Čistota dat aneb Z provozních na analytická

Miloš Uldrich

Pokud se řekne čistota dat, málokdo si představí něco jasně uchopitelného. V tomto příspěvku si popíšeme různé kroky, které se při přípravě dat pro další zpracování používají.

Konsolidace dat

Většina větších společností potřebuje pro své fungování více než jeden software. Samozřejmě, že existují rozsáhlá řešení, která pokrývají všechny agendy společnosti od účetnictví až po autopark. Praxe je často taková, že spousta firem vlastní různé softwary pro různé oblasti, každý takový software generuje své vlastní historické záznamy. Z těchto obvykle decentralizovaných systémů jsou potom data konsolidována v datovém skladu společnosti. Konsolidace dat není jednoduchý proces, ale v některých případech jde téměř o nutnost, neboť roztříštěná a nekonzistentní klientská data mohou velmi brzdit práci konzultantů a všech ostatních, kteří pravidelně jednotlivé informace využívají. Máme-li zákaznická data (obchodní a účetní data, nabídnuté či zakoupené produkty, technické supporty, kontaktní historie, výsledky dotazníků, reakce na soutěže zaslané formou newsletteru atd.) z různých systémů konsolidovaná v jednom celku, schopnost vytěžení netriviálních informací se zvětšuje, neboť tato data mají již za sebou určitou standardizaci. Ještě to ale neznamená, že máme vyhráno. Data v datovém skladu totiž vůbec nemusejí být jednoduše využitelná, a obvykle ani nejsou. Důležitým faktorem pro jejich využitelnost je jejich čistota.

Příčinou vzniku nekorektních záznamů v databázi je špatně nastavený proces, ať už na úrovni softwaru, který firma používá, nebo na úrovni jeho obsluhy. Úprava softwaru, jenž generuje data ve špatném formátu či s nedostatečnou informací, je z dlouhodobého hlediska mnohem efektivnější. Stejné je to i s uživateli CRM a dalších programů. Pokud se definují pravidla, jak každodenní informace (data, jména, poznámky) jednotně zadávat, využitelnost surových dat se zvýší.

Čistota dat z pohledu dalšího zpracování dat

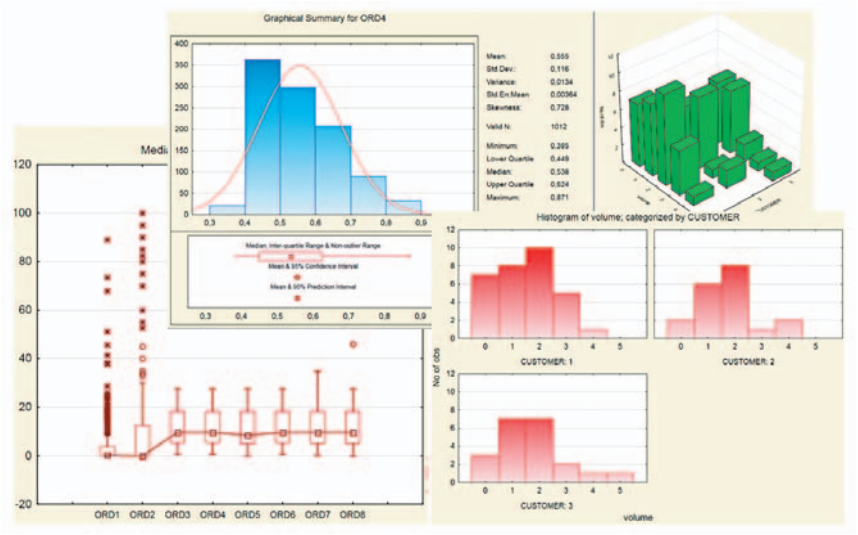
Data v syrové podobě nemají žádnou vypovídací hodnotu. Pohled na nekonečnou tabulku údajů nám nic neřekne, údaje musejí být zpracovány tak, aby byly cenné. Souhrnné informace z konkrétního objemu dat dostáváme obvykle ve formě grafických výstupů, agregovaných počtů, reportovacích tabulek apod. Operativní řízení a podpora rozhodovacích procesů obvykle vyžadují provádět nad datovým souborem složitější analýzy než výpočet průměrných hodnot a tabulek četností, například různé predikční modely, jež nám dají odhad budoucího stavu, a poskytnou tak podklad pro manažerská rozhodnutí. V určitých odvětvích jsou modely chování (klienta, pohledávek, trhu) vytvořeny nad historickými daty a používají se pro data nová. Pomocí původních dat se odhaduje, jak se bude klient chovat (zda mu půjčit, zda opustí firmu, zda zareaguje na další nabídku), jak se bude chovat poptávka po komoditě při zlevnění jejího substitutu, jak se bude vyvíjet návštěvnost v nejbližším období na základě známých vstupů. Možností je hodně, ale mají jedno společné: přesné a použitelné výstupy získáme pouze z přesných a použitelných

vstupních dat. Nepřesnosti, překlepy či nekorektní informace (záporná cena, chybně připojené produkty zákazníkovi, desetinná čárka na špatném místě, zaměněné pohlaví, ...) jsou tím, co nám nejvíce podlamuje spolehlivost závěrů, které nad daty děláme. Náš model bude tak přesný, jak přesný bude vzorek dat, na jehož základě jsme ho vytvořili. Jak je obecně známo, v rámci statistického modelování zabere samotná příprava dat přibližně osmdesát procent času.

Příprava datového souboru

Prvním krokem před vlastní analýzou je zajištění přesnosti záznamu, kde se snažíme zkontrolovat (validace dat) správnost jednotlivých dat. Sleduje se počet chybějících hodnot, duplicitní záznamy, kontrolují se jednotlivé varianty znaku, srovnatelnost jednotlivých proměnných v různých tabulkách apod. K odhalení těchto nepřesností v datovém souboru slouží celá řada technik, jež mají analytické softwary implementovány. Jednotlivé přístupy určuje typ atributu (numerický, kategorický, textový). Podle povahy dat potom aplikujeme vhodné metody, které ve své podstatě jako celek spadají do tzv. deskriptivní statistiky. Díky těmto výstupům, jimiž jsou například tabulka četností, kategorizované sloupcové grafy, 3D grafy, charakteristiky polohy a rozptýlenosti dat apod., snadno odhalíme chybná pozorování v datovém souboru, neboť každá proměnná má své logické omezení, které musí jednotlivá pozorování splňovat. Ze surových, nic neříkajících dat získáme jednak smysluplné informace o konkrétním datovém souboru, například za určité období, jež poslouží jako podklad pro další rozhodování, ale také se nám ukážou nekorektní, případně nekompletní záznamy (např. špatně nadefinovaný zákazník), které vznikly chybami při zadávání do informačního systému společnosti. Chyba bývá často na straně uživatele, který systém používá a data zadává. V konsolidovaných tabulkách je potom častým nedostatkem odlišný název kategorické proměnné, neboť co systém, to jiná konvence. Na výsledky výše zmíněných analýz navazuje překódování jednotlivých proměnných, standardizace názvů, kategorizace numerických hodnot, tvorba podmnožin a jejich následná úprava apod.

Další kapitolou jsou záznamy, které na první pohled nemohou vyvrátit příslušnost do konkrétního atributu tabulky, ale svou hodnotou jsou mimo těžiště dané proměnné. Jinak řečeno, jejich hodnota je výrazně odchýlená od průměru všech hodnot, ale



Graf pro detekci odlehlých hodnot a další grafické a číselné výstupy v rámci deskriptivní statistiky, které zachycují datový soubor z různých úhlů pohledu umožňujících odhalit nepřesná a nekorektní data

na první pohled tato hodnota vypadá, že do konkrétní proměnné patří. Tato pozorování nazýváme odlehlá, respektive extrémní. K jejich identifikaci slouží celá řada statistických technik, jež vycházejí z vlastností datových distribucí konkrétní proměnné. Extrémní pozorování mají vliv na spolehlivost výsledků většiny pokročilých statistických technik, a proto je obecně dobré konkrétní jednotky ze souboru vyřadit.

Explorační analýzou (jednorozměrnou, vícerozměrnou) provedeme identifikaci těchto hodnot a následně jejich prověření. U hodnot, které jsou obvykle identifikovány jako extrémní (nepřiměřená sleva, cena apod.), často zjistíme, že nepřísluší ke konkrétnímu souboru a konkrétní jedinec (pacient, zákazník) nebude ten, pro kterého bychom výsledky modelu používali, a proto ho ze souboru vyřadíme. Často jsou tato pozorování způsobena špatným zadáním hodnot na vstupu. Po identifikaci a případném vyřazení těchto hodnot je obvykle vhodné podívat se na vztahy mezi jednotlivými proměnnými, které bychom v konkrétním pokročilém modelu chtěli použít. Tedy jestli se proměnné, kterými chceme vysvětlovat transakční historii u klienta nebo objem prodeje našeho zboží v kamenných prodejnách, vzájemně příliš nevysvětlují (nekorelují mezi sebou), což by pro určitý typ analýz bylo nevýhodné.

Analytická data

Po aplikaci vhodných technik, překódování a odstranění chybných záznamů, máme k dispozici datový soubor, který je použitelný pro další analýzu. Otázkou je, jak analyzovat tzv. big data. Pro většinu modelovacích technik, jež se v odvětví, kde máme rozsáhlé soubory dat, používají, nám postačí pracovat se stratifikovanými vzorky, které jsou náhodným výběrem z původního extrémně rozsáhlého souboru. Někdy je však žádoucí vytvořit model na celém souboru, k tomu je potřeba mít softwarové vybavení, jež není omezené velikostí vstupního souboru a pokročilými algoritmy (např. postupný import z databáze) dokáže postupně zpracovat i zdánlivě nekonečné soubory dat.

Velké společnosti čelí potřebě uchovávat rozsáhlé datové soubory, včetně nestrukturovaných dat. Výše naznačené techniky se často aplikují již při transportu dat do datového skladu z provozních softwarů společnosti. Při automatizaci těchto procesů částečně odpadá potřeba kontroly datových vlastností a vztahů při přebírání dat k dalším účelům. Analytické softwary mohou svou práci vykonávat jednak nad samotným datovým skladem, ale také již při vlastním importu, v pravidelných intervalech, do datového skladu. Automatizované analýzy, které pomohou k identifikaci nekorektních položek, potom výrazně šetří čas na přípravu vlastního „normovaného“ souboru analytických dat.

Zachytit případné odlehlé, nečekané nebo nelogické hodnoty je zdlouhavá práce, ale při průběžném přístupu a automatizaci se dá výrazně zjednodušit. V dnešní době je kladen velký důraz na čištění a úpravu historických dat z provozních systémů společnosti, v menší míře se však hledá možnost odstranění příčiny vzniku nekorektních dat, ať už na úrovni softwaru nebo jeho obsluhy, čímž by se mělo naopak vždy začít. ■

Autor je odborným konzultantem společnosti Statsoft CR.