PROBLÉMY LOKALIZÁCIE TWEETOV

Martin ZAJAC[1], Jiří HORÁK[1]

[1] Katedra geoinformatiky, Hornicko-geologická fakulta, Vysoká škola báňská – Technická univerzita Ostrava, 17. listopadu 15, 708 33 Ostrava, Česká republika
martin.zajac.st@vsb.cz, jiri.horak@vsb.cz

doi: https://doi.org/10.31490/9788024845081-124

Abstrakt

Využívanie on-line sociálnych sietí je stále viac populárne. Často je študovaná reakcia komunikujúcej verejnosti na rôzne témy, obsahová analýza ich príspevkov a ich sentiment. Príspevok sa zaoberá problémom lokalizácie sťahovaných príspevkov zo siete Twitter, kde len časť obsahuje priamo súradnice. Na príklade tweetov k verejnej doprave z Londýna, Madridu a Prahy dokumentuje možnosti geokódovania, možnosti využitia toponým, presnosť lokalizácie, rozdiely medzi polohou získanou geokóidovaním a skutočnou polohou menovaného objektu a diskutuje problémy lokalizácie a možnosti spresnenia.

Abstract

The use of on-line social networking sites is becoming more and more popular. The reaction of the communicating public to various topics, the content analysis of their posts and its sentiment are frequently studied. The presentation is focused on problems of localization of downloaded posts from the Twitter from which only a part directly contains coordinates. Using the example of tweets related to public transport from London, Madrid and Prague the paper provides an overview of different possibilities of geocoding, possibilities of using toponyms, accuracy of localization, differences between the geocoding-based location and the real location of the relevant object and it enables to discuss localization problems and possibilities of their solution.

Klíčová slova: sociálna sieť; lokalizácia; geokódovanie; verejná doprava

Keywords: social site; localization; geocoding; public transport

1. ÚVOD

Denne je publikované po rôznych platformách nespočetné množstvo príspevkov. Ľudia sú schopní takmer okamžite reagovať na udalosti, šíriť informácie či viesť debaty na aktuálne témy. Vo veľkých mestách bývajú predmetom takýchto debát aj udalosti a problémy vo verejnej doprave. Vyjadrenia bývajú spravidla expresívne a podfarbené sentimentom. Čo si však ľudia možno neuvedomujú, je že pri napísaní komentáru pod nejaký príspevok alebo opísaní dobrej či zlej skúsenosti, ktorá sa im stala, prispievajú do obrovského zdroja nového druhu geodát. Treba si však o týchto dátach vytvoriť ucelený obraz, nakoľko pri práci s nimi vyskytuje aj niekoľko problémov. Reklamy, nadbytočný obsah, problém z dohľadaním kontextu a problém lokalizácie príspevkov. V polovici roka 2019 bola ukončená možnosť lokalizovať tweet pomocou dvojice súradníc a polohy mobilného zariadenia. Stále je však možné príspevky lokalizovať pomocou „twitter places“ a to pre rôzne úrovne lokalít (od POI – Point of Interest až po mesto či štát). Stále je tu však možnosť geokódovať príspevku cez informáciu v texte. Tu je však nutné zvoliť sofistikovaný prístup a mať aspoň čiastočnú znalosť lokálnych pomerov.

2. PRIESTOROVÉ INFORMÁCIE V DÁTACH

Dáta zo sociálnych sietí môžu obsahovať priestorové informácie v rôznych podobách. Najpraktickejšou formou je vyjadrenie polohy v geografických súradniciach. Než sa však príspevku priradí súradnica, je nutné splniť nasledujúce podmienky: Užívateľ musí mať povolené vo svojom zariadení pristupovať príslušnej aplikácii k polohe zo zabudovaného GNSS prijímača. Užívateľ geotaguje príspevok tak, že vyberie jeho polohu zo zoznamu kandidujúcich miest v blízkosti jeho polohy. Môže ísť o bodové vyjadrenie (napr. zastávka MHD, reštaurácia), alebo takzvané ohraničené územie (anglicky „bounding box“) čo môže

reprezentovať celú ulicu, mestskú časť poprípade kraj alebo štát. Pritom musí použiť správne označenie - užívateľ musí geotagovať svoju presnú polohu tak, aby korešpondovala z takzvaným „Twitter Place“, čo je vo svojej podstate zoznam miest poprípade inštitúcii alebo podnikov ktoré sú známe Twittru a priradí im polohu v podobe súradníc. Je treba upozorniť, že výber zo zoznamu je obmedzujúci, i keď je zoznam obsiahly a aktualizovaný. Twittru ho poskytuje Foursquare. Pri voľbe miesta samozrejme závisí na užívateľovi aký geotag priradí k svojmu príspevku. Preto sa v takejto polohe skrýva veľké množstvo neistoty, nakoľko užívateľ nemusí vyplniť správnu polohu alebo ju vyplní natoľko generalizovanú, že bude pre ďalšie spracovanie nepoužiteľná (napr. na úrovni mesta, kraja, krajiny). V metadátach tweetu sa objavujú tieto údaje o polohe (Tabuľka 1).

Tab. 1 Lokalita tweetu v metadátach

Názov	**Popis**
place_url	JSON s informáciami o „Place name“ a „place full name“
place_name	Názov miesta
place_full_name	Celý názov miesta aj s názvom miesta o jednu vyššiu úroveň (mestská časť + mesto, mesto + krajina)
place_type	Typ miesta, ktoré bolo označené v tweete, môže nadobúdať hodnôt „country“, „neighborhood“, „city“, „admin“ a „POI“ (point of interest).
country	Názov krajiny
country_code	Kód krajiny
geo_coords	Súradnice vo formáte Longitude, Latitude
coords_coords	Súradnice vo formáte Latitude, Longitude
bbox_coords	Súradnice ohraničujúceho štvoruholníka
location	Označenie miesta užívateľa

Polia geo_coords a coords_coords obsahujú súradnicu, kde sa zariadenie nachádzalo v čase, keď bol tweet odoslaný. Možnosť takto precízne lokalizovať tweet však bola pozastavená v júni 2019 (https://twittercommunity.com/t/recent-tweet-compose-update-might-affect-some-geo-use-cases/126982). Oficiálne stanovisko Twitteru je, že málo ľudí takto presnú lokalizáciu používalo, a preto chcú, aby bolo používanie Twitteru a s tým aj určovanie polohy pre príspevok jednoduchšie (https://twitter.com/TwitterSupport/status/1141039841993355264). Domnievame sa, že chcú týmto krokom podporiť ľudí, aby lokalizovali príspevky na úrovni place_name nakoľko v Twitter API v.2 bude možné získavať tweety lokalizované na úrovni jednotlivých miest.

Obsah poľa bbox_coords je vyplnený pri oboch prípadoch určenia polohy. Je potrebné brať do úvahy, ide o sériu LON súradníc a potom sériu LAT súradníc.

Uvádza sa, že pre sociálnu sieť Twitter sa takto podarí získať približne len 1% tweetov (developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects).

Location popisuje domácu polohu užívateľa a nie tweetu, napriek tomu má svoj význam aj táto polohová informácia. Umožňuje napr. rozlíšiť miestnych užívateľov od návštevníkov a cudzincov.

Inou možnosťou, ako lokalizovať tweet, je využiť priamo obsah správy. Ľudia majú sami o sebe väčšiu tendenciu označovať miesta záujmu (ďalej len miesta), o ktorých hovoria ich názvom a nie súradnicami. Preto by mal byť tento prístup pre priraďovania dát zo sociálnych sietí k miestam vhodnejší ako vyššie spomínaný prístup. Ako zoznam miest, ktoré hľadať v jednotlivých záznamoch možno použiť databázu miestopisných názvov. Tento postup však nie je bez problémov. Ľudia v svojich príspevkoch na sociálne

siete používajú prevažne slangové poprípade nárečové názvy jednotlivých miest rovnako ako ich skratky. Takéto názvy sa spravidla značne líšia od oficiálnych miestopisných názvov.

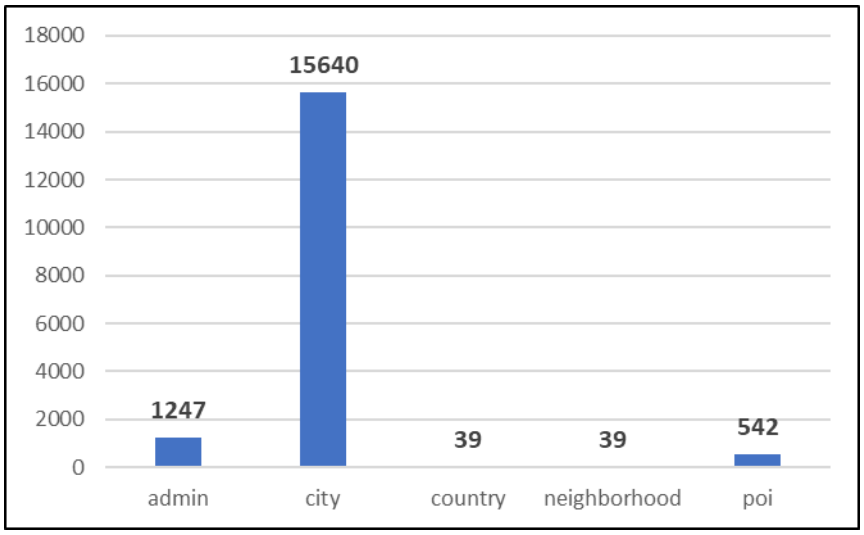
3. ZBER DÁT

V rámci diplomovej práce bol prevedený zber dát v období od 26.3.2020 do 31.1.2021 pre twitterový účet dopravného podniku Londýna (@TfL – Transport for London). Pre oblasť Hlavného mesta Prahy (@DPPOficialni) začal zber o niečo pozdejšie a to 23.4.2020 a prebiehal rovnako do 31.1.2021. Ako posledný sa začal zber pre Ostravu (@DPOstrava) a to 31.7.2020 do 31.1.2021. Zber prebiehal v týždňových periódach.

4. SPÔSOBY LOKALIZÁCIE

Pri voľbe prístupu k lokalizácii je potrebné si dobre uvedomiť čo má byť výsledkom lokalizácie, nakoľko každá z možností nám vo výsledku dá úplne odlišné informácie o polohe. Dáta obsahujú rôzne informácie a preto je treba rozlíšiť či chceme lokalizovať:

- Tweet - Pre lokalizovanie tweetu je najvhodnejšie použiť informácie z metadát. Hoci Twitter svojím rozhodnutím zastaviť možnosť precíznej lokalizácie tweetov zhoršil presnosť takejto lokalizácie, stále je možné tweety lokalizovať pomocou place_name až na úroveň záujmového bodu (POI). Pre Londýn sa takto podarilo úspešne lokalizovať až 3% (17 507) z celkových 545 295 tweetov. Problém je že až 90% z týchto tweetov bolo lokalizovaných len na úrovni mesta (Obrázok 1), čo nám neposkytuje dostatočné priestorové rozlíšenie.



Obr. 1 Prehľad počtu tweetov na rôznych úrovniach lokalizácie

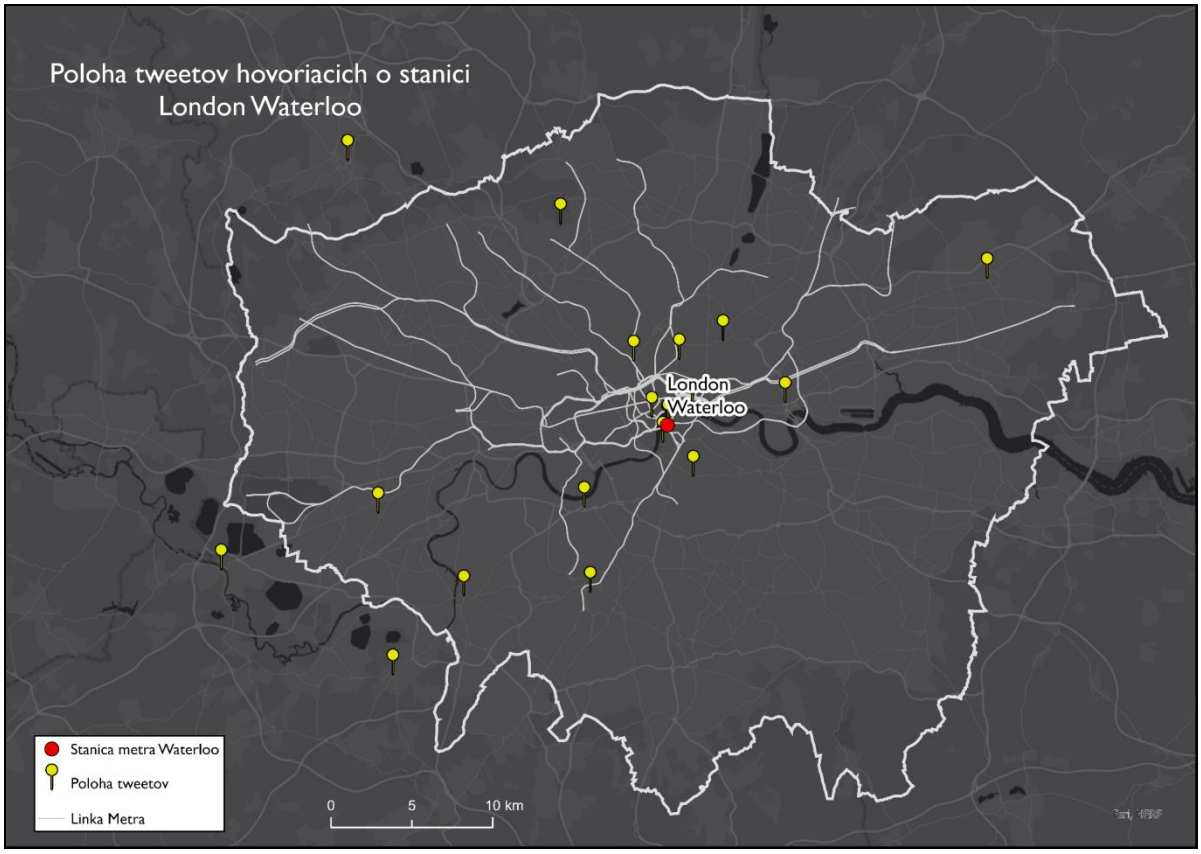
- Text - Lokalizovanie textu tweetu, nám umožní priradiť miesto k emócii, sťažnosti resp. problému, ktorý sa zmieňuje v samotnom texte. Poloha tweetu z metadát je pravdepodobne poloha autora tweetu v čase, keď tweet uverejnil. Samotný text môže niesť informáciu o úplne inom mieste. Príkladom zmienky takéhoto miesta (geokódu) v texte môže byť názov mesta, zastávky mestskej hromadnej dopravy, poštovej doručovacej adresy apod. (Rapant, 2006).

Pre lepšiu demonštráciu je možné uviesť príklad. Text tweetu rozoberá situáciu ohľadne Oyster karty pre dopravu v Londýne: „Coming from the same TFL that increased the congestion charge to £15/day, extended its operational hours and days, scrapped free buses for Zip oysters and scrapped free travel in peak hours for over 60s“. Problém však je tweet má v metadátach určenú polohu v meste Rím (Obrázok 2).

Rome	Rome, Lazio	city	Italy	IT

Obr. 2 Metadáta o polohe tweetu

Dôkazom že nejde o ojedinelý jav sme vzali množinu tweetov lokalizovaných pomocou textu na stanicu London Waterloo a vykreslili sme do mapy miesta, ku ktorým boli nahlásené pomocu place_name (Obrázok 3). Záznamy lokalizované na úrovni mesta, boli zobrazené ako stred územia daného mesta.



Obr. 3 Poloha tweetov k stanici Waterloo z metadát

5. TESTOVANIE LOKALIZÁCIE NA PRÍKLADE LONDÝNA

Pre sťahovanie dát z Londýna sa zvažovalo niekoľko prístupov:

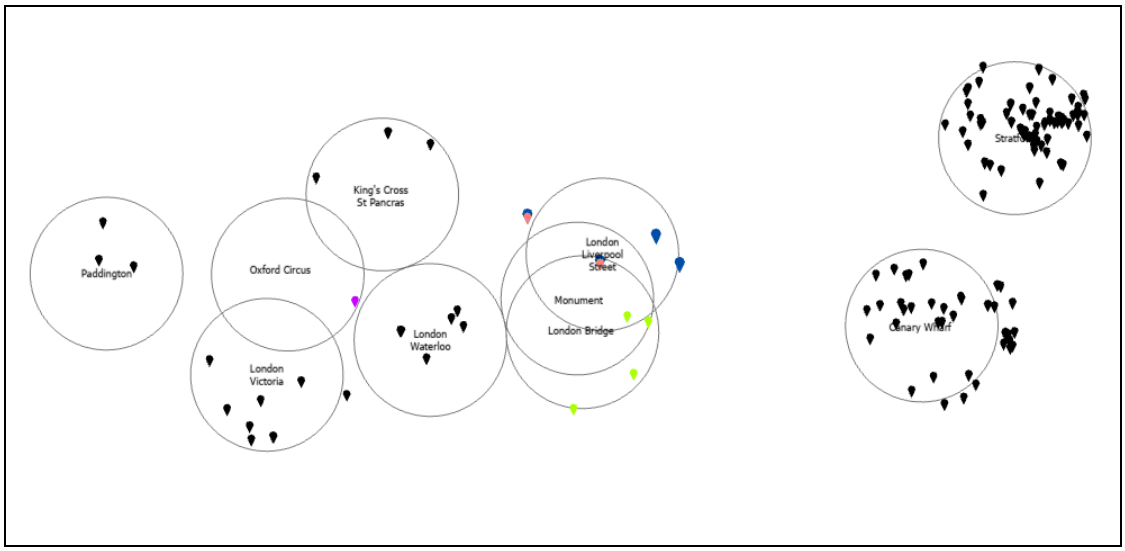
- Všetky tweety obsahujúce zmienku twitterového účtu londýnskeho dopravného podniku Transport for London „@TfL“. Výhodou tohto spôsobu je nadobudnutie veľkého objemu dát a vďaka tomu aj úplné a po časovej stránke súvislé spektrum diskutovaných problémov. Naopak nevýhodou je lokalizácia - prakticky lokalizovať je možné len zlomok a to buď súradnicami v metadátach alebo pomocou kľúčových slov (typicky názov stanice alebo linky, prípadne iného označenia miesta).
- Všetky tweety obsahujúce názov jednej z 10 najľudnatejších staníc v Londýne. Výhodou je že všetky tweety sú lokalizované (predpokladá sa lokalizácia v mieste stanice, o ktorej tweet hovorí). Nevýhodou je možná chyba z opomenutia (nie je možné zistiť všetky používané názvy stanice prípadne slang). Rovnako je nevýhodou fakt že takto frekventovaná stanica je aj komplexná zóna nákupných zón a reštaurácii a preto nie je možné zabezpečiť aby všetky tweety obsahujúce názov stanice boli aj relevantné k doprave. V neposlednom rade takáto selekcia podľa názvu stanice neposkytne súvislé spektrum diskutovaných problémov.

- Všetky dáta nachádzajúce sa do určitého okolia od stredu stanice. Výhodou je že všetky dáta sú presne lokalizované (minimálne na úrovni POI a v najlepšom prípade presnou súradnicou). Nevýhodou je že takto stiahnuté dáta nemusia mať vôbec nič spoločné s príslušnou stanicou a už vôbec nie s dopravou.

5.1 Kontrola polohovej presnosti tweetov získaných geografickým výberom

Poloha súradnicovo lokalizovaných tweetov bola vizuálne kontrolovaná voči súradnici stanice metra získanej z mapy.cz a bufferu o veľkosti 1 míle okolo tejto súradnice. Lokalizované tweety je možné vidieť na Obrázku č.4. Pri niektorých staniciach sa na prvý pohľad môže zdať, že obsahujú len jeden tweet;v skutočnosti sa ich ale viacej prekrýva na tom istom mieste. Smerodajná odchýlka bodov okolo stanice metra je 177 m.

Aj keď je vidieť u niektorých staníc väčší rozptyl bodov, nemôžeme hovoriť o zlej alebo nepresnej lokalizácii tweetov. Väčšina (80%) sa nachádzajú v danom rádiuse. Zvyšných 343 (20%) je mimo uvedený rádius a pre stiahnutie aj týchto tweetov nie je zrejmý dôvod. Väčším problémom ale je, že ani jeden takto získaný tweet nemal nič spoločné s dopravou. Často ide o tweety vztiahnuté k POI v blízkosti stanice. Takýto prístup k získavaniu dát je nevhodný z dôvodu veľkého šumu.



Obr. 4 Znázornenie polohy tweetov v porovnaní s polohou stanice a rádiusom 1 míle

Po otestovaní a zvážení všetkých výhod a nevýhod jednotlivých variant k získavaniu dát bol ako najlepší variant vyhodnotený variant 1. Na úroveň 10 najfrekventovanejších staníc a liniek londýnského metra sa podarilo z celkového počtu lokalizovať 3% tweetov (Tabuľka 2, 3).

Tab. 2 Prehľad úspešne lokalizovaných tweetov na úroveň najfrekventovanejších staníc

Stanica	**Počet tweetov**
Bank Monument	9
Cannary Wharf	296
Oxford Circus	237
Paddington	675
Stratford	949
Victoria	1870
Waterloo	1134
Liverpool Street	560
London Bridge	636
King’s Cross St. Pancras	676

Tab. 3 Prehľad úspešne lokalizovaných tweetov na úroveň linky metra

Stanica	**Počet tweetov**
Waterloo & City Line	7
Victoria Line	843
Piccadilly Line	1107
Northern Line	1527
Metropolitan Line	318
Jubilee Line	1466
Hammersmith & City Line	12
District Line	1506
Circle Line	392
Central Line	2495
Bakerloo Line	820

6. ZÁVER

V tejto práci nebolo možné využitie súradnicovo lokalizovaných tweetov, nakoľko Twitter túto možnosť zastavil ešte pred začatím tvorby tejto práce. Ponechaná zostala možnosť lokalizovať tweet využitím twitter place. Ukázalo sa že takáto lokalizácia nemá vhodné priestorové rozlíšenie z dôvodu cca 90 % zastúpenia tweetov lokalizovaných len na úroveň mesta. Najvhodnejší sa ukázal byť prístup geokódovania z textu tweetov. V tejto práci bolo úspešne geokódovaných približne 3% tweetov zo získanej vzorky. Tweety boli geokódované na úroveň liniek metra a 10 najrušnejších staníc. Pri geokódovaní dát takéhoto charakteru je nutné počítať s chybou z vynechania. Ako z ďalšou možnosťou lokalizácie je možné pracovať s lokalitou uvedenou v profile užívateľa. Táto práca identifikuje významnosť stanovenia účelu lokalizácie a samotného predmetu lokalizácie. Pre problémy spojené s dopravou bol ako najvhodnejší identifikovaný geoparsing, teda lokalizácia pomocou textu. Ako výzvu do budúcna vidíme využitie Twitter API druhej generácie, ktorá svojimi novými funkciami rozšíri možnosti získavania dát a tým vytvorí priestor pre nové druhy analýz. Rovnako veľkou výzvou do budúcna je spracovanie informácii z užívateľských profilov a následná kategorizácia jednotlivých užívateľov resp. ich sociálne zaradenie a vzťah k doprave.

7. LITERATURA

1. Osorio-Arjona J., Horak J., Svoboda R., García-Ruíz Y. (2021): Social Media Semantic Perceptions on Madrid Metro System: Using Twitter Data to Link Complaints to Space. Sustainable Cities and Society 64): 102530. https://doi.org/10.1016/j.scs.2020.102530
2. Rapant, P. (2006): Geoinformatika a geoinformační technologie. Ostrava: VŠB - Technická univerzita Ostrava, Hornicko-geologická fakulta, Institut geoinformatiky.

Odkaz na www stránku:

1. https://twittercommunity.com/t/recent-tweet-compose-update-might-affect-some-geo-use-cases/126982
2. https://twittercommunity.com/t/recent-tweet-compose-update-might-affect-some-geo-use-cases/126982
3. https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects