

ІНТЕЛЕКТУАЛЬНА АВТОМАТИЗАЦІЯ ГЕОПРИВ'ЯЗКИ ПОВІДОМЛЕНЬ З СОЦМЕРЕЖ ДО МАСИВІВ ВОД ЗА ДОПОМОГОЮ ЗВАЖЕНОЇ JACCARD-МІРИ

¹Вінницький національний технічний університет, Україна

Abstract

Робота присвячена підвищенню точності автоматизованої геоприв'язки текстів новин із соцмереж та ЗМІ про конкретні локальні об'єкти до описів, що характеризують масиви вод заданої великої річки. Запропоновано алгоритм на основі зваженої Jaccard-міри, який дозволяє точно геоприв'язати такі тексти. Випробування дозволили досягти максимальної точності 1.0, що є суттєвим підвищенням у порівнянні з метрикою cosine_similarity. Це відкриває можливості для автоматизованого оброблення даних та підвищення ефективності моніторингу та управління водними ресурсами.

Ключові слова: оброблення природномовного тексту, зважена Jaccard-міра, геоприв'язування тексту, україномовний природномовний текст, штучний інтелект, масив вод, р. Південний Буг.

Abstract

The work is devoted to increasing the accuracy of automated georeferencing of news texts from social networks and mass media about specific local objects to descriptions characterizing the water bodies of a given large river. An algorithm based on the weighted Jaccard measure is proposed, which allows for accurate georeferencing of such texts. The tests allowed achieving a maximum accuracy of 1.0, which is a significant increase compared to the cosine_similarity metric. This opens up opportunities for automated data processing and increasing the efficiency of monitoring and management of water resources.

Keywords: Natural Language Processing (NLP), Weighted Jaccard measure, text geolocation, Ukrainian-language natural text, artificial intelligence (AI), water body, Southern Bug River.

Вступ

Оскільки кількість водних масивів у межах річкового басейну, зазвичай, є значною (наприклад, для басейну р. Південний Буг їх 1091), якісний моніторинг кожного з них є складним завданням. Для вирішення цієї задачі у роботі [1] запропоновано використовувати україномовні тексти, які можна зібрати в першу чергу, із соцмереж (хоча, в загальному випадку, це можуть бути й матеріали звітів чи тексти зі сторінок сайтів установ екологічного профілю тощо). Після чого можна здійснити автоматизовану геоприв'язку текстів до територій масивів вод з подальшим узагальненням цих текстів по них, що описано у статті [2]. Зокрема, кожен масив вод (водозбірна площа розташованих на його території ділянок площинних і лінійних водних об'єктів) описується наборами ключових слів щодо розташованих на їх територіях географічних об'єктів. Ці набори ключових слів можуть бути результатом роботи алгоритмів NER (англ. Named Entity Recognition), просторового аналізу з використанням геоінформаційних систем, різноманітних автоматизованих алгоритмів Data Mining для оброблення природномовного тексту (NLP), які наприклад, наводились авторами у роботах [3, 4], та результатом роботи експертів. Потім аналізується збіг тексту новин чи звітів з цими ключовими словами і вибирається той масив вод, збіг з яким був найбільшим. Якщо збіг – майже рівнозначний, тоді робиться висновок про релевантність одночасно декільком масивам вод. Наприклад, новина про річку Соб дійсно може стосуватись усіх масивів вод, які розташовані в її басейні. Однак, практичні випробування запропонованих у статті [2] алгоритмів з метрикою cosine_similarity для новин, які описують локальні об'єкти, розташовані тільки на одному масиві вод, показали, що вони рідко

працюють ефективно. Варто знайти кращу метрику та спробувати використати потенціал сучасних інформаційних технологій на основі великих мовних моделей.

Метою даного дослідження є підвищення точності автоматизованої геоприв'язки тексту новин із соцмереж та ЗМІ про конкретні локальні об'єкти до ключових слів, що характеризують масиви вод заданої великої річки.

Алгоритм розв'язання задачі

Після проведення дослідження різноманітних метрик виявлено основну проблему – ключові слова нерівнозначні. Для того, щоб результат був максимально точним під час прив'язування текстових даних про об'єкти, необхідно провести класифікацію ключових слів для конкретної локації з вагою $w = 3$ (наявність пляжу, ресторану на березі водойми, ГЕС, греблі), регіону з вагою $w = 2$ (назва міста, району, наявність водосховища) та глобально з вагою $w = 1$ (відношення до національного парку, адміністративної області тощо). Алгоритм роботи полягає в наступному:

1. Генерація ключових слів для масивів вод.
2. Провести класифікацію та визначення ваги w згенерованих ключових слів за допомогою сучасних великих мовних моделей (LLM).
3. Отримання ключових слів з новин за рахунок LLM.
4. Застосування зваженої Jaccard-міри та обчислення ступеня збігу класифікованих ключових слів зі словами із новин:

$$S = \frac{\sum w_i \cdot |K_{\text{новина}} \cap K_{\text{масив}}|}{|K_{\text{новина}} \cup K_{\text{масив}}|},$$

де w_i - вага ключового слова, а K - множина ключових слів у новинах та в описі масивів вод.

5. Пошук масиву води з максимальним показником Jaccard-міри (або усіх зі значенням, більшим за 0).

Зразок роботи інтелектуального алгоритму

Для проведення дослідження обрано 10 масивів вод зі статті [2] для гіпотетичних новин з об'єктами у масивах вод, розташованих на річці Південний Буг (рис. 1).

Новина1. У Хмельницькому нечистоти потрапили в канал, з'єднаний з Південним Бугом: деталі. Витік нечистот з колектора каналізаційно-насосної станції трапився в мікрорайні Гречани, що в Хмельницькому. Стоки потрапили в канал, що сполучається з Південним Бугом. «Дуже велика площа була забруднена»

Новина2. Біля пляжу "Хімік" у Вінниці у Бугу була помічена велика пляма бруду, яка запливла в очерет біля пляжу і там осіла. Екоінспектори досліджують наслідки і можливі причини цього забруднення.

Новина3 Біля села Новоселівка на Ладизинському водосховищі велика кількість риби у серпні загинули і спливла. Екологи роблять аналізи води та намагаються якось боротись з наслідками цього вкрай негативного явища.

Рисунок 1 – Гіпотетичні новини про об'єкти у масивах вод, розташованих на річці Південний Буг

Jaccard-Міра Для 3 Новин

	code	news_1	news_2	news_3
1	UA_M5.4_0002	0.0	0.0	0.0
2	UA_M5.4_0003	7.6923076923076925	0.0	0.0
3	UA_M5.4_0004	0.0	0.0	0.0
4	UA_M5.4_0011	0.0	0.0	0.0
5	UA_M5.4_0013	0.0	3.0303030303030303	0.0
6	UA_M5.4_0152	0.0	0.0	0.0
7	UA_M5.4_0019	0.0	0.0	3.125
8	UA_M5.4_0507	0.0	0.0	0.0
9	UA_M5.4_0028	0.0	0.0	0.0
10	UA_M5.4_0969	0.0	0.0	0.0
11	UA_M5.4_0970	0.0	0.0	0.0

Рисунок 2 – Результат геоприв'язки новин до масивів вод за запропонованим алгоритмом з використанням моделі ChatGPT 4o

З рисунку 2 видно, що точність отриманих результатів (номери ділянок, які відповідають максимальному значенню міри) дорівнює 1.0. Це означає повну відповідність очікуванням.

Під час класифікації ключових слів за ступенем відношення до глобальних, регіональних, локальних доцільно враховувати різні типи ієрархічні структури, передусім, державні реєстри, наприклад класифікатор КОАТУУ (Класифікатор об'єктів адміністративно-територіального устрою України), довідник гідрографічної мережі (річки, їх притоки, приток їх приток тощо), а також наявність певних назв у реєстрах назв підприємств, назв національних парків тощо. Також, варто враховувати те, що ці дані можуть бути представлені у вигляді електронних таблиць, баз даних або шарів геоінформаційних систем.

Висновки

Запропоновано удосконалити запропонований раніше авторами метод автоматизованої геоприв'язки текстів новин із соцмереж та ЗМІ про конкретні локальні об'єкти до описів, що характеризують масиви вод заданої великої річки, за рахунок використання класифікації ключових слів за ступенем глобальності/локальності та їх врахування з різними вагами для оцінювання ступеню збігу зваженої Jaccard-міри замість метрики cosine_similarity, що дозволило підвищити точність цієї геоприв'язки.

Наведено приклад, який довів, що розроблений алгоритм дозволяє отримати максимальну точність 1.0 і є досконалішим рішенням, ніж метрика cosine_similarity у статті [2].

Запропоновано підходи щодо класифікації глобальності/локальності ключових слів на основі державних та інших реєстрів у різних форматах, що прискорить цю класифікацію і зробить її більш точною. Все це, в цілому, дасть змогу підвищити ефективність моніторингу водних ресурсів та їх управління.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Бондалетов К. О., Мокін В. Б., Григорчук М. В., Джура С. В., Кищук М. О., Неруцький О. В., Невольа С. Д., Фурман А. М., Гіжєвський В. В. Побудова датасету для тренування інтелектуальних моделей веб-системи з інформацією про екологічні проблеми та заходи у масивах вод басейну р. Південний Буг WISEST-SBB // Матеріали ЛІІ Науково-технічної конференції факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного

університету, Вінниця, 21 – 23 червня 2023 р. – Електрон. текст. дані. – 2023. – Режим доступу: <https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2023/paper/view/18958/15723>

2. 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>.

3. 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>

4. Bondalietov, K., Mokin, V. (2023). Notation System for Comparing and Synthesis of Intelligent Key Phrase Extraction Methods for Ontological Models in Information Systems. In: Dovgyi, S., Trofymchuk, O., Ustimenko, V., Globa, L. (eds) Information and Communication Technologies and Sustainable Development. ICT&SD 2022. Lecture Notes in Networks and Systems, vol 809. Springer, Cham. https://doi.org/10.1007/978-3-031-46880-3_11.

5. Інформаційна інтелектуальна технологія автоматизованої геоприв'язки екологічної текстової природно-мовної інформації [Електронний ресурс] / В. Б. Мокін, М. А. Гораш, Є. М. Крижановський, Т. Є. Вуж // Наукові праці ВНТУ. – 2020. – № 4. – Режим доступу: <https://praci.vntu.edu.ua/index.php/praci/article/view/624/585>.

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

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

Bondalietov Kostiantyn O. — Post-graduate student of the Chair of System Analysis and Information Technology, e-mail: bondalietov.k@gmail.com;

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