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

Dobrý den, tohle je moje nové TODO. Pěkný, no ne?

Finish the sentence.

(Christen, 2012)

1.2 Cíl

Cílem práce je vymyslet něco ultra dobrýho

```
1  import numpy as np
2
3  def incmatrix(genl1,genl2):
4      m = len(genl1)
5      n = len(genl2)
6      M = None #to become the incidence matrix
7      VT = np.zeros((n*m,1), int)  #dummy variable
8      a = "Asdasdasasdasdas"
9      #compute the bitwise xor matrix
10     M1 = bitxormatrix(genl1)
11     M2 = np.triu(bitxormatrix(genl2),1)
12
13     for i in range(m-1):
14         for j in range(i+1, m):
15             [r,c] = np.where(M2 == M1[i,j])
16             for k in range(len(r)):
17                 VT[(i)*n + r[k]] = 1;
18                 VT[(i)*n + c[k]] = 1;
19                 VT[(j)*n + r[k]] = 1;
20                 VT[(j)*n + c[k]] = 1;
21
22                 if M is None:
23                     M = np.copy(VT)
24                 else:
25                     M = np.concatenate((M, VT), 1)
26
27                 VT = np.zeros((n*m,1), int)
28
29     return M
```

2 Literární rešerše

2.1 Problematika detekce duplicit v datech

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

2.2 Přístupy k rozpoznání duplicitních objektů

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

3 Literatura

CHRISTEN, PETER *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Berlin Heidelberg, 2012. 270 s. ISBN 978-3-642-43001-5 978-3-642-31163-5..

HUYEN, CHIP *Designing machine learning systems: an iterative process for production-ready applications*. Beijing Boston Farnham Sebastopol Tokyo, 2022. 367 s. ISBN 978-1-09-810796-3..

MCCLAIN, BONNY P. *Python for geospatial data analysis: theory, tools, and practice for location intelligence*. Beijing Boston Farnham Sebastopol Tokyo, 2023. 262 s. ISBN 978-1-09-810479-5..

MCGREGOR, SUSAN E. *Practical Python data wrangling and data quality*. Sebastopol, CA, 2022. 395 s. ISBN 978-1-4920-9150-9..

NAUMAN, FELIX; HERSCHEL, MELANIE *An Introduction to Duplicate Detection*. , 2022. 84 s. ISBN 978-3-031-01835-0..

WITTEN, IAN H.; FRANK, EIBE; HALL, MARK A.; PAL, CHRISTOPHER J. *Data mining: practical machine learning tools and techniques*. Amsterdam Boston Heidelberg London New York Oxford Paris San Diego San Francisco Singapore Sydney Tokyo, 2017. 621 s. ISBN 978-0-12-804291-5..

SCIKIT-LEARN DEVELOPERS *scikit-learn: machine learning in Python — scikit-learn 1.4.2 documentation* [online]. 2024 [cit. 2024-04-30]. Dostupné z `https://scikit-learn.org/stable/index.html`.