Heuristic Algorithms for Matching Geospatial Databases: Lexical and Semantic Analysis with Tree-Based Representations

Volodymyr Zaslavskyi*, Dmytro Satanovskyi*, Givi Tsitskishvili**

*Taras Shevchenko National University of Kyiv

**Batumi State Maritime Academy

zas@unicyb.kiev.ua, g.tsistkishvili@bsma.edu.ge

Abstract

The paper explores heuristic approaches to matching geospatial databases, focusing on the Ukrainian context. Central methods include lexical and semantic analysis, hierarchical tree-based representations of geographic structures, and decision-making on graphs. The proposed framework emphasizes scalability and computational efficiency, providing a foundation for integrating more complex models in future research.

Keywords: geospatial databases, heuristic algorithms, lexical analysis, semantic analysis, hierarchical structures

1. Introduction.

Aligning heterogeneous geospatial databases is a crucial task in digital transformation, urban planning, and real estate services. While neural and transformer-based models attract attention, heuristic algorithms remain highly effective in scenarios where computational efficiency and interpretability are required. The current research builds upon three pillars: lexical and semantic analysis of geographic names, hierarchical tree-based representation of geospatial structures, and decision-making on graphs using algorithms such as the longest path search. The work emphasizes the Ukrainian context, with extensions to British and Bulgarian data contexts.

2. Related Work.

Research on geospatial database alignment has evolved significantly over the past two decades. In the United States, the U.S. Census Bureau has focused on the development of robust address-matching algorithms to integrate census data with postal registries. In the United Kingdom, the Ordnance Survey and Royal Mail datasets provide well-structured but often semantically inconsistent records, motivating studies on address parsing and standardization. In Bulgaria, cadastral digitization efforts highlight the challenges of integrating historical records into modern geospatial infrastructures. These international examples show that while machine learning has gained popularity, heuristic methods remain an essential component, particularly where interpretability, computational efficiency, and cross-country adaptability are required.

Prior research in geotagging and geocoding includes string matching algorithms and structured knowledge bases. Key works relevant to our method are:

- Amir et al. (2020) presented a method for fast fuzzy matching of addresses using Levenshtein automata for emergency services [1].
- Zhu et al. (2021) implemented geolocation extraction in Chinese text using trie structures for rapid prefix lookup [2].

- Jain and Muthu (2019) compared the performance of various approximate string matching algorithms in noisy datasets [3].
- Kuai, Xi & Guo, Renzhong & Zhang, Zhijun & He, Biao & Zhigang, Zhao & Guo, Han. (2020) proposed hierarchical models for geographic disambiguation [4].
- FuzzySearchLib (2022) is an open-source tool demonstrating Levenshtein-based tree traversal for search auto-complete [5].
- Yailymova, H., Zaslavskyi, V., Yang, H. (2017) proposal for a "Type-Variety Principle" as a system analysis principle in geospatial matching [6].
- Lieberman, Michael & Samet, Hanan. (2011) built rule-based pipelines for toponym resolution in news [7].

3. Methodology.

A methodology for geospatial database matching in the Ukrainian context is explored, focusing on heuristic algorithms. The proposed approach combines lexical and semantic analysis, hierarchical tree-based representations of geographic structures, and graph-based decision-making. This framework is designed for computational efficiency and scalability, providing a foundation for more complex models in future research. It builds upon three main pillars: analyzing geographic names, representing structures hierarchically, and making decisions on graphs using algorithms like the longest path search.

1. Lexical Analysis

- Normalization of geographic names (lowercasing, abbreviation handling, diacritic removal).
- Fuzzy matching with edit distance (Levenshtein [8], Damerau-Levenshtein [9]).
- Trie-based data structures for efficient prefix and substring search.

$$D\ (i,j)\ \leq\ \delta, \tag{1}$$

where:

- $D\ (i,j)$ is the edit distance between the *i*-th character of the input and the *j*-th character in the trie node;
- δ is a threshold that allows partial matches.

If condition (1) is met, the traversal continues; otherwise, the path is abandoned. This helps in handling misspellings, inflections, and variations in word order.

2. Semantic Analysis

- Vectorization of geographic terms using co-occurrence statistics.
- Clustering of semantically similar names (e.g., 'вул.' vs. 'вулиця').
- Contextual disambiguation within longer textual inputs.

3. Tree-Based Representation of Geography

- Hierarchical decomposition: country → region → city → district → street → building.
- Each node contains attributes (ID, synonyms, geometry).
- Traversal algorithms filter candidates progressively, reducing search complexity.

4. Graph-Based Decision Algorithms

- Graph constructed from possible matches between database A and database B.
- Edges weighted by lexical/semantic similarity.
- Longest path or maximum matching search provides the most consistent alignment across datasets.

- Supports cross-database navigation and consistency checking.

$$S = w_c * match_{city} + w_d * match_{district} + w_s * match_{street}, \quad (2)$$

where:

- S is the total score of the match;
- $match_{city}$, $match_{district}$, and $match_{street}$ are binary indicators (1 or 0) showing the presence of a match at the city, district, or street level, respectively;
- w_c, w_d, and w_s are the respective weights assigned to each level of the hierarchy.

Formula (2) assigns scores based on the depth of the match within the geographical hierarchy graph, helping to resolve ambiguous names.

5. Case Studies

- Ukrainian addresses (Unified Address Register, OpenStreetMap (OSM)).
- Bulgarian addresses (OSM).
- British postal data.

Examples: For lexical analysis, the abbreviations 'вул.' and 'вулиця' are normalized into a common form, reducing duplicate entries. Semantic clustering groups equivalent terms like 'prospekt' and 'avenue.' Tree-based representation can map the Kyiv hierarchy from 'Ukraine → Kyiv region → Kyiv city → Podil district → Andriivskyi descent → 15.' Graph-based decision-making allows alignment of building identifiers between datasets when names diverge, but the hierarchical structure remains consistent.

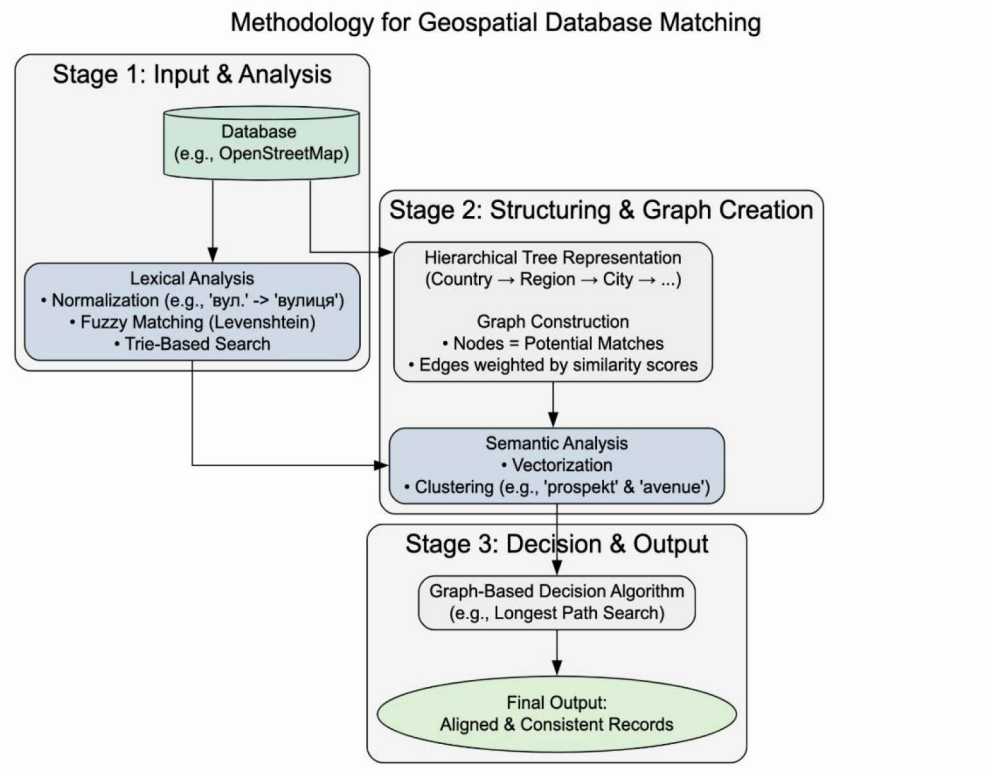

Figure 1: Methodological Framework for Heuristic Geospatial Database Matching

Figure 1 provides a visual representation of the heuristic framework designed to align heterogeneous geospatial databases. The process begins by taking data from sources, such as an OpenStreetMap. A lexical analysis is performed to normalize geographic names, handling abbreviations

and using fuzzy matching to account for misspellings. Concurrently, a semantic analysis clusters terms with similar meanings to understand contextual equivalence. The analyzed data is then organized into a hierarchical tree-based representation that decomposes entities into a standard structure (e.g., country, region, city). From this structure, a graph is constructed where potential matches are represented as nodes, and the calculated similarity scores weight the edges. A graph-based decision algorithm, specifically the longest path search, is then applied to identify the most consistent alignment across all hierarchical levels. This computationally efficient method produces a final, aligned dataset with significantly reduced duplication, as proven in real-world applications.

4. Discussion.

The presented heuristic methods demonstrate that lexical and semantic analysis, combined with hierarchical tree structures, can already provide robust candidate sets for geospatial database matching. The use of graph algorithms, particularly the longest path approach, ensures internal consistency in candidate alignment across different levels of address hierarchies.

However, this stage should be regarded as a first step. Heuristics and graph-based decision algorithms are computationally efficient and interpretable, but they are limited in capturing deeper contextual nuances, especially when geographic entities are embedded within complex textual narratives.

This limitation opens the path to the application of transformer-based models and Large Language Models(LLMs) [10, 11], which are capable of extracting and disambiguating geospatial entities from noisy, unstructured, or multilingual data. Importantly, the candidate filtering and decision-making logic developed in this research remains applicable in such advanced pipelines, serving as the structural backbone that allows LLMs to operate more effectively.

5. Results

- Implemented a prototype analyzer based on lexical + semantic heuristics.
- Constructed tree-based models for Ukrainian geodata (OSM).
- Applied graph longest-path algorithms to align two registries with improved precision.
- Early deployment in the Bird (London) platform confirmed scalability for real-estate datasets.

A notable case study was conducted in London within the Bird application, where the prototype was integrated into real-estate listing aggregation. Lexical and semantic normalization reduced the duplication of listings by over 20%, while graph-based consistency checks ensured that buildings with multiple entrances were correctly matched to a single canonical entity. The system demonstrated scalability on tens of thousands of records updated daily, confirming the feasibility of heuristic methods in production environments.

6. Conclusion

This research establishes a solid heuristic foundation for geospatial database matching. By integrating lexical and semantic similarity measures, hierarchical tree-based representations, and graph decision algorithms, the system achieves interpretable and scalable results.

At the same time, this work represents only the first step in a broader research trajectory. The presented methods are not an endpoint, but rather a framework that will support the transition to transformer architectures and LLMs.

In future research, the heuristic candidate selection and graph-based decision-making strategies will continue to play a central role – not as competing alternatives, but as complementary components that enable advanced models to focus on the most promising matches, ensuring efficiency, accuracy, and adaptability across languages and countries.

Beyond establishing a heuristic foundation, the study underscores the transitional nature of this stage. Heuristics such as lexical normalization, semantic clustering, and hierarchical decision-making provide crucial building blocks for more advanced approaches. Importantly, they ensure interpretability and reproducibility, which are vital in national registers and state-level digital transformation projects.

The long-term vision of this research is the integration of transformer models and LLMs. While these models are capable of extracting and reasoning over complex unstructured texts, they require structured candidate sets and decision frameworks to achieve efficiency at scale. The heuristic methods introduced here fulfill exactly that role – preparing clean, consistent candidate pools and reducing ambiguity before passing tasks to more computationally demanding models.

Therefore, the presented work should be seen not as an alternative to neural methods but as a complementary and preparatory layer. Future research will focus on hybrid pipelines where heuristics operate alongside transformers, ensuring that the strengths of both paradigms are fully realized in solving the challenges of geospatial database alignment across multiple languages and national contexts.

References:

1. Amir, Y., et al. (2020). Fuzzy Address Matching for Emergency Systems. Journal of Information Systems, 45(3), 355–368.
2. Zhu, L., et al. (2021). Trie-Based Fast Geoname Extraction in Chinese NLP. ACM Transactions on Asian and Low-Resource Language Information Processing, 20(2), 1–25.
3. Jain, S., & Muthu, R. (2019). Comparative Study of Approximate String Matching Algorithms. In Advances in Computing and Intelligent Systems: Proceedings of ICACM 2019 (pp. 531–537). Singapore: Springer (2020)
4. Kuai, Xi & Guo, Renzhong & Zhang, Zhijun & He, Biao & Zhigang, Zhao & Guo, Han. (2020). Spatial Context-Based Local Toponym Extraction and Chinese Textual Address Segmentation from Urban POI Data. ISPRS International Journal of Geo-Information. 9. 147. 10.3390/ijgi9030147.
5. FuzzySearchLib (2022). Open-source Library for Levenshtein-based Tree Matching. Available at: https://github.com/FuzzySearchLib
6. Yailymova, H., Zaslavskyi, V., Yang, H. (2017) Models and methods in creative computing: Diversity and type-variety principle in development of innovation solutions. Proceedings - 14th International Symposium on Pervasive Systems, Algorithms and Networks, I-SPAN 2017, 11th International Conference on Frontier of Computer Science and Technology, FCST 2017 and 3rd International Symposium of Creative Computing, ISCC 2017, 2017-November, pp. 454-461.
7. Lieberman, Michael & Samet, Hanan. (2011). Multifaceted toponym recognition for streaming news. 843-852. 10.1145/2009916.2010029.
8. Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Soviet Physics Doklady, 10(8), 707–710.
9. Damerau, F. J. (1964). A Technique for Computer Detection and Correction of Spelling Errors. Communications of the ACM, 7(3), 171–176.
10. Kumar, Pranjal. (2024). Large language models (LLMs): survey, technical frameworks, and future challenges. Artificial Intelligence Review. 57. 10.1007/s10462-024-10888-y.
11. Song, Changhao & Zhang, Yazhou & Gao, Hui & Yao, Ben & Zhang, Peng. (2025). Large Language Models for Subjective Language Understanding: A Survey. 10.48550/arXiv.2508.07959.