DOR: 20.1001.1.2538161.1403.9.1.4.3

DOI: 10.22034/csj.2024.197227

بهار ۱۴۰۳، دوره ۹، شماره ۱، پیاپی ۳۲

دریافت مقاله: ۱۴۰۲/۱۱/۱۳

پذیرش مقاله: ۱۴۰۳/۰۱/۱۸

نوع مقاله: پژوهشی

یک سیستم توصیه‌گر وب برای پیش‌بینی صفحات موردنظر کاربر با استفاده از رویکردی مبتنی بر وزن‌دهی ترکیبی

رضا مولایی فرد*

دانشکده مهندسی کامپیوتر، دانشگاه علوم و تحقیقات، تهران، ایران

پست الکترونیکی: molaefard@gmail.com

پیام یاراحمدی

دانشکده مهندسی کامپیوتر، دانشگاه ملّی تاواس شوچنکو، کیف، اکراین

پست الکترونیکی: Yarahmadi.pay@gmail.com

جواد محمدزاده

گروه مهندسی کامپیوتر، دانشکده هوش مصنوعی، واحد کرج، دانشگاه آزاد اسلامی، کرج، ایران

پست الکترونیکی: j.mohammadzadeh@kiau.ac.ir

چکیده

با توجه به رشد روزافزون صفحات موجود در سطح وب، وجود سیستمی که اطلاعات موردنیاز را از میان حجم عظیم داده‌ها که روزبه‌روز نیز در حال افزایش می‌باشند استخراج کند ضروری به نظرمی‌رسد. سیستم‌های توصیه‌گر ابزار یا تکنیک‌هایی برای فیلتر کردن انبوه اطلاعات هستند و به کاربران اقلامی را پیشنهاد می‌کنند که برای آن‌ها رضایت‌بخش و موردعلاقه هستند. در سیستم‌های توصیه‌گر با چالش‌هایی مانند شروع سرد و پراکندگی داده‌ها مواجه هستیم که در این مقاله با استفاده از روش مبتنی بر وزن‌دهی و خوشه‌بندی، سعی در رفع این چالش‌ها نمودیم. در این تحقیق به ارائه روش جدیدی به‌منظور بهبود سیستم‌های توصیه‌گر در زمینه وب پرداخته می‌شود که می‌تواند صفحات موردنظر کاربر را پیش‌بینی کند و پیشنهادهای مناسبی را به کاربر ارائه دهد. اساس کار این سیستم توصیه‌گر استفاده از فیلترینگ مشارکتی و موارد مورد جستجوی کاربر درگذشته می‌باشد که با انجام عملیات‌های خوشه‌بندی ترکیبی از دو الگوریتم K-means و C-means و سپس وزن‌دهی صفحات موردعلاقه کاربر، لیستی از صفحات را در اختیار کاربر قرار می‌دهد که کاربر در نظر دارد آن را جستجو کند. طبق تحقیقات صورت گرفته این سیستم توصیه‌گر تا ۸۵ درصد می‌تواند صفحات موردنیاز کاربر را به‌درستی تشخیص و مورد پیش‌بینی قرار دهد.

واژه‌های کلیدی: سیستم توصیه‌گر، وزن‌دهی، داده کاوی، وبکاوی، فیلترینگ مشارکتی

مقدمه

اخیراً وب به منبع بزرگی از اطلاعات تبدیل شده است که با افزایش استفاده از اینترنت و همچنین افزایش وبگاه‌ها باعث بروز مشکلاتی برای کاربران شده است. یکی از مشکلات

* نویسندهٔ مسئول

پیش‌آمده برای کاربران یافتن اطلاعات مورد علاقه و یا مورد نیاز آن‌ها در این حجم انبوه اطلاعات می‌باشد. در این میان رقابت ایجاد شده در بخش تجارت الکترونیکی نیز نیازمند ایجاد وبگاه‌هایی مورد علاقه کاربران می‌باشد [۱,۲,۳]. به همین دلیل روش‌های شخصی‌سازی وب به‌منظور حل این مشکلات پیشنهاد می‌شود تا بتوان با کمک این روش‌ها وبگاه‌هایی موردعلاقه و مطابق با نیازهای کاربران ایجاد نمود. هدف شخصی‌سازی وب تولید پیشنهادهایی پویا مبتنی بر الگوهای رفتاری کاربران می‌باشد. اخیراً به‌منظور کشف الگوهای رفتاری کاربران از تکنیک‌های وب‌کاوی استفاده می‌شود. از میان تکنیک‌های وب‌کاوی، تکنیک وب‌کاوی مبتنی بر کاربرد به‌طور گسترده‌ای به‌منظور کشف الگوهای رفتاری کاربران و مدل‌سازی این رفتارها استفاده می‌شود [۴,۵]. این تکنیک از داده‌های موجود در فایل‌های سوابق کارسازها استفاده می‌کند. در حقیقت بدون درخواست صریح کاربران تکنیک‌های وب‌کاوی مبتنی بر کاربرد قادر است از اطلاعات موجود در فایل‌های سوابق علایق کاربر را استخراج نموده و بر اساس آن‌ها به مدل‌سازی رفتار کاربران پردازد و الگوهای رفتاری آن‌ها را تولید کند [۶,۷,۸]. در این زمینه کارهای تحقیقاتی بسیاری انجام شده است که دارای معایب و چالش‌های مختلفی هستند مانند شروع سرد که مشکل اکثر سیستم‌های توصیه‌گر می‌باشد و همچنین مشکل پراکندگی داده‌ها که باعث به دست آوردن نتایج نادرست می‌شود، که در این تحقیق با استفاده از روش‌های خوشه‌بندی و وزن‌دهی سعی در حل این چالش‌ها نمودیم. در این تحقیق پیشنهاد می‌شود از الگوریتم‌های وب‌کاوی و سیستم‌های توصیه‌گر برای این منظور استفاده شود تا بتوان درخواست‌های آینده کاربر را پیش‌بینی کرد و سپس لیستی از صفحات موردعلاقه کاربر تولید شود؛ به‌عبارت دیگر بتوان نمایه‌ای دقیق از رفتار کاربران به دست آورده و صفحه‌ای پیش‌بینی شود که کاربر در حرکت بعدی آن را انتخاب خواهد کرد. این پیشنهادهای پویا می‌توانند همان پیوندهای صفحات باشند. در این تحقیق با بهره‌گیری از روش‌ها داده‌کاوی و وب‌کاوی و در نظر گرفتن ویژگی‌هایی که رفتار کاربر را به صورت دقیق مشخص می‌کند نمایه‌های کاربران ساخته می‌شود، سپس با تکنیک خوشه‌بندی الگوهای حرکتی کاربران را به دست آورده می‌شود؛ به عبارت دیگر با تلفیق خوشه‌بندی و ویژگی‌هایی مانند تاریخ رویت صفحه می‌توان بهترین نمایه‌ای که توصیف‌گر رفتار کاربر می‌باشد را تولید نمود. پس از پیدا کردن الگوهای حرکتی کاربران، با استفاده از خوشه‌بندی، الگوی مناسب برای کاربر پیدا شده و پیشنهادهای مناسب برای درخواست‌های آینده کاربر تولید خواهد شد، سپس با استفاده از سیستم توصیه‌گر مبتنی بر پالایش مشارکتی صفحات مورد علاقه به کاربر توصیه خواهند شد.

ادبیات تحقیق

سیستم‌های توصیه‌گر

سیستم توصیه‌گر یا پیشنهادگر زیرمجموعه‌ای از سامانه پالایش اطلاعات است که به دنبال پیش‌بینی امتیاز یا اولویتی است که کاربر به یک مورد خواهد داد. در سال‌های اخیر سیستم‌های توصیه‌گر بسیار متداول شده و در حوزه‌های مختلفی مورد استفاده قرار گرفته‌اند. علاوه بر این، سیستم‌های توصیه‌گر برای متخصصان، گروه‌های همکاران، طنزپردازی‌ها، رستوران‌ها، خدمات مالی، بیمه عمر، مسائل عاطفی و صفحات توئیتر نیز ارائه شده است. به طورکلی وظیفه استفاده از سیستم‌های توصیه‌گر و فیلترینگ اطلاعات، تبدیل اطلاعات موجود در مورد کاربران و اولویت‌های آن‌ها جهت پیش‌بینی علاقه آنها در آینده می‌باشد. در واقع در این سیستم از بین موردهای مختلف بر اساس این که چه موردی برای کاربر جذاب است، از یک مجموعه بزرگ از موارد و کاربران فیلتر می‌کنند [۹].

داده‌کاوی

«**داده‌کاوی**» ترجمه عبارت «Data Mining» و به معنای

«**کاویدن معادن داده**» اســت. داده‌کاوی یعنی اســتخراج اطلاعــات گرانبها از حجم عظیم معادن داده. کلمهٔ Mining در معنای تحت اللفظی خود یعنی «اســتخراج از معدن» به کار می‌رود و در واقع عبارت Data Mining نشان می‌دهد کــه حجم انبوه اطلاعات مانند یــک معدن عمل می‌کند و از ظاهر آن مشــخص نیســت چه عناصر گرانبهایی در عمق این معدن وجود دارد. تنها با کندوکاو و استخراج این معدن است که می‌توان به آن عناصر گرانبها دست پیدا کرد[۱۰].

وب‌کاوی

کاوش محتوای وب فرآیند اســتخراج اطلاعات مفید از محتوای مستندات وب است. محتوای یک سند وب متناظر با مفاهیمی است که آن سند در صدد انتقال آن به کاربران است. این محتوا می‌تواند شامل متن، تصویر، ویدئو، صدا و یا رکوردهای ساخت‌یافته مانند لیست‌ها و جداول باشد [۱۱]. در حقیقت وب‌کاوی ســعی بــر پیش‌بینی آینده دارد در حالی که مدل تحلیل پیش‌نگر فقط به بررســی ســوابق داده‌ها می‌پردازد. وب‌کاوی با دســته‌بندی اسناد اینترنتی و شناســایی صفحات وب، قدرت موتور جستجو را بهبود می‌بخشد. در فرایند جستجوی اینترنتی نظیر گوگل و یاهو و در فرایند جســتجوی عمودی نظیر FatLens و Become مورد اســتفاده قرار می‌گیرد. از وب‌کاوی برای پیش‌بینی رفتار کاربر نیز استفاده می‌شود. همچنین وب‌کاوی برای یک وبــگاه و خدمات الکترونیکی خــاص، به طور مثال به منظور بهینه‌سازی صفحه فرود، بسیار سودمند است. به طورکلی، وب‌کاوی را می‌توان به ســه نوع تکنیک مختلف داده‌کاوی تقسیم کرد: کاوش در محتوای وب (محتواکاوی وب)، کاوش در ساختار وب (ساختارکاوی وب) و کاوش در استفاده از وب (استفاده‌کاوی وب).

پیشینه پژوهش

معضوضی و همکاران در مقاله خود در ســال ۲۰۲۰ بــه ارائه روش جدیدی به منظــور بهبود توصیه صفحات وب بــه کاربران پرداختند. ایــن محققان در مقاله خود یک سیستم پیشنهادی موثر ارئه دادند که مبنای آن بر گفتگوی کاربران و گروه‌بندی آنها در گروه‌های مختلف و ســپس توصیه موثر به یک گروه که ویژگی‌های مشــابهی داشتند بود. نویســندگان از روش ضریب همبســتگی پیرسون و مذاکرات TED کاربر استفاده کردند. سپس آنها صفحات را با استفاده از روش خوشه‌بندی K-Means برای گروه‌بندی کاربران و سپس پیشــنهاد به کاربر هدف، مورد استفاده قرار دادند. نتایج حاصل از تحقیق محققان که با اســتفاده از RMSE صورت گرفت، دقت و فراخوانی بهتری نسبت با سایر روش‌های موجود به دست آورد [۱۲].

سوراچی در مقاله خود در سال ۲۰۱۸ به ارائه روشی به مظور بهبود سیستم‌های توصیه‌گر برای شخصی‌سازی صفحات وب پرداخت. این محقق در مقاله خود از ترکیبی از الگوریتم‌های ژنتیک و اعتماد به URLهای کلیک شده و قابل اعتماد برای توصیه صفحات وب اســتفاده نمود. صفحات وب مورد اعتماد کاربران بر اســاس جلسات پرس‌وجوی خوشــه‌ای برای رتبه‌بندی بهینه با GA[1] اســتفاده شــد تا اســناد با ارتباط بیشتر در رتبه‌بندی بازیابی شوند و دقت نتایج بهبود گردد. رتبه‌بندی مطلوب URLهای کلیک شــده قابل اعتماد، اســناد مربوط را به کاربران وب جهت هدف جستجوی خود توصیه می‌کند و نیازهای اطلاعاتی کاربر را به طور گسترده‌ای برآورده می‌کند [۱۳].

بورکوکو و عمر در مقاله خود در ســال ۲۰۱۸ به ارائه یک ســیتم توصیه‌گر به منظور بهبود نتایج با اســتفاده از تاریخچه جستجوی قبلی کاربران پرداختند. این محققان در رویکرد پیشنهادی خود سعی کردند تا منابع یادگیری را به فــرد یادگیرنده با در نظر گرفتن ترجیحات وی و تاریخچه جســتجوهای قبلی وی پیشــنهاد دهند، تاریخچه‌ای که از فایل‌های log اســتخراج شده است. این رویکرد، سبک‌های یادگیــری و روش‌هــای فیلتر کردن مشــارکتی را ترکیب می‌کند تا کیفیت پیشــنهادات را ارتقا دهد. نتایج تحقیق این محققان حاکی از بهبود سیســتم‌های توصیه‌گر در زمینه پیشنهاد صفحات به کاربران بود [۱۴].

1-Genetic algorithm

ریاحی و سهرابی در مقاله خود در سال ۲۰۲۰ به ارائه روشی به منظور بهبود توصیه صفحات وب با استفاده از یک سیستم توصیه‌گر 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 براساس اهمیت معنایی آنها در یک ساختار سلسله‌مراتبی سازماندهی کردند. ساختار سلسله‌مراتبی برای جستجوی برچسب‌های مربوطه در بخش فیلتر محتوامحور مورد استفاده قرار گرفت و درخواست‌های کاربران با استفاده از وب معنایی مرتبط با هم 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 [۱۵].

مولایی فرد و مصلح در مقاله خود درسال ۲۰۲۳ به ارائه روش جدیدی به منظور بهبود سیستم‌های توصیه‌گر در زمینه وب پرداختند. در روش پیشنهادی از الگوریتم خوشه‌بندی DBSCAN جهت خوشه‌بندی داده‌ها استفاده شد. سپس با استفاده از الگوریتم Page rank، صفحات موردعلاقه کاربر وزن‌دهی می‌شوند. سپس با استفاده از روش SVM [۲] ، داده‌ها را دسته‌بندی و جهت تولید پیش‌بینی به کاربر به یک سیستم توصیه‌گر ترکیبی داده شد که درنهایت این سیستم توصیه‌گر 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 درصد را در قسمت فراخوانی و امتیاز ۹۹ درصد را در قسمت دقت به دست آورد که این نتایج اثبات می‌کند که این سیستم توصیه‌گر تا بیش از ۹۰ درصد می‌تواند صفحات موردنظر کاربر را به‌درستی تشخیص داده و تا حدود زیادی نقاط ضعف سایر سیستم‌های پیشین را برطرف سازد[۱۶].

وو و همکاران در مقاله خود در سال ۲۰۲۲ به ارائه روشی به منظور بهبود سیستم توصیه‌گر وب با استفاده از نمودار پیچیدگی کاربر- مورد 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 تشکیل شده است: یک رمزگذار، لایه پیچیدگی گراف و یک رمزگشا. رمزگذار کاربران، موارد و زمینه‌ها را به بردارهای جاسازی می‌کند که به لایه‌های GC منتقل می‌شوند که جاسازی‌های کاربر و مورد را با پیچیدگی‌های گراف آگاه از زمینه در نمودار کاربر- مورد اصلاح می‌کنند. رمزگشا تعبیه‌های تصفیه شده را هضم می‌کند تا با در نظر گرفتن تعاملات بین کاربر، مورد و جاسازی‌های زمینه، امتیاز پیش‌بینی را به دست آورد. آزمایش‌هایی بر روی سه مجموعه داده واقعی از Yelp و Amazon صورت گرفت، که اثربخشی GCM و مزایای انجام پیچش‌های نمودار برای CARS را تأیید می‌کند[۱۷].

تنور و ویشواکارما در مقاله‌ای 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. راه حل این مشکل استفاده از یک سیستم توصیه‌کننده (RS)[۳] است که به شما کمک می‌کند مورد 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)[۴] و سیستم‌های ترکیبی فعلی می‌پردازد و در عین حال 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

2-Support Vector Machines

3-Recommender systems
4-Collaborative filtering

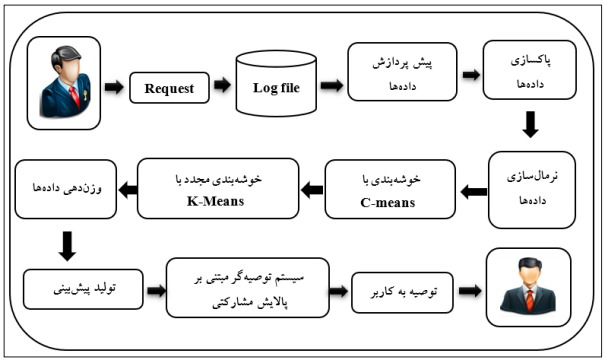
بر شبکه عصبی عمیق استفاده شد که از بردارهای کاربر و مورد برای کپسوله کردن داده‌های کاربران و موارد برای آموزش داده‌های غیرخطی با ابعاد بالا برای ارائه توصیه‌های دقیق‌تر استفاده می‌کند. شبکه‌های کاربر- کاربر برای ارائه یک همکاری و جنبه هم‌افزایی بهتر به این مدل استفاده می‌شوند. در این رویکرد، ترکیب شبکه‌های کاربر- کاربر با شبکه‌های عصبی عمیق، دقت پیش‌بینی بالاتر و زمان اجرای بهتری را نسبت به سایر روش‌های پیشرفته به دست می‌دهد. آزمایش‌های گسترده بر روی مجموعه داده‌های Flixster و MovieLens در دسترس عموم به این نتیجه رسید که این تکنیک با دستیابی به بهبود ۱۹ درصد در RMSE، ۹/۲ درصد در MAE از روش‌های برتر فعلی بهتر عمل می‌کند[۱۸].

الحیجاوی و نایمات در مقاله خود در سال ۲۰۲۲ به بررسی روش‌های بهبوددهنده در زمینه سیستم توصیه‌گر وب پرداختند. این محققان معتقد بودند که فیلتر مشارکتی از نظر دقت موفقیت قابل توجهی 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 چند لایه مثبت (PMLG-RS) را پیشنهاد کردند. این روش شامل یک نمودار چند لایه مثبت و یک الگوریتم جستجوی مسیر برای تولید توصیه‌ها است. نمودار چند لایه مثبت شامل دو لایه متصل است: لایه کاربر و مورد .PMLG-RS نیازمند 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-RS با روش‌های توصیه شناخته شده مبتنی بر سه مجموعه داده معیار نشان‌دهنده برتری PMLG-RS و قابلیت بالای آن در ارائه توصیه‌های مرتبط، جدید و متنوع برای کاربران است[۱۹].

در روش‌های پیشنهادی سابق چالش‌هایی وجود دارد از جمله شروع سرد که مشکل اکثر سیستم‌های توصیه‌گر پیشین است که با استفاده از رویکرد وزن‌دهی ترکیبی سعی در حل این مشکل نمودیم. چالش‌های دیگر پراکندگی داده‌ها و مقیاس‌پذیری داده‌ها بود که با استفاده از روش خوشه‌بندی ترکیبی پیشنهادی این مشکلات مرتفع گردید. از جمله سایر چالش‌ها نیز درصد پایین میزان فراخوانی و دقت سیستم‌های توصیه‌گر می‌باشد که در روش پیشنهادی، سیستم مورد استفاده در این تحقیق و ترکیب الگوریتم‌های خوشه‌بندی، پراکندگی داده‌ها را کاهش داد و پس از خوشه‌بندی مجدد داده‌ها درصد دقت نیز افزایش پیدا کرد که این کار به همراه روش وزن‌دهی ترکیبی باعث به دست آوردن میزان دقت و فراخوانی بیشتری نسبت به روش‌های مشابه به دست آورد.

روش پیشنهادی

در روش پیشنهادی به ارائه روش جدیدی به منظور بهبود سیستم‌های توصیه‌گر در زمینه استخراج اطلاعات در زمینه وب و پیشنهاد آنها به کاربر پرداخته می‌شود. در روش پیشنهادی پس از استخراج فایل سوابق مربوط به کاربران باید این اطلاعات را آماده‌سازی کنیم. از مهمترین مراحل آماده‌سازی داده‌ها می‌توان به عملیاتی مانند پیش‌پردازش داده‌ها، پاکسازی داده‌ها و نرمال‌سازی داده‌ها اشاره کرد. پس از پیش‌پردازش داده باید داده‌ها را خوشه‌بندی کرد. برای انجام خوشه‌بندی از یک روش خوشه‌بندی ترکیبی استفاده می‌کنیم که شامل دو الگوریتم خوشه‌بندی K-means و C-Means می‌باشد که روش ترکیبی پراکندگی داده‌ها را کاهش داد و پس از خوشه‌بندی مجدد داده‌ها، خوشه‌های دقیق‌تری نسبت به مرتبه اول خوشه‌بندی به دست آمد. درصد دقت خوشه در مرتبه اول خوشه‌بندی ۷۱ درصد بود که پس از خوشه‌بندی مجدد داده‌ها درصد دقت خوشه به ۷۸ درصد رسید. سپس خوشه‌های به دست آمده از مرحله قبل را وزن‌دهی می‌کنیم، سپس با استفاده از سیستم توصیه‌گر پالایش مشارکتی، صفحاتی به کاربر پیشنهاد می‌شود که می‌تواند مورد علاقه وی باشد.



شکل (۱): نمایی از روش پیشنهادی

داده‌ها

برای پیاده‌سازی سیستم پیشنهادی ابتدا یک Log file، از درخواست‌های کاربران مختلف را جمع‌آوری کردیم. برای این کار از Log file مربوط به ناسا که در وبگاه http://ita.ee.lbl.gov/html/contrib/NASAHTTP، قابل‌استخراج می‌باشد استفاده شد. فایل ثبت ناسا شامل بیش از ۸۰۰۰ جستجو مربوط به کاربران در یک بازه مشخص بود که پس از عملیات پیش‌پردازش داده‌ها فایلی حاوی ۷۳۶۷ 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 ۳۴۰۸ نشست استخراج گردید که با پاکسازی این نشست‌ها به ۲۹۹۸ نشست تبدیل شدند.

داده‌های فایل سوابق شامل ۵ ویژگی مشترک می‌باشند که این ویژگی‌ها عبارتند از:

- Host: میزبانی که درخواست موردنظر از جانب آن ارسال شده است. این بخش دارای ۲ نوع میزبان می‌باشد. در صورت وجود میزبانی، نشانی آن نشان داده می‌شود و در غیر این صورت نشانی IP مربوطه نشان داده می‌شود.
- 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.

پیش‌پردازش داده‌ها

پس از جمع‌آوری Log fileهای مربوط به جستجوی کاربران در گذشته ابتدا باید عملیات پیش‌پردازش داده‌ها را انجام دهیم. زیرا نمی‌توان داده‌های خام را به صورت مستقیم وارد الگوریتم‌های داده کاوی کرد. پیش‌پردازش داده‌ها یکی از مهمترین مراحل وب‌کاوی می‌باشد که امری ضروری است. وظیفه این بخش از سیستم پاکسازی داده، شناسایی نشست‌های کابر می‌باشد [۲۲].

پاکسازی داده‌ها

به دلیل آن که همه داده‌های موجود در فایل سوابق مربوط به کاربران مناسب نیستند، باید مقادیر اضافی مورد حذف قرار گیرند تا از فایل‌های بدون استفاده پاکسازی گردند.

نرمال‌سازی داده‌ها

برای نرمال‌سازی داده‌ها از روش Min-Max استفاده می‌کنیم. در این روش هر مجموعه‌ای از داده‌ها به بازه‌ای دلخواه، که کمترین و بیشترین مقدار آن از قبل مشخص است نگاشت می‌شود. در این روش می‌توان هر بازه دلخواه را تنها با یک تبدیل ساده به بازه جدید نگاشت کرد. به طور مثال فرض کنیم می‌خواهیم ویژگی A، از مجموعه داده‌هایی که در بازه بین Min-A تا Max-A قرار دارد، به بازه جدید new-Min تا new-Max نگاشت شود. برای این منظور، هر مقدار اولیه مانند V در بازه جدید اولیه، طبق رابطه زیر به مقدار جدید V′ در بازه جدید تبدیل خواهد شد [۲۰].

رابطه (۱)

$$V'=(v-minA)\frac{newMax-newMin}{maxA-minA}+(newMin)$$

خوشه‌بندی

در مرحله بعد باید داده‌های پیش‌پردازش شده از مرحله قبل را خوشه‌بندی کنیم. روش خوشه‌بندی مورد استفاده در تحقیق استفاده از یک روش خوشه‌بندی ترکیبی می‌باشد. بدین‌صورت که ابتدا داده‌ها را با استفاده از الگوریتم خوشه‌بندی C-means، خوشه‌بندی می‌کنیم. سپس با استفاده از خوشه‌بندی K-means یک خوشه‌بندی مجدد روی داده‌ها انجام می‌دهیم که باعث افزایش دقت خوشه‌بندی می‌شود.

خوشه‌بندی با c-means

در این الگوریتم تعداد خوشه‌ها (c) از قبل مشخص شده است. تابع هدفی که برای این الگوریتم تعریف شده به صورت زیر می‌باشد [۲۱].

رابطه(۲)

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

در فرمول فوق m یک عدد حقیقی بزرگتر از ۱ است که اکثر موارد برای m عدد ۲ انتخاب می‌شود. X_k هم، رتبه K ام است و V_i نماینده یا مرکز خوشه iام است. U_{ik} میزان تعلق نمونه iام در خوشه kام را نشان می‌دهد. علامت $\|*\|$میزان تشابه (فاصله) نمونه از مرکز خوشه می‌باشد که می‌توان از هر تابعی که بیانگر تشابه نمونه و مرکز خوشه باشد را استفاده کرد. از روی U_{ik} می‌توان یک ماتریس U تعریف کرد که دارای c سطر و n ستون می‌باشد و مولفه‌های آن هر مقداری بین ۰ و ۱ را اختیار کنند اما مجموع مولفه‌های هر یک از ستون ها برابر ۱ باشد و داریم:

رابطه (۳) $\sum_{i=1}^{c} u_{ik} = 1, \forall k = 1, \dots, n$

معنای این شرط این است که مجموع تعلق هر نمونه به c خوشه باید برابر ۱ باشد. با استفاده از شرط فوق و مینیمم کردن تابع هدف خواهیم داشت:

رابطه (٤) $$u_{ik} = \frac{1}{\sum_{k=1}^{n}\left(\frac{d_{ik}}{d_{jk}}\right)^{2/(m-1)}}$$

رابطه (٥) $$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m}}{\sum_{k=1}^{n} u_{ik}^{m}}\, x_k$$

خوشه‌بندی مجدد با k-means

در مرحله بعد خوشه‌های به دست آمده از مرحله قبل را به صورت مجدد خوشه‌بندی می‌کنیم تا دقت خوشه افزایش یابد. برای این کار از الگوریتم خوشه‌بندی K-means استفاده می‌کنیم. اساس کار این الگوریتم براساس حداقل فاصله داده‌ها از مرکز خوشه‌هایی که به صورت تصادفی و از بین داده‌های موجود انتخاب می‌شود، عمل خوشه‌بندی را انجام می‌دهد. گام‌های این الگوریتم به صورت زیر می‌باشد:

- ابتدا k میانگین یعنی، $(\mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}, \dots, \mu_k^{(0)})$ را نماینده خوشه‌ها هستند، به صورت تصادفی مقداردهی می‌کنیم. سپس، دو مرحله زیر را چندین بار اجرا می‌کنیم تا میانگین‌ها به یک ثبات کافی برسند و یا مجموع واریانس‌های خوشه‌ها تغییر چندانی نکنند [۱۰]:
- از میانگین‌ها k خوشه می‌سازیم، خوشه iام در زمان t تمام داده‌هایی هستند که از لحاظ اقلیدسی کمترین فاصله را با میانگین $\mu_i^{(t)}$ یعنی میانگین iام در زمان t دارند. خوشه iام در زمان t برابر خواهد بود با:

رابطه (٦)

$$\left(\|x_p - \mu_i^{(t)}\|^2 \le \|x_p - \mu_j^{(t)}\|^2 \ \forall j, 1 \le j \le k\right)$$

- حال میانگین‌ها را بر اساس خوشه‌های جدید به این شکل بروز می‌کنیم:

رابطه (۷) $$\mu_i^{t+1} = \frac{1}{|S_i^t|}\sum_{x_j \in S_i^t} X_j$$

وزن‌دهی صفحات

پس از پیش‌پردازش logهای وب و خوشه‌بندی داده‌ها برای اهداف پیشنهادی، داده‌ها به شکل‌های مناسبی تبدیل یا ادغام می‌شوند [۲۳].

ما برای این منظور، وزن رتبه‌دهی به هر فعالیت یادگیری را با استفاده از تابع امتیاز زیر تعریف نموده‌ایم:

رابطه (۸)

$$P(\theta) = EXP(\theta) + \alpha * IMP(\theta) + \beta * S(\theta)$$

که EXP امتیاز صریح داده شده توسط فرد یادگیرنده بـــه هر شـــئ یادگیری θ اســـت. IMP نیز امتیـــاز ضمنی و S امتیـــاز بُعد اجتماعی اســـت. پارامترهای α و β برای نرمال‌ســـازی توابع IMP و S با واحد ۵ انتخاب شـــده‌اند. امتیاز صریح به صورت زیر ارائه شده است:

$$IMP(\theta) = A(\theta) + B(\theta) + C(\theta)$$ رابطه (۹)

A وقتی برابر یک اســـت که صفحات نشان شده ذخیره شـــده باشـــد، در غیر این صورت برابر با صفر می‌باشد. تابع $B(\theta) = 1 - e^{-t}$ که t مدت زمان سپری شده توسط فرد یادگیرنده در طول شـــئ یادگیری θ است، C فرکانس دسترسی فعالیت یادگیری است.

در نهایـــت تابع S که بُعد رتبه‌بندی اجتماعی اســـت، به صورت زیر تعریف می‌شود:

$$S(\theta) = c * e^{t'}$$ رابطه (۱۰)

که t' مدت زمان سپری شـــده در طول تمام ارتباطات همزمان و غیرهمزمان با استفاده از ابزار ارتباطی و انجمنی است، c تعداد مشارکت‌ها و تعاملات با این ابزار می‌باشد.

پـــس از وزن‌دهی به منابع یادگیری، ما یک مدل ترجیح برای هر فرد یادگیرنده به دست می‌آوریم که به صورت یک ماتریس رتبه‌بندی شئ یادگیری فرد یادگیرنده (LLOR) با n ردیف و m ستون تعریف می‌شود که n نشان‌دهنده تعداد افـــراد یادگیرنده $L = \{l1, l2, \dots, ln\}$، و m نشـــان‌دهنده تعداد اشیاء یادگیری $J = \{j1, j2, \dots, jm\}$ است.

جدول ۱ مثالی از ماتریس رتبه‌بندی شئ یادگیری فرد یادگیرنده (LLOR) را نشان می‌دهد.

جدول ۱. نحوه امتیازدهی به داده‌ها

کاربران	صفحه ۱	صفحه ۲	صفحه ۳	صفحه ۴	صفحه n
کاربر ۱	۱	۱	۶	۶	۴
کاربر ۲	۹	۸	۰	۳	۷
کاربر ۳	۳	۲	۳	۳	۱۰
کاربر ۴	۸	۱	۰	۸	۱

ایـــن ماتریس از یک مقیاس رتبه‌بندی ۰ تا ۱۰ اســـتفاده می‌کند: ۱۰ بدین معنی است که فرد یادگیرنده به شدت از شئ یادگیری انتخاب شـــده راضی است، ۵ نشان‌دهنده این است که فرد یادگیرنده نسبتا راضی نیست، ۱ نشان می‌دهد که فرد یادگیرنده از شـــئ یادگیری به هیچ وجه راضی نیست و در نهایت امتیاز صفر نیز نشان می‌دهد که شئ یادگیری هنوز به صورت صریح رتبه‌بندی نشده یا اصلا هنوز مورد استفاده قرار نگرفته است.

محاسبهٔ شباهت با فاصله اقلیدسی

خوشه‌بندی فرایند خودکارســـازی است که در طی آن، اشـــیاء به دســـته‌هایی که اعضای آن از نظر شـــاخص‌های موردنظر مشـــابه یکدیگر باشند تقســـیم می‌شوند، بنابراین برای سنجش شباهت بین اشیای داده از اندازه‌گیری فاصله استفاده می‌شود. روش‌های مختلفی برای اندازه‌گیری فاصله بین دو شـــیء وجود دارد که فاصله اقلیدسی معروف‌ترین و پرکاربردترین گونه فاصله است که به صورت رابطه زیر تعریف می‌شود.

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (xi - yi)^2}$$ رابطه (۱۱)

فرض کنید در فضای اقلیدسی،n بعدی مجموعه‌ای از n شـــیء به وســـیله مجموعه $S = \{x1, X2, X3, \dots XN\}$ و k خوشه به وسیله مجموعه $C = \{c1, c2, c3, \dots, cc\}$ نشان داده شود، از این رو خوشه‌ها باید شرایط زیر را داشته باشد:

هر خوشه حداقل باید شامل یک شیء باشد، یعنی:

$$Ci \neq \varphi \forall \in \{1,2,3, \dots, K\}$$

دو خوشـــه مختلف نباید اشیای مشترک داشته باشند، یعنی:

$$Ci \cap Cj = \varphi \forall \neq j \; and \; i,j \in \forall \in \{1,2,3, \dots, K\}$$

هر شیء به یک خوشه تخصیص پیدا کند، یعنی:

$$U_{i=1}^{k} \; C_i = S$$

برای پیداکردن مرکز K خوشـــه، مســـاله بـــه عنوان یک بهینه‌ســـازی (حداقل) یک تابع عملکرد تعریف می‌شود. تابع عملکرد مورد استفاده برای این هدف، کل میانگین مربع خطا تدریج است که به صورت رابطه زیر تعریف می‌شود:

رابطه (۱۲)

$$f(X, C) = \sum_{i=1}^{n} Min\{X_{i-} \; C_I^2 \; | j = 1,2, \dots, K\}$$

هدف این معیـــار کمینه کردن مجموعه فواصل داده‌های موجود درون خوشه ها از مراکز خوشه می‌باشد.

عبارت $||Ci, Xi||$ فاصله بین داده Xi و مرکز مربوطه Ci را نشان می‌دهد.

رابطه (۱۳)
$$SC(u,v) = \frac{\sum_j^m (r_{u,j} - \bar{r}_u)(r_{v,j} - \bar{r}_v)}{\sqrt{\sum_j^m (r_{u,j} - \bar{r}_u)^2 (r_{v,j} - \bar{r}_v)^2}}$$

در رابطـــه با معادلـــه بالا: $\bar{r}_u$ و $\bar{r}_v$ به ترتیب متوســـط رتبه‌بندی یادگیرنده‌های u و v هستند؛ $r_{u,j}$ و $r_{v,j}$ به ترتیب رتبه‌بندی یادگیرنده‌های u و v برای شئ یادگیری j می‌باشند.

پس از این که شـــباهت بین دو فرد یادگیرنده محاســـبه شد، یک ماتریس شـــباهت $N \times N$ تولید می‌شود، N تعداد افراد یادگیرنده اســـت. سپس، برای پیش‌بینی شئ یادگیری رتبه‌بندی نشده در ماتریس رتبه‌بندی توسط فرد یادگیرنده‌ی فعال K، u عدد از مشابه‌ترین افراد یادگیرنده انتخاب خواهند شد و به عنوان ورودی برای محاسبهٔ پیش‌بینی u بر روی j مورد استفاده قرار خواهند گرفت.

تولید پیشنهادها

بـــرای ایجاد یک پیش‌بینی و تولید پیشـــنهادها برای یک فرد یادگیرنده فعال u در مورد اشـــیاء یادگیری خاص j، ما می‌توانیم مطابق با فرمول زیر، یک متوسط وزن‌دار از تمام رتبه‌بندی‌هـــای صورت گرفته بر روی این اشـــیاء یادگیری بگیریم[۱۱]:

رابطه (۱۳)
$$P_{u,j} = \bar{r}_u + \frac{\sum_{v=1}^{k} SC(u,v)(r_{v,j} - \bar{r}_v)}{\sum_{v=1}^{k} |SC(u,v)|}$$

در رابطه (۵)، $r_{v,j}$ نشان‌دهنده مقدار رتبه‌بندی داده شده توسط کاربر v برای شئ یادگیری انتخاب شدهٔ j است.

پراکندگی رتبه‌بندی به صورت زیر محاسبه می‌شود:

رابطه (۱۴)
$$\text{پراکندگی} = 1 - \frac{\sum \text{رتبه‌بندی‌ها}}{\sum \text{اشیاء یادگیری} * \sum \text{افراد یادگیرنده}} \times 100$$

رتبه‌بندی‌های صریح، ضمنی و اجتماعی اشیاء یادگیری برای وزن‌دهی اشـــیاء یادگیری به صورت مقادیر عددی از صفر تا ۵ نمایش داده می‌شـــوند. ما در آزمایش‌های خود از الگوریتم KNN (K نزدیکترین همســـایه) استفاده کرده‌ایم تا بهترین مقدار از K عدد همسایه را در مجموعه دادهٔ افزایشی بیابیم.

ارزیابی نتایج

برای ارزیابی سیستم‌های توصیه‌گر معمولا از دو روش دقت و فراخوانی استفاده می‌شود. میزان دقت و فراخوانی با استفاده از دو فرمول زیر به دست می‌آید. برای ارزیابی روش پیشنهادی، مقایسه‌ای بین روش پیشنهادی و دو روش قوانین انجمنی و الگوریتم RBF صورت گرفت و این دو الگوریتم با روش مورد استفاده در روش پیشنهادی مورد پیاده‌سازی و شبیه‌سازی قرار گرفت که نتایج به دست آمده در شکل‌های ۳ و ٤ قابل مشاهده می‌باشد.

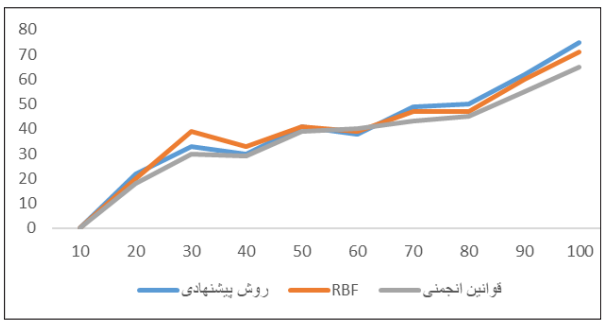
دقت با استفاده از رابطه زیر محاسبه می‌گردد:

رابطه (۱۵)
$$Precision = \frac{|\{relevant\ pages\} \cap \{retrved\ pages\}|}{|\{retrived\ pages\}|}$$

فراخوانی با استفاده از رابطه محاسبه می‌گردد:

رابطه (۱۶)
$$Recall = \frac{|\{relevant\ pages\} \cap \{retrved\ pages\}|}{|\{relevant\ pages\}|}$$

(شکل ۲): نمودار مقایسه روش پیشنهادی با سایر روش ها در قسمت دقت



(شکل ۳): نمودار مقایسه روش پیشنهادی با سایر روش‌ها در قسمت فراخوانی

بحث و نتیجه‌گیری

در ایـــن تحقیق به ارائه روش جدیـــدی به منظور بهبود سیستم‌های توصیه‌گر در زمینه وب پرداخته شد. سیستم‌های توصیه‌گر یا پیشنهاددهنده سیستم‌هایی هستند که با گرفتن اطلاعات محدودی از کاربر می‌توانند که پیشنهادها مناسبی به کاربـــر ارائه دهند. در این تحقیق به ارائه روشـــی مبتنی بر وزن‌دهی صفحات برای پیشـــنهاد به کاربر استفاده شد. بدین‌صورت کـــه پس از جمع‌آوری فایل‌های ثبت مربوط به کاربران، عملیات پیش‌پردازش داده‌ها صورت گرفت سپس با اســـتفاده یک روش خوشه‌بندی ترکیبی که ترکیبی از دو الگوریتـــم K-MEANS و C-MEANS بود، داده‌های موجود در فایل ثبت خوشه‌بندی گردیدند سپس با استفاده از رفتارهای کاربر مانند علایق یک کاربر به یک صفحه، الگوهای موجود در داده‌ها اســـتخراج گردیدند. سپس صفحات به دست آمده وزن‌دهی گردیدند و صفحاتی که بیشـــتر توســـط کاربران جستجو شده بود مشخص شدند. در نهایت با استفاده از یک سیستم توصیه‌گر، صفحات موجود به کاربر پیشنهاد خواهند شد. در ارزیابی صورت گرفته بین روش پیشنهادی و سایر روش‌های موجود مشـــخص شـــد که روش پیشنهادی در قســـمت دقت، امتیاز ۷۵ درصد و در قسمت فراخوانی امتیاز ۷۵ را به دست آورد که در مجموع می‌توان گفت که سیستم پیشـــنهادی به صورت کلی می‌تواند تـــا ۸۵ درصد نیازهای کاربر را به درستی پیش‌بینی و صفحات مورد علاقه کاربر را با دقت بالاتری به کاربر پیشنهاد دهد.

منابع

1.Gheisari, M., Hamidpour, H., Liu, Y., Saedi, P., Raza, A., Jalili, A., ... & Amin, R. (2023). Data Mining Techniques for Web Mining: A Survey. In Artificial Intelligence and Applications (Vol. 1, No. 1, pp. 3-10).

2.Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H. Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. Education and Information Technologies, 28(1), 905-971.

3.Dol, S. M., & Jawandhiya, P. M. (2023). Classification Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining—A survey. Engineering Applications of Artificial Intelligence, 122, 106071.

4.Yu, B., Mao, W., Lv, Y., Zhang, C., & Xie, Y. (2022). A survey on federated learning in data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(1), e1443.

5.Yates, D., & Islam, M. Z. (2022). Data mining on smartphones: An introduction and survey. ACM Computing Surveys, 55(5), 1-38.

6.Hamdi, A., Shaban, K., Erradi, A., Mohamed, A., Rumi, S. K., & Salim, F. D. (2022). Spatiotemporal data mining: a survey on challenges and open problems. Artificial Intelligence Review, 1-48.

7.Xiao, W., Ji, P., & Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. Engineering Reports, 4(5), e12482.

8.Riehle, D., Harutyunyan, N., & Barcomb, A. (2021). Pattern discovery and validation using scientific research methods. arXiv preprint arXiv:2107.06065.

9. Alhijawi, B. & Kilani, Y. (2020). The recommender system: A survey. International Journal of Advanced Intelligence Paradigms, 15(3), 229-251.

10. Pradhan, N. & Dhaka, V. S. (2020). Comparison-Based Study of PageRank Algorithm Using Web Structure Mining and Web Content Mining. In Smart Systems and IoT: Innovations in Computing (pp. 719-729). Springer, Singapore.

11. Leskovec, J. Rajaraman, A. & Ullman, J. D. (2020). Mining of massive data sets. Cambridge university press.

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

13. Chawla, S. (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.

15. Bourkoukou, Outmane, and Omar Achbarou. "Weighting based approach for learning resources recommendations." JOIV: International Journal on Informatics Visualization 2, no. 3 (2018): 104-109.

15. Riyahi, M. & Sohrabi, M. K. (2020). 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.

۱۶. مولایی فرد، مصلح. ارائه یک سیســـتم توصیه گر وب برای پیش‌بینی صفحات مورد علاقه کاربر با اســـتفاده از الگوریتم خوشـــه‌بندی DBSCAN و روش SVM یادگیری ماشین. فصلنامه فناوری اطلاعات و ارتباطات ایران. ۵۷(۵۷)، ۷۷.

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

18.Tanwar, A., Vishwakarma, D.K. A deep neural network-based hybrid recommender system with user-user networks. Multimed Tools Appl (2022).

19.Alhijawi, B., AL-Naymat, G. Novel Positive Multi-Layer Graph Based Method for Collaborative Filtering Recommender Systems. J. Comput. Sci. Technol. 37, 975–990 (2022).

20. Kiran, A. & Vasumathi, D. (2020). Data Mining: Min–Max Normalization Based Data Perturbation Technique for Privacy Preservation. In Proceedings of the Third International Conference on Computational Intelligence and Informatics (pp. 723-734). Springer, Singapore.

21. J. C. Bezdek, R. Ehrilch, W. Full, 1984, "FCM: THE FUZZY c-MEANS CLUSTERING ALGORITHM ", Elsevier Computer and Geoscience, Vol. 10, No. 2-3, PP. 191-203.

22. Luengo, J. García-Gil, D. Ramírez-Gallego, S. García, S. & Herrera, F. (2020). Big Data Preprocessing: Enabling Smart Data. Springer Nature.

23. Bourkoukou, O., & Achbarou, O. (2018). Weighting based approach for learning resources recommendations. JOIV: International Journal on Informatics Visualization, 2(3), 104-109.