دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

سال پانزدهم، شماره ۵۷ و۵۸ ، پاییز و زمستان ۱۴۰۲
صفحات:۷۷ الی ۹۲

A web recommender system for user nose pages using DBSCAN clustering algorithm and machine learning SVM method

Reza Molaee Fard[*], Mohammad Mosleh[**]
[*]Instructor, Department of Computer Engineering, Islamic Azad University, Dezful Branch, Dezful, Iran
[**]Assistant Professor, Department of Computer Engineering, Islamic Azad University, Dezful Branch, Dezful, Iran

Abstract

Today, due to the increasing growth of web pages, the existence of a system that can extract the information needed by users from the huge amount of data available on the web seems necessary and necessary. To do this, we need to personalize the desired systems. One of the best ways to personalize the system is to use recommender systems. Recommender systems are systems that can provide suitable suggestions to the user by getting limited information from the user. Recommender systems can predict future user requests and then generate a list of the user's favorite pages. In other words, it is possible to obtain a detailed profile of users' behavior and predict a page that the user will choose in the next move, which can solve the problem of the cold start of the system and improve the quality of the search. In this research, a new method is presented in order to improve recommender systems in the field of the web, which uses the DBSCAN clustering algorithm to cluster data, and this algorithm obtained an efficiency score of 99%. Then, using the Page rank algorithm, the user's favorite pages are weighted. Then, using the SVM method, we categorize the data and give the user a combined recommender system to generate predictions, and finally, this recommender system will provide the user with a list of pages that may be of interest to the user. The evaluation of the results of the research indicated that the use of this proposed method can achieve a score of 95% in the recall section and a score of 99% in the accuracy section, which proves that this recommender system can reach more than 90%. It detects the user's intended pages correctly and solves the weaknesses of other previous systems to a large extent.

Keywords: Recommender system, data mining, DBSCAN algorithm, SVM algorithm, machine learning

رضا مولایی فرد، محمد مصلح، دوفصلنامه فناوری اطلاعات و ارتباطات ایران، سال پانزدهم، شماره های ۵۷و ۵۸، پاییز و زمستان ۱۴۰۲، صفحه ۷۷ الی ۹۲

ارائه یک سیستم توصیه گر وب برای پیش‌بینی صفحات موردعلاقه کاربر با استفاده از الگوریتم خوشه‌بندی DBSCAN و روش SVM یادگیری ماشین

رضا مولایی فرد *×، محمد مصلح **

* مربی، گروه مهندسی کامپیوتر، ، دانشگاه آزاد اسلامی، واحد دزفول، دزفول، ایران

** استادیار، گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد دزفول، دزفول، ایران

تاریخ دریافت:۱۴۰۱/۰۶/۱۱ تاریخ پذیرش:۱۴۰۱/۱۱/۰۸

نوع مقاله: پژوهشی

چکیده

امروزه با توجه به رشد روزافزون صفحات وب، وجود سیستمی که بتواند اطلاعات موردنیاز کاربران را از میان حجم عظیم داده‌های موجود در سطح وب استخراج کند لازم و ضروری به نظر می‌رسد. سیستم‌های توصیه گر می‌توانند درخواست‌های آینده کاربر را پیش‌بینی و سپس لیستی از صفحات موردعلاقه کاربر را تولید کند. در این تحقیق به ارائه روش جدیدی به‌منظور بهبود سیستم‌های توصیه گر در زمینه وب پرداخته می‌شود که از الگوریتم خوشه‌بندی DBSCAN جهت خوشه‌بندی داده‌ها استفاده می‌شود که این الگوریتم امتیاز کارایی ۹۹٪ را به دست آورد که این الگوریتم توانست بالاترین میزان کارایی را از بین الگوریتم‌های خوشه‌بندی به دست آورد. سپس با استفاده از الگوریتم Page rank، صفحات موردعلاقه کاربر وزن دهی می‌شوند. سپس با استفاده از روش SVM، داده‌ها را دسته‌بندی کردیم که ماشین بردار پشتیبان بالاترین میزان دقت را نسبت به سایر الگوریتم‌های مشابه به دست آورد سپس داده‌ها را جهت تولید پیش‌بینی به کاربر به یک سیستم توصیه گر ترکیبی داده می‌دهیم که درنهایت این سیستم توصیه گر لیستی از صفحات را در اختیار کاربر قرار خواهد داد که می‌تواند موردعلاقه وی باشند. ارزیابی نتایج حاصل از تحقیق حاکی از آن بود که استفاده از این روش پیشنهادی می‌تواند امتیاز ٪۹۵ را در قسمت فراخوانی و امتیاز ٪۹۹ را در قسمت دقت به دست آورد که این نتایج اثبات می‌کند که این سیستم توصیه گر تا بیش از ۹۰٪ می‌تواند صفحات موردنظر کاربر را به‌درستی تشخیص داده یعنی وقتی کاربر اقدام به جستجو نماید از هر ۲۰ جستجو توسط کاربر قریب به ۱۸ جستجو مطابق و موردعلاقه کاربر خواهد بود، همچنین این سیستم می‌تواند مشکلاتی از قبیل مشکل شروع سرد که مشکل اکثر سیستم ای توصیه گر می‌باشد را با استفاده از یک سیستم توصیه گر ترکیبی که این سیستم ترکیبی از دو سیستم فیلترینگ مشارکتی و سیستم مبتنی بر دانش است را تا حدود زیادی برطرف نماید و می‌توان گفت که سیستم پیشنهادی از موفق‌ترین سیستم‌های توصیه گر بوده است.

واژگان کلیدی: سیستم توصیه گر، داده‌کاوی، الگوریتم **DBSCAN**، الگوریتم **SVM**، یادگیری ماشین

×**نویسنده مسئول: رضا مولایی فرد**molaefard@gmail.com

۱. مقدمه

با پیشرفت سریع تکنولوژی و اینترنت و افزایش منابع وب وجود مکانیزمی که بتواند نیازها و خواسته‌های کاربران را پیش‌بینی کند لازم و ضروری به نظر می‌رسد. به این منظور سیستم‌های توصیه گر یا پیشنهاددهنده به وجود آمدند. سیستم‌های پیشنهاددهنده سیستم‌هایی هستند که در پیدا کردن و انتخاب نمودن آیتم‌های موردنظر کاربران به آن‌ها کمک می‌کنند. طبیعی است که این سیستم‌ها بدون در اختیار داشتن اطلاعات کافی و صحیح در مورد کاربران و آیتم‌های موردنظر آن‌ها و همچنین سابقه جستجوی کاربر قادر به پیشنهاد دادن نمی‌باشند؛ بنابراین یکی از اساسی‌ترین اهداف آن‌ها جمع‌آوری اطلاعات گوناگون در رابطه با سلایق کاربران و آیتم‌های موجود در سیستم است. یکی از روش‌هایی که سیستم‌های توصیه گر از آن بهره می‌برند استفاده از رفتارها، فعالیت‌ها و سوابق کاربر همچون صفحات بازدید شده، علایق کاربر و یا ارتباطات او با سایر کاربران اشاره کرد. یک سیستم توصیه گر باقابلیتی که در جمع‌آوری اطلاعات مربوط به سلایق، علایق و اولویت‌های کاربران، دسته‌بندی و تفسیر آن‌ها دارد، امکانی را فراهم می‌آورد که کاربران با صرف زمان و انرژی کمتر به اطلاعات موردنظر خود دسترسی پیدا کنند. سیستم‌های توصیه گر یا پیشنهاددهنده زیرمجموعه ایی از سامانه پالایش اطلاعات که به دنبال پیش‌بینی امتیاز یا اولویتی است که کاربر به یک آیتم خواهد داد. در سال‌های اخیر سیستم‌های توصیه گر بسیار متداول شده و در حوزه‌های مختلفی مورداستفاده قرارگرفته‌اند. برخی از کاربردهای معروف این سیستم‌ها می‌توان به موارد زیر اشاره کرد: موسیقی، صفحات وب، اخبار، کتاب‌ها و مقالات، جستجو و شبکه های اجتماعی. سیستم توصیه گر ازجمله ابزارهایی است که می تواند کاربران را در محیط‌های الکترونیکی به سمت یافتن اطلاعات، خدمات و آیتم‌های موردنظرشان هدایت کند. سیستم‌های توصیه گر باقابلیت کشف علایق کاربران و پیش‌بینی اولویت آن‌ها، آیتم‌هایی که احتمال می‌رود موردتوجه کاربر باشد را از بین حجم بالای داده‌ها پالایش کرده و یا آن‌ها را پیشنهاد آن‌ها، در زمان او صرفه‌جویی می‌کند. از طرف دیگر این سیستم‌ها توانایی ذخیره و تحلیل رفتارهای گذشته کاربر، خدمات و اطلاعاتی را که موردتوجه کاربران نبوده و احتمالاً به آن‌ها علاقه‌مند هستند را نیز استنتاج کرده و نتایج جالب‌توجهی به کاربران ارائه می‌کند. درواقع سیستم‌های توصیه گر یکی از ابزارهای اصلی غلبه بر مشکل افزونگی اطلاعات بوده و با داشتن قدرت تحلیل رفتارهای کاربر، مکمل هوشمندی برای مفاهیم بازیابی و پالایش اطلاعات است؛ اما این سیستم ها علاوه بر مزایای زیادی که دارند، دارای معایبی نیز هستند که می‌توانند جستجو را با مشکلاتی همراه نمایند. یکی از اصلی‌ترین مشکلات سیستم‌های توصیه گر مشکل شروع سرد می‌باشد. این مشکل وقتی رخ می‌دهد که کاربران جدیدی به سیستم وارد می‌شوند یا اقلام جدیدی به کاتالوگ اضافه می‌شوند. در چنین مواردی نه سلیقه کاربر جدید را می‌توان پیش‌بینی نمود و نه اقلام جدید می‌توانند توسط کاربران رتبه‌بندی یا خریداری شوند که این موارد به ارائه پیشنهادهایی نامناسب و با دقت کمتر منجر می‌شود. مشکل شروع سرد می‌تواند با راه‌حل‌های زیادی حل شود، ازجمله: الف) در همان ابتدا از کاربر درخواست شود که برخی از اقلام را رتبه‌بندی کند، ب) به‌صورت کلی از کاربر درخواست شود که به‌طور واضح سلیقه‌ی خود را بیان نماید و ج) بر اساس اطلاعات جمعیت‌شناسی[1] جمع‌آوری‌شده، اقلام به کاربر جدید پیشنهاد شود. اطلاعات جمعیت‌شناسی می‌توانند برای آگاهی از مکان، کد پستی و غیره مورداستفاده قرار گیرند، این اطلاعات در حین تعاملات کاربر جدید با سیستم جمع‌آوری می‌شوند و برای پیشنهاد اقلامی مورداستفاده قرار می‌گیرند که توسط کاربران مشابه دیگری رتبه بندی شده‌اند، کاربرانی که اطلاعات جمعیت‌شناسی مشابهی با این کاربر دارند. در برخی حوزه‌ها ممکن است اقلام خفته[2] وجود داشته باشند، این موارد درواقع اقلامی هستند که آیتم‌های خوبی می باشند ولی رتبه‌بندی نشده‌اند. مشکل اقلام خفته می‌تواند با یکی از روش‌های زیر حل شود: استفاده از روش‌های متاداده یا مبتنی برمحتوا یا با استفاده از محبوبیت آیتم، آنتروپی آیتم و شخصیت کاربر و همچنین با استفاده از داده‌های باز پیوندی[3]، با استخراج داده‌هایی در مورد اقلام، با استفاده از منابع داده‌های بازپیوندی بدون منتظر ماندن برای رتبه‌بندی واضح توسط کاربر [۱]. از دیگر مشکلات سیستم‌های توصیه گر می‌توان به مشکل داشتن معنای یکسان اشاره نمود. مشکل داشتن معنای یکسان[4] وقتی رخ می‌دهد که یک آیتم با دو یا چند نام نشان داده شود، نام‌هایی که معنای مشابهی دارند [۲]. در چنین مواردی، سیستم توصیه‌گر نمی‌تواند تشخیص دهد که این نام‌ها نشان‌دهنده‌ی اقلام متفاوتی هستند یا همگی به یک آیتم یکسان اشاره دارند. به‌عنوان‌مثال، یک رویکرد فیلتر کردن مشارکتی مبتنی بر حافظه با دو عنوان " comedy movie" و "comedy film" به‌صورت متفاوت برخورد خواهد نمود، درصورتی‌که هر دو به معنای فیلم کمدی هستند. امکان وجود انواع این‌گونه تغییرات در هنگام استفاده از شرح و توصیف اقلام بسیار بیش‌ازحد معمول است و استفاده‌ی زیاد از واژگان مترادف و هم‌معنا باعث کاهش عملکرد سیستم‌های توصیه‌گر فیلتر کننده‌ی مشارکتی می‌شود. ازآنجایی‌که محتویات اقلام به‌طور کامل نادیده گرفته می‌شوند، سیستم توصیه‌گر وجود ارتباط مخفی بین

[1] demographic information

[2] sleepers

[3] Linked Open Data (LOD)

[4] synonymy

اقلام را در نظر نمی‌گیرد. به همین دلیل است که اقلام جدید تا زمانی که توسط کاربران رتبه‌بندی نشده‌اند، پیشنهاد داده نمی شوند. برای رفع مشکل داشتن معنای یکسان، روش‌های مختلفی ازجمله هستی‌شناسی [۳]، روش‌های تجزیه‌ی مقدار تکین[1] و شاخص‌گذاری معنای مخفی[2] مورداستفاده قرار بگیرند. مشکل دیگر سیستم‌های توصیه گر حمله‌های Shiling است. اگر یک کاربر مخرب یا یک رقیب وارد یک سیستم شود و شروع به دادن رتبه بندی‌های نادرست به اقلام شود تا محبوبیت اقلام افزایش یا کاهش یابد [۴]. چنین حمله‌هایی می‌توانند باعث ایجاد عدم اعتماد به سیستم توصیه‌گر شوند و همچنین عملکرد و کیفیت پیشنهادات را نیز کاهش دهند. از سایر مشکلات سیستم‌های توصیه گر می‌توان به پراکندگی داده‌ها داده‌ها اشاره نمود. وجود حجم عظیمی از داده ها در مورد اقلام موجود در دسته‌ها و همچنین عدم تمایل کاربران به رتبه‌بندی اقلام باعث به وجود آمدن پراکندگی در ماتریس نمایه می‌شود که این امر به ارائه‌ی پیشنهاداتی با دقت کمتر و نامناسب منجر می‌گردد [۵]. رتبه‌بندی پراکنده در سیستم‌های توصیه‌گر فیلتر کننده‌ی مشارکتی پیش‌بینی‌های دقیق در مورد اقلام را دشوار می‌سازد. سیستم‌های توصیه‌گر فیلترکننده‌ی مشارکتی از نزدیک‌ترین همسایه‌ها برای پیشنهاد اقلام استفاده می‌کند و تعداد کم رتبه‌بندی‌ها باعث می‌شود که محاسبه‌ی همسایه‌ها ازنظر محاسباتی دشوار باشد. این مشکل می‌تواند در سیستم‌های توصیه گر آگاه به محتوا شدیدتر نیز باشد، زیرا این سیستم‌ها از بردارهای چندبُعدی استفاده می‌کنند و بدین ترتیب برای اقلامی با رتبه‌بندی بسیار کم ارائه‌ی پیشنهادات دقیق بسیار دشوار می‌شود [۲]. برای مقابله با چنین وضعیتی، رویکردهای متعددی ازجمله مدل توصیه ی چندبُعدی، فیلتر کردن آگاه به جمعیت افراد و الگوریتم‌های فیلترینگ مشارکتی تقویت‌شده با محتوا می‌توانند مورداستفاده قرار گیرند. اکثر این مشکلات را می‌توان با استفاده صحیح از الگوریتم‌های داده‌کاوی و خوشه‌بندی برطرف نمود. در این تحقیق نیز سعی شده است که برخی از این مشکلات برطرف گردد ازجمله مشکل شروع سرد و مشکل پراکندگی که مشکل اکثر سیستم‌های توصیه گر می‌باشند. برای حل مشکل شروع سرد از ترکیب سیستم توصیه گر مبتنی بر دانش و ترکیب آن با الگوریتم فیلترینگ مشارکتی استفاده‌شده است همچنین از الگوریتم خوشه‌بندی جهت مشکلات پراکندگی داده‌ها مورداستفاده قرارگرفته است [۳،۴]. در این پژوهش از یک سیستم توصیه گر متشکل از الگوریتم‌های خوشه‌بندی و تکنیک‌های یادگیری ماشین ارائه می‌شود که قادر خواهد بود مشکل شروع سرد را که موجب کاهش کارایی سیستم‌های توصیه گر می‌شود برطرف نماید. برای این منظور از الگوریتم‌های خوشه‌بندی DBSCAN به‌منظور کشف الگوهای موجود در داده‌ها مورداستفاده قرار می‌گیرد.DBSCAN الگوریتم پایه روش‌های خوشه‌بندی مبتنی بر چگالی است. این الگوریتم قابلیت کشف خوشه‌های با اندازه متفاوت را از حجم زیادی داده دارد و در مقابل نویز نیز مقاوم می‌باشد. مزیت الگوریتم خوشه‌بندی DBSCAN این است که می‌تواند داده‌هایی با شکل‌های مختلف و نامنظم را خوشه‌بندی کند، کاری که الگوریتم‌های دیگر مانند K-MEANS و C-MEANS قادر به انجام آن نیستند. علاوه بر این با استفاده از تکنیک‌های ماشین جهت تولید پیش‌بینی و دسته‌بندی داده‌ها استفاده می‌شود. روش مورداستفاده، SVM یا ماشین بردار پشتیبان می‌باشد که با معیار قرار دادن بردارهای پشتیان، بهترین دسته‌بندی و تفکیک داده‌ها را انجام خواهد دهد. این سیستم توصیه گر می‌تواند درخواست‌های آینده کاربر را پیش‌بینی کند و سپس لیستی از صفحات موردعلاقه‌ی کاربر تولید شود. به‌عبارت‌دیگر بتوان نمایه‌ای دقیق از رفتار کاربران به دست آورده و صفحه‌ای پیش‌بینی شود که کاربر درحرکت بعدی آن را انتخاب خواهد کرد که این کار می‌تواند مشکلات کاربر هنگام مواجه با شروع سرد را حل کند و کیفیت جستجو را بالا ببرد.

۲. پیشینه پژوهش

معضوضی و همکاران در مقاله خود در سال ۲۰۲۰ به ارائه روش جدیدی به‌منظور بهبود توصیه صفحات وب به کاربران پرداختند. روش پیشنهادی این محققان که TDM نام‌گذاری شد توانست تا حدودی مشکلات مربوط به پراکندگی داده‌ها را حل نماید. این محققان در مقاله خود یک سیستم پیشنهادی مؤثر ارائه دادند که مبنای آن بر گفتگوی کاربران و گروه‌بندی آن‌ها در گروهای مختلف سپس توصیه مؤثر به یک گروه که ویژگی‌های مشابهی داشتند. نویسندگان از روش ضریب همبستگی پیرسون و مذاکرات TED کاربر استفاده کردند. سپس آن‌ها صفحات را با استفاده از روش خوشه‌بندی K-Means برای گروه‌بندی کاربران و سپس پیشنهاد به کاربر هدف، مورداستفاده قراردادند. [۶].

وو و همکاران در مقاله خود در سال ۲۰۲۲ به ارائه روشی به‌منظور بهبود سیستم توصیه گر وب با استفاده از نمودار پیچیدگی کاربر-آیتم پرداختند. در این روش، مزایای پیچیدگی نمودار را به سیستم توصیه‌کننده آگاه از متن که نشان‌دهنده یک نوع عمومی از مدل‌ها است که می‌تواند اطلاعات جانبی مختلف را مدیریت کند، گسترش داده شد. این محققان ماشین پیچیدگی گراف را پیشنهاد کردند که یک چارچوب سراسری است که از سه جزء تشکیل‌شده است: یک رمزگذار، لایه پیچیدگی گراف و یک رمزگشا. رمزگذار کاربران، آیتم‌ها و زمینه‌ها را به بردارهای جاسازی می‌کند که به لایه‌های GC منتقل می‌شوند که جاسازی‌های کاربر و آیتم را با

[1] Single Value Decomposition (SVD)

[2] Latent Semantic Indexing (LSI)

پیچیدگی‌های گراف آگاه از زمینه در نمودار کاربر-مورد اصلاح می‌کنند. رمزگشا تعبیه‌های تصفیه‌شده را هضم می‌کند تا با در نظر گرفتن تعاملات بین کاربر، آیتم و جاسازی‌های زمینه، امتیاز پیش‌بینی را به دست آورد. آزمایش‌هایی را بر روی سه مجموعه داده واقعی از Yelp و Amazon صورت گرفت که اثربخشی GCMو مزایای انجام پیچش‌های نمودار برای CARS را تأیید می‌کند [۷].

الحیجاوی و نایمات در مقاله خود در سال ۲۰۲۲ به بررسی روش‌های بهبوددهنده در زمینه سیستم توصیه گر وب پرداختند. این محققان معتقد بودند که فیلتر مشارکتی از نظر دقت موفقیت قابل‌توجهی دارد و به یکی از محبوب‌ترین روش‌های توصیه تبدیل می‌شود اما به ارائه روشی برای بهبود توصیه‌ها ارائه کردند. در این روش یک روش فیلتر مشترک مبتنی بر نمودار جدید، یعنی سیستم توصیه‌گر مبتنی بر نمودار چندلایه مثبت (PMLG-RS) را پیشنهاد کردند. این روش شامل یک نمودار چندلایه مثبت و یک الگوریتم جستجوی مسیر برای تولید توصیه‌ها است. نمودار چندلایه مثبت شامل دولایه متصل است: لایه کاربر و آیتم . PMLG-RSنیازمند توسعه یک روش جستجوی مسیر جدید است که کوتاه‌ترین مسیر را با بالاترین هزینه از یک گره منبع به هر گره دیگر پیدا می‌کند. مجموعه‌ای از آزمایش‌ها برای مقایسه PMLG-RSبا روش‌های توصیه شناخته‌شده مبتنی بر سه مجموعه داده معیار نشان‌دهنده برتری PMLG-RS و قابلیت بالای آن در ارائه توصیه‌های مرتبط، جدید و متنوع برای کاربران است [۸].

چاولا سوراچی در مقاله خود در سال ۲۰۱۸ به ارائه روشی به‌منظور بهبود سیستم‌های توصیه گر برای شخصی‌سازی صفحات وب پرداخت. این محقق در مقاله خود از ترکیبی از الگوریتم‌های ژنتیک و اعتماد به URL های کلیک شده و قابل‌اعتماد برای توصیه صفحات وب استفاده نمود. صفحات وب مورد اعتماد کاربران بر اساس جلسات پرس و جوی خوشه ایی برای رتبه‌بندی بهینه با GA استفاده شد تا اسناد با ارتباط بیشتر در رتبه‌بندی بازیابی شوند و دقت نتایج بهبود گردد. رتبه‌بندی مطلوب URL های کلیک شده قابل‌اعتماد، اسناد مربوط را به کاربران وب جهت هدف جستجوی خود توصیه می‌کند و نیازهای اطلاعاتی کاربر را به‌طور گسترده ایی برآورده می‌کند [۹].

تنور و ویشواکارما در مقاله ایی که در سال ۲۰۲۲ ارائه کردند معتقد بودند که در عصر دیجیتال امروزی، انتخاب محصول مناسب، صفحه وب، مقاله خبری یا حتی یک مقاله تحقیقاتی مانند این از بین گزینه‌های فراوان، یکی از خسته‌کننده‌ترین کارها است. راه‌حل این مشکل استفاده از یک سیستم توصیه‌کننده (RS) است که به شما کمک می‌کند آیتم مناسب را با توجه به مشخصات خود انتخاب کنید. در این تحقیق، یک سیستم توصیه‌کننده ترکیبی مبتنی بر شبکه عصبی عمیق جدید ارائه شد که به خلأهای فیلتر مشترک سنتی (CF) و سیستم‌های ترکیبی فعلی می‌پردازد و درعین‌حال دقت بالاتری را در توصیه‌ها ارائه می‌دهد. به دلیل داده‌های آموزشی ناکافی، سیستم‌های توصیه‌کننده CF از دقت پایین، عامل پنهان خطی و مشکل شروع سرد رنج می‌برند. برای غلبه بر این مشکلات، از یک رویکرد مبتنی بر شبکه عصبی عمیق استفاده شد که از بردارهای کاربر و آیتم برای کپسوله کردن داده‌های کاربران و آیتم‌ها برای آموزش داده‌های غیرخطی با ابعاد بالا برای ارائه توصیه‌های دقیق‌تر استفاده می‌کند. شبکه‌های کاربر-کاربر برای ارائه یک همکاری و جنبه هم‌افزایی بهتر به این مدل استفاده می‌شوند. در این رویکرد، ترکیب شبکه‌های کاربر-کاربر با شبکه‌های عصبی عمیق، دقت پیش‌بینی بالاتر و زمان اجرای بهتری را نسبت به سایر روش‌های پیشرفته به دست می‌آورد. [۱۰].

بوپانا و ثاندهاوا در مقاله خود در سال ۲۰۲۱ به ارائه روشی به‌منظور بهبود سیستم‌های توصیه گر وب پرداختند. در این مقاله یک مدل پیشنهادی مبتنی بر یادگیری عمیق مؤثر پیشنهادشده است که می‌تواند با نشان دادن حداقل خطا در حین توصیه، به‌عنوان یک سیستم توصیه گر وب کارآمد عمل کند. در ابتدا، مجموعه داده با استفاده از کیت ابزار زبان طبیعی (NLTK) در پلتفرم پایتون پیش‌پردازش شده است. پس از پیش‌پردازش، مدل TF–IDF و جاسازی کلمه برای هر بررسی از پیش‌پردازش شده برای استخراج ویژگی‌ها و اطلاعات متنی استفاده می‌شود. ویژگی استخراج‌شده به‌عنوان ورودی خوشه‌بندی مبتنی بر چگالی برای گروه‌بندی احساسات منفی، خنثی و مثبت نظرات کاربران در نظر گرفته می‌شود. درنهایت، شبکه عصبی بازگشتی عمیق (DRNN) برای به دست آوردن ارجح‌ترین کاربر از هر خوشه استفاده می‌شود. مقادیر پارامتر مدل شبکه عصبی بازگشتی از طریق محاسبه تناسب الگوریتم جستجوی عقاب طاس (BES) مقداردهی اولیه می‌شوند. مدل پیشنهادی که با استفاده از پلتفرم برنامه‌نویسی پایتون پیاده‌سازی شده و عملکرد بر اساس معیارهای دقت، یادآوری و مقایسه با مدل‌های موجود ارزیابی شد و توانست عملکرد قابل قبولی را به نمایش بگذارد [۱۱].

ویشنبرت و همکاران در مقاله ایی که در سال ۲۰۲۱ ارائه کردند به بررسی سیستم‌های توصیه گر وب و ارائه روشی به‌منظور بهبود توصیه‌ها پرداختند. این محققان معتقد بودند که سیستم‌های توصیه گر به ابزاری قدرتمند برای بهبود شخصی‌سازی در وب تبدیل‌شده‌اند. بااین‌حال، بسیاری از وب‌سایت‌های محبوب فاقد

چنین عملکردی هستند، پیاده‌سازی آن معمولاً به مهارت‌های فنی خاصی نیاز دارد و مهم‌تر از همه، معرفی آن خارج از محدوده و کنترل کاربران نهایی است. برای کاهش این مشکلات، این مقاله ابزار جدیدی را برای توانمندسازی کاربران نهایی بدون مهارت‌های برنامه‌نویسی، بدون دخالت ارائه‌دهندگان وب‌سایت، برای جاسازی توصیه‌های شخصی از موارد در وب‌سایت‌های دلخواه در سمت مشتری ارائه می‌دهد. برای این کار یک متا-مدل عمومی برای گرفتن پارامترهای پیکربندی سیستم توصیه‌گر به‌طورکلی و همچنین درزمینه تقویت وب ایجاد شد. پس‌ازآن، یک قالب افزونه مرورگر با استفاده آسان پیاده‌سازی شد که به تولید به‌اصطلاح اسکریپت‌های کاربر اجازه می‌دهد که در مرورگر اجرا شوند تا عملکرد فیلتر مشترک را از یک سرویس استراحت خارجی ارائه‌شده درگیر کنند. نتایج حاصل از ارزیابی این روش حاکی از دقت ۹۳ درصدی این سیستم بود [12].

بورکوکو و عمر در مقاله خود در سال ۲۰۱۸ به ارائه یک سستم توصیه گر به‌منظور بهبود نتایج با استفاده از تاریخچه‌ی جستجوی قبلی کاربران پرداختند. این محققان در رویکرد پیشنهادی خود سعی کردند تا منابع یادگیری را به فرد یادگیرنده با در نظر گرفتن ترجیحات وی و تاریخچه‌ی جستجوهای قبلی وی پیشنهاد دهند، تاریخچه‌ای که از فایل‌های log استخراج‌شده است. این رویکرد، سبک‌های یادگیری و روش‌های فیلتر کردن مشارکتی را ترکیب می‌کند تا کیفیت پیشنهادات را ارتقا دهد. [۱۳].

ریاحی و سهرابی در مقاله خود در سال ۲۰۲۰ به ارائه روشی به‌منظور بهبود توصیه صفحات وب با استفاده از یک سیستم توصیه گر ترکیبی و استفاده از برچسب‌گذاری داده‌ها پرداختند. این محققان ارتباط معنای برچسب‌ها را با استفاده از بانک اطلاعاتی واژگان WORDNET استخراج کردند سپس برچسب‌ها را بر اساس اهمیت معنایی آن‌ها در یک ساختار سلسله مراتبی سازمان‌دهی کردند. ساختار سلسله مراتبی برای جستجوی برچسب‌های مربوطه در بخش فیلتر محتوا محور مورداستفاده قرار گرفت و درخواست‌های کاربران با استفاده از وب معنایی مرتبط باهم گسترش یافت و در قسمت فیلترینگ مشارکتی محاسبه گردید. نتایج حاصل از ترکیب این دو بخش یک سیستم پیشنهادی ترکیبی بود که می‌توانست صفحات را به کاربران پیشنهاد دهد [۱۴].

سیسودیا و همکاران در مقاله خود در سال 2017 از خوشه‌بندی جهت ایجاد یک سیستم پیشنهاددهنده استفاده‌شده است. در این مقاله یک معیار عدم شباهت بر اساس ارتباط بین دسترسی به صفحات و ساختار نحوی URL های وبگاها ارائه‌شده سپس از الگوریتم K_MEANS برای خوشه‌بندی کاربران وب استفاده‌شده است. در ادامه ارزش خوشه‌های تولیدشده توسط دو شاخص اعتبارسنجی خوشه ایی ارزیابی‌شده است. نتایج این مقاله نشان می‌دهد که اندازه‌گیری غیرمستقیم نسبت به سایر روش‌های عدم شباهت مستقل در مورد شاخص‌های اعتبار خوشه ایی برتر است [۱۵].

ژی و وانگ در مقاله خود در سال 2018 به ارائه روشی به‌منظور پیشنهاد صفحات وب پرداختند. این محققان در مقاله خود برای پیشنهاد صفحات از خوشه‌بندی دوگانه استفاده‌شده است. در این روش نقاط قوت خوشه‌بندی مبتنی بر k-means ترکیب‌شده و ایده اصلی آن استفاده از خوشه‌بندی مبتنی بر تراکم جهت شناسایی تعداد خوشه و مراکز اولیه هر خوشه است. سپس صفحات وب به‌دست‌آمده را به کاربران پیشنهاد می‌کردند [۱۶].

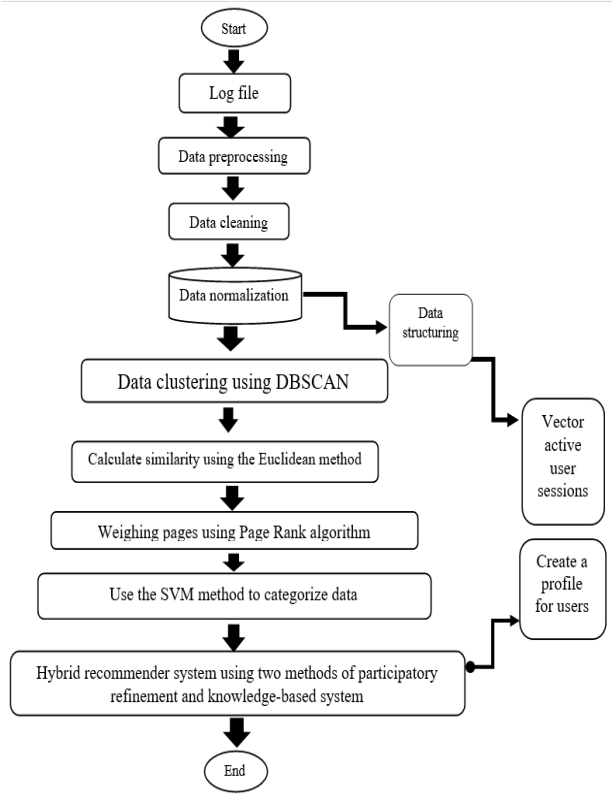
واق و پاتیل در سال ۲۰۱۹ در [۱۷] برای شخصی‌سازی وب و توصیه صفحات وب از تکنیک‌های وب‌کاوی استفاده می‌کند. این تکنیک‌ها برای پیدا کردن رابطه بین صفحات وب، مرحله خوشه‌بندی و طبقه‌بندی در روش های داده‌کاوی و تجزیه‌وتحلیل داده‌ها استفاده می‌شود. این دو محقق برای رابطه بین صفحات از اقدامات جدیدی مانند ماتریس فاصله و ماتریس فرکانس وقوع و ماتریس رابطه مدل کرده‌اند. برای رابطه بین صفحات وب، نمودار مجازی متناسب با ماتریس رابطه ایجاد می‌شود. آن‌ها به ارائه یک الگوریتم جستجوی پیشرفته، نمودار مجازی را به خوشه‌های مختلف یعنی الگوهای ناباروری تقسیم می‌کند. این روش یک الگوریتم تقسیم‌بندی بر اساس نمودار است. با استفاده از الگوریتم LCS، کاربر فعال موردنظر را در یکی از خوشه‌ها طبقه‌بندی می‌شود و در آخر یک مقدار آستانه استفاده کرده‌اند تا فقط صفحات بهینه را به کاربر پیشنهاد کنند.

۲،۱ سیستم توصیه گر

سیستم‌های توصیه گر در سال‌های اخیر اهمیت زیادی کسب کرده‌اند. هدف هر سیستم پیشنهاددهنده این است مصرف‌کنندگان بتوانند کالاها یا خدمات جدیدی ازجمله، وب، کتاب، موسیقی، رستوران یا حتی افراد را بر اساس اطلاعات درباره مصرف‌کننده یا مورد توصیه‌شده پیدا کنند [۱۸،۱۹]. سیستم توصیه گر سیستمی است که با توجه به ترجیحات کاربر، موارد را به‌طور مشترک به گروهی از کاربران توصیه می‌کند [۲۰]. سیستم‌های توصیه گر سیستم هایی هستند که به یافتن موارد موردعلاقه کاربر در موقعیت‌های بیش‌ازحد اطلاعات کمک می‌کنند. جایی که ترجیحات کاربر بر اساس رفتار مشاهده‌شده وی درگذشته تخمین زده می‌شود و می‌تواند لیستی رتبه‌بندی شده از پیشنهادات را در اختیار کاربر قرار دهد.

۳. روش پیشنهادی

در روش پیشنهادی به ارائه روشی جهت بهبود سیستم‌های توصیه گر وب پرداخته می‌شود بدین‌صورت که ابتدا باید به جمع‌آوری فایل‌های ثبت کاربران بپردازیم. فایل ثبت کاربران شامل اطلاعاتی ازجمله آدرس IP سرویس‌گیرنده، زمان درخواست، URLدرخواست شده، نوع سیستم‌عامل مورداستفاده، تاریخ بازید و... می‌باشد. در مرحله اول باید عمل پیش‌پردازش داده‌ها را انجام داد تا داده‌ها را آماده کنیم زیرا نمی‌توان به‌صورت خام داده‌ها را به الگوریتم‌های داده‌کاوی تزریق کرد. سپس باید داده‌ها را پاکسازی کرد، عملیات پاکسازی بدین منظور انجام می‌گیرد که تمام دادهای مورداستفاده مناسب نیستند و باید داده‌های اضافی پاکسازی شوند. در مرحله بعد باید داده‌ها را نرمالسازی کنیم، نرمالسازی یعنی داده‌ها را از یک رنج به رنج دیگر ببریم، هدف نرمالسازی حذف افزونگی و باقی نگه‌داشتن وابستگی بین داده‌هاست. پس از نرمالسازی داده‌ها باید آن‌ها را خوشه‌بندی کرد، برای خوشه‌بندی داده‌ها از الگوریتم خوشه‌بندی DBSCAN استفاده می‌کنیم سپس باید داده‌های خوشه‌بندی‌شده را وزن دهی کنیم. وزن دهی بدین منظور که کاربر چه رتبه ایی به کالاهای موردنظر داده است انجام می‌شود. سپس شباهت داده‌ها را با استفاده از الگوریتم اقلیدسی محاسبه می‌کنیم. در مرحله بعد با استفاده از روش یادگیری ماشین داده‌ها دسته‌بندی می‌شوند و با استفاده از سیستم توصیه گر صفحات را به کاربر هدف پیشنهاد می‌دهد. در شکل (۱) نمایی از روش پیشنهادی را ملاحظه می‌کنید.



شکل ۱. نمایی از روش پیشنهادی سیستم مورد

۳,۱ آماده‌سازی داده‌ها

در سیستم پیشنهادی مدنظر ابتدا باید به آماده‌سازی داده‌ها پرداخت زیرا داده‌های وب جمع‌آوری‌شده معمولاً دارای حجم زیاد، بسیار ناهمگن و ساختار نیافته می‌باشند. این داده‌ها باید به داده‌های سازگار و یکپارچه تبدیل شوند تا بتوانند برای مرحله کشف الگو مفید باشند. در این ماژول ابتدا فایل‌های ثبت وقایع از سرورهای موردنظر جمع‌آوری می‌گردند و در پایگاه داده جهت بررسی‌های آتی ذخیره می‌شوند.

۳,۱,۱ داده ها

برای پیاده‌سازی سیستم پیشنهادی ابتدا باید یک فایل ثبت یا Log file، از درخواست‌های کاربران مختلف را جمع‌آوری کرد. برای این کار از Log file مربوط به ناسا که در سایت http://ita.ee.lbl.gov/html/contrib/NASAHTTP، قابل‌استخراج می‌باشد را جمع‌آوری می‌کنیم. فایل ثبت ناسا شامل بیش از ۱۱۷۵۰۰۰ جستجو مربوط به کاربران در یک بازه مشخص بود که پس از عملیات پیش‌پردازش داده‌ها فایلی حاوی ۸۷۳۲۳۳ نشست استخراج گردید که فایل ثبت موردنظر را با استفاده از قانون ۳۰ دقیقه، تبدیل به نشست‌های فعال کاربران تبدیل کردیم. در پایان این مرحله، از فایل ثبت اولیه فایلی حاوی ۱۰۴۸۷۳ نشست استخراج گردید که با پاکسازی این نشست‌ها به ۷۴۳۵۲ نشست تبدیل شدند. نمونه ایی از قایل ثبت اولیه مستخرج از سایت ناسا را در شکل ۲ مشاهده می‌کنید.

شکل۲. نمونه ایی از فایل ثبت مستخرج از NASA Log file

اطلاعات موجود در این فایل ثبت بدون ساختار و بدون محتوا هستند و باید ابتدا مورد پیش‌پردازش قرار گرفته سپس سایر عملیات‌های داده‌کاوی بر روی آن‌ها قرار گیرد. در جدول زیر نمونه ایی مربوط به ۵ نمونه از منابع مربوط به یک روز را مشاهده می‌کنید.

همان‌گونه که در جدول ۱ مشاهده می‌کنید همه این داده‌ها شامل ۵ ویژگی مشترک می‌باشند که این ویژگی‌ها عبارت‌اند از:

Host: هاستی که درخواست موردنظر از جانب آن ارسال‌شده است. این بخش دارای ۲ نوع هاست می‌باشد. در صورت وجود هاست، آدرس آن نشان داده می‌شود و در غیر این صورت آدرس IP، مربوطه نشان داده می‌شود.

جدول ۱. اطلاعات مربوط به فایل ثبت بعد از تفکیک

آدرس هاست	زمان درخواست	درخواست	کد پاسخ از طرف HTTP	میزان بایت در پاسخ
tagoss.clark.net	/۰٤Sep/1995:00:00:27	"GET /html/cgi.html HTTP/1.0"	200	1834
ix-dc9—19.ix.netcom.com	/۰٤Sep/1995:00:00:28	"GET /html/cgi.html HTTP/1.0"	200	2217
evins-ppp.clark.net	04/Sep/1995:00:00:28	"GET /pub/tcarpent/tlc-home.html HTTP/1.0"	200	1969
mitchdas.liii.com	/۰٤Sep/1995:00:00:28	"GET /pub/chinatom/home.html HTTP/1.0"	200	1432
s211-177.qns.com	/۰٤Sep/1995:00:00:28	"GET /pub/sshay/interact.html HTTP/1.0"	200	2541

- زمان ارائه درخواست: در این بخش زمان ارائه دقیق درخواست کاربر نشان داده می‌شود.
- درخواست: این بخش درخواست موردنظر کاربر را نشان می‌دهد.
- کد پاسخ: در این بخش کد پاسخ به درخواست کاربر از طرف HTTP را نشان می‌دهد. این ویژگی کد پاسخ به درخواست موردنظر از طرف NASA را نمایش می‌دهد.
- میزان بایت پاسخ: در این بخش میزان بایت پاسخ به درخواست کاربر را نشان می‌دهد.

همچنین نتایج مربوط به تعداد نشست ها پس از پاکسازی را در جدول ۲ مشاهده می‌کنید.

جدول ۲: تعداد نشست های کاربران

مجموع داده ها	تعداد کل نشست ها	نشست های پاکسازی شده
فایل ثبت استخراج شده از سایت ناسا	۱۰۴۸۷۳	۷۴۳۵۲

جدول ۳: تعداد کل نشست ها بعد از خوشه بندی

مجموع نشست ها	۷۴۳۵۲
تعداد خوشه ها	۷۰۹

۳،۲ پیش‌پردازش داده‌ها

در مرحله اول روش پیشنهادی ابتدا باید عمل پیش‌پردازش داده‌ها را انجام دهیم. اگر داده های مختلف پیش پردازش شوند، عملکرد قابل‌اعتماد و مؤثر یکسانی در تمام مجموعه‌های داده ایی به وجود خواهد آمد. در فرایند داده‌کاوی مانند طبقه‌بندی و خوشه‌بندی نیاز داریم تا داده‌ها برای الگوریتم آماده شوند، زیرا معمولاً نمی‌توان داده‌ها را به‌صورت خام به الگوریتم‌های داده‌کاوی و یادگیری ماشین تزریق کرد [۱۸،۱۹]. برای آماده‌سازی داده‌ها، نیاز است تا آن‌ها را از شکل و حالت اولیه، خارج کرده و به شکلی که برای الگوریتم مناسب باشد تبدیل کرد [۲۰]. پیش‌پردازش داده‌ها شامل مراحل اصلی زیر می‌باشد:

۳،۲،۱ پاکسازی داده‌ها

در این مرحله باید داده‌های موجود را پاکسازی کنیم. پاکسازی داده‌ها فرایند از بین بردن خطاها و ناسازگاری‌ها در داده‌هاست و درواقع مرحله کنترل کیفی قبل از انجام تحلیل داده‌ها می‌باشد. اغلب به جهت خطاهای عملیاتی و پیاده‌سازی سیستم‌ها، داده‌های برآمده از منابع دنیای واقعی پر غلط، ناقص و ناسازگار هستند. لازم است در ابتدا چنین داده‌هایی پاکسازی شوند. این کار شامل برخی عملیات پایه مانند نرمالسازی، حذف نویز یا اغتشاش، مواجهه با داده‌های مفقوده، کاهش افزونگی و برطرف کردن داده‌هاست.

سه فاز اصلی تعریف‌شده برای فرایند پاکسازی داده‌ها بدین شرح است:

- تعریف و تعیین نوع خطا
- جستجو و شناسایی موارد خطا
- تصحیح خطاهای کشف‌شده

۳،۲،۲ یکپارچه‌سازی داده‌ها

در مرحله بعد باید داده‌های خود را یکپارچه کنیم. یکپارچه‌سازی داده‌ها باعث کاهش هزینه‌های ادغام و بهبود دقت می‌شود [۲۱]. یکپارچه‌سازی، به دلایل گوناگونی منابع و همپوشانی داده‌ها در منابع گوناگون، تنوع در نحوه ذخیره‌سازی داده‌ها و قابلیت پردازش تقاضاهای متفاوت در منابع مختلف انجام می‌شود. در این گام مشکلات مربوط به تضاد و افزونگی داده‌ها بررسی و رفع می‌گردد؛ بنابراین درصورتی‌که نگاره‌ها در منابع اطلاعاتی مختلفی ثبت‌شده باشند، نیاز به یکپارچه‌سازی دارند.

۳،۲،۳ کاهش داده‌ها

مرحله دیگر پیش‌پردازش داده‌ها، کاهش داده‌ها می‌باشد. هدف از کاهش داده، دست‌یابی به حجم کوچک‌تری از داده‌هاست. یکی از مهم‌ترین دلایل کاهش داده‌ها، حجم بالای داده‌هاست که تحلیل آن‌ها را پیچیده، زمان‌بر و گاهی غیرممکن می‌کند. استخراج دانش از داده‌هایی با حجم بالا مستلزم زمان زیادی است بنابراین لازم است روش‌هایی برای کاهش اندازه داده‌ها به‌کاربرده شود. هدف تکنیک کاهش داده‌ها در داده‌کاوی، استخراج زیرمجموعه ایی کوچک از حجم انبوهی از داده‌ها با حفظ خصوصیات داده‌های اصلی می‌باشد. این امر باعث می‌شود عملیات سخت و یا غیرممکن داده‌کاوی را به‌صورت کارا و مؤثری انجام شود.

۳،۲،۴ تبدیل داده‌ها

عملیاتی همچون نرمالسازی داده‌ها، تغییر و تبدیل داده‌ها در این گام انجام می‌شود، زیرا ویژگی‌های نگهداری شده در منابع اطلاعاتی، داده‌های خام هستند، یعنی ویژگی‌ها، متناسب با حوزه کاری خاص و یا نتیجه کار سامانه ایی مشخص، طراحی و نگهداری شده‌اند. این داده‌ها مناسب پردازش نیستند و لازم است به استانداردی متناسب با پروژه موردنظر تبدیل شوند.

۳،۲،۵ نرمال‌سازی داده‌ها

در این مرحله باید داده‌ها را نرمال کنیم، نرمال‌سازی داده‌ها تغییر داده‌ها به گونه ایی است که آن‌ها را به یک دامنه کوچک و معین مانند فاصله بین ۱- و ۱ نگاشت کنند. هدف نرمال‌سازی حذف افزونگی داده‌ها و باقی نگه‌داشتن وابستگی بین داده‌های مرتبط می‌باشد. این فرایند اغلب باعث ایجاد جداول بیشتر می‌شود ولی اندازه‌گیری پایگاه داده را کاهش داده و بهبود کارایی را تضمین می‌کند. روش‌های مختلفی جهت نرمال‌سازی داده‌ها وجود دارد که از معروف‌ترین آن‌ها می‌توان به روش Min-Max Normalization اشاره کرد. در این روش هرکدام از داده‌ها را می‌توان به یک بازه دلخواه تبدیل کرد. فرمول کلی این روش برای تبدیل داده‌ها به بازه بین ۰ تا ۱ به‌صورت زیر می‌باشد:

(۱)

$$Z = \frac{X - MIN(X)}{MAX(X) - MIN(X)}$$

۳،۲،۶ ساختار سازی داده‌ها

در این مرحله باید عملیات ساختار سازی داده‌ها را بروی داده‌های به‌دست‌آمده از مراحل قبل، انجام دهیم، بدین‌صورت که داده نرمال‌سازی شده از فایل‌های ثبت را به نشست‌های کاربران تبدیل کنیم. نشست‌های کاربران نشان‌دهنده علایق و رفتارهای کاربران هستند که در سیستم توصیه از آن‌ها برای استخراج الگوهای رفتاری کاربران مورداستفاده قرار می‌گیرند [۲۲]. بدین‌صورت که با استفاده از یک تایم زمانی مشخص برای شناسایی نشست‌ها مورداستفاده قرار می‌گیرد، صفحاتی که در آن تایم زمانی مشخص‌شده مشاهده شوند به‌عنوان یکی از نشست‌های کاربر در نظر گرفته می‌شود.

۳،۳ بردارسازی نشست‌های فعال کاربران

به‌منظور خوشه‌بندی اطلاعات کاربران ابتدا آن‌ها را به کمک تکنیک هایی به یک بردار تبدیل می‌کنیم که مجموع این بردارها تشکیل یک ماتریس می‌دهد. هر سطر ماتریس یک کاربر می‌باشد و در هر ستون صفحه ایی که کاربر دیده است با عددی نمایش داده می‌شود که دفعات بازدید آن صفحه توسط کاربر را نشان می دهد. S_i نشست i-ام یک کاربر است که به‌صورت رابطه زیر تعریف می‌شود:

$$Si = < w(p1,si), w(p2,si), \dots, w(pk,si), \dots w(pn,si) > \quad (۲)$$

n تعداد صفحات وبی است که در همه جلسات دسترسی کاربران بازدید شده‌اند. Pk صفحه k-ام است و $w(pk, si)$ نشان‌دهنده وزن و عددی است که باید در خانه k-ام بردار مربوط به Si قرار گیرد. این وزن با توجه به معیار بسامد محاسبه می‌شود. بسامد درواقع تعداد بازدید از یک صفحه وب است. فرض بر این است که صفحات با بسامد بالاتر محبوبیت بیشتری نزد کاربران دارند که این امر طبیعی به نظر می‌رسد. بسامد هر صفحه در هر نشست از طریق رابطه زیر محاسبه می‌شود:

$$Frequency(page) = \frac{number\ of\ visits\ in\ the\ session(page)}{\sum number\ of\ visits\ in\ the\ session(page)} \quad (۳)$$

صورت این کسر نشان‌دهنده تعداد بازدیدهای کاربر از یک صفحه در یک نشست مشخص است. مخرج آن بیانگر تعداد کل بازدیدها از صفحات وب در همان نشست مشخص است. در پایان نیز تمام بردارهای مربوط به نشست های دسترسی کاربران در کنار هم قرارگرفته و ماتریس $m * n$ بعدی از وزن‌های صفحات وب تشکیل می‌شود (m تعداد کل نشست‌های کاربران). سطرهای این ماتریس نشان‌دهنده نشست‌های کاربران است و ستون‌های آن نشان‌دهنده صفحاتی هست که در نشست‌های مختلف وب، بازدید شده‌اند.

۴. استفاده از الگوریتم خوشه‌بندی DBSCAN

به‌منظور درک الگوریتم $DBSCAN$ لازم است ابتدا برخی از تعاریف مورداستفاده در این الگوریتم معرفی شوند:

الگوریتم $DBSCAN$ نیاز به تعیین ۲ پارامتر Eps و $Minpts$دارد. این دو پارامتر برای تعیین حداقل چگالی یک خوشه مورداستفاده قرار می‌گیرند.

تعریف ۱: همسایه‌های شعاع Eps یک نقطه: همسایه‌های موجود در شعاع Eps یک نقطه مثل p که به $NEPs(p)$نشان داده می‌شوند مجموعه‌ای از نقاط هستند که فاصله‌شان از p کمتر از شعاع Eps باشد یعنی:

$$NEPs(p) = \{q \in \frac{D}{Dist(p,q)} \leq EPs\} \quad (۴)$$

تعریف ۲: شیء مرکزی، به شیء ایی که حداقل تعداد $Minpts$ شیء در همسایگی شعاع Eps خود را داشته باشد شیء مرکزی گفته می شود.

تعریف ۳: دسترسی‌پذیر چگالی مستقیم، نقطه p دسترسی‌پذیر چگالی مستقیم از نقطه q است اگر اولاً p جزء همسایه‌های شعاع Eps شیء q باشد و ثانیاً شیء q یک شیء مرکزی باشد.

تعریف ۴: دسترسی‌پذیر چگالی، نقطه p دسترسی‌پذیر چگالی از نقطه q است اگر یک زنجیره از نقاط $p1, p2, p3, \dots, pn$ دسترسی‌پذیر چگالی مستقیم از pi باشد.

تعریف ۵: متصل چگالی، نقطه p متصل چگالی از نقطه q است اگر یک نقطه مثل o وجود داشته باشد به گونه ایی که هردوی p و q دسترسی پذیر چگالی از o باشند

تعریف ۶: خوشه: فرض کنید که D یک پایگاه داده از نقاط باشد. خوشه C یک زیر مجموعه غیر تهی از D است به گونه ایی که شرط‌های زیر را ارضاء کند:

- به ازای همه جفت نقاط q و p اگر $p \in C$ ، یعنی p از اعضای خوشه C باشد(شرط حداکثر بودن).
- به ازای همه جفت نقاط q وp باید متصل چگالی از q باشد (شرط اتصال).

تعریف ۷: نویز: فرض کنید که $C1, C2, \dots, Ck$ خوشه‌های یافت شده از پایگاه داده D باشند. به مجموعه ایی از نقاط که در پایگاه D وجود دارند ولی متعلق به هیچ‌یک از خوشه‌های یافت شده یافت $Cie/to\ k$نباشند می گویند.

$$Noise = \{p \in D | \forall i: p \in Ci\} \quad (۵)$$

تعریف ۸: شیء حاشیه ایی: شیء حاشیه ایی به شی ایی گفته می شود که شیء مرکزی نباشد منتها از یک شیء مرکزی دیگر دسترسی‌پذیر چگالی باشد.

۴،۱ نحوه عملکرد الگوریتم

نحوه کار این الگوریتم بدین‌صورت می‌باشد که DBSCAN با یک نقطه اولیه دلخواه شروع می‌شود که بازدید نشده است. محدوده این نقطه با استفاده از فاصله epsilon استخراج می‌شود (تمام نقاطی که در فاصله ε هستند نقاط هم‌گروه یا همسایه هستند). باید به یادداشت که الگوریتم برای پیدا کردن همسایگی در یک فضای دوبعدی و سه‌بعدی از فاصله اقلیدسی استفاده می کند به‌این‌ترتیب همسایگی توسط کمترین مقدار فاصله از نقطه اصلی تعریف می‌شود. اگر تعداد کافی نقاط ($minpoints$) در این محدوده وجود داشته باشد، فرایند خوشه سازی شروع می‌شود (border point) و نقطه داده فعلی به اولین نقطه خوشه در خوشه جدید تبدیل می‌شود و در غیر این صورت، نقطه به‌عنوان نویز تلقی می‌شود (بعدها این نقطه نویز ممکن است بخشی از خوشه شود). در هر دو مورد این نقطه به‌عنوان بازدید شده مشخص می‌شود. برای این اولین نقطه در خوشه جدید، نقاط در محدوده ε فاصله آن نیز بخشی از یک خوشه است. این روش برای ساخت همه نقاط در گروه ε متعلق به یک خوشه مشابه است و سپس برای همه نقاط جدید که فقط به گروه خوشه اضافه‌شده‌اند تکرار می‌شود. این فرایند در مراحل ۲ و ۳ تکرار می‌شود تا تمام نقاط در خوشه‌ها وارد شوند یعنی همه نقاط در محدوده ε خوشه‌ای بازدید شده و برچسب‌گذاری شده‌اند [۲۳،۲۴].

کد الگوریتم خوشه‌بندی DBSCAN به‌صورت زیر می‌باشد:

DBSCAN Algorithm

$D_{unprocessed} \leftarrow D$ // D: data-set
$no_of_clusters \leftarrow 0$
while $D_{unprocessed} \neq \emptyset$ do
  Arbitrarily select a $p \in D_{unprocessed}$
  if p is a non-core point, then
    - Mark p as noise point
    - $D_{unprocessed} \leftarrow D_{unprocessed} - \{p\}$
  else //p is a core point
    - $no_of_clusters$ ++
    - $D_{DR}(p) \leftarrow$ Determine all ***Density-Reachable*** points in D from p
    //note: The border points that may have been marked as noise, now belong to $D_{DR}(p)$.
    - $Cluster_{no_of_clusters} \leftarrow \{p\} + D_{DR}(p)$
    - $D_{unprocessed} \leftarrow D_{unprocessed} - Cluster_{no_of_clusters}$
  end-if
end-while

شکل ۳. کد الگوریتم DBSCAN

۴،۲ استفاده از روش اقلیدسی برای محاسبه شباهت بین خوشه‌ها

خوشه‌بندی فرایند خودکارسازی است که در طی آن، اشیا به دسته‌هایی که اعضای آن‌ها ازنظر شاخص‌های موردنظر مشابه یکدیگر باشند تقسیم می‌شوند، بنابراین برای سنجش شباهت بین اشیای داده از اندازه‌گیری فاصله استفاده می‌شود. روش‌های مختلفی برای اندازه‌گیری فاصله بین دو شی وجود دارد که فاصله اقلیدسی معروف‌ترین و پرکاربردترین گونه فاصله است که به‌صورت رابطه زیر محاسبه می‌شود:

$$d = \sqrt{(X_{i1} - X_{c1})^2 + (X_{i2} - X_{c2})^2 + \cdots + (X_{in} - X_{cn})^2} \quad (۶)$$

۵. وزن دهی صفحات

در این مرحله باید صفحات وب به‌دست‌آمده را وزن دهی کنیم. وزن دهی را با استفاده از الگوریتم page rank انجام می‌دهیم. در این روش یک‌بار به هر سند وب امتیاز اختصاص می‌دهد و از این امتیاز، با در نظر گرفتن یا بدون در نظر گرفتن معیاری با توجه به پرس و جوی کاربر جهت رتبه‌بندی اسناد استفاده می‌کند. این الگوریتم رتبه هر صفحه را با اختصاص وزن به پیوندی که به آن صفحه داده‌شده است به دست می‌آورد. مقدار این وزن به کیفیت صفحه‌ای که پیوند در آن قرارگرفته، بستگی دارد. در این صورت پیوندهای صفحات مهم‌تر وزن بیشتری می‌گیرند. جهت مشخص کردن کیفیت صفحه‌های رجوع کننده، در Page Rank از رتبه آن صفحه که به‌صورت بازگشتی تعیین و مقدار اولیه آن اختیاری است، استفاده می‌شود. اگر n سند در دسترس باشد، مقدار اولیه رتبه سند را می‌توان برابر 1/n در نظر گرفت. رتبه هر صفحه مانند P طبق فرمول زیر محاسبه می‌شود که B_p مجموعه همه صفحات اشاره‌کننده به P می‌باشد در این رابطه ε مقدار ثابتی بین ۰،۱ و ۰،۲، n تعداد گره‌ها در گراف G (تعداد صفحات وب در مجموعه) و Outdegree(Q) تعداد پیوندهای خروجی موجود در صفحه Q است. رتبه مرحله j صفحه P_i طبق فرمول زیر محاسبه می‌شود:

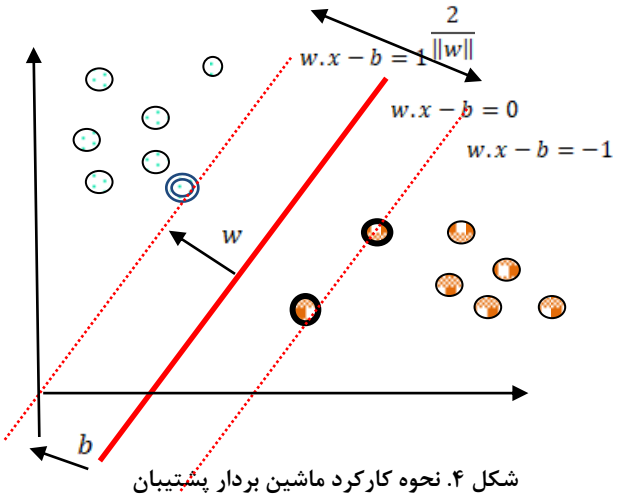
$$r_j = \frac{\varepsilon}{n} + (1-\varepsilon)\sum_{Q \in B_{pi}} \frac{r_{j-1}(Q)}{Outdgree(Q)} \quad (۷)$$

۶. روش طبقه‌بندی ماشین بردار پشتیبان

در این مرحله باید داده‌های خود را با استفاده از svm دسته‌بندی کنیم. ماشین بردار پشتیبانی به دلیل عملکرد کارآمد آن در طبقه‌بندی داده‌ها، به‌طور گسترده در کاربردهای واقعی مورداستفاده قرارگرفته است [16]. روش کار svm بدین‌صورت می

باشد که ما فرض می‌کنیم مجموعه نقاط داده $\{(x_1, c_1), (x_2, c_2), \dots, (x_n. c_n)\}$ را در اختیارداریم و می خواهیم آن‌ها را به دوطبقه $c_i = \{-1,1\}$ تفکیک کنیم. هر x_i یک بردار p بعدی از اعداد حقیقی است که درواقع همان متغیرهای بیانگر رفتار نرم‌افزار هستند[۲۵،۲۶]. روش‌های طبقه‌بندی خطی، سعی دارند که با ساختن یک ابر سطح (که عبارت است از یک معادله خطی)، داده‌ها را از هم تفکیک کنند. روش طبقه‌بندی ماشین بردار پشتیبان که یکی از روش‌های طبقه بندی خطی است، بهترین ابر سطحی را پیدا می‌کند که با حداکثر فاصله (maximum margin)، داده‌های مربوط به دو طبقه را از هم تفکیک کند. به‌منظور درک بهتر مطلب، در شکل ۴، تصویری از یک مجموعه داده متعلق به دو کلاس نشان داده‌شده که روش ماشین بردار پشتیبان بهترین ابر سطح را برای جداسازی آن‌ها انتخاب می‌کند.

درروش ماشین بردار پشتیبان، بردارهای ورودی به یک فضای چندبعدی نگاشت می‌شوند. پس‌ازآن، یک ابر سطح ساخته خواهد شد که با حداکثر فاصله ممکن، بردارهای ورودی را از هم جدا خواهد کرد. به این ابر سطح، ابر سطح با حداکثر مرز جداکننده گفته می‌شود. همان‌گونه که در شکل ۳، نشان داده‌شده است، دو ابر سطح موازی در دو سمت ابر سطح با حداکثر مرز جداکننده ساخته خواهد شد که داده‌های مربوط به دو طبقه را به‌گونه‌ای از هم مجزا می‌کنند که هیچ داده‌ای در مرز بین این دو ابر سطح قرار نمی‌گیرد. ابر سطح با حداکثر مرز جداکننده، ابر سطحی است که فاصله بین دو ابر سطح موازی را به حداکثر می‌رساند. فرض بر این است که هرچقدر مرز جداکننده یا در واقع، فاصله بین دو ابر سطح موازی بیشتر باشد، خطای طبقه‌بندی هم کمتر خواهد بود.



شکل ۴. نحوه کارکرد ماشین بردار پشتیبان

۷. ایجاد نمایه کاربران

سیستم توصیه گر برای پیش‌بینی علایق و ارائه توصیه به کاربران، طراحی و پیاده‌سازی شده است. در هر یک از این سیستم‌ها با توجه به حوزه کاری و اهداف، مجموعه ایی از تکنیک‌های ساخت، به‌روزرسانی و استخراج داده‌ها به کار گرفته‌شده است ولی محور اساس در تمامی این سیستم‌ها پروفایل کاربر است. چگونگی ساخت پروفایلی که در ساخت توصیه‌ها استفاده خواهد شد، پروفایل پیش‌فرض سیستم برای کاربران، نحوه بروز رسانی اطلاعات پروفایل و منبع این به‌روزرسانی فاکتورهایی هستند که در طراحی یک سیستم توصیه گر جایگاه مهمی دارند. فرض می‌کنیم $s1, s2, \dots sk$ مجموعه نشست‌های مربوط به کاربر i-ام (ui) باشد. جهت ایجاد نمایه کاربران، بردار میانگین sui برای کاربر ui به‌عنوان نماینده محاسبه می‌شود و در واقع‌نمایشی از صفحات موردعلاقه کاربر می‌باشد. وزن هر صفحه وب در بردار میانگین، از میانگین وزن آن صفحه وب در همه نشست‌های کاربر $(s1, s2, \dots sk)$ به دست می‌آید.

۸. تولید لیست پیش‌بینی و پیشنهاد به کاربر با استفاده از سیستم توصیه گر ترکیبی

در سیستم‌های توصیه گر ترکیبی برای رسیدن به بالاترین کارایی، بر اساس یک استراتژی معین تکنیک‌های مختلف با یکدیگر ترکیب می‌شوند [۲۷]. دو الگوریتم پالایش مشارکتی و پالایش مبتنی بر دانش را باهم ترکیب می‌کنیم، نتیجه سیستمی خواهد بود که به‌واسطه جزء مبتنی بر دانش، می‌تواند بر شروع سرد الگوریتم پالایش مشارکتی غلبه کند و باوجود جزء مبتنی بر پالایش مشارکتی و قدرت بالای آن دریافتن اولویت‌های مشابه کاربران می تواند توصیه‌هایی کند که هیچ سیستم توصیه گری قادر به توانایی ساخت آن نخواهد بود. با توجه به اینکه الگوریتم‌های پالایش مشارکتی و مبتنی بر محتوا به گونه ایی با مشکل شروع سرد روبرو هستند که این مشکل هم به دلیل نبود داده‌های کافی از آیتم‌ها و هم به دلیل نبود داده‌های کافی از کاربران و یا هردو مورد بروز می کند و ازیک‌طرف تا زمانی که کاربران به آیتم‌های جدید امتیاز ندادند نمی‌توان آن‌ها را وارد لیست‌های توصیه کرد. از طرف دیگر کاربران جدیدی که خریدی نکرده‌اند و به آیتم‌ها امتیاز نداده‌اند نیز این مشکل را دارند که این مشکل را می‌توان با ترکیب الگوریتم‌های توصیه گر مختلف از بین برد. به‌هرحال مشکل شروع سرد مانع از این می‌شود که کاربران جدید از مزایای الگوریتم‌های مشارکتی و مبتنی بر محتوا به‌طور کامل بهره‌مند شوند. سیستم‌های مبتنی بر الگوریتم‌های پالایش مشارکتی و مبتنی بر محتوا، از بهترین سیستم‌ها برای کاربران اختصاصی هستند که تمایل دارند با صرف زمان، سیستم را بر اساس اولویت‌ها و علایقشان سازگار و شخصی کنند. در این زمینه سیستم‌های مبتنی بر دانش مشکل کمتری دارند، زیرا این‌گونه سیستم‌ها به پیشینه کاربر توجهی ندارند و از دیدگاه آن‌ها یک کاربر جدید با کاربری که سابقه فعالیت‌ها و

خریدهایش در سیستم ثبت‌شده تفاوت چندانی ندارد. در این میان سیستم‌های توصیه گر ترکیبی به آن دسته از سیستم‌هایی گفته می‌شود که از ترکیب یک یا چند الگوریتم برای رسید به بالاترین کارایی استفاده می‌کنند. تحقیقات صورت گرفته نشان از آن دارد که سیستم‌های ترکیبی سیستم‌های بسیار موفقی هستند.

۹. ارزیابی نتایج روش پیشنهادی

اغلب برای اعتبارسنجی سیستم‌های توصیه گر به مواردی همچون کارایی روش خوشه‌بندی و دقت سیستم و فراخوانی استفاده می‌شود. در این تحقیق نیز از این معیارها برای ارزیابی سیستم استفاده گردیده است. دقت و فراخوانی در سیستم‌های توصیه گر با استفاده از دو رابطه زیر محاسبه می‌شود.

دقت با استفاده از رابطه زیر محاسبه می‌گردد:

دقت برابر با تعداد تشخیص‌های درست سیستم بر تعداد مجموعه بازیابی شده است:

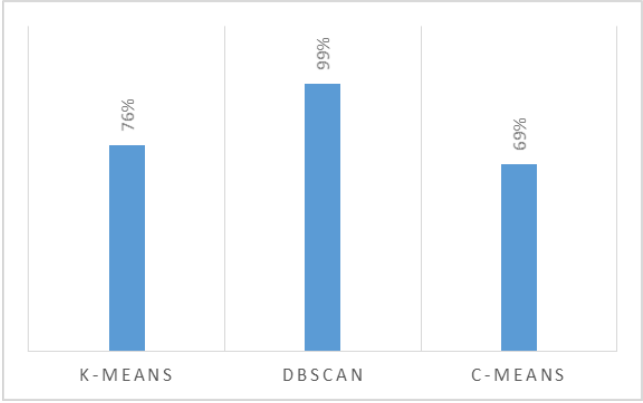
$$Precision = \frac{|\{relevant\ pages\} \cap \{retrved\ pages\}|}{|\{retrived\ pages\}|} \quad (۸)$$

فراخوانی با استفاده از رابطه زیر محاسبه می‌گردد:

فراخوانی برابر با نسبت تعداد تشخیصات درست سیستم بر تعداد کل مجموعه سیستم معیار است:

$$Recall = \frac{|\{relevant\ pages\} \cap \{retrved\ pages\}|}{|\{relevant\ pages\}|} \quad (۹)$$

برای ارزیابی میزان کارایی الگوریتم DBSCAN مقایسه ایی بین این روش و الگوریتم k-means صورت گرفت که نتایج حاصل از ارزیابی میزان کارایی حاکی از آن است که کارایی الگوریتم خوشه‌بندی پیشنهادی حدود ۰/۹۹ می‌باشد این در حالی است که الگوریتم خوشه‌بندی K-MEANS امتیاز کارایی ۰/۷۶ را به دست آورد که نشان از کارایی الگوریتم DBSCAN نسبت به سایر الگوریتم‌های خوشه‌بندی می‌باشد.

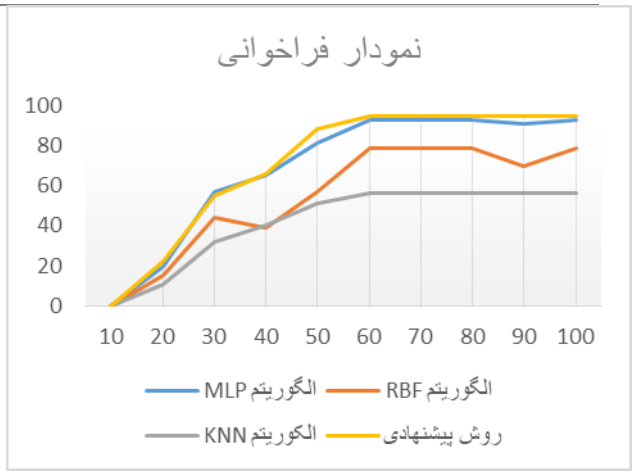


شکل۵. نمودار کارایی الگوریتم خوشه‌بندی DBSCAN و K-MEANSو C_MEANS

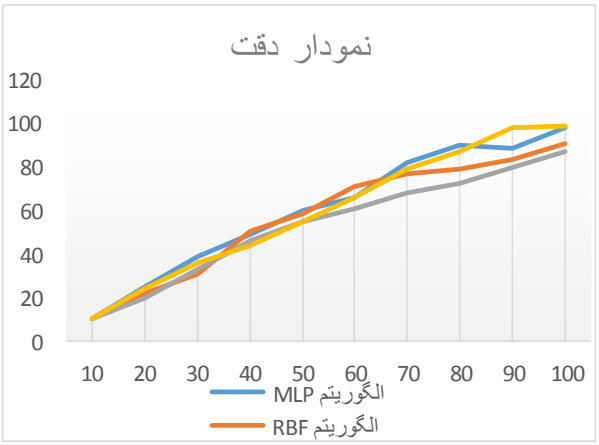
همچنین برای ارزیابی میزان دقت و میزان فراخوانی روش پیشنهادی با سایر روش‌های موجود، مقایسه ای بین روش پیشنهادی و الگوریتم های KNN، RBF، MLP، TDM و LCS (مراجع ۶ و ۱۷) صورت گرفت که نتایج ارزیابی حاکی از آن بود که روش پیشنهادی از روش های موجود توانست میزان دقت و فراخوانی بالاتری را کسب کند.

جدول ۴. مقایسه روش پیشنهادی با سایر روش‌ها

نام الگوریتم	فراخوانی	دقت
الگوریتم KNN	56.04%	87.04%
الگوریتم RBF	79.01%	91.08%
الگوریتم MLP	94%	98%
الگوریتم TDM	93%	92%
الگوریتم LCS	90%	89%
روش پیشنهادی	95%	99%

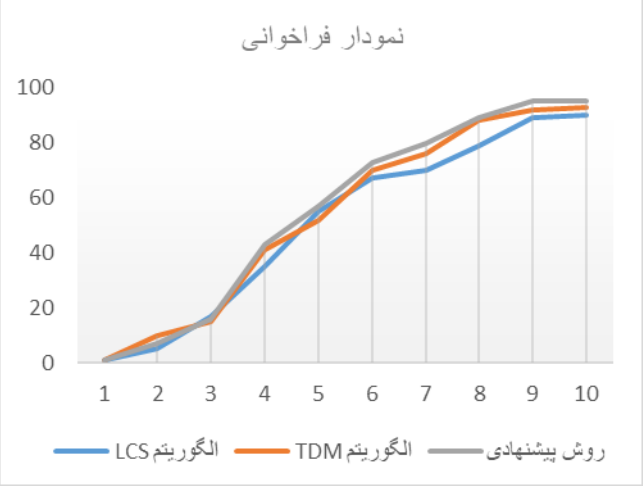


شکل ۶. نمودار مقایسه فراخوانی روش پیشنهادی با سایر روش ها

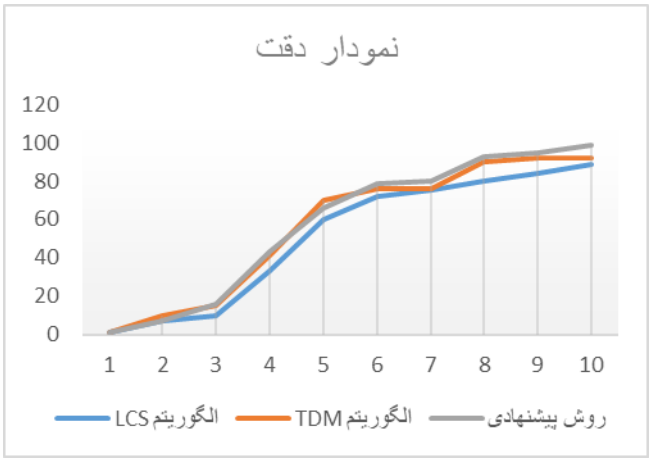


شکل ۷. نمودار مقایسه دقت روش پیشنهادی با سایر روش ها

در جدول زیر مقایسه ای بین روش پیشنهادی و دو مورد از روش‌های پیشنهادی موجود در پیشنیه تحقیق مربوط به الگوریتم‌های TDM و LCS (مراجع ۶ و ۱۷) برای مقایسه در قسمت‌های میزان دقت و میزان فراخوانی صورت گرفت که در این قسمت روش پیشنهادی عملکرد بالاتری را نسبت به این دو روش به دست آورد. نتایج حاصل از این مقایسه بدین‌صورت است که روش پیشنهادی در قسمت فراخوانی توانست عملکرد ۹۵٪ را به دست آورد و الگوریتم‌های TDM و LCSبه ترتیب ۹۳٪ و ۹۰٪ را به دست آوردند. در قسمت میزان دقت نیز روش پیشنهادی توانست عملکرد ۹۹٪ را به دست آورد این در حالی است که الگوریتم‌های TDM و LCSبه ترتیب ۹۲٪ و ۸۹٪ را به دست آوردند. نتایج حاصل از این مقایسه را در شکل ۸ و ۹ مشاهده می‌کنید.



شکل ۸. نمودار مقایسه میزان فراخوانی روش پیشنهادی با سایر روش‌ها

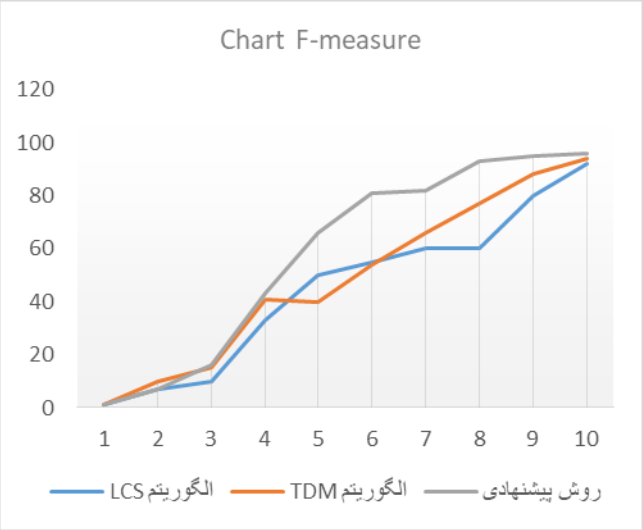


شکل ۹. نمودار مقایسه میزان دقت روش پیشنهادی با سایر روش‌ها

درنهایت برای ارزیابی کلی سیستم از معیار ارزیابی F-measure استفاده خواهیم کرد. این معیار که از ترکیب دو معیار Recall یا فراخوانی و Precision یا دقت، می‌باشد برای ارزیابی کل سیستم استفاده می‌گردد. معیار F-measure را از طریق فرمول زیر محاسبه می‌کنیم.

$$F - measure = \frac{2 * (Recall * Precision)}{(Recall + Precision)} \quad (۱۰)$$

نتایج حاصل مقایسه بین روش پیشنهادی و دو مورد از روش‌های پیشنهادی موجود در پیشنیه تحقیق مربوط به الگوریتم‌های TDM و LCS (مراجع ۶ و ۱۷) با استفاده از معیار F-measure را در شکل زیر مشاهده می‌کنید. در ارزیابی کلی سیستم با استفاده از معیار F-measure مشخص‌شده که روش پیشنهادی در این بخش نسبت به الگوریتم‌های TDM و LCS عملکرد بهتری را به دست آورد. بدین‌صورت که الگوریتم روش پیشنهادی توانست عملکرد ۹۶٪ را به دست آورد، این در حالی است که الگوریتم‌های TDM و LCS به ترتیب عملکرد ۹۴٪ و ۹۲٪ را به دست آوردند. نتایج حاصل از این مقایسه را در شکل (۱۰) مشاهده می‌کنید.



شکل ۱۰. نمودار مقایسه روش پیشنهادی با سایر روش‌ها با استفاده از معیار F-measure

با توجه به نتایج به‌دست‌آمده به این نتیجه می‌رسیم که روش پیشنهادی هم از جانب الگوریتم خوشه‌بندی و هم از جانب دو فاکتور اساسی دقت و فراخوانی، همچنین در ارزیابی کلی سیستم با استفاده از معیار F-measure از سایر الگوریتم‌های موجود نتایج بهتری را کسب کرده است.

۱۰. نتیجه‌گیری

سیستم‌های توصیه گر یا پیشنهاددهنده سیستم‌هایی هستند که با گرفتن اطلاعات محدودی از کاربر می‌توانند پیشنهادات مناسبی به کاربر ارائه دهند و کاربر را در هنگام مواجه با مشکلات سیستم‌های توصیه گر یاری دهند. این سیستم‌ها با شناسایی رفتار کاربران می‌توانند علائق آن‌ها را کشف کنند و با استفاده از این اطلاعات پیشنهادات مناسبی را به کاربر پیشنهاد دهند. در این تحقیق به ارائه روش جدیدی به‌منظور بهبود سیستم‌های توصیه گر در زمینه وب پرداخته شد و سعی شد که با پوشش قرار دادن مشکلات و ایرادات سیستم‌های پیشین، به ارائه روشی جهت بهبود سیستم های توصیه گر پرداخته شود. با مقایسه روش پیشنهادی و موردمطالعه و بررسی نتایج به‌دست‌آمده با روش‌های پیشین، سیستم موردنظر از عملکرد قابل‌قبول‌تری برخوردار هست. نتایج حاکی از تحقیقات صورت گرفته در هنگام استفاده از روش پیشنهادی نشان داد که خوشه‌بندی با استفاده از الگوریتم خوشه‌بندی DBSCAN با توجه به ویژگی‌هایی که نسبت به سایر الگوریتم‌های خوشه‌بندی مانند K_MEANS و C_MEANS دارد می‌تواند خوشه‌بندی را با عملکرد بهتری انجام دهد تا جایی که الگوریتم خوشه‌بندی DBSCAN، امتیاز کارایی ۰/۹۹ را به دست آورد اما الگوریتم خوشه‌بندی K_MEANS، امتیاز ۰/۷۶ را به دست آورد. همچنین با توجه به عملکرد سیستم موردنظر و نتایج به‌دست‌آمده مشخص شد که سیستم توصیه گر ترکیبی مورداستفاده در تحقیق از عملکرد قوی‌تری نسبت به سایر سیستم‌های توصیه گر که به‌صورت انفرادی مورداستفاده قرار می‌گیرند ازلحاظ دقت و فراخوانی عملکرد بهتری داشته است. در قسمت پیش‌بینی نیز استفاده از روش یادگیری ماشین موردبحث در تحقیق یعنی SVM، می‌تواند به‌صورت قابل قبولی جهت پیش‌بینی صفحات به کاربر مورداستفاده قرار گیرد. نتایج کلی آزمایش نیز حاکی از آن می‌باشد که سیستم مورداستفاده در تحقیق در خوشه‌بندی و دقت و پیشنهاد صفحات نسبت به سایر روش‌های پیشین از عملکرد بهتر و قابل قبولی برخوردار می‌باشد.

مراجع

[1] M. Heupel, L. Fischer, Bourimi, M. Scerri, S.: Ontology-enabled access control and privacy recommendations. In: Mining, Modeling, and Recommending ‘Things’ in Social Media, pp. 35–54. Springer,2015.

[2] L. Simões, L. Shah, V. Silva, J. Rodrigues, N., Leite, N., Lopes, N. New Performance Metrics for Offline Content-Based TV Recommender System. In: Boratto, L., Faralli, .2021.

[3] B. Twardowski, Zawistowski, P. Zaborowski, S. Metric Learning for Session-Based Recommendations. In: Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science, vol 12656. Springer, Cham.2021.

[4] R. Alhajj, J. Rokne, (eds). Recommender Systems. In:Encyclopedia of Social Network Analysis and Mining. Springer,2018.

[5] J. Neidhardt, Kuflik, T. Wörndl, W. Special section on recommender systems in tourism. Inf Technol Tourism 19, 83–85 ,2018.

[6] F. Maazouzi, Zarzour, H. Jararweh, Y.. An effective recommender system based on clustering technique for ted talks. International Journal of Information Technology and Web Engineering (IJITWE), 15(1), 35-51,2020.

[7] J. Wu, He, X Wang, X. et al. Graph convolution machine for context-aware recommender system. Front. Comput. Sci. 16, 166614 ,2022.

[8] B. Alhijawi, G. AL-Naymat,. Novel Positive Multi-Layer Graph Based Method for Collaborative

Makeup Recommendation," Proc. 31st AAAI Conference on Artificial Intelligence, pp.941–947, 2017.

[18] K.D. Gupta,. A Survey on Recommender System. International Journal of Applied Engineering Research, 14(14), 3274-3277, 2019.

[19] S. Dara, Chowdary, C. R., Kumar.. A survey on group recommender systems. *Journal of Intelligent Information Systems, 54*(2), 271-295, 2020.

[20] G. De Giacomo, D. Lembo, Lenzerini, M., Poggi, A., & Rosati, R.. Using ontologies for semantic data integration. In *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years* (pp. 187-202). Springer, Cham, 2018.

[21] A. Kumar, Sangwan, S. R. Nayyar,.. Multimedia social big data: Mining. In *Multimedia Big Data Computing for IoT Applications* (pp. 289-321). Springer, Singapore, 2020.

[22] A. Zatni, K. abdelkarim . .Document text Detection in video frames acquired by a smartphone based on line segment detector and DBSCAN clustering .Journal of engineering science and technology,vol.13,no.2,540-557,2018.

[23] A. Siddharth Agrawal. Machine learning-DBSCAN. Toward Data Science, 2019.

[24] K.Shao, Fu, W. Tan, J. Wang, K.. Coordinated approach fusing time-shift multiscale dispersion entropy and vibrational Harris hawks optimization-based SVM for fault diagnosis of rolling bearing. *Measurement*, *173*, 108580, 2021.

[25] X. Zhang, Li, C. Wang H. XWu. A novel fault diagnosis procedure based on improved symplectic geometry mode decomposition and optimized SVM. *Measurement*, *173*, 108644,2021.

[26] M. Riyahi, M.K. Sohrabi, Providing effective recommendations in discussion groups using a new hybrid recommender system based on implicit ratings and semantic similarity. *Electronic Commerce Research and Applications*, *40*, 100938, 2020.

Filtering Recommender Systems. J. Comput. Sci. Technol. 37, 975–990 ,2022.

[9] S. Chawla, (2018). Web page recommender system using hybrid of genetic algorithm and trust for personalized web search. *Journal of Information Technology Research (JITR)*, *11*(2), 110-127, 2018.

[10] A. Tanwar, Vishwakarma, D.K. A deep neural network-based hybrid recommender system with user-user networks. Multimed Tools Appl ,2022.

[10] V. Boppana, P. Sandhya, Web crawling based context aware recommender system using optimized deep recurrent neural network. J Big Data 8, 144 ,2021.

[11] M. Wischenbart, S. Firmenich, Rossi, G. et al. Engaging end-user driven recommender systems: personalization through web augmentation. Multimed Tools Appl 80, 6785–6809 ,2021.

[12] O. Bourkoukou, Outmane, and Omar Achbarou. "Weighting based approach for learning resources recommendations." *JOIV: International Journal on Informatics Visualization* 2, no. 3 .104-109, 2018.

[13] M. Riyahi, M.K. Sohrabi,. Providing effective recommendations in discussion groups using a new hybrid recommender system based on implicit ratings and semantic similarity. *Electronic Commerce Research and Applications*, *40*, 100938, 2020.

[14] D. S. Sisodia, S. Verma, and O. P. Vyas, "Augmented intuitive dissimilarity metric for clustering of web user sessions," Journal of Information Science, vol. 43, pp. 480-491, 2017.

[15] X. Xie and B. Wang, "Web page recommendation via twofold clustering: considering user behavior and topic relation," Neural Computing and Applications, vol. 29, pp. 235-243, 2018.

[16] R. Wagh, J. Patil.. A Novel Web Page Recommender System for Anonymous Users Based on Clustering of Web Pages. *Asian Journal For Convergence In Technology (AJCT)*, 2019.

[17] T. Alashkar, S. Jiang, S.Wang, and Y. Fu, "Examples-Rules Guided Deep Neural Network for