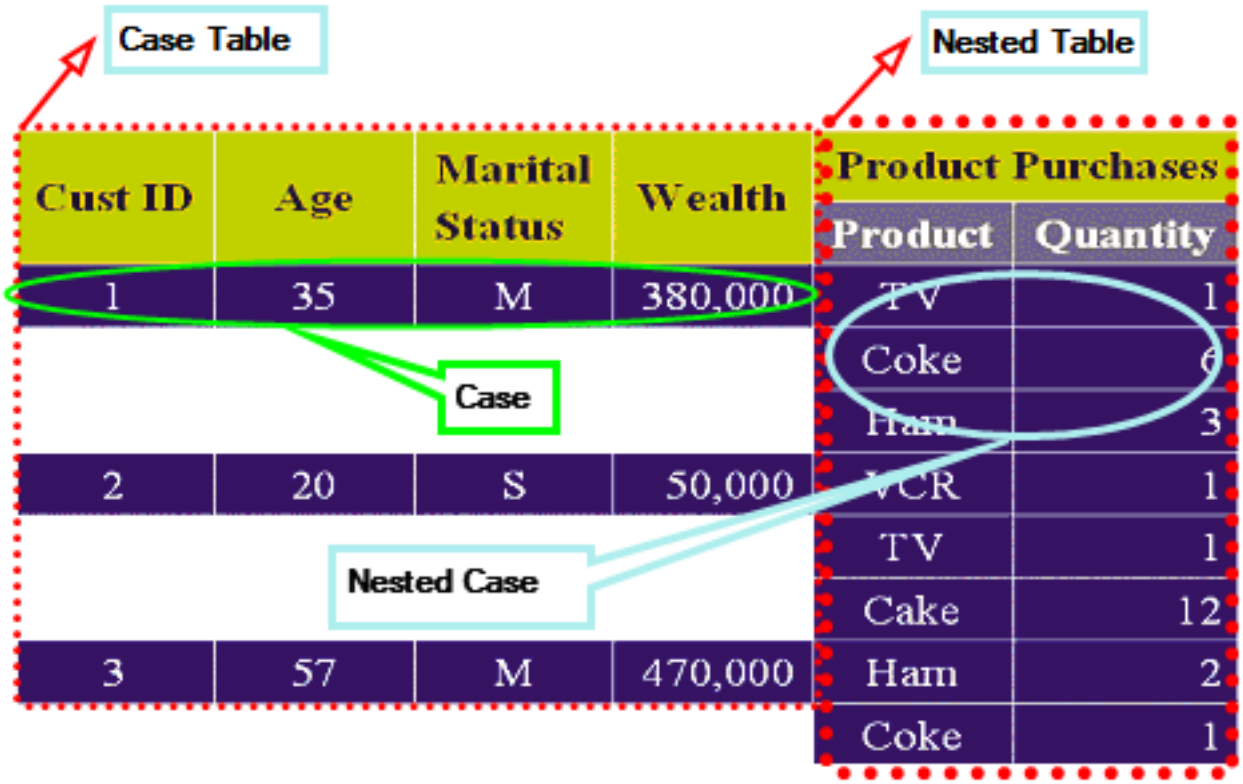
عنوان: **مفاهیم کلیدی مورد استفاده در مدل سازی داده کاوی (بخش دوم)**
نویسنده: محمد رجبی
تاریخ: ۱۳۹۳/۰۸/۱۸ ۱۲:۳۳
آدرس: www.dotnettips.info
گروه‌ها: Analysis Services, data mining, Microsoft SQL Server

مفاهیم کلیدی

Case مهمترین مفهومی است که در تحلیل یک مسئله داده کاوی می‌بایست شناسائی شود و تشخیص اشتباه در شناسائی آن منجر به عدم موفقیت پروژه داده کاوی خواهد شد. Case به معنای یک موجودیت پایه از اطلاعات می‌باشد که عملیات داده کاوی بر روی آن انجام می‌شود و هدف از معرفی آن، معرفی ساختار مسئله به موتور داده کاوی است. هر Case شامل مجموعه ای از ویژگی‌ها (Attributes) می‌باشد؛ مانند سن، جنسیت. ویژگی‌ها می‌توانند دارای یک مجموعه از مقادیر ممکن باشند که به آنها وضعیت یا مقدار (State/Value) می‌گویند؛ مانند جنسیت که دارای دو وضعیت زن یا مرد می‌باشد.
Case می‌تواند ساده باشد؛ برای نمونه زمانیکه قصد دارید «از اطلاعات آماری مشتریان به منظور تحلیل ریسک وام گرفتن» استفاده کنید، بدین ترتیب هر Case شامل اطلاعات یک مشتری و یا ردیفی از داده مشتریان است.

Case می‌تواند کمی پیچیده‌تر باشد؛ برای مثال زمانیکه می‌خواهید «رفتار خرید مشتری را بر اساس تاریخچه خرید مشتری» تحلیل کنید، که در این صورت هر Case شامل یک رکورد از اطلاعات مشتری به همراه لیستی از محصولاتی که خریداری کرده است، می‌باشد. (توجه کنید تعریف رفتار به طور ضمنی، بیانگر عملکرد در طول زمان می‌باشد)

Case مثال فوق نمونه ای از **Nested Case** است، که به اطلاعات Details در ساختار Master/Details اشاره دارد. چنانچه Case ای از نوع Nested باشد، الگوریتم‌ها به Case ای به عنوان ورودی فرمت مجموعه ردیف سلسله مراتبی (Hierarchical Row-set) نیاز دارند.



Case Key مشخصه ای است که یکتا بودن هر Case را مشخص می‌کند و اغلب Primary Key یک جدول رابطه ای است، همچنین ممکن است یک کلید ترکیبی باشد. ذکر این نکته ضروری است که بدانیم Case Key فقط یک شناسه است و شامل هیچ الگویی نمی‌باشد و بدین ترتیب غالباً بوسیله الگوریتم‌های داده کاوی نادیده گرفته می‌شود.

Nested Key مهمترین مشخصه ویژگی از بخش Nested هر Case است و در واقع کلید معنایی تحلیل می‌باشد که شامل اطلاعات مفیدی درباره‌ی الگوهاست. به بیان دیگر ویژگی است که عناصر مختلف موجود در Nested Case را به ازای هر Case تفکیک می‌کند. همچنین در نظر داشته باشید که Nested Key یک شناسه نیست و دارای مفهومی متفاوت با Foreign Key است، بدین ترتیب سایر مشخصه‌های دیگر در بخش Nested؛ جهت توصیف Nested Key بکار می‌روند. برای نمونه چنانچه مدلی برای یادگیری الگوهایی درباره رفتار خرید مشتری داشته باشیم، Nested Key برابر با محصول و میزان خرید است.

به همین ترتیب **Case Table** جدولی است شامل اطلاعات Case و بطور مشابه **Nested Table** جدولی است که شامل اطلاعات مرتبط با قسمت Nested از Case می‌باشد. از اپراتور **Shape** به منظور پیوند میان Case Table و Nested Table استفاده می‌شود.

در خصوص **Attribute** ها (ویژگی ها) از آنجا که هر ویژگی؛ توصیف کننده مسئله داده کاوی از یک منظر خاص می‌باشد، می‌توان اینگونه بیان نمود که هر چه تعداد ویژگی‌ها در یک پروژه بیشتر باشد، توان تحلیل در آن پروژه افزایش می‌یابد. انواع ویژگی‌ها به دو دسته **Discrete** (گسسته) و **Continuous** (پیوسته) تقسیم می‌شوند. برای نمونه ویژگی جنسیت، تحصیلات و ... گسسته و همچنین ویژگی سن، درآمد و ... پیوسته هستند. به مقادیر موجود در یک ویژگی پیوسته **Value** و بطور مشابه به وضعیت‌های موجود در یک ویژگی گسسته **State** گفته می‌شود. ویژگی‌ها در یک الگوریتم از حیث کاربرد (Attribute Usage) به دو دسته **Input** و **Output** تقسیم می‌شوند.

یک الگوریتم از ویژگی‌های ورودی (Input) استفاده می‌کند تا الگویی برای پیش بینی ویژگی‌های خروجی (Output) پیدا کند. همچنین لازم است در نظر داشته باشید که برخی الگوریتم‌ها نظیر Naïve Bayes صرفاً با داده‌های گسسته و بطور مشابه الگوریتم هایی نظیر Logistic Regression تنها با مقادیر پیوسته کار می‌کنند.