Text mining

aneb Kladivo na nestrukturovaná data

Miloš Uldrich

Trendem dnešní doby je obrovský nárůst počtu dat uložených v databázích. Je obecně známo, že až osmdesát procent uložených dat v databázích po celém světě má podobu textu, tedy nestrukturovaných dat. Chceme-li obsaženou informaci získat, musíme je všechny číst?

Kam s dokumenty

Téměř každý z nás si musel všimnout rychlého vývoje technologií v oblasti ukládání dat (a tím nemyslím přechod od děrných štítků). Někteří z nás ocenili výhody několikagigového MP3 přehrávače za cenu, kterou měl před pár lety přehrávač o desítkách megabytů. Jiní se zase radují nad velikostí dostupných datových úložišť a různých disků. Dalším oceňovaným bonusem souvisejícím s nárůstem prostoru pro data je často téměř nekonečná velikost e-mailových schránek, jejichž nabídkou poskytovatelé bojují o své zákazníky. A s možností ukládání velkého objemu dat také souvisí naše dnešní povídání.

Je obecně známo, že většina dat v databázích a datových úložištích má nestrukturovanou podobu. Pro společnosti na celém světě je proto žádoucí, aby se v těchto nestrukturovaných informacích, které jsou uloženy v diskových polích, neztratili a udržovali v nich pořádek, případně aby z nich vytěžili další zisk v podobě zajímavé a včasné informace.

Co je text mining

Informace obsažená v tisících či milionech textových dokumentů a různých formulářů (stížnosti, reklamace, žádosti) lze obvykle vyjádřit mnohem stručněji, protože klíčové sdělení je podstatně menší. Text mining je nástroj, který takové záznamy umí zpracovávat, poskytne nám stěžejní informaci o obsahu dokumentu a setřídí dokumenty podle podobnosti, aniž by je musel někdo číst. A to je samozřejmě v dnešní uspěchané době velice žádoucí.

Text mining obecně spadá pod soubor dataminingových metod, tyto metody však pracují s čísly, případně s nominálními či ordinálními proměnnými, jako jsou názvy kategorií apod. Text mining pracuje s nestrukturovaným textem, lze ho tedy definovat jako proces vytěžení cenné informace z textu, tato metoda však může pomoci i při samotné dataminingové analýze. Pojďme se nyní podívat na různé typy využití tohoto algoritmu.

Extrakce významu z nestrukturovaného textu

Nástroj text mining kvantifikuje jednotlivé objekty v textu (obr. 1). Objektem rozumíme jednotlivá slova nebo důležitá spojení – termy (eskontní úvěr, cystická fibróza, gotické památky), například slovo traumacentrum indikuje vyšší pojistné plnění, neboť klient byl pravděpodobně vážně zraněn. Termy se pak zobrazí v matici slov, která je vytvořena na základě frekvenční analýzy. Podle počtu a struktury slov lze identifikovat téma a smysl čteného dokumentu, přitom nemusí jít pouze o mnohastránkovou ročenku nebo diplomovou práci, ale například o webovou stránku. Zajímavější možností je potom definice konkrétních hledaných slov nebo spojení. Nástroj tak může prohledávat obsah webových stránek a nacházet ty s pro vás klíčovým sdělením.

	Words summary		
Stem / Phras	Count	Number of documents	Example
work	55	1	
custom	40	1	
statement	35	1	
model	19	1	
library	17	1	
program	17	1	
print	16	1	
name	14	1	
line	12	1	
file	11	1	
syntax	11	1	

Obr. 1

Automatické třídění textů

Ještě zajímavější vlastností textminingových nástrojů je potom identifikace specifických či podobných textových záznamů na základě shlukové analýzy. Textové záznamy jsou klasifikovány a tříděny do shluků podle podobnosti. Jak vypadá v praxi nejčastěji takový záznam?

Marketingové dotazníky s otevřenou odpovědí, reklamace

Text mining je často využíván pro analýzu otevřených odpovědí z webového průzkumu, analýzu reklamací apod., tedy všude tam, kde má klient prostor pro vlastní názor. Číst takové odpovědi je někdy velmi časově náročné a kolikrát téměř nemožné. Text mining tyto texty sám klasifikuje a roztřídí například podle typu odpovědi (stížnosti kladné, záporné, irelevantní apod.) a dalších podobných znaků. Marketingové oddělení se tak může zabývat pouze určitými typy

odpovědí, které mu dají cennou informaci pro tvorbu další prodejní strategie, a s nerelevantními odpověďmi neztrácet čas. Navíc má okamžitě přehled o četnostech kritických odpovědí za určité období.

Třídění textových záznamů

Obrázek 2 ukazuje jednotlivé textové záznamy (dokumenty, formuláře, žádosti atd.), které byly podrobeny analýze. Záznamy, které jsou mimo hlavní shluk, se nějakým způsobem od většiny dokumentů odlišují, a proto by jim analytické oddělení mělo věnovat pozornost. Mohou to být nepřesné žádosti o úvěr, falešné pojistné události, neobvyklý názor klienta na naši společnost apod.

Dalším typem nestrukturovaného textu je dodatečná informace o klientovi uložená v databázi. V moderních databázích jsou uchovávány nejen číselné informace, které předhazujeme data miningu, ale i další textové záznamy, které číselné informace upřesňují, často potom tvoří klíčovou část záznamu. Jsou to údaje o stavu pacienta, o aplikaci konkrétního léku, popis reakce na lék, záznamy o smlouvách klienta apod., záleží na odvětví. Díky text miningu můžeme v těchto klientských kartách nejen efektivně vyhledávat, ale také je automaticky roztřídit a podrobit další analýze.

Automatická identifikace podobných textových záznamů je neocenitelným pomocníkem při správě rozsáhlých úložišť. Také nám umožňuje efektivně analyzovat zpětnou vazbu od klientů či návštěvníků webových stránek, což pomáhá při analýze poptávky po službě nebo produktu a šetří čas pracovníka, který má analýzu nákupního chování na starosti.

Vylepšení dataminingové analýzy

Jak již bylo naznačeno, dataminingové metody s nestrukturovaným textem pracovat nedokážou, přitom textová pole jsou dnes běžnou součástí klientských záznamů uložených v databázích. Jak může tyto záznamy analytické oddělení využít?

Prvním příkladem je již zmiňovaná segmentace těchto textových řetězců dle povahy problémů, které danou oblast reprezentují. Klíčová slova pro danou oblast roztřídí záznamy podle typu reklamace, typu úrazu, druhu pojistné události atd., lze také shlukovat podobné výsledky vstupních prohlídek, záznamy z pohovorů, obdobné reakce na léky atd. Na univerzitách je frekvenční analýza využívána pro kontrolu podobnosti diplomových prací.

Jednotlivé shluky podobných textových záznamů jsou klasifikovány a zařazeny zpět do databáze ve formě numerické proměnné. Ta je potom využita v dataminingové analýze – například v technice, která se nazývá rozhodovací stromy. Tou například farmaceutické společnosti zkoumají nežádoucí účinky svých léků. V pojišťovnictví a bankovnictví se celosvětově ve velké míře využívají výsledky textminingové analýzy pro zpřesnění dataminingových modelů, které klasifikují a neustále automaticky vyhodnocují jednotlivé žádosti o pojistná plnění, záznamy o škodách, úvěrové dokumenty apod. Tím se dostáváme do oblasti detekce podvodů, kde má pokročilá dataminingová analýza své místo.

Fraud management

Fraud management neboli detekce podvodů je oblast, která se zaměřuje na včasné odhalení podvodného jednání. Do fraud managementu spadá celá řada opatření, neboť podvodné jednání je zde od počátku věků. S rozvojem elektroniky a počítačů se objevily i nové generace podvodníků, kteří hledají skuliny v IT systémech firem a velkých společností. Podvody přitom nepáchají pouze lidé z vnějšku společnosti (jednotlivci a organizované skupiny), ale také zaměstnanci. Ve velké míře jde potom o kooperaci zaměstnance (např. likvidátora) a člověka, který stojí mimo firmu. Pojišťovny, banky a další finanční instituce však nemusejí být vůči takovému jednání úplně bezbranné. S rozvojem podvodů a cílených útoků na infrastrukturu společností jde totiž ruku v ruce vývoj aplikací, které se tomuto jednání snaží zabránit.

Textminingový nástroj v této oblasti slouží pro potřeby interní kontroly. Automaticky čte e-maily zaměstnanců, a pokud detekuje určité slovo nebo spojení, které ukazuje na podvodné jednání, je e-mail označen a příslušné oddělení mu potom věnuje zvýšenou pozornost. Stejným způsobem textminingový nástroj analyzuje také elektronické žádosti, objednávky přes internet apod., které do firmy přicházejí z vnějšku. Vstupy jsou tříděny do smysluplných shluků, a lze tak odhalit například nepřesnou žádost, podezřelou objednávku apod.

Kvalitní fraud management nezahrnuje pouze softwarové nástroje, ale celou řadu opatření, která začínají firemní kulturou a uvědomělými zaměstnanci. Bez nich by prevence podvodů neměla zdaleka takovou úspěšnost. Pracovníci na různých úrovních mohou sami poznat a ohlásit podezřelé jednání (enormní výběr z účtu, podivné pojistné plnění vzhledem k věku pojistníka). Zaměstnanci však nemohou uhlídat všechno, a tak softwarové nástroje ostatní opatření nenahrazují, ale naopak vhodně doplňují. Komplexní přístup k detekci podvodů přináší zmiňovaným institucím výrazné úspory. Mezi organizovanými skupinami je dobrá informovanost. Pokud se zjistí, že nějaký podvod již nefunguje, rychle se to dozvědí ostatní. A proto se zavedením těchto opatření postupně také klesá počet pokusů o podvod. Ale to jen předznamenává to, že nepoctivci vymýšlejí jiný druh podvodu... ■

Autor je odborným konzultantem společnosti StatSoft CR.

Obr. 2