
بسم الله الرحمن الرحیم

دانشکده مهندسی صنایع و مدیریت

گروه مدیریت

رشته : MBA

تجزیه و تحلیل سبد خرید مشتریان با رویکردهای داده کاوی بر اساس مجاورت کالاها

دانشجو :

افشین میرزایی

استاد راهنما :

دکتر رضا شیخ

پایان نامه ارشد جهت اخذ درجه کارشناسی ارشد

آبان ۱۳۹۱

دانشگاه صنعتی شاهرود

دانشکده مهندسی صنایع و مدیریت

گروه مدیریت

پایان نامه کارشناسی ارشد آقای افشین میرزایی

تحت عنوان: تجزیه و تحلیل سبد خرید مشتریان با رویکردهای داده کاوی بر اساس مجاورت کالاها

در تاریخ ..... توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی ارشد مورد ارزیابی و با درجه ..... مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
دکتر رضا شیخ			

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء

به پاس تعبیر عظیم و انسانی شان از کلمه ایثار و از خودگذشتگی

به پاس قلب های بزرگشان که فریادرس است و سرگردانی و ترس در پناهشان به شجاعت می گراید

به پاس عاطفه سرشار و گرمای امید بخش وجودشان که در این سردترین روزگاران بهترین پشتیبان است

و به پاس محبت های بی دریغشان که هرگز فروکش نمی کند

این مجموعه را به خانواده عزیزم تقدیم می کنم.

تشکر و قدردانی

نخست بر خود می دانم که پروردگار خود را شکر گویم به پاس

لبخند های محبت بار

دستان یاری رسان

همه آنچه را که از رحمت او دریافت کرده ام.

درود فراوان خدمت پدر و مادر عزیز، دلسوز و فداکارم که پیوسته جرعه نوش جام تعلیم و تربیت، فضیلت و انسانیّت آنها بوده ام و همواره چراغ وجودشان روشنگر راه من در سختی ها و مشکلات بوده است، از ایشان تقدیر و تشکر فراوان دارم.

از زحمات بی دریغ استاد فرهیخته و بزرگوارم در دانشگاه صنعتی شاهرود جناب آقای دکتر رضا شیخ که با راهنمایی های دلسوزانه خود در پیشبرد این پروژه راهگشای اینجانب بودند، کمال تشکر و سپاسگزاری را دارم.

و برای همه این عزیزان آرزوی سربلندی و روزگاری سبز دارم.

دانشجو تأیید می نماید که مطالب مندرج دراین پـایان نامه نتیجه تحقیقات خودش می باشد و در صورت استفاده از نتایج دیگران مرجع آن را ذکر نموده است.

کلیه حقوق مادی مترتب از نتایج مطالعات ، آزمایشات و نو آوری ناشی از تحقیق موضوع این پایان نامه متعلق به دانشگاه صنعتی شاهرود می باشد .

ماه و سال

چکیده :

تجزیه و تحلیل سبد خرید یک واژه عمومی در روش شناسی هایی است که به مطالعه ساختار اجزای تشکیل دهنده سبد محصولات خریداری شده به وسیله مشتریان در یک خرید می پردازد.این ایده وجود دارد که سبد محصولات خریداری شده توسط مشتریان منعکس کننده وابستگی بین محصولات یا خرید هایی که بین دسته های مختلف محصولات صورت گرفته است می باشد که تعیین این وابستگی ها می تواند پشتیبان مناسبی برای تصمیم گیری های بازاریابی و فروش باشند.

در محیط رقابتی کنونی که در حال تغییر مستمر است استفاده از داده کاوی برای بهره جستن از پتانسیل دانش و تصمیمات حاصل از آن اهمیت حیاتی پیدا کرده است که می تواند کاربردهای بسیاری در زمینه های مختلف مانند تصمیمات بازاریابی،کاهش هزینه ها و همچنین افزایش وفاداری مشتریان داشته باشد، با توسعه فن آوری اطلاعات شرکت ها و سازمان ها حجم عظیمی از اطلاعات را ذخیره می کنند که دارای پتانسیل ارزشی بالایی است که بدون علم داده کاوی بی ثمر می باشند در دهه اخیر تعدادی از تکنیک های داده کاوی مانند قوانین پیوند فرصت های جدیدی برای تجریه و تحلیل چنین داده هایی را فراهم آورده اند،از جمله موارد استفاده از قوانین پیوند ترتیب چیدمان اجناس در فروشگاه ها و یا همان مجاورت کالاها است،از دیگر موارد استفاده، تدوین استراتژی های فروش و ... می باشند.

بعد از تحلیل داده ها، ساختار ها و الگوهای متفاوتی تشکیل می شوند و باید ساختارها و الگوهای بهینه را که هر کدام ازآنها می توانند به عوامل بسیار زیادی وابسته باشند، از میان کل ساختارها و الگوها انتخاب کرد که این نوع انتخاب یک نمونه از تصمیم گیری های چند معیاره می باشد.

در این تحقیق بعد از تجزیه و تحلیل داده ها با استفاده از الگوریتم های متفاوت، از تعریف دو عامل که درجه اطمینان[۱] و میزان پوشش[۲] می باشند به عنوان دو معیار جهت تعیین میزان اعتبار و سنجش دقت و صحت قوانین تولید شده استفاده شده است، همچنین با تعریف این دو معیار میزان قابلیت اطمینان نتایج بدست آمده در این تحقیق مورد سنجش قرار گرفته است. از آنجا که از حجم زیادی از داده های مشتریان برای انجام این پژوهش استفاده شده است نتایج بدست آمده از این پژوهش دارای میزان اعتبار و قابلیت اطمینان بالایی هستند و مقدار درجه اطمینان در قوانین تولید شده تا ۰/۹۹۴و میزان درجه پوشش در قوانین تولید شده تا ۰/۹۷۴ می باشد که این نشان دهنده میزان اعتبار بالای این قوانین تولید شده است که از نقاط قوت این پژوهش می باشند، بعد از تجزیه و تحلیل داده ها و تعیین میزان اعتبار قوانین تولید شده، با استفاده از تکنیک های تصمیم گیری چند معیاره نظیر تاپسیس[۳] و دیمتل[۴] به تجزیه و تحلیل سبدهای تولید شده و اولویت بندی آنها پرداخته شده است.

روش تحقیق به این صورت بوده‌است که پس از مرور ادبیات و تعاریف ، از تکنیک های تصمیم گیری چند معیاره و الگوهای داده کاوی برای مدل سازی بهینه سبدهای خرید و اولویت بندی آنها استفاده شده است.

کلمات کلیدی: تجزیه و تحلیل سبد خرید، مجاورت کالاها،داده کاوی.

[۱] Confidence
[۲] Support
[۳] TOPSIS
[۴] DEMATEL

لیست مقالات استخراج شده از پایان نامه

۱. Mirzaei, A., Sheikh, R., (۲۰۱۲). *A Data Mining Framework for Extracting Product Sales Patterns in Retail Store Transactions Using Association Rules: A Case Study*. ۲۰۱۲; J Am Sci;۸(۹):۳۰۴-۳۰۸ (ISI)

۲. Mirzaei, A., Sheikh, R., Moodi, N., (۲۰۱۲). *Multi-criteria Group Purchasing Decision-making Process Based on Marketing Mix (۴Ps) Using Fuzzy TOPSIS*.۲۰۱۲; J Am Sci;۸(۹):۳۰۹-۳۱۳ (ISI)

۳. Mirzaei, A., Sheikh, R., (۲۰۱۲). *Retail Market Basket Analysis Using Multi-Criteria Decision Making Techniques A Case Study*.Submitted

فهرست مطالب

فهرست جداول

فهرست اشکال

فصل اول

کلیات تحقیق

۱-۱ مقدمه

امروزه در دنیای فناوری حجم بسیار زیادی از اطلاعات خام وجود دارد که به تنهایی هیچ کاربردی ندارند، اما میتوان با استفاده از تکنیک های داده کاوی بهترین بهره برداری را از این اطلاعات خام به عمل آورد. داده کاوی یک ابزار مطلوب است که میتوان با استفاده از آن بهترین الگوها و اطلاعات را از داده های خام استخراج کرد. با افزایش رقابت در سر تا سر جهان شرکت ها باید برای حفظ بقای خود با استفاده از داده های خام و تکنولوژی اطلاعات به پیش بینی وضعیت بازار طی چند ماه آینده بپردازند. .این فرایند منجر می شود شرکت ها تصمیمات مهمی را در محیط کاری خود اخذ کنند که در پیشرفت آنها بسیار موثر است. تا به حال کارهای زیادی در این زمینه انجام شده است، مواردی مثل کاربرد داده کاوی در بازاریابی، نویسندگان با ارائه تکنیک های خاص و بدست آوردن الگوهای مناسب دریافته اند که گسترش چه محصولاتی و به چه میزان مفید است و نیز چه تبلیغاتی در فروش بیشتر محصولات موثر است، بنابراین برای تصمیم گیری بهتر و ارائه خدمات بیشتر، بهتر است اطلاعات به دانش تبدیل شود .آنها نتیجه گیری کرده اند که رفتار مشتریان گذشته به جهت پیدا کردن الگو برای ساختن استراتژی جدید موثر است .در مواردی دیگر نویسندگان به کمک داده کاوی سیستمی را ارائه داده اند که جهت تعیین نوع درمان سنگ های کلیوی راهکار مطلوبی را بیان کرده است .در گزارش دیگری بیان شده است نویسندگان روشی را ارائه داده اندکه طبق آن می توان با تکنیک های داده کاوی به تشخیص پول شویی پرداخت .آنها با استفاده از تحلیل پارامترهای مرتبط به نتایج مطلوبی رسیده اند .به عقیده این نویسندگان، پولشویی نه تنها بر روی شرکت ها اثر بدی دارد، بلکه بر ثبات اقتصادی و سیاسی ملت ها نیز تاثیر گذار است .آنها با استفاده از تکنیک های خوشه بندی و طبقه بندی به تشخیص پولشویی پرداخته اند .

تغییر گرایش استراتژی تولید از تولید انبوه به تولید انبوه سفارشی باعث شده که استراتژی‌های بازاریابی نیز از بازاریابی انبوه به بازاریابی فرد به فرد متمایل شود. لذا استراتژی‌های بنگاه‌های اقتصادی در بازاریابی و معرفی محصول و خدمات به مشتریان شناخته شده‌ی خود و در نظر گرفتن خواسته‌ها و انتظارات آنها و نیز فراهم نمودن تسهیلات مناسب و ارائه خدمات مناسب پس از فروش برای آنها و در کنار آن حفظ روابط با مشتریان و تاکید بر نگه داشتن آنها به عنوان سرمایه‌های سازمان خلاصه می شود. به واسطه‌ی دستاوردهای تکنولوژی ، محیط های خدماتی امروزه نسبت به گذشته با سرعت بیشتری متحول شده است. در حال حاضر محیط‌های خدماتی ، براساس رقابت شدید شرکت‌های داخلی و خارجی ، ازدیاد روزافزون ادغام‌ها جهت کسب سود بیشتر، وجود پیچیدگی‌های بیشتر و مشتریان سخت گیر که تجربیات بیشتری در مورد محصولات خریداری شده و مصرف آن دارند تعریف شده‌اند. (Dabholkar et al. ۱۹۹۶; Sellers, ۱۹۸۹; Smith, ۱۹۸۹) از این رو شرکت‌ها جهت حضور موفق در بازارهای جهانی، نیازمند به دست آوردن تعداد زیادی مشتری جهت ایجاد برتری استراتژیک بادوام هستند.

۱-۲ بیان مسأله

امروزه مسئله مشتری مداری و توجه به نیازهای مشتریان از اهمیت ویژه ای درصنعت بازاریابی برخوردار است، چنانکه تلاش موفق ترین مدیران، توجه به نیازهای مشتریان است در این میان برخی از مشتریان که آنها را مشتریان بالقوه می نامیم از اهمیت بیشتری برخوردارند این مشتریان عمدتاً با دو دیدگاه به سراغ یک فروشگاه خواهند رفت .۱- توصیه سایر مشتریان ۲- نیاز مشتری .که در اولی با توجه به توصیه ای که دیگران به آنها داشته اند برای جستجو و در نهایت خرید یک محصول مراجعه می کنند . در این پارامتر توجه به مشتریان قبلی و خصوصاً توجه به مشتریان دائم از اهمیت بالایی برخوردار است . نیازمندی مشتری عاملی است که طی آن یک مشتری بر اساس تعاملی که فروشنده با مشتریان در نتیجه معرفی محصول دارد ایجاد می شود که در این وضعیت تبلیغات بهترین پارامتر تلقی می شود .در این

مطالعه تراکنش های رخ داده شده برای مشتریانی که خرید خود را قبلا تکمیل کرده اند مورد توجه قرار گرفته است چنانچه با تکنیک های داده کاوی به تحلیل اقلام خریداری شده توسط مشتریان و خصوصاً مشتریان ویژه سعی در ارائه یک پیشنهاد مطلوب برای مشتریان بالقوه داریم به این ترتیب هر مشتری جدید از یک یا چند راهنمای دائم و همراه برخوردار خواهند بود که با انتخاب هر محصول می توانند پیشنهادهای کارآمدی را برای سایر محصول ها ارائه کند .بطور مثال طی تجربه ای که از پیاده سازی این سیستم داشته ایم دریافتیم که اغلب مشتریانی که یک لپ تاپ خریده اند، یک عدد حافظه قابل حمل نیز خریداری کرده اند لذا پیشنهاد آن برای هر مشتری جدیدی که قصد خریدن لپ تاپ را دارد می تواند از جذابیت خاصی برخوردار باشد.

۱-۳ اهمیت مساله

از آنجا که مهمترین وظیفه بخش بازرگانی داخلی هر کشور، توزیع کالا در سطح جامعه است، یکی از روش های اصلاح نظام توزیع داخلی ایجاد فروشگاههای زنجیره ای است. در جهت پیشبرد اهداف این فروشگاه ها شناخت و پیش بینی نیازهای مشتریان اهمیت ویژه ای دارد.

در گذشته فروشندگان شغل خود را به دید خرید یک سری محصولات و قرار دادن آنها در کنار یکدیگر و فروش آنها به مردم می دیدند.اگر یک محصول بیشتر خریداری می شد سفارش آنها نیز بیشتر می شد و اگر خریداری نمی شد از ردیف ها خارج می شدند.اینگونه فروش ها به بیان بلیسچک[۱] کسب و کار محصول محور نامیده می شوند.امروزه با وجود محیط پر رقابت ،برای دستیابی به موفقیت، فروشندگان دیگر نمی توانند تنها با دید محصول محور به کسب و کار نگاه کنند.آنها برای دستیابی به موفقیت باید به اصطلاح مشتری محور باشند و خدمات بهتر به مشتریان از داشتن علم و دانش بیشتر نسبت به مشتریان و درک بهتر رفتار آنها ناشی می شود و علم به درک رفتار مشتریان از نوع خریدهای آنها مشخص می

[۱] Blischok

شود به عنوان مثال از تجزیه و تحلیل سبد خرید مشتریان.در حال حاضر با پیشرفت تکنولوژی حجم انبوهی از اطلاعات در اختیار می باشد.هدبرگ[1] بیان می کند که فروشگاه های زنجیره ای در آمریکا به نام وال-مارت[2] روزانه بیست میلیون فاکتور را ذخیره می کنند.این رشد بسیار زیاد در اطلاعات، فروشندگان را در جایگاه بسیار دشواری برای دریافت علم و درک رفتار مشتریان از اطلاعات موجود قرار می دهد و روش های قدیمی تحلیل داده ها دیگر کارایی لازم را ندارند،از این رو کشف دانش در پایگاه داده به کمک فروشندگان آمده است.امروزه از میان تکنیک های کشف دانش در پایگاه داده معروفترین آنها استخراج قوانین انجمنی یا همان قوانین پیوند است.

در محیط رقابتی کنونی که در حال تغییر مستمر است استفاده از داده کاوی برای بهره جستن از پتانسیل دانش و تصمیمات حاصل از آن اهمیت حیاتی پیدا کرده است که می تواند کاربردهای بسیاری در زمینه های مختلف مانند تصمیمات بازاریابی،کاهش هزینه ها و همچنین افزایش وفاداری مشتریان داشته باشد،با توسعه فن آوری اطلاعات،شرکت ها و سازمان ها حجم عظیمی از اطلاعات را ذخیره می کنند که دارای پتانسیل ارزشی بالایی است که بدون علم داده کاوی بی ثمر می باشند لذا ضرورت توجه به آن آشکارا احساس می شود.

۱-۴ اهداف تحقیق

این پژوهش اهداف متعددی را دنبال می کند. از چشم انداز نظری پژوهش حاضر می تواند به فهم آسان تری از عوامل تاثیرگذار بر تصمیم مشتری جهت تداوم رابطه با یک کالا یا خدمت کمک کند.

به طور کلی اهداف این تحقیق را این گونه می توان بیان کرد:

۱. تعیین الگوهای بهینه مجاورت کالاهادر فروشگاه.

[1] HEDBERG
[2] WAL-MART

۲. تشخیص و درک سبد های خرید مشتریان.

۳. تعیین قوانین پیوند و انجمنی بین اجناس فروشگاه.

۱-۵ سوال های تحقیق

۱-چگونه می توان با استفاده از سبد خرید مشتریان مجاورت کالاها را تعیین کرد؟

۲- چگونه می توان با استفاده از تکنیک داده کاوی چیدمان و مجاورت کالاها را تعیین کرد؟

۱-۶ جنبه های جدید و نوآوری تحقیق

در این تحقیق تلاش می‌شود تا با رویکردی متفاوت و جامع‌نگری بیشتر به مقوله‌ی سبد خرید مشتریان نگاه شود و با چندین روش به تجزیه و تحلیل سبد خرید مشتریان می پردازیم، همچنین از روش‌های نوین تصمیم گیری های چند معیاره در بهینه سازی و تجزیه و تحلیل سبد خرید مشتریان بهره می‌بریم که هرگز پیش از این در ایران صورت نگرفته است که همگی از ابعاد نوین این پژوهش‌اند.

در این تحقیق بعد از تشکیل سبدهای متفاوت توسط الگوریتم های طراحی شده و اجرای آنها با استفاده از نرم افزار MATLAB ، از روش های متفاوت تصمیم گیری چندمعیاره برای اولویت بندی و انتخاب سبد های پر اهمیت از بین آنها استفاده می شود که شامل روش تاپسیس[۱] و دیمتل[۲] می باشد. در هر کدام از این روش ها اصطلاحات جدیدی تعریف شده است و در ادامه به تفسیر راجب آنها صحبت خواهیم کرد.

[۱] TOPSIS
[۲] DEMATEL

۱-۷ روش ها و ابزار تجزیه و تحلیل داده ها

در این تحقیق اطلاعات مربوط به مشتریان و خرید آنها استخراج شده است و پس از حذف فیلدهای غیر اطلاعاتی از طریق پیاده سازی الگوریتم های داده کاوی به شناسایی الگوهای مشترک در سبدهای خرید مشتریان پرداخته می شود و در نهایت بر اساس دانش بدست آمده می توان توصیه های لازم و مناسب در مورد سبد خرید جدید و یا سایر عوامل مربوطه به مسئولان فروشگاه توصیه کرد.

فصل دوم

مروری بر ادبیات موضوع

۲-۱ مروری بر داده کاوی

۲-۲ آشنایی با ادبیات موضوع

۲-۲-۱ داده کاوی

۲-۲-۲ چه چیزی سبب پیدایش داده کاوی شده است؟

۲-۲-۳ مراحل کشف دانش

۲-۲-۴ جایگاه داده کاوی در میان علوم مختلف

۲-۲-۵ داده کاوی چه کارهایی نمی‌تواند انجام دهد؟

۲-۲-۶ داده کاوی و انبار داده‌ها

۲-۲-۷ توصیف داده‌ها در داده کاوی

۲-۳ سابقه و تاریخچه داده کاوی

۲-۴ شناخت کسب و کار (Business Understanding)

۲-۵ شناخت داده ها (Data Understanding)

۲-۶ آماده سازی داده ها (Data preparing)

۲-۷ مدلسازی (Modeling)

۲-۸ فرآیند داده کاوی

۲-۸-۱ پیش پردازش داده های خام

۲-۸-۲ کاوش داده ها

۲-۱ مروری بر داده کاوی

به تازگی داده کاوی موضوع بسیاری از مقالات کسب و کار و مجلات نرم افزار بوده است. با این حال، تنها چند سال پیش، فقط چند نفر واژه داده کاوی را شنیده بودند. هر چند داده کاوی شکل تکامل یافته رشته ای با سابقه طولانی است، اما این واژه به خودی خود به تازگی و در دهه ۹۰ معرفی شده است.

داده‌کاوی پایگاه‌ها و مجموعه‌های حجیم داده‌ها را در پی کشف واستخراج دانش مورد تحلیل و کند و کاوهای ماشینی (و نیمه‌ماشینی) قرار می‌دهد. این گونه مطالعات و کاوش‌ها را به واقع می‌توان همان امتداد و استمرار دانش کهن و همه جا گیر آمار دانست. تفاوت عمده در مقیاس، وسعت و گوناگونی زمینه‌ها و کاربردها، و نیز ابعاد و اندازه‌های داده‌های امروزین است که شیوه‌های ماشینی مربوط به یادگیری، مدل‌سازی، و تعلّم را طلب می‌نماید.

اصلاح داده کاوی[۱] همان طور که از ترجمه آن به معنی داده کاوی مشخص می‌شود به مفهوم استخراج اطلاعات نهان و یا الگوها وروابط مشخص در حجم زیادی از داده‌ها به یک یا چند بانک اطلاعاتی بزرگ است.بسیاری از شرکت‌ها و موسسات دارای حجم انبوهی از اطلاعات هستند. تکنیک‌های داده‌کاوی به طور تاریخی به گونه‌ای گسترش یافته‌اند که به سادگی می‌توان آنها را با ابزارهای نرم‌افزاری امروزی و موجود در این موسسات تطبیق داده و از اطلاعات جمع آوری شده فعلی بهترین بهره را برد. در صورتی که سیستم‌های داده کاوی بر روی سکوهای سرویس گیرنده/ سرویس دهنده[۲] قوی نصب شده باشد و دسترسی به بانک‌های اطلاعاتی بزرگ فراهم باشد، به کمک چنین سیستم‌هایی می‌توان به سوالاتی از قبیل :کدامیک از مشتریان ممکن است خریدار کدامیک از محصولات آینده شرکت باشد (چرا،در کدام مقطع زمانی)و بسیاری از موارد مشابه پاسخ داد.

---

۱ Data Mining
۲ Client/Server

در تعریفی، داده کاوی را یک رشته میان رشته ای دانسته و آنرا حاصل تلاش و همکاری مجموعه ای از رشته های آمار، هوش مصنوعی، یادگیری ماشین، پایگاه داده، بازیابی اطلاعات[۱] ، تشخیص الگو[۲] و تکنیک های تصویرسازی داده ها[۳] در طی یک پروسه دانسته اند (Kamber,۲۰۰۱ & Han).

همچنین در تعریفی دیگر از داده کاوی، آنرا فرایندی نامیده اند که طی آن الگوهای جالبی که به صورت آشکار جرئی از داده ها نیستند کشف می شوند (Frank & Witten). از این الگوها می توان به منظور پیش بینی استفاده نمود.

ریشه های داده کاوی از طریق دو مسیر به نیاکان خود بر می گردد. آمار کلاسیک یکی از این مسیر هاست. بدون آمار، داده کاوی وجود نخواهد داشت، آمار پایه اکثر فناوری هایی است که داده کاوی بر مبنای آنها ساخته شده است. آمار کلاسیک در بستر مفاهیمی مانند تحلیل رگرسیون، توزیع استاندارد، واریانس استاندارد، تجزیه و تحلیل تفکیکی[۴]، تجزیه و تحلیل خوشه ای و فاصله اطمینان[۵] و ... به مطالعه داده ها و روابط داده ها می پردازد. این مفاهیم و روش ها بلوک های تشکیل دهنده ساختمان بسیار پیشرفته تری هستند که با تجزیه و تحلیل آماری کلاسیک نقش مهمی ایفا می کند و می توان آمار کلاسیک را یکی از پایه های اصلی و بنیادین داده کاوی دانست.

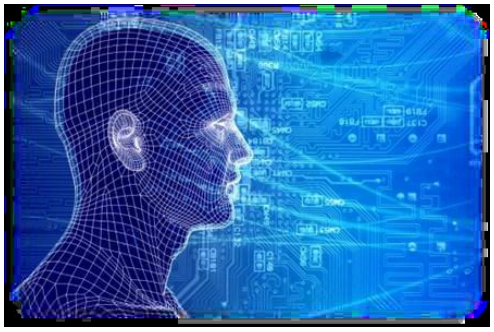
هوش مصنوعی یکی دیگر از نیاکان داده کاوی است. هوش مصنوعی که اساس آن تکنیک های اکتشافی[۶] است و در برابر تکنیک های آماری قرار دارد، تلاشی است برای اعمال روشی مانند روش

---

۱ Information Retrieval
۲ Pattern Recognition
۳ Data Visualization
۴ Discriminate Analysis
۵ Confidence Intervals
۶ Heuristic

پردازش تفکر انسان بر مشکلات آماری از بین متدهای هوش مصنوعی، تکنیک های یادگیری ماشین در زمینه داده کاوی بیشتر مورد استفاده قرار گرفتند.

در آنها تکنیک های اکتشافی هوش مصنوعی با تکنیک های پیشرفته آمار ترکیب می شوند. در یادگیری ماشین تلاش می شود تا به برنامه های کامپیوتری امکان داده شود از داده های خود جهت آموزش استفاده کنند و اصطلاحا آموزش[1] داده شوند، مانند برنامه هایی که بر اساس کیفیت داده های ورودی تصمیمات مختلفی می گیرند.



داده کاوی، در بسیاری از زمینه ها اساسا اقتباس روش های یادگیری ماشین و اعمال آن ها به یک برنامه کسب و کار است. به عنوان بهترین توصیف، داده کاوی را می توان ترکیب پیشرفت های اخیر و قدیمی آمار، هوش مصنوعی و یادگیری ماشیت دانست. این تکنیک ها بصورت تجمعی مورد استفاده قرار می گیرند تا به مطالعه داده ها بپردازند تا روندها و الگوهای ناشناخته قبلی را کشف نمایند. داده کاوی در زمینه هایی مورد اقبال قرار گرفته است که می بایست داده های زیادی مورد تجزیه و تحلیل قرار گیرند تا روندهای خاصی کشف شوند و این روندها به هیچ روش یا ابزار دیگری قابل کشف شدن نباشند.

به طور عام می توان داده کاوی را فرآیند استخراج الگوها ار داده ها دانست. همچنان که داده ها به طور پیوسته در حال گرد آوری هستند، با نرخ دو برابر شدن حجم داده ها هر سه سال یکبار (Lyman&

[1] Learn

(Hal R, ۲۰۰۳) و یا به برآوردی دیگر هر نه ماه یکبار (Fayyad, g & Uthurusamy, ۲۰۰۳) ، داده کاوی به طور فزاینده ای در حال تبدیل شدن به یک ابزار مهم برای تبدیل این داده ها به اطلاعات است. در تحقیقی که در سال ۲۰۰۵ ارایه گردید برآورد شد که حدود ۶ بیلیون دلار در فعایت های متن کاوی[۱] و داده کاوی سرمایه گذاری انجام گردد (Ebecken, ۲۰۰۵ & Zanasi, Brebbia). داده کاوی معمولا در طیف گسترده ای از شیوه های پروفایلینگ مانند بازاریابی، نظارت، ردیابی و کشف تقلب و اکتشافات علمی مورد استفاده قرار می گیرد.

تکنیک های داده کاوی را بر روی بازه گسترده ای از انواع مجموعه های داده ها نظیر پایگاه داده ای، انبار داده ای، داده های جغرافیایی، داده های مالتی مدیا، اینترنت، وب، داده های متنی و ... قابل اعمال است. شاید عبارت کشف دانش برای کل فرآیند استخراج الگو از داده ها مناسب تر باشد ولی واژه داده کاوی انتخاب شده و مورد اقبال قرار گرفته است (Andrassoya & Paralic, ۱۹۹۹).

زمانی که به داده کاوی به عنوان ابزاری برای کشف الگوهای موجود در مجموعه داده ای[۲] می نگریم، توجه به سه نکته بسیار مهم حائز اهمیت است. اولین نکته این است که با استفاده از داده ها و نمونه هایی[۳] که در دامنه (محدوده مساله مورد بررسی) حضور ندارد ممکن است نتایجی را تولید کند که از آن دامنه نباشند و دومین نکته این است که در صورتیکه داده ها موجود در دامنه در قالب مشخص و قابل کاوشی ارایه شده نباشند داده کاوی قادر نخواهد بود که اقدام به کشف الگو از آن داده ها نماید. این نکته به این معنی

[۱] Text Mining
[۲] Training Set
[۳] Sample

است که برای انجام یک داده کاوی موفق می بایست داده های مناسبی ارایه شود. داده کاوی نوعی ابزار است و مانند هر ابزار دیگری می باست برای نتیجه بخش بودن مواد اولیه مناسبی آماده شود.بدین منظور کاربر می بایست به عنوان اولین قدم از فرآیند داده کاوی مجموعه ای از داده های نماینده دامنه جمع آوری کند.سومین نکته مهم در فرآیند داده کاوی این است که کشف یک الگوی خاص در مجموعه ای خاص از داده ها لزوما به این معنی نیست که این الگو به طور عام نماینده کل جمعیت آماری مورد بررسی است.

بعضی از کاربردهای داده کاوی را می توان در کاربردهای معمول تجاری (مثل تحلیل و مدیریت بازار، تحلیل سبد بازار، پیش بینی قیمت نفت، بازاریابی هدف، فهم رفتار مشتری و تحلیل و مدیریت ریسک)، مدیریت و کشف فریب (کشف فریب تلفنی، کشف فریب های بیمه ای اتومبیل، کشف حقه های کارت اعتباری، کشف تراکنش های مشکوک مالی و پول شویی)،متن کاوی (خلاصه سازی، یافتن متون مشابه و کلمات کلیدی، پالایش نامه های الکترونیکی، گروه های خبری و غیره)، پزشکی (کشف ارتباط علامت و بیماری، تحلیل آرایه های DNA ، تصاویر پزشکی)، وب کاوی (پیشنهاد صفحات مرتبط، بهبود ماشین های جستجوگر یا شخصی سازی حرکت در وب سایت) و یافتن روندهای فرهنگی سیاسی در وب، تحلیل شبکه های اجتماعی وب (وبلاگها، ویکی ها) ، آنالیر ترافیک وب، تشخیص نفوذی به شبکه، متن کاوی، بیوانفورماتیک، سیستم پیشنهاد دهنده برای آموزش مجازی و کاربردهای بسیار دیگری در شاخه های مختلف مهندسی دانست. البته داده¬کاوی هر کاری را انجام نمی دهد و هر کار آماری را داده کاوی نمی نامند. برای داده کاوی شناخت و تحلیل داده ها مورد نیاز است، به طوری که بتوان روابط و الگوهای بین داده ها را با کمک افراد خبره پیدا کرد.

امروزه، بیشترین کاربرد داده کاوی در بانکها، مراکز صنعتی و کارخانجات بزرگ، مراکز درمانی و بیمارستانها، مراکز تحقیقاتی، بازاریابی هوشمند و بسیاری از موارد دیگر می باشد.

داده کاوی پل ارتباطی میان علم آمار ، علم کامپیوتر ، هوش مصنوعی ، الگوشناسی ، فراگیری ماشین و بازنمایی بصری داده می باشد. داده کاوی فرآیندی پیچیده جهت شناسایی الگوها و مدل های صحیح، جدید و به صورت بالقوه مفید، در حجم وسیعی از داده می باشد، به طریقی که این الگوها و مدلها برای انسانها قابل درک باشند. داده کاوی به صورت یک محصول قابل خریداری نمی باشد، بلکه یک رشته علمی و فرآیندی است که بایستی به صورت یک پروژه پیاده سازی شود.

کاوش داد ه ها به معنی کنکاش داده های موجود در پایگاه داده و انجام تحلیل های مختلف بر روی آن به منظور استخراج اطلاعات است.

داده کاوی فرایندی تحلیلی است که برای کاوش داده ها (معمولا حجم عظیمی از داده ها - در زمینه های کسب وکار و بازار) صورت می‌گیرد و یافته‌هابابه‌کارگیری الگوهایی،احراز اعتبار می‌شوند . هدف اصلی داده کاوی پیش بینی است. و به صورت دقیق تر میتوان گفت:

"کاوش داده ها شناسایی الگوهای صحیح، بدیع، سودمند و قابل درک از داده های موجود در یک پایگاه داده است که با استفاده از پرداز شهای معمول قابل دستیابی نیستند"

۲-۲ آشنایی با ادبیات موضوع

۲-۲-۱ داده کاوی

استفاده گسترده از بارکد برای تولیدات تجاری، به خدمت گرفتن کامپیوتر در کسب و کار، علوم، خدمات دولتی و پیشرفت در وسائل جمع‌آوری داده، از اسکن کردن موتون و تصاویر تا سیستم‌های سنجش از دور ماهواره‌ای، در این تغییرات نقش مهمی دارند.

به طور کلی استفاده همگانی از وب و اینترنت به عنوان یک سیستم اطلاع‌رسانی جهانی ما را مواجه با حجم زیادی از داده و اطلاعات می‌کند. این رشته انفجاری در داده‌های ذخیره شده، نیاز مبرم وجود تکنولوژی‌های جدید و ابزارهای خودکاری را ایجاد کرده که به صورت هوشمند به انسان، یاری می‌رسانند تا این حجم زیاد داده را به صورت اطلاعات و دانش تبدیل کند. داده کاوی به عنوان یک راه حل برای این مسائل مطرح می‌باشد. در یک تعریف غیررسمی داده کاوی فرایندی است، خودکار برای استخراج الگوهایی که دانش را بازنمایی می‌کنند، که این دانش به صورت قسمتی در پایگاه داده‌های عظیم انباره داده‌ها و دیگر مخازن بزرگ اطلاعات، ذخیره شده است. داده کاوی به طور همزمان از چندین رشته علمی بهره‌ها می‌برد. نظیر تکنولوژی ایجاد پایگاه داده، هوش مصنوعی، یادگیری ماشین، شبکه‌های عصبی، آمار، شناسایی الگو، سیستم‌های مبتنی بر دانش، حصول دانش، بازیابی اطلاعات، محاسبات سرعت بالا و بازنمایی بصری داده . داده کاوی در اواخر دهه ۱۹۸۰ پدیدار گشته، در دهه ۱۹۹۰ گام‌های بلندی در این شاخه از علم برداشته شده است. و انتظار می‌رود در این قرن به رشد و پیشرفت خود ادامه دهد. واژه‌های داده کاوی و کشف دانش در پایگاه داده اغلب به صورت مترادف یکدیگر مورد استفاده قرار می‌گیرد. کشف دانش در پایگاه داده فرایند شناسایی درست، ساده مفید و در نهایت الگوها و مدل‌های قابل فهم در داده‌ها می‌باشد. داده کاوی، مرحله‌ای از فرایند کشف دانش می‌باشد و شامل الگوریتم‌های مخصوص داده کاوی است. به طوری که تحت محدودیت‌های موثر محاسباتی قابل قبول، الگوها و یا مدل‌ها را در داده کشف می‌کند. به بیان ساده‌تر، داده کاوی به فرایند استخراج دانش ناشناخته، درست و بالقوه مفید از داده اطلاق می‌شود. تعریف دیگر اینست که داده کاوی گونه‌ای از تکنیک‌ها برای شنسایی اطلاعات و یا دانش تصمیم‌گیری از قطعات داده می‌باشد، به نحوی که با استخراج آنها، در حوزه‌های تصمیم‌گیری، پیش‌بینی، پیش‌گویی و تخمین مورد استفاده قرار گیرند. داده‌ها اغلب حجیم، اما بدون ارزش می‌باشند. داده به

تنهایی قابل استفاده نیست. بلکه دانش نهفته در داده‌ها قابل استفاده می‌باشد. به این دلیل اغلب به داده کاوی، تحلیل داده‌ای ثانویه گفته می‌شود.

۲-۲-۲ چه چیزی سبب پیدایش داده کاوی شده است؟

اصلی‌ترین دلیلی که باعث شده داده کاوی کانون توجهات در صنعت اطلاعات قرار بگیرد، مساله در دسترس بودن حجم وسیعی از داده‌ها و نیاز شدید به اینکه از این داده‌ها اطلاعات و دانش سودمند استخراج کنیم. اطلاعات و دانش بدست آمده در کاربردهای وسیعی از مدیریت کسب و کار کنترل تولید و تحلیل بازار تا طراحی مهندس و تحقیقاتی علمی مورداستفاده قرار می‌گیرد. داده کاوی را می‌توان حاصل سیر تکاملی طبیعی تکنولوژی اطلاعات دانست، که این سیر تکاملی ناشی از یک سیر تکاملی در صنعت پایگاه داده می‌باشد، نظیر عملیات جمع‌آوری داده‌ها و ایجاد پایگاه داده، مدیریت داده و تحلیل و فهم داده.

تکامل تکنولوژی پایگاه داده و استفاده فراوان آن در کاربردهای مختلف سبب جمع‌آوری حجم فراوانی داده شده است. این داده‌های فراوان باعث ایجاد نیاز برای ابزارهای قدرتمند برای تحلیل داده‌ها گشته زیرا در حال حاضر به لحاظ داده ثروتمند هستیم ولی دچار کمبود اطلاعات می‌باشیم.

ابزارهای داده کاوی داده‌ها را آنالیز می‌کنند و الگوهای داده‌های را کشف می‌کنند که می‌توان از آن در کاربردهایی نظیر تعیین استراتژی برای کسب و کار، پایگاه دانش و تحقیقات علمی و پزشکی، استفاده کرد. شکاف موجود بین داده‌ها و اطلاعات سبب ایجاد نیاز برای ابزارهای داده کاوی شده است تا داده‌های بی‌ارزش را به دانشی ارزشمند تبدیل کنیم. به طور ساده داده کاوی به معنای استخراج یا مدل ؟ داشتن از مقدار زیادی داده خام است، البته این نام‌گذاری برای این فرایند تا حدی نامناسب است، زیرا به طور مثال عملیات معدن کاری برای استخراج طلا از صخره و ماسه را طلاکاوی می‌نامیم. نه ماسه کاوی یا

صخره کاوی، بنابراین بهتر بود به این فرایند نامی شبیه به استخراج دانش از داده می‌دادیم که متاسفانه بسیار طولانی است. دانش کاوی به عنوان یک عبارت کوتاه‌تر به عنوان جایگزین نمی‌تواند بیانگر تاکید و اهمیت بر معدن کاری زیاد داده باشد. معدن‌کاری عبارتی است که بلافاصله انسان را به یاد فرایندی می‌اندازد که به دنبال یافتن مجموعه کوچکی از قطعات ارزشمند از حجم بسیار زیادی از مواد خام هستیم. با توجه به مطالب عنوان شده، با اینکه این فرایند تا حدی دارای نامگذاری نقص است ولی این نامگذاری یعنی داده کاوی بسیار عمومیت پیدا کرده است. البته اساسمی دیگری برای این فرایند پیشنهاد شده است که بعضاً بسیار متفاوت با واژه داده کاوی است. نظیر استخراج دانش از پایگاه داده‌ها، استخراج دانش، آنالیز داده و الگو، باستان‌شناسی داده، و لایروپی داده‌ها.

۳-۲-۲ مراحل کشف دانش

کشف دانش دارای مراحل تکراری زیر است:

۱. پاکسازی داده‌ها (از بین بردن نویز و ناسازگاری داده‌ها)

۲. یکپارچه‌سازی داده‌ها (چندین منبع داده ترکیب می‌شوند)

۳. انتخاب داده‌ها (داده‌ها مرتبط با آنالیز داده بازیابی می‌شوند)

۴. تبدیل کردن داده‌ها (تبدیل داده‌ها به فرمی که مناسب برای داده کاوی باشد مثل خلاصه‌سازی و همسان‌سازی

۵. داده کاوی (فرایند اصلی که روال‌های هوشمند برای استخراج الگوها از داده‌ها به کار گرفته می‌شود)

۶. ارزیابی الگو (برای مشخص کردن الگوهای صحیح و موردنظر به وسیله معیارهای اندازه‌گیری)

۷. ارائه‌ی دانش (یعنی نمایش بصری، تکنیک‌های بازنمایی برای ارائه دانش کشف شده به کاربر استفاده می‌شود).

در هر مرحله داده کاوی باید به کار بر با پایگاه دانش تعامل داشته باشد. الگوهای کشف شده به کاربر ارائه می‌شوند و در صورت خواست او به عنوان دانش به پایگاه دانش اضافه می‌شوند. توجه شود که بر طبق این دیدگاه داده کاوی تنها یک مرحله از کل فرایند است. البته به عنوان تک مرحله اساسی که الگوهای منحنی را آشکار می‌سازد. با توجه به مطالب عنوان شده، در اینجا تعریفی از داده کاوی ارائه می‌دهیم: داده کاوی عبارتست از فرایند یافتن دانش از مقادیر عظیم داده‌های ذخیره شده در پایگاه داده انبار داده و یا دیگر مخازن اطلاعات.

براساس این دیدگاه یک سیستم داده کاوی به طور نمونه دارای اجزای اصلی زیر است:

این اجزاء اصلی شامل:

۱- پایگاه داده، انبار داده یا دیگر مخازن اطلاعات: که از مجموعه‌ای از پایگاه داده‌ها و انبار داده‌ها، صفحات گسترده یا دیگر انواع مخازن اطلاعات، پاکساز ی داده‌ها و تکنیک‌های یکپارچه‌سازی روی این داده‌ها انجام می‌شود.

۲- سرویس‌دهنده پایگاه داده یا انبار داده، که مسئول بازیابی داده‌های مرتبط براساس نوع درخواست داده کاوی کاربر می‌باشد.

۳- پایگاه دانش: این پایگاه از دانش زمینه تشکیل شده تا به جستجو کمک کند، یا برای ارزیابی الگوهای یافته شده از آن استفاده می‌شود.

۴- موتور داده کاوی: این موتور جزء اصلی از سیستم داده کاوی است و به طور ایده‌آل شامل مجموعه‌ای از پیمانه‌هایی نظیر توصیف، تداعی، کلاسبندی، آنالیز خوشه‌ها و آنالیز تکامل و انحراف است.

۵- پیمانه ارزیابی الگو: این جزء معیارهای جذابیت

۶- واسط کاربر گرافیکی

۴-۲-۲ جایگاه داده کاوی در میان علوم مختلف

کاربردهای معمول تجاری: از قبیل تحلیل و مدیریت بازار، تحلیل سبد بازار، بازاریابی هدف، فهم رفتار مشتری، تحلیل و مدیریت ریسک

مدیریت و کشف ضریب: کشف ضریب تلفنی، کشف ضریب‌های بیمه‌ای و اتومبیل، کشف حقه‌های کارت اعتباری، کشف تراکنش‌های مشکوک مالی (پونستویی)

متن کاوی: پالایش متن (نامه‌های الکترونیکی و گروه‌های خبری و غیره)

پزشکی: کشف ارتباط علامت و بیماری، تحلیل آرایه‌های DNA (تصاویر پزشکی)

ورزش: آمارهای ورزشی

وب کاوی: پیشنهاد صفحات مرتبط بهبود ماشین‌های جستجوگر یا شخصی سازی حرکت در وب‌سایت

۲-۲-۵ داده کاوی چه کارهایی نمی‌تواند انجام دهد؟

داده کاوی فقط یک ابزار است و نه یک عصای جادویی داده کاوی به این معنی نیست که شما راحت به کناری بنشینید و ابزارهای داده کاوی همه کار را انجام دهد.

داده کاوی نیاز به شناخت داده‌ها و ابزارهای تحلیل و افراد خبره در این زمینه‌ها را از بین نمی‌برد. داده کاوی فقط به تحلیلگران برای پیدا کردن الگوها و روابط بین داده‌ها کمک می‌کند و در این مورد نیز روابطی که یافته می‌شود باید به وسیله داده‌های واقعی دوباره بررسی و تست گردد.

۲-۲-۶ داده کاوی و انبار داده‌ها

معمولاً داده‌هایی که در داده کاوی مورد استفاده قرار می‌گیرد از یک انبار داده استخراج می‌گردند و یک پایگاه داده یا مرکز داده‌ای ویژه برای داده کاوی قرار می‌گیرند.

اگر داده‌های انتخابی جزیی از انبار داده‌ها باشد بسیار مفید است چون بسیاری از اعمالی که برای ساختن انبار داده‌ها انجا می‌گیرد با اعمال مقدماتی داده کاوی مشترک است و در نتیجه نیاز به انجام مجدد این اعمال وجود ندارد.از جمله این اعمال پاکسازی داده‌ها می‌باشد. پایگاه داده مربوط به داده کاوی می‌تواند جزیی از سیستم انبار داده‌ها باشد و یا می‌تواند یک پایگاه داده جدا باشد.

ولی با این حال وجود انبار داده‌ها برای انجام داده کاوی شرط لازم نیست و بدون آن هم اگر داده‌ها در یک و یا چندین پایگاه داده باشند می‌توان داده کاوی را انجام دهیم و بدین منظور فقط کافیست داده‌ها را در یک پایگاه داده جمع‌آوری کنیم و اعمال جامعیت داده‌ها و پاکسازی داده‌ها را روی آن انجام دهیم. این پایگاه داده جدید مثل یک مرکز داده‌ای عمل می‌کنند.

۷-۲-۲ توصیف داده‌ها در داده کاوی

۱) خلاصه‌سازی و به تصویر درآوردن داده‌ها

۲) خوشه‌بندی

۳) تحلیل لینک

الف) خلاصه‌سازی و به تصویر در آوردن داده‌ها

قبل از اینکه بتوان روی مجموعه‌ای از داده‌ها، داده کاوی انجام داد و یک مدل پیش‌بینی مناسب ایجاد کرد، باید بتوان داده‌ها را به خوبی شناخت که برای شروع این کار می‌توان از پارامترهایی مثل میانگین، انحراف معیار و... استفاده کرد.

ابزارهای تصویرسازی داده‌ها و گراف‌سازی برای شناخت داده‌ها بسیار مفید می‌باشند و نقش آنها در آماده‌سازی داده‌ها بسیار مفید و غیر قابل انکار است، مثلاً با استفاده از این ابزار می‌توان توزیع مقادیر مختلف داده‌ها را در یک نمودار مشاهده کرد و میزان داده‌های دارای خطا را به طور تقریبی حدس زد. مهمترین مشکل این ابزار این است که معمولاً تحلیل ها دارای تعداد زیادی پارامتر هستند که به هم مربوطند و باید رابطه این پارامترها را که چند بعدی می‌باشد. در دو بعد نمایش دهند که این کار اگر هم عملی باشد برای استفاده از آنها نیاز به افراد خبره است.

ب)خوشه‌بندی

هدف از خوشه‌بندی این است که داده‌های موجود را به چند گروه تقسیم کنند و در این تقسیم‌بندی

داده‌های گروه‌های مختلف باید حداکثر تفاوت ممکن را به داشته باشند و داده‌های موجود در یک گروه باید بسیار به هم شبیه باشند.

برخلاف کلاس‌بندی (که در ادامه خواهیم دید) در خوشه‌بندی، گروه‌ها از قبل مشخص نمی‌باشند و همچنین معلوم نیست که بر حسب کدام خصوصیات گروه‌بندی صورت می‌گیرد. در نتیجه پس از انجام خوشه‌بندی باید یک فرد خبره خوشه‌های ایجاد شده را تفسیر کند و در بعضی مواقع لازم است که پس از بررسی خوشه‌ها بعضی از پارامترهایی که در خوشه بندی در نظر گرفته شده‌اند ولی بی‌ربط بوده یا اهمیت چندانی ندارند حذف شده و جریان خوشه‌بندی از اول صورت گیرد.

پس از اینکه داده‌ها به چند گروه منطقی و توجیه‌پذیر تقسیم شدند از این تقسیم‌بندی می‌توان برای کسب اطلاعات در مورد داده‌ها یا تقسیم داده‌ها جدید استفاده کنیم.

از مهمترین الگریتم‌هایی که برای خوشه‌بندی استفاده می‌شوند می‌توان Kohnen و الگوریتم K-means را نام برد.

ج) تحلیل لینک

تحلیل داده‌ها یکی از روش‌های توصیف داده‌هاست که به کمک آن داده ها را بررسی کرده و روابط بین مقادیر موجود در بانک اطلاعاتی کشف می شود. از مهمترین راههای تحلیل لینک کشف وابستگی و کشف ترتیب می‌باشد.

منظور از کشف وابستگی یافتن قوانینی در مواردی است که با هم اتفاق می‌افتند مثلاً اجناسی که در یک فروشگاه احتمال خرید همزمان آنها زیاد است.

کشف ترتیب نیز بسیار مشابه می‌باشد ولی پارامتر زمان نیز در آن دخیل می‌باشد.

وابستگی‌ها به صورت B نمایش داده می‌شوند که A مقدم و به B موخر یا نتیجه گفته می‌شود. مثلا اگر یک قانون به صورت زیر باشد:

"اگر افراد چکش بخرند آنگاه آنها میخ خواهند خرید"

در این قانون مقدم خرید جکش و نتیجه خرید میخ می‌باشد.

۳-۲ سابقه و تاریخچه داده کاوی

داده کاوی[۱] به بهره گیری از ابزارهای تجزیه و تحلیل داده‌ها به منظور کشف الگوها و روابط معتبری که تا کنون ناشناخته بوده‌اند اطلاق می‌شود. این ابزارها ممکن است مدلهای آماری،الگوریتم‌های ریاضی و روش‌های یاد گیرنده[۲] باشند. داده کاوی منحصر به گردآوری و مدیریت داده‌ها نبوده و تجزیه و تحلیل اطلاعات و پیش بینی را نیز شامل می‌شود.

علاوه بر پیشرفت ابزارهای مدیریت داده، افزایش قابلیت دسترسی به داده و کاهش نرخ نگهداری داده نقش ایفا می‌کند. در طول چند سال کذشته افزایش سریع جمع آوری و نگه داری حجم اطلاعات وجود داشته‌است. با پیشنهادهای برخی از ناظران مبنی بر آنکه کمیت داده‌های دنیا به طور تخمینی هر ساله دوبرابر می‌گردد. در همین زمان هزینه ذخیره سازی داده‌ها بطور قابل توجهی از دلار برای هر مگابایت به پنی برای مگابایت کاهش پیدا کرده‌است. مطابقا قدرت محاسبه‌ها در هر ۱۸ – ۲۴ ماه به دوبرابر ارتقاء پیدا کرده‌است این در حالی است که هزینه قدرت محاسبه رو به کاهش است. داده کاو به طور معمول در

---

[۱] Data Mining
[۲] Machine Laming Method

دو حوزه خصوصی و عمومی افزایش پیدا کرده‌است. سازمانها داده کاوی را به عنوان ابزاری برای بازدید اطلاعات مشتریان کاهش تقلب و اتلاف و کمک به تحقیقات پزشکی استفاده می‌کنند. با اینهمه ازدیاد داده کاوی به طبع بعضی از پیاده سازی و پیامد اشتباه را هم دارد.اینها شامل نگرانی‌هایی در مورد کیفیت داده‌ای که تحلیل می‌گردد، توانایی کار گروهی پایگاههای داده و نرم‌افزارها بین ارگانها و تخطی‌های بالقوه به حریم شخصی می‌باشد.همچنین ملاحظاتی در مورد محدودیتهایی در داده کاوی در ارگان‌ها که کارشان تاثیر بر امنیت دارد، نادیده گرفته می‌شود.

در حالیکه محصولات داده کاوی ابزارهای قدرتمندی می‌باشند، اما در نوع کاربردی کافی نیستند.برای کسب موفقیت، داده کاوی نیازمند تحلیل گران حرفه‌ای و متخصصان ماهری می‌باشد که بتوانند ترکیب خروجی بوجود آمده را تحلیل و تفسیر نمایند.در نتیجه محدودیتهای داده کاوی مربوط به داده اولیه یا افراد است تا اینکه مربوط به تکنولوژی باشد.

اگرچه داده کاوی به الگوهای مشخص و روابط آنها کمک می‌کند، اما برای کاربر اهمیت و ارزش این الگوها را بیان نمی‌کند.تصمیماتی از این قبیل بر عهده خود کاربر است.برای نمونه در ارزیابی صحت داده کاوی، برنامه کاربردی در تشخیص مظنونان تروریست طراحی شده که ممکن است این مدل به کمک اطلاعات موجود در مورد تروریستهای شناخته شده، آزمایش شود.با اینهمه در حالیکه ممکن است اطلاعات شخص بطور معین دوباره تصدیق گردد، که این مورد به این منظور نیست که برنامه مظنونی را که رفتارش به طور خاص از مدل اصلی منحرف شده را تشخیص بدهد.

از سال ۱۹۵۰ رایانه‌ها در تحلیل و ذخیره‌سازی داده‌ها به کار گرفته شدند. پس از حدود ۲۰ سال حجم داده‌ها دو برابر شد و پس از آن تقریبا هر دو سال یک‌بار همزمان با پیشرفت فناوری اطلاعات، حجم داده‌ها هم به دو برابر افزایش یافت. این پیشرفت آن قدر زیاد بود که تعداد رکوردهای برخی از پایگاه

داده‌ها به چند صد میلیارد رسید. پدیده شبکه جهانی وب، استفاده گسترد ه از بارکد برای تولیدات تجاری، به خدمت گرفتن کامپیوتر در کسب و کار، خدمات الکترونیکی دولتی و پیشرفت در وسایل جمع-آوری داد ه ، انفجاری را در مجموعه‌های اطلاعاتی سازمان‌ها و موسسات ایجاد کرد ه است. حجم زیاد اطلاعات، مدیران این مجموعه ها را در تحلیل و یافتن اطلاعات مفید دچار چالش کرده است. داده‌کاوی، ابزار مناسب را برای تجزیه و تحلیل اطلاعات و کشف و استخراج روابط پنهان در مجموعه‌های داده‌ای سنگین فراهم می‌کند.

داده کاوی، فرآیند کشف الگوهای پنهان، جالب توجه، غیر منتظره و با ارزش از داخل مجموعه وسیعی از داده-هاست و فعالیتی در ارتباط با تحلیل دقیق داده های سنگین بی ساختار است که علم آمار ناتوان از تحلیل آنهاست. بعضی مواقع دانش کشف شده توسط داده کاوی عجیب به نظر می رسد؛ مثلا ارتباط افراد دارای کارت اعتباری و جنسیت با داشتن دفترچه تامین اجتماعی یا سن، جنسیت و درآمد اشخاص با پیش بینی خوش حسابی او در بازپرداخت اقساط وام. داده کاوی از علو می مانند یادگیری ماشین، هوش مصنوعی، آمار، پایگاه داده و شناسایی الگو به طور همزمان بهره گرفته و در حوزه های تصمیم گیری، پیش بینی، و تخمین مورد استفاده قرار می گیرد.

واژه کشف دانش در پایگاه داده ها[1] در اوایل دهه ۸۰ در مراجعه به مفهوم کلی، گسترده، سطح بالا و به دنبال جستجوی دانش در اطلاعات شکل گرفته است. این لغت به بیان دیگر به همه شیوه هایی اشاره دارد که هدف آنها پی بردن به ارتباط و نظم بین اطلاعات قابل مشاهد است. لغت KDD برای توصیف همه مراحل استخراج اطلاعات از پایگاه داده و نیز بیان اهداف کارهای اولیه کاربرد قوانین تصمیم گیری است. این واژه به طور رسمی اولین بار توسط Usama Fayaad در اولین کنفرانس بین المللی داده کاوی و کشف دانش که در سال ۱۹۹۵ در مونترال برگزار شده بود، معرفی شد که به بیان ارتباط تکنیکهای آنالیز

[1] Knowledge Discovery in Database (KDD)

در چندین مرحله با هدف استخراج دانش های ناشناخته قبلی از داده های در دسترس می پرداخت. داده هایی که ارتباط منظم و پراهمیت آنها قبلاً به نظر نمی رسید. کم کم واژه داده کاوی جای خود را پیدا کرد و مترادفی برای همه مراحل استخراج دانش شد. هر چند که داده کاوی مرحله ای از KDD است، اما در کل KDD فرآیند یافتن اطلاعات و الگوهای مفید از داده را گویند و داده کاوی بهره گیری از الگوریتم هایی برای یافتن اطلاعات مفید در فرآیند KDD است.

فرآیند KDD عبارت است از:

۱) پاک سازی و یکپارچه سازی داده(پیش پردازش داده)

۲) ایجاد یک انبار داده مشترک برای تمام منابع

۳) داده کاوی

۴) بصری سازی نتایج تولید شده

که مرحله پیش پردازش غالباً یکی از مراحل زمان بر و در عین حال بسیار مهم در کسب نتیجه مطلوب است.

در تعاریف قبلی جنبه بسیار مهمی که همان هدف نهایی داده کاوی است حذف شد ه است. هدف نهایی داده کاوی به دست آوردن نتایجی است که می تواند منافع کاری داشته باشد.

داده کاوی کاربرد سطح بالای فنون و ابزار به کار برده شده برای معرفی و تحلیل داده های تصمیم گیرندگان است. اصطلاح داده کاوی را متخصصین آمار، تحلیلگران داده ها و انجمن سیستم های اطلاعات مدیریت به کار برده اند در حالی که پژوهشگران یادگیری ماشین و هوش مصنوعی بیشتر از KDD

استفاده می کنند. از نقطه نظر محققان، داده کاوی یک نظم نسبتاً جدید است که به طور عمد ه از میان مطالعاتی که به منظور نظم بخشیدن به برخی از فعالیتها همچون تخمین زدن، بازاریابی و سرشماری و آمار انجام گرفته، توسعه یافته است. اید ه ای که مبنای داده کاوی است یک فرآیند با اهمیت از شناخت الگوهای بالقوه مفید، تازه و درنهایت قابل درک در داده هاست. کشف دانش در پایگاه داده ها برای کشف اطلاعات مفید از مجموعه بزرگ داده هاست. دانش کشف شده می تواند قاعده ای باشد که با کمک آن ویژگی های داده ها، الگوهایی که به طور متناسب رخ می دهند، خوشه بندی موضوع های درون پایگاه داده ها و غیره را توصیف کند.

یک کاربر سیستم KDD به منظور انتخاب زیر مجموعه صحیحی از داده ها باید درک بالایی از قلمرو داده ها، رده مناسبی از الگوها و معیار خوبی برای الگوهای جالب داشته باشد. بنابراین سیستم KDD باید ابزارهایی با اثر تعاملی داشته باشد نه سیستم های تجزیه و تحلیل خودکار.

پژوهش جدی روی موضوع داده کاوی از اوایل دهه ۹۰ شروع شد. پژوهش ها و مطالعه های زیادی در این زمینه صورت گرفته؛ همچنین سمینارها، دوره های آموزشی و کنفرانس هایی نیز برگزار شد ه است. نتایج پایه های نظری داده کاوی در تعدادی از مقاله های پژوهشی آورده شد ه است. سال ۱۹۹۵ با استفاد ه از داده کاوی، انباره های داده بانک های آمریکا را بررسی کرده و بیان کردند که چگونه این سیستم ها برای بانک های آمریکا قدرت رقابت بیشتری ایجاد می کنند. در این سال انجمن داده کاوی همزمان با اولین کنفرانس بین المللی «کشف دانش و داده کاوی» شروع به کار و یک سازمان علمی به نام ACM- SIGKDD را تاسیس کرد. سال ۱۹۹۶ دیدگاهی از داده کاوی به عنوان «پرس و جو کننده از پایگاه های استنتاجی» پیشنهاد شد و فیاض و شاپیرو پیشرفت های کشف دانش و داده کاوی را اعلام کردند. همان سال دیدگا ه اقتصاد سنجی روی داده کاوی و عملکرد داده کاوی به عنوان یک مسأله بهینه

ارائه و کنفرانس های ناحیه ای و بین المللی در مورد داده کاوی برگزار شد که از جمله می توان به کنفرانس آسیا و اقیانوسیه درباره کشف دانش و داده کاوی اشاره کرد. سال ۲۰۰۰ بحث های مقایسه ای بین آمار و داده کاوی و نیز استفاده از وب در کاوش داده ها و کاربردهای آن ارائه شد. سال ۲۰۰۲ «داده کاوی ساختارهای پیوند برای مدل رفتار مصرف کننده» عرضه شد.

از لحاظ تاریخی، توسعه داده کاوی را در طول زمان می توان به مراحل زیر تقسیم کرد:

مرحله اولیه: گردآوری وایجاد پایگاه اطلاعاتی (تا دهه ۱۹۶۰)

مرحله دوم: نظام های مدیریتی مبنی بر پایگاه اطلاعاتی (دهه۱۹۷۰ واوایل دهه ۱۹۸۰)

مرحله سوم: نظام های پایگاه اطلاعاتی پیشرفته (اواسط دهه ۱۹۸۰ تازمان حاضر)

مرحله چهارم: انبارش اطلاعات و داده کاوی (اواخر دهه ۱۹۸۰ تا به امروز)

مرحله پنجم: نظام پایگاه اطلاعاتی مبنی برشبکه (دهه ۱۹۹۰ تا کنون)

مرحله ششم: نسل نوین نظام های اطلاعاتی یکپارچه شده (از ۲۰۰۰ به بعد)

بدین ترتیب فعالیتی که از دهه ۱۹۶۰ شروع شده بود، در دهه ۱۹۹۰ گام های بلندی برداشت و انتظار می رود در این قرن به رشد و بالندگی خود ادامه دهد.

از هنگامی که رایانه در تحلیل و ذخیره‌سازی داده‌ها به کار رفت پس از حدود ۲۰ سال، حجم داده‌ها در پایگاه داده‌ها دو برابر شد، ولی پس از گذشت دو دهه و همزمان با پیشرفت فن‌آوری اطلاعات[۱] هر دو سال یکبار حجم داده‌ها، دو برابر شد. همچنین تعداد پایگاه داده‌ها با سرعت بیشتری رشد

[۱] Information Technology

نمود.حال با وجود سیستم‌های یکپارچه اطلاعاتی سیستم‌های یکپارچه بانکی و تجارت الکترونیک، لحظه به لحظه به حجم داده‌ها در پایگاه داده‌های مربوط اضافه شده است.

انسان ها فرآیند استخراج الگوها از داده ها را برای قرن ها بصورت "دستی" انجام داده اند، اما با افزایش حجم اطلاعات در عصر مدرن ، لزوم استفاده از شیوه های خودکار امری کاملا بدیهی به نظر می رسد. متد های قدیمی و اولیه ای که در تشخیص الگوها در داده ها مورد استفاده قرار می گرفته اندعبارتند از قضیه بیز (دهه ۱۷۰۰) و تجزیه و تحلیل رگرسیون (دهه ۱۸۰۰) تکثیر سریع ، دسترسی آسان و قدرت و ظرفیت های روزافزون تکنولوژی های رایانه ای، حجم ذخیره سازی آنها را افزایش داده و این امکان را فراهم می آورد تا بتوان حجم بسیار زیادی از مجموعه های داده ای را ذخیره سازی کرد. با گذشت زمان همانطور که مجموعه داده ها در بعد اندازه و پیچیدگی رشد کرده اند، تجزیه و تحلیل مستقیم دستی[۱] داده ها به طور فزاینده با پردازش های غیرمستقیم و خودکار داده ها تکمیل (و نه جایگزین) شده است.این فرآیند به کمک سایر اکتشافات در علوم رایانه از قبیل شبکه های عصبی، خوشه بندی[۲]، الگوریتم های ژنتیکی (دهه ۱۹۵۰)، درخت تصمیم گیری (دهه ۱۹۶۰) و ماشین های بردار پشتیبانی[۳] (دهه ۱۹۸۰) انجام شده است. در تعریفی از داده کاوی فرآیند اعمال این روش ها به داده ها با هدف کشف الگوه های پنهان معرفی شده است (Kantradzic, ۲۰۰۳) داده کاوی برای سال های زیادی در کسب و کار های مختلف، دانشمندات و دولت ها استفاده شده است تا حجم انبوهی از داده ها را نظیر سوابق و تاریخچه سفر مسافران خطوط هوایی ، داده های سرشماری، داده های تولید شده توسط اسکنر سوپرمارکت ها و ... غربال کرده تا گزارش های متنوعی را تولید کنند. با این حال می بایست توجه شود که اینگونه گزارش ها همیشه نمی تواند به عنوان داده کاوی تلقی شود.

---

۱ Manual
۲ Clustering
۳ Support Vector Machine

در جهت تعریف استاندارد های مربوط به داده کاوی تلاش هایی انجام گرفته است، برای مثال می توان به

European Cross Industry Standard Process for Data Mining (CRISP-DM ۱،۰) اشاره کرد که در سال ۱۹۹۹ انجام پذیرفت و یا Java Data Mining Standard (JDM ۱،۰) که در سال ۲۰۰۴ انجام شد. لازم به ذکر است که این دو استاندارد هایی هستند که در حال تکامل اند و نسخه های بعدی و کاملتر در دست توسعه است. مستقل از تلاش های انجام گرفته برای این استاندارد ها ، سیستم های نرم افزاری با منابع باز[۱] مانند RapidMiner، Weka، Knime و R project که آزادانه در دسترس هستند به یک استاندارد غیر رسمی برای تعریف فرآیند های داده کاوی تبدیل شده اند.بسیاری از این سیستم ها قادر به دریافت و ارسال[۲] مدل ها به PMML (Predictive Model Markup Language) هستند. PMML یک روش استاندارد برای نمایش مدل های داده کاوی ارایه می کند به طوری که این مدل ها بتواند در برنامه های مختلف آماری بصورت اشتراکی مورد استفاده قرار گیرند. PMML یک زبان مبتنی بر XML است که توسط گروه داده کاوی[۳] (DMG) که یک گروه مستقل و متشکل از بسیاری شرکت های داده کاوی است، توسعه داده شده و نسخه ۴،۰ آن در ژوئن ۲۰۰۹ منتشر گردید

.(Guazzelli, lin,& (PMLL project page, Williams, ۲۰۰۹))

---

[۱] Open-Source
[۲] import & export
[۳] Data Mining Group

۲-۴ شناخت کسب و کار[۱]

در این مرحله لازم است که اهداف پروژه و نیازها از دیدگاه کسب وکار بطور مشخص تبیین شود. توجه به این مرحله بسیار ضروری است و با شناخت کامل اهداف می توان از قابلیت های داده کاوی بهترین استفاده را کرد و الگوریتم مناسب را انتخاب نمود. گامهای این مرحله عبارت اند از:

- تعیین اهداف کسب و کار (مفاهیم اولیه، اهداف و معیار های موفقیت)
- ارزیابی موقیت (موجودی منابع، نیازمندیها، فرضیه ها و محدودیت ها، ریسک ها و روابط، اصطلاحات و هزینه و فایده)
- تعیین اهداف داده کاوی (اهداف داده کاوی و معیار موفقیت)
- تولید طرح پروژه (طرح پروژه و ارزیابی اولیه ابزارها و تکنیکها)

۲-۵ شناخت داده ها[۲]

در این مرحله مواردی از قبیل موجود بودن داده ها، محرمانگی داده ها، چگونگی دسترسی به داده ها و میزان سیستمی بودن داده ها بررسی می شود. تحلیل صحیح در این مرحله ممکن است به تجدید نظر در مرحله شناخت کسب وکار بیانجامد. جمع آوری و آشنایی با داده ها ،شناخت مسائل مربوط به کیفیت داده ها و مشاهده اطلاعات اولیه نهانی یا زیرمجموعه های داده ها که ممکن است برای آنالیز جالب باشند از وظایف این مرحله است. گام های این مرحله عبلرتند از:

- جمع آوری اولیه داده ها
- توصیف داده ها

[۱] Business understanding
[۲] Data Understanding

- شناسایی داده ها
- تایید کیفیت داده ها

۲-۶ آماده سازی داده ها[۱]

در این مرحله داده های موجود جهت استفاده الگوریتم های داده کاوی آماده می شود.هدف این مرحله رسیدن به چشم انداز قابل کاوش است. در این مرحله یکپارچگی ، انتخاب ، پاکسازی و تغییر شکل داده ها انجام می شود که گام های این مرحله عبارتند از :

- انتخاب داده ها (با دلایل حذف یا انتخاب داده ها)
- پاکسازی داده ها
- ساختاردهی داده ها (انتخاب برخی صفات و ایجاد رکوردها)
- یکپارچه نمودن داده ها (داده های ترکیبی)
- فرمت کردن داده ها

۲-۷ مدلسازی[۲]

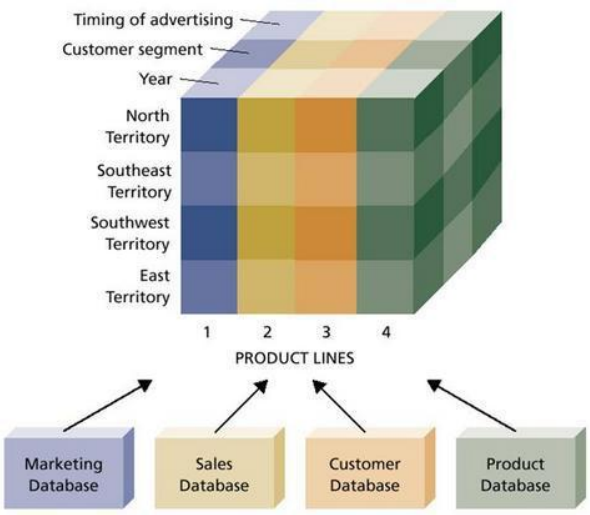
این مرحله پیاده سازی تکنیک های مدلسازی یا داده کاوی بر روی دیدگاه قابل کاوش ایجاد شده در مرحله قبل است، گام های این مرحله عبارتند از:

[۱] Data preparing
[۲] Modeling

- انتخاب تکنیک های مدلسازی (فرضیه ها و تکنیک های مدلسازی)
- ارزیابی طراحی (طراحی آزمایش)
- ساخت مدل (پارامترهای انتخاب شده و توصیف مدل)
- ارزیابی مدل (سنجش مدل و بازنگری در پارامترهای انتخاب شده)

۲-۸ فرآیند داده کاوی

کشف دانش در پایگاه داده نامی است که توسط Gregory Piatetsky – Shapiro در سال ۱۹۸۹ برای توصیف فرایند پیدا کردن داده های جالب ، تفسیر شده ، مفید و جدید گذاشته شد. در موارد و پروژه های مختلف داده کاوی بسیاری از تفاوت ها و جزئیات ظریف در این فرایند وجود دارد و بنا به شرایط این پروژه این فرآیند ممکن است تغییر می کند ، اما تقریبا به طور کلی مراحل این فرایند شامل پیش پردازش داده های خام ، کاوش داده ها و تفسیر نتایج می شود (Fayyad, Piatetsky- Shapiro, & Smyth, ۱۹۹٦)



فرایند کشف دانش در پایگاه داده که منجر به یافتن الگو در پایگاه داده می شود با فرایند گزارش گیری یا استخراج بخشی از داده ها از پایگاه داده کاملا نتفاوت است.با استفاده از گزارش گیری از پایگاه داده ای می توان به این سوال پاسخ داد که "چه شرکت هایی در سال قبل کالاهایی با ارزش بیشتر از ۱۰۰ هزار دلار

را فروخته اند؟" ، اما با بهره گیری از داده کاوی قادر خواهیم بود به این سوال پاسخ دهیم که "چه شرکت هایی احتمالا در سال آینده کالاهایی با ارزش بیش از ۱۰۰ هزار دلار را خواهند فروخت و به چه دلیل؟" (Benoit, ۲۰۰۲)

۲-۸-۱ پیش پردازش داده های خام

پس از آنکه هدف برای فرایند KDD شناخته شد ، مجموعه داده های[۱] هدف هدف باید گردآوری شود. داده کاوی صرفا قادر است الگوهایی را کشف کند که در داده ها وجود داشته باشند ، از این رو مجموعه داده های هدف باید به اندازه کافی بزرگ باشد تا شامل این الگوها باشد در حالی که می بایست به اندازه کافی کوچک باشد تا در بازه زمانی قابل قبول کاوش و استخراج انجام پذیرد. بدین معنی که اندازه مجموعه داده ها نقش بسیار مهمی را در این فرایند بازی می کند و با انتخاب نادرست داده ها احتمال بروز اشکالاتی در این پروسه وجود خواهد داشت.معمولا منابع رایج برای این داده های خام[۲] و انبار داده ها[۳] است.یکی از مهمترین قدم ها در پیش پردازش داده های خام تمیز کردن داده ها[۴] است. حذف داده های نامربوط[۵] و داده های نامشخص[۶] از مجموعه داده را عمل تمیز کردن می نامند. بدیهی است که طی این فرآیند می بایست به دقت و با استانداردهای لازم انجام شود زیرا که در داده های واقعی معمولا حجم اینگونه داده ها (داده های نویزی و داده های گم شده[۷]) زیاد بوده و حذف آنها می تواند منجر به کاهش چشمگیر مجموعه داده ای شود.

---

۱ Data Set
۲ Data Mart
۳ Data Warehouse
۴ Clearing
۵ Noisy Data
۶ Missing Values
۷ Missing Values & Noisy Data

غیر از حذف کردن این داده ها، مقدار دهی دستی این داده ها[۱] ، ثبت داده های ثابت[۲] و تعیین مقادیر پیش بینی شده[۳] از جمله روش هایی هستند که در تمیز کردن داده ها مورد استفاده قرار می گیرند. (Smyth, ۱۹۹٦ & Fayyad, piatetsky-shapiro).

در برخی موارد پس از پاکسازی داده ها می بایست آنها را به فرمت های قابل قبول و قابل کاوش تبدیل کرد. معمولا داده ها با فرمت های مختلفی برای داده کاوی جمع آوری می شوند. داده هایی در فرمت های Text، Execel، Word و ... هستند که می بایست یکسان سازی شوند. همچنین از مکانیزم های اجتماع[۴] (مثلا در نظر گرفتن داده ها در بازه ماهیانه به جای روزانه)، عمومی سازی[۵](مثلا تبدیل مقادیر عددی سن به جوان، بزرگسال به پیر)، نرمال کردن[۶] (تصحیح مقیاس داده ها و تبدیل آنها به بازه های ثابت) و ساختار صفت[۷] (اضافه کردن ویژگی ها و خصوصیات جدید به مجموعه داده ها) می توان نام بردکه در این بخش انجام می پذیرد (Kamber, ۲۰۰۱ & Han).

سپس داده های پاک شده به بردار ویژگی[۸] کاهش می یابد. بردار ویژگی نسخه خلاصه شده از داده های خام است. به عنوان مثال یک تصویر سیاه و سفید با سایز ۱۰۰ در ۱۰۰ پیکسل شامل ۱۰۰۰۰ بیت ار داده های خام است. این مجموعه داده ها ممکن است توسط جانمایی مکان های چشم ها و دهان در تصویر به بردار ویژگی تبدیل شود.انجام این کار باعث می شود که داده ها برای هر بار از ۱۰۰۰۰ بیت به سه بیت کد برای مکان ها کاهش یابد ، به طرز چشم گیری کاهش اندازه مجموعه داده های قابل استخراج را در پی خواهد داشت و از این رو زمان و هزینه پردازش کاهش داده می شود. انتخاب ویژگی ها

---

[۱] Manual Values
[۲] Constant Values
[۳] Estimated Values
[٤] Aggregation
[٥] Generalization
[٦] Normalization
[۷] Attribute construction
[۸] Feature Vector

به هدف بستگی دارد. واضح است که انتخاب درست ویژگی ها برای یک داده کاوی موفق امری بنیادی است.

بردارهای ویژگی به دو مجموعه، "آموزشی" و "آزمون" تقسیم می شود، مجموعه آموزشی به منظور "یادگیری" الگوریتم های داده کاوی استفاده می شود، در حالی که مجموعه آزمون به منظور تایید دقت الگوها پیدا شده مورد استفاده قرار می گیرد.

تکنیک هایی که به منظور کاهش حجم داده ها مورد استفاده قرار می گیرند عبارتند از

(Kamber, ۲۰۰۱ & Han)

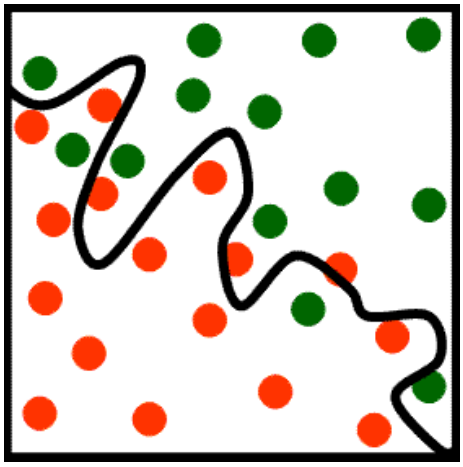
- Data Cube Aggregation
- Dimension Reduction
- Data Compression
- Discretization & Concept Hierarchy generation

۲-۸-۲ کاوش داده ها

پس از آنکه مجموعه داده ها از مرحله پیش پردازش عبور کردند و برای کاوش اماده شدند بخش دوم فرآیند که همان کاوش داده ها و کشف الگوهای موجود در این داده هاست آغاز می گردد. بر اساس تعریف انجام شده، داده کاوی را می توان به طور کلی به چهار دسته کلی از عملیات تقسیم بندی کرد که می توان یک یا ترکیبی از چند عمل مختلف را در یک مجموعه داده جهت کشف الگو ها مورد استفاده قرار داد : (Smyth, ۱۹۹٦ & Fayyad, Piatetsky - Shapiro)

- طبقه بندی[۱]: داده ها را در گروه های از پیش تعریف شده قرار می دهد. به عنوان مثال یک برنامه ایمیل ممکن است یک را به مشروع یا هرزنامه[۲] طبقع بندی نماید. الگوریتم های معمول در این زمینه عبارتند از

درخت تصمیم[۳]، نزدیک ترین همسایه[۴]، طبقه بندی ساده بیز[۵] و شبکه های عصبی[۶] .



شکل ۲-۱ طبقه بندی در داده کاوی

- خوشه بندی[۷] : شبیه طبقه بندی است با این تفاوت که گروه ها از پیش تعریف نشده اند، بنابراین الگوریتم سعی می کند اقلام مشابه با هم را در یک گروه قرار دهد.

---

۱ Classification
۲ Spam
۳ Decision Tree
٤ Nearest Neighbor
٥ Naïve Bayesian classification
٦ Neural Network

۷ Clustering

شکل ۲-۲ خوشه بندی در داده کاوی

- رگرسیون[۱]: تلاش می کند یک تابع پیدا کند که داده ها را با حداقل خطا فرموله و مدل کند.
- یافتن قوانین وابستگی[۲]: جستجو برای روابط بین متغییر هاست. برای مثال یک سوپر مارکت ممکن است اطلاعات مربوط به عادات خرید مشتری را جمع آوری کند. با استفاده از این روش ، سوپرمارکت می تواند محصولاتی که به صورت مکرر با هم خریداری می شوند را شناسایی کرده و از این اطلاعات برای اهداف بازاریابی استفاده نماید. به همین دلیل است که گاهی این روش را "تجزیه و تحلیل خرید سبد خرید" نیز می نامند.

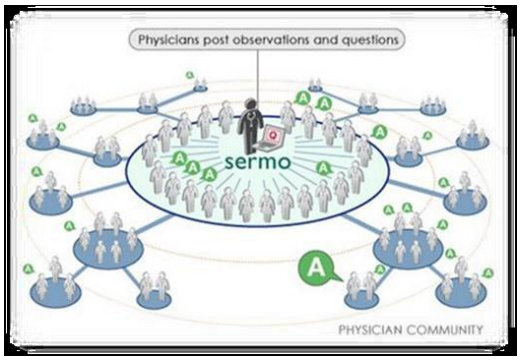
---

۱ Regression
۲ Association Rule Learning

۹-۲ سیستم های توصیه گر

در زندگی روزمره هر فرد موقعیت هایی رخ می دهد که وی را در شرایطی قرار می دهد که مجبور است علیرغم کمبود اطلاعات اقدام به تصمیم گیری نماید. گاهی اوقلت این تصمیم گیری ها برای انتخاب موضوع یا پیشنهادی است. به عنوان مثال فردی را تصور کنید که قصددیدن فیلمی را دارد که به تازگی اکران شده است، مسلما این فرد چندین انتخاب مختلف (فیلم های به تازگی اکران شده) دارد که اطلاعات و دانش کافی را جهت انتخاب مناسب ندارد. در دنیای واقعی معمول است که فرد به دوستان خود مراجعه می کند و از بین آنها تعدادی را انتخاب می کند که در این زمینه (دیدن فیلم) سلایقی مشابه (به طور نسبی) داشته باشند و از آنها می خواهدتا بهترین فیلم را (انتخاب) به وی پیشنهاد دهند. بدیهی است که هر یک از دوستان ممکن است فیلمی را توصیه کنند و خصوصیات برجسته آن فیلم را (دلیل توصیه فیلم) بیان کنند. طبق روال معمول، فرد مورد نظر این توصیه ها را تجمیع کرده و با معیار های خود اندازه گیری کرده و آنگاه اقدام به تصمیم گیری می نماید.

سیستم های توصیه گر نیز به طور عام همین روال را دنبال می کنند و با استفاده از توصیه های افراد به عنوان ورودی و پردازش آنها اقدام به کشف الگوهای موجود در آنها نموده و در صورت نیاز و درخواست توصیه از سیستم، بر اساس شرایط و الگوهای در خواست اقدام به ارائه توصیه های لازم و مناسب می کند.



طراحان و توسعه دهندگان اولین سیستم توصیه گر به نام Tapestry عبارت" فیلتر کردن اشتراکی[۱]" را برای سیستم توصیه گر خود انتخاب کردند و به تبع آنها عده ای نیز از همین عبارت برای توصیف سیستم های توصیه گر خود استفاده نموده اند. اما بدیهی است که استفاده و بکارگیری از این عبارت برای کلیه سیستم های توصیه گر درست و منطقی نیست زیرا هر سیستم توصیه گر لزوما بر مبنای استفاده از کمک و همکاری دیگران[۲] و توصیه ها و تمایلات آنها ساخته نمی شود. در ادامه به توصیف انواع سیستم های توصیه گر پرداخته خواهد شد.

طبق تعریفی از ویکی پدیا سیستم های توصیه گر نوعی از تکنیک های فیلتر کردن اطلاعات محسوب می شوند که تلاش می کنند تا اقلام اطلاعاتی نظیر فیلم، کتاب، موسیقی، خبر، تصویر، صفحه وب، و ... که بیشترین تطابق را با سلیقه کاربر دارد را به وی پیشنهاد کند. در واقع سیستم های توصیه گر جایگزین بسیار مناسبی برای سیستم های جستجو[۳] هستند و به کاربران کمک می کنند که علاوه بر یافتن آیتم های اطلاعاتی مورد نظر، آنهایی را بیابند که به سلیقه و تمایلات آنها نزدیک تر باشد.

۲-۹-۱ تبدیل بازدید کنندگان به خریدار

سیستم های توصیه گر در فروشگاه های وب سایتها با ارائه و پیشنهاد کالاهای جدید مرتبط به سبد خرید مشتریان باعث فروش کالای بیشتری می شوند. در صورتی که این پیشنهادات صحیح و مناسب باشند می توانند منجر به افزایش چشمگیر فروش محصولات در فروشگاه وب سایت شوند.

---

۱ Collaborative Filtering
۲ Collaboration
۳ Search

وفاداری

در دنیای وب و کسب و کارهای الکترونیکی فاصله هر فروشگاه با فروشگاه دیگر به اندازه یک یا دو کلیک است و این بدین مفهوم است که در صورتیکه مشتری با صرف اندکی وقت کالای مورد نظر خود را نتواند پیدا کند به آسانی به فروشگاه دیگری مراجعه می کند.سیستم های توصیه گر با ایجاد یک رابطه ارزش افزوده بین مشتری و وب سایت باعث افزایش وفاداری مشتریان می شود. وب سایت ها بر روی درک نیازهای مشتریان خود و یادگیری سلایق آنها از طریق سفارشی سازی ها[۱] و شخصی سازی ها[۲] سرمایه گذاری می کنند و مشتریان نیز با برگشت به سایت هایی که نیاز آنها را به درستی شناخته اند به این سرمایه گذاری پاسخ می دهند.در واقع ایجاد یک رابطه بین مشتریان و سایت باعث افزایش وفاداری آنها می گردد و این وفاداری منجر به ایجاد یک مزیت رقابتی می شود. به تعبیر پین حتی اگر رقیب یک سایت قابلیت ها و امکانات دقیقا مشابه ارائه دهد، یک مشتری باید زمان و انرژی قابل توجهی را صرف آموزش آن سایت نماید تا چیزهایی را به آن سایت بیاموزد که قبلا به سایت اولی آموخته است.

(Your Customers forever?, ۱۹۹۵, Do you want to keep &, pine, Peppers)

پس مشخص است که این ارتباط در صورتیکه به درستی مدیریت شود می تواند بسیار ارزشمند باشد.

---

۱ Customization
۲ Personalization

۲-۱۰ تکنیک های مورد استفاده در سیستم های توصیه گر

تکنیک های مختلف مورد استفاده در انواع سیستم های توصیه گر را می توان در چهار دسته کلی قرار داد. توصیه های غیر شخصی، توصیه های مبتنی بر خصوصیات، توصیه های آیتم های مرتبط و توصیه های اشخاص مرتبط این طبقه بندی را تشکیل می دهند که در ادامه به شرح مختصری از هریک پرداخته شده است.

۲-۱۰-۱ توصیه های غیر شخصی[۱]

در این تکنیک، سیستم های توصیه گر اقدام به ارائه پیشنهادات خود بر اساس میانگین نظرات سایر مشتریان می نماید. پیشنهادات ارائه شده کاملا مستقل از مشتریان بوده و به طبع تمامی مشتریان توصیه های مشابهی را دریافت می نمایند. معمولا اینگونه سیستم ها کاملا خودکار بوده زیرا به کمترین کمک مشتریان برای تولید توصیه ها احتیاج دارد و نیز به دلیل استقلال پیشنهادات از مشتریان، توصیه های ارائه شده موقتی و گذرا بوده و پایدار نمی باشند و احتمالا برای یک مشتری در یک بازه زمانی یک سری توصیه ارائه شود و در بازه زمانی بعدی سری جدیدتری و این به این دلیل است که پیشنهادات بر اساس میانگین نظرات سایر مشتریان است. اینگونه متدهای توصیه معمولا در فروشگاه های فیزیکی ارائه می شود که در آنها تابلوهایی نصب می گردد تا پیشنهادات را به مشتریان اعلام نماید بدون اینکه با تغییر مشتریان تغییر کند. به عنوان مثال می توان به سیستم نمایش نظرات و ارزیابی های مشتریان در Amazon.com و MovieFinder.com اشاره نمود که بر اساس میانگین نظرات کلیه مشتریان می باشد. همچنین سیستم های Text Comment در آمازون و Feedback Profile در Ebay از این دسته سیستم های توصیه گر هستند.

[۱] Non-Personalized Recommendations

۲-۱۰-۲ توصیه های مبتنی بر خصوصیت[۱]

در این نوع از سیستم های توصیه گرپیشنهادات به مشتریان بر اساس خصوصیات محصولات (آیتم ها) انجام می پذیرد. به عنوان مثال در صورتی که مشتری عبارت "رمان های تاریخی" راجستجو نماید و سیستم به وی چندین کتاب معرفی نماید، این فرآیند در این طبقه دسته بندی می شود و به آن توصیه های مبتنی بر خصوصیت اطلاق می شود. اینگونه سیستم ها عموما دستی انجام می پذیرد، زیرا که مشتری می بایست دقیقا مشخص نماید که به دنبال چه چیزی می گردد. به عنوان مثال می توان به Movie Map در Reel.com اشاره نمود.در این سیستم به دلیل اینکه مشتری باید دقیقا مشخص نماید که به دنبال چه چیزی می گردد دستی و به دلیل اینکه سیستم تمایلات مشتری را به خاطر نمی سپارد گذرا محسوب می شود.هرچند می توان با شناسایی مشتریان و به خاطر سپاری تاریخچه جستجوهای انجام شده و نیز گزینه های انتخاب شده توسط مشتری اینگونه سیستم ها را به سمت پایداری پیش برد. کاری که در سیستم های جستجوی گوگل نیز اخیرا راه اندازی شده است. در سیستم های جدید گوگل در صورتی که عضو گروه های خدماتی گوگل باشید (مثل Gmail, Picasa, Google App و ...)و در لحظه جستجو به سیستم وارد شده باشید، گوگل شما را شناسایی کرده و جستجوها و لینک هایی را که برای مشاهده انتخاب می نمایید ذخیره می نماید. همچین امکانی را ارئه می دهد که شما می توانید مشخص کنید که کدام لینک از نظر شما بهتر است، در این صورت در مراجعت های بعدی و جستجوهای آتی این ترجیهات در نظر گرفته می شود و نتایجی به شما ارائه می شود بیشتر شبیه آنهایی باشد که قبلا انتخاب کرده اید. در برخی مقالات (Mobashera &, Ramezania, Bergmanb, Thompsonb, Burkea) اشاره شده است که به این تکنیک Content-Based Filtering یا به اختصار CB گفته می شود.

همچنین در مقاله دیگری تحت عنوان به سوی نسل جدید سیستم های توصیه گر

[۱] Attribute-Based Recommendations

(Adomavisius & Tuzhilin, ۲۰۰۵) این تکنیک به دو گروه اکتشاف پایه[۱] و مدل پایه[۲] تقسیم بندی شده است که متدهایی نظیر TF-IDF (IR) و خوشه بندی[۳] را در گروه اول و طبقه بندی بیزی[۴] ، خوشه بندی، درخت تصمیم[۵] و ANN را در گروه دوم قرار داده است.

۳-۱۰-۲ توصیه های آیتم های مرتبط[۶]

در اینگونه سیستم های توصیه گر، توصیه ها بر مبنای مجموعه ای از آیتم ها و محصولاتی است که کاربر به آن علاقه نشان داده است. به عنوان مثال مشتری را در نظر بگیرید که در یک فروشگاه در سبد خرید خود تعدادی محصول قرار داده است. در این وضعیت سیستم توصیه گر می تواند بر اساس محصولات خریداری شده (موجود در سبد خرید مشتری) به وی سایر محصولات تکمیلی مشابه را نیز پیشنهاد کند. این نوع از سیستم های توصیه گر می تواند خودکار باشد، اگر سیستم بر اساس رفتارهای مشتریان خود اقدام به استخراج الگوها نماید و نیز می تواند دستی باشد در صورتی که از مشتریان خواسته شود تا اطلاعاتی راجع به خود و ترجیهات (تمایلات) خود به سیستم ارائه دهند. همچنین این سیستم ها معمولا گذرا بوده و سابقه ای از توصیه های ارائه شده به مشتریان ثبت و نگهداری نمی نمایند، حال اینکه همانطور که در بخش قبلی نیز توصیف شد می توان با ثبت سابقه و تاریخچه توصیه ها سیستمی پایدار داشت.

Movie Matches در Rell.com، Match Maker در Movie Finder و Customer Who Bought در آمازون از جمله این سیستم ها می باشند. این سه سیستم گذرا و خودکار می باشند اما Album

---

۱ Heuristic-Based
۲ Model Based
۳ Clustering
۴ Bayesian Classifiers
۵ Decision Tree
۶ Item to Item Correlation

Advisor در CDNOW گذرا و دستی است زیرا که بر اساس در خواست کاربر اقدام به ارائه توصیه ها می نماید که این مقدار تعامل مشتری را با سیستم طلب می نماید.

۲-۱۰-۴ توصیه های اشخاص مرتبط[۱]

در اینگونه سیستم ها توصیه ها به مشتریان بر اساس مشابهت های مشتریان (افراد) انجام می گیرد. به عنوان مثال یک فروشگاه آنلاین را درن نظر بگیرید ، در صورتی که مشتری که در حال خرید است بر اساس معیارهای خاص دارای مشابهت هایی با گروه دیگری از مشتریان داشته باشد، توصیه های ارائه شده به وی بر اساس تمایلات (ترجیهات) و خریدهای اشخاص آن گروه در این سیستم می باشد. این نوع از سیستم های توصیه گر، فیلترینگ مشارکتی[۲]، نام دارندکه به اختصار CF گفته می شود. این نام گذاری به این دلیل است که در این سیستم ها از تکنیک های تصفیه اطلاعات و نظرات اعضای گروه به منطور تولید توصیه ها استفاده می شود.به عبارت دیگر در این تکنیک ها از دانش اجتماعی[۳] استفاده می شود. همچنین واژه ارتباط[۴] استفاده شده در نام این نوع سیستم ها بیانگر تکنیک های نزدیکترین همسایه[۵] است که در اینگونه سیستم ها مورد استفاده قرار می گیرد و پایه آن ارتباط خطی[۶] است. این نوع از سیستم ها خودکار بوده و به دلیل اینکه آنها به صورت خودکار اقدام به ارائه توصیه به مشتریان خود می کنند. همچنین در این سیستم ها رفتار مشتریان به صورت خودکار ثبت و ظبط می گردد تا الگوه های موجود در آنها شناسایی شوند. گرچه در برخی سیستم های CF سعی می شود تا از نظرات کاربران در

---

۱ People to People Correlation
۲ Collaborative Filtering
۳ Social Knowledge
۴ Correlation
۵ Nearest-Neighbor
۶ Linear Correlation

موقعیت های مختلف پرسیده شود که اغلب این کار توسط مشخص نمودن یک مقدار رتبه ای ۵ ارزشی یا ۷ ارزشی انجام می شود که در این وضعیت سیستم مزبور به سمت دستی بودن حرکت می کند. به طور کلی می توان گفت که دستی بودن یا خودکار بودن و نیز گذرا بودن یا پایدار بودن سیستم های توصیه گرکاملا نسبی است و بر روی طیفی از حداکثر خودکار بودن تا حداقل آن و نیز حداگثر پایداری در مقابل آن می توانند قرار بگیرند.

Book Matcher در آمازون، We Predict در MovieFInder و Style Finder در Levis از اینگونه سیستم ها به شمار می روند.

۲-۱۱ انواع روش های دریافت اطلاعات از کاربران در سیستم های توصیه گر

همانطور که در بحث قبل اشاره شد در برخی سیستم های توصیه گر لازم است تا از کاربران اطلاعاتی دریافت شود تا در تولید توصیه ها مورد استفاده قرار گیرد. معمولا به روش های مختلفی می توان این اطلاعات را از کاربران (مشتریان) جمع آوری کرد و بدست آورد، اما می توان به طور کلی روش ها را به چهار دسته طبقه بندی کرد که در ادامه به تشریح هر یک پرداخته شده است.

- **داده های خرید**

یکی از منلبع و روش هایی که برای دریافت اطلاعات مورد استفاده قرار می گیرد داده های خرید و یا همان سبد خرید مشتریان است. Customer who bought در آمازون و My CDNOW نیز از جمله سیستم هایی هستند که از این روش برای جمع آوری اطلاعات لازم و ارائه توصیه ها بهره می برند که مبنی بر شناسایی الگوهایی خرید بین مشتریان است.

- **Likert**

در برخی سیستم ها برای اگاهی از نظر مشتریان راجع به یک محصول یا موضوع از طیف Likert استفاده می کنند که در آن مشتری می بایست ارزشی از بین مقادیر ۵ گانه و یا ۷ گانه مرتب شده انتخاب نماید. این ارزش ها می تواند عددی باشد و یا متنی ولی در هر صورت می بایست از بی ارزش ترین تا با ارزش ترین مرتب شده باشد. سیستم هایی نظیر Feedback Profile در ebay و Style Finder در Levis از این مکانیزم برای دریافت نظرات مشتریان خود استفاده می کنند.

- **Text**

یکی دیگر از روش های دریافت اطلاعات از مشتریان ارئه امکان نوشتن نظرات آنها بصورت متن است. در این روش معمولا می بایست این اطلاعات توسط اشخاص (مدیران و اپراتورهای سیستم) مورد بررسی و بهره برداری قرار گیرد زیرا که پردازش این اطلاعات توسط سیستم های گامپیوتری بسیار دشوار بوده و هنوز در مراحل اولیه و بسیار ابتدایی است. به عنوان مثال Customer Comment در آمازون از جمله این سیستم ها به شمار می رود.

- **Editor's Choice**

آخرین روشی که در سیستم های توصیه گر مورد استفاده قرار می گیرد انتخاب گزینه مناسب توسط مشتریان از بین گزینه هایی است که مدیر یا مسئول آن بخش در نظر گرفته است. Match Maker در MovieFinder و نیز Movie Matchدر Rell.com از این روش استفاده نموده اند.

۱۲-۲ فرایند داده کاوی در مدیریت ارتباط با مشتری

داده کاوی یکی از عناصر مدیریت ارتباط با مشتری است و می‌تواند به حرکت شرکت‌ها به سمت مشتری محوری کمک کند.

در فرایند داده کاوی در مدیریت ارتباط با مشتری داده‌های خام از منابع مختلفی جمع‌آوری می‌شوند و از طریق استخراج، ترجمه و فرایندهای فراخوانی به انبار داده این مدیریت وارد می‌شوند. در بخش مهیاسازی داده، داده‌ها از انبار خارج شده و به صورت یک فرمت مناسب برای داده کاوی در می‌آیند، بخش کشف الگو شامل چهار لایه است:

۱-سوال‌های تجاری مانند توصیف مشتری

۲-کاربردها مانند امتیازدهی، پیش‌گویی

۳-روش‌ها مانند خوشه‌بندی، طبقه‌بندی، درخت تصمیم

۴-الگوریتم‌ها ، در این بخش روش‌های داده کاوی با کاربرد مخصوص خود برای پاسخ به سوال‌های تجاری که به ذهن می‌رسند، الگوریتم‌هایی را استخراج می‌کنند و از این الگوریتم‌ها برای ساخت الگو استفاده می‌شود.

در بخش تجزیه و تحلیل الگو، الگوها به یک دانش مفید و قابل استفاده تبدیل می‌شوند و پس از بهبود آنها، الگوهایی که کارا محسوب می‌شوند در یک سیستم اجرایی به کار گرفته خواهند شد.

داده کاوی می‌تواند سودآوری مشتری‌های بالقوه را که می‌توانند به مشتریان بالفعل تبدیل شوند، پیش‌بینی کند و اینکه تا چه مدت به صورت مشتریان وفادار خواهند ماند و چگونه احتمالاً ما را ترک خواهند کرد. بعضی از مشتریان مرتباً مراجعاتشان را به شرکت‌ها برای کسب مزیت‌هایی که طی رقابت

میان آنها به وجود می‌آید، تغییر می‌دهند. در این صورت شرکت‌ها می‌توانند هدفشان را روی مشتریانی متمرکز کنند که سودآوری بیشتری دارند. بنابراین می‌توان از طریق داده کاوی ارزش مشتریان را تعیین، رفتار آینده آنها را پیش‌بینی و تصمیمات آگاهانه‌ای را دراین رابطه اتخاذ کرد.

یکی از نمونه‌های بارز داده کاوی را می‌توان در فروشگاه‌های زنجیره‌ای مشاهده نمود . فروشگاه‌های زنجیره‌ای مشتاق هستند بدانند که چه محصولاتی با یکدیگر به فروش می‌روند. برای مثال طی یک عملیات داده کاوی در یک فروشگاه زنجیره‌ای در آمریکا که بر روی حجم عظیمی از داده‌های فروش صورت گرفت، مشخص گردید که مردانی که برای خرید قنداق بچه به فروشگاه می‌روند معمولاً آب جو نیز خریداری می‌کنند. همچنین مشخص گردید مشتریانی که تلویزیون خریداری می‌کنند، غالبا گلدان کریستالی نیز می خرند. در یک عملیات داده کاوی در یک شرکت بزرگ تولید و عرضه پوشاک در اروپا مشخص شد که افرادی که کراوات‌های ابریشمی خریداری می‌کنند، در همان روز یا روزهای آینده گیره کراوات مشکی رنگ نیز خریداری می‌کنند. این نوع استفاده از داده کاوی می‌تواند فروشگاه‌ها را در برگزاری هوشمندانه فستیوال‌های فروش و نحوه ارائه اجناس به مشتریان یاری رساند. داده کاوی در یک شرکت بزرگ تکثیر و عرضه فیلم‌های سینمایی در آمریکا نشان داد که روابط مشتریان و هنرپیشه‌های سینمایی و نیز گروه‌های مختلف مشتریان بر اساس سبک فیلم‌ها (ترسناک، رمانتیک، حادثه‌ای و ...) مشخص می گردد.بنابراین آن شرکت به صورت کاملا هوشمندانه می‌توانست مشتریان بالقوه فیلم‌های سینمایی را بر اساس علاقه مشتریان به هنرپیشه‌های مختلف و سبک‌های سینمایی شناسایی کند.

۲-۱۳ فروشگاه های زنجیره ای

۲-۱۳-۱ تاریخچه فروشگاه های زنجیره ای

قرون هیجدهم و نوزدهم میلادی در اروپا مبدا تحول در روشهای تولید، حمل ونقل و ... بوده است. سیستم توزیع نیز که رابط بین تولید و مصرف است از این تغییرات بی نصیب نماند. لذا، نهادهای مدرن توزیع شکل گرفتند که فروشگاه های زنجیره ای یکی از آنهاست. فروشگاه‌های زنجیره ای به شبکه ای از خرده فروشی های کوچک و بزرگ اطلاق می شود که کالای یکسانی را با مالکیت انحصاری ارایه می نمایند. به بیان دیگر، فروشگاه های زنجیره ای را می توان مجموعه فروشگاه هایی در نظر گرفت که دارای مدیریت واحد بوده و در سطح یک شهر، کشور یا حتی چندین کشور پراکنده اند. این فروشگاه ها از عنوان تجاری یکسانی استفاده می کنند و معمولا محصولات مشابه یا تقریبا مشابهی ارایه می کنند. وظایف و کارکردهای عمده یک فروشگاه زنجیره ای (خرید، تبلیغات، اجاره و ...) توسط یک واحد فرماندهی مرکزی یا یک مدیریت مرکزی کنترل می شوند.

آغاز تاسیس فروشگاههای زنجیرهای به سال ۱۷۹۲ در انگلستان باز می گردد. این فروشگاه که واقع در لندن بود، به فروش کتاب، مجله، روزنامه و وسایل سرگرمی اختصاص داشت. از آن زمان به بعد، ایده بهره گیری از چنین واحدهای زنجیره ای در سایر کشورها، به ویژه کشورهای اروپایی و ایالات متحده آمریکا، مورد استقبال قرار گرفت.

۲-۱۳-۲ مزایای فروشگاه های زنجیره ای

به تدریج این نهاد نوین توزیع نشان داد که علاوه بر سودآوری (برای صاحبان سرمایه در بخش خصوصی)، برای جامعه نیز مزایای عمده ای به ارمغان می‌آورد. مهم ترین این مزایا را می توان به شرح موارد زیر برشمرد:

- تسهیل دسترسی به کالاهای مورد نیاز جامعه؛

• سهولت نظارت دولت بر عملکرد بخش خرده فروشی؛

• تثبیت قیمت کالا در مناطق مختلف شهری باتوجه به یکسان بودن قیمت انواع کالاهای عرضه شده در کلیه شعب؛

• کاهش هزینه های توزیع از طریق حذف واسطه های غیرضروری میان تولیدکننده و مصرف کننده؛

• امکان فروش کالاها در قیمت پایین تر به سبب مقیاس بالای خرید فروشگاه در مقایسه با خرده فروشی های کوچک.

• آزادی انتخاب بیشتر با توجه به تنوع کالایی؛

• ترویج فرهنگ استانداردکردن کالاهای تولید داخلی (از طریق ضرورت استفاده از بارکد ملی و اجتناب از فروش کالاهای غیراستاندارد واحدهای زنجیره ای).

۲-۱۳-۳ ویژگی های فروشگاه های زنجیره ای

فروشگاههای زنجیره ای از ویژگیهای خاصی نیز برخوردار هستند که تاثیر مستقیم آن بر بخش عرضه وارد میشود. این ویژگی ها عبارتند از:

• **عملکرد توام با شفافیت بالا و امکان برخورداری از منافع برنامه ریزی جامع در این واحدها**. نظر به وسعت عمل و قدرت مالی واحدهای زنجیره ای، این واحدها میتوانند برای کنترل موجودی، ثبت سفارشات مشتری، پیش بینی فروش، تجزیه وتحلیل اطلاعات و سازوکار فروش از مدرن ترین تجهیزات و فناوری ها استفاده کنند. لذا، با ایجاد دسترسی به اطلاعات به روز در این واحدها، امکان برنامه ریزی دقیق تر فراهم می آید.

• **افزایش سطح تقاضای مقابل واحدها در نتیجه خرید همراه**. وقتی که اجناس و کالاهای مختلف و متعدد در زیر یک سقف کنار هم قرار گرفته باشند. به فروش یکدیگر کمک می کنند. به

عبارتی، در بسیاری مواقع، بخشی از فروش در واحدها به صورت برنامه ریزی نشده است که درصد بالایی را به خود اختصاص می دهد.

• **برخورداری از قدرت چانه زنی بالا.** حجم بسیار بالای خرید در واحدهای زنجیره ای، آنها را قادر می سازد که از موضع قدرت، با تولیدکنندگان و عمده فروشی ها، معامله کنند. همچنین این واحدها به واسطه حجم فروش بالایی که دارند، با سرعت و سهولت بیشتری می توانند از بانک ها و سایر موسسات اعتباردهنده تقاضای وام کنند.

۲-۱۳-۴ فروشگاه های زنجیره ای در ایران

سابقه فروشگاه های زنجیره ای در ایران به سال ۱۳۲۸ و تشکیل شرکت تعاونی مصرف ارتش (تعاونی مصرف سپه) باز می گردد. پس از آن با شکل گیری فروشگاه هایی باعنوان شهر و روستا در سال ۱۳۳۲ ، کوروش (قدس) در سال ۱۳۵۰ و فروشگاه بزرگ ایران، توسعه فروشگاه های زنجیره ای در کشور شکل گرفت.

در دوران پس از پیروزی انقلاب اسلامی، وقوع جنگ تحمیلی و مشکلات ناشی از آن موجب شد تا ایجاد فروشگاه های زنجیره ای جدید در اولویت نباشد. با این حال، با شروع دوران سازندگی و به دلیل وجود مشکلات فراوان در بخش توزیع کالاها که بیشتر متاثر از دو عامل وضعیت نابسامان اقتصادی و ساختار نامناسب نظام توزیع سنتی بود، دولت برآن شد تا با تاسیس فروشگاه های نوین زنجیره ای، وضعیت توزیع را در کشور بهبود بخشد و از آشفتگی در بازار جلوگیری کند. در این راستا شهرداری تهران به عنوان متولی ارایه و اشاعه خدمات مطلوب شهری پیش قدم شد. لذا، در آبان ماه سال ۱۳۷۴ فروشگاه شماره یک شهروند در میدان آرژانتین تحت عنوان فروشگاه بیهقی افتتاح شد. پس از آن فروشگاه های دیگری از جمله فروشگاه‌های رفاه، اتکا و کوثر برای عرضه مایحتاج عمومی، و فروشگاه های زنجیره ای شیلات (عرضه کننده انواع ماهی وآبزیان دریایی)، پاتریس (لوازم صوتی و تصویری)، صنعت نفت

(محصولات شیمیایی)، زندگی بهتر (لوازم خانگی) و ... به عنوان فروشگاه های عرضه کننده محصولات خاص شکل گرفتند.

باید توجه داشت که با وجود گسترش و تعدد فروشگاه های زنجیره ای در کشور و پیشرفتی که در این زمینه پدید آمده است، همچنان نقایصی در این ارتباط وجود دارد. بر این اساس و نظر به اهمیت موضوع ، یکی از محورهای طرح تحول اقتصادی "ساماندهی شبکه توزیع کالا و خدمات" قرار داده است که یکی از پروژه های دهگانه ذیل آن "ایجاد شبکه های زنجیره ای خرده فروشی اصناف" می‌باشد. در این راستا، اقداماتی همچون"صدور ۵ فقره مجوز قطعی تاسیس فروشگا ه زنجیره ای"، "دریافت ۹ فقره درخواست برای تاسیس فروشگاه زنجیره ای خرده فروشی" و"اطلاع رسانی به مجامع امور صنفی کل کشور توسط شورای اصناف"صورت گرفته است.

بدیهی است توسعه فروشگاه های زنجیره ای خرده فروشی در کشور با یکسان سازی قیمت ها ، تشدید تقارن اطلاعات در بازار، کاهش هزینه کنترل نرخ اجناس و تامین و توزیع متمرکز کالا، زمینه ساز عملکرد هرچه کاراتر نظام توزیع در اقتصاد ایران خواهد بود.

فصل سوم

روش شناسی تحقیق

۳-۱ مقدمه

۳-۱-۱ شناخت کسب و کار (Business Understanding)

۳-۱-۲ شناخت داده ها (Data Underestanding)

۳-۱-۳ آماده سازی داده ها (Data Prepration)

۳-۲ تعریف قوانین وابستگی (Assosiation Rule)

۳-۳ الگوریتم Apriori

۳-۴ الگوریتم TOPSIS

۳-۵ الگوریتم DEMATEL

۳-۶ داده های تحقیق

۳-۶-۱ شیوه و ابزار جمع آوری داده ها

۳-۶-۲ نوع داده ها و مقیاس آنها

۳-۶-۳ جامعه آماری و نمونه گیری

۳-۷ ساختار اجرایی تحقیق

۳-۷-۱ درک مساله کسب و کار

۳-۷-۲ درک داده ها

۳-۷-۳ آماده سازی داده ها

۳-۸جمع بندی

۳-۱ مقدمه

در این فصل ساختار مفهومی روش تحقیق تشریح می شود، برای همین منظور ابتدا مباحث مرتبط با شناخت کسب و کار و داده ها بررسی شده، سپس الگوریتم های استفاده شده برای تجزیه و تحلیل داده ها و تئوری های آنها بیان شده است و نهایتا ساختار اجرایی تحقیق بیان شده است.

۳-۱-۱ شناخت کسب و کار[۱]

همانطور که در بخش های قبلی هم ذکر شد، با گسترش روزافزون فن آوری اطلاعات و تکنولوژی های نرم افزاری شاهد نفوذ آن در کلیه جنبه های اجتماعی و اقتصادی هستیم.یکی از بخش های موثر و قابل توجه هم در اقتصاد و هو در حوزه خانواده، مساله خرید است. عموما خانواده ها برای تامین منابع مورد نیاز خود به فروشگاه های بزرگ و زنجیره ای مراجعه کرده و خرید انجام می دهند، این خرید گاه موثر نیست و خانواده ها مجبور هستندتا برای تامین منابع مورد نیاز خود (که از خرید اولیه بدست نیامده است) مجددا به فروشگاه های جزئی مراجعه کنند. این اقدام باعث اتلاف وقت خانواده ها، کاهش صرفه جویی های مالی ناشی از خرید از فروشگاه های زنجیره ای از طرف خانواده ها و عدم فروش بهینه و در نتیجه کاهش سود از طرف فروشگاه های زنجیره ای می شود. بدیهی است که در صورت ارائه راه حلی برای بهینه کردن خرید که منجر به سود آوری برای هر دو طرف یعنی خانواده ها و فروشگاه های زنجیره ای بسیار مطلوب خواهد بود. سیستم های توصیه گر خرید به نوعی راه حل این مساله می تواند باشد. سیستم های توصیه گر خرید به نوعی با تحلیل رفتار خرید مشتریان و شناسایی الگوهای خرید، به مشتریان کالاهایی را توصیه می کنند که احتمالا در سبد خرید فراموش شده اند که این عمل باعث

[۱] Business Understanding

افزایش حجم سبد خرید و به دنبال آن فروش بیشتر و بهینه تر را در فروشگاه ها خواهد داشت. اکنون در هیچ فروشگاه زنجیره ای و غیر زنجیره ای فعال در ایران این ساز و کار وجود نداشته و می تواند شروع خوبی برای راه اندازی سیستم های توصیه گر در فروشگاه های زنجیره ای باشد.

با توجه به ماهیت و هدف تحقیق از روش های بدون ناظر[۱] در داده کاوی استفاده خواهد شد که (روش های خوشه بندی[۲] و کشف قوانین پیوند[۳] جزوه این دسته قرار می گیرند) و ما نیز با استفاده از همین تکنیک های کشف قوانین پیوند اقدام به یافتن توصیه های لازم خواهیم کرد.

۳-۱-۲ شناخت داده ها[۴]

داده های مورد استفاده در این تحقیق از فروشگاه زنجیره ای اتکا بدست آمده است که در یک دوره یکساله جمع آوری شده است. این داده ها به دلیل جمع آوری از طریق صندوق های مکانیزه و نرم افزار های خاص فروشگاهی تقریبا تمیز[۵] بوده و اشتباهات ناشی از عامل انسانی که موثرترین عامل در کیفیت داده هاست به حداقل رسیده است. این داده ها در پایگاه های داده ای اتکا ذخیره سازی شده است که متاسفانه به دلیل کمبود امکانات کافی در نرم افزار مورد استفاده در اتکا امکان تهیه گزارشات با کیفیت مناسب برای داده کاوی وجود نداشته و صرفا گزارش فاکتورها در قالب فایل اکسل[۶] موجود می باشد.

---

۱ Un-Supervised
۲ Clustering
۳ Associate Rule Mining
۴ Data Underestanding
۵ Clean
۶ Excel

ردیف گردش ترمینال	زمان کنترل	فروش ریالی	پرداختی نقد	بن کاغذی	کارت اعتباری	کل تخفیف	کل مالیات	اعتبار حکمت کار	بن حکمت کار	بارکد کالا	شرح کالا
54248_54249_		19450,95	0	0	0	0	3919,02	125377,95	0	6,26007E+12	کره حیوانی 50 رامک
54248_54249_		0,93	0	0	0	0	3919,02	125377,95	0	2,00173E+12	نایلکس بزرگ سوپری
54248_54249_		23482,5	0	0	0	0	3919,02	125377,95	0	6,26111E+12	لپه 900 گرم گلستان
54248_54249_		25038,39	0	0	0	0	3919,02	125377,95	0	6,26032E+12	ماست1,5٪چربی2,5ک مان
54248_54249_		12787,5	0	0	0	0	3919,02	125377,95	0	6,26029E+12	خیارشوردرجه1شیشه700آ
54248_54249_		10641,06	0	0	0	0	3919,02	125377,95	0	6,26007E+12	خامه عسلی 100 گ رامک
54248_54249_		9746,4	0	0	0	0	3919,02	125377,95	0	6,26007E+12	خامه مربایی100گ رامک
54248_54249_		20311,2	0	0	0	0	3919,02	125377,95	0	6,26027E+12	کاسه پلاستیک0514 پرد
64016_		42256,41	0	0	712631	52280,69	18044,269	0	0	8,80112E+12	چوکوبای20تایی اوریون
64016_		9932,4	0	0	712631	52280,69	18044,269	0	0	6,26032E+12	آویشن شاسه 20 گرم
64016_		3504,24	0	0	712631	52280,69	18044,269	0	0	6,2601E+12	کرمدار پاکتی پرتغال
64016_		3504,24	0	0	712631	52280,69	18044,269	0	0	6,2601E+12	کرمدا پاکت مریم موز
64016_		9820,8	0	0	712631	52280,69	18044,269	0	0	6,26027E+12	کاسه کوچک102سپتا آر
64016_		7560,9	0	0	712631	52280,69	18044,269	0	0	6,26027E+12	سبد دورنگ کوچک202
64016_		6863,4	0	0	712631	52280,69	18044,269	0	0	6,2605E+12	پتی بورکاکائو90گ ناد
64016_		92628	0	0	712631	52280,69	18044,269	0	0	6,26028E+12	روغن مایع سرخ پت900گ
64016_		8048,22	0	0	712631	52280,69	18044,269	0	0	6,26053E+12	ماکارونی شلز سبزیجات
64016_		8226,78	0	0	712631	52280,69	18044,269	0	0	6,26033E+12	اسپاگتی700گ انسی ناص

شکل۳-۱ نمایی از گزارش حاوی داده های تحقیق

۳-۱-۳ آماده سازی داده ها[1]

به منظور آماده سازی داده ها برای استفاده در فرآیند داده کاوی لازم است تا داده های خام موجود در گزارش فاکتورها را ساختاربندی نمود و بعد از اتمام پروسه آماده سازی داده ها آنها را به شکل مطلوب برای وارد کردن به الگوریتم های داده کاوی آماده می کنیم. در این پروژه به منظور رسیدن به جواب بهینه داده های خام اولیه توسط الگوریتم های مختلفی اجرا شده است که آماده سازی آنها برای هر الگوریتم به صورت جداکانه به تفصیل در ادامه خواهد آمد. برای رسیدن به جواب های بهینه از حجم داده های بسیار بالایی استفاده شده است که انجام پروسه داده کاوی را عملا با استفاده از برنامه های نرم افزاری آماده غیر ممکن کرده است و برای تجزیه و تحلیل این حجم زیاد داده ها از کد نویسی در محیط نرم افزار MATLAB استفاده شده است که با جزئیات ارائه خواهد شد.

[1] Data Prepration

۲-۳ تعریف قوانین وابستگی

قوانین وابستگی، مطالعه صفات یا ویژگی‌هایی می‌باشد که با همدیگر وابستگی زیادی دارند. روش‌های تحلیل وابستگی، وابستگی بین مشخصه‌ها را کشف می‌کنند، به این معنی که آن‌ها قواعدی را پیدا می‌کنند که با استفاده از آن‌ها بتوان روابط بین یک یا چند مشخصه را کمی‌سازی نمود. همچنین قوانین وابستگی روابط و وابستگی‌های متقابل بین مجموعه بزرگی از اقلام داده‌ای را نشان می‌دهند. پیدا کردن چنین قوانینی می‌تواند در حوزه‌های مختلف مورد توجه بوده و کاربردهای متفاوتی داشته باشد. بعنوان مثال کشف روابط وابستگی بین حجم عظیم تراکنش‌های کسب و کار می‌تواند در تشخیص تقلب، در حوزه پزشکی و شخصی‌سازی مورد استفاده قرار گیرد. یا در طراحی کاتالوگ، بازاریابی و دیگر مراحل فرایند تصمیم‌گیری کسب و کار موثر باشد .

برای روشن شدن مطلب یک فروشگاه خرده فروشی را در نظر بگیرید. مشخصات اجناس خریده شده توسط هر مشتری در یک رکورد پایگاه داده ذخیره می‌شود. به هر رکورد یک شناسه (TID) نسبت داده می‌شود. فرض کنید که مجموعه I شامل تمام آیتم‌ها(اجناس) فروشگاه باشد. اگر مجموعه آیتم‌های $x, y \subset I$ باشد به‌طوری‌که $x \cap y = \emptyset$ (یعنی x و y هیچ اشتراکی نداشته باشند) باشد، آنگاه $x \Longrightarrow y$ یک قانون وابستگی است که بیان می‌کند اگر یک مشتری اجناس مجموعه x را بخرد، اجناس مجموعه y را نیز می‌خرد. این چنین قوانین، تأثیر مهمی در تعیین استراتژی‌های فروش، بخش‌بندی مشتریان، تنظیم کاتالوگ‌ها و... دارد.

به طور مثال فرض کنید اطلاعات مربوط به سبد خرید پنج مشتری یک فروشگاه را در اختیار داریم که در جدول (۲-۱) نمایش داده شده‌اند.

جدول ۳-۱ داده‌های مربوط به سبد خرید مشتریان یک فروشگاه

شماره مشتری	نان	شیر	گوشت	آبمیوه	تخم-مرغ	نوشابه
۱	۱	۱	۰	۰	۰	۰
۲	۱	۰	۱	۱	۱	۰
۳	۰	۱	۱	۱	۰	۱
۴	۱	۱	۱	۱	۰	۰
۵	۱	۱	۱	۰	۰	۱

در جدول ۳-۱ همانطور که ملاحظه می‌کنید، هر ردیف نمایانگر یک تراکنش (هر بار خرید) می‌باشد. در این جدول، ۱ ها به معنی این هستند که مشتری آن جنس را خریده و ۰ ها به این معنی هستند که مشتری آن جنس را نخریده است. با توجه به داده‌های موجود در جدول، برای مشتری ۱ به طور ساده می‌توان گفت اگر مشتری شماره ۱، نان خریده باشد آنگاه شیر هم خریده یا می‌توان گفت اگر مشتری شماره ۲ گوشت خریده آنگاه آبمیوه هم خریده است.

حال برای پیدا کردن قواعد اگر و آنگاهی که برای همه این رکوردها صادق باشند می‌توان این قواعد را بر اساس داده‌های موجود در جدول ۳-۱ استخراج نمود. به طور مثال قاعده الف می‌تواند این باشد، " اگر هر مشتری نان بخرد آنگاه شیر هم می خرد" البته با پشتیبانی[۱] ۶۰٪ از رکوردها و اطمینان[۲] ۷۵٪.

[۱] Support
[۲] Certainty

در قاعده الف، پشتیبانی ۶۰٪ به این معنی است که این قاعده فقط بر روی ۳ مشتری (مشتری ۱و۴و۵) از ۵ مشتری صادق می‌باشد. یعنی فقط همین مشتریان هستند که نان و شیر را با هم خریده‌اند. منظور از اطمینان این است که فقط ۳ مشتری (مشتری ۱و۴و۵) از ۴ مشتری (مشتری ۱و۲و۴و۵) که نان خریده‌اند، شیر هم خریده‌اند. مشتری ۲ نان خریده اما شیر نخریده است.

شاخص‌های پشتیبانی و اطمینان اغلب با قوانین وابستگی ذکر می‌شوند. هرچقدر میزان این دو شاخص به عدد ۱ (۱۰۰٪) نزدیک‌تر باشد، قاعده بدست آمده مورد اطمینان‌تر، محکم‌تر و قابل اعمال کردن می‌باشد. در حالت کلی برای استخراج قوانین وابستگی از یک مجموعه داده، شاخص‌های پشتیبان و اطمینان به صورت زیر تعریف می‌شوند.

تعریف ۳-۱:

فرض کنید $I = \{i_1, i_2, \dots, i_m\}$ مجموعه آیتم‌های موجود باشد که شامل m عضو است. به هر زیر مجموعه از مجموعه آیتم‌های (I) نیز' یک مجموعه آیتم ' می‌گوییم.

تعریف ۳-۲:

مجموعه آیتم $X \subseteq I$ را در نظر بگیرید، اگر تعداد مشتریانی که اقلام X را خریداری کرده‌اند برابر با ℓ باشد آنگاه آن‌را به‌صورت رابطه‌ی ۳-۱ نشان می‌دهیم.

$$\rho(x) = \ell \qquad (۱-۳)$$

مثلاً اگر مجموعه X شامل اقلام نان، شیر و گوشت باشد آنگاه $\rho(X) = ۲$ می‌باشد. یعنی اقلام نان، شیر و گوشت توسط دو مشتری با هم خریداری شده‌اند.

تعریف ۳-۳:

فرض کنید دو مجموعه آیتم X و Y که زیرمجموعه‌های مجموعه آیتم I هستند را داشته باشیم. قانون وابستگی تولید شده توسط این دو مجموعه آیتم را به‌صورت رابطه‌ی ۳-۲ نشان می‌دهند.

$$X \Longrightarrow Y \qquad (۲-۳)$$

تعریف ۳-۴:

مجموعه I شامل تمامی اقلام، و مجموعه آیتم‌های X و Y را در نظر بگیرید. پوشش قانون $X \Longrightarrow Y$ به صورت رابطه‌ی ۳-۳ تعریف می‌شود.

$$Support(X \Longrightarrow Y) = S(X \Longrightarrow Y) = \frac{\rho(X \,\&\, Y)}{m} \qquad (۳-۳)$$

که در آن منظور از $\rho(X \,\&\, Y)$ ، تعداد مشتریانی است که مجموعه اقلام X و Y را با هم خریداری کرده‌اند و m تعداد کل مشتریان می‌باشد. فاکتور پوشش بیانگر نسبت تعداد مشتریانی که مجموعه اقلام X و Y را با هم خریداری کرده‌اند به کل مشتریان است.

تعریف ۳-۵:

مجموعه I شامل تمامی اقلام، و مجموعه آیتم‌های X و Y را در نظر بگیرید. درجه اطمینان قانون $X \Longrightarrow Y$ به‌صورت رابطه‌ی ۳-۴ تعریف می‌شود.

$$Confidence(X \Longrightarrow Y) = \; C(X \Longrightarrow Y) \; = \frac{\rho(X \,\&\, Y)}{\rho(X)} \qquad (۴-۳)$$

فاکتور درجه اطمینان بیانگر نسبت تعداد مشتریانی است که مجموعه اقلام X و Y را با هم خریداری کرده‌اند به تعداد مشتریانی که اقلام X را خریداری کرده‌اند. برای مثال اگر مجموعه اقلام X را اقلام {نان

و شیر} و مجموعه اقلام Y را {گوشت و آبمیوه} در نظر بگیریم، آنگاه فاکتورهای پوشش و اطمینان برابرند با:

$$\rho(X) = ۳ \quad , \quad \rho(\mathrm{Y}) = ۳ \quad , \quad \rho(X \,\&\, \mathrm{Y}) = ۱$$

$$S(X \Longrightarrow \mathrm{Y}) = \frac{۱}{۵} \quad , \quad C(X \Longrightarrow \mathrm{Y}) = \frac{۱}{۳}$$

به همین صورت می‌توان برای تمامی قوانین وابستگی تولید شده فاکتورهای پوشش و اطمینان را محاسبه کرد. هدف کلی در کشف قوانین وابستگی استخراج قانون‌هایی از جدول اطلاعات است که دارای میزان پوشش و اعتبار قابل قبول باشند. میزان فاکتورهای پوشش و اطمینان مورد نظر توسط کارشناسان مربوطه تعیین می‌شوند. بنابراین تمامی قانون‌ها باید دارای حداقل پوشش و حداقل اطمینان تعیین شده باشند. برای شناسایی این قانون‌ها روش‌های مختلفی وجود دارد که یکی از سنتی‌ترین این روش‌ها روش Brute-Force می‌باشد. در این روش ابتدا تمامی قانون‌های ممکن تولید می‌شوند و از میان قانون‌های تولید شده، قانون‌هایی که در شرط حداقل پوشش و حداقل اطمینان صدق می‌کنند نگه داشته شده و بقیه قانون‌ها حذف می‌شوند. مشکل اساسی که این روش دارد این است که تعداد قانون‌هایی که در ابتدا باید از مجموعه آیتم‌ها تولید شود بسیار زیاد می‌باشد. به‌عنوان مثال اگر مجموعه آیتمی با ۱۰ عضو داشته باشیم آنگاه ۵۷۰۰۰ قانون از این مجموعه آیتم تولید می‌شود. بنابراین در عمل این روش ناکارآمد خواهد بود. ازاین‌رو الگوریتم‌هایی طراحی شدند که به کمک آنها قانون‌های مورد نظر را در زمان کمتر و هزینه‌ی کمتر استخراج می‌کند. ازجمله این الگوریتم‌ها می‌توان به الگوریتم اپریوری[۱]، الگوریتم ایکلات[۲] و الگوریتم

[۱] Apriori
[۲] Eclat

اف‌پی‌گروس[1] اشاره کرد که در ادامه به معرفی الگوریتم اپریوری می‌پردازیم. در همه‌ی این الگوریتم‌ها مسئله‌ی کشف قوانین وابستگی به سه مسئله‌ی زیر تقسیم می‌شود:

۱. پیدا کردن تمامی زیر مجموعه‌های مجموعه آیتم‌های I که تکرار (وقوع) آنها در پایگاه بیشتر از یک حد تعیین شده از لحاظ فاکتور پوشش است. "به مجموعه آیتم‌هایی که تعداد وقوع آنها در پایگاه داده بزرگتر (یا مساوی) حد تعیین شده است، مجموعه آیتم‌های بزرگ، و به بقیه مجموعه آیتم‌های کوچک می‌گویند."

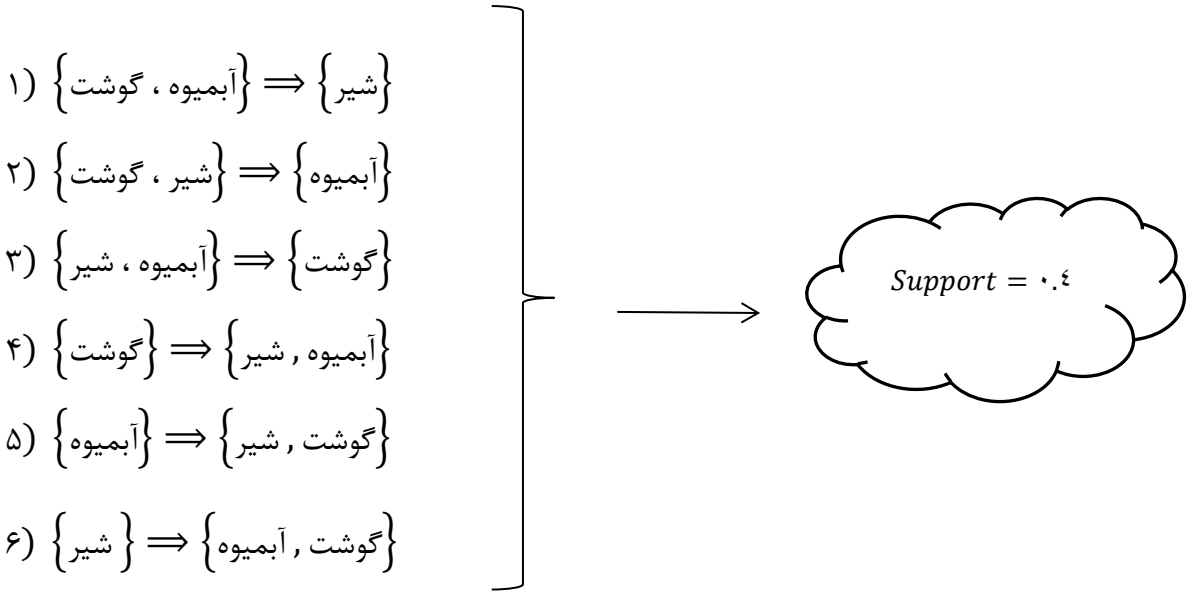
۲. بکارگیری مجموعه آیتم‌های بزرگ برای تولید قوانین.

۳. انتخاب قانون‌هایی که در شرط حداقل اطمینان صدق کنند.

به عبارت دیگر در همه‌ی این الگوریتم‌ها ابتدا مجموعه آیتم‌هایی که فاکتور پوشش آنها از حداقل میزان پوشش در نظر گرفته شده بزرگتر یا مساوی باشد (مجموعه آیتم‌های بزرگ) انتخاب می‌شوند و سپس برای هر کدام از مجموعه آیتم‌های بزرگ قانون‌های وابستگی ممکن را تولید می‌کند.

به عنوان مثال اگر برای استخراج قانون از در جدول (۲-۱) حداقل پوشش مورد نظر را ۰,۴ در نظر بگیریم آنگاه یکی از مجموعه آیتم‌هایی که در این شرط صدق می‌کند مجموعه آیتم {شیر، گوشت، آبمیوه} می‌باشد که دارای پوششی برابر با ۰/۴ می‌باشد. حال قانون‌های ممکن را از این مجموعه آیتم استخراج می‌کنیم که عبارتند از:

[1] FP-growth



به همین صورت قوانین وابستگی از سایر مجموعه آیتم‌هایی که در شرط حداقل پوشش صدق می‌کنند استخراج می‌شوند.

در مرحله‌ی بعد از میان تمام قانون‌های استخراج شده که دارای شرایط حداقل پوشش بودند، قانون-هایی انتخاب می‌شوند که در شرط حداقل اطمینان که توسط کارشناس معین می‌شود صدق کنند. به عنوان مثال اگر کارشناس مربوطه شرط حداقل اطمینان ۰,۸ را برای قانون‌های استخراج شده در نظر بگیرد، آنگاه برای مثال از میان قانون‌هایی که از مجموعه آیتم {شیر، گوشت، آبمیوه} استخراج شدند، خواهیم داشت:

۱) {آبمیوه ، گوشت} ⟹ {شیر} , $support = ۰.۴$, $confidence = ۰.۶۶$

۲) {شیر ، گوشت} ⟹ {آبمیوه} , $support = ۰.۴$, $confidence = ۰.۶۶$

۳) {آبمیوه ، شیر} ⟹ {گوشت} , $support = ۰.۴$, $confidence = ۱$

۴) {گوشت} ⟹ {آبمیوه , شیر} , $support = ۰.۴$, $confidence = ۰.۵$

۵) {آبمیوه} ⟹ {گوشت , شیر} , $support = ۰.۴$, $confidence = ۰.۶۶$

۶) {شیر} ⟹ {گوشت , آبمیوه} , $support = ۰.۴$, $confidence = ۰.۵$

بنابراین از میان قانون‌های تولید شده، فقط قانون (۳) در شرط حداقل اطمینان صدق کرده است و بقیه قانون‌ها حذف می‌شوند. به همین صورت این عمل را برای سایر قانون‌های تولید شده توسط سایر مجموعه آیتم‌های انتخابی در شرط اول، تکرار می‌کنیم و نهایتاً شاهد یک دسته قانون وابستگی هستیم که در دو شرط حداقل پوشش ۰,۴ و حداقل اطمینان ۰,۸ صدق کرده‌اند.

دو مسئله‌ی اصلی که هر کدام از الگوریتم‌ها برای آن راه‌حلی ارائه کرده‌اند، "نحوه‌ی استخراج مجموعه آیتم‌های دارای شرط حداقل پوشش" و " نحوه‌ی تولید قانون‌های دارای حداقل اطمینان" می-باشد. در ادامه نحوه‌ی شناسایی مجموعه آیتم‌های بزرگ و مجموعه قانون‌های با اطمینان بالا را توسط الگوریتم اپریوری معرفی می‌کنیم.

۳-۳ الگوریتم Apriori

الگوریتم اپریوری در سال ۱۹۹۴ توسط آگراوال[۱] و همکاران پیشنهاد شد. این الگوریتم به منظور شناسایی مجموعه آیتم‌های بزرگ به‌صورت زیر عمل می‌کند.

۱. ابتدا مجموعه آیتم‌های تک‌عضوی که در شرط حداقل پوشش صدق می‌کنند را انتخاب می‌کند.

۲. با استفاده از مجموعه آیتم‌های تک‌عضوی، مجموعه آیتم‌های با ۲ عضو می‌سازد. سپس شرط حداقل پوشش را برای مجموعه آیتم‌های تولید شده در این مرحله بررسی می‌کند و هر کدام از مجموعه آیتم‌های تولید شده در این مرحله در شرط حداقل پوشش صدق نکنند، حذف می‌کند.

۳. عمل انجام شده در مرحله‌ی دوم را برای مجموعه آیتم‌های تولید شده در مرحله‌ی دوم که در شرط حداقل پوشش صدق کردند تکرار می‌کند با این تفاوت که در این مرحله باید به این نکته توجه شود که مجموعه‌های سه عضوی ساخته شده نباید شامل مجموعه آیتم‌هایی باشد که در مرحله‌ی دوم حذف

[۱] Agrawal

شدند. مرحله‌ی سوم به هین صورت تا تولید بزرگ‌ترین آیتم‌های دارای شرط حداقل پوشش، تکرار می‌شود. بدیهی است اگر در مرحله‌ای هیچ‌یک از مجموعه آیتم‌ها در شرط حداقل پوشش صدق نکردند، الگوریتم متوقف می‌شود.

مثال ۳-۱:

برای جدول ۳-۱، مجموعه آیتم‌های بزرگی که دارای شرط حداقل پوشش ۰,۶ می‌باشند با استفاده از الگوریتم اپریوری به‌صورت زیر تولید می‌شوند.

مرحله‌ی اول:

کوچکترین مجموعه آیتم موجود در جدول ۳-۱، به‌صورت مجموعه‌های تک عضوی در جدول ۲-۲ می‌باشند.

جدول (۳-۲)- نتایج مرحله‌ی اول الگوریتم اپریوری برای شناسایی مجموعه آیتم‌های بزرگ جدول ۳-۱

فاکتور پوشش	مجموعه آیتم
۰,۶	{آبمیوه}
۰,۸	{نان}
۰,۴	{نوشابه}
۰,۸	{گوشت}
۰,۸	{شیر}
۰,۲	{تخم‌مرغ}

از میان مجموعه آیتم‌های جدول ۳-۲، آیتم {تخم‌مرغ} و {نوشابه} دارای شرط حداقل پوشش ۰,۶ نیستند و از جدول آیتم‌ها حذف می‌شوند.

مرحله‌ی دوم:

با استفاده از مجموعه آیتم‌هایی که در مرحله‌ی اول که در شرط حداقل پوشش ۰,۶ صدق کردند، مجموعه آیتم‌های دو عضوی می‌سازیم که به صورت جدول ۳-۳ می‌باشند.

جدول (۳-۳)- نتایج مرحله‌ی دوم الگوریتم اپریوری برای شناسایی مجموعه آیتم‌های بزرگ جدول ۳-۱

فاکتور پوشش	مجموعه آیتم
۰,۴	{نان، آبمیوه}
۰,۶	{گوشت، آبمیوه}
۰,۴	{شیر، آبمیوه}
۰,۶	{نان، گوشت}
۰,۶	{نان، شیر}
۰,۶	{گوشت، شیر}

با توجه به جدول ۳-۳، مجموعه آیتم‌های {نان ، آبمیوه} و {شیر ، آبمیوه} در شرط حداقل پوشش ۰,۶ صدق نکرده و از این جدول حذف می‌شوند.

مرحله‌ی سوم:

با استفاده از مجموعه آیتم‌هایی که در مرحله‌ی دوم که در شرط حداقل پوشش ۰,۶ صدق کردند، مجموعه آیتم‌های سه عضوی می‌سازیم که به‌صورت جدول ۳-۴ می‌باشند.

جدول (۳-۴)- نتایج مرحله‌ی سوم الگوریتم اپریوری برای شناسایی مجموعه آیتم‌های بزرگ جدول ۳-۱

فاکتور پوشش	مجموعه آیتم
۰,۶	{نان، گوشت، شیر}

علت اینکه مثلاً مجموعه آیتم سه عضوی {نان، گوشت، آبمیوه} در مرحله‌ی سوم تولید نشد این است که در مرحله‌ی دوم مجموعه آیتم دو عضوی {نان ، آبمیوه} دارای شرایط حداقل پوشش نبوده‌اند، بنابراین هیچ مجموعه آیتم سه عضوی شامل {نان ، آبمیوه} نمی‌توان یافت که در شرط حداقل پوشش صدق کنند. بنابراین تنها مجموعه‌ی آیتم سه عضوی که می‌توان از اجتماع مجموعه آیتم‌های دو عضوی ساخت، مجموعه آیتم {نان، گوشت، شیر} می‌باشد. اما با توجه به فاکتور پوشش این مجموعه آیتم، می-بینیم که در شرط حداقل پوشش صدق نکرده است. بنابراین بزرگترین مجموعه آیتمی که می‌توان از جدول ۳-۱ استخراج کرد و دارای شرط حداقل پوشش است، مجموعه آیتم‌های دو عضوی {گوشت، آبمیوه}، {نان، گوشت}، {نان، شیر} و {گوشت، شیر} می‌باشند.

پس از شناسایی مجموعه آیتم‌های بزرگ، نوبت به استخراج قوانین وابستگی دارای حداقل اطمینان مورد نظر می‌رسد. فرض کنید اگر مجموعه آیتم بزرگ را f بنامیم و s را مجموعه آیتمی زیرمجموعه‌ی

f در نظر بگیریم آنگاه قانون‌هایی که دارای حداقل اطمینان مورد نظر هستند به‌صورت زیر تولید می-شوند:

قانون $s \Longrightarrow f$ یک قانون وابستگی دارای حداقل اطمینان لازم است اگر رابطه‌ی ۳-۵ برقرار باشد:

$$\frac{Support(f)}{Support(s)} \geq \min Confidence \qquad (۵-۳)$$

مثال ۳-۲:

برای مجموعه آیتم‌های بزرگ استخراج شده در مثال ۳-۱، قانون‌های وابستگی که دارای حداقل اطمینان ۰٫۸ هستند را می‌یابیم.

برای مجموعه آیتم {گوشت، آبمیوه} اگر $f = \{$آبمیوه ، گوشت$\}$ در نظر بگیریم آنگاه خواهیم داشت:

۱. $if\ s = \{$آبمیوه$\}$; $s \Longrightarrow (f - s)$, $\frac{Support(f)}{Support(s)} = \frac{۰.۶}{۰.۶} = ۱ > ۰.۸$

۲. $if\ s = \{$گوشت$\}$; $s \Longrightarrow (f - s)$, $\frac{Support(f)}{Support(s)} = \frac{۰.۶}{۰.۸} = ۰.۷۵ \not> ۰.۸$

برای مجموعه آیتم {نان، گوشت} اگر $f = \{$نان ، گوشت$\}$ در نظر بگیریم آنگاه خواهیم داشت:

۳. $if\ s = \{$نان$\}$; $s \Longrightarrow (f - s)$, $\frac{Support(f)}{Support(s)} = \frac{۰.۶}{۰.۸} = ۰.۷۵ \not> ۰.۸$

۴. $if\ s = \{$گوشت$\}$; $s \Longrightarrow (f - s)$, $\frac{Support(f)}{Support(s)} = \frac{۰.۶}{۰.۸} = ۰.۷۵ \not> ۰.۸$

برای مجموعه آیتم {نان، شیر} اگر $f = \{$نان ، شیر$\}$ در نظر بگیریم آنگاه خواهیم داشت:

۵. $if\ s = \{$نان$\}$; $s \Longrightarrow (f - s)$, $\frac{Support(f)}{Support(s)} = \frac{۰.۶}{۰.۸} = ۰.۷۵ \not> ۰.۸$

$$۶.\ if\ s = \{\text{شیر}\} \quad ;\quad s \Longrightarrow (f-s) \quad ,\quad \frac{Support(f)}{Support(s)} = \frac{۰.۶}{۰.۸} = ۰.۷۵ \not> ۰.۸$$

برای مجموعه آیتم {گوشت، شیر} اگر $f = \{\text{گوشت ، شیر}\}$ در نظر بگیریم آنگاه خواهیم داشت:

$$۷.\ if\ s = \{\text{گوشت}\} \quad ;\quad s \Longrightarrow (f-s) \quad ,\quad \frac{Support(f)}{Support(s)} = \frac{۰.۶}{۰.۸} = ۰.۷۵ \not> ۰.۸$$

$$۸.\ if\ s = \{\text{شیر}\} \quad ;\quad s \Longrightarrow (f-s) \quad ,\quad \frac{Support(f)}{Support(s)} = \frac{۰.۶}{۰.۸} = ۰.۷۵ \not> ۰.۸$$

بنابراین تنها قانونی که در شرط حداقل اطمینان صدق می‌کند قانون ۱ می‌باشد.

$$\{\text{آبمیوه}\} \Longrightarrow \{\text{گوشت}\} \qquad , \qquad support = ۰.۴ \qquad , \qquad confidence = ۱$$

به این ترتیب قوانین وابستگی را می‌توان با استفاده از الگوریتم اپریوری از یک پایگاه داده استخراج کرد.

۴-۳ الگوریتم TOPSIS

مدل TOPSIS توسط هوانگ و یون در سال ۱۹۸۱ پیشنهاد شده است. این تکنیک بر اساس این مفهوم ایجاد شده است که گزینه های مناسب گزینه هایی هستند که حداقل فاصله را نسبت به راه حل ایده آل مثبت (بهترین حالت ممکن) و دورترین فاصله را نسبت به راه حل ایده آل منفی (بدترین حالت ممکن) داشته باشند (Lev, ۲۰۰۱). در این روش m گزینه به وسیله n شاخص، مورد ارزیابی قرار می گیرد. فرض بر این است که مطلوبیت هر شاخص، به طور یکنواخت افزایشی یا کاهشی است. حل مساله با این روش، مستلزم طی مراحل زیر است:

تاپسیس یکی از روش‌های MADM است که m گزینه برتر را با توجه به n معیار, رتبه‌بندی می‌کند. اساس این روش، انتخاب گزینه‌ای است که کمترین فاصله را از جواب ایده آل مثبت و بیشترین

فاصله را از جواب ایده آل منفی دارد. در این روش شاخصی تحت عنوان "نزدیکی نسبی گزینه iام بـه راه حل ایده آل $(C_i^+ \ or \ C_i^-)$"، معرفی می‌گردد وگزینه‌ای که دارای بیشترین C_i^+ است، انتخاب می‌گردد

.

$$C_i^+ = \frac{\sqrt{\sum_{j=1}^{n}(w_j \cdot \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}^2} - v_j^-)^2}}{\sqrt{\sum_{j=1}^{n}(w_j \cdot \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}^2} - v_j^+)^2} + \sqrt{\sum_{j=1}^{n} w_j \cdot \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}^2} - v_j^-)^2}} \qquad (۱)$$

و یا گزینه‌ای که دارای کمترین C_i^- است، انتخاب می‌شود.

(۲)

$$C_i^- = \frac{\sqrt{\sum_{j=1}^{n}(w_j \cdot \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}^2} - v_j^+)^2}}{\sqrt{\sum_{j=1}^{n}(w_j \cdot \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}^2} - v_j^+)^2} + \sqrt{\sum_{j=1}^{n} w_j \cdot \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}^2} - v_j^-)^2}}$$

که i مبین گزینه‌ها و j مبین شاخص‌هاست. w_j وزن شـاخص jام، x_{ij} مقـدار ارزش شـاخص iام بـرای گزینه jام، v_j^+ مقدار ایده آل مثبت برای شاخص jام، (برای شاخص‌هایی که جنبه منفی دارنـد، حـداقل و برای شاخص‌هایی که جنبه مثبت دارند، حداکثر) و v_j^- مقـدار ایـده آل منفـی بـرای شـاخص jام (بـرای شاخص‌هایی که جنبه مثبت دارند، حداقل و برای شاخص‌هایی که جنبه منفی دارند، حداکثر) است.

بنابراین مراحل کلی این روش به شرح ذیل می‌باشد :

گام۱: ساخت ماتریس تصمیم

$$D = \begin{array}{c} \\ A_1 \\ \vdots \\ A_i \\ \vdots \\ A_m \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & \cdots & x_j & \cdots & x_n \end{array} \\ \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & & & & \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & & \vdots & & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{bmatrix} \end{array}$$

$$W = (w_1, \ldots, w_j, \ldots, w_n)$$

گام ۲ : تبدیل ماتریس تصمیم‌گیری موجود, به یک ماتریس "بی مقیاس شده".

در ابتدا باید ماتریس تصمیم نرمالیزه گردد تا عناصر آن" بی مقیاس" شود. در این تحقیق نیازی به نرمالایز کردن ماتریس تصمیم نیست، زیرا تمامی لایه‌ها در کلاس یکسان ۰ تا ۱۰ قرار دارند و دارای مقیاس یکسان هستند.

گام ۳ : ایجاد ماتریس "بی مقیاس" وزین با مفروض بودن بردار w (به عنوان ورودی الگوریتم).

عناصر ماتریس نرمالیزه شده وزن دار (v_{ij}) با استفاده از رابطه زیر بدست می آید:

$$v_{ij} = r_{ij}.w_j \qquad (۳)$$

گام۴ : مشخص نمودن راه حل ایده آل مثبت و ایده آل منفی.

$$A^+ = [v_1^+, \ldots, v_j^+, \ldots, v_n^+]; \qquad v_j^+ = \max_i \{v_{ij}\}$$

$$A^- = [v_1^-, \ldots, v_j^-, \ldots, v_n^-]; \qquad v_j^- = \max_i \{v_{ij}\} \qquad (۴)$$

گام۵ : محاسبه اندازه فاصله از ایده آل مثبت و منفی.

برای بدست اوردن فاصله هر گزینه از ایده‌آل‌های مثبت و منفـی، دو روش وجـود دارد: روش اقلیدسـی و ورش بلوکی. در اینجا رابطه مربوط به روش اقلیدسی بیان می‌گردد:

(۵) فاصله گزینه iام از ایده آل مثبت

$$S_i^+ = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_{s_j^+})^2}$$

(۶) فاصله گزینه iام از ایده آل منفی

$$S_i^- = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_{s_j^-})^2}$$

گام ۶ : محاسبه نزدیکی نسبی گزینه‌ها به راه حل ایده آل.

این شاخص را جهت ترکیب کردن مقادیر S_i^- و S_i^+ و در نتیجه مقایسه گزینه‌ها نسبت به هـم تعریـف می‌کنیم، که با رابطه زیر قابل محاسبه است:

$$C = \frac{S_i^-}{S_i^+ + S_i^-} \qquad (۷)$$

گام ۷ : رتبه بندی گزینه‌ها براساس نزدیکی نسبی آنها به راه حل‌های ایده‌آل.(هر گزینـه ای کـه C آن بزرگ تر باشد، بهتر است)

۳-۵ الگوریتم DEMATEL

تکنیک دیمتل[1] توسط برنامه علوم و بشر انستیتو Battelle Memorial ژنو، بین سالهای ۱۹۷۲ و ۱۹۷۶ ایجاد شد و برای مطالعه و حل مسائل پیچیده و در هم تنیده مورد استفاده قرار گرفت. روش دیمتل مبتنی بر گرافهای جهت داری (دیاگرافهایی) است که می توانند عوامل دخیل در یک مسئله را به دو گروه علت و معلول تفکیک نمایند. این دیاگرافها، رابطه وابستگی میان عناصر یک سیستم را به تصویر می کشند، بطوریکه اعداد روی هر دیاگراف، بیانگر شدت تأثیر یک عنصر بر عنصر دیگر است. از این رو، روش دیمتل می تواند رابطه میان علتها و معلولهای عوامل را به یک ساختار سلسله مراتبی قابل درک از سیستم تبدیل نماید.

در سالهای اخیر روش دیمتل بدلیل نمایش ساختار روابط پیچیده عناصر در ژاپن بسیار مورد استفاده قرار گرفته است. در تحقیقات چیو و همکارانش، هوری و شی میزو ، لیو و همکارانش و لین و وو درباره کاربردهای موفق و متفاوت دیمتل در حوزه‌های یافتن استراتژی‌های بازاریابی، سیستم‌های کنترل، مسائل امنیتی خطوط هوایی ، گسترش صلاحیت مدیران جهانی و تصمیم گیری گروهی بحث شده است. تسایی چو نیز در سال ۲۰۰۹ در تحقیق خود کاربرد دیمتل را در انتخاب سیستم‌های مدیریتی در صنایع کوچک شرح داد.

اجرای تکنیک دیمتل شامل مراحل ذیل است:

۱. تعیین معیارهای موجود در عامل مورد بررسی
۲. تعیین روابط حاکم بین معیارها بر اساس مقایسات زوجی آنها
۳. تعیین روابط نهایی بین معیارها با توجه به توافق جمعی خبرگان و رسم دیاگراف روابط
۴. تعیین شدت روابط نهایی میان معیارهای مرتبط

[1] DEMATEL

۵. تشکیل ماتریس شدت روابط (ماتریس $\hat{M}$)

۶. تشکیل ماتریس شدت نسبی حاکم بر روابط مستقیم ($M = \alpha * \hat{M}$ بیشترین مجموع ردیفی ماتریس($\alpha = \hat{M}$)

۷. تشکیل ماتریس شدت نسبی موجود از روابط مستقیم و غیر مستقیم (ماتریس $S = M(I - M)^{-1}$)

۸. مشخص نمودن سلسله مراتب یا ساختار ممکن معیارها

در ماتریس S ، جمع سطری دارایه ها (R_k) و جمع ستونی درایه ها (J_k) و مجموع ($R_k + J_k$) و تفاضل ($R_k - J_k$) به ازای معیار kاُم محاسبه می گردد. مجموع ($R_k + J_k$) برای هر یک از معیارهای تشکیل دهنده سیستم، اهمیت (وزن) آن معیار در سیستم را نشان می دهد. مقدار (R_k) برای هر معیار نشانگر میزان تأثیر گذاری آن معیار بر سایر معیارهای سیستم و (J_k) مقدار متناظر با آن بیان کننده شدت تأثیر پذیری معیار مذکور از سایر معیارهای سیستم است. بنابراین ($R_k + J_k$) مشخص کننده مجموع تأثیر گذاری و تأثیر پذیری معیار مورد نظر در سیستم می باشد. بعبارتی معیاری که بیشترین مقدار ($R_k + J_k$) را داراست، بیشترین تعامل را با سایر معیارهای سیستم دارد. مقدار نهایی اثرگذاری هر معیار بر مجموعه معیارهای دیگر سیستم نیز از تفاضل ($R_k - J_k$) حاصل می شود. اگر مقدار ($R_k - J_k$) مثبت باشد، آن معیار متعلق به گروه علت است و چنانچه مقدار ($R_k - J_k$) منفی باشد آن معیار متعلق به گروه معلول است .

۳-۶ داده های تحقیق

داده های این تحقیق شامل موارد زیر است:

- فهرست محصولات (آیتم ها) موجود در فروشگاه
- فاکتورهای فروش جزئی در فروشگاه

در ادامه، شیوه و ابزار جمع آوری داده ها، نوع داده ها و جامعه آماری مورد مطالعه در این پایان نامه بیان شده است.

۳-۶-۱ شیوه و ابزار جمع آوری داده ها

داده ها در قالب فایل های اکسل (گزارش های خروجی از نرم افزار مخصوص فروشگاهی) از فروشگاه اتکا شیراز تهیه شده است.

۳-۶-۲ نوع داده ها و مقیاس آنها

نوع داده ها و مقیاس داده ها به شرح زیر است:

- فهرست محصولات (آیتم ها) موجود در فروشگاه از رکوردهای یکساله فروشگاه بدست آمده است که صرفا از نوع اسمی[۱] هستند.
- فاکتورهای فروش جزئی فروشگاه نیز از رکوردهای یکساله فروشگاه بدست آمده است که هم از نوع اسمی و هم عددی[۲] هستند.

[۱] Nominal
[۲] Numeric

۳-۶-۳ جامعه آماری و نمونه گیری

جامعه آماری در این پژوهش مشتریان فروشگاه اتکا شیراز در یک دوره یکساله می باشند. با توجه به ماهیت داده کاوی که در حجم داده های بالا نتایج بهتری را ارائه می دهد و به منظور کسب نتایج با اعتبار بالاتر، تمامی داده های موجود در مدل های مورد استفاده قرار گرفته اند و نمونه گیری در خصوص این جامعه آماری صورت نگرفته است.

۳-۷ ساختار اجرایی تحقیق

ماهیت این تحقیق داده محور است و پایه اصلی تحقیق حاضر بر کشف دانش از پایگاه داده های اتکا نهاده شده است.

۳-۷-۱ درک مساله کسب و کار

در این مرحله گام های زیر دنبال شده است که جزئیات آنها در ادامه بیان می شود:

- تعیین اهداف کسب وکار
- ارزیابی موقعیت
- تعییم اهداف داده کاوی

در این فاز ابتدا اهداف کسب وکار تعیین و تبیین گردیده که ارائه توصیه های مناسب خرید به خریداران اصلی ترین آنها محسوب می شود که برای این کار از معیار شباهت محصولات (آیتم ها) به یکدیگر استفاده می شود. سپس موقعیت های کسب و کار در فروشگاه زنجیره ای بررسی شده و اتکا به عنوان مورد کاوی انتخاب شد. این انتخاب به دلیل دارا بودن زیر ساخت های قوی تر و به روزتر این فروشگاه ها نسبت به سایر فروشگاه های زنجیره ای در داخل ایران است.

در گام بعدی یعنی تعیین اهداف داده کاوی مشخص شده که تحقیق و فرآیند داده کاوی از نوع یافتن و شناسایی الگوها مرتبط[۱] می باشد، از این رو هدف ایت تحقیق کشف روابط بین آیتم ها از نوع یادگیری بدون نظارت[۲] می باشد.

۳-۷-۲ درک داده ها

در این مرحله گام های زیر دنبال شده است که جزئیات آنها در ادامه آورده شده است:

- جمع آوری داده های اولیه
- توصیف داده ها
- تصدیق کیفیت داده ها و شناسایی داده های هدف

در مراحل قبل یک تصویر کلی از وضعیت کسب و کار و اهداف مورد نظر از لحاظ کسب و کار و داده کاوی برای تحقیق حاصل گردید. در این مرحله با توجه به درک صورت گرفته از این اهداف و وضعیت موجود، لازم است تا درک صحیحی از وضعیت داده ها، نحوه جمع آوری و اطمینان از کیفیت داده های مورد بررسی حاصل گردد.

در مجموع، مهمترین اقداماتی که در این بخش به آنها پرداخته شده تعیین داده های مورد نیاز و نحوه به آن داده ها می باشد.

داده ها در قالب گزارش های اکسل آماده سازی شد و تحویل گرفته شد.این فایل ها دارای ساختار نامناسبی بوده که برای استفاده بایستی آماده می شدند.فایل های متعددی دریافت شد که یک فایل آن مربوط به محصولات به همراه کد و نام (شرح) آنهاست و باقی فایل ها فاکتورهای فروش روزانه است. این

[۱] Associated Rule Mining
[۲] Un-Supervised

فایل ها دارای فیلدهای متفاوتی است که از بین آنها شماره فاکتور و زمان خریداری شدن محصول مورد نیاز بوده که پس از جداسازی به پایگاه داده ای تحقیق منتقل شده است.

۳-۷-۳ آماده سازی داده ها

در این مرحله گام های زیر دنبال شده است که جزئیات آنها در ادامه آورده شده است:

جمع آوری و انتخاب داده ها

یکپارچه سازی داده ها

پاکسازی داده ها (کاهش داده ها)

شکل دادن و ساخت داده ها (تغییر داده ها)

قالب بندی داده ها (برای مدلسازی در نرم افزارها و الگوریتم ها)

۳-۸ جمع بندی

در این فصل متدولوژی های تحقیق صورت گرفته تشریح گردید. روش انجام تحقیق بر مبنای استفاده از یک مدل استاندارد داده کاوی می باشد که مراحل اجرایی تحقیق با الگوگیری از این مدل بصورت گام به گام مورد بررسی قرار گرفت. همچنین کلیاتی در خصوص داده های تحقیق بیان گردیدند. در نهایت اینکه این فصل تمامی مراحلی را که بر روی داده های تحقیق در فصل بعد صورت خواهد گرفت را پوشش داده است.

فصل چهارم

تجزیه و تحلیل داده ها

۴-۱ مقدمه

در این فصل ساختار اجرایی و پیاده سازی روش تحقیق تشریح می گردد. به این منظور، مراحل اجرایی روش تحقیق که به صورت یک مدل فرآیندی مطرح گردید، به صورت عملیات تجزیه و تحلیل که بر روی مجموعه داده های واقعی انجام شده است، بیان می گردند. از آنجا که لازم است تا یک شناخت مناسب از داده های تحقیق و چگونگی آماده سازی آنها قبل از اعمال مدل ها و الگوریتم ها مختلف کسب گردد، در این بخش هم به ویژگی های داده ها و هم به مراحل آماده سازی داده ها پرداخته شده و سپس به اعمال داده ها در مدل ها و بررسی نتایج حاصل از آن خواهیم پرداخت.

۴-۲ آماده سازی داده ها

در این قسمت گام های انجام شده برای آماده سازی داده ها برای استفاده در مدل ها و الگوریتم ها به صورت کامل بین شده است. اولین گام در این بخش شناسایی داده های خام و تبدیل آنها به داده های تمیز و مورد استفاده در سیستم است.

۴-۲-۱ داده های خام اولیه

همانطور که در فصل قبل نیز اشاره شد، داده های دریافتی از فروشگاه اتکا به صورت فایل های اکسل[۱] بوده که در قالب گزارش های سیستمی ارائه شده است. این فایل ها حاوی اطلاعات فاکتور ها (سبد های خرید مشتریان) است. این فایل ها دارای فیلد های متفاوتی از قبیل شرح کالاها، بارکدهای

[۱] Excel

کالاها، زمان خرید هر محصول، قیمت هر محصول، مالیات مخصوص به هر محصول، پرداخت نقدی، میزان تخفیف و ... است.

این فایل ها فیلد های متعددی دارند که در این تحقیق تنها از فیلد های خاصی استفاده می شود که در ادامه بیان خواهند شد.

۴-۲-۲ حذف فیلد های بدون استفاده

همانطور که بیان شد داده های خام دریافت شده از فروشگاه اتکا دارای فیلدهای متعددی می باشند که در کل مراحل تجزیه و تحلیل داده ها تنها از فیلد های خاصی جهت بدست آوردن اطلاعات استفاده می شود، پس لازم است که فیلدهای غیر ضروری حذف شود تا بدین گونه هم از حجم اطلاعات بسیار بالایی که بدون استفاده است کاسته شود و هم مراحل اجرای الگوریتم ها بسیار کوتاه تر می شود همه فیلدها غیر از فیلد های مورد استفاده حدف می شوند.

با یک بررسی اجمالی از بین فیلدهای موجود ، تنها بارکد اجناس که برای هر جنس منحصر به فرد می باشد و زمان خریداری آن اجناس فیلدهایی هستند که دقت آن تا بیان ثانیه می باشد حفظ و باقی فیلدها حذف می شوند.

۴-۲-۳ حذف آیتم های بی ارزش و کم ارزش

بعضی از اجناس در فروشگاه ها وجود دارند که فروش بسیار پایینی نسبت به باقی اجناس دارند و از نظر مشتریان دارای اهمیت بسیار پایینی می باشند. با کد نویسی در محیط نرم افزار MATLAB و تعیین این اجناس، آنها را از لیست اقلام مورد بررسی حذف می کنیم همچنین بعد از بررسی داده ها و تشکیل سبد های خرید مشتریان از روی داده های خرید آنها، به منظور تجزیه و تحلیل سریعتر و راحتتر داده ها

باید تعدادی از سبدهای خرید مشتریان را از مجموعه سبد ها حذف کرد. آن دسته از سبدهایی که دارای تعداد اقلام بسیار کمی می باشند مثلا دارای یک یا دو قلم جنس می باشند چون کمک چندانی به تعیین قوانین پیوند نمی کنند به منظور تسهیل در مراحل تجزیه و تحلیل سبدها و کم اهمیت بودن آنها حذف می شوند.

اجرای این مرحله و کاهش حجم داده ها در زمان اجرای الگوریتم ها بسیار پر اهمیت است و یکی از مهمترین و کلیدی ترین مراحل آماده سازی داده ها تعیین اقلام کم اهمیت و حذف آنها به منظور بهینه کردن پاسخ است.

۳-۴ تجزیه و تحلیل داده ها

بعد از آماده سازی داده ها که شامل مراحل دریافت داده های خام، حذف فیلدهای بدون استفاده و همچنین حذف آیتم های کم ارزش و بی ارزش از میان داده ها و تمیز کردن آنها است در این مرحله با استفاده از الگوریتم های ذکر شده در مراحل قبل و نرم افزارهای مربوط به آنها و کد نویسی در محیط نرم افزار MATLAB به تجزیه و تحلیل آنها خواهیم پرداخت.

۱-۳-۴ استفاده از نرم افزار RapidMiner

بعد از مراحل آماده سازی داده ها و تمیز کردن آنها، باید آماده ورود به نرم افزار RapidMiner شوند. این نرم افزار یکی از قویترین نرم افزار های داده کاوی است که توانایی بسیار زیادی در تجزیه و تحلیل داده ها دارد. شکل داده های ورودی این نرم افزار یک ماتریس است که سطرهای آن سبدهای هر مشتری می باشد. این سبد ها با توجه به شکل ورودی فایل های Excel دریافت شده از فروشگاه و با

استفاده از نرم افزار Matlab تشکیل شده اند. با کد نویسی در محیط نرم افزار Matlab و خواندن این فایل های Excel ، خروجی الگوریتم نوشته شده در Matlab ، ماتریس ورودی نرم افزار RapidMiner می باشد.این ماتریس تشکیل شده دارای تعداد بسیار زیادی سطر و ستون است که سطرهای این ماتریس سبدها است و ستون های آن اجناس موجود در فروشگاه می باشند. در این ماتریس در هر سطر تعدادی صفر و یک وجود دارد، یک های موجود در هر سطر نشان دهنده این است که آن سبد، جنس متناظر با آن ستونی که دارای یک می باشد را دارا می باشد و صفرها نشان دهنده این است که آن سبد اجناس متناظر با آن ستون هایی که صفر هستند را دارا نمی باشد.

بعد از اجرای نرم افزار خروجی آن شامل آن دسته از قوانینی است که ما به دنبال آنها هستیم. هر جنس در فروشگاه یک بارکد مخصوص به خود دارد و تجزیه و تحلیل با توجه به آن بارکد صورت می پذیرد که نمونه از خروجی نرم افزار به صورت زیر است.برای اطمینان از صحت نتایج دو پارامتر را به نام های Support و Confidence تعریف می کنیم که میزان عددی آنها درجه اعتبار هر قانون را نمایش خواهد داد که در فصل بعدی در مرحله اعتبار سنجی نتایج آنها را تعریف و به بحث در مورد آنها خواهیم پرداخت.

```
[6260136920021, 6260074050026] --> [6261715202347] (confidence: 0.974)
[6260281414093] --> [6260608980317, 6261715202347] (confidence: 0.974)
[6260608980317, 6261715202347, 6260281411061] --> [6260136920021] (confidence: 0.974)
[6260608980317, 6260281411061] --> [6260136920021] (confidence: 0.974)
[6260281411061, 2001726000037, 2001726000020] --> [6260608980317, 6261715202347] (confidence: 0.974)
[6260608980317, 6261715202347] --> [6260136920021] (confidence: 0.974)
[6260608980317] --> [6260136920021] (confidence: 0.974)
[6260281411061, 2001804000010] --> [6261715202347] (confidence: 0.974)
[6261715202347, 6260281411061] --> [6260136920021] (confidence: 0.974)
[6260281411061] --> [6260136920021] (confidence: 0.974)
[6261715202347] --> [6260136920021] (confidence: 0.974)
[6260136920021, 6260281411061, 2001804000010] --> [6261715202347] (confidence: 0.974)
[6260136920021, 2001726000037, 2001726000020] --> [6260608980317, 6261715202347] (confidence: 0.974)
[6260136920021] --> [6260608980317, 6261715202347] (confidence: 0.974)
[2001804000010] --> [6261715202347] (confidence: 0.974)
[6260136920021, 6260074023181] --> [6260608980317, 6261715202347] (confidence: 0.974)
[6260136920021, 2001804000010] --> [6261715202347] (confidence: 0.974)
```

شکل ۴-۱ نمونه ای از قوانین تولید شده توسط نرم افزار RapiMiner

به عنوان مثال قانون اول تولید شده که در شکل بالا با رنگ زرد مشخص شده است به این معناست که اجناسی که بارکدهای مربوط به آنها ۶۲۶۰۱۳۶۹۲۰۰۲۱، ۶۲۶۰۰۷۴۰۵۰۰۲۶ و ۶۲۶۱۷۱۵۲۰۲۳۴۷ با درجه Confidence ۹۷,۴ درصد که مفهوم این عدد را در فصل بعدی به تفصیل ارائه خواهیم داد با هم خریداری می شوند و در یک سبد قرار می گیرند.

۴-۳-۲ استفاده از الگوریتم TOPSIS و نرم افزار MATLAB

یکی دیگر از الگوریتم هایی که با استفاده از آن به تجزیه و تحلیل داده ها پرداختیم، الگوریتم تاپسیس[۱] است که پیشتر مراحل اجرای آن را شرح دادیم. این الگوریتم دارای این خاصیت است که با توجه به نوع تعریف ما از ایده آل مثبت و ایده آل منفی ما گزینه ای را از بین گزینه های موجود (که در

[۱] TOPSIS

اینجا سبد های مشتریان هستند) به ما پیشنهاد می دهد که بیشترین فاصله را از ایده آل منفی و کمترین فاصله را از ایده آل مثبت دارد. بعد از تشکیل ماتریس که سطر های آن سبدهای مشتریان و ستون های آنها اجناس می باشند، اکنون به منظور انتخاب یا اولویت بندی سطر ها یا همان سبد های مشتریان دو اصطلاح را به صورت زیر تعریف می کنیم.

۴-۳-۲-۱ آیتم های اساسی Basic Items و آیتم های اضافه شدهAdded Items

اجناس فروشگاه را به دو دسته تقسیم می کنیم، آیتم های اساسی و آیتم های اضافی. برای این منظور اجناسی که بیشترین فروش و کمترین فروش را دارند مشخص می کنیم، برای اینکه مشخص کنیم کدام یک از اجناس اساسی و یا اضافی هستند میزان فروش آنها را با میزان فروش پر فروش ترین جنس مقایسه می کنیم اگر آن جنس خاص میزان فروشی از ۷۰ درصد میزان جنس پرفروش به بالا فروش داشته باشد آن را جزو اجناس اساسی و اگر کمتر باشد آن را جزو اجناس اضافی قرار می دهیم.

بعد از تقسیم اجناس به دو دسته که اساسی و اضافی می باشند،در مراحل حل الگوریتم اجناس اساسی را ایده آل منفی و اجناس اضافی را ایده آل مثبت فرض می کنیم و دلیل آن با توجه به تعاریف الگوریتم پر واضح است که هدف نزدیکتر شدن به ایده آل مثبت و دور شدن از ایده آل منفی است. چون اجناس پر فروش تحت هر شرایطی فروش خود را خواهند داشت و اکثرا از کالاهایی هستند که نیازهای اساسی مردم هستند به خودی خود دارای فروش کافی هستند و تلاش ما با رویکرد این الگوریتم این است که هر چه بیشتر به سوی فروش کالاهای اضافی بریم و ان را ایده آل مثبت خود قرار می دهیم. در نهایت با اجرای الگوریتم با استفاده از کد نویسی در محیط نرم افزار MATLAB با این رویکرد سبد های مناسب را بدست آوردیم و قسمتی از خروجی حاصل به شکل زیر است و آن سطر یا سبدی که دارای C* بیشتر است سبد مناسبتر و اولویت بالاتر ما است.

2.5E+11	5E+11	6.34E+11	7.62E+11	7.62E+11	1E+12	1E+12	1E+12	1E+12	2E+12	2E+12	2E+12	2E+12	C*
1	8	1	3	53	9	5	3	2	85	77	163	47	0.59919
107	1	1	1	25	5	2	5	2	86	92	165	61	0.589619
41	0	1	1	7	2	0	2	1	27	30	53	10	0.307708
21	2	0	1	15	1	3	1	1	19	22	55	13	0.280511
22	2	0	0	1	5	0	2	1	21	20	38	15	0.264834
39	3	0	1	10	1	0	2	1	13	17	35	11	0.263512
18	1	0	0	8	3	1	0	1	11	13	30	11	0.221761
19	0	0	0	11	2	0	1	1	8	14	23	10	0.185773
17	1	0	0	9	1	2	1	1	11	11	33	7	0.166019
21	0	0	0	7	0	0	3	0	1	8	19	7	0.165261
13	0	0	0	9	0	0	2	0	11	1	17	7	0.159707
14	1	0	1	5	3	0	1	1	21	18	1	13	0.157763
8	1	0	0	1	1	0	1	0	9	4	17	1	0.151814

شکل ۴-۲ قسمتی از اولویت با استفاده از الگوریتم TOPSIS

۴-۳-۳ استفاده از الگوریتم DEMATEL و نرم افزار MATLAB

یکی دیگر از الگوریتم هایی که با استفاده از آن به تجزیه و تحلیل داده ها پرداختیم، الگوریتم DEMATEL است که پیشتر مراحل اجرای آن را شرح دادیم. این الگوریتم دارای این خاصیت است که با توجه به رابطه داخلی خرید اجناس با یکدیگر و مراحل اجرای الگوریتم به تعیین تاثیر گذارترین و تاثیر پذیرترین گزینه از بین گزینه های موجود (که در اینجا سبد های مشتریان هستند) می پردازد و به ترتیب اولویت آنها را به ما پیشنهاد می دهد. بعد از تشکیل ماتریس که سطر های آن سبدهای مشتریان و ستون های آنها اجناس می باشند، اکنون به منظور انتخاب مناسب ترین سبدها آنها را به الگوریتم دیمتل[۱] که در محیط MATLAB کد نویسی شده است وارد می کنیم و نتایج را با الگوریتم های دیگر مقایسه خواهیم کرد که در ادامه خواهد آمد.

[۱] DEMATEL

۴-۴ اعتبار سنجی مدل ها

برای تعیین اعتبار مدل ها همانطور که قبلا هم گفته شد دو اصطلاح پشتیبان[۱] و اطمینان[۲] را به صورت زیر تعریف می کنیم.

۴-۴-۱ Support (پوشش):

اگر سبد I شامل تمامی اقلام باشد، مجموعه آیتم‌های X و Y را در نظر بگیرید. پوشش قانون $X \Longrightarrow Y$ به‌صورت رابطه‌ی زیر تعریف می‌شود.

$$Support(X \Longrightarrow Y) = S(X \Longrightarrow Y) = \frac{\rho(X \,\&\, Y)}{m}$$

که در آن منظور از $\rho(X \,\&\, Y)$ ، تعداد مشتریانی است که مجموعه اقلام X و Y را با هم خریداری کرده‌اند و m تعداد کل مشتریان می‌باشد. فاکتور پوشش بیانگر نسبت تعداد مشتریانی که مجموعه اقلام X و Y را با هم خریداری کرده‌اند به کل مشتریان است.

۴-۴-۲ Confidence (درجه اطمینان):

اگر سبد I شامل تمامی اقلام باشد، مجموعه آیتم‌های X و Y را در نظر بگیرید. درجه اطمینان قانون $X \Longrightarrow Y$ به‌صورت رابطه‌ی زیر تعریف می‌شود.

$$Confidence(X \Longrightarrow Y) = \; C(X \Longrightarrow Y) \; = \frac{\rho(X \,\&\, Y)}{\rho(X)}$$

فاکتور درجه اطمینان بیانگر نسبت تعداد مشتریانی است که مجموعه اقلام X و Y را با هم خریداری کرده‌اند به تعداد مشتریانی که اقلام X را خریداری کرده‌اند.

---

[۱] Support
[۲] Confidence

با توجه به این دو آیتم که تعریف آنها در بالا آورده شده است و محاسبه آنها که در زیر نمونه ای از آنها آورده شده است بالا درجه بالایی از پوشش و اطمینان به سبدهای بهیته دست پیدا کرده ایم که نشان می دهند چه اجناسی با چه درجه ای از اطمینان و پوشش با یکدیگر خریداری شده اند.

در ادامه مقدار Support و Confidence که در زیر نمونه ای از آنها آورده شده است را محاسبه می کنیم.

Premises	Conclusion	Support	Confidence
6261715202347, 6260136920021, 6260281411061, 20017260	6260608980317	0.268	0.993
6260136920021, 6260074050026	6260608980317	0.119	0.993
6260136920021, 6260281411061, 2001726000020	6260608980317	0.273	0.993
6261715202347, 6260074050026	6260608980317	0.120	0.993
6261715202347, 6260281411061, 2001726000020	6260608980317	0.276	0.993
6261715202347, 6260136920021, 2001726000020	6260608980317	0.277	0.993
6260281411061, 2001726000020	6260608980317	0.281	0.993
6260136920021, 2001726000020	6260608980317	0.282	0.993
6260074050026	6260608980317	0.124	0.993
6261715202347, 2001726000020	6260608980317	0.285	0.993
2001726000020	6260608980317	0.290	0.993
6260136920021, 6260281411061	6260608980317	0.933	0.994
6261715202347, 6260136920021, 6260281411061	6260608980317	0.914	0.994
6260136920021	6260608980317	0.968	0.994
6261715202347, 6260136920021	6260608980317	0.949	0.994
6260281411061	6260608980317	0.958	0.994
6261715202347, 6260281411061	6260608980317	0.939	0.994
6261715202347, 6260074023136	6260608980317	0.101	0.994
6261715202347	6260608980317	0.974	0.994
6260074023136	6260608980317	0.104	0.994

شکل ۴-۳ نمونه ای از Support و Confidence قوانین تولید شده

۴-۵ مقایسه مدل ها

با انجام الگوریتم ها و تعیین سبدهای بهینه و اولویت بندی آنها و تعیین میزان اعتبار هر قانون بدست آمده و همچنین با مقایسه نتایج بدست آمده از مدل های موجود شاهد این هستیم در که بالای ۹۰ ٪ شرایط سبدهای حاصل شده از تمامی الگوریتم ها همپوشانی دارند و تقریبا همه الگوریتم ها سبدهای یکسان و مشابه ای را پیشنهاد می کنند که این نشان از دقت بالای این تحقیق دارد که با اکثر الگوریتم ها با درجه اطمینان و پوشش بالا به نتایج بسیار مشابه رسیده ایم.

۶-۴ جمع بندی

در این فصل داده های تحقیق توصیف و تحلیل شدند و مراحل انجام تحقیق بطور عملی تشریح گردید. اعمال مدل ها با پارامتر های مختلف صورت پذیرفت و نتایج حاصله مورد بررسی قرار گرفت. شاخص های درستی و دقت مدل ها در خصوص داده های تحقیق مورد مطالعه قرار گرفتند و در فصل بعد نتایج حاصل از تحقیق را ارائه می دهیم.

فصل پنجم

نتیجه گیری و ارائه پیشنهادات

۵-۱ مقدمه

در این فصل مطالب در دو بخش ارائه شده است. ابتدا نتایج و یافته های حاصل از تحقیق بیان می شود و در بخش دوم با توجه به مطالعات انجام شده که شامل مطالعه ادبیات موضوع، انجام روش تحقیق و کسب نتایج می باشد، پیشنهادهایی برای تحقیقات آتی و بهره وران از این تحقیق ارائه می گردد.

۵-۲ نتایج و یافته های تحقیق

در این بخش نتایج تحقیق بصورت پاسخ به سوال تحقیق و نیز نتایج مستقیم حاصل از اجرای تحقیق بیان می گردد.

۵-۲-۱ پاسخ به سوال های تحقیق

همانطور که از بررسی نتایج حاصل از اعمال مدل های ارائه شده در داده های تحقیق بر می آید می توان بر اساس سابقه مشتریان و کالاهایی که در سبد خرید خود دارند اقدام به ارائه توصیه های مفید نمود.همانطور که از بخش اعتبار سنجی مدل ها بر می آید حتی این مدل ها قادر هستند با درجه اطمینان و پوشش بسیار بالا کالاهای مورد نیاز مشتریان را توصیه نمایند. بدیهی است که این دقت نسبت به سلایق بی شمار مشتریان و نیازهای متعدد آنها و نیز تعداد مشتریان کم یا زیاد باشد. با توجه به بررسی مدل ها مشخص شده است که در صورتیکه سبد خرید مشتری حاوی کالاهای مرتبط با یکدیگر باشد مدل های این تحقیق محصولات مناسب را توصیه می کنند. به عنوان مثال در اکثر خریدها شیر و ماست و پنیر با یکدیگر خریداری می شود، حال آنکه یک مشتری در سبد خرید خود شیر و سبزی خرد شده را قرار دهد بدیهی است که اگر چند درصد مشتریان این دو محصول را در سبد خرید خود قرار دهند، این ترکیب توسط مدل ها کشف می شود و جهت توصیه مورد استفاده قرار خواهد گرفت.

در واقع فرآیند خرید و به نوعی انتخاب محصولات و قرار دادن آنها در سبد خرید توسط مشتریان بسیار پیچیده و غیر قابل پیش بینی است، زیرا که پارامترهای متعددی هستند که در این امر دخالت دارند، قیمت شناور محصولات (که با عوامل مختلف تغییر می کند)، ورود و خروج رقبای جدید محصولات، نحوه نمایش[1] محصول جهت خرید در محل، میزان تبلیغات ارائه شده و خرید، متوسط زمان های خرید مشتری (روزانه، هفتگی، ماهیانه و ...)، عوامل فرهنگی، شاخص های خانوادگی (نظیر تعداد فرزند، سن فرزندان، پراکندگی سن و نوع آنهاو ...) و ... از جمله عواملی هستند که منجر به رفتارهای متفاوت خرید می شوند. حتی یک مشتری خاص با نیازهای ثابت و شرایط یکسان در خریدهای متوالی ممکن است اقدام به خرید محصولات مشابه قبل ننماید. بدین صورت است که عموما سیستم های توصیه گر خرید نظیر Amazon و ... با بهترین مدل ها و پیچیده ترین الگوریتم ها قادر هستند فقط درصد کوچکی از خریدها را توصیه نمایند.

از نگاهی دیگر فرآیند توصیه، پیشنهاد خرید محصولی است که در سبد خرید وجود نداشته و در اغلب موارد حتی اگر کاملا هم دقیق نباشد (به اندازه کافی دقت داشته باشد) ممکن است مورد توجه مشتری قرار گیرد و محصول (محصولات) توصیه شده را نیز در کنار محصولات سبد خرید خود قرار دهند، زیرا که در برخی موارد مشتریان برخی محصولات (کالاهای) مورد نیاز خود را فراموش می نمایند که در این صورت این سیستم به آنها در این زمینه کمک خواهد نمود،خصوصا این امر در فروشگاه های زنجیره ای که عموما خریدهای سنگین از آنها انجام می پذیرد و مشتریان به قصد خرید ماهیانه به آنجا می روند بیشتر صدق می کند.

در این تحقیق با استفاده از تجزیه و تحلیل سبد های خرید مشتریان و درک رفتار آنها با توجه به محصولاتی که خریداری کرده اند و همچنین با توجه به تعاریف میزان پوشش و اطمینان که پیشتر آورده

[1] Present

شده اند مشخص کردیم که کدام اجناس با چه درجه ای از پوشش و اطمینان با یکدیگر خریداری شده اند و به هم وابسته اند و با داشتن این علم که هدف از انجام این تحقیق نیز نائل شدن به همین علم است می توان تصمیمات فروش و بازاریابی متفاوتی را با توجه به سیاست های فروش و نظر مدیران مسئول گرفت. به عنوان مثال فرض کنید با تجزیه و تحلیل سبد خرید مشتریان به این علم رسیده ایم که دو قلم کالا با درجه اطمینان و پوشش بالا به یکدیگر وابسته اند، یکی از سیاست هایی که می تواند مورد توجه مدیران باشد این است که چون این دو قلم کالا با توجه به شدت درجه پوشش و اطمینان رابطه قوی با یکدیگر دارند پس این دو قلم کالا بسیار به هم وابسته اند و این بدین معنی است که مشتریان اکثرا این دو قلم کالا را با هم خریداری می کنند و خرید آنها با یکدیگر را فراموش نمی کنند، با داشتن این علم ما می توانیم یکی از این دو قلم کالا را در ابتدای فروشگاه و دیگری را در انتهای فروشگاه قرار دهیم و با علم به میزان شدت وابستگی آنها با یکدیگر از این مطمئن هستیم که مشتریان خرید آنها با یکدیگر را فراموش نمی کنند و از این علم به عنوان یک مزیت برای فروش بیشتر استفاده می کنیم چون مشتری را به صورت نا خودآگاه مجبور می کنیم تا تمام طول فروشگاه را برای پیدا کردن اجناس وابسته طی کند و در حین مسیر اجناس دیگر موجود در فروشگاه را مشاهده می کند و اگر قلم کالای دیگری را فراموش کرده باشد آن را در مسیر مشاهده می کند و حتی اگر کالایی را هم فراموش نکرده باشد تصویر کالاهای بین مسیر دو قلم اصلی را به طور ناخودآگاه در ذهن نگاه می دارد و هر زمان که به آن کالاها نیاز پیدا کند اولین گزینه برای خرید آن کالاها فروشگاهی است که یکبار در آن کالای مورد نیاز خود را مشاهده کرده است و این باعث فروش بیشتر فروشگاه می شود.

همچنین اگر با تجزیه و تحلیل سبد خرید مشتریان به این علم دست یافتیم که دو قلم کالا با درجه پوشش و اطمینان کمی به هم وابسته هستند به این نتیجه می رسیم شاید اگر این اجناس در کنار

هم قرار نداشته باشند، احتمال اینکه مشتریان خرید آنها با یکدیگر را فراموش کنند می باشد پس آنها را در مجاورت هم قرار می دهیم تا فروش آنها و ئر نتیجه فروش فروشگاه را بالا ببریم.

در نهایت پاسخ قطعی به سوال تحقیق منوط است به همبستگی محصولات موجود در سبد خرید مشتریان مورد نظر و شباهت نسبی آنها به سایر خریدها. بدیهی است که در صورتیکه این همبستگی در حد مطلوب باشد و سیستم توصیه گر بتواند توصیه های مفیدی ارائه دهد منجر به خرید بیشتر مشتریان و افزایش حجم سبد خرید آنها و به تبع آن فروش بیشتر محصولات خواهد شد.

۵-۲-۲ نتایج مستقیم حاصل از اجرای تحقیق

- در صورتیکه رفتار خرید مشتری مشابه با بخشی از سایر مشتریان باشد، سیستم می تواند با استفاده از مدل های خود اقدام به ارائه توصیه های مناسب و موثر بنماید. بدیهی است هرچقدر این نسبت بزرگتر باشد یعنی رفتار خرید مشتری مورد نظر به بخش بزرگتری از مشتریان شبیه باشد (رفتار خرید عمومی تری داشته باشد) درصد دقت توصیه های ارائه شده بیشتر خواهد بود.
- در مدل های ارائه شده پارامترهای تعیین دقت نقش اساسی ایفا می کنند و مشاهده می شود که با بزرگ شدن مقدار این متغییر ها از تعداد کالاهایی که سیستم قادر خواهد بود بر روی آن توصیه نماید کاسته می شود و بر دقت مدل ها افزوده می گردد و با کوچک شدن این معیارها نیز بر تعداد کالاهایی که سیستم قادر است بر روی آنها توصیه انجام دهد افزوده شده ولی دقت نسبی کلیه مدل ها کمتر می شود.

۳-۵ پیشنهادات

- اعمال مدل های دیگر از تکنیک های تصمیم گیری چند معیاره و مقایسه آنها با یکدیگر
- اجرای الگوریتم های موجود برای داده های خام مربوط به مکان ها و جاهای مختلف و تعیین اینکه با توجه به شرایط آن مکان ها و داده ها کدام الگوریتم ها مناسب تر می باشند.
- غنی سازی داده ها با افزودن طبقه بندی کالا ها و امکان ارایه پیشنهاد برای محصولات آینده.

منابع و مراجع

Andrássoyá, E., & Paralič, J. (۱۹۹۹). Knowledge discovery in databases - a comparison of different views. ۱۰th International Conference on Information and Intelligent Systems.

Adomavicius, G., & Tuzhilin, A. (۲۰۰۵). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. ۱۷.

About : SIGKDD Explorations. (n.d.). (ACM) Retrieved from http://www.kdd.org/explorations/about.php

Bay Vo, Bac Le, Thang N. Nguyen: Mining Frequent Item sets from Multidimensional Databases. ACIIDS ۲۰۱۱:۱۷۷-۱۸۶

Benoît, G. (۲۰۰۲). Annual Review of Information Science and Technology. Silver Spring.

David Hand, Heikki Mannila, Padhraic Smyth. Principles of Data Mining, MIT Press, ۲۰۰۱.

De Almeida, P. (۲۰۰۸). Master Thesis References stylesheet for Word ۲۰۰۷. Retrieved March ۱۲, ۲۰۰۹, from Software Engineering group: http://diuf.unifr.ch/softeng/student projects/templates/soften_en_msword_bib.xsl

Fayyad, U., G., P.-S., & Uthurusamy, R. (٢٠٠٣). Data mining: The next ١٠ years. SIGKDD Explorations.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (١٩٩٦). Retrieved from From Data Mining to Knowledge Discovery in Databases: http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-١٩٩٦-Fayyad.pdf.

Guazzelli, A., Zeller, M., Lin, W.-C., & Williams, G. (٢٠٠٩, May). PMML: An Open Standard for Sharing Models. ١.

Han, J., & Kamber, M. (٢٠٠١). Data mining: concepts and techniques (Morgan-Kaufman Series of Data Management Systems). San Diego: Academic Press.

International Conferences on Knowledge Discovery and Data Mining. (n.d.). New York: ACM.

Julio Ponce, Adem Karahoca: Data Mining and Knowledge Discovery in Real Life Applications: In-The, Vienna: ٢٠٠٩

Jayanthi Ranjan, Vishal Bhatnagar: Application of data mining techniques in the financial sector for profitable customer relationship management. IJICT ٢(٤):٣٤٢-٣٥٤ (٢٠١٠).

Kantardzic, M. (٢٠٠٣). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons.

Lyman, P., & Hal R., V. (٢٠٠٣). How Much Information. Retrieved from http://www.sims.berkeley.edu/how-much-info-٢٠٠٣.

Longbing Cao, Jiang Zhong, Yong Feng: Advanced Data Mining and Applications - ٦th International Conference, ADMA ٢٠١٠, Chongqing, China, November ١٩-٢١, ٢٠١٠, Proceedings, Part II ADMA ٢٠١٠

Michael J. Shaw, Chandrasekar Subramaniam, Gek Woo Tan, Michael Welge: Knowledge management and data mining for marketing. Decision Support Systems (DSS) ٣١(١):١٢٧ ١٣٧ (٢٠٠١).

Manolis Maragoudakis, Dimitrios N. Serpanos: Towards Stock Market Data Mining Using Enriched Random Forests from Textual Resources and Technical Indicators. AIAI ٢٠١٠:٢٧٨-٢٨٦] Maher Ragheb

Michael J. Shaw, Chandrasekar Subramaniam, Gek Woo Tan, Michael Welge: Knowledge management and data mining for marketing. Decision Support Systems (DSS) ٣١(١):١٢٧-١٣٧ (٢٠٠١)

Nhien-An Le-Khac, Sammer Markos, M. Tahar Kechadi: A Data Mining-Based Solution for Detecting Suspicious Money Laundering Cases in an Investment Bank. DBKDA ٢٠١٠:٢٣٥- ٢٤٠

Nhien-An Le-Khac, Sammer Markos, M. Tahar Kechadi: A Data Mining-Based Solution for Detecting Suspicious Money Laundering Cases in an Investment Bank. DBKDA ٢٠١٠:٢٣٥-٢٤٠

Parisa Rahnama, Mohammad Mehdi Sepehri, Pejman Shadpour, Babak Teimurpour: a decision support system for selecting type of treatment for kidney stone patients with data mining approach. IDMC (٢٠٠٧).

Peng, Y., Kou, G., Shi, Y., & Chen, Z. (٢٠٠٨). A Descriptive Framework for the Field of Data Mining and Knowledge Discovery. ٧, ٦٣٩ – ٦٨٢.

Pine, B. J. (١٩٩٣). Mass Customization. Boston, Massachusetts: Harvard Business School Press.

Pine, B. J., Peppers, D., & Rogers, M. (١٩٩٥). Do you want to keep your customers forever? Harvard Business School Review.

PMML Project Page. (n.d.). Retrieved from http://sourceforge.net/projects/pmml

Ramezania, M., Bergmanb, L., Thompsonb, R., Burkea, R., & Mobashera, B. (n.d.). Selecting and Applying Recommendation Technology.

Retrieved from IEEE International Conference on Data Mining: (٢٠٠٩). (IEEE) http://www.cs.umbc.edu/ICDM٠٩/

(n.d.). Retrieved from The Data Mining Group (DMG): http://www.dmg.org

(n.d.). Retrieved from International Conference on Data Mining: http://www.dmin-٢٠٠٨.com

Seonyoung Shim, Byungtae Lee: Applying experimental online auctions in marketing research for multi-channel firms. Expert Syst. Appl. (ESWA) ٣٧(٣):٢٥٠٥-٢٥١٦ (٢٠١٠)

Seonyoung Shim, Byungtae Lee: Applying experimental online auctions in marketing research for multi-channel firms. Expert Syst. Appl. (ESWA) ٣٧(٣):٢٥٠٥-٢٥١٦ (٢٠١٠)

Witten, I., & Frank, E. (n.d.). Data mining: practical machine learning tools and techniques(٢nd ed, Morgan-Kaufman Series of Data Management Systems).

Zanasi, A., Brebbia, C., & Ebecken, N. (٢٠٠٥). Preface. In Zanasi, A., Brebbia, C.A., Ebecken, N.F.F.(Eds.). England: WIT Press.

Abstract

Market basket analysis is a generic term for methodologies that study the composition of a basket of products (i.e. a shopping basket) purchased by a household during a single shopping trip. The idea is that market baskets reflect interdependencies between products or purchases made in different product categories, and that these interdependencies can be useful to support retail marketing decisions.

Due to the recent competition in retailer industry, retailers are striving to improve their operations in order to increase the efficiency and profitability. Therefore, this condition forced retail companies to consider more about fundamental issues such as marketing stimuli, products to be displayed, and space to display assorted products.

Retailers collect terabytes of data every day such as transactional data, customer demographics and product sales based on parameters such as seasons and festivals. This data alone cannot enable good decision making for a retailer. It is necessary to discover and understand the underlying patterns involved in the organization's operations from these data. Hence, there is a need present for accurate, timely information to react to changing market conditions, identify new customer segments, improve inventory management, and optimize overall store performance. Recently, a number of advances in data mining and statistics (association rules) offer new opportunities to analyses such data.

In the first phase of this thesis, different algorithms are used in order to analysis and discover the relationship and patterns of product's adjacency. In the next phase, we take advantage of Multi Criteria Decision Making methods (MCDM) like TOPSIS and DEMATEL to rank the patterns. In the third phase of the thesis, by the use of the baskets generated by the algorithm in the first phase and their rank generated by the MCDM methods in the second phase, an approach for shelf allocation is presented. Finally, for the purpose of evaluation, an experimental study using real data was conducted with definition of "*Confidence*" and "*Support*" phrases as benchmark. The result shows *0.994* for "*confidence*" and *0.974* for "*support*" .The algorithm is implemented in C++ language with the help of MATLAB software.

Key Words: Data Mining, Retail Market Basket Analysis, MCDM, Product assortment, Shelf-space management.

Shahrood University of Technology

Department of Industrial Engineering and Management

Retail Market Basket Analysis: Data Mining Approach with Consideration of Shelf-Space Adjacency

Thesis submitted to obtain the degree of

Master of Business Administration

By:

Afshin Mirzaei

Supervisor:

Dr. Reza. Sheikh

November 2012