Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

Pokročilé přístupy k DZD

Zpoždění vlaků ČD

Zpráva o specifikaci podoby umělých dat

Tomáš Pokorný

Jan Žítek

ZS/2015

1 Popis domény a motivace generování dat

V naší semestrální práci se budeme zabývat generováním umělých dat z domény vlakové dopravy, konkrétně dat o zpoždění vlakových spojů. Všechny získané doménové znalosti, které budou zmiňovány a využívány dále, se vztahují k vlakovým spojům Českých drah (ČD). Motivem pro vybrání tohoto tématu bylo jednak to, že oba máme s tímto druhem zpoždění mnoho zkušeností, ale také to, že je velmi obtížné dohledat ucelená skutečná data. Existují sice aplikace, pomocí kterých lze sledovat polohu a zpoždění vlaků, ale jakékoli další automatické zpracování těchto dat je bez souhlasu Českých drah zakázáno. Jedná se přitom o fenomén s velkými dopady na denní aktivity obrovského množství lidí. Pravidelní cestující sice obvykle počítají s možným výskytem malého zpoždění (řekněme do 15 minut) a vytváří si časové rezervy, větší zpoždění již ale může způsobit velké problémy v souvislosti s jejich absencí na pracovištích, školách, schůzkách atp. Podle zpráv [1] bylo za rok 2014 převezeno největším domácím dopravcem přes 170 milionů lidí. Při velmi hrubém odhadu, jestliže bylo zpožděno přes 6 procent vlaků, tak se zpoždění dotklo více než 10 milionů cestujících.

Skutečná data by mohla sloužit širokému spektru subjektů. Jedním z nich by byly i samotné České dráhy. V současnosti existuje rozdíl mezi způsobem vyplácení odškodného cestujícím za zpoždění u nás a v členských státech EU řídících se nařízením z roku 2007 [1], jelikož v ČR platí až do roku 2019 výjimka, kterou v roce 2014 prodloužila vláda. ČD sice vyplácejí odškodné 25 procent z ceny jízdenky při zpoždění nad 60 minut (50 procent při zpoždění nad 120 minut), pokud výše odškodného přesáhne 100 korun (jinak není propláceno), to se ale nevztahuje na případy způsobené „vyšší mocí“, tedy třeba počasím (na rozdíl od unijních pravidel). Použití metod k objevování skrytých vztahů na data týkající se vlakové dopravy obohacená o data zachycující vnější vlivy by tak mohlo poskytnout Českým drahám cenné informace. Může se jednat například o dopady kombinací různých vnějších vlivů na délky zpoždění a s nimi spojenými odškodněními, potažmo nevyplácenými odškodněními v době výjimky z nařízení EU.

Na základě informací z veřejně dostupných zdrojů, jako jsou stránky Českých drah [2] [3] [4], nebo články na zpravodajských webech [1] [6], případně nám známých skutečnostech, lze o vlakových spojích Českých drah tvrdit následující základní fakta:

- Celkový počet spojů z každého dne je rozdělen mezi 5 typů vlaků tak, že zhruba 89 % tvoří osobní vlaky (Os), 3,5 % spěšné vlaky (Sp), 6 % rychlíky (Rj), 1 % vlaků dalších vlaků na dlouhé vzdálenosti (Ec, Ic, En, Ex), ty budou dále označovány pouze jako EC, a necelých 0,5 % tvoří Pendolina (Sc).
- Vlaky jezdící na delší vzdálenost jsou častěji zpožděné (cca 22 % případů) než vlaky spadající do regionální dopravy (cca 9 % případů) a jejich zpoždění nabírají vyšších hodnot. Výjimku představují vlaky Pendolino, které mají méně zpoždění než regionální vlaky.
- Přibližně 40 % zpoždění je spojeno s vnějšími vlivy (mimořádné události), které ČD nemohou ovlivnit.
- V ranních a večerních hodinách jezdí více vlaků než v ostatních denních dobách (CF-Miner).

2 Popis navržené struktury dat a rozdělení hodnot

2. 1 Hlavní data

Skupina	Sloupec	Datový typ	Popis
Vlak	Typ vlaku	Text	O jaký typ vlaku se jedná (Os/Sp/Ec/Rj/Sc).
Vlak	Zpoždění	Integer	Délka zpoždění vlaku.
Vlak	Zrušen	Boolean	Zda vlak vůbec jel.
Trať	Výluka	Boolean	Zda je či není na trati výluka. Například z důvodu stavebních činností
Trať	Mimořádná událost	Boolean	Zda se na trati vyskytla či nevyskytla událost způsobená vnějšími vlivy.
Trať	Provoz na trati	Text	Rozlišení vytíženosti trati.
Datum/Čas	Předpokládaný příjezd	Time	Určuje dobu, kdy měl vlak přijet do cílové stanice. Slouží pro rozlišení denní doby.
Datum/Čas	Datum	Date	Kalendářní datum dne, kdy vlak jel.

2. 2 Externí data

Skupina	Sloupec	Datový typ	Popis
Datum/Čas	Datum	Date	Kalendářní datum dne.
Datum/Čas	Den v týdnu	Text	O jaký konkrétní den v týdnu se jedná. Slouží například pro rozlišení víkendů.
Datum/Čas	Svátek	Boolean	Zda je tento den státním svátkem.
Počasí	Teplota	Decimal	Průměrná teplota dne ve stupních Celsia.
Počasí	Srážky	Decimal	Denní úhrn srážek v milimetrech

Průměrné denní teploty a srážky byly získány ze stránek NOAA Satellite and Information Service [5]. Jedná se o hodnoty pro všechny dny roku 2014 naměřené v pražském Klementinu. Po jejich získání bylo nutné převést je z původních hodnot (stupňů Fahrenheita a palců) na námi používané jednotky (stupně Celsia a milimetry).

3 Popis navržených vztahů v datech

V první části byly vypsány některé obecné vztahy vyplývající z doménových znalostí. Kromě výše zmíněných pravidel budou v datech zahrnuta i pravidla následující:

3. 1 4FT-Miner

- Při mimořádné události na trati je zpoždění vlaku více než 30 minut.

- Mimořádné události nastávají, mimo jiné, při kombinaci průměrných denní teplot menších než 5 stupňů a vysokých srážek.
- Mimořádné události často nastávají také v ranních hodinách, pokud průměrná denní teplota klesne pod 5 stupňů.
- Pokud se jedná o trať s velkým provozem, pak výluka na trati znamená zpoždění u všech vlaků.
- Pokud se jedná o trať s velkým provozem (a nemusí dojít k výluce ani mimořádné události), tak vznikají zpoždění u více než 30 % osobních vlaků.
- Pokud se jedná o trať se středním provozem, pak výluka na trati znamená zpoždění u osobních vlaků a spěšných vlaků.
- V pracovní dny dochází k více zpožděním než v nepracovní dny (víkendy a svátky)

3. 2 CF-Miner

- Výluky na trati vznikají nejčastěji v létě.
- Mimořádné události vznikají nejčastěji v zimně.

3. 3 KL-Miner

- Čím větší jsou průměrné denní srážky, tím delší je zpoždění vlaků.

4 Zdroje

[1] Aktuálně.cz: Zpoždění přes hodinu? Kdy vám České dráhy dají odškodnění.
Dostupné z: http://zpravy.aktualne.cz/finance/zpozdeni-pres-hodinu-drahy-zavedly-dobrovolne-odskodneni/r~17ef74547ae311e4840b002590604f2e/

[2] České dráhy: O společnosti. [online]
Dostupné z: https://www.cd.cz/infoservis/o-spolecnosti/-3540/

[3] České dráhy: Tiskové zprávy [online]
Dostupné z: http://www.ceskedrahy.cz/tiskove-centrum/tiskove-zpravy/-14775/

[4] České Dráhy: Standardy kvality společnosti České dráhy a.s.
Dostupné z: https://www.cd.cz/assets/infoservis/cim-se-ridime/standardy-kvality-spolecnosti-ceske-drahy--a-s-.pdf

[5] NOAA Satellite and Information Service, National Climatic Data Center [online]
Dostupné z: http://www7.ncdc.noaa.gov/CDO/cdo

[6] Železničář.cz: ČD loni přepravily rekordní počet cestujících v pětileté historii [online]
Dostupné z: https://zeleznicar.cd.cz/zeleznicar/hlavni-zpravy/cd-loni-prepravily-rekordni-pocet-cestujicich-v-petilete-historii/-6499/