

РОЗРОБЛЕННЯ ПРОСТОРОВО-ЗВАЖЕНОГО КРИТЕРІЮ ГЕОПРИВ'ЯЗКИ ПРИРОДОМОВНИХ ТЕКСТІВ ДО МАСИВІВ ВОД

¹Вінницький національний технічний університет, Україна

Анотація

У роботі розглядається проблема автоматичної геоприв'язки природомовних текстів до масивів вод. Попередні дослідження показали необхідність враховувати просторовий аспект географічних назв для покращення точності роботи. З цією метою розроблено новий просторово-зважений критерій на основі адаптації методу TF-IDF до просторової прив'язки даних (Multiplicity-Length Term with Inverse Spatial Object Frequency). Таким чином поєднуємо оцінку унікальності просторових геоназв та їх просторову протяжність з конкретними водними масивами. За рахунок комплексного аналізу суттєво підвищується точність процесу автоматичної прив'язки.

Ключові слова: геоприв'язка текстів, масиви вод, просторовий аналіз, TF-IDF, штучний інтелект

Abstract

The paper considers the problem of automatic georeferencing of natural language texts to water bodies. Previous studies have shown the need to take into account the spatial aspect of geographical names to improve the accuracy of the work. For this purpose, a new spatially weighted criterion has been developed based on the adaptation of the TF-IDF method to spatial data referencing (Multiplicity-Length Term with Inverse Spatial Object Frequency). Thus, we combine the assessment of the uniqueness of spatial geonames and their spatial extent with specific water bodies. Due to the complex analysis, the accuracy of the automatic referencing process is significantly increased.

Keywords: text georeferencing, water bodies, spatial analysis, TF-IDF, artificial intelligence.

Вступ

У попередніх дослідженнях [1, 2, 3] розглянуто проблему інтелектуальної автоматизації прив'язки довільних текстів до масивів вод (МВ) на прикладі р. Південний Буг. Хоча такі методи демонструють ефективність, у них є певні недоліки. Наприклад, як cosine_similarity, так і Jaccard-міра, не розглядають питання проєкції географічних назв на лінійну вісь водних масивів (на усереднену течію основної річки), оскільки всі згадки топонімів прирівнюються між собою, що призводить до нехтування фактичною протяжністю цих об'єктів, через що підвищується ризик топономічної полісемії.

Метою дослідження є підвищення точності просторової прив'язки природомовних текстів до масивів вод, через адаптацію одного з найбільш ефективних методів семантичної прив'язки TF-IDF до задачі просторової прив'язки текстів. Ми пропонуємо модифікований критерій Multiplicity-Length Term with Inverse Spatial Object Frequency (далі – MLT-ISOF), що оснований на урахуванні довжини проєкцій геоназв на лінійні осі масивів вод (усереднену течію річки).

Методика

Для реалізації запропонованого критерію встановимо наступні вхідні дані:

- $M = \{M_i\}$ – множина масивів вод із довжинами лінійних осей L_i ;
- $G = \{G_j\}$ – множина географічних назв із довжинами проєкцій $L_{g(i,j)}$ на лінійні осі відповідних M_i ;

- $L_{g(i,j)}$ - відображає, яку частину масиву вод M_i охоплює геоназва, та чи охоплює довшу відстань, ніж має лінійна вісь заданого масиву вод;
- $T = \{T_k\}$ - множина довільних текстів, що містять слова з G .

Для адаптації TF-IDF вводяться дві складові – Multiplicity-Length Term (1), де замість частоти терміна у документі використовується просторова кратність довжини топоніма до лінійної осі масиву вод, а замість кількості документів Inverse Spatial Object Frequency представляє кількість просторових об'єктів, на яких зустрічається топонім, що враховує просторові характеристики.

Для кожної пари «масив – геоназва» (M_i, G_j) визначаємо MLT через r :

$$MLT_{i,j} = \begin{cases} r \times \frac{L_i}{L_{g(i,j)}}, & \text{якщо } L_{g(i,j)} > L_i, \\ r, & \text{якщо } L_{g(i,j)} \leq L_i, \end{cases}$$

де r – кількість входжень G_j у текст, L_i це – довжина лінійної осі масиву M_i , а $L_{g(i,j)}$ – довжина проєкції G_j на лінійну вісь масиву M_i .

Отже, за умови, що проєкція топоніму більш протяжна, ніж лінійна вісь масиву вод, тоді її вага множиться на коефіцієнт співвідношення довжин, інакше вагою є кількість згадувань геоназви.

ISOF це – просторовий аналог класичного IDF, який відображає унікальність G_j :

$$ISOF_j = \log \frac{N}{n_j},$$

де N – загальна кількість масивів вод у річці, а n_j – к-ть масивів, до яких належить географічна назва G_j . Таким чином, поширені топоніми, що присутні у багатьох масивах (наприклад, «водосховище»), мають меншу вагу, а вищі значення $ISOF_j$ вказують на рідкісніші топоніми.

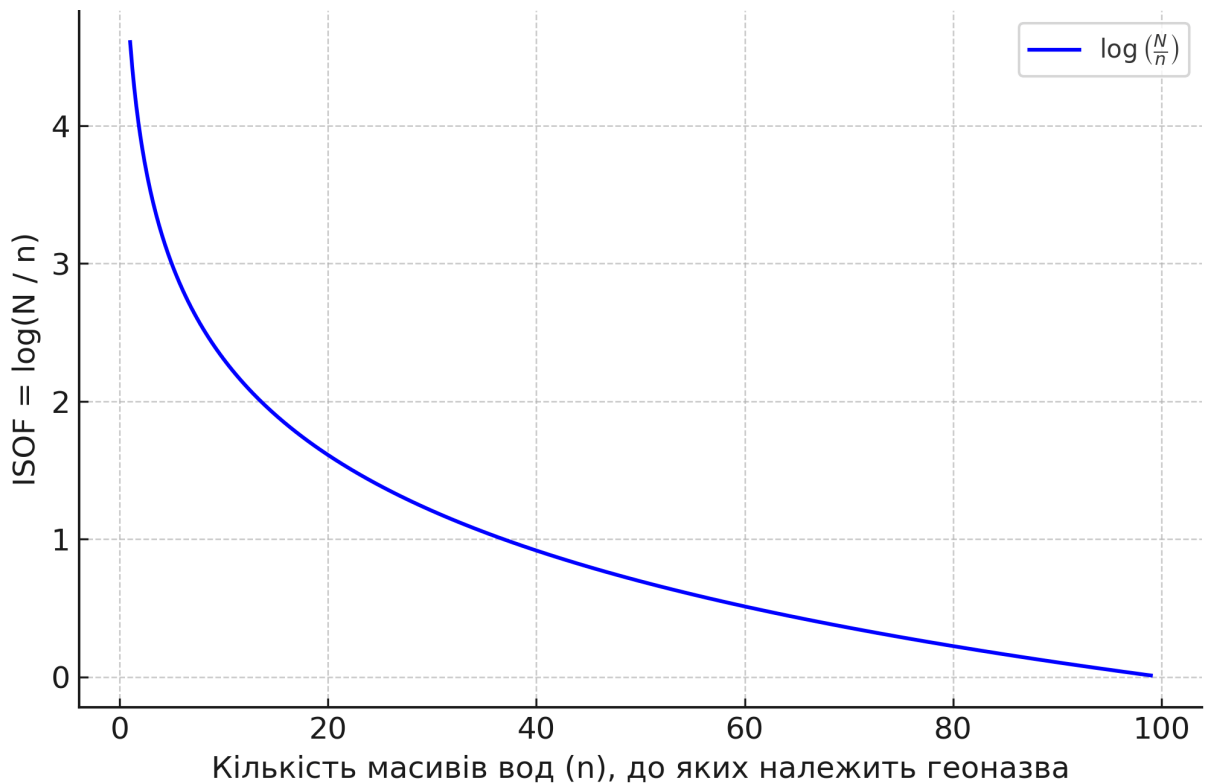


Рис.1 – Залежність $ISOF$ від к-ті МВ, до яких належить геоназва

Визначимо зв'язок між текстом та масивом вод, обчисливши інтегральний показник, що і є аналогом TF-IDF для просторової задачі:

$$MLT - ISOF_{i,j} = MLT_{i,j} \times ISOF_j.$$

Алгоритм застосування

1. $G^T = \{G_j | G_j \in T\}, \quad G_j \subseteq G,$
2. $MLT - ISOF_{i,j} = \begin{cases} MLT_{i,j} \times ISOF_j, & G_j \in M_i, \\ 0, & G_j \notin M_i, \end{cases}$
3. $S_i = \sum_{j \in G^T} MLT - ISOF_{i,j},$
4. $M^* = \{M_i | S_i = \max(S_1, S_2, \dots, S_n)\},$

де M^* - множина масивів вод з максимальною оцінкою релевантності, до яких прив'язується текст новин.

Висновки

В даному дослідженні запропоновано новий критерій MLT-ISOF для покращення точності прив'язки природомовних текстів до масивів вод, завдяки використанню просторових параметрів та врахування фактичної протяжності лінійних проєкцій геоназв на усереднену течію річки.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Мокін В. Б., Бондалетов К. О., Крижановський Є. М., і Караваєв В. О. Метод аугментації текстів про стан масивів вод на основі інтелектуальної прив'язки до багатозв'язних геоінформаційних систем іменованих сутностей, Вісник Вінницького політехнічного інституту, вип. 3, с. 55–65, Черв. 2023. URL: <https://doi.org/10.31649/1997-9266-2023-168-3-55-65>.
2. Мокін В.Б., Бондалетов К.О. Інтелектуальні методи видобування ключових словосполучень із тексту для побудови онтологічних моделей інформаційно-пошукових систем. С. 197-199. – Режим доступу: https://itgip.org/wp-content/uploads/2022/11/tezy_70.pdf. Інформаційно-комунікаційні технології та сталий розвиток: колективна монографія за матеріалами XXI Міжнародної науково-практичної конференції (14-16 листопада 2022 р.). Київ, 2022. ТОВ «Видавництво «стон», 2022, 242. URL: <https://itgip.org/kolektivna-monografiya-za-materialam>
3. Бондалетов, Костянтин, AND Мокін, Віталій. "Інтелектуальна автоматизація геоприв'язки повідомлень з соцмереж до масивів вод за допомогою зваженої JACCARD-міри" ВНТКП ВНТУ. Факультет інтелектуальних інформаційних технологій та автоматизації (2025): н. стор. Веб-ресурс. 13 Jun. 2025

Бондалетов Костянтин Олегович — аспірант кафедри системного аналізу та інформаційних технологій ВНТУ; e-mail: bondaletov.k@gmail.com;

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри системного аналізу та інформаційних технологій ВНТУ; e-mail: vbmokin@vntu.edu.ua;

Bondaletov Kostiantyn O. — Post-graduate student of the Chair of System Analysis and Information Technologies VNTU, e-mail: bondaletov.k@gmail.com;

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technologies VNTU, e-mail: vbmokin@vntu.edu.ua;