Mendelova univerzita v Brně
Provozně ekonomická fakulta

Detekce duplicit v geoprostorových datech

Diplomová práce

Vedoucí práce:
Ing. Pavel Turčínek, Ph.D.

Bc. Adam Prchal

Brno 2024

Poděkování

Velké poděkování patří vedoucímu diplomové práce Ing. Pavlovi Turčínkovi, Ph.D. za užitečné rady, vedení a ochotu konzultovat v jakoukoliv hodinu. V neposlední řadě patří poděkování také všem, kteří se jakkoliv podíleli na zlepšení kvality této práce.

Čestné prohlášení

Prohlašuji, že jsem práci **Detekce duplicit v geoprostorových datech** vypracoval samostatně a veškeré použité prameny a informace uvádím v seznamu použité literatury. Souhlasím, aby moje práce byla zveřejněna v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách ve znění pozdějších předpisů a v souladu s platnou Směrnicí o zveřejňování závěrečných prací.

Jsem si vědom, že se na moji práci vztahuje zákon č. 121/2000 Sb., autorský zákon, a že Mendelova univerzita v Brně má právo na uzavření licenční smlouvy a užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

Dále se zavazuji, že před sepsáním licenční smlouvy o využití díla jinou osobou (subjektem) si vyžádám písemné stanovisko univerzity, že předmětná licenční smlouva není v rozporu s oprávněnými zájmy univerzity, a zavazuji se uhradit případný příspěvek na úhradu nákladů spojených se vznikem díla, a to až do jejich skutečné výše.

Brno 2024 .....

podpis

Abstract

.

.

Abstrakt

.

.

Obsah

Todo list

Seznam obrázků

1 Úvod a cíl

1.1 Úvod

Společnosti poskytující služby v oblasti Location Intelligence využívají své platformy pro komplexní geoprostorové analýzy, což je zásadní pro efektivní rozhodování v různých typech podnikání. Jako součást těchto platforem, některé společnosti, včetně CleverMaps, nabízí služby zvané jako Data Marketplace, které umožňují uživatelům získávat a integrovat různorodé datové sady. Tyto sady zahrnují např. demografické informace, obchodní statistiky a infrastrukturu měst, což pomáhá klientům lépe cílit své služby a strategická rozhodnutí. (CleverMaps, 2024)

Vzhledem k tomu, že při práci s geoprostorovými daty dochází ke kombinaci dat z různých zdrojů, vzniká problém duplicitních záznamů. Duplicity v datech mohou mít různé podoby a vznikají z několika důvodů:

- **Rozdílné názvy stejných míst** – např. *"Velký městský park"* vs. *"Park v centru města"*.
- **Typografické chyby a nejednotné formáty** – např. ulice *"Masarykova"* vs. *"Masarykova tř."*.
- **Rozdílné souřadnicové systémy** – jeden dataset může používat *WGS84*, zatímco jiný *S-JTSK*.

Tyto duplicity snižují kvalitu dat a mohou způsobit chyby v rozhodovacích procesech, např. při plánování dopravy, marketingových analýzách nebo při geokódování obchodních poboček. Proto je detekce a eliminace duplicit v geoprostorových datech klíčová. K identifikaci duplicit mohou být využity různé metody, od jednoduchých textových porovnání až po složitější techniky strojového učení, které analyzují podobnost dat v širším kontextu. (Nauman a Herschel, 2022; Christen, 2012)

Existují komerční i open-source nástroje pro detekci duplicit, které často využívají cloudové technologie a pokročilé algoritmy. Mezi ně patří např. Data Ladder, Tilores nebo Melissa. Ačkoliv tyto nástroje zlepšují kvalitu dat, často představují vysoké náklady nebo nejsou dostatečně přizpůsobitelné konkrétním datovým sadám. To motivuje společnosti k hledání vlastních řešení. (Christen, 2012)

1.2 Cíl

Cílem této práce je prozkoumat a otestovat různé metody detekce duplicit na geoprostorových datech, včetně metod založených na strojovém učení. Na základě analýzy výsledků testů doporučit nejvhodnější metody pro konkrétní typy sad geoprostorových dat, přičemž ověření těchto metod proběhne na datových sadách poskytnutých společností CleverMaps.

Výsledná doporučení by měla společnosti CleverMaps pomoci v rámci zvyšování automatizace a zkvalitnění procesů kontrol kvality dat.

2 Literární rešerše

2.1 Problematika detekce duplicit v datech

Data jsou v rámci různých ekosystémů ukládána v různých formách. Liší se ve struktuře, pojmenování atributů a hlavně ve způsobu využívání unikátních identifikátorů datových entit. Ve chvíli, kdy se pokoušíme data z takových různých zdrojů agregovat například do centrální databáze, vzniká problém. Tím je rozhodnout, které záznamy jsou unikátní a které se ve skutečnosti opakují a mají pouze jinou podobu kvůli odlišné struktuře dat nebo způsobu jejich získání.

Nekvalitní data jsou komplikací, která může vést k zásadním chybám v datové analýze a rozhodovacích procesech. Například ve zdravotnictví může mít existence duplicitních záznamů o pacientech vážné důsledky – od nesprávně předepsaných léků až po chybné zdravotní statistiky.(Bess, 2024) V oblasti e-commerce může duplicita zákaznických účtů znamenat špatně cílené marketingové kampaně nebo mylné vyhodnocení chování uživatelů.(Brown, 2019) A v geoprostorových datech mohou duplicity vést k nesprávné identifikaci lokací, chybným navigačním trasám nebo nesrovnalostem v mapových podkladech.

Abychom mohli přesně popsat proces detekce duplicit, je nutné nejprve vyjasnit základní pojmy související s datovými entitami a jejich reprezentací v databázích.

Entita je obecný pojem označující reálný objekt nebo koncept, který má své vlastnosti a může být reprezentován v datech. V závislosti na kontextu může entitou být např. osoba, firma, geografický bod zájmu (POI) nebo administrativní oblast.

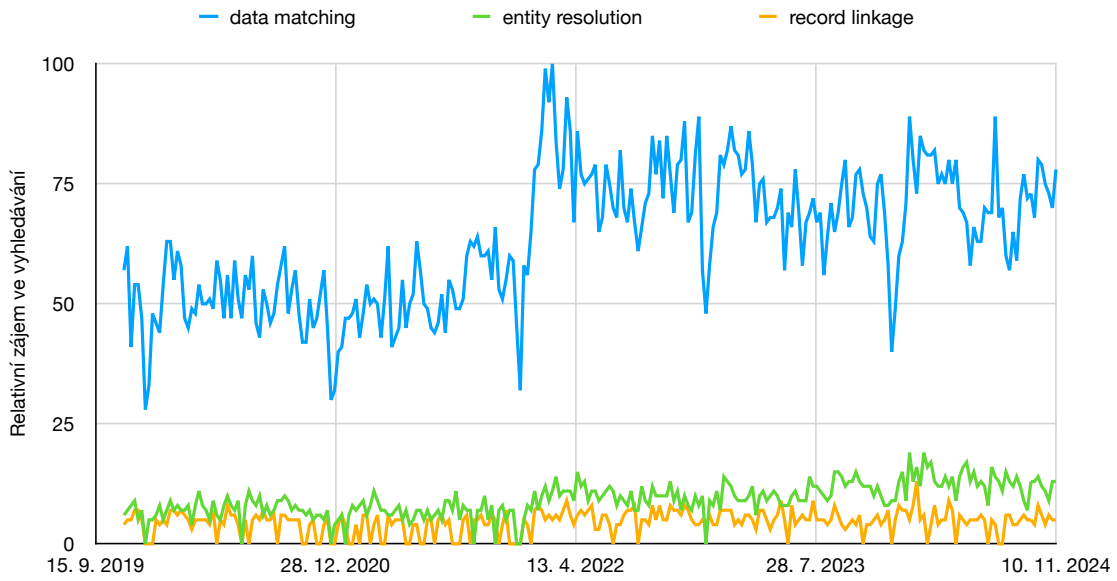
Záznam představuje konkrétní reprezentaci entity v databázi. Jedna entita tak může být v různých databázích reprezentován více různými záznamy, například s mírně odlišnými údaji nebo formátem.

Duplicitní záznamy vznikají tehdy, když databáze obsahuje více reprezentací stejné entity, ať už kvůli chybám v zápisu, rozdílnému formátu dat, nebo rozdílným zdrojům dat. V literatuře se však termíny označující proces identifikace těchto duplicit ne vždy používají jednotně a často se jejich významy překrývají. Zatímco některé zdroje považují následující pojmy za synonyma, jiné mezi nimi rozlišují:

- **Data matching** – proces porovnávání záznamů za účelem nalezení těch, které odpovídají stejné entitě, i když mají různé atributy nebo formáty. (Christen, 2012)
- **Entity resolution** – proces slučování duplicitních záznamů do jednoho sjednoceného záznamu reprezentujícího danou entitu. (Quantexa, 2024)
- **Record linkage** – propojení odpovídajících záznamů napříč různými databázemi, aniž by došlo ke sloučení do jedné reprezentace. (Stepanenko, 2024)

To, že existují různé termíny pro podobné procesy, ukazuje, že problematika detekce duplicit vznikala nezávisle v různých oborech, a proto je důležitější se soustředit na samotné postupy, než na přesné označení. (Christen, 2012)

V této práci bude hlavní pozornost věnována *data matching*, tedy samotnému procesu hledání duplicitních záznamů v datech, spíše než jejich následnému slučování (*entity resolution*). Pro zajímavost, podle Google Trends je termín *data matching* v posledních pěti letech globálně vyhledávanější než ostatní zmiňované termíny. Viz obr. 1.



Obrázek 1: Zájem o termíny ”data matching”, ”entity resolution” a ”record linkage”. *Zdroj*: (Google Trends, 2024)

2.2 Příčiny vzniku duplicit

Zazněli zde již některé důvody vzniku duplicit v datech, pojdmě se blíže podívat na nejčastější z nich. Rozdělil jsem je zde na duplicity v datech, které mohou vzniknout v rámci jednoho datového zdroje, a duplicity vznikající spojením více datových zdrojů.

Duplicity v samostatném zdroji

Přestože by se na první pohled zdálo, že v rámci jednoho datového zdroje šance na vznik duplicity minimiální, je zde i tak několik případů, které vedou ke vzniku duplicit. Tady jsou některé z nich:

- Lidský faktor – při zadávání dat člověkem ručně je velká pravděpodobnost zadání typografických chyb, různých nekonzistentních formátů a nebo přímo neplatných informací. Při zadávání dalších záznamů si pak daný člověk nemusí všimnout již existujícího záznamu vloženého dříve.
- Nedostatečné standardizace formátů – různé formuláře, dotazníky, nebo části/-moduly jednoho systému mohou využívat různé formáty pro datumy, telefonní čísla, nebo rodná čísla a dalších typů dat. Pokud není správně ošetřen převod těchto dat do jednotného formátu před vložením záznamu, může opět docházet ke vzniku duplicit.
- Chybějící integritní omezení databáze – kdyby byly sebevíce ošetřeny předchozí zmíněné způsoby, v případě, že daná databáze nemá nastavená integritní omezení (např. na unikátní hodnoty v rámci rodného čísla, IČO nebo třeba unikátnosti dvou a více atributů v rámci jednoho záznamu), bude každý vložený záznam brán jako nový neexistující záznam, přestože se bude jednat o identickou duplicitu.

Duplicity ve sloučeném zdroji

Při pokusu o spojení více zdrojů do jednoho vzniká několik výzev. V rámci business inteligence se tento proces nazývá ETL[1] v rámci něj může docházet k duplicitám v několika případech:

- Integrace různých systémů – v případě systémů, které mají data o entitách kvůli různým důvodům (např. inventární systém školy ukládá záznamy o vybavení, online bazar má záznamy o prodávaném vybavení), může vzniknout situace, kdy záznamy z více systémů nelze spojit na základě společného unikátního klíče, protože má každý systém svůj unikátní pro danou entitu.
- Průběžné spojování dat – některé datasety jsou pravidelně aktualizované, aby odráželi co nejvíce realitu. Pokud však entita z realného světa v rámci aktuality

[1] Extract, transform, load

změnila zásadně své atributy (např. pekárna se přestěhovala na jinou adresu a přejmenovala se), může vzniknout situace, kdy některý ze zdrojů bude mít stále staré údaje, a jiný nové, a tím pádem nepůjde tento rozdíl spojit s původním záznamem a bude se milně jevit, že jde o novou entitu

- Ještě jeden...

Dopady duplicit na analýzu a rozhodování

2.3 Přístupy k rozpoznání duplicitních entit

Detekce duplicit bez využití strojového učení zahrnuje například algoritmus pro výběr párů Sorted-Neighborhood-Method (SNM). Je možné provést porovnání všech možných kombinací záznamů, ale SNM efektivně snižuje počet nutných porovnání tím, že seřadí data a porovnává pouze sousední záznamy. Pro následné měření podobnosti mezi záznamy se často používají metody jako Jaro-Winkler a Levenshtein, které posuzují, do jaké míry se záznamy podobají. Na základě takové podobnosti lze určit, zda jsou porovnané záznamy duplicity. Tyto metody jsou obzvláště užitečné pro identifikaci menších rozdílů a překlepů ve zpracovávaných datech. (Draisbach a Naumann, 2013)

Metody strojového učení nabízí pokročilejší možnosti, které umožňují identifikovat složitější vzory, které by metody bez využití strojového učení nemuseli odhalit. Například, hluboké neuronové sítě mohou být trénovány na rozsáhlých datových sadách a díky tomu mohou nalézt v datech složité vzory. Takto natrénované neuronové sítě mohou označit za duplicitní záznamy, které na první pohled jako duplicity nevypadají. (Pašek et al., 2022)

2.4 Specifika geoprostorových dat

Jde o data, která představují místa, oblasti nebo třeba předměty ze skutečného světa, a udávají jejich přesnou lokalitu pomocí souřadnic na Zemi. Můžeme mít například geodata představující státní hranice evropských zemí, geodata představující všechny lampy veřejného osvětlení v Brně.

Souřadnicové systémy

Geodata používají souřadnice pro popsání pozice konkrétního objektu. Těchto souřadnic však existuje hned několik a při práci s geodaty je potřeba vědět, které to jsou. Existují různé typy souřadnicových systémů, které se používají podle konkrétních požadavků. Mezi nejpoužívanější v kontextu České republiky patří:

Globální a celosvětové použití

Geografický souřadnicový systém (GCS) - WGS 84

Systém využívající úhlových souřadnic, konkrétně zeměpisné šířky a délky. Nadmořská výška se udává v metrech nad povrchem referenčního elipsoidu. Je vhodný pro geodata představující entity na celosvětové úrovni. Díky tomu je využíván například i v GPS technologii, kde je zapotřebí pracovat v rámci celé Země. Není však příliš vhodná pro lokální úroveň (města, ulice), pokud je vyžadována vysoká přesnost. Je tomu tak protože při výpočtu vzdáleností nebo ploch z úhlových souřadnic může docházet k nepřesnostem.

Regionální a národní použití

Projekční souřadnicový systém (PCS) - S-JTSK (Systém Jednotné Trigonometrické Sítě Katastrální)

Tento konkrétní systém je navržen pro Českou republiku a umožňuje tak na jejím území mapovat s vysokou přesností. Je proto používán jako jeden ze závazných souřadnicových systémů pro orgány České republiky. Souřadnice jsou vyjádřeny pomocí kartézského souřadnicového systému (X - sever, Y - východ) a jednotkou jsou metry. Pro nadmořskou výšku používá metry nad střední hladinou Jaderského moře.

Evropský souřadnicový systém - ETRS89 (European Terrestrial Reference System 1989)

Souřadnicový systém ETRS89 slouží jako standard pro evropské projekty a díky tomu i usnadňuje spolupráci vícero zemí na projektech s mezihraničním rozsahem. Stejně jako S-JTSK používá kartézský souřadnicový systém v jednotkách metru. Výška v tomto systému je počítáná v metrech od povrchu referenčního elipsoidu GRS80.

Toto nejsou zdaleka všechny souřadnicové systémy. Každý z nich slouží pro jiné potřeby a je důležité si uvědomit, jak se liší, a v jakém kontextu se s nimi lze setkat.

K čemu slouží

Geodata sama o sobě ale popisují vždy specifický objekt. Aby se dali z těchto dat získat nějaké znalosti, je potřeba je dále zpracovávat. K tomu slouží například geografické informační systémy (GIS). Tyto systémy dokáží poskytnutá data analyzovat a vizualizovat a dále s nimi pracovat dle potřeby. Mohou tak umožnit například přehledně vidět, v jakém úseku silnice je v daný čas nejvíce aut, a zároveň vidět, ze kterých navazujících silnic tam tyto auto přijíždí. Díky takové vizualizaci a analýze lze následně vycházet například při plánování dopravní infrastruktury.

Každý kdo otevře služby jako Google Maps, Seznam Mapy nebo Apple Maps se dívá na několika vrstvou vizualizaci geodat. Na první pohled se jedná o běžnou

mapu, avšak vizualizace v těchto službách je běžně tvořena několika samostatnými vrstvami:

- vrstvy komunikace,
- vrstvy staveb,
- vrstvy řek,
- vrstvy zeleně a dalších.

Každá taková vrstva je sada geodat, která popisuje daný typ předmětu/lokace s přesnými souřadnicemi dalšími pomocnými atributy/vlastnostmi.

2.5 Existující nástroje

Literatura

Bess, Oleg (12. led. 2024). *The problem with duplicate and mismatched patient records*. Physicians Practice. URL: https://www.physicianspractice.com/view/the-problem-with-duplicate-and-mismatched-patient-records (cit. 09. 02. 2025).

Brown, Forrest (13. dub. 2019). *8 Problems That Result from Data Duplication • Profisee*. Enterprise Master Data Management • Profisee. URL: https://profisee.com/blog/8-business-process-problems-that-result-from-data-duplication/ (cit. 09. 02. 2025).

CleverMaps (2024). *Location Insights for Business*. CleverMaps. URL: https://www.clevermaps.io (cit. 28. 04. 2024).

Draisbach, Uwe a Felix Naumann (2013). *On Choosing Thresholds for Duplicate Detection*. URL: https://hpi.de/oldsite/fileadmin/user_upload/fachgebiete/naumann/publications/PDFs/2013_draisbach_on.pdf.

Google Trends (2024). *Google Trends*. Google Trends. URL: https://trends.google.com/trends/explore?date=today%205-y&q=data%20matching,entity%20resolution,record%20linkage&hl=cs (cit. 14. 11. 2024).

Christen, Peter (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Data-centric systems and applications. Berlin Heidelberg: Springer. 270 s. ISBN: 978-3-642-43001-5 978-3-642--31163-5.

Nauman, Felix a Melanie Herschel (2022). *An Introduction to Duplicate Detection*. Google-Books-ID: DYdyEAAAQBAJ. Springer Nature. 84 s. ISBN: 978-3-031--01835-0.

Pašek, Jan et al. (2022). *MQDD – Pre-training of Multimodal Question Duplicity Detection for Software Engineering Domain*. URL: https://www.researchgate.net/publication/359518098_MQDD_--_Pre-training_of_Multimodal_

Question_Duplicity_Detection_for_Software_Engineering_Domain#fullTextFileContent.

Quantexa (2024). *What is Data Matching & How Does it Work?* Quantexa. URL: https://www.quantexa.com/resources/data-matching/ (cit. 13. 11. 2024).

Stepanenko, Roman (21. ún. 2024). *What is Entity Resolution.* RecordLinker. URL: https://recordlinker.com/what-is-entity-resolution/ (cit. 14. 11. 2024).