مؤسسه آموزش عالی غیرانتفاعی مجازی
نور طوبی

پایان نامه کارشناسی ارشد

مهندسی فناوری اطلاعات (تجارت الکترونیک)

بررسی کاربردهای سیستم‌های توصیه‌گر در محیط‌های آموزش الکترونیکی

استاد راهنما:

دکتر مهرگان مهدوی

استاد مشاور:

مهندس علی‌رضا قنادان

نگارش:

آیلر سخاوتیان

بهمن ماه ۱۳۸۹

مؤسسه آموزش عالی غیرانتفاعی مجازی
نور طوبی

پایان نامه کارشناسی ارشد

مهندسی فناوری اطلاعات (تجارت الکترونیک)

بررسی کاربردهای سیستم‌های توصیه‌گر در محیط‌های آموزش الکترونیکی

استاد راهنما:

دکتر مهرگان مهدوی

استاد مشاور:

مهندس علی‌رضا قنادان

نگارش:

آیلر سخاوتیان

بهمن ماه ۱۳۸۹


بسم الله الرحمن الرحيم

بسمه تعالی

اظهارنامه(اصالت اثر)

اینجانب **آیلر سخاوتیان** دانشجوی رشته **فناوری اطلاعات** گرایش **تجارت الکترونیک** مؤسسه **آموزش عالی نورطوبی** اظهار می‌کنم که این پایان نامه حاصل پژوهش خودم بوده و در جاهایی که از منابع دیگران استفاده کرده‌ام، نشانی دقیق و مشخصات کامل آن را نوشته‌ام. همچنین اظهار می‌کنم که تحقیق و موضوع پایان نامه‌ام تکراری نیست و تعهّد می‌نمایم که بدون مجوز دانشگاه دستاوردهای آن را منتشر ننموده و یا در اختیار غیر ندهم. کلیه حقوق این اثر مطابق با آیین نامه مالکیت فکری و معنوی متعلق به مؤسسه آموزش عالی نورطوبی است.

نام و نام خانوادگی: آیلر سخاوتیان

تاریخ و امضاء: ۱۳۸۹/۱۱/۲۷

تقدیم به

بهترین‌های زندگیم

پدر و مادرم

که تمام لحظات و موفقیت‌های زندگی خود را مدیون فداکاری‌های بی دریغ آنان می‌دانم که همواره مشوق من بوده‌اند و شرایط مناسب جهت تحصیل مرا فراهم نموده‌اند.

و برادرانم

که همواره در طول تحصیل متحمل زحماتم بودند و تکیه‌گاه من در مواجهه با مشکلات، و وجودشان مایهٔ دلگرمی من می‌باشد.

سپاسگزاری

از استاد ارجمند جناب آقای دکتر مهرگان مهدوی که در طور مراحل تدوین و مطالعهٔ پایان نامه، راهنمایی‌ها و حمایت‌های خود را دریغ ننمودند، صمیمانه سپاسگزارم.

از استاد بزرگوار جناب آقای علی‌رضا قنادان نیز که زحمت مشاوره این پایان نامه را بر عهده گرفتند، سپاسگزارم.

چکیده

بررسی کاربردهای سیستم‌های توصیه‌گر در محیط‌های آموزش الکترونیکی

نگارنده

آیلر سخاوتیان

در سال‌های اخیر، پیشرفت‌های بسیاری در سیستم‌های آموزشی، در راستای معرفی تکنولوژی‌های نوینی همچون آموزش تحت وب رخ داده است. در طی چند سال گذشته هزاران رشته تحت وب ثبت و گسترش یافته است. اما در بسیاری از رشته‌های تحت وب مطالب بدون توجه به تفاوت یادگیری فراگیران، به صورت یکنواخت آموزش داده می‌شود. به منظور جلوگیری از این مسأله و افزایش کارایی این محیط‌ها و همچنین بالا بردن انگیزه و کیفیت آموزش فراگیران از سیستم‌های توصیه‌گر استفاده می‌گردد.

در این پژوهش به معرفی سیستم توصیه‌گری می‌پردازیم که قادر به ارائهٔ مطالب آموزشی متناسب با سطح فراگیران و در نتیجه جلوگیری از مردود شدن و افزایش پیشرفت تحصیلی آنهاست. این سیستم، از تکنیک‌های داده‌کاوی استفاده می‌نماید. داده‌های مطالعهٔ فراگیران را تجزیه و تحلیل نموده و قبل از اتمام دورهٔ تحصیلی وی نتیجهٔ نهایی‌اش را با توجه به عملکرد فراگیران مشابه پیش بینی خواهد نمود. در صورت تشخیص ضعیف و یا قوی بودن فراگیر از رویکردهای مختلف جهت ارائه پیشنهادها استفاده می شود. نتایج حاکی از آن است که سیستم پیشنهادی به طور مؤثری مفید بوده است. به منظور ارزیابی عملکرد سیستم، دقت و احتمال پیش بینی صحیح وضعیت فراگیران محاسبه می‌شود. احتمال پیش بینی قبولی فراگیران هنگامیکه از پیشنهادهای ارائه شده توسط سیستم استفاده نمودند، نسبت به حال عادی افزایش یافته است.

فهرست مطالب

فصل سوم: مرور ادبیات و تحقیقات پیشین

فصل چهارم: روش تحقیق: طراحی و پیاده‌سازی سیستم توصیه‌گر پیشنهادی

فهرست شکل‌ها

فهرست جداول

فصل ۱- مقدمه

۱-۱ مقدمه

با گسترده شدن فناوری اطلاعات و نفوذ وسایل ارتباط از راه دور به عمق جامعه، ابزارها و روش‌های آموزش نیز دچار تحول شدند. امروزه، افراد بسیاری از انواع برنامه‌های آموزش الکترونیکی[۱] بهره‌مند شده‌اند. اگرچه در بسیاری از سیستم‌های آموزشی آنلاین، مجموعه واحدی از منابع آموزشی، بدون توجه به تفاوت یادگیری فراگیران به صورت یکنواخت آموزش داده می‌شود. در محیط‌های آموزش الکترونیکی تلاش بر این است تا با تهیهٔ طرحی از اهداف شخصی، علایق و معلومات فراگیران، گونه ای از آموزش اختصاصی به ایشان ارائه داده شود.

حجم بسیار بالا و فزایندهٔ اطلاعات در شبکهٔ جهانی موجب توجه روز افزون به فرآیند شخصی سازی[۲] شده است. در این راستا، تحلیل و بررسی داده‌های موجود بر اساس تکنیک‌های داده‌کاوی[۳] بسیار مفید خواهد بود. سیستم‌های پیش‌بینی کننده و یا توصیه‌گر[۴] برای این منظور به طور روز افزونی توسعه داده شده‌اند. در سیستم‌های توصیه‌گر تلاش می‌شود تا از دانش علایق کاربر که از گذشتهٔ گردش وی در وب بدست آمده، برای پیدا کردن کالاها یا صفحات وب مورد علاقهٔ وی استفاده شود.

تاکنون تحقیقات بسیاری در حوزهٔ سیستم‌های توصیه‌گر انجام پذیرفته است و روش‌های گوناگونی برای پیش‌بینی و توصیهٔ کالا مورد بررسی قرار گرفته‌اند. اگر توصیه‌های ارائه شده با سلایق کاربر هم خوانی نداشته باشد، آنگاه کاربر دیگر اعتمادی به این سیستم‌ها نخواهد داشت و خود به دنبال اطلاعات مورد علاقهٔ خود خواهد گذشت. به همین دلیل در سال‌های اخیر الگوریتم‌های گوناگونی در

[1] E-learning
[2] Personalization
[3] Data Mining
[4] Recommender Systems

جهت ارتقای کیفیت این سیستم‌ها ابداع شده اند. سیستم‌های توصیه‌گر در تجارت الکترونیک به منظور تسهیل فرایند خرید کالا بسیار استفاده گردیده، ولی به کاربرد و پیاده‌سازی آن در محیطهای آموزشی کمتر توجه شده است. کاربرد این سیستم‌ها در محیطهای آموزشی توجه ویژه‌ای می‌طلبد که در حوزهای دیگر به این شدت نیست و از مهمترین آنها در نظر گرفتن ویژگی‌های آموزشی خاص فراگیر و سیستم است.

در این پژوهش سعی شده است کلیه فرآیندهای شخصی سازی، انواع رویکردها در سیستم‌های توصیه‌گر، تکنیک‌های داده کاوی و وب‌کاوی، و چگونگی شناسایی اولویت و علایق کاربران بررسی شود. در ادامه به بحث پیرامون چگونگی بکارگیری و کارایی این سیستم‌ها در محیطهای آموزش الکترونیک پرداخته، سپس به ارائه و معرفی یک سیستم توصیه‌گر با استفاده از تکنیک‌های داده‌کاوی در محیطهای آموزش الکترونیک خواهیم پرداخت. پارامترهای کارایی و ارزیابی نیز معرفی شده و مورد بررسی قرار خواهند گرفت تا میزان کارایی روش پیشنهادی تعیین گردد.

۱-۲ ضرورت و اهمیت پژوهش

تحقیقات نشان داده شده است که بازدهی آموزش الکترونیکی بر روی همهٔ افراد یکسان نیست و در بسیاری از مواقع دوره‌های آموزش الکترونیکی، پس از صرف زمان و هزینهٔ زیاد توسط افراد، با شکست مواجه خواهند شد. در این سیستم‌ها کاربران سیستم (فراگیران) از لحاظ ویژگی‌های فردی مانند سطح دانش، سرعت پیشرفت در یادگیری، سن، تخصص، انگیزه و هدف یادگیری با یکدیگر متفاوت هستند. اگر سیستم آموزش الکترونیکی، شیوهٔ آموزش مطالب درسی را با توجه به ویژگی‌های کاربر انتخاب کرده و به او ارائه کند، تأثیر بسزائی در ارتقاء کیفیت آموزش الکترونیکی خواهد داشت.

بنابراین ارائهٔ یک سیستم شخصی که بتواند به طور خودکار با سطح و علایق فراگیر منطبق شود از اهمیت بسیاری برخوردار است. در سیستم‌های آموزش الکترونیکی شخصی می‌بایست مشخصات کاربر از جمله علایق، سطوح و الگوهای یادگیری وی در طول فرایند یادگیری ارزیابی شده، سپس منابع آموزشی شخصی بر اساس پروفایل کاربر و با مطابقت با مشخصات و سطوح فردی وی تولید گردد. علاوه بر آن فراگیران با علایق و سطوح یکسان گروه بندی شده و بازخورد یکی از افراد گروه، به عنوان راهنما جهت ارائهٔ اطلاعات به سایر اعضای گروه به کار گرفته شود. از جمله دلایل نیاز به سیستم‌های توصیه‌گر در محیطهای آموزش الکترونیک می توان به موارد زیر اشاره نمود:

۱-۲-۱ گرانبار شدن اطلاعات در محیط‌های آموزش الکترونیکی

با دسترسی بیشتر اطلاعات به صورت الکترونیکی و از طریق وب گستردهٔ جهانی، کاربران هنگام جستجو در اینترنت با مشکل حجم عظیم اطلاعات مواجه می‌شوند. به طور کلی، هدف استفاده از سیستم‌های توصیه‌گر در کاربردهای وب، ارائهٔ اطلاعات مفید و متناسب با سلیقه و اولویت‌های کاربران با تلاش کمتر می‌باشد. در برخی مواقع از این سیستم‌ها جهت مخفی نمودن و فیلتر اطلاعاتی خاص استفاده می‌گردد. هدف استفاده از سیستم های توصیه گر در کاربردهای آموزش الکترونیکی (LMS[1]) فهرست نمودن «نزدیکترین محتویات آموزشی به نیاز کاربر و با توجه به توصیهٔ مربی است». زیرا سیستم مدیریت آموزش شامل هزاران دوره بوده که با مشکل گرانبار شدن اطلاعات مواجه می باشد. (Itmazi, 2008)

۱-۲-۲ ارتقاء کارایی محیط‌های آموزش الکترونیکی

یکی از چالش‌های جدی در مدیریت امور آموزشی پیش بینی وضعیت تحصیلی دانشجویان به منظور شناسایی دانشجویانی است که دچار افت تحصیلی شده و ادامه تحصیل آنها با مشکل روبرو خواهد شد. سیستم‌های توصیه‌گر نه تنها به مربیان در شناسایی نقاط ضعف و مشکلات فراگیران در فرایند یادگیری کمک می‌نماید، بلکه توصیه‌های مفیدی را به فراگیران جهت شناخت نقاط ضعف خود ارائه می‌دهد. به طور کلی، مربیان به راحتی متوجهٔ تعداد فراگیران رد شده از آزمون، می‌گردند. اما در شناسایی مشکلات واقعی آنها در یادگیری ضعف دارند. سیستم با توجه به رکوردهای آزمون فراگیر، به شناسایی و کشف مشکلات آنها پرداخته و سپس پیشنهاداتی را جهت طراحی استراتژی‌های آموزشی جدید ارائه می‌دهد. علاوه بر آن، اطلاعات تولید شده، برای فراگیران نیز مفید خواهد بود. و آنها را قادر ساخته تا فرایند یادگیری خود را بهبود بخشند. (Hsu, 2008)

۱-۲-۳ افزایش انگیزه و کیفیت آموزش فراگیران

به منظور بالا بردن انگیزهٔ یادگیری، کمک به فراگیران در انتخاب دروس مطالعه که متناسب با علاقه‌های متفاوت آنها باشد، از ارزش بسیاری برخوردار است. مطالعات متعدد نشان می‌دهد که با استفاده از تکنولوژی‌های آموزش مبتنی بر کامپیوتر CAI[2]، قادر به کمک فراگیران با سطح عملکرد

[1] Learning Management System
[2] Computer-Assisted Instruction

پایین خواهیم شد. در واقع CAI، اصطلاحی است که اغلب برای فعالیت‌های تمرینی و اصلاحی به کار می‌رود. مخصوصاً به این دلیل که بکارگیری این تکنولوژی ها، الهام بخش پشتکار و انگیزه در فراگیران شده که خود منجر به افزایش اعتماد به نفس و علاقهٔ آنها در یادگیری خواهد گردید. استفاده از کامپیوتر و آخرین فناوری‌ها، انگیزهٔ فراگیرانی که برای انجام تکالیف خود نیاز به آموزش اصلاحی دارند را افزایش می دهد. (Hancock, 1992)

۱-۳ هدف پژوهش

هدف اصلی این پژوهش بررسی تکنیک‌ها و الگوریتم‌های مورد استفاده و نحوهٔ عملکرد سیستم‌های توصیه‌گر در محیط‌های آموزش الکترونیک می‌باشد. همچنین ارائهٔ یک مدل تلفیقی جهت بهبود عملکرد این سیستم‌ها، با توجه به وضعیت و پیشرفت تحصیلی هر فراگیر.

در این پژوهش، سیستم توصیه‌گری طراحی و ارائه گردیده است که با دسته بندی فراگیران به گروه‌های مختلف، به بررسی عملکرد و رفتار آنها پرداخته و سپس به توصیهٔ مطالب و یا آزمون های مناسب به وی می‌پردازد. بدین ترتیب پس از ورود فراگیر جدید، سیستم داده‌های مطالعهٔ فراگیران را تجزیه و تحلیل نموده و نتیجهٔ نهایی‌اش را قبل از اتمام دورهٔ تحصیلی با توجه به عملکرد فراگیران مشابه پیش بینی خواهد نمود. در صورت تشخیص ضعیف و یا قوی بودن وی از رویکردهای مختلف جهت ارائه پیشنهادها استفاده می شود.

۱-۴ سؤالات کلیدی پژوهش

پرسش‌های کلیدی این پژوهش شامل:

- دلایل تفاوت سیستم‌های توصیه گر در محیط های آموزش الکترونیکی نسبت به سایر حوزه‌ها در چیست؟
- تأثیر و کاربرد سیستم‌های توصیه‌گر در محیط‌های آموزش الکترونیکی چیست؟
- مزایا و روش‌های شناسایی فراگیران و ارائهٔ خدمات شخصی شده چیست؟
- استفاده از چه تکنیک‌هایی می‌تواند در شخصی سازی محیط‌های آموزش الکترونیکی مفید باشد؟
- آیا گروه هایی از فراگیران که از منابع برخط به طور مشابه استفاده نموده و رفتار یکسانی داشته، وجود دارد؟

- ❖ آیا امکان بهبود فرایند یادگیری یک فراگیر با توجه به فعالیت سایر فراگیران در گروه مربوطه، وجود دارد؟
- ❖ آیا قادر به پیشنهاد محتویات آموزشی متناسب با نیاز هر فراگیر، به وی هستیم؟
- ❖ آیا امکان توسعه و طراحی یک مدل توصیهٔ بهینه وجود دارد؟

۱-۵ رویکرد بکار رفته در تحقیق

منابع بکار رفته در پروژه را می‌توان به دو دسته تقسیم کرد. دستهٔ اول منابعی هستند که دانش پیش‌زمینه ای موضوع کار را در اختیار قرار می‌دهند. رویکرد بکار رفته در مطالعهٔ این منابع رویکردی کاملاً کاربردی است به این معنی که هدف از مطالعهٔ آنها آشنایی با ابزارها و تکنیک ها بوده تا ارائهٔ یک روش جدید. دستهٔ دوم منابعی می‌باشند که در زمینهٔ صورت مساله مطرح شده روشی را ارائه کرده اند. در مطالعهٔ این گونه منابع رویکرد بکار رفته علاوه بر مروری مختصر بر ابزارها و جزئیات پیاده سازی، تأکید بر ایده و روش آنها و نحوهٔ ارزیابی و سعی در بهبود آن بوده است.

۱-۶ معرفی و مرور فصل‌های پایان نامه

این پایان‌نامه در برگیرنده ۶ فصل به شرح ذیل می باشد:

فصل ۱. مقدمه: بطور خلاصه به شناخت موضوع و معرفی پژوهش می پردازد و به گونه‌ای تدوین شده که خواننده با مطالعه آن اطلاعات جامعی را در ارتباط با حوزه تحقیق، ضرورت و اهداف تحقیق درمی‌یابد.

فصل ۲. مبانی نظری تحقیق: در این فصل ابتدا در مورد سیستم‌های آموزش تحت وب و همچنین سیستم‌های توصیه‌گر و کاربردها، مزایا و رویکردهای مختلف آن به تفصیل مطالبی ارائه می‌گردد. داده‌کاوی و تکنیک‌های مربوط به آن مورد بررسی قرار گرفته، سپس در خصوص وب کاوی، تکنیک های آن و مراحل شخصی سازی بحث می‌شود.

فصل ۳. مروری بر تحقیقات انجام شده: این فصل شامل دو بخش است. در بخش اول کارهای انجام شده در زمینهٔ کاربرد سیستم‌های توصیه‌گر در محیط‌های آموزش الکترونیک مورد بررسی قرار می‌گیرند و بخش دوم حاوی مروری بر الگوریتم‌های مورد استفاده در مدل پیشنهادی ارائه شده و نیز بخش اول این فصل، می‌باشد.

فصل ۴. روش تحقیق: در این فصل یک مدل تلفیقی جهت بهبود عملکرد سیستم‌های توصیه‌گر در محیط‌های آموزش الکترونیک معرفی می‌شود. همچنین جزئیات مربوط به پیاده‌سازی روش پیشنهادی، مجموعه داده‌ها، ابزارها و غیره ارائه می‌گردد.

فصل ۵. نتایج (بحث و بررسی و تحلیل داده‌ها): در این فصل، معیارهای ارزیابی معرفی شده و نتایج آزمایشات انجام شده و مقایسهٔ آنها با سایر روش‌ها و تحلیل آنها بررسی می‌شود.

فصل ۶. نتیجه گیری و پیشنهادات: در این فصل دستاوردهای پروژه و زمینه‌های کاری آیندهٔ آن ارائه می‌شود.

فصل ۲- مبانی نظری تحقیق:

مبانی، مفاهیم و معرفی سیستم‌های توصیه‌گر

۲-۱ مقدمه

در این فصل مطالب پیش‌زمینه‌ای را که برای درک روش ارائه شده در پایان‌نامه ضروری است توضیح می‌دهیم. ابتدا در مورد سیستم‌های آموزش تحت وب و همچنین سیستم های توصیه گر به تفصیل مطالبی ارائه می گردد. داده کاوی و تکنیک های مربوط به آن مورد بررسی قرار گرفته، سپس در خصوص وب کاوی، تکنیک های آن و مراحل شخصی سازی بحث می‌شود.

پیش بینی می‌شد که تا اوایل سال ۲۰۰۷ میلادی در سایت دانشنامه اینترنتی ویکی‌پدیا چیزی حدود ۵.۱ میلیون مقاله به ثبت رسیده باشد و یا سایت مدیریت و به اشتراک‌گذاری تصاویر Flickr بالغ بر ۲۵۰ میلیون تصویر را در خود جای دهد. از این رو، می‌توان گفت که ما در میان حجم عظیمی از داده و اطلاعات قرار گرفته‌ایم که بدون راهنمایی و ناوبری درست ممکن است انتخاب‌هایی غلط و یا غیر بهینه از میان آن‌ها داشته باشیم. سیستم‌های توصیه‌گر، سیستم‌های تأثیرگذار در راهنمایی و هدایت کاربر، در میان حجم عظیمی از انتخاب‌های ممکن، برای رسیدن به گزینه مفید و مورد علاقه وی هستند به گونه‌ای که این فرآیند، برای نفس همان کاربر، شخصی‌سازی شده باشد.

سیستم‌های توصیه‌گر در واقع سیستم‌های اطلاعاتی هستند که، توانایی تحلیل رفتارهای گذشته و ارائه توصیه‌هایی برای مسائل جاری را دارا هستند (Ting peng Liang, 2008). به زبان ساده‌تر در سیستم‌های توصیه‌گر تلاش بر این است تا با حدس زدن شیوه تفکر کاربر (به کمک اطلاعاتی که از نحوهٔ رفتار وی یا کاربران مشابهٔ وی و نظرات آن‌ها داریم) به وی مناسب‌ترین و نزدیک‌ترین کالا به سلیقه او را شناسایی و پیشنهاد کنیم. این سیستم‌ها در حقیقت همان فرآیندی که ما در زندگی روزمره خود به کار می‌بریم را شبیه‌سازی و به صورت اتوماتیک اجرا می‌کنند و این همان فرآیندی

است در زندگی عادی و روزمره خود، طی آن تلاش می‌کنیم تا افرادی با سلایق نزدیک به خود را پیدا کرده و از آنها در مورد انتخاب‌هایمان نظر بخواهیم.

۲-۲ آموزش الکترونیکی

با گسترش فناوری اطلاعات و نفوذ وسایل ارتباط از راه دور به عمق جامعه، ابزار ها و روش های آموزش نیز دچار تحول شدند. اکنون با به صحنه آمدن دانشگاه ها و مراکز آموزش الکترونیکی، امکان یادگیری در هر زمینه ای، برای هر فردی در هر زمان و مکانی به صورتمادام العمر فراهم شده است. برنامهٔ آموزشی فوق رسانه[1] ای از منابع و مراجع وب جهانی[2] (www) استفاده می کند تا محیطی هدفمند برای یادگیری ایجاد کند. منابع وب توسط متون، تصاویر و رسانه های مختلف دیگر به دانشجویان عرضه می شود. همچنین وب به دانشجویان اجازه می دهد تا با یکدیگر و اساتید خود درمکان های مختلف به صورت همزمان یا غیرهمزمان تعامل کنند.

به طور خلاصه می‌توان گفت که مراکز آموزش الکترونیکی، محدودیتهای مراکز آموزش سنتی را نداشته و در بسیاری از جهات مزایای زیادی نیز نسبت به آنها دارند.

مزایای این آموزش به شرح زیر می باشد:

- ✓ محاوره ای: دانشجویان می‌توانند با سایر دانشجویان و اساتید به صورت برخط گفتگو کنند؛
- ✓ چند رسانه ای: اطلاعات فراگیری برخطی که به شکل محیط مناسب چند رسانه ای مانند متن، تصویر، انیمیشن، نوارهای سمعی و بصری و غیره است؛
- ✓ باز: دانشجویان میتوانند آزادانه به فراگیری برخط بپردازند یا از آن دست بکشند؛
- ✓ همزمان یا غیرهمزمان بودن ارتباط غیرمستقیم رایانه ای؛
- ✓ استقلال از مسافت و زمان بودن: دانشجویان می توانند دوره های برخط را در هر مکان و در هر زمانی بگذرانند؛

امروزه، بسیاری از افراد از انواع برنامه های آموزش الکترونیکی بهره مند شده اند. اگرچه تنوع بالای سلایق کاربران در اینترنت چالش های را در مدل یادگیری «one size fit all» ایجاد نموده که در آن مجموعهٔ واحدی از منابع آموزشی به تمام فراگیران ارائه می گردد. در حقیقت ممکن است

[1]. Hyper Media
[2]. World Wide Web

فراگیران سلایق متفاوتی داشته باشند، حتی در صورت داشتن سلایق مشترک، ممکن است سطوح تخصص متفاوتی داشته باشند. بنابراین نباید با آنها به روش یکسانی برخورد کرد.

سیستم مدیریت آموزش LMS/CMS پلتفرم آموزش الکترونیکی می باشد که از نقطه نظر دانشگاه به عنوان بخش مهمی از راه حل آموزش الکترونیکی در نظر گرفته شده است. در واقع نوعی نرم افزار مدیریت محتوای سایت است. برای اینکه راحت تر بتوانیم محتویات سایتمان را مدیریت کنیم و آنرا به روز رسانی کنیم، از اینگونه نرم افزارها کمک می گیریم.

این نرم افزارهای مدیریت محتوا به طور کلی،CMS[1] نامیده می شوند. که LMS ها نیز گونه ای از این نرم افزارها هستند. با این تفاوت که CMS کاربرد عام دارد و LMS کاربرد خاص. علاوه بر آن، مفاهیم دیگری مشابه LMS وجود دارند مانند سیستم مدیریت محتوای آموزشی (LCMS) و پورتال آموزش. در هر صورت، LMS نرم افزاری است که وقایع آموزشی را به طور خودکار مدیریت می نماید. بطور خاص سیستمهای مدیریت یادگیری به مجموعه ای از فرایند ها منتهی میشود که مدیران آموزشی را قادر می‌سازد واحد های آموزشی مجازی متناسب با نیازهای خود را طراحی و اجرا نمایند. از وظایف LMS، ادارهٔ ورود کاربران ثبت نام شده، مدیریت کاتالوگ های دوره، پی گیری نتایج و فعالیت های فراگیر و ارائهٔ گزارشات به مدیریت می باشد.

بازار LMS به سرعت در حال رشد بوده و بیش از ۷۰ فروشنده در این حوزه وجود دارد. برخی از LMSها نرم افزارهای تجاری و برخی دیگر کدباز و رایگان می‌باشند. که در زیر به چند نمونه اشاره می‌نماییم:

- LMS تجاری

WebCT <www.WebCT.com> و eCollege <www.ecollege.com>

- LMS کدباز

MOODLE<http://moodle.org> و ILIAS <www.ilias.de>

۲-۳ معرفی سیستم های توصیه‌گر

با توجه به اینکه یکی از مهمترین خصوصیات دنیای وب، سهل الوصول بودن جابجایی از یک منطقه به منطقهٔ دیگر است، کاربران وب را می توان بسیار سیار تلقی کرد. و با در نظر گرفتن این خصوصیت

[1] Content Management System

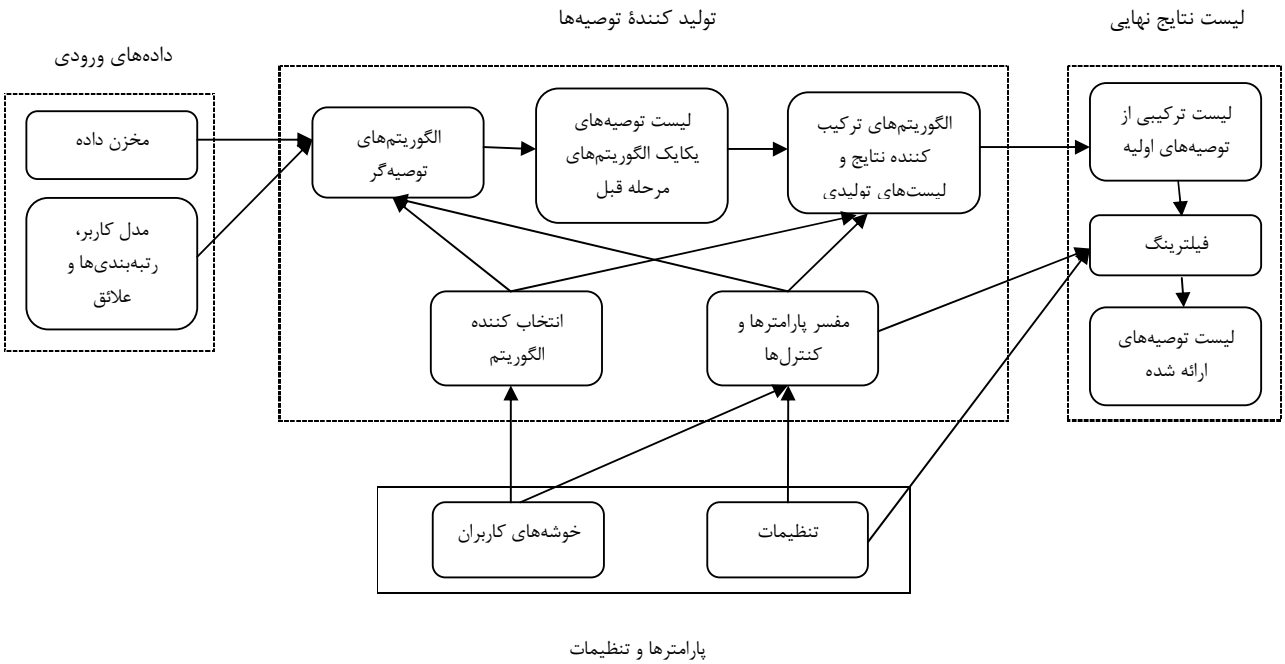
سیار بودن کاربران وب، اگر یک وب سایت نتواند به نیاز های یک مشتری، در مدت کوتاهی از زمان پاسخ مفیدی بدهد، کاربر به راحتی و به سرعت به وب سایت دیگری مراجعه می کند. بنابراین فهم و کشف نیازها و خصوصیات کاربران و استفاده کنندگان وب، برای طراحان وب سایت بسیار مهم است.

برای حل این مشکل، شخصی کردن وب یک پدیدهٔ محبوب به منظور سفارشی کردن محیط های وب تبدیل شده است. استفاده از این سیستم‌ها که با شناخت از کاربران و مشتریان همراه است، منجر به ایجاد خدمات با کیفیت و مناسب برای ارائهٔ فعالیت‌ها و فروش محصولات می‌گردد. در داقع دانش اولیهٔ این سیستم‌ها، همان اطلاعات موجود در مدل رفتار و مدل اولویت کاربر است.

هدف سیستم‌های شخصی ساز فراهم کردن نیازهای کاربران، بدون اینکه به طور صریح آنها را بیان کنند یا نشان بدهند، می باشد. ایدهٔ سیستم توصیه‌گر از ارائهٔ اطلاعات شخصی می آید. توصیه‌هایی که از سوی سیستم‌های توصیه‌گر ارائه می‌شوند به طور کلی می‌توانند دو نتیجه در برداشته باشند:

- ✓ کاربر را در اخذ تصمیمی یاری می‌کنند (که مثلاً از میان چندین گزینه پیش رو کدام بهتر است و آن را انتخاب کند و ...).
- ✓ موجب افزایش آگاهی کاربر، در زمینهٔ مورد علاقهٔ وی می‌شود (مثلاً در حین ارائه توصیه به کاربر موجب می‌شود تا وی با اقلام و اشیاء جدیدی را که قبلاً آنها را نمی‌شناخته، آشنا شود).

یک سیستم توصیه‌گر، سیستمی است که اطلاعات مفید و یا استراتژی‌های خاص را که کاربران برای رسیدن به اهدافشان به کار می‌گیرند، به آنها توصیه می‌نماید. چنین سیستمی ممکن است با توجه به یک رویداد مانند یک خطا و یا مشاهدهٔ رفتار کلی کاربران فعال گردد. شکل ۲-۱ نمای کلی از فرآیند سیستم‌های توصیه‌گر را نمایش می‌دهد. (Azene, Ozok, Norcio, 2005)



شکل ۲-۱ مدل کلی فرآیند توصیه

سیستم توصیه‌گر یا سامانه پیشنهاد دهنده، با تحلیل رفتار کاربر خود، اقدام به پیشنهاد مناسب‌ترین اقلام (داده، اطلاعات، کالا و...) می‌نماید. رویکردی است که برای مواجهه با مشکلات ناشی از حجم فراوان و رو به رشد اطلاعات ارائه شده‌است و به کاربر خود کمک می‌کند تا در میان حجم عظیم اطلاعات سریع‌تر به هدف خود نزدیک شوند. یک شرکت تجارت الکترونیکی باید به طراحی سیستم توصیه گر جهت جذب بازار مورد هدف خود بپردازد این کار نه تنها منجر به حفظ ترافیک سایت خواهد شد بلکه باعث افزایش رضایتمندی کاربران و در نهایت افزایش سوددهی خواهد گردید.

در سال های اخیر، سیستم توصیه گر توجه زیادی را مخصوصاً در حوزهٔ فناوری اطلاعات، به خود جلب نموده است (Wang, Chuang, Hsu, & Keh, 2004). یک سیستم توصیه گر، سیستمی است که اطلاعات مفید را توصیه نموده و یا استراتژی هایی که کاربران به منظور رسیدن به اهداف خود به کار می گیرند را پیشنهاد می نماید. یک توصیه ممکن است با انجام یک رویداد مانند یک خطا و یا مشاهدهٔ رفتار کلی کاربر فعال گردد. موتور جستجو یک نمونهٔ ساده از این سیستم ها می باشد که هنگامی که هیچ نتیجه ای برای یک جستجو یافت نشد، کلمات کلیدی و یا جستجوهای جایگزین را با امکان دریافت نتایج بهتر، پیشنهاد می نماید (Diamond Bullet, 2004). سیستم های توصیه گر موفق بسیاری هستند که به طور گسترده در حوزه های تجارت الکترونیک، فیلم،

موسیقی و کتاب و صفحات وب مورد استفاده قرار می گیرند. (Chen & Chen, 2001) یک سیستم توصیه گر موسیقی (MRS[1]) را طراحی کردند که سرویس شخصی ارائهٔ موسیقی را بر اساس گروه بندی داده های موسیقی و علایق کاربر فراهم می نمود.

(Carenini, Smith & Poole, 2003) مجموعه ای از تکنیک ها، جهت انتخاب هوشمندانه در خصوص اینکه چه اطلاعاتی را از کاربر استنباط نمایند، یافتند. از جمله سیستم‌های توصیه‌گر می‌توان به آمازون (Linden et al., 2003)، سیستم توصیه‌گر کتاب اشاره نمود. شرکت آمازون یکی از سایت های موفق در زمینه فروش کتاب و برخی محصولات مشابه است. این شرکت سایت خود را به گونه ای طراحی نموده است که هر فرد که به سایت وارد می شود داده های خود را به سیستم وارد می کند. بدین ترتیب نیمرخی از کاربران ایجاد می کند که به عنوان منبع دانش مشتری مورد استفاده قرار می گیرد و هر بار که کاربر وارد سایت این شرکت می شود، این شرکت کاربر خود را شناخته و به او سلام می کند و بر اساس الگوهای رفتاری استخراج شده از وی، کتابهایی را به او پیشنهاد می نماید که در بخش توصیهٔ کتاب[2] آنها را به مصرف کننده پیشنهاد می کند. باید توجه داشت که پیشنهاد کتاب به کاربران بر اساس الگوی رفتاری استخراج شده می باشد و تصادفی نیست.

Netflix، imdb، Yahoo Movies، سیستم توصیه‌گر فیلم، Flicker Explore، سیستم توصیه‌گر عکس‌های برگزیده، Google News، سیستم توصیه‌گر اخبار، Face Book، سیستم توصیه‌گر دوست، PhotoTree، سیستم توصیه‌گر عکس، Google Reader Discover، سیستم توصیه‌گر RSS، Grundy، اولین سیستم توصیه‌گری که از الگویی مرسوم به Stereotype به عنوان مکانیزمی برای ساخت مدل‌هایی از کاربران بر اساس مقدار محدودی از اطلاعات فردی کاربر استفاده نموده و با استفاده از این الگو، مدل‌های کاربر فردی را ساخته و از آنها در توصیهٔ کتب مربوط به هر کاربر استفاده می‌نمود، Tapestry، سیستم توصیه‌گر به کاربران همفکر و متجانس و سایر سیستم‌ها مانند Grouplens، Video Recommender و Ringo اشاره نمود.

Ringo به کاربران این امکان را می دهد که توصیه های موسیقی را به صورت آنلاین دریافت و با سایر علاقمندان موسیقی ارتباط برقرار نمایند. (Shardanand, Maes, 1995)

[1] Music Recommender System
[2] Book Recommendations

۲-۳-۱ تعاریف اولیه در سیستم‌های توصیه‌گر

لازم است برای درک مفهوم سیستم توصیه‌گر، مفاهیم ابتدایی زیر را بررسی کنیم.

- در سیستم‌های توصیه گر به کاربری که توصیه جاری در سیستم، برای وی در حال پردازش و آماده شدن است، کاربر فعال یا کاربر هدف گفته می‌شود.
- الگوریتم‌های به کار رفته در این سیستم‌ها، از ماتریسی به نام ماتریس رتبه‌دهی یا نرخ‌دهی[1] استفاده می‌کنند؛ ساختار ماتریس رتبه‌ها بدین گونه است که در آن، هر سطر ماتریس نمایانگر یک کاربر و هر ستون آن معرف یک آیتم یا قلم خاص است. نمونه‌ای از این ماتریس در جدول ۲-۱ آمده است. این جدول، مثالی از ماتریس نرخ‌دهی کاربر-آیتم برای پیشنهاد اقلام است که نرخ‌ها بین ۱ تا ۵ است. نماد Ø، به این معنا است که کاربر نرخی به آن آیتم نداده است، بنابراین سیستم توصیه‌گر بایستی بتواند اقلام و آیتم‌های نرخ داده نشده و انواع مناسب پیشنهاد بر اساس این نرخ‌ها را تخمین زده و پیش بینی نماید.
- از عبارت *مصرف کردن* در سیستم‌های توصیه‌گر، زمانی استفاده می‌کنند که کاربر توصیه ارائه شده توسط سیستم را می‌پذیرد. به عبارت دیگر، وقتی کاربری توصیه‌ای را که توسط سیستم به وی شده است را می‌پذیرد، گفته می‌شود کاربر آن توصیه را مصرف کرده است، این پذیرش می‌تواند به شکل‌های مختلفی باشد، مثلاً کاربر، کتاب پیشنهادی را می‌خرد، سایت پیشنهادی را مرور می‌کند و یا به شرکت خدماتی ای که به او پیشنهاد شده مراجعه می‌کند.
- مورد بعدی، مفهوم تابع سودمندی است. به منظور توضیح نحوهٔ پیاده‌سازی سیستم‌های توصیه‌گر، به لحاظ اینکه قصد داریم به کمک آن، یک مدل کلی ریاضی از سیستم‌های توصیه‌گر را ارائه دهیم، با مفهوم تابع سودمندی آشنا می‌شویم. در واقع یک سیستم توصیه‌گر را می‌توان با این نگاشت همسان دانست و مدل کرد:

$$u: C \times S \rightarrow R \qquad (۱-۲)$$

[1] Rating Database/Preference Database

جدول۲-۱ ماتریس نرخ‌دهی کاربر- آیتم

	آیتم A	آیتم B	آیتم C	آیتم D
کاربر A		۴	۵	۳
کاربر B	۵	۲		۱
کاربر C	۲	۴	۵	

فرض کنید C مجموعه تمامی کاربران و S مجموعه اقلام در دسترس باشند. تابعی را که میزان مفید و متناسب بودن کالای $s \in S$ را برای کاربر $c \in C$ را محاسبه می‌کند با u نشان می‌دهیم، که در آن R، مجموعه‌ای کاملاً مرتب براساس میزان اهمیت است. R_{cs} نرخ‌گذاری و رتبه‌ای است که کاربر c بر روی آیتم s انجام می‌دهد. هرکدام از عناصر S را می‌توان با مجموعه‌ای از خصوصیات خود (پروفایل)، توصیف نمود. برای مثال، محصولی مثل فیلم را می‌توان با مشخصه‌هایی چون عنوان فیلم، کارگردان، طول زمانی فیلم، تاریخ تولید و غیره ثبت کرد. همچنین عناصر مجموعه C را نیز می‌توان بر اساس ویژگی‌های مثل سن، جنسیت و غیره ثبت کرد. سپس برای هر کاربر $c \in C$ کالایی مانند $s \to S$ را که سودمندی کاربر را ماکزیمم نماید، به صورت قراردادی به صورت زیر تعریف می‌کنیم:

$$\forall\, c \in C,\ S_c = argmax_{s \in S} u(c,s) \qquad (۲-۲)$$

- مورد آخر، فرآیند شناخت و طبقه بندی کاربران است. باید توجه داشت، اطلاعاتی که یک سیستم توصیه‌گر از کاربران بدست می‌آورد، هم از لحاظ نوع و هم از لحاظ نحوه بدست آمدن متفاوتند. به دریافت اطلاعات توسط سیستم و به طور کلی فرآیند شناخت کاربران، نرخ گذاری گفته می‌شود. در سیستم‌های توصیه‌گر، سودمندی یک آیتم کالا معمولاً بوسیلهٔ نرخ گذاری مشخص می‌شود و بیان می‌کند چگونه کاربر خاصی، محصولی را دوست دارد. به عنوان مثال کاربر شماره ۱ به محصول شماره ۵، امتیاز ۶ از ۱۰ را می‌دهد.

به طور کلی فرآیند شناخت کاربران به دو روش زیر انجام می‌شود:

۱. روش مبتنی بر دانش (صریح)[1]

در این روش، مدل‌های ثابتی از کاربران ایجاد شده و سپس کاربران در نزدیکترین گروه‌های مشابه و هم سلیقه گروه‌بندی می‌شوند. اطلاعات دانش در مورد کاربران از راه‌های مختلفی بدست می‌آید. (Nayak, Seow, 2002)

اولین راه، پرسش از کاربر برای انتخاب بین انواع مفاهیم و خدمات است. علی رغم مزیت کسب اطلاعات مستقیم و دقیق از کاربر، این روش همیشه و برای همهٔ کاربران پاسخگو نیست. در مواردی که کاربر حوصلهٔ پاسخگویی به سوالات را ندارد، این روش ممکن است که کاملاً اطلاعات اشتباهی را ارائه نموده و گاهی مجبور به استناد به اطلاعات قدیمی و منسوخ نیز شود. لذا ممکن است بهتر باشد این اطلاعات، به خصوص اطلاعات نا ثابت و متغیر از کاربر، به طور غیر مستقیم گرفته شده و به صورت مفهومی پردازش و سپس از آن به عنوان پایه‌ای برای توصیه استفاده شود.

۲. روش مبتنی بر رفتار (ضمنی)[2]

این روش، از مدل رفتار کاربر استفاده می‌نماید. در این روش با استفاده از تکنیک‌های یادگیری بر اساس فعالیت‌های کاربر در وب‌سایت‌ها، رفتارهای با قاعدهٔ او شناسایی و سپس یک مدل به نام مدل کاربر ساخته می‌شود. مدل کاربر مجموعه‌ای از اطلاعات درباره رفتار کاربر است و به شناسایی بیشتر کاربر، که چه کرده و حدس اینکه در آینده چه خواهد کرد، کمک می‌نماید. به طور نمونه، در بعضی از سیستم‌های توصیه، ماشین‌های یادگیری تعداد خریدها و جستجوهای هر کاربر را ذخیره و سپس بررسی نموده و علایق او را بر این اساس استخراج می‌نماید.

۳. روش ترکیبی

برخی از سیستم‌های توصیه‌گر، از هر دو روش برای تکمیل فرآیند شناسایی کاربران خود و سپس ارائهٔ توصیه بهره می‌برند. از جمله این سیستم‌های توصیه گر می‌توان سیستم EUP[3]را نام برد. EUP یک سیستم توصیه‌گر است که توصیه‌های شخصی در مورد صفحاتی که دارای کاتالوگ‌های جذاب برای کاربران است، را ارائه می‌دهد. در این سیستم،

[1] Knowledge-based (Explicit)
[2] Behavioral-based (Implicit)
[3] User Profile Engine

خصوصیات کاربران، هم از نوع متغیر و هم از نوع ثابت است. خصوصیات ثابت، اطلاعاتی هستند که برای هر کاربر مشخص بوده و تغییر نمی‌کند و یا به ندرت تغییر می‌کند و از فرم‌های ثبت نام و اطلاعات درخواستی از کاربر دریافت می‌شود (مانند نام کاربر). اطلاعات متغیر یک کاربر، خصوصیاتی از کاربر هستند که ممکن است حتی در هر بازدید، تغییر نماید (مانند اولویت‌های کاربر). (Clare-Marie, Jan, 2004)

در جدول ۲-۲ نمونه‌ای از انواع داده‌ها و اطلاعاتی که به کمک روش‌های فوق به دست می‌آیند، نشان داده شده است.

جدول۲-۲ نمونه هایی از روش‌های جمع آوری اطلاعات و شناخت کاربران

روش	نحوهٔ کسب اطلاعات	نمونهٔ آیتم‌ها
مبتنی بر دانش	داده‌های شخصی	نام، جنس، سن، شغل،درآمد، آدرس
	پرس و جو های شخصی	سطح تخصص کاربر، حوزه‌های مورد علاقهٔ کاربر، علایق مرتبط
مبتنی بر رفتار	جنبه‌های آماری	تعداد بازدیدها، میزان زمان صرف شده در هر های بازدید شده، URLبازدید/صفحه، ترتیب فرآیند جستجو
ترکیبی	تکنیک‌های مفهومی در ترکیب داده‌های صریح و ضمنی	اعتماد میان کاربران، بازدیدهای مشابه، میزان تخصص در حوزهٔ خدمات، خریدهای محتمل و غیره

۲-۳-۲ مزایای سیستم‌های توصیه‌گر

سیستم‌های توصیه‌گر مزایایی را فراهم می‌آورند که در ذیل به مهمترین آنها اشاره شده است:

- ✓ در یک تعامل تجاری، کاربران از این جهت که عمل جستجو و انتخاب بهترین گزینه متناسب با خصوصیات شخصی‌اشان، در میان حجم عظیم اطلاعات تسهیل و تسریع می‌شود، استفاده از این سیستم‌ها را مفید می‌دانند.
- ✓ این سیستم‌ها موجب افزایش آگاهی کاربر و اکتشاف فضاهای جدید، در زمینهٔ مورد علاقهٔ وی می‌شود. (به عنوان مثال، در حین ارائهٔ توصیه به کاربر، وی با اقلام جدیدی که قبلاً آنها را نمی‌شناخته است، آشنا می‌شود.)
- ✓ طرف‌های تجاری به کمک این سیستم‌ها می‌توانند رضایت مشتریان را بالا برده و فروش خود را نیز افزایش دهند.

- ✓ در سیستم‌های توصیه‌گر امروزی، محدودیت‌های سیستم‌های قدیمی مبتنی بر جستجو و بازیابی اطلاعات که قادر به تشخیص و تفکیک اقلام با کیفیت و بی‌کیفیت در ارائهٔ پیشنهاد برای یک موضوع یا کالا نبودند، نیز حل شده است.
- ✓ به لحاظ اینکه بر اساس فعالیت کاربران و گردآوری رفتار و سلایق آن‌ها عمل می‌کند. به این ترتیب پیشنهادها بر اساس حدس و گمان نخواهد بود.
- ✓ با استفاده از هوش جمعی که چکیده رفتار هزاران کاربر است، می‌توان از جدیدترین و به روزترین اتفاقات در زمینه‌های مورد علاقه مطلع گردید.
- ✓ در سیستم‌های توصیه‌گر، با توجه به این‌که کاربران اطلاعات را بر اساس سلایق خود شکل می‌دهند، دستکاری و مدیریت اطلاعات وب سایت‌ها و به روزرسانی آن توسط خود کاربران انجام می‌شود. به این ترتیب هزینه نگهداری و سازماندهی آنها در مقایسه با وب‌سایت‌های معمولی بسیار کمتر است.

۲-۳-۳ کاربردهای سیستم توصیه‌گر

سیستم های توصیه گر به طور گسترده در بسیاری از فعالیت های اینترنتی استفاده شده اند. که در این بخش برخی از نمونه هایی که در حال حاضر از این سیستم ها استفاده می نمایند، ذکر می گردد:

- ➢ تجارت الکترونیک: تجارت الکترونیک رسانهٔ مهمی در تبادل محصولات و خدمات می باشد که از سیستم های توصیه گر جهت پیشنهاد محصولات به مشتریان و ارائهٔ اطلاعات کمکی به آنها به منظور تصمیم گیری بهتر در خرید استفاده می کنند. به عنوان نمونه سایت Amazon.com و Barnesnobel.com از این دست می باشند.
- ➢ صفحات وب: محققان به منظور غلبه بر مشکل حجم عظیم اطلاعات، به طور مؤثر از سیستم های توصیه گر در این حوزه استفاده نموده اند. در هنگام استفاده از موتورهای جستجو مانند yahoo و Google که برای یک جستجوی موردی چندین هزار صفحه فراهم می نمایند این مشکل به وضوح دیده می شود. اکثر صفحات یافت شده نیز ارتباط کمی با محتوی جستجو شده دارند.
- ➢ سیستم های سانسور: از سیستم های توصیه گر نیز به منظور حفاظت در حوزه های زیر استفاده می گردد:
  - جلوگیری از دسترسی کودکان به موارد نامطلوب در اینترنت مانند cyberpatrol.com

- جلوگیری دسترسی و کاوش شهروندان در برخی وب سایت ها. که در برخی از دولت ها پیاده سازی شده است.

- اینترانت‌های بنگاهی: برای پیدا کردن افراد خبره در یک زمینه خاص و یا افرادی که در رویارویی با شرایط مشابه، تجاربی کسب کرده و راه حل‌هایی یافته‌اند(بیشتر داخل یک سازمان کاربرد دارد).
- کتابخانه دیجیتال: پیدا کردن کتاب، مقاله و غیره؛ سایت www.elibraryhub.com
- کاربردهای پزشکی: انتخاب پزشک متناسب با شرایط (مکان، نوع بیماری، زمان) بیمار، انتخاب دارو و غیره.
- مدیریت ارتباط با مشتری CRM[1]: برای ارائه راهکارهایی برای حل مشکلات تولیدکننده و مصرف‌کننده در زنجیره تأمین.

سایر بخش ها مانند:

- اخبار؛ سایت www.lemonde.fr
- دایرۀ المعارف؛ سایت http://en.wikipedia.org
- نرم افزار؛ سایت www.download.com
- فروشگاه ها؛ سایت www.drugstore.com
- اطلاعات گردشگری؛ سایت www.viamichelin. com

۲-۳-۴ دلایل تفاوت سیستم‌های توصیه‌گر در محیط‌های آموزش الکترونیکی نسبت به سایر حوزه‌ها:

ارائۀ پیشنهادات در محیط های آموزشی نسبت به حوزه های دیگر (به طور مثال حوزۀ پیشنهاد فیلم که تاکنون بسیار مورد بررسی قرار گرفته است)، متفاوت می باشد. موارد خاصی برای سیستم توصیه گر آموزش الکترونیکی وجود دارد که عبارتند از:

- ممکن است مطالب مورد علاقۀ فراگیران الزاماً از نظر آموزشی مناسب آنها نباشد. به طور نمونه، یک فراگیر بدون مطالعه و پیشینۀ قبلی در حوزۀ تکنیک ها و روش های وب کاوی، ممکن است به جدیدترین و پیشرفته ترین تکنولوژی ها در این حوزه علاقمند باشد. در این صورت سیستم به این فراگیر مطالعۀ تعدادی مقالات مروری و دوره ای را پیشنهاد می دهد. اگرچه مقالات فنی با کیفیت بالای بسیاری نیز با توجه به علاقۀ وی موجود می باشد. از طرف دیگر، برای فراگیری که

[1] Customer Relationship Management

از حوزه های صنعتی وارد شده و تجربه و دانش قبلی در این زمینه دارد و تمایل به دانستن چگونگی استفاده از وب کاوی در حل مشکلات تجارت الکترونیکی دارد، مقالات فنی با کیفیت بالاتری پیشنهاد می گردد. در مقابل در حوزه های دیگر، پیشنهادات صرفاً براساس علایق کاربر ساخته شده اند.

- سفارشی سازی نباید فقط در مورد انتخاب عناوین اقلام و موارد آموزشی تأکید داشته باشد، بلکه می بایست در مورد نحوهٔ ارائه و ترتیب آنها نیز مورد توجه قرار گیرد. به طور نمونه، برخی از مدرسان به فراگیران خود پیشنهاد می نمایند که قبل از مطالعهٔ یک مقالهٔ فنی، یک مقالهٔ جذاب مرتبط را از یک مجله مطالعه نمایند. به این دلیل که آنها اعتقاد بر این دارند که این کار به درک بهتر فراگیر و همچنین کاهش سردرگمی و ترس وی کمک شایانی می نماید. که همان طور که می دانیم این مورد در پیشنهادات تجارت الکترونیک مشاهده نمی گردد. در آنجا مدیران سایت ترجیح می دهند که لیستی از اقلام را به طور نامرتب و بدون هیچ ترتیبی به کاربر پیشنهاد نمایند و از توصیهٔ یک پیشنهاد خاص به عنوان بهترین انتخاب، اجتناب می نمایند.

بدین ترتیب سیستم های توصیه گر در حوزهٔ آموزش الکترونیکی می بایست جوانب بیشتری را در نظر گیرند. به عنوان مثال در مورد مقالات، علاوه بر در نظر گرفتن دسته های اصلی پژوهشی، سطوح فنی آنها و یا مقالات دوره ای (مروری)، مقالات کارگاه آموزشی، مقالات فنی با کیفیت بالا و غیره را نیز مد نظر داشته باشند.

از طرف دیگر، ارائهٔ پیشنهادات در زمینهٔ سیستم های هوشمند کمک درسی، نسبت به حوزه ها دیگر دست یافتنی تر خواهد بود. بدین ترتیب که علاقه، هدف و سطح دانش فراگیر در یک محیط آموزشی، قابل پیش بینی خواهد بود. (Tang, McCalla, 2005)

۲-۴ رویکرد های مختلف سیستم های توصیه گر

در رایج‌ترین تقسیم‌بندی، سیستم‌های توصیه‌گر به طور کلی به پنج دسته تقسیم می‌شوند:

۱. سیستم‌های مبتنی بر دانش[1]؛
۲. سیستم‌های مبتنی بر محتوا[2]؛
۳. سیستم‌های مبتنی بر آمارگیری؛
۴. سیستم‌های مشارکت جمعی[1]؛

[1] Knowledge-based
[2] Content-based

۵. سیستم های ترکیبی[۲]

بسیاری از سیستم های توصیه گر از رویکرد مبتنی بر محتوا و یا مشارکت جمعی جهت ارائهٔ پیشنهادات به کاربران استفاده می نمایند. (Resnick, Varian, 1997)

۲-۴-۱ سیستم مبتنی بر دانش

این سیستم‌ها با ادراک و یادگیری نیازها، اولویت‌ها و علایق کاربر و همچنین ویژگی‌های اقلام سعی بر ارائهٔ توصیه دارند. به نوعی همه روش‌های توصیه و پیش بینی، زیر مجموعه این روش هستند، زیرا با بررسی نیازها و علایق مشتری با روش‌های متفاوت، توصیهٔ کالا را انجام می‌دهند. به عبارتی در این گونه از سیستم‌های توصیه‌گر مواد اولیه مورد استفاده برای تولید لیستی از پیشنهادها، دانش سیستم در مورد مشتری و کالا است. سیستم‌های مبتنی بر دانش از متدهای مختلفی مانند الگوریتم‌های ژنتیک، فازی، شبکه‌های عصبی، درخت‌های تصمیم و به ویژه استدلال‌های مبتنی بر نمونه نیز برای تحلیل دانش بهره می‌برند. (Nayak, Seow, 2002)

۲-۴-۲ سیستم مبتنی بر محتوا

یک مجموعه از اسناد و اطلاعات که توسط یک کاربر رتبه بندی و ارزش گذاری شده است، مورد تحلیل و بررسی قرار می‌گیرد و از مفاهیم و محتوای این اسناد و اطلاعات و بر اساس رتبه‌های داده شده، مدلی که می‌تواند برای توصیه و پیش‌بینی کالاها و خدمات مورد علاقهٔ مشتری استفاده شود، استخراج و استنباط می‌گردد. (Glover, Lawrence, Gordon et al., 2000)

در واقع، در این روش، اقلام پیشنهادی، به این دلیل که با اقلامی که کاربر فعال (کاربری که قرار است به او توصیه کنیم) نسبت به آنها ابراز علاقه کرده‌است شباهت‌هایی دارند، به وی توصیه می‌شوند. در حقیقت، اقلام با توجه به ارتباط میان محتوایشان و اولویت های کاربر انتخاب می گردند. به طور نمونه: http://infofinder.cgiar.org. وب سایت‌های زیادی برای استفاده از این روش جهت ارائهٔ محصولات و خدمات خود رشد و توسعه پیدا کرده‌اند. ساده‌ترین آنها در ابتدا با دریافت نکات و خصوصیا از کاربر و آنچه کاربر می‌پسندد، مرحله به مرحله همانطور که پیش از این در شکل ۲-۱ توضیح داده شد، کالاها و خدمات را برای او فیلتر نموده و انتخاب را برای او ساده می‌نماید.

---

[3] Collaborative filtering
[4] Hybrid Recommender System

سایت مشهور Amazon.com نیز بر اساس خریدهای قبلی مشتریانش و محصولات ارزیابی شده مشابه توسط او، فرآیند فیلتر را انجام داده و به مشتریانش بر این اساس توصیه ارائه می‌دهد.

در مورد LMSها، سیستم های مبتنی بر محتوا جهت پیشنهاد محتویات آموزشی و به عنوان یک رویکرد آغازگر با شناسایی شباهات میان صفات دورهٔ فعلی (نام، کلمات کلیدی، چکیده و غیره) و سایر دوره‌ها به کار گرفته می‌شوند.

سیستم‌های توصیه‌گر مبتنی بر محتوا محدودیت‌هایی نیز دارند (Balabanovic, Shoham, 1997) که در ذیل بدان‌ها اشاره شده است:

- این روش در زمانی که تعداد کاربران زیاد شده و آیتم‌های زیادی نیز بایستی از نظر محتوا مورد بررسی قرار گیرند، زیاد مناسب نیست. (Papagelis, Plexousakis, 2005)
- در رویکرد مبتنی بر محتوا، سیستم فقط و فقط آیتم‌هایی شبیه به پروفایل کاربر را به کاربر پیشنهاد می‌دهد، در نتیجه کاربر فقط آیتم‌هایی مشابه به آنچه را که قبلاً نرخ‌گذاری کرده است، می‌تواند مشاهده کند.
- الگوریتم مبتنی بر محتوا توسط مشخصه‌های اقلام محدود می‌شود و بنابراین نیاز به مجموعه کافی و متناسبی از مشخصه‌ها است تا سیستم توصیه‌گر یا به صورت دستی و یا اتوماتیک، مشخصه‌ها را تجزیه نموده و خصوصیات آنها را به منظور پیشنهاد آیتم کشف نماید. تکنیک‌های بازیابی اطلاعات تا زمانی که آیتم‌ها به صورت متنی باشند، به خوبی می‌تواند مشخصه‌ها را استخراج کنند اما انواع دیگری از آیتم‌ها به طور ذاتی با مسألهٔ استخراج اتوماتیک مشخصه‌ها مشکل دارند. به عنوان مثال، متدهای استخراج اتوماتیک مشخصه‌ها در مورد داده‌های مالتی مدیا مانند عکس‌های گرافیکی، داده‌های صوتی و داده‌های ویدیویی، با مشکلات فراوانی مواجه هستند و علاوه بر این اغلب امکان پذیر نیست که خصوصیات را به صورت دستی وارد نماییم. از طرفی این روش در مورد توصیهٔ بعضی از آیتم‌ها که تحلیل محتوا و خصوصیات آنها (از جمله تعریف عقاید و ایده‌های کاربران در مورد این آیتم‌ها)، برای کامپیوتر مشکل است، دارای کاستی‌هایی است و کیفیت توصیه را پایین می‌آورد.
- سیستم مبتنی بر محتوا فقط زمانی می‌تواند به کاربر پیشنهاد قابل اعتماد بدهد که کاربر به تعداد کافی آیتم از قبل نرخ داده باشد و سیستم بتواند علایق کاربر را تشخیص دهد.

بنابراین، در خصوص یک کاربر جدید، که نرخ‌های کمتری به آیتم‌ها داده است، سیستم توصیه‌گر قادر نیست به وی توصیهٔ درستی ارائه کند.

همانطور که ذکر گردید، در رویکرد مبتنی بر محتوا از اطلاعات و خصوصیات اقلام در ارائهٔ پیشنهادات استفاده می‌شود. در واقع در روش محتوا محور، اقلام پیشنهادی، به این دلیل که با اقلامی که کاربر فعال نسبت به آنها ابراز علاقه کرده است شباهت‌هایی دارند، به وی توصیه می‌شوند. (Wang, Chuang, Hsu, & Keh, 2004) از رویکرد مبتنی بر محتوا جهت مقایسهٔ محتویات و پروفایل کاربران استفاده نمودند و اقلام مشابه با آن چیزی که کاربر در گذشته بدان علاقمند بود را به وی پیشنهاد کردند. اغلب از برخی طرح‌های توزین استفاده می‌گردد که به کلمات مشخص وزن بالایی اختصاص داده می شود. هنگامی که آیتمی انتخاب گردید، به کاربر نمایش داده شده و بازخورد ها جمع آوری و سپس تجزیه وتحلیل و استنباط می گردد. اگر کاربر به آن علاقه داشت، وزن آن قلم به پروفایل کاربر افزوده می‌شود. به چنین فرایندی بازخورد مربوط به بروز رسانی پروفایل کاربر گفته می‌شود. همانطور که Mooney و Roy اشاره کردند، رویکرد مبتنی بر محتوا، قادر به توصیهٔ اقلامی به کاربر با علایق منحصر به فرد می باشد که در گذشته امتیاز دهی نشده‌اند. به همین دلیل است که این رویکرد بسیار در کاربردهای عملی به کار گرفته می‌شود. (Mooney & Roy, 2000)

مثال های زیادی وجود دارد. به طور نمونه (Gauch, Gauch, Bouix & Zhu, 1999) یک سیستم بلادرنگ تشخیص و طبقه بندی صفحات ویدئویی را پیشنهاد نمودند.

۲-۴-۳ سیستم مبتنی بر آمارگیری

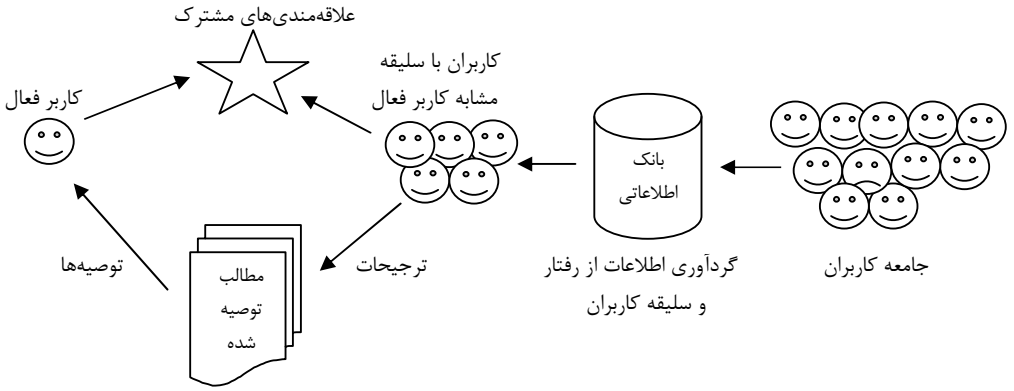
کاربران را بر اساس خصوصیات فردی تقسیم بندی نموده و توصیه را بر اساس تقسیم بندی‌های آماری ارائه می‌دهد. در حقیقت این سیستم، از دانش قبلی بدست آمده از اطلاعات دموگرافیک فراگیر و نظراتش به منظور ارائهٔ پیشنهادات استفاده می‌نماید. هدف آن گروه بندی فراگیران با توجه به خصوصیات و داده های صریح شخصی و در نهایت تولید پیشنهادات براساس گروه خاص دموگرافیکی است که فراگیر بدان تعلق دارد. مانند درآمد، سن، سطح یادگیری، منطقهٔ جغرافیایی و غیره و یا ترکیبی از این گروه‌ها.

به عنوان مثال، Grundy یک سیستم توصیه گر کتاب است که از توضیحات افراد از خود به منظور ساخت یک مدل کاربر استفاده نموده و سپس کتاب‌هایی را با ویژگی‌های مورد علاقهٔ آنها به

آنها پیشنهاد می‌نماید. و همچنین عرضه‌کنندگان ایمیل رایگان تبلیغات خود را بر اساس اطلاعات دموگرافیک کاربران ارائه می‌دهند. مانند سیستم توصیه گر موجود در Hotmail و yahoo.

۲-۴-۴ سیستم‌های مشارکت جمعی

در این رویکرد به جای استفاده از محتوای اقلام، از نظرات و رتبه‌بندی‌های انجام شده توسط سایر کاربران برای ارائه توصیه به کاربر، استفاده می‌شود. در این رویکرد لیست اقلام پیشنهادی بر اساس این اصل که کاربرانی مشابه کاربر هدف از آن‌ها رضایت داشته‌اند، تهیه می‌شود. لذا تمرکز روی یافتن شباهت بین کاربران است. به عبارت دیگر، الگوریتم‌های این رویکرد، بر این پایه استوارند که کسانی که در گذشته با هم توافق داشته اند، به احتمال زیاد در آینده نیز توافق خواهند داشت. یعنی کاربرانی که در گذشته رفتار مشابهی از خود ابراز داشته‌اند، می‌توانند در مورد اقلام نرخ گذاری نشده نیز در مورد یکدیگر به عنوان توصیه‌گر رفتار کنند. عملکرد این سیستم بدین گونه است که با تجزیه و تحلیل آماری اطلاعات یا استخراج داده های کاربر، رفتار گذشتهٔ وی در سایت و سایر اطلاعات، یک محدودهٔ همسایگی از افراد با سلایق و علائق مشترک ایجاد نموده و سپس با یافتن نزدیکترین همسایه‌ها برای هر کاربر به توصیهٔ نرخ‌گذاری‌ها و انتخاب‌های این همسایگان به کاربر هدف می پردازد. شکل ۲-۲ شمای ساده عملکرد یک سیستم توصیه‌گر مبتنی بر مشارکت جمعی را نشان می‌دهد. (Chen, McLeod, 2004)



شکل۲-۲ توصیف سادهٔ نحوهٔ عملکرد یک سیستم توصیه‌گر مبتنی بر مشارکت جمعی

این رویکرد برای هر گونه محتوایی پر کاربردتر است زیرا قادر به استخراج مفاهیمی مانند کیفیت اقلام است که برای نمایش مشکل هستند. همچنین برای محتواهایی که قابل توضیح و قابل ارتباط با کالاها نبوده و یا قابل تحلیل به کمک کامپیوتر نیستند، بسیار مؤثر است. علاوه بر این رویکرد

مشارکت جمعی محدودهٔ توصیه‌ها را به موارد مشابهی که کاربر قبلاً آنها را نرخ گذاری کرده، محدود نمی‌کند و به عنوان جدید ترین و گسترده ترین تکنیک توصیه مورد استفاده قرار می‌گیرد. (ضرغامی، عمادی، ۱۳۸۸)

یک سیستم مبتنی بر مشارکت جمعی، توانایی ارائهٔ توصیه‌های سخت و غیر مترقبه را نیز دارد و شامل محتوای کالا و خصوصیات مورد علاقهٔ فرد نمی‌شود، به همین دلیل از این روش در اکثر حوزه‌ها استفاده می‌شود. (Kohrs, Merialdo, 2001) بنابراین لزوم بکارگیری رویکرد مبتنی بر مشارکت جمعی پر رنگ تر است.

در خصوص محیط‌ها و سایت‌های آموزش الکترونیکی، رفتار فراگیران مورد تحلیل قرار گرفته، فراگیرانی که رفتار و سطح دانش مشابه فراگیر جاری دارند پیدا شده (بر اساس رتبه بندی) و فعالیت های آنها به فراگیر هدف پیشنهاد می شود. به مجموعه فراگیران مشابه و هم سطح، همسایگی فراگیر جاری گفته می‌شود. نگاشت بین سابقهٔ یک فراگیر به همسایگانش می‌تواند بر مبنای شباهت رتبه-بندی دروس (واحدهای درسی)، مطالب، دسترسی به صفحات با محتوای مشابه و یا انجام تکالیف و آزمون های مشابه انجام شود. همسایگی بدست آمده سپس برای توصیهٔ آیتم ها و اقداماتی که توسط فراگیر جاری مورد دسترسی قرار گرفته، استفاده می‌گردد.

(Sarwar, Karypis, Konstan & Riedl, 2002) اظهار داشتند که سیستم های توصیه گر از تکنیک های کشف دانش مانند مشارکت جمعی، جهت توصیهٔ محصولات در طول تعامل زندهٔ کاربر استفاده می‌نمایند.

همانطور که گفته شد در رویکرد مشارکت جمعی به نرخ‌گذاری مناسب یک مورد که قبلاً توسط کاربر هدف نرخ‌گذاری نشده است، توجه ویژه‌ای مبذول می‌گردد. برای مثال در یک برنامه توصیهٔ فیلم به کاربر، سیستم توصیه‌گر مبتنی بر مشارکت جمعی سعی می‌کند کاربرانی که سلیقهٔ مشابهی در انتخاب فیلم به کاربر C داشته‌اند را پیدا کرده و فقط فیلم‌هایی که کاربران مشابه دوست دارند به کاربر C توصیه نماید. روش فیلترینگ همکارگونه و یا مشارکت جمعی، اولین اقدام جهت استفاده از هوش مصنوعی برای دستیابی به فرآیند شخصی سازی فروش کالاها است. (Liu, Shih, 2005)

با تمام این مزایا این روش دارای ایراداتی نیز می‌باشد:

- چالش نخست مشارکت جمعی، مسألهٔ نهفتگی[1] است. این مسأله زمانی رخ می‌دهد که آیتمی تازه به سیستم وارد شده باشد در این صورت آن آیتم توسط کاربری نرخ داده نشده است،

[1] Latency

بنابراین آیتم نمی‌تواند انتخاب شود و به کاربران پیشنهاد داده شود، بنابراین کلاً حذف می‌شود. اگر از دید دیگری به این مسأله نگاه کنیم، کاربری که تازه به سیستم وارد شده باشد و به هیچ کالایی نرخ نداده باشد، در این صورت، سیستم نمی‌تواند علایق کاربر را تشخیص دهد و آیتمی را به کاربر پیشنهاد دهد. (Adomavicius, Tuzhilin, 2005)

- چالش دوّم، در مورد زمان حضور یک کاربر متفاوت و ناهمسان با گروه‌های کاربران است. چون این کاربر نیاز به توجهی متفاوت نسبت به سایرین دارد، تا توصیهٔ درست به او ارائه گردد. همچنین اگر یک مشتری تعداد کالاهایی که رتبه داده است، کم باشد، ممکن است همسایگان مناسبی برای او تعیین نشده باشند و در نتیجه توصیه‌ای نامناسب به او ارائه گردد.
- چالش دیگر الگوریتم‌های مشارکت جمعی، مبحث توسعه پذیری[1] آن است. (Das, Datar, Grag, 2007) الگوریتم‌های مشارکت جمعی، زمانی که تعداد کاربران، صدها یا هزاران نفر باشد، به خوبی پاسخ می‌دهد. اما زمانی که تعداد کاربران بیش از میلیون‌ها تن باشد و داده‌های کاربران پویا بوده و در یک دورهٔ زمانی کوتاه تغییر کند، کاربران جاری نیز ممکن است الگوهای رفتاری خود را تغییر دهند و یا در هر زمانی وارد سیستم گردند. میلیون‌ها دادهٔ کاربری که همسایه نامیده می‌شوند، برای تأمین توصیه باید در زمان واقعی (بلادرنگ) مورد محاسبه قرار گیرند که باعث می‌شود زمان پاسخگویی[2] بسیار طولانی شود. جستجو میان میلیون‌ها همسایه، یک فرآیند زمان‌بر بوده و این سیستم‌ها دیگر پاسخ‌گو نیستند. برای رفع این مشکل راه حلی ارائه شده است. (Sarwar, Karypis, Konstan, Riedl, 2002; Wang, Vries, Reinders, 2006)
- چالش دیگر، مسألهٔ پراکندگی[3] است. با توجه به اینکه این امکان وجود ندارد که همهٔ محصولات توسط همهٔ کاربران رتبه بندی و ارزش گذاری گردند، بنابراین ماتریسی که برای نشان دادن اولویت مشتریان برای کالا تشکیل می‌شود، برای خیلی از کالاهای موجود خالی بوده، لذا احتمال پیدا کردن مشتریان هم سلیقه و مناسب کم است. این مشکل در

[1] Scalability
[2] Response Time
[3] Sparsity

فروشگاه‌های با کالاهای زیاد رخ می‌دهد. همچنین در زمانی که سیستم در مراحل ابتدایی و در مرحله یادگیری است نیز این مشکل زیاد دیده می‌شود. (Si, Jin, 2003)

- چالش‌های دیگری را نیز می‌توان به این روش نسبت داد. یکی از این‌ها، مسألهٔ کامل نبودن اطلاعات در مورد اولویت‌های کاربر است. در این روش تنها معیار تعیین اولویت، خریدهای گذشتهٔ کاربر است و به الگوهای رفتاری او در فروشگاه از جمله سبد خرید او و همچنین محصولاتی که او کلیک و یا مطالعه نموده است، توجهی نمی‌نماید. در نتیجه تعداد اطلاعات در نظر گرفته شده از کاربران و مشتریان بسیار محدود بوده که در مقایسه با مشخصات مورد نیاز جهت پیش‌بینی خریدهای کالای بعدی آنها بسیار ناچیز است. لذا میزان صحت ویژگی‌های پیش‌بینی شده به طور چشمگیری کاهش می‌یابد. (Soo et al., 2005) یکی دیگر اینکه، مشتریان با سابقه علاوه بر خریدهای گذشتهٔ خود، اطلاعات دیگری از جمله میزان مراجعه به سایت و تناوب زمانی خریدها، وفاداری به فروشگاه و میزان پولی که خرج می‌کنند، را در اختیار سیستم توصیه‌گر قرار می‌دهند که در صورت استفادهٔ بهینه از این اطلاعات، می‌توان گروه‌بندی مناسب تری انجام داده و کیفیت توصیه را افزایش داد. (Liu, Shih, 2005) نقص دیگر این روش در عدم توجه به نوع، مشخصات و ویژگی‌های کالای مورد توصیه است. کاربران ممکن است در یک زمینه از محصولات و یا خدمات هم سلیقه باشند اما در سایر زمینه‌ها با یکدیگر متفاوت بوده و در نتیجه کیفیت توصیه کاهش می‌یابد. توجه به مشخصات کالاها و قواعد وابستگی[1] و یا ارتباطی بین کالاها مسألهای است که در توصیه و انتخاب محصول بسیار مؤثر بوده اما در این روش لحاظ نشده است. اگر کالاها بر اساس خصوصیات و ویژگی‌های آنها بطور مناسبی گروه‌بندی شوند، با استفاده از قواعد وابستگس می‌توان ارتباط بین هر گروه را با سایر گروه‌ها استخراج نموده و در نتیجه علاوه بر نظر کاربران هم سلیقه، از روابط بین کالاها هم جهت ارائهٔ توصیه استفاده نمود.

با توجه به مشکلاتی که در مورد رویکرد سیستم‌های مبتنی بر مشارکت جمعی گفته شد، در سال ۱۹۹۸ برای رفع کم بودن تعداد رتبه‌ها و ارزش‌گذاری ها، توجه به گذشتهٔ مشتری از جمله خریدهای

[1] Association Rules

او نیز مد نظر قرار گرفته شد و این روش دستخوش تغییراتی گردید. (Melville, Mooney, Nagarajan, 2002)

۲-۴-۴-۱ مراحل پیاده سازی مشارکت جمعی

❖ ارزیابی کاربران[1]:

اولین گام جهت ایجاد یک سیستم مبتنی بر مشارکت جمعی، جمع آوری داده ها به منظور پیش بینی ترجیحات فراگیر است. اطلاعاتی که سیستم از فراگیران بدست می آورد هم از لحاظ نوع و هم از لحاظ نحوه بدست آمدن متفاوتند. تجزیه و تحلیل مخازن داده و سیستم های جستجو در آموزش الکترونیک منجر به ایجاد مجموعه ای از روش های جمع آوری داده صریح و ضمنی خواهد گردید:

الف)داده های صریح: داده هایی که مستقیما بوسیله کاربر برای سیستم تعریف می شود مثل داده های مربوط به نام، رشته و مقطع تحصیلی، رتبه بندی (زمانیکه فراگیر به هر یک از آیتم های آموزشی با توجه به مفید و مناسب بودنشان رتبهٔ خاصی می‌دهد).

ب)داده های ضمنی: داده هایی که از تعامل کاربر با سایت بدست می آیند مثل سابقه بازدید فراگیر از آیتم های آموزشی، انتخاب (زمانیکه فراگیر در لیست نتیجه بر روی لینک خاصی کلیک می نماید و صفحه ای با توضیحات گسترده در مورد آن آیتم، نمایش داده می‌شود)، آزمون های انجام شدهٔ قبلی و دانلودهای فراگیر.

همان طور که توضیح داده شد به دو صورت صریح و ضمنی ماتریسی همانند شکل ۲-۳ از نرخ هایی که کاربران به شیء ها داده اند تشکیل می‌شود.

❖ تشکیل گروه:

هستهٔ اصلی مشارکت جمعی محسوب می‌شود. پس از تشکیل ماتریس مرحله قبل می‌توان به چند صورت گروه بندی را انجام داد. در ادامه به توضیح مشارکت جمعی مبتنی بر کاربر[2]، مبتنی بر شیء[3] و مشارکت جمعی مبتنی بر k نزدیکترین همسایه[4] (K-NN) پرداخته می‌شود.

[1] Rating
[2] User-based
[3] Item-based
[4] K-Nearest Neighbores

- مشارکت جمعی مبتنی بر کاربر: کاربران ستون ها و اشیاء سطرها را تشکیل می‌دهند و میزان شباهت اشیاء بر اساس کاربرانی که به آن اشیاء رأی داده اند مورد بررسی قرار می‌گیرد. در واقع در این روش، هر فرد به یک گروه بزرگ از کاربران هم سلیقه تعلق دارد و کالاهایی که توسط این گروه‌ها خریداری می‌شود، می‌تواند به بقیه نیز ارائه گردد. این روش همان مشکلات پراکندگی و پیچیدگی در تعداد زیاد کاربر را به همراه دارد. (Wang, Vries, Reinders, 2006) مشکل این روش در زمانیکه تعداد شیء ها زیاد می‌شود بروز می‌کند. برای مثال در سایتی مانند amazon.com که دارای اشیاء زیادی است احتمال اینکه تمام اشیاء حداقل توسط دو کاربر استفاده شده باشد بسیار کم و می توان گفت نا ممکن است. به این دلیل از روش مبتنی بر شیء استفاده می‌شود.
- مشارکت جمعی مبتنی بر شیء: در سال ۱۹۹۹ برای انعطاف پذیری بیشتر سیستم‌های توصیه گر مبتنی بر مشارکت جمعی با تعداد زیاد کاربر و ارائهٔ بهتر توصیه معرفی شد. (Sarwar et al., 2002; Idayes, Cimningham, 2004) در این روش یک ماتریس برای کاربران و کلیهٔ موارد تشکیل می‌شود و سپس با بررسی این ماتریس، ارتباط بین کالاها تعیین و با استفاده از این ارتباط، توصیهٔ کالا صورت می‌گیرد. روش مشارکت جمعی مبتنی بر شیء بر اساس این اصل که کاربر دوست دارد چیزی را بخرد که مشابه و یا در ارتباط با چیزی است که قبلاً خریده است، و دیگر نیازی به همسایه‌ها هم نیست، توصیه ارائه می‌دهد. روش مشارکت جمعی مبتنی بر شیء، با اینکه دو مشکل روش مبتنی بر کاربر را رفع می‌کند، اما به دلیل عدم توجه به کاربران، گاهی توصیهٔ خوبی ارائه نمی دهد و کیفیت توصیه‌اش پایین‌تر است. با این حال مشارکت جمعی مبتنی بر کاربر هم اگر کاربران علایق متفاوتی داشته باشند، توصیهٔ خوبی ارائه نمی‌دهد. (Li, Lu, Xuefeng, 2005)

در واقع در این روش، همان طور که در شکل۲-۳ نمایش داده شده است برخلاف روش قبل از روی میزان شباهت itemها به شباهت کاربران می‌رسند. به عبارت دیگر برای یافتن شباهت دو کاربر به بررسی اشیائی که هر دو، آنها را ارزیابی کرده‌اند پرداخته و از طریق فرمول‌های نظیر آنچه در ادامه ذکر شده میزان شباهت کاربران محاسبه می‌شود:

- شباهت بر اساس کسینوس[1]

[1] Cosine-based similarity

ابتدا شیء های i,j را به صورت بردار در نظر گرفته و سپس شباهت دو بردار می‌تواند بوسیلهٔ محاسبهٔ زاویهٔ کسینوس بین آنها بدست آید:

$$\text{Sim (i, j)} = \cos(\vec{\imath}, \vec{\jmath}) = \frac{\vec{\imath}.\vec{\jmath}}{\|\vec{\imath}\|_2 * \|\vec{\jmath}\|_2} \qquad (۳-۲)$$

در حقیقت، بردارهای $\vec{\imath}$ و $\vec{\jmath}$ بردارهای اولویت دو کاربر بوده که زاویهٔ بین دو بردار، سطح اولویت مشترک دو کاربر را مشخص می‌کند. جائیکه $\vec{\imath}.\vec{\jmath}$ مشخص می‌کند، نقطهٔ حاصلضرب بین دو بردار $\vec{\imath}$ و $\vec{\jmath}$ می‌باشد.

- شباهت بر اساس ضریب همبستگی[1]

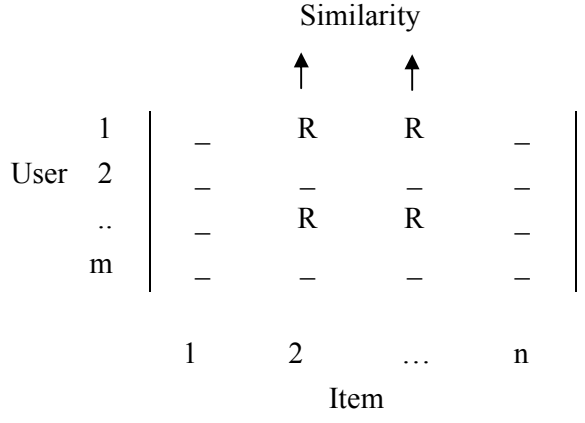
در روش تشابه براساس ضریب همبستگی، از ضریب همبستگی پیرسون برای محاسبهٔ مشابهت و همسانی استفاده می‌شود. (Adomavicius, Tuzhilin, 2005)

(۴-۲)

$$sim(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R}_i)^2}\sqrt{\sum_{u \in U}(R_{u,j} - \bar{R}_j)^2}}$$

دو نماد $\bar{R}_i$ و $\bar{R}_j$ بیانگر میانگین کاربرانی است که کالاهای i و j را خریداری کرده‌اند. همچنین متغیر U مجموعه کاربرانی است که اشیاء i,j را انتخاب کرده اند.

$R_{u,i}$ معادل ارزیابی‌هایی است که کاربران عضو u روی شیء i تعریف کرده اند. یعنی آیا کاربر شیء i را خریداری کرده است یا خیر؟



شکل ۲-۳ ماتریس رتبهٔ کاربران به اقلام

[1] Correlation-based similarity

- مشارکت جمعی مبتنی بر k نزدیکترین همسایه: ابتدا به بررسی رفتار خرید افراد در درون هر گروه خریداران هم سلیقه می‌پردازد و بر اساس الگوی رفتاری اعضای این گروه، کالاها را به افراد آن ارائه می‌کند. این فرآیند شامل دو مرحله است، که در مرحلهٔ اول ابتدا به شناسایی کاربران و پیشینهٔ آنها پرداخته و در مرحلهٔ دوم با استفاده از روش‌های تعیین وابستگی و سپس گروه‌بندی، کاربر را در گروه کاربران هم سلیقه‌اش قرار می‌دهد و در نهایت بر اساس الگوهای رفتاری آن گروه، به کاربر توصیه ارائه می‌دهد.

سیستم GroupLenz که اخبار Usenet را فیلتر می نماید، یک سیستم مشارکت جمعی بوده که از این روش برای توصیهٔ اخبار و فیلم‌هایش استفاده می کند. در این الگوریتم، پروفایل کاربر براساس زیر مجموعهٔ مناسبی از n کاربر مشابه با این کاربر ارزیابی می گردد. در نسخهٔ اولیهٔ GroupLenz بازخورد کاربران فقط به روش رتبه بندی صریح جمع آوری می گردید. اگرچه با مشاهدهٔ هزینه های اضافهٔ رتبه بندی صریح، در آخرین نسخهٔ این سیستم از زمان مطالعه نیز به عنوان شاخص ضمنی استفاده شده است.

❖ پیش بینی[1] اشیاء:

از مهمترین قسمت رویکرد مشارکت جمعی، انجام پیش بینی اشیاء محسوب می‌شود. پس از تشکیل همسایگی‌ها برای یک کاربر، می‌توان میزان علاقهٔ او را برای یک شیء مورد نظر پیش بینی کرد. این مقدار بر اساس نرخ‌های همسایگان کاربر اندازه گیری می‌شود و در نتیجه از پیشنهاد دادن شیء غیر مرتبط به کاربر پیشگیری می گردد.

❖ پیشنهاد اشیاء:

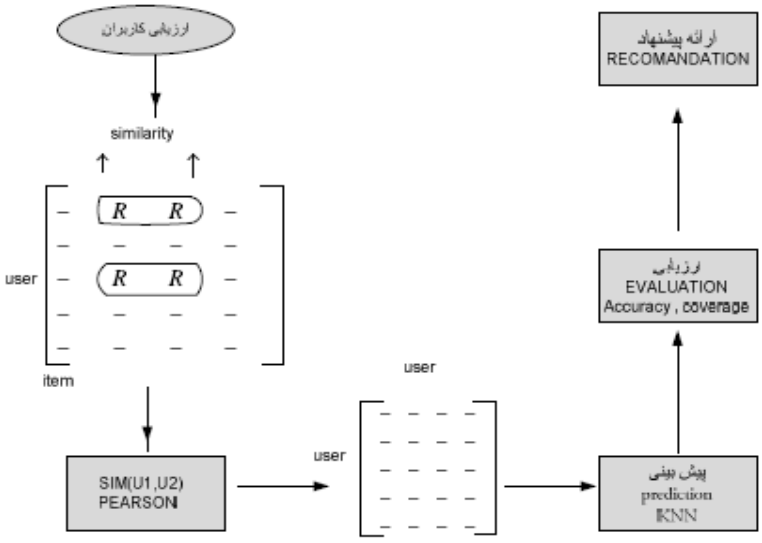
به منظور پیشنهاد شیء باید میزان سازگاری آن را با آنچه که کاربر انتخاب می‌کند محاسبه کرد. اگر این میزان از آستانه ای بالاتر بود، می توان از این روش برای پیشنهاد دادن استفاده نمود. معیارهای زیادی برای ارزیابی روش مشارکت جمعی وجود دارد اما می توان به دو عامل دقت و میزان پوشش[2] به عنوان مهمترین عامل‌ها نام برد. منظور از دقت معیاری است که نشان دهندهٔ میزان تطابق پیشگویی انجام شده با نظر کاربر و منظور از میزان پوشش

[1] Prediction
[2] Coverage

معیاری است که نشان می‌دهد چه میزانی از اشیاء مناسب برای پیشنهاد دادن از بین تمامی شیء های مناسب پیدا شده است. برای بررسی این دو معیار روش‌های مختلفی موجود است از جملهٔ این روش‌ها می‌توان به MAE[1] اشاره نمود. در این روش ابتدا مراحل فوق را روی یک مجموعه داده که می‌توان از سایت‌هایی نظیر Movielens تهیه نمود انجام داده و سپس با استفاده از فرمول زیر میزان خطا و یا به عبارت دیگر میزان عدم تطابق بین خواستهٔ کاربر و آنچه به او پیشنهاد شده محاسبه می‌شود:

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \qquad (۵-۲)$$

در حقیقت pi بیانگر پیش بینی که صورت گرفته و qi بیانگر خواستهٔ کاربر است و N نیز تعداد اشیاء را مشخص می‌کند. در انتها اگر MAE از یک آستانه ای پایین تر بود می‌توان از این روش برای پیشنهاد دادن استفاده نمود. مراحل فوق به اختصار در (شکل۲-۴) نشان داده شده است:



شکل۲-۴ مراحل پیاده سازی مشارکت جمعی

۲-۴-۴-۲ بررسی انواع مشارکت جمعی

علاوه بر تفکیک انواع سیستم‌های توصیه‌گر مبتنی بر مشارکت جمعی که در فوق (بخش تشکیل گروه) به آنها اشاره شد، این سیستم‌ها به خصوص نوع مبتنی بر نزدیکترین همسایه را از دیدگاهی دیگر نیز می‌توان به دو دستهٔ اصلی زیر تقسیم کرد:

[1] Mean Absolute Error

❖ بر اساس مدل[1]: این روش با ساختن یک مدل از امتیازات (رتبه‌های) کاربران به آنها پیشنهاد می‌دهد. در این الگوریتم دسته بندی بر اساس احتمال بوده و از الگوریتم های یادگیری ماشین[2] نظیر چند الگوریتم زیر استفاده می‌شود:

- شبکه‌های Bayesian[3]: مدل احتمال را به صورت فرمول در می آورد تا در مشارکت جمعی استفاده کند. این قابلیت را دارد تا به صورت آفلاین ساعت ها یا روزها بر روی یک مسأله کار کند اما برای محیط هایی که به بروز کردن سریع نیاز دارد مناسب نیست. در مدل شبکه Bayesian، هر آیتم در دامنه را به عنوان گره‌ای از ساختار Bayesian، جائیکه نرخ هر گره به آیتم‌های دیگر شبیه است، در نظر می‌گیرند. (Condliff, Lewis, 2000)
- کلاستر[4]: به صورت یک مسأله از نوع دسته‌بندی[5] کار می کند. به این صورت که احتمال قرار گرفتن کاربری در یک کلاس را حساب کرده و بر حسب عدد بدست آمده تصمیم می گیرد و کاربران مشابه را در یک کلاس قرار می‌دهد.
- براساس قانون[6]: از ارتباطاتی نظیر فروش یک شیء به چند کاربر قوانین را استخراج می‌کند.

هم در ساختار شبکه و هم در احتمالات شرطی از داده‌ها آموخته می‌شوند. یکی از محدودیت‌های این روش در این است که هر کاربر عضو یک گروه است در حالیکه ممکن است برخی از برنامه‌های توصیه‌گر، قابلیت عضویت یک کاربر در چندین گروه را داشته باشند. برای مثال در برنامهٔ پیشنهاد کتاب، یک کاربر ممکن است به خاطر کارش، به موضوعی مثل برنامه‌نویسی علاقه مند باشد، در حالیکه برای اوقات فراغتش کتابی در مورد آشپزی بخواند.

❖ بر اساس حافظه[7]:

در روش‌های مبتنی بر حافظه، بر روی کل بانک اطلاعاتی کاربران برای یافتن نزدیکترین همسایگی‌ها به کاربر هدف و تأثیر توصیه‌های آنان براساس میزان شباهت‌هایشان، عملیات

[1] Model-based
[2] Machine learning
[3] Bayesian networks
[4] Clustering
[5] Classification
[6] Rule based
[7] Memory-based

صورت می‌گیرد. این شباهت می‌تواند به صورت صریح با نظر سنجی و یا از طریق بررسی فعالیت هایی که کاربر انجام می دهد (مانند خرید مشابه) و فرمولی که تعریف شده، (به صورت ضمنی) استخراج شود. الگوریتم‌های مبتنی بر حافظه ذاتاً هیوریستیک است. (Breese, Heckerman, Kadie, 1998) و پیش بینی بر اساس مجموعه کل آیتم‌های نرخ داده شده بوسیلهٔ کاربر انجام می‌شود. لذا بنیادی ترین الگوریتم در رویکرد مبتنی بر حافظه، الگوریتم نزدیکترین همسایگی می‌باشد که همانطور که قبلاً اشاره گردید، یکی از پر کاربردترین الگوریتم‌های مشارکت جمعی می‌باشد.

تفاوت اصلی بین تکنیک‌های مشارکت جمعی مبتنی بر مدل و مبتنی بر حافظه یا هیوریستیک این است که تکنیک‌های مبتنی بر مدل، نرخ سودمندی را به شیوهٔ مبتنی بر هیوریستیک محاسبه نمی‌کنند بلکه آنها براساس یادگیری از داده‌ها، آمار و تکنیک‌های یادگیری ماشین هستند.

۲-۴-۵ سیستم های ترکیبی

دو سیستم مبتنی بر محتوا و فیلترینگ همکارگونه تضادی با هم ندارند و می‌توانند با هم در یک سیستم ترکیبی، ترکیب شوند و تعریف این سیستم شامل هر دو سیستم بالا می‌شود. این سیستم ترکیبی در ابتدا بر اساس مجموعه‌های تعریف شده در رویکرد مبتنی بر محتوا بنا نهاده می‌شود و سپس با رویکرد مبتنی بر فیلترینگ همکارگونه تکمیل می‌گردد. (Dastani, et al., 2005)

البته معدود سیستم‌های ترکیبی دیگری نیز وجود دارد که از ترکیب دو سیستم پایه به همراه سایر تکنیک‌ها به وجود می آیند. طراحان این نوع سیستم‌ها، دو یا چند گونه از انواع مذکور را غالباً به دو منظور با هم ترکیب می‌کنند:

۱. افزایش عملکرد سیستم؛
۲. کاهش اثر نقاط ضعف زمانی که آن مدل‌ها به تنهایی به کار گرفته شوند؛

نمونه ای از این سیستم ها Tapestry است که از هر دو رویکرد مبتنی بر محتوا و مشارکت جمعی استفاده می نماید. سیستم Fab نیز از مدل مشارکت جمعی در عین معرفی تحلیل محتوا با فیلترینگ موضوع (عنوان) استفاده می نماید. صفحات وب ابتدا توسط فیلتر موضوع رتبه بندی شده و سپس به فیلترهای شخصی کاربر فرستاده می‌شوند. رتبه های صریح و بازخورد هایی که از کاربران دریافت می گردند، منجر به تغییر فیلتر شخصی و همچنین فیلتر موضوع (عنوان) اصلی خواهند شد.

روش‌های مختلفی برای ترکیب دو یا چند رویکرد پیشنهاد شده است که در ادامه آنها را دسته بندی می‌کنیم:

- وزندار[1]: نتایج (نرخ یا امتیاز) چندین مدل توصیه‌گر با هم ترکیب می‌شوند تا یک توصیه ساده تولید شود.
- راهگزینی[2]: در این روش با توجه به شرایط جاری سیستم، یکی از مدل‌ها را انتخاب می‌کند.
- آمیخته[3]: پیشنهاد از چندین مدل توصیه‌گر متفاوت که در یک زمان نمایش داده شده‌اند، ایجاد می‌شود.
- ترکیب خصوصیات[4]: خصوصیات از منابع دادهٔ مدل‌های توصیه‌گر متفاوت با هم در یک الگوریتم ساده قرار می‌گیرند.
- آبشاری[5]: سیستم، توصیه‌های دیگر مدل‌ها را پالایش می‌کند.
- افزایش خصوصیات[6]: خروجی یک مدل، به عنوان خصوصیت ورودی مدل دیگر استفاده می‌شود.
- فرا سطح[7]: مدلی که یک سیستم یاد گرفته، به عنوان ورودی دیگری استفاده می‌شود.

در سال های اخیر، علاوه بر تجزیه و تحلیل مبتنی بر محتوا و مشارکت جمعی، داده کاوی نیز به روش های سیستم توصیه گر پیوست. (Lee, Kim, Rhee 2001) نظریهٔ کارشناس شخصی وب سایت را با استفاده از روش مشارکت جمعی و تکنیک قوانین انجمنی ارائه نمودند. به عنوان نمونه ای دیگر، Surflen سیستم توصیه گری است که از تکنیک های داده کاوی جهت کشف قواعد انجمنی در صفحات وب استفاده می نماید. و این عمل را فقط از طریق مشاهدهٔ پیشینهٔ کاربر (بدون دریافت بازخورد) انجام می دهد. البته تشخیص دقیق علایق کاربر فقط براساس مشاهدهٔ پیشینهٔ وی امری دشوار بوده زیرا ممکن است کاربر صفحهٔ وبی را از روی اشتباه باز نماید و به محتویات آن علاقه ای نداشته باشد. این مشکل در مواقعی که از سیستم کم استفاده می گردد، وخیم تر می شود. (Cho, Kim & Kim, 2002)یک سیستم توصیه گر را بر اساس وب کاوی کاربرد و الگوریتم های درخت

[1] Weighted
[2] Switching
[3] Mixed
[4] Feature combination
[5] Cascade
[6] Feature augmentation
[7] Meta-level

تصمیم گیری پیشنهاد نمودند. سیستم های توصیه گر به طور گسترده در بسیاری از فعالیت های اینترنتی استفاده شده است.

۲-۵ داده‌کاوی

در سال‌های اخیر، پیشرفت‌های بسیاری در سیستم‌های آموزشی، در راستای معرفی تکنولوژی‌های نوینی همچون آموزش تحت وب رخ داده است. به جرأت می توان گفت بار بسیاری از این پیشرفت‌ها بر دوش یکی از حوز ه های علوم کامپیوتر یعنی داده‌کاوی است.

نگاهی به ترجمه تحت اللفظی داده‌کاوی، به ما در درک بهتر این واژه کمک می‌کند. همانطور که می‌دانیم، Mine به معنای استخراج از منابع نهفته و با ارزش زمین اتلاق می‌باشد. پیوند این کلمه با کلمه داده، جستجویی عمیق جهت پیدا کردن اطلاعات اضافی مفید که قبلاً نهفته بودند، از داده های قابل دسترس حجیم، را پیشنهاد می‌کند. داده‌کاوی یک رشتهٔ نسبتاً جدید علمی می‌باشد که از انجام تحقیقات در رشته‌های آمار، یادگیری ماشینی، علوم کامپیوترخصوصاً مدیریت پایگاه داده‌ها شکل گرفته است. در زمینهٔ داده کاوی تعاریف متعددی وجود دارد که در زیر به برخی از آنها پرداخته شده است:

- داده کاوی به فرآیند جستجوی اطلاعات با ارزش، از میان حجم بزرگی از داده ها و اطلاعات یا به عبارت دیگر کاوش در انبار داده ها اطلاق می‌شود.
- داده کاوی، استخراج و تحلیل بوسیله ابزار اتوماتیک یا شبه اتوماتیک از میان تعداد بسیار زیادی از داد ه‌ها (انباره داده) به منظور استخراج الگوهای با معنی و قوانین می‌باشد.
- داده کاوی فرایند استخراج دانش از بانک اطلاعات به منظور شناسایی الگو های معتبر، مفید و قابل فهم در داده ها می باشد.

تعریفی که در اکثر مراجع به اشتراک ذکر شده عبارت است از «استخراج اطلاعات و دانش و کشف الگو های پنهان از یک پایگاه داده های بسیار بزرگ و پیچیده». عموماً روش‌های داده کاوی به دو دسته کلی توصیف کننده و پیش‌بینی کننده تقسیم می‌گردد.

۱. کاوش توصیف کننده: در این روش ویژگی‌ها و روابط موجود بین داده‌ها، در یک پایگاه داده مشخص می‌شود.

۲. کاوش پیش‌بینی کننده: در این روش به منظور پیشگویی موارد جدید، عمل استنتاج بر روی داده‌های موجود انجام می‌گردد و مدلی آموزش دیده جهت پیش بینی در آینده ایجاد می‌شود.

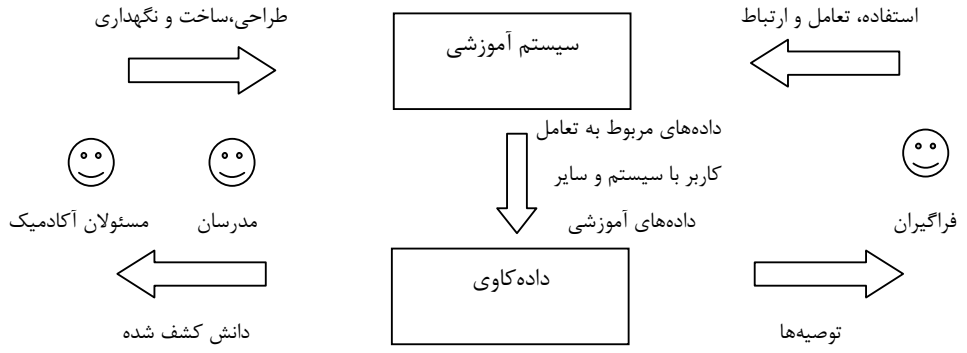
در این پژوهش از هر دو روش داده‌کاوی به منظورکشف روابط موجود بین داده‌ها، و دسته بندی و پیشگویی موارد جدید استفاده شده است.

۲-۵-۱ داده‌کاوی آموزشی

در سال‌های اخیر تحقیقات زیادی در زمینهٔ بکارگیری فرآیند داده‌کاوی در امر آموزش صورت می‌گیرد. این زمینهٔ تحقیقاتی جدید، داده‌کاوی آموزشی نامیده می‌شود که به امر توسعهٔ روش‌های کشف دانش از داده‌های محیط‌های آموزشی خصوصاً دانشجویان می‌پردازد. داده‌های جمع آوری شده در مورد دانشجویان می تواند شخصی یا آموزشی باشد که از طریق دفاتر و پایگاه داده‌های موجود در مدارس یا دانشکده‌ها جمع آوری می‌شوند. این نوع داده‌ها همچنین از طریق سیستم‌های آموزش الکترونیکی قابل دستیابی هستند. با بکارگیری تکنیک‌های داده‌کاوی روی داده‌های آموزشی می‌توان اطلاعات و دانش مفیدی را از آنها استخراج کرد که این دانش نیز به نوبه خود می تواند برای درک و فهم رفتار دانشجویان، کمک در امر آموزش و تدریس، ارزیابی و بهبود سیستم‌های آموزش الکترونیکی، بهبود برنامهٔ آموزشی، افزایش بازدهی و کارایی دانشجویان و اهداف دیگری بکار گرفته شود. داده‌کاوی آموزشی به چندین حوزهٔ تحقیقاتی از جمله آموزش الکترونیکی، سیستم‌های آموزشی هوشمند، وب کاوی و غیره مرتبط می‌شود. (یقینی، اکبری و شریفی، ۱۳۸۷).

کاربرد داده‌کاوی در سیستم های آموزشی توجه ویژه‌ای می‌طلبد که در حوزهای دیگر به این شدت نیست که از مهمترین آنها در نظر گرفتن ویژگی‌های آموزشی خاص فراگیر و سیستم است. در یک دههٔ گذشته مهمترین نوآوری‌ها در سیستم‌های آموزشی، در راستای معرفی تکنولوژی‌های نوینی همچون آموزش تحت وب بوده است. این نوع از آموزش به کمک رایانه و کاملاً بدون وابستگی زمانی و مکانی خاص است. از مزایای دیگر آن می‌توان به انعطاف پذیری و راحتی آموزش اشاره نمود. با وجود تمام مزایایی که آموزش الکترونیکی دارد، آمار مختلف نشان داده است که در مقایسه با فراگیران دوره‌های درسی سنتی و غیر الکترونیکی، درصد بالایی از فراگیران که دوره‌های درسی را به صورت برخط و الکترونیکی می‌گذرانند، تمایل به نا تمام رها کردن دورهٔ درسی دارند. در بسیاری از رشته‌های تحت وب مطالب بدون توجه به تفاوت یادگیری دانشجویان بصورت یکنواخت آموزش داده

می‌شود. سیستم آموزشی تحت وب هوشمند و سازگار در محیط‌هایی که نیاز به یادگیری شخصی غنی‌تری احساس می شود، راهگشا خواهد بود. این سیستم تلاش می‌کند به وسیلهٔ تهیه طرحی از اهداف شخصی، علایق و معلومات فراگیران، گونه ای از آموزش اختصاصی را به ایشان ارائه دهد. با تخمین احتمال موفقیت فراگیر در دورهٔ آموزشی پیش از اتمام آموزش و انجام آزمون نهایی، و سپس هدایت و مشاورهٔ فعالیت‌ها، تکالیف و دروس آموزشی مناسب به وی، می‌توان در بسیاری از موارد از عدم موفقیت فراگیران و نیمه تمام رها شدن دوره‌های آموزش الکترونیکی، پیش‌گیری به عمل آورد. در محیط‌های آموزشی معمولی، اساتید اطلاعات مورد نیاز خود را در زمینهٔ نحوهٔ یادگیری فراگیران، از طریق تبادلات رودرروی حضوری با آنها بدست می‌آورند به هر حال زمانی که فراگیر در محیطی الکترونیکی فعالیت می‌کند این گونه نظارت غیر قابل اجرا است و استاد باید به دنبال روش‌های دیگری برای دستیابی به اطلاعات مورد نیازش باشد. سازمان‌هایی که پایگاه‌های آموزش از راه دور را راه اندازی می‌کنند بطور خودکار حجم عظیمی از اطلاعات را توسط سرورهای وب تولید و در قسمت ثبت وقایع[1] سرورها جمع آوری می‌کنند. این محیط‌ها قادرند بیشتر رفتارهای یادگیری فراگیران را ثبت و حجم عظیمی از پروفایل ها را فراهم کنند. داده‌کاوی استخراج خودکار از الگوهای مفید از مجموعه داده‌های بزرگ است و با جستجو و یافتن اطلاعات آموزشی سودمند بر مبنای اسناد آموزشی، در ارزیابی و بهبود سیستم آموزش نیز کاربرد دارد. از مدل‌های کاوشی استخراج شده و الگوهای بدست آمده بر اساس داده های آموزشی، می توان در تعیین راه کارهای مناسب، جهت جلوگیری از افت تحصیلی فراگیران، بهره گرفت (یقینی، اکبری و شریفی، ۱۳۸۷). عملکرد داده‌کاوی در سیستم آموزش الکترونیک، یک سیکل تعاملی طرح فرضیه، آزمودن و اصلاح است (شکل۲-۵). دانش کاوش شده باید وارد چرخهٔ سیستم و هدایت شده، و بطور کلی باعث تسهیل و افزایش سطح یادگیری شود.



شکل ۲-۵ چرخهٔ استفاده از داده کاوی در سیستم آموزش الکترونیک

[1] Log Files

همانطور که در شکل ۲-۵ می‌بینیم آموزش دهندگان و مسئولان آکادمیک، مسئول طراحی، برنامه‌ریزی و حفظ سیستم آموزشی هستند که فراگیران از آن استفاده می‌کنند و با آن تعامل دارند. با استفاده از تمام داده‌های موجود دربارهٔ دوره‌های آموزشی، فراگیران و تعاملات، تکنیک‌های متفاوتی از کاوش اطلاعات به کار گرفته می‌شود تا اطلاعات مفیدی برای پیشرفت و بهبود فرآیند آموزش الکترونیکی بدست آید. دانش بدست آمده نه تنها می‌تواند به وسیلهٔ تعلیم‌گر جهت بالا بردن کیفیت آموزشی سیستم و برنامه‌ریزی بهتر برای آینده، استفاده شود بلکه می‌تواند در غالب سیستم آموزشی هوشمند و سازگار برای بهبود فرآیند یادگیری فراگیران، استفاده گردد.

اخیراً، علاقهٔ زیادی به آنالیز خودکار اطلاعات تعاملات فراگیر با محیط آموزش تحت وب بوجود آمده است. داده‌کاوی می تواند برای کاراتر ساختن محیط آموزشی، مؤثر باشد. در داده کاوی آموزشی ممکن است بخواهیم با در نظر گرفتن مقدار تلاش یک دانشجو، نمره نهایی او را پیش بینی کنیم. با استفاده از یکی از تکنیک‌های داده‌کاوی (دسته‌بندی) می‌توان به عنوان مثال فراگیرانی را که احتمال مردود شدن آنها در امتحان نهایی وجود دارد قبلاً تشخیص داد و روی آنها کار کرد و جلوی این امر را گرفت. (خادم القرانی، سرائی و مصطفوی، ۱۳۸۸)

۲-۵-۲ مراحل داده‌کاوی

داده‌کاوی شامل مراحل مختلفی می‌باشدکه عبارتند از:

۱. تعیین اطلاعات گذشته؛

۲. تمیز کردن داده‌ها و پردازش اولیه؛ در این مرحله خطاهای داده‌ها تصحیح می‌شوند و داده‌های اشتباه جایگزین می‌شوند. این مرحله ممکن است تا ۶۰ درصد از زمان داده‌کاوی را در برگیرد.

۳. یکپارچه‌سازی داده‌ها؛ معمولا داده‌ها از منابع متفاوتی جمع آوری می‌شوند باید به صورتی درآیند که یک مخزن داده‌های[1] مناسب ایجاد شود تا بتوان عملیات داده‌کاوی را بهتر انجام داد.

۴. انتخاب مجموعه داده‌های هدف؛

۵. یافتن ویژگیهای مورد استفاده و تعیین ویژگی‌های جدید؛

۶. نمایش داده ها به صورتیکه بتوان برای داده‌کاوی استفاده نمود؛

[1] Data Warehouse

۷. انتخاب عملیات داده‌کاوی (دسته‌بندی، خوشه‌بندی، پیش‌بینی وغیره)؛

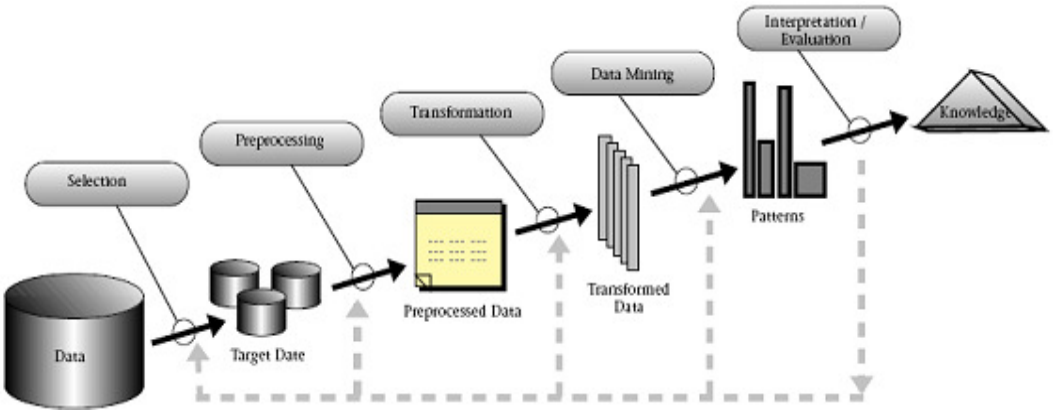
۸. انتخاب روش داده‌کاوی (شبکه‌های عصبی، درخت تصمیم و نظایر آن)؛

۹. داده‌کاوی و جستجو برای یافتن الگوی مناسب؛

۱۰. ارزیابی و تحلیل الگوی به دست آمده و حذف الگوهای نامناسب؛

۱۱. تفسیر نتایج داده‌ها و استنتاج از اطلاعات با ارزش؛

باید توجه داشت که جمع آوری و محافظت از داده ها نکته بسیار مهمی می‌باشد. اصولاً چون قالب و نوع داده‌ها در طول زمان تغییر می‌کند ممکن است بسیاری از داده‌های موجود در قالب‌های متفاوت باشند و همچنین بسیاری از داده‌های قدیمی از بین رفته و دور ریخته شوند. در حالیکه ممکن است اهمیت این داده‌ها از داده‌های جدید به هیچ وجه کمتر نباشد. همچنین به علت اینکه ممکن است داده‌ها از منابع مختلف داخلی و خارجی باشند، باز هم ممکن است قالب داده ها با هم یکسان نباشد. به همین دلیل انتخاب داده‌های درست و یکپارچه سازی قالب آن ها به منظور استفاده در داده‌کاوی از اهمیت بسیار بالایی برخوردار می‌باشد که در فصول آینده به شرح مراحل آماده‌سازی و پیش‌پردازش داده‌ها می‌پردازیم. در شکل۲-۶ می‌توان مراحل داده‌کاوی را به اختصار نشان داد.



شکل ۲-۶ مراحل داده‌کاوی

۲-۵-۳ تکنیک‌های داده‌کاوی

مجموعه عملیاتی را که روش داده‌کاوی قادر به انجام آن است در ذیل به صورت کامل تشریح شده اند:

۲-۵-۳-۱ دسته بندی و پیشگویی[1]

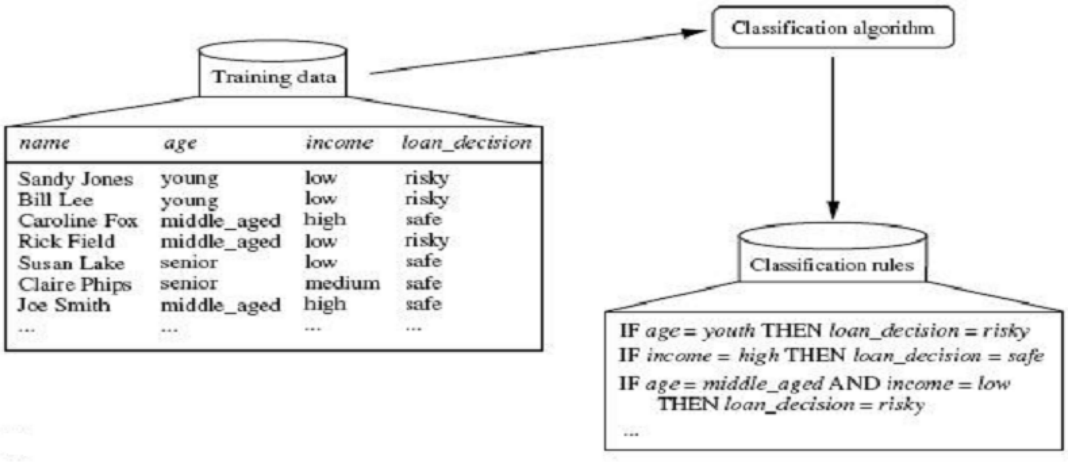
دسته بندی یکی از عملیات رایج و مورد استفاده در داده کاوی است که عضویت یک نمونه داده را در یکی از گروه‌های از قبل مشخص شده پیش بینی می‌کند. دسته بندی عملیاتی است که سازمان ها را قادر می سازد که در حل مسائل خاص در مجموعه های بزرگ و پیچیده به کشف الگوها دست یابند. دسته بندی فرآیندی می باشد که مجموعه داده ها را به قسمت‌های مشخص تقسیم می‌کند. برای مثال مشتریان یک شرکت بیمه بر اساس خصوصیاتشان به دو گروه با ریسک بالا و ریسک پایین تقسیم می‌شوند. با این کار در واقع مشتریان این شرکت دسته بندی شده اند. ساده ترین روشی که برای دسته بندی به نظر می رسد گذاشتن حدی برای دسته‌ها می‌باشد، مثلاً افراد با درآمد بالای مقداری مشخص را به یک دسته، و افراد با درآمد پایین تر از آن را به یک دستهٔ دیگر تخصیص دهیم. تعدادی از روش‌هایی که می‌توانند جهت داده‌کاوی مسائل دسته بندی به کار برده شوند، شامل: درخت تصمیم و شبکه های عصبی و نظیر این ها را ارائه کردند. این روش‌ها در دامنهٔ گسترده‌ای از زمینه‌های مهندسی به کار برده می‌شوند. (کیوان پور، خلعتبری، ۱۳۸۸)

دسته‌بندی داده‌ها یک فرآیند دو مرحله‌ای است:

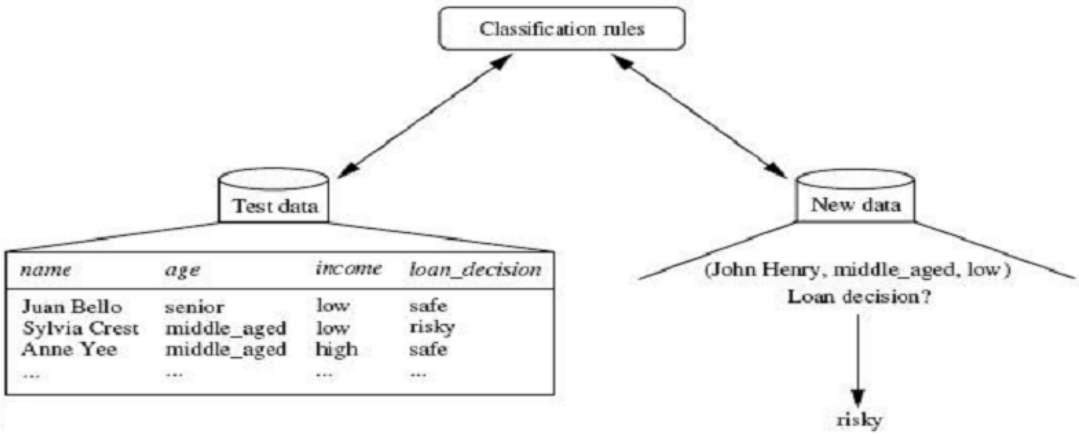
۱. در مرحله اول یک مدل ساخته می‌شود که مجموعه‌ای از کلاس‌های داده‌ای یا مفاهیم را مشخص می‌کند. این مرحله را مرحله یادگیری[2] گوییم که در آن یک الگوریتم کلاسه‌بندی یک مدل را با تحلیل یک مجموعه آموزشی[3] که مجموعه‌ای از تاپل‌های پایگاه است می سازد و برچسب کلاس‌های مربوط به این تاپل‌ها را مشخص می‌کند. چنین مدلی می‌تواند به فراهم کردن یک درک بهتر از داده های گمشده کمک کند. به طور معمول، این مدلها به فرم‌هایی از درخت تصمیم، یا فرمول‌های ریاضی نمایش داده می شود. یک تاپل X با یک بردار صفت $X=(x_1,x_2,\ldots,x_n)$ نمایش داده می‌شود. فرض می شود که هر تاپل به یک کلاس از پیش تعریف شده متعلق است و کلاس با یک صفت که به آن صفت برچسب کلاس می‌گوییم مشخص می‌شود. مجموعه آموزشی به صورت تصادفی از پایگاه انتخاب می‌شود. (شکل۲-۷ (الف)) سپس مدل می تواند به کمک قوانین اگر-آنگاه جهت پیش‌گویی برچسب‌های کلاس داده‌های جدید که دارای برچسب کلاس نامعلوم هستند، مورد استفاده قرار گیرد.

[1] Classification and Prediction
[1] Learning
[2] training set

۲. در مرحلهٔ دوم، یادگیری از طریق یک تابع $y=f(X)$ انجام می‌شود که می‌تواند برچسب کلاس هر تاپل X از پایگاه را پیش‌بینی کند. این تابع به صورت قواعد کلاسه‌بندی، درخت‌های تصمیم‌گیری یا فرمول‌های ریاضی است. (شکل۲-۷(ب)). در واقع قواعد یادگیری که از تحلیل داده های کاربران موجود حاصل شده است، می توانند برای پیش‌گویی کلاس اعتبار کاربران جدید مورد استفاده قرار گیرند.



شکل۲-۷(الف) پروسهٔ دسته‌بندی داده‌ها- مرحلهٔ یادگیری



شکل۲-۷(ب) پروسهٔ دسته‌بندی داده‌ها- مرحلهٔ کلاسه‌بندی

درخت‌های تصمیم از طریق جداسازی متوالی داده‌ها به گروه‌های مجزا ساخته می‌شوند و هدف در این فرآیند افزایش فاصله بین گروه‌ها در هر جداسازی است. یکی از تفاوت‌ها بین متدهای ساخت درخت تصمیم اینست که این فاصله چگونه اندازه‌گیری می‌شود. درخت‌های تصمیمی که برای پیش‌بینی متغیرهای دسته‌ای استفاده می‌شوند، درخت‌های دسته‌بندی نامیده می‌شوند زیرا نمونه‌ها

را در دسته‌ها یا رده‌ها قرار می‌دهند. درخت‌های تصمیمی که برای پیش‌بینی متغیرهای پیوسته استفاده می‌شوند درخت‌های رگرسیون[1] نامیده می‌شوند.

به عبارتی، از نقطه نظر کلی، دسته بندی و رگرسیون دو نوع اصلی از مسائل پیش‌گویی هستند، که دسته‌بندی جهت پیشگویی مقادیر گسسته و اسمی مورد استفاده قرار می گیرد، در حالی که رگرسیون جهت پیشگویی مقادیر پیوسته مورد استفاده قرار می گیرد. دسته بندی و پیشگویی کاربردهای زیادی در بازرگانی، بانکداری، پزشکی، ارتباطات، کشاورزی و غیره دارد. (مظهری، ایمانی، ۱۳۸۸)

روش‌های دسته‌بندی در داده‌کاوی عبارتند از:

۱. رگرسیون خطی چندگانه
۲. رگرسیون لجستیک
۳. تحلیل ممیزی
۴. بیز ساده
۵. شبکه‌های عصبی
۶. درخت‌های تصمیم
۷. K نزدیکترین همسایگی

که در فصل بعد مروری بر عملکرد این روش‌ها و الگوریتم‌ها خواهیم داشت.

۲-۵-۳-۲ خوشه‌بندی

تکنیک خوشه بندی به دنبال یافتن گروهی از آیتم‌های مشابه در میان حجم عظیمی از داده‌ها بر اساس ایدهٔ عمومی تابع فاصله[2] که میزان شباهت را بین گروه‌ها محاسبه می‌کنند، هستند. قابلیت خوشه بندی وظیفه تقسیم یک گروه ناهمجنس را در چندین زیر گروه بر عهده دارد. به این ترتیب یک خوشه حاوی یکسری داده‌های متشابه می‌باشد که همانند یک گروه واحد رفتار می‌کنند. این فرایند یک تفاوت اساسی با دسته بندی دارد. زیرا در این مدل هیچ گونه الگوی آموزشی نداریم. و دسته‌ها از پیش تعریف شده و معین نمی باشند. خوشه بندی به طور خودکار ویژگی‌های متمایزکنندهٔ زیر گروه ها را تعریف می کند و زیر گروه‌ها را سازماندهی می‌نماید. و به عنوان نوعی

[1] Regression
[1] Distance Function

قابلیت داده کاوی غیر مستقیم مطرح است. برای اکثر روش های داده کاوی مثل درخت تصمیم گیری و شبکه های عصبی، با یک مجموعه آموزشی شروع کرده و به کمک این مجموعه سعی می‌شود یک مدل برای بخش بندی داده‌ها ، ایجاد گردد. سپس از آن مدل برای پیش بینی داده های جدید استفاده شود. در روش خوشه بندی هیچ دسته ای از قبل وجود ندارد و در واقع متغیر ها به صورت مستقل و وابسته تقسیم نمی شوند. بلکه ما در اینجا به دنبال گروههایی از داده ها هستیم که به هم شباهت دارند و با کشف این شباهت ها می‌توان رفتارها را بهتر شناسایی کرد و بر مبنای آنها طوری عمل کرد که نتیجه بهتری حاصل شود.

در هر خوشه یک خود شباهتی[1] بین اقلام آن خوشه وجود دارد. در واقع یک عملیات غیر نظارتی[2] می‌باشد. این عملیات هنگامی استفاده می‌شود که ما به دنبال یافتن گروه هایی از داده‌های مشابه می باشیم بدون اینکه از قبل پیش بینی در مورد شباهت های موجود داشته باشیم. خوشه بندی معمولاً هنگامی استفاده می‌شود که به دنبال یافتن گروه هایی از افراد هستیم که قبلاً شناخته نشده اند.

در خوشه بندی توجه به نکات زیر ضروری می‌باشد:

۱. می‌بایست داده‌های داخل یک خوشه بیشترین تشابه و داده‌های خوشه‌های متمایز بیشترین تفاوت را با یکدیگر داشته باشند.
۲. کیفت خوشه‌بندی به معیارهای تشابه به کار گرفته شده در متدها و پیاده‌سازی آنها بستگی دارد.
۳. کیفیت خوشه‌بندی به تعریف و نمایش خوشه‌ها بستگی دارد.
۴. کیفیت خوشه‌بندی به واسطه توانائی استخراج الگوهای مخفی تعیین می‌گردد.

۲-۵-۳-۳ کاوش قواعد انجمنی

هدف از کاوش قوانین، یافتن ارتباطات بین اجزاء یک مجموعه می‌باشد. به این ترتیب، جستجو و یافتن وابستگی‌ها، همبستگی‌ها و ساختارهای علّی و معلولی موجود بین یک‌سری اجزاء و یا اشیای موجود در بانک‌های اطلاعات تراکنشی رابطه‌ای و سایر انبارهای اطلاعاتی در این مقوله جای می‌گیرد. (Han, Kamber, 2001).

قواعد انجمنی ارتباطات میان اقلام را بر مبنای الگوهای وقوع آن‌ها با یکدیگر در تراکنش ها (بدون در نظر گرفتن ترتیب آن‌ها) نشان می‌دهند. در مورد تراکنش‌های وب، قواعد انجمنی ارتباطات بین

[1] Self Similarity
[2] Unsupervised Modelling

مشاهده صفحه ها بر مبنای الگوهای گردشی کاربران را نشان می‌دهند. بیشتر رویکردهای کشف قواعد انجمنی بر مبنای الگوریتم Apriori می‌باشند که یک استراتژی تولید و آزمایش را بکار می‌برد. این الگوریتم، گروه های اقلام (مشاهده صفحات ظاهر شده در ثبت پیش‌پردازش شده) را که با یکدیگر بطور مکرر در تراکنش های زیادی ظاهر شده‌اند، پیدا می‌کند. چنین گروه هایی از اقلام به مجموعه اقلام مکرر معروفند.

فرآیند کاوش قوانین انجمنی، قوانینی را تولید می‌نماید که انجام یک جزء به انجام جزیی دیگر منوط است. شکل این قوانین به صورت زیر می‌باشد:

(ضریب اطمینان[1]، ضریب پشتیبانی[2]) تالی[3] ← مقدم[4]؛

به عبارت دیگر، قانون انجمنی $X \rightarrow Y$ به مفهوم زیر است:

«یک رکورد بانک اطلاعات که شرایط X را ارضاء کند آنگاه شرایط Y را نیز ارضاء می نماید».

هر قانون انجمنی به وسیلهٔ ضریب پشتیبان و اطمینانش به صورت زیر تعریف می‌شود:

ضریب پشتیبانی، احتمال اینکه شمول مقدم و تالی در یک تراکنش می‌باشد. این ضریب را با علامت S نمایش می‌دهیم. ضریب اطمینان، احتمال اینکه اگر تراکنشی، شرایط مقدم را ارضاء کند، آنگاه شرایط تالی را نیز ارضاء می‌نماید، می‌باشد. این ضریب را با علامت C نمایش می دهیم. در واقع ضریب اطمینان یک قانون میزان همبستگی[5] بین اجزاء (itemset) را اندازه گیری می‌کند در حالی که ضریب پشتیبان یک قانون، اهمیت همبستگی بین این اجزاء را اندازه‌گیری می‌کند.

کاوش قوانین انجمنی اساساً همهٔ قوانین انجمنی را که ضریب پشتیبان و اطمینانشان بالاتر از حداقل ضریب پشتیبان و حداقل ضریب اطمینان مشخص شده باشد، پیدا می‌کند.

تحلیل وابستگی ها یک حالت غیر نظارتی داده کاوی می‌باشد که به جستجو برای یافتن ارتباط در مجموعه داده‌ها می‌پردازد. یکی از کاربردی ترین حالات تحلیل وابستگی ها «تجزیه و تحلیل سبد بازار» می‌باشد که در آن هدف یافتن کالاهایی است که معمولاً به طور همزمان خریداری می‌شوند. این کار کمک می‌کند که خرده فروشان بهتر بتوانند کالاهای خود را سازماندهی کرده و چیدمان بهتری از محصولات خود داشته باشند.

---

[1] Confidence
[2] Support
[3] Consequent
[4] Antecedent
[5] Correlation

کشف قواعد انجمنی در داده‌های تراکنشی وب نیز مزایای زیادی دارد. به عنوان مثال یک قانون با اطمینان بالا مانند *{special-offers/,/products/software/}=>{shopping-cart/}* ممکن است حاکی از این باشد که عملیات بهبود در محصولات نرم‌افزاری بطور مثبتی بر روی فروش برخط تاثیر می‌گذارد. چنین قواعدی همچنین می‌توانند برای بهینه سازی ساختار سایت مورد استفاده قرار گیرند. به عنوان مثال، اگر در یک سایت لینک مستقیم بین دو صفحه مثل A و B وجود نداشته باشد، کشف قاعدهٔ *{A}=>{B}* نشان می‌دهد که قرار دادن یک لینک مستقیم بین این دو صفحه ممکن است به کاربران در یافتن اطلاعات مورد علاقهٔ خود کمک کند.

باید توجه داشت که دسته‌بندی متفاوت از قوانین وابستگی می باشد. قوانین وابستگی شرایط موجود را نشان می دهند در حالیکه دسته بندی شرایط و وضعیت آینده را پیش بینی می کند.

۲-۵-۳-۴ کشف الگوهای ترتیبی[1]

الگوهای ترتیبی در داده های کاربرد وب، صفحاتی را که توسط کاربر دیده شده اند به همان ترتیب دیده شدن آن‌ها در نظر می‌گیرند. الگوهای ترتیبی یا SP، آن دسته از دنباله هایی از اقلام هستند که در بخش نسبتاً زیادی از تراکنش ها بطور مکرر ظاهر شده‌اند. می‌گوییم یک دنبالهٔ $< s_1, s_2, ..., s_n >$ در یک تراکنش $t =< p_1, p_2, ..., p_m >$ (که $n \le m$) رخ داده است اگر n عدد مثبت $1 \le a_1 \le a_2 \le ... \le a_n \le m$، و به ازای هر i داشته باشیم $s_i = p_{a_i}$.

الگوریتم Apriori که در کاوش قواعد انجمنی استفاده می‌شود را می‌توان به منظور کاوش الگوهای مکرر تغییر داد. این کار معمولا با تغییر تعریف پشتیبانی انجام می‌شود که می‌توان آن را بر مبنای تعداد تکرار وقوع زیردنباله های اقلام بجای زیرمجموعه ها تعریف کرد.

۲-۶ وب‌کاوی[2]

وب‌کاوی، استخراج اطلاعات قابل توجه و الگوهای قوی استفاده شده از محتویات صفحات وب، اطلاعات قابل دسترس وب، اتصالات[3] صفحات وب، و منابع تجارت الکترونیکی با استفاده از تکنیک‌های داده‌کاوی است، که نتیجه کاربردی این تکنیک‌ها، کمک به کاربران در استخراج هر چه بهتر اطلاعات، بهبود بخشیدن طراحی و شخصی سازی وب سایت است.

[1] Sequential patterns
[2] Web mining
[3] Web link

محدودهٔ جستجوی وب‌کاوی بسیار گسترده است. مفاهیم زیادی در وب‌کاوی وجود دارد که به برخی از آنها اشاره می‌کنیم:

- لاگ فایل[1]

لاگ فایل‌ها به اصطلاح گزارش هایی است که در سرور ذخیره می شود و می تواند یکی از بهترین منابع اطلاعاتی برای ردگیری کاربران وبگاه در بین صفحات وب باشد و می تواند رفتار بینندگان را در بین صفحات ذخیره کند.

- صفحات وب[2]

بیشتر متدهای وب‌کاوی موجود، بر روی صفحات وب، که بر پایه استانداردهای Xml وHtml می باشند کار می‌کنند.

- ساختار ابرلینک[3] وب

همه صفحات وب بوسیله ابرلینک‌ها با یکدیگر ارتباط برقرار می کنند. ابرلینک یک جزء ساختاری است که محتوای یک صفحه را به صفحهٔ دیگری متصل می کند.

وب‌کاوی به استفاده از تکنیک‌های داده‌کاوی به منظور کشف و استخراج خودکار اطلاعات از سرویس‌ها و اسناد وب گفته می شود. اخیراً استفاده از تکنیک‌های کاوش وب برای کشف الگوهای مفید از رفتار کاربران، رو به رشد است و کاوش وب تأثیر زیادی روی پیشرفت شخصی سازی وب داشته است. از این طریق، قادر هستیم که فهم بهتری هم از وب و هم از تمایلات کاربران وب بدست آوریم.

شخصی سازی مبتنی بر وب از الگوهای رفتاری کاربر که از کاوش وب استخراج شده است، به منظور تشخیص نیازها و علایق هر کاربر و نتیجتاً ارائهٔ یک ساختار و محتوای شخصی شده بر اساس نیازهای کاربر، استفاده می کند. چندین سیستم شخصی سازی وب بر اساس کاوش وب به وجود آمده اند که همگی شامل دو مرحلهٔ اصلی هستند: در مرحلهٔ اول که به صورت آفلاین انجام می شود، داده‌های آموزشی گرفته شده از رفتار کاربران روی وب به منظور کشف الگوهای دستیابی و استخراج مدل کاربران، کاوش می شوند. در مرحلهٔ دوم که به صورت آنلاین (برخط) انجام می شود، از مدل استخراج شده در مرحلهٔ اول برای تفسیر و مقایسه با الگوی پیمایشی کاربر فعال وب استفاده می‌شود و توصیه هایی نیز بر اساس این مقایسه ارائه می‌شود.

[1] Web Log
[2] Web page
[3] Hyperlink

فهم دقیق تمایلات کاربران وب و نحوهٔ پیمایش آنان در سایت، یک قدم اساسی در مطالعهٔ کیفیت و کارایی یک وب سایت است. به عنوان مثال، کشف الگوهای دستیابی مفید به ارائه کنندگان سرویس تجارت الکترونیکی این اختیار و توانایی را می دهد که کیفیت ساختار سایت را با جابجایی و تغییر ابرلینک ها یا با تغییر پویای یک صفحهٔ وب، به منظور راهنمایی کاربر برای بدست آوردن اطلاعات مورد نظرش، بالا ببرند. در زمینهٔ آموزش الکترونیکی نیز با ارائهٔ یک شخصی سازی مناسب، می توان برای یادگیرندگان مختلف روش های متفاوت آموزشی را به کار برد.

۲-۶-۱ کاربرد وب کاوی در سیستم های آموزش الکترونیکی

سیستم های توصیه گر را می توان تکنولوژی شخصی سازی برای فیلتر کردن اطلاعات دانست. یکی از روش های مورد استفادهٔ سیستم های توصیه گر، وب کاوی است. هر چند ابتدا در سیستم‌های توصیه گر یکی از شاخه های وب کاوی یعنی وب کاوی کاربرد بیشتر مورد توجه بود. اخیراً ترکیبی از روش های وب کاوی از جمله دانش کشف شده از محتوا و ساختار صفحات وب استفاده شده، عملکرد بهتری را نشان داده است. یکی از زمینه‌های مهم در یادگیری الکترونیکی توجه به توانمندی‌ها و سلایق فراگیران در فرآیند یادگیری است. برای این منظور لازم است رفتار فراگیران را در تعامل با محتوای الکترونیکی ردگیری نماییم. فرارسانه های تطبیق پذیر[1] به ما این امکان را می‌دهند که با ردگیری تعامل فراگیران در ارتباط با آنها بتوانیم خدمات آموزشی را، با یک فراگیر خاص مطابقت دهیم. برای تطبیق محتوا بر اساس رفتار فراگیر تاکنون مدل های متنوع و الگوریتم‌های به هنگام سازی مختلفی مطرح شده است که اغلب بر ارائهٔ اقدامات بعدی محتوای آموزشی برای کاربر خاص متناسب با رفتار وی تمرکز دارند. اغلب الگوریتم‌های موجود بیشتر به ثبت مسیریابی و دسترسی کاربران به منابع مورد نظرشان می‌پردازند. (Ha, Bae & Park, 2000)

۲-۶-۲ تکنیک‌های وب‌کاوی

بسته به اینکه چه بخشی از دادهٔ وب مورد کاوش قرار می‌گیرد، وب‌کاوی به سه دستهٔ کاوش محتوای وب، کاوش ساختار وب و کاوش کاربرد وب تقسیم بندی می‌شود. (خنشا، صدرالدینی، ۱۳۸۷)

[1] Hyper Media

۲-۶-۲-۱ کاوش محتوای وب

کاوش محتوای وب به پروسهٔ کشف دانش و استخراج اطلاعات کاربردی از محتوای صفحات وب گفته می شود. انواع محتوا شامل عکس، صدا، فیلم و متن می تواند مورد کاوش قرار گیرد. محتوای وب، داده‌های «ساخت‌یافته[۱]» مانند جدول، «نیمه ساخت‌یافته[۲]» مانند برچسب‌های HTML و «غیرساخت‌یافته[۳]» مانند متن معمولی را شامل می‌شود. از آنجا که محتوای متنی بیشتر مورد کاوش قرار می گیرد، کاوش محتوا را کاوش متن[۴] نیز می نامند.

تکنولوژی هایی که عموماً در کاوش محتوای وب به کار گرفته می شوند، شامل پردازش زبان طبیعی[۵] و بازیابی اطلاعات [۶] می باشد. کاوش محتوای وب کاربرد بسیاری دارد، به عنوان مثال دسته بندی صفحات وب براساس محتوا یکی از کاربردهای تکنیک‌های کاوش محتوا است. در وب‌کاوی محتوا تکنیک‌های داده‌کاوی معمولی مثل دسته‌بندی، خوشه‌بندی و کاوش قوانین انجمنی را نیز می‌توان به کار گرفت.

۲-۶-۲-۲ کاوش کاربرد وب

کاوش کاربرد وب به پروسهٔ کشف الگوهای جالب و مفید از داده های کاربرد کاربران وب، گفته می‌شود. داده های کاربرد، داده هایی است که در زمان پیمایش فراگیران در وب سایت، توسط کارگزاران در فایل های ثبت وقایع[۷] ذخیره می شوند. با کمک این روش می‌توان رفتار کاربران مختلف را پیش‌بینی کرده و بر اساس آن، صفحات وب را برای آن کاربر خاص، شخصی‌سازی نمود.

کاوش کاربرد وب عمومی ترین و پر کاربرد ترین روش برای استخراج الگوهای رفتاری کاربران از این فایل هاست. کشف الگوهای خطی، کاوش قوانین وابستگی و خوشه بندی، الگوهای پیمایشی را از فایل های ثبت وقایع استخراج می کنند که می توانند برای الگوسازی رفتار کاربران و ارائهٔ سرویس های شخصی شده به کار گرفته شوند.

در حقیقت، با استفاده از وب‌کاوی کاربرد می‌توان به شخصی‌سازی[۸] صفحات وب پرداخت، سایت‌ها را اصلاح نمود[۱] و یا کارائی سایت‌ها[۲] افزایش داد.

---

[1] Structured
[2] Semi Structured
[3] Unstructured
[4] Text Mining
[1] Nutural Language Processing
[2] Information Retrieval
[3] Log Files
[8] Personalization

۲-۶-۲-۳ کاوش ساختار وب

کاوش ساختار وب به پروسهٔ آنالیز گره ها و اتصالات ساختار یک وب سایت با استفاده از تئوری گراف گفته می شود. بسته به اینکه چه نوع داده ساختاری وب مورد کاوش قرار می گیرد، کاوش ساختار وب به دو دستهٔ Inter page و Intra page تقسیم می شود: در روش Inter page ساختار صفحات وب در ارتباط با یکدیگر مورد توجه است. به این روش، تحلیل ابرلینک‌ها[3] نیز گفته می‌شود. ابرلینک یک جزء ساختاری است که محتوای یک صفحه را به صفحهٔ دیگری متصل می کند. در این روش معمولاً صفحات وب و لینک‌های میان آنها به شکل گراف جهت‌دار نمایش داده می‌شود. هر گره از این گراف، نشان دهنده یک صفحهٔ وب بوده و هر یال جهت‌دار از گره «الف» به گره «ب» وجود لینکی از صفحه «الف» به صفحه «ب» را نمایش می‌دهد.

در مورد روش Intra page، ساختار درونی یک صفحهٔ وب، در نظر گرفته می‌شود. در حقیقت، کاوش ساختار یک سند وب است که از یک ساختار درختی برای توصیف و آنالیز تگ های HTML یا XML درون یک صفحه استفاده می کند.

هر چند ابتدا در سیستم های توصیه گر، وب کاوی کاربرد بیشتر مورد توجه بود. اخیراً ترکیبی از روش های وب کاوی از جمله دانش کشف شده از محتوا و ساختار صفحات وب استفاده شده، عملکرد بهتری را نشان داده است.

۲-۶-۳ دلایل نیاز به استفاده از محتوا

رویکرد تنها مبتنی بر کاربرد در شخصی‌سازی وب یک عیب مهم دارد و آن این است که فرآیند توصیه به کاربر تنها براساس داده‌های تراکنشی موجود او صورت می‌گیرد و از این رو اقلام یا صفحاتی که اخیراً به سایت اضافه شده اند نمی‌توانند به او توصیه شوند. این مشکل عموماً *مشکل قلم جدید* نامیده می‌شود. از سوی دیگر اگرچه الگوهای کشف شدهٔ مربوط به کاربرد منابع وب از طریق وب کاوی کاربرد وب در کشف ارتباطات اقلام با یکدیگر یا کاربران با یکدیگر و نیز تعیین شباهت در جلسات کاربر مفیدند، اما بدون استفاده از دانش عمیق تری از دامنه‌ی وب سایت مورد نظر، چنین الگوهایی درک اندکی از دلایل آن که چرا اقلام یا کاربران در گروه هایی با هم قرار می‌گیرند در

[1] Site Modification
[2] Performance Improvement
[3] Hyperlink Analysis

اختیار ما قرار می‌دهند. یک رویکرد معمول برای حل این مشکل در فیلتر کردن جمعی آن است که مشخصات محتوای صفحات را با رتبه‌بندی ها و قضاوت های کاربر ادغام کنیم. بطور کلی در این رویکردها کلمات کلیدی از محتوای وب‌سایت استخراج می‌شوند و برای اندیس‌گذاری صفحات براساس محتوا یا طبقه‌بندی آن‌ها به دسته‌های مختلف مورد استفاده قرار می‌گیرند. در حوزهٔ شخصی‌سازی وب این رویکرد به سیستم اجازه می‌دهد تا صفحات را نه تنها براساس افراد مشابه بلکه براساس شباهت محتوایی آن‌ها به صفحاتی که کاربر اخیرا بازدید کرده است به او توصیه کند.

۲-۶-۴ دلایل نیاز به استفاده از معنا

رویکردهای مبتنی بر کلمات کلیدی، در درک ارتباطات پیچیده تر بین ویژگی‌های اشیا در عمق معنایی بیشتر ناتوان هستند. به عنوان مثال اطلاعات ارزشمند بین دانشجویان، دروس و اساتید در صورت استفاده از صرفا کلمات کلیدی برای توصیف این موجودیت ها از دست می‌روند. به منظور توصیهٔ انواع مختلفی از اشیای پیچیده با استفاده از خصوصیات و ویژگی های آن‌ها سیستم باید قادر باشد خصوصیات آن‌ها را در عمق معنایی بالاتری نسبت به کلمات کلیدی در نظر بگیرد. به طور نمونه، سیستم شخصی‌سازی سنتی وب سایت یک دانشگاه ممکن است درس جاوا را به علت این که دانشجویی قبلاً به این درس علاقه نشان داده است به او توصیه کند. از سوی دیگر سیستمی که از *دانش دامنه‌ی مربوطه* بهره می‌برد ممکن است تشخیص دهد که این دانشجو ابتدا باید دروس پیش‌زمینهٔ درس جاوا را بگذراند یا ممکن است قادر باشد مناسب ترین استاد ارائه کنندهٔ این درس برای این دانشجو را به او توصیه کند. استفاده از دانش معنایی در حوزه شخصی‌سازی وب به تعامل عمیق‌تر مشتریان و کاربران وب سایت با آن منجر می‌شود. ادغام دانش دامنه در این سیستم ها این امکان را می‌دهد که توصیه های مفید بیشتری برای کاربران بر اساس مشخصات عمیق تر اشیاء تولید شوند و امکان استنتاج در مورد دلایل اقدامات کاربران را فراهم می‌کند.

۲-۶-۵ کاربردهای وب‌کاوی

۱. نگهداری تاریخچه کاربر: سیستم صفحاتی را که کاربر قبلا مورد دسترسی قرار داده را می داند و هنگام ورود بعدی کاربر به سیستم لیستی از آنها را به کاربر ارائه می دهد.
۲. ارائه دادن لینک: سیستم لینک‌های مورد علاقه کاربر را تشخیص داده و به کاربر نمایش می‌دهد. این لینک ها می تواند در یک قاب جداگانه یا در همان صفحه به کاربر نمایش داده شود. این روش برای جلب رضایت مشتری می‌باشد.
۳. شخصی سازی اطلاعات: در این روش بعد از درک علاقمندی‌های کاربر، محتوای متفاوتی برای کاربران متفاوت نمایش می دهد. این تغییرات شامل رنگ، نحوه چینش کنترل هایی که در صفحه وجود دارد می باشد. این روش بیشتر بوسیله درگاههای وب[1] مثل Yahoo و Altavista انجام می گیرد که می خواهند آنچه کاربر می خواهد به آنها ارائه دهد.
۴. فرستادن پیغام: سیستم می تواند به کاربران خاصی پیغام های مناسب را بفرستد و آنها را از تغییراتی که در سایت ایجاد شده و مرتبط با علاقمندی آنهاست مطلع نماید.
۵. شخصی سازی نتایج جستجو: در بسیاری از سایت ها امکاناتی برای جستجوی اطلاعات برای کاربران فراهم شده است و نشان دادن اطلاعات مرتبط با نیاز کاربر از چالش های مرتبط با این موضوع است. با داشتنن علاقمندی کاربر میتوان نتایج جستجو را با تجه به آن تغییر داد و نتیجه کار را بهبود بخشید.

۲-۶-۶ شخصی سازی براساس وب‌کاوی کاربرد وب

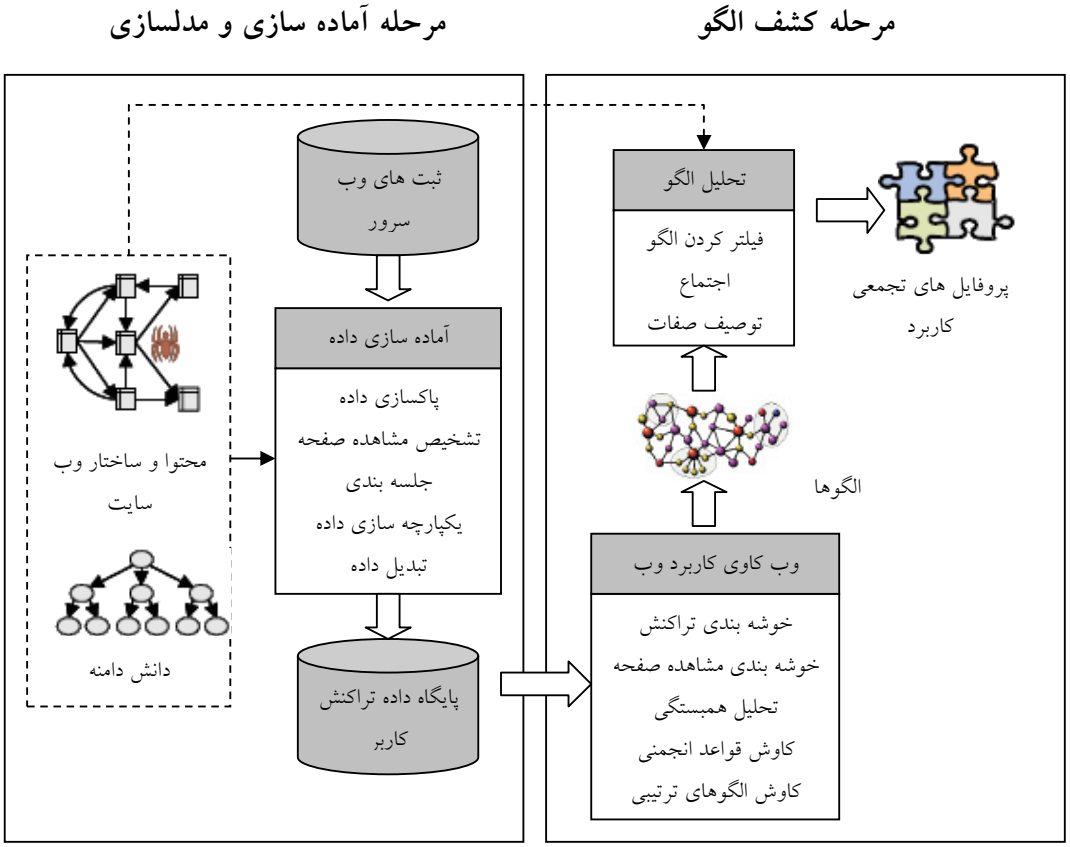
هدف شخصی‌سازی وب براساس وب‌کاوی کاربرد وب توصیهٔ مجموعه ای از فعالیت‌ها به فراگیر جاری شامل لینک، مقاله، متن، آزمون، تکلیف و غیره با جهت گیری به سمت ترجیحات و علایق وی می‌باشد. این عمل با تطابق جلسه جاری فراگیر (و احتمالا همراه با پروفایل ذخیره شده او) با الگوهای کاربردی کشف شده از طریق وب‌کاوی کاربرد وب صورت می‌گیرد. به این الگوهای کاربردی پروفایل های تجمعی کاربرد گفته می‌شود، چون یک نمایش تجمعی از فعالیت ها و علایق مشترک گروهی از فراگیران فراهم می‌کند. این فرآیند توسط موتور توصیه انجام می‌شود که مولفهٔ برخط سیستم شخصی‌سازی است. (قادریان، ۱۳۸۷)

[1] Portal

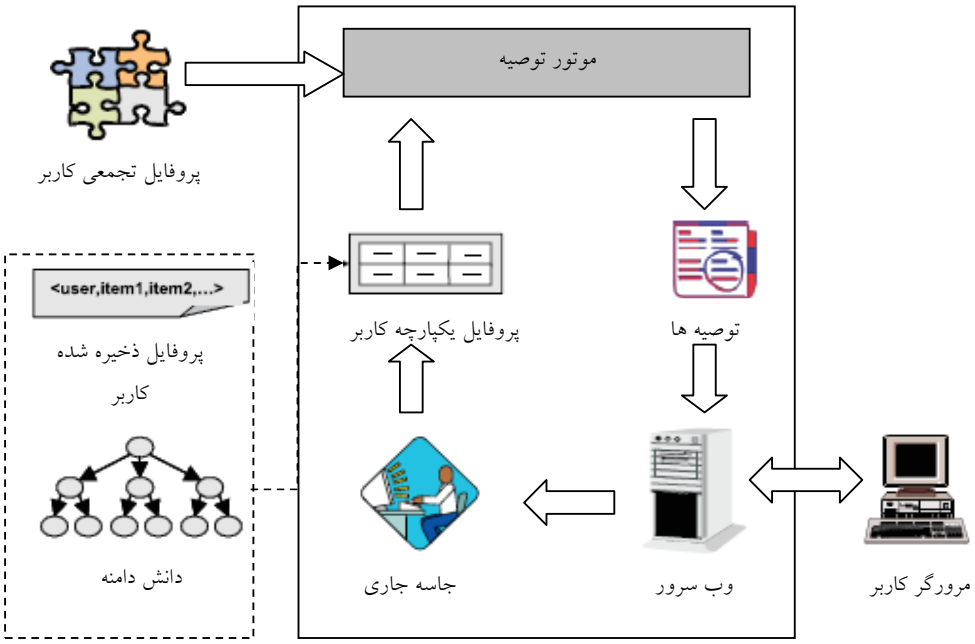
فرآیند کلی شخصی‌سازی براساس وب‌کاوی کاربرد وب شامل سه مرحله است: آماده سازی و مدل‌سازی داده، کشف الگو از داده‌های کاربرد وب و استفاده از الگوهای کشف شده برای شخصی‌سازی وب.

از بین این مراحل تنها مرحلهٔ سوم بصورت بلادرنگ انجام می‌شود. مرحله آماده سازی داده، ثبت های خام وب را به دادهٔ تراکنشی تبدیل می‌کند که می‌تواند در داده کاوی مورد استفاده قرار گیرد. این مرحله همچنین شامل یکپارچه سازی داده از منابع مختلف مانند پایگاه های داده، سرورهای خدمات کاربردی و محتوای سایت می‌باشد. در مرحله کشف الگو وظیفه های مختلف داده‌کاوی مانند خوشه بندی، کاوش قواعد انجمنی و کشف الگوهای ترتیبی را می‌توان بر روی این دادهٔ تراکنشی اجرا کرد. نتایج فاز کاوش به پروفایل های تجمعی کاربرد تبدیل می‌شوند که برای مرحله توصیه مناسب می‌باشد. موتور توصیه جلسهٔ جاری کاربر را همراه با الگوهای کشف شده به منظور فراهم نمودن محتوای شخصی‌سازی شده مورد استفاده قرار می‌دهد.

در این بخش یک نمای جامع از فرآیند شخصی‌سازی براساس وب کاوی کاربرد وب ارائه می‌دهیم. یک چارچوب کلی برای این فرآیند در شکل های ۲-۸ و ۲-۹ نشان داده شده است.



شکل ۲-۸ مولفه های برون خطی آماده سازی داده و کشف الگو



شکل ۲-۹ مولفه برخط شخصی‌سازی وب

۲-۶-۶-۱ آماده‌سازی و مدل‌سازی داده

یک مرحلهٔ مهم در هر عمل داده‌کاوی ایجاد مجموعه دادهٔ مناسبی است که بر روی آن الگوریتم های داده‌کاوی عمل کنند. این فرایند می‌تواند شامل پیش‌پردازش داده‌های اولیه، یکپارچه‌سازی داده‌ها از منابع گوناگون و تبدیل داده‌های یکپارچه به فرمتی مناسب به عنوان ورودی عملیات داده‌کاوی باشد. به مجموعه‌ی این عملیات آماده‌سازی داده گفته می‌شود.

فرایند آماده‌سازی داده اغلب پرهزینه‌ترین مرحله از نظر زمان و محاسبات در فرآیند کشف دانش محسوب می‌شود و در مورد وب‌کاوی کاربرد وب نیز این موضوع صادق است. مرحلهٔ آماده‌سازی داده در وب‌کاوی کاربرد وب در استخراج موفقیت‌آمیز الگوهای مفید از داده‌ها بسیار حیاتی است.

۲-۶-۶-۱-۱ منابع و انواع داده

منابع عمدهٔ داده در وب‌کاوی کاربرد وب فایل‌های ثبت وب‌سرور می‌باشد که شامل ثبت‌های دسترسی وب‌سرور و ثبت‌های سرویس‌دهندهٔ خدمات وب است. سایر منابع داده که آنها نیز هم در آماده‌سازی داده و هم در مرحله‌ی کشف الگو مهم می‌باشند شامل فایل‌های وب‌سایت و متاداده‌ها، پایگاه‌های داده، قالب‌های کاربردی و دانش دامنه است. در کل داده‌های بدست آمده از این منابع را می‌توان به چهار دسته تقسیم کرد که در ادامه هریک را بصورت مختصر شرح می‌دهیم.

۲-۶-۶-۱-۱-۱ داده‌های کاربرد

پیش از آنکه به انواع این‌گونه داده‌ها بپردازیم ابتدا تعاریفی از متاداده‌هایی که توسط وب‌سرورها تولید و استفاده می‌شوند ارائه می‌کنیم. شکل۲-۱۰ یک تراکنش HTTP را بین یک مشتری HTTP و یک سرور HTTP نشان می‌دهد. برای سادگی فرض کنید که مشتری HTTP یک مشتری وب است و یک سرور HTTP نیز یک وب‌سرور می‌باشد. یک مشتری وب که برای کاربران انسانی طراحی شده است یک *مرورگر وب* نامیده می‌شود مانند Mozilla Firefox، Netscape Navigator و Microsoft Internet Explorer. مثال‌هایی از وب‌سرور عبارتند از IBM HTTP Server، Apache HTTP Server و Microsoft Internet Information Server (IIS).

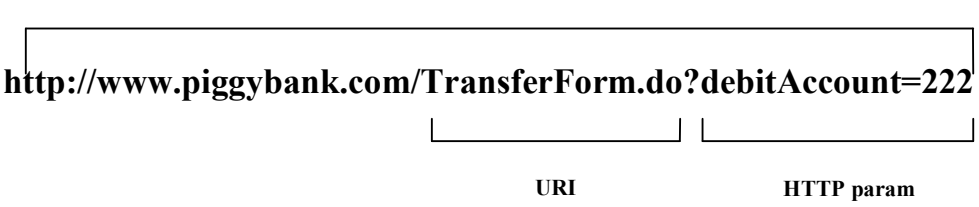
شکل ۲-۱۰ تراکنش HTTP

در یک تراکنش HTTP داده‌های کاربرد اساسی با متاداده‌های زیر تعریف می‌شوند:

- آدرس IP ماشین مشتری
- شناسه کاربر در صورتی که فرایند تصدیق HTTP را انجام می‌دهد.
- زمانی که سرور پردازش درخواست را انجام می‌دهد.
- متد HTTP (GET، POST، ...)
- URI درخواست
- پروتکل و نسخه‌ی پروتکل مانند HTTP 1.0، HTTP 1.1 و غیره.
- کد وضعیت HTTP که به مشتری پس فرستاده می‌شود.
- اندازه‌ی پاسخ برحسب بایت
- ارجاع‌دهنده که URI ای است که گزارشات مشتری از آن ارجاع شده اند.

- عامل کاربر که شامل اطلاعاتی است که مرورگر مشتری در مورد خود گزارش می‌کند. این اطلاعات شامل این موارد است: نام مرورگر، نسخه‌ی آن و سیستم عاملی که مرورگر بر روی آن در حال اجراست.

یک تفاوت مهم باید بین URL و URI در نظر گرفت. URL آدرس کامل یک درخواست HTTP مشتری است که شامل پیشوند پروتکل (http://)، نام وب‌سرور، منبع مورد درخواست از سایت (URI) و پارامتر HTTP است (شکل ۲-۱۱).



شکل ۲-۱۱ URL و URI

۲-۶-۶-۱-۱-۲ داده‌های محتوا

داده‌های محتوا در یک وب سایت مجموعه اشیاء و ارتباطاتی است که به کاربر تحویل داده می‌شود. قسمت عمده‌ی آن را ترکیبی از منابع متنی و تصویری تشکیل می‌دهد. منابع داده‌ای که برای تحویل یا تولید این داده‌ها استفاده می‌شوند شامل صفحات ایستای HTML و XML، تصاویر، کلیپ های ویدیویی، فایل‌های صوتی، قسمت هایی از صفحات که بصورت پویا از اسکریپت‌ها یا سایر برنامه‌ها تولید می‌شوند و مجموعه رکوردهای تولید شده از پایگاه های داده می‌باشند. محتوای سایت همچنین شامل معنا یا متادادهٔ ساختاری تعبیه شده در سایت یا صفحات می‌باشد مانند کلمات کلیدی توصیفگر، ویژگی‌های مستندات، تگ‌های معنایی یا متغیرهای HTTP.

۲-۶-۶-۱-۱-۳ داده‌های ساختار

داده‌های ساختار در واقع دید طراح وب از سازماندهی محتوای وب سایت را نشان می‌دهند. این سازماندهی از طریق لینک‌های بین صفحات بدست می‌آید. داده های ساختار همچنین شامل ساختار محتوای داخل صفحه که با تگ های HTML و XML نشان داده می‌شود نیز می‌باشد. به عنوان مثال سندهای HTML و XML را می‌توان در یک ساختار درختی نمایش داد.

داده های ساختاری برای یک وب سایت معمولا از طریق *نقشه های سایت* بدست می‌آیند. نقشه های سایت بصورت خودکار تولید می‌شوند و ساختار ارتباطی داخل سایت را نشان می‌دهند. یک ابزار تولید

نقشه از سایت باید قابلیت نمایش ارتباطات داخل صفحات و ارتباطات بین صفحات را داشته باشد. این لازمه مخصوصا در سایت های مبتنی بر فریم که هر بخش از صفحه در واقع نمایانگر یک مشاهده صفحه‌ی مستقل می‌باشد، اهمیت بیشتری پیدا می‌کند. برای صفحاتی که بصورت پویا ایجاد می‌شوند، ابزار تولید نقشه از سایت باید یا دانش نهفته در اسکریپت ها و سایر برنامه هایی که این صفحات را تولید می‌کنند، را داشته باشد و یا توانایی تولید محتوا با استفاده از دادن پارامتر به این اسکریپت‌ها و برنامه‌ها را داشته باشد.

۲-۶-۶-۱-۱-۴ داده‌های کاربران

پایگاه های دادهٔ وب سایت می‌توانند اطلاعات اضافی از پروفایل کاربران را نیز داشته باشند. چنین داده هایی شامل اطلاعات دموگرافیک مانند سن، جنس، شغل و غیره و یا سایر اطلاعات شناسایی کاربران ثبت نام شده، رأی هایی که کاربران به اشیای مختلف مانند صفحات، دروس آموزشی، مقالات، آزمون ها یا بازدیدهای قبلی خود داده‌اند و نیز نمایش صریح یا ضمنی از علایق کاربران می-باشند. واضح است که که دریافت چنین داده هایی به تعامل مستقیم با کاربران سایت نیاز دارد. برخی از این داده ها را می‌توان بصورت ضمنی بدست آورد مثلا اطلاعات موجود در کوکی های سمت مشتری را می‌توان به عنوان بخشی از پروفایل به حساب آورد. بسیاری از سیستم های شخصی‌سازی از ابتدا به ذخیرهٔ اطلاعات پروفایل کاربران نیاز دارند، به عنوان مثال سیستم های فیلترینگ جمعی معمولا رأی های کاربران به اشیا را ذخیره می‌کنند.

۲-۶-۶-۱-۲ پیش‌پردازش نهایی داده‌های کاربرد

اعمال زیر باعث بهبود در دقت توصیه های سیستم شخصی‌سازی بر مبنای وب‌کاوی کاربرد وب می‌شوند.

- ❖ فیلترکردن ارزشی: تعیین درجهٔ ارزش هر مشاهده صفحه یا قلم اهمیت زیادی دارد. به عنوان مثال یک کاربر ممکن است به یک قلم مانند i رجوع کند تا بفهمد که آیا به آن علاقه دارد یا خیر، سپس بلافاصله به بخش دیگری از سایت مراجعه کند. از این رو می‌توان این نوع دسترسی کاربر به قلم i را یک دسترسی غیرارزشمند به حساب آورد. به حذف صفحات یا اقلامی که توسط کاربر درخواست شده اند و غیرارزشمند بوده اند، فیلترکردن ارزشی گفته می‌شود. هدف فیلترکردن ارزشی حذف اقلام نامربوطی است که بطور چشمگیری دارای مدت زمان مشاهدهٔ

کمتری از یک مقدار آستانه در تراکنش می‌باشند. بطور معمول سنجه های آماری مانند میانگین و واریانس را می‌توان برای تعریف این مقادیر آستانه جهت فیلترکردن استفاده کرد.

- ❖ نرمال‌سازی: مقادیر خام داده‌ها مانند مدت زمان صرف شده بر روی صفحات ممکن است برای ارزشمندی یک مشاهده صفحه مناسب نباشد. این به آن علت است که فاکتورهای زیادی مانند ساختار، طول و نوع مشاهده صفحه و همچنین علاقهٔ کاربر به یک قلم خاص در صفحه می‌توانند مدت زمان صرف شده بر روی آن را تحت تاثیر قرار دهند. نرمال‌سازی وزن ها بطور مناسب می‌تواند نقش اساسی در تصحیح این فاکتورها ایفا کند. در کل دو نوع نرمال‌سازی را می‌توان بکار برد که عبارتند از نرمالسازی در طول مشاهده صفحات هر تراکنش بطور منفرد و نرمال‌سازی وزن مشاهده صفحه ها در تمامی تراکنش‌ها. این دو نوع به ترتیب نرمال‌سازی تراکنش و نرمال‌سازی مشاهده صفحه نامیده می‌شوند. نرمالسازی مشاهده صفحه برای تعیین وزن یک مشاهده صفحه برای یک کاربر نسبت به وزن همان مشاهده صفحه برای کاربران دیگر بکار می‌رود. نرمالسازی تراکنش، درجهٔ اهمیت یک مشاهده صفحهٔ یک کاربر خاص را نسبت به سایر اقلام مشاهده شده توسط آن کاربر در همان تراکنش تعیین می‌کند. نوع دوم خصوصاً برای تمرکز بر صفحات *هدف* در تاریخچه‌های کوتاه از کاربر مفید است.

۲-۶-۶-۲ کشف الگو از داده‌های کاربرد

در این بخش تمرکز خود را به وظایف داده‌کاوی که اغلب بر روی داده‌های کاربرد وب بکار می‌روند معطوف می‌کنیم. در بخش قبل (بخش مربوط به داده‌کاوی) اطلاعات پیش‌زمینه‌ای لازم و نحوهٔ کاربرد این تکنیک ها بر روی داده‌های کاربرد وب شرح داده شد.

تکنیک هایی مانند خوشه‌بندی می‌توانند منجر به کشف خوشه‌هایی از کاربران و یا فراگیران شوند. بطور کلی، دو نوع خوشه بندی می‌تواند روی داده‌های تراکنشی کاربرد وب انجام شود که عبارتند از خوشه بندی تراکنش‌ها (یا کاربران) و خوشه بندی مشاهده صفحه ها. هر یک از این رویکردها کاربردهای مختلفی دارند و بطور خاص، هر دو رویکرد را می‌توان برای شخصی‌سازی استفاده کرد. هدف نهایی در خوشه‌بندی تراکنش‌های کاربر فراهم کردن قابلیت تحلیل خوشه ها به منظور استخراج هوش تجاری یا استفاده از آن‌ها برای اعمالی نظیر شخصی‌سازی و ارائهٔ توصیه می‌باشد.

سایر تکنیک ها مانند خوشه‌بندی اقلام (مثلاً مشاهده صفحه ها)، کاوش قواعد انجمنی یا کشف الگوهای ترتیبی نیز می‌توانند برای یافتن ارتباطات مهم بین اقلام بر مبنای الگوهای گردشی کاربران

سایت مورد استفاده قرار گیرند. نتایج کاوش قواعد انجمنی به منظور تولید یک مدل برای سیستم‌های شخصی‌سازی مورد استفاده قرار می گیرد. در ابتدا می‌توان تمامی قواعد انجمنی را بر روی اطلاعات فراگیر و فعالیت‌هایش کشف کرد. سپس قوانین مفید را برحسب درجهٔ اطمینانش پیدا کرد. تمامی عبارات سمت راست این قواعد را می‌توان بر حسب اطمینان مرتب کرد و N عبارت و قاعدهٔ اول آن را به عنوان *مجموعهٔ توصیه* برای کاربر انتخاب کرد

در مورد خوشه‌بندی و کشف قواعد انجمنی عموماً ترتیب بین مشاهده صفحه‌ها در نظر گرفته نمی‌شود، از این رو برای این اعمال، یک تراکنش به صورت یک مجموعه از صفحات در نظر گرفته می‌شود. در مورد کشف الگوهای ترتیبی نیاز به در نظر گرفتن ترتیب بین مشاهده صفحات در تراکنش می‌باشد.

۲-۶-۶-۳ استفاده از الگوهای کشف شده

همان طور که اشاره شد، هدف موتور توصیه، تطبیق جلسهٔ کاربر جاری با پروفایل تجمعی کشف شده از طریق وب‌کاوی کاربرد وب و توصیهٔ یک مجموعه از اشیاء و فعالیت‌ها به کاربر است. به مجموعهٔ اشیای توصیه شده *مجموعهٔ توصیه* گفته می‌شود.

از هر یک از رویکردهای خوشه‌بندی، کاوش قواعد انجمنی و کاوش الگوهای ترتیبی می توان برای ایجاد توصیه استفاده کرد. در روش مبتنی بر خوشه‌بندی، می‌توان جلسه جاری را بصورت برداری از اشیاء نشان داد و پروفایل تجمعی را که بیشترین شباهت را با آن دارد به عنوان مبنا برای توصیه استفاده کرد. در رویکرد مبتنی بر قواعد انجمنی، می‌توان جلسهٔ جاری کاربر را با مجموعهٔ اقلام مکرر مقایسه کرد و از قوانینی که شامل آن مجموعه از اقلامی هستند که با جلسهٔ جاری بیشترین شباهت را دارند، برای توصیه استفاده کرد. روش مبتنی بر الگوهای ترتیبی نیز با در نظر گرفتن ترتیب و تغییر روش مبتنی بر قواعد انجمنی عمل می‌کند.

فصل۳- مرور ادبیات و تحقیقات پیشین

۳-۱ مقدمه

این فصل شامل دو بخش می باشد. در بخش اوّل تحقیقات انجام شده در خصوص نحوهٔ استفاده از سیستم‌های توصیه‌گر در محیط های آموزش الکترونیک مورد بررسی قرار می‌گیرد. و در بخش دوم، الگوریتم‌های مورد استفاده در کارهای انجام گرفته و همچنین در روش پیشنهادی، معرفی و بررسی می‌گردد. این الگوریتم‌ها مربوط به تکنیک‌های داده‌کاوی از جمله دسته‌بندی، خوشه‌بندی و کاوش قوانین انجمنی می‌باشد. که در ادامه هریک توضیح داده خواهد شد.

سیستم‌های توصیه گر در تجارت الکترونیک به منظور تسهیل فرآیند خرید کالا بسیار استفاده گردیده ولی به کاربرد و پیاده سازی آن در محیط‌های آموزشی کمتر توجه گردیده است. آموزش الکترونیکی حوزهٔ جدیدی برای بکارگیری سیستم‌های توصیه‌گر می‌باشد که به منظور پیشنهاد مناسب‌ترین محتوای درسی به فراگیر استفاده می‌گردد. در سال‌های اخیر در زمینهٔ راهنمایی کاربران سیستم آموزش الکترونیکی در مسیر آموزش، مطالعات زیادی صورت گرفته است. که در ادامه به معرفی برخی از این تحقیقات می‌پردازیم.

یک توصیه گر تکلیف در آموزش الکترونیکی، سیستم توصیه گری است که یک تکلیف آموزشی را بر اساس تکالیف انجام شده توسط فراگیر و موفقیت هایش و همچنین براساس تکالیف انجام شده توسط فراگیران مشابه، به وی پیشنهاد می کند. تشابه میان فراگیران از طریق بررسی پروفایل آنها و یا بر اساس الگوهای رایج دسترسی‌هایشان قابل تشخیص خواهد بود. در حقیقت، دو بخش عمده در طراحی چنین عاملی وجود دارد:

یک ماژول یادگیری که از الگوهای دسترسی قبلی یاد می گیرد و یک مدل دسترسی فردی یا عمومی استنتاج می نماید. و یک ماژول مشاوره که از آن مدل یاد گرفته شده در زمان های مشخص جهت پیشنهاد فعالیت ها استفاده می نماید. به عبارت دیگر طراحی و پیاده سازی چنین سیستمی شامل دو گام به ظاهر مستقل (ولی کاملا مرتبط) است. گام نخست، که سیستم در آن باید به جمع آوری داده‌های فراگیران بپردازد و گام دوم که با تجزیه و تحلیل داده‌های گام نخست فهرستی از فعالیت های مرتبط با نیاز های فراگیر تهیه و به وی معرفی می‌گردد.

از سیستم های توصیه گر در سیستم‌های مدیریت آموزش/یادگیری[1] استفاده می گردد زیرا دلایل و انگیزه های استفاده از این سیستم ها در سایر حوزه ها در یک LMS نیز موجود است. برخی از این دلایل عبارتند از:

- سیستم مدیریت آموزش یک سیستم تطبیقی بوده که قادر به ایجاد محیط شخصی متناسب با نیاز فراگیر می‌باشد.
- سیستم مدیریت آموزش یک سیستم تعاملی و با اثر متقابل (فعل و انفعالی) می‌باشد.
- سیستم مدیریت آموزش شامل هزاران دوره بوده که با مشکل گرانبار شدن اطلاعات مواجه می‌باشد.

۳-۲ کاربرد سیستم‌های توصیه‌گر در محیط‌های آموزش الکترونیک

این بخش، به بررسی کاربرد سیستم های توصیه گر در محیط های آموزش الکترونیک می‌پردازد. نحوۀ ارائۀ کارهای انجام شده به این صورت است که ابتدا روش ومدل آن معرفی می‌شود و سپس نتایج حاصل از کاربرد آن ارائه می‌گردد.

برخی از محققان به توانایی ها و نحوۀ استفاده از سیستم های توصیه گر در سیستم های آموزش الکترونیکی اشاره نموده و مدل هایی را به منظور عملکرد این سیستم ها در یک محیط آموزش الکترونیک ارائه نمودند. به طور نمونه:

- (Hsu, 2008, 683–688) یک سیستم توصیه گر آموزش زبان انگلیسی شخصی را برای فراگیران ESL توسعه داد که قادر به ارائۀ دروس آموزشی متناسب با علایق فراگیران و در نتیجه افزایش انگیزۀ یادگیری آنهاست.

[1] Learning Management System

- (Itmazi, Megias, 2008, 234–240) به قابلیت استفاده از سیستم‌های توصیه‌گر در سیستم مدیریت آموزشی پرداخته و چارچوبی جهت پیشنهاد لیست کوتاهی از محتویات آموزشی متناسب با یک فراگیر خاص و زمینهٔ یادگیری اش ارائه گردید.
- (Hsu, 2008, 2102–2110) یک سیستم آموزش و یادگیری توصیه گر ESL ارائه نمود که قادر به تولید اطلاعات سودمند برای مربیان به منظور حل مسائل مربوط به دستور زبان فراگیران بوده همچنین توصیه های مفیدی را به فراگیران جهت شناخت نقاط ضعف خود ارائه می‌دهد.
- (Chang, Li, 2005, 109-114) به شرح سیستم آموزش الکترونیکی شخصی سازی شده ای پرداختند که به طور خودکار با علایق و سطوح فراگیران منطبق خواهد بود. این سیستم بر اساس معماری سیستم های فناوری یادگیری IEEE (IEEE LISA[1]) و جهت دستیابی به مقیاس پذیری بالا و قابلیت استفادهٔ مجدد طراحی شده است. یک استخراج کنندهٔ بازخورد نیز با قابلیت ادغام، به منظور ترکیب مقادیر بازخوردهای متعدد (زمان مطالعه، تعداد حرکات[2] ، ذخیره/چاپ، شاخص نسبی) و استنباط اولویت‌های کاربر به منظور ارزیابی نهایی، پیشنهاد گردید. پروفایل کاربر که سطح تخصص و اولویت های کاربر را ذخیره می نماید، توسط profiler کاربر جهت ارائه اطلاعات شخصی و با استفاده از الگوریتم مشارکت جمعی، جمع آوری می‌گردد.
- (Bobadilla et al., 2009, 261–265) در زمینهٔ سیستم های توصیه گر آموزش الکترونیک، پیشنهاد می‌کنند که فراگیران با دانش بالاتر، وزن بیشتری را در محاسبهٔ پیشنهادات نسبت به کاربران با دانش پایین تر به خود اختصاص دهند.

در زیر به شرح و بررسی چند نمونه از این تحقیقات ذکر شده که در مدل تلفیقی ارائه شده در این پایان نامه تأثیر بسزایی داشته اند، می‌پردازیم:

۳-۲-۱ معرفی یک سیستم توصیه‌گر محتوای آموزشی مبتنی بر مشارکت جمعی

عملکرد سیستم‌های مبتنی بر مشارکت جمعی بدین گونه است که با تجزیه و تحلیل آماری اطلاعات و یا استخراج داده‌های فراگیر، رفتار گذشتهٔ وی و سایر اطلاعات، یک محدوده همسایگی از افراد با

[1] IEEE Learning Technology Systems Architecture
[2] Scroll

سلایق، تخصص و سطح دانش مشترک ایجاد نموده و سپس با یافتن نزدیکترین همسایه‌ها برای هر فراگیر به توصیهٔ نرخ‌گذاری‌ها و انتخاب‌های این همسایگان به فراگیر هدف می‌پردازد.

برخی محققان در زمینهٔ سیستم‌های توصیه گر آموزش الکترونیک، پیشنهاد می‌کنند که فراگیران با دانش بالاتر، وزن بیشتری را در محاسبهٔ پیشنهادات نسبت به کاربران با دانش پایین تر به خود اختصاص دهند. و به منظور نیل به این هدف، معادلات جدیدی را طراحی و ارائه نمودند. (Bobadilla et al., 2009) یکی از ایده‌های تأکید شده در فلسفهٔ پیاده‌سازی سیستم‌های توصیه‌گر، برابری میان کاربران است. یک سیستم توصیه گر متداول از رتبه‌های ارائه شده توسط کاربران با بیشترین شباهت به کاربر فعال، به منظور ایجاد پیشنهادات استفاده می نمایند. هیچ دلیلی وجود ندارد که در پیشنهاد یک فیلم، کتاب، وبلاگ و غیره، کاربری را شایسته تر و واجد شرایط تر از سایر کاربران بدانیم. اگرچه این وضعیت در برخی حوزه ها صدق نمی کند که سیستم توصیه گر در آموزش الکترونیک نمونهٔ بارزی از این مورد می باشد. زیرا در این حوزه، ایجاد تمایز میان فراگیران تازه کار و حرفه ای امکان پذیر می باشد.

در مدل مشارکت جمعی پیشنهاد شده، از دو ماتریس دو بعدی R و C استفاده می گردد. یک ماتریس نسبت دو بعدی (R) از U کاربر برای I قلم و همچنین یک ماتریس دو بعدی دیگر (C)، برای نمرات U کاربر و سطح آزمون یا امتحان T. بدین ترتیب، هر کاربر سیستم آموزشی توسط I رتبهٔ احتمالی از اقلام (که در اینجا بر اساس ارزیابی مطالب آموزشی، مربیان، مقالات، جزوات و غیره خواهد بود) و T، نمرات آزمون های سطح تحصیلی (که بیانگر دانش وی از مواد آموزشی می باشد) مشخص خواهد شد.

ایدهٔ اصلی، وزن پیشنهادات ناشی شده از جزئیات فراگیران است. نه تنها مانند تشابه متداول میان رتبهٔ فراگیر جاری با سایرین، بلکه با در نظر گرفتن این مهم که پیشنهادات فراگیران با نمرات بهتر و دانش بالاتر نسبت به پیشنهادات فراگیران با نمرات پایین تر از وزن بیشتری برخوردار است. به منظور محاسبهٔ میزان اهمیت دانش فراگیر x (C_x) در طول پیشنهاداتی که به فراگیر y با دانش C_y ارسال و یا دریافت می گردد، فرمول و معادلات بسیاری قابل اعمال می باشد. که در اینجا از دو معادلهٔ ساده که توسط تابع f پیاده سازی می گردد، استفاده شده است.

(۱-۳)

$$f = \begin{cases} C_x - C_y, & C_x > C_y \\ 0, & C_x \leq C_y \end{cases}$$

بنابراین، در فرمول (۳-۱) اگر دانش فراگیر x، ۰/۷ در بازهٔ (۱-۰) و دانش فراگیر y، ۰/۲ (در همان بازه) باشد، وزن دانش فراگیر x نسبت به y، ۰/۵ خواهد بود. در حالیکه وزن دانش فراگیر y نسبت به x صفر خواهد بود.

$$f = C_{\mathrm{x}} - \mathrm{C_y},\ f = (C_{\mathrm{x}} - \mathrm{C_y})^2,\ \ f = \mathrm{e}^{C_{\mathrm{x}} - \mathrm{C_y}} \qquad (۲-۳)$$

فرمولی که حاوی تابع نمایی است (۳-۲)، نتایج بهتر و با دقت بالاتری را ارائه می نماید. معادلهٔ جدیدتری می تواند جهت ارزیابی میزان تشابه میان دو فراگیر x و y به کار گرفته شود که آنرا میزان اهمیت می نامیم. این معادله توسط فرمول (۳-۳) تعریف می گردد. عبارت نخست معادله، بیانگر اهمیت نمرات و عبارت دوم نیز بیانگر تشابه فراگیران با توجه به رتبه هایشان و نیز اعمال برخی از معیارهای متداول مانند پیرسون، همبستگی، MSD و غیره می‌باشد.

مجموع جهت کشف میانگین حسابی T نمره که دانش فراگیر را ارزیابی می کند، به کار می رود. همچنین، C_{xt} نمایانگر دانش کاربر x بر روی موضوع، آزمون و یا مقالهٔ t می باشد.

(۳-۳)

$$imp\ (\mathrm{x,y}) = \left[\frac{1}{T}\sum_{t=1}^{T} f\,(C_{xt}, C_{yt})\right] \times sim\ (x, y)$$

مقادیر اهمیت بدست آمده میان جفت فراگیران، به منظور کشف K همسایهٔ هر فراگیر به کار می‌رود. در آزمایشات انجام شده برای هر آیتم انتخاب شده توسط کاربر، مقدار میانگین رتبه های k همسایهٔ کاربر مورد نظر، به آن آیتم خاص محاسبه گردیده و مقدار پیش بینی شده با رتبهٔ کاربر مقایسه می‌گردد. بنابراین مقداری تحت عنوان میانگین خطای مطلق یا MAE[1] بدست می‌آید.

به منظور ارزیابی اهمیت دانش کاربر x با توجه به پیشنهادات دریافت شده از کاربر y، از معیار کسینوس[2] و تابع f تعریف شده در فرمول (۳-۱) استفاده می گردد. نتایج بدست آمده با نتایج حاصل از سیستم های توصیه گر متداول که فقط با معیار کسینوس ارزیابی می گردند، مقایسه می‌شود. معیار های میزان دقت و یا مقدار MAE، تعداد پیش بینی های درست و تعداد پیش بینی های غلط، مورد بررسی قرار می‌گیرند.

نتایج حاصله حاکی از بهبود قابل ملاحظهٔ هر سه معیار فوق در مدل پیشنهادی نسبت به سیستم های توصیه گر متداول می باشد. بدین ترتیب که میانگین خطا در مدل پیشنهادی (معادلات ارائه

[1] Mean Absolute Error
[2] Cosine

شدهٔ جدید) نسبت به سیستم های توصیه گر متداول پایین تر بوده، همچنین درصد پیش بینی های درست بالا رفته و نیز درصد پیش بینی های غلط پایین می‌آید.

۳-۲-۲ معرفی یک سیستم توصیه‌گر ترکیبی در سیستم مدیریت آموزش

با توجه به اشکالاتی که هر کدام از دو روش توصیه گر مبتنی بر محتوا و مشارکت جمعی در پیش‌بینی و ارائهٔ توصیه دارند، تحقیقات بسیاری در زمینه ارائهٔ روش‌هایی ترکیبی از این دو روش انجام شده است که در اکثر این روش‌ها، به دلیل ترکیب شدن مزایای دو روش در کنار یکدیگر، کیفیت توصیه افزایش یافته است. از جمله این تحقیقات در حوزهٔ آموزش الکترونیکی، (Itmazi, Megias, 2008, 234-240) بوده است که به مطالعهٔ قابلیت استفاده از سیستم های توصیه گر در سیستم مدیریت آموزش پرداخته و الگوریتمی جهت توصیهٔ دوره های مناسب به فراگیر فعال هنگام ورود وی به دوره اش طراحی نموده اند. که قادر به توصیهٔ محتویات آموزشی و دروس متناسب با فراگیر از میان فهرست عظیم محتویات آموزشی می باشد. این الگوریتم پیشنهادی به عنوان یک سیستم توصیه گر ترکیبی عمل می نماید که از چندین رویکرد تشکیل می‌شود:

سیستم مبتنی بر محتوا، مشارکت جمعی، فیلترینگ مبتنی بر قاعده[1] (RBF) و سیستم مبتنی بر آمارگیری[2] (DBS).

در خصوص سیستم‌های مبتنی بر محتوا، مبتنی بر آمارگیری و مشارکت جمعی به تفضیل مطالبی ارائه گردید. اکنون به معرفی رویکرد فیلترینگ مبتنی بر قاعده می‌پردازیم:

فیلترینگ مبتنی بر قاعده: که اطلاعات را با توجه به مجموعه ای از قوانین که سیاست فیلترینگ اطلاعات است، فیلتر می نماید. این قوانین ممکن است بخشی از محتویات پروفایل کاربر و یا سیستم باشد که به ویژگی های مختلف اقلام داده رجوع داده شده باشد. معمولاً این سیستم‌ها به طور گسترده در موارد زیر استفاده می‌گردد:

- ✓ سانسور: RBF در محافظت از حوزه و دامنه مفید می باشد. و نیز محافظت کودکان از دسترسی به برخی صفحات اینترنتی. مانند Cyberpatrol.com و Cybersitter.com.
- ✓ فیلترینگ هرزنامه[3]: RBF جهت بکارگیری در مقابل ایمیل های هرز بسیار مفید می باشد. مانند spamassassin.apache.org و www.gfi. com.

[1] Rule Based Filtering
[2] Demographic-Based system
[3] spam

در سیستم های مدیریت آموزش، RBF بر اساس برخی قوانین سیستم و کاربر جهت فیلتر نمودن لیست پیشنهادات محتویات آموزشی و دروس به کار گرفته می‌شود.

سیستم توصیه گر متناسب با سیستم مدیریت آموزش فقط شامل یک رویکرد خالص نمی‌باشد. بلکه ترکیبی از چندین رویکرد، گزینهٔ مناسبتری خواهد بود.

برخی از ملاحظات سیستم پیشنهادی (شکل ۳-۱) به شرح زیر است:

- سیستم مبتنی بر محتوا به عنوان رویکرد اولیه به کار گرفته خواهد شد زیرا پیشنهادات جامع، کافی و مرتبطی را با استفاده از خصوصیات اقلام در فرایند توصیه ارائه می دهد. در این مرحله، دوره ها با شناسایی تشابه میان اقلام دورهٔ فعلی (دوره ای که فراگیر در حال حاضر وارد آن شده) و اقلام سایر دوره ها انتخاب می شوند. این اقلام دوره ها شامل نام، کلمات کلیدی، چکیده و ... هستند. بنابراین در مرحلهٔ نخست، سیستم مبتنی بر محتوا دوره های مرتبط را از پایگاه دادهٔ LMS[1] بازیابی می‌نماید.
- « توصیه‌های مربی» منابعی می‌باشند که مربی آنها را در دورهٔ خود به عنوان منابع پیشنهادی قرار می‌دهد. منابع می‌توانند داخلی (دوره‌هایی از همان سیستم مدیریت آموزش) و یا خارجی باشند.
- از سیستم‌های مشارکت جمعی به عنوان یک رویکرد مکمل جهت سازماندهی اولویت پیشنهادات استفاده می گردد. مکانیزم کلی این سیستم ها بر اساس تعریف زیر گروه هایی می باشد که اولویت ها و سلایقشان مشابه به کاربر فعال بوده؛ بنابراین نزدیکترین همسایه ها به فراگیر فعال، فراگیرانی از مؤسسات مشابه (حوزه/مدرسه) می‌باشند. سپس میانگین رتبهٔ زیر گروه محاسبه می گردد تا پیشنهادات با بالاترین رتبه ارائه شوند. سیستم مدیریت آموزش، می بایست راهی برای دریافت رتبه ها به روش ضمنی و یا صریح و یا ترکیبی از هر دو روش داشته باشد. رتبهٔ فراگیران از دوره ها در جدولی به صورت یک ماتریس دو بعدی ذخیره می گردد. که ردیف ها نمایانگر رتبهٔ هر فراگیر به دوره های مختلف آموزشی و ستون های آن نمایانگر رتبهٔ تمام فراگیران به یک دورهٔ آموزشی می باشد. (جدول ۳-۱)

جدول ۳-۱ ماتریس رتبه‌دهی

دوره / فراگیر	دورهٔ۱	دورهٔ۲	...	دورهٔm
فراگیر۱				۲
فراگیر۲	۵	۳	۳	
...			۳	۵
فراگیرn	۳		۵	

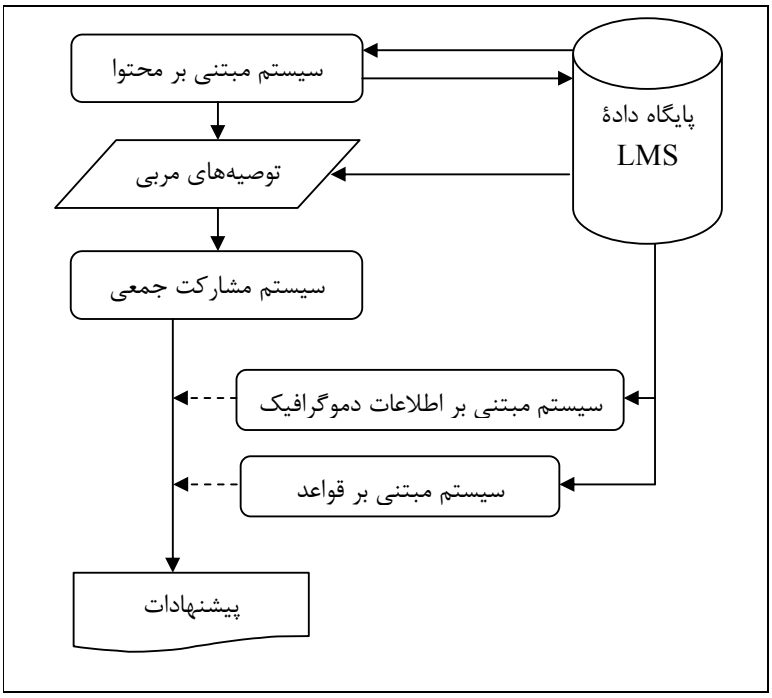
- سیستم مبتنی بر قواعد و همچنین سیستم مبتنی بر اطلاعات دموگرافیک به عنوان رویکرد مکمل به کار گرفته می شوند. زیرا اطلاعات دموگرافیک DBS و قواعد RBF جهت استفاده به عنوان رویکرد اولیه مفید نمی باشند. به طور علمی نقش DBF در یک سیستم مدیریت آموزش فیلتر نمودن پیشنهادات وارد شده از مرحلهٔ قبل بر اساس داده های دموگرافیک و شخصی فراگیران که مرتبط با مسائل آموزشی آنهاست، می باشد.

به طور مثال، داده های شخصی/دموگرافیک زیر می توانند به مسائل آموزشی فراگیر مرتبط باشند: زبان مورد نظر، تخصص دانشجویی، سطح و سال تحصیلی، بخش و حوزهٔ فراگیر.

همچنین RBF دوره های پیشنهادی وارد شده را بر اساس قواعد موجود در پروفایل سیستم و یا فراگیر فیلتر خواهد کرد. انواع قوانین زیر می تواند در پروفایل کاربر یا سیستم جهت فیلتر نمودن دوره های ورودی به کار گرفته شوند:

۱. لینک ها: سیستم هر دوره ای را که پیوندش در قواعد یافت شود را فیلتر خواهد نمود.
۲. عبارات و یا کلمات: سیستم هر دوره ای را که نام، کلمات کلیدی و یا چکیده اش مطابق با عبارت و یا کلمه ای در قواعد باشد را فیلتر خواهد کرد.
۳. تاریخ: سیستم دوره ای را که در زمان و تاریخ مقرر نباشد را نشان نخواهد داد و فیلتر خواهد نمود.

- زمانیکه فراگیر وارد دوره اش می گردد، پیشنهادات در صفحه نمایش نشان داده خواهد شد.



شکل ۳-۱ معماری پیشنهادی سیستم توصیه گر در LMS

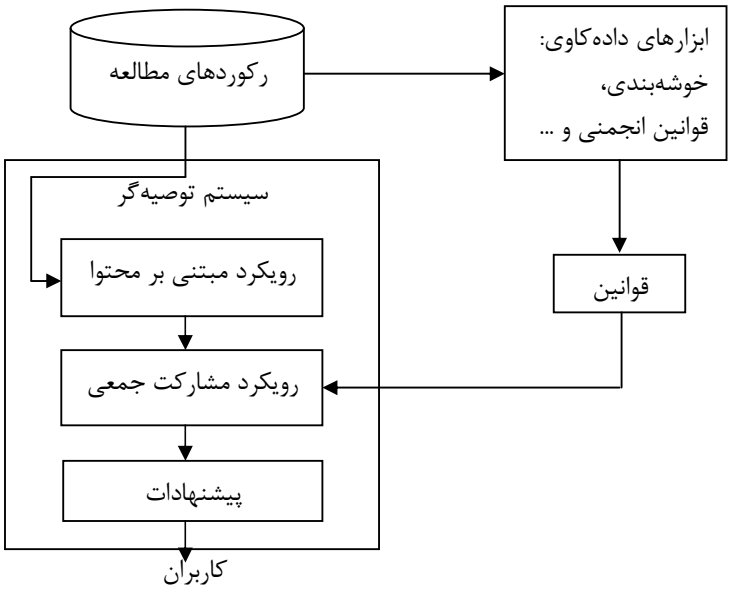
۳-۲-۳ معرفی یک سیستم توصیه‌گر آموزش زبان انگلیسی شخصی

یافتن قواعد وابستگی و یا ارتباطی، نوعی روش بازیابی اطلاعات است که می توان از آن برای توصیهٔ محصولات در وب سایت‌های فروش الکترونیکی استفاده نمود. قواعد وابستگی بر اساس تعداد مشابه و تکراری از مبادلات اطلاعات می‌توانند نتایج مفیدی را استخراج نموده و با آنالیز این اطلاعات کسب شده، باعث اجتناب از یک‌سری مبادلات تکراری شده و رده‌بندی و الگوسازی محصولات و ارتباطات بین آنها را روشن می‌نمایند.

(Hsu, 2008) یک سیستم توصیه گر آموزش زبان انگلیسی شخصی آنلاین را برای فراگیران ESL توسعه داده است. که قادر به ارائهٔ دروس آموزشی متناسب با علایق فراگیران و در نتیجه افزایش انگیزهٔ یادگیری آنهاست. این سیستم، از تجزیه و تحلیل مبتنی بر محتوا، مشارکت جمعی و تکنیک های داده کاوی استفاده می نماید. عملکرد این سیستم در دورهٔ زمانی بیش از یک سال پی‌گیری گردیده و تأثیر مثبت آن در بالا بردن انگیزه و علاقه به مطالعهٔ فراگیران ESL ثابت شده است.

معماری سادهٔ سیستم توصیه گر مورد بحث در شکل ۳-۲ نشان داده شده است.

ابتدا یک الگوریتم خوشه بندی جهت بخش بندی فراگیران به گروه های مختلف استفاده می شود. سپس الگوریتم قواعد انجمنی به منظور تجزیه و تحلیل وابستگی دروس هر بخش به کار گرفته می شود. برای هر درس در هر گروه، به کمک روش مبتنی بر محتوا، یک نمرهٔ اولیه گذاشته می شود.



شکل ۳-۲ معماری سادهٔ سیستم توصیه گر

سپس از رویکرد مشارکت جمعی- الگوریتم قواعد انجمنی- برای تنظیم نمرات دروس هر فراگیر استفاده می گردد. در رویکرد مبتنی بر محتوا ابتدا یک نمرهٔ پایه برای هر درس گذاشته می شود. در اینجا، فراگیران با انتخاب های متفاوتشان از دروس، مشخص شده و یک فرمول ساده نیز به عنوان فاکتور شخصی به کار گرفته می‌شود.

$$Weight(c) = \frac{CF}{TSC} \qquad (۴-۳)$$

که weight(c) وزن درس c، CF تعداد دفعاتی که درس c انتخاب می شود، TSC تعداد دفعاتی که تمام دروس توسط یک خوشهٔ خاص انتخاب می گردد. مقدار وزن که تلاش در محاسبهٔ اهمیت یک انتخاب خاص در مقایسه با انتخاب کلی دروس در یک خوشه دارد، به عنوان سطح علاقهٔ فراگیر برای هر درس در نظر گرفته می شود. اگر یک درس خاص از وزن بالایی برخوردار باشد، بدین معناست که فراگیران آن خوشه، احتمال بیشتری جهت انتخاب آن درس داشته، بنابراین می بایست به آنها پیشنهاد گردد. بعد از مشخص شدن وزن، الگوریتم قواعد انجمنی جهت تعیین قاعدهٔ $A \rightarrow B$ که به کار گرفته می شود. و سپس نمرات توصیه شده برای درس های مربوطه تنظیم می‌گردد. این روش در فرمول زیر بیان شده است:

$$f = \begin{cases} 0, & |B| > 0 \\ Weight(B), & |B| = 0 \hat{} |A| = 0 \\ Weight\,(B) \times Lift\,[A \to B], & |B| = 0 \hat{} |A| > 0 \end{cases} \qquad (۵-۳)$$

که x، ID فراگیر و $Weight(B)$ مقدار وزن درس B از خوشهٔ فراگیر x می باشد. $|A|$تعداد مشاهدات و جستجوهای قبلی درس A و $|B|$تعداد مشاهدات درس B می باشد. $Lift\,[A \to B]$ مقدار پیشرفت قاعدهٔ انجمنی موجود A→B برای این خوشه است.

$|B| > 0$ بدین معناست که فراگیر درس B را انتخاب نموده و نمرهٔ توصیه شده برای این درس صفر می باشد. هنگامیکه $|B| = 0$ و $|A| = 0$ باشد، بدین معناست که فراگیر درس A و B را برنگزید. و نمرهٔ توصیه گر برای درس B، وزن B است. هنگامیکه $|B| = 0$ و $|A| > 0$ باشد در آن صورت فراگیر درس A را انتخاب نموده ولی درس B را برنگزیده است. و نمرهٔ توصیه گر برای درس B عبارتند از:

$Weight\,(B) \times Lift\,[A \to B]$.

بنابراین هنگامیکه درس B را به این دست از فراگیران پیشنهاد می نماییم احتمال موفقیت بسیار بالاست. عملکرد این سیستم در دورهٔ زمانی بیش از یک سال (از سال ۲۰۰۳ تا ۲۰۰۴) پی گیری گردیده است.

با مقایسهٔ نتایج به این مطلب می رسیم که اطمینان اکثر قواعد در سال ۲۰۰۴ نسبت به سال ۲۰۰۳ افزایش یافته است. به طور نمونه اطمینان قاعدهٔ {۱۶۶}→{۱۶۷}، در سال ۲۰۰۳، ۷۶.۶۷ می باشد که در سال ۲۰۰۴ به ۹۷.۴۶ افزایش یافته است. پس هنگامیکه فراگیر درس{۱۶۷} را مطالعه می نماید، احتمال مطالعهٔ درس {۱۶۶} توسط وی ۲۰.۷۹٪ افزایش می یابد. این بدان معناست که ۲۰.۷۹٪ فراگیران توصیهٔ سیستم را در خصوص مطالعهٔ درس {۱۶۶} پذیرفته اند. بنابراین مشاهدهٔ عملکرد مؤثر سیستم توصیه گر، کاملاً واضح می باشد.

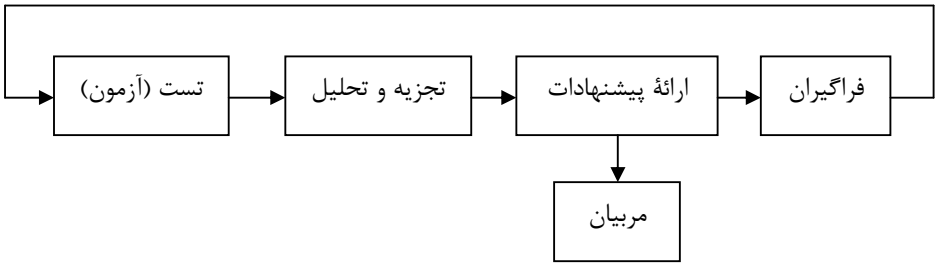
۳-۲-۴ معرفی یک سیستم توصیه‌گر مبتنی بر آموزش و یادگیری

در محیط‌های آموزشی معمولی، اساتید اطلاعات مورد نیاز خود را در زمینه نحوهٔ یادگیری فراگیران، از طریق تبادلات رودرروی حضوری با آنها بدست می‌آورند. به هر حال زمانی که فراگیر در محیطی الکترونیکی فعالیت می کند، این گونه نظارت غیر قابل اجرا است و استاد باید به دنبال روش‌های

دیگری برای دستیابی به اطلاعات مورد نیازش باشد. داده‌کاوی، استخراج خودکار الگوهای مفید از مجموعه داده‌های بزرگ است و با جستجو و یافتن اطلاعات آموزشی سودمند بر مبنای اسناد آموزشی، در ارزیابی و بهبود سیستم آموزش نیز کاربرد دارد. سازمان‌هایی که پایگاه های آموزش از راه دور را راه اندازی می‌کنند بطور خودکار حجم عظیمی از اطلاعات را توسط خدمتگذار[1]های وب تولید و در قسمت ثبت وقایع[2] خدمت گذارها جمع آوری می‌کنند. محیط‌های آموزشی تحت وب قادرند بیشتر رفتارهای یادگیری فراگیران را ثبت و حجم عظیمی از پروفایل‌ها را فراهم کنند. داده‌کاوی می‌تواند برای کاراتر ساختن محیط آموزشی، مؤثر باشد (Baker). در یکی دیگر از سیستم‌های توصیه‌گر ترکیبی، روش سیستم، از ترکیب روش مبتنی بر الگوهای رفتاری فراگیران و نیز دسته‌بندی منابع و مفاد آموزشی تشکیل شده است.

(Hsu, 2008, 2102-2110) یک سیستم آموزش و یادگیری توصیه گر ESL را بر اساس روند چرخشی فرآیند «ارزیابی- آموزش- ارزیابی مجدد» (شکل ۳-۳) پیشنهاد نمود. مفهوم اصلی، تجزیه و تحلیل خودکار نتایج آزمون گرامری است که توسط فراگیران انجام شده و نیز ارائهٔ پیشنهادات بلادرنگ جهت بهبود توانایی های آنها در حوزه هایی است که ضعیف می باشند. پس از مطالعهٔ مطالب تکمیلی، فراگیر آزمونی مشابه آزمون نخست داده و تجزیه و تحلیل ها و فرایند توصیه مجدداً به صورت خودکار انجام می پذیرد.

ابتدا سوالات آزمون طبقه بندی گردیده و هر سوال به صورت (Ai,Bi) برچسب گذاری می شود. که Ai به معنای شمارهٔ سوال و Bi نوع سوال است.



شکل ۳-۳ چرخشهٔ توصیهٔ آموزش و یادگیری

[1] Servers
[2] Log files

سیستم برای هر فراگیر، یک جدول آماری پاسخ صحیح/غلط ایجاد می نماید. جداول آماری پاسخ صحیح/غلط تمام فراگیران با هم جمع شده و در یک جدول کلی وارد می گردد. سپس، مجموع مقادیر جدول به ترتیب نزولی رتبه بندی شده تا یک ترتیب نزولی از نقاط ضعف فراگیران را نشان دهد. با توجه به نقاط ضعف مشاهده شده، سیستم قادر به ایجاد پیشنهادات آموزشی مناسب و دقیقی خواهد بود.

سپس الگوریتم خوشه بندی سلسله مراتبی برای داده های جمع آوری شده به کار گرفته می‌شود. براساس چنین اطلاعاتی، مربی درک بهتری از فراگیر پیدا نموده و قادر به بهبود وضعیت وی خواهد بود.

به منظور ارزیابی عملکرد سیستم، از داده‌های ۵۰ فراگیر ESL با مهارت کم در زبان انگلیسی که از آزمون گرامر ابتدایی سیستم رد شده اند، استفاده می‌نماییم. از این ۵۰ فراگیر، ۲۵ نفر در گروه آزمایشی و ۲۵ نفر دیگر در گروه شاهد هستند. به گروه آزمایشی با توجه به برنامهٔ توصیه گر و روش ارائه شده توسط سیستم آموزش داده می شود. در حالیکه آموزش گروه شاهد به روش نرمال صورت می پذیرد. بعد از سه ماه آموزش اصلاحی، از همان ۵۰ فراگیر آزمونی مشابه دیگری گرفته می شود و نتایج آزمون نمایانگر آن است که فراگیران گروه آزمایشی بهبود قابل ملاحظه ای داشته اند. این فراگیران نمرهٔ میانگین ۱۰۴ را در آزمون نخست و نمرهٔ میانگین ۱۲۴ را در آزمون دوم که سه ماه بعد برگزار گردید، کسب نمودند. افزایش قابل توجهی را مشاهده می نماییم. در مقابل فراگیران گروه شاهد، نمرهٔ میانگین ۱۱۱ را هم در آزمون نخست و هم در آزمون دوم کسب نمودند که هیچ پیشرفتی را مشاهده نمی کنیم.

نتایج آزمون گروه های آزمایشی و شاهد بیانگر بهبود شایانی در وضعیت فراگیرانی است که از سیستم توصیه گر و آموزش اصلاحی آن استفاده نمودند. سیستم نه تنها به شناسایی و کشف نقاط ضعف و مشکلات فراگیران در یادگیری کمک می نماید، بلکه با ارائهٔ پیشنهاداتش مربیان را قادر می سازد تا طرح‌ریزی‌های استراتژی اصلاحی مؤثر و مطابق با نیاز فراگیر داشته باشند. و علاوه بر آن پیشنهادات مفیدی را برای بهبود وضعیت فراگیرانی که در شناسایی نقاط قوت و به خصوص نقاط ضعف اشان در فرایند یادگیری زبان به آموزش اصلاحی نیاز دارند، ارائه می دهند.

۳-۳ بررسی الگوریتم‌های مورد استفاده در پژوهش

در این بخش الگوریتم‌های مورد استفاده در کارهای انجام گرفته و همچنین روش پیشنهادی، مورد بررسی قرار می گیرند. این الگوریتم‌ها مربوط به تکنیک‌های داده‌کاوی از جمله دسته‌بندی، خوشه‌بندی و کاوش قوانین انجمنی می‌باشد. در ادامه در مورد هریک توضیح داده خواهد شد.

۳-۳-۱ الگوریتم‌های دسته‌بندی

الگوریتم‌های دسته‌بندی، به رده‌بندهای trees، Bayesian، functions، lazy، meta، misc و rules تقسیم شده‌اند (شکل ۳-۴(الف) و (ب)). (Data Mining, witten et Al., 2005)

دراین بخش برخی از این کلاسه‌بندها معرفی می‌شوند. (مظهری، ایمانی، ۱۳۸۸)

	Name	Function
Bayes	*AODE*	Averaged, one-dependence estimators
	BayesNet	Learn Bayesian nets
	ComplementNaiveBayes	Build a Complement Naïve Bayes classifier
	NaiveBayes	Standard probabilistic Naïve Bayes classifier
	NaiveBayesMultinomial	Multinomial version of Naïve Bayes
	NaiveBayesSimple	Simple implementation of Naïve Bayes
	NaiveBayesUpdateable	Incremental Naïve Bayes classifier that learns one instance at a time
Trees	*ADTree*	Build alternating decision trees
	DecisionStump	Build one-level decision trees
	Id3	Basic divide-and-conquer decision tree algorithm
	J48	C4.5 decision tree learner (implements C4.5 revision 8)
	LMT	Build logistic model trees
	M5P	M5′ model tree learner
	NBTree	Build a decision tree with Naïve Bayes classifiers at the leaves
	RandomForest	Construct random forests
	RandomTree	Construct a tree that considers a given number of random features at each node
	REPTree	Fast tree learner that uses reduced-error pruning
	UserClassifier	Allow users to build their own decision tree
Rules	*ConjunctiveRule*	Simple conjunctive rule learner
	DecisionTable	Build a simple decision table majority classifier
	JRip	RIPPER algorithm for fast, effective rule induction
	M5Rules	Obtain rules from model trees built using M5′
	Nnge	Nearest-neighbor method of generating rules using nonnested generalized exemplars
	OneR	1R classifier
	Part	Obtain rules from partial decision trees built using J4.8
	Prism	Simple covering algorithm for rules
	Ridor	Ripple-down rule learner
	ZeroR	Predict the majority class (if nominal) or the average value (if numeric)
Functions	*LeastMedSq*	Robust regression using the median rather than the mean
	LinearRegression	Standard linear regression
	Logistic	Build linear logistic regression models
	MultilayerPerceptron	Backpropagation neural network
	PaceRegression	Build linear regression models using Pace regression
	RBFNetwork	Implements a radial basis function network
	SimpleLinearRegression	Learn a linear regression model based on a single attribute
	SimpleLogistic	Build linear logistic regression models with built-in attribute selection
	SMO	Sequential minimal optimization algorithm for support vector classification

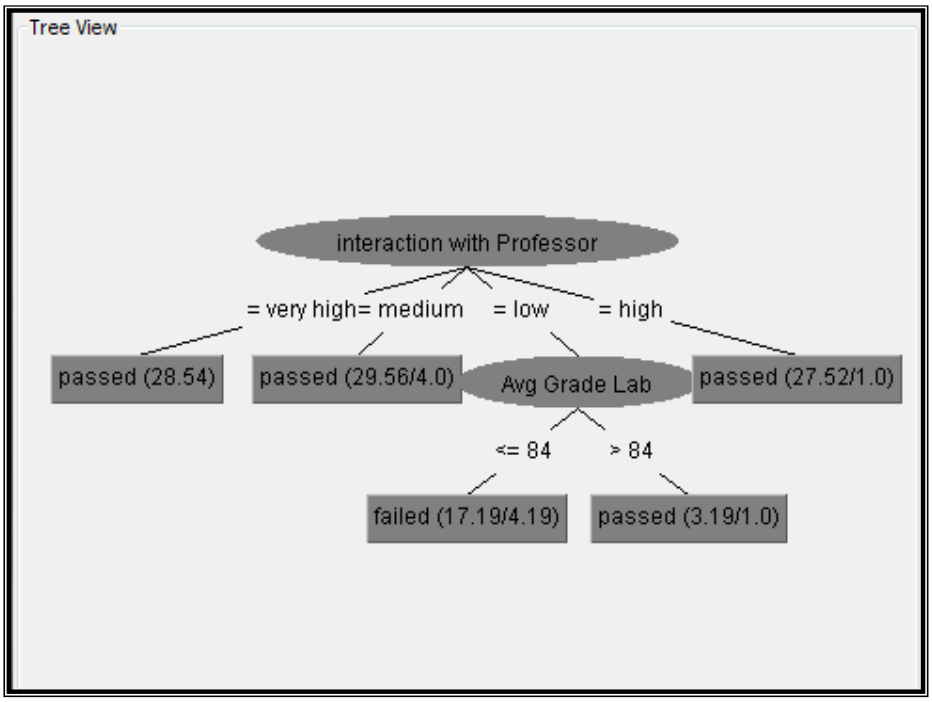
شکل ۳-۴(الف) الگوریتم‌های دسته‌بندی

	Name	Function
	SMOreg	Sequential minimal optimization algorithm for support vector regression
	VotedPerceptron	Voted perceptron algorithm
	Winnow	Mistake-driven perceptron with multiplicative updates
Lazy	*IB1*	Basic nearest-neighbor instance-based learner
	IBk	*k*-nearest-neighbor classifier
	KStar	Nearest neighbor with generalized distance function
	LBR	Lazy Bayesian Rules classifier
	LWL	General algorithm for locally weighted learning
Misc.	*Hyperpipes*	Extremely simple, fast learner based on hypervolumes in instance space
	VFI	Voting feature intervals method, simple and fast

شکل ۳-۴(ب) الگوریتم‌های دسته‌بندی

۳-۳-۱-۱ درختان تصمیم‌گیری

بر اساس مجموعهٔ آموزشی یک درخت ایجاد می‌کنیم که در این درخت هر گرهٔ داخلی یک آزمون را روی یک صفت نشان می‌دهد، هر شاخه نتیجه‌ای از تست را نشان می‌دهد و هر برگ برچسب یک کلاس را نگهداری می‌کند. نمونه‌ای از درخت تصمیم‌گیری در شکل ۳-۵ آمده است:



شکل ۳-۵ نمونه‌ای از یک درخت تصمیم

درخت تصمیم در شکل ۳-۵ مفهوم پیش بینی قبولی و یا رد یک فراگیر را نشان می‌دهد، اینکه آیا یک فراگیر، موفق به گذراندن یک دورهٔ تحصیلی است یا خیر. همانطور که در شکل دیده می شود، این درخت دارای دو برچسب کلاس متمایز در برگ‌های خود است که passed و failed می‌باشند. نحوهٔ استفاده از درخت تصمیم گیری به این صورت است: اگر تاپلی چون X که برچسب کلاس آن نامشخص است داشته باشیم صفات این تاپل در درخت مورد آزمون قرار می گیرند و یک مسیر از ریشه به سمت یک برگ که برچسب یک کلاس را دارد ایجاد می‌شود.

از دلایل محبوبیت درختان تصمیم گیری می‌توان به موارد زیر اشاره نمود:

- ✓ ساختار کلاسه بندهای درختان تصمیم گیری به هیچ قلمرو دانش و یا تنظیم پارامتری نیاز ندارند؛
- ✓ درختان تصمیم گیری می توانند داده هایی با حجم زیاد را مدیریت کنند؛
- ✓ نمایش آنها از دانش کسب شده به صورت یک درخت، برای انسان قابل درک است و به راحتی شبیه سازی می شود؛
- ✓ مراحل یادگیری و کلاسه بندی درختان تصمیم گیری سریع و ساده هستند؛
- ✓ در حالت کلی، کلاسه بندهای درختان تصمیم گیری میزان صحت قابل قبولی دارند؛

با این حال، میزان موفقیت در استفاده از درختان بستگی به داده های مورد استفاده نیز دارد.

الگوریتم‌های مورد استفاده در درختان تصمیم گیری، درختان خود را توسط عملیات بازگشتی تصمیم و غلبه می‌سازند و ساختاری از بالا به پایین دارند. الگوریتم با یک مجموعه آموزشی از تاپل ها و برچسب کلاس مرتبط با آنها شروع می‌کند. مجموعه آموزشی به صورت بازگشتی به زیر مجموعه های کوچکتری تجزیه شده و درخت را تشکیل می‌دهد. از جمله الگوریتم‌های درختی می‌توان به نمونه‌های زیر اشاره نمود:

- ❖ AD Tree

الگوریتم AD Tree یک ساختار درختی بهینه برای پاسخ‌دهی به گزارش‌های شمارشی ایجاد می‌کند. بدین معنا که تعداد رکوردهایی را که در ترکیب خاصی از صفات و مقادیر آنها صدق می‌کنند را، بر می‌گردانند. برای ساختن این درخت به هر صفت یک اندیس نسبت داده می‌شود. گرهٔ ریشه دارای اندیس صفر بوده و تعداد تمامی رکوردهای موجود در مجموعهٔ داده را در خود نگه می‌دارد. از هر گره با اندیس i، گره‌هایی با اندیس j گسترش داده می‌شوند به طوریکه $j > i$. بنابراین تمامی صفات از گرهٔ ریشه گسترش داده می‌شوند.

❖ CART[1]

CART درختانی با تنها دو شاخه در هر نود ایجاد می‌کند. هر شاخه منجر به نود تصمیم دیگر یا یک نود برگ می‌شود. با پیمایش یک درخت تصمیم از ریشه به پایین به یک مورد یک رده یا مقدار نسبت می‌دهیم. هر نود از داده‌های یک مورد برای تصمیم‌گیری درباره آن انشعاب استفاده می‌کند. در واقع درخت تصمیم گیری CART یک روال بخش بندی بازگشتی و دودوئی است که قادر به پردازش صفت‌های خاصه با مقادیر پیوسته و گسسته است. هدف، تولید سلسله ای از درخت‌های تودرتو و هرس شده است که هر یک از آنها درختانی بهینه و کاندید هستند.

درخت‌های تصمیم از طریق جداسازی متوالی داده‌ها به گروه‌های مجزا ساخته می‌شوند و هدف در این فرآیند افزایش فاصله بین گروه‌ها در هر جداسازی است.

یکی از تفاوت‌ها بین متدهای ساخت درخت تصمیم این است که این فاصله چگونه اندازه‌گیری می‌شود. درخت‌های تصمیمی که برای پیش‌بینی متغیرهای دسته‌ای استفاده می‌شوند، درخت‌های classification نامیده می‌شوند زیرا نمونه‌ها را در دسته‌ها یا رده‌ها قرار می‌دهند. درخت‌های تصمیمی که برای پیش‌بینی متغیرهای پیوسته استفاده می‌شوند درخت‌های regression نامیده می‌شوند.

❖ C۴.۵

یک معیار استاندارد در یادگیری ماشین است. انتخاب صفت در ID۳[2] و C۴.۵ بر اساس حداقل کردن مقیاس اطلاعات در یک گره است. هر مسیر از ریشه به سمت یک گره، نمایان-گر یک قانون دسته‌بندی می‌باشد.

تئوری بر این اساس است که تعداد آزمون‌هایی که باعث می‌شود یک نمونه جدید در داخل پایگاه داده، دسته‌بندی شود، حداقل گردد. بخش انتخاب صفت در C۴.۵ بر این اساس است که پیچیدگی درخت تصمیم به شدت وابسته به مقدار اطلاعاتی است که با آن صفت در ارتباطند. با انتخاب آن صفت، اطلاعات بیشتر از هر صفت دیگری، جدا و تقسیم می‌شوند.

الگوریتم C۴.۵ دامنهٔ دسته‌بندی را علاوه بر صفات قیاسی در انواع صفات عددی نیز توسعه می‌دهد. الگوریتم اصولا صفتی را که حداکثر درجه جداسازی بین دسته‌ها را دارد را انتخاب می‌کند و درخت تصمیم را بر اساس آن می‌سازد. تولید درخت تصمیم اولیه از روی مجموعه

[1] Classification and Regression Trees
[1] Inducing Decision trees

داده‌ای، مهم‌ترین بخش الگوریتم C۴.۵ می‌باشد. الگوریتم در نهایت یک دسته‌بند را در قالب یک درخت تصمیم تولید می‌کند که دارای دو نوع گره است. یک گره بصورت برگ که یک دسته را مشخص می‌کند و یک گره تصمیم که آزمون‌هایی روی یک صفت انجام می‌دهد تا یک شاخه یا زیر درخت به ازای هر خروجی آزمون تولید می‌کند. روش ساخت درخت مشابه‌ای، بصورت بازگشتی به هر زیر مجموعه از نمونه‌ها اعمال می‌شود. این رویه ادامه می‌یابد تا زیر مجموعه‌ها شامل نمونه‌هایی باشند که به یک دسته تعلق داشته باشند. (نصیری، کاردان، هادیان و مینایی، ۱۳۸۸)

۳-۳-۱-۲ رده بندهای Rules

- جدول تصمیم‌گیری[1]

یک رده بند بر اساس اکثریت جدول تصمیم‌گیری می‌سازد. این الگوریتم، با استفاده از جستجوی اولین بهترین، زیر دسته‌های ویژگی‌ها را ارزیابی می‌کند و می‌تواند از اعتبارسنجی تقاطعی برای ارزیابی بهره ببرد (Kohavi ،۱۹۹۵). یک امکان این است که به جای استفاده از اکثریت جدول تصمیم‌گیری که بر اساس دسته ویژگی‌های مشابه عمل می‌کند، از روش نزدیکترین همسایه برای تعیین رده هر یک از نمونه‌ها که توسط مدخل جدول تصمیم‌گیری پوشش داده نشده‌اند، استفاده شود.

- Conjunctive Rule

قاعده‌ای را یاد می‌گیرد که مقادیر رده‌های عددی را پیش‌بینی می‌کند. نمونه‌های آزمایشی به مقادیر پیش فرض رده نمونه‌های آموزشی، منسوب می‌شوند. سپس تقویت اطلاعات (برای رده‌های رسمی)، یا کاهش واریانس (برای رده‌های عددی) مربوط به هر والد محاسبه شده و به روش هرس کردن با خطای کاهش یافته[2]، قواعد هرس می‌شوند.

- ZeroR

برای رده‌های اسمی، اکثریت داده‌های مورد آزمایش و برای رده‌های عددی، میانگین آنها را پیش‌بینی می‌کند. این الگوریتم بسیار ساده است.

- M5Rules

به کمک M5 از روی درختهای مدل، قواعد رگرسیون استخراج می‌کند.

[1] Decision Table
[2] Reduced-error pruning

۳-۳-۱-۳ الگوریتم‌های بیز

❖ بیز ساده[1]

استفاده از تئوری بیز، ابزاری قدرتمند برای تصمیم گیری در شرایط عدم قطعیت[2] است. یک شکل خیلی ساده از دسته بندِ بیز، تحت عنوان نیو-بیز نامیده می شود که به صورت زیر عمل می نماید.

اگر D را مجموعه رکوردها در نظر بگیریم که هر رکورد دارای برداری شامل n صفت باشد، هدف یافتن مقداری برای صفت دسته است که مقدار عبارت (۳-۶) را حداکثر کند.

$$P\ (Ci|A1, A2, \ldots, An \qquad (۳-۶)$$

که این امر، هم ارز با حداکثر شدن فرمول (۳-۷) است:

$$P(X|C_i) = \prod^{n}_{k=1} P(x_k|C_i) = P(x_1|C_1)\times P(x_2|C_2)\times\ldots\times P(x_n|C_i) \qquad (۳-۷)$$

در اینجا p احتمال پسین[3] می باشد. (کائدی، برآنی و قاسم آقایی، ۱۳۸۷)

۳-۳-۱-۴ الگوریتم‌های Functions

❖ Simple Linear Regression

مدل رگرسیون خطی یک ویژگی مشخص را یاد می‌گیرد، آنگاه مدل با کمترین خطای مربعات را انتخاب می‌کند. در این الگوریتم، مقادیر از دست رفته و مقادیر غیرعددی مجاز نیستند.

❖ Linear Regression

رگرسیون خطی استاندارد با کمترین خطای مربعات را انجام می‌دهد و می‌تواند به طور اختیاری به انتخاب ویژگی بپردازد، این کار میتواند به صورت حریصانه، با حذف عقب رونده[4] انجام شود، یا با ساختن یک مدل کامل از همه ویژگیها و حذف یکی یکی جمله‌ها با ترتیب نزولی ضرایب استاندارد شده آنها، تا رسیدن به شرط توقف مطلوب انجام گیرد.

❖ Least Med sq

یک روش رگرسیون خطی مقاوم است که به جای میانگین مربعات انحراف از خط رگرسیون، میانه را کمینه می‌کند. این روش به طور مکرر رگرسیون خطی استاندارد را به

[3] Naïve Bayesian
[4] Uncertainty
[1] Posterior probability
[2] Backward elimination

زیرمجموعه‌هایی از نمونه‌ها اعمال می‌کند و نتایجی را بیرون می‌دهد که کمترین خطای مربع میانه را دارند.

- Logistic Regression

رگرسیون لجستیک یک روش رگرسیون غیرخطی است که هرکدام از نمونه ها داده هایش بااحتمال شرطی رابطه دارند.

- SMOreg

الگوریتم بهینه سازی حداقل ترتیبی را روی مسایل رگرسیون اعمال می‌کند.

- Pace Regression

با استفاده از تکنیک رگرسیون pace، مدل‌های رگرسیون خطی تولید می‌کند. رگرسیون pace، زمانی که تعداد ویژگی‌ها خیلی زیاد است، به طور ویژه‌ای در تعیین ویژگی‌هایی که باید صرف‌نظر شوند، خوب عمل می‌کند. در واقع در صورت وجود نظم و ترتیب خاصی، ثابت می‌شود که با بی‌نهایت شدن تعداد ویژگی‌ها، این الگوریتم بهـینه عمل می‌کند.

- RBF Network

یک شبکه با تابع پایه‌ای گوسی شعاعی را پیاده سازی می‌کند. مراکز و عرض‌های واحدهای مخفی به وسیله روش میانگین K[1] تعیین می‌شود. سپس خروجی‌های فراهم شده از لایه‌های مخفی[2]، با استفاده از رگرسیون منطقی در مورد رده‌های اسمی و رگرسیون خطی در مورد رده‌های عددی، با یکدیگر ترکیب می‌شوند. فعال سازی‌های توابع پایه پیش از ورود به مدل‌های خطی، با جمع شدن با عدد یک، نرمالیزه می‌شوند. در این الگوریتم می‌توان K، تعداد خوشه‌ها، بیشترین تعداد تکرارهای رگرسیون‌های منطقی برای مسأله‌های رده‌های رسمی، حداقل انحراف معیار خوشه‌ها، و مقدار بیشینه رگرسیون را تعیین نمود. اگر رده‌ها رسمی باشد، میانگین K به طور جداگانه به هر رده اعمال می‌شود تا K خوشه مورد نظر برای هر رده استخراج گردد.

۳-۳-۱-۵ رده بندهای Lazy

یادگیرنده‌های lazy نمونه‌های آموزشی را ذخیره می‌کنند و تا زمان رده بندی هیچ کار واقعی انجام نمی‌دهند.

[1] k-means
[2] Hidden layer

- IB1

یک یادگیرنده ابتدایی بر پایه نمونه است که نزدیکترین نمونه‌های آموزشی به نمونه‌های آزمایشی داده شده را از نظر فاصله اقلیدسی پیدا کرده و نزدیکترین رده‌ای مشابه رده همان نمونه‌های آموزشی را تخمین میزند.

- IBK

یک رده بند با K همسایه نزدیک است که معیار فاصله ذکر شده را استفاده می‌کند. تعداد نزدیکترین فاصله‌ها (پیش فرض$k=1$ است)، می‌تواند به طور صریح در ویرایشگر شیء تعریف شود. پیش‌بینی‌های متعلق به پیش از یک همسایه میتواند بر اساس فاصله آنها تا نمونه‌های آزمایشی، وزندار گردد. دو فرمول متفاوت برای تبدیل فاصله به وزن، پیاده سازی شده‌اند. تعداد نمونه‌های آموزشی که به وسیله رده بند نگهداری می‌شود، می‌تواند با تنظیم گزینه اندازه پنجره محدود گردد. زمانی که نمونه‌های جدید اضافه می‌شوند، نمونه‌های قدیمی حذف شده تا تعداد کل نمونه‌های آموزشی در اندازه تعیین شده باقی بماند.

- Kstar

یک روش نزدیکترین همسایه است که از تابع فاصله ی عمومی شده بر اساس تبدیلات استفاده میکند. این الگوریتم یک یادگیر مبتنی بر نمونه[1] می باشد و هر رکورد جدید را توسط مقایسه آن با رکوردهای کلاس بندی شده موجود در پایگاه داده، کلاس بندی می کند. مبنای مفروض این الگوریتم، این است که نمونه های شبیه به هم دارای کلاس های شبیه به هم هستند. دو مؤلفه اساسی یادگیرهای مبتنی بر نمونه عبارتند از تابع فاصله که میزان شباهت نمونه ها به یکدیگر را تعیین می‌کند و تابع کلاس بندی که مشخص می کند چگونه تشابه نمونه‌ها منجر به یک کلاس نهایی برای یک نمونه جدید خواهد شد.

- LWL

یک الگوریتم کلی برای یادگیری وزن دار شده به صورت محلی است. این الگوریتم با استفاده از یک روش بر پایه نمونه، وزنها را نسبت می‌دهد و از روی نمونه‌های وزندار شده، رده‌بند را می‌سازد. رده‌بند نیو- بیز، در ویرایشگر شیءLWL انتخاب می‌شود. برای مسایل رده بندی و رگرسیون خطی برای مسایل رگرسیون، انتخاب‌های خوبی هستند. می‌توان در این الگوریتم، تعداد همسایه‌های مورد استفاده را که پهنای باند هسته و شکل هسته مورد

[1] Instance based

استفاده برای وزن دار کردن را (خطی، معکوس، یا گوسی) مشخص میکند، تعیین نمود. نرمال سازی ویژگیها به طور پیش فرض فعال است.

۳-۳-۲ الگوریتم‌های خوشه‌بندی

با استفاده از این تکنیک می توان فراگیران را بر اساس ویژگیهای مختلفی به عنوان مثال میزان فعال بودن آنها گروه بندی کرد و با هر گروه از آنها بصورت مناسبی رفتار کرد.

دو دسته‌ی کلی از الگوریتم‌های خوشه‌بندی، سلسله‌مراتبی و تفکیکی می‌باشند. الگوریتم‌های سلسله مراتبی خوشه‌ها را به تدریج می‌سازند ولی الگوریتم‌های تقسیم کننده، مستقیماً خوشه‌بندی را انجام می‌دهند. آنها سعی می‌کنند که خوشه‌ها را با جایگذاری مجدد نقطه‌ها بین زیرمجموعه‌ها کشف کنند. در یک تقسیم‌بندی جزیی‌تر این الگوریتم‌ها به صورت زیر دسته‌بندی می‌گردند:

۱. الگوریتم‌های تفکیک

یکی از انواع الگوریتم‌های خوشه‌بندی است که درابتدا مجموعهٔ داده‌ها را به بخش‌هایی تبدیل کرده، سپس با استفاده از برخی معیارها آن دسته‌بندی را مورد ارزیابی قرار می‌دهد و در صورت لزوم در دسته‌بندی اولیه تغییراتی ایجاد می‌نماید. رایج‌ترین الگوریتم‌های خوشه‌بندی در این دسته قرار می‌گیرند. این الگوریتم‌ها داده‌ها را به چندین زیر مجموعه تقسیم می‌کنند. به علت این‌که چک کردن همهٔ زیر مجموعه‌های ممکن، امکانپذیر نیست. تابع‌های مکاشفه‌ای حریصانهٔ خاصی به کار گرفته می‌شوند. در این الگوریتم‌ها به صورت تکراری نقاط بین k خوشه جابجا می‌شوند تا در نهایت به بهترین خوشه ممکن نسبت داده شوند. بر خلاف متدهای سلسله مراتبی که خوشه‌ها بعد از ساخته شدن بازبینی نمی‌شود، این الگوریتم‌ها مرتبا خوشه‌ها را به منظور بهبود بخشیدنشان تغییر می‌دهند، به همین دلیل در این روش‌ها نهایتاً خوشه‌هایی با کیفیت بالا خواهیم داشت.

- ❖ الگوریتم k-means

الگوریتم k-means یکی از رایج ترین الگوریتم های خوشه بندی می باشد. این الگوریتم یک بانک داده D را که از n شی ء تشکیل شده به یک مجموعه که دارای K خوشه می باشد، تقسیم می کند و عمل مذکور را با استفاده از معیارهای خاصی به انجام می‌رساند. در این روش هر شی ء با نقطه ای روی صفحه متناظر می‌گردد و هر خوشه توسط مرکز آن نمایش داده می‌شود. گام‌های الگوریتم به صورت به شرح زیر می‌باشد:

گام اول: اشیاء در ابتدا به K زیر مجموعه غیر تهی تقسیم می‌شود.

گام دوم: نقاط Seed به عنوان مرکز هر خوشه محاسبه می‌گردد.

گام سوم: هر شیء به خوشه‌ای تخصیص می‌یابد که نزدیکترین فاصله را با مرکز (نقطه میانه) آن خوشه داشته باشد.

گام چهارم: برگشت به گام ۲، توقف در صورتیکه تخصیص جدیدی وجود نداشته باشد.

این روش خوشه بندی که رایج ترین روش در این زمینه می‌باشد کارائی بالایی دارد. همچنین این الگوریتم غالبا در یک بهینه محلی پایان می‌یابد که این مورد نیز از ویژگی های مثبت روش می‌باشد. در این روش پارامتر هایی نظیر K و میانه به عنوان ورودی معرفی می‌شوند که منظور از میانه، نقطه میانگینی است که نحوه محاسبه آن بر اساس فرمول خاصی برای الگوریتم تعریف می‌شود. (شکورنیاز، حاجی علی اکبری، ۱۳۸۷)

- الگوریتم EM[1] که در سال ۱۹۹۷ ارائه شده است، یک چارچوب کلی برای الگوریتم‌های آماری می‌باشد که یکی از حالت‌های خاص آن الگوریتم k-means محبوب می‌باشد. این الگوریتم با توزیع نرمال کار می‌کند. (Mitchell,1997)

۲. الگوریتم‌های سلسله‌مراتبی

نوع دیگری از الگوریتم‌های خوشه‌بندی است که در ابتدا با در نظر گرفتن برخی معیارها به تجزیهٔ سلسله مراتبی داده‌ها می‌پردازد و سپس با روش‌های اجماع و تقسیم تغییراتی در دسته‌بندی اولیه ایجاد می‌نماید. الگوریتم‌های BIRCH و CHRE در این گروه جای دارند. الگوریتم‌های سلسله مراتبی به دو گونه تقسیم می‌شوند. الگوریتم‌های تجمیعی (پایین به بالا) و تقسیمی (بالا به پایین). در خوشه‌بندی تجمیعی، کار با خوشه‌هایی با یک داده شروع می‌شود (تعداد خوشه‌ها در ابتدا به اندازه‌ی تعداد داده‌های موجود می‌باشد). در هر مرحله دو یا چند خوشهٔ مناسب با هم ترکیب شده و خوشهٔ جدیدی را بوجود می‌آورند. در خوشه‌بندی تقسیمی عمل خوشه‌بندی با یک خوشه شروع می‌شود. این خوشه به صورت بازگشتی به دو یا چند خوشه تقسیم می‌گردد و به همین ترتیب عمل خوشه‌بندی ادامه پیدا می‌کند.

برای هر دو نوع از الگوریتم‌های بالا ما نیاز به‌یک شرط پایانی داریم این شرط اغلب رسیدن به k خوشه می‌باشد.

[1] Expectation-Maximization

ادغام یا تقسیم خوشه‌ها به شباهت یا عدم شباهت عناصرخوشه‌ها وابسته است یکی از مهمترین ملاک‌های شباهت، فاصله‌ی بین عناصر دو خوشه باشد. یعنی، فاصله‌ی دو زیر مجموعه از یک خوشه (برای هر ترکیب دوتایی از عناصر آن زیرمجموعه از خوشه‌ها) محاسبه می‌گردد.

۳. روش‌های متکی برچگالی

یکی از روش‌های خوشه‌بندی است که مجموعه‌ی داده‌ها را بر اساس معیارهایی همچون توابع همسایگی و چگالی دسته‌بندی کرده، مورد ارزیابی قرار می‌دهد.

OPTICS، CLIQUE و DBScon از جمله مثال‌های این نوع خوشه‌بندی می‌باشند. این الگوریتم‌ها از لحاظ تشکیل شکل‌های نامنظم الگویی، انعطاف پذیرتر هستند ولی اغلب فقط بر روی داده‌های عددی و با ابعاد پایین کار می‌کنند.

۴. روش‌های متکی بر گرید

در این متدها، داده‌ها ابتدا خلاصه شده و سگمنت می‌شوند، سپس عمل افراز روی فضاهای بدست آمده انجام می‌پذیرد و نهایتا فضاهای خوشه‌بندی شده منجر به داده‌های خوشه‌بندی شده می‌شود. هدف از این عملیات بالا بردن کارایی می‌باشد چون دیگر لازم نیست که با کل فضاهای داده‌ای کار کنیم و فقط باید به سگمنت‌ها کار کنیم. بعد از بدست آمدن سگمنت‌ها، کل کار به همان صورت قبلی دنبال می‌شود. یعنی می‌توان این متدها را معادل سایر متدها منتها با یک مرحله‌ی پیش‌پردازش در نظر گرفت.

۵. روش‌های متکی بر مدل

یک نوع از روش‌های خوشه‌بندی است که برای هر خوشه، مدلی فرضی را در نظر می‌گیرد و هدف آن یافتن مناسب‌ترین مدل برای هر خوشه می‌باشد. Cod Web ، Denclue وAutoClass مثال‌هایی از این نوع روش خوشه‌بندی می‌باشند.

۳-۳-۳ الگوریتم‌های کاوش قوانین انجمنی

مفاهیم پایه در یافتن قوانین وابستگی یکسان می‌باشد. الگوریتم‌های متفاوتی برای این‌کار پیشنهاد شده‌اند که خروجی همه‌ی آن‌ها یکسان می‌باشد. (رحمانی، میبدی، ۱۳۸۸)

- الگوریتم Apriori

هدف در این الگوریتم، پیدا کردن بزرگ‌ترین مجموعه آیتم است که حداقل Support و Confidence را رعایت کند. دو فرض زیر در این الگوریتم در نظر گرفته می‌شود:

۱ .هر زیر مجموعه از یک مجموعه آیتم تکرار شونده، تکرار شونده است. (یعنی اگر فرضاً مجموعهٔ {a,b,c} تکرار شونده باشد، آنگاه مجموعهٔ {a,b} نیز تکرار شونده است).

۲ .هر فوق مجموعه از یک مجموعه آیتم تکرار نشونده، است. (یعنی اگر فرضاً مجموعهٔ {a,b}تکرار شونده نباشد، آنگاه مجموعهٔ {a,b,c} نیز تکرار شونده نیست).

الگوریتم Apriori به این صورت است که در هر بار، یک سری مجموعه آیتم بزرگ با طول K +۱ را از روی مجموعه آیتم های کاندید با طول K می‌سازد و این کار را تا رسیدن به یک مجموعه آیتم با بیشترین طول انجام می‌دهد. مجموعه آیتم های کاندید در هر دفعه با ضرب مجموعه کاندید در خودش به دست می‌آید. از مشکلات این روش می توان به حجم بسیار بالای تراکنش های موجود در پایگاه داده، طولانی بودن زمان جست و جوی آنها در هر بار و تعداد زیاد کاندیدها در هر مرحله اشاره کرد.

- DHP

این روش مشابه با الگوریتم Apriori بوده و تنها تفاوت آن در ایجاد مجموعه کاندید در هر مرحله است. در روش Apriori مجموعه کاندید، با ضرب مجموعه آیتم بزرگ به دست آمده تا این مرحله در خودش، به وجود آمد. اما در روشDHP برای ساخت مجموعه کاندید در هر مرحله ، از یک جدول hash استفاده می شود و تنها یک سری از مجموعه آیتم های موجود در حاصلضرب به عنوان مجموعه کاندید پذیرفته می شود. (مجموعه آیتم هایی که دارای Support بالاتری هستند). الگوریتم DHP با استفاده از کاهش تعداد کاندیدها، الگوریتم Apriori را بهبود می‌بخشد.

فصل۴- روش تحقیق:
طراحی و پیاده‌سازی سیستم توصیه‌گر پیشنهادی

۴-۱ مقدمه

وقتی که مؤسسات از داده کاوی برای تشخیص دانشجویان تحت ریسک استفاده می کنند، می‌توانند از شکست و حذف آنها جلوگیری کنند، قبل از اینکه حتی دانشجویان خودشان از اینکه تحت ریسک هستند مطلع باشند. ساخت یک مدل داده کاوی در این مورد نیازمند دانش کافی در حوزه گذراندن دروس و شناخت انواع فراگیران دارد. کاراترین روش برای افزایش سطح فراگیران، آماده‌سازی و کمک به افرادی است که احتمال بیشتری برای گذراندن دروس دارند و همچنین شناسایی و ارائهٔ مشاوره به فراگیرانی که در بدو ورود و در ادامه با افت تحصیلی مواجه می شوند که این پدیده در نهایت منجر به رد آنها از آزمون نهایی می‌گردد.

در این فصل، یک مدل تلفیقی جهت بهبود عملکرد سیستم‌های توصیه‌گر در محیط‌های آموزش الکترونیک ارائه می‌شود. هدف از طراحی مدل پیشنهادی، افزایش عملکرد سیستم‌های توصیه‌گر با توجه به وضعیت و پیشرفت تحصیلی هر فراگیر می‌باشد. در ابتدا معماری سیستم پیشنهادی ارائه شده و مراحل آن بیان می‌شود. سپس جزئیات مربوط به پیاده‌سازی مورد بررسی قرار می‌گیرد. در نهایت، مجموعه داده و نرم افزار مورد استفاده معرفی و جزئیات پیاده سازی فازهای مختلف سیستم، شرح داده می‌شود.

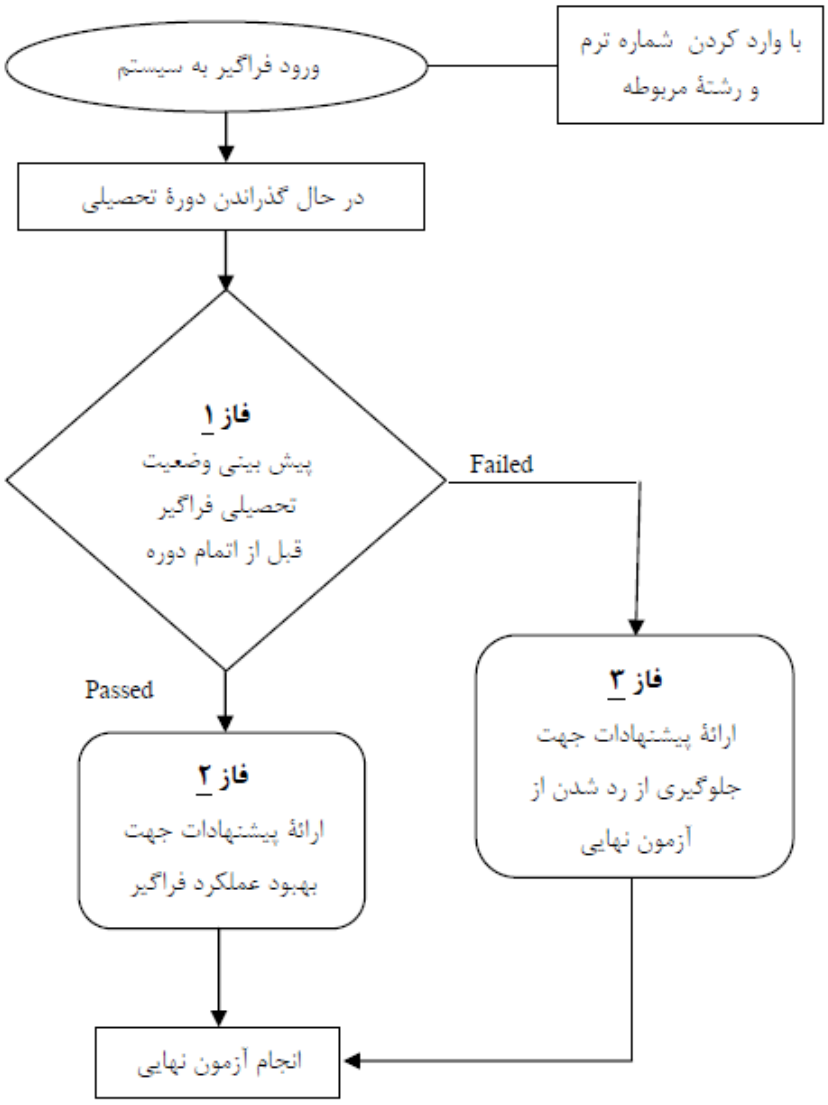
۴-۲ فرآیند و روش سیستم توصیه‌گر پیشنهادی

با توجه به کاستی هایی که در سیستم‌های توصیه‌گر و پیش‌بینی کنندهٔ مطالب و مفاد آموزشی در حوزهٔ آموزش الکترونیک مشاهده می‌گردد، لزوم طراحی سیستم‌های توصیه‌گر بهینه جهت مواجهه با سیل عظیم اطلاعات و هدایت فراگیران، در چنین محیط‌هایی ضروری به نظر می‌رسد. بر همین اساس در این پژوهش، سیستم توصیه‌گر بهینه‌ای طراحی و ارائه گردیده است که از مزایای چندین

روش معرفی شده در فصل گذشته بهره می‌برد. در سیستم توصیه‌گر ترکیبی پیشنهاد شده، روش سیستم، از ترکیب روش مبتنی بر الگوهای رفتاری فراگیران و نیز استفاده از تکنیک‌های داده‌کاوی در دسته‌بندی فراگیران و نیز تشخیص وابستگی دروس و فعالیت‌های مختلف در قبولی یک فراگیر، تشکیل شده است.

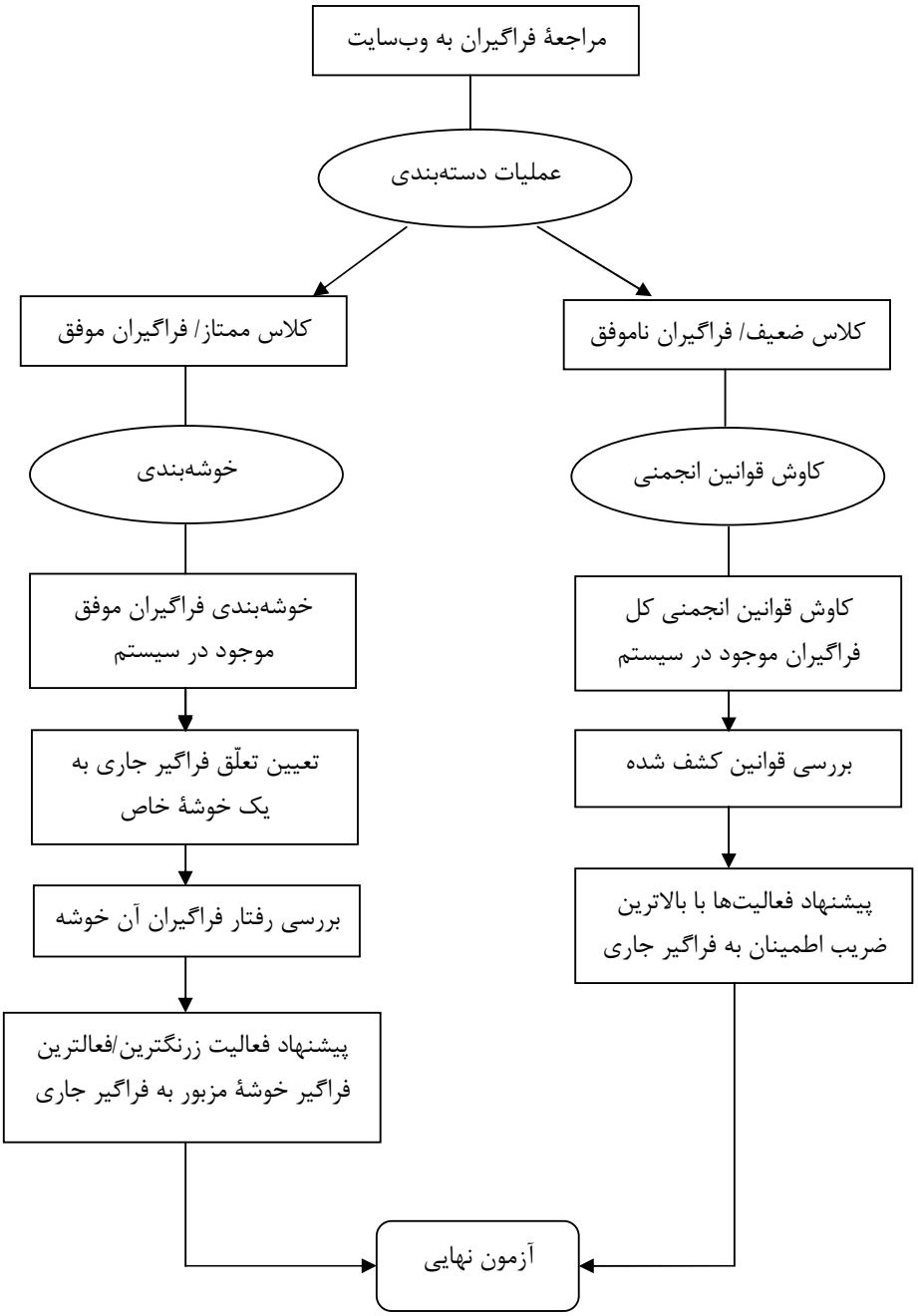
این سیستم با تقسیم بندی فراگیران به گروه‌های مختلف، به بررسی عملکرد و رفتار آنها در سایت پرداخته و سپس به توصیهٔ مطالب و یا آزمون‌های مناسب به وی می‌پردازد. مزیت این روش در دسته‌بندی فراگیران به لحاظ سطح علمی‌اشان می‌باشد. بدین ترتیب پس از ورود فراگیر جدید، سیستم قادر به پیش بینی نتیجهٔ نهایی وی قبل از اتمام دورهٔ تحصیلی‌اش، با توجه به رفتار وی و عملکرد فراگیران مشابه و موجود در سیستم خواهد بود. توانایی تعداد گروه‌ها و کلاسه‌ها در عملیات دسته‌بندی فراگیران متفاوت بوده، اگرچه با مقایسهٔ نتایج تعداد گروه‌های مختلف و با توجه به مجموعه دادهٔ[1] موجود، استفاده از دو گروه ضعیف و قوی (موفق و ناموفق)، از دقت بالاتری برخوردار است. در صورتی که سیستم تشخیص دهد فراگیر قوی بوده و قادر به اتمام موفقیت آمیز دوره می‌باشد، از یک رویکرد و در صورت تشخیص ضعف و یا ناموفق بودن فراگیر، از یک رویکرد دیگر که خاص فراگیران ضعیف می باشد، استفاده می‌نماید.

[1] Data set



شکل ۴-۱ نمای کلی از فازهای سیستم توصیه‌گر طراحی شده

معماری سیستم پیشنهادی در شکل ۴-۱ نشان داده شده است. همان طور که در شکل مشاهده می‌شود، سیستم شامل سه فاز است که در ادامه به شرح مراحل و فرضیه‌های ایجاد شده در هر فاز می‌پردازیم. شکل ۴-۲ نیز نمای کلی از نحوهٔ عملکرد سیستم توصیه‌گر پیشنهادی و تکنیک‌های مورد استفاده در هر فاز را نمایش می‌دهد.



شکل ۴-۲ نمای کلی از نحوهٔ عملکرد و تکنیک‌های مورد استفادهٔ سیستم توصیه‌گر طراحی شده

همانطور که در شکل دیده می‌شود، در هر فاز سیستم پیشنهادی، از یکی از تکنیک‌های داده‌کاوی استفاده خواهیم کرد. که عبارتند از دسته‌بندی، خوشه‌بندی و کاوش قوانین انجمنی.

دانش داده‌کاوی دارای ابزارهای قدرتمندی است که با بکارگیری این ابزارها مدل‌هایی را ارائه می‌کند که می‌تواند به توصیف، اکتشاف، کنترل و پیش بینی وقایع بپردازد. به طور کلی تکنیک‌های

داده‌کاوی فهم عمیقی از الگوهایی را که قبلاً ناشناخته بودند ارائه می‌دهند. در این پژوهش از تکنیک‌های داده‌کاوی نظارتی و غیر نظارتی بهره می‌بریم. (حاتم لو، هاشمی نژاد، ۱۳۸۷)

داده کاوی نظارتی[1] در شرایطی به کار می‌رود که رکوردها خروجی شناخته شده‌ای دارند. به عنوان مثال یک پایگاه داده مربوط به فراگیران، شامل رکوردهای دانشجویانی است که دروسشان را گذرانده‌اند و در آزمون نهایی قبول و یا مردود شده اند. داده‌کاوی نظارتی در اینجا از طریق اتصال الگوهای رفتاری به سوابق تحصیلی و سایر اطلاعات ذخیره شده، برای مطالعهٔ رفتار هر دو گروه به کار می رود. پس از ساخت مدل نهایی می‌توان از آن، جهت پیش‌بینی رفتار دانشجویان جدید و احتمال قبولی آنها استفاده کرد. تکنیک‌های پیش‌بینی احتمال خروجی‌های مختلف از فراگیران مانند ماندگاری و موفقیت آنها در آزمون نهایی را برآورد می‌کنند. بعلاوه، پیش بینی در داده‌کاوی به سیستم‌های آموزشی این امکان را می‌دهد که با کمک اطلاعات دریافتی از تعداد فراگیرانی که درس خاصی را اخذ می کنند یا تعداد فراگیرانی که در درس خاصی موفق می شوند، قبل از تخصیص منابع به درستی عمل نمایند و یا با شناخت الگوهای فراگیران مشروطی و یا فراگیرانی که موفق به گذراندن دروس خاصی نخواهند شد قبل از اینکه فراگیری با مشکل مواجه شود، اقدامات لازم را به عمل آورند.

داده‌کاوی غیر نظارتی نیز در شرایطی به کار می‌رود که الگوها یا گروه‌های ویژه ناشناخته‌اند. به عنوان مثال، در پایگاه داده انتخاب واحد دانشجویان اطلاعات بسیار کمی در مورد اینکه کدام دروس معمولاً با هم اخذ می شوند، یا اینکه چه نوع دروسی با چه نوع دانشجویانی مرتبط هستند وجود دارد. داده‌کاوی غیر نظارتی معمولاً در چنین حالاتی در ابتدای فرآیند مدل‌سازی به کار می‌رود تا الگوهایی را که از قبل پنهان می‌باشند کشف کند و از این طریق به فهم، توصیف و دسته‌بندی داده‌ها قبل از اعمال فرضیه‌ها کمک کند و تحلیل جامعی از خصوصیات دانشجویان ارائه دهد.

آنالیز قوانین همبستگی یکی از مهمترین تکنیک‌های داده کاوی محسوب می‌شود که به کشف قوانین و وابستگی‌ها بین عوامل و وقایع می پردازد. آنالیز خوشه نیز یکی از رایج ترین تکنیک‌های داده‌کاوی محسوب می شود که مجموعه داده را به خوشه‌های مختلفی تقسیم می‌کند بطوریکه بیشترین شباهت در خود خوشه‌ها و بیشترین تفاوت بین خوشه‌ها وجود دارد. خروجی تحلیل کلاستر عموماً به عنوان ورودی سایر تکنیکهای داده کاوی استفاده می گردد.

[1] Supervised Modelling

۴-۳ فازهای سیستم توصیه‌گر پیشنهادی

اساس کار فاز اول (پیش بینی وضعیت یک فراگیر قبل از اتمام دورهٔ تحصیلی‌اش)، دسته بندی می‌باشد. در فاز دوم از تکنیک خوشه بندی به منظور گروه‌بندی فراگیران مشابه استفاده خواهیم نمود. و در نهایت در فاز سوم با به کارگیری تکنیک کاوش قوانین انجمنی و با توجه به شناختی که از دسترسی های فراگیران پیدا نمودیم، اقدام به پیشنهاد فعالیت‌های مناسب به فراگیر جاری می‌نماییم.

۴-۳-۱ فاز اول: دسته‌بندی فراگیران

دسته‌بندی فراگیران و همینطور یافتن روندها و الگوهای پنهان در داده‌های سیستم آموزشی به منظور پی بردن به وضعیت آموزشی، پیدا کردن فراگیران تحت ریسک، ایجاد راهکارها و ارائهٔ مشاوره‌های صحیح به فراگیران در جهت بهبود وضعیت آینده‌اشان و مسائلی از این دست، از اهداف مورد انتظار در این پژوهش می‌باشد.

در این فاز از تکنیک‌های مختلف دسته بندی به منظور تحلیل معدل نهایی فراگیران و پیش‌بینی آن استفاده گردیده است. که در ادامه با بررسی نتایج حاصل از آنها به مقایسه این مدل ها می‌پردازیم و مناسبترین مدل با توجه به مجموعه دادهٔ موجود انتخاب می‌گردد. در حقیقت تکنیک‌های مختلف دسته‌بندی مانند شبکه‌های عصبی، درخت تصمیم، رگرسیون و غیره به پیدا کردن الگوها و دانش نهفته در داده‌های سیستم آموزش کمک می‌کنند. می‌توان مسیر تحصیلی فراگیر و وضعیت وی در نیمسال‌های بعدی (مشروطی، ممتازی) را به منظور تسهیل اقدامات آموزشی پیش‌بینی کرد. همچنین با کشف روند تحصیلی فراگیران موجود در سیستم، آنرا برای فراگیران جدید و در حال تحصیل به کار بست. در فاز دسته بندی فرضیه‌های مختلفی ایجاد می شود که در ادامه و در مرحلهٔ پیاده‌سازی، به آنها پاسخ می‌دهیم.

۴-۳-۲ فاز دوم: نحوهٔ برخورد با فراگیران موفق

در این فاز فراگیران موفق موجود در سیستم، با استفاده از تکنیک خوشه بندی به خوشه‌های مجزا تقسیم می‌شوند. با استفاده از این تکنیک می‌توان فراگیران را بر اساس ویژگی‌های مختلفی به عنوان مثال میزان فعال بودن آنها (انجام تکالیف، حضور در سایت، انجام تست‌ها و غیره) گروه‌بندی کرد و با هر گروه از آنها بصورت مناسبی رفتار کرد. پس از ورود فراگیر جدید به سیستم و پیش‌بینی موفقیت وی، تعلّق فراگیر به یک خوشهٔ خاص مشخص می‌گردد. جزئیات رفتار و عملکرد آن خوشه مورد

بررسی قرار می گیرد. به طور نمونه اگر فراگیر متعلق به خوشهٔ شمارهٔ سه است، برخی از صفات آن خوشه مورد بررسی قرار می‌گیرد. در نهایت فعالیت فعال ترین و زرنگترین فراگیر موجود در آن خوشه به وی پیشنهاد می‌گردد. همانطور که ذکر گردید در سیستم‌های توصیه‌گر موجود در سایر حوزه‌ها و همچنین آموزش الکترونیک فعالیت و عملکرد کاربران کل خوشه مربوط به کاربر جاری پیشنهاد می‌شود. به این دلیل که یکی از ایده‌های تأکید شده در فلسفهٔ پیاده‌سازی سیستم‌های توصیه‌گر، برابری میان کاربران است. یک سیستم توصیه گر متداول از رتبه‌ها و انتخاب‌های کاربران با بیشترین شباهت به کاربر فعال، به منظور ایجاد پیشنهادات استفاده می‌نمایند. هیچ دلیلی وجود ندارد که در پیشنهاد یک فیلم، کتاب، وبلاگ و غیره، کاربری را شایسته تر و واجد شرایط تر از سایر کاربران بدانیم. اگرچه این وضعیت در برخی حوزه ها مانند محیط‌های آموزش الکترونیک صدق نمی‌کند. زیرا در این حوزه، ایجاد تمایز میان فراگیران معمولی، ممتاز و با دانش بالاتر امکان پذیر می باشد.

در این فاز از رویکرد مبتنی بر مشارکت جمعی استفاده می‌گردد که با تجزیه و تحلیل آماری اطلاعات و یا استخراج داده‌های فراگیران موفق، رفتار گذشته و سایر اطلاعات، یک محدودهٔ همسایگی از افراد با تخصص و سطح دانش مشترک ایجاد نموده و سپس با یافتن نزدیکترین همسایه‌ها برای هر فراگیر به توصیهٔ انتخاب‌های این همسایگان به فراگیر هدف می‌پردازد. که البته به منظور بالا بردن کارایی و بهبود وضعیت آموزشی فراگیران در این طرح پیشنهاد می‌شود که عملکرد بهترین و زرنگترین فراگیر موجود در خوشهٔ مربوطه، به فراگیر جاری ارائه گردد.

۴-۳-۳ فاز سوم: نحوهٔ برخورد با فراگیران ناموفق

وقتی که مؤسسات از داده کاوی برای تشخیص فراگیران تحت ریسک استفاده می‌کنند، می‌توانند از شکست و حذف آنها جلوگیری کنند، قبل از اینکه حتی خودشان از اینکه تحت ریسک هستند مطلع باشند. با استفاده از تکنیک‌های داده‌کاوی، قادر به تشخیص فراگیران با عملکرد پایین خواهیم گردید. داده‌های در اختیار شامل اطلاعات دموگرافیک فراگیران و داده‌های مربوط به عملکرد آنها در سایت قبل از آزمون نهایی می‌باشند.

همانطور که ذکر گردید در مرحلهٔ اول فراگیران ضعیف بالقوه مشخص می شوند. مرحلهٔ اول یک فرآیند دسته بندی می باشد که در فاز نخست توضیح داده شد. متدلوژی فاز سوّم به این صورت می‌باشد که از تکنیک‌های کاوش قواعد انجمنی جهت کشف روابط میان فعالیت‌ها و ترتیب‌اشان با قبولی فراگیر استفاده می‌گردد. بنابراین در سمت چپ این قوانین معدل نهایی به عبارتی وضعیت

تحصیلی موفقیت آمیز در آزمون نهایی و در سمت راست آن فعالیت‌ها، تکالیف آموزشی و ترتیب آزمون‌ها قرار می‌گیرد. بدین ترتیب فعالیت‌های آموزشی به منظور ارتقاء آموزش الکترونیکی از طریق پیشنهاد دروس و یا جهش‌هایی به برخی منابع مرتبط، به عنوان راهنمایی جهت گذراندن مواد درسی به فراگیران ضعیف توصیه می‌گردد. به طور نمونه پیشنهاد فعالیت‌هایی مانند مطالعهٔ دروس خوانده نشده، ترتیب انجام تکالیف و آزمون‌ها و یا افزایش میزان تعامل با استادان به وی ارائه می‌گردد.

در این پژوهش با استفاده از آنالیز قوانین همبستگی، عملکرد کل فراگیران مجموعه داده، مورد تحلیل قرار گرفته و ارتباطات و همبستگی‌های بین دروس و فعالیت‌های آموزشی با موفقیت و قبولی فراگیران کشف گردید. روابط و قوانین کشف شده به سیستم در خصوص شناخت عوامل مؤثر و همچنین ارائهٔ پیشنهادات به فراگیرانی که دچار افت تحصیلی شده و قادر به اتمام موفقیت آمیز دوره نبوده، کمک شایانی می‌کند. بدیهی است که در این فاز استفاده از تکنیک خوشه‌بندی راه حل مناسبی نمی‌باشد زیرا در این صورت فراگیران ضعیف که عملکرد مشابهی داشته‌اند در خوشه‌هایی مجزا قرار گرفته و عملکرد کل خوشه به فراگیر هدف پیشنهاد می‌گردد. از آنجائیکه بدنبال روشی جهت بهبود وضعیت تحصیلی فراگیران در محیط آموزش الکترونیکی هستیم پیشنهاد عملکرد و رفتار فراگیران ضعیف به فراگیر جدیدی که دچار افت تحصیلی شده است، راه حلی منطقی نمی‌باشد.

۴-۴ جزئیات پیاده‌سازی

پس از بررسی جنبه‌های نظری و تکنیک‌های مختلف سیستم توصیه‌گر طراحی شده، و همچنین با استفاده از الگوریتم‌های شرح داده شده در فصل قبل، حال برای انجام فرایند توصیه و نیز امکان ارزیابی آن، به پیاده‌سازی بخش‌های مختلف این سیستم پیشنهادی می‌پردازیم.

۴-۴-۱ نرم افزار داده‌کاوی مورد استفاده

با بررسی تعدادی از نرم افزارهای رایج و پرکاربرد داده‌کاوی نظیر DBMiner، IBM ، Clemetine Intelligent Miner، Microsoft SQL Server 2000، XLMiner، SAS Enterprise Miner و غیره و با مقایسهٔ میزان توانایی‌ها، تعداد و انواع تکنیک‌های مورد پشتیبانی در هر نرم افزار، الگـوریتم های پیاده سـازی شده، روش‌های ارزیابی نتایج، روش‌های مصـور سـازی، واسط‌های کاربر پسند، پلت‌فرمهای سازگار برای اجرا و در دسترس بودن نرم افزار، در نهایت نرم افزار Weka با داشتن امکانات بسیار گسترده، امکان مقایسه خروجی روش‌های مختلف با هم، راهنمای خوب، واسط

گرافیگی کارا، سازگاری با سایر برنامه‌های ویندوزی، و از همه مهم‌تر وجود کتابی بسیار جامع و مرتبط با آن، به عنوان نرم افزار اصلی مورد استفاده در این پژوهش انتخاب گردید(Data Mining, witten et Al., 2005).

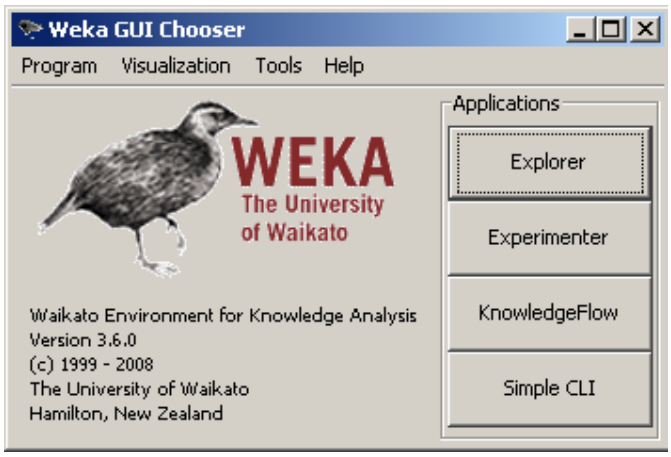
این نرم‌افزار به گونه‌ای طراحی شده است که می‌توان به سرعت، روش‌های موجود را به صورت انعطاف‌پذیری روی مجموعه‌های جدید داده، آزمایش نمود. این نرم‌افزار، پشتیبانی‌های ارزشمندی را برای کل فرآیند داده کاوی های تجربی فراهم می‌کند. این پشتیبانی‌ها، آماده سازی داده‌های ورودی، ارزیابی آماری چارچوبهای یادگیری و نمایش گرافیکی داده‌های ورودی و نتایج یادگیری را در بر می‌گیرند. همچنین، هماهنگ با دامنه وسیع الگوریتم‌های یادگیری، این نرم‌افزار شامل ابزارهای متنوع پیش پردازش داده‌هاست. این جعبه ابزار متنوع و جامع، از طریق یک واسط متداول در دسترس است، به نحوی که کاربر می‌تواند روش‌های متفاوت را در آن با یکدیگر مقایسه کند و روش هایی را که برای مسایل مدنظر مناسب‌تر هستند، تشخیص دهد.

نرم‌افزار Weka در دانشگاه Waikato واقع در نیوزلند توسعه یافته است و اسم آن از عبارت «Waikato Environment for knowledge Analysis» استخراج گشته است. همچنین Weka، نام پرنده‌ای با طبیعت جستجوگر است که پرواز نمی کند و در نیوزلند، یافت می‌شود.

این سیستم به زبان جاوا نوشته شده و بر اساس لیسانس عمومی و فراگیر GNU[1] انتشار یافته است. Weka تقریباً روی هر پلت فرمی اجرا می‌شود و نیز تحت سیستم عامل‌های لینوکس،ویندوز، و مکینتاش، و حتی روی یک منشی دیجیتالی شخصی، آزمایش شده است.

این نرم افزار، یک واسط همگون برای بسیاری از الگوریتم‌های یادگیری متفاوت، فراهم کرده است که از طریق آن روش‌های پیش پردازش، پس از پردازش و ارزیابی نتایج طرح‌های یادگیری روی همه مجموعه‌های داده موجود، قابل اعمال است. شکل ۴-۳، راه‌های انتخاب واسط‌های مختلفWeka را نشان می‌دهد. آسانترین راه استفاده از Weka از طریق واسطی گرافیکی است کهExplorer خوانده می‌شود. این واسط گرافیکی، به وسیله انتخاب منوها و پر کردن فرم‌های مربوطه، دسترسی به همه امکانات را فراهم کرده است. پیاده سازی الگوریتم‌های مختلف یادگیری را فراهم می کند و به آسانی می‌توان آنها را به مجموعه های داده خود اعمال کرد.

[1] GNU is Not Unix



شکل ۴-۳ وضعیت انتخاب واسط در weka

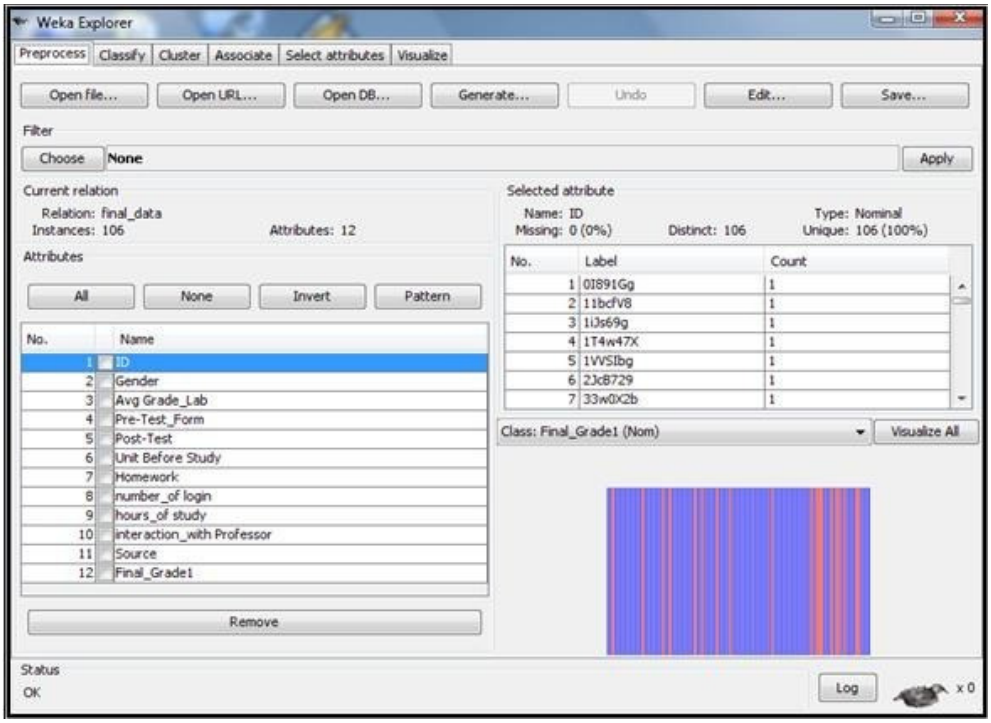
همچنین، این نرم افزار شامل مجموعه متنوعی از ابزارهای تبدیل مجموعه‌های داده‌ها، همانند الگوریتم‌های گسسته سازی می باشد. در این محیط می‌توان یک مجموعه داده را پیش پردازش کرد، آن را به یک طرح یادگیری وارد نمود، و دسته‌بندی حاصله و کارآیی‌اش را مورد تحلیل قرار داد. (همه این کارها، بدون نیاز به نوشتن هیچ قطعه برنامه‌ای میسر است). این محیط، شامل روش‌هایی برای همه مسایل استاندارد داده کاوی مانند رگرسیون، رده‌بندی، خوشه‌بندی، کاوش قواعد انجمنی و انتخاب ویژگی می باشد. با در نظر گرفتن اینکه، داده‌ها بخش مکمل کار هستند، بسیاری از ابزارهای پیش پردازش داده‌ها و مصورسازی آنها فراهم گشته است. همه الگوریتم ها، ورودی‌های خود را به صورت یک جدول رابطه‌ای به فرمتARFF[1] دریافت می‌کنند. این فرمت داده‌ها، می‌تواند از یک فایل خوانده شده یا به وسیله یک درخواست از پایگاه دادهای تولید گردد.

یکی از راه‌های به کارگیری Weka، اعمال یک روش یادگیری به یک مجموعه داده و تحلیل خروجی آن برای شناخت چیزهای بیشتری راجع به آن اطلاعات می‌باشد. راه دیگر استفاده از مدل یادگیری شده برای تولید پیش‌بینی‌هایی در مورد نمونه‌های جدید است. سومین راه، اعمال یادگیرنده‌های مختلف و مقایسه کارآیی آنها به منظور انتخاب یکی از آنها برای تخمین میباشد. علاوه بر موارد فوق، Weka شامل پیاده سازی الگوریتم‌هایی برای یادگیری قواعد انجمنی، خوشه‌بندی داده‌ها در جایی که هیچ دست‌های تعریف نشده است، و انتخاب ویژگی‌های مرتبط در داده‌ها می‌باشد.

زمانی که Weka فعال می‌شود، امکان انتخاب بین چهار واسط کاربری وجود دارد: Explorer، knowledge، Experimenter و واسط خط فرمان. اکثر کاربران، حداقل در ابتدای کار

[1] Attribute-Relation File Format

Explorer را به عنوان واسط کاربری انتخاب می‌کنند. شکل ۴-۴، نمای Explorer را نشان می‌دهد. در این واسط، شش پانل مختلف وجود دارد که از طریق نوار بالای صفحه قابل انتخاب هستند و با وظایف داده‌کاوی پشتیبانی شده توسط Weka متناظر می‌باشند.



شکل ۴-۴ واسط گرافیکی Explorer

به طور خلاصه، کارکرد تمام گزینه‌ها به شرح ذیل است.

- Preprocess انتخاب مجموعه داده و اصلاح آن از راه‌های گوناگون؛
- Classify آموزش برنامه‌های یادگیری که رده‌بندی یا رگرسیون انجام می‌دهند و ارزیابی آنها؛
- Cluster یادگیری خوشه‌ها برای مجموعه‌های داده؛
- Associate یادگیری قواعد انجمنی برای داده‌ها و ارزیابی آنها؛
- Select attributes انتخاب مرتبط‌ترین جنبه‌ها در مجموعه‌های داده؛
- Visualize مشاهده نمودارهای مختلف دوبعدی داده‌ها و تعامل با آنها؛

۴-۴-۲ مجموعه داده‌ها

یکی از مشکلات بزرگ تحقیقاتی در حوزهٔ سیستم‌های توصیه‌گر و شخصی سازی محیط های آموزش الکترونیک، کمبود مجموعه داده‌های استاندارد می‌باشد. به دلیل مسأله خصوصی بودن و محرمانگی، معمولاً ثبت های وب سرورها و اطلاعات فراگیران در دسترس عموم قرار نمی‌گیرد. و تمامی مجموعه داده های موجود، متعلق به چندین سال پیش می‌باشند.

از سوی دیگر در تمامی مقالاتی که از مجموعه داده‌های فراگیران به منظور پیش بینی و بهبود مدل فراگیر استفاده کرده اند، داده‌ها مربوط به وب سایت دانشکده و دانشگاه نویسندگان مقالات و یا یک وب سایت تجارت الکترونیکی فروش محصولات بصورت برخط می‌باشد که هیچ یک از این دو مورد در دسترس عموم قرار ندارد.

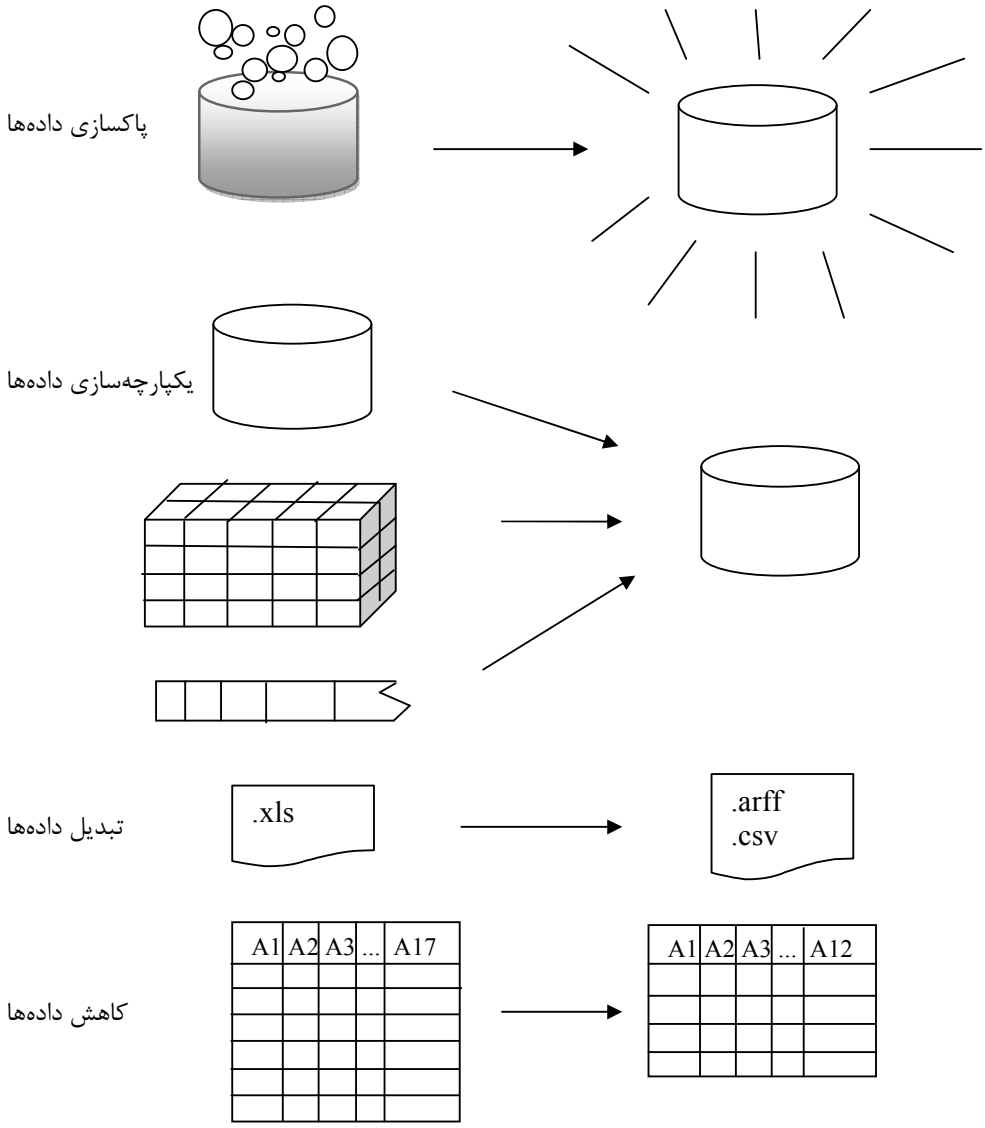
با توجه به محدودیت‌های ذکر شده، در این پایان نامه از اطلاعات ۱۰۶ فراگیر در یک سیستم آموزش الکترونیک به‌عنوان مجموعه داده استفاده شده است. این مجموعه، از یک مرکز داده[1] به آدرس اینترنتی *https://pslcdatashop.web.cmu.edu/* گرفته شده است. ثبت‌ها متعلق به فراگیران رشتهٔ ریاضی در درس جبر بوده که به مدت یک ترم در بهار سال ۲۰۰۷ جمع‌آوری شدند.

۴-۴-۲-۱ پیش پردازش داده‌ها[2]

وجود مسائلی نظیر ناقص بودن داده‌ها، ناسازگاری آنها و وجود ناخالصی‌هایی همچون خطاها، مقادیر تقریبی و مقادیر خارج از محدوده نرمال در پایگاه داده‌های واقعی، باعث کاهش کیفیت داده‌کاوی می شود. برای دستیابی به نتایج مطلوب تر، نیاز به داده‌های با کیفیت بالاتر وجود دارد.

پیش‌پردازش، گامی مهم در راستای داده‌کاوی موفقیت آمیز است. اعمالی که در پیش‌پردازش انجام می‌شوند عبارتند از حذف ناخالصی‌ها و اصلاح داده‌های نادرست، یکپارچه سازی داده‌ها، تغییر داده‌ها و کاهش داده‌ها. در اولین و کلی ترین دسته بندی، تکنیک‌های به کار رفته در پیش‌پردازش به چهار دسته کلی تقسیم می‌شوند. نمایش کلی از روش‌های پیش پردازش داده در شکل ۴-۵ نشان داده شده است. (بابائی، صفار یزدی و سرائی، ۱۳۸۷)

[1] Data center
[2] Preprocess



شکل۴-۵ روش‌های اصلی پیش پردازش داده‌ها

۴-۴-۲-۱-۱ پاکسازی داده‌ها[1]

اولین دسته از فعالیت‌های انجام گرفته، پاکسازی داده‌ها می‌باشد. پاکسازی داده‌ها بر مقدار دهی به ویژگی‌های فاقد مقدار، متعادل سازی مقادیر دارای نویز، شناسایی و حذف مقادیر خارج از محدوده[2] و رفع ناسازگاری داده ها متمرکز است. پاکسازی داده، به ویژه هنگام مجتمع سازی داده های ناهمگن مورد نیاز است و باید همراه با تبدیلات داده‌های مرتبط با الگو مورد استفاده قرار گیرد.

[1] Data cleaning
[2] Outlier

در مخازن داده، پایگاه‌داده‌های وابسته و سیستم‌های اطلاعاتی مبتنی بر وب عمومی، نیاز به پاکسازی داده به شدت افزایش می‌یابد. این افزایش به این علت است که منابع، به علت در بر گرفتن نمایش‌های مختلف از داده‌های یکسان، دارای اطلاعات افزونه می‌باشند. به منظور فراهم نمودن دسترسی به داده‌های صحیح و سازگار، ترکیب نمایش‌های مختلف داده و حذف اطلاعات تکراری، ضروری است. بسته به تعداد منابع داده، درجه ناهمگن بودن آنها و «کثیفی» داده ها، ممکن است اجرای چندین گام تبدیل و پاکسازی داده‌ها ضروری باشد.

الگوریتم‌های داده‌کاوی اغلب به خصوصیات ویژهٔ داده‌ها حساس هستند. لذا با بررسی اولیهٔ داده‌ها، تعدادی از فیلدهای در بردارندهٔ اطلاعات لازم برای شناسایی هویت فراگیر را حذف نمودیم. همچنین ستون‌هایی که با یکدیگر تغییر می‌کنند (مانند ترم تحصیلی با سال ورود و سن با تاریخ تولد)، بطور مثال ستون‌های تاریخ تولد و ترم تحصیلی، از مجموعه دادهٔ اصلی حذف شدند. با تحلیل‌های آماری، فیلدهای سن و سال ورود جزء داده‌های پرت محسوب شدند. لذا از ورود آنها برای آنالیز و کشف الگوها توسط تکنیک‌های داده‌کاوی اجتناب شده‌اند.

۴-۴-۲-۱-۲ یکپارچه‌سازی داده‌ها[1]

با توجه به اینکه در فرایند داده کاوی، داده های مورد بررسی از منابع مختلفی به دست می آیند، نیاز به یکپارچه سازی داده های این منابع در یک مجموعه داده کلی وجود دارد که این عمل، دومین بخش از فعالیت های پیش‌پردازش را تشکیل می‌دهد. مسئلهٔ یکپارچه‌سازی داده، به سه دلیل پیچیده است: اول اینکه منابع گوناگون، داده‌های همپوشان و وابسته به یکدیگر را در برمی‌گیرند. دوم، داده‌ها در مدل‌ها و طرح‌های مختلفی ذخیره می‌شوند، و سومین دلیل این است که منابع مختلف داده، قابلیت پردازش انواع متفاوتی از درخواست‌ها را دارند.

در این گام داده های ذخیره شده در جداول و فایل‌های مختلف، یکپارچه شده و در یک فایل اکسل واحد ذخیره می‌گردند.

مهمترین مزیت سیستم یکپارچه‌سازی، فراهم نمودن این امکان است که کاربران بدون نگرانی در مورد چگونگی به دست آوردن پاسخ‌ها، بر بخشی از داده‌ها که مورد نیاز آنها می‌باشد متمرکز شوند. در نتیجه از جستجوی منابع مرتبط داده، تعامل با هر یک از این منابع به صورت مجزا و ترکیب داده‌های منابع مختلف بی نیاز می‌گردند.

[1] Data integration

۴-۴-۲-۱-۳ تبدیل داده‌ها[1]

ممکن است داده های موجود برای فرایند داده‌کاوی مناسب نباشند. در این حالت، باید عمل تبدیل داده ها، روی آنها انجام شود. در نرم افزار weka، در ابتدا تنها فایل‌های با پسوند arff در مرورگر[2] فایل نمایش داده می‌شود. به منظور خواندن فایل مربوطه و تبدیل آن به فرمتARFF می‌بایست از یکی از دو روش ذیل استفاده نماییم:

الف) با توجه به اینکه فرمت مجموعه دادهٔ مورد نظر ما اکسل بوده، به منظور خواندن فایل‌ها می‌بایست ابتدا پسوند xls. را به csv. تبدیل نماییم. زیرا weka قادر به خواندن فایل های صفحه گسترده با پسوند CSV[3] می باشد. در این قالب هر رکورد از اطلاعات در یک سطر فایل ذخیره می‌شود. و در هر سطر هم فیلدها به وسیله کاراکتر کاما «،» از هم جدا می‌شوند.

ب) روش دیگر استفاده از نرم افزار تبدیل کنندهٔ CeDiyez به منظور تبدیل فرمت فایل‌ها با پسوند xls به پسوند arff می‌باشد.

۴-۴-۲-۱-۴ کاهش داده‌ها[4]

همانطور که قبلاً نیز گفته شد، یکی از دلایل عمده نیاز به انجام پیش پردازش، حجم بسیار بالای مجموعه داده‌هایی است که باید عمل داده‌کاوی روی آن انجام شو.د تحلیل این حجم بالای اطلاعات، نیازمند زمان زیادی است که عملاً داده کاوی را غیر ممکن می سازد.

کاهش داده، مبحث مهمی در زمینه داده کاوی است. هدف تکنیک‌های کاهش داده در داده‌کاوی، استخراج زیرمجموعه های کوچک از حجم انبوهی از مجموعه داده‌ها با حفظ خصوصیات داده‌های اصلی می‌باشد. اینکار باعث می شود عملیات سخت یا غیرممکن داده‌کاوی به صورت کارا و مؤثری انجام شوند.

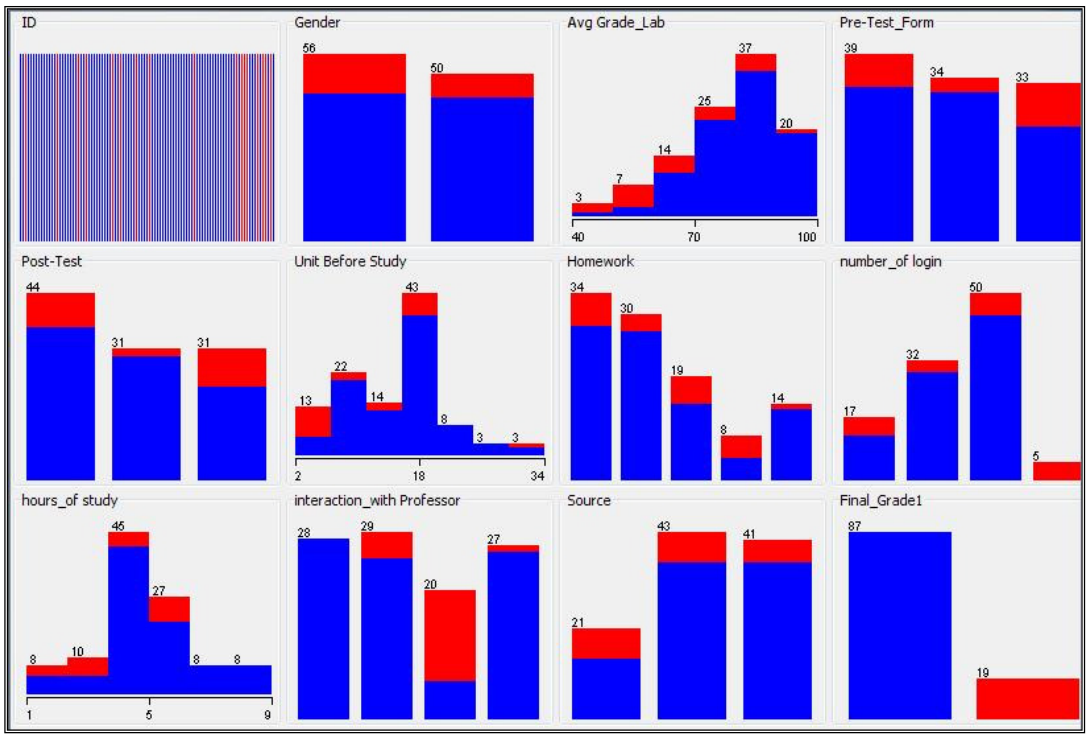
در این گام با توجه به حوزهٔ کارکردی مورد مطالعه در این تحقیق برخی از صفات و داده های نامربوط حذف می گردند. به طور مثال اطلاعاتی در خصوص محل زندگی فراگیر، سن، تاریخ ثبت نام در سایت و غیره. در ابتدا مجوعه دادهٔ مورد نظر شامل ۱۷ صفت بود که برخی از آنها به دلیل بی ربط بودن به این تحقیق حذف گردیدند. در نهایت ۱۱ صفت شرطی و یک صفت کلاس مورد بررسی قرار گرفت. که این صفات و شرح آنها را در جدول شماره ۴-۱ مشاهده می‌نماییم.

[1] Data transformation
[2] browser
[3] Comma Separated Value
[4] Data reduction

جدول ۴-۱ شرح صفات مجموعه دادهٔ موجود

صفت	شرح	مقادیر ممکن
ID	شماره دانش آموزی	{0I891Gg,…, zW2b0zD}
Gender	جنیست فراگیران	M,F
Avg Grade	میانگین نمرات ترم قبل	{40,…,100}
Pre test form	فرم آزمون نخست	A,B,C
Post test form	فرم آزمون دوم	A,B,C
Unit before study	آخرین فصل مطالعه شده	{3,4,5,6,…,30,31,32,33,34}
Homework	انجام تکالیف	very low, low, middle, high, very high
Number of Login	تعداد دفعات ورود به سیستم	very low, low, medium, high
Hours of study	مدت حضور در هر ورود (ساعت)	1,2,3,4,5,6,7,8,9
Interaction with Professor	میزان تعامل با استاد	low, medium, high, very high
Source	منبع مطالعه دروس	Site, e-book, both (site &e-book)
Final Grade	نمره نهایی	Passed, failed

توزیع مقادیر مختلف مربوط به هر کدام از ویژگی‌ها نیز در شکل ۴-۶ نمایش داده شده است. این شکل از دوازده بخش که نمایانگر دوازده صفت (یازده صفت شرطی و یک صفت کلاس) است، تشکیل شده است. هر بخش نمایانگر فراوانی فراگیران در مقادر مختلف صفت مربوطه می‌باشد. به طور مثال اگر صفت Gender و یا جنسیت را در نظر بگیریم، از ۱۰۶ فراگیر، ۵۶ نفر آنها پسر و ۵۰ نفر دختر می‌باشند. رنگ قرمز به معنی فراگیران موفق و قوی که قادر به گذراندن دوره گشته‌اند و رنگ آبی فراگیران ضعیف را نشان می‌دهد.



شکل۴-۶ توزیع مقادیر مربوط به ویژگی‌های مختلف

در شکل ۴-۷ نیز نمونه‌ای از مقادیر مجموعه دادهٔ مورد استفاده نشان داده شده است.

Relation: final_data

No.	ID Nominal	Gender Nominal	Avg Grade_Lab Numeric	Pre-Test_Form Nominal	Post-Test Nominal	Unit Before Study Numeric	Homework Nominal	number_of login Nominal	hours_of study Numeric	interaction_with Professor Nominal	Source Nominal	Final_G Nom
1	0I891Gg	M	100.0	B	A	27.0	Middle	low	2.0	very_high	site	passed
2	11bcfV8	F	77.0	B	C	11.0	Middle	medium	5.0	medium	both	passed
3	1iJs69g	M	54.0	A	B	4.0	high	high	6.0	low	site	failed
4	1T4w47X	F	100.0	A	C	22.0	high	high	5.0	very_high	both	passed
5	1VVSIbg	M	91.0	B	A	18.0	Middle	medium	5.0	high	both	passed
6	2JcB729	F	76.0	A	C	10.0	high	high	5.0	high	e-book	passed
7	33w0X2b	M	60.0	C	B	6.0	Middle	medium	4.0	medium	site	passed
8	35jI12H	M	88.0	B	A	17.0	low	medium	5.0	high	both	passed
9	3cjD21W	F	100.0	A	B	20.0	Middle	high	4.0	very_high	e-book	passed
10	4gJnw14	F	74.0	A	B	8.0	Middle	high	5.0	medium	both	passed
11	53BUR...	M	79.0	B	A	13.0	very_low	high	1.0	high	e-book	passed
12	59v7i97	F	78.0	B	A	16.0	Middle	medium	5.0	high	e-book	passed
13	619gQzH	M	90.0	B	C	17.0	high	medium	6.0	high	e-book	passed
14	6zkEsmR	F	74.0	A	B	8.0	Middle	high	4.0	high	both	passed
15	7BJfpoV	M	100.0	B	A	34.0	Middle	low	3.0	high	both	failed
16	7gi78ec	M	86.0	B	A	17.0	high	medium	6.0	high	both	passed
17	7LZr10z	M	100.0	B	A	34.0	Middle	low	6.0	very_high	site	passed
18	7OalbuD	F	80.0	C	A	15.0	Middle	high	6.0	very_high	e-book	passed
19	883KL7w	F	75.0	B	C	8.0	high	low	7.0	medium	both	passed
20	8EWe...	M	74.0	A	B	9.0	high	medium	6.0	high	site	passed
21	9SHYrfx	F	54.0	A	C	4.0	Middle	high	5.0	medium	both	passed
22	A1MG...	M	74.0	A	C	9.0	Middle	high	5.0	high	both	passed
23	A6478k2	M	52.0	C	B	4.0	low	low	3.0	low	e-book	failed
24	a8YLu01	M	82.0	C	A	16.0	Middle	high	5.0	high	both	passed
25	axjI69v	F	98.0	B	A	21.0	Middle	low	4.0	very_high	e-book	passed
26	BU25p0d	M	60.0	A	B	5.0	very_low	high	6.0	low	e-book	failed
27	c71bR5y	F	90.0	B	A	17.0	Middle	medium	6.0	high	both	passed
28	C76m01z	M	76.0	B	A	16.1	low	low	5.0	very_high	site	passed
29	C8siiRn	F	86.0	B	C	18.0	very_high	high	8.0	high	both	passed
30	D49Uk8t	F	61.0	A	B	9.0	Middle	low	6.0	medium	e-book	passed
31	dIUTKe8	M	82.0	A	C	11.0	Middle	high	5.0	very_high	e-book	passed

شکل ۴-۷ نمونه ای از مقادیر مجموعه داده

۴-۴-۳ روش ارزیابی مورد استفاده

انتخاب فاکتورهای مؤثر برای ارزیابی روش‌های یادگیری ماشین اعمال شده یکی از مسایل مهم در انجام این پژوهش بوده است که ما از روش ارزیابی متقاطع[1] ده دسته استفاده کردیم که در آن میزان مقدار داده تست نگاه داشته شده ۱۰ درصد کل داده آموزشی است. در ارزیابی متقاطع برای تعیین میزان کیفیت یک طبقه بندی کننده مقداری از داده آموزشی به عنوان مجموعه تست استفاده می‌شود. به عنوان مثال ۹۰٪ مجموعه داده به عنوان داده آموزشی استفاده می‌شود و ۱۰٪ بقیه برای تست روش استفاده می‌شود. تعداد نمونه‌هایی که به طور صحیح طبقه بندی می‌شود به عنوان دقت آن روش لحاظ می‌شود. این فرآیند ده بار تکرار می‌شود و هر مرتبه با نگاه داشتن ۱۰٪ داده آموزشی اجرا می‌شود. هنگامی که این فرایند تکمیل می‌شود، میانگین مقادیر دقت مراحل به عنوان میزان دقت نهایی آن روش لحاظ می‌شود. (احمدی، شاکری اسکی و علیشاهی۱۳۸۷)

۴-۴-۴ پیاده‌سازی فازهای سیستم توصیه‌گر پیشنهادی

مدل پیشنهادی متشکل از سه فاز بوده که در فاز نخست فراگیران ضعیف و قوی از هم جدا می شوند. در واقع سیستم قبل از اتمام دورهٔ تحصیلی فراگیر، موفق و یا ناموفق بودن وی را پیش بینی می‌نماید. فاز دوم و سوم به بیان نحوهٔ برخورد با یک فراگیر موفق و یا ناموفق می‌پردازد و بدین ترتیب برای هر گروه از یک رویکرد خاص استفاده می‌گردد. هدف ارائهٔ پیشنهاداتی است که کارایی فراگیران را در سیستم آموزشی بالا ببرد.

۴-۴-۴-۱ پیاده‌سازی فاز نخست

همانطور که ذکر گردید، در فاز اوّل هنگامی که فراگیر وارد سیستم می‌شود، سیستم پیش‌بینی می‌کند که وی ضعیف است و یا قوی. در این فاز فرضیه های مختلفی ایجاد می شود که با استفاده از نرم افزار weka به آنها پاسخ می دهیم:

۱. اگر از دو کلاس passed و failed استفاده گردد، دقت پیش بینی بالاتر است یا هنگام استفاده از سه کلاس low، middle و high؟
۲. کدام الگوریتم دسته بندی از دقت بالاتری برخوردار است؟
۳. صفات مناسب برای عملیات دسته‌بندی کدامند؟

[1] Cross Validation

نحوهٔ کار بدین شکل است که ابتدا بایستی اطلاعات ورودی به مدل داده شود و سپس با اضافه شدن تکنیک ها به آن، مدل تکمیل شده و آماده دادن خروجی‌های متنوع می‌باشد. به منظور پاسخ به فرضیهٔ نخست فراگیران را در دو کلاس passed و failed و همچنین سه کلاس low، middle و high تقسیم بندی می‌نماییم. در جداول ۴-۲ و ۴-۳ درصد و تعداد فراگیران را در کلاس‌های متفاوت مشاهده می‌نماییم.

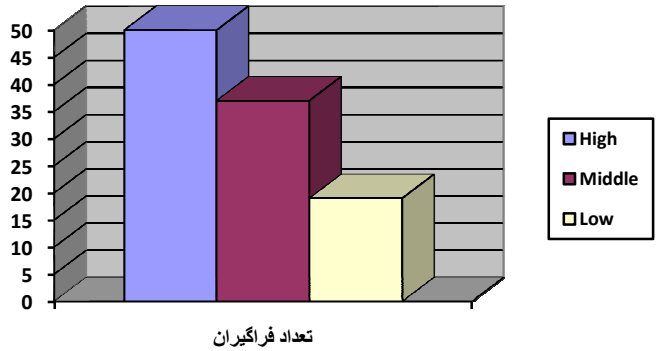
جدول ۴-۲ درصد و تعداد فراگیران در سه کلاس low، middle و high

کلاس	نمرهٔ نهایی	تعداد فراگیران	درصد
High	>= ۸۰	۵۰	۴۷.۱۷
Middle	۶۵< x< ۸۰	۳۷	۳۴.۹۱
Low	<= ۶۵	۱۹	۱۷.۹۲

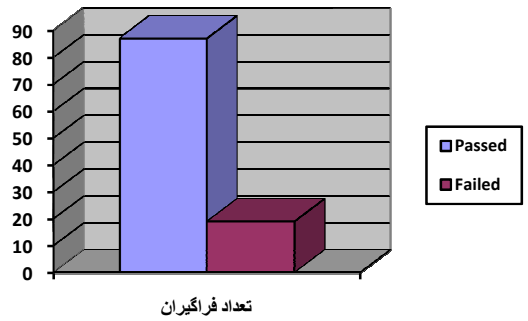
جدول ۴-۳ درصد و تعداد فراگیران در دو کلاس passed و failed

کلاس	نمرهٔ نهایی	تعداد فراگیران	درصد
Passed	>۶۵	۸۷	۸۲.۰۸
Failed	<= ۶۵	۱۹	۱۷.۹۲

همچنین شکل‌های ۴-۸ و ۴-۹ نمایانگر فراوانی فراگیران در کلاس‌های مذکور می‌باشد.



شکل۴-۸ نمودار فراوانی فراگیران در سه کلاس low، middle و high



شکل۴-۹ نمودار فراوانی فراگیران در دو کلاس passed و failed

۴-۴-۴-۱-۱ انتخاب تعداد کلاسه‌ها و الگوریتم مناسب

با استفاده از نرم افزار داده‌کاوی weka دقت پیش‌بینی را در دو حالت به کارگیری دو کلاس passed و failed، و همچنین به کارگیری سه کلاس low، middle و high محاسبه نمودیم. (جدول ۴-۴)

جدول ۴-۴ دقت پیش بینی الگوریتم‌های دسته بندی متفاوت در تعداد کلاسه‌های مختلف

		میزان دقت (درصد)	
الگوریتم دسته بندی		۲-کلاس	۳-کلاس
الگوریتم‌های درختی	REP Tree	۸۲.۰۷۵	۴۷.۱۷
	J48	۸۶.۷۹	۷۶.۴۱۵
	J48graft	۸۸.۶۸	۷۶.۴۱۵
	Random Forest	۸۲.۰۷۵	۶۷.۹۲
	Random Tree	۸۳.۹۶	۵۳.۷۷
	Simple Cart	۸۲.۰۷۵	۴۷.۱۷
الگوریتم‌های غیر درختی	OneR	۸۲.۰۷۵	۴۷.۱۷
	Bayes Net	۸۳.۹۶	۷۱.۷
	SimpleLogistic	۸۹.۶۲	۷۹.۲۴۵
	JRip	۸۹.۶۲	۷۴.۵۳
	IBk	۸۳.۹۶	۶۰.۳۸

همانطور که در جدول شماره ۴-۴ نشان داده شده است دقت پیش‌بینی دو کلاس، بدون در نظر گرفتن الگوریتم دسته بندی مورد استفاده، در تمام موارد بالاتر است. بنابراین توانایی تعداد دو کلاسه

ضعیف و قوی (موفق و ناموفق)، در عملیات دسته‌بندی فراگیران با توجه به مجموعه دادهٔ موجود، بیشتر بوده و از دقت بالاتری برخوردار است.

همچنین با مقایسه نتایج الگوریتم‌های مختلف دسته بندی موجود در این جدول، الگوریتم درختی J48graft و همچنین الگوریتم غیر درختی SimpleLogistic به عنوان مناسب‌ترین مدل انتخاب می‌گردد.

همانطور که اشاره گردید الگوریتم‌های درختی مزایای فراوانی دارند. بدلیل سادگی، خروجی قابل فهم و میزان صحت قابل قبول این الگوریتم‌ها، الگوریتم نهایی مورد استفاده جهت انجام پیش‌بینی‌های بعدی، الگوریتم درختی J48graft خواهد بود. که در فصل ۵ (فصل ارزیابی) مدل‌های ایجاد شده از الگوریتم‌های منتخب با یکدیگر مقایسه شده، نتایج آنها مورد تحلیل قرار می‌گیرد و نهایتاً دلیل استفاده از این الگوریتم با توجه به پارامترهای ارزیابی شرح داده می‌شود.

۴-۴-۴-۱-۲ انتخاب صفات و ویژگی‌های مناسب

حذف و نادیده گرفتن برخی صفات در پایگاه داده‌ها، جهت تصمیم گیری و تحلیل سریع و همچنین پیش‌بینی‌هایی با دقت بالاتر می تواند مؤثر واقع شود. زیرا در این کاربردها تعداد زیادی ویژگی وجود دارد، که بسیاری از آنها یا بلااستفاده هستند و یا اینکه بار اطلاعاتی چندانی ندارند. حذف نکردن این ویژگی‌ها مشکلی از لحاظ اطلاعاتی ایجاد نمی‌کند ولی بار محاسباتی را برای کاربرد مورد نظر بالا می‌برد. و علاوه بر این باعث می‌شود که اطلاعات غیر مفید زیادی را به همراه داده‌های مفید ذخیره کنیم. اخیراً تحقیقات زیادی بر روی انتخاب صفات خاصهٔ مناسب جهت دسته‌بندی داده‌ها انجام شده است. (اسماعیلی، طرفدار، ۱۳۸۸)

به طور کلی روش‌های مختلف انتخاب ویژگی را بر اساس نوع جستجو به دسته‌های مختلفی تقسیم بندی می‌کنند. در بعضی روش‌ها تمام فضای ممکن جستجو می‌گردد. در سایر روش‌ها که می‌تواند مکاشفه‌ای و یا جستجوی تصادفی باشد، در ازای از دست دادن مقداری از کارآئی، فضای جستجو کوچکتر می‌شود.

برای اینکه بتوانیم تقسیم بندی درستی از روش‌های مختلف انتخاب ویژگی داشته باشیم، به این صورت عمل می‌کنیم که فرآیند انتخاب ویژگی در تمامی روش‌ها را به این بخش‌ها تقسیم می‌کنیم:

- تابع تولید کننده[1]: این تابع زیر مجموعه‌های کاندید را برای روش مورد نظر پیدا می‌کند.
- تابع ارزیابی[2] : زیرمجموعهٔ مورد نظر را بر اساس روش داده شده، ارزیابی و یک عدد به عنوان میزان خوبی روش باز می‌گرداند. روش‌های مختلف سعی در یافتن زیرمجموعه‌ای دارند که این مقدار را بهینه کند.
- شرط خاتمه: برای تصمیم‌گیری در مورد زمان توقف الگوریتم.
- تابع تعیین اعتبار[3]: تصمیم می‌گیرد که آیا زیر مجموعه انتخاب شده معتبر است یا خیر؟

۴-۴-۱-۲-۱ تابع ارزیابی

یک تابع ارزیابی، میزان خوب بودن یک زیرمجموعه تولید شده را بررسی کرده و یک مقدار به عنوان میزان خوب بودن زیرمجموعه مورد نظر باز می‌گرداند. این مقدار با بهترین زیرمجموعهٔ قبلی مقایسه می‌شود. اگر زیر مجموعهٔ جدید، بهتر از زیرمجموعه‌های قدیمی باشد، زیرمجموعه جدید به عنوان زیرمجموعه بهینه، جایگزین قبلی می‌شود.

توابع ارزیابی را می‌توان به طرق مختلفی دسته بندی کرد. در اینجا ما دسته بندی‌ای که توسط Dash و Liu ارائه شده است را بیان می‌کنیم. آنها این معیارها را به پنج دست تقسیم کرده‌اند:

۱. معیارهای مبتنی بر فاصله[4]

در این معیارها، مثلاً برای یک مساله دو کلاسه، یک ویژگی یا یک مجموعه ویژگی مثل X بر یک ویژگی یا یک مجموعه ویژگی دیگر مثل Y ارجحیت دارد، اگر که با آن مجموعه ویژگی مقادیر بزرگتری برای اختلاف بین احتمالات شرطی دو کلاس داشته باشیم. نمونه‌ای از این معیارها همان معیار فاصله اقلیدسی می‌باشد.

۲. معیارهای مبتنی بر اطلاعات[5]

این معیارها میزان اطلاعاتی را که بوسیله یک ویژگی بدست می‌آید را در نظر می‌گیرند. ویژگی X در این روش‌ها بر ویژگی Y اولویت دارد، اگر اطلاعات بدست آمده از ویژگی X بیشتر از

[1] Generation procedure
[2] Evaluation function
[3] Validation procedure
[1] Distance Measures
[2] Information Measures

اطلاعاتی باشد، که از ویژگی Y بدست می‌آید. نمونه‌ای از این معیارها همان معیار آنتروپی[1] می باشد.

۳. معیارهای مبتنی بر وابستگی[2]

این معیارها که با عنوان معیارهای همبستگی[3] نیز شناخته می‌شوند، قابلیت پیشگوئی مقدار یک متغیر بوسیله یک متغیر دیگر را اندازه‌گیری می‌کنند. ضریب[4] یکی از معیارهای وابستگی کلاسیک است و می‌توانیم آنرا برای یافتن همبستگی بین یک ویژگی و یک کلاس به کار ببریم. اگر همبستگی ویژگی X با کلاس C بیشتر از همبستگی ویژگی Y با کلاس C باشد، در اینصورت ویژگی X بر ویژگی Y برتری دارد. با یک تغییر کوچک، می‌توانیم وابستگی یک ویژگی با ویژگی‌های دیگر را اندازه‌گیری کنیم. این مقدار درجه افزونگی این ویژگی را نشان می‌دهد. همه توابع ارزیابی بر پایه معیار وابستگی را می‌توانیم بین دو دسته معیارهای مبتنی بر فاصله و اطلاعات تقسیم کنیم. اما به خاطر اینکه این روش‌ها از یک دید دیگر به مساله نگاه می‌کنند، این کار را انجام نمی‌دهیم.

۴. معیارهای مبتنی بر سازگاری[5]

این معیارها جدیدتر هستند و اخیراً توجه بیشتری به آنها شده است. این معیارها خصوصیات متفاوتی نسبت به سایر معیارها دارند، زیرا که به شدت به داده‌های آموزشی تکیه دارند و در انتخاب یک زیرمجموعه از ویژگی‌ها تمایل دارند، که مجموعه ویژگی‌های کوچکتری را انتخاب کنند. این روش‌ها زیرمجموعه‌های با کمترین اندازه را بر اساس از دست دادن یک مقدار قابل قبول سازگاری که توسط کاربر تعیین می‌شود، پیدا می‌کنند.

۵. معیارهای مبتنی بر خطای طبقه بندی کننده[6]

روش‌هایی که این نوع از تابع ارزیابی را استفاده می‌کنند، با عنوان «methods wrapper» شناخته می‌شوند. دقت عملکرد در این روش‌ها برای تعیین کلاسی که نمونه داده شده متعلق به آن است، برای نمونه‌های دیده نشده بسیار بالا است، اما هزینه‌های محاسباتی در آنها نیز نسبتاً زیاد است.

---

[1] Entropy
[4] Dependence Measures
[5] Correlation
[6] Coefficient
[1] Consistency Measures
[2] Classifier Error Rate Measures

۴-۴-۱-۲-۲ توابع تولید کننده

اگر تعداد کل ویژگی‌ها برابر N باشد، تعداد کل زیرمجموعه‌های ممکن برابر ۲^{N} می‌شود. این تعداد برای N های متوسط هم خیلی زیاد است. بر اساس نحوه جستجو در میان این تعداد زیر مجموعه، روش‌های مختلف انتخاب ویژگی را می‌توان به سه دسته زیر تقسیم‌بندی نمود:

۱. جستجوی کامل
۲. جستجوی مکاشفه‌ای
۳. جستجوی تصادفی

در ادامه به معرفی هر کدام از این دسته‌ها می‌پردازیم.

❖ جستجوی کامل

در روش‌هایی که از این نوع جستجو استفاده می‌کنند، تابع تولید کننده بر اساس تابع ارزیابی استفاده شده، تمام فضای جواب (زیرمجموعه‌های ممکن) را برای یافتن جواب بهینه جستجو می‌کند. البته Schlimmer استدلال آورده است که: «کامل بودن جستجو به این معنی نیست که جستجو باید جامع باشد».

توابع مکاشفه‌ای مختلف زیادی طراحی شده‌اند، تا جستجو را بدون از دست دادن شانس پیدا کردن جواب بهینه، کاهش دهند. اما با توجه به بزرگی فضای جستجو، $O(۲^{N})$ ، این روش‌ها باعث می‌شوند که فضای کمتری جستجو شود. روش‌ها و تکنیک‌های مختلفی برای اینکار استفاده شده اند، بعضی از آنها از تکنیک بازگشت به عقب[1] نیز در جریان کار استفاده کرده-اند، مانند: branch and bound، best first search و beam search.

❖ جستجوی مکاشفه ای

در روش‌های با این نوع جستجو، در هر بار اجرای الگوریتم، یک ویژگی به مجموعه ویژگی انتخاب شده، اضافه و یا از آن حذف می‌شود. به همین دلیل پیچیدگی زمانی آنها محدود و کمتر از $O(N^{۲})$ می‌باشد. در اینگونه موارد، اجرای الگوریتم خیلی سریع می‌باشد و پیاده سازی آنها نیز بسیار ساده است.

[1] Back Tracking

❖ جستجوی تصادفی

روش‌هایی که از این نوع جستجو استفاده می‌کنند، محدوده کمتری از فضای کل حالات را جستجو می‌کنند، که اندازه این محدوده به حداکثر تعداد تکرار الگوریتم بستگی دارد. در این روش‌ها پیدا شدن جواب بهینه به اندازه منابع موجود و زمان اجرای الگوریتم بستگی دارد. در هر بار تکرار، تابع تولید کننده تعدادی از زیرمجموعه‌های ممکن از فضای جستجو را به صورت تصادفی انتخاب می‌کند و در اختیار تابع ارزیابی قرار می‌دهد. تابع تولید کننده تصادفی پارامترهایی دارد که بایستی تنظیم شوند، تنظیم مناسب این پارامترها در سرعت رسیدن به جواب و پیدا شدن جواب‌های بهتر مؤثر است.

۴-۴-۴-۱-۲-۳ روش های جستجو[1]

❖ جستجوی نخست-بهترین[2]

در این روش در هر مرتبه گرهی که بهترین ارزیابی را داشته باشد ابتدا بسط داده می‌شود به عبارت دیگر گرهی انتخاب می‌شود که تابع ارزیابی بهترین مقدار را برای آن باز گرداند. برحسب اینکه تابع ارزیابی چگونه پیاده سازی شود. روش‌های گوناگونی از الگوریتم اول بهترین خواهیم داشت جستجوی اول بهترین دو نوع کلی دارد که در یکی تابع ارزیابی هزینه رسیدن از گره فعلی به سمت هدف را حداقل می‌کند و در دومی هزینه کل مسیر از گره شروع تا هدف حداقل می‌شود.

❖ الگوریتم جستجوی ژنتیک[3]

تکنیک جستجویی برای یافتن راه‌حل تقریبی برای بهینه‌سازی و مسائل جستجو است. الگوریتم ژنتیک نوع خاصی از الگوریتم‌های تکامل است که از تکنیک‌های زیست‌شناسی فرگشتی مانند وراثت و جهش استفاده می‌کند.

❖ الگوریتم حریصانهٔ مرحله‌ای[4]

جستجوی حریصانه یکی از روش‌های جستجوی *نخست-بهترین* است در این روش هدف به حداقل رساندن هزینه رسیدن به هدف با استفاده از تابع تخمین می‌باشد ($h(n)$). بدین صورت که گرهی که به هدف نزدیکتر است ابتدا بسط داده می‌شود تابع ارزیابی که هزینه

[1] Search Method
[2] Best First
[3] Genetic search
[4] Greedy stepwise

رسیدن از یک حالت (حالت فعلی) به حالت هدف را تخمین می‌زند *تابع اکتشافی*[1] نام دارد و با حرف H بیان می‌گردد.

h(n)): هزینه تخمین زده شده از ارزان ترین مسیر از گره n به هدف.

❖ الگوریتم جستجوی کاشف[2]

الگوریتم جستجوی کاشف، به دسته‌ای از راه‌حل‌ها گویند که به دنبال جوابی معقول و قابل قبول برای یک مسأله دشوار می‌گردند که الزاماً بهترین جواب برای آن مسأله نیست.

کشف‌کنندگی از نظر دانشمندان هوش مصنوعی در دو وضعیت پایه می‌تواند صورت گیرد:

۱. مسئله‌ای وجود داشته باشد که فاقد راه حل دقیق باشد چرا که در تعریف مسئله و یا داده‌های موجود برای آن ابهام دیده می‌شود.

۲. مسئله‌ای ممکن است پاسخ دقیقی داشته باشد اما هزینه یافتن این پاسخ به قدری سنگین باشد که در عمل مقرون به صرفه نباشد.

خواص کشف کنندگی به‌وسیله جهت دادن به جستجو بخش قابل ملاحظه‌ای از این فضا را حذف می‌کند. متأسفانه روش‌های کشف کنندگی متکی بر تجربه و یا حس هستند و به همین علت استفاده از آنها در الگوریتم‌ها دشوار است. باید توجه داشت که خاصیت کشف کنندگی به علت حذف بخش قابل توجهی از فضای حالت ممکن است بعضی از جواب‌های بهینه را نیز از دست بدهد و در نهایت به جواب شبه بهینه دست یابد و یا اینکه در این امر توفیقی نداشته باشد. الگوریتم‌های کاشف شامل دو بخش هستند:

الف) ملاک کشف کنندگی

ب) الگوریتمی که بر پایه ملاک کشف کنندگی برای جستجوی فضای حالت مورد استفاده قرار گیرد.

۳-۱-۴-۴-۴ انتخاب صفات مناسب در مجموعه دادهٔ موجود

مهمترین صفات در مجموعه دادهٔ موجود انتخاب می‌گردند. از این صفات منتخب در پیاده‌سازی فاز دوّم به عنوان معیاری جهت خوشه‌بندی و شناسایی فراگیران ممتاز در هر خوشه بهره می‌بریم. در این پژوهش با استفاده از الگوریتم‌های جستجوی گوناگون (Vasantha, Bharathy, 2010) در دو روش به انتخاب صفات مناسب در میان مجموعه داده، پرداختیم:

[5] Heuristic Search
[6] Exhaustive search

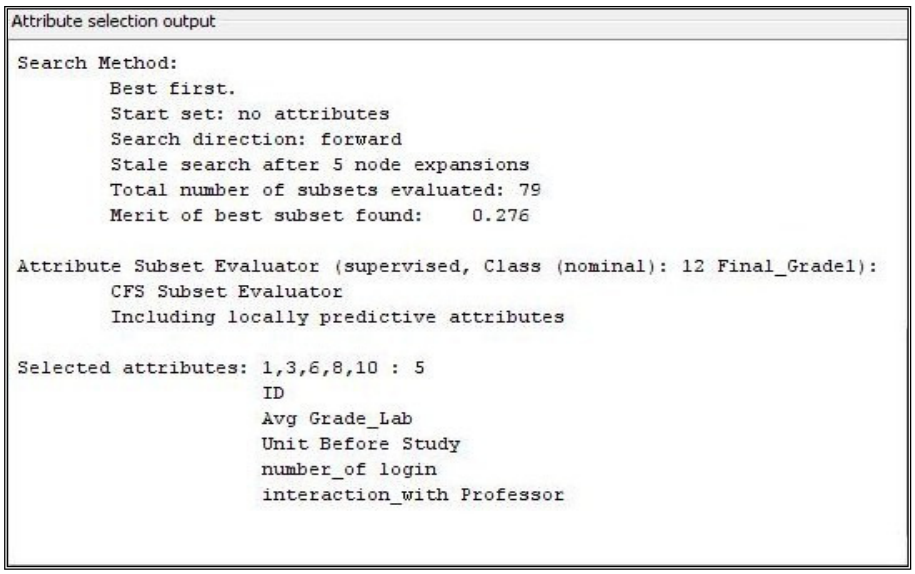
الف) روش اول: در این بخش با بکارگیری روش‌های جستجو و توابع ارزیابی مختلف، صفات منتخب را گروه‌بندی نمودیم. از ترکیب پنج روش جستجوی مختلف و دو تابع ارزیابی مبتنی بر سازگاری و مبتنی بر عوامل حیاتی[1] (گونه‌ای از معیارهای مبتنی بر وابستگی/همبستگی)، دو گروه A و B ایجاد می‌گردد. که گروه A شامل پنج صفت (ID، Interaction with professor، Number of login، Avg grade، Unit before study) و گروه B نیز شامل یک صفت واحد ID می‌باشد (جدول ۴-۵).

جدول ۴-۵ گروه‌بندی صفات منتخب با ترکیب روش‌های جستجو و توابع ارزیابی مختلف

شماره گروه	تعداد صفات منتخب	روش جستجو	تابع ارزیابی
A	۵	Best first	CfsSubset Evaluator
A	۵	Exhaustive search	CfsSubset Evaluator
A	۵	Genetic search	CfsSubset Evaluator
A	۵	Greedy stepwise	CfsSubset Evaluator
A	۵	Linear forward search	CfsSubset Evaluator
B	۱	Best first	Consistency Subset Evaluator
B	۱	Exhaustive search	Consistency Subset Evaluator
B	۱	Genetic search	Consistency Subset Evaluator
B	۱	Greedy stepwise	Consistency Subset Evaluator
B	۱	Linear forward search	Consistency Subset Evaluator

شکل شماره ۴-۱۰ نیز، خروجی فرآیند انتخاب ویژگی، با روش جستجوی نخست بهترین و تابع ارزیابی عوامل حیاتی (مبتنی بر وابستگی) را نشان می‌دهد.

[1] Critical Factors



شکل ۴-۱۰ خروجی فرایند انتخاب ویژگی

در مرحلهٔ بعد، میزان دقت پیش‌بینی را با در نظر گرفتن صفات منتخب ارزیابی می‌نماییم. همان‌طور که در قسمت‌های پیشین اشاره شد، الگوریتم درختی J48graft مدل مناسبی برای عملیات دسته‌بندی می‌باشند. با استفاده از این الگوریتم، دقت پیش‌بینی صفات منتخب دو گروه A و B و همچنین گروه C که حاوی تمام صفات بوده، محاسبه می‌گردد. (جدول ۴-۶) معیار کارایی را نیز روش جذر میانگین مربع خطاها[1] در نظر می گیریم. RMSE یکی از روش‌های اعتبارسنجی است که در فصل بعدی (فصل ارزیابی روش پیشنهادی)، شرح داده خواهد شد.

بررسی اعتبار روش جذر میانگین مربع خطاها بدین صورت است که RMSE با مقادیر پایین، نسبت به سایر روش‌ها، از اعتبار بیشتری برخوردار می‌باشد.

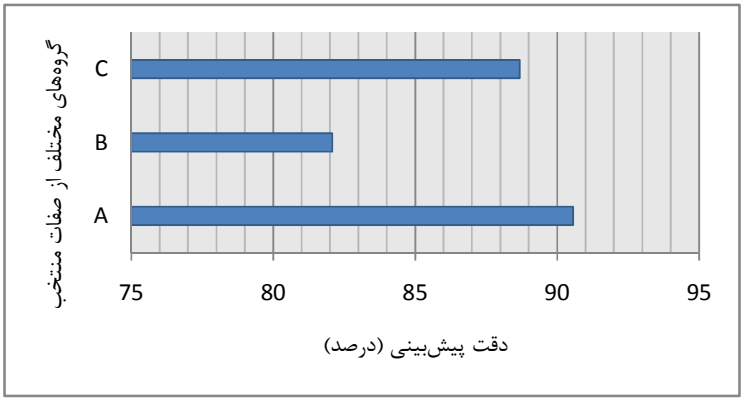
جدول ۴-۶ دقت پیش‌بینی صحیح در گروه‌های مختلف

شماره گروه	دقت پیش‌بینی (درصد)	Root mean square error
A	۹۰.۵۶۶	۰.۲۸۵۳
B	۸۲.۰۷۵۵	۰.۳۸۳۸
C	۸۸.۶۸	۰.۳۰۶۸

بدیهی است که تابع ارزیابی مبتنی بر سازگاری با انتخاب تنها یک صفت، گزینهٔ مناسبی جهت انتخاب ویژگی با توجه به مجموعه دادهٔ موجود نخواهد بود. همانطور که از جدول برآورد می‌گردد، گروه A که حاوی پنج صفت از مجوعه دادهٔ ما بود، بالاترین دقت پیش‌بینی را در عملیات دسته‌بندی

[1] Root mean square error

دارد. شکل ۴-۱۱ نیز نمایانگر مقایسه‌ای بین دقت پیش‌بینی نهایی گروه‌های مذکور در جدول ۴-۶ می‌باشند.



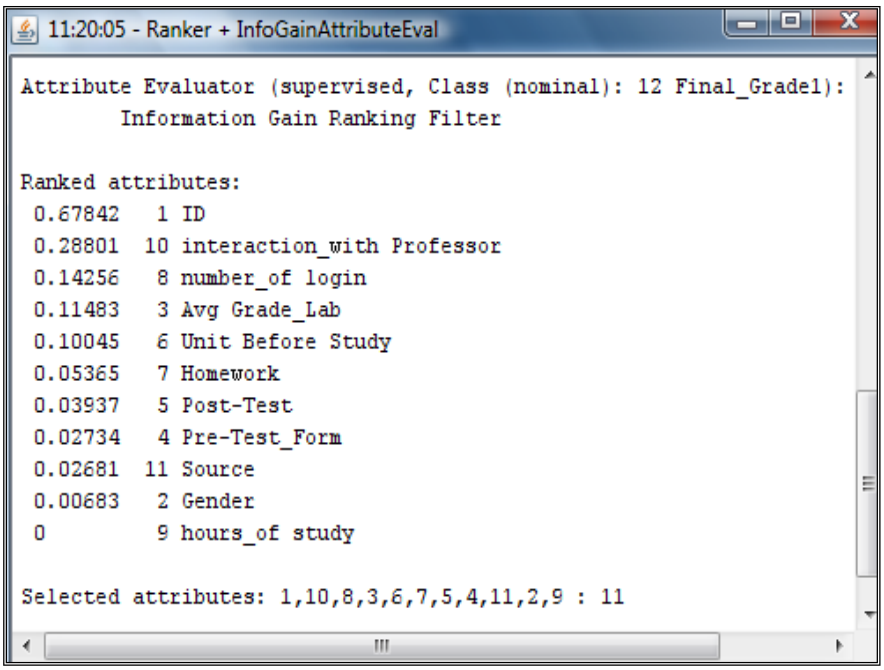
شکل۴-۱۱ مقایسهٔ دقت پیش‌بینی گروه‌های مختلف از صفات منتخب

ب) روش دوم: روش دیگر استفاده از الگوریتم رتبه‌بندی[1] صفات خاصه می باشد که مهمترین و مرتبط‌ترین صفات را با توجه به درجهٔ ارتباطش در لیستی فهرست می نماید. این روش جستجو از تابع ارزیابی مبتنی بر اطلاعات استفاده می‌نماید. (اسماعیلی، طرفدار، ۱۳۸۸)

همان‌طور که اشاره گردید، این تابع ارزیابی، میزان اطلاعاتی را که بوسیله یک ویژگی بدست می‌آید را در نظر می‌گیرد. ویژگی X در این روش‌ها بر ویژگی Y اولویت دارد، اگر اطلاعات بدست آمده از ویژگی X بیشتر از اطلاعاتی باشد که از ویژگی Y بدست می‌آید. همان طور که در شکل شماره ۴-۱۲ ملاحظه می نماییم صفات شماره ۱، ۱۰، ۸، ۳ و ۶ که به ترتیب ID، Interaction with professor، Number of login، Avg grade و Unit before study هستند، مهمترین و مؤثرترین صفات در عملیات دسته بندی می‌باشند. دقت پیش‌بینی با استفاده از این روش جستجو و الگوریتم درختی J48graft، ۹۰.۵۶۶ درصد می‌باشد. (شکل۴-۱۳)

از این صفات منتخب در پیاده‌سازی فاز دوم به عنوان معیاری جهت خوشه‌بندی و شناسایی فراگیران ممتاز در هر خوشه بهره می‌بریم. بدین معنی که فراگیران موفق بر اساس یک و یا چندین صفت از صفات منتخب خوشه‌بندی می‌شوند. و به منظور شناسایی زرنگترین و فعالترین فراگیران در یک خوشه، مقادیر پنج صفت منتخب را در رکورد فراگیران یک خوشه بررسی نموده و یا از روی نمودارهای مربوطه این مقادیر را مشاهده و با سایر فراگیران خوشه مقایسه می‌نماییم. فراگیرانی که مقادیر این صفات آنها در سطح بالاتری باشند به عنوان فراگیران ممتاز آن خوشه شناخته می‌شوند.

[1] Ranker



شکل ۴-۱۲ لیست صفات منتخب روش جستجوی Ranker به ترتیب الویت‌اشان

در واقع، از سه معیار ارزیابی مبتنی بر سازگاری، وابستگی و اطلاعات استفاده نمودیم. دو روش مبتنی بر سازگاری و اطلاعات از دقت بالاتری برخوردار بودند. در روش دوم (روش رتبه‌بندی) که از معیار مبتنی بر اطلاعات استفاده می‌نمود، مزایایی دیده می‌شود که باعث کاربرد بیشتر این روش در انتخاب ویژگی‌ها و خصایص یک مجموعه داده شده است. با مقایسهٔ دو شکل ۴-۱۰ و ۴-۱۲ در می‌یابیم که با بکارگیری روش جستجوی رتبه‌بندی و تابع ارزیابی مبتنی بر اطلاعات، صفات منتخب به ترتیب اولویت‌اشان با مقادیری رتبه‌بندی و وزن‌دهی می‌شوند. در نتیجه میزان اهمیت صفات مختلف قابل مقایسه می‌باشد.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48graft -C 0.25 -M 2
Relation:     final_data-weka.filters.unsupervised.attribute.Remove-R2,4-5,7,9,11
Instances:    106
Attributes:   6
              ID
              Avg Grade
              Unit Before Study
              number_of login
              interaction_with Professor
              Final_Grade1
Test mode:    10-fold cross-validation

Correctly Classified Instances          96               90.566  %
Incorrectly Classified Instances        10                9.434  %
Kappa statistic                          0.6794
Mean absolute error                      0.1591
Root mean squared error                  0.2853
Relative absolute error                 53.2719 %
Root relative squared error             74.3249 %
Total Number of Instances              106

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.943     0.263       0.943     0.943     0.943      0.861    passed
                 0.737     0.057       0.737     0.737     0.737      0.861    failed
Weighted Avg.    0.906     0.226       0.906     0.906     0.906      0.861

=== Confusion Matrix ===

  a  b   <-- classified as
 82  5 |  a = passed
  5 14 |  b = failed
```

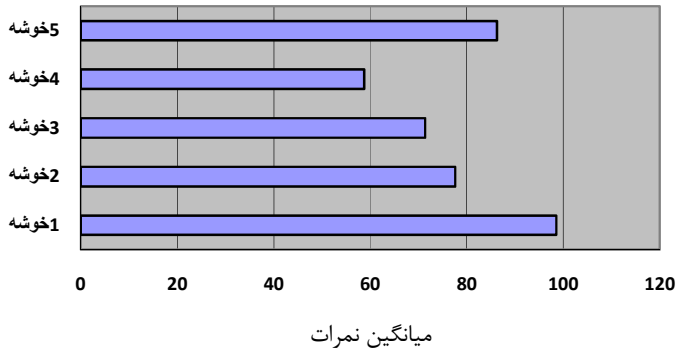
شکل ۴-۱۳ خروجی و دقت پیش‌بینی عملیات دسته‌بندی با در نظر گرفتن صفات منتخب الگوریتم رتبه‌بندی

قسمت فوقانی شکل ۴-۱۳نمایانگر نرم افزار مورد استفاده، تکنیک داده‌کاوی مورد استفاده و همچنین الگوریتم به کار گرفته شده در عملیات دسته‌بندی می‌باشد. سپس صفات منتخبی که توسط الگوریتم رتبه‌بندی برگزیده و در این پیش‌گویی استفاده شد، مشخص می‌گردد. قسمت میانی شکل، درصد پیش‌بینی های صحیح و غلط و یا به عبارتی دقت پیش‌بینی الگوریتم را نشان داده و در نهایت در بخش آخر پارامترهای ارزیابی در هر دو کلاس و ماتریس آشفتگی را مشاهده می‌نماییم که پارامترهای ارزیابی و ماتریس آشفتگی در فصل بعدی (فصل۵) به تفصیل شرح داده خواهد شد.

۴-۴-۴-۲ پیاده‌سازی فاز دوم

در این پژوهش از الگوریتم k-means به منظور خوشه‌بندی فراگیران استفاده نمودیم. تعداد فراگیرانی که بر اساس رفتار و عملکردشان در محیط آموزش الکترونیکی خوشه‌بندی گردیدند، ۸۷ نفر می‌باشد. به عبارتی فقط فراگیران موفق را خوشه‌بندی نمودیم. این فراگیران بر اساس نمرات ترم قبل‌اشان در ۵ خوشه گروه‌بندی شدند. الگوریتم k-means با تعداد خوشه های مختلف از ۲ تا ۶ اجرا شده است و در نهایت تعداد ۵ خوشه به بهترین شکل واقعیت را منعکس کرده است. به‌عنوان راه حلی دیگر، چون در الگوریتم k-means تعداد خوشه‌ها از قبل مشخص نبوده، می‌توان از الگوریتم EM بهره گرفته که تعداد خوشه‌های بهینه را خود الگوریتم مشخص می‌نماید. همانطور که ذکر گردید، در عملیات خوشه‌بندی فراگیران، تمامی صفات را در نظر نمی‌گیریم بلکه تنها صفت و یا صفاتی را که از قبل مشخص بوده و در طول تعامل فراگیر با سیستم آموزشی تغییر نمی‌کند و همچنین موفقیت وی در وهلهٔ اول بدان بستگی دارد، مد نظر قرار می‌دهیم. بدین ترتیب فراگیران با توجه به میانگین نمرات ترم‌های قبل‌اشان خوشه‌بندی می‌شوند.

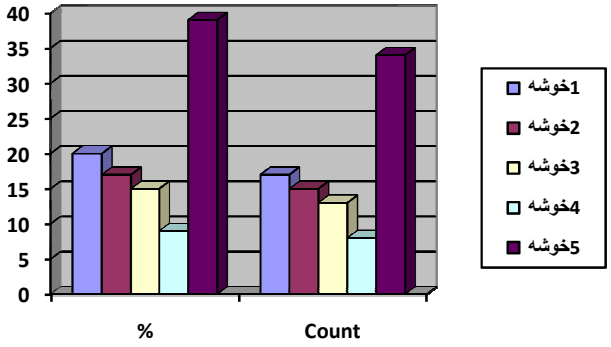
معیار کارایی را نیز مجموع مربعات خطا[1] در نظر می گیریم که در فصل ارزیابی توضیحاتی در خصوص این معیار کارایی داده خواهد شد. نتایج حاصله نشان داد که مجموع مربعات خطا در هر خوشه هنگام در نظر گرفتن مهمترین صفات، نسبت به زمانیکه تمامی صفات در نظر گرفته شوند، پایین می‌آید. مقدار میانگین نمرات فراگیران در خوشه‌های مختلف، در شکل ۴-۱۴ نشان داده می‌شود. خوشهٔ شمارهٔ یک شامل ممتازترین فراگیران می‌باشد.



شکل ۴-۱۴ نمودار میانگین نمرات فراگیران در خوشه‌های مختلف

[1] sum of squared errors

شکل ۴-۱۵ نیز فراوانی هر خوشه را مشخص می‌نماید. با توجه به نمودار، خوشهٔ ۵ بیشترین فراوانی را دارد و خوشهٔ ۴ که پایین‌ترین نمرات را دارد، دارای کمترین فراوانی است و فقط حدود ۹ درصد از فراگیران را شامل می شود.



شکل۴-۱۵ فراوانی و درصد فراگیران در هر خوشه

۴-۴-۴-۳ پیاده‌سازی فاز سوم

یکی از چالش‌های جدی در مدیریت امور آموزشی پیش‌بینی وضعیت تحصیلی فراگیران قبل از اتمام دورهٔ تحصیلی، به منظور شناسایی فراگیرانی است که دچار افت تحصیلی شده و ادامه تحصیل آنها با مشکل روبرو خواهد شد. (ایرجی، مینایی و شکورنیاز، ۱۳۸۷)

از الگوریتم Apriori به‌منظور کاوش قوانین انجمنی استفاده نمودیم. این الگوریتم را به کل مجموعه داده و به عبارتی کل فراگیران موجود در سیستم اعمال می‌کنیم. با بکارگیری این مدل، سیستم آموزشی می‌تواند مشاوره‌ها و پیشنهادات لازم را برای پیشگیری از رسیدن دانشجویان به وضعیت بحرانی بکار گیرد. نمونه‌ای از این قواعد در شکل شماره ۴-۱۶ نشان داده شده است. قواعد با ضریب اطمینان بالا از اهمیت بیشتری برخوردارند.

```
Best rules found:

 1. interaction_with Professor=very_high 28 ==> Final_Grade1=passed 28    conf:(1)
 2. number_of login=high interaction_with Professor=very_high 20 ==> Final_Grade1=passed 20    conf:(1)
 3. Unit Before Study=ch3 19 ==> Final_Grade1=passed 19    conf:(1)
 4. Post-Test=C number_of login=high 19 ==> Final_Grade1=passed 19    conf:(1)
 5. Unit Before Study=ch2 interaction_with Professor=high 18 ==> Final_Grade1=passed 18    conf:(1)
 6. Pre-Test_Form=A Post-Test=C 17 ==> Final_Grade1=passed 17    conf:(1)
 7. Avg Grade=D 16 ==> Unit Before Study=ch2 16    conf:(1)
 8. interaction_with Professor=high 27 ==> Final_Grade1=passed 26    conf:(0.96)
 9. Unit Before Study=ch2 number_of login=medium 20 ==> Final_Grade1=passed 19    conf:(0.95)
10. Pre-Test_Form=A Source=e-book 19 ==> Final_Grade1=passed 18    conf:(0.95)
```

شکل ۴-۱۶ نمونه‌ای از قواعد انجمنی کشف شده

فصل۵ - ارزیابی روش پیشنهادی

۵-۱مقدمه

در این فصل آزمایشات انجام شده به منظور ارزیابی روش پیشنهادی ارائه می‌شوند. ابتدا پارامترهای ارزیابی معرفی شده و در ادامه، آزمایشات و نتایج آن‌ها گزارش می‌شود. در پایان نیز نتایج آزمایشات مورد بررسی و تحلیل قرار می‌گیرند.

انتخاب فاکتورهای مؤثر برای ارزیابی روش‌های یادگیری ماشین اعمال شده یکی از مسایل مهم در انجام این پژوهش بوده است که ما از روش ارزیابی متقاطع[1] ده دسته استفاده کردیم که در آن میزان مقدار داده تست نگاه داشته شده ۱۰ درصد کل داده آموزشی است. در ارزیابی متقاطع برای تعیین میزان کیفیت یک طبقه بندی کننده مقداری از داده آموزشی به عنوان مجموعه تست استفاده می‌شود. اگر میزان مقدار داده تست نگاه داشته شده ۱۰٪ کل دادهٔ آموزشی باشد به این شیوه، ارزیابی متقاطع ده-دسته گفته می‌شود. می‌توان تعداد دسته‌ها را به غیر از ده نیز انتخاب نمود. به عنوان مثال اگر ۵٪ از داده به عنوان تست نگاه داشته شود، آنگاه ارزیابی متقاطع ۲۰-دسته را خواهیم داشت.

۵-۲ پارامترهای ارزیابی

انتخاب فاکتورهای مؤثر به منظور ارزیابی روش‌های یادگیری ماشین اعمال شده، از اهمیت بسیاری برخوردار است. در ادامه به معرفی فاکتورهایی که در ارزیابی روش‌های مختلف مخصوصاً فاز نخست (دسته‌بندی)، مورد استفاده واقع شده‌اند، می‌پردازیم.

[1] Cross Validation

۵-۲-۱ ماتریس آشفتگی[1]

یک ماتریس آشفتگی شامل اطلاعاتی در مورد دسته‌بندی واقعی و پیش بینی شده توسط یک سیستم دسته‌بندی می باشد.

در زمینهٔ مطالعهٔ ما درایه‌های ماتریس آشفتگی دارای معانی زیر می باشند: (شکل۵-۱)

Predicted			
Positive	Negative		
b	a	Negative	Actual
d	c	Positive	

شکل۵-۱ نمایش مفهوم درایه های ماتریس آشفتگی

هر یک از درایه های این ماتریس معانی زیر را دارد:

- a، تعداد پیش بینی های درستی که یک نمونه را منفی تعبیر کرده اند.
- b، تعداد پیش بینی های نادرستی که یک نمونه منفی را مثبت تعبیر کرده اند.
- c، تعداد پیش بینی های نادرستی که یک نمونه مثبت را منفی تعبیر کرده اند.
- d، تعداد پیش بینی های درستی که یک نمونه را مثبت تعبیر کرده اند.

۵-۲-۲ صحت[2] : نسبت تعداد پیش بینی های درست به تعداد کل پیش بینی ها.

$$AC = \frac{a+d}{a+b+c+d} \qquad (۱-۵)$$

۵-۲-۳ یادآوری (فراخوان)[3] یا نرخ مثبت درست[4] : نسبت حالات مثبتی که درست تعیین شده اند.

$$TP = \frac{d}{c+d} \qquad (۲-۵)$$

۵-۲-۴ نرخ مثبت کاذب[5] : نسبت حالات منفی که بصورت اشتباه کلاسه بندی شده اند.

$$FP = \frac{a}{a+b} \qquad (۳-۵)$$

[1] Confusion Matrix
[2] Accuracy
[3] Recall
[4] True Positive rate
[5] False Positive rate

۵-۲-۵ نرخ منفی درست[1] : نسبت تعداد حالات منفی که درست کلاسه بندی شده اند.

$$TN = \frac{a}{a+b} \qquad (۴-۵)$$

۵-۲-۶ نرخ منفی کاذب[2] : نسبت حالات مثبتی که بصورت حالات منفی کلاسه بندی شده اند.

$$FN = \frac{c}{c+d} \qquad (۵-۵)$$

۵-۲-۷ دقت[3] : نسبت حالات مثبت پیش‌بینی شده‌ای که درست بوده اند.

$$P = \frac{d}{b+d} \qquad (۶-۵)$$

مقدار صحت بیان شده در فرمول ۵-۱ ممکن است معیار کارا و کاملی نباشد اگر تعداد حالات منفی بسیار بیشتر از حالات مثبت باشد. فرض کنید ۱۰۰۰ حالت داریم که ۹۹۵ تای آنها حالات منفی هستند و فقط ۵ تای آنها حالات مثبت هستند. اگر سیستم همه آنها را بعنوان حالت منفی کلاسه بندی کند میزان دقت ۹۹/۵٪ خواهد بود حتی اگر طبقه بندی کننده[4] همه حالات مثبت را از دست داده باشد.

۵-۲-۸ معیار F-Measure

مقیاسی است که از تلفیق دو معیار دقت و یادآوری بدست می آید:

$$F = \frac{2\times recall\times precision}{recall+precision} \qquad (۷-۵)$$

این مقیاس می تواند برای ارزیابی دقت یک طبقه بندی، مورد استفاده قرار گیرد. راه دیگر برای تست و بررسی عملکرد طبقه بندی کننده استفاده از گراف ROC می باشد.

۵-۲-۹ گراف ROC[5]

گراف های ROC روش دیگری برای بررسی عملکرد طبقه‌بندی کننده می‌باشند (Swets,1988). ROC یک ابزار مدل سازی قوی است که در تصمیم گیری های پژشکی، روانشناسی، مخابرات و در زمانی که نیاز به ارزش های آستانه ای مد نظر است استفاده می شود. منحنی ROC یک نمودار

[1] True Negative rate
[2] False Negative rate
[3] Precision
[4] Classifier
[5] Receiver Operating Characteristic

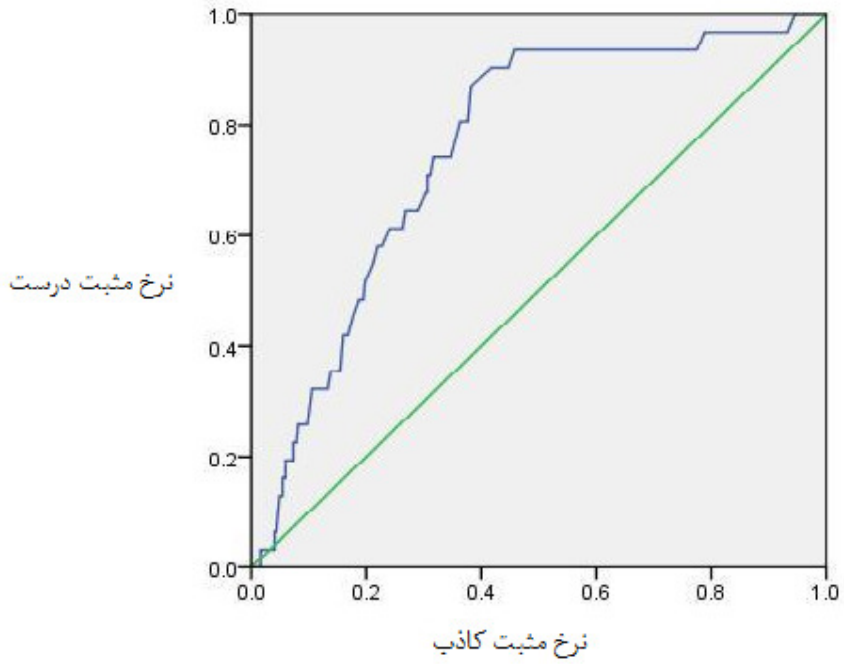
پراکندگی از حساسیت[1] برای یک سیستم طبقه بندی کنندهٔ باینری است که آستانهٔ تمیز آن متغیر است. برای تشکیل نمودارهای ROC به دو طبقه بندی نیاز است. یکی طبقه بندی واقعی و دیگری طبقه بندی پیش بینی شده بر اساس مدل. یک گراف ROC نموداری با نرخ مثبت کاذب روی محور جداکننده X و نرخ مثبت درست روی محور Y می باشد. مختصات نقطه (۱و۰) طبقه‌بندی کنندهٔ کامل و حالت ایده آل است یعنی نقطه‌ای است که تمام حالات مثبت و منفی را درست کلاسه‌بندی می‌کند. . این نقطه نشان می دهد که آنچه مدل پیش بینی ارائه می دهد انطباق کامل بر مدل واقعی دارد.هر چه نقاط به سمت بالا و چپ نزدیک تر باشد مناسب تر است و مدل پیش بینی به حالت ایده آل خود نزدیک‌تر است. نقطه (۰و۰) طبقه‌بندی کننده‌ای را نشان می‌دهد که تمام حالات را به منفی تعبیر می‌کند در حالیکه نقطه (۱و۱) با طبقه‌بندی کننده‌ای تطابق دارد که همه حالات را مثبت تفسیر می‌کند. نقطه (۰و۱) مبین طبقه‌بندی کننده‌ای است که برای تمام کلاسه‌بندی‌ها نادرست است و به این معنا است که هر چه مدل پیش بینی ارائه داده عکس مدل واقعی است. این قسمت از فضای نمودار ROC نیز جالب توجه است و باید نتیجه ی سوالاتی که منجر به این حالت شده را عکس کرد. اما در صورتی که مدل تصادفی عمل کند نقاط در اطراف خط y=x قرار می‌گیرند.

$$Tp\ rate \approx \frac{positives\ correctly\ classified}{total\ positives} \qquad (۸-۵)$$

$$Fp\ rate \approx \frac{negatives\ incorrectly\ classified}{total\ negatives} \qquad (۹-۵)$$

هر پارامتر یک زوج (FP,TP) را فراهم می کند و یک مجموعه از چنین زوج هایی می‌تواند برای رسم منحنی ROC استفاده شود. شکل ۵-۲ مثالی از یک منحنی ROC را نشان می‌دهد.

[1] Sensitivity



شکل ۵-۲ یک منحنی ROC

اندازه‌گیری صحت مبتنی بر ناحیه[1] : پیشنهاد شده است که ناحیه زیر منحنی Roc می تواند بعنوان معیار صحت در بسیاری کاربردها در نظر گرفته شود (Swets,1988). بر اساس مساحت بین منحنی ROC و خط y=x قدرت منحنی های ROC مشخص می شود. هر چه این مساحت کمتر شود نشان می دهد که مدل پیش بینی قدرت ضعیف تری دارد و هرچه مساحت افزایش یابد نشان از قدرت مدل پیش‌بینی و نزدیکی آن به مدل حقیقی است.

۵-۲-۱۰ ارزیابی پیش‌بینی عددی[2]

معیارهایی که در جدول ۵-۱آمده‌اند می‌توانند برای ارزیابی موفقیت پیش‌بینی عددی استفاده شوند. مقادیر پیش‌بینی روی نمونه های تست عبارتند از p_1 , p_2 , ... , p_n و مقادیر واقعی عبارتند از $a_1, a_2, \dots, a_n$. در اینجا p_i به مقدار عددی پیش‌بینی برای i امین نمونه تست اشاره می کند.

خطای میانگین مربع[3] اصلی ترین و معمولترین معیار مورد استفاده است. بسیاری از تکنیک‌های ریاضیاتی (مانند رگرسیون خطی) از خطای میانگین مربع استفاده می‌کنند چرا که آسانترین معیار

[1] Area-based Accuracy Measure
[2] Evaluating Numeric Prediction
[3] Mean Squared error

دستکاری[1] ریاضیاتی می‌باشد و در اصطلاح ریاضیات به آن خوش‌رفتار[2] گفته می‌شود. به هر حال در اینجا آن را بعنوان معیار عملکرد در نظر می‌گیریم: همه معیارهای عملکرد برای محاسبه آسان هستند بنابراین خطای میانگین مربع مزیت خاصی ندارد.

خطای میانگین مطلق[3] روش دیگری است که اندازهٔ یک خطا را بدون در نظر گرفتن علامت آن نظر دارد. خطای میانگین مربع تمایل به بزرگ‌نمایی تأثیر نمونه‌هایی که خطای پیش‌بینی آنها بزرگتر از سایرین است[4]، دارد اما خطای مطلق این تأثیر را ندارد و با همه اندازه‌های خطا بطور یکسان بر طبق بزرگی آنها رفتار می‌شود.

گاهی اوقات خطای نسبی مهم‌تر از خطای مطلق است. مثلاً، اگر یک خطای ۱۰٪ در اینکه آیا یک خطای ۵۰ در یک پیش بینی از ۵۰۰ است یا یک خطای ۰.۲ در یک پیش بینی از ۲ است از اهمیت یکسانی برخوردار باشد، میانگین‌های خطای مطلق بی‌معنی خواهند بود و خطاهای نسبی مناسب واقع خواهند شد. این تأثیر با استفاده از خطاهای نسبی در محاسبات خطای میانگین مربع یا محاسبات خطای میانگین مطلق به کار گرفته می شود.

خطای نسبی مربع[5] در جدول ۵-۱ به چیز کاملاً متفاوتی اشاره می‌کند. خطا وابسته به آن چیزی است که اگر یک پیش‌بینی کنندهٔ ساده استفاده می‌شد، می‌بود. پیش‌بینی کننده ساده میانگین مقادیر واقعی داده های آموزش[6] است. بنابراین خطای نسبی مربع، خطای مجموع مربع را می گیرد و آن را با تقسیم بر خطای مجموع مربع پیش بینی کننده پیش فرض نرمال می کند.

معیار خطای بعدی خطای نسبی مطلق است و همان خطای مجموع مطلق با همان نرمالیزاسیون می‌باشد. در این سه معیار خطای نسبی، خطاها بوسیله خطای پیش بینی کننده ساده که مقادیر میانگین را پیش بینی می کند نرمال می‌شوند.

[1] manipulate
[2] well - behaved
[3] Mean Absolute error
[4] Outlier
[5] Relative Squared error
[6] Training Data

جدول ۵-۱ معیارهای عملکرد برای پیش بینی عددی

معیارهای عملکرد	فرمول
mean-squared error	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$
root mean-squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}}$
mean absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{n}$
relative squared error	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}$, where $\bar{a}=\frac{1}{n}\sum_i a_i$
root relative squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}}$
relative absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{\lvert a_1-\bar{a}\rvert+\ldots+\lvert a_n-\bar{a}\rvert}$

۵-۳ آزمایشات انجام شده در فاز نخست

همانطور که در مرحلهٔ پیاده‌سازی در فصل قبل اشاره گردید، الگوریتم‌های بهینه به منظور انجام فرآیند پیش‌گویی عبارتند از: j48 graft، SimpleLogistic، JRip. در این مطالعه نهایتاً از روش کلاسه‌بندی درخت تصمیم‌گیری به عنوان مدل استفاده می‌شود. هدف اصلی، شناسایی فراگیران تحت ریسک و نیز فراگیران با انگیزه و با علاقه‌مندی بالا به ادامه تحصیل است.

الگوریتم j48 graft، یک الگوریتم درختی با دقت پیش‌گویی بالا و صحت قابل قبول است که خروجی آن به صورت یک درخت بوده و برای انسان قابل درک است. از این الگوریتم به منظور طبقه‌بندی فراگیران و پیش‌گویی نتیجهٔ تحصیلی آنها استفاده شده است. از ۱۰۶ نمونهٔ موجود در مجموعه داده، تعداد ۹۴ عدد یعنی ۸۸.۶۸ درصد کلاس مربوطه، به شکل صحیح طبقه‌بندی شده‌اند. که نتایج این روش در جدول۵-۲ تشریح شده است. شکل ۵-۳ نمایانگر ماتریس آشفتگی حاصل از اعمال این الگوریتم است. که یکی از پارامترهای ارزیابی مورد نظر ما در این پژوهش می‌باشد.

جدول ۵-۲ نتایج بدست آمده از اعمال تکنیک j48 graft

مورد	میزان
دقت کلاس قبول	۰.۹۲۱
دقت کلاس مردود	۰.۷۰۶
یادآوری کلاس قبول	۰.۹۴۳
یادآوری کلاس مردود	۰.۶۳۲
F-Measure کلاس قبول	۰.۹۳۲
F-Measure کلاس مردود	۰.۶۶۷
Roc Area کلاس قبول	۰.۸۳
Roc Area کلاس مردود	۰.۸۳
تعداد نمونه صحیح اختصاص داده شده از کلاس قبول	۸۲
تعداد نمونه صحیح اختصاص داده شده از کلاس مردود	۱۲
دقت نهایی روش (درصد)	**۸۸.۶۸**

```
=== Confusion Matrix ===

  a  b   <-- classified as
 82  5 |  a = passed
  7 12 |  b = failed
```

شکل ۵-۳ نمایش ماتریس آشفتگی حاصل از الگوریتم J48 graft

درآیه‌های ماتریس آشفتگی بدست آمده با توجه به شکل ۵-۱دارای مقادیر زیر می‌باشند:

a=۱۲, b= ۷, c=۵, d=۸۲

بنابراین به منظور محاسبهٔ دقت نهایی روش و یا صحت کافیست ارقام بالا را در معادلهٔ مربوط به صحت (معادلهٔ ۵-۱) جایگذاری نماییم.

$$\text{دقت نهایی} = \frac{a+d}{a+b+c+d} = \frac{۱۲+۸۲}{۱۲+۷+۵+۸۲} = \frac{۹۴}{۱۰۶} = ۰.۸۸۶۸ = ۸۸.۶۸٪$$

دقت کلاس قبول که عبارتند از نسبت حالات مثبت پیش‌بینی شده‌ای که درست بوده‌اند (معادلهٔ ۵-۲)، از رابطهٔ زیر محاسبه می‌گردد:

$$\text{دقت کلاس قبول} = \frac{d}{b+d} = \frac{۸۲}{۷+۸۲} = \frac{۸۲}{۸۹} = ۰.۹۲۱ = ۹۲.۱٪$$

بدین ترتیب سایر پارامترهای ارزیابی که در خروجی نرم افزار و در جدول ۵-۲ قرار گرفته‌اند نیز از طریق جایگذاری درآیه‌های مربوط به ماتریس آشفتگی در معادلات ذکر شده، قابل محاسبه می‌باشند. روش رگرسیون ساده، روش دیگری است که برای طبقه‌بندی فراگیران و پیش گویی نتیجهٔ تحصیلی آنها استفاده شده است. در این روش۹۵ عدد یعنی ۸۹.۶۲ درصد کلاس مربوطه به شکل صحیح طبقه‌بندی شده‌اند. نتایج این روش در جدول ۵-۳ تشریح شده است.

جدول ۵-۳ نتایج بدست آمده از اعمال تکنیکSimpleLogistic

مورد	میزان
دقت کلاس قبول	۰.۹۴۲
دقت کلاس مردود	۰.۷
یادآوری کلاس قبول	۰.۹۳۱
یادآوری کلاس مردود	۰.۷۳۷
F-Measure کلاس قبول	۰.۹۳۶
F-Measure کلاس مردود	۰.۷۱۸
Roc Area کلاس قبول	۰.۸۰۳
Roc Area کلاس مردود	۰.۸۰۳
تعداد نمونه صحیح اختصاص داده شده از کلاس قبول	۸۱
تعداد نمونه صحیح اختصاص داده شده از کلاس مردود	۱۴
دقت نهایی روش (درصد)	**۸۹.۶۲**

روش آخر دسته‌بندی مورد نظر، روش JRip است. این روش نیز در فرآیند انجام شده موفق به طبقه‌بندی صحیح ۹۵ عدد یعنی ۸۹.۶۲ درصد از مجموعه داده شده است. جزییات مربوط به نتایج بدست آمده از این روش در جدول ۵-۴ نشان داده شده است. شکل ۵-۴ نمایانگر ماتریس آشفتگی حاصل از اعمال الگوریتم‌های SimpleLogistic و JRip می‌باشد.

```
=== Confusion Matrix ===

  a  b   <-- classified as
 81  6 |  a = passed
  5 14 |  b = failed
```

شکل ۵-۴ نمایش ماتریس آشفتگی حاصل از الگوریتم SimpleLogistic و JRip

جدول ۵-۴ نتایج بدست آمده از اعمال تکنیکJRip

مورد	میزان
دقت کلاس قبول	۰.۹۴۲
دقت کلاس مردود	۰.۷
یادآوری کلاس قبول	۰.۹۳۱
یادآوری کلاس مردود	۰.۷۳۷
F-Measure کلاس قبول	۰.۹۳۶
F-Measure کلاس مردود	۰.۷۱۸
Roc Area کلاس قبول	۰.۷۵۶
Roc Area کلاس مردود	۰.۷۵۶
تعداد نمونه صحیح اختصاص داده شده از کلاس قبول	۸۱
تعداد نمونه صحیح اختصاص داده شده از کلاس مردود	۱۴
دقت نهایی روش (درصد)	**۸۹.۶۲**

۵-۳-۱ مقایسهٔ نتایج مدل‌ها و اجرای مدل منتخب

با مقایسهٔ عملکرد کلی مدل های ساخته شده در مرحله قبل، از بین سه مدل مذکور، مدلی که دقت بیشتری دارد را برای انجام تحلیل‌های بیشتر به کار می گیریم. هر سه مدل منتخب از دقت نسبتاً بالایی بر خوردار می‌باشند. گرچه دقت دو روش SimpleLogistic و JRip کمتر از یک درصد از روش دسته‌بندی j48 graft بیشتر است، اما به منظور ارزیابی دقت و صحت طبقه‌بندی کننده‌های مورد استفاده، از ناحیه زیر منحنی ROC[1] استفاده می‌نماییم. هرچه مساحت زیر این منحنی بیشتر باشد نشان‌دهندهٔ قدرت مدل پیش‌بینی و نزدیکی آن به مدل حقیقی است. با توجه به نتایج جداول ۵-۲، ۵-۳ و ۵-۴ این مساحت در مدل دسته‌بندی j48graft نسبت به دو مدل دیگر بالاتر بوده، همچنین با در نظر گرفتن نقاط قوت روش‌های درختی در پیش بینی و خروجی درختی قابل درک آن، برتری مدل ساخته شده با تکنیک j48graft کاملاً مشهود می باشد.

در نتیجهٔ اجرای مدل، مجموعه‌ای از مهمترین قوانین کشف شده از دید نرم‌افزار ارائه خواهد شد. که در جدول شماره ۵-۵، چند قانون تولید شده، به تفسیر در آمده است.

[1] Roc Area

قوانین بدست آمده از درخت تصمیم j48graft با فرض final-grade (نمرهٔ نهایی) به عنوان فیلد هدف:

Interaction with Professor = very high: passed (28.54)

Interaction with Professor = medium: passed (29.56/4.0)

Interaction with Professor = low

|Avg Grade <= 92.5

| |Homework = very high: passed (0.0|14.0/1.0)

| |Homework != very high: failed (20.4/6.38)

|Avg Grade > 92.5: passed (0.0|18.0/1.0)

Interaction with Professor = high: passed (27.52/1.0)

جدول ۵-۵ قوانین بدست آمده از درخت تصمیم j48graft با فرض final-grade به عنوان فیلد هدف

شماره قانون	مجموعه قانون
۱	If Interaction with Professor = very high, then final grade= passed If Interaction with Professor = medium, then final grade= passed If Interaction with Professor = high, then final grade= passed
	جمعیتی از فراگیران که میزان تعاملشان با استاد متوسط، بالا و خیلی بالا بوده ، به احتمال خیلی زیاد قادر به گذراندن موفقیت آمیز دورهٔ تحصیلی خواهند بود.
۲	If Interaction with Professor = low and Avg Grade > 92.5 then final grade= passed
	گروهی از فراگیران که میزان تعامل‌اشان با استاد کم بوده و میانگین نمرات ترم قبل آنها بیشتر از ۹۲.۵ به احتمال خیلی زیاد قادر به گذراندن موفقیت آمیز دورهٔ تحصیلی خواهند بود.
۳	and Avg Grade <= 92.5 and If Interaction with Professor = low (If Homework = very high then final grade= passed Or If Homework != very high then final grade= failed)
	گروهی از فراگیران که میزان تعامل‌اشان با استاد کم بوده و میانگین نمرات ترم قبل آنها کمتر از ۹۲.۵ است. اگر تمامی تکالیف خود را انجام دهند قادر به گذراندن دوره و در غیر این صورت قادر به گذراندن دوره نخواهند بود و از آزمون نهایی رد می‌شوند.

۵-۳-۲ نتایج آزمایش فاز اول

به منظور تست سیستم، فرض می‌گردد که فراگیر جدیدی وارد سیستم شد، احتمال پیش‌بینی قبولی و یا رد وی از آزمون نهایی قبل از اتمام دورهٔ تحصیلی‌اش محاسبه می‌گردد. (شکل۵-۵ و ۵-۶) مجموعه دادهٔ موجود شامل رکوردهایی است که هر رکورد داده‌های مربوط به یک فراگیر را مشخص می‌نماید. داده‌های مربوط به هر فراگیر به صورت زیر نمایش داده می‌شود:

ID, Gender, Avg Grade, Pre-Test Form, Post-Test, Unit Before Study, Homework, number of login, hours of study, interaction with Professor, Source, Final Grade.

به عنوان مثال داده‌های مربوط به فراگیر X به شرح زیر است:

X,M,88,B,A,17,low,medium,5,high,both,passed

این عبارت بدین معناست که جنسیت فراگیر X پسر و میانگین نمرات ترم قبلش ۸۸ بوده، وی در ابتدا فرم تست B و سپس تست A را انجام داده، درس مطالعه شده توسط وی قبل از آزمون نهایی درس شماره ۱۷ بوده است. تکالیف خود را کم انجام می‌دهد و تعداد دفعات ورود به سیستم آموزشی متوسط بوده، در هر بار ورود تقریباً به مدت ۵ ساعت به مطالعه و گشت و گذار در سایت مشغول است. میزان تعاملش با استادان و مربیان آموزشی بالا بوده و دروس را هم از طریق سایت آموزشی و هم از طریق کتب الکترونیکی که به وی معرفی می‌گردد مطالعه می‌نماید. نتیجهٔ نهایی وی هم در آزمون آخر دوره موفقیت آمیز بوده و در امتحان نهایی قبول می‌شود.

بدین ترتیب داده‌های مربوط به فراگیر Y که دورهٔ تحصیلی‌اش را به اتمام نرسانده و آزمون نهایی را انجام نداده می‌تواند به شکل زیر باشد:

Y,F,93,A,?,8,Middle,high,4,high,both,?

که علامت ؟ به معنای نامشخص بودن فیلد مربوطه می‌باشد.

```
=== Predictions on test set ===

inst#,     actual, predicted, error, probability
    1          ?   1:passed          *0.964
```

شکل ۵-۵ احتمال پیش‌بینی نتیجهٔ نهایی فراگیر جدید Y

همانطور که در شکل ۵-۵ ملاحظه می‌نماییم، خروجی نرم‌افزار نشان می‌دهد که هنگامی‌که فراگیر جدید Y وارد سیستم می‌شود، سیستم با احتمال ۹۶.۴ درصد قادر به پیش‌بینی صحیح نتیجهٔ نهایی وی که قبولی می‌باشد، است. به منظور آزمایش و تست مجدد سیستم داده‌های مربوط به فراگیر Z را نیز که دورهٔ تحصیلی‌اش را به اتمام نرسانده و آزمون نهایی را انجام نداده، وارد سیستم نموده و نتیجه را در شکل ۵-۶ مشاهده می‌نماییم.

داده‌های مربوط به فراگیر Z به شرح زیر است:

Z,F,63,C,B,8,high,very_low,4,low,site,?

```
=== Predictions on test set ===

inst#,     actual, predicted, error, probability
    1          ?   2:failed              *0.687
```

شکل ۵-۶ احتمال پیش‌بینی نتیجهٔ نهایی فراگیر جدید Z

بدین ترتیب سیستم، در مورد فراگیر جدید Z با احتمال ۶۸.۷ درصد نتیجهٔ نهایی وی که مردود شدن از آزمون نهایی است را پیش‌بینی می‌نماید.

۵-۴ آزمایشات انجام شده در فاز دوم

همانطور که ذکر گردید، از الگوریتم k-means به منظور خوشه بندی فراگیران استفاده نمودیم. تعداد فراگیرانی که بر اساس رفتار و عملکردشان در محیط آموزش الکترونیکی خوشه‌بندی گردیدند، ۸۷ نفر می‌باشد. به عبارتی فقط فراگیران موفق را خوشه‌بندی نمودیم. در خصوص مقدار میانگین خوشه‌ها برای هر یک از ویژگی‌های مورد نظر و فراوانی خوشه‌ها در مرحلهٔ پیاده‌سازی بحث گردید. خروجی نرم افزار پس از عملیات clustering در شکل ۵-۷ قابل مشاهده می‌باشد.

```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 0.23818841628959275
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4
                  (87)       (17)       (15)       (13)        (8)       (34)
=============================================================================
Avg Grade      82.4253    98.5294       77.6    71.3846      58.75    86.2941


Clustered Instances

0      17 ( 20%)
1      15 ( 17%)
2      13 ( 15%)
3       8 (  9%)
4      34 ( 39%)
```

شکل۵-۷ خروجی عملیات خوشه‌بندی و میانگین خوشه‌ها

در این شکل مقدار میانگین خوشه‌ها را برای هر یک از ویژگی‌ها، همچنین تعداد و درصد فراگیران در هر خوشه را مشاهده می‌نماییم. به منظور برآورد میزان صحت عملیات خوشه‌بندی، به عبارتی تشخیص اینکه آیا هر فراگیر در خوشهٔ مربوط به خودش قرار گرفته است یا خیر؟ از الگوریتم J48 graft استفاده نموده و آنرا بر روی این ۸۷ نمونه (فراگیران موفق) اعمال می‌کنیم. در این بخش فیلد هدف به جای معدل نهایی، شمارهٔ خوشه می‌باشد.

از ۸۷ نمونهٔ موجود در مجموعه داده، تعداد ۸۴ عدد یعنی ۹۶.۵۵ درصد کلاس مربوطه به شکل صحیح طبقه‌بندی شده و یا به عبارتی در خوشهٔ مربوط به خود قرار گرفته‌اند (جدول ۵-۶). خروجی نرم افزار weka را نیز در شکل ۵-۸ مشاهده می‌نماییم. داده‌های جدول ۵-۶ از این شکل استخراج شده است.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          84               96.5517 %
Incorrectly Classified Instances         3                3.4483 %
Kappa statistic                          0.9539
Mean absolute error                      0.0138
Root mean squared error                  0.1174
Relative absolute error                  4.579  %
Root relative squared error             30.2902 %
Total Number of Instances               87

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.014      0.944     1         0.971      0.993    cluster1
                 0.933     0          1         0.933     0.966      0.967    cluster2
                 1         0.014      0.929     1         0.963      0.993    cluster3
                 0.875     0          1         0.875     0.933      0.938    cluster4
                 0.971     0.019      0.971     0.971     0.971      0.976    cluster5
Weighted Avg.    0.966     0.012      0.967     0.966     0.965      0.977
```

شکل ۵-۸خروجی نرم‌افزار پس از اعمال تکنیک j48 graft در خوشه‌های مختلف

جدول ۵-۶ نتایج بدست آمده از اعمال تکنیک j48 graft در خوشه‌های مختلف

مورد	میزان				
	خوشه۱	خوشه۲	خوشه۳	خوشه۴	خوشه۵
دقت	۰.۹۴۴	۱	۰.۹۲۹	۱	۰.۹۷۱
یادآوری	۱	۰.۹۳۳	۱	۰.۸۷۵	۰.۹۷۱
F-Measure	۰.۹۷۱	۰.۹۶۶	۰.۹۶۳	۰.۹۹۳	۰.۹۷۱
تعداد نمونه صحیح اختصاص داده شده	۱۷	۱۴	۱۳	۷	۳۳
Roc Area	۰.۹۹۳	۰.۹۶۷	۰.۹۹۳	۰.۹۳۸	۰.۹۷۶
دقت نهایی روش (درصد)	**۹۶.۵۵**				

۵-۴-۱ تحلیل خوشه‌ها

با استفاده از الگوریتم دسته‌بندی درصد پیش‌بینی صحیح تعلق یک فراگیر به یک خوشهٔ خاص محاسبه گردید. شکل ۵-۹ نمایانگر ماتریس آشفتگی حاصل از اعمال این الگوریتم در خوشه‌های مختلف می‌باشد.

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 17  0  0  0  0 |  a = cluster1
  0 14  0  0  1 |  b = cluster2
  0  0 13  0  0 |  c = cluster3
  0  0  1  7  0 |  d = cluster4
  1  0  0  0 33 |  e = cluster5
```

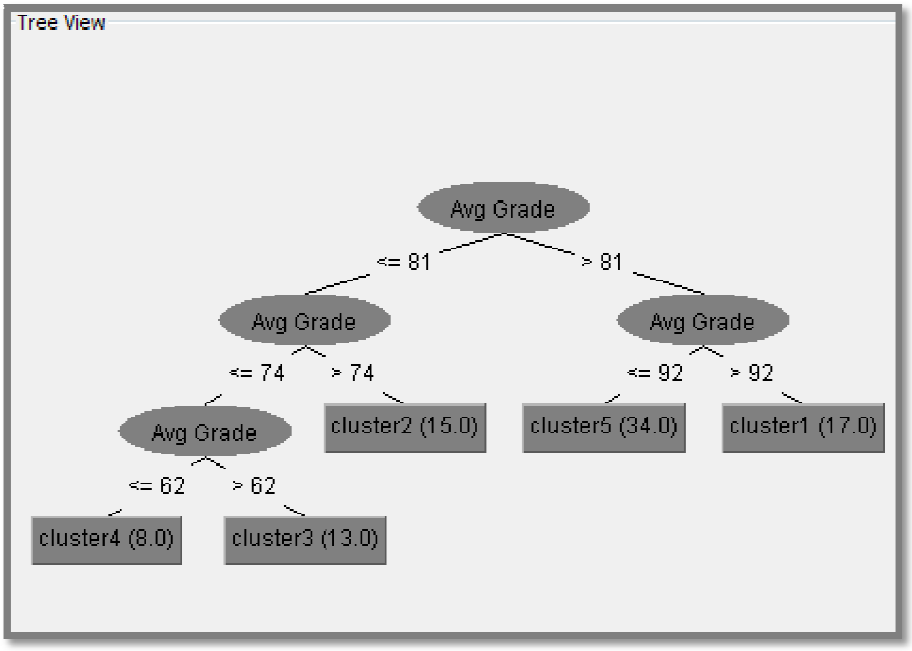
شکل ۵-۹ نمایش ماتریس آشفتگی حاصل از الگوریتم J48 graft در خوشه‌های مختلف

این بدان معناست که از ۱۷ فراگیری که متعلق به خوشهٔ یک بودند، مقدار پیش بینی معادل مقدار حقیقی بود و همهٔ آنها در خوشهٔ مربوط به خود قرار گرفتند. و از ۸ فراگیری که متعلق به خوشهٔ شمارهٔ چهار بودند، تعلّق ۷ نفر آنها به خوشهٔ مربوطه به درستی پیش‌بینی شده امّا یکی از آنها اشتباهاً به خوشهٔ شمارهٔ سه تخصیص داده شده است.

در نتیجهٔ اجرای مدل به دست آمده با توجه به ویژگی میانگین نمرات ترم قبل، قوانین بدست آمده از درخت تصمیم j48graft با فرض Cluster به عنوان فیلد هدف به شکل زیر خواهد بود:

```
Avg Grade <= 81
|   Avg Grade <= 74
|   |   Avg Grade <= 62: cluster4 (8.0)
|   |   Avg Grade > 62: cluster3 (13.0)
|   Avg Grade > 74: cluster2 (15.0)
Avg Grade > 81
|   Avg Grade <= 92: cluster5 (34.0)
|   Avg Grade > 92: cluster1 (17.0)
```

درخت بدست آمده بعد از هرس شدن در شکل۵-۱۰ نشان داده شده است. این درخت دارای ۵ برگ است که هر کدام از برگ‌ها نمایانگر یک خوشه و اندازه این درخت ۹ است.



شکل۵-۱۰ درخت بدست آمده از عملیات دسته‌بندی خوشه‌ها

۵-۴-۲ نتایج آزمایش فاز دوم

به منظور تست و ارزیابی عملکرد سیستم در این فاز، احتمال پیش‌بینی قبولی یک فراگیر قبل از اتمام دوره تحصیلی‌اش، در دو حالت گذراندن عادی دوره و عدم دریافت پیشنهادها و حالتی که در آن فراگیر پیشنهادهای سیستم توصیه گر را دریافت و به آن عمل نموده، مقایسه می‌گردد. به دلیل اینکه دسترسی به فراگیران و انجام آزمون مجدد از آنها امکان پذیر نبود، از این معیار (احتمال پیش بینی قبولی) به منظور بررسی بهبود عملکرد فراگیران استفاده نمودیم. واضح است که چون فراگیر موفق بوده در هر دو حالت، اتمام موفقیت آمیز دوره توسط وی پیش‌بینی می‌شود. هدف ما در این فاز بررسی افزایش احتمال قبولی وی پس از دریافت پیشنهادات و توصیه‌های سیستم است. نتایج حاکی از آن است که احتمال پیش بینی قبولی فراگیر افزایش یافته است. این احتمال قبولی برای فراگیران مختلف، متفاوت بوده، ولی در تمام موارد بهبودی را در درصد احتمال پیش‌بینی موفقیت آمیز دوره تحصیلی نشان می‌دهد.

پس از آنکه که فراگیر X وارد سیستم شد و سیستم با استفاده از تکنیک دسته‌بندی تشخیص داد که این فراگیر موفق است، خوشه‌ای که فراگیر بدان تعلق دارد مشخص می‌گردد.

به عنوان مثال داده‌های مربوط به فراگیر X به شرح زیر است:

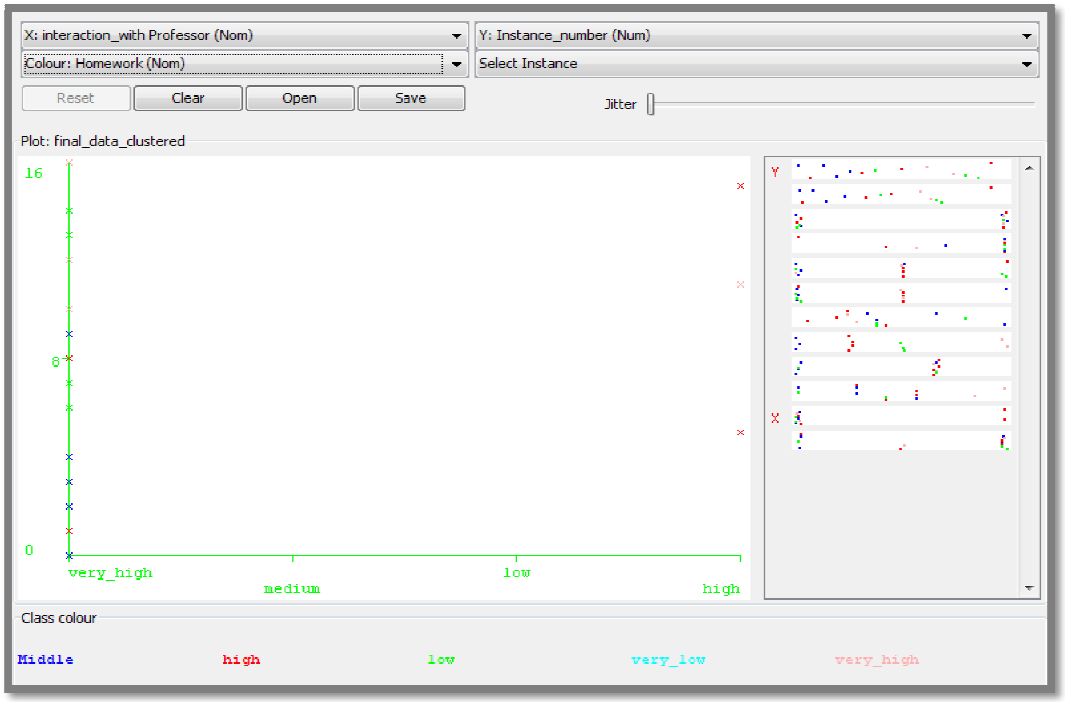
X,F,94,?,?,?,high,high,4,high,site,?

که علامت ؟ در انتهای رکورد فراگیر X، به معنای نامشخص بودن فیلد شمارهٔ خوشه می‌باشد. همچنین فیلد مربوط به فرم‌های تست اول و دوم و فیلد مربوط به درس مطالعه شدهٔ قبلی خالی بوده زیرا فراگیر در حین گذراندن دوره می‌باشد و هنوز این تست‌ها را انجام نداده است. با اعمال تکنیک‌های دسته‌بندی خوشهٔ مربوطه پیش‌بینی می‌گردد. (شکل۵-۱۱)

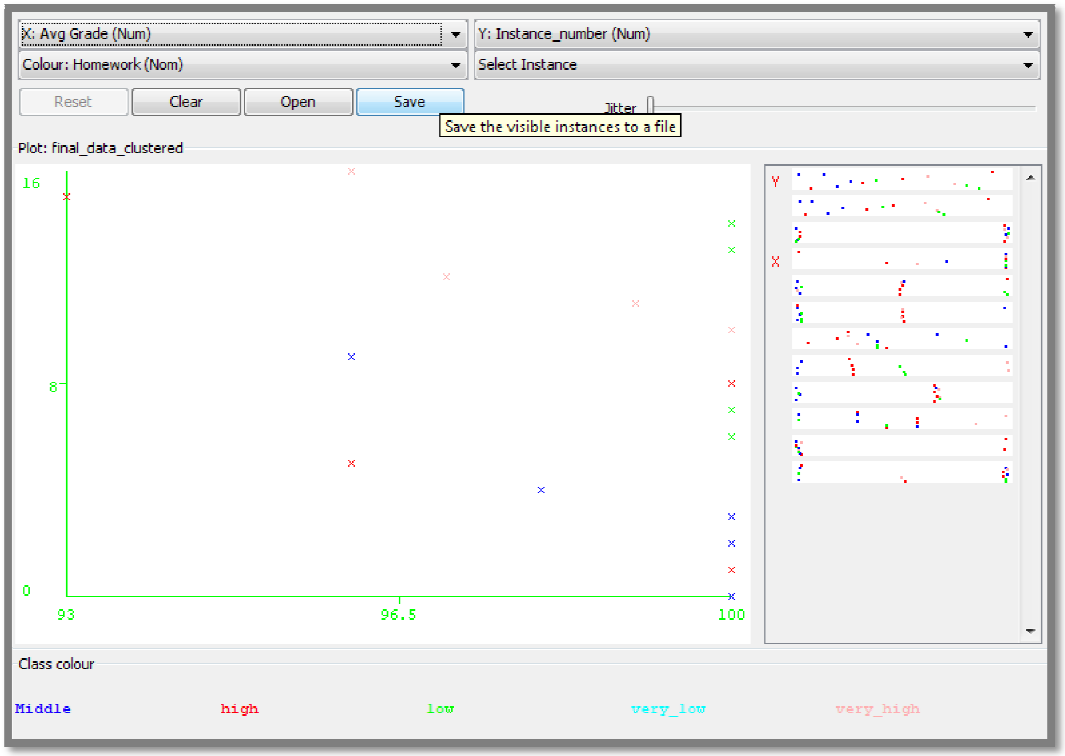
```
=== Predictions on test set ===

 inst#,     actual, predicted, error, probability
     1          ? 1:cluster1         *1
```

شکل ۵-۱۱ احتمال پیش‌بینی خوشهٔ فراگیر جدید X

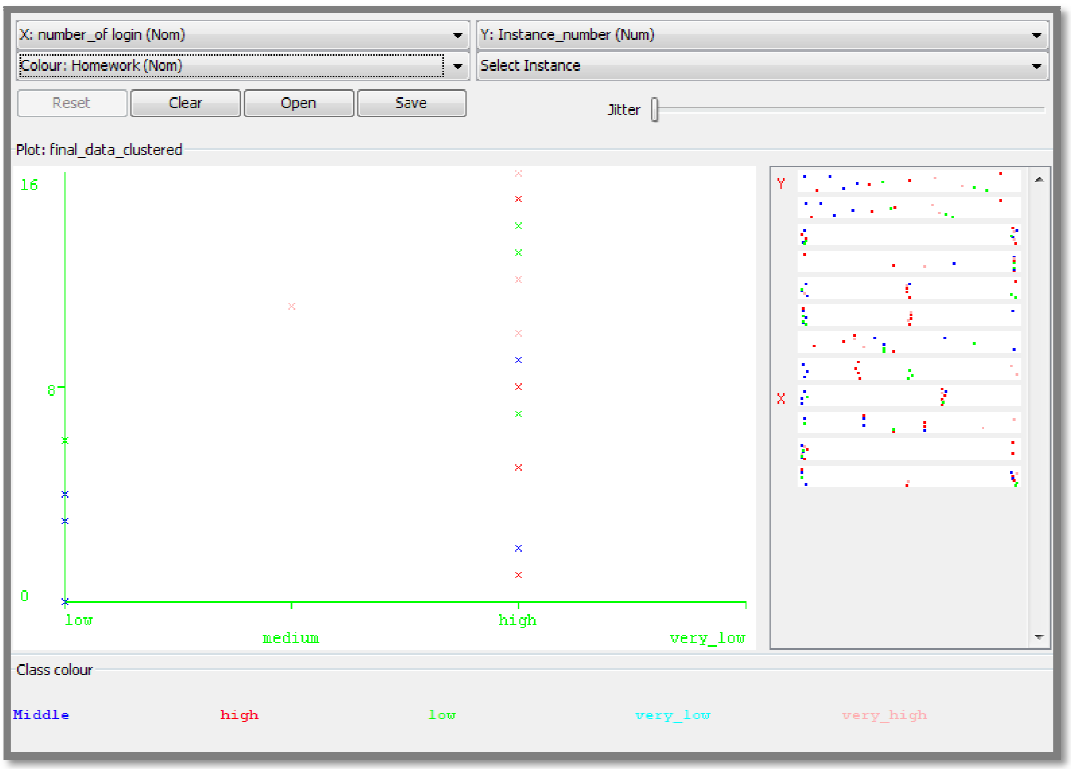
با توجه به رکورد و داده‌های فراگیر X، نرم افزار با احتمال ۱۰۰ درصد تشخیص می‌دهد که وی متعلق به خوشهٔ شمارهٔ یک است. در گام بعدی فعالیت و رفتار خوشه‌ای که فراگیر بدان تعلق دارد، بررسی می‌شود. زرنگترین و فعالترین فراگیران آن خوشه بر اساس مهمترین صفاتی که رابطهٔ مستقیم با موفقیت فراگیر دارد، مشخص می‌گردند. بدین ترتیب می‌بایست در خوشهٔ شمارهٔ یک، فراگیرانی مشخص گردند که صفات منتخب آنها از همه بهتر و بالاتر باشد. با توجه به نمودارهای موجود در نرم‌افزار Weka، زرنگترین فراگیران خوشه با توجه به صفاتی مانند میزان تعامل با استاد، میانگین نمرات ترم قبل و تعداد دفعات ورود به سایت برگزیده شده‌اند. (شکل۵-۱۲، شکل ۵-۱۳ و ۵-۱۴)



شکل ۵-۱۲ خروجی نرم‌افزار، نمودار وضعیت میزان تعاملات فراگیران با مربیان آموزشی در خوشهٔ یک



شکل ۵-۱۳ خروجی نرم‌افزار، نمودار وضعیت میانگین نمرات ترم قبل فراگیران خوشهٔ یک



شکل ۵-۱۴ خروجی نرم‌افزار، نمودار وضعیت تعداد دفعات ورود به سایت فراگیران در خوشهٔ یک

در این نمودارها محور X صفاتی مانند میزان تعامل با استاد، میانگین نمرات ترم قبل و تعداد دفعات ورود به سایت و محور Y نیز فراگیران را نشان می‌دهد. رنگ نمونه‌ها نیز بیانگر کلاس انجام تکالیف می‌باشد. به عنوان مثال فراگیرانی تکالیف خود را انجام نداده با رنگی آبی و فراگیرانی که تمام تکالیف خود را به نحو احسنت انجام داده با رنگ صورتی مشخص می‌شوند. و بازه‌های دیگر آن با رنگ‌هایی مانند قرمز و سبز که هر کدام نشان‌دهندهٔ میزان انجام تکالیف توسط فراگیران است قابل تشخیص و شناسایی می‌باشد. سپس فعالیت‌های موفق‌ترین فراگیران به وی (فراگیر X) پیشنهاد می‌گردد. به طور مثال ترتیب انجام آزمون‌ها، مطلب و فصل آموزشی مورد مطالعه، همچنین نحوهٔ مطالعهٔ دروس که از طریق سایت باشد یا کتب الکترونیکی و یا هر دو و هشدارهایی مانند افزایش تعامل با استاد و یا توصیهٔ تکالیف و تأکید بر انجام آنها به وی توصیه می‌گردد.

در نهایت درصد پیش‌بینی قبولی فراگیر X در دو حالت عادی و حالتی که فراگیر پیشنهادها را مد نظر قرار داده محاسبه می‌گردد. واضح است که در هر دو حالت، قبولی فراگیر در آزمون نهایی پیش‌بینی می‌گردد، زیرا همانطور که اشاره نمودیم این فاز مربوط به فراگیران موفق بوده و هدف ما مشاهدهٔ افزایش احتمال پیش‌بینی قبولی فراگیران می باشد. بدین ترتیب احتمال پیش‌بینی قبولی فراگیر X در حالت عادی (حالتی که فراگیر پیشنهادات را دریافت ننماید و به مسیر آموزشی خود ادامه دهد) محاسبه می‌گردد.

X,F,94,?,?,?,high,high,4,high,site,?

? در آخر رکورد فراگیر، به معنای نامشخص بودن وضعیت نهایی و نتیجهٔ نهایی وی در آزمون آخر دوره است. و علامت ? در سایر فیلدهای فراگیر X به معنای نامشخص بودن فیلد مربوط به برخی از فعالیت‌های وی می‌باشد. با اعمال تکنیک‌های دسته‌بندی و الگوریتم J48 graft نتیجهٔ نهایی وی پیش‌بینی می‌گردد. (شکل۵-۱۵)

```
=== Predictions on test set ===

 inst#,     actual, predicted, error, probability
     1          ?   1:passed          *0.865
```

شکل ۵-۱۵ احتمال پیش‌بینی قبولی فراگیر جدید X در حالت عادی

اگر فراگیر به مشاوره‌ها و توصیه‌های سیستم توجهی ننماید و به روند آموزشی معمولی خود ادامه دهد، از آنجائیکه وضعیت تحصیلی وی خوب بوده، با احتمال ۸۶.۵ درصد قادر به گذراندن دوره می‌باشد. در گام بعدی احتمال پیش‌بینی قبولی فراگیر X در حالتی که پیشنهادهای سیستم را مد

نظر قرار داده محاسبه می‌گردد. حال با توجه به نمودارهای مربوط به شکل ۵-۱۱، ۵-۱۲ و ۵-۱۳، و مشخص شدن ممتازترین فراگیران خوشهٔ یک و همچنین بررسی لاگ فایل‌های مربوط به این فراگیران توسط سیستم، به وی توصیه می‌گردد که ابتدا فرم تست A را انجام دهد و سپس فرم تست C، همچنین توصیه می‌گردد که قبل از آزمون نهایی درس شمارهٔ ۲۰ را مطالعه نماید. این هشدارها و پیشنهادات می‌تواند از طریق باز شدن پنجره‌ای در هنگام ورود فراگیر به سیستم به وی نشان داده شود. در این صورت عملکرد فراگیر مورد نظر (رکورد فراگیر)، با جایگزینی مقادیر جدید در فیلدهای مورد نظر بدین صورت تغییر می‌یابد:

X,F,94,A,C,20,high,high,4, high,site,?

بار دیگر نتیجهٔ آزمون نهایی وی پیش‌بینی می‌گردد. (شکل ۵-۱۶)

```
=== Predictions on test set ===

 inst#,     actual, predicted, error, probability
     1          ?   1:passed        *0.964
```

شکل ۵-۱۶ احتمال پیش‌بینی قبولی فراگیر جدید X با ملاحظه و دریافت پیشنهادات سیستم

همانطور که ملاحظه می‌گردد احتمال قبولی فراگیر X به ۹۶.۴ درصد افزایش یافته است. نتایج آزمایشات حاکی از آن است که فراگیرانی که زیر نظر سیستم توصیهٔ ما بوده‌اند، با توجه به مطالب آموزشی اصلاحی و مشاوره‌هایی که بدنبال پیشنهادات سیستم ارائه می‌گردد، در عملکرد یادگیری خود پیشرفت داشته‌اند.

۵-۵ آزمایشات انجام شده در فاز سوم

از الگوریتم Apriori به‌منظور کاوش قوانین انجمنی استفاده نمودیم. این الگوریتم را به کل مجموعه داده و به عبارتی کل فراگیران موجود در سیستم اعمال می‌کنیم. از آنجائیکه الگوریتم Apriori قادر به کشف قوانین با مقادیر عددی نمی‌باشد. لذا فیلدهای مربوط به ویژگی‌های دارای مقادیر پیوسته را تبدیل به مقادیر گسسته می‌نماییم.

در نرم افزار weka می‌توان به لیستی از فیلترها دست یافت. می‌توان از فیلترها برای حذف ویژگی‌های مورد نظر از یک مجموعه داده و یا انتخاب دستی ویژگی‌ها استفاده نمود. یکی دیگر ازفیلترهای موجود، Discretize است که با استفاده از آن می‌توان مقادیر یک صفت پیوسته را به

تعداد دلخواه بازه گسسته تبدیل کرد. جدول ۵-۷ تقسیم مقادیر صفت Ave Grade را به ۱۰ بازه A، B ، C، D، F، G، H ، I و J نشان می‌دهد.

جدول ۵-۷ تقسیم مقادیر صفت میانگین نمرات ترم‌های قبل

مقادیر	بازه	فراوانی
۹۴ –۱۰۰	A	۱۷
۸۸ –۹۴	B	۱۰
۸۲ –۸۸	C	۲۴
۷۶ –۸۲	D	۱۶
۷۰ –۷۶	E	۱۵
۶۴ –۷۰	F	۷
۵۸ –۶۴	G	۹
۵۰ –۵۸	H	۵
۴۶ –۵۰	I	۲
۴۰ –۴۶	J	۱

جدول ۵-۸ نیز بیانگر تقسیم مقادیر صفت Unit before study به ۴ بازهٔ گسستهٔ ch1، ch2 ch3 و ch4 می‌باشد. مقادیر دروس از ۲ تا ۳۴ بوده که ما آنرا به ۴ فصل تقسیم نمودیم. در حقیقت، هر کدام از این بازه‌ها فصول مطالعه شده توسط فراگیر را نشان می‌دهند.

جدول ۵-۸ تقسیم مقادیر صفت دروس مطالعه شده

مقادیر	بازه	فراوانی
۲۶ –۳۴	Ch4	۵
۱۸ –۲۶	Ch3	۱۹
۱۰ –۱۸	Ch2	۵۴
۲–۱۰	Ch1	۲۸

۵-۵-۱ تحلیل قوانین استخراج شده

نمونه ای از این قواعد در شکل شماره ۴-۹ در فصل قبل نشان داده شده است. قواعد با ضریب اطمینان بالا از اهمیت بیشتری برخوردارند.

این قوانین ارتباطات جالب و پنهان را مابین خصایص مجموعه داده نشان می دهند. آنها امکان پیدا کردن روابطی شبیه اگر←مقدمه←در اینصورت←تالی، را می دهند که مقدمه و تالی خصایصی از

مجموعه داده هستند. ما دنبال خصایصی می گردیم که وضعیت نهایی فراگیر را مشخص و تعیین می کنند تالی این قوانینx= نمره نهایی می‌باشد که x یک مقدار برای نمرهٔ نهایی مثل قبول و مردود می باشد.

جدول ۵-۹ قوانینی را نشان می‌دهد که برای نمرهٔ نهایی «قبول» فراگیران استخراج گردیده اند. این قوانین بر اساس مقدار اطمینان مرتب گردیده اند. مقدار اطمینان کسری از موارد است که وقتی مقدم قانون در آنها ظاهر شده است، تالی نیز در آنها وجود دارد. به عنوان مثال اولین قانون جدول ۵-۹ با مقدار اطمینان یک به این معنی است که یک ارتباط مستقیم قوی ما بین مقدم تعامل بالا با استاد با تالی نمره نهایی قبولی برقرار است.

جدول ۵-۹ قوانین وابستگی داده‌های فراگیران با تالی نمرهٔ نهایی قبولی

شماره قانون	اطمینان	مقدم
۱	۱	میزان تعامل با استاد = بسیار بالا
۲	۱	دفعات ورود به سیستم = بالا // میزان تعامل با استاد = بسیار بالا
۳	۱	دروس مطالعه شده = دروس فصل ۳
۴	۱	فرم آزمون دوم = C // دفعات ورود به سیستم = بالا
۵	۱	دروس مطالعه شده = دروس فصل ۲ // میزان تعامل با استاد = بالا
۶	۱	فرم آزمون نخست = A // فرم آزمون دوم = C
۷	۰.۹۶	میزان تعامل با استاد = بالا
۸	۰.۹۵	دروس مطالعه شده = دروس فصل ۲ // دفعات ورود به سیستم = متوسط
۹	۰.۹۵	فرم آزمون نخست = A // منبع مطالعه دروس = کتب الکترونیکی
۱۰	۰.۹۴	میانگین نمرات ترم قبل = A (بین ۹۴ تا ۱۰۰)
۱۱	۰.۹۴	فرم آزمون دوم = A // دفعات ورود به سیستم = متوسط
۱۲	۰.۹۴	فرم آزمون دوم = C
۱۳	۰.۹۲	فرم آزمون نخست = A // دفعات ورود به سیستم = بالا
۱۴	۰.۹۲	میانگین نمرات ترم قبل = C (بین ۸۲ تا ۸۸)

۵-۵-۲ نتایج آزمایش فاز سوم

حال به منظور تست سیستم، نتیجهٔ آزمون نهایی یک فراگیر مشکوک به رد شدن از آزمون نهایی را پس از دریافت پیشنهادات و به عبارتی اعمال یکی از قوانین کشف شده به مجموعه فعالیت‌ها و عملکرد وی می‌سنجیم. رکورد فراگیر مورد نظر به شرح زیر است:

Y,F,D,?,?,ch2,very_high,medium,3,medium,both

این عبارت بدین معناست که جنسیت فراگیر Y دختر و میانگین نمرات ترم قبلش بین ۷۶ تا۸۲ بوده، وی هنوز تستی را انجام نداده، فصل مطالعه شدهٔ قبل از آزمون نهایی فصل ۲ که شامل دروس ۱۰ تا ۱۸ بوده، می‌باشد. تکالیف خود بسیار خوب و کامل انجام می‌دهد و تعداد دفعات ورود به سیستم آموزشی متوسط بوده، در هر بار ورود تقریباً به مدت ۳ ساعت به مطالعه و گشت و گذار در سایت مشغول است. میزان تعاملش با استادان و مربیان آموزشی متوسط بوده و دروس را هم از طریق سایت آموزشی و هم از طریق کتب الکترونیکی که به وی معرفی می‌گردد مطالعه می‌نماید.

پیش‌بینی می‌گردد که فراگیر Y در آزمون نهایی رد شود. حال با توجه به قانون کشف شدهٔ شمارهٔ ۶ به وی توصیه می‌گردد که ابتدا فرم تست A را انجام دهد و سپس فرم تست C. در این صورت عملکرد فراگیر مورد نظر (رکورد فراگیر) بدین صورت تغییر می‌یابد:

Y,F,D,A,C,ch2,very_high,medium,3,medium,both

این نمونه را وارد نرم افزار weka نموده و خروجی آن را مشاهده می‌نماییم. (شکل ۵-۱۷)

بدین ترتیب که با استفاده از الگوریتم‌های دسته‌بندی و ترجیحاً استفاده از الگوریتم J48graft بار دیگر نتیجهٔ آزمون نهایی وی پیش‌بینی می‌گردد.

```
=== Predictions on test set ===

inst#,     actual, predicted, error, probability
    1          ?   1:passed          *0.865
```

شکل ۵-۱۷ خروجی نرم افزار weka (پیش‌بینی نتیجهٔ نهایی فراگیر ضعیف)

همانطور که در خروجی نرم افزار مشاهده می‌گردد، فراگیر Y که فراگیری ضعیف پیش‌بینی شده بود و احتمال گذراندن موفقیت آمیز دوره را نداشت، با دریافت پیشنهادات ارائه شده به کمک قوانین وابستگی، اکنون با احتمال ۸۶.۵٪ قادر به قبولی از آزمون نهایی خواهد گردید. حال با توجه به قانون کشف شدهٔ شمارهٔ ۷ به وی توصیه می‌گردد که میزان تعامل و ارتباطش را با استادان و مربیان آموزشی بالا ببرد. در این صورت عملکرد فراگیر مورد نظر بدین ترتیب تغییر می‌یابد:

Y,F,D,A,C,ch2,very_high,medium,3,high,both

بار دیگر نتیجهٔ آزمون نهایی وی پیش‌بینی می‌گردد. (شکل ۵-۱۸)

```
=== Predictions on test set ===

 inst#,     actual, predicted, error, probability
     1          ?   1:passed        *0.964
```

شکل ۵-۱۸ خروجی نرم افزار weka (پیش‌بینی نتیجهٔ نهایی فراگیر ضعیف)

با توجه به خروجی نرم‌افزار ملاحظه می‌نماییم که احتمال قبولی فراگیر Y از ۸۶.۵٪ به ۹۶.۴٪ افزایش می‌یابد. نتایج بیانگر بهبود قابل توجهی در وضعیت فراگیرانی است که از پیشنهادهای ارائه شده توسط سیستم استفاده نمودند. بدین ترتیب فراگیری که تاکنون رد شدنش از آزمون نهایی پیش‌بینی می‌گردید، اکنون با دریافت پیشنهادات سیستم قبل از اتمام دوره تحصیلی، قادر به جلوگیری از افت تحصیلی‌اش خواهد بود.

فصل۶- نتیجه‌گیری و پیشنهادات

۶-۱ بحث و نتیجه‌گیری

سیستم‌های توصیه‌گر به طور گسترده در بسیاری از فعالیت های اینترنتی عمدتاً جهت غلبه بر مشکل حجم بالای اطلاعات به کار گرفته می شوند. برخی از این فعالیت‌ها مربوط به سایت‌های تجارت الکترونیک، جستجوی صفحات وب، پرتال های خبری، کتابخانه های دیجیتال و سیستم های سانسور می باشد. علاوه بر این، محققان در خصوص استفاده از سیستم های توصیه گر در راه حل های آموزش الکترونیکی، تحقیقاتی انجام داده اند.

سیستم توصیه گر آنچه را که فراگیر و یا کاربر انجام می دهد را مشاهده نموده و سعی در پیشنهاد فعالیت ها و تکالیف سودمند به وی دارد. مدل ایجاد شده با نمایش رفتار فراگیر آنلاین، قادر به پیشنهاد فعالیت ها، تکالیف و میانبرهایی به وی می باشد. این پیشنهادات فراگیر را در گذراندن مواد و واحدهای درسی هدایت نموده، منابع مرتبط و مفید را با سرعت و دقت بالا به وی معرفی کرده، سردرگمی و ترس فراگیر را کاهش داده و با توصیهٔ تکالیف مناسب با توجه به رفتار فراگیر انگیزه‌اش را افزایش می‌دهد. همچنین با شناسایی نقاط ضعف و قوت فراگیران و پشتیبانی آنها از طریق ارائهٔ محتوای تکمیلی (مواردی همچون تنظیم و ارائهٔ مثالها و تمرینات مرتبط از یک یا چندین منبع)، تعاملات فراگیران را در محیط یادگیری ردگیری نموده و مدل رفتار وی شناسایی خواهد شد. سپس فرآیند آموزش توسط سیستم‌های توصیه‌گر ادامه پیدا می‌کند. بنابراین فرآیند یادگیری در محیط‌های آموزش الکترونیکی بهبود می‌یابد.

۶-۲ نتایج حاصل از پژوهش

این پژوهش به مطالعه و جستجوی قابلیت استفاده از سیستم‌های توصیه‌گر در محیط‌های آموزش الکترونیکی پرداخته و مدلی را جهت توصیهٔ مفاد و فعالیت‌های آموزشی مناسب به فراگیر فعال قبل

از اتمام دوره‌اش طراحی نموده است. الگوریتم پیشنهادی از چندین رویکرد داده‌کاوی تشکیل می‌شود: دسته‌بندی، خوشه‌بندی و کاوش قوانین وابستگی.

روش پیشنهادی از این قرار است که ابتدا فراگیران به لحاظ سطح علمی‌اشان دسته‌بندی می‌گردند. بدین ترتیب پس از ورود فراگیر جدید، سیستم قادر به پیش بینی نتیجهٔ نهایی وی قبل از اتمام دورهٔ تحصیلی‌اش، با توجه به رفتار وی و عملکرد فراگیران مشابه و موجود در سیستم خواهد بود. در صورتی که سیستم تشخیص دهد فراگیر قوی بوده و قادر به اتمام موفقیت آمیز دوره می‌باشد، از یک رویکرد (خوشه‌بندی) و در صورت تشخیص ضعف و یا ناموفق بودن فراگیر، از یک رویکرد دیگر (کاوش قوانین وابستگی) که خاص فراگیران ضعیف می‌باشد، استفاده می‌نماید.

در حقیقت، این تحقیق تلاشی برای پیاده سازی مدل‌های داده کاوی پیش بینی کننده، به منظور پیش بینی وضعیت تحصیلی دانشجویان براساس رفتار و عملکرد تحصیلی آنان بوده است. با توجه به نتایج آماری که از ساخت مدل‌های پیش بینی کننده وضعیت دانشجو در این تحقیق بدست آمده است، می‌توان با اطمینان بالایی از آیندهٔ تحصیلی دانشجویان بر مبنای داده‌های گذشتهٔ آنها اطلاع حاصل نمود. دقت و صحت این مدل‌های پیش بینی کننده در این تحقیق، بر روی مجموعه دادهٔ فراگیران مربوط به یک سیستم آموزش الکترونیک آزموده شده و به اثبات رسیده است. اگرچه نتایج این تحقیق منتهای هدف در این حوزه نیست و می توان با بسط مدل‌ها و درگیر کردن پارامترهای جدید به نتایج دقیق‌تر و قابل اطمینان‌تری دست یافت. میزان اهمیت این پیش بینی‌ها همانطور که پیش از این گفته شد بسیار واضح و روشن است. بنابراین نتایج حاصل از این تحقیق و موارد مشابه می‌تواند به صورت جدی در سایت‌ها و مؤسسات آموزشی آنلاین پیاده‌سازی و مورد استفاده قرار گیرد. استفاده از چنین مدل‌هایی این محیط‌ها را در ارتقاء سطح علمی فراگیران و هدفمند نمودن محیط‌های آموزشی آنلاین یاری می‌نماید.

در این پژوهش، تحقیقات خود را بر روی سیستم آموزش الکترونیکی شخصی شده شرح دادیم. طرح اولیهٔ سیستم و نتایج آن نیز نشان داده شد. نتایج بیانگر بهبود قابل توجهی در وضعیت فراگیرانی است که از پیشنهادهای ارائه شده توسط سیستم استفاده نمودند. سیستم با پیش‌بینی نتیجهٔ تحصیلی فراگیر قبل از اتمام دوره تحصیلی‌اش، قادر خواهد بود که برنامه آموزشی متناسب با فراگیر را به وی ارائه نماید، از افت تحصیلی فراگیران ضعیف جلوگیری نموده و همچنین شکوفایی بیشتر فراگیران مستعد را فراهم آورد.

۳-۶ دستاوردهای پژوهش

مهم‌ترین دستاورد پروژه ارائهٔ یک سیستم جدید برای ساخت مدل فراگیر در یک محیط آموزش الکترونیکی می‌باشد. در این راستا، سایر دستاوردهای پروژه عبارتند از:

- طراحی سیستمی تلفیقی در وب سایت‌های آموزشی با استفاده از تکنیک‌های داده‌کاوی، به منظور ارائهٔ محتویات آموزشی‌ای که متناسب با نیاز هر فراگیر باشد. معماری پیشنهادی در این پایان‌نامه، حاصل مطالعه و بررسی کارهای مشابه و شناسایی ایرادات آن‌ها و سعی در برطرف کردن آن‌ها در طراحی سیستم بوده است.
- دسته‌بندی فراگیران بر اساس سطح علمی و امکان بهبود فرایند یادگیری یک فراگیر با توجه به فعالیت سایر فراگیران. با دسته‌بندی فراگیران به دو گروه موفق و نا موفق، کیفیت پیشنهادات ارائه شده به آنها را بالا می‌بریم.
- معرفی و انتخاب مدل j48graft به عنوان یک الگوریتم قابل فهم و با دقت پیش‌بینی بالا در حوزهٔ شخصی سازی محیط‌های آموزش الکترونیکی: در تحقیقات انجام شده در این پایان‌نامه دقت الگوریتم‌های گوناگون مورد بررسی قرار گرفت (رجوع کنید به فصل۴و۵) و نهایتاً الگوریتم درختی j48graft به عنوان مدل منتخب معرفی شد.
- خوشه‌بندی فراگیران موفق با توجه به وضعیت تحصیلی‌اشان. در سیستم‌های توصیه‌گر موجود در سایر حوزه‌ها و همچنین آموزش الکترونیک فعالیت و عملکرد کاربران کل خوشهٔ مربوط، به کاربر جاری پیشنهاد می‌شود. اما به منظور بالا بردن کارایی و بهبود وضعیت آموزشی فراگیران در این طرح پیشنهاد می‌شود که عملکرد بهترین فراگیران موجود در خوشهٔ مربوطه، به فراگیر جاری ارائه گردد.
- با کشف روابط موجود و با استفاده از آنالیز قوانین همبستگی در این پژوهش، عملکرد کل فراگیران مجموعه داده مورد تحلیل قرار گرفته و ارتباطات و همبستگی‌های بین دروس و فعالیت‌های آموزشی با موفقیت و قبولی فراگیران کشف گردید. روابط و قوانین کشف شده به سیستم در خصوص شناخت عوامل مؤثر و همچنین ارائهٔ پیشنهادات به فراگیرانی که دچار افت تحصیلی شده و قادر به اتمام موفقیت آمیز دوره نبوده، کمک شایانی می‌کند. به توصیهٔ فعالیت‌های آموزشی مناسب به فراگیر جاری پرداخته، از افت تحصیلی وی جلوگیری نموده و همچنین امکان شکوفایی بیشتر وی را فراهم می‌آوریم.

۶-۴ کارهای آینده

- یکی از کارهایی که بلافاصله می‌توان از نتایج این پایان نامه انجام داد تعبیهٔ مدل بدست آمده از این روش در یک سیستم توصیه‌گر آموزش الکترونیک و ارزیابی توصیه‌های حاصل از آن می‌باشد. در این پایان‌نامه به دلیل محدودیت مجموعه داده‌ها امکان انجام این کار میسر نشد.
- امکان پیشنهاد ثبت نام دوره که به فراگیر جهت انتخاب دوره کمک می نماید.
- امکان افزایش دقت روش‌های پیش بینی به کمک ترکیب مدل‌ها و الگوریتم‌های مختلف.

فهرست منابع و مراجع

اسماعیلی، مهدی، طرفدار، منصور(۱۳۸۸). "ارائه و بررسی روشی مؤثر جهت انتخاب صفات خاصهٔ مناسب در ساخت درخت تصمیم"، *سومین کنفرانس داده کاوی ایران*.

احمدی، حسن، شاکری اسکی، بهارک، علیشاهی، محمد(۱۳۸۷). "غنی سازی محیط های آموزش الکترونیکی با استفاده از تکنیک های طبقه بندی داده"، *دومین کنفرانس داده کاوی ایران*.

ایرجی، اعظم، مینایی، بهروز، شکورنیاز، ونوس(۱۳۸۷). " استخراج قوانین تصمیم با استفاده از الگوریتم درخت تصمیم جهت هدایت تحصیلی دانش‌آموزان به کمک دسته بندی داده‌های آموزش و پرورش"، *دومین کنفرانس داده کاوی ایران*.

ایرجی، اعظم، مینایی، بهروز، شکورنیاز، ونوس(۱۳۸۷). " بکارگیری فن آوری داده کاوی به منظور آسیب شناسی افت تحصیلی هنرجویان هنرستانی و استخراج نمایه ساز توصیفی در ارائهٔ تمایز دانش آموزان ضعیف و ممتاز"، *دومین کنفرانس داده کاوی ایران*.

ایرجی، اعظم، مینایی، بهروز، شکورنیاز، ونوس(۱۳۸۷). "بکارگیری داده کاوی برای کشف تاثیرعامل جنسیت و مدرسه در موفقیت تحصیلی رشته‌های مختلف"، *دومین کنفرانس داده کاوی ایران*.

بابائی، مریم، صفار یزدی، زهرا، سرائی، محمدحسین(۱۳۸۷). "معرفی و مقایسه روش‌های پیش‌پردازش داده برای کاربردهای مختلف داده‌کاوی"، *دومین کنفرانس داده کاوی ایران*.

برهان، احسان، عظیمی، کاوه، شهرابی، جمال(۱۳۸۶). "بکارگیری قوانین همبستگی و آنالیز کلاستر در آسیب شناسی افت تحصیلی دانشجویان دانشگاه صنعتی امیرکبیر"، *اولین کنفرانس داده کاوی ایران*.

پاینده‌فر، هومن، سید رضی، حسن، رهگذر، مسعود، فراهی، احمد(۱۳۸۷). "مقایسهٔ تکنیک‌های داده‌کاوی جهت شخصی‌سازی مطلوب‌تر در آموزش الکترونیک"، *دومین همایش ملّی مهندسی کامپیوتر، برق و فناوری اطلاعات*.

حاتم لو، عبدالرضا، هاشمی نژاد، سید جواد(۱۳۸۷). "تحلیل رفتار آموزشی دانشجویان با استفاده از تکنیک های داده کاوی"، *دومین کنفرانس داده کاوی ایران*.

خادم القرانی، فریبا، سرائی، محمد حسین، مصطفوی، سید ابولفضل(۱۳۸۸). "کاربرد داده‌کاوی در هدفمند کردن انتخاب رشتهٔ دانشگاهی و بهبود کیفیت برنامه‌ریزی آموزشی"، *سومین کنفرانس داده کاوی ایران*.

خنشا، سمیرا، صدرالدینی، محمد هادی(۱۳۸۷). "بررسی تکنیک‌های ترکیبی کاوش وب برای شخصی‌سازی مؤثرتر"، *دومین کنفرانس داده کاوی ایران*.

رحمانی، سمیّه، میبدی، محمدرضا(۱۳۸۸). "شخصی‌سازی وب با استفاده از قوانین انجمنی توسعه یافته"، *سومین کنفرانس داده کاوی ایران*.

زمانیان، مهناز، مقدم چرکری، نصرالله(۱۳۸۷). "بررسی رفتار و بخش بندی مشتریان با استفاه از داده‌کاوی در دانشگاه‌ها"، *دومین کنفرانس داده کاوی ایران*.

شکورنیاز، ونوس، حاجی علی اکبری، آرش(۱۳۸۷). "خوشه بندی داده‌های آماری دانشجویان دانشگاه علم و صنعت و استخراج نمایه ساز توصیفی برای دانشجویان موفق"، *دومین کنفرانس داده کاوی ایران*.

صحافی زاده، ابراهیم(۱۳۸۸). "تحلیل عوامل مؤثر بر نمرات دانشجویان دانشگاه پیام نور بوشهر با استفاده از داده‌کاوی"، *سومین کنفرانس داده کاوی ایران*.

ضرغامی، شیرین، عمادی، سید پیمان(۱۳۸۸). "بررسی مشکلات Collaborative Filtering مبتنی بر شباهت و ارائهٔ راهکارهایی در این زمینه"، *سومین کنفرانس داده کاوی ایران*.

طهماسبی، حمیدرضا، احمدی، حسن(۱۳۸۸). "افزایش دقت کلاسه بندی در داده کاوی با استفاده از ترکیب کلاسه بندها"، *سومین کنفرانس داده کاوی ایران*.

غضنفری، مهدی، فتحیان، محمد، دشتی، یاسر(۱۳۸۶). "استخراج الگوی حرکتی کاربران وبگاه با استفاده از تکنیک خوشه بندی و ارائه ساختار ماشین پیشنهاد دهنده"، *پنجمین کنفرانس بین‌المللی مهندسی صنایع*.

قادریان، میثم(۱۳۸۷). " بهبود مدل کاربر در وب‌سایت بصورت خودکار با استفاده از معناشناسی با مفاهیم خاص دامنه"، تحت راهنمایی دکتر احمد عبداله زاده بارفروش، دانشگاه صنعتی امیرکبیر، دانشکده مهندسی کامپیوتر و فناوری اطلاعات.

کائدی، مرجان، برآنی، احمد، قاسم آقایی، ناصر(۱۳۸۷). "تخمین موفقیت کاربران آموزش الکترونیکی با استفاده از ترکیب دسته بندی کنندهٔ بیز و روش k-NN"، *دومین کنفرانس داده کاوی ایران*.

کیوان پور، محمدرضا، خلعتبری، لیلا(۱۳۸۸). "مقایسهٔ الگوریتم‌های کلاس بندی در تشخیص دیابت و نارسایی قلبی"، *سومین کنفرانس داده کاوی ایران*.

مظهری، نیلوفر، ایمانی، مهدی، جودکی، مجید، قلیچ پور، احمد(۱۳۸۸). "مروری بر کلاسه‌بندی و الگوریتم‌های آن"، *سومین کنفرانس داده کاوی ایران*.

نصیری، مهدی، کاردان، نوید، هادیان، علی، مینایی، بهروز(۱۳۸۸). "مقایسهٔ روش‌های نزدیکترین همسایهٔ مجاور، SVM، C۴.۵ و نیو-بیز برای دسته‌بندی داده‌ها"، *سومین کنفرانس داده کاوی ایران*.

یقینی، مسعود، حیدری، سمیّه(۱۳۸۷). "داده‌کاوی جهت ارتقاء و بهبود فرآیندهای سیستم آموزش عالی"، *دومین کنفرانس داده کاوی ایران.*

یقینی، مسعود، اکبری، امین، شریفی، سید محمد مهدی(۱۳۸۷). "پیش بینی وضعیت تحصیلی دانشجویان با استفاده از تکنیک های داده کاوی"، *دومین کنفرانس داده کاوی ایران.*

یقینی، مسعود، اکبری، امین، شریفی، سید محمد مهدی(۱۳۸۷). "دسته بندی دانشجویان و استخراج روابط موجود در سیستم آموزشی دانشگاه"، *دومین کنفرانس داده کاوی ایران.*

Al-Radaideh, Q.A., Al-Shawakfa, E.M., Al-Najjar, M.I. (2006). "Mining Student Data Using Decision Trees", *the 2006 International Arab Conference on Information Technology*.

Adomavicius, G.,Tuzhilin, A. (2005). "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", *IEEE Transactions on Knowledge and Data Engineering*, Vol.17,No.6.

[Aftab]http://www.aftab.ir/articles/science_education/education_training/c3c1181216248p1.php.

Azene, Z., Ozok. A., Norcio. A.F. (2005). "Personalized Recommender Systems in e-commerce and m-commerce: A Comparative Study", Department of Information Systems, University of Maryland Baltimore County (UMBC) Baltimore, MD 21250 USA.

Baker, R.S.J.d. (in press) Data Mining for Education. To appear in McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier.

Balabanovic, M., Shoham, Y. (1997). "Fab: Content-Based, Collaborative Recommendation", *Comm. ACM*, Vol.40, No.3, 62–72.

Bobadilla, J., Serradilla, F., Hernando, A., & MovieLens. (2009). Collaborative filtering adapted to recommender systems of e-learning. *Knowledge-Based Systems*, 22, 261–265.

Breese, J.S., Heckerman, D., Kadie, C. (1998). "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", *Proc. 14th conf. Uncertainty in Artificial Intelligence*.

Burdescu, D.D., Mihăescu, M., C. (2006). "How Learner's Proficiency May Be Increased Using Knowledge about Users within an E-Learning Platform", *Informatica*, 30, 433–438

Carenini, G., Smith, J., & Poole, D. (2003). Towards more conversational and collaborative recommender systems. In Proc IUI'03 (pp. 12–18). Miami, FL.

Chen, A.Y., McLeod, D. (2004). "Collaborative Filtering for Information Recommendation Systems", *Encyclopedia of e-Commerce, e-Government and m-Commerce*, Idea Group Inc.

Chen, H. C. & Chen, A. L. P. (2001). A music recommendation system based on music data grouping and user interests. In Proc. CIKM'01 (pp. 231–238). Atlanta, Georgia.

Cho, Y. H., Kim, J. K., & Kim, S. H. (2002). A personalized recommender system based on web usage mining and decision tree induction. *Expert Systems with Applications*, 23, 329–342.

Clare-Marie, K., Jan, B. (2004). "Designing Personalized User Experiences in E-Commerce", Boston: Kluwer Academic Publishers.

Condliff, M.K., Lewis, D.D. (2000). "Bayesian Mixed-Effects Models for Recommender Systems", *AT&T Labs-Research*.

Das, A., Datar, M., Grag, A. (2007). "Google News Personalization: Scalable Online Collaborative Filtering", Banff, Alberta, Canada.

Dastani, M., Jacobs, N., Jonken C.M., & Treur, J. (2005). "Modeling user preferences and mediating agents in electronic commerce", *Knowledge-Based Systems* 18, 335-352.

Dash, M., Liu, H. (1997). "Feature selection for Classification", *In: Intelligent Data Analysis*, Vol. 1, No. 3.

Dekker, G.W. (2009). "Predicting students drop out: a case study". In T. Barnes, M. Desmarais, C. Romero, S. Ventura (Eds.), *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009, July 1-3, 2009. Cordoba, Spain. (pp. 41-50).*

Diamond Bullet (2004). Provider of usability & web design services. Retrieved May 20, 2004, from the World Wide Web: http://www.usabilityfirst.com/glossary/term_962.txl.

Esposito, F., Licchelli, O., Semeraro, G. (2004). "Discovering Student Models in e-learning Systems". *Journal of Universal Computer Science*, 10(1), 47–57.

Felfernig, A., Friedrich, G., & Schmidt-Thieme, L. (2007). "Introduction to the IEEE Intelligent Systems Special Issue: Recommender Systems", 22(3), 18-21.

Gauch, J. M., Gauch, S., Bouix, S., & Zhu, X. (1999). "Real time video scene detection and classification". *Information Processing and Management*, 35, 401-420.

Ha, S., Bae, S., & Park, S. (2000). "Web mining for distance education". *IEEE international conference on management of innovation and technology*, 2(2), 715-719.

Halees, A. El. (2008). "Mining Students data to analyze learning behavior: a case study". Department of Computer Science, Islamic University of Gaza P.O.Box 108 Gaza, Palestine.

Han, J., Kamber, M. (2001). "Data Mining – Concepts and Techniques" , JimGray Series Editor Morgan Kaufmann Publishers.

Han, J., Kamber, M., and Tung, A. (2001). "Spatial Clustering Methods in Data Mining: A Survey", *In Miller, H., and Han, J., eds., Geographic Data Mining and Knowledge Discovery*. Taylor and Francis.

Hancock, V. E. (1992/1993). "The at-risk student". *Educational Leadership*, 50(4), 84-85.

Hsu, M. H. (2008). "Proposing an ESL recommender teaching and learning system". *Expert Systems with Applications*, 34, 2102-2110.

Hsu, M. H. (2008). "A personalized English learning recommender system for ESL students". *Expert Systems with Applications*, 34, 683-688.

Idayes, C., Cimningham, P. (2004). "Context boosting collaborative recommendations", *Knowledge-Based Systems* 17, 131-138.

Itmazi, J., Megias, M. (2008). "Using Recommendation Systems in Course Management Systems to Recommend Learning Objects". *The International Arab Journal of Information Technology*, Vol.5, No. 3, 234-240.

Kim, Y.S. (2008). "Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size", *Elsevier, Expert Systems with Applications*, 1227-1234.

Kontkanen, P., Myllymaki, P., & Tirri, H. (1996). "Predictive data mining with finite mixtures.", *In Proceeding 2nd international Conference "Knowledge Discovery and Data Mining (KDD,96)*.

Kohrs, A., Merialdo, B. (2001). "Creating user-adapted websites by the use of Collaborative filtering", Interacting with computers13, 695-716.

Lee, C. H., Kim, Y. H., & Rhee, P. K. (2001). "Web personalization expert with combining collaborative filtering and association rule mining technique". *Expert Systems with Applications*, 21, 131-137.

Liang, T. P. (2008). "Recommendation systems for decision support", An editorial introduction. *Decision Support Systems (DSS)*, 45(3):385-386.

Li, X., Chang, S. K. (2005). " A Personalized E-Learning System Based on User Profile Constructed Using Information Fusion", *The eleventh International Conference on Distributed Multimedia Systems*, 109-114.

Li, X., Lu, L., Xuefeng, L. (2005). "A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-Commerce", *Expert Systems with Applications* 28, 67-77.

Linden, G., Smith, B., & York, J. (2003). "Amazon.com recommendations". *IEEE Internet Computing 7. no. 1, (Jan.-Feb. 2003)* , 76-80.

Liu, D., Shih, Y. (2005). "Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences", *The Journal of Systems and Software*77, 181-191.

Merceron, A., Yasef, K. (2005). "Educational Data Mining: a Case Study", *Proceedings of the 12th International Conference on Artificial Intelligence in Education AIED 2005*, Amsterdam, The Netherlands,IOS Press.

Melville, P., Mooney, R.J., Nagarajan, R. (2002). "Content-Boosted Collaborative Filtering for Improved Recommendations", *Proceedings of the 18th National Conference on Artificial Intelligence*, 187-192, Edmonton,Canada.

Minaei, B. (2004). "Data mining for A Web-Based Educational System", Ph.D. Thesis, Department of Computer Science and Engineering, Michigan State University.

Minaei, B., Kashy, D.A., Kortemeyer, G., & Punch, W.F. (2003). "Predicting student performance: An Application of data mining methods with the educational Web-Based System LON-CAPA", *In Proceedings of ASEE/IEEE Frontiers in Education Conference*, Boulder, CO: IEEE.

Mobasher, B. (2004). "Web Usage Mining and Personalization", *Practical Handbook of Internet Computing*, Chapman Hall and CRC Press.

Mooney, R. J. & Roy, L. (2000). "Content-based book recommending using learning for text categorization", *In Proc. Digital Libraries*, 195-204. San Antonio, TX.

Nayak, R., Seow, L. (2002). "Knowledge Discovery in Mobile Business Data", Queensland University of Technology, Australia.

Orzechowski, T., Ernst, S., Dziech, A. (2007). "Profiled Search Methods for e-Learning Systems". *First International Workshop on Learning Object Discovery & Exchange*.

Papagelis, M., Plexousakis, D. (2005). "Qualitative analysis of user-based and item-based prediction algorithms for recommendation agents", *Engineering Applications of Artificial Intelligence*18,781-789.

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3), 56–60.

Ribeiro, B., Cardoso, A. (2007). “Behavior Pattern Mining during the Evaluation Phase in an e-Learning Course”. *In: Proceedings of the International Conference on Engineering and Education (ICCE)*.

Romero Morales, C., Ventura Soto, S., Zafra Gómez, A. & Bro, P. (2009). “Applying Web usage mining for personalizing hyperlinks in Web based adaptive educational systems”, *Computers & Education*, 53, 828–840.

Romero Morales, C., Porras Pérez, A.R., Ventura Soto, S., Hervás Martínez, C, & Zafra Gómez, A. (2006). “Using sequential pattern mining for links recommendation in adaptive hypermedia educational systems”. *Current Developments in Technology-Assisted Education*.

Sarwar, B., Kerypis, G., Konstan, J. A., & Riedl, J. (2002). “Item-based collaborative filtering recommendation algorithms”. In Proc. WWW10. University of Minnesota, Minneapolis. (pp. 285–295). Hong Kong.

Soo, Y., Bong, K.,Yum, J., Song, J., & Kim, S.M. (2005). “Development of a recommender system based on navigational and behavioral patterns of customers in e-commerce sites”, *Expert Systems with Applications*28, 381-393.

Shardanand, U., & Maes, P. (1995). “Social information filtering: Algorithms for automating "word of mouth"”. *In Proc. CHI’95* (pp. 210–217). Denver, CO.

Si, L., Jin, R. (2003). “Flexible Mixture Model for Collaborative Filtering”. Proc. 20th Int’1 Conf. of Machine Learning.

Tang, T. & McCalla, G. (2005). “Smart Recommendation for an Evolving E-Learning System”, *Architecture and Experiment.* International Journal on *E-Learning*, Vol.4, No.1, 105–129.

Vasantha, M., Subbiah Bharathy, V. (2010). “Evaluation of Attribute Selection Methods with Tree based Supervised Classification: A Case Study with Mammogram Images”, *International Journal of Computer Applications,* Vol.8, No.12, 35–38.

Wang, J., Vries, A.P., Reinders, M.J.T. (2006). “Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion”, Information and Communication Theory Group, Delf University of Technology.

Wang, Y. F., Chuang, Y. L., Hsu, M. H., & Keh, H. C. (2004). A personalized recommender system for the cosmetic business. *Expert Systems with Applications*, 26, 427–434.

WEKA: Waikato environment for knowledge analysis .http://www.cs.waikato.ac.nz/ml/weka.

Witten, I.H., Frank, E. (2005). “Data Mining: Practical Machine Learning Tools and Techniques”. 2nd Edition, San Francisco: Morgan Kaufmann.

Zaiane, O.R. (2002). “Building a Recommender Agent for e-Learning Systems”. *In: Proceedings of the International Conference on Computers in Education (ICCE).*

ABSTRACT

Application of Recommender Systems on E-Learning Environment

By

Ailar Sekhavatian

In recent years, many advances in educational systems have occurred in order to introduce new technologies such as web based training. Nowadays more and more people have benefited from various e-learning applications. However, high diversity of the learners on the Internet poses new challenges to the traditional "one-size-fit-all" learning model, in which a single set of learning resource is provided to all learners. In fact, the learners may have different levels of expertise, and hence they can not be treated in a uniform way. Recommender Systems are used to avoid this problem, increase the efficiency of e-learning environment and enhance the quality of education and motivation of learners.

In this research we have developed a recommender system capable of providing students with appropriate educational materials that suit their different levels, therefore prevent at-risk students from failing and improves their academic achievement. The system uses data mining techniques. So it analyzes students' reading data and predicts their final results based on the similar students' records before completing their course. After the process of recognizing weak and strong students, different approaches are then used to provide recommendations. The proposed system has proved to be effective. In order to evaluate system performance, probability of correct prediction results of students is calculated. The experimental results show that the students under our recommendation system have made good progress in their performance and the probability of successfully passing their courses has increased.

Nooretouba University

A thesis for the degree of M.S.
……

Application of Recommender Systems on E-Learning Environment

Supervised by:
Dr. Mehregan Mahdavi

Advised by:
Eng. Alireza Ghanadan

By:
Ailar Sekhavatian

February / 2011