Lietuviškų adresų geokodavimo problemos ir jų sprendimo būdai

Viktoras Paliulionis

Matematikos ir informatikos instituto mokslo darbuotojas, daktaras
Institute of Mathematics and Informatics, Researcher, PhD
Akademijos g. 4-327, LT-08663 Vilnius
Tel. (8 5) 210 93 40
El. paštas: vikpal@ktl.mii.lt

Geokodavimas yra procesas, kai tekstinis vietos aprašas transformuojamas į geografines koordinates. Vienas iš dažniausiai naudojamų vietos aprašymo būdų yra pašto adresas, kurį sudaro gyvenvietės pavadinimas, gatvės pavadinimas, namo numeris ir kiti adreso elementai. Šiame straipsnyje nagrinėjamos lietuviškų adresų geokodavimo problemos, atsirandančios dėl adreso formatų įvairovės, netiksliai ir su rašybos klaidomis užrašomų adresų. Straipsnyje aprašyti geokodavimo proceso etapai ir juose naudojamų algoritmų principai. Pasiūlytas lietuvių kalbai pritaikytas LT-Soundex algoritmas, leidžiantis indeksuoti adreso elementus pagal fonetinį panašumą ir atlikti apytikslę paiešką.

Įvadas

Geokodavimas yra tekstinio vietos aprašo transformavimo į geografines koordinates procesas. Vietą aprašyti galima skirtingais būdais, pavyzdžiui, nurodant gatvę ir pastato numerį (pašto adresą), gatvių susikirtimą, pastato pavadinimą, vietovardį. Nustatyti vietą kartais įmanoma net pagal vietinio telefono numerį. Tačiau dažniausiai vieta identifikuojama adresu, sudarytu iš gyvenamosios vietovės pavadinimo, gatvės pavadinimo, pastato numerio ir kitų adreso elementų. Adresai naudojami įvairiuose informacijos šaltiniuose, duomenų bazėse, elektroniniuose dokumentuose, interneto tinklalapiuose. Geokodavimo metu gauta geografinė informacija leidžia adresu aprašomus objektus pavaizduoti žemėlapyje, atlikti jų erdvinę analizę (Goldberg, Wilson, Knoblock, 2007).

Geokodavimas yra viena iš pagrindinių GIS funkcijų, jis plačiai naudojamas navigacinėse sistemose adresų paieškai. Kai kurios geokodavimo paslaugos yra teikiamos internetu (pavyzdžiui, *Google Maps* geokodavimo paslauga). Tačiau dažnai jos turi ribotas galimybes, nes reikalauja pateikti duomenis tik tam tikru formatu ir ne visada atpažįsta Lietuvoje naudojamus adresų formatus. Elektroniniuose šaltiniuose adresas dažnai nurodomas ne visas, su rašybos klaidomis arba nekorektiškai suformatuotas. Tai labai sunkina geokodavimo procesą.

Šiame straipsnyje nagrinėjamos lietuviškų adresų geokodavimo problemos, atsirandančios dėl adreso formatų įvairovės, netiksliai ir su rašybos klaidomis užrašomų adresų. Aprašyti geokodavimo proceso etapai ir juose naudojamų algoritmų principai. Straipsnyje pasiūlytas lietuvių kalbai pritaikytas *LT-Soundex* algoritmas, leidžiantis indeksuoti adreso elementus (pavyzdžiui, gatvių, gyvenviečių pavadinimus) pagal fonetinį panašumą ir atlikti apytikslę paiešką.

Pašto adreso formatai

Norint tiksliai geokoduoti adresą, svarbu žinoti adreso struktūrą, t. y. kokie elementai sudaro adresą, kokia jų užrašymo tvarka. Skirtingų šalių adresų sistemos turi savo ypatumų, tačiau

dažniausiai pašto adresą sudaro tos šalies hierarchine administracine struktūra grindžiami elementai, pavyzdžiui: šalis, regionas (valstija, rajonas), gyvenvietė, gatvė, pastato numeris (Davis, Fonseca, 2007). Tarp gyvenvietės ir šalies pavadinimų paprastai nurodomi vienas arba keli tarpinių lygių administraciniai vienetai. Adrese nurodytas pašto indeksas taip pat identifikuoja adresato vietą ir padeda išvengti skirtingų adreso elementų interpretacijų.

Lietuvoje galioja mažiausiai du adresų užrašymo standartai, o iš tiesų naudojama dar daugiau. Jie skiriasi adresą sudarančiais elementais ir jų užrašymo tvarka. Lietuvos Respublikos Vyriausybės patvirtintos *Adresų formavimo taisyklės* (Žin., 2002, Nr. 127-5753) numato tokį adreso sudedamųjų dalių išdėstymą:

1) savivaldybė (išskyrus atvejus, kai adreso objektas yra savivaldybės centre);
2) seniūnija (išskyrus miestų seniūnijas);
3) gyvenamoji vietovė (miestas, miestelis, kaimas ir kt.);
4) gatvė (išskyrus gyvenamąsias vietoves, kuriose gatvės nesudaromos);
5) žemės sklypo, pastato ar pastatų komplekso numeris gatvėje ar gyvenamojoje vietovėje.

Taip pat adrese gali būti nurodytas buto numeris ir kiti adreso elementai, kurie adreso vietai nustatyti nėra svarbūs, bet geokodavimo metu juos taip pat reikia atpažinti. Šios adresų formavimo taisyklės yra privalomos valstybės ir savivaldybių informacinėms sistemoms, kur reikia nurodyti fizinio asmens adresą arba juridinio asmens buveinę, tačiau kitoms informacinėms sistemoms ir rašytiniams bendriesiems dokumentams šių taisyklių taikymas yra rekomendacinio pobūdžio.

Pagal Lietuvos pašto taisykles (www.post.lt) adresuojant laišką ar siuntą savo šalyje adresą reikia rašyti taip:

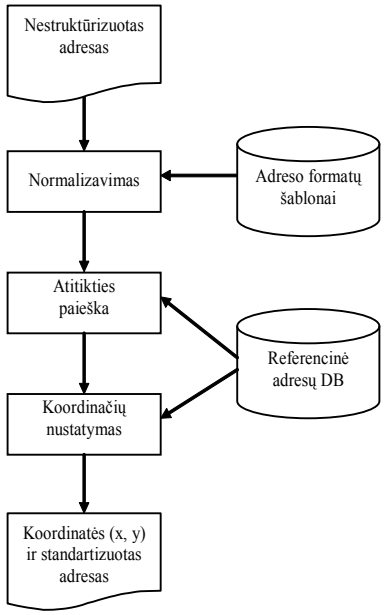
- siunčiant į savivaldybės centrą: gatvė, pastato numeris, pašto indeksas, miesto pavadinimas;
- siunčiant į gyvenamąją vietovę, kuri nėra savivaldybės centras, bet ten yra paštas ir jo pavadinimas sutampa su gyvenamosios vietovės pavadinimu: gatvė, pastato numeris, gyvenamoji vietovė, pašto indeksas, savivaldybė;
- siunčiant į gyvenamąją vietovę, kurioje nėra pašto: [gatvė, pastato numeris,] gyvenamoji vietovė, pašto pavadinimas, pašto indeksas, savivaldybė (gatvių gali nebūti).

Siunčiant pašto siuntą į užsienį siuntėjo adreso gale papildomai nurodomas šalies pavadinimas anglų arba prancūzų kalbomis (LITHUANIA arba LITUANIE). Kaip matome, Lietuvos pašto nustatytas adreso formatas skiriasi nuo Vyriausybės patvirtinto adreso formato sudedamųjų dalių rašymo tvarka ir tuo, kad pagal pašto taisykles papildomai nurodomas pašto indeksas, o vietoj seniūnijos nurodomas pašto pavadinimas. Dar kitokie adreso formatai naudojami informaciniuose kataloguose. Papildomos problemos iškyla adresuojant objektus kaimo vietovėse, kuriose nėra gatvių ir adresų numeracijos (Čypas, 2001).

Minėti adreso formatai gana painūs, nes priklauso nuo to, ar adreso objektas yra savivaldybės centre, ar tai miesto vietovė, ar kaimo, ar turi paštą. Todėl jiems sudėtinga pritaikyti tarptautinius adresavimo standartus. Pavyzdžiui, *Google Maps* geokodavimo sistema lietuviškuose adresuose gyvenvietėms identifikuoti naudoja keturis hierarchijos lygius: *Country*, *SubAdministrativeArea*, *Locality* ir *DependentLocalityName*. *Country* žymi šalį, *SubadministrativeArea* – apskritį, *Locality* vienais atvejais savivaldybę, kitais – miestą (jei sutampa savivaldybės ir savivaldybės centro pavadinimas), *DependentLocalityName* – gyvenvietės pavadinimą arba miesto seniūnijos pavadinimą (priklausomai nuo to, ką žymi *Locality*). Formuodama iš sudėtinių adreso dalių visą adresą, *Google Maps* sistema neįtraukia *DependentLocalityName* komponento, kadangi pagal prasmę tai turėtų būti tik gyvenamosios vietovės dalis, todėl visas adresas kartais nurodomas nekorektiškai. Pavyzdžiui, pateikus *Google Maps* užklausą „Troškūnai, Vytauto 15“, vieta parodoma teisingai, o visas adresas pagal *Google Maps* standartą nurodomas nekorektiškai – „15 Draugystės gatvė, Anykščiai 29033, Lithuania“. Kaip matome, vietoje „Troškūnai, Anykščių r. sav.“ nurodyta „Anykščiai“.

Geokodavimo procesas

Geokodavimo metu pagal laisvo formato vietos aprašą (adresą) yra nustatoma standartizuota adreso forma ir vietos geografinės koordinatės, kurias paskui galima panaudoti vietos vaizdavimui žemėlapyje arba tolesnei analizei. Galima išskirti tris geokodavimo proceso etapus: pradinio adreso normalizavimas, atitikties paieška ir koordinačių nustatymas (žr. pav.).



P a v. ***Apibendrinta geokodavimo proceso schema***

Normalizavimo etapas. Pirmajame geokodavimo etape pradinis adreso tekstas suskaidomas į adreso komponentus (t. y. normalizuojamas). Tam naudojami adreso formatų šablonai ir atpažinimo taisyklės. Normalizavimo metu tekstas iš pradžių suskirstomas į leksemas, o paskui bandoma nustatyti, kokiam adreso atributui priskirti kiekvieną leksemą. Adreso elementams atpažinti naudojami reikšminiai žodžiai, žymintys gatvės tipą (pavyzdžiui *gatvė*, *aikštė*, *prospektas*, jų trumpiniai *g*., *a.*, *pr*. ir kt.), gyvenvietės tipą (pavyzdžiui *m*., *mstl*., *k.*), pašto povardį (*paštas*, *pšt*.), seniūnijos povardį (*sen*.), savivaldybės povardį (*sav*.), apskrities povardį (*apskr*.) ir kt. Pirmoje lentelėje pateiktas suskaidyto į komponentus adreso „Vytauto g. 15, Troškūnų m., Anykščių r. sav.“ pavyzdys. Normalizavimas sudėtingėja, kai yra praleisti reikšminiai žodžiai. Tuomet jis gali būti nevienareikšmis. Tokiais atvejais antrajam etapui bus perduoti visi galimi normalizavimo variantai.

1 l e n t e l ė. ***Normalizuoto adreso pavyzdys***

Šalis	
Apskritis	
Savivaldybė	Anykščių r. sav.
Seniūnija	
Paštas	
Pašto kodas	
Gyvenvietės pavadinimas	Troškūnų
Gyvenvietės tipas	m.
Gatvės pavadinimas	Vytauto
Gatvės tipas	g.
Pastato numeris	15

Atitikties paieškos etapas. Antrajame geokodavimo etape normalizuotam adresui ieškoma atitikties referencinėje adresų duomenų bazėje, sutapdinant atskirus adreso komponentus. Adreso formato šablonai apima įvairius užklausose pasitaikančius adreso elementus, net ir tuos, kurių nėra naudojamoje adresų duomenų bazėje. Jei ieškomas adresas turi tokių perteklinių elementų, tai jie ignoruojami (pavyzdžiui, pašto pavadinimas). Vienareikšmiškam gatvės radimui dažniausiai pakanka gatvės, gyvenvietės ir savivaldybės pavadinimo, o kartais ir savivaldybės pavadinimas nėra būtinas (pavyzdžiui, kai gatvės ieškoma mieste). Pertekliniai elementai gali būti pravartūs tikrinant, ar gautas rezultatas teisingas.

Problemų ieškant atitikčių atsiranda, kai nurodytas ne visas adresas arba adresas yra su rašybos klaidomis. Pirmuoju atveju galime gauti daugiau negu vieną rezultatą, tenkinantį užduotas sąlygas, o antruoju atveju – negauti jokio rezultato.

Esant daugiaprasmiškumui, į rezultatą galima įtraukti visus įmanomus variantus (kad vartotojas galėtų patikslinti adresą) arba suteikti galimiems variantams prioritetus ir pateikti didžiausią prioritetą turintį variantą. Pavyzdžiui,

jei užklausą tenkina dvi gatvės skirtingose gyvenvietėse, be to, viena gatvė yra mieste, o kita kaime, tai aukštesnį prioritetą galima suteikti mieste esančiai gatvei.

Užrašant adresus dažnai daromos tokios rašybos klaidos:

1) įvedama tik adreso elemento (gatvės pavadinimo, gyvenvietės pavadinimo ar kt.) pradžia,
2) lietuviškos raidės nurodomos be diakritinių ženklų,
3) kitos rašybos klaidos (pvz., praleista arba sukeista raidė ir kt.).

Iš pradžių lyginame raides su diakritiniais ženklais konvertavę į atitinkamas raides be diakritinių ženklų ir ieškome atskirų adreso elementų pagal žodžio pradžią. Neradę tinkamo atitikmens pagal tikslią paiešką, paieškai pagal klaidingai užrašytą žodį naudosime apytikslės paieškos algoritmą, kuris bus aptartas kitame skyrelyje.

Koordinačių nustatymo etapas. Geokodavimo rezultatas priklauso ne tik nuo naudojamų geokodavimo algoritmų, bet ir nuo naudojamos adresų duomenų bazės, kurioje saugomos adresų taškų koordinatės, tikslumo ir išsamumo. Tačiau net ir neradus gatvės pastato numerio kaip rezultatą galima pateikti artimiausio pastato numerio koordinates arba interpoliuoti dviejų artimiausių adresų koordinates reikiamoje gatvės pusėje. Jei adresų duomenų bazėje pastatų numerių nėra visai, tai galima adresui priskirti kurio nors gatvės taško koordinates. Geokodavimo paslaugos rezultatai pateikiami XML, KML, CSV, JSON arba kuriuo nors kitu formatu.

Apytikslės paieškos algoritmas

Paieškos pagal klaidingai užrašytą žodį algoritmų klasės yra dvi: vieni algoritmai nagrinėja žodžių panašumą pagal rašybą, kiti – pagal tarimą.

Pirmojo tipo algoritmo pavyzdys yra Levenšteino atstumo skaičiavimas (Levenshtein, 1965). *Levenšteino atstumu* **d(u, v)** tarp dviejų eilučių **u** ir **v** vadinamas minimalus redagavimo operacijų skaičius, reikalingas tam, kad pervestume **u** į **v**. Redagavimo operacijomis laikomi ženklo įterpimas, pašalinimas ir pakeitimas. Pavyzdžiui, Levenšteino atstumas tarp žodžių „vilnius“ ir „vilkas“ yra trys (du pakeitimai: *n* į *k*, *i* į *a* ir vienas raidės *u* šalinimas). Apskaičiuoti Levenšteino atstumą galima naudojant dinaminio programavimo metodą. Esant didelei duomenų bazei šis algoritmas nėra efektyvus.

Žodžių panašumui pagal tarimą nustatyti naudojami vadinamieji *fonetiniai algoritmai*. Dauguma žinomų fonetinių algoritmų yra tinkami tik anglų kalbai. Plačiausiai naudojamas *Soundex* algoritmas, kurį XX a. pradžioje sukūrė ir 1918 m. užpatentavo Robertas C. Russellas ir Margaret K. Odell (Hall, Dowling, 1980). Vėliau šis algoritmas buvo šiek tiek modifikuotas ir daugiausia buvo naudojamas JAV gyventojų surašymo duomenų bazėse pavardėms indeksuoti. Algoritmas (dar vadinamas *American Soundex*) dėl savo paprastumo iki šiol naudojamas ir geokodavimui. *Soundex* indeksavimo sistemoje kiekvienam žodžiui yra priskiriamas keturių ženklų kodas, sudarytas iš raidės ir trijų skaitmenų. *Soundex* kodas sudaromas pagal tokias taisyklės:

1. *Soundex* kodo pirmas ženklas yra pirma pavadinimo raidė.
2. Kiti trys skaitmenys nustatomi žodžio raides pakeičiant *Soundex* kodo skaitmenimis nuo 1 iki 6 (žr. 2 lentelę). Visos balsės ir priebalsės H ir W nekoduojamos.
3. Jei žodyje dvi gretimos raidės turi tą patį *Soundex* raktinį kodą arba jeigu jos atskirtos raidėmis H arba W, tai pakeičiamos vienu kodo skaitmeniu.
4. Jei gautą kodą sudaro daugiau negu keturi ženklai, imami tik pirmi keturi. Jei kodą sudaro mažiau negu keturi ženklai, tai gale pridedamas reikiamas nulių skaičius.

Soundex kodavimo sistema netinka lietuvių kalbai, nes pagal šią sistemą priebalsė H ir specifinės lietuviškos priebalsės (Č, Š, Ž) neturėtų kodo, o dviraidžių CH, DZ, DŽ kodavimas neatspindėtų jų tarimo. Yra pasiūlyta ir dau-

2 lentelė. ***Soundex kodavimo lentelė***

Raidės	Kodo skaitmuo
A, E, I, O, U, H, W, Y	Nekoduojama
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

giau *Soundex* algoritmo variantų, pavyzdžiui, *Daitch-Mokotoff Soundex*, *Metaphone*, NYSIIS (Rajkovic, Jankovic, 2007), tačiau nė vienas nėra pritaikytas lietuvių kalbai. Mes siūlome lietuvių kalbai pritaikytą fonetinį algoritmą *LT-Soundex*. Šiame algoritme naudojami kodo skaitmenys nuo 1 iki 8, visos balsės ir priebalsė J ignoruojamos (žr. 3 lentelę). Be to, įvedėme papildomas taisykles, skirtas pirmai raidei koduoti:

1. Raidės su diakritiniais ženklais (Ą, Ę, Ė, Į, Ų, Ū, Č, Š, Ž) keičiamos į atitinkamas raides be diakritinių ženklų.
2. Keičiama CH → H; DZ, DŽ → Z; Y, J → I.

3 lentelė. ***LT-Soundex kodavimo lentelė (pasiūlymas)***

Raidės	**Kodo skaitmuo**
A, Ą, E, Ę, Ė, I, Į, Y, O, U, Ų, Ū, J	Nekoduojama
B, P	1
C, Č, S, Š, Z, Ž, DZ, DŽ	2
D, T	3
L	4
M, N	5
R	6
CH, G, H, K	7
F, V	8

Mes siūlome tokį lietuviško adreso elemento apytikslės paieškos algoritmą, derinantį *LT-Soundex* algoritmą ir Levenšteino atstumo algoritmą:

1. Turėdami netiksliai užrašytą adreso elementą, apskaičiuojame jo *LT-Soundex* kodą.
2. Vietovardžių duomenų bazėje, indeksuotoje pagal *LT-Soundex* kodą, randame pavadinimus su tuo pačiu *LT-Soundex* kodu. Tai bus panašūs pagal tarimą pavadinimai, kurie sudarys pradinių paieškos kandidatų aibę. Tačiau pavadinimų su tuo pačiu kodu gali būti gana daug ir ne visi jie yra tinkami.
3. Apskaičiuojame Levenšteino atstumą tarp netiksliai užrašyto adreso elemento ir kiekvieno vietovardžio iš antrame žingsnyje gautos aibės.
4. Randame vietovardį su mažiausiu Levenšteino atstumu. Jis ir bus apytikslės paieškos rezultatas. Jei tokių yra keli, tai pateikiami visi.

Pavyzdžiui, ieškodami pagal pavadinimą su rašybos klaidomis „Pan**ia**vė**š**ys“, pirmiausia apskaičiuojame jo *LT-Soundex* kodą. Jis yra P582. Tą patį kodą turi vietovardžiai „Panevėžys“, „Panevėžiukas“, „Panevėžė“ ir „Pamavys“. Jie bus pradiniai paieškos kandidatai. Iš jų randame vietovardį su mažiausiu Levenšteino atstumu nuo „Paniavėšys“. Šiuo atveju tai būtų vietovardis „Panevėžys“.

Išvados

Straipsnyje išnagrinėtos laisva nestruktūrizuota forma užrašytų adresų geokodavimo problemos. Aprašytas geokodavimo procesas, leidžiantis atpažinti Lietuvoje naudojamus įvairius adreso formatus, atsižvelgti į galimus netikslumus ir klaidas užrašant adresus. Pasiūlytas lietuvių kalbai pritaikytas *LT-Soundex* fonetinis algoritmas kartu su Levenšteino atstumo tarp žodžių skaičiavimu leidžia sėkmingai realizuoti apytikslį adreso elementų sutapdinimą.

Ateityje reikėtų atlikti daugiau praktinių eksperimentų, kad būtų ištirtas siūlomų geokodavimo metodų ir *LT-Soundex* algoritmo efektyvumas ir tinkamumas įvairiems adresavimo atvejams.

LITERATŪRA

ČYPAS, K. (2001). Adresai geoinformacinėse duomenų bazėse gyvenamajai vietai nustatyti. *Geodezija ir kartografija*, t. XXVII, Nr. 2, p. 81–86.

DAVIS, C.; and FONSECA, F. (2007). Assessing the Certainty of Locations Produced by an Address Geocoding System. *GeoInformatica*, vol. 11(1), p. 103–129.

GOLDBERG, D. W.; WILSON, J. P.; and KNOBLOCK, C. A. (2007). From Text to Geographic Coordinates: The Current State of Geocoding. *URISA Journal*, vol. 19(1), p. 33–46.

HALL, P.; and DOWLING, G. (1980). Approximate string matching. *Computing Surveys*, vol. 12(4), p. 381–402.

LEVENSHTEIN, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals (in russian). *Doklady Akademii Nauk SSSR*, vol. 4(163), p. 845–848.

NICOARA, G. (2005). Exploring the geocoding process: a municipal case study using crime data. Master's thesis, University of Texas at Dallas, Dallas, TX.

RAJKOVIC, P.; JANKOVIC, D. (2007). Adaptation and Application of Daitch-Mokotoff Soundex Algorithm on Serbian Names. Iš *XVII Conference on Applied Mathematics*, Kragujevac, September, p. 193–204.

LITHUANIAN ADDRESS GEOCODING: PROBLEMS AND SOLUTIONS

Viktoras Paliulionis

S u m m a r y

Geocoding is the process of converting of a textual description of a location into geographic coordinates. One of the most frequently used way to describe a place is its postal address that contains a city name, street name, house number and other address components. The paper deals with the problems of the geocoding of Lithuanian addresses. The main problems are variety of used address formats and possible typing and spelling errors. The paper describes the steps of the geocoding process and used algorithms. We propose a phonetic algorithm called LT-Soundex, adapted for the Lithuanian language and enabling to index addresses components by phonetic similarity and perform approximate address searching. It is used with Levenshtein distance for effective approximate address searching.