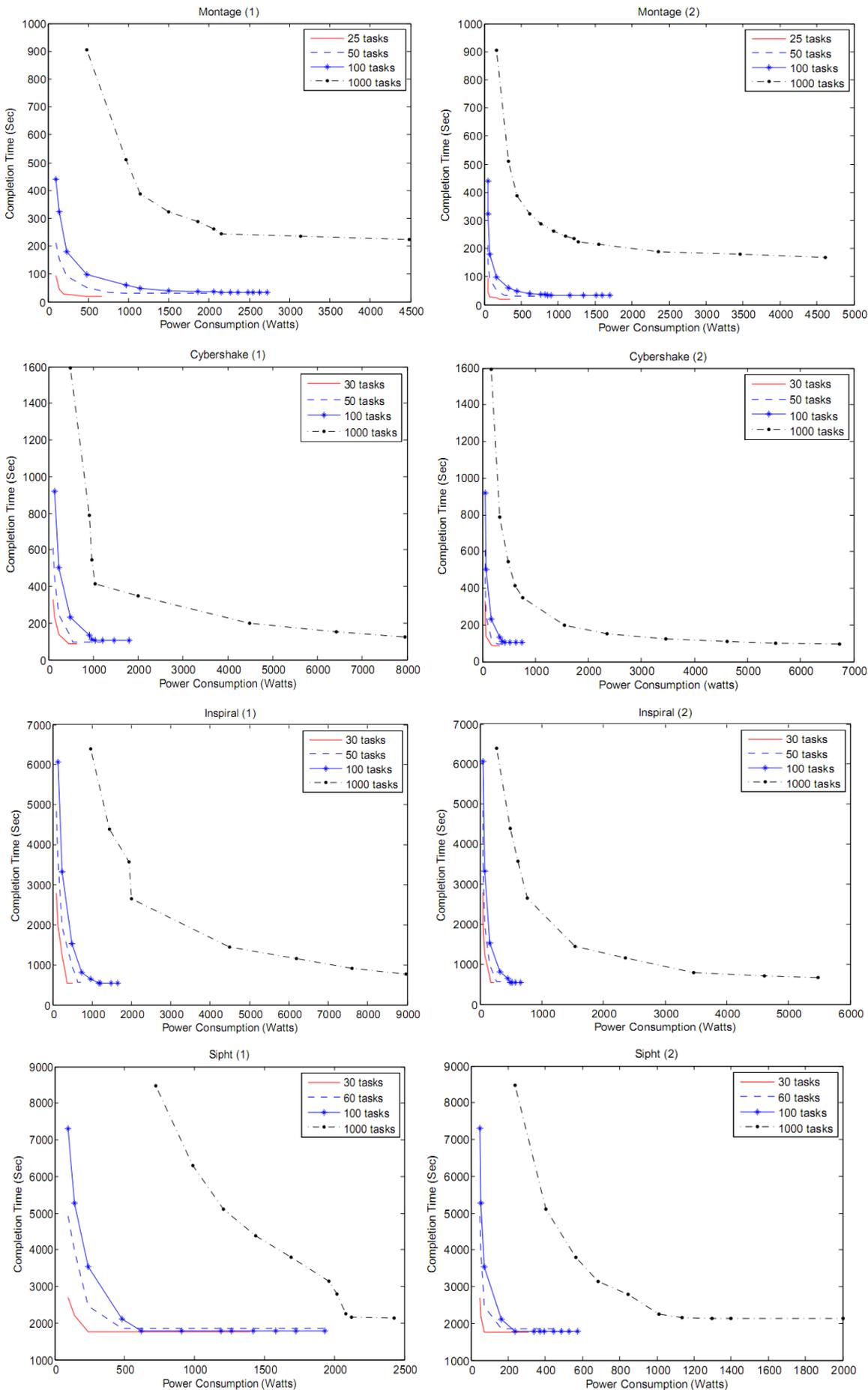
همانطور که در نمودار شکل ۷ مشاهده می‌شود، در هر دو الگوریتم، زمان تکمیل وظایف یکسان است، چرا که برای زمانبندی وظایف، در هر دو مورد از الگوریتم HEFT استفاده کرده‌ایم. به علاوه براساس رابطه ۷، برای محاسبه‌ی زمان انتقال وظایف بین ماشین‌های مجازی، میانگین پهنای باند ماشین‌های مجازی محاسبه می‌شود و ما در این شبیه‌سازی، پهنای باند میان ماشین‌های مجازی را یکسان در نظر گرفته‌ایم.

از طرفی هر ماشین مجازی MIPS مخصوص خود را دارد و در الگوریتم HEFT، انتخاب ماشین مجازی برای زمانبندی وظایف (همانطور که در بخش ۷ مطرح شد) به الگوریتم تخصیص ماشین‌های مجازی بر روی میزبان‌ها بستگی ندارد، بنابراین الگوریتم‌های تخصیص متفاوت که در سطح زیرساخت به کار می‌روند تاثیری بر زمان تکمیل جریان‌های کاری ندارند.

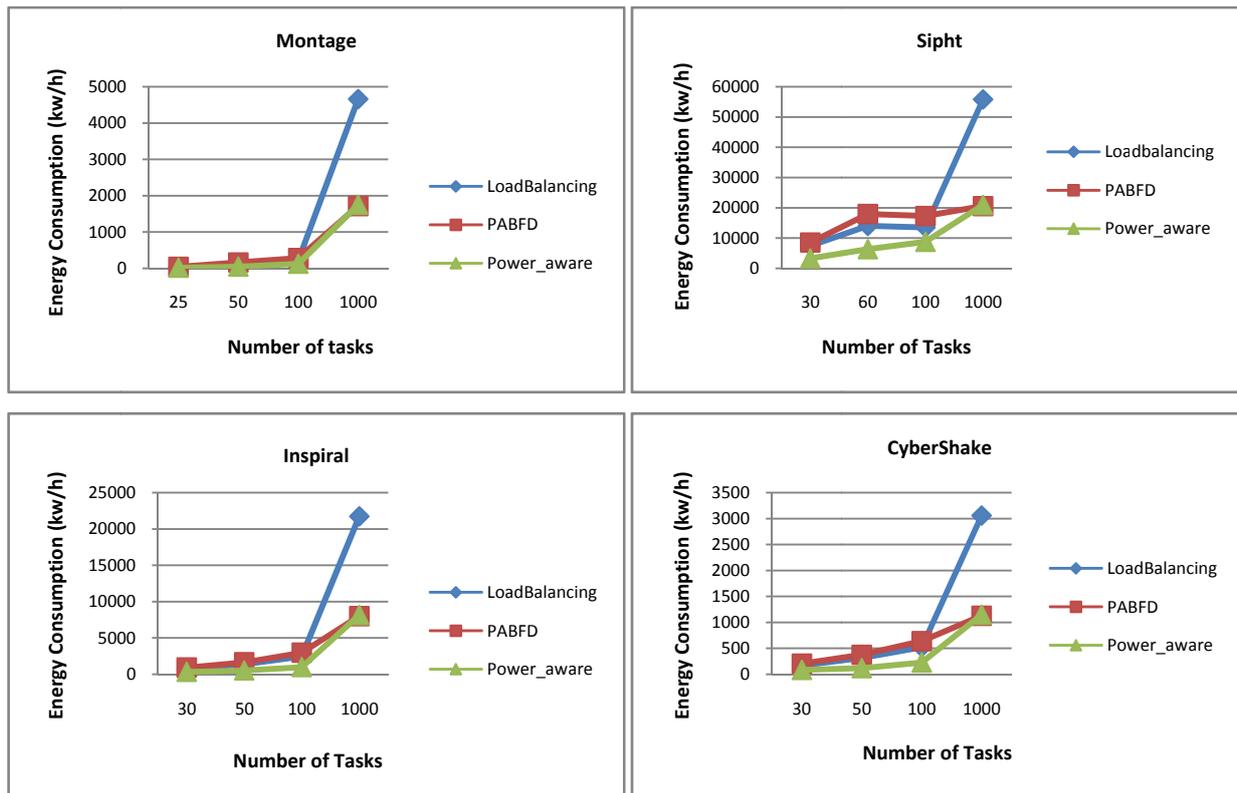
نمودارهای شکل ۸ میزان انرژی مصرفی مرکز داده را برای ۴ نوع جریان کاری مختلف با سه الگوریتم تخصیص منبع تعدیل بار، تخصیص منبع PABFD و تخصیص منبع آگاه به توان نشان می‌دهد. با توجه به نتایج حاصل از شبیه‌سازی که در شکل ۸ مشاهده می‌شود، می‌توان دریافت که با افزایش تعداد وظایف یک گراف، تعداد ماشین‌های مجازی مورد نیاز برای اجرای وظایف افزایش یافته و این امر موجب افزایش تعداد میزبان‌های مشغول به کار در مرکز داده شده است که در نتیجه‌ی آن میزان انرژی مصرفی مرکز داده افزایش یافته است.

همانطور که مشاهده می‌شود به‌طور میانگین این افزایش انرژی در الگوریتم آگاه به توان کمتر می‌باشد. چرا که با افزایش تعداد ماشین‌های مجازی مورد نیاز، تعداد میزبان‌های مشغول افزایش چشمگیری پیدا نمی‌کنند. بدلیل اینکه ما سعی کرده‌ایم که آنها را بر روی تعداد کمتری از میزبان‌ها جاسازی کنیم.

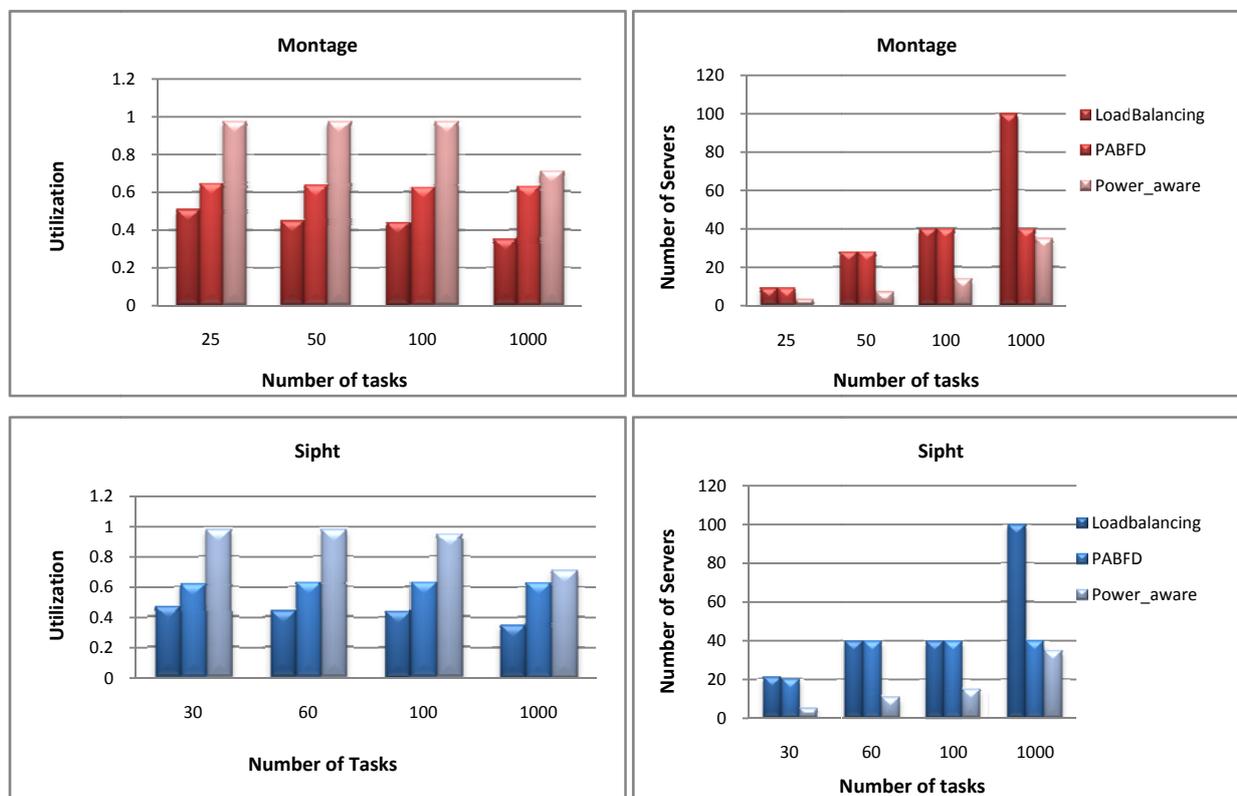
نمودارهای شکل ۹ میانگین میزان بهره‌وری میزبان‌های فعال (سمت چپ) و تعداد منابع فیزیکی مورد استفاده برای اجرای جریان‌های کاری مختلف (سمت



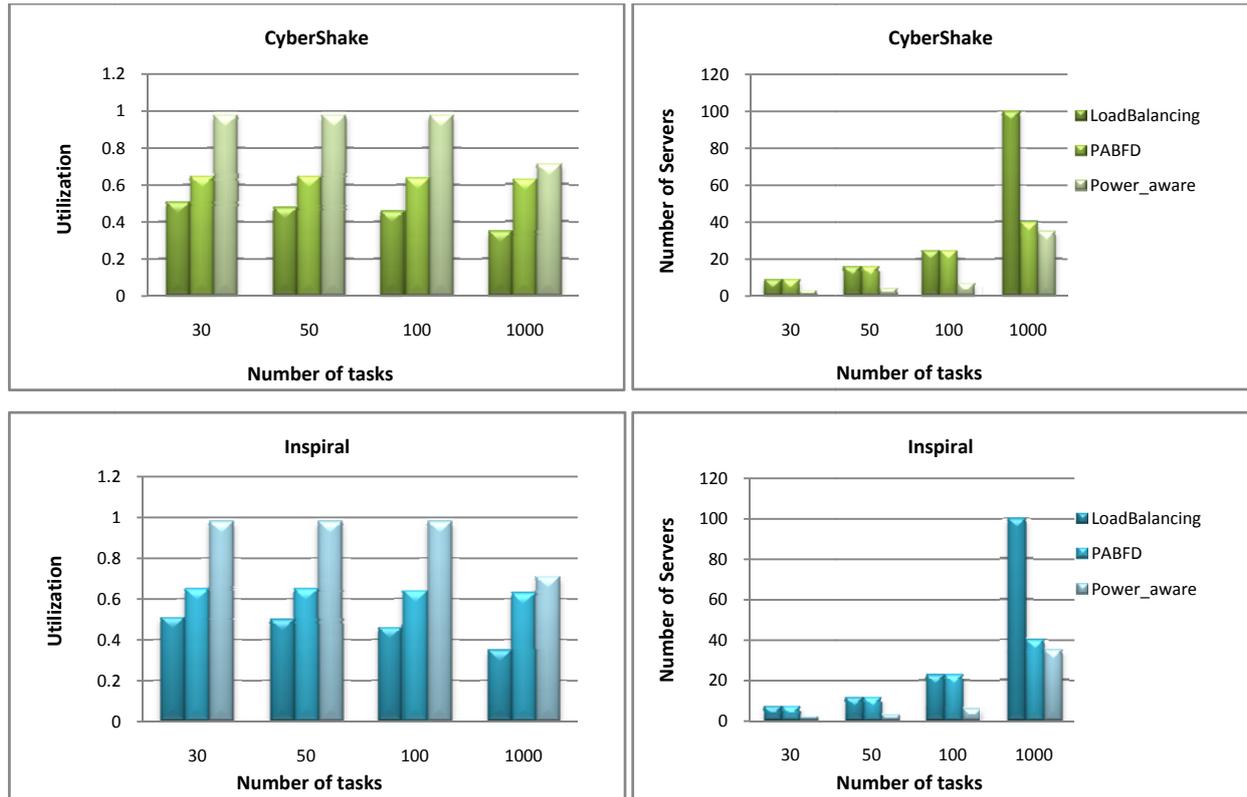
شکل ۷- نمودار زمان تکمیل وظایف (ثانیه) در مقابل توان مصرفی مرکز داده (وات) در جریان‌های کاری مختلف در الگوریتم تخصیص منبع تعدیل بار (۱) و الگوریتم تخصیص منبع پیشنهادی آگاه به توان (۲)



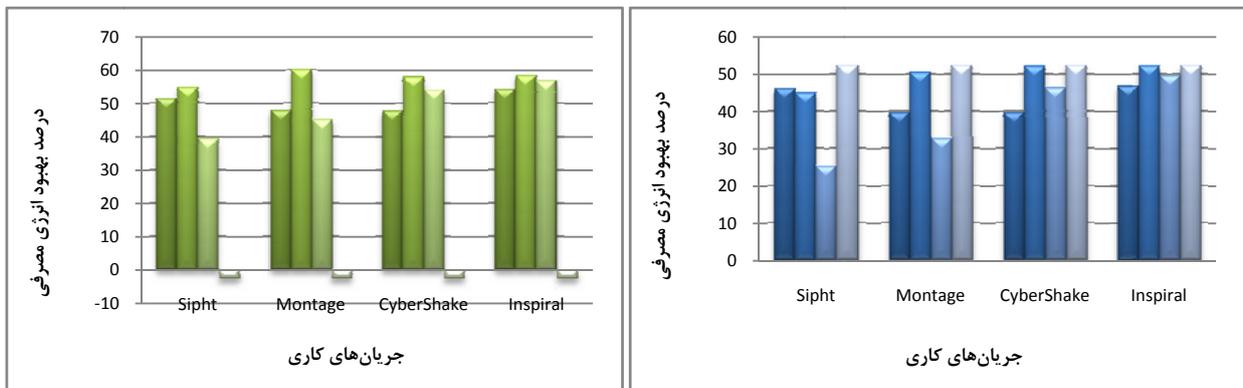
شکل ۸- نمودار انرژی مصرفی مرکز داده (کیلو وات بر ساعت) با افزایش تعداد وظایف جریان کاری در سه الگوریتم تخصیص متفاوت



شکل ۹- میانگین بهره‌وری میزبان‌ها (سمت چپ) و تعداد میزبان‌های فعال (سمت راست) با افزایش تعداد وظایف جریان‌های کاری



ادامه شکل ۹- میانگین بهره‌وری میزبان‌ها (سمت چپ) و تعداد میزبان‌های فعال (سمت راست) با افزایش تعداد وظایف جریان‌های کاری



شکل ۱۰- درصد بهبود انرژی مصرفی الگوریتم پیشنهادی در مقایسه با الگوریتم تعدیل بار (سمت راست) و الگوریتم PABFD (سمت چپ)

مقابل توان مصرفی مرکز داده است، می‌توان دریافت که نوع گراف ورودی و تعداد وظایف آن در میزان توان مصرفی مرکز داده تاثیرگذار است.

براساس نمودار شکل ۸ که نمایان‌گر درصد بهبود انرژی مصرفی الگوریتم پیشنهادی در مقایسه با الگوریتم تعدیل بار و PABFD است، می‌توان دریافت که نوع گراف ورودی در میزان انرژی مصرفی مرکز داده تاثیرگذار بوده و الگوریتم پیشنهادی ما در مقایسه با الگوریتم تعدیل بار، حداقل برابر با ۲۵.۳۱٪ و حداکثر برابر با ۵۲.۴۵٪ کاهش انرژی را داشته و در مقایسه با الگوریتم PABFD

## ۹- نتیجه‌گیری

با توجه به نتایج حاصل از شبیه‌سازی می‌توان گفت که الگوریتم زمانبندی ایستای HEFT در مقایسه با دیگر الگوریتم‌های زمانبندی معرفی شده بهترین نتیجه را داشته است و براساس شکل ۶ مشاهده می‌شود که با افزایش تعداد وظایف گراف‌های ورودی، این الگوریتم نسبت به دیگر الگوریتم‌ها بهتر عمل می‌کند. همچنین با توجه به نمودارهای شکل ۷ که نشان‌دهنده‌ی زمان تکمیل وظایف در

*International Conference, Bangalore, India, vol. 4297, pp. 353-362, December 2006.*

[11] J. Durillo, H. Mohammadi Fard, and R. Prodan, "MOHEFT: A Multi-Objective List-based Method for Workflow Scheduling," *2012 IEEE 4th International Conference on Cloud Computing Technology and Science, 2012.*

[12] W. Chen, and E. Deelman, "WorkflowSim: A toolkit for simulating scientific workflows in distributed environments," *IEEE Xplore: 11 January 2013.*

[13] P. S. Rawat, P. Dimri, and G. P. Saroha, "Tasks Scheduling in Cloud Computing Environment using Workflowsim," *Research Journal of Information Technology, 8: 98-104, October 2016.*

[14] "WorkflowGenerator," <https://confluence.pegasus.isi.edu/display/pegasus/WorkflowGenerator>, September 2015.

[15] F. Soleimani, M. Ahadi, R. Habibpour, and A. Kamalinia, "Analysis of Scheduling Algorithms in Grid Computing Environment," *International Journal of Innovation and Applied Studies ISSN 2028-9324 vol. 4, no. 3, Nov. 2013, pp. 560-567.*

[16] A. Beloglazov, and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers," *Journal Concurrency and Computation: Practice and Experience vol. 24, pp. 1397-1420, September 2012.*

## پیوست

### الگوریتم PABFD

این الگوریتم، تمامی ماشین‌های مجازی را براساس میزان بهره‌وری پردازنده‌ی آنها به صورت نزولی مرتب می‌کنیم و هر ماشین‌های مجازی را به میزبانی تخصیص می‌دهیم که کمترین افزایش توان مصرفی را به ازای آن تخصیص داشته باشیم. شبه کد زیر الگوریتم PABFD را نشان می‌دهد.

Algorithm Power Aware Best Fit Decreasing (PABFD)

```

1. Input: hostlist, vmlist Output: allocation of VMs
2. vmlist.SortDecreasingUtilization
3. foreach vm in vmlist do
4.   minpower ← MAX
5.   allocatedhost ← NULL
6.   foreach host in hostlist do
7.     if host has enough resources for vm then
8.       power ← estimatepower(host, vm)
9.       if power < minpower then
10.        allocatedhost ← host
11.        minpower ← power
12.   if allocatedhost ≠ NULL then
13.     allocation.add(vm, allocatedhost)
14. return allocation

```

حداقل ۲.۸۵٪ افزایش انرژی و حداکثر ۰.۲٪ کاهش انرژی را به همراه داشته است.

بنابراین به‌طور میانگین الگوریتم پیشنهادی ما در مقایسه با الگوریتم تعدیل بار، ۰.۲٪ و در مقایسه با الگوریتم PABFD، ۳۶.۳۳٪ کاهش انرژی را داشته است.

## مراجع

[1] M. Abu Sharkh, M. Jammal, A. Shami, and A. Ouda, "Resource Allocation in a Network-Based Cloud Computing Environment: Design Challenges," in *Proceeding IEEE Communications Magazine*, pp. 46-52, November 2013.

[2] H. Topcuoglu, S. Hariri, and Min-You Wu, "Performance-effective and low-complexity task scheduling for heterogeneous computing. Parallel and Distributed Systems," *IEEE Transactions on, 13(3): 260-274, march 2002.*

[3] P. Akilandeswari, and H. Srimathi, "Survey and Analysis on Task Scheduling in Cloud Environment," *Indian Journal of Science and Technology, vol. 9(37), Issue 37, October 2016.*

[4] S. Shruthi, and V. Nagaveni, "A Survey on Various Parallel Power Aware Task Scheduling Algorithms for Reducing Power Consumption," *International Journal on Recent and Innovation Trends in Computing and Communication, vol. 2, Issue: 4, April 2014.*

[5] L. Ma, Y. Lu, F. Zhang, and S. Sun, "Dynamic Task Scheduling in Cloud Computing Based on Greedy Strategy," *Trustworthy Computing and Services Communication in Computer and Information Science, vol. 320, pp. 156-162, 2013.*

[۶] ه. رضایی و ق. حاجین، "مدیریت منابع مجازی‌سازی شده در محیط‌های رایانش ابر،" پایان‌نامه جهت اخذ درجه‌ی کارشناسی‌ارشد، دانشگاه علم و صنعت ایران، دانشکده مهندسی کامپیوتر، بهمن ۱۳۹۰.

[7] A. Beloglazov, and R. Buyya, "Energy Efficient Allocation of Virtual Machines in Cloud Data Centers," *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, pp. 577-578, 2010.*

[۸] م. نجف‌زاده، "مقیاس‌پذیری خودمختار محیط‌های توزیعی مجازی براساس توافق‌های سطح سرویس،" دانشگاه علم و صنعت ایران، دانشکده مهندسی کامپیوتر، پایان‌نامه کارشناسی‌ارشد، ۱۳۸۹.

[9] X. Fan, W. D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," *Proceedings of the 34th annual international symposium on Computer architecture, vol. 35, Issue 2, pp. 13-23, May 2007.*

[10] S. Baskiyar, and K. Kumar Palli, "Low Power Scheduling of DAGs to Minimize Finish Times," *13th*

جدول ۴- میزان توان مصرفی سرورها بر حسب وات در سطوح متفاوت بهره‌وری

سرور/سطوح متفاوت بهره‌وری	%۰	%۱۰	%۲۰	%۳۰	%۴۰	%۵۰	%۶۰	%۷۰	%۸۰	%۹۰	%۱۰۰
HP ProLiant ML110 G3	۱۰۵	۱۱۲	۱۱۸	۱۲۵	۱۳۱	۱۳۷	۱۴۷	۱۵۳	۱۵۷	۱۶۴	۱۶۹
HP ProLiant ML110 G4	۸۶	۸۹.۴	۹۲.۶	۹۶	۹۹.۵	۱۰۲	۱۰۶	۱۰۸	۱۱۲	۱۱۴	۱۱۷
HP ProLiant ML110 G5	۹۳.۷	۹۷	۱۰۱	۱۰۵	۱۱۰	۱۱۶	۱۲۱	۱۲۵	۱۲۹	۱۳۳	۱۳۵
IbmX3250XeonX3470	۴۱.۶	۴۶.۷	۵۲.۳	۵۷.۹	۶۵.۴	۷۳	۸۰.۷	۸۹.۵	۹۹.۶	۱۰۵	۱۱۳
IbmX3250XeonX3480	۴۲.۳	۴۶.۷	۴۹.۷	۵۵.۴	۶۱.۸	۶۹.۳	۷۶.۱	۸۷	۹۶.۱	۱۰۶	۱۱۳

## اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۰۸/۰۱

تاریخ اصلاح: ۱۳۹۶/۰۱/۲۶

تاریخ قبول شدن: ۱۳۹۷/۰۲/۰۵

نویسنده مرتبط: عاطفه یکتا اول، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران.

**عاطفه یکتا اول** متولد سال ۱۳۶۹، مدرک کارشناسی

خود را در رشته‌ی مهندسی کامپیوتر گرایش نرم‌افزار در

سال ۱۳۹۱ از دانشگاه علم و صنعت ایران گرفته است و

سپس در سال ۱۳۹۳ مدرک کارشناسی ارشد را در رشته

مهندسی فناوری اطلاعات گرایش شبکه‌های کامپیوتری از دانشگاه علم و صنعت

ایران دریافت نمود. زمینه‌های تحقیقاتی مورد علاقه ایشان شبکه‌های کامپیوتری،

سیستم‌های توزیع شده، رایانش ابری و امنیت شبکه می‌باشد.

آدرس پست الکترونیکی ایشان عبارت است از:

atefe.yekta@gmail.com



**محمود فتحی** مدرک کارشناسی خود را در رشته‌ی

مهندسی برق - الکترونیک در سال ۱۳۶۳ از دانشگاه علم و

صنعت ایران گرفته است، و سپس در سال ۱۳۶۶ مدرک

کارشناسی ارشد را در گرایش معماری سیستم‌های

کامپیوتری از دانشگاه برادفورد انگلستان، برادفورد غربی و

دکتر را از دانشگاه یومیست منچستر انگلستان - موسسه‌ی

علوم و فناوری در سال ۱۳۷۰ در حوزه‌ی معماری کامپیوتر سیستم‌های پردازش

تصویر دریافت کرده است. از سال ۱۳۷۰ تا سال ۱۳۹۶ وی به عنوان هیئت علمی

در دانشکده کامپیوتر دانشگاه علم و صنعت ایران مشغول به آموزش و پژوهش بوده

و در حال حاضر بازنشسته دانشکده‌ی مهندسی کامپیوتر می‌باشد. علامه‌مندی‌های

تحقیقاتی وی شامل: مدلسازی سخت‌افزار، شبکه‌های کامپیوتری (کیفیت سرویس،

شبکه‌های خودرویی، همپوشان، نظارت سلامت (محاسبات زیستی و زیست

انفورماتیک)، سیستم‌های حمل و نقل هوشمند، پردازش توزیع شده‌ی ویدیو و

تصویر، رایانش موازی، توزیع شده و ابری، فناوری اطلاعات سبز است. وی دارای

تالیفات متعددی در زمینه کتاب (۷ کتاب)، مقالات مجلات (۱۳۲ مقاله)، مقالات

کنفرانس‌های بین‌المللی (۱۷۶ مقاله) و مقالات کنفرانس‌های بین‌المللی و ایرانی

(۱۸۴ مقاله) می‌باشد.

آدرس پست الکترونیکی ایشان عبارت است از:

mahfathy@iust.ac.ir

<sup>1</sup>Data Center<sup>2</sup>Scientific Workflow<sup>3</sup>Application<sup>4</sup>Utilization<sup>5</sup>Workload<sup>6</sup>Directed Acyclic Graph<sup>7</sup>Nodes<sup>8</sup>Edges<sup>9</sup>Static Task Scheduling<sup>10</sup>Dynamic Task Scheduling<sup>11</sup>Non-Preemptive<sup>12</sup>Task Duplication Heuristic<sup>13</sup>Guided Random Search Base<sup>14</sup>Dynamic Greedy Scheduling<sup>15</sup>Greedy<sup>16</sup>Dynamic Voltage Frequency Scaling (DVFS)<sup>17</sup>Multi-Threading<sup>18</sup>Migration<sup>19</sup>Liza Wang<sup>20</sup>Earliest Task First<sup>21</sup>Control Flow<sup>22</sup>Data Flow<sup>23</sup>Parent<sup>24</sup>Haluk Topcuoglu

نسخه نهائی مقالات ارسالی برای چاپ در نشریه "علوم رایانش و فناوری اطلاعات" باید بر طبق اصول مطرح شده در این راهنما تهیه شده باشد. رعایت این اصول در نسخه اولیه نیز قویاً توصیه می‌شود. مقالات به زبان فارسی ارسال گردد.

### ۱- ساختار مقاله

- عنوان: کوتاه و معرف محتوای مقاله باشد و از ۱۵ کلمه تجاوز نکند.
- نام نویسندگان و مؤسسه محل اشتغال آنان: از ذکر عناوین خودداری شود.
- چکیده فارسی: حاوی تعریف مسأله، روش حل، و نتایج مهم باشد و از ۱۵۰ کلمه تجاوز نکند.
- واژه‌های کلیدی: حداکثر ۱۰ کلمه
- بدنه اصلی مقاله: بدنه اصلی با "مقدمه" شروع و با "نتیجه‌گیری" خاتمه می‌یابد. بخش‌ها و زیربخش‌های بدنه اصلی باید شماره‌گذاری شوند. شماره "مقدمه" یک خواهد بود.
- تشکر و قدردانی (در صورت نیاز).
- مراجع: مراجع به ترتیبی که در متن به آنها رجوع می‌شود آورده شوند. نام مؤلفان مراجع در صورت لزوم در متن بصورت فارسی آورده شود. رجوع به مراجع با ذکر شماره آنها در داخل کروشه ([ ]) انجام پذیرد.
- پیوست‌ها (در صورت نیاز)
- واژه‌نامه (در صورت نیاز)
- برای مقالات فارسی، عنوان مقاله، نام نویسندگان، مؤسسه محل اشتغال، چکیده، و کلمات کلیدی به زبان انگلیسی نیز در صفحه‌ای جداگانه داده شود.
- بیوگرافی کامل نویسندگان به زبان فارسی به همراه عکس

### ۲- معادله‌ها، شکل‌ها، جدول‌ها، و عکس‌ها

- معادله‌ها باید با فاصله کافی از بالا و پائین تایپ و به صورت متوالی شماره‌گذاری شوند. شماره معادله در پرانتز در انتهای سمت راست سطر حاوی معادله قرار داده شود. معادلات دستنویس به هیچ شکل قابل قبول نیستند.
- شکل‌ها و جدول‌ها باید دارای شماره و عنوان باشند. در شکل‌ها شماره و عنوان در زیر شکل و در جدول‌ها در بالای شکل قرار می‌گیرد. اعداد و متون روی شکل‌ها و جدول‌ها باید دارای اندازه مناسب و کاملاً خوانا باشند.
- اعداد و کلمات روی شکل‌ها و جدول‌ها در مقالات فارسی به زبان فارسی باشند.
- عکس‌ها سیاه و سفید، برقی، و با کیفیت عالی باشند.

### ۳- نحوه نگارش مراجع

در لیست مراجع انواع مختلف مرجع‌ها به شکل زیر نوشته شوند:

- [۱] ب. مقدم، ا. تقوی، و ن. طاهری، آشنائی با شبکه‌های کامپیوتری، چاپ دوم، انتشارات نصر، تهران، ۱۳۷۵.
- [۲] ی. براون، مقدمه‌ای بر شبکه‌های عصبی، ترجمه م. ع. آرام، انتشارات فجر، مشهد، ۱۳۷۰.
- [۳] راهنمای کاربران حسابر، شرکت پردازش رایانه‌ای ایران، تهران، ۱۳۶۵.

[۴] ج. عارف، استنتاج فازی بوسیله شبکه‌های عصبی، پایان‌نامه کارشناسی ارشد هوش مصنوعی، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شهر، ۱۳۷۴.

[۵] ج. حسینی، و. ح. ربانی، "تشخیص چهره انسان در تصویر"، نشریه امیرکبیر، سال هشتم، شماره ۴۲، ص ۱۲۵-۱۴۷، ۱۳۷۷.

[۶] ج. حسینی، و. ح. ربانی، "تشخیص چهره انسان در تصویر"، در مجموعه مقالات هفتمین کنفرانس سالانه انجمن کامپیوتر ایران، ص ۲۲۴-۲۳۲، ۱۳۸۰.

[7] M. A. Ahmadi, and M. H. Rahimi, *Fuzzy Set Theory*, New Jersey: Prentice-Hall, 1995.

[8] M. A. Ahmadi, M. H. Rahimi, and A. Fatemi, "Evidence-Based Recognition of 3D Objects," *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 12, no. 10, pp. 811-835, 1994.

[9] A. Taheri, "On-Line Fingerprint Verification," *Proc. IEEE Intl Conf. Pattern Recognition*, pp. 752-758, 1992.

[10] M. A. Ahmadi, *On-line Fingerprint Verification*, Ph. D. Dissertation, MIT, Cambridge, MA, 1982.

[11] A. J. Washington, "The Fingerprint of MalcomX," <http://www.dermatoglyphics.com>, June 2003.

[12] International Biometrics Group, <http://www.biometricgroup.com>, May 2003.

#### ۴- واژه‌نامه

هر واژه خارجی در واژه‌نامه در انتهای مقاله با شماره‌ای مشخص شود و شماره در معادل فارسی آن واژه در متن، بصورت بالانویس آورده شود.

#### ۵- آماده‌سازی مقاله

• مقاله را با نرم‌افزار Word تایپ نمائید.

• متن چکیده به صورت تک ستونی با طول سطر ۱۸ سانتیمتر و متن مقاله به صورت دو ستونی با طول هر ستون ۸۷ میلیمتر و فاصله دو ستون ۶ میلیمتر تایپ شود. حاشیه‌ها از بالا و پائین برابر ۲۰ میلیمتر و از طرفین برابر ۱۵ میلیمتر اختیار شود.

• فاصله عنوان مقاله در صفحه اول از بالای صفحه برابر ۸۵ میلیمتر باشد و عنوان وسط چین شود.

• کلیه عناوین بصورت پررنگ با قلم "B Nazanin" تایپ شوند، اندازه قلم عنوان مقاله ۱۸، عناوین سطح اول ۱۴، و عناوین سطح دوم و سوم ۱۲ انتخاب شوند.

• متن چکیده‌ها با قلم "B Nazanin" اندازه ۹، متن مقاله با قلم "B Nazanin" اندازه ۱۰، و کلمات و متن انگلیسی با قلم Times New Roman اندازه ۹ تایپ شوند.

• تمام متن بصورت تک فاصله تایپ شود. اسامی نویسندگان از عنوان مقاله و اسامی نویسندگان از عناوین محل اشتغال نویسندگان دو خط فاصله داشته باشد. بالای هر عنوان یک سطر فاصله قرار داده شود.

• سعی شود تعداد صفحات مقاله از ۳۰ صفحه بیشتر نباشد.

#### ۶- نحوه ارسال مقاله

• ارسال مقاله فقط از طریق ایمیل مجله ([csitjour@gmail.com](mailto:csitjour@gmail.com)) انجام شود.

• مقاله ارسالی برای نشریه علوم رایانش و فناوری اطلاعات نباید در جای دیگری به چاپ رسیده باشد و یا در زمان بررسی توسط نشریه برای چاپ به نشریه دیگری ارسال گردد.

• پس از قبول مقاله، نسخه نهائی تصحیح شده مقاله باید در قالب‌های Word و PDF به نشریه ارسال گردد.

• در نسخه نهائی باید بیوگرافی کلیه نویسندگان (به زبان فارسی) و عکس آنها در انتهای مقاله قرار داده شود، همچنین عنوان مقاله، نام نویسندگان، مؤسسه محل اشتغال، چکیده، کلمات کلیدی به زبان انگلیسی در فایل جداگانه ارسال شود.

# Improvement of Power Consumption and Execution Time in Cloud Computing Approach to Resource Allocation and Tasks Scheduling

**Atefe Yekta Aval**

**Mahmood Fathi**

Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran

## ABSTRACT

Scientific workflows are one of the famous ways of modeling cloud applications. One of the most challenging research issues in this context is how to schedule the different tasks onto the available resources so that we can reduce completion time of workflow's tasks. On the other hand one of the major and challenging issues in cloud computing is to reduce the power consumption of data centers. The problem is contradicts these qualitative criteria that many algorithms try to optimize them. In this article to reduce power consumption and reduce completion time of tasks, resource allocation has been done on two levels: At first, at the infrastructure level, we allocate physical resources to virtual machines so that we have increase resource utilization and reduce power consumption in the data center. Next we schedule applications on these virtual machines so that we decrease completion time of tasks. Our Experiments show that this method is successful.

**Keywords:** Cloud Computing, Tasks' Graph, Tasks Scheduling, Resource Allocation, Power Management.

## Evolutionary Rule-Based Recommender System

Behzad Soleimani Neysiani<sup>1,2</sup>

Nasim Soltani<sup>3</sup>

Reza Mofidi<sup>3</sup>

Mohammad H Nadimi-Shahraki<sup>4</sup>

<sup>1</sup>Faculty of Engineering, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran

<sup>2</sup>Faculty of Electrical & Computer Engineering, University of Kashan, Kashan, Iran

<sup>3</sup>Faculty of Engineering, Allame Naeini Higher Education Institute, Naein, Iran

<sup>4</sup>Faculty of Computer Engineering, Najafabad branch, Islamic Azad University, Najafabad, Iran

### ABSTRACT

Recommender systems are designed to deliver the products to potential customers as well as collaborative filtering is a perfect method for this purpose. This system will generate suggestions by identifying similar users based on the time of arrival and previous transactions. The low accuracy of suggestions due to sparseness is one of the major concerns about collaborative filtering method, to solve this problem, many researchers used the association rules mining methods. Creating fast high quality association rules can lead to higher quality and fast offers. In this regard, this paper proposes a rule-based recommender system using by Genetic algorithm which requires rules with constant vertices and finally produces rules with constant vertices. The experiments performed on the Movie Lensdataset show a higher rate in rule production than related works, and convergence is faster than the particle swarm algorithm while maintaining quality.

**Keywords:** Recommender System, Collaborative Filtering, Association Rule Mining, Particle Swarm Optimization, Multi-Objective Optimization.

# Modeling the Evolution of Class Relationships of Object-Oriented Software Systems Networks

Mojtaba Sadeghian

Mohammad Khansari

Farid Dehghan

Faculty of New Sciences and Technologies, University Of Tehran, Tehran, Iran

## ABSTRACT

Software class relationships networks evolve due to development and maintenance. Studying the changes in the characteristics of these networks helps to understand the design process and the hidden patterns of software development. Our goal in this paper is to model the evolution of the object-oriented software class network by examining network characteristics in the process of software changes (creation and removal of classes). Object-oriented software systems can be presented as directed networks in which the nodes correspond to the classes and edges corresponding to the connections between them (including associative, inheritance, realization and aggregation). Our studies have shown that the software network has small world characteristics and classes' creation and deletion do follow power-law in-degrees and normal out-degrees. The proposed model is evaluated based on the actual data of the EGit open source project (with 13 versions and an average of 720 classes). The results can help developers to predict the process of software development and extract the rules governing software development.

**Keywords:** Mining Software Code Repository, Network Analysis, Network Evaluation, Class Relationships of Object-Oriented Software Systems Networks.

# Using Social Networks to Improve Customer Relationship Management

**Elnaz Mazaheri**

**Seyed Abolghasem Mirroshandel**

Faculty of Engineering, University of Guilan, Rasht, Iran

## ABSTRACT

Nowadays, social networks have attracted attention of many people. Lots of resources are presented by companies in social media, and selection will be difficult in these large numbers of customers. Recommender systems are an important part of e-commerce, because they help customers to choose among few products accordance with their taste. The works done in this area propound solutions that companies can win customers loyalty by identifying customers' preferences and offer them sincere suggestions. In this article, a hybrid recommender system is presented for movie domain. This system combines collaborative filtering and content based filtering techniques to take advantage of their best features and overcome their limitations. Target of proposed algorithm is anticipation of like or dislike of a movie by user. This algorithm has been evaluated by cross validation and has demonstrated reasonable results.

**Keywords:** Hybrid Recommender System, Social Network, Customer Relationship Management.

# **A Social Image Collection Summarization Method Based on Content and Attractiveness of Images**

**Zahra Riahi Samani**

**Mohsen Ebrahimi Moghaddam**

Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran

## **ABSTRACT**

Automatic image summarization becomes very important in using large volume image collections available these days. They attempt to select a small set of the most representative images to highlight larger amounts of images. Recent studies have shown that summary images should cover different information aspect of the whole collection; and they should meet other factors like image quality or aesthetic features. In this paper, we propose an automatic approach for social image collection summarization that applies two different sets of features named attractiveness and content features. In the proposed approach, the introduced features are extracted using social media platform and domain ontology and their effectiveness in image collection summarization is modeled through a used feedback process. The final summary images are then selected by combing summaries coming from the two sets of features. The proposed approach was evaluated on a set of human generated summaries of a collection of Flickr images. The results show that the proposed method reaches a 35% improvement of information coverage in summarizing image collection to 1% of its original volume and 4% improvement in summarizing image collection to 5% of its original volume. It reaches a 39% improvement in making summary images that are like human made summaries as well.

**Keywords:** Image Collection Summarization, Social Images, User Feedback.

# Deep Learning for Recommender Systems

Omid Abbasi

Mahdieh Soleymani Baghshah

Computer Engineering Department, Sharif University of Technology, Tehran, Iran

## ABSTRACT

Collaborative filtering (CF) is one of the most widely employed approaches in Recommender systems (RS). This approach tries to find some latent features for users and items to predict user rates from these features. Early CF methods used matrix factorization to learn users and items latent features. However, these methods face cold start and sparsity problem. Recent methods usually employ side information along with rating matrix to learn users and items latent features. On the other hand, deep learning models show great potential for learning effective representations in the recent years. Due to this capability of deep learning, we use this approach to learn proper representation for items. In particular, we propose a hybrid method utilizes deep learning alongside matrix factorization which creates a two-way interaction between latent features learnt from matrix factorization and extracted features from the content of the items with deep learning methods and simultaneously optimize the parameters of these models. Experiments on real-world datasets show that the proposed method outperforms state-of-the-art RS methods.

**Keywords:** Recommender System, Deep Learning, Collaborative Filtering.

# The CSI Journal on Computing Science and Information Technology

---

Vol. 15

No. 2

2018

---

## ABSTRACTS

- **Deep Learning for Recommender Systems** ..... 1  
Omid Abbasi and Mahdiah Soleymani Baghshah
  
- **A Social Image Collection Summarization Method Based on Content and Attractiveness of Images** ..... 2  
Zahra Riahi Samani and Mohsen Ebrahimi Moghaddam
  
- **Using Social Networks to Improve Customer Relationship Management** ..... 3  
Elnaz Mazaheri and Seyed Abolghasem Mirroshandel
  
- **Modeling the Evolution of Class Relationships of Object-Oriented Software Systems Networks** ..... 4  
Mojtaba Sadeghian, Mohammad Khansari and Farid Dehghan
  
- **Evolutionary Rule-Based Recommender System** ..... 5  
Behzad Soleimani Neysiani, Nasim Soltani, Reza Mofidi and Mohammad H Nadimi-Shahraki
  
- **Improvement of Power Consumption and Execution Time in Cloud Computing Approach to Resource Allocation and Tasks Scheduling** ..... 6  
Atefe Yekta Aval and Mahmood Fathi

# The CSI Journal on Computing Science and Information Technology

*A Semiannual Publication of Computer Society of Iran (CSI)*

## Editor-in-Chief

A. Khonsari, Associate Professor, University of Tehran, Tehran, Iran.

## Editorial Board

H. R. Rabiee, Professor, Sharif University of Technology, Iran  
H. Sarbazi-azad, Professor, Sharif University of Technology, Iran  
K. Faez, Professor, Amirkabir University of Technology, Iran  
A. Ghaffarpour Rahbar, Professor, Sahand University of Technology  
E. Kabir, Professor, Tarbiat Modares University, Iran  
K. Navi, Professor, Shahid Beheshti University, Iran  
N. Yazdani, Professor, University of Tehran, Iran  
M. H. Yaghmaee Moghaddam, Professor, Ferdowsi University of Mashhad, Iran  
M. Analoui, Associate Professor, Iran University of Science & Technology, Iran  
M. Ebrahimi Moghaddam, Associate Professor, Shahid Beheshti University, Iran  
H. Asadi, Associate Professor, Sharif University of Technology, Iran  
A. Akbari, Associate Professor, Iran University of Science & Technology, Iran  
R. Berangi, Associate Professor, Iran University of Science & Technology, Iran  
H. Pedram, Associate Professor, Amirkabir University of Technology, Iran  
N. Moghadam Charkari, Associate Professor, Tarbiat Modares University, Iran  
G. Jaberipur, Associate Professor, Shahid Beheshti University, Iran  
J. Habibi, Associate Professor, Sharif University of Technology, Iran  
A. H. Jahangir, Associate Professor, Sharif University of Technology, Iran  
S. Hesabi, Associate Professor, Sharif University of Technology, Iran  
S. H. H. S. Javadi, Associate Professor, Shahed University, Iran  
M. Rahgozar, Associate Professor, University of Tehran, Iran  
M. Sedighi, Associate Professor, Amirkabir University of Technology, Iran  
H. Faili, Associate Professor, University of Tehran, Iran  
A. Ghasemi, Associate Professor, K.N. Toosi University of Technology, Iran  
M. Abbaspour, Associate Professor, Shahid Beheshti University, Iran  
M. Abdollahi Azgomi, Associate Professor, Iran University of Science & Technology, Iran  
M. Kargahi, Associate Professor, University of Tehran, Iran  
M. Goudarzi, Associate Professor, Sharif University of Technology, Iran  
N. Mozayani, Associate Professor, Iran University of Science & Technology, Iran

## Assistants

L. Nourani, Publication Assistant  
A. Ghaffari Sheshjavani, Editorial Assistant

**Disclaimer:** Publication of papers in CSI-CSIT does not imply that the editorial board, reviewers, or CSI-CSIT accept, approve or endorse the data and conclusions of authors.