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 sektorech podnikání. Jako součást těchto platforem některé společnosti, včetně CleverMaps, nabízí služby zvané jako Data Marketplace, které umožňují uživatelům získávat a integrovat různorodé datové sady. Takový marketplace poskytuje širokou škálu dat. Příkladem takových dat mohou být demografické informace, které klientům umožňují získat přesné informace například o jejich cílové skupině. (CleverMaps, 2024)

Vzhledem k tomu, že při práci (nejen) s geodaty se často zapracovávají informace získané z různých zdrojů, nastává to, že se totožné údaje ve shromážděných datových sadách mohou objevovat vícekrát. Tento jev se označuje jako duplicity. Tyto duplicity mohou nastat z mnoha důvodů a mohou mít různé podoby. Příkladem může být situace, kdy jsou v různých databázích stejná místa nebo objekty zaznamenány pod různými názvy. Například, jedna databáze může obsahovat bod zájmu (point of interest) jako "Velký městský park", zatímco druhá jej může uvádět jako "Park v centru města". Obě tato označení se mohou vztahovat ke stejnému místu, ale různé názvy komplikují jejich rozpoznání jako duplicit. Další běžnou příčinou je chybné zadání dat, kde se například dva různé záznamy pro stejnou ulici mohou lišit pouze malými typografickými chybami nebo odlišnými způsoby zkrácení názvů. (Nauman a Herschel, 2022)

Pro zajištění integrity a přesnosti geodat je detekce a řešení duplicit nezbytná. K identifikaci duplicit mohou být využity různé metody, od jednoduchých porovnání řetězců až po sofistikovanější techniky, jako jsou algoritmy strojového učení a neuronové sítě. Tyto pokročilé metody umožňují rozpoznávat a porovnávat podobnosti na základě kontextu a pravděpodobnosti, což vede k efektivnějšímu a přesnějšímu detekování duplicit. (Christen, 2012)

V oblasti detekce duplicit existuje řada služeb, které nabízejí pokročilé řešení tohoto problému. Tyto služby jsou často založené na cloud technologiích a strojovém učení. Jsou schopné efektivně identifikovat a eliminovat duplicity v rozsáhlých datových sadách. Mezi poskytovatele těchto služeb patří například služby jako Data Ladder, Tilores nebo Melissa, které nabízejí různé nástroje pro automatickou detekci duplicitních záznamů.

Ačkoliv tyto služby přinášejí významné výhody v podobě úspory času a zlepšení kvality dat, mohou být spojeny s vysokými náklady, zejména v případech, kdy je zapotřebí zpracovat velké objemy dat. Finanční zátěž z těchto služeb může být pro některé společnosti značná, což může motivovat k implementaci vlastního 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

Základní pojmy, metriky, typy algoritmů, problémy s kvalitou dataaasdasdaa

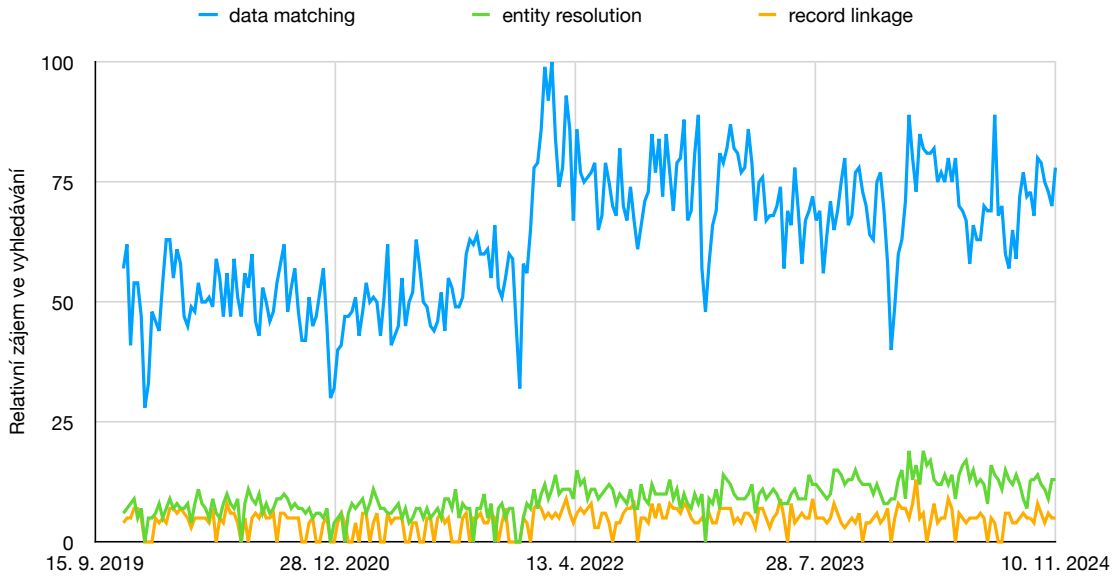
Data v rámci různých ekosystémů jsou ukládáná v různých formách. Liší se ve struktuře, pojmenování atributů a hlavně ve způsobu využívání unikátních identifikátorů datových entit. Ve chvíli kdy se pokoušíme data z takových různých zdrojů agregovat například do centrální databáze, vzniká problém. Tím je rozhodnout, které záznamy (entity) jsou unikátní, a které jsou například duplikované z toho důvodu, že prostředí ze kterých jsou získána mají odlišnou strukturu a nelze snadno rozhodnout o tom, zda je daná entita shodná s entitou z jiného zdroje na základě jejího identifikátoru.

V literatuře se objevuje několik termínů, které označují procesy pro nalezení duplicit a vypořádání se s nimi. Přestože se některé z nich používají zaměnitelně, některé zdroje popisují procesy rozdílně. Zde je stručný přehled.

- *Data matching* je obecný pojem označující různé techniky pro porovnávání a rozpoznávání duplicit mezi různými zdroji dat.(Christen, 2012)
- *Entity resolution* zahrnuje proces nalezení shodných entit s důrazem na spojení těchto entit do jedné entity. (Quantexa, 2024)
- *Record linkage* zahrnuje proces nalezení a propojení entit. Narozdíl od *entity resolution* je však nespojuje do jedné entity, ale vytváří se např. propojovací tabulky. (Stepanenko, 2024)

Nicméně, tyto rozdíly se objevují pouze v některých zdrojích a to svědčí o tom, že většina těchto a další podobných termínů vznikla v různých oborech a v rámci řešení různých problémů a proto není příliš důležité věnovat se označení procesu, ale spíš procesu jako takovému a krokům, které vedou k nalezení duplicit.(Christen, 2012)

V této práci bude používán termín *data matching*, převážně z toho důvodu, že podle Google Trends (Google Trends, 2024), je celosvětově za posledních 5 let termín *data matching* relativně vyhledávanější než ostatní alternativní termíny. Viz. obr. 1.



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

Příkladem kdy k takovému problému nedojde může být případ z prostředí České republiky a předpokládejme, že jde o data čistě občanů České republiky, kdy máme první datový zdroj s pacienty ordinace a druhý zdroj je databáze klientů zdravotní pojišťovny. V tomto případě lze snadno rozpoznat dané entity (osoby) na základě rodného čísla, které se u záznamů v těcho zdrojích pravděpodobně nachází. V tomto případě můžeme následně data zagregovat a dále s nimi pracovat.

2.2 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)

Algoritmy a techniky

2.3 Geoprostorová data

Typy geodat(vector, body, linie a rastry), formáty uložení, souřadnicové systémy

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.

Obrázek s GIS mapama

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.

Tvary geodat (points, lines, and polygons)

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.

Obrázek map z aplikací

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

2.4 Existující nástroje

Popište dostupné software, jejich výhody a nevýhody

Literatura

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).